VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA PODNIKATELSKÁ

FACULTY OF BUSINESS AND MANAGEMENT

ÚSTAV INFORMATIKY

INSTITUTE OF INFORMATICS

VYUŽITÍ STROJOVÉHO UČENÍ PRO PREDIKCI ODCHODU ZÁKAZNÍKA

MACHINE LEARNING IN CUSTOMER CHURN PREDICTION

DISERTAČNÍ PRÁCE
DOCTORAL THESIS

AUTOR PRÁCE — **Ing. Martin Fridrich, MSc**
AUTHOR

VEDOUCÍ PRÁCE — **prof. Ing. Petr Dostál, CSc.**
ADVISOR

BRNO 2023

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA PODNIKATELSKÁ

Zadání dizertační práce

Ústav:	Ústav informatiky
Student:	**Ing. Martin Fridrich, MSc**
Vedoucí práce:	**prof. Ing. Petr Dostál, CSc.**
Akademický rok:	2022/23
Studijní program:	Ekonomika a management
Studijní obor:	Řízení a ekonomika podniku

Využití strojového učení pro predikci odchodu zákazníka

Charakteristika problematiky úkolu:

Úvod
Teoretická východiska
Literární rešerše
Cíle práce a užité metody
Návrh a implementace řešení
Dosažené výsledky
Shrnutí a diskuse
Přínosy práce
Závěr
Literární zdroje

Cíle, kterých má být dosaženo:

Hlavním cílem disertační práce je návrh, implementace a zhodnocení systému strojového učení, který bude předpovídat odchod zákazníka v prostředí elektronického maloobchodu. Představené řešení by mělo reflektovat potřeby retenčního managementu, kam autor řadí především odhad ekonomického dopadu retenční kampaně, a bližší porozumění modelovanému jevu.
Dílčí cíle disertační práce:
Dílčí cíl 1: Popsat teoretická východiska, zahrnující prostředí elektronického maloobchodu, problematiku řízení vztahů se zákazníky, a strojové učení.
Dílčí cíl 2: Zanalyzovat současné poznatky v oblasti predikce ztráty zákazníka s využitím metod výpočetní lingvistiky i tradiční rešerše.
Dílčí cíl 3: Navrhnout a vytvořit systém strojového učení, zaměřený na předpověď odchodu zákazníka v prostředí elektronického maloobchodu v intencích vymezených hlavním cílem práce.
Dílčí cíl 4: Zhodnotit schopnosti navrženého systému strojového učení, včetně interpretace zachycených znalostí.

Základní literární prameny:

Ascarza, E., Neslin, S. A., Netzer, O., Anderson, Z., Fader, P. S., Gupta, S., Hardie, B. G. S., Lemmens, A., Libai, B., Neal, D., Provost, F., & Schrift, R. (2018). In Pursuit of Enhanced Customer Retention Management: Review, Key Issues, and Future Directions. Customer Needs and Solutions, 5(1-2), 65-81.

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd). Springer.

Lundberg, S., & Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems. Curran Associates.

Tamaddoni Jahromi, A., Stakhovych, S., & Ewing, M. (2014). Managing B2B customer churn, retention and profitability. Industrial Marketing Management, 43(7), 1258-1268.

Termín odevzdání dizertační práce je stanoven časovým plánem akademického roku 2022/23.

V Brně, dne 12. 6. 2020

L. S.

prof. Ing. et Ing. Stanislav Škapa, Ph.D. předseda oborové rady	doc. Ing. Vojtěch Bartoš, Ph.D. děkan

Abstrakt

Disertační práce se zaměřuje na predikci odchodu zákazníků v prostředí elektronického maloobchodu. Text představuje současný stav vědeckého bádání, analyzuje klíčové trendy a identifikuje příležitosti pro další výzkum. Literární rešerše je dílem realizována prostřednictvím metod pro zpracování přirozeného jazyka. Cílem práce je navrhnout, implementovat a zhodnotit systém strojového učení pro predikci odchodu zákazníků v elektronickém maloobchodě, který reflektuje perspektivy ekonomického dopadu navazujících retenčních aktivit a umožňuje bližší porozumění modelovanému jevu.

Vlastní řešení je strukturováno do částí vymezení problému, porozumění a zpracování dat, modelování, vyhodnocení, interpretace a produkční nasazení systému. Nad rámec klasického pojetí odchodu zákazníka, jako absence transakce v budoucím období, je představeno nové pojetí inkrementálního ekonomického dopadu retenční kampaně. Přístupy jsou ověřeny na dvou datových souborech. V rámci modelování je uvažováno o GLM, SVM, ANN, rozhodovacích stromech a meta-algoritmech. Vnější parametry vlastního zpracování dat a konstrukce modelu jsou odhadnuty s pomocí Bayesovské optimalizace. Porozumění modelovaným jevům je podpořeno s pomocí SHAP nástrojů, které jsou rozšířeny v oblastech odhadu a vizuální prezentace.

Z pohledu přirozených ukazatelů prediktivních schopností vyčnívají řešení využívající náhodné lesy nebo gradient boosting, v klasickém pojetí vynikají i ANN. Z hlediska ekonomického výsledku retenční aktivity vyčnívá nové pojetí úlohy, pozoruhodné jsou především systémy postavené na rozhodovacích stromech nebo meta-algoritmech. Jako klíčové nezávislé proměnné se podařilo identifikovat reprezentace stáří a frekvenci interakcí a transakcí, v novém pojetí vyčnívá i hodnota zákazníka. Určení a porozumění zákaznickým shlukům, na které je vhodné cílit, pak přímo podporuje související retenční aktivity.

Disertační práce tak představuje ucelený přehled nových přístupů a nástrojů pro predikci odchodu zákazníka, využitelných jak pro další výzkum, tak v podnikové nebo pedagogické praxi.

Klíčová slova

predikce odchodu zákazníka, elektronický maloobchod, řízení vztahů se zákazníky, retenční řízení, strojové učení

Abstract

The dissertation examines customer churn prediction in e-commerce retail settings, presenting the current research landscape, analyzing key trends, and pinpointing opportunities for further investigation. The literature review is conducted using language processing. The study aims to develop, implement, and evaluate a machine learning system for predicting customer churn in the e-commerce environment, considering the economic implications of retention efforts, and facilitating a deeper understanding of the modeled phenomenon.

The solution is organized into sections covering problem definition, data comprehension and processing, model development, evaluation, interpretation, and deployment. The author extends the traditional concept of customer churn as the lack of a transaction in a future period with a novel idea of the incremental economic impact of a retention campaign. The notions are validated using two datasets. The modeling framework incorporates GLM, SVM, ANN, decision trees, and meta-algorithms. Bayesian optimization estimates external parameters related to data processing and model building. The understanding of the phenomena is enhanced using SHAP tools, which are improved in terms of computation and visual representation.

From the perspective of natural prediction performance, random forests and gradient boosting dominate; in the original task, ANN also performs well. When considering the financial results of the retention campaign, the novel approach functions excellently, mainly when coupled with decision trees or meta-learning. Recency and frequency representations of interactions and transactions are identified as key features; the feature importance of customer value emerges in the novel approach. Identifying and comprehending customer segments to target directly supports subsequent retention initiatives.

In summary, the thesis offers an extensive overview of novel methods and tools for predicting customer churn, which can be valuable for future research and practical applications in business or educational settings.

Keywords

customer churn prediction, e-commerce retail, customer relationship management, retention management, machine learning

Bibliografická citace

Fridrich, M. (2023). *Využití strojového učení pro predikci odchodu zákazníka* [Disertační práce]. Vysoké učení technické v Brně, Fakulta podnikatelská, Ústav informatiky. Vedoucí práce Petr Dostál.

Prohlášení autora o původnosti díla

Jméno a příjmení autora:	Ing. Martin Fridrich. MSc
VUT ID studenta:	101005
Typ práce:	Disertační práce
Akademický rok:	2022/2023
Téma závěrečné práce:	Využití strojového učení pro predikci odchodu zákazníka

Prohlašuji, že svou závěrečnou práci jsem vypracoval samostatně pod vedením vedoucího závěrečné práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené práce dále prohlašuji, že v souvislosti s vytvořením této závěrečné práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

V Praze, dne 29. srpna 2023 _____

podpis autora

Poděkování

Na tomto místě bych rád poděkoval všem, kteří mi pomohli na cestě k dokončení disertační práce. Mé díky náleží především školiteli, prof. Ing. Petru Dostálovi, CSc., jehož neúnavná podpora a odborné vedení byly nedocenitelné. Rovněž bych rád poděkoval Ing. Karlu Doubravskému, Ph.D., MSc, za připomínky ke zpracování této práce.

Dále bych rád poděkoval doc. RNDr. Bedřichu Půžovi, CSc., Mgr. Veronice Novotné, Ph.D., a dalším kolegům z Fakulty podnikatelské Vysokého učení technického v Brně, především za podnětné diskuse a zajištění podmínek pro realizaci doktorského studia.

V neposlední řadě bych chtěl vyjádřit vděk svým nejbližším, za podporu, trpělivost a milá rozptýlení.

Obsah

Úvod

Během posledních desetiletí je možné pozorovat nebývalý příklon podniků k vnímání individuálního zákazníka jako středobodu aktivit, což firmám umožňuje pružně reagovat na změny v zákaznických požadavcích a tržních podmínkách při udržení ziskových vztahů. Mezi katalyzátory naznačeného posunu řadíme vysoce konkurenční prostředí a postupující technologické inovace. Kumar & Reinartz (2018) považují zákaznickou orientaci za nezbytnou pro růst nabízené subjektivní hodnoty a tím i pro hospodářský výsledek společnosti.

Část aktivit směřovaných k prevenci odchodu a udržení stávajících zákazníků bývá označována jako retenční management. Gronwald (2017) akcentuje význam činností s ohledem na značný rozdíl v efektivitě prostředků vynaložených na získání, respektive udržení zákazníka. Vztah mezi retenčními schopnostmi a úspěchem firmy dovozují Gupta et al. (2004), Kumar et al. (2018), Umashanjar et al (2017) a další. Není tak překvapením, že podpora retenčních aktivit bývá jednou z podnikových priorit. Daunis & Iwan (2014) poukazují na nespokojenost vrcholového managementu se schopností tuto prioritu naplňovat. Handley (2013) upozorňuje na skutečnost, že i zákazníci jsou z úrovně retenčních snah rozmrzelí. Rozpor mezi důležitostí podnikových aktivit směřujícího k udržení zákaznických vztahů, a vnímanou úrovní realizace ilustruje relevanci a aktuálnost řešeného tématu.

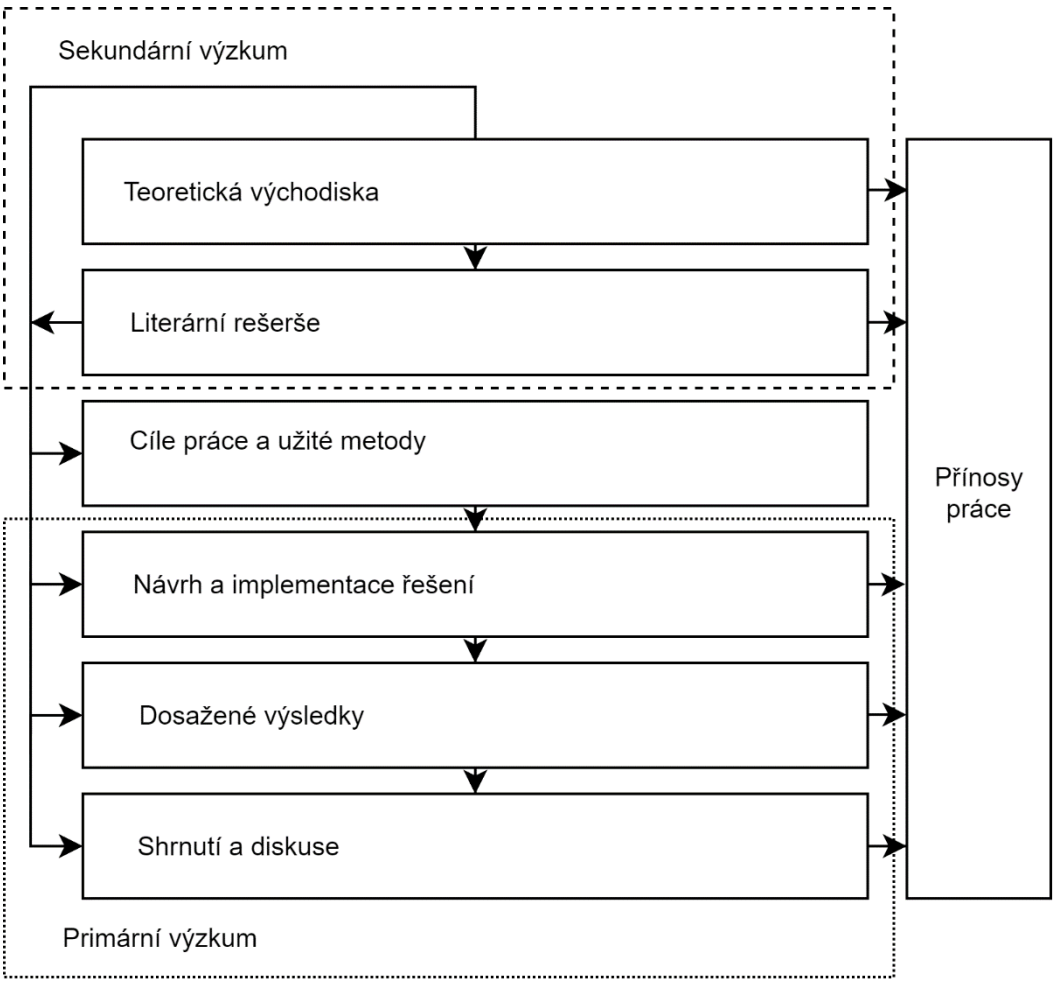
Úspěch retenčního úsilí vychází ze schopnosti předvídat jací zákazníci se chystají vzájemný vztah přerušit, a jejich úmyslu předcházet prostřednictvím individuální pobídky nebo jiné intervence. Prvotní úlohou je predikce odchodu zákazníka. S ohledem na rozsah datových souborů popisujících interakce zákazníka, podniku, a ostatních relevantních entit, bývá k problému přistupováno s využitím metod strojového učení, jenž umožňuje exploataci komplexních struktur chování, které se v datech ukrývají. Neslin et al. (2006) varují před nedostatečnými predikčními schopnostmi značné části přístupů. Odpovědí se zdá být využití velkých dat, nebo nových přístupů strojového učení. Ascarza et al. (2018) však poukazují na některé opomíjené aspekty návrhu retenční kampaně jako jsou porozumění zákaznickému chování, výběr cílové skupiny, časový rámec a způsob intervence, případně vyhodnocení realizované kampaně.

Předchozí odstavce představují posun firem k zákaznické orientaci, význam retenčního managementu a úlohy predikce ztráty zákazníka. Za společný jmenovatel změn považuje autor překotný vývoj v oblastech komunikačních a informačních technologiích. Disertační práce se

tak přirozeně soustředí na řešení dané úlohy v odvětví elektronického maloobchodu, jehož existence je v zásadě jedním z důsledků nastíněných změn (Chaffey, 2015).

Struktura disertační práce

Disertační práce je tematicky členěna do kapitol, vzájemné vazby a povaha výzkumného úsilí jsou obsahem Obr. 1. Autor v následujících odstavcích stručně uvádí, co je náplní příslušných sekcí.



Obr. 1 Koncepce disertační práce

Teoretická východiska jsou jedním ze základních stavebních kamenů práce. Kapitola pokrývá oblasti elektronického obchodování, správu zákaznických vztahů a strojové učení. Účelem je vymezit některé zásadní pojmy a východiska, poskytnout podklady a motivaci pro další výzkum. Autor zde nepředstavuje úplný přehled literatury, ale připravuje půdu pro navazující sekce disertační práce.

Literární rešerše naopak představuje rozsáhlý vhled do vědecké domény prostřednictvím dvou větví, kde první větev předkládá obsahovou analýzu vědeckých článků zabývajících se predikcí odchodu zákazníka, prostřednictvím metod zpracování přirozeného jazyka. Druhou větev zaměřuje autor na podmnožinu prací relevantních pro elektronické obchodování, které analyzuje tradičním způsobem, což vede k bližšímu porozumění výzkumných problémů, dat, metod aj. Kontrastování obou větví umožňuje popsat některá hlavní témata, trendy, ale i příležitosti dalšího výzkumu. Sekce je hlavním podkladem pro formulaci cílů a výzkumných otázek disertační práce, informuje také vlastní postup autora.

Cíle práce a užité metody slouží k vymezení směřování primárního výzkumu disertační práce prostřednictvím cílů a výzkumných otázek, které vychází z dříve popsaných slepých skvrn. Kapitola rozřazuje a rozšiřuje paletu nástrojů prezentovaných v předchozích kapitolách o metody nezbytné k adresování některých dalších aspektů představeného výzkumu. První tří kapitoly poskytují čtenáři jasnou představu o rozsahu a směřování výzkumu, včetně použitých metod.

Návrh a implementaci řešení zaměřujeme na vlastní přístup ke návrhu a konstrukci systému strojového učení, který pokrývá cíle a vědecké otázky vymezené v předchozí kapitole. Při strukturování textu vychází autor z referenčního metodického rámce pro organizaci projektu dobývání znalostí a strojového učení (Chapman et al., 2000), tj. věnuje se vymezení úlohy, porozumění a zpracování dostupných datových souborů, výběru algoritmů, přístupu k vyhodnocení a ověření schopnosti řešení, a v neposlední řadě také některými praktickým aspektům produkčního nasazení.

Dosažené výsledky prezentují detailní zhodnocení a porovnání jednotlivých přístupů strojového učení v intencích dostupných datových sad. Soustředíme se nejen na přirozené ukazatele úspěšnosti, ale i očekávaný ekonomický dopad zamýšlených retenčních aktivit. Znalosti reflektované úspěšnými systémy jsou interpretovány s ohledem na význam a charakter vztahu mezi závislými a nezávislými proměnnými, případně s pomocí zákaznických skupin vykazujících podobnou afinitu k závislé proměnné.

Shrnutí a diskuse obsahují ucelený souhrn postupu realizace výzkumu, na který je navázáno výzkumnými otázkami, v jejichž rámci se autor zabývá organizací dosažených výstupů, jejich zasazení do kontextu retenčního řízení, ale i relevantní vědecké literatury. Dále autor kriticky hodnotí limity realizovaného úsilí a poukazuje na možné náměty výzkumu budoucího.

Přínosy práce jsou závěrečnou kapitolou, představující zamyšlení nad pozitivní dopady disertační práce do oblastí vědy a výzkumu, podnikatelské, ale i pedagogické praxe. Autor zde staví na významu a relevanci zkoumaného tématu, současně nastiňuje možnosti dalšího využití prezentovaných závěrů.

Jazyková a citační konvence

Disertační práce je psána v českém jazyce. U artefaktů, které autor zpracoval svépomocí, není na další literární prameny odkazováno. Materiály vytvořené výhradně pro tuto práci, u kterých autor neuvažuje o dalším užití, jsou anotovány v českém jazyce. V disertační práci však lze nalézt původní schémata, obrázky nebo tabulky s popisky v anglickém jazyce; zde se jedná o podkladové materiály pro navazující publikační činnost autora. Perspektiva dalších publikací stojí i za rozhodnutím využívat anglické číselné konvence a citační normy APA.

1 Teoretická východiska

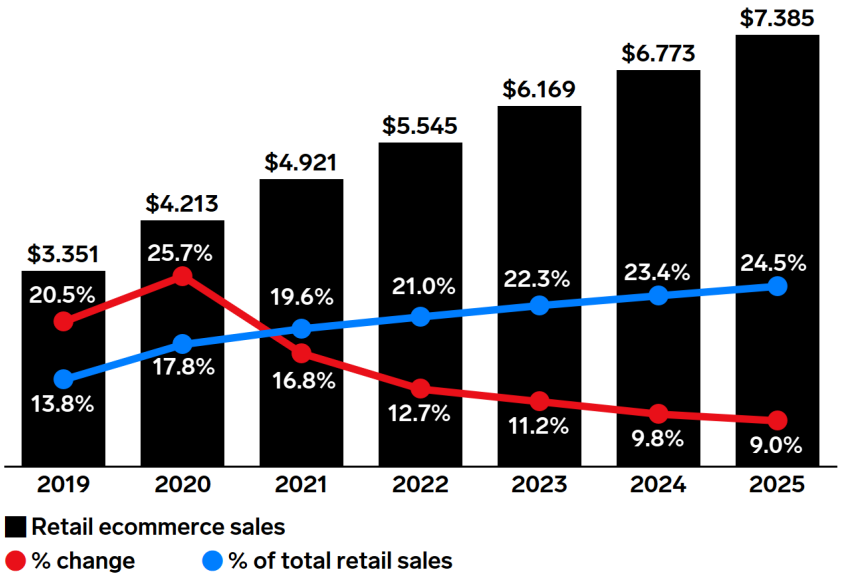
1.1 E-commerce retail

Termín e-commerce je často pojímán jako prodej a nákup realizovaný skrz internetové připojení, ve své široké podobě ovšem pokrývá všechny elektronicky realizované výměny informací mezi organizací a třetí stranou (Chaffey, 2015). Podobně vymezuje e-commerce i britská vláda, která pojmem rozumí elektronickou výměnu informací v rámci dodavatelského řetězce, uvnitř i mimo podniky, mezi podnikem a spotřebitelem nebo mezi entitami veřejného a soukromého sektoru, bez ohledu, zda se jedná o transakce finanční, či nikoliv (Cabinet Office, 1999). Šíře vybraných definic umožňuje pozorovat aspekty e-commerce v každé organizaci, které využívá moderních komunikačních technologií. Faktory, které motivují společnosti k adopci takových technologií dělí Perrott (2005) na ekonomické (zvýšení tržeb, snížení marketingových nákladů, snížení nákladů na řízení dodavatelsko-odběratelského řetězce), konkurenční (udržení trhu před konkurenty, kteří již využívají e-commerce), tržní výhody („first-mover advantage“) a přidanou hodnotu (zlepšení zákaznické spokojenosti a budování dlouhodobých vztahů se zákazníky). Elektronický maloobchod je potom podmnožinou elektronicky realizovaných transakcí, kde dochází k prodeji zboží a služeb. Transakce v tomto případě označuje objednávku zadanou konečným spotřebitelem, finanční část transakce nemusí být zajištěna elektronicky. Transakce je možné odlišit dle zainteresovaných stran na zákaznické a podnikové, možných kombinací ale existuje víc (Chaffey, 2015).

Celosvětový maloobchod postupně snižuje meziroční tempo růstu tržeb, s celkovými tržbami za rok 2020 v objemu 23.7 biliónu USD. Na vině jsou nejistota a zhoršující se ekonomické podmínky. Navzdory celkovému zpomalení pozorujeme meziroční růst tržeb elektronického maloobchodu na úrovni 25,7 %, s celkovými tržbami za rok 2020 v objemu 4.2 bilionu USD. E-commerce tak pokračuje v kanibalizaci tržeb tradičního maloobchodu; analytici společnosti Insider Intelligence předpokládají, že v roce 2022 bude až více než pětina maloobchodních tržeb realizována online kanály. Nejrychleji rostoucím regionem elektronického obchodování zůstává Latinská Amerika, následovaná Střední a východní Evropou. Trhy Spojených států, Asie a Evropské unie naopak rostou nejpomaleji (Insider Intelligence, 2020).

Vzestup elektronického obchodování je označován za jeden z faktorů zániku mnoha tradičních maloobchodních řetězců. Ve Spojených státech je tento jev v médiích často popisován termíny „retail apocalypse“ nebo „Amazon effect“. Mezi známé americké řetězce, které

v posledních letech procházely insolvenčním řízení patří Sears Holding, Borders Books nebo Toys R Us (Ovide, 2011; Geeter, 2018; Valinsky, 2019).



Obr. 2 Celosvětové tržby odvětví e-commerce retail, 2019-2025

Zdroj: Insider Intelligence (2020)

Největší maloobchodní společností současnosti je e-commerce gigant Amazon.com Inc., s tržní kapitalizací 1.515 bilionů USD (Investopedia, 2022). Mezi lety 2019 a 2020 došlo k nadpolovičnímu růstu čistého zisku (YOY) na 22,9 miliard USD, provozní zisk rostl ještě rychleji a dosáhl 21,3 miliard USD v roce 2020. Tržby potom zaznamenaly meziroční růst o 37 %, na 386,1 miliard USD v roce 2020. Páteřní aktivitou společnosti zůstává retail s 88 % tržeb, doplněk potom náleží cloudovým službám Amazon Web Services. Technologický segment je odpovědný za více než polovinu provozního zisku organizace, AWS patří mezi významné poskytovatele cloudových služeb (Amazon Inc, 2020). Právě zaměření na rozvoj technologií, strojové učení a zákaznickou zkušenost/spokojenost bývá považováno za pilíř úspěchu společnosti (Mackenzie et al., 2013; Morgan, 2018; Terdiman, 2018).

1.2 Řízení vztahů se zákazníky

Porozumění marketingu a řízení vztahů se zákazníky (customer relationship management – CRM) se díky technologickému pokroku posunulo od Petera Druckera a jeho marketingu jako „vnímání podniku očima zákazníka“, k moderní „zákaznické koncepci, vymezené realizací všech marketingových aktivit s přesvědčením, že středobodem jakékoliv analýzy nebo akce je

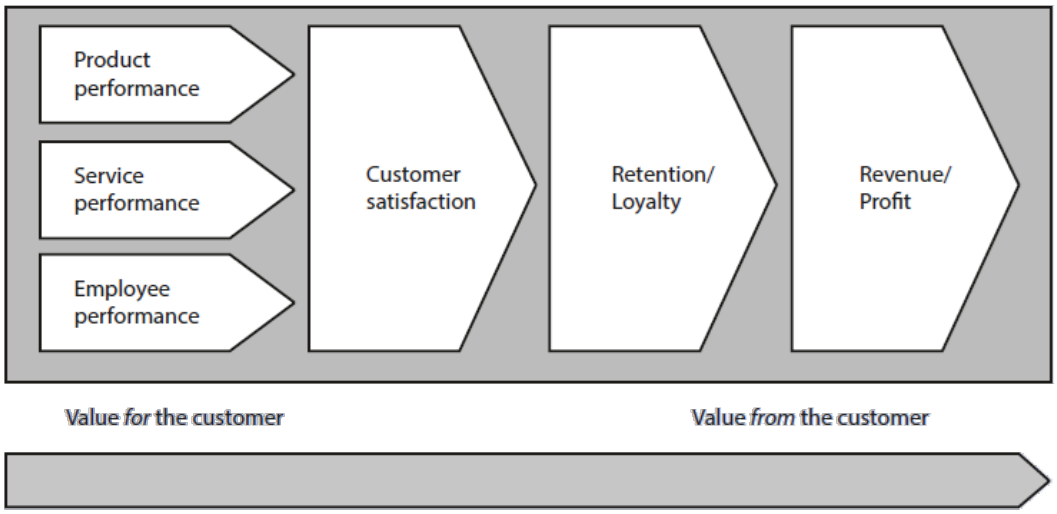
individuální zákazník“ (Kumar & Reinartz, 2018). Takové chápaní marketingu umožňuje formování vztahů mezi jednotlivcem a společností napříč prodejními kanály, včetně sociálních sítí. Kumar & Reinartz (2018) dále vymezují CRM jako „strategický proces výběru zákazníků, s kterými společnost dokáže interagovat, současně dokáže tyto zákazníky obsloužit při dosažení maximální ziskovosti. Konečným cílem podniku je optimalizace současné a budoucí hodnoty zákaznické báze“. Buttle & Maklan (2019) řadí uvedený přístup k CRM po bok tradičních podnikových orientací. Podniky se zákaznickým/tržním zaměřením potom líčí jako společnosti, které sdílí přesvědčení, že zákazník má být středobodem snažení. Takové podniky reagují na změny v zákaznických požadavcích a tržních podmínkách tak, aby dokázaly zákazníkům nabídnout co možná nejvyšší přidanou hodnotu a současně utvářely profitabilní vztahy. Doplňující perspektivy CRM lze popsat jako provozní, která se soustředí na integraci a automatizaci procesů jako jsou prodej, marketing nebo zákaznický servis, a analytickou, jenž se zabývá transformací zákaznických dat do poznatků využitelných napříč marketingovými aktivitami.

Závažnost řízení vztahů se zákazníky spatřují Kumar & Reinartz (2018) především ve schopnosti adaptace na postupné, ale zásadní změny v oblastech spotřebitele, trhu a marketingových funkcí. V rámci demografického vývoje dochází ke stárnutí populace rozvinutých zemí (European Comission, 2020), nárůstu etnické diverzity a individualismu. Z hlediska spotřebitelského chování upozorňují autoři na využívání sociálních platforem a prevalenci mobilních zařízení (Chaffey, 2015), s čímž úzce souvisí očekávání okamžité interakce nebo rozmach samoobslužných řešení. Mezi další trendy lze řadit potřebu autentičnosti nebo zájem o zdraví a udržitelnost (Solomon, 2015). Tržní změny zahrnují postupující globalizaci, v jejímž důsledku mizí hranice místních trhů a stoupají požadavky na logistické systémy a distribuční partnerství. V rozvinutých zemích dochází k převisu nabídky nad poptávkou, a tříštění poptávky dle individuálních potřeb spotřebitele. Sbližování kvality výrobků vede k obtížné diferenciaci. Společenské a technologické změny se promítají i v očekávání plynoucích z marketingových funkcí. Dochází k rozostření komunikačních kanálů, tradiční média jsou nahrazována médii novými. Lze také pozorovat pokles účinnosti marketingových výdajů (Teixeira, 2014). Uvedené trendy ilustrují nutnost soustředit se na zákaznické preference, nabízenou hodnotu a přizpůsobení výrobků a služeb.

1.2.1 Koncept zákaznické hodnoty

Tvorba hodnoty je premisou existence podniku, pouze společnost nabízející dobré výrobky a služby upoutá pozornost zákazníků. Na vazbu mezi úspěchem a misí podniku, která se

orientuje na tvorbu hodnoty pro zákazníka, případně akcionáře poukazují Kumar & Reinartz (2016). Klasická ekonomická teorie předpokládá maximalizaci užitku spotřebitele díky výběru souboru produktů a služeb přinášejících co možná nejvyšší subjektivní hodnotu. Růst této hodnoty je možný především díky strategickému přístupu ke správě zákaznických vztahů a moderním technologiím, uvádí Kumar & Reinartz (2018).



Obr. 3 Souvislost mezi zákaznickou spokojeností, loajalitou a ekonomickými výsledky organizace

Zdroj: Kumar & Reinartz (2018), Anderson & Mittal (2000)

Vzájemné vztahy mezi zákaznickou spokojeností, loajalitou a ekonomickými výsledky společnosti ilustruje Obr. 3. Koncepce vychází z předpokladu, že zdokonalení výrobků a služeb vede k vyšší spokojenosti, tj. že zvýšení zákazníkem vnímané hodnoty vede k vyšší retenci, která dále vede k růstu loajality, jenž společnost přetaví v nárůst vlastní ziskovosti. Kumar & Reinartz (2018) upozorňují na neprůkazné výsledky empirických studií, které vztahy mezi komponenty v minulosti zkoumaly. Komplexita prvků, síla, asymetrie a nelinearita vztahů vedou často k neefektivní alokaci marketingových výdajů. Autoři proto doporučují využít hlediska ekonomických výsledků společnosti, kterou lze případně doplnit dalšími perspektivami.

Za jednu ze zásadních myšlenek CRM označují Buttle & Maklan (2019) vnímání zákazníka jako souvislý tok ekonomických příjmů, nikoliv skrz izolované transakce; perspektiva implicitně akcentuje význam kontinuity takového vztahu. U individuálního zákazníka označujeme tento tok jako celoživotní hodnotu zákazníka („customer lifetime value“), u zákaznické báze hovoříme o součtu těchto hodnot, tzv. zákaznickém kapitálu („customer equity“). Koncept

celoživotní hodnoty zákazníka je v odborné literatuře pevně ukotven jako čistá současná hodnota zisků, přijatých od individuálních zákazníků nebo kohorty zákazníků, během doby trvání vzájemného vztahu (viz Gupta et al., 2006; Fader, 2012; Chaffey, 2015; Kumar & Reinartz, 2018; Buttle & Maklan, 2019). Vazbu mezi ekonomickými výsledky podniku a CLV prokazují práce Reichheld & Sasser (1990), Umashanjar et al (2017), Gupta et al. (2004), Kumar et al. (2018) a další. Příčiny vazby spatřují Buttle & Maklan (2019) v postupném navyšování tržeb, poklesu nákladů, snížení citlivosti na cenu a sílu doporučení.

Mezi další perspektivy zákaznické hodnoty řadí Kumar et al. (2010) vliv, jehož nejběžnější formou je osobní komunikace („word of mouth“). V širším pojetí jde o sdílení informací, znalostí a pomoci mezi existujícími i potenciálními zákazníky. Pro příjemce takové komunikace roste subjektivní hodnota nabídky díky lepšímu porozumění atributům, snížení rizika, nebo poklesu nákladů na transakci a adopci. Pozitivní dopady zákaznického vlivu lze pozorovat v růstu konverze, pokračujících interakcí nebo délce zákaznického životního cyklu. Moderní technologie umožňují společnostem detekovat témata nebo sentiment obsažený v komentářích hodnotící kvalitu produktu na elektronickém tržišti nebo sociálních platformách, výzvou ovšem zůstává mapování interakcí napříč dostupnými komunikačními kanály. Kumar et al (2010) upozorňují na specifickou formu zákaznického vlivu, tzv. systém referencí, ve kterém podniky odměňují doporučení vedoucí k uskutečnění transakce. Mezi další formy hodnoty uvádí znalost zákazníka, která vychází z informací a dat, poskytovaných společnosti ve formě hodnocení, stížností, doporučení nebo participací při tvorbě nabídky. Kumar & Reinartz (2018) popisují význam uvedeného aspektu v kontextu inovačního úsilí, kde stojí za vyšší mírou úspěchu. V běžném provozu firmy může taková orientace vést k vyšší produktivitě procesů a poklesu výdajů na retenční aktivity.

Pojetí výpočtu celoživotní hodnoty zákazníka se liší především v uvažovaných složkách, které mohou zahrnovat pravděpodobnost retence, přímé náklady transakce, náklady na marketingové aktivity, ale i výnosy plynoucí z doporučení atp. S ohledem na zaměření práce uvádí autor dva typy propočtu, tj. základní individuální CLV a individuální CLV rozšířené o pravděpodobnost retence. Kumar & Reinartz (2018) vymezují základní výpočet celoživotní hodnoty individuálního zákazníka jako sumu diskontovaných hrubých příspěvků daného zákazníka ve sledovaném období T.

$$CLV_i = \sum_{t=1}^{T} GC_{it} \left(\frac{1}{1+\delta}\right)^t, \quad (1)$$

kde i označuje individuálního zákazníka, t odpovídá časovému období, δ popisuje diskontní sazbu případně náklady na kapitál, GC_{it} reprezentuje hrubý příspěvek zákazníka v daném období, T je časový horizont, v kterém kalkulaci uvažujeme. Výsledkem je čistá současná hodnota zákazníka CLV_i v čase $t = 0$.

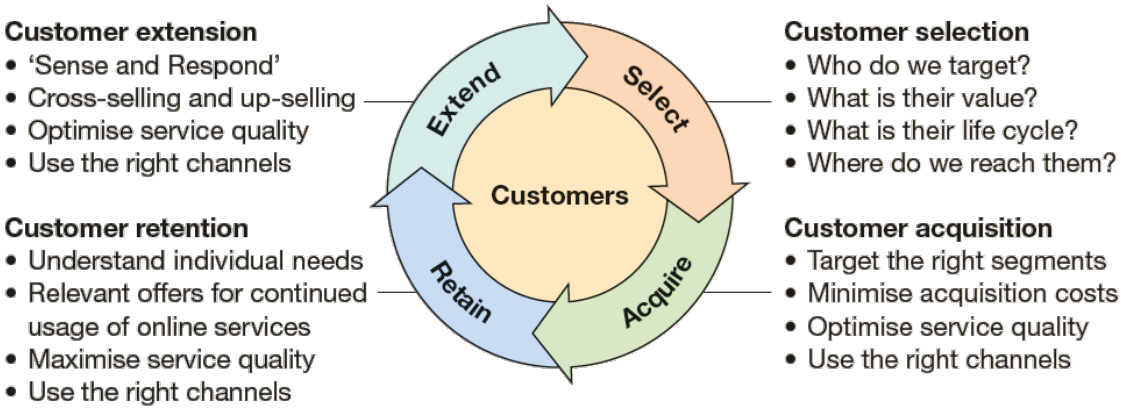
Uvedené základní pojetí předpokládá, že zákazník zůstane aktivní po celou dobu uvažovaného období T. Opak je však pravdou, s plynoucím časem opouští společnost stále více zákazníků, je tedy vhodné uvažovat o pravděpodobnosti retence, případně o jejím doplňku pravděpodobnosti ztráty. Kumar & Reinartz (2018) definují individuální CLV v nastíněných intencích takto.

$$CLV_i = \left(\sum_{t=1}^{T} \left(\prod_{k=1}^{K} Rr_k\right) GC_{it} \left(\frac{1}{1+\delta}\right)^t\right) - AC_i, \quad (2)$$

kde je původní notace rozšířena o průměrnou pravděpodobnost retence Rr_k v časovém období k, součin $\prod_{k=1}^{K} Rr_k$ popisuje pravděpodobnost, že je zákazník v období t aktivní, AC_i.zahrnuje náklady na akvizici. Autoři doporučují využít průměrnou pravděpodobnost především s ohledem na komplexitu odhadu parametru na individuální úrovni.

1.2.2 Retenční management

Nákupní chování spotřebitele lze rozdělit do etap zvažování nákupu, jeho realizace, a etapy po-nákupové. CRM spojuje tyto kroky s marketingovými aktivitami výběru, akvizice, udržení a rozšíření vztahů; integraci činností ilustruje Obr. 4. Chaffey (2015) zde popisuje výběr zákazníka jako vymezení skupiny zákazníků, na které bude společnost cílit s pomocí zbylých aktivit, obvykle s využitím segmentace a předpokládaného zákaznického životního cyklu. Akvizice odkazuje na snahy o formování nových vztahů s hodnotnými zákazníky při co nejmenších nákladech. Retenční aktivity společnosti směřují k udržení zákaznické báze s pomocí relevantních, individuálních pobídek. Podstatná je i kvalita služby a volba vhodných komunikačních kanálů. Rozšíření vztahů potom rozumíme jako podporu spotřeby nabízených produktů a služeb, v šíři i objemu.

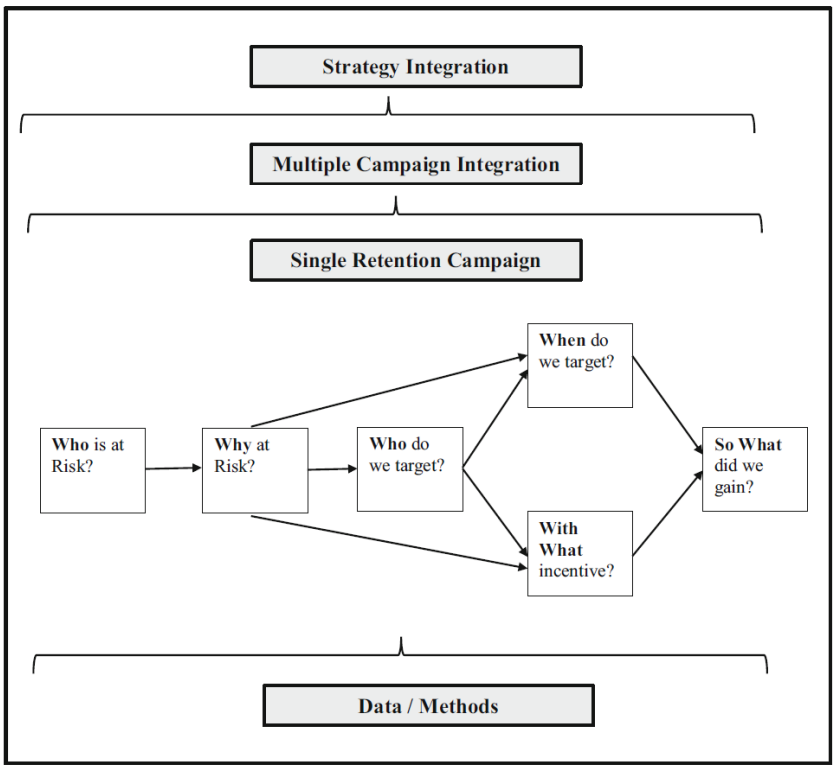


Obr. 4 Tradiční marketingové aktivity využívané při správě vztahů se zákazníky

Zdroj: Chaffey (2015)

Ascarza et al. (2018) charakterizují retenci jako takový stav, kdy zákazník a podnik souvisle interagují. Tato definice zahrnuje jak finanční, tak i nefinanční interakce; tj. vystihuje jak smluvní, mimosmluvní i hybridní vztahy mezi zákazníkem a organizací. Perspektivou kontinuity přistupují k vymezení retence i Chaffey (2015), Kumar & Reinartz (2018) nebo Buttle & Maklan (2019). Protipólem retence zákazníka je pak jeho odchod, respektive ztráta.

Význam retenčního managementu napříč odvětvími akcentuje práce Dawkins & Reichheld (1990); autoři dovozují, že 5% nárůst zákaznické retence vede k navýšení celoživotní hodnoty zákazníka o 25-95 %. Pro e-commerce odhadují Gupta et al. (2004) průměrnou retenční elasticitu 4.9 %, tj. zvýšení retence o 1 % vede k růstu celkového zákaznického kapitálu o 4.9 %. Uvedené výsledky přisuzují Buttle & Maklan (2019) postupně rostoucímu objemu tržeb, poklesu nákladů na správu vzájemného vztahu, nižší citlivosti na cenu a síle doporučení, tj. stejným faktorům které stojí za úspěšnou prací s celoživotní hodnotou zákazníka. Konkrétní aspekty retenčního managementu ohraničujeme v následující části textu s pomocí metodického rámce představeného v Ascarza et al. (2018), který ilustruje Obr. 5.



Obr. 5 Řízení zákaznické retence

Zdroj: Ascarza et al. (2018)

Individuální retenční kampaň

„Kteří zákazníci nás opustí?“ – Úvodní úloha retenční kampaně bývá zpravidla řešena prostředky prediktivního modelování. Cílem je odhalit zákazníky, kteří inklinují k přerušení vztahu se společností. V mimosmluvních vztazích, kde je odchod zákazníka skrytý, však bývá daný problém řešen zřídka. Výzkumníci běžně uvažují vysvětlující proměnné popisující zákaznickou spokojenost, chování a využívání služeb, vlastnosti zákazníka, náklady spojené s ukončením vzájemného vztahu nebo marketingové aktivity. Mezi méně běžné proměnné lze řadit emoce nebo společenské vazby. Vlastní modelování bývá řešeno širokou paletou metod. Neslin et al. (2006) však poukazují na problémy s přesností, díky kterým nemůže být značná část řešení využita v praxi.

„Proč nás chtějí zákazníci opustit?“ – Porozumění příčinám je pro prevenci ztráty zákazníka podstatné. Rozdíl mezi prediktivní a příčinnou vazbou dobře ilustruje demografie zákazníka, která může nést informaci pro predikci daného jevu významnou, příčinou přerušení vztahu však bude zřídkakdy. Potřebná je i úvaha o úrovni detailu, jenž by měla reflektovat nejen

celkové tendence zákaznické báze, ale i specifika individuálního zákazníka. U konkrétních faktorů je nezbytné zvážit, zda je podnik dokáže kontrolovat či nikoliv (Braun & Schweidel, 2011).

„Na jaké zákazníky cílit?“ – Ascarza (2018) navrhuje zaměřit úsilí na zákazníky, kteří jsou ohrožení a podnik je dokáže účinně oslovit. Je třeba zvážit nepřesnost úvodní predikce a případy kdy kampaň směřuje na inertní zákazníky, tj. jejich budoucí chování je neměnné, bez ohledu na retenční kampaň. Autoři upozorňují na paradoxní případ selhání, kdy u spokojených, „spících“ zákazníků vede retenční aktivita k přehodnocení existujícího vztahu se společností. Významné mohou být i mezilidské vztahy, tj. silně propojený uživatel může vykazovat vyšší retenci, nebo naopak vede k vyššímu riziku ztráty zákazníků v sousedství silně propojených vrcholů sociálního grafu. Typickým příkladem takového chování jsou služby mobilních operátorů nebo sociální sítě. Ohled je třeba brát i na smysl retenčních aktivit v kontextu ziskovosti a cílit retenční kampaň na takové zákazníky, u kterých dojde k pokrytí nákladů na retenci i z hlediska dalšího vývoje individuální CLV.

„Jakou pobídku zákazníkovi nabídneme?“ – Porozumění příčinám odlivu zákazníků je pro vymezení konkrétní retenční reakce nezbytné. Zpracování strukturovaných i nestrukturovaných dat s využitím strojového učení umožní společnostem navrhnout a optimalizovat retenční nabídky každému zákazníkovi na míru. Charakter incentivy ovlivňuje i povahu retence, kde pobídky orientované na cenu mají zpravidla krátkodobý vliv, zatímco pobídky orientované na službu nebo produkt mají dopad dlouhodobý (Dodson et al., 1978); z celkového zlepšení úrovně služeb může také těžit celá zákaznická báze. Originálním přístupem je ponechat volbu konkrétní pobídky na každém z oslovených zákazníků.

„Kdy zákazníka oslovíme?“ – Spuštění kampaně lze rámovat jako hledání kompromisu mezi reaktivní retenční kampaní, tj. po rozvázání vzájemného vztahu, a příliš brzkou kampaní, která je v nejlepším případě irelevantní, může ale zákazníka přimět k přehodnocení vzájemného vztahu (Ascarza, 2018). Problém navrhují autoři řešit s pomocí odhadu vývoje pravděpodobností ztráty a záchrany zákazníka v čase, díky kterým je možné popsat funkci pravděpodobnosti záchrany ohroženého zákazníka a nalézt její maximum.

V závěru je nezbytné retenční úsilí vyhodnotit. S pomocí kontrolních skupin je možné odhadnout přímý dopad na příjmy společnosti, ziskovost nebo poměr udržených zákazníků. Ascarza et al. (2017) doporučují zkoumat i konečný dopad na individuální zákazníky, případně společné rysy úspěšných retenčních kampaní. V dlouhodobém horizontu je možné se zabývat

příčinami změn v zákaznickém chování a vlivem retenčních aktivit na změny ziskovosti zákazníka.

Integrace více retenčních kampaní

Postup vedoucí k integraci řeší otázky předestřené v předchozí sekci, v kontextu více kampaní. Blattberg et al. (2008) doporučuje rozložení aktivit v čase s ohledem na cyklus individuálních kampaní, tj. růstu efektu („wear-in“), dosažení maxima, a opadnutí („wear-out“). Mezi zajímavé aspekty integrace je možné řadit dynamickou optimalizaci aktivit s ohledem na úspěch předchozích kampaní, případně rozhodnutí o oddělených a průběžných retenčních programech.

Integrace s marketingovou strategií

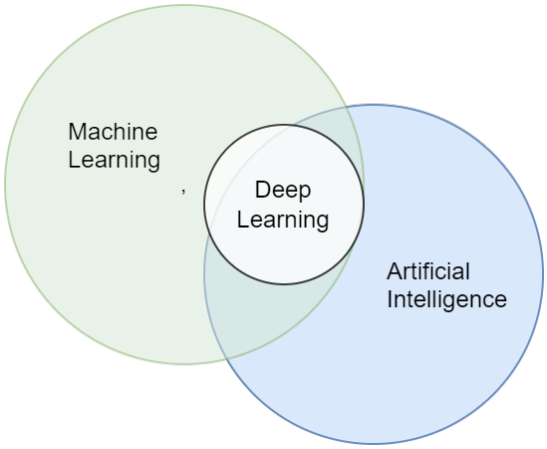
Ukotvení retenčního managementu v marketingové strategii doporučují Ascarza et al. (2018) realizovat koordinací akvizičních a retenčních aktivit, shodou retenčních výdajů s marketingovou strategií a společným přístupem k zákaznické segmentaci, cílení a pozicování. Alokací marketingových výdajů se mimo jiné zabývají práce Blattberg & Deighton (1996) nebo Reinartz et al. (2005), pozoruhodným postřehem je negativní dopad suboptimální alokace nákladů retenčních aktivit na délku životního cyklu zákazníka. Mezi další faktory, které ovlivňují konkrétní podobu alokace výdajů, řadí Ascarza et al. (2017) úroveň konkurence, nebo omezení nabídky. Stahl et al. (2012) poukazuje na nezbytnost koordinace retenčních programů se strategií firmy, tj. společnost která se zabývá výrobou a prodejem luxusního spotřebního zboží by neměla v rámci retenčních aktivit využívat promočních slev atp.

Data a metody

Podniky disponují strukturovanými i nestrukturovanými daty, které mohou při správě vztahů se zákazníky využít. Prevalentní úlohou je předpověď ztráty zákazníka, která bývá adresována s pomocí vysvětlujících proměnných popisujících zákaznickou aktivitu a tradičními metodami, jako jsou logistická regrese, pravděpodobnostní modely nebo analýza přežití. Rozvoj internetového obchodování, sociálních sítí a kontaktů, spolu s technologiemi pro uložení a zpracování velkých dat a strojovým učením rámují nové příležitosti výzkumu zákaznického chování. Aktuální stav poznání v těchto oblastech je předmětem literární rešerše.

1.3 Strojové učení

Strojové učení popisuje systémy, které počítačům umožňují učit se s pomocí dat. Samuel (1959) představuje strojové učení jako „obor, který zkoumá, jak předat počítačům schopnost učit se bez toho, aby byly explicitně naprogramovány“. Více technický pohled předkládá Alpaydin (2020), jenž uvažuje o strojovém učení jako o „programování počítačů tak, aby optimalizovaly dané kritérium s ohledem na data nebo předchozí zkušenosti“. Uplatnění takového pojetí je výhodné všude tam, kde je nemožné dosáhnout řešení s pomocí pevně stanovených pravidel, sledované jevy se vyvíjejí v čase nebo je nezbytné získat vhled do rozsáhlého fenoménu (Géron, 2019).



Obr. 6 Strojové učení v kontextu příbuzných vědeckých disciplín

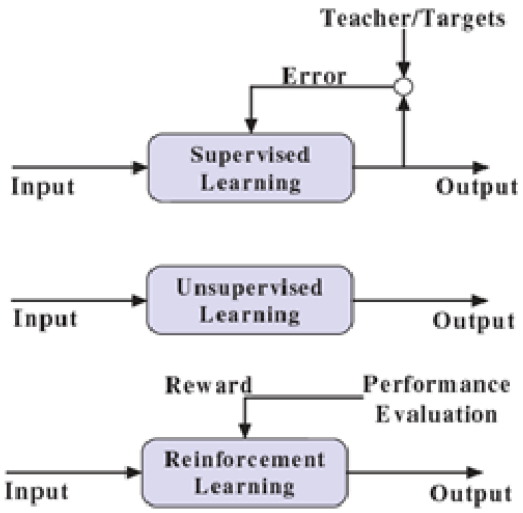
Zdroj: Raschka et al. (2022)

Z hlediska ukotvení mezi ostatními disciplínami bývá část strojového učení pojímána jako součást umělé inteligence, tedy oboru počítačové vědy zabývajícím se řešením komplikovaných úloh, v kterých vynikají lidé. Naznačený průnik pokrývají především hluboké architektury neuronových sítí. Systémy strojového učení ale využívají i metod z oblasti matematické statistiky nebo matematické analýzy, hranice mezi obory jsou v tomto nejasné (Raschka et al., 2022). Nad rámec řešení dobře popsaných úloh v rámci zpracování přirozeného jazyka, rozpoznání řeči nebo počítačového vidění se výzkum umělé inteligence zaměřuje i na tzv. obecnou umělou inteligenci, jejímž cílem je schopnost porozumět a řešit jakýkoliv problém (Hodson, 2019).

1.3.1 Rozlišení úloh strojového učení

S ohledem na širokou paletu přístupů považuje Géron (2019) za užitečné členit systémy strojového učení s ohledem na (1) úroveň lidské supervize, (2) přístupu k novým pozorováním ve fázi učení, a (3) pojetí generalizace. V rámci první z perspektiv, dělí Wang et al. (2012) strojové učení na učení s učitelem („supervised learning“), učení bez učitele („unsupervised learning“), a agentní učení („reinforcement learning“), viz Obr. 7.

Učení s učitelem se zakládá na konstrukci řešení z dat, která obsahují i cílový jev. Dle třídy modelovaného jevu rozlišujeme regresní a klasifikační úlohy. Mezi oblíbené algoritmy náleží lineární regrese, logistická regrese, metoda nejbližších sousedů, metoda podpůrných vektorů, rozhodovací stromy, náhodné lesy nebo neuronové sítě (Géron, 2019).



Obr. 7 Převládající přístupy k úlohám strojového učení s ohledem na úroveň lidské supervize

Zdroj: Wang et al. (2012)

Učení bez učitele využívá dat bez informace o modelovaném jevu, záměrem bývá zpravidla shlukování, detekce anomálií, redukce počtu dimenzí nebo popis asociací. Populární algoritmy zahrnují metodu K-průměrů, metodu podpůrných vektorů s jednou třídou, asociační pravidla, analýzu hlavních komponent a další způsoby projekce dat do prostoru s nižším počtem dimenzí ad. (Géron, 2019)

Agentní učení je zásadně odlišné, učící se systém (agent) čerpá informace z prostředí, vybírá a realizuje jednotlivé akce a je následně odměněn nebo penalizován. Cílem je naučit agenta

takovou strategii chování, která vede k maximalizaci odměn v čase. Dobrým příkladem užití agentního učení je systém AlphaGo, který v roce 2017 porazil ve hře Go úřadujícího světového šampiona Ke Jie. Vítězná strategie byla výsledkem učení s využitím milionů partií, včetně partií mezi agentními systémy (Metz, 2017).

Představme si dostupnou databázi finančních výkazů, kde část historických zpráv je označeno za pozitivní, část za negativní. Tato data mohou být využita k označení zpráv budoucích, pomocí metod založených na učení s učitelem. Pokud jsou výkazy nestrukturované a cílem je odhalení vztahů a témat, je možné využít řešení založené na učení bez učitele. Finanční výkazy také z části charakterizují současný stav finančních trhů, systém může agentního učení k určení vhodné obchodní strategie pro maximalizaci zisku. Prostředí je v tomto případě tvořeno finančními trhy (Koshiyarna et al., 2020).

Systémy strojového učení lze odlišit dle přístupu k novým pozorováním ve fázi učení na dávkové („batch learning“) nebo průběžné („incremental learning“). V rámci dávkového učení se systém nedokáže adaptovat inkrementálně, tj. pro zahrnutí nových pozorování je třeba proces učení opakovat na celé datové sadě. Průběžné učení naopak umožňuje postupnou úpravu parametrů systému, je tedy vhodné v případě souvislého přísunu nových pozorování nebo v případě omezených výpočetních prostředků.

Přístup ke zobecnění zachycených znalostí lze rozlišit na generalizaci dle pozorování („instance-based learning“) a generalizaci modelem. První ze jmenovaných vychází z prostého porovnání nových pozorování s pozorováními stávajícími, v druhém případě dochází k popisu daného vztahu modelem, tj. stávající pozorování slouží k odhadu vnitřních parametrů modelu, s jejichž pomocí dochází k určení pozorování nových (Russell & Norvig, 2022).

1.3.2 Selekce a posouzení modelu

Dělení datového souboru

Dělení datového souboru si klade za cíl přiblížit hodnocení přístupu k modelovanému problému skutečným podmínkám aplikace daného řešení. Výstup takového hodnocení je klíčový pro odhad predikčních schopnosti modelu, výběr vhodného modelu, související hodnoty vnějších parametrů modelu atp.

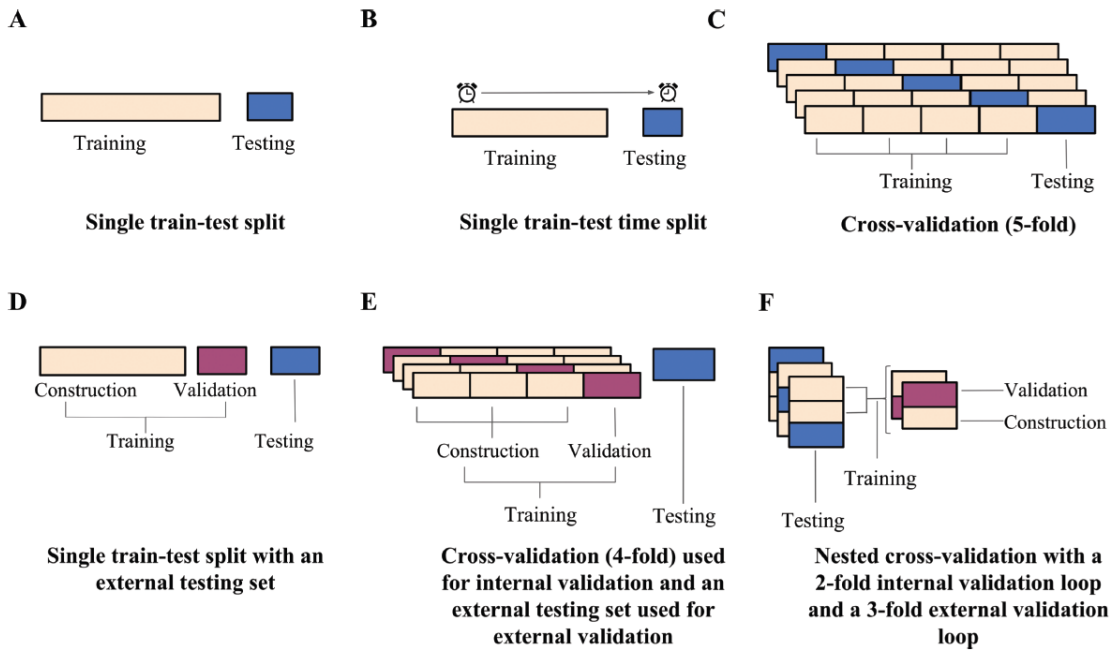
Naivní řešením takové situace by byla konstrukce i evaluace prediktivních metod s využitím jediné množiny dat. Takový postup nadhodnocuje schopnosti modelu a je ospravedlnitelný

pouze v případě, že kýžená množina dat obsahuje všechna přípustná pozorování. Neduh je možné adresovat prostým rozdělením datového souboru na trénovací a testovací množinu dat, kde trénovací množina slouží ke konstrukci modelu, jehož generalizační schopnosti jsou hodnoceny množinou testovací. Konkrétní velikost jednotlivých množin se může lišit, trénovací množina dat však obsahuje zpravidla alespoň polovinu prvků původního souboru. Uvedená technika rozdělení dat je vhodná, pokud dostupná datová sada pokrývá dostatečnou paletu přípustných pozorování, a použitý model je výpočetně náročný.

V případě nedostatečné velikosti dat a nižší komplexitě prediktivního řešení bývá využíváno křížové validace. Procedura dělí soubor dat na k stejně velkých množin. Pro každou z množin K_i pak existuje doplněk K_i^-, kde K_i^- je využit ke konstrukcí modelu a K_i k dílčímu hodnocení generalizace. Postup se opakuje pro každou z k množin, výsledná předpokládaná chyba je potom střední hodnotou dílčích odhadů. Hastie et al. (2009) popisují prvky testovací množiny K_i pomocí vysvětlující proměnné y_i a vektorem vysvětlovaných proměnných x_i. Odchýlení predikcí modelu $\hat{f}^{K_i^-}(x_i)$, od pozorovaných hodnot y_i, vyjadřují jako $L\left(y_i, \hat{f}^{K_i^-}(x_i)\right)$, odhad testovací chyby s pomocí křížové validace je následně možné formulovat jako

$$CVErr(\hat{f}) = \frac{1}{k}\sum_{i=1}^{k} L\left(y_i, \hat{f}^{K_i^-}(x_i)\right). \tag{3}$$

Zvláštní případ křížové validace můžeme definovat jako $k = N$, kde N odpovídá počtu pozorování ve výchozím souboru dat. Důsledkem bývá nízká variabilita trénovací množiny, což vede k příliš optimistickým výsledkům (Varma & Simon, 2006; Arlot & Celise, 2010). Je obvyklé využívat počty testovacích množin $k = 5$ nebo $k = 10$ (Mathai et al., 2020). Vlastní volba počtu testovacích množin však závisí na velikosti výchozího souboru dat, výpočetní náročnosti modelu i nároky na přesnost odhadu střední hodnoty testovací chyby.



Obr. 8 Vybrané přístupy k dělení datového souboru: (A) náhodné rozdělení na trénovací a testovací množinu dat, (B) trénovací a testovací množina dat je oddělena časově, (C) křížová validace, (D) náhodné rozdělení na trénovací, validační a testovací množinu dat, (E) křížová validace trénovací množiny dat, (F) vnořená křížová validace

Zdroj: Mathai et al. (2020)

Hastie et al. (2009) dále doporučují zavést množiny trénovací, validační a testovací, motivací je oddělení výběru modelu (validační data) od konečného odhadu generalizace (testovací data). Pro dobrý odhad chyby v rámci výběru modelu je možné nejprve soubor dat rozdělit na trénovací a testovací množiny, následně aplikovat křížovou validaci na trénovací data. Problematický je v takovém případě odhad na testovací množině dat. Řešením může být vnořená křížová validace, kde vnitřní validace slouží k výběru modelu a vnější k odhadu chyby. Převládající přístupy k dělení datového souboru znázorňuje Obr. 8.

Uvedené přístupy využívají náhodné přiřazení prvků do dílčích množin, jsou tedy vhodné pro případy, kde nedochází k zásadním změnám dat v čase (průřezová data). V opačném případě je třeba omezit tvorbu trénovací množiny na data dostupná do vybraného časového okamžiku a testovací množinu konstruovat pomocí dat nových (Mathai et al., 2020). Takový přístup umožňuje do odhadu chyby generalizace zahrnout nejistotu způsobenou změnami v datové reprezentaci, bývá všem častým zdrojem průsaku informace z trénovací do testovací množiny dat.

Mezi další neduhy náhodného dělení pozorování náleží zkreslení vysvětlované proměnné, které bývá adresováno s pomocí tzv. stratifikovaného výběru. Prvky jsou v takovém případě nejprve rozřazeny dle strat cílové proměnné (např. ztracení a udržení zákazníci). Z každého strata je náhodně vybírán takový počet pozorování, aby bylo dosaženo potřebného zastoupení ve výsledné datové množině (Botev & Ridder, 2017).

Ukazatele úspěšnosti v klasifikačních úlohách

K hodnocení predikční úspěšnosti klasifikátoru bývá užíváno matice záměn. Matice má dvě dimenze, první je indexována pomocí pozorovaných tříd, zbývající potom dle tříd přiřazených, individuální prvky obsahují četnosti odpovídajících kombinací. Speciálním případem je binární klasifikace, kde třídy označujeme jako pozitivní a negativní, viz Obr. 9. Kombinace prvků potom nazýváme jako skutečně pozitivní, falešně negativní, falešně pozitivní a skutečně negativní.

		Assigned Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

Obr. 9 Matice záměn pro binární klasifikační úlohy

Zdroj: Sammut & Webb (2017)

Accuracy (ACC) popisuje podíl správně klasifikovaných pozorování k počtu všech klasifikovaných pozorování (Sammut & Webb, 2017). Oblíbenost ACC pramení především ze srozumitelnosti (v jakém podílu instancí se klasifikátor nemýlil), není však spolehlivá při řešení úloh s nevyváženými třídami. Ve vztahu k matici záměn definujeme metriku jako

$$ACC = \frac{TP + TN}{TP + FN + FP + TN}. \tag{4}$$

Precision (PRE) určuje kolik z pozorování zařazených do pozitivní třídy bylo klasifikováno správně (Sammut & Webb, 2017). PRE se uplatňuje jako doplněk ACC, který dovoluje přesnost klasifikátoru interpretovat v rámci řešeného problému (jaký podíl zákazníků označených jako ztracení byl klasifikován správně). Pomocí prvků matice záměn vymezujeme míru takto

$$PRE = \frac{TP}{TP + FP}. \quad (5)$$

Recall (REC) říká kolik z pozorovaných instancí pozitivní třídy bylo klasifikováno správně (Sammut & Webb, 2017). Také REC obohacuje ukazatele ACC a PRE o srozumitelnost v mezích řešené úlohy (jaký podíl ze všech ztracených zákazníku klasifikátor zachytí). S využitím prvků konfúzní matice charakterizujeme ukazatel následovně

$$REC = \frac{TP}{TP + FN}. \quad (6)$$

F_β Score sdružuje míry PRE a REC do jediného ukazatele, pomocí váženého harmonického průměru (Sammut & Webb, 2017), který zajišťuje, že vysoké hodnoty jedné složky nekompenzují nízké hodnoty složky druhé. V obecné podobě definujeme ukazatel jako

$$F_\beta Score = (1 + \beta)^2 \cdot \frac{PRE \cdot REC}{\beta^2 \cdot PRE + REC}, \quad (7)$$

kde β označuje význam perspektivy REC, nejčastější forma míry je nepreferenční tzn. $\beta = 1$, což vede na obyčejný harmonický průměr PRE a REC.

Lift popisuje relativní užitečnost klasifikačního modelu, obvykle ve srovnání s náhodným určením třídy (Sammut & Webb, 2017). Pomocí předešlých metrik definujeme $Lift$ takto

$$Lift = \frac{ACC}{OPR}, \quad (8)$$

kde OPR odpovídá pozorovanému poměru pozitivní třídy. Populární obměnou je metrika kalkulující $Lift$ pro 10 % zákazníků, jimž model přiřadí nejvyšší pravděpodobnost příslušnosti k pozitivní třídě, důvodem jsou omezené možnosti retenčních aktivit a cílení na ohrožené zákazníky. Zřejmou nevýhodou úpravy je arbitrární určení cílového poměru zákazníků (Verbeke et al., 2012).

Area Under Curve of the Receiver Operating Characteristic (AUCROC) je univerzálním a oblíbeným ukazatelem přesnosti klasifikačních modelů (Bradley, 1997). Máme-li model produkující hodnotu příslušnosti $s = s(x)$, která je funkcí vektoru vysvětlujících proměnných x; hustotu pravděpodobnosti souvisejících hodnot $f_k(s)$, s kumulativním rozdělením pravděpodobnosti $F_k(s)$, a dvě třídy $k \in \{0, 1\}$, pak míru spočteme jako

$$AUCROC = \int_{-\infty}^{\infty} F_0(s) f_1(s) ds\,. \tag{9}$$

$AUCROC$ je možné interpretovat jako pravděpodobnost, že pozorování náhodně vybrané z pozitivní třídy přiřadí klasifikátor vyšší skóre příslušnosti k této třídě než u náhodně vybraného pozorování negativní třídy. Výhodou je nezávislost ukazatele na hranici mezi třídami. Mezi úskalí náleží samotný koncept hodnocení modelu měrou, která vychází z vlastností modelu nikoliv z vlastností klasifikačního problému (Hand, 2009).

Ukazatele úspěšnosti v regresních úlohách

Koeficient determinace (Coeficient of Determination – R^2) charakterizuje úspěšnost modelu jako podíl variability závislé proměnné vysvětlený regresním modelem (Holčík & Komenda, 2015). Ukazatel definujeme jako

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \lambda(x_i))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{10}$$

kde pro n pozorování x_i značíme skutečnou hodnotu vysvětlované proměnné y_i, hodnotu predikce $\lambda(x_i)$, a střední hodnotu vysvětlované proměnné $\bar{y}$. Menšitel popisuje podíl součtu kvadratických chyb modelu a rozptylu závislé proměnné. Neužitečný model může nabývat $R^2 \leq 0$, naopak dokonalý regresní model dosahuje $R^2 = 1$. Ukazatel je oblíbený pro srozumitelnost a přenositelnost mezi datovými soubory. Kritizován je naopak pro neopodstatněný nárůst při zahrnutí nových vysvětlujících proměnných.

Průměrná absolutní chyba (Mean Absolute Error – MAE) vystihuje úspěšnost modelu pomocí průměru absolutních chyb modelu, kde chybou modelu rozumíme rozdíl mezi skutečnou a předpovídanou hodnotou vysvětlované proměnné (Sammut & Webb, 2017). S pomocí představené notace vystihujeme ukazatel následovně

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \lambda(x_i)|}{n}. \tag{11}$$

K prevalenci MAE přispívá srozumitelný odhad chyby v přirozených jednotkách vysvětlované proměnné. Mezi úskalí řadíme podhodnocení extrémních selhání modelu, případně ztrátu informace o směru chyby.

Průměrná kvadratická chyba (Mean Squared Error – MSE) je počítána jako průměr kvadratické chyby modelu, což umožňuje penalizovat extrémní selhání modelu (Sammut & Webb, 2017). S pomocí představené notace vymezujeme ukazatel takto

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \lambda(x_i))^2}{n}. \tag{12}$$

Perspektiva je oblíbená i díky vztahu s vychýlením a rozptylem odhadů. Mezi neduhy náleží odlišnost jednotek při srovnání s vysvětlovanou proměnnou, ztráta informace o směru selhání nebo citlivost na jednotlivá pozorování.

Srozumitelnost modelu

Srozumitelnost vymezují Kim et al. (2016) jako míru, do jaké je člověk schopen předpovědět chování prediktivního systému. Miller (2019) definuje daný termín skrze pochopení příčin konkrétních rozhodnutí zkoumaného systému. Mezi klíčové vlastnosti srozumitelného systému řadí Masís (2021) transparenci modelu, transparencí návrhu experimentu a jeho hodnocení, a reprodukovatelnost. Interpretovatelná řešení umožňují bližší poznání fenoménu, vedou k vyšší spolehlivosti, bezpečnosti a konzistenci, umožňují také reflexi etických aspektů problému.

Molnar (2022) rozlišuje mezi srozumitelností transparentních modelů nebo metod na modelu nezávislých. Transparentní modely bývají zastoupeny zobecněnými lineárními modely nebo rozhodovacími stromy. Oddělení modelovaného problému a přístupu k interpretaci může být výhodné, především z hlediska flexibility, svobody při volbě prediktivního algoritmu, dostupné palety nástrojů, ale i volnosti při volbě nezávislých proměnných (Ribeiro et al., 2016). Molnar (2022) dále rozeznává úrovně detailu na globální a lokální. Globální interpretací se rozumí vnitřní fungování systému, včetně zachycení charakteru asociací mezi závislou a nezávislými proměnnými. Lokální interpretace vysvětluje predikce modelu pro konkrétní instanci datového souboru, což je přínosné pro hlubší porozumění modelované jevu, nebo pro diagnostiku problematických rozhodnutí. Následující odstavce shrnují významné agnostické přístupy k interpretaci.

Permutační významnost proměnných (Permutation Feature Importance) je přístup k hodnocení důležitosti vysvětlujících proměnných, představený spolu s náhodnými lesy v Breiman (2001), a zobecněný napříč rodinami algoritmů v Fisher et al. (2019). Přístup je založený na sledování změn úspěšnosti modelu poté co je pořadí hodnot hodnocené vysvětlované

náhodně zamícháno, čímž dojde k přerušení vazby mezi vysvětlující a vysvětlovanou proměnnou, na kterou se prediktivní model spoléhá, tj. pro významnou vysvětlující proměnnou předpokládáme vyšší pokles prediktivních schopností modelu.

Výhodou je srozumitelný vhled do chování modelu, který je srovnatelný napříč modely i problémy, bere v úvahu interakce mezi proměnnými, dále potom není opakovat fázi trénování modelu. Mezi nevýhody můžeme řadit potřebu relevantního datového souboru, včetně známých vysvětlovaných proměnných, potřebu adresovat kompromis mezi výpočetním časem a stabilitou výsledného odhadu, permutace některých proměnných vede k tvorbě nerealistických instancí, nebo nerealistické odhady významnosti u silně korelovaných vysvětlujících proměnných (Masís,2021; Molnar, 2022).

Graf částečné závislosti (Partial Dependece Plot) zobrazuje marginální efekt množiny vybraných vysvětlujících proměnných na predikce modelu, umožňuje tak kvalifikovat povahu vztahů mezi vysvětlovanou a vysvětlujícími proměnnými. Mějme množinu zvolených vysvětlujících proměnných X_S, a doplněk těchto proměnných X_C, pak částečnou závislost predikovaných hodnot f v bodě x_S definujeme jako

$$pd_{X_S}(x_S) = E_{X_C}[f(x_S, X_C)] = \int f(x_S, x_C)\, p(x_C)\, dx_C, \tag{13}$$

kde $f(x_S, x_C)$ je prediktivní funkcí modelu pro daný soubor dat, který je vymezený zvolenými proměnnými $x_S \epsilon\, X_S$ a doplňkem $x_C \epsilon\, X_C$. Spočtením integrálu napříč různými hodnotami zvolené proměnné umožňuje vynést graf částečné závislosti.

Mezi výhody řadíme jasnou, intuitivní srozumitelnost nástroje, které vyplývá z definice funkce částečné závislosti, tj. graf závislosti popisuje, změnu hodnoty predikce modelu, ve vztahu ke změně vysvětlované proměnné. Neposlední výhodou je kauzální interpretace vztahu zachyceného modelem, což ovšem nemusí platit pro modelovaný jev. Mezi možné nedostatky řadíme nutnost vizuální inspekce a z ní plynoucí omezení pro vynášení interakcí více proměnných. Metoda dále předpokládá nezávislost vysvětlujících proměnných, což je často naivní a vede k výpočtu marginálního efektu pro nerealistická pozorování. Metoda také může skrýt různorodé, protichůdné chování jednotlivých pozorování. Poslední dva neduhy je možné adresovat s využitím akumulace místní efektů (Molnar, 2022; Thampi, 2022).

Lokální agnostická vysvětlení (Local Interpretable Model-Agnostic Explanations – LIME) umožňují interpretovat predikce těžko nesrozumitelného modelu, pro konkrétní instanci datového souboru (Ribeiro et al., 2016). Metoda využívá predikcí vysvětlovaného modelu pro zčásti simulovaná vstupní data, na kterých se konstruují modely lokální. Cílem je vytvořit soubor vysvětlujících modelů, které úspěšně zachycují lokální vlastnosti vysvětlovaného modelu. Vysvětlujícími modely jsou zpravidla zobecněné lineární modely nebo rozhodovací stromy. Matematicky je možné zástupný model popsat následovně

$$explanation(x) = arg \min_{g \in G} L(f, g, \pi_x) + \Omega(\mathrm{g}), \tag{14}$$

kde se vysvětlení predikce pro instanci x pomocí lokálního modelu g, který minimalizuje součet účelové funkce L, popisující vzdálenost od výchozího modelu f a regularizačního prvku $\Omega(\mathrm{g})$, cílem je totiž přesný a jednoduchý lokální model. G označuje množinu lokálních modelů, π_x potom definici regionu v kterém k výběru zástupného modelu dochází.

Lokální agnostická vysvětlení jsou oblíbená díky jasné kontrastní interpretaci, i možnosti odhadnout věrnost ve vztahu k původnímu modelu. Zajímavou možností je také interpretace na jiném souboru vysvětlujících proměnných, než byl použitý pro konstrukci vysvětlovaného modelu. Mezi nevýhody je řadíme potřebu stanovit dodatečné vnější parametry, jako jsou způsob konstrukce umělých datových instancí, velikost regionu, počet lokálních modelů atp. S tímto souvisí i nestabilita existujících implementací i pro opakované experimenty, které tak mohou být manipulovány (Masís, 2021; Molnar, 2022).

Shapleyho hodnoty (Shapley Values) označují koncept teorie her, který kvantifikuje spravedlivé rozdělení celkové odměny mezi koalici hráčů v kooperativní hře. (Shapley, 1953). V kontextu interpretace modelu rozumíme „hrou“ predikci pro konkrétní pozorování, „celková odměna“ označuje hodnotu predikce pro danou instanci, zmenšenou o průměrnou hodnotu predikce pro zbylé instance, „hráčem“ jsou konkrétní hodnoty vysvětlujících proměnných formující koalice, na jejichž základě, dochází k odhadu dílčích spravedlivých odměn. Formálně můžeme definovat Shapleyho hodnotu proměnné i

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!} \left(v(S \cup \{i\}) - v(S)\right), \tag{15}$$

kde v značí funkci modelu mapující koalici na konkrétní hodnotu predikce, S značí podmnožinu vysvětlujících proměnných, které neobsahují i, N značí množinu dostupných

proměnných a n její velikost. Vzorec můžeme interpretovat jako situaci, kdy dochází k postupnému formování koalice, kde každý z hráčů požaduje spravedlivou kompenzaci ve výši $v(S \cup \{i\}) - v(S)$. Pro každého hráče pak získáme průměrnou spravedlivou kompenzaci napříč všemi permutacemi, v jakých může daná koalice vzniknout. Shapleyho hodnoty se vyznačují následujícím vlastnostmi. Efektivita hodnot spočívá v rozdělení celkové odměny mezi hráče, tj. součet dílčích odměn v koalici všech hráčů odpovídá celkové výši odměny. Symetrie zajišťuje stejnou úroveň Shapleyho hodnot pro stejně přispívající hráče. Neplatnost zabezpečuje nulovou Shapleyho hodnotu pro hráče, který nepřispívá k odměně koalice. Linearita Shapleyho hodnot zajišťuje možnost sčítání dílčích odměn hráče, napříč funkcemi mapování.

Výhody Shapleyho hodnot tkví v teoretickém základu postaveném na koaliční teorii her, tj. spravedlivé rozdělení odměn mezi vysvětlující proměnné a kontrastní vysvětlení. Problematické jsou především s ohledem na výpočetní čas, tvorbou nerealistických datových instancí nebo možnosti manipulace s interpretací. Z praktického hlediska může být problematická potřeba mít k dispozici potřebná data, využití všech vysvětlujících proměnných, výsledné přiřazení není model, který je možné dotazovat jako u LIME (Masís, 2021; Molnar, 2022).

Shapleyho aditivní vysvětlení (Shapley Additive Explanations – SHAP) označují soubor metod jejichž cílem je vysvětlit predikce konkrétních datových instancí (Lundberg & Lee, 2017). SHAP vychází z teorie her. Nabízí alternativní přístupy k aproximaci Shapleyho hodnot, ale i konceptuální rozšíření pro navazující globální interpretaci. Cílem je vysvětlit predikce pro vybranou instanci s pomocí odhadu příspěvků každé z vysvětlujících proměnných. Jednou z inovací, kterou SHAP přináší oproti klasickým Shapleyho hodnotám, je reprezentace s pomocí aditivních charakteristik atribučního modelu, což lze považovat jako spojovací prvek mezi LIME a Shapleyho hodnotami. Formálně pak můžeme SHAP zapsat takto

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \tag{16}$$

kde g je vysvětlující model, z' koaliční vektor, M je maximální velikost koalice a $\phi_i \epsilon R$ je přiřazený efekt vysvětlující proměnné. K odhadu Shapleyho hodnot dochází s pomocí simulace, která uvažuje pouze část koaličních proměnných. S výhodou je možné spočíst odpovídající ϕ_i s pomocí lineárního modelu.

Klíčové vlastnosti SHAP hodnot považujeme lokální přesnost, absenci a konzistenci. Při zahrnutí všech vysvětlujících proměnných a $\phi_0 = E_x(\hat{f}(x))$, chápeme místní přesnost podobně efektivitu Shapleyho hodnot, pouze s využitím celého koaličního vektoru.

$$\hat{f}(x) = g(x') = \phi_0 + \sum_{j=1}^{M} \phi_j x'_j \tag{17}$$

$$\hat{f}(x) = \phi_0 + \sum_{j=1}^{M} \phi_j x'_j = E_x(\hat{f}(x)) + \sum_{j=1}^{M} \phi_j. \tag{18}$$

Absencí rozumíme situaci, kdy chybějící proměnné přiřazujeme nulovou odměnu, přestože by s ohledem na koaliční vektor mohla nabývat jakýchkoliv hodnot, bez vlivu na výslednou lokální přesnost, v praxi má tato vlastnost vliv na dopad konstantních proměnných. Konzistence říká, že pokud se mezní příspěvek vysvětlující proměnné zvýší nebo zůstane stejný, pak se zvýší nebo zůstane stejný i přiřazený příspěvek dané proměnné, bez ohledu na proměnné ostatní. Lundberg & Lee (2017) prokazují, že díky těmto charakteristikám platí symetrie, neplatnost a linearita pro SHAP, podobně jako je tomu u Shapleyho hodnot.

Mezi přednosti řadíme, podobně jako u Shapleyho hodnot, pevné teoretické základy a spravedlivé rozdělení příspěvků vysvětlujících proměnných, dále také možnost jednotným způsobem interpretovat globální i lokální chování prediktivního systému. Za zmínku stojí i efektivní implementace pro některé algoritmy strojového učení. Mezi nedostatky uvažujeme výpočetní komplexitu, tvorbu nerealistických datových instancí u univerzálních metod, dále také náročnost interpretace nebo možnost manipulace výsledků (Masís,2021; Molnar, 2022).

1.3.3 Vybrané algoritmy

Zobecněné lineární modely (Generalized Linear Models – GLM) jsou širokou třídou přístupů postavených na tradičních lineárních modelech, umožnují však modelovat vysvětlované proměnné s pomocí exponenciálních rozdělení pravděpodobnosti, čímž umožní zachytit i nelineární vztahy. GLM se sestávají z náhodné složky, charakterizující rozdělení pravděpodobnosti vysvětlované proměnné, systematické složky popisující lineární kombinaci vysvětlujících proměnných, a transformační funkce, jenž zachycuje vazbu mezi uvedenými složkami (Hastie et al., 2009). Prostý lineární model předpokládá normální rozdělení náhodné komponenty a transformační funkci identity. Rovnici modelu pak můžeme vyjádřit jako

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p, \tag{19}$$

kde $E(Y|X)$ označuje funkci identity, X_p popisuje vysvětlující proměnnou p, β_p koresponduje s interním parametrem modelu.

Pro klasifikační úlohy bývá využívána logistická regrese. U LR uvažujeme binomické rozdělení vysvětlované proměnné a transformační funkcí logaritmu šancí. Mějme střední hodnotu vysvětlované proměnné $\mu(X) = P(Y = 1|X)$, rovnici modelu potom zapíšeme jako

$$\log\left(\frac{\mu(X)}{1-\mu(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p, \tag{20}$$

kde $\log\left(\frac{\mu(X)}{1-\mu(X)}\right)$ označuje funkci logaritmu šancí, X_p popisuje vysvětlující proměnnou p, β_p koresponduje s interním parametrem modelu. Koeficienty modelu bývají odhadnuty metodou maximální věrohodnosti (Hastie et al., 2009). GLM předpokládají nezávislost pozorování i lineární nezávislost vysvětlujících proměnných. Obliba metod souvisí především se srozumitelností odhadnutých parametrů, možnostmi regularizace a nízkou výpočetní složitostí. Za stinné stránky považujeme nároky na zpracování datového souboru, omezené možnosti zachycení komplexních vztahů nebo vysokou variabilitu predikcí.

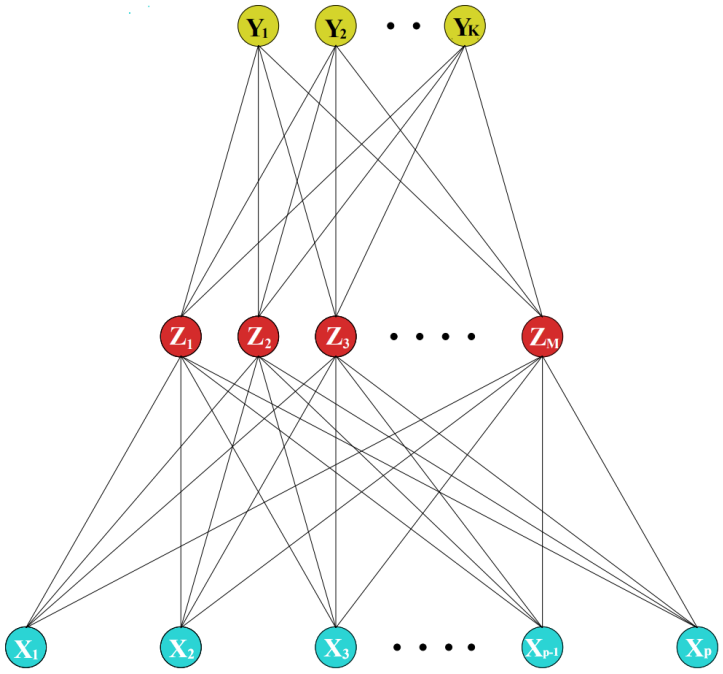
Metody podpůrných vektorů (Support Vector Machines – SVM) přistupují k problému klasifikace pomocí promítání původního prostoru vysvětlujících proměnných do nadprostoru, pomocí jádrové transformace („kernel function“), kde dochází k nalezení nadroviny jenž maximalizuje minimální vzdálenosti okrajových bodů tříd („support vectors“). Lineární separace v nadprostoru vede k tvorbě nelineární hranice mezi třídami v prostoru původním. Mějme pozorování x pro které odhadujeme třídu $\hat{G}(x)$ pomocí

$$\hat{G}(x) = sign\left(\sum_{i=1}^{N} \hat{\alpha}_i y_i K(x, x_i) + \hat{\beta}_0\right), \tag{21}$$

kde vysvětlovaná proměnná nabývá hodnot $\{-1,1\}$, x_ia y_i charakterizují podpůrné vektory, parametry $\hat{\alpha}_i$ a $\hat{\beta}_0$ popisují nadrovinu a jádrová funkce $K(x, x_i)$ slouží k výpočtu vnitřního součinu vektorů v transformovaném nadprostoru. Určení podpůrných vektorů a parametrů nadroviny bývá dosaženo s pomocí duální účelové funkce, která umožňuje formulovat problém jako lineárně omezenou kvadratickou úlohu. Metoda je oblíbená díky minimalizaci strukturálního rizika a existenci globálního řešení problému. Zachycené zákonitosti je však obtížné

interpretovat, běžné implementace se navíc vyznačují kubickou výpočetní a kvadratickou prostorovou složitostí (Bishop, 2006; Hastie et al., 2009).

Umělé neuronové sítě (Artificial Neural Networks – ANN) označují matematické modely volně inspirované biologickými neuronovými sítěmi. Základní myšlenou umělých neuronových sítí je konstrukce odvozených proměnných, z lineární kombinace existujících vysvětlujících proměnných, a následné modelování vysvětlované proměnné pomocí nelineární transformace odvozených vstupů. Nejpoužívanějším typem neuronové sítě je dopředná síť s jednou skrytou vrstvou, využívající zpětného šíření chyby. Uvedená topologie předmětem Obr. 10.



Obr. 10 Schématické vyobrazení dopředné neuronové sítě s jednou skrytou vrstvou

Zdroj: Hastie et al. (2009)

V případě klasifikační úlohy, je třeba konstruovat síť tak, že počet prvků svrchní vrstvy odpovídá počtu tříd K, kde pak každý prvek odhaduje pravděpodobnost příslušnosti k dané třídě. V regresních úlohách bývá užívána jednoprvková svrchní vrstva. Počet prvků spodní vrstvy určuje délka vektoru atributů x. Odvozené proměnné, jež jsou konstruovány ve skryté vrstvě sítě, označujeme jako Z_m. Výsledné pravděpodobnosti $\hat{f}(x)$ získáme aplikací aktivační funkce na lineární kombinaci Z_m. Formálně je možné postup zapsat následovně

$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T x), \mathrm{m} = 1, \ldots, \mathrm{M}, \tag{22}$$

$$T_k = \beta_{0k} + \beta_k^T Z, k = 1, \dots, K, \quad (23)$$

$$\hat{f}(x) = g_k(T), k = 1, \dots, K, \quad (24)$$

kde α_{0m} a α_m^T označují váhy skryté vrstvy, β_{0k} a β_k^T reprezentují váhy výstupní vrstvy, aktivační funkce skryté vrstvy σ bývá sigmoid, hyperbolický tangens, radiální nebo rektifikovaná lineární funkce (Bishop, 2006; Hastie et al., 2009). Aktivační funkcí výstupní vrstvy g_k bývá softmax, výsledný vektor pak můžeme interpretovat jako pravděpodobnost příslušnosti k třídě. V regresních úlohách využívá svrchní vrstva zpravidla lineární aktivační funkce. Stanovení vhodných vah bývá dosaženo iterativní úpravou vah s ohledem na změny účelové funkce („gradient descent back-propagation"), u klasifikačních úloh se typicky jedná o křížovou entropii, u regresních úloh o střední kvadratickou chybu. Umělé neuronové sítě jsou univerzálním nelineárním modelem, který může díky interní extrakci vysvětlujících proměnných, zachytit velmi složité vzorce chování. Popis vztahů, které natrénovaná síť reprezentuje, však není přímo srozumitelný. Mezi další úskalí patří stanovení vhodných parametrů modelu, vysoká variabilita predikcí i výpočetní komplexita (Bishop, 2006; Hastie et al., 2009).

V posledních letech zažívá rozmach specifická podmnožina umělých neuronových sítí, tzv. hluboké učení. Jedná se třídu sítí, které obsahují velký počet skrytých vrstev a jsou postavené na komplikovaných architekturách. Hluboké sítě jsou úspěšné především v oblastech rozpoznání řeči, zpracování přirozeného jazyka, počítačového vidění, nebo agentního učení (Goodfellow et al., 2016).

Metody nejbližších sousedů (k-Nearest Neighbors – kNN) popisují přístupy založené na odhadu vysvětlované proměnné s pomocí nejbližších *k* známých pozorování. Přístup závisí na kalkulaci vzdálenosti mezi instancemi a předpokládá vysvětlující proměnné ve srovnatelných jednotkách. KNN je používaná především s ohledem na jednoduchost implementace, a možností interpretovat predikce pro konkrétní pozorování. S ohledem na výpočetní složitost může být obtížné využít metodu na souborech dat s vysokým počtem pozorování a dimenzí, případně v situacích, kdy je třeba generovat predikce v reálném čase (Hastie et al., 2009; Molnar, 2022).

Rozhodovací stromy (Decision Trees – DT) popisují soubor metod které dělí prostor vysvětlujících proměnných do množin obdélníků, následně pro každý takový útvar odhadnou vysvětlovanou třídu (v regresní úloze konstantu). Stromy jsou obvykle konstruovány rekurzivně, hladový algoritmus rozhoduje o vhodné proměnné a dělícím prahu uzlu, což určuje topologii výsledného grafu. V klasifikačních problémech je kritérium tvorby uzlu založeno na míře

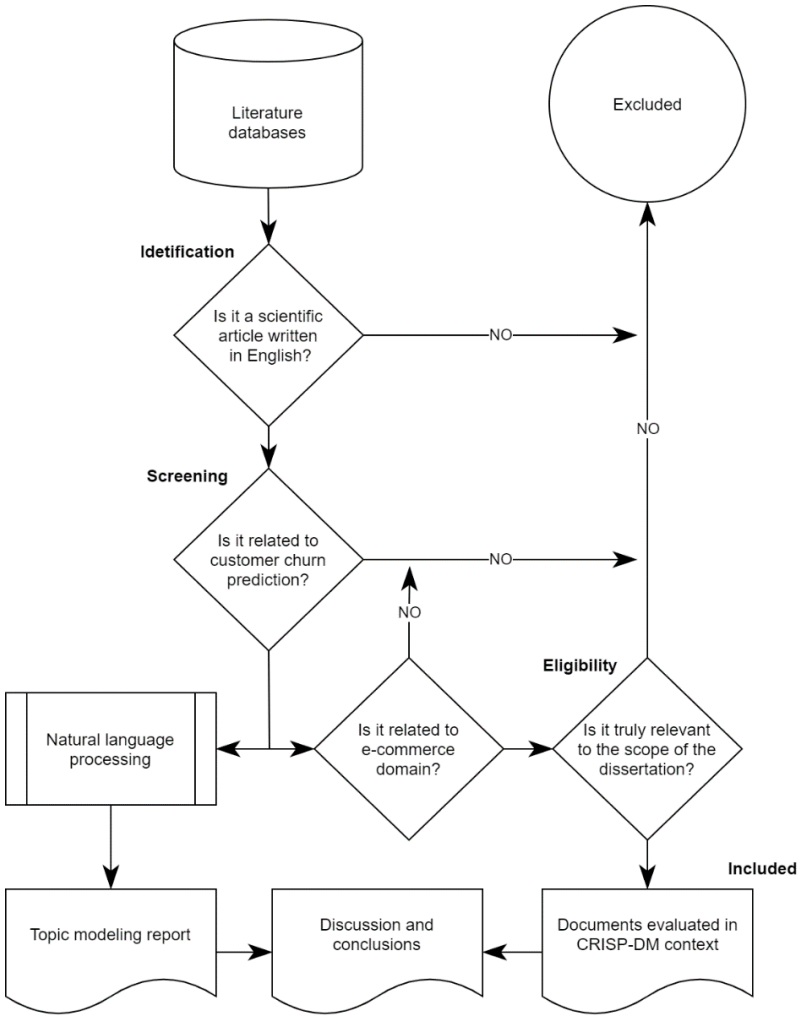
selhání, Gini indexu nebo křížové entropii. Mezi užívané implementace patří algoritmy CART, C4.5, C5.0, CHAID nebo ID3. Klíčovou předností rozhodovacích stromů je možnost interpretace zachycených znalostí, které připomíná rozhodovací proces experta. Mezi pozitiva dále řadíme schopnost zachytit komplexní funkce, doprovodnou selekcí vysvětlujících proměnných nebo výpočetní složitost. Problematická je, podobně jako u logistické regrese, citlivost na jednotlivá pozorování (Hastie et al., 2009).

Bootstrap aggregating (Bagging) je technikou kombinace dílčích modelů do tzv. ansámblu modelů, což vede u nestabilních algoritmů ke zlepšení přesnosti (Breiman, 1996). Motivací je hledání kompromisu mezi schopností modelu zachytit skutečnou funkci jevu a citlivostí na konkrétní pozorování. Metoda spočívá v konstrukci nových sad, které pochází z náhodného výběru původní datové množiny s opakováním („bootstrap“). Každý z vytvořených datových souborů pak slouží k trénovaní dílčího modelu, predikce modelů je možné agregovat s pomocí většinového hlasování, průměrů pravděpodobnosti tříd ad. Oblíbenou variací popsané procedury jsou náhodné lesy („random forests“), jejichž základním dílčím modelem je rozhodovací strom. Lesy jsou oproti stromům méně citlivé na konkrétní pozorování, čímž zabraňují přetrénování modelu, redukují však možnost interpretace a navyšují složitost řešení (Bishop, 2006; Hastie et al., 2009).

Boosting je další z přístupů spojující více modelů. Metoda staví na sekvenčním učení dílčích modelů, které jsou přidávány do finálního souboru. Každý další model bere v potaz chybu stávajícího souboru, tj. soustředí se na především na pozorování u kterých existující ansámbl modelů selhává. Mezi populární implementace lze řadit AdaBoost (Freund & Shapire, 1997), případně moderní škálovatelné přístupy XGBoost (Chen & Guestrin, 2016) nebo LightGBM (Ke et al., 2017). Algoritmy jsou velmi citlivé na jednotlivá pozorování, což vede k možnosti zachycení složitých funkcí i nebezpečí vysoké variability predikcí. Mezi úskalí dále náleží redukce srozumitelnosti i vyšší výpočetní náročnost.

2 Literární rešerše

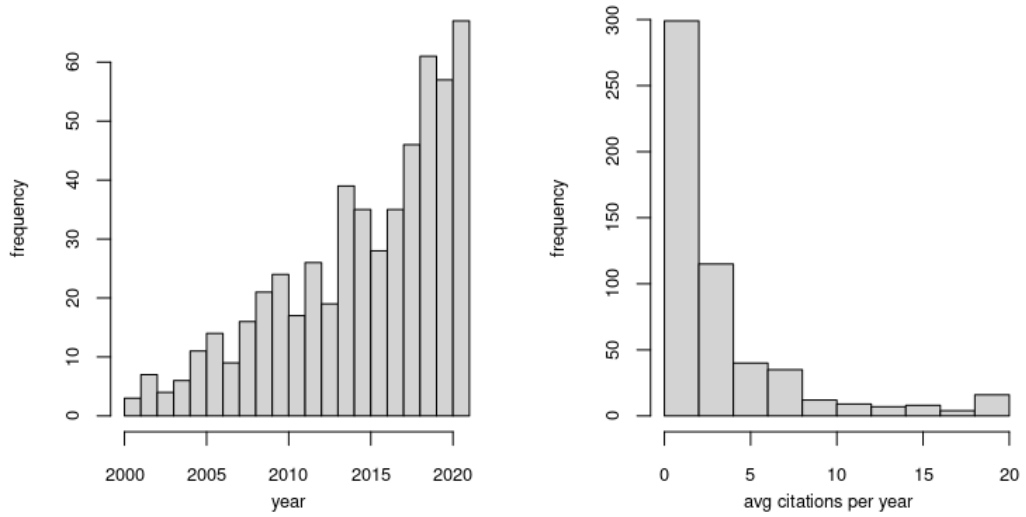
Cílem této sekce je popsat aktuální stav poznání modelování odchodu zákazníka, a ten dále porovnat se specifiky prostředí e-commerce. Pro zajištění transparentního výběru a hodnocení literárních zdrojů vycházíme z principů PRISMA (Moher et al., 2009); postup je dokumentován vývojovým diagramem Obr. 11.



Obr. 11 Metodický rámec pro výběr a hodnocení literárních zdrojů

Jako pro disertační práci relevantní byly identifikovány anglicky psané vědecké články publikované v recenzovaných časopisech indexovaných v databázích Web of Science nebo Scopus. Oblast širšího zájmu je vymezena jako průnik množiny článků zacílených na ztrátu zákazníka (klíčová slova: „customer churn“, „customer attrition“, „customer defection“, „customer retention“) a množiny článků prediktivního modelování (klíčová slova: „prediction“, „forecasting“, „modeling“, „machine learning“, „data mining“). V období 01/2000–09/2021 bylo v databázích indexováno 595 relevantních článků, plný text jsme získali u 549 z nich.

Prostředí e-commerce potom chápeme jako podmnožinu uvedeného průniku (klíčová slova: „electronic commerce“, „e-commerce“, „online“, „internet“, „web“). Zde se podařilo identifikovat 34 článků, z nichž je 29 v souladu se zaměřením práce.



Obr. 12 Vývoj počtu článků zaměřených na modelování ztráty zákazníka v čase (nalevo) a průměrný počet citací za rok (napravo)

Množina textů popisující modelování odchodu zákazníka je studována s využitím metod pro zpracování přirozeného jazyka. Cílem autora je popsat témata, odhadnout jejich prevalenci, identifikovat trendy výzkumu, a probádat vztahy mezi tématy. Podmnožina dokumentů, zaměřených na sektor elektronického obchodování je studovány prostředky tradiční rešerše. Poznatky obou přístupů jsou v závěru shrnuty a kriticky zhodnoceny.

2.1 Predikce ztráty zákazníka

2.1.1 Modelování témat

K utřídění a porozumění rozsáhlému souboru vědeckých prací, zaměřených na predikci ztráty zákazníka, se autor rozhodl využít metod zpracování přirozeného jazyka, respektive výpočetní lingvistiky. Oblastí zájmu je především modelování témat, tj. vysvětlení části podobnosti v pozorované množině dokumentů pomocí nepozorovaných strukturálních faktorů (témat). Takový přístup umožňuje členění, sumarizaci i anotaci rozsáhlých souborů nestrukturovaných dat; užitečnost takového přístupu pro porozumění souboru vědeckých textů prokazují Griffiths & Steyvers (2004), Wang & Blei (2011), Blei (2012), Bohr & Dunlap (201), Fridrich (2020), ad.

Obvykle bývá modelování témat adresováno s pomocí nelineárního pravděpodobnostního generativního modelu – latent Dirichlet allocation (LDA). Předpokládaný proces tvorby dokumentů se skládá z odhadu distribuce pravděpodobnosti pro výskyt témat v dokumentech, odhadu distribuce pravděpodobnosti slov v tématech, a přiřazení tématu a související distribuce pravděpodobnosti obsahu témat pro každé slovo napříč dokumenty. Generativní proces určí společné rozdělení pravděpodobnosti skrz pozorovanými a nepozorovanými náhodnými proměnnými. Toho je následně užito k odhadu rozdělení podmíněné pravděpodobnosti skrytých témat za v dokumentech pozorovaných slovech.

Mějme témata $\beta_{1:K}$, kde každé β_k je distribuce pravděpodobnosti slovníku. Směs témat dokumentu d jest θ_d. Označení tématu dokumentu d je poté z_d, přiřazení tématu n-tému slovu v dokumentu d jest $z_{d,n}$. Pozorovaná slova dokumentu d jsou f_d, n-té slovo v dokumentu d jest $f_{d,n}$. Kýžené rozdělení podmíněné pravděpodobnosti pak je (Blei et al., 2003):

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|f_{1:D}) = \frac{\prod_{i=1}^{K} p(\beta_i) \prod_{i=1}^{D} p(\theta_i)(\prod_{n=1}^{N} p(z_{d,n}|\theta_d)p(f_{d,n}|\beta_{1:K}, z_{d,n}))}{p(f_{1:D})}. \quad (25)$$

Čitatel zlomku reprezentuje spojené rozdělení pravděpodobnosti latentních proměnných a vymezuje vztahy mezi jednotlivými prvky výpočtu, je spočitatelný. Jmenovatel popisuje pravděpodobnost pozorování daného slovníku pro všechny modely témat, je nespočitatelný. V praxi se proto odhaduje s pomocí vzorkovacích nebo variačních metod (Blei, 2012).

Jiným algoritmem z rodiny nelineárních pravděpodobnostních generativních modelů je strukturální model témat (STM), jenž umožňuje zavést funkční závislost dílčích odhadovaných rozdělení pravděpodobnosti na dalších, nezávislých proměnných. Předpokládaný proces tvorby dokumentů se skládá z odhadu distribuce pravděpodobnosti výskytu témat jako funkce externích nezávislých proměnných, volby distribuce pravděpodobnosti obsahu témat, jako funkce externích nezávislých proměnných, a observačního modelu, který urovnává výsledky předchozích zdrojů variace odpovídající kompozicí slov napříč dokumenty.

Pro formální popis čitatele posteriorního rozdělení pravděpodobnosti rozšiřme notaci LDA následujícím způsobem. Nechť η značí vzorkovanou hodnotu normálního rozdělení, které je podkladem pro konstrukci log-normálního rozdělení θ_d s rozptylem Σ. Matice X a Y popisují nezávislé proměnné výskytu témat a obsahu. Matice Γ a K reprezentují apriorní koeficienty modelu výskytu a obsahu (Roberts et al., 2014; Roberts et al., 2016).

$$p(\eta, z, \kappa, \gamma, \Sigma | f, X, Y) \propto$$

$$\left(\prod_{d=1}^{D} Norm(\eta_d | X_d \gamma, \Sigma) \left(\prod_{n=1}^{N} Mult(z_{n,d} | \theta_d) \times Mult(f_d | \beta_{d,k=z_{d,n}}) \right) \right) \times \prod p(\mathrm{K}) \prod p(\Gamma) \quad (26)$$

Rozdílem mezi odhady podmíněné pravděpodobnosti získané pomocí LDA a STM je skutečnost, že STM umožňuje odlišit výskyt i obsah tématu napříč všemi dokumenty. Posteriorní rozdělení pravděpodobnosti je odhadováno s pomocí variačního algoritmu pro hledání maximálně věrohodného odhadu. Kompletní matematický aparát představují autoři této metody v publikaci Roberts et al. (2016).

Kvalita témat

Frekvence výskytu a výjimečnost slov jsou významnými faktory nastiňující sémantický obsah textu. Frekventovaná slova bývají rozprostřena napříč mnoha tématy a méně častá slova nemusí nést potřebnou informaci, proto je třeba faktory vyvážit. Bischof & Airoldi (2012) navrhují kompenzaci frekvence výskytu a exkluzivity slov s pomocí váženého harmonického průměru následovně. Nechť f značí slovo v tématu k, pak

$$FREX = \sum_{k=1}^{K} \sum_{f=1}^{F} \left(\frac{w}{E_{f,k}} + \frac{1-w}{F_{f,k}} \right)^{-1}, \quad (27)$$

kde w reprezentuje váhu exkluzivity E, která popisuje empirické kumulativní rozdělení pravděpodobnosti normované frekvence výskytu slova f v tématu k, F označuje empirické kumulativní rozdělení frekvence výskytu slova f v tématu k. Vysoké hodnoty výsledného skóre lze dosáhnout vysokým počtem ohraničených témat.

Sémantická soudržnost (Mimmo et al., 2011) vychází z ideje vzájemné informace a předpokládá, že pravděpodobná slova nesourodých témat by se měla vyskytovat ve stejném dokumentu. Mějme slova f_ia f_j, pak pro množinu nejpravděpodobnějších slov v tématu k M vyjádříme významovou koherenci jako

$$C_k = \sum_{i=2}^{M} \sum_{j=1}^{i-1} \log\left(\frac{D(f_i, f_j) + 1}{D(f_j)}\right), \tag{28}$$

kde $D(f_i, f_j)$ popisuje společný výskyt slov f_i *a* f_j, analogicky $D(f_i)$ reprezentuje výskyt slova f_j. Praktické aplikace ukazatele naznačují, že vysoké soudržnosti lze dosáhnout při malém počtu témat, která mají mnoho společných slov (Roberts et al., 2014).

Podobnost témat

Podobnost rozdělení pravděpodobnosti výskytu jednotlivých témat napříč dokumenty je odhadnuta s pomocí Jensen-Shannon divergence (JSD), jež vychází z Kullback-Leibler divergence (KL). Pokud máme vektory pravděpodobnosti x, pak můžeme vzdálenost mezi rozděleními P a Q odhadnout s pomocí KL divergence $D(P \parallel Q)$ následovně

$$D(P \parallel Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}. \tag{29}$$

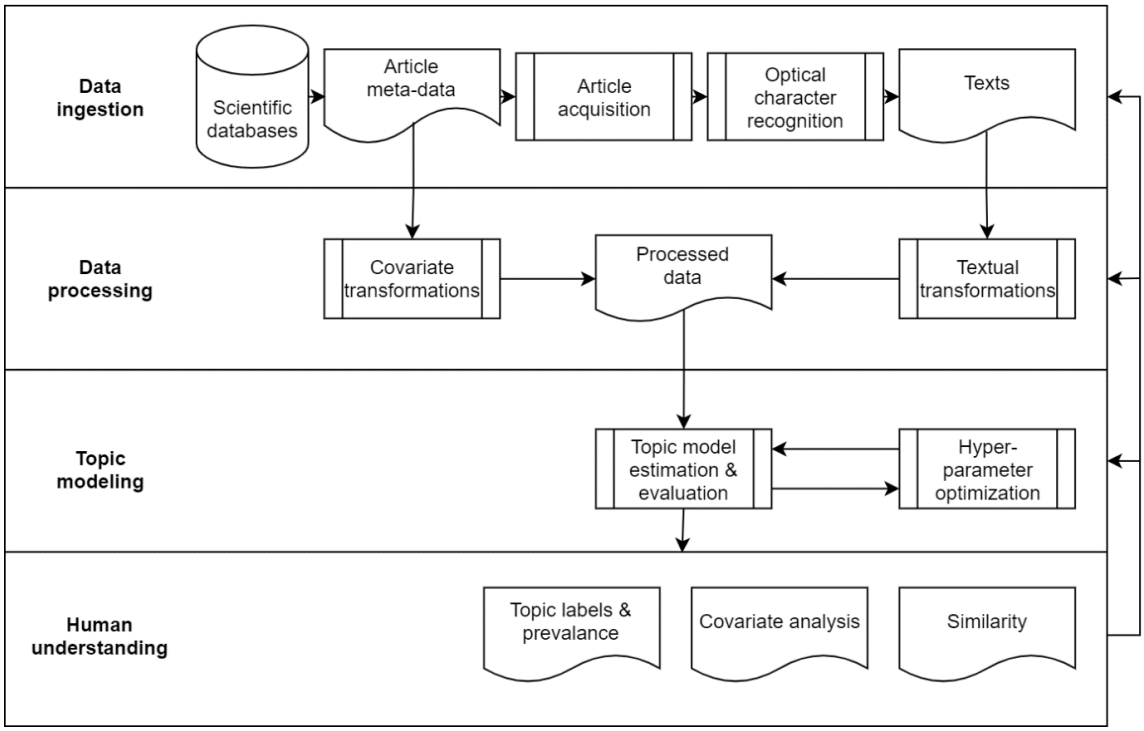
KL divergence není definovaná pro $P(x) \neq 0 \wedge Q(x) = 0$, platí také $D(P \parallel Q) \neq D(Q \parallel P)$, tzn. divergence není symetrická. JSD proto odhaduje podobnost rozdělení pravděpodobnosti jako

$$JSD(P \parallel Q) = \frac{1}{2} D\left(P \parallel \frac{1}{2}(P + Q)\right) + \frac{1}{2} D\left(Q \parallel \frac{1}{2}(P + Q)\right). \tag{30}$$

Jedná se tedy o prostý průměr KL divergencí P a Q od rozdělení $\frac{1}{2}(P + Q)$, což zajišťuje potřebnou symetrii. Motivací k užití druhé odmocniny JSD (Jensen-Shannon vzdálenost), před populární kosinovou vzdáleností, náleží bližší popis sémantiky prostoru (Hockenmaier, 2020) i experimentální výsledky dosažené Lee (2001).

2.1.2 Návrh řešení a implementace

Vlastní zpracování vědeckých článků je iterativním procesem, který je složen z bloků: získání textové reprezentace a souvisejících dat, zpracování dat, modelování témat, a interpretace modelů. Vazby mezi jednotlivými bloky popisuje vývojový diagram na Obr. 13. Navržené řešení staví na metodickém přístupu k tvorbě strukturálních modelů témat představeném Roberts et al. (2019). Tento je rozšířen v oblastech syntaktické analýzy textu, určení kandidátních modelů nebo interpretace.



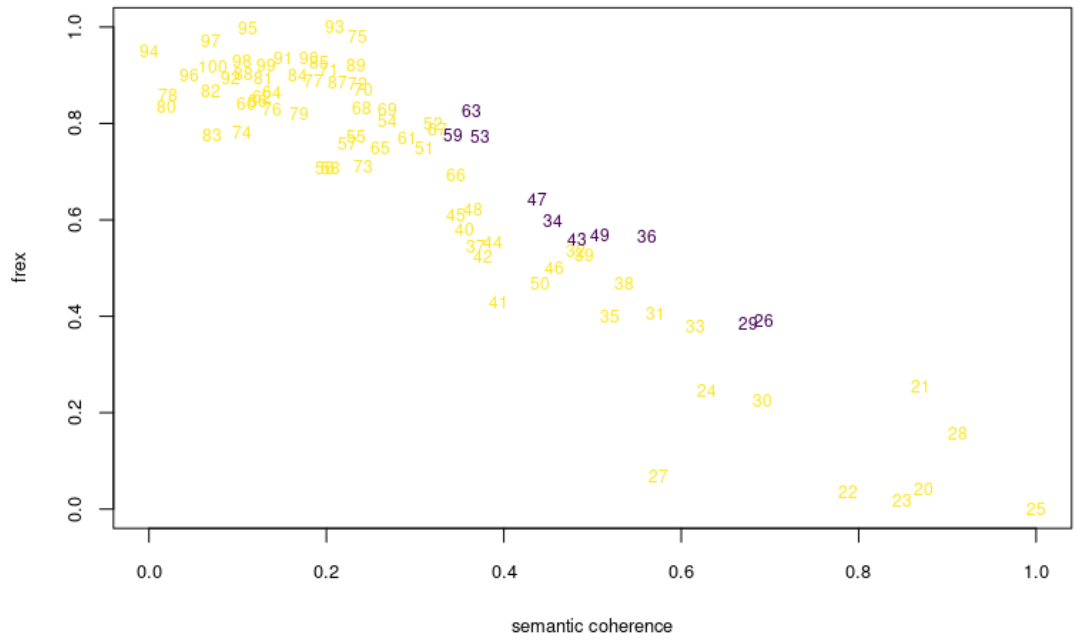
Obr. 13 Proces zpracování vědeckých článků s pomocí metod výpočetní lingvistiky

Úvodní blok činností popisuje opatření datové reprezentace, jež vychází z identifikace relevantních článků ve vědeckých databázích a exportu souvisejících metadat. Tato data jsou unifikována a očištěna o duplicitní záznamy, především s ohledem na digitální identifikátor a text abstraktu. Výsledná metadata slouží jako podklad pro získání úplného znění dokumentů ve formátu PDF, které jsou převedeny do prostého textu s pomocí nástroje Tessaract (Kay, 2007).

Transformace vstupních dat spojuje větve zpracování externích nezávislých proměnných a zpracování textů. V první větvi je konstruována proměnná popisující význam dokumentu jako zastropovaný průměrný roční počet citací, důvodem je především nutnost kompenzace prostého počtu citací s ohledem na stáří dokumentu. Proměnná popisující rok vydání zůstává beze změny. V druhé větvi dochází k transformaci textů, kde zachováváme výhradně alfanumerické znaky. Následně je využito vektorového modelu en_core_web_lg knihovny spaCy (Honnibal & Montani, 2017) pro tokenizaci, parsování a syntaktickou analýzu textu. Autor předpokládá, že pouze lemmata určitého druhu a minimální pozorované frekvence jsou významnými nositeli informace. Tento přístup snižuje nezbytnost manuálních zásahů do slovníku i výpočetní čas navazujících kroků; může však negativně ovlivnit kvalitu výsledného modelu.

V rámci modelování témat je definován vztah mezi externími nezávislými proměnnými a způsob odhadu interních parametrů modelu a dochází k hledání modelů vhodných pro další

interpretaci. Výstupem bloku je množina kandidátních modelů. Pro prevalenci témat s ohledem na rok vydání a průměrný počet citací je předpokládán prostý lineární vztah s interakcí prvního řádu, důvodem je především srozumitelnost. Parametry modelu jsou inicializovány pseudonáhodně pomocí LDA, následně dochází během ~ 750 EM iterací k jejich upřesnění. S ohledem na počet dokumentů a rozsah slovníku lze očekávat skutečný počet skrytých témat v rozmezí 20 až 100, což vede na výběr mezi 81 STM modely. Kvalita modelů je hodnocena dvěma instrumentálními ukazateli, průměrnou kompozitní metrikou FREX a průměrnou sémantickou koherencí. K výběru je přistupováno jako k nepreferenční vícekriteriální optimalizaci s lokálním bodem utopie. Vhodné modely pak leží na Pareto-optimální hranici, nedaleko utopistického řešení. Výsledná množina kandidátů obsahuje 10 modelů, v rozsahu témat od 26 do 63. Graficky je tento proces ztvárněn na Obr. 14, kde lze pozorovat umístění modelů s různými počty témat mezi rozsahem škálovanými ukazateli; kandidátní modely jsou vyneseny ve fialové barvě.



Obr. 14 Výpočetní hodnocení strukturálního modelu témat

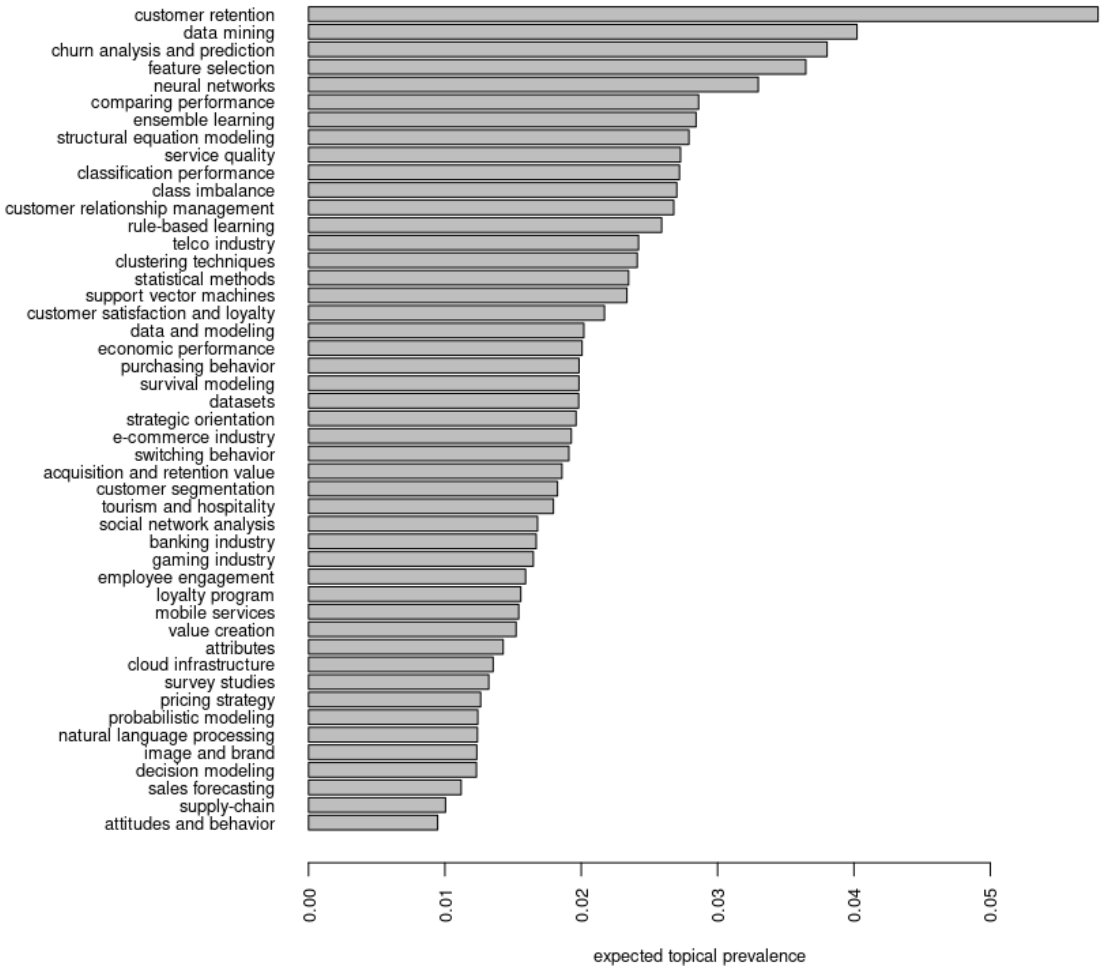
Pro porozumění obsahu kandidátních modelů následuje autor postup předestřený Roberts et al. (2019); určení prevalence a popisných jmen témat, s ohledem na asociované dokumenty a slova, porozumění vztahu mezi prevalencí témat a externími nezávislými proměnnými, a zkoumání podobnosti výskytu témat napříč dokumenty, tj. jaká témata jsou řešena pospolu.

Nejprve je pro každé téma určena množinu asociovaných slov a dokumentů. Vhodná slova jsou zjištěna především dle pravděpodobnosti výskytu, a celkovou frekvencí a exkluzivitou v rámci tématu (viz rovnice 27). Pro úplnost je využito i podílu frekvencí slova v tématu a mimo něj a podíl logaritmů těchto frekvencí; obě metriky přirozeně zeslabují význam slov rozprostřených skrz více témat. Vhodné dokumenty jsou získány dle očekávané míry výskytu témat napříč texty. Dané téma pak lze uchopit a pojmenovat právě pomocí asociovaných slov a dokumentů; část těchto perspektiv je obsahem příloh A1 a A2.

Závislost míry výskytu tématu na externích vysvětlujících proměnných předpokládá prostý lineární vztah s interakcí prvního řádu. Určení parametrů lineárního modelu je založené na simulaci s globálním odhadem nejistoty (viz Roberts et al., 2019). Vztahy popsané tímto modelem jsou zkoumány vizuálně, s moderací pro nevynesenou proměnnou na úrovni mediánu, a 95% intervaly spolehlivosti pro prevalenci; perspektivy jsou ilustrovány přílohami A3 a A4.

Dále je analyzována podobnost mezi tématy skrz množinu zkoumaných dokumentů. Pro určení vzdálenosti mezi faktory je využita vzdálenost založenou na Jensen-Shannon divergenci. Vztahy mezi tématy autor dále interpretuje s pomocí hierarchického aglomerativního shlukování, kde vzdálenost mezi skupinami určujeme Wardovou metodou.

2.1.3 Dosažené výsledky



Obr. 15 Označení detekovaných témat, včetně očekávané míry výskytu

Na základě hodnocení a interpretace kandidátních modelů, s použitím dříve uvedených nástrojů, vybral autor instanci modelu formující 47 témat, jenž dosahuje sémantické soudržnosti -25.744 a hodnot FREX 9.711. Mezi identifikovanými tématy lze pozorovat skupiny prvků popisujících některé z aspektů predikce ztráty zákazníka jako jsou proces modelování („data mining", „churn analysis and prediction", „feature selection", „neural networks"), řízení vztahů se zákazníky („customer retention", „service quality", „customer relationship management", „customer satisfaction and loyalty"), odvětví („telco industry", „tourism and hospitality", „banking") a některé ostatní („survey studies").

Nejčastěji se v textech objevuje téma „customer retention" s prevalencí ~ 5.8 %. Takový výsledek je v souladu s očekáváním autora; řešená problematika je součástí snah vedoucích k udržení zákazníka. Obecně lze tvrdit, že znalost modelované domény je pro úspěšné využití

metod strojového učení nezbytná. Mezi další populární témata patří „data mining" s očekávanou proporcí ~ 4.0 % a „churn analysis and prediction" s očekávaným výskytem ~ 3.8 %. Jedná o široká témata, která jsou blíž metodám a způsobu hodnocení prediktivních modelů než podnikovému kontextu řízení vztahů se zákazníkem.

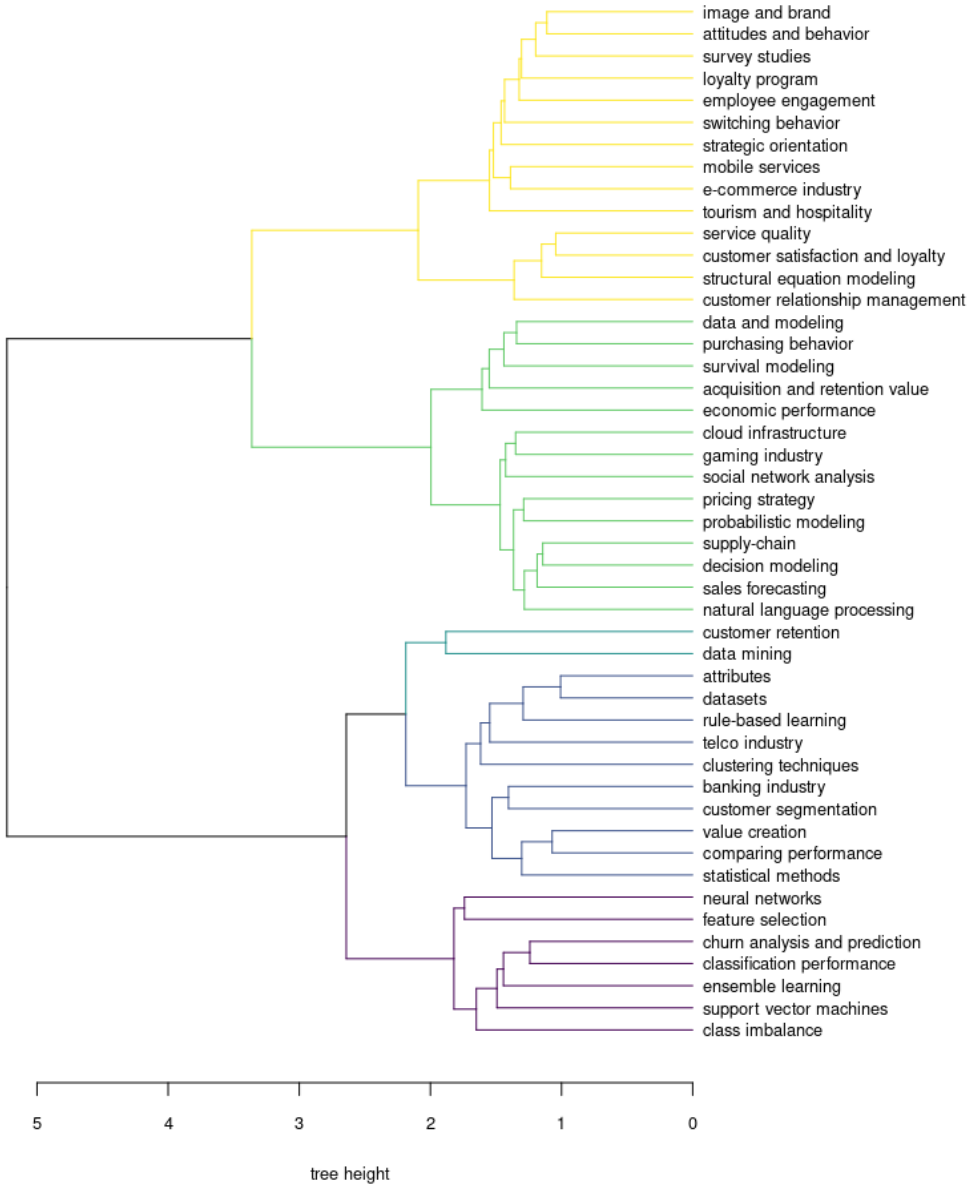
Pozoruhodný vzestup v prevalenci napříč lety demonstrují některé aspekty prediktivního modelování jako „feature selection", „ensemble learning" nebo „class imbalance". Na druhou stranu dochází k poklesu „rule-based learning" a „survival modeling". Témata „ensemble learning" a „class imbalance", se dokonce zdají být často citovaná. S ohledem na šíři intervalů spolehlivosti je ale takový závěr spíše neprůkazný. V zásadě lze tuto situaci číst jako úkrok vědecké komunity směrem ke strojovému učení. Příkladem budiž práce Poornappriya & Durairaj (2019), která představuje nový způsob výběru vysvětlujících proměnných založeném na minimalizaci redundance a maximalizaci relevance s pomocí nejasných množin a optimalizace hejnem částic; navržený algoritmus je pouze ověřen na problému predikce ztráty zákazníka.

Dalším zajímavým úkazem je pokles tématu „purchasing behavior", tj. zdá se, že články z posledních let využívají skupiny vysvětlujících proměnných zaměřených nad rámec nákupního chování zákazníka. Skupinu takových textů je možné najít mezi „customer satisfaction and loyalty", kde se výzkumníci obvykle zabývají vnímanými aspekty produktů a služeb, nebo u „social network analysis", jež s využitím síťové analýzy popisuje interakce mezi subjekty a jejich vliv na modelovaný fenomén. Je zřejmé, že skupiny nezávislých proměnných jsou odlišné i mezi odvětvími.

Perspektiva významu jednotlivých témat dle průměrného počtu citací za rok odhaluje silný pozitivní vztah u „classification performance" a „economic performace". Jinými slovy, výzkum podepřený životaschopným návrhem experimentu nebo zaměřený na ekonomické aspekty ztráty zákazníka bývá často citován. Zajímavostí je, že míra výskytu u těchto témat neroste v čase příliš rychle, tj. jedná se o potenciálně zajímavé oblasti predikce ztráty zákazníka. Konkrétními příklady obou témat mohou být vědecké práce Ahmed & Maheswari (2019) a Devriendt et al. (2020), v kterých autoři využívají ekonomické perspektivy reakce zákazníka na retenční kampaň.

V rámci odvětví lze pozorovat úpadek zájmu o tradiční „banking industry" v čase. Při bližším pohledu na, v citacích význačné, „mobile services" se, ale ukazuje, že se tento zájem spíše tříbí. Ukázkou může být článek Kumar et al. (2018), který s pomocí strukturálního modelu

rovnic zkoumá, jak uživatelé vnímají službu mobilních plateb z perspektiv vlastní zkušenosti, bezpečnosti nebo nápravy stížností. Rostoucí v čase je „gaming industry", kde výzkumníci řeší výzvy nového sektoru, obvykle s využitím velkých dat. Castro & Tsuzuki (2015) prezentují užitečnost vlnkové transformace frekvenční časové řady přihlášení pro predikci odchodu uživatele napříč několika herními tituly. Význačný růst citací v odvětví „tourism and hospitality" lze v posledním měsících spojit i s dopady pandemie COVID-19, např. Yu et al. (2021) s pomocí strukturálního modelu rovnic ověřuje perspektivu hygienických opatření jako nástroje úspěšné zákaznické retence.



Obr. 16 Podobnost rozdělení pravděpodobnosti témat napříč vědeckými publikacemi

Pro zjištění, jaká témata se v článcích objevují pospolu zkonstruoval autor matici vzdálenosti založené na Jensen-Shannon divergenci. Na tuto matici nahlížíme prostřednictvím stromovém grafu vyneseném na Obr. č. 16; pro účely intepretace arbitrárně dělí latentní faktory do následujících pěti skupin (od shora dolů), orientace podniku na zákazníka (žlutá), modelování nákupního a dalšího chování (zelená), zaštiťující pojmy (akvamarína), explorace dat a tradiční prediktivní metody (modrá), a strojové učení (fialová). Postupem ke kořenu stromu je možné snadno odhadnout primární výzkumné zaměření skupin dokumentů, tj. podnikový kontext ztráty zákazníka (žlutý a zelený shluk) a prediktivní modelování (akvamarínový, modrý a fialový shluk). Při postupu opačným směrem, je v některých případech možné identifikovat dílčí úzce zaměřené podmnožiny shluků, které popisují jak zkoumaný problém, tak charakter vstupních dat nebo užité metody. Pěknou ukázkou takového shluku je spodní část žluté skupiny témat zaměřená na kvalitu služby, zákaznickou spokojenost, řízení vztahu se zákazníkem, a strukturální modely rovnic, případně svrchní část zeleného shluku zaměřená na nákupní chování, specifické statistické modely, a ekonomickou hodnotu akvizice nebo retence zákazníka.

2.1.4 Shrnutí a diskuse

Napříč změnami prevalence v čase můžeme konstatovat, že dochází k odklonu od obecných a dominantních témat jako „customer retention" nebo „data mining" a příklonu k specifickým oblastem prediktivního modelování, přičemž dochází i k posunu v použité terminologii. Tento závěr naznačuje určitou zralost vědecké domény, která využívá diseminaci souvisejícího výzkumu i postupující zákaznické orientace mnohých odvětví. Pro účely disertační práce považuje autor za podstatnou především identifikaci velmi citovaných a méně prevalentních témat „classification performance" a „economic performance", která se vypořádávají s návrhy experimentů, hodnocením klasifikačních modelů, a v neposlední řadě také diskrepancí mezi modelovanými problémy a podnikovým kontextem jejich řešení. Dále se podařilo identifikovat některé běžně adresované podproblémy („class imbalance", „feature selection") nebo skupiny populárních klasifikačních algoritmů („rule-based learning", „neural networks", „support vector machines", „ensemble learning"). Pří pohledu na výskyt skrytých faktorů skrz zkoumané texty, autor identifikoval dvě převažující skupiny prací, tj. články zaměřené na podnikový kontext ztráty zákazníka a prediktivní modelování. Toto zjištění je v souladu s vymezeným polem zájmu.

Pro srovnání dosažených závěrů s existující literaturou autor upravil stávající dotaz do vědeckých databází zacílením na texty označené jako „review". Takto získané práce využívají

klasického přístupu k rešerši literatury, což se odráží na odlišné granularitě problému. Ngai et al. (2009) ve své analýze literatury představují čtyři pilíře, v kterých dobývání znalostí může podpořit řízení vztahů se zákazníky, kde udržení zákazníka je oblastí nejobsáhlejší. Nižší počet zkoumaných textů umožnil autorům popsat modelové případy využití klasifikačních a jiných metod; zajímavým postřehem je celkové podcenění vizuálních prvků pro komunikaci modelovaných fenoménů. Ze srovnání s předestřenou rešerší lze identifikovat třídy algoritmů, které jsou populární dodnes („neural networks“), případně se dostaly do popředí zájmu během poslední dekády („ensemble learning“). Jain et al. (2021) zkoumají literaturu popisující predikci ztráty zákazníka v prostředí telekomunikačních společností. Sektorové zaměření umožnilo autorům popsat specifika fenoménu, veřejně dostupné datové sady, i některé aspekty prediktivního modelování jako výběr nezávislých proměnných, klasifikační algoritmy nebo způsob hodnocení jednotlivých řešení. Překvapující je absence důrazu na síťové aspekty odchodu zákazníka. Při komparaci s vlastní rešerší můžeme konstatovat, že se autorovi nepodařilo identifikovat třídu metod založených na fuzzy logice, její nízkou prevalenci naznačuje i práce Britto & Gobinath (2020).

Limitace představeného výzkumu vychází především ze způsobu vymezení zkoumané domény, užitých metod a perspektivě. Pro analýzu byla vybrána široká paleta textů, kde část textů nemusí být pro zkoumaný problém zásadní; ukázkou mohou být deskriptivní modely zákaznické spokojenosti a loajality. V budoucnu by bylo možné specifikovat zkoumanou oblast s pomocí úžeji zaměřeného dotazu, který by byl omezen i ve smyslu konkrétních časopisů. Dále by stálo za zvážení uvést i dokumenty vícero typů; v oblasti strojového učení je značná část aktuálních vědeckých výstupů prezentována formou konferenčních příspěvků. Externích nezávislé proměnné by pak bylo možné rozšířit o metadata textů jako typ vědecké práce, příslušný časopis, vydavatelský dům nebo kolekce vědecké databáze.

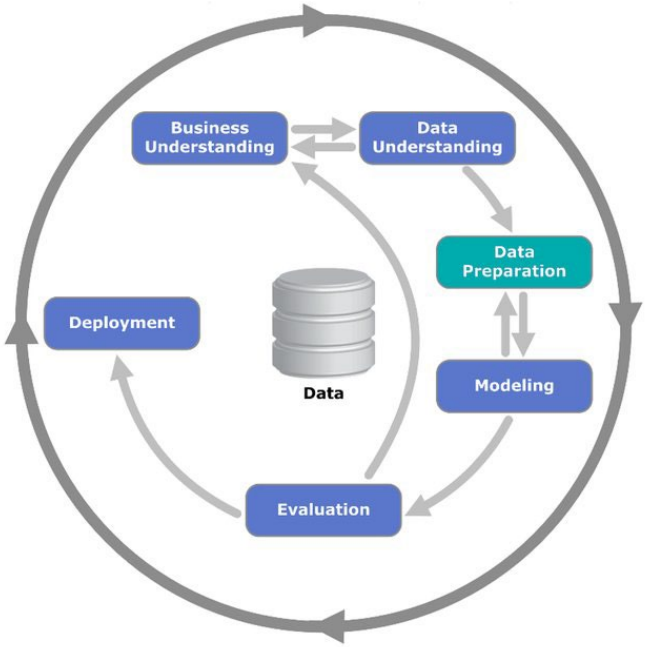
Pro získání textové reprezentace by bylo možné využít rozdílných přístupů k transformaci obrazových dokumentů do potřebné podoby, třeba s pomocí existujících programatických rozhraní některého z poskytovatelů cloudových služeb. K zpracování textů by bylo možné uvažovat o neinvazivní korektuře textu, alternativním způsobům filtrování význačných tokenů s pomocí kolokací, síťové analýzy, nebo vektorové reprezentaci dokumentů. Pro konstrukci modelů témat pomocí STM by bylo možné uvažovat zavedení nových nezávislých proměnných, případně nelineární závislosti prevalence na těchto proměnných. Určení kandidátního modelu lze rozšířit v účelové funkci, prohledávaném prostoru, optimalizačním algoritmu. Dále by bylo

možné prozkoumat možnosti pokročilých metod modelování témat založených na hlubokých neuronových sítích.

V neposlední řadě je limitujícím faktorem úroveň detailu, v jaké je daná aplikace NLP zpracována; v popředí zájmu jsou obecné stránky modelování ztráty zákazníka jako zákaznická orientace, modelování nákupního chování, explorace dat nebo prediktivní modelování; nikoliv však konkrétní přístup autorů k těmto aspektům. Uvedené omezení je adresováno v další sekci práce, která zevrubně analyzuje relevantní podmnožinu dokumentů popisující předpověď ztráty zákazníka ve vybrané vertikále.

2.2 Ztráta zákazníka v e-commerce

Pro roztřízení odborné literatury v oblasti predikce ztráty zákazníka bylo využito referenčního modelu CRISP-DM (Cross-Industry Standard Process for Data Mining), který popisuje fáze životního cyklu prediktivního modelování a související logické vazby (Chapman et al., 2000). V současnosti se jedná o stěžejní metodiku využívanou napříč projekty aplikujícími metody dobývání znalostí nebo strojového učení (Schröer et al., 2021).



Obr. 17 Fáze životního cyklu prediktivního modelování

Zdroj Chapman et al. (2000)

Životní cyklus modelování se dle CRISP-DM skládá ze šesti fází, které vyobrazuje Obr. 17. Ze směru vazeb mezi kroky je patrné, že se jedná o iterativní proces, vnější kružnice naznačuje kontinuální charakter prediktivního modelování – doručení projektu často vede k navazujícím, specifickým otázkám, které je třeba adresovat. Vymezení jednotlivých kroků ve vztahu ke zkoumané doméně je obsahem následujících podkapitol.

2.2.1 Vymezení problému

Úvodní fáze prediktivního modelování se zaměřuje na porozumění cílům projektu a podnikové nebo výzkumné perspektivě problému, získané znalosti jsou pak transformovány do jasně ohraničené úlohy strojového učení a rámcového plánu implementace takového řešení. Pro potřeby práce se autor omezuje na perspektivu úlohy a hodnocení výstupu modelování v daném kontextu.

Vybrané články cílí na podnikovou perspektivu problému, modelování a strojové učení, nebo analýzu odborné literatury. Podnikovou perspektivou rozumíme snahu adresovat otázky zákaznické retence nad rámec prosté identifikace rizikových zákazníků. Příkladem může být snaha Hengliang & Weiwei (2012) a Rachid et al. (2018) o využití vlastností popisujících aktuálnost, frekvenci a peněžní hodnotu zákaznických transakcí k vymezení vlastního jevu ztráty, případně k určení hodnotných zákazníků, na které je třeba při modelování soustředit pozornost.

Modelováním a strojovým učením postihujeme výzkum, který se věnuje vybraným technickým aspektům předpovědi ztráty zákazníka. Chen (2016) předpovídá ztrátu zákazníka s pomocí generativního hierarchického pravděpodobnostního modelu, který autor vhodně kombinuje s regulačními diagramy; zdá se, že takový přístup umožňuje dosáhnout uspokojivých výsledků i pro velmi omezenou sadu vysvětlujících proměnných. Yu et al. (2011) využívají problému nejmenší opsané sféry k aproximaci duální formy SVM, což vede k zvýšení úspěšnosti řešení a snížení časové i prostorové složitosti. Wang et al. (2019) představují architekturu neuronové sítě založené na kombinaci konvoluční extrakce dynamických atributů zákaznického chování a statických atributů, které jsou propojené ve svrchních vrstvách navržené sítě; životaschopnost takového řešení je prokázána srovnáním s jinými populárními klasifikačními algoritmy. Neuronové sítě jsou rovněž předmětem zájmu Venkatesh & Jeyakarthic (2020), jenž se zabývají hledáním vhodné kombinace vnějších parametrů rekurentní neuronové sítě; navržený přístup je inspirován sociálními interakcemi v sloních stádech. Je s podivem, že datová sada užitá pro ověření predikčních schopností řešení neodpovídá charakteru neuronové sítě.

Prolnutí obou perspektiv demonstruje práce Song et al. (2004), zaměřená na porozumění změnám zákaznického chování vzhledem k možné ztrátě i udržení zákazníka. Za tímto účelem jsou zákazníci shlukováni pomocí Kohonenových map; výsledné skupiny jsou anotovány s pomocí rozhodovacích stromů; významné trajektorie přechodů zákazníka mezi shluky v čase jsou identifikovány asociačními pravidly. Kim et al. (2005) využívají analogické řešení, přechody mezi shluky jsou však modelovány Markovovými řetězci s absorpčními stavy.

Zevrubnou analýzu odborné literatury napříč odvětvími předkládají Ahn et al. (2020). Zajímavý je předpoklad růstu prevalence metod založených na hlubokých neuronových sítích, především s ohledem na implicitní extrakci významných atributů i velikost dostupných dat. Delgosha et al. (2020) analyzují abstrakty a klíčová slova výzkumu zaměřeného na modelování podnikových procesů s využitím síťové analýzy a metod pro zpracování přirozeného jazyka. Síťová analýza výskytu slov vedla k identifikaci tří klíčových oblastí, tj. analytických metod, praktických aplikací a tvorby přidané hodnoty. Současně byl popsán růst prevalence využití velkých dat a metod strojového učení. LDA formuje komponenty popisující analýzu sociálních sítí, dodavatelský řetězec, velká data a související infrastrukturu, dále pak dobývání znalostí. Singh et al. (2020) mapují užití strojového učení napříč procesy řízení vztahů se zákazníky; práce poukazuje na rostoucí popularitu metod založených na neuronových sítích. Zdá se, že adopce napříč sektory je hnána potřebou obsloužit rozsáhlou zákaznickou základnu, při vysoké zákaznické spokojenosti i udržitelné profitabilitě.

Predikce ztráty zákazníka bývá vymezena jako binární klasifikace, jejíž cílem je přiřadit pozorování (zákazníky) do jedné ze dvou tříd (ztracení, udržení zákazníci). Pro hodnocení predikčních schopností klasifikačních modelů jsou užívány ukazatele vycházející z matice záměn, výskyt ve vybrané literatuře zachycuje Tab. 2. Míra ACC určuje podíl správně klasifikovaných pozorování, ukazatel je srozumitelný, není však vhodný pro problémy s dominantní třídou. Proto bývá doplňován ukazateli PRE a REC. Pro výpočet těchto ukazatelů je nezbytné stanovit práh mezi třídami, což není případ oblíbeného ukazatele AUCROC. Ekonomickou perspektivu retenčního řízení reflektují práce Coussement & De Bock (2013), Castro & Zsuzuki (2015) a Tamaddoni et al. (2014) a Lee et al. (2020).

K hodnocení efektivity systému slouží odhady výpočetní složitosti, což je významné především s ohledem na praktickou využitelnost daného řešení. Pouze Yu et al. (2011) uvádí alespoň empirický odhad časové náročnosti. Opomíjenou stránkou je i srozumitelnost zachycených znalostí. Interpretace systému strojového učení je adresována především s pomocí

hodnocení významu vysvětlujících proměnných. Song et al. (2004) a Kim et al. (2005) ale demonstrují nezbytnost porozumění modelovanému problému s pomocí transparentních modelů.

2.2.2 Porozumění datovému souboru

Porozumění datovému souboru bývá zahájeno určením faktorů, které řešenou úlohu ovlivňují, následné kroky zahrnují akvizici příslušné datové reprezentace, explorativní analýzu a ověření datové kvality. V rámci vybrané literatury se autor zaměřuje především na definici ztráty zákazníka a prvotní rozpoznání významných proměnných, viz Tab 1.

Tab. 1 Ztráta zákazníka v prostředí e-commerce

Reference	Odvětví	Definice ztráty zákazníka
Abbasi et al. (2015)	multimédia a retail	ukončení smlouvy nebo nedokončení nákupního procesu
Ahn et al. (2020)	přehled literatury	-
Almuqren et al. (2021)	telekomunikace	ukončení smlouvy
Castro & Zsuzuki (2015)	herní	bez návštěvy po časové období
Coussement & De Bock (2013)	hazardní hry	bez návštěvy po časové období
Delgosha et al. (2020)	přehled literatury	-
Ding et al. (2015)	herní	uživatel nenavštíví platformu po 1 měsíc
Gordini & Veglio (2017)	retail	bez transakce během následujícího kalendářního roku
Hengliang & Weiwei (2012)	retail	-
Chen (2016)	služby	nízký počet přístupů k platformě během období
Chou & Chuang (2018)	rezervační služby	zákazník nevytvoří novou rezervaci po 3 měsíce
Kim et al. (2005)	herní	uživatel nenavštíví platformu po 1 měsíc
Lee et al. (2017)	telekomunikace	odliv zákazníků ke konkurenci
Lee et al. (2020)	herní	uživatel nenavštíví platformu
Li & Li (2019)	retail	zákazník nerealizuje transakci po 2 měsíce
Li et al. (2013)	služby	-
Llave Montinel & Lopez (2020)	prodej potravin	uživatel nerealizuje transakci po 1 měsíc
Milosevic et al. (2017)	herní	bez návštěvy po dobu 14 dní
Perisic & Pahor (2021)	herní	více perspektiv
Rachid et al. (2018)	retail	přesun mezi třídami zákaznického chování
Rothmeier et al. (2021)	herní	uživatel platformu nenavštíví
Shih & Fang (2005)	retail	respondent s nízkým zájmem o opakovaný nákup
Singh et al. (2020)	přehled literatury	-
Song et al. (2004)	herní	uživatel nenavštíví platformu po 1 měsíc
Tamaddoni et al. (2014)	-	zákazník nerealizuje transakci po časové období

Reference	Odvětví	Definice ztráty zákazníka
Tsai & Chen (2010)	multimédia	ukončení smlouvy nebo neuhrazení závazků
Venkatesh & Jeyakarthic (2020)	telekomunikace	-
Wang et al. (2019)	bankovnictví a finance	zákazník s nízkou bilancí účtu neinteraguje s platformou
Yu et al. (2011)	retail	-

V mimosmluvním prostředí vystihuje vysvětlovaná proměnná změnu uživatelského chování v budoucím období; zpravidla se jedná o útlum návštěv, což platí především pro odvětví online her a služeb (Song et al., 2004; Kim et al., 2005; Coussement & De Bock, 2013; Castro & Zsuzuki, 2015; Ding et al., 2015; Milosevic et al., 2017; Lee et al., 2020; Rothmeier et al., 2021); v retailu je naopak běžné nahlížet na ztrátu zákazníka skrze transakční chování (Gordini & Veglio; 2017; Li & Li, 2019; Llave Montinel & Lopez, 2020). Délka budoucího období bývá stanovena heuristicky. Smluvní prostředí zastupují platformy poskytující multimédiální obsah (Tsai & Chen, 2010; Abbasi et al., 2015). Některé z prací definici závislé proměnné bohužel neuvádí.

Modelování ztráty zákazníka se opírá především o objektivní (pozorované) nezávislé proměnné. Oblíbené množiny proměnných zahrnují způsob užívání aplikací a služeb nebo transakční historii uživatele. Pro sektor online her je navíc příznačná, podobně jako u mobilních operátorů, reflexe mezilidských vztahů (Ding et al., 2015; Lee et al., 2020). Dále bývá uvažováno o sociálně-demografických (Wang et al., 2019; Almuqren et al., 2021; Llave Montinel & Lopez, 2020) nebo místních odlišnostech mezi uživateli (Chou & Chuang, 2018; Venkatesh & Jeyakarthic, 2020; Llave Montinel & Lopez, 2020). Zdrojem těchto dat bývají informační systémy podniku.

Způsob užívání aplikací a služeb je zastoupen nesourodě; za jednotící prvek považujeme perspektivu uživatelské relace. Zvolená datová reprezentace zpravidla reflektuje perspektivu vysvětlované proměnné, tj. pro odvětví online her a služeb zachycuje návštěvnost a způsob užívání služby. Příkladem muže být Castro & Zsuzuki (2015), kde autoři využívají časově-frekvenční dekompozici denního počtu návštěv. Chen (2016) vychází výhradně z časových intervalů mezi uživatelskými relacemi, a stáří poslední z nich. Lee et al. (2020) uvažují i o délce trvání herní seance. Pro maloobchod je obvyklá reflexe vztahu mezi vlastnostmi uživatelské relace a transakčním chováním. Abbasi et al. (2015) a Rachid et al. (2018) navíc rozlišují chování v jednotlivých krocích nákupu, např. míru opuštění prohlíženého produktu, míru opuštění

produktu přidaného do košíku nebo míru opuštění platby (podíl relací ukončených po přístupu na webovou stránku se shrnutím objednávky a platby).

Transakční historie bývá reprezentována prostřednictvím stáří, frekvence a peněžní hodnoty operací (Recency, Frequency, Monetary – RFM). Stáří představuje dobu uplynulou od poslední transakce; nízké hodnoty naznačují vyšší pravděpodobnost opakovaného nákupu. Frekvence určuje, jak často k transakcím dochází; vyšší hodnoty ukazují spokojeného zákazníka. Peněžní hodnota operací potom odpovídá celkovému objemu prostředků, které zákazník utratí během stanového časového období; vyšší hodnoty naznačují, jaké zákazníky je vhodné udržet (Liu & Shih, 2005). Další užití transakční historie zahrnuje mapování zákaznických preferencí napříč kategoriemi produktů (Gordini & Veglio, 2017; Li & Li, 2019) nebo prevalenci platebních metod (Rachid et al., 2018).

Lidské počínání je podněcováno směsí objektivních (pozorovaných) a subjektivních (vnímaných) faktorů. Mezi subjektivní faktory řadí Brusilovsky (1996) vnímání, názory, předchozí zkušenosti, hodnoty a důvěru. Jejich reprezentaci lze získat z dotazníkových šetření, rozhovorů, sociálních médií, recenzí, nebo jiným uživatelsky generovaným obsahem. Shih & Fang (2005) užívají dotazníkového šetření ověřují vztah mezi atributy zákaznické spokojenosti, plánovaným chování a záměrem opakovaného nákupu. Almuqren et al. (2021) konstruují dílčí odhad zákaznické spokojenosti s pomocí nestrukturovaných textových dat sociální sítě Twitter. Vyspělý přístup k datové reprezentaci modelovaného problému předvádí Abbasi et al. (2015), kde tvůrci uvažují nejen s širokou paletou objektivních a subjektivních faktorů, ale i se strukturou průchodu nákupním procesem společnosti.

2.2.3 Zpracování datového souboru

Zpracování datového souboru pokrývá sadu aktivit, vedoucích k vytvoření finální datové množiny. V praxi se často jedná i mnohokrát opakovaný proces, který zahrnuje čištění, transformace a výběr významných nezávislých proměnných.

První kroky zahrnují redukci, případně imputaci řídkých vysvětlujících proměnných, odstranění neúplných instancí datové množiny, volbu náležitých transformací, eliminaci proměnných s nízkou variabilitou atp. V mezích zkoumané literatury je tento krok přecházen. Nezpracovaná data užívají Tsai & Chen (2010), což se zdá být jednou z příčin nižší predikční schopnosti dopředných neuronových sítí při srovnání s rozhodovacími stromy. Na vině by mohla být pomalá konvergence učení nebo uvíznutí v lokálním extrému. Význam datové reprezentace

vyzdvihuje práce Coussement et al. (2017), ve které autoři konfrontující schopnosti logistické regrese a vhodně zpracovaných vstupních dat s moderními metodami a nezpracovanými daty.

K omezení počtu vysvětlujících proměnných bývá přistupováno pomocí výběru proměnných nebo extrakce vlastností tak, aby výsledná podmnožina proměnných vedla k lepšímu rozlišení vysvětlované proměnné nebo alespoň věrně popisovala vlastnosti původní množiny. Postupy založené na extrakci promítají původní prostor vysvětlujících proměnných do prostoru s nižším počtem dimenzí, což může vést k horší srozumitelnosti. Snížení počtu nezávislých proměnných bývá nejčastěji nezáměrné, tj. jedná se o doprovodný jev klasifikačního algoritmu.

Milosevic et al. (2017), Li & Li (2019) a Rothmeier et al. (2021) využívají k selekci vysvětlujících proměnných nižší vzájemnou korelací. Li & Li (2019) hodnotí důležitost atributů s pomocí statistické významnosti parametrů LR. Tsai & Chen (2010) podobně využívají asociačních pravidel. Rothmeier et al. (2021) vybírají podstatné charakteristiky s pomocí algoritmu náhodných kapradin („random ferns“), jenž jsou typické spíše pro oblast počítačového vidění. Výsledky těchto prací naznačují, že výběr podstatných proměnných může vést k vyšší stabilitě a predikční schopnosti celkového řešení. Příkladem práce zaměřené na extrakci charakteristik může být Castro & Zsuzuki (2015), prokazující užitečnost vlnkové dekompozice do časově-frekvenční domény. Wang et al. (2019) těží dynamické aspekty chování uživatele s pomocí konvolučních vrstev neuronové sítě.

Obvyklým problémem predikce ztráty zákazníka je nevyváženost tříd vysvětlované proměnné, neošetření vede zpravidla ke konstrukci vadného klasifikátoru upřednostňujícího dominantní třídu. Problém lze zmírnit specifickou konstrukcí modelu nebo úpravou datového souboru, druhá z možností je oblíbená díky nezávislosti na doméně úlohy. Úpravu konstrukce modelu ilustruje práce Yu et al. (2011), díky reformulací duální formy SVM. Mezi oblíbené postupy úprav datového souboru náleží podvzorkování nebo převzorkování instancí (Gordini & Veglio, 2017; Milosevic et al., 2017; Lee et al., 2020; Perisic & Pahor, 2021). Rozsáhlou studii vzorkovacích technik a jejich dopadu na modelování ztráty zákazníka předkládají Zhu et al. (2018); závěry práce naznačují potřebu rozdílných přístupů v návaznosti na datovou množinu, klasifikační algoritmus, způsob hodnocení predikční schopnosti řešení.

2.2.4 Modelování

Fáze zahrnuje výběr a aplikaci technik modelování, včetně určení vnějších parametrů modelu. Některé prediktivní modely předpokládají specifické vlastnosti vstupních atributů, což je

jeden z důvodů častých iterací mezi kroky zpracování vstupních dat a modelováním. V následujících odstavcích shrnuje autor užití některých algoritmů a způsoby hodnocení predikční úspěšnosti.

Tab. 2 Aplikace strojového učení pro predikci ztráty zákazníka v prostředí elektronického maloobchodu

Reference	Dělení datového souboru	Ukazatele úspěšnosti	Metody
Abbasi et al. (2015)	trénovací a testovací množiny dat	ACC, PRE, REC, F1	kompozitní SVM, Bayesovské sítě
Ahn et al. (2020)	-	-	-
Almuqren et al. (2021)	trénovací a testovací množiny dat	PRE, REC, F1	Bi-GRU, AraBERT, Transfer learning
Castro & Zsuzuki (2015)	opakované dělení na trénovací a testovací množinu dat	AUCROC, Lift, GAIN	vlnková dekompozice, k-NN
Coussement & De Bock (2013)	trénovací a testovací množiny dat, křížová validace	Lift, GAIN	DT, GAM, Bagging
Delgosha et al. (2020)	-	-	-
Ding et al. (2015)	-	-	Síťová analýza, LR
Gordini & Veglio (2017)	trénovací a testovací množiny dat, křížová validace	ACC, AUC, Lift	LR, SVM, MLP
Hengliang & Weiwei (2012)	-	-	K-means, DT, MLP
Chen (2016)	trénovací a testovací množiny dat	ACC, TPR, FPR, AUCROC	hierarchický Bayesovský model
Chou & Chuang (2018)	trénovací a testovací množiny dat	AUCROC	GAM, Bagging, Boosting
Kim et al. (2005)	trénovací a testovací množiny dat	ACC, FPR, FNR	Kohonenovy mapy, DT, MLP
Lee et al. (2017)	křížová validace	ACC, R^2	LR, DT, MLP, SEM-PLS
Lee et al. (2020)	trénovací, validační a testovací množiny dat	ACC, PRE, REC, F1, AUCROC, Profit	Bagging, Boosting
Li & Li (2019)	trénovací, validační a testovací množiny dat	ACC, PRE, REC	LR, Boosting
Li et al. (2013)	křížová validace	ACC	Lineární programování, DT
Llave Montinel & Lopez (2020)	trénovací a testovací množiny dat	1-ACC, AUCROC	LR, GAM
Milosevic et al. (2017)	křížová validace	PRE, REC, F1, AUCROC	LR, NB, DT, Bagging, Boosting
Perisic & Pahor (2021)	křížová validace	AUCROC	LR, Bagging
Rachid et al. (2018)	křížová validace	ACC, REC, PRE, F1	DT, ANN, Bagging
Rothmeier et al. (2021)	křížová validace	AUCROC	LR, DT, k-NN, SVM, Bagging, Boosting
Shih & Fang (2005)	trénovací a testovací množiny dat	ACC	PCA, DT, MDA, MLP
Singh et al. (2020)	-	-	-
Song et al. (2004)	trénovací a testovací množiny dat	ACC, PRE, REC	Kohonenovy mapy, Apriori, MLP
Tamaddoni et al. (2014)	křížová validace	Lift, Profit	Pareto/NBD, LR, SVM, Boosting
Tsai & Chen (2010)	trénovací, validační a testovací množiny dat	ACC, PRE, REC, F1	Apriori, DT, MLP

Reference	Dělení datového souboru	Ukazatele úspěšnosti	Metody
Venkatesh & Jeyakarthic (2020)	trénovací data	ACC, PRE, REC, F1, Kappa	LR, DT, SVM, NB, MLP, LSTM, Bagging, Boosting, OGA, EHO
Wang et al. (2019)	trénovací, validační a testovací množiny dat	ACC, AUC, TDL	LR, SVM, RF, MLP, CNN
Yu et al. (2011)	trénovací a testovací množiny dat	ACC, PRE, REC, Lift	DT, SVM, MLP

Dělení datového souboru

Přístup k dělení datového souboru by měl reflektovat podmínky aplikace modelu v reálném prostředí, užité přístupy jsou prezentovány v Tab. 2. Populární je rozlišení trénovacích a testovacích množin, případně křížová validace. Mezi neduhy lze řadit kombinaci nedostatečně velkého výchozího souboru dat a prosté oddělení množin, což vede k nestabilním odhadům. Obvykle absentuje i časové rozlišení. K selekci vnějších parametrů a konečnému posouzení predikčních možností modelu dochází s využitím jedné množiny dat. Důsledkem uvedených nedostatků bývá přecenění schopností navržených řešení.

Časové rozlišení využívají při konstrukci potřebných množin práce Gordini & Veglio (2017), Tsai & Chen (2010) a Coussement & De Bock (2013). V rámci počátečního časového období dochází ke konstrukci modelů, jejichž predikční schopnosti jsou následně ověřeny s pomocí dat ze stávajícího i budoucího období. U všech dokumentů pozorujeme nárůst chyb spojený s novými daty, pořadí modelů se však zásadně nemění. Zdá se tedy, že z pro některé dílčí kroky jako výběr důležitých proměnných nebo nastavení parametrů modelu, je možné využít data z úvodního časového období.

Rothmeier et al. (2021) příkladně využívají časově odlišenou křížovou validaci pro manuální výběr vnějších parametrů modelu i ověření dosažených výsledků na datové sadě z následného období. Zhoršení predikčních schopností ověřené na nových datech je nižší, tj. časové odlišení v rámci křížové validace vede k rozumnému odhadu schopností modelů. Dosažené výsledky jsou stabilní a dostatečně podložené. Vybočuje také práce Milosevic et al. (2017), kde autoři určí vhodný model s pomocí křížové validace bez časového rozlišení, celková úspěšnost navrženého řešení je však vhodně hodnocena s pomocí konverzních schopností realizované retenční kampaně.

Ukazatele úspěšnosti

Ve vybrané literatuře převládá hodnocení klasifikačních schopností pomocí ukazatelů založených na matici záměn. Oblíbená je především kombinace ACC, PRE, REC, jenž umožňuje srozumitelnou interpretaci predikce ztráty zákazníka i pro datové sady s nevyváženými třídami. Další populární metrikou je AUCROC, jenž na rozdíl od předchozích metrik, pracuje přímo s predikovanou pravděpodobností příslušnosti k dané třídě. Perspektiva blízká praktickému užití modelu je prezentovaná ukazatelem Lift, jenž zohledňuje přesnost i kalibraci klasifikátoru. Většina textů se soustředí na hodnocení predikční způsobilosti s pomocí přirozených klasifikačních měr, což bohužel nereflektuje podnikový kontext modelovaného problému.

Coussement & De Bock (2013) posuzují úspěšnost jednotlivých klasifikátorů s pomocí rozšíření konceptu metriky Lift, směrem k očekávané hodnotě životního cyklu zákazníka. Nový koncept vychází z popisu dynamiky retenčních nákladů a výnosů představené Neslin et al. (2006). Autoři hodnotí očekávané zlepšení profitability pro různé hodnoty dílčích parametrů jako jsou průměrná hodnota životního cyklu zákazníka, případně úspěšnost retenční kampaně. Podobný přístup k hodnocení ekonomické úspěšnosti klasifikace prezentují Castro & Zsuzuki (2015), kde však nedochází k dílčím odhadům jednotlivých parametrů. Lee et al. (2020) navazují na obě uvedené práce, naproti nim však zavádějí individuální úroveň budoucí hodnoty životního cyklu zákazníka. Hlavním přínosem je však určení prahu mezi třídami tak, aby klasifikace zákazníků současně maximalizovala očekávaný zisk retenční kampaně. Tamaddoni et al. (2014) rozšiřují pohled předchozích prací s pomocí konzervativního odhadu budoucí hodnoty zákazníka a zavedením nejistoty simulací pravděpodobnosti přijetí retenční nabídky. Související citlivostní analýza poukazuje především na význam střední hodnoty simulovaného rozdělení pravděpodobnosti.

Konceptuálně problematické jsou metody odhadu hodnoty zákaníka pro hodnocení klasifikačních modelů, které očekávanou budoucí hodnotu nediferencují napříč zákaznickou bází (Coussement & De Bock, 2013; Castro & Zsuzuki, 2015), případně přejímají historickou individuální úroveň zákaznické hodnoty (Tamaddoni et al., 2014; Lee et al., 2020). Dalším omezením je deterministický pohled na jednotlivé parametry, kde pouze Tamaddoni et al. (2014) využívají simulační prostředky k odhadu dopadů nahodilých změn v úrovni pravděpodobnosti přijetí retenční nabídky, včetně analýzy citlivosti. Z pohledu na dílčí kroky v procesu modelování je zřejmé, že ekonomická perspektiva je využita výhradně při hodnocení daného řešení, nikoliv při jeho konstrukci. V tomto ohledu je nejvyspělejší prací Lee et al. (2020),

kde autoři navrhují stanovit hranice mezi třídami tak, aby došlo k maximalizaci očekávaného zisku retenční kampaně. Předchozí dílčí kroky jako výběr proměnných nebo konstrukce klasifikátoru však zůstávají netknuty.

Klasifikační metody

Logistická regrese bývá používána jako výchozí klasifikační model, s kterým jsou často prezentované techniky srovnávány. Mezi hlavní nedostatky aplikací náleží opomíjení předpokladů a požadavků na použití metody, což dobře ilustruje práce Coussement et al. (2017). Využití LR za účelem výběru význačných proměnných, s pomocí statistické významnosti koeficientů regresního modelu, demonstrují Li & Li (2019).

Dalším referenčním modelem jsou rozhodovací stromy; bývají využity samostatně, nebo jako základní metoda meta–algorimů. Tsai & Chen (2010) nepřímo prokazují sílu algoritmu při aplikaci na nezpracovaný soubor dat. Kim et al. (2005) a Hengliang & Weiwei (2012) se naopak soustředí na srozumitelnost metody, a popisují s pomocí rozhodovacích stromů společné vlastnosti zákaznických shluků.

Klasifikační schopnosti metody podpůrných vektorů dokládá práce Gordini & Veglio (2017), kde autoři hledají vhodné parametry jádrové funkce a penalizace chyb s pomocí mřížkové optimalizace a různými účelovými funkcemi. Zajímavý je úspěch optimalizace směrem k AUCROC. Predikce generované algoritmem se totiž sestávají pouze z výsledné třídy. Pro odhad míry příslušnosti k třídě, nezbytný pro AUCROC, je třeba zkonstruovat soubor modelů a následně pravděpodobnost kalibrovat, což text zcela opomíjí. Yu et al. (2011) navrhují reformulaci duální formy SVM pomocí problému nejmenší opsané koule; přístup byl původně představen v textu Tsang et al. (2005). Modifikace zajišťuje dostatečnou predikční úspěšnost, spolu s lineární výpočetní a konstantní prostorovou složitostí, což umožňuje aplikaci metody na rozsáhlé datové soubory. Abbasi et al. (2015) zavádějí složenou hierarchickou jádrovou funkci algoritmu, která umožňuje lépe reflektovat očekávanou strukturu zákaznického chování.

Současné klasifikační metody bývají ve vědeckých pracích reprezentovány dopřednými vícevrstvými neuronovými sítěmi. V rámci zkoumané literární oblasti jsou aplikace v zásadě referenční, bez podrobného popisu topologie, aktivačních funkcí, způsobu učení, optimalizace uvedených parametrů atp. V posledních letech pozorujeme práce rozmach prací zaměřených na hluboké neuronové sítě. Wang et al. (2019) využívají dvou konvolučních vrstev pro hierarchickou extrakci jádrových funkcí, popisujících dynamiku uživatelského chování. Výstupy jsou

skrze sdružovací vrstvu propagovány do plně propojené vrstvy, jenž je sdílená se statickými vysvětlujícími proměnnými. Klasifikační hlava sítě se sestává z další plně propojené a výstupní vrstvy. Řešení je inspirováno architekturami počítačového vidění. Venkatesh & Jeyakarthic (2020) optimalizují vnější parametry obousměrné rekurentní neuronové sítě s pomocí algoritmu sloních stád, za tímto účelem využívají statický tabulární soubor dat, což není v souladu s charakterem použité sítě. Almuqren et al. (2021) kombinují obousměrné rekurentní neuronové sítě s jazykovým modelem AraBERT pro predikci zákaznické spokojenosti. Vertikální šíření signálu v rekurentních sítích umožňuje zachytit vztahy mezi sekvencí pozorování, což nachází uplatněné především při predikci časových řad, rozpoznávání řeči, zpracování přirozeného jazyka atp.

Z hlediska klasifikační úspěšnosti jednoznačně dominují moderní meta-algoritmy. Coussement & De Bock (2013) demonstrují užitečnost ansámblu modelů při srovnání s dílčím modelem rozhodovacího stromu nebo zobecněným aditivním modelem. Rachid et al. (2018) předestírají komparaci klasifikátorů založených na rozhodovacích stromech, neuronových sítích a ansámblech. Rothmeier et al. (2021) hodnotí klasifikační schopnosti mnoha oblíbených algoritmů s ohledem na různé netransakční definice ztráty zákazníka. Coussement & De Bock (2013), Rachid et al. (2018) i Rothmeier et al. (2021) shodně označují ansámbly modelů konstruované metodou bagging (např. náhodné lesy) za nejvíce přesné. Tamaddoni et al. (2014) srovnávají predikčních schopnosti pravděpodobnostních modelů, logistické regrese, podpůrných vektorů a metody boosting napříč simulovanými vlastnostmi souboru dat jako jsou velikost vzorku, nákupní chování zákazníka nebo rozložení vysvětlované proměnné. Snahy Milosevic et al. (2017) míří na predikci ztráty zákazníka krátce po založení uživatelského účtu v rámci herní online platformy. Za tímto účelem porovnávají autoři značnou část oblíbených metod. Tamaddoni et al. (2014) i Milosevic et al. (2017) prokazují predikční schopnosti meta-algoritmu boosting v různých situacích.

Vhodné nastavení vnějších parametrů modelu je spíše opomíjeno, případně k němu autoři přistupují manuálně (Song et al., 2004; Castro & Zsuzuki, 2015; Rothmeier et al., 2021). Úplné prohledání kartézského součinu hyperparametrů využívají práce Gordini & Veglio (2017) a Perisic & Pahor (2021). Omezení takového přístupu tkví především v rychlosti růstu počtu prvků v prohledávaném prostoru i neefektivním využití výpočetních prostředků. Informované prohledávání prostoru s pomocí algoritmu sloních stád představují Venkatesh & Jeyakarthic (2020). Optimalizace je založená na iterativním pohybu slonů (konkrétní kombinace parametrů) dle vedoucí samice klanu (nejlepší kombinace parametrů klanu), kde část dospělých mužských

samců (nejslabší kombinace parametrů) na konci iterace klan opouští a je v další iteraci nahrazena novými slony (náhodně vygenerované kombinace parametrů, které leží v prostoru klanu). Popsanou metodu je možné použít výhradně pro spojitý prostor vnějších parametrů, což není v souladu s prezentovanou aplikací, která trpí i dalšími neduhy.

2.2.5 Vyhodnocení a interpretace

Ve rámci hodnocení a interpretace dochází k ověření kandidátních řešení, která se zdají být vhodná z pohledu úspěšnosti. Východiska jsou zasazena do perspektivy organizace a je vytyčeno budoucí směřování projektu. Autor v rámci podkapitoly člení poznatky dle zaměření studované literatury na podnikové hledisko problému, a témata dalšího výzkumu.

Úspěšnost řešení je hodnocena zpravidla přirozenými ukazateli klasifikace, jenž neuvažují ekonomické důsledky retenčních aktivit. Výjimkou budiž Coussement & De Bock (2013) a Castro & Zsuzuki (2015), kteří hodnotí kýžený dopad s pomocí dynamiky retenčních nákladů a výnosů. Limitem přístupu autorů je zavedení průměrné očekávané hodnoty zákazníka. Tamaddoni et al. (2014) a Lee et al. (2020) proto představují individuální úrovně dané veličiny. Závěry uvedených prací dobře ilustrují nesoulad mezi obvyklým přístupem k tvorbě, hodnocení a využití prediktivních řešení.

Pro bližší porozumění modelovanému fenoménu bývá využito analýzy vysvětlujících proměnných, na které se vybrané modely spoléhají. V odvětví online her a služeb se prosazují především faktory popisující návštěvnost, případně způsob užívání služeb. V retailu je kladen důraz na různé aspekty interakcí transakčních. Tuto dichotomii dobře ilustrují práce Chen (2016) a Li & Li (2019), jejichž řešení lpí především na vysvětlujících proměnných popisujících frekvenci a stáří uživatelských interakcí. Rothmeier et al. (2021) a Perisic & Pahor (2021) shodně poukazují i na důležitost některých specifických aktivit v rámci zkoumaných herních platforem. Abbasi et al. (2015) a Rachid et al. (2018) nepřímo popisují podobný fenomén v maloobchodním prostředí, tj. význam sběru dat o průchodu zákazníka nákupním procesem. Možný přínos místních odlišností ilustrují práce Chou & Chuang (2018) a Llave Montinel & Lopez (2020). Rothmeier et al. (2021) se naopak vztah mezi lokalitou uživatele a ztrátou zákazníka nedaří prokázat. Předpokládáme, že se v tomto případě jedná o odlišnost mezi příslušnými odvětvími. Dopady subjektivních faktorů jsou bohužel zkoumány na malých souborech dat. Závěry Shih & Fang (2005) a Hengliang & Weiwei (2012) naznačují možný význam vnímání organizace, spolehlivosti služby a cenové dostupnosti produktů. Abbasi et al. (2015)

bohužel určují důležitost proměnných dle množin prvků, není tedy možné hodnotit přispění individuálních subjektivních proměnných.

Song et al. (2004) a Kim et al. (2005) kontrastují sousedící shluky setrvávajících a ohrožených uživatelů, čímž je možné odhalit problematické faktory a na ty potom reagovat v rámci retenční kampaně. Pokud rizikový shluk pokulhává v průměrné době herní seance, je například možné navrhnout systém generování odměn ve formě herních předmětů v závislosti na odehraném čase. Případným problémem přístupu je interpretace odlišností mezi shluky ve vícerozměrném prostoru, případně mezi různými typy vysvětlujících proměnných. Obě práce také neověřují životaschopnost daného řešení na dostatečně rozsáhlých datových sadách.

Předestřený budoucí výzkum se zaměřuje zejména na další odvětví. V rámci procesu prediktivního modelování bývají mezi oblastmi dalšího zkoumání uváděny nové datové sady i vysvětlující proměnné, nové klasifikační metody nebo přístupy k optimalizaci parametrů modelu. Tsai & Chen (2010), Tamaddoni et al. (2014) a Milosevic et al. (2017) navíc vhodně zasazují další výzkum do kontextu retenčních činností podniku, tj. doporučují bližší porozumění fenoménu ztráty zákazníka, které pak lze reflektovat v návrhu kampaně. Úvahy o ztrátě zákazníka je možné doplnit o předpokládané dopady retenčních aktivit.

2.2.6 Aplikace řešení

Projekt obvykle není uzavřen vytvořením prediktivního modelu, ale formováním a přístupnou prezentací znalostí, které jsou využívány v konkrétních procesech podniku. V závislosti na typu datového produktu potom dochází k realizaci řešení, kde výstupem může být text vědeckého článku, ale i komplexní aplikace.

Na tomto místě je třeba vyzdvihnout práci Milosevic et al. (2017), kde je úspěšnost daného řešení hodnocena s pomocí A/B testu konverzních schopností realizované retenční kampaně. Autoři uvádějí, že se danou aktivitu podařilo zcela automatizovat, což je ojedinělé. Převládajícím neduhem je potom absence důrazu na transparenci výzkumu. Texty využívají pouze privátních souborů dat, výjimkou je Coussement & De Bock (2013). Doprovodný kód není součástí žádné z prací.

2.2.7 Shrnutí a diskuse

V rámci studované literatury rozlišujeme texty dle zaměření na podnikovou perspektivu problému, modelování a strojové učení, nebo analýzu odborné literatury. Podnikovou

perspektivou autor rozumí snahu adresovat otázky zákaznické retence nad rámec identifikace rizikových zákazníků. Je s podivem, že této perspektivě není věnováno více pozornosti.

Z hlediska definice ztráty zákazníka a prvotního rozpoznání významných faktorů je možné pozorovat dichotomii napříč odvětvími, kde se online hry a služby zaměřují na změny v netransakčním chování uživatele, v retailu je naopak v centru pozornosti chování transakční. Zajímavá je nízká prevalence faktorů popisujících firemní aspekty ztráty zákazníka jako je úroveň služby. Některé práce využívají i subjektivních vysvětlujících proměnných, komplikací však bývá nákladný sběr dat a nedostatečné pokrytí zákaznické báze. Podceňovanou partií je zpracování vstupních dat. Výběr důležitých faktorů bývá realizován prostřednictvím odhadu vzájemné korelace s vysvětlovanou proměnnou, nebo je doprovodným jevem použitého klasifikačního algoritmu. Nerovnoměrné zastoupení cílových tříd je zpravidla adresováno vzorkováním.

Systémy strojového učení jsou hodnoceny pomocí trénovacích a testovacích množin dat, případně s využitím křížové validace. Značná část autorů neuvažuje časové rozlišení, což vede k příliš optimistickým odhadům klasifikačních schopností. Posouzení systému je obvykle zpracováno s pomocí ukazatelů vycházejících z matice záměn, kontext retenčních aktivit je reflektován zřídka. V rámci úspěšných klasifikačních technik pozorujeme postupující odklon od tradičních přístupů strojového učení k metodám založeným na meta-algoritmech nebo hlubokých neuronových sítích. K systematické optimalizaci vnějších parametrů modelů běžně autoři nepřistupují.

Pro bližší pochopení ztráty zákazníka dochází k dalšímu zkoumání nezávislých faktorů, na které představená řešení spoléhají, využití transparentních modelů je však na okraji zájmu. Zdá se, že v sektoru online her a služeb vynikají faktory zachycující návštěvnost nebo způsob užívání služeb, v maloobchodu naopak převládají faktory transakční. Další způsob reflexe podnikového kontextu modelovaného problému je možné demonstrovat pomocí hodnocení modelu s využitím ekonomické reality retenčních aktivit, takový přístup je bohužel ojedinělý. Kroky budoucího výzkumu většiny prací směřují k novým odvětvím, souborům dat a metodám.

Aplikací řešení lze ve studovaném kontextu rozumět ověření předestřených návrhů na dostupném souboru dat a prezentace vědecké práce formou článku. Problematická je v tomto ohledu transparentnost výzkumu, texty nedoprovází veřejné sady dat ani programový kód.

Komparace dosažených závěrů s existující literaturou je možná s pomocí přehledových článků. Ahn et al. (2020) prezentují zevrubnou analýzu vědeckých prací napříč odvětvími, která

je v úrovni detailu nejbližší představené rešerši. Za významné způsoby hodnocení ekonomických dopadů modelu považují autoři perspektivu nákladů na akvizici a perspektivu očekávané celkové hodnoty životního cyklu zákazníka. V rámci metod pozorují příklon k hlubokým neuronovým sítím, což je způsobeno menšími nároky na konstrukci specifických vysvětlujících proměnných a rostoucí objemy dat. Delgosha et al. (2020) se soustředí na výzkum v oblasti podnikové analýzy, kde nejprve s pomocí síťového přístupu k textu identifikují shluky analytických metod, praktických aplikací a tvorby přidané hodnoty. Následně užívají LDA pro odhalení skrytých faktorů, které anotují jako sociální sítě, dodavatelský řetězec, velká data a infrastruktura a dobývání znalostí. Singh et al. (2020) se zabývají aplikacemi strojového učení při identifikaci, akvizici, udržení a rozvoji zákazníků. Popularita hlubokých neuronových sítích je identifikována i v této práci.

Při zasazení do širšího výzkumu ztráty zákazníka pozorujeme rozdíly mezi odvětvími i užší vazbu na aplikace prediktivního modelování. Prostředí podniku je spíše upozaděno, což se odráží i v souvislosti s ekonomickou realitou retenčních aktivit nebo srozumitelnosti řešení. Z hlediska jednotlivých kroků prediktivního modelování odpovídají texty směřování vědecké domény, sporné jsou však přístupy k návrhu a realizaci experimentů nebo k transparenci a reprodukovatelnosti výzkumu. Omezení předestřených závěrů vychází především z definice oblasti výzkumného zájmu, struktury využité pro analýzu, případně z úrovně detailu.

3 Cíle práce a užité metody

3.1 Cíle práce

s. Hlavní cíl bude naplněn prostřednictvím cílů dílčích.

Dílčí cíle disertační práce:

Dílčí cíl 1: Popsat teoretická východiska, zahrnující prostředí elektronického maloobchodu, problematiku řízení vztahů se zákazníky, a strojové učení.

Dílčí cíl 2: Zanalyzovat současné poznatky v oblasti predikce ztráty zákazníka s využitím metod výpočetní lingvistiky i tradiční rešerše.

Dílčí cíl 3: Navrhnout a vytvořit systém strojového učení, zaměřený na předpověď odchodu zákazníka v prostředí elektronického maloobchodu v intencích vymezených hlavním cílem práce.

Dílčí cíl 4: Zhodnotit schopnosti navrženého systému strojového učení, včetně interpretace zachycených znalostí.

3.2 Výzkumné otázky

Zajímavé aspekty řešených problémů jsou v rámci vybraných dílčích cílů dále rozpracovány s pomocí výzkumných otázek. Vztah mezi dílčími cíli a výzkumnými otázkami ilustruje Tab. 3.

Tab. 3 Struktura cílů a výzkumných otázek disertační práce

Hlavní cíl	Dílčí cíle	Výzkumné otázky
Návrh a implementace systému strojového učení, který bude předpovídat odchod zákazníka v prostředí elektronického maloobchodu. Řešení bude reflektovat podnikový kontext problému, čímž autor rozumí ekonomický dopad uvažované retenční kampaně a srozumitelnost zachycených znalostí.	DC 1: Popsat teoretická východiska, zahrnující prostředí elektronického maloobchodu, problematiku řízení vztahů se zákazníky, a strojové učení.	
	DC 2: Zanalyzovat současné poznání v oblasti predikce ztráty zákazníka s využitím metod výpočetní lingvistiky i tradiční rešerše.	VO1: Jaké jsou výzkumné mezery současného poznání v oblasti predikce ztráty zákazníka v daném kontextu?
	DC3: Navrhnout a vytvořit systém strojového učení, zaměřený na předpověď odchodu zákazníka v prostředí elektronického maloobchodu v intencích vymezených hlavním cílem práce.	
	DC4: Zhodnotit schopnosti navrženého systému strojového učení, včetně interpretace zachycených znalostí.	VO2: Jaké třídy modelů vedou k lepším predikčním schopnostem řešení?
		VO3: Jaké třídy modelů vedou k lepším ekonomickým výsledkům retenční kampaně?
		VO4: Jaké vysvětlující proměnné jsou klíčové pro predikci modelů?
		VO5: Jaké společné znaky vykazují zákazníci, na které je vhodné retenční aktivity cílit?

Význam úvodní výzkumné otázky VO1, tedy určení slepých skvrn v oblasti predikce ztráty zákazníka v prostředí elektronického maloobchodu, vyplývá z potřeby identifikace takových částí problému, jejichž zkoumání posune existující stav poznání, a současně povede k vývoji účinnějších přístupů zákaznické retence. Odpovědi pomohly vymezit směřování, respektive cíle disertační práce.

Další výzkumné otázky jsou orientovány na vybrané aspekty zamýšleného systému strojového učení. VO2 a VO3 směřují k hodnocení tříd modelů určených k predikci odchodu zákazníka z hlediska prediktivních schopností, respektive ekonomických výsledků souvisejících retenčních aktivit. Bude tak možné porozumět přesnosti a spolehlivosti řešení, případně nahlížet

na rozdíly mezi nastíněnými hledisky, což může podpořit výběr modelů pro budoucí aplikace v prostředí podniku a další směřování výzkumného úsilí. Nasazení, rozvoj a výzkum účinných a spolehlivých modelů povede v konečném důsledku k efektivnější realizaci retenčních aktivit.

Zbývající výzkumné otázky VO4 a VO5 cílí na porozumění zákaznickému chování prostřednictvím vybraných modelů strojového učení. Smyslem počínání je srozumitelnost jednotlivých přístupů, způsobů tvorby predikcí, a souvisejících limitů. Zachycené znalosti je možné využít k interpretaci zákaznického chování, ať už z hlediska obecného charakteru vztahů k modelovanému jevu, tak z hlediska dílčích segmentů. Bude tak možné porozumět chování zákazníků i prediktivním modelům, což dále podpoří výběr a konstrukci vysvětlujících proměnných pro zamýšlené aplikace v prostředí podniku, zacílení retenčního úsilí, ale i budoucí směřování výzkumu.

3.3 Užité metody

3.3.1 Metody vědeckého zkoumání

Empirické metody

Molnár et al. (2012) vymezuje empirické metody jako metody založené na přímém pozorování a měření reality. Takové přístupy zahrnují záznam a vnímání jevů prostřednictvím různých úrovní vnímání, díky čemuž je možné identifikovat specifické a jedinečné vlastnosti objektu nebo jevu v realitě. Pro práci významná je především podmnožina metod experimentálních, která umožňuje systematický a kontrolovaný přístup k nastíněným problémům.

Logické metody

Časté je užití párových metod jako jsou indikce a dedukce, analýza a syntéza, abstrakce a konkretizace. Vlastní aplikace metod se pojí s prokazováním platnosti zvolených hypotéz pomocí empiricky získaných poznatků (Bryman, 2012).

Indukce značí postup od konkrétního k obecnému, *dedukce* potom od obecného ke konkrétnímu (Kumar, 2019). Příkladem užití indukce v rámci disertační práce může být vyhodnocení prediktivních schopností systému, kde na základě pozorované úrovně ukazatelů formujeme závěry o schopnostech jednotlivých tříd modelů. Jako dedukci můžeme označit soubor doporučení plynoucích z relevantní vědecké literatury, který byl brán při konstrukci výsledného systému strojového učení v potaz.

Analýza představuje logickou metodu využívanou k rozložení logického celku na dílčí prvky a zkoumání vazeb a vlastností. Související metodou je *syntéza*, která individuální prvky kreativně skládá, transformuje zpět do nového celku (Kumar, 2019). Příkladem užití této párové metody v rámci disertační práce může být literární rešerše, kde zkoumáme dílčí aspekty jednotlivých prací, které pak organizujeme do rozpoznaných vzorů a trendů.

Abstrakce odděluje nepodstatné atributy úkazu tak, aby byly uvažovány pouze zásadní charakteristiky úkazů a objektů, *konkretizace* naopak aplikuje charakteristiky třídy jevů na jev konkrétní (Molnár, 2020). Ukázkou využití abstrakce může být vymezení zákaznických shluků a jejich charakteristik v rámci interpretace modelu strojového učení, příkladem konkretizace může být snaha o porozumění shluku prostřednictvím individuálního pozorování.

Počítačové modely, simulace a experimenty

Molnár et al. (2012) řadí počítačové modely, simulace a experimenty mezi nejvýznamnější metody vědeckého zkoumání, které staví na vlastních objevech, rozvíjí tvořivost a uvažování.

Počítačové modely zahrnují vytvoření matematické reprezentace systému nebo procesu a využití počítačových prostředků k simulaci jeho chován, což umožňuje studovat systém nebo proces v kontrolovaném prostředí, a předvídat chování ve skutečném světě. Využití počítačových modelů v rámci disertační práce budiž ilustrováno navrženými a implementovanými systémy strojového učení.

Počítačové simulace oproti počítačovým modelům staví na konstrukci umělého prostředí reflektujícího realitu zkoumaného fenoménu, díky čemuž je možné testovat hypotézy, studovat složité systémy nebo ojedinělé jevy. V kontextu představené práce jsou simulace využívány k odhadu chování různých instancí datového souboru, a to jak ve fázi konstrukce datové reprezentace zákazníka, tak při interpretaci prediktivních modelů.

Počítačové experimenty zpravidla zahrnují manipulaci s jednou nebo více proměnnými v kontrolovaném prostředí a pozorování dosažených výsledků, což dovoluje testovat hypotézy a vyvozovat závěry o vztahu mezi proměnnými. S touto problematikou se zaobírají kapitoly věnované sestavení datové reprezentace zákazníka a modelování, těží z ní ale i kapitoly zhodnocení a interpretace řešených systémů.

3.3.2 Matematická statistika

Yeo-Johnsonova mocninná transformace je zobecněním tradiční Box-Cox transformace, jejíž definiční obor odpovídá kladné části reálných čísel. Yeo & Johnson (2000) zavádějí novou rodinu mocninných transformací, pro kterou toto omezení neplatí, a současně disponuje užitečnými vlastnostmi Box-Cox transformací. Transformace jsou definovány jako

$$\psi(\lambda, y) = \begin{cases} \frac{(y+1)^{\lambda} - 1}{\lambda} & if\ \lambda \neq 0, y \geq 0 \\ \log(y+1) & if\ \lambda = 0, y \geq 0 \\ -\frac{(-y+1)^{2-\lambda} - 1}{2-\lambda} & if\ \lambda \neq 2, y < 0 \\ -\log(-y+1) & if\ \lambda = 2, y < 0 \end{cases} \tag{31}$$

Pokud je y striktně pozitivní, pak transformace odpovídá Box-Cox transformaci pro $(y + 1)$, pokud je y striktně negativní, pak transformace odpovídá Box-Cox transformaci $(-y + 1)^{2-\lambda}$. Pokud y nabývá pozitivních i negativních hodnot, pak je výsledné rozdělení směsí obou transformací, tj. pozitivní a negativní hodnoty jsou transformované s pomocí různých mocnin, což vede ke komplikovanější interpretaci parametru λ.

Wilcoxonův test je využíván především jako neparametrická alternativa k párovému Studentovu t-testu, kde distribuce párových rozdílů mezi veličinami nemusí sledovat normální rozdělení. V rámci nulové hypotézy očekáváme, že párové rozdíly jsou symetrické okolo nuly, alternativní hypotéza předpokládá rozdíly nesymetrické kolem nuly, případně směr odlišnosti. Mějme párové rozdíly $X_1, \dots, X_n$, a vzestupné pořadí absolutních hodnot těchto rozdílů $R_1, \dots, R_n$, pak testovací statistiku T vypočteme následovně.

$$T = \sum_{i=1}^{n} sign(X_i)\, R_i \tag{32}$$

$$sign(x) = \begin{cases} -1 & if\ x < 0 \\ 0 & if\ x = 0 \\ 1 & if\ x > 0 \end{cases} \tag{33}$$

K testovací statistice T pak, ze srovnání s rozložením veličiny při platnosti nulové hypotézy, získáme odpovídající p-hodnotu, kterou dále porovnáváme s předem určenou hladinou

významnosti. Dalším praktickým aspektům testu jako jsou nulové a stejné hodnoty rozdílů, nebo určení rozložení testovací statistiky se zevrubně věnují Pratt & Gibons (1981).

3.3.3 Strojové učení

Shlukování

Hierarchické aglomerativní shlukování (Hierarchical Agglomerative Clustering) je založené na párovém porovnání podobnosti mezi shluky, kde podobnost je určena mírou vzdálenosti a způsobem porovnání vzdálenosti skupinami instancí. V úvodní iteraci je každá instance datového souboru považována za individuální shluk, následně jsou tyto shluky postupně slučovány, od nejvyšší podobnosti, až je nakonec vytvořen jeden shluk obsahující všechny instance datového souboru. Přístup bývá reprezentován jako neorientovaný strom (Hastie et al., 2009). Hierarchické aglomerativní shlukování je oblíbené s ohledem na přímočarost, flexibilitu a možnosti vizuální reprezentace. Mezi problematické aspekty lze řadit výpočetní složitost nebo nutnost určit způsob porovnání vzdálenosti mezi shluky a související citlivost.

Doporučující systémy

Doporučující systémy zahrnují souhrn přístupů, jejímž cílem je generovat užitečná doporučení pro uživatele, respektive zákazníka. Příkladem takového doporučení může být jaké další zboží koupit, jaký film shlédnout nebo jaký vědecký článek přečíst.

Kolaborativní filtrování (Collaborative filtering) je nejrozšířenějším pojetím tvorby doporučení a vychází z předpokladu, že užitečný návrh další interakce by měl vycházet z historie interakcí uživatelů s podobnými preferencemi (Aggarwal, 2016). Interakce rozlišujeme na explicitní a implicitní, kde explicitní zpětnou vazbou rozumíme otevřeně komunikované preference, implicitní hodnocení reflektuje upřednostnění prvků skrze pozorované chování. Doporučení další interakce může být výstupem metody nejbližších sousedů, využívající podobnost mezi uživateli, případně prvky interakce, jako míru vzdálenosti. S ohledem na nízkou hustotu bývá matice interakcí uživatele s příslušnými prvky rozložena do sdíleného latentního podprostoru. Pro konstrukci latentního podprostoru bývá zpravidla užíváno přístupů vycházejících z faktorizace matic, nebo hlubokých neuronových sítí.

Nezáporná faktorizace matic (Non-negative Matrix Factorization) označuje přístupy k rozkladu matic s nezápornými prvky. Mějme matici $V \in \mathbb{R}_{+}^{n \times m}$ a hodnost $k \leq \min(m, n)$, pak

hledáme takové matice $W \in \mathbb{R}_{+}^{n \times k}$ a $H \in \mathbb{R}_{+}^{k \times m}$, jejichž součin přibližně rekonstruuje původní matici V, tj.

$$V \approx WH. \tag{34}$$

Klíčový je způsob hodnocení podobnosti původní matice V a aproximované matice $\hat{V} = WH$, mezi oblíbené přístupy k hodnocení podobností patří Frobeniova norma nebo zobecněná Kullback-Leiber divergence (Lee & Seung, 2000). V doporučujících systémech bývá úloha řešena s pomocí metody střídajících se nejmenších čtverců (Ricci et al., 2011), především díky efektivnímu přístupu k řídkým datovým souborům, stabilním implementacím a dobré výpočetní složitosti. Mezi neduhy lze řadit předpoklad užitečnosti projekce do nízko rozměrného podprostoru, citlivost na vnější parametry modelu, případně sklon k přeučení u více dimenzionálních latentních podprostorů.

Optimalizace vnějších parametrů modelu

Optimalizací vnějších parametrů modelu rozumíme proces určení vhodných parametrů vzhledem k vybrané účelové funkci a datovému souboru, kde vnější parametry kontrolují proces konstrukce a učení modelu. K odhadu vnitřních parametrů dochází v procesu učení. Výhody programatického řešení optimalizace vnějších parametrů modelu shrnují Feurer & Hutter (2019) jako snížení potřeby lidského zásahu, zlepšení úspešnosti systému strojového učení skrz přizpůsobení danému problému, a zlepšení transparentnosti v rámci studií různých přístupů strojového učení. Nejrozšířenější přístupy k optimalizaci zahrnují úplné hledání, náhodné hledání, Bayesovské metody, evoluční metody, aj.

Bayesovská optimalizace (Bayesian Optimization – BO) je sekvenční strategie globálního prohledávání, pohlíží na účelovou funkci jako na závislou proměnnou a zachycuje současnou představu o jejím chování s pomocí apriorního rozdělení pravděpodobnosti. Při ohodnocení nové instance prohledávaného prostoru dochází k aktualizaci apriorního rozdělení pravděpodobnosti na posteriorní rozdělení pravděpodobnosti závislé proměnné, kde dané mapování zajišťuje proxy model. Posteriorní rozdělení pravděpodobnosti a akviziční funkce pak určují další instanci prohledávaného prostoru k ohodnocení (Bergstra et al., 2013; Shahriari et al, 2016). Oblíbené proxy modely zahrnují Gaussovské procesy, stromy, případně meta-algoritmy. V rámci akvizičních funkcí bývá uvažováno o pravděpodobnosti zlepšení účelové funkce, očekávaném zlepšení účelové funkce nebo hranici spolehlivosti. BO je vhodná pro problémy, v kterých není možné účelovou funkcí a její derivace vyjádřit analyticky, ohodnocení účelové

funkce je výpočetně nákladné, nebo v případech s omezeným přístupem k prohledávanému prostoru. Na druhou stranu je přístup citlivý na výběr a nastavení proxy modelu; jeho konstrukce a učení stojí další výpočetní prostředky, a pro dobré odhady je třeba dostatek pozorování.

Algorithm 1: Bayesian Optimization

for search space samples $t = 1, 2, \ldots$ do:

1. select sampling point x_t by optimizing the acquisition function u over the surrogate model, where $x_t = argmax_x u(x|D_{1:t-1})$,
2. evaluate noisy objective function f, $y_t = f(x_t) + \epsilon_t$,
3. add the sampling point to the sampling set such as $D_{1:t} = \{D_{1:t-1}, (x_t, y_t)\}$,
4. re-fit the surrogate model,
5. go to 1.

Obr. 18 Algoritmus Bayesovské optimalizace

Zdroj: Shahriari et al. (2016)

Kalibrace pravděpodobností klasifikačního modelu

V úlohách klasifikace není předmětem zájmu pouze příslušnost k třídě, ale i její pravděpodobnost. Prediktivní modely produkují zpravidla zkreslené odhady pravděpodobnosti, které reflektují spíše jistotu než úspěšnost systému, případně není možné potřebné odhady generovat vůbec. Naznačené nedostatky je možné adresovat kalibrací pravděpodobnosti, mezi oblíbené přístupy lze zahrnout škálování odhadnutých pravděpodobností příslušnosti k třídě s pomocí sigmoidové funkce (Platt, 1999), případně s využitím isotonické regrese (Zadrozny & Elkan, 2002). Zavedení kalibrace vede k zpřesnění odhadů nejistoty, a v konečném důsledku také k lepšímu řazení datových instancí s ohledem na související podnikovou perspektivu. Za problematické lze považovat další výpočetní náklady, růst komplexity systému a v některých případech i pokles prediktivních schopností.

3.3.4 Ostatní

PRISMA (Preferred Reporting Items for Systematic review and Meta-Analysis) označuje metodiku systematického rešeršního procesu, zajišťuje že proces je důkladně naplánován a explicitně dokumentován před započetím vlastní rešerše, což vede ke konzistentnímu postupu vědeckého týmu, zodpovědnosti, integritě a transparentnosti dosažených výsledků. Mezi další dopady využití metodiky náleží eliminace svévolných rozhodnutí a identifikace možných problémů výzkumu. Motivací k vytvoření metodiky byla nutnost systematizovat rešeršní proces a tvorbu meta-analýz v prostředí zdravotnického výzkumu (Moher et al., 2009). Pro potřeby

disertační práce využívá autor především strukturu informačního toku rešerše, pro diskusi konkrétních atributů výzkumu je užito rámce CRISP-DM.

CRISP-DM (Cross-Industry Standard Process for Data Mining) popisuje referenční model určující fáze životního cyklu prediktivního modelování a odpovídající vazby, které demonstrují iterativní a kontinuální charakter aktivit. Model se v základní podobě skládá ze šesti kroků, porozumění podnikovému/výzkumnému problému, porozumění souboru dat, zpracování dat, modelování, hodnocení a interpretace výsledků a vlastní nasazení řešení (Chapman et al., 2000). Autor využívá představené perspektivy k analýze současného stavu poznání i k organizaci vlastního řešení.

4 Návrh a implementace řešení

4.1 Vymezení problému

Systém strojového učení je navrhován s ohledem na cíle disertační práce, má tedy sloužit k predikci odchodu zákazníka v prostředí elektronického maloobchodu, při reflexi ekonomického dopadu retenčních aktivit a dalšímu porozumění modelovanému jevu. Cíle práce a výzkumné otázky jsou zevrubně popsány v kapitolách 3.1 a 3.2.

4.1.1 Retenční management

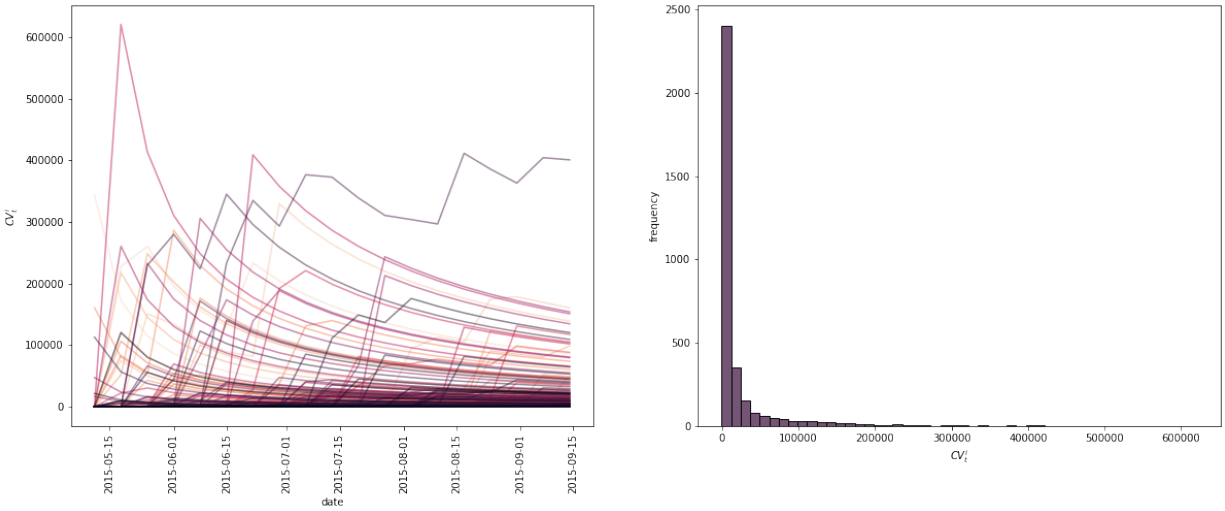
Za účelem zhodnocení ekonomického dopadu zamýšlené retenční kampaně realizované na základě predikcí ztráty zákazníka, ale za i účelem nové formulace úlohy v rámci systému strojového učení, vymezuje autor v následujících odstavcích potřebné veličiny a související vztahy.

Hodnota zákazníka

Pro potřeby disertační práce definuje autor očekávanou hodnotu zákazníka pomocí individuální úrovně kumulativního průměru zisku, která je dále upravena tak, aby odrážela délku časového úseku cílové proměnné. Pokud máme zákazníka i v čase t, pak jeho hodnotu spočteme jako

$$CV_i^t = \frac{n_t}{t} \sum_{n=1}^{t} m_p^n r_p^n, \tag{35}$$

kde n_t jest délka časového okna cílové proměnné, m_p^n označuje marži produktu p v čase n, podobně potom r_p^n reprezentuje výdaje zákazníka na produkt p v čase n. Ukazatel umožňuje zohlednit rozdílné úrovně zisku napříč zákaznickou bází, indikuje změny zákaznického chování v čase a je užitečný v kontextu časově ohraničených retenčních aktivit. Kumulativní průměr snižuje očekávanou hodnotu u zákazníků s nižším počtem transakcí, tento aspekt je možné dále rozvinout pomocí jiné funkce rozpadu. Mezi možná omezení řadí autor úzké spojení s časovým oknem závislé proměnné a omezenou reflexi životního cyklu zákazníka.



Obr. 19 Očekávaná hodnota zákazníka pro časový úsek 4 týdnů, vývoj v čase a celkové rozložení veličiny v datovém souboru Retail Rocket

Přístup k výpočtu hodnoty zákazníka vystihuje Obr. 19, ve kterém levá část grafu ilustruje vývoj zákaznické hodnoty v čase, převládající vzorec chování se sestává z počátečního vrcholu (první transakce) a postupného poklesu (žádné další transakce). Pouze hrstka zákazníků nakupuje opakovaně, což potvrzuje význam retenčního řízení. Napravo lze pozorovat asymetrii celkového rozložení očekávané hodnoty zákazníka, s těžištěm okolo nuly a významným pravým chvostem. Tvar rozložení je výsledkem uvedeného dominantního vzoru.

Maximální očekávaný zisk

Idea maximálního očekávaného zisku kampaně reflektuje obvyklé využití prediktivních modelů k řazení zákaznické báze, dle očekávané pravděpodobnosti odchodu, a následnému cílení retenčních aktivit na nejvíce ohroženou část zákazníků. Klíčový je příklon k řazení zákaznické báze, dle očekávaní inkrementální hodnoty zisku nebo ztráty, které společnost získá zahrnutím individuálního zákazníka do dané kampaně. Autorova perspektiva rozšiřuje práci Tamaddoni et al. (2014) o časově ohraničený, konzervativní odhad očekávané hodnoty zákazníka popsaný v předchozí sekci. Předpokládaný ekonomický důsledek zařazení zákazníka i do retenční kampaně pak spočteme jako

$$\pi_i^{expected} = p_i[\gamma_i(CV_i - \delta)] + (1 - p_i)[-\psi_i\delta], \tag{36}$$

kde p_i jest predikovanou pravděpodobností ztráty zákazníka v budoucím období, γ_i určuje pravděpodobnost, že retenční nabídka přiměje zákazníka zůstat aktivním, ψ_i reprezentuje pravděpodobnost, že zákazník, který neměl v úmyslu odejít akceptuje retenční nabídku a δ označuje

jednotný náklad incentivy. První sčítanec odpovídá předpokladu, že zdárné oslovení ohroženého zákazníka povede k novým transakcím a zisku alespoň ve výši CV_i. Takový odhad je konzervativní, především s ohledem na očekávanou hodnotu u zákazníků s nižším počtem nákupů. Druhý sčítanec koresponduje se zacílením retenční aktivity na zákazníka, u kterého odchod nehrozí. Výše incentivy δ je zvolena arbitrárně. V případě cílení na ohrožené a hodnotné zákazníky nabývá ukazatel hodnot kladných, při cílení na věrné zákazníky, potom hodnot záporných. Celkový očekávaný zisk retenční kampaně potom můžeme snadno spočíst jako

$$\Pi^{expected} = \sum \pi_i^{expected}, \tag{37}$$

kde $\pi_i^{expected}$ značí očekávaný zisk nebo ztrátu spojené se zahrnutím příslušného zákazníka do retenční aktivity. Maximální očekávaný zisk kampaně pak odpovídá cílení na zákazníky, u nichž očekáváme kladný inkrementální výsledek zařazení do retenční aktivity. Výstupem takového přístupu je tedy odhad dopadu kampaně i složení cílové skupiny zákazníků. Pojetí však zohledňuje spíš jistotu než úspěšnost klasifikačního řešení.

Maximální dosažený zisk

Přístup popsaný v předchozí sekci naznačuje, jak s pomocí klasifikace cílit na vhodné zákazníky, nereflektuje však prediktivní schopnosti řešení. Tento neduh autor adresuje s pomocí maximálního dosaženého zisku retenční kampaně, který využívá informaci o vztahu k pozorovaným třídám. Pokud by byl zákazník i zahrnut v retenční kampani, pak skutečný inkrementální příspěvek k výsledku kampaně spočteme s využitím představené notace jako

$$\pi_i^{actual} = y_i[\gamma_i(CV_i - \delta)] + (1 - y_i)[-\psi_i\delta], \tag{38}$$

kde y_i označuje binární závislou proměnnou, jenž nabývá hodnoty 0 pokud zákazník setrvá a 1 pokud bude ztracen. Podobně pak skutečný dosažený zisk určíme jako součet individuálních zisků zákazníků zahrnutých do kampaně, tj.

$$\Pi^{actual} = \sum \pi_i^{actual}, \tag{39}$$

kde π_i^{actual} určuje dosažený zisk nebo ztrátu spojené se zahrnutím daného zákazníka do retenční aktivity. Maximální dosažený zisk kampaně pak odpovídá cílení na zákazníky, u nichž očekáváme kladný inkrementální výsledek zařazení do retenční aktivity, tj. na zákazníky kteří dosahují kladného $\pi_i^{expected}$. Nespornou předností nastíněného rámce je hodnocení

prediktivních modelů s ohledem na ekonomický dopad retenční kampaně, vlastní odhad ekonomického dopadu i vhodné složení cílové skupiny zákazníků. Mezi možné nedostatky řadíme především nutnost stanovení dílčích parametrů kampaně.

4.2 Porozumění datovému souboru

V rámci sekce se autor věnuje dostupným datovým souborům, jejich vlastnostem a nezbytným úpravám, dále se zabývá návrhem a konstrukcí datové reprezentace zákaznické modelu, včetně související explorativní analýzy.

4.2.1 Datové soubory

Pro ověření navrženého přístupu využíváme dvou volně dostupných souborů dat, Retail Rocket (2017) a REES46 (2020), které popisují uživatelské interakce s webovou aplikací podniku, včetně transakční historie.

Retail Rocket Dataset, zveřejnila společnosti Retail Rocket na platformě Kaggle, za účelem řešení konstrukce doporučovacího systému. Datový soubor pokrývá období mezi daty 05/09/2015 až 09/17/2015, tj. přibližně 19 týdnů. Data popisují 2,331,222 interakcí generovaných 1,160,164 uživateli. Tyto interakce vedly k 14,252 transakcím v průměrné výši 155,644.7 peněžních jednotek (Currency Units – CU), které generovalo 10,890 zákazníků.

Podobně i REES46 Dataset, zveřejnila společnost REES46 na platformě Kaggle, za účelem zvýšení povědomí o vlastních datových produktech a službách. Tento datový soubor pokrývá období 10/01/2019 až 05/01/2020, tj. přibližně 30 týdnů. Data popisují 411,709,736 interakcí generovaných 15,639,803 uživateli. Tyto interakce vedly k 5,449,933 transakcím, v průměrné výši 377.5 CU, které realizovalo 2,064,899 zákazníků.

Úroveň detailu u obou datových souborů odpovídá interakci uživatele s nabízeným produktem, kde známe identifikátor zákazníka, čas interakce, identifikátor časové relace, typ interakce, identifikátor produktu, zařazení produktu v katalogu kategorií, a cenu produktu během interakce. Množina popsaných vlastností limituje možné směry, kterými se v následných krocích věnovat, především s ohledem na další vlastnosti zákazníka (mezilidské vztahy, sociálně-ekonomické nebo místní odlišnosti), podniků (mikroprostředí, makroprostředí), atp. Pro další práci s daty jsou oba zdroje sjednoceny tak, aby k nim bylo možné přistupovat jednotně ve smyslu významu i datového formátu. Autor se v rámci následného zpracování soustředí na

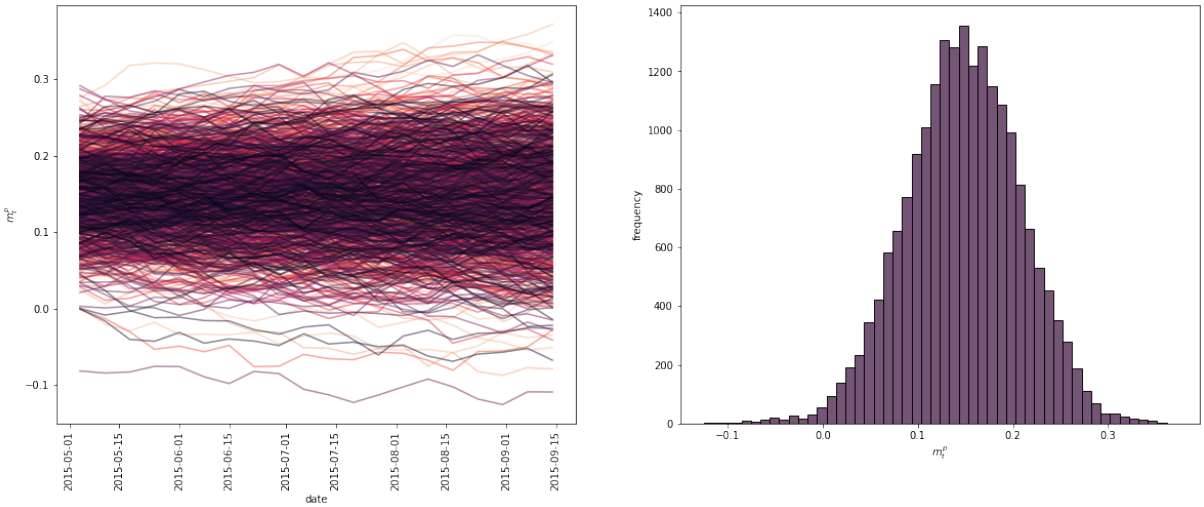
podmnožiny zákazníků, které v prvním datovém souboru vymezuje realizací alespoň dvou, v druhém datovém souboru, alespoň dvaceti transakcí.

4.2.2 Marže produktu

Dostupné datové soubory neobsahují informace o realizovaných maržích. Za účelem analýzy ekonomického výsledku retenční kampaně navrhuje autor přístup založený na počítačové simulaci, který zachycuje možný vývoj úrovně marže produktu v čase. Pro každý z produktů je nejprve určena základní úroveň marže tažením z prvotního náhodného rozdělení. Toto tažení slouží jako počáteční bod náhodné procházky, jejíž kroky jsou simulovány dalšími tahy z odpovídajícího rozdělení, cílovou polohu marže pak je možné spočíst kumulativním součtem těchto kroků. Jinými slovy, pro simulaci marže je využita jednorozměrná náhodná procházka s Gaussovskými kroky. Pro produkt p v čase t spočteme úroveň marže m_p^t následovně

$$m_p^t = X_{Normal(\mu_0,\sigma_0)} + \sum_{n=1}^{t} X_{Normal(\mu_{diff},\sigma_{diff})}, \tag{40}$$

kde první prvek odpovídá úvodnímu určení polohy marže, a druhý prvek reprezentuje kumulativní součet dílčích kroků. Pro jednoduchost autor předpokládá, že úvodní tah napříč produkty a jednotlivé kroky v rámci produktu mají stejný rozptyl. Pak je možné úroveň parametrů stanovit jako $\mu_{diff} = 0$ a $\sigma_{diff} = \frac{\sigma_0}{\sqrt{t}}$, tj. jsme schopni určit marži produktu pouze stanovením atributů počátečního náhodného rozdělení μ_0 a σ_0.



Obr. 20 Simulovaná úroveň marže pro vzorek produktového portfolia datového souboru Retail Rocket

Přístup k počítačové simulaci marže produktu je ilustrován s pomocí Obr. 20, kde, kde lze v levém grafu pozorovat vývoj simulované marže v průběhu času, vpravo potom vidíme celkové rozložení simulovaných marží. Oba grafy jsou v souladu s parametry počítačové simulace a odrážejí některé zákonitosti běžného maloobchodu, tj. většina produktů vykazuje kladnou marži, jsou zde ale i produkty s marží zápornou.

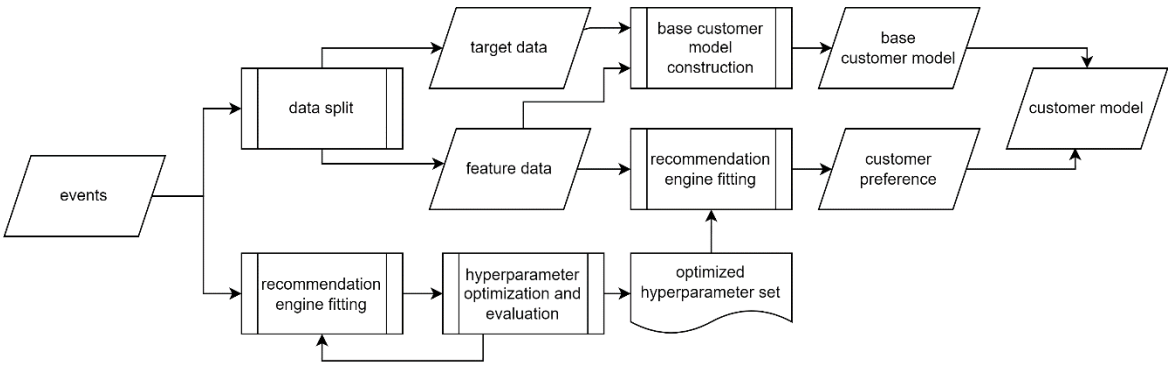
Pro účely práce využívá autor vstupních parametrů na úrovni $\mu_0 = 0.15$, $\sigma_0 = 0.05$, spolu s denními kroky náhodné procházky. Pro úplnost je však vhodné popsat vztah mezi vstupními parametry počítačové simulace a výslednou očekávanou hodnotou zákazníka pomocí analýzy citlivosti. Pro každou z vybraných kombinací μ_0 a σ_0 byl proces opakován v tisíci krocích a byla spočteny odpovídající střední očekávaná hodnota zákaznické hodnoty, které jsou uvedeny v Tab. 4. Z výsledků je zřejmé, že pro změny v $E(CV_i)$ je významná především úroveň μ_0, změny σ_0 mají na svědomí růst variability, úroveň střední hodnoty však zásadně neovlivňují. Jinými slovy, zákaznická hodnota přímo souvisí s celkovou schopností podniku udržovat v rámci portfolia služeb nebo produktů zdravou úroveň marže.

Tab. 4 Očekávaná úroveň zákaznické hodnoty pro různé vstupní parametry počítačově simulované marže produktu

		Retail Rocket	REES46
μ_0	σ_0	$E(CV_i)$	$E(CV_i)$
0.05	0.05	6885.2 (6053.6, 7749.5)	19.0 (14.1, 24.2)
0.05	0.10	6892.1 (5228.9, 8620.6)	19.0 (9.3, 29.4)
0.05	0.15	6898.9 (4404.2, 9491.7)	19.1 (4.4, 34.7)
0.15	0.05	20641.9 (19810.3, 21506.2)	56.9 (52.0, 62.1)
0.15	0.10	20648.8 (18985.6, 22377.3)	57.0 (47.2, 67.4)
0.15	0.15	20655.6 (18160.9, 23248.4)	57.0 (42.4, 72.6)
0.25	0.05	34398.6 (33567.0, 35262.9)	94.8 (90.0, 100.0)
0.25	0.10	34405.5 (32742.3, 36134.0)	94.9 (85.1, 105.3)
0.25	0.15	34412.3 (31917.6, 37005.1)	94.9 (80.3, 110.5)

4.2.3 Model zákazníka

Úvodním krokem tvorby datové reprezentace je sjednocení užitých souborů dat z hlediska významu atributů i datového formátu. Konstrukce modelu zákazníka staví na dostupných uživatelských interakcích, které jsou transformovány ve větvích základního modelu a modelu preferencí. V rámci větvě preferencí dochází k optimalizaci vnějších parametrů doporučovacího systému, parametry jsou využity pro konkrétní časový řez, kde na základě dostupných dat a zvolených parametrů konstruován doporučovací systém, jehož latentní faktory jsou využity jako implicitní reprezentaci zákaznických preferencí. Ve větvi základního modelu dochází k akvizici dat potřebných časových řezů, za účelem oddělené konstrukce vysvětlovaných a vysvětlujících proměnných. Výstupy obou větví jsou spojeny do dílčí reprezentace, které jsou v závěru sloučeny do kýženého modelu zákazníka. Výsledná data i programový kód jsou obsahem přílohy B1.



Obr. 21 Proces konstrukce datové reprezentace modelu zákazníka

Vysvětlované proměnné

Pro vymezení vysvětlovaných proměnných je uvažováno v intencích otázek, které je třeba zodpovědět pro návrh a realizaci úspěšné retenční kampaně. Tradiční událost odchodu zákazníka je charakterizována jako absence nákupu po dobu následujících čtyř týdnu, tj. jedná se o binární závislou proměnnou. Pojetí umožňuje adresovat identifikaci rizikových zákazníků a reflektuje obvyklé vymezení ztráty zákazníka (viz podkapitola 2.2.1). Perspektiva je rozšířena o inkrementální příspěvek k ekonomickému výsledku retenční kampaně plynoucího ze zahrnutí individuálního zákazníka do uvažované aktivity během následujících čtyř týdnů, tj. jedná se o spojitou závislou proměnnou. Původní hledisko přímo adresuje odhad ekonomického výsledku zamýšlené kampaně, ale i vhodnou velikost a složení cílové skupiny zákazníků. Význačný je také dopad na konstrukci dílčích prvků systému, včetně interpretace prediktivních modelů.

Pro konstrukci inkrementálního příspěvku k ekonomickému výsledku retenční kampaně π_i^{actual} je využito především rovnic 35 a 38, přestavených v kapitole 4.1.1. Z dostupných dat jsou získány hodnoty tradiční vysvětlované proměnné y_i a hodnoty zákazníka CV_i. Náklad incentivy individuální retenční kampaně δ vymezuje autor na úrovni přibližně třetiny průměrného CV_i, tj. pro datový soubor Retail Rocket $\delta = 7000\ CU$, pro REES46 $\delta = 20\ CU$. Uplatnění incentivy je charakterizováno u ohrožených zákazníků parametry $E(\gamma_i) = 0.01$, a současně $\gamma_i \sim Beta(2.04, 202.04)$, u věrných zákazníků pak $E(\psi_i) = 0.66$ a zároveň $\psi_i \sim Beta(6.12, 3.15)$. S pomocí počítačové simulace je losováno z příslušných náhodných rozdělení pravděpodobnosti, na jejichž základě dochází ke stanovení dílčího odhadu závislé proměnné. Očekávaná hodnota veličiny je pak spočtena jako prostý průměr tisíce losování. Popsaný přístup umožňuje odhadnout π_i^{actual} i s omezenými znalostmi citlivosti zákazníka na marketingové aktivity, současně zohledňuje i prvek nejistoty.

Počítačová simulace je doplněna o analýzu citlivosti očekávané hodnoty vysvětlované proměnné ve vztahu k parametrům náhodného rozdělení pravděpodobnosti popisujícího uplatnění incentivy u ohrožených zákazníků γ_i. Veličina byla vybrána především s ohledem na přímý dopad do příjmové části retenční kampaně. Pro zvolené rozdělení pravděpodobnosti byl proces simulace opakován v tisíci krocích, díky čemuž bylo možné odhadnout odpovídající střední hodnoty závislé proměnné. Z výsledků, popsaných v Tab. 5, vyplývá, že pro změny $E(\pi_i^{actual})$ je významná především úroveň $E(\gamma_i)$, $var(\gamma_i)$ ovlivňuje spíše rozptyl veličiny. Za povšimnutí stojí záporné hodnoty $E(\pi_i^{actual})$, které akcentují potřebu správného řazení zákazníků pro

uvažované retenční aktivity. Při náhodném zařazení zákazníků do retenční kampaně zvolených parametrů, by totiž individuální kampaň generovala ztrátu, nikoliv zisk.

Tab. 5 Očekávaná úroveň individuálního příspěvku k ekonomickému výsledku retenční kampaně pro různé vstupní parametry počítačové simulace uplatnění incentivy ohroženými zákazníky

			Retail Rocket	REES46
Distribution of γ_i	$E(\gamma_i)$	$ln[var(\gamma_i)]$	$E(\pi_i^{actual})$	$E(\pi_i^{actual})$
Beta(0.001, 1.1)	0.001	-7.65	-459.8 (-488.1, -375.1)	-8.85 (-8.92, -8.75)
Beta(0.02, 19.7)	0.001	-9.94	-457.7 (-481.4, -419.6)	-8.85 (-8.89, -8.81)
Beta(0.41, 407.0)	0.001	-12.92	-457.8 (-474.9, -440.8)	-8.85 (-8.88, -8.82)
Beta(0.20, 19.5)	0.010	-7.65	-318.4 (-388.8, -224.0)	-7.81 (-7.90, -7.72)
Beta(2.0, 202.1)	0.010	-9.94	-316.4 (-347.9, -283.7)	-7.81 (-7.85, -7.77)
Beta(40.5,4 005.7)	0.010	-12.92	-317.1 (-333.7, -301.2)	-7.81 (-7.84, -7.78)
Beta(18.7, 168.4)	0.100	-7.65	1094.0 (1002.4, 1181.4)	2.58 (2.49, 2.67)
Beta(186.4, 1677.7)	0.100	-9.94	1093.0 (1063.0, 1123.9)	2.58 (2.54, 2.62)
Beta(3679.2, 33112.5)	0.100	-12.92	1093.6 (1077.8, 1108.5)	2.58 (2.55, 2.61)

Vysvětlující proměnné

Vlastní datová reprezentace vychází především z objektivních proměnných, popisujících způsob využití webové aplikace a transakční historii uživatele. Další aspekty úlohy jako reflexe mezilidských vztahů, sociálně-demografické, nebo místní odlišnosti mezi uživateli nejsou využívány, především s ohledem na limitace spojené s dostupnými soubory dat. Pro přehlednost členíme vysvětlující proměnné do skupin popisujících aktualitu, frekvenci, peněžní hodnotu, případně datum a čas interakcí mezi podnikem a zákazníkem; dále jsou zahrnuty zákaznické preference a některé další vlastnosti. Výsledkem je amalgam jak transakčních, tak behaviorálních vysvětlujících proměnných.

Nosnou částí modelu je perspektiva běžně charakterizující především transakční historii, tj. stáří, frekvence a peněžní hodnota transakcí. Takto však nahlížíme na transakční i netransakční chování. Stáří zde vystihuje dobu uplynulou od poslední známé interakce; vysoké hodnoty naznačují nižší pravděpodobnost opakovaní interakce. Frekvence vyjadřuje, jak často k interakcím dochází; nižší hodnoty mohou naznačovat méně loajálního zákazníka. Peněžní hodnota operací potom odpovídá objemu prostředků, které zákazník zamýšlí utratit nebo utratí během stanového časového období (Liu & Shih, 2005).

Preference zákazníka jsou zahrnuty s ohledem na předpoklad, že zákazník se zájmem o širokou paletu produktů bude podniku věrný (Mozer et al., 2000), naopak zájem o problematickou část portfolia může indikovat zákazníka ohroženého (Buckinx & Dirk, 2005). Datum a čas je uvažován s ohledem na reflexi vývoje zákaznické zkušenosti (Buckinx & Dirk, 2005), příkladem může být pravidelná polední zátěž služby prodejce vedoucí ke zpomalení odezvy a zhoršení této zkušenosti. Ostatní atributy popisují zákaznické chování uvnitř relace; předpokládáme, že věrní zákazníci interagují s aplikací jiným způsobem než zákazníci, kteří zvažují přerušení vzájemného vztahu.

Tab. 6 Model zákazníka v perspektivě vysvětlujících proměnných

Set	Attribute	Description	Variable name
Recency	session recency	time duration from the last session [days]	session_recency
	purchase recency	time duration from the last transaction [days]	purchase_recency
	time to session	time between sessions [days]	inter_session_time
	time to purchase	time between purchases [days]	inter_purchase_time
Frequency	session number	user-session number [n]	session_number
	purchase number	user-purchase number [n]	purchase_number
	session daily frequency *	session count per day	session_count_ratio
	interaction daily frequency *	interaction count per day	click_count_ratio
	transaction daily frequency *	transaction count per day	transaction_count_ratio
	interaction frequency	user–application interaction (view/add-to-cart/purchase) count [n]	click_count
	view frequency	product views count [n]	view_count
	add-to-cart frequency	products added to a carts count [n]	cart_count
	purchase frequency	product purchased count [n]	purchase_count
	total session frequency, with lags *	monthly total session count, with lags [n]	session_count_month_lag
	ma of total session frequency *	moving average of monthly total session count	session_count_month_ma
	total purchase frequency, with lags *	monthly total purchase count, with lags [n]	purchase_count_month_lag
	ma of total purchase frequency *	moving average of monthly total purchase count	purchase_count_month_ma
	has-purchase indicator	indicator, whether the session includes purchase	haspurchase
Monetary	viewed revenue	potential revenue from viewed products [CU]	view_revenue
	added-to-cart revenue	potential revenue from products added to cart [CU]	cart_revenue
	transactional revenue	revenue from realized purchases [CU]	purchase_revenue
	transactional revenue, with lags *	monthly transactional revenue, with lags [CU]	purchase_revenue_month_lag
	ma of transactional revenue *	moving average of monthly transactional revenue [CU]	purchase_revenue_month_ma
	customer value, with lags *	monthly customer value, with lags [CU]	customer_value_month_lag
	ma of customer value *	moving average of monthly total customer value [CU]	customer_value_month_ma
Preference	view latent factor *	latent factor characterizing preference manifested through product views	view_latent_factor

Set	Attribute	Description	Variable name
	purchase latent factor *	latent factor characterizing preference manifested through transactions	purchase_latent_factor
Date & time	year	year of a session start	start_year
	month	month of a session start	start_month
	week	week of a year of a session start	start_week
	day of year	day of a year of a session start	start_yearday
	day of month	day of a month of a session start	start_monthday
	day of week	day of a week of a session start	start_weekday
	weekend indicator	indicator, whether the session began during a weekend	start_isweekend
	hour	hour of a session start	start_hour
Others	time to interaction	average duration between interactions [min]	time_to_click
	time to view	average duration between product views [min]	time_to_view
	time to add-to-cart	average duration between adding products to cart [min]	time_to_cart
	time to purchase	average duration between purchases [min]	time_to_purchase
	time to interaction revenue	average duration to interact with a product of value 1 CU [min/CU]	time_to_click_revenue
	time to view revenue	average duration to view a product of value 1 CU [min/CU]	time_to_view_revenue
	time to add-to-cart revenue	average duration to add a product of value 1 CU to a cart [min/CU]	time_to_cart_revenue
	time to revenue	average duration to a purchase of value 1 CU [min/CU]	time_to_purchase_revenue
	session length	session duration [min]	length

* hodnoty proměnných jsou počítány přímo na úrovni zákazníka, tj. není využito agregačních funkcí

Základní vlastnosti zvolené datové reprezentace ilustruje tabulka Tab. 6. Převládají atributy spočtené na úrovni zákaznické relace, pro každého zákazníka a časový výřez jsou pak agregovány funkcemi minima, maxima, průměru, sumy, standardní odchylky výběru a variačního koeficientu výběru. Konstrukce zákaznických preferencí a časově rozlišených proměnných je realizována na úrovni individuálního zákazníka, tj. nebylo třeba úpravy detailu.

Explorace dat

Datovou reprezentaci je třeba prozkoumat, především pro porozumění vlastnostem jednotlivých atributů i vzájemných vztahů. Za tímto účelem je využito nástrojů popisné statistky, jako jsou ukazatele polohy, rozptylu, nebo zakřivení. Pohled napříč vztahy mezi proměnnými ilustrují perspektivy významných korelací mezi závislou a nezávislými proměnnými, ale i distribucí korelací mezi nezávislými atributy datových sad.

Tab. 7 Popisné statistiky vybraných proměnných datového souboru Retail Rocket

Variable name	μ	median	min	max	σ	skew	kurt
target_event	0.90	1.00	0.00	1.00	0.30	-2.62	4.87
target_actual_profit	-317.35	-23.52	-4,744.99	9,360.41	1,587.21	-1.43	4.82
session_recency_min	31.33	25.38	0.00	97.39	25.81	0.73	-0.46
purchase_recency_min	33.71	28.52	0.00	97.39	25.60	0.65	-0.57
session_number_max	8.61	3.00	1.00	311.00	21.37	6.65	56.83
purchase_number_max	3.05	2.00	1.00	124.00	5.55	9.83	138.70
view_revenue_sum	7.77E+06	5.21E+05	4,560.00	1.25E+09	4.57E+07	19.44	468.18
purchase_revenue_sum	6.29E+05	1.49E+05	4,320.00	5.66E+07	2.07E+06	12.15	241.36

Zákaznický model pro datový soubor Retail Rocket se sestává z 2722 pozorování, 2 vysvětlovaných a 222 vysvětlujících proměnných. Cílová proměnná klasifikace, tj. absence transakce po následující 4 týdny, je silně nevyvážená, pozorujeme odliv ~ 90 % všech zákazníků. Cílová proměnná regrese, tj. skutečný inkrementální zisk retenční kampaně, dosahuje střední hodnoty −317.35 CU a vykazuje znaky mírně záporně sešikmeného rozdělení s těžšími chvosty.

Vybrané vysvětlované proměnné reprezentují tradiční dimenze RFM analýzy, aplikované na uživatelské relace i transakce. Stáří poslední relace i transakce mají podobná rozdělení, která jsou více symetrická a plošší. V průměru došlo k poslední relaci/transakci zákazníka před přibližně 30 dny. Průměrný počet uživatelských relací je přibližně 9, průměrný počet transakcí zákazníka jsou 3. Rozdělení obou proměnných jsou pozitivně sešikmené, s tlustým pravým chvostem. Zákazník si v průměru prohlédl produkty v hodnotě 7.77E+06 CU a nakoupil zboží a služby za 6.29E+05 CU. Podobně jako u frekvence, vykazují obě rozdělení znaky pozitivního sešikmení s tlustým pravým chvostem.

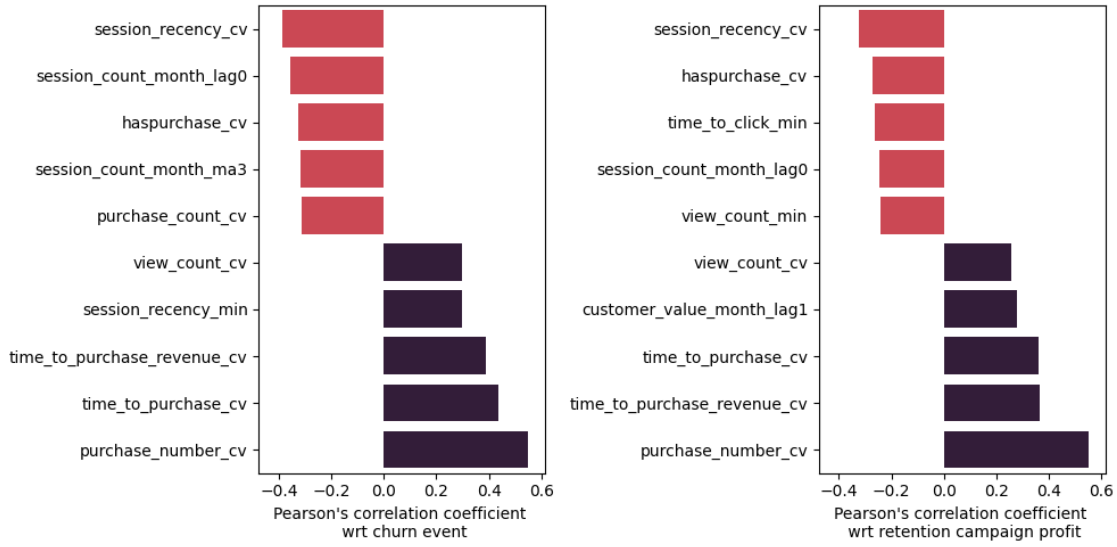
Zákaznický model pro datovou sadu REES46 se skládá z 112610 instancí, 2 vysvětlovaných a 267 vysvětlujících proměnných. Cílová proměnná klasifikace je více vyvážená, s odlivem °~ 31 % zákazníků. Cílová proměnná regrese dosahuje střední hodnoty −7.81 CU, vykazuje znaky pozitivně sešikmeného rozdělení tlustým pravým chvostem.

Průměrná doba uplynulá od poslední relace je přibližně 13 dní, průměrná uplynulá doba od poslední transakce je 18 dní, rozdělení veličin vykazují znaky asymetrie. Průměrný počet uživatelských relací je 54, průměrný počet transakcí je 23. Rozdělení obou proměnných je pozitivně sešikmené, s tlustým pravým chvostem. Zákazník si v průměru prohlédl produkty v hodnotě 7.36E+04 CU a nakoupil zboží a služby za 11407.35 CU. Podobně jako u frekvence, vykazují obě rozdělení znaky pozitivního sešikmení s tlustým pravým chvostem.

Tab. 8 Popisné statistiky vybraných proměnných datového souboru REES46

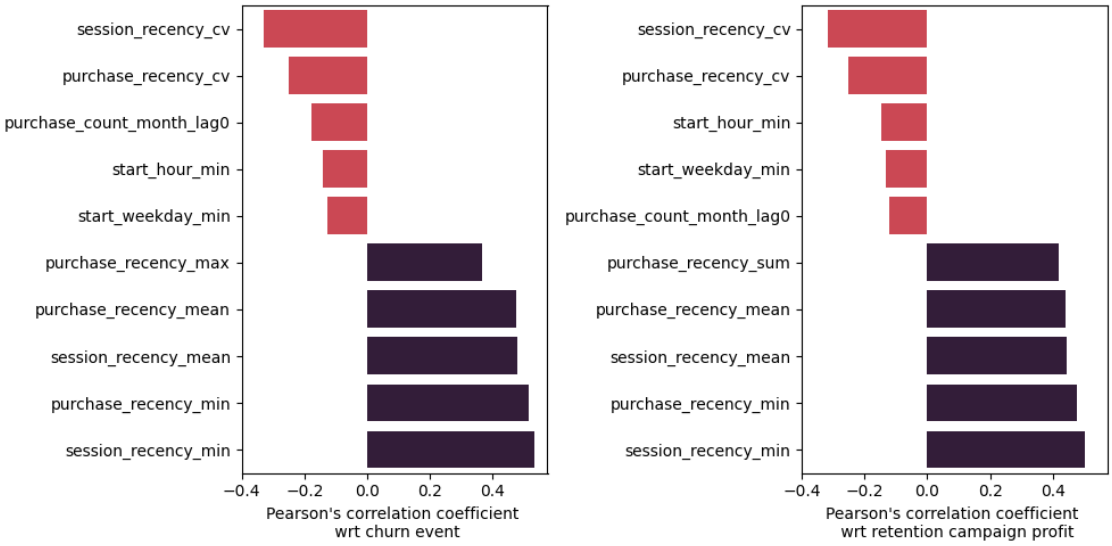
Variable name	μ	median	min	max	σ	skew	kurt
target_event	0.32	0.00	0.00	1.00	0.47	0.77	-1.41
target_actual_profit	-7.81	-13.15	-13.59	189.97	8.47	2.03	14.64
session_recency_min	13.33	3.88	0.00	174.56	23.57	2.90	9.22
purchase_recency_min	17.91	7.46	0.00	176.47	25.45	2.37	6.16
session_number_max	53.62	40.00	1.00	2.15E+04	116.35	103.31	1.63E+04
purchase_number_max	22.71	20.00	1.00	1742.00	23.97	11.35	506.61
view_revenue_sum	7.36E+04	4.26E+04	0.00	1.34E+07	1.22E+05	24.50	1990.62
purchase_revenue_sum	11407.35	6083.23	1.13	7.50E+05	2.01E+04	10.25	212.47

Na datové modely je dále nahlíženo prostřednictvím pěti nejsilnějších pozitivních a negativních korelací mezi vysvětlovanými a vysvětlujícími proměnnými, viz Obr. 22 a 23. V Retail Rocket se ukazuje, že událost ztráty zákazníka vykazuje negativní, středně silnou korelaci s variačním koeficientem doby uplynulé od posledního nákupu a počtu uživatelských relací; vyšší hodnoty koeficientů naznačují věrného zákazníka. Pozitivní korelace středně silné a naznačují nižší počet nákupů i čas mezi nákupními interakcemi u věrných zákazníků. Pro inkrementální zisk retenční kampaně lze pozorovat slabší negativní korelaci s variačním koeficientem stáří relace a konverzním poměrem, tj. pokud zákazník navštěvuje obchodníka pravidelně a často, případně vykazuje vyšší konverzní poměr relací, pak je možné očekávat lepší výsledek retenční aktivity. Pozitivní korelace jsou středně silné, vynikají především variační koeficienty počtu nákupů a doby v relaci vedoucí k nákupu 1 CU.



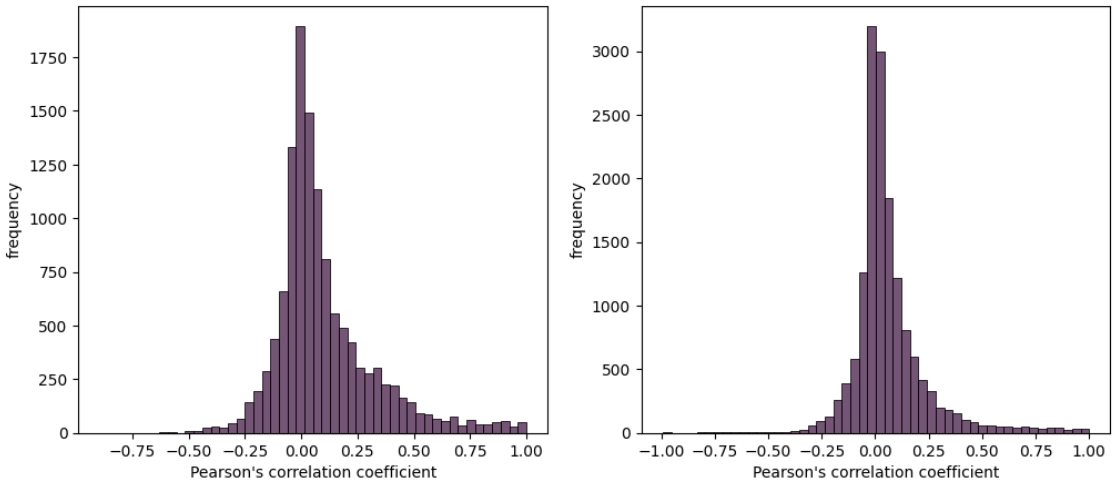
Obr. 22 Významné korelace mezi vysvětlovanými a vysvětlujícími proměnnými v datovém souboru Retail Rocket

V REES46 vykazuje událost ztráty zákazníka slabou negativní korelaci s variačním koeficientem stáří poslední uživatelské relace i transakce. Dále je lze pozorovat slabou negativní korelaci s počtem transakcí v aktuálním měsíci, tj. nízký počet transakcí naznačuje vyšší riziko ztráty zákazníka. Naopak pozitivní středně silnou korelaci vykazují stáří relace i nákupu, tj. nedávné a stabilní interakce naznačují zákazníka méně ohroženého. Pro inkrementální zisk retenční kampaně je možné pozorovat slabé negativní korelace s variačními koeficientem stáří poslední relace i transakce. Dále se ukazují středně silné pozitivní korelace se stářím relací/transakcí, což značí vyšší předpokládanou ziskovost u zákazníků ohrožených.



Obr. 23 Významné korelace mezi vysvětlovanými a vysvětlujícími proměnnými v datovém souboru REES46

S ohledem na předpoklady některých algoritmů nebo metod interpretace zkoumáme i rozdělení korelací mezi jednotlivými vysvětlujícími proměnnými, která ilustruje Obr. 24. Zdá se, že v obou datových sadách pozorujeme především slabé korelace, nezanedbatelná část pozorování však vykazuje korelace silné.



Obr. 24 Rozdělení koeficientů korelace mezi vysvětlujícími proměnnými pro datové soubory Retail Rocket (vlevo) a REES46 (vpravo)

4.3 Zpracování datového souboru

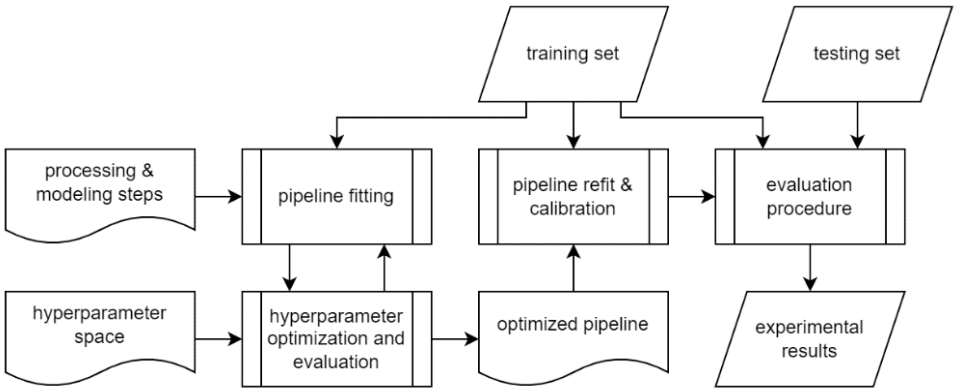
S ohledem na nastíněné vlastnosti jako jsou počet, nízká hustota, asymetrie, nebo vnitřní korelace nezávislých proměnných, případně nevyvážené třídy nebo asymetrie závislé proměnné, je proces zpracování datového souboru složen z kroků škálování, eliminace proměnných s nízkou variabilitou, a výběru významných nezávislých proměnných. Pro klasifikační úlohy zvažujeme vzorkování instancí datového souboru, u regresních úloh naproti tomu uvažujeme o transformaci závislé proměnné. U systémů využívající rozhodovací stromy opomíjíme výběr proměnných, důvodem je mechanismus fungování algoritmu i dosažené experimentální výsledky (Fridrich, 2019).

Úvodním krokem je sjednocení škály nezávislých proměnných, za tímto účelem je aplikována jedna z následujících metod – mocninná transformace, jenž u výstupu zaručí vlastnosti podobné Gaussovskému rozdělení; kvantilová transformace, jenž zajistí uniformní rozdělení výsledných dat, nebo robustní transformace, postavená na vzdálenosti mezi kvartily pro škálování dat i úpravy extrémních hodnot. Následným krokem je filtrování proměnných s nízkou variabilitou, na základě prahu nejmenšího přijatelného rozptylu. Výběr významných vysvětlujících proměnných je realizován prostřednictvím hierarchického aglomerativního shlukování, které si klade za cíl snížit kolinearitu nezávislých proměnných, klíčovým parametrem kroku je maximální počet proměnných. V rámci klasifikační úlohy je zvažováno vzorkování datového souboru, kde dochází k výběru mezi podvzorkováním nebo převzorkováním instancí, případně

zda nebude datový soubor ponechán v původním stavu. Volba a nastavení jednotlivých kroků je předmětem optimalizace vnějších parametrů řešení, čímž je zajištěna flexibilita systému.

4.4 Modelování

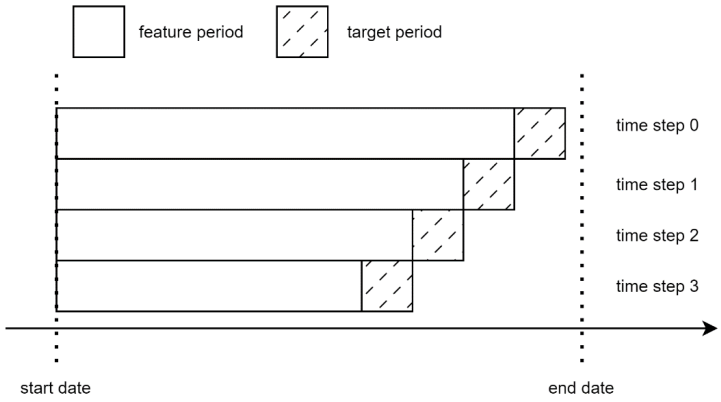
Modelováním postihuje autor vybrané aspekty návrhu, hodnocení a sestavení systému strojového učení. Nejprve je pro vybraný typ úlohy, algoritmus a časový výřez vymezen skelet řešení. Na základě dílčích prvků zpracování datového souboru, a modelování je pak sestaven prostor vnějších parametrů, který je s cílem dosažení přijatelné prediktivní schopnosti prohledáván. Optimalizovaný model je následně kalibrován a využit k tvorbě predikcí pro trénovací i testovací množiny dat. Na hodnocení modelu je pohlíženo prizmatem přirozených a ekonomických ukazatelů úspěšnosti, ale i dobou potřebnou ke konstrukci řešení. Dílčí kroky nastíněného procesu a vzájemné vazby jsou vyobrazeny na Obr. 25



Obr. 25 Proces konstrukce modelu strojového učení

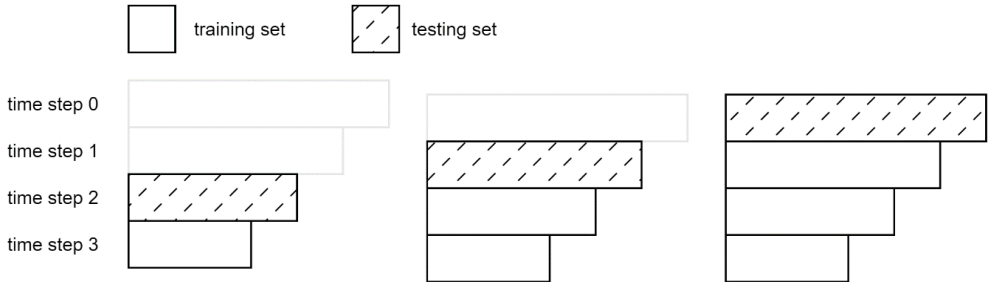
4.4.1 Dělení datového souboru

Za přístupem k dělení datového souboru je snaha napodobit podmínky aplikace prediktivního systému v reálném prostředí. Návrh experimentu je inspirován křížovou validací časových řad (Hyndman & Athanasopoulos, 2013), tj. zabezpečuje časové odlišení trénovací a testovací množiny dat. Autor staví na časově ohraničených řezech dat („time step“), jejichž rozsah je snižován kroky v délce časového období vysvětlované proměnné, směrem do minulosti. V rámci řezu dochází k oddělení dat pro sestavení vysvětlovaných („feature period“) a vysvětlujících („target period“) proměnných. Přístup k sestavení časových řezů je ilustrován s pomocí Obr. 26.



Obr. 26 Časově ohraničené řezy vstupního datového souboru, užité pro konstrukci zákaznického modelu

Časově ohraničené řezy jsou dále využity k vlastnímu dělení datového souboru, tj. pro testovací množinu z libovolného výřezu je konstruována odpovídající trénovací množina seskupením všech starších výřezů, viz Obr. 27. Zahrnutí více výřezů je motivováno experimentálními výsledky Gattermann-Itschert & Thonemann (2021), vede ale k porušení nezávislosti datových instancí.



Obr. 27 Využití časově ohraničených řezů při dělení datového souboru na trénovací a testovací množiny dat

4.4.2 Ukazatele úspěšnosti

Zhodnocení schopností systémů strojového učení je realizováno přirozenou, ale i podnikovou perspektivou řešených úloh. Predikční schopnosti klasifikačních modelů jsou odhadovány s pomocí ukazatelů založených na matici záměn (ACC, F1), nebo s ohledem na predikovanou pravděpodobnost příslušnosti k dané třídě (AUCROC). Predikční schopnosti regresních modelů jsou posuzovány s přihlédnutím k variabilitě závislé proměnné vysvětlené modelem (R^2), a úspěšnosti s ohledem na absolutní (MAE) a kvadratické (MSE) odchylky. Přirozené ukazatele úspěšnosti jsou popsány v podkapitole 1.3.2.

Podnikovou perspektivu retenční kampaně reflektuje autor originálním přístupem, který umožňuje odhadnout ekonomický dopad kampaně s ohledem na inkrementální zařazení zákazníka do retenční aktivity. Rozhodnutí o členství v cílové skupiny dané kampaně vychází z odhadu závislých proměnných. Pojetí staví na ukazatelích maximálního očekávaného zisku a maximálního dosaženého zisku, které jsou zevrubně popsány v podkapitole 4.1.1.

4.4.3 Klasifikační a regresní metody

Základním stavebním kamenem každého z prediktivních systémů jsou zvolené klasifikační a regresní algoritmy. Výběr vychází z teoretických východisek práce a rešerše vědecké literatury, zahrnuje zobecněné lineární modely, podpůrné vektory, neuronové sítě, rozhodovacích stromy a meta-algoritmy. Zvolené implementace jsou obsahem Tab. 9.

Tab. 9 Vybrané algoritmy strojového učení

Family	Algorithm	Abbr.	Implementation
generalized linear models	logistic regression	lr	Pedregosa et al. (2011)
	elastic net regression	lr	Pedregosa et al. (2011)
support vector machines	support vector machine with linear kernel	svm-lin	Pedregosa et al. (2011)
	support vector machine with approx. radial basis function kernel	svm-rbf	Williams & Seeger (2000), Pedregosa et al. (2011)
artificial neural networks	multi-layer perceptron	mlp	Chollet et al. (2015) Abadi et al. (2015)
decision trees	decision tree	dt	Pedregosa et al. (2011)
meta-learning (bagging)	random forest	rf	Pedregosa et al. (2011)
meta-learning (boosting)	gradient boosting machine	gbm	Ke et al. (2017)

Autor spoléhá na implementace obsažené v knihovně scikit-learn (Pedregosa et al., 2011), výjimku tvoří podpůrné vektory s radiální jádrovou funkcí, které staví na kombinaci explicitní jádrové transformace a lineárních podpůrných vektorů (Williams & Seeger, 2000), což vede k zásadnímu snížení výpočetní složitosti. Vybočují i umělé neuronové sítě, pro které je využito knihoven Keras (Chollet et al., 2015), respektive Tensorflow (Abadi et al., 2015). Sítě autor staví na dopředné vícevrstvé architektuře, další prvky zahrnují nenasycené aktivační funkce a inicializaci vnitřních vah (Glorot & Bengio, 2010), dávkovou normalizaci (Ioffe & Szegedy, 2015), ale i různé přístupy k iterativní optimalizaci vah (Kingma & Ba, 2014; Hinton et al., 2018). Cílem je snížení počtu epoch potřebných ke konvergenci účelové funkce, při eliminaci odpojených uzlů, i mizejícího/explodujícího gradientu. Další výjimkou jest meta-algoritmus boosting, pro který autor volí knihovnu LightGBM (Ke et al., 2017), která rozšiřuje běžné metody o vzorkování instancí v závislosti na velikosti gradientu a redukci vylučujících se

vysvětlujících proměnných, což vede ke zlepšení prediktivních schopností, i snížení výpočetní složitosti.

K nastavení vnějších parametrů řešení je přistupováno s pomocí Bayesovské optimalizace (Bergstra et al., 2013) o 25 iteracích, s proxy modelem parzenového stromu, a akviziční funkcí maximalizující očekávané zlepšení účelové funkce. Vyhodnocení účelové funkce zahrnuje konstrukci systému s dílčí sadou parametrů na 60 % instancí dat a ohodnocení na zbývajících 40 % instancí. Metrikou optimalizovanou v rámci klasifikační části úlohy je F1, pro regresní úlohu potom uvažujeme R^2. Prostory vnějších parametrů jsou obsahem přílohy B2.

4.5 Vyhodnocení a interpretace

V podkapitole se zabýváme jak přístupem k posouzení prediktivních schopností kandidátních řešení, tak i porozuměním modelovanému jevu. Obě aktivity reflektují jak přirozené prediktivní schopnosti systémů, tak ekonomické dopady souvisejících aktivit.

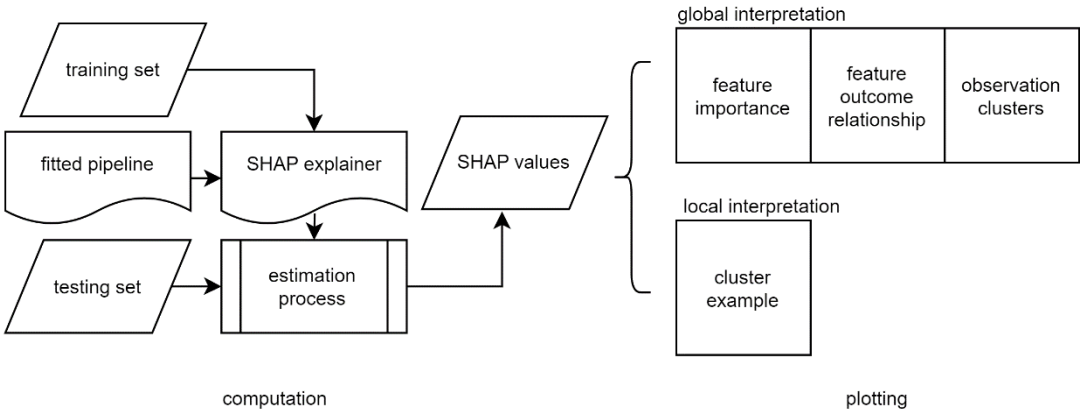
4.5.1 Vyhodnocení prediktivní schopnosti modelů

Posouzení z hlediska přirozených ukazatelů prediktivních schopností a doby potřebné pro konstrukci systému je realizováno prostým srovnáním odhadů středních hodnot a příslušných intervalů spolehlivosti na testovacích řezech dat. Ověření významnosti rozdílů s pomocí statistických testů autor opomíjí, především s ohledem na nízký počet časových řezů a nezbytnost kompenzace opakovaného testování (viz Mittelhammer et al., 2000). Prediktivní řešení jsou dále diagnostikována z hlediska kompromisu mezi vychýlením a rozptylem predikcí.

Ekonomický dopad na modelech založených retenčních aktivit je doložen s pomocí očekávaného a skutečného zisku, respektive odpovídajícími odhady středních hodnot a intervalů spolehlivosti. Statistický význam rozdílů v dosaženém zisku úspěšných klasifikačních a regresních přístupů je vyhodnocen s pomocí Wilcoxonova testu. Pro bližší porozumění schopnostem řešení řadit zákazníky dle inkrementálního zisku je na úlohu nahlíženo s pomocí kumulativních křivek očekávaného a skutečného kumulativního zisku zařazení daného zákazníka do kampaně v aktuálním časovém řezu, tj. neporovnáváme pouze celkového hodnoty maximálního zisku, ale i vlastní mechanismus řazení.

4.5.2 Interpretace vybraných modelů

Pro další interpretaci jsou pro každou datovou sadu vybrány takové systémy, které spolehlivě identifikují rizikové zákazníky, případně vedou k maximalizaci ekonomického výsledku zamýšlené retenční kampaně. Vlastní porozumění je založeno na SHAP metodách (Lundberg & Lee, 2017), motivací je teoretické ukotvení Shapleyho hodnot, ucelené pojetí globální a lokální srozumitelnosti, tj. porozumění vlivu vysvětlujících proměnných na predikce modelu zůstávají konzistentní. Výzvou je naopak výpočetní složitost. Zevrubný popis metod je obsahem podkapitoly 1.3.2.



Obr. 28 Proces konstrukce artefaktů pro interpretaci systému strojového učení

Přístup k interpretaci prediktivních modelů lze rozlišit na výpočetní a prezentační část., kde si první ze jmenovaných klade za úkol efektivní odhad Shapleyho hodnot, druhá pak přístupnou vizuální prezentaci zachycených znalostí. Podobu jednotlivých kroků ilustruje Obr. 28.

Výpočetní část řešení vychází z SHAP přístupů, k odhadu Shapleyho hodnot dochází skrze permutace antiteticky vzorkovaných vysvětlujících proměnných, což zajišťuje kýžené vlastnosti výsledných odhadů, včetně interakcí mezi atributy. S ohledem na rozsah datových souborů a výpočetní složitost je celý proces výpočtu implementován tak, aby jej bylo možné efektivně paralelizovat na distribuovaném výpočetním clusteru. Současně omezujeme velikost testovacího časového řezu na tisíc náhodně vybraných pozorování, alternativně lze použít podmnožinu zákazníků, která je v popředí zájmu, tj. zákazníků s kladným predikovaným výsledkem zařazení do retenční kampaně.

V prezentační části je na spočtené SHAP hodnoty nahlíženo perspektivami, které charakterizují celkové latentní vztahy reflektované modelem (globální interpretace), dále také skrze

predikci modelu pro vybranou datovou instanci (lokální interpretace). Užité vizuální prvky rozšiřují původní nástroje představené především o souvislosti s vlastnostmi původního datového souboru a zaměření na shluky pozorování.

V globální perspektivě představuje práce reflexi vztahů skrze atributy, ale i skupiny instancí datového souboru. Nejprve dochází k odhadu dopadu vysvětlujících proměnných s pomocí průměrné absolutní hodnoty SHAP, tj. určení celkového významu vysvětlující proměnné je nezávislé na směru předpokládaného vztahu. Pro nejvýznamnější proměnné autor konstruuje bodové diagramy vztahu mezi SHAP hodnotami a zkoumanou veličinou, pro bližší porozumění charakteru vztahu a robustnosti odhadu SHAP hodnot je uvažováno i o rozložení veličin. Pohled skrze instance datového souboru využívá hierarchického aglomerativního shlukování pro tvorbu zákaznických shluků, které jsou podobné napříč SHAP hodnotami, vhodný počet shluků je pro potřeby práce odhadnut pomocí siluety (Rousseeuw, 1987). Na shluky nahlížíme prostřednictvím rozložení SHAP hodnot členských pozorování a současně určením pěti nejvýznamnějších faktorů shluku, včetně směru působení daného faktoru.

V lokální perspektivě se autor zaměřuje na reprezentativní pozorování odlišných shluků, tj. zákazníků u nichž předpokládáme, že setrvají nebo budou ztraceni. Reprezentativním rozumíme takové pozorování, které je nejblíže těžišti shluku. Pro vlastní interpretaci predikce je využito horizontálního sloupcového grafu síly a směru působení vysvětlujících proměnných, který je doplněn o bodový kobercový graf, s pozicí daného pozorování. Díky tomuto přístupu je možné identifikovat jednak směr působení dané veličiny na dané pozorování i polohu pozorování v rámci zkoumaného datového souboru.

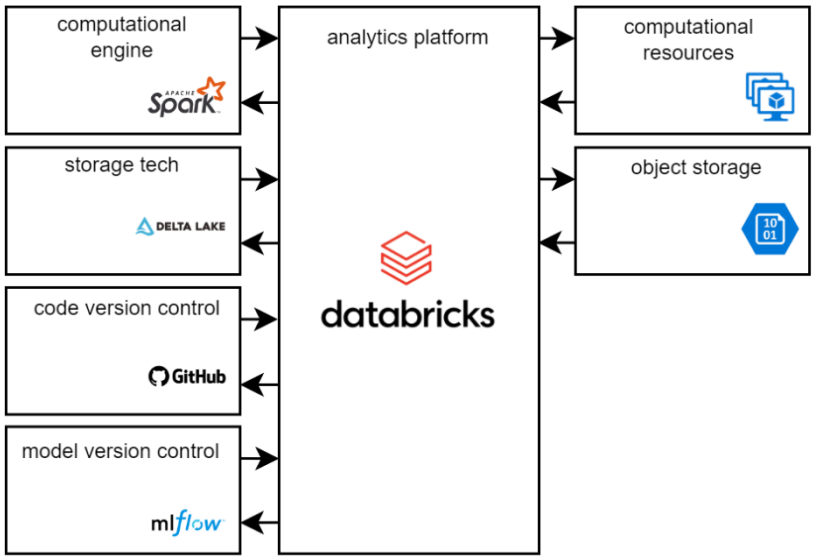
4.6 Aplikace řešení

Aplikací řešení rozumíme praktické výstupy realizovaného výzkumného úsilí, kam lze řadit datové soubory, programový kód datového produktu, a v širším smyslu i text disertační práce. Na datovou reprezentaci i kód odkazuje příloha B1. V aktuální sekci je věnována pozornost praktickým aspektům konstrukce systému strojového učení jako je technologická koncepce řešení nebo uvažované provozní náklady.

S ohledem na rozsah vstupních dat a výpočetní náročnost potřebných operací je datový produkt realizován v cloudovém prostředí Microsoft Azure. Základním stavebním kamenem je analytická platforma Databricks, spojující nástroje potřebné pro vývoj, nasazení, provoz a údržbu rozsáhlých datových produktů (Etaati, 2019). Oblíbená je především pro agnostický

přístup k poskytovatelům cloudových služeb, bezpečnost, integraci cloudových uložišť, případně alokaci a řízení výpočetních prostředků. Srdcem platformy je distribuovaný výpočetní systém Apache Spark (Zaharia et al., 2016), který využíváme za účelem zpracování velkých dat, i pro distribuci náročných výpočtů.

Podpůrné komponenty zahrnují běžné objektové úložiště Azure Blob Storage, pro surová vstupní data; vrstvu Delta Lake, zajišťující potřebné vlastnosti transakcí, vynucení schémat, správu metadat i verzování objektů, kterou využíváme pro organizaci zpracovaných tabulárních dat; nástroj Mlflow pro správu životního cyklu systémů strojového učení, včetně monitorování, verzování, a produkčního nasazení; v neposlední řadě také GitHub pro distribuované verzování kódu.



Obr. 29 Technologická koncepce řešení

Použité typy virtuálních strojů, včetně přibližné doby běhu a souvisejících nákladů, jsou obsahem Tab. 10. Na veličiny je nahlíženo odpovídajícími kroky navrženého řešení, tj. skrz zpracování dat (bez optimalizace parametrů pro doporučovací systémy), modelování a interpretaci vybraných modelů. Pro každý z kroků jsme zvolili přiměřenou konfiguraci clusteru, kde první prvek odpovídá řídícímu uzlu, druhý potom počtu a typu uzlů pracovních. Součástí nákladů jsou také licence analytické platformy a datové úložiště (viz Microsoft, 2022).

Tab. 10 Zvolené výpočetní prostředky, doba výpočtu jednotlivých kroků a odhad nákladů navrženého systému

	Retail Rocket			REES46		
Step	**ingestion**	**modeling**	**interpret.**	**ingestion**	**modeling**	**interpret.**
Cluster setup	F8 + 4x F4	DS4 + 5x DS3	DS3 + 10x DS3	F8 + 4x F4	DS4 + 5x DS3	DS3 + 10x DS3
Computational time [h]	0.25	1.5	0.10	0.75	39	0.42
Compute costs [USD/h]	2.42	3.73	5.86	2.42	3.73	5.86
Licence costs [USD/h]	0.90+	1.58+	2.48+	0.90+	1.58+	2.48+
Costs per run [USD]	0.83	7.96	1.39	2.49	206.93	3.47
Storage costs [USD/month]		9.60*			38.40**	

+Databricks, premium-tier scheduled job; * standard 128 GB SSD; ** standard 512 GB SSD

Pokud bychom uvažovali opakované trénování všech modelů na měsíční bázi, se stejným počtem a velikostí časových řezů, pak pro datový soubor Retail Rocket můžeme předpokládat náklady ve výši ~ 20 USD, pro REES46 by se jednalo o ~ 250 USD. Náklady by bylo možné snížit díky využití specializovaných výpočetních prostředků pro konstrukci umělých neuronových sítí, které nebyly autorovi k dispozici, úplnému vypuštění neuronových sítí ze srovnání (90 % výpočetního času ve fázi modelování pro datový soubor REES46), případně uzavření smlouvy se závazkem k čerpání služeb Microsoft Azure. Je zřejmé, že v případě implementace systému v prostředí podniku, leží nosná část nákladů v integraci se stávajícími technickými a organizačními systémy.

5 Dosažené výsledky

5.1 Retail Rocket

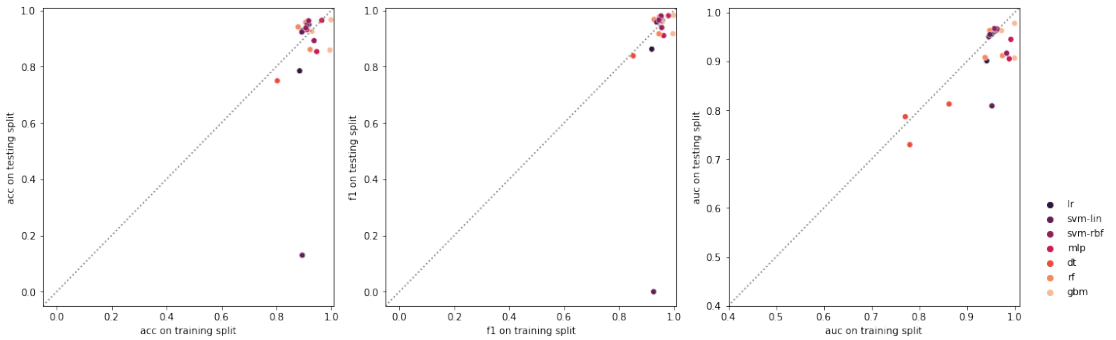
5.1.1 Přirozené ukazatele úspěšnosti

Pokud je na schopnosti klasifikačních systémů nahlíženo perspektivami ACC a F1, pak se nejúspěšnějším řešením zdají být podpůrné vektory s radiální jádrovou funkcí a náhodné lesy, dobrých výsledků dosahují i umělé neuronové sítě a gradient boosting. Nevhodným přístupem se naopak ukazuje metoda podpůrných vektorů s lineární jádrovou funkcí. V rámci AUCROC dominují podpůrné vektory s radiální jádrovou funkcí a gradient boosting, zajímavý je propad schopností rozhodovacích stromů ve srovnání s ukazateli ACC a F1. S ohledem na dobu potřebnou pro konstrukci řešení na jednom časovém řezu se zdají být jednotlivá řešení srovnatelná, vyčnívají pouze umělé neuronové sítě.

Tab. 11 Ukazatele klasifikační úspěšnosti – Retail Rocket

Algorithm	ACC	F1	AUCROC	Optimization runtime [min]	Refit runtime [min]
lr	0.888 (0.665, 1.110)	0.932 (0.781, 1.084)	0.939 (0.857, 1.021)	0.6 (0.2, 0.9)	0.2 (0.1, 0.2)
svm-lin	0.668 (-0.490, 1.825)	0.644 (-0.742, 2.030)	0.905 (0.698, 1.111)	0.5 (0.5, 0.6)	0.2 (0.1, 0.2)
svm-rbf	0.931 (0.842, 1.020)	0.962 (0.910, 1.014)	0.950 (0.878, 1.022)	0.5 (0.5, 0.6)	0.2 (0.2, 0.2)
mlp	0.917 (0.775, 1.059)	0.952 (0.861, 1.043)	0.938 (0.864, 1.012)	3.6 (-1.1, 8.2)	1.1 (-0.2, 2.4)
dt	0.884 (0.595, 1.172)	0.928 (0.735, 1.122)	0.776 (0.671, 0.882)	0.5 (0.5, 0.5)	0.2 (0.1, 0.2)
rf	0.920 (0.792, 1.049)	0.954 (0.873, 1.035)	0.928 (0.851, 1.004)	0.6 (0.6, 0.7)	0.2 (0.2, 0.3)
gbm	0.917 (0.783, 1.052)	0.953 (0.871, 1.035)	0.949 (0.856, 1.042)	0.9 (0.3, 1.5)	0.2 (0.1, 0.3)

Pro bližší porozumění schopnostem jednotlivých klasifikačních přístupů autor konstruuje Obr. 30, který umožňuje porovnat predikční schopnosti na trénovacích a testovacích množinách dat, osu prvního kvadrantu označuje tečkovaná čára. Perfektní řešení leží v pravém horním rohu, body pod osou kvadrantu značí příliš složitý model (rozptyl predikce), body nad osou kvadrantu naopak značí příliš jednoduchý model (vychýlení). Zdá se, že řešení selhávají v rozptylu predikcí na nejzazším časovém řezu, což se následně projevuje i šíři intervalů spolehlivosti. Ke zmírnění neduhu by bylo možné přistoupit s využitím více dat, restriktivním pojetím výběru vysvětlujících proměnných i vnitřní struktury modelů. Problematické výsledky rozhodovacích stromů v perspektivě AUCROC lze pozorovat napříč časovými řezy, s ohledem na uspokojivé výsledky v perspektivách ACC a F1 může být na vině kalibrace pravděpodobnosti.



Obr. 30 Kompromis vychýlení a rozptylu klasifikačních řešení napříč časovými řezy – Retail Rocket

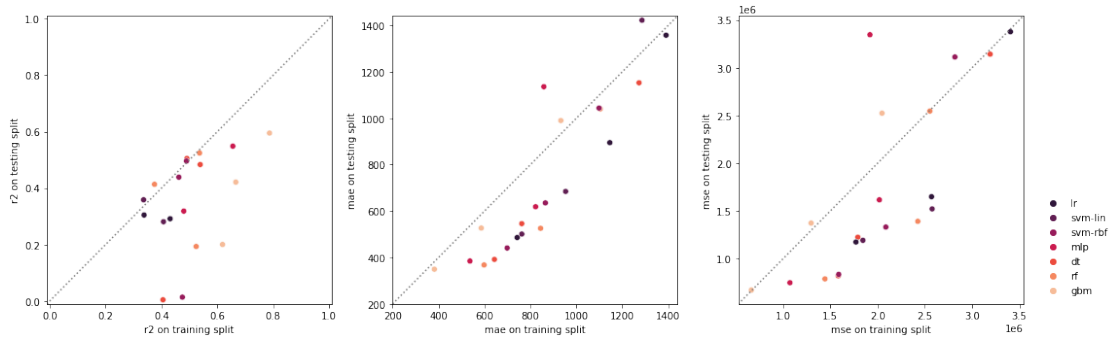
Uchopíme-li úspěšnost regresního systému perspektivami R^2, MAE i MSE, pak se nejúspěšnějšími zdají být gradient boosting a náhodné lesy. Problematické se ukazují podpůrné vektory s lineární jádrovou funkcí a lineární regrese. Z pohledu na řády MAE a MSE se zdá, že regresní modely selhávají při predikci prominentních instancí testovací množiny dat, což je možné blíže diagnostikovat analýzou residuí. Doba potřebná pro konstrukci řešení na jednom časovém řezu odpovídá klasifikační větvi řešení, tj. pozorované hodnoty jsou srovnatelné pro všechny algoritmy, výjimku tvoří umělé neuronové sítě.

Tab. 12 Ukazatele regresní úspěšnosti – Retail Rocket

Algorithm	R^2	MAE	MSE	Optimization time [min]	Refit time [min]
lr	0.175 (-0.353, 0.704)	912.9 (-170.8, 1996.6)	2.07E+06 (-8.12E+05, 4.95E+06)	0.5 (0.5, 0.6)	0.2 (0.2, 0.2)
svm-lin	0.147 (-0.601, 0.896)	869.3 (-342.2, 2080.8)	2.17E+06 (-1.33E+06, 5.66E+06)	0.5 (0.5, 0.6)	0.2 (0.1, 0.2)
svm-rbf	0.316 (-0.338, 0.971)	706.6 (-58.1, 1471.4)	1.76E+06 (-1.21E+06, 4.73E+06)	0.5 (0.5, 0.6)	0.2 (0.1, 0.2)
mlp	0.269 (-0.494, 1.033)	713.1 (-241.9, 1668.1)	1.90E+06 (-1.38E+06, 5.19E+06)	3.8 (-3.4, 11)	2.2 (-3.3, 7.8)
dt	0.331 (-0.372, 1.034)	696.8 (-302.8, 1696.5)	1.73E+06 (-1.35E+06, 4.81E+06)	0.5 (0.5, 0.5)	0.2 (0.1, 0.2)
rf	0.377 (-0.042, 0.796)	644.5 (-228.9, 1518)	1.58E+06 (-6.43E+05, 3.79E+06)	0.6 (0.5, 0.7)	0.3 (0, 0.5)
gbm	0.406 (-0.086, 0.897)	622.1 (-199.7, 1443.8)	1.52E+06 (-8.00E+05, 3.85E+06)	0.8 (0.2, 1.4)	0.2 (0, 0.4)

Další pohled na způsobilost jednotlivých regresních přístupů ilustruje Obr. 31. Perfektní řešení leží pro R^2 v pravém horním rohu, pro MAE a MSE naopak v levém dolním rohu, což upravuje vztah mezi povahou selhání a osou kvadrantu. Perspektiva R^2 odkrývá příliš složité modely na nejzazším časovém řezu, současně také neschopnost velké části řešení vysvětlit více než 60 % variability i na trénovací množině dat. Pohled na MAE a MSE naznačuje problémy všech přístupů na nejzazším časovém řezu (viz pravá horní část grafu), u zbylých řezů dochází

naopak k nedostatečnému využití trénovací množiny dat. Úspěšnost na obtížném časovém řezu je možné zlepšit využitím více dat, restriktivním přístupem k výběru vysvětlujících proměnných i vnitřní struktuře modelů. Nedostatečné využití trénovací množiny dat je možné adresovat konstrukcí dalších vysvětlujících proměnných, případně složitější vnitřní strukturou modelu.



Obr. 31 Kompromis vychýlení a rozptylu regresních řešení napříč časovými řezy – Retail Rocket

5.1.2 Ekonomický dopad retenční kampaně

Z pohledu na ekonomický dopad řešení je možné pozorovat značný propad mezi očekávaným a dosaženým ziskem retenční kampaně pro oba typy úloh. Příčinou je nesoulad mezi modelovanými veličinami i nedostatečné schopnosti některých prediktivních systémů. Při srovnání modelů stejného typu lze upozornit na vyšší střední hodnoty skutečného ekonomického dopadu retenční kampaně u regresních modelů založených na rozhodovacích stromech. 95% intervaly spolehlivosti jsou u této veličiny velmi široké, tj. může nabývat i záporných hodnot.

Klasifikačním modelům jednoznačně dominují podpůrné vektory s radiální jádrovou funkcí, při jejichž použití pro řazení zákaznické báze by bylo dosaženo střední hodnoty zisku 8.40E+04 CU, při zahrnutí 15.2 % zákazníků do kýžené retenční aktivity. Následují umělé neuronové sítě dosahující střední hodnoty zisku 7.95E+04 CU, při oslovení 11.0 % zákazníků, případně gradient boosting dosahující střední hodnoty zisku 7.92E+04 CU, při oslovení 14.6 % zákazníků.

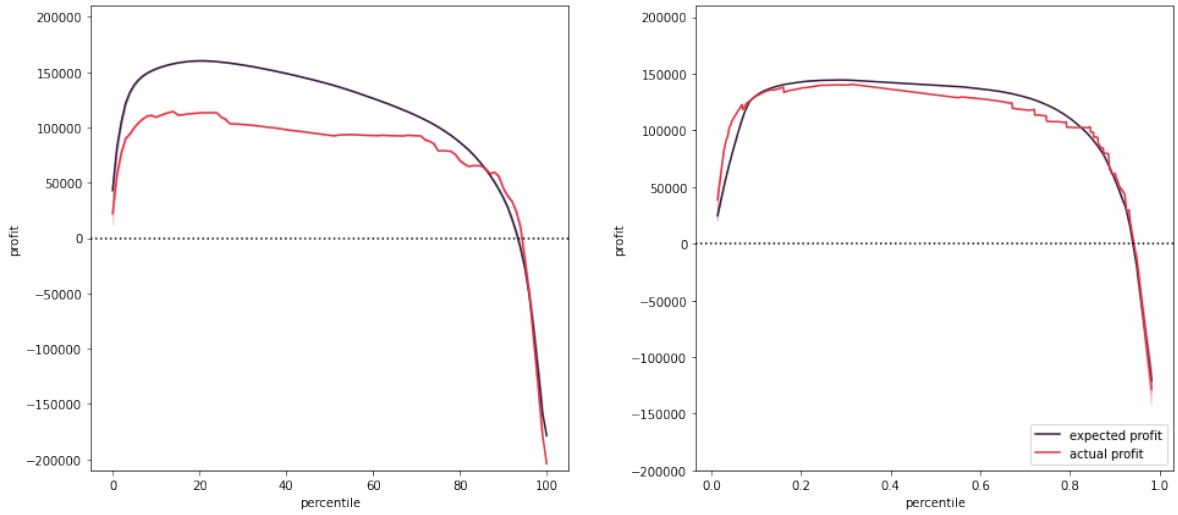
Mezi regresními modely vyčnívají rozhodovací stromy, se střední hodnotou zisku 9.54E+04 CU při zahrnutí 25.4 % zákazníků do retenční kampaně. Následují gradient boosting dosahující střední hodnoty zisku 8.69E+04 CU při oslovení 34.9 % zákazníků, případně náhodné lesy dosahující střední hodnoty zisku 8.04E+04 CU při zahrnutí 29.9 % zákazníků.

Tab. 13 Ukazatele ekonomického dopadu retenční kampaně – Retail Rocket

Algorithm	classification		regression	
	$\Pi^{expected}$	Π^{actual}	$\Pi^{expected}$	Π^{actual}
lr	1.11E+05 (-8.06E+04, 3.02E+05)	7.31E+04 (-5.22E+04, 1.98E+05)	1.69E+05 (-1.91E+05, 5.28E+05)	6.44E+04 (-7.51E+04, 2.04E+05)
svm-lin	1.11E+05 (-1.33E+05, 3.55E+05)	6.39E+04 (-6.33E+04, 1.91E+05)	1.48E+05 (-1.06E+05, 4.02E+05)	4.91E+04 (-4.64E+04, 1.45E+05)
svm-rbf	1.11E+05 (-7.13E+04, 2.93E+05)	8.40E+04 (-2.54E+04, 1.93E+05)	1.08E+05 (-9.75E+04, 3.14E+05)	7.67E+04 (-5.75E+04, 2.11E+05)
mlp	1.03E+05 (-7.79E+04, 2.84E+05)	7.95E+04 (-5.46E+04, 2.14E+05)	1.41E+05 (-3.58E+04, 3.18E+05)	7.93E+04 (-3.04E+04, 1.89E+05)
dt	1.07E+05 (-4.98E+04, 2.64E+05)	4.73E+04 (-2.11E+04, 1.16E+05)	8.78E+04 (-8.75E+04, 2.63E+05)	9.54E+04 (-5.13E+04, 2.42E+05)
rf	1.04E+05 (-8.08E+04, 2.89E+05)	5.95E+04 (-1.69E+04, 1.36E+05)	7.99E+04 (-3.74E+04, 1.97E+05)	8.04E+04 (-4.57E+04, 2.06E+05)
gbm	1.09E+05 (-6.98E+04, 2.88E+05)	7.92E+04 (-4.51E+04, 2.04E+05)	1.17E+05 (-3.77E+04, 2.72E+05)	8.69E+04 (-3.90E+04, 2.13E+05)

Ze srovnání nejlepších regresních a klasifikačních přístupů vyplývá, že využití regresního rozhodovacího stromu oproti klasifikačním podpůrným vektorům s radiální jádrovou funkcí, vede v průměru ke zlepšení dosaženého zisku o ~ 13.6 %, což odpovídá ~ 11415.4 CU. Modely dále porovnáváme s pomocí párového Wilcoxonova testu, napříč časovými řezy. Nulová hypotéza předpokládá, že střední hodnota rozdílu zisků vybraného regresního a klasifikačního modelu je symetrická kolem nuly. Alternativní jednostranná hypotéza předpokládá, že střední hodnota tohoto rozdílu je vyšší než u rozdělení symetrického kolem nuly. Hladina významnosti byla stanovena na úrovni $\alpha = 0.1$. Pro regresní rozhodovací strom a klasifikační podpůrné vektory s radiální jádrovou funkcí dosahuje test součtu pořadí rozdílů zisku, které jsou vyšší než nula, hodnoty 5, což odpovídá p-hodnotě 0.250. Nepodařilo se tedy zamítnout hypotézu alternativní, tj. pozorované kladné rozdíly mezi modely mohou být nahodilé.

Pro bližší porozumění způsobu řazení zákazníků dle očekávaného inkrementálního zisku kampaně konstruuje autor Obr. 32, který porovnává očekávaný a dosažený kumulativní zisk kampaně v aktuálním časovém řezu pro uvažované klasifikační a regresní přístupy, tj. v tomto případě se jedná o řešení postavená na metodě podpůrných vektorů a rozhodovacích stromech. Klasifikační model vede k enormnímu podhodnocení odhadu až do 90 percentilu zákaznické báze, regresní model naproti tomu těsně popisuje kumulativní křivku skutečného zisku kampaně. U obou přístupů pozorujeme akumulaci chyb, která vede k posunu křivek nad 95 percentilem.



Obr. 32 Očekávaný a skutečný zisk retenční kampaně pro klasifikační podpůrné vektory s radiální jádrovou funkcí (vlevo), a regresní rozhodovací strom (vpravo) – Retail Rocket

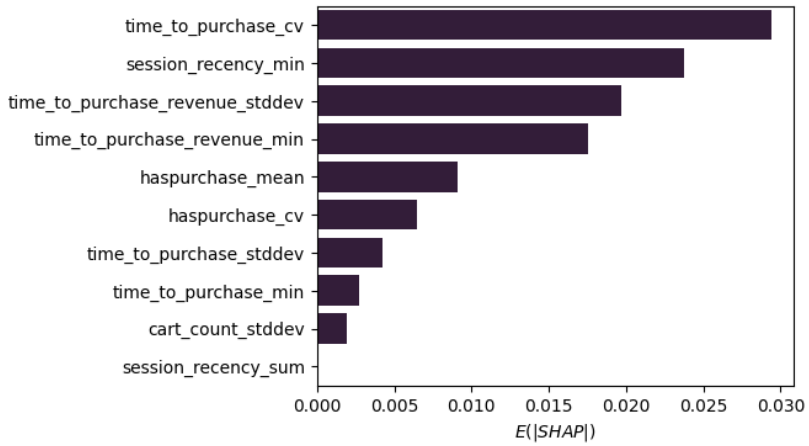
5.1.3 Interpretace vybraných modelů

Predikce ztráty zákazníka

Na základě výsledků dosažených v přirozených ukazatelích predikčních schopností byl pro další interpretaci vybrán systém strojového učení využívající podpůrné vektory s radiální jádrovou funkcí. Pro daný časový řez byl optimalizací vnějších parametrů konstruováno řešení využívající kvantilové transformace, striktní výběr vysvětlujících proměnných a převzorkování dostupných pozorování. Vlastní klasifikační model je komplexní ve smyslu počtu dimenzí aproximace jádrové transformace; L2 regularizace je slabší, tj. přeučení brání především nízký počet vysvětlujících proměnných.

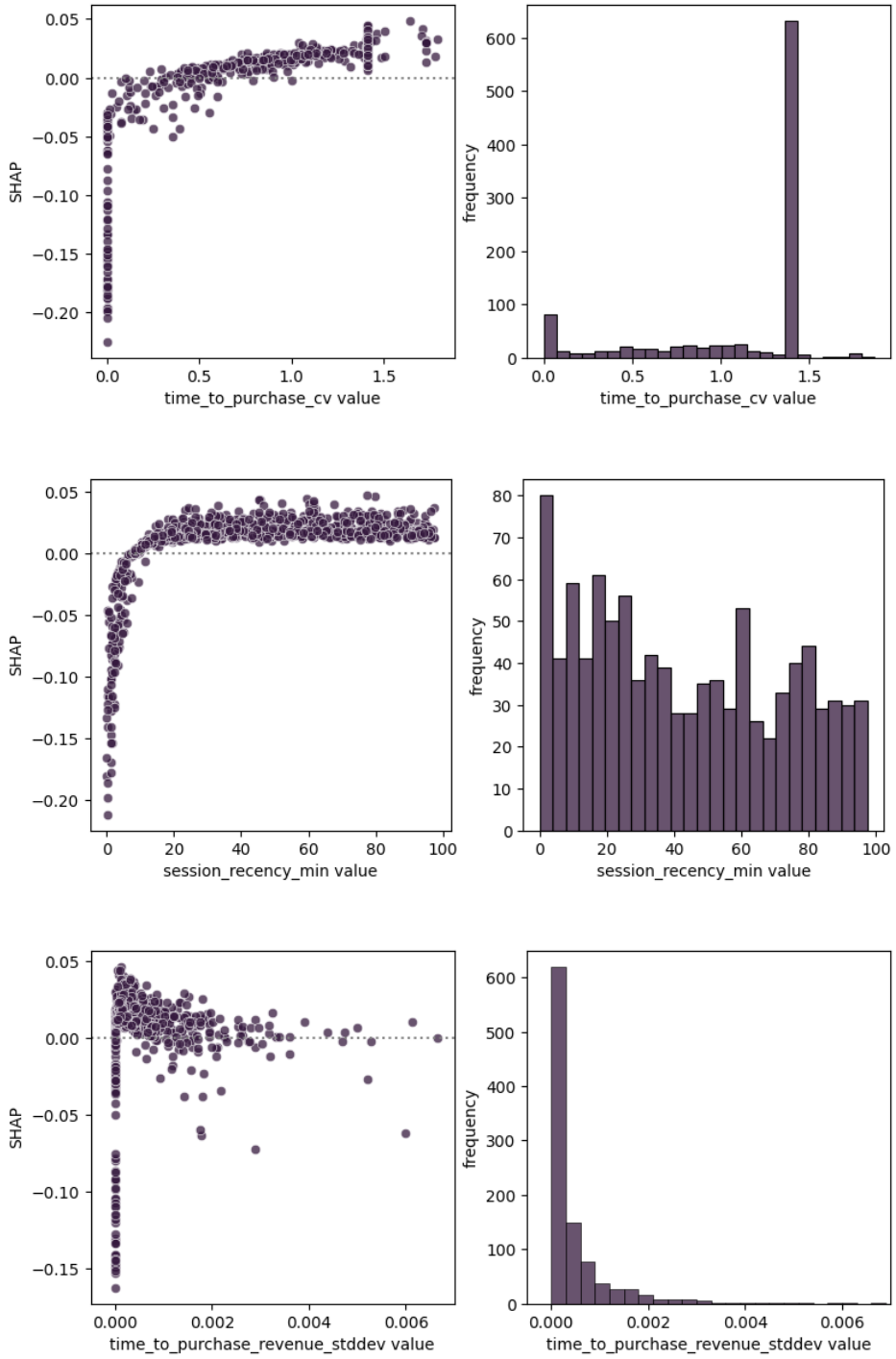
Vysvětlující proměnné, které mají v průměru nejvyšší absolutní dopad na predikci ztráty zákazníka, ilustruje horizontální sloupcový graf na Obr. 33. Vynikají především variační koeficient průměrného počtu minut mezi nákupními interakcemi, doba uplynulá od poslední návštěvy a směrodatné odchylky výběru pro průměrnou dobu vedoucí k nákupu v objemu 1 CU. Z hlediska zastoupení množin vysvětlujících proměnných pozorujeme stáří (session_recency_min) a frekvenci uživatelských interakcí (haspurchase_mean, haspurchase_cv, cart_count_stddev), případně ostatní atributy (time_to_purchase_cv, time_to_purchase_stddev, time_to_purchase_min, time_to_purchase_revenue_stddev, time_to_purchase_revenue_min). Dopad posledního z atributů, session_recency_sum, je nulový, tj. podpůrné vektory využívají

pouze devět vysvětlujících proměnných. Zdá se, že podpůrné vektory spoléhají na některé vlastnosti uživatelských interakcí, zásadní je však chování uvnitř relace.



Obr. 33 Vysvětlující proměnné významné pro predikci odchodu zákazníka, s využitím podpůrných vektorů – Retail Rocket

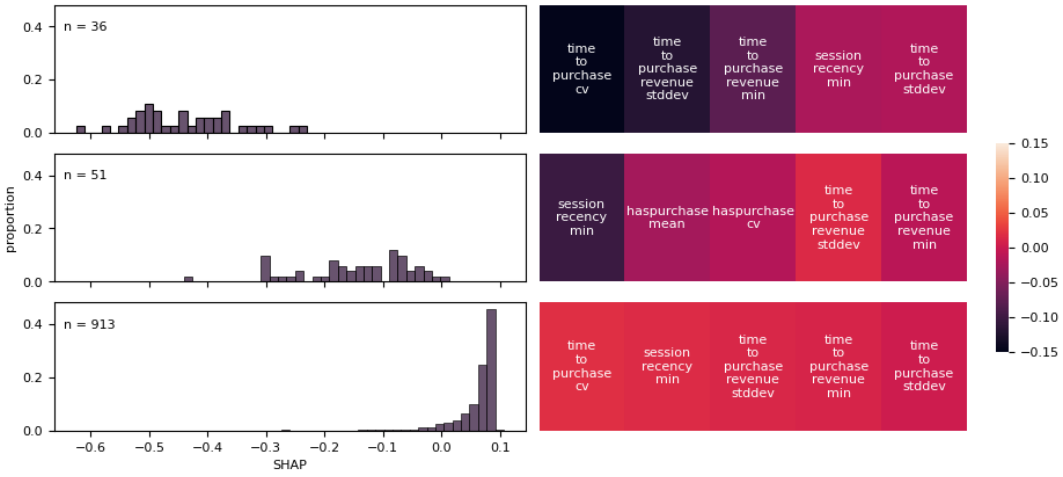
Charakter vztahu mezi vysvětlujícími proměnnými a jejich vlivem na predikce klasifikačního modelu, ilustruje Obr. 34, využívající bodových diagramů SHAP hodnot, robustnost ilustrují histogramy veličin. Vysvětlující proměnnou se zásadním dopadem na predikci modelu je variační koeficient doby mezi transakčními interakcemi. Zdá se, že s rostoucím variačním koeficientem dochází k poklesu rizika odhodu zákazníka. Věrní zákazníci tíhnou k realizaci krátce po sobě jdoucích transakčních interakcí uvnitř uživatelské relace (time_to_purchase_cv<0.15), u ohroženějších zákazníků naopak dochází k prodlevám (time_to_purchase_cv>1.25). Další významnou proměnou je stáří poslední uživatelské relace. Ukazuje se, že s rostoucím stářím relace roste i pravděpodobnost odchodu zákazníka, která se přibližně po třech týdnech stabilizuje. U setrvávajících zákazníků není zpravidla poslední relace starší než týden, u ohrožených zákazníků uvažujeme o třech a více týdnech. Význačná je také směrodatná odchylka času, sloužícího k realizaci transakční interakce v objemu 1 CU. Pro velmi nízké hodnoty nelze považovat charakter vztahu za funkční (time_to_purchase_revenue_stddev~0.000), nad touto hranicí pozorujeme s rostoucí hodnotou směrodatné odchylky pokles rizika ztráty zákazníka.



Obr. 34 SHAP hodnoty proměnných, významných pro klasifikaci ohrožených zákazníků, včetně pozorovaného rozdělení – Retail Rocket

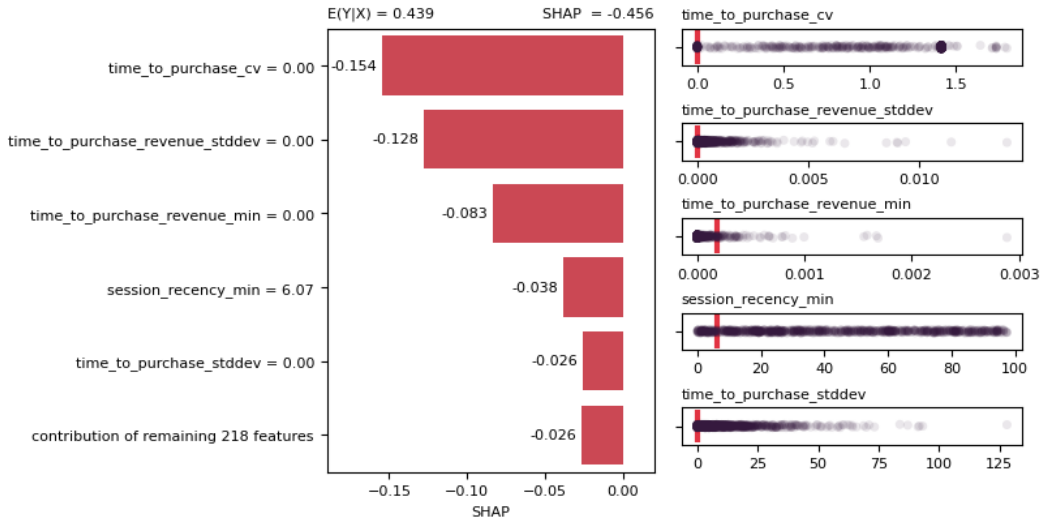
Cenný vhled umožňují i shluky zákazníků, jejichž SHAP hodnoty napříč vysvětlujícími proměnnými sledují podobné vzorce chování. Shluk tak odpovídá pozorováním, u kterých model uvažuje podobný charakter vztahů mezi vysvětlovanou a vysvětlujícími proměnnými. Zákaznické shluky pro klasifikační model ilustruje Obr. 35, skupiny jsou charakterizovány

histogramy součtů SHAP hodnot, a teplotními mapami, které popisují strukturu a směr působení významných vysvětlujících proměnných. První shluk odpovídá zákazníkům, u kterých lze předpokládat nižší pravděpodobnost ztráty, daný shluk vykazuje vysoký rozptyl SHAP hodnot. Za poklesem rizika stojí především nižší variační koeficient časových úseků mezi transakčními interakcemi, směrodatná odchylka doby vedoucí k interakci ve výši 1 CU a nejmenší doby vedoucí k interakci ve výši 1 CU. Druhý ze shluků obsahuje ohrožené zákazníky, u nichž si je klasifikační model méně jistý, shluk také trpí vysokým rozptylem SHAP hodnot. Za nejistotou modelu stojí nízké stáří poslední relace, současně s vyšším konverzním poměrem a variačním koeficientem indikátoru nákupů. Třetí shluk popisuje zákazníky s vyšší pravděpodobností odchodu v následujícím období, SHAP hodnoty zde dosahují nižšího rozptylu. Za růstem rizika stojí především nižší variační koeficient doby od poslední transakční interakce, vyšší stáří poslední relace, a směrodatná odchylka doby vedoucí k interakci ve výši 1 CU.



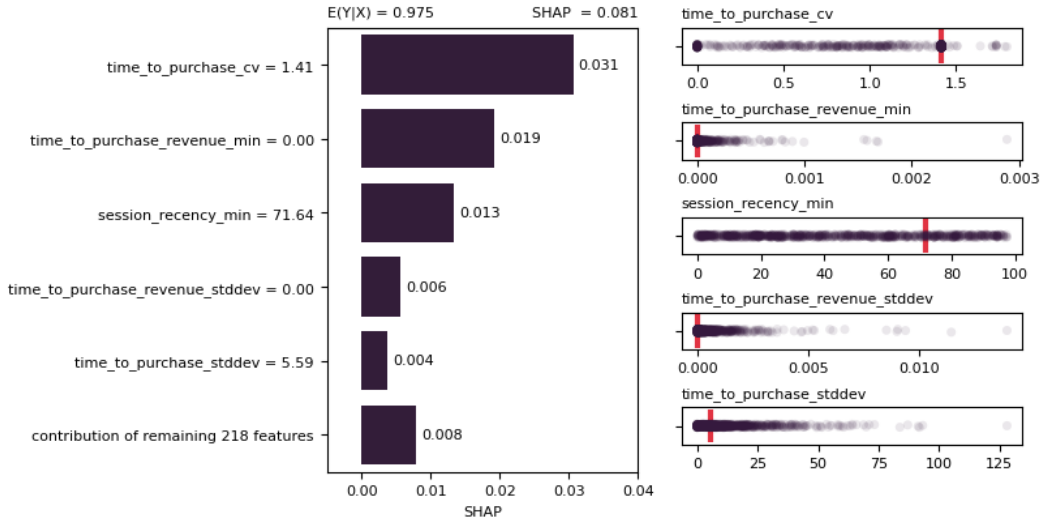
Obr. 35 Zákaznické shluky SHAP hodnot v klasifikaci ohrožených zákazníků, včetně klíčových vysvětlujících proměnných – Retail Rocket

Pro další porozumění zákaznickým shlukům je využito lokální interpretovatelnosti, kde je pro každou skupinu zákazníků určeno pozorování, které je nejblíže jejímu těžišti. Na vybraného zákazníka pak nahlížíme prostřednictvím významných SHAP hodnot i skutečných hodnot vysvětlujících proměnných. Přístup je ilustrován protiklady zákazníků setrvávajících (první shluk) a ohrožených (třetí shluk). Pozorováním reprezentativní pro první shluk zobrazuje Obr. 36, sestrojený pro zákazníka s identifikátorem user_id 1000093, u kterého předpokládáme ztrátu s pravděpodobností 0.439. Pokles pravděpodobnosti oproti očekávané hodnotě vychází především z chování uvnitř uživatelské relace, v kterých uživatel realizoval nákup právě jednoho produktu.



Obr. 36 Dopady významných vysvětlujících proměnných na predikci pravděpodobnosti ztráty zákazníka, který je těžištěm shluku setrvávajících zákazníků – Retail Rocket

Pozorování reprezentativní pro čtvrtý shluk ilustruje Obr. 37, kde pro zákazníka user_id 429213 předpovídáme pravděpodobnost ztráty 0.975. Nárůst pravděpodobnosti oproti očekávané hodnotě vychází především z vysokého variačního koeficientu doby mezi transakčními interakcemi, nejnižší doby potřebné k transakční interakci v objemu 1 CU a stáří poslední uživatelské relace.



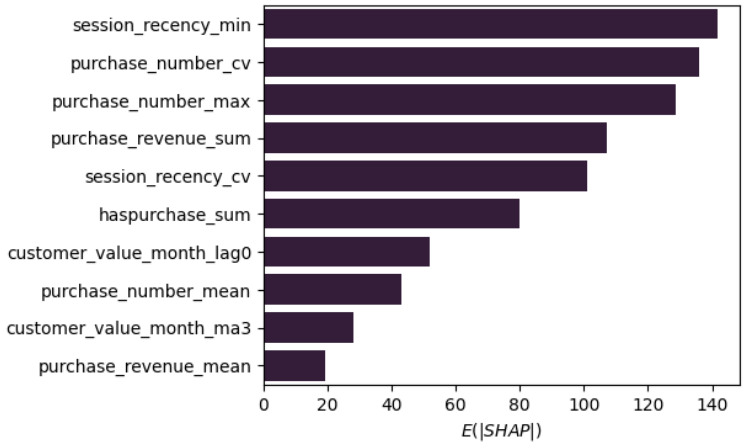
Obr. 37 Dopady významných vysvětlujících proměnných na predikci pravděpodobnosti ztráty zákazníka, který je těžištěm shluku ohrožených zákazníků – Retail Rocket

Nastíněné srovnání naznačuje diferenciaci mezi setrvávajícími a ohroženými zákazníky na základě chování v rámci uživatelské relace a stáří poslední relace. Možným nedostatkem je

způsob konstrukce a výběru vysvětlujících proměnných, který vede k horší srozumitelnosti/praktické využitelnosti předestřených závěrů.

Predikce ekonomického dopadu retenční kampaně

S ohledem na výsledky dosažené v ekonomickém hodnocení predikčních schopností regresních řešení byl pro tuto sekci vybrán systém strojového učení využívající rozhodovací strom. Pro daný časový řez byl optimalizací vnějších parametrů konstruován systém využívající kvantilové transformace vysvětlujících proměnných a robustní transformace vysvětlované proměnné. Vlastní rozhodovací strom není příliš hluboký, přeučení brání i vyšší minimální počet pozorování potřebných pro konstrukci rozvětvení stromu.

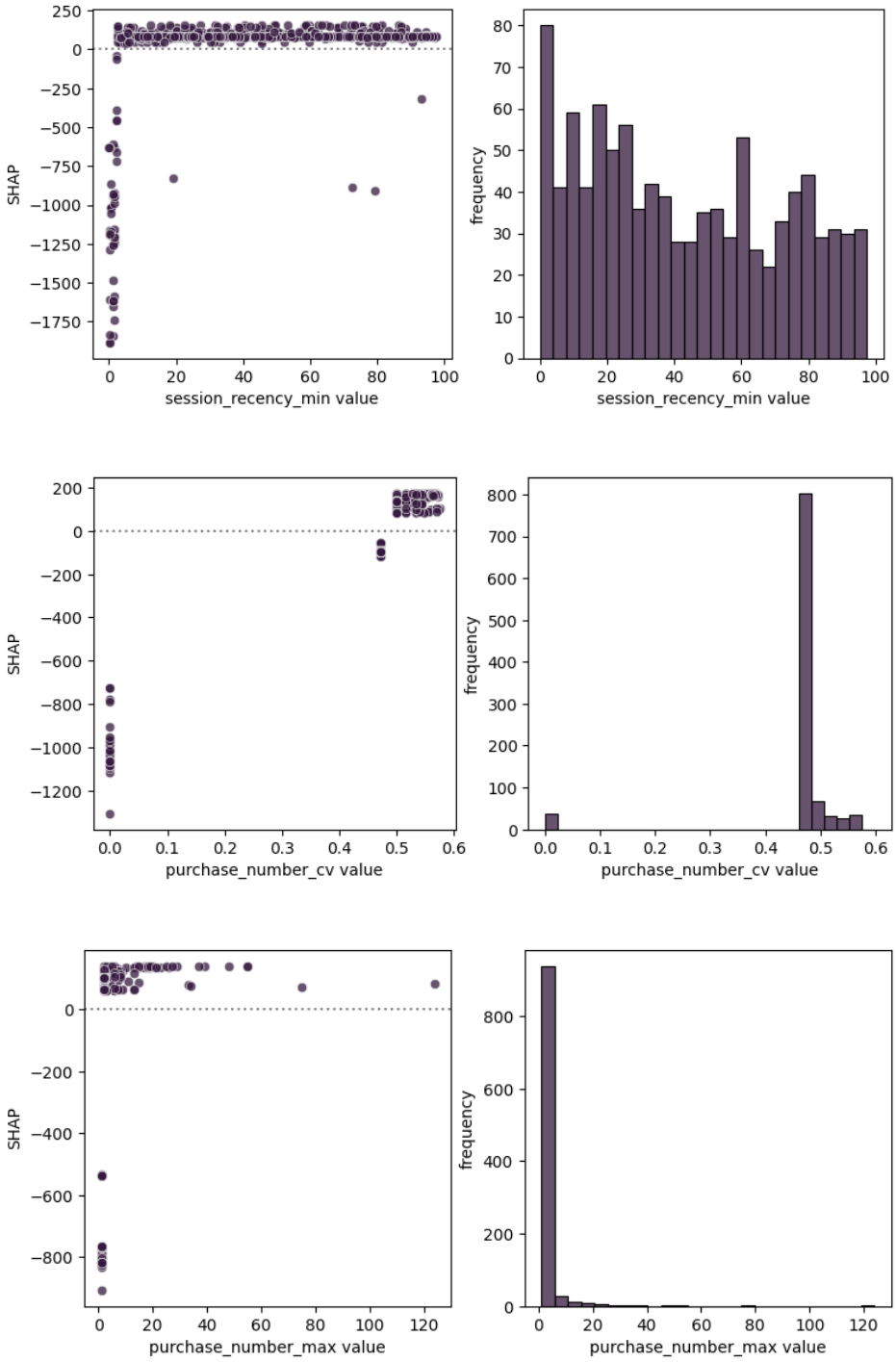


Obr. 38 Proměnné významné pro predikci inkrementálního zisku retenční kampaně, s využitím rozhodovacího stromu – Retail Rocket

Vysvětlující proměnné, které mají v průměru nejvyšší absolutní dopad na predikce ekonomického výsledku retenční kampaně, ilustruje horizontální sloupcový graf na Obr. 38. Vynikají zejména stáří poslední uživatelské relace, variační koeficient pořadí transakcí a celkový počet transakcí. Z hlediska zastoupení množin vysvětlujících proměnných pozorujeme stáří (session_recency_min, session_recency_cv), frekvenci (purchase_number_cv, purchase_number_max, haspurchase_sum, purchase_number_mean) a peněžní hodnotu uživatelských interakcí (purchase_revenue_sum, customer_value_month_lag0, customer_value_month_ma3, purchase_revenue_mean). Ukazuje se, že rozhodovací strom spoléhá na především na strukturu návštěv a frekvenci i objem transakcí.

Povahu vztahu mezi vysvětlujícími proměnnými a jejich vlivem na predikce regresního modelu ilustruje Obr. 39, využívající především bodových diagramů SHAP hodnot, stabilitu

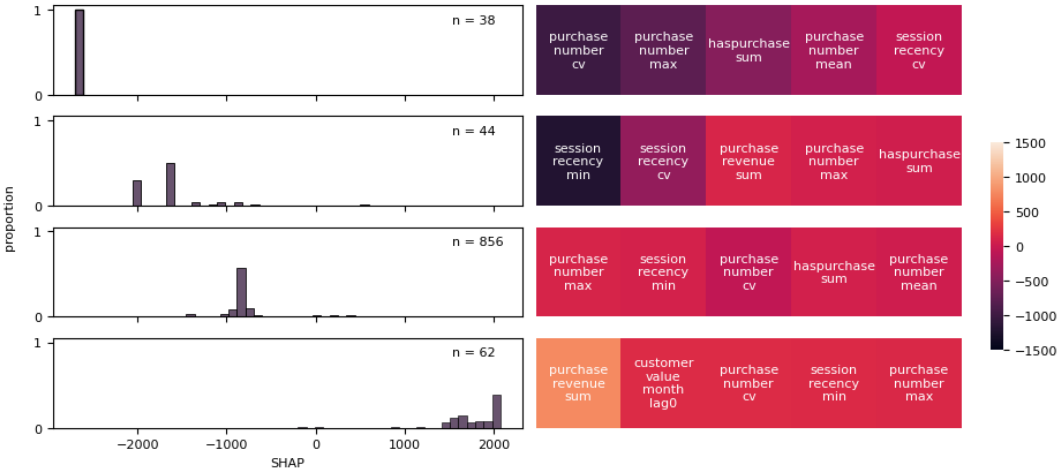
odhadů doplňují příslušné histogramy. Vysvětlující proměnnou se zásadním dopadem na predikci modelu je doba uplynulá od poslední návštěvy. Charakter vztahu není přímočarý, podobně jako v případě klasifikačního modelu, se však zdá být výhodné aktivity soustředit na zákazníky, jejichž poslední relace je starší než jeden týden. Naopak cílit na zákazníky s velmi nízkým stářím poslední uživatelské relace nemusí být prospěšné. Další významnou proměnou je variační koeficient pořadí nákupů. Zdá se, že s rostoucím variačním koeficientem dochází k růstu očekáváného ekonomického dopadu kampaně, tj. cílit na zákazníky s vyšším počtem realizovaných transakcí je pro organizaci výhodné. Podobně i počet nákupů odděluje nižší očekávaný výsledek retenční kampaně při oslovení zákazníků s jedním nebo více nákupy. Regresní rozhodovací strom v tomto případě exploatuje minimální uvažovaný počet transakcí datového souboru.



Obr. 39 SHAP hodnoty proměnných, významných pro predikci inkrementálního zisku retenční kampaně, včetně pozorovaného rozdělení – Retail Rocket

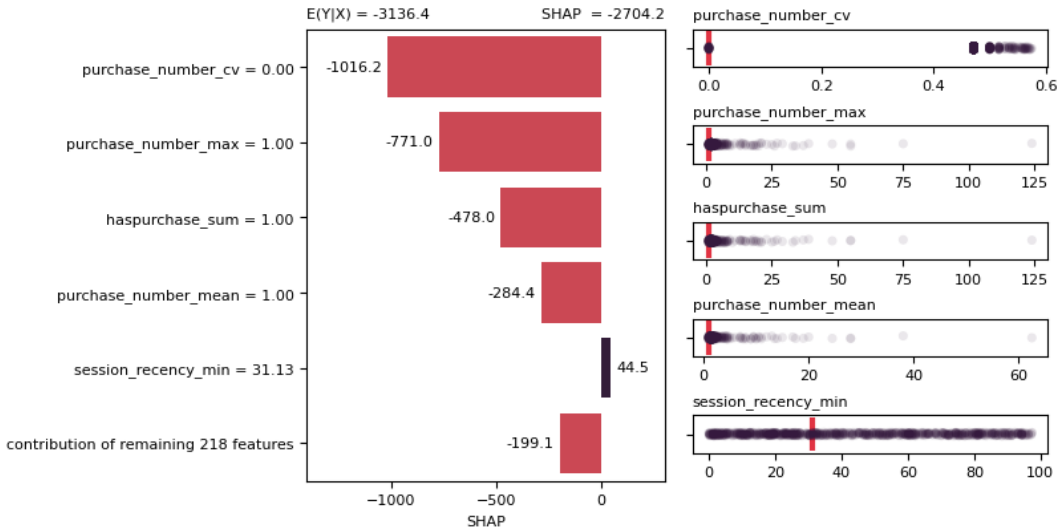
Další vhled umožňují skupiny zákazníků charakterizované podobnými vztahy mezi vysvětlovanou a vysvětlujícími proměnnými, viz Obr. 40. První tři shluky popisují zákazníky, jejichž zahrnutí do retenční kampaně by vedlo k spíše záporným výsledkům, odlišují se ale v rozptylu SHAP hodnot, pořadím i strukturou významných atributů. První shluk je velmi kompaktní, rozhodujícím faktorem je transakční chování. Druhý a třetí shluk vykazují vyšší rozptyl a obsahují

zákazníky s podobnými nákupními zvyklostmi, odlišují se ale stářím poslední uživatelské relace. Poslední ze shluků popisuje zákazníky s vyšším než očekávaným inkrementálním ekonomickým dopadem kampaně. Ekonomický dopad je ovlivněn objemem peněžních transakcí, aktuální hodnotou zákazníka a variačním koeficientem pořadí transakcí.



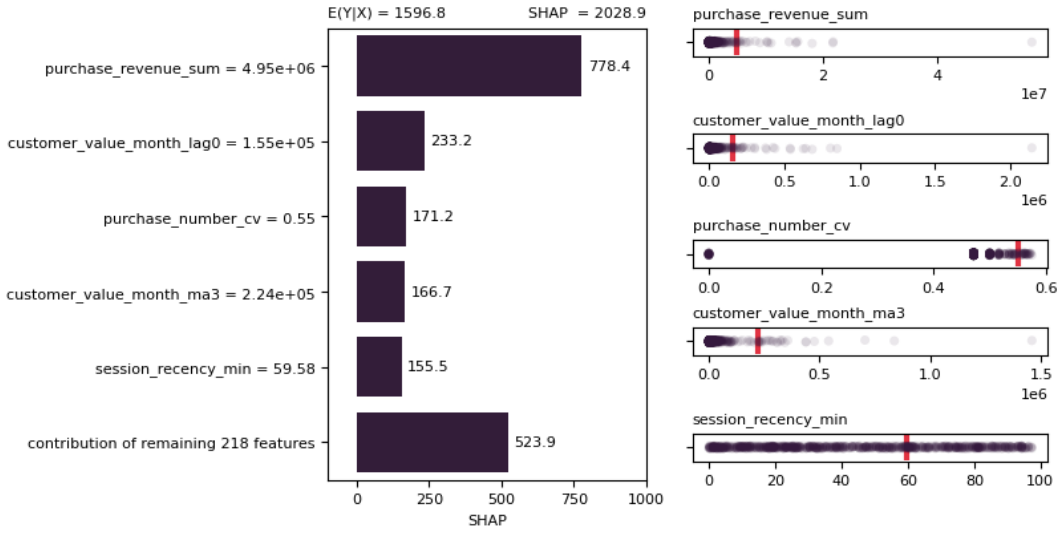
Obr. 40 Zákaznické shluky SHAP hodnot v predikci inkrementálního zisku retenční kampaně, včetně klíčových vysvětlujících proměnných – Retail Rocket

Pro pochopení zákaznických skupin je využito lokální interpretovatelnosti, tj. pro každý shluk je vybráno pozorování nejblíže příslušnému těžišti. Na vybrané pozorování je pak nahlíženo prostřednictvím významných SHAP hodnot i skutečných hodnot vysvětlujících proměnných. Pro demonstraci přístupu využíváme protiklad zákazníků, které je ekonomicky výhodné (čtvrtý shluk) nebo nevýhodné (první shluk) do retenční aktivity řadit. Pozorování reprezentativní pro první shluk ilustruje zákazník user_id 294088, u kterého předpokládáme inkrementální dopad zahrnutí do retenční kampaně ve výši −3136.4 CU. Ztráta je oproti očekávané inkrementální ekonomické hodnotě ovlivněna především nízkým počtem realizovaných transakcí, což se projevuje napříč významnými vysvětlujícími proměnnými, viz Obr. 41.



Obr. 41 Dopady významných vysvětlujících proměnných na predikci inkrementálního zisku retenční kampaně, který je těžištěm shluku zákazníků, které je nevýhodné uvažovat v retenční kampani – Retail Rocket

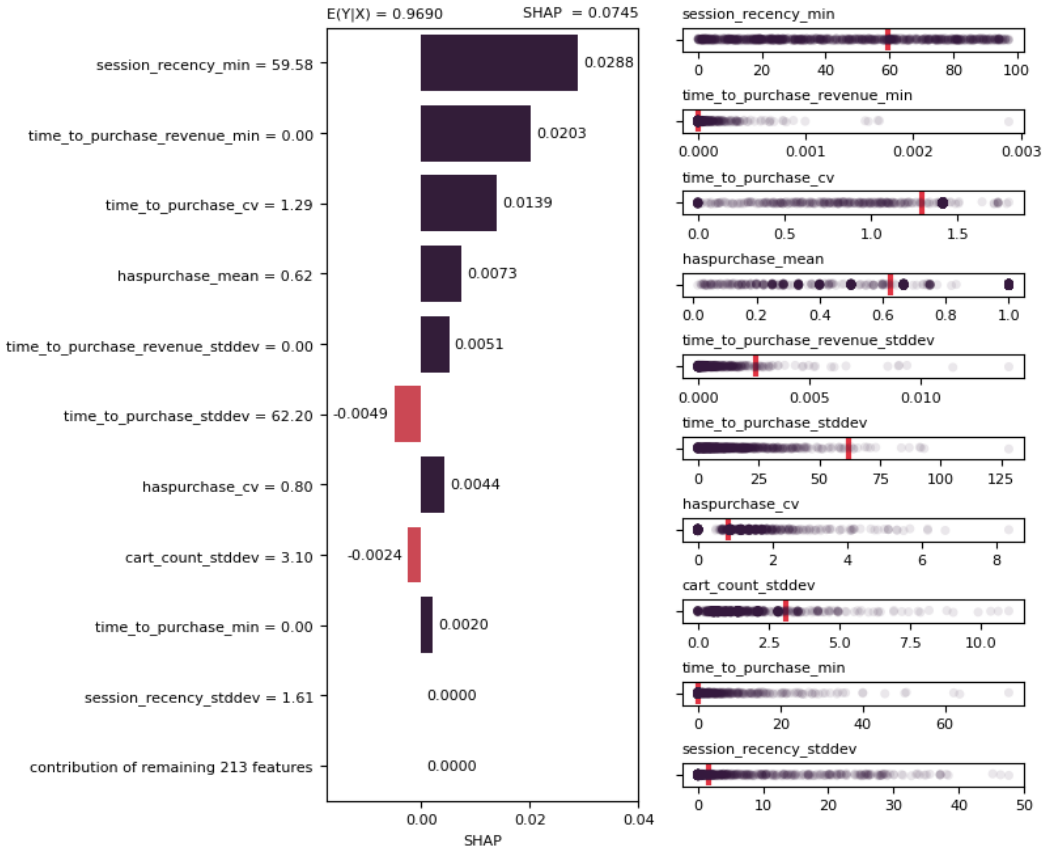
Pozorování reprezentativní pro čtvrtý shluk ilustruje zákazníka user_id 1065255, u kterého předpokládáme inkrementální dopad zahrnutí do retenční kampaně ve výši 1596.8 CU. Zisk je oproti očekávané základní inkrementální ekonomické hodnotě vyšší, především díky objemu realizovaných nákupů, hodnotě zákazníka, počtu nákupů, viz Obr. 42.



Obr. 42 Dopady významných vysvětlujících proměnných na inkrementálního zisku retenční kampaně, který je těžištěm shluku zákazníků, které je vhodné oslovit v rámci retenční kampaně – Retail Rocket

Nastíněné srovnání odhaluje u zákazníků, u kterých je zařazení do retenční aktivity nevýhodné, exploataci pravidel pro konstrukci datové sady, tj. pokud zákazník realizoval právě jednu transakci, je jisté, že dojde alespoň k jedné další. Naproti tomu u zákazníků, které je vhodné oslovit, lze pozorovat vysokou celkovou peněžní hodnotu transakcí, počet transakcí, i zákaznickou hodnotu; současně také značné stáří poslední uživatelské relace.

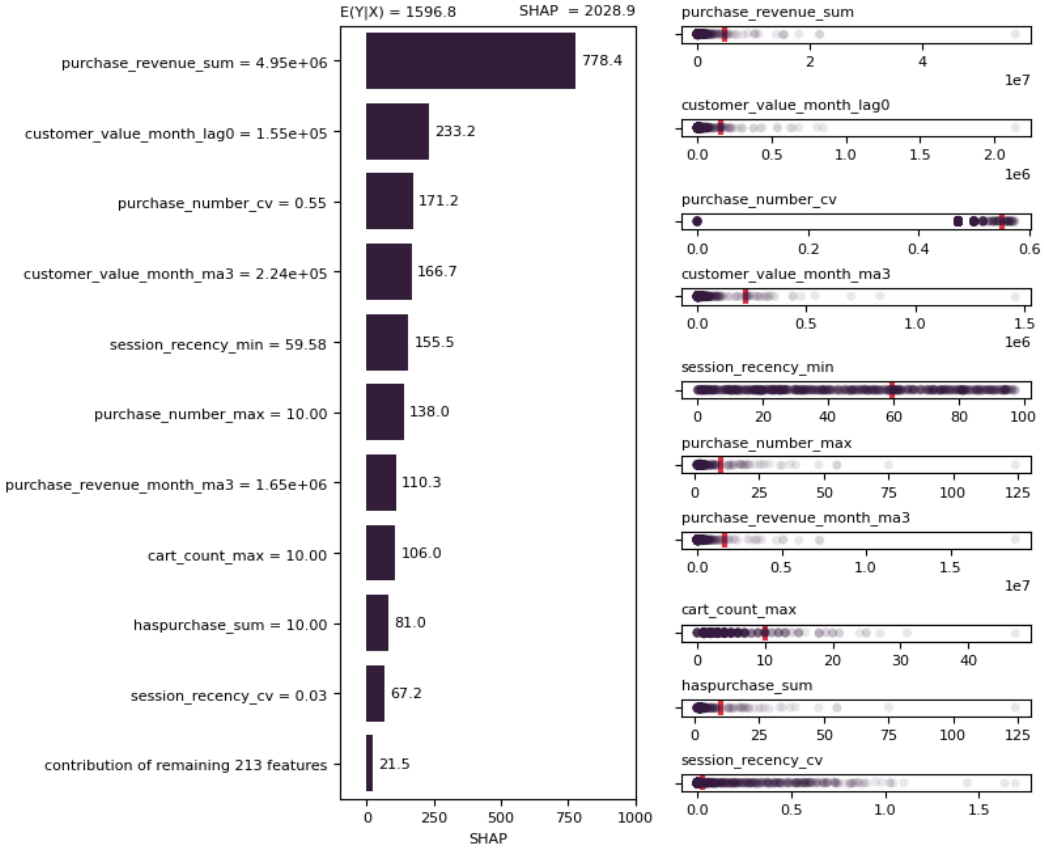
Retenční kampaň však bude směřovat na zákazníky, u nichž je možné předpokládat kladný inkrementální ekonomický výsledek aktivity. Podrobněji proto nahlédneme na těžiště posledního shluku, tj. zákazníka s user_id 1065255, a to perspektivou systému predikujícího pravděpodobnost ztráty i řešení predikujícího očekávaný inkrementální dopad zařazení daného zákazníka. Z klasifikačního hlediska, tj. z pohledu na pravděpodobnost ztráty zákazníka, spoléhá systém podpůrných vektorů s radiální jádrovou funkcí na vyšší stáří návštěv, specifické chování v rámci zákaznické relace, i vyšší konverzní poměr relace, viz Obr. 43.



Obr. 43 Detail těžiště shluku zákazníků, které je výhodné zařadit do retenčních aktivit pohledem klasifikačního modelu – Retail Rocket

Regresní systém využívající rozhodovacího stromu, který je zaměřen na predikci inkrementálního dopadu zařazení zákazníka do retenční kampaně, identifikuje vysoké hodnoty

tohoto dopadu především s ohledem na celkový obrat generovaný zákazníkem, a současnou i minulou hodnotu zákazníka. Za další významné faktory je možné považovat různé reprezentace počtu nákupů, nebo vyšší stáří poslední interakce, viz Obr. 44.



Obr. 44 Detail těžiště shluku zákazníků, které je výhodné zařadit do retenčních aktivit pohledem regresního modelu – Retail Rocket

Představená interpretace prediktivních modelů nastiňuje případné směřování retenční kampaně. V rámci uživatelské relace se ukazuje jako významný vysoký konverzní poměr i dlouhý čas rozhodnutí zákazníka. Retenční aktivita by tak měla cílit jednak na podporu zvýšení počtu seancí, případně na další produktový obsah nebo služby. Vysoké stáří zákaznických návštěv ohraničuje vhodné kontaktní kanály. Zákaznické shluky odkrývají i další faktory, především v souvislosti s hodnotu zákazníka. Vysvětlující proměnné bohužel reflektují především zákaznické chování, tj. nepodařilo se identifikovat skutečné příčiny odlivu zákazníků.

5.2 REES46

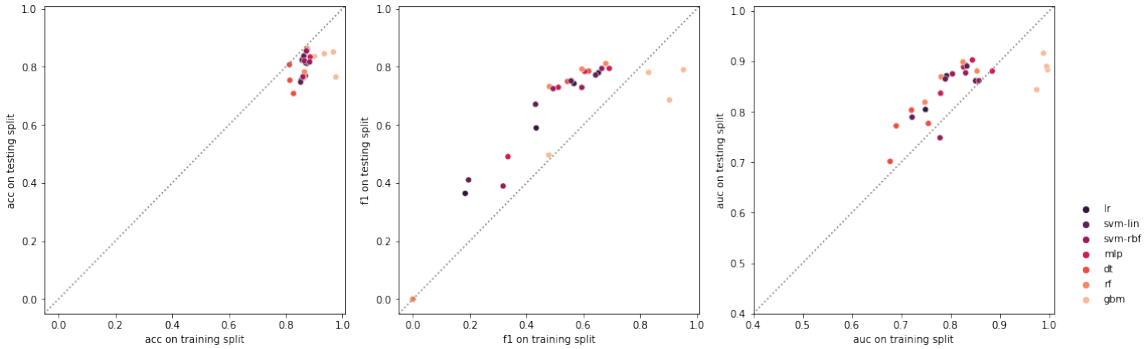
5.2.1 Přirozené ukazatele úspěšnosti

Mezi klasifikačními systémy vynikají řešení postavená na umělých neuronových sítích a gradient boosting, a to napříč perspektivami ACC, F1 i AUCROC. Zajímavý je výsledek náhodných lesů, které dosahují dobrých výsledků v perspektivách ACC a AUCROC; ovšem v perspektivě F1 lesy selhávají, na vině se zdá být především komponenta přesnosti předpovědi pozitivní třídy. Rozhodovací stromy se zdají být zcela nevhodné. S ohledem na dobu potřebnou pro konstrukci řešení na jednom časovém řezu jsou jednotlivá řešení srovnatelná, vyčnívají pouze meta-algoritmy a umělé neuronové sítě.

Obr. 45 Ukazatele klasifikační úspěšnosti – REES46

Algorithm	ACC	F1	AUCROC	Optimization time [min]	Refit time [min]
lr	0.803 (0.751, 0.856)	0.619 (0.319, 0.919)	0.858 (0.798, 0.917)	1.6 (0.8, 2.5)	0.4 (0.2, 0.5)
svm-lin	0.806 (0.742, 0.869)	0.651 (0.386, 0.916)	0.852 (0.782, 0.921)	1.2 (0.7, 1.7)	0.4 (0.2, 0.7)
svm-rbf	0.814 (0.756, 0.873)	0.660 (0.369, 0.950)	0.841 (0.742, 0.939)	1.4 (0.7, 2.1)	0.7 (0.5, 1.0)
mlp	0.824 (0.762, 0.886)	0.700 (0.473, 0.926)	0.877 (0.832, 0.923)	182.3 (56, 308)	81.8 (-14, 177)
dt	0.758 (0.693, 0.823)	0.384 (-0.322, 1.089)	0.764 (0.694, 0.833)	1.2 (0.6, 1.8)	0.3 (0.1, 0.5)
rf	0.814 (0.736, 0.892)	0.584 (-0.038, 1.206)	0.867 (0.813, 0.921)	3.6 (-0.1, 7.3)	1.5 (0.2, 2.7)
gbm	0.824 (0.760, 0.888)	0.688 (0.471, 0.905)	0.883 (0.836, 0.931)	2.3 (1.0, 3.6)	4.2 (0.2, 8.3)

Pro bližší porozumění schopnostem jednotlivých přístupů byl zkonstruován Obr. 46, který porovnává predikční schopnosti na trénovacích a testovacích množinách dat. Pohled perspektivou ACC naznačuje, že klasifikační systémy trpí přílišnou variabilitou predikcí, ukázkovým případem je gradient boosting, jehož přeučení by bylo možné adresovat restriktivním přístupem k výběru vysvětlujících proměnných i vnitřní struktuře modelů. Naproti tomu perspektiva F1 naznačuje vysoké vychýlení predikcí napříč modely na nejzazším časovém řezu, kde pro část dosahují řešení hodnot nižších než jedna polovina, a to i pro trénovací data. Ukázkovým případem selhání na obtížném časovém řezu jsou náhodné lesy. Řez je odpovědný i šíři intervalů spolehlivosti. Ke zmírnění neduhu by bylo možné přistoupit s využitím více dat, konstrukcí nových proměnných nebo rozšířením vnitřní struktury modelů.



Obr. 46 Kompromis vychýlení a rozptylu klasifikačních řešení napříč časovými řezy – REES46

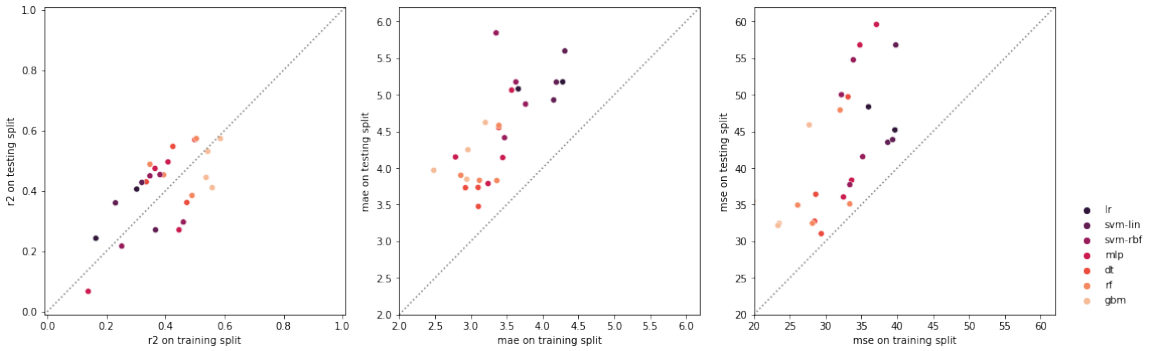
Nahlížíme-li na úspěšnost regresních systému perspektivou R^2, pak se nejúspěšnější zdají být přístupy založené na rozhodovacích stromech. Naproti tomu podpůrné vektory s lineární jádrovou funkcí významně selhávají. Rozhodovací stromy, náhodné lesy i gradient boosting dominují i v perspektivách MAE a MSE. Z pohledu na řády příslušných chyb se zdá, že regresní modely jsou v tomto případě robustní, což je možné blíže diagnostikovat analýzou residuí. Doba potřebná pro konstrukci řešení na jednom časovém řezu odpovídá klasifikační větvi řešení, tj. pozorované hodnoty jsou srovnatelné pro všechny algoritmy, výjimky tvoří podobně jako v klasifikační větvi meta-algoritmy a umělé neuronové sítě.

Tab. 14 Ukazatele regresní úspěšnosti – REES46

Algorithm	R^2	MAE	MSE	Optimization time [min]	Refit time [min]
lr	-3.63 (-15.79, 8.54)	7.37 (2.1, 12.6)	320.1 (-511, 1151)	1.1 (0.5, 1.7)	0.6 (-0.2, 1.3)
svm-lin	-2.0E+26 (-8.3E+26, 4.3E+26)	4.1E+13 (-9.0E+13, 1.7E+14)	1.2E+28 (-2.8E+28, 5.3E+28)	1.1 (0.5, 1.7)	1.1 (0.5, 1.8)
svm-rbf	0.354 (0.168, 0.541)	5.08 (4.12, 6.03)	46.0 (33.7, 58.4)	1.2 (0.7, 1.7)	1.0 (0.2, 1.9)
mlp	0.327 (0.007, 0.647)	4.29 (3.42, 5.15)	47.7 (28.3, 67.1)	142.3 (46, 239)	122.4 (62, 183)
dt	0.477 (0.320, 0.634)	3.87 (3.13, 4.62)	37.5 (24.0, 50.9)	1.3 (0.6, 2.0)	0.8 (0.3, 1.4)
rf	0.475 (0.350, 0.600)	4.04 (3.45, 4.62)	37.6 (26.5, 48.7)	3.9 (0.3, 7.6)	9.2 (0.5, 17.9)
gbm	0.490 (0.370, 0.610)	4.17 (3.63, 4.72)	36.5 (26.3, 46.7)	2.3 (0.6, 4.0)	2.3 (0.3, 4.3)

Další pohled na způsobilost jednotlivých regresních přístupů ilustruje Obr. 47. Perfektní řešení leží pro R^2 v pravém horním rohu, pro MAE a MSE naopak v levém dolním rohu, což upravuje vztah mezi povahou selhání a osou kvadrantu. Perspektiva R^2 naznačuje neschopnost velké části řešení vysvětlit více než 50 % variability i na trénovací množině dat. Pohled na MAE a MSE odhaluje problémy vychýlení všech řešení na nejzazším časovém řezu, u lineární regrese i podpůrných vektorů pozorujeme tento problém napříč trénovacími daty. Úspěšnost na

obtížném časovém řezu je možné zlepšit využitím více dat, restriktivním přístupem k výběru vysvětlujících proměnných i vnitřní struktuře modelů. Nedostatečné využití trénovací množiny dat je možné adresovat konstrukcí dalších vysvětlujících proměnných, případně složitější vnitřní strukturou modelu.



Obr. 47 Kompromis vychýlení a rozptylu regresních řešení napříč časovými řezy – REES46

5.2.2 Ekonomický dopad retenční kampaně

Z pohledu na ekonomický dopad retenční kampaně postavené na jednotlivých řešeních vyčnívají značné propady mezi očekávaným a dosaženým ziskem. Při srovnání klasifikačních a regresních modelů stejného typu lze sledovat vyšší dosažený zisk regresních modelů postavených na rozhodovacích stromech.

Klasifikačním modelům dominuje gradient boosting, při jehož použití pro řazení zákaznické báze by bylo dosaženo střední hodnoty zisku 1.32E+04 CU, při zahrnutí 12.5 % zákazníků do kýžené retenční aktivity. Následují umělé neuronové sítě, které dosahují střední hodnoty zisku 1.27E+04 CU, při oslovení 13.8 % zákazníků, případně podpůrné vektory s radiální jádrovou funkcí dosahující střední hodnoty zisku 1.24E+04 CU při zahrnutí 12.1 % zákazníků do retenční kampaně. 95% intervaly spolehlivosti pro skutečný ekonomický dopad kampaně obsahují výhradně nezáporné hodnoty.

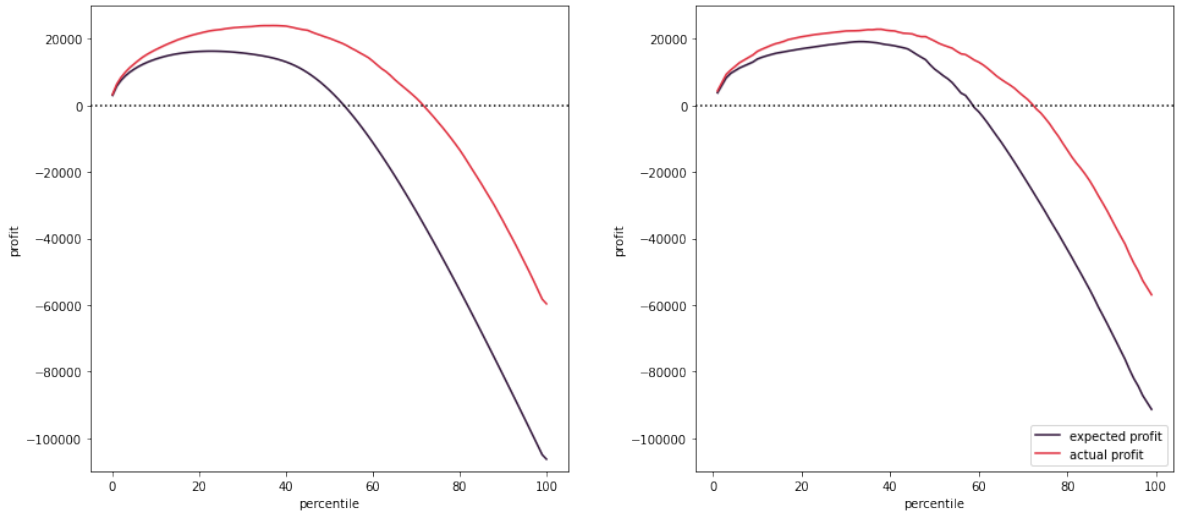
Mezi regresními modely vyčnívají rozhodovací stromy a náhodné lesy se střední hodnotou zisku 1.40E+04 CU při zahrnutí 19.8 %, respektive 18.4 % zákazníků do retenční kampaně. Následuje gradient boosting se střední hodnotou zisku 1.37E+04 CU a oslovení 15.6 % zákazníků. 95% intervaly spolehlivosti pro skutečný ekonomický dopad kampaně obsahují výhradně nezáporné hodnoty.

Ze srovnání nejlepších regresních a klasifikačních přístupů vyplývá, že využití regresního modelu vede v průměru ke zlepšení dosaženého zisku o ~ 6.1 %, což odpovídá ~ 798.2 CU. Modely jsou dále porovnány s pomocí párového Wilcoxonova testu, napříč časovými řezy. Nulová hypotéza předpokládá, že střední hodnota rozdílu zisků vybraného regresního a klasifikačního modelu je symetrická kolem nuly. Alternativní jednostranná hypotéza říká, že střední hodnota tohoto rozdílu je vyšší než u rozdělení symetrického kolem nuly. S ohledem na počet pozorování byla stanovena hladina významnosti na úrovni $\alpha = 0.1$. Pro regresní rozhodovací strom a klasifikační gradient boosting dosahuje test součtu pořadí rozdílů zisku, které jsou vyšší než nula, hodnoty 10, což odpovídá p-hodnotě 0.0625. Je tedy možné přijmout hypotézu alternativní, tj pozorované kladné rozdíly nejsou nahodilé a využití regresního modelu vede k lepším ekonomickým výsledkům.

Tab. 15 Ukazatele ekonomického dopadu retenční kampaně – REES46

Algorithm	**classification**		**regression**	
	$\Pi^{expected}$	Π^{actual}	$\Pi^{expected}$	Π^{actual}
lr	7992.2 (-491, 16475.5)	1.11E+04 (5.68E+02, 2.16E+04)	8831.7 (-11139.6, 28803.1)	7422.0 (1532.9, 13311.1)
svm-lin	1.06E+04 (4.99E+03, 1.62E+04)	1.16E+04 (1.79E+03, 2.14E+04)	3392.2 (-3064, 9848.4)	7960.0 (-5201.9, 21121.9)
svm-rbf	1.03E+04 (-2.20E+03, 2.27E+04)	1.24E+04 (5.90E+02, 2.42E+04)	8690.9 (-1175, 18556.8)	1.14E+04 (6.88E+02, 2.21E+04)
mlp	1.15E+04 (-3.36E+02, 2.33E+04)	1.27E+04 (6.21E+02, 2.47E+04)	8283.6 (-3800.6, 20367.8)	1.13E+04 (-2.99E+03, 2.56E+04)
dt	1757.1 (-1156.6, 4670.9)	4993.6 (-1832.4, 11819.7)	1.29E+04 (4.34E+03, 2.15E+04)	1.40E+04 (3.67E+03, 2.43E+04)
rf	7155.5 (-4097.2, 18408.2)	1.04E+04 (-2.85E+03, 2.36E+04)	1.13E+04 (3.04E+03, 1.96E+04)	1.40E+04 (2.91E+03, 2.50E+04)
gbm	9705.2 (1255.6, 18154.7)	1.32E+04 (2.21E+03, 2.42E+04)	1.19E+04 (8.32E+02, 2.30E+04)	1.37E+04 (2.90E+03, 2.46E+04)

Bližší pohled na způsob řazení zákazníků dle očekávaného inkrementálního zisku kampaně zobrazuje Obr. 48, porovnávající očekávaný a skutečný kumulativní zisk kampaně v aktuálním časovém řezu pro vybrané klasifikační (gradient boosting) a regresní přístupy (rozhodovací strom). Klasifikační přístup vede k výraznému podhodnocení očekávaného zisku napříč celým oborem hodnot. Regresní přístup také podhodnocuje skutečný výsledek kampaně, ale uspokojivě odhaduje tvar křivky až do 40 percentilu, mezi 40-60 percentilem dochází k akumulaci chyb, což vede k posunu chvostu křivky. Regresní model tedy přesně určí zákazníky, kteří přinesou vysoký inkrementální zisk nebo ztrátu, predikce mezi těmito extrémy jsou však problematické.



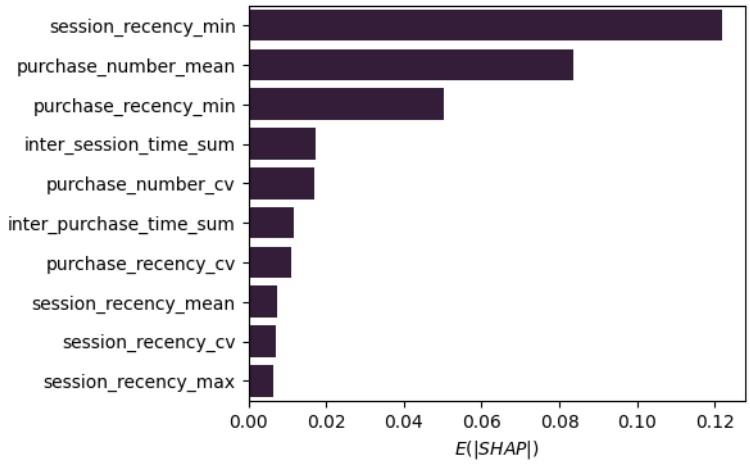
Obr. 48 Křivka kumulativního zisku retenční kampaně pro metody klasifikační gradient boosting (vlevo) a regresní rozhodovací strom (vpravo) – REES46

5.2.3 Interpretace vybraných modelů

Predikce ztráty zákazníka

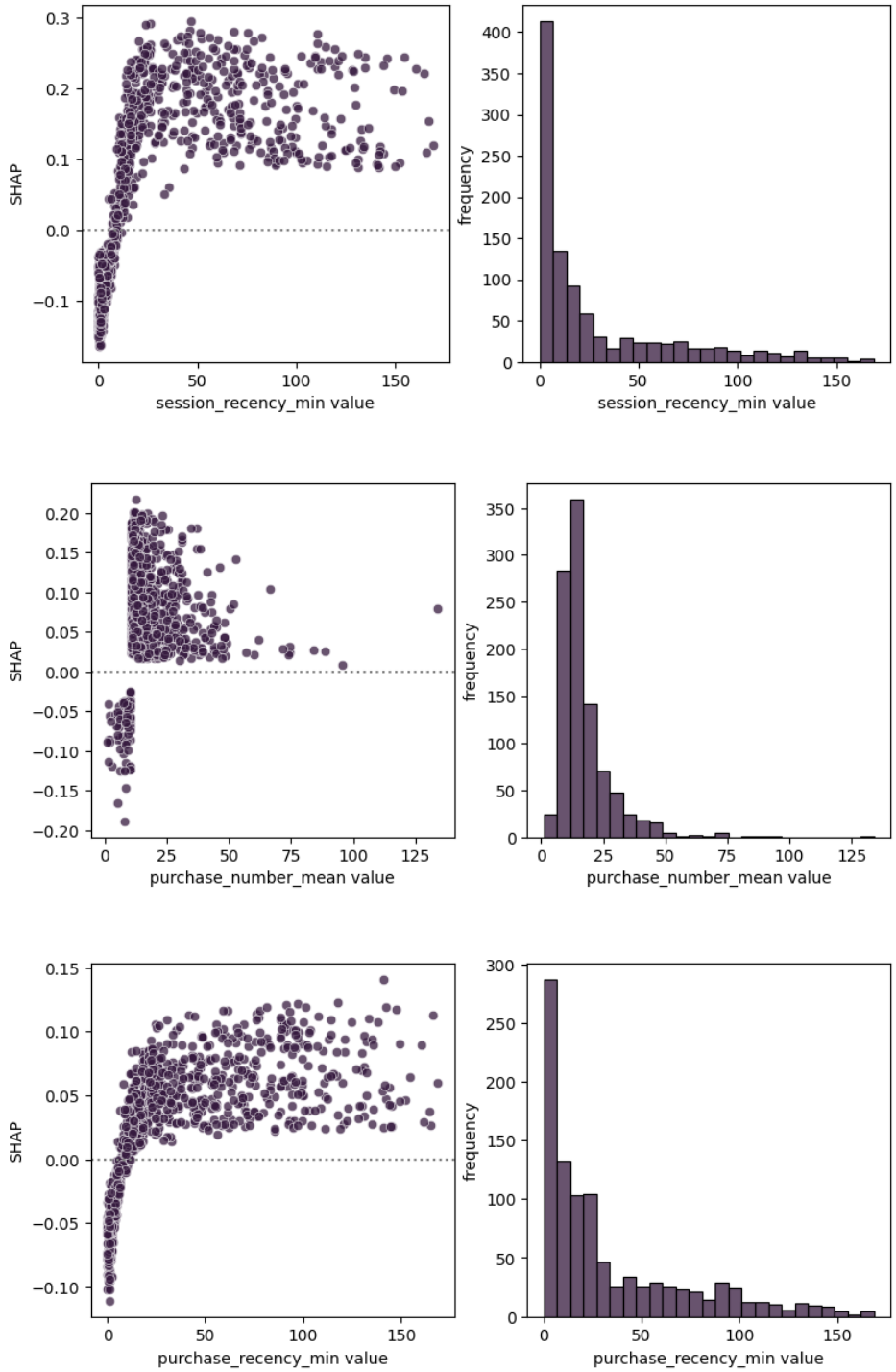
Na základě výsledků dosažených v přirozených ukazatelích predikčních schopností jsme pro další interpretaci vybrali systém strojového učení využívající gradient boosting. Pro daný časový řez byl optimalizací vnějších parametrů konstruováno řešení využívající mocninné transformace vysvětlujících proměnných a převzorkování dostupných pozorování. Vlastní klasifikační model je komplexní ve smyslu počtu a hloubky dílčích rozhodovacích stromů, využívá i vysoké adaptability; přeučení brání vyšší minimální počet pozorování v lístkové úrovni, silné vzorkování pozorování i vysvětlujících proměnných, i silná regularizace vah dílčích rozhodovacích stromů.

Vysvětlující proměnné, které mají v průměru nejvyšší absolutní dopad na predikci ztráty zákazníka, ilustruje horizontální sloupcový graf na Obr. 49. Vynikají doba uplynulá od poslední návštěvy, průměrné pořadí nákupu a doba uplynulá od posledního nákupu. Z hlediska zastoupení množin vysvětlujících proměnných pozorujeme stáří (session_recency_min, purchase_recency_min, purchase_recency_cv, session_recency_mean, session_recency_cv, session_recency_max) a frekvenci uživatelských interakcí (purchase_number_mean, purchase_number_cv). Významné se zdají být i některé ostatní proměnné, popisující vlastnosti uživatelské relace (inter_session_time_sum, inter_purchase_time_sum). Zdá se, že gradient boosting spoléhá na tradiční aspekty uživatelských interakcí, ale i na některé atributy chování uvnitř relace.



Obr. 49 Proměnné významné pro klasifikaci odchodu zákazníka, s využitím metody gradient boosting – REES46

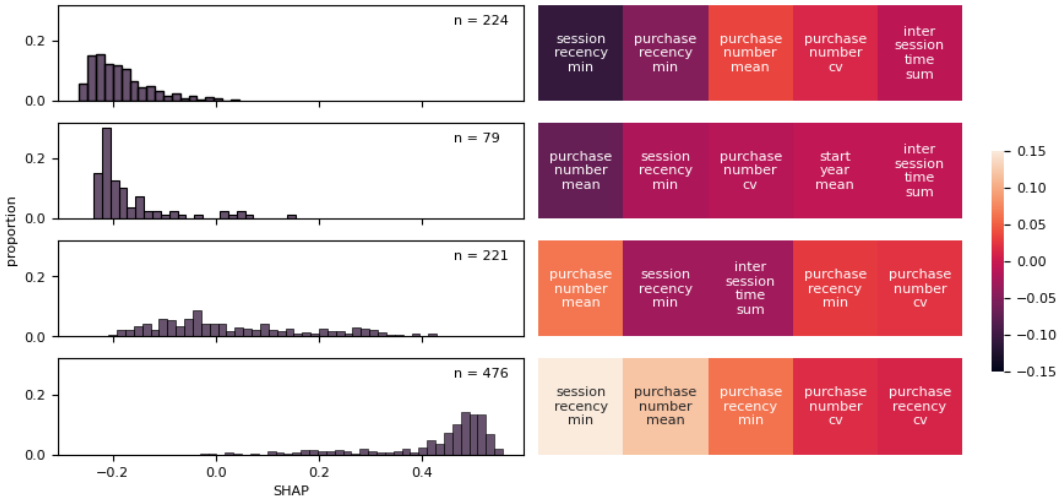
Charakter vztahu mezi vysvětlujícími proměnnými a jejich vlivem na predikce klasifikačního modelu, ilustruje Obr. 50, využívající bodových diagramů SHAP hodnot, robustnost ilustrují histogramy veličin. Vysvětlující proměnnou se zásadním dopadem na predikci modelu je doba uplynulá od poslední uživatelské relace. S rostoucí dobou od poslední návštěvy dochází k růstu pravděpodobnosti odchodu. K poslední interakci věrných zákazníků došlo zpravidla v uplynulém týdnu. U zákazníků, kteří s platformou neinteragovali déle než měsíc, se vliv atributu stabilizuje. Další výstižnou proměnnou je průměrné pořadí transakcí. Ukazuje se, že průměrné pořadí transakcí dělí příspěvek k pravděpodobnosti ztráty zákazníka na dvě části, kde pokud zákazník realizoval méně než dvacet transakcí (purchase_number_mean<10.5) pak je riziko ztráty nižší, pro vyšší počet transakcí (purchase_number_mean>10.5) je naopak riziko vyšší. I v tomto případě model exploatuje podmínky pro výběr zákazníků. Význačná je i doba uplynulá od poslední transakce. Zdá se, že s rostoucí dobou od posledního nákupu roste i pravděpodobnost ztráty zákazníka, kdy během prvních tří týdnů dochází ke skokovému růstu pravděpodobnosti ztráty, která se následně stabilizuje.



Obr. 50 SHAP hodnoty proměnných, významných pro klasifikaci ohrožených zákazníků, včetně pozorovaného rozdělení – REES46

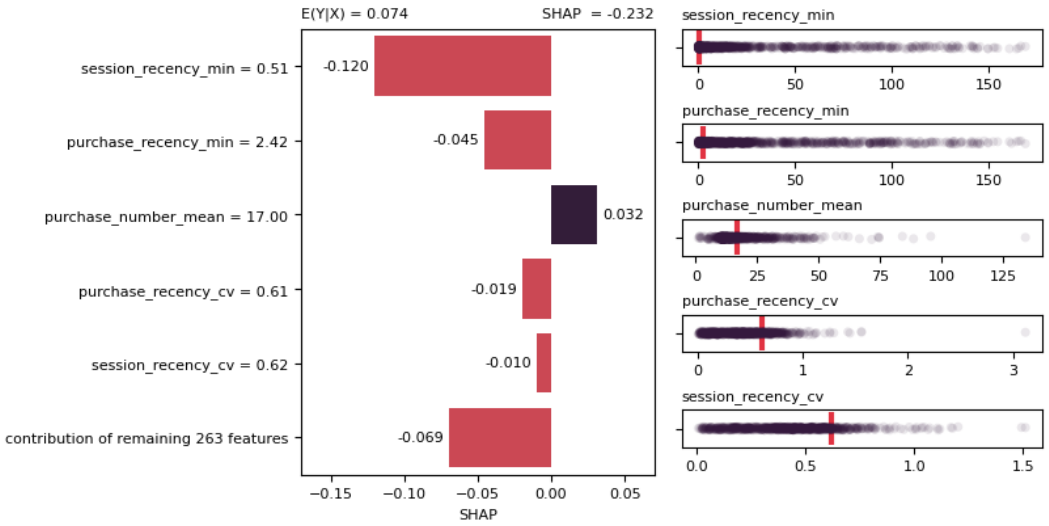
Hlubší pochopení umožňují shluky zákazníků, jejichž SHAP hodnoty napříč vysvětlujícími proměnnými sledují obdobné vzorce chování. Skupiny zákazníků pro klasifikační model znázorňuje Obr. 51, postavený na histogramech součtů SHAP hodnot, a teplotních mapách, popisujících strukturu a směr působení významných vysvětlujících proměnných. První shluk

seskupuje zákazníky, u kterých lze očekávat nižší pravděpodobnost ztráty. Za poklesem stojí především stáří poslední uživatelské relace a realizovaného nákupu. Druhý ze shluků odpovídá zákazníkům, u kterých je možné uvažovat nižší pravděpodobnost ztráty. Jedná se o zákazníky, kteří realizovali nižší počet nákupů, a jejich poslední relace je starší než u zákazníků identifikovaných v předchozím shluku. Třetí ze shluků popisuje zákazníky, u nichž je model méně jistý, což reflektuje i vysoký rozptyl SHAP hodnot. Shluk popisuje zákazníky, kteří realizovali vyšší počet nákupů, současně ale v nedávné době interagovali s platformou prodejce. Poslední ze skupin sdružuje zákazníky s vyšší pravděpodobností odchodu, kteří s webem společnosti dlouho neinteragovali, nenakupovali a současně v minulosti uskutečnili vyšší počet nákupů.



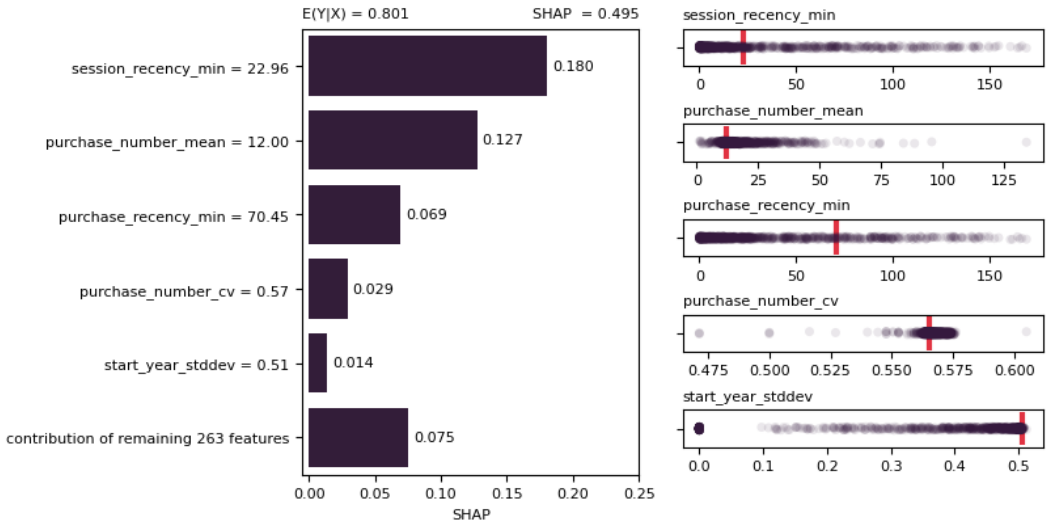
Obr. 51 Zákaznické shluky SHAP hodnot v klasifikaci ohrožených zákazníků, včetně klíčových vysvětlujících proměnných – REES46

Pro další porozumění zákaznickým shlukům je využito lokální interpretovatelnosti, kde pro každou skupinu zákazníků určíme pozorování nejblíže těžišti. Na vybraného zákazníka pak nahlížíme prostřednictvím významných SHAP hodnot i pozorovaných hodnot vysvětlujících proměnných. Pojetí je ilustrováno protiklady zákazníků setrvávajících (první shluk) a ohrožených (čtvrtý shluk). Pozorování reprezentativní pro první shluk zastupuje Obr. 52, kde u zákazníka user_id 578096346 předpokládáme ztrátu s pravděpodobností 0.074. Pokles pravděpodobnosti oproti očekávané hodnotě vychází ze stáří poslední uživatelské relace realizované před přibližně dvanácti hodinami a ze stáří posledního nákupu, realizovaného v uplynulých třech dnech.



Obr. 52 Dopady významných vysvětlujících proměnných na predikci pravděpodobnosti ztráty zákazníka, který je těžištěm shluku setrvávajících zákazníků – REES46

Pozorování reprezentativní pro čtvrtý shluk ilustruje Obr. 52, kde pro zákazníka s user_id 540119256 předjímáme pravděpodobnost ztráty 0.801. Nárůst pravděpodobnosti oproti očekávané hodnotě vychází především ze stáří poslední uživatelské relace i vyššího počtu nákupů, které byly provedeny před více než dvěma měsíci.



Obr. 53 Dopady významných vysvětlujících proměnných na predikci pravděpodobnosti ztráty zákazníka, který je těžištěm shluku ohrožených zákazníků – REES46

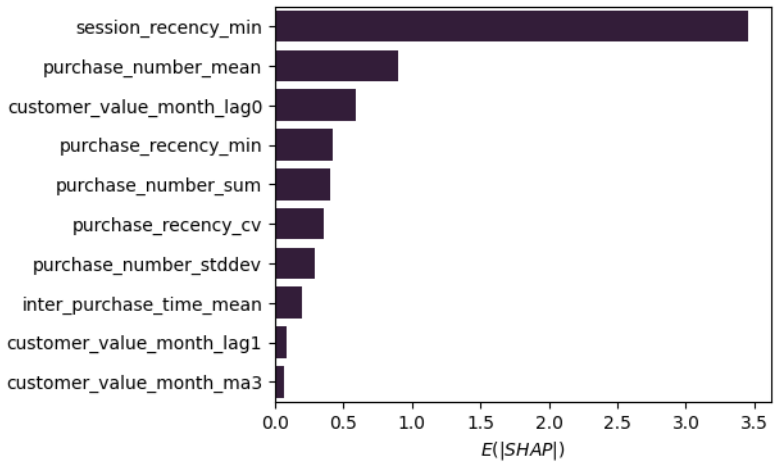
Představené srovnání odkrývá rozdíly mezi setrvávajícími a ohroženými zákazníky v tradičních aspektech uživatelských relací i transakční historie, především v dimenzích frekvence

a stáří. Možným nedostatkem je omezená praktická využitelnost předestřených skutečností při návrhu retenční kampaně.

Predikce ekonomického dopadu retenční kampaně

S ohledem na výsledky dosažené v ekonomickém hodnocení predikčních schopností regresních řešení v této sekci interpretujeme systém strojového učení využívající rozhodovací strom. Pro daný časový řez byl optimalizací vnějších parametrů konstruováno řešení využívající kvantilové transformaci vysvětlujících proměnných a robustní transformaci vysvětlované proměnné. Vlastní rozhodovací strom není příliš hluboký, přeučení brání i vyšší minimální počet pozorování v lístkové úrovni stromu.

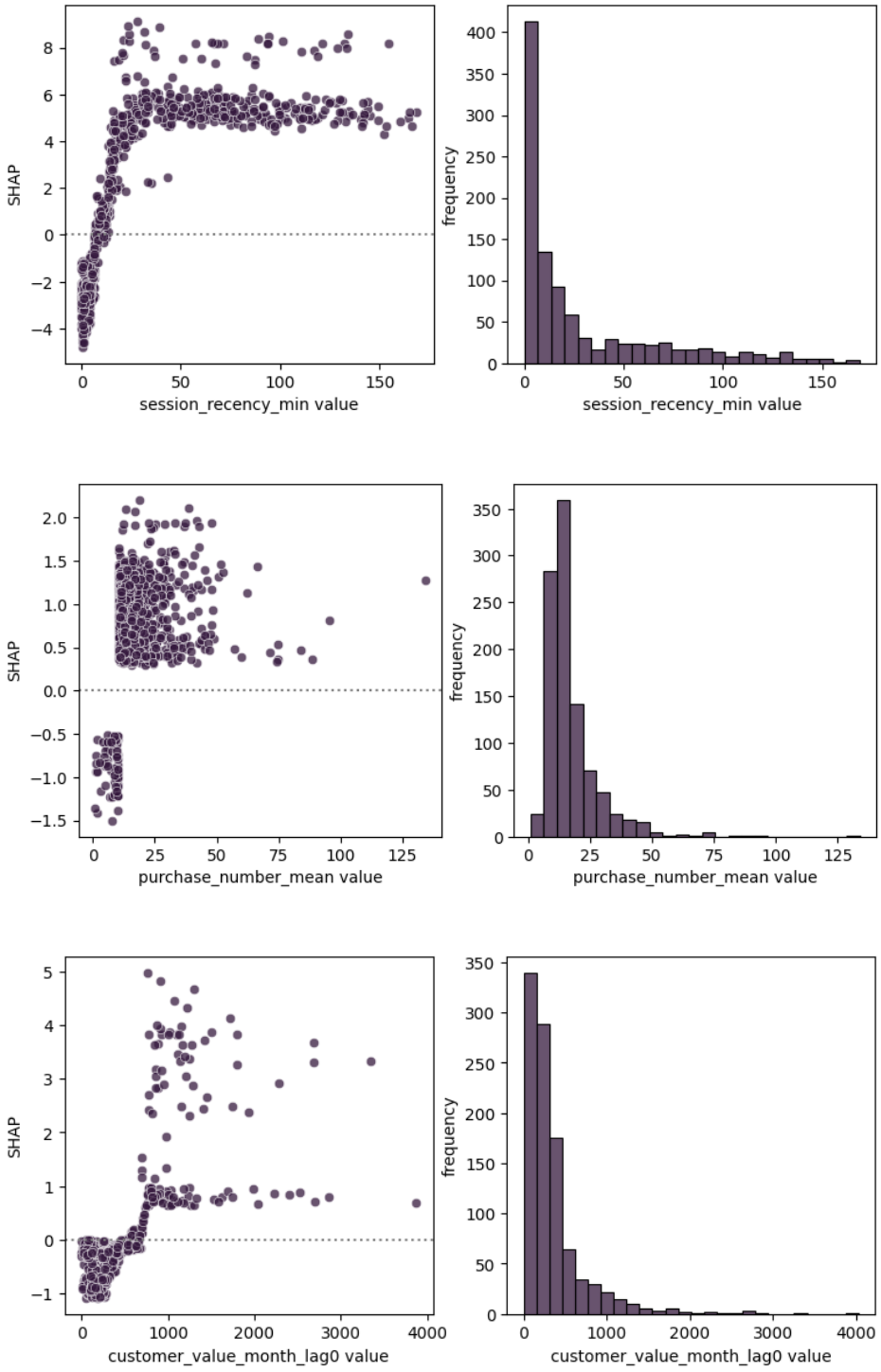
Vysvětlující proměnné, které mají v průměru nejvyšší absolutní dopad na predikce ekonomického výsledku retenční kampaně, ilustruje horizontální sloupcový graf na Obr. 54. Vynikají především doba od poslední návštěvy, počet realizovaných nákupů a celkový peněžní objem transakcí zákazníka. Z hlediska zastoupení množin vysvětlujících proměnných pozorujeme stáří (session_recency_min, purchase_recency_min, purchase_recency_cv), frekvenci (purchase_number_mean, purchase_number_sum, purchase_number_stddev) a peněžní hodnotu interakcí (customer_value_month_lag0, customer_value_month_lag1, customer_value_month_ma3) a ostatní (inter_purchase_time_mean). Ukazuje se, že rozhodovací strom spoléhá především na různé reprezentace zákaznické hodnoty a stáří relací, případně nákupů.



Obr. 54 Proměnné významné pro predikci inkrementálního zisku retenční kampaně, s využitím rozhodovacího stromu – REES46

Povahu vztahu mezi vysvětlujícími proměnnými a jejich vlivem na predikce regresního modelu, ilustrujeme s pomocí Obr. 55. Vysvětlující proměnnou se zásadním dopadem na

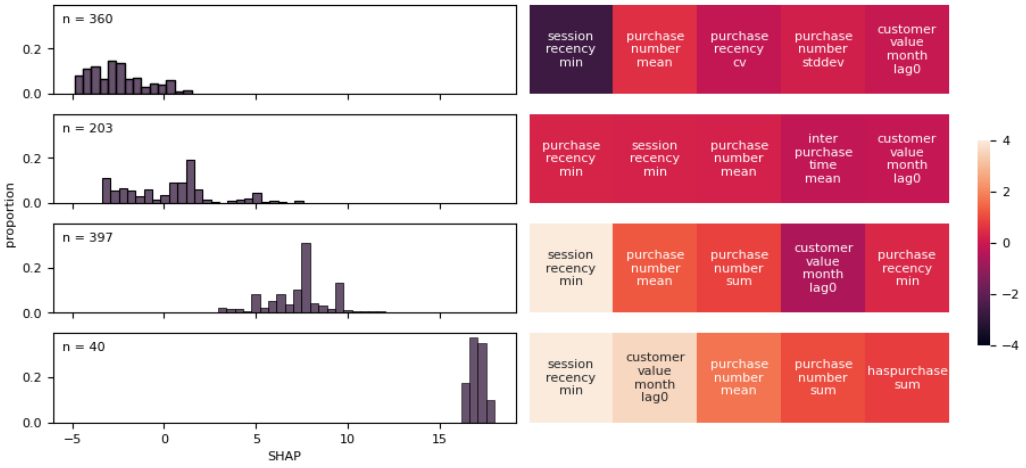
predikci modelu je doba uplynulá od poslední návštěvy. Zdá se, že inkrementální dopad zařazení do retenční aktivity roste spolu s dobou od poslední návštěvy. K poslední interakci věrných zákazníků došlo zpravidla v uplynulém týdnu. U zákazníků, kteří s platformou neinteragovali déle než měsíc, se vliv atributu ustálí. Další vlivnou proměnnou jest průměrné pořadí nákupů, tj. průměrný počet nákupů. Také v tomto případě exploatuje model podmínky pro výběr zákazníků. Průměrné pořadí transakcí dělí příspěvek k inkrementálnímu ekonomickému výsledku na dvě části, kde pokud zákazník realizoval méně než dvacet transakcí (purchase_number_mean<=10.5), pak očekáváme spíše výsledek záporný, pro vyšší počet transakcí (purchase_number_mean>10.5) naopak očekáváme spíše kladný výsledek. Významný je i objem realizovaných transakcí, kde je možné pozorovat silný pozitivní vztah. S ohledem na predikovaný inkrementální zisk aktivity je tedy vhodné cílit na zákazníky s vyšší současnou hodnotou (customer_value_month_lag0>750), což je v souladu s představeným výpočtem zisku retenční kampaně.



Obr. 55 SHAP hodnoty proměnných, významných pro predikci inkrementálního zisku retenční kampaně, včetně pozorovaného rozdělení – REES46

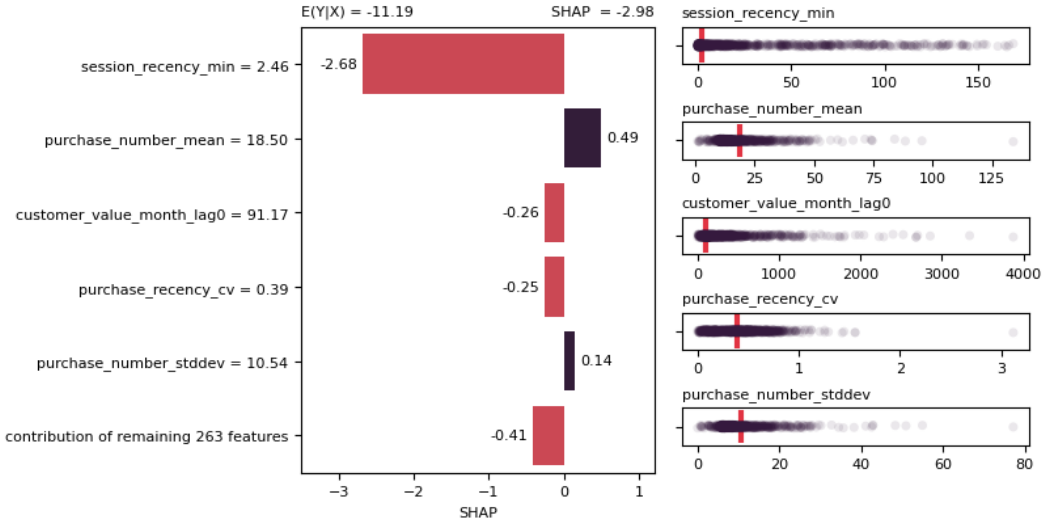
Další pochopení jevu umožňují shluky zákazníků, které pro regresní model ilustruje Obr. 56. První shluk obsahuje zákazníky, jejichž zařazení do kampaně by vedlo ke ztrátě. Jedná se zpravidla o zákazníky, kteří v nedávné době navštívili platformu společnosti a realizovali vyšší počet transakcí v minulosti, aktuálně tak vykazují nižší zákaznickou hodnotu. Druhý shluk

obsahuje zákazníky, u nichž není zcela jasné, zda je do retenční aktivity zařadit. Tito zákazníci vykazují starší interakce i nákupy, než je tomu u prvního shluku, vyšší objem těchto nákupů však vede k vyšší zákaznické hodnotě. Třetí a čtvrtý shluk popisuje zákazníky, jejichž zařazením to retenční kampaně je možné dosáhnout pozitivního ekonomického výsledku aktivity. Ekonomický dopad je ovlivněn především vysokým stářím poslední návštěvy, velmi vysokým peněžním objemem realizovaných transakcí, vysokým počtem nákupů, a související zákaznickou hodnotou.



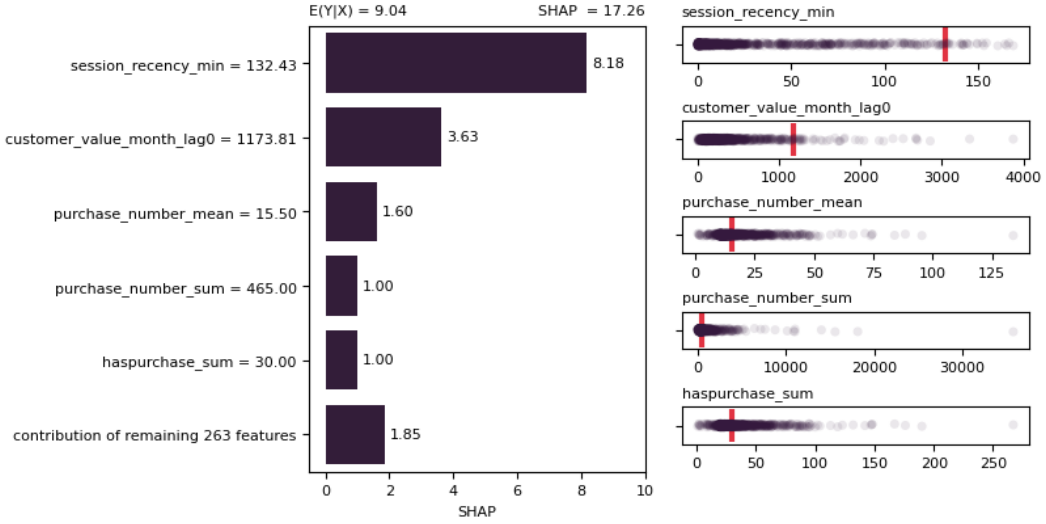
Obr. 56 Zákaznické shluky SHAP hodnot v predikci inkrementálního zisku retenční kampaně, včetně klíčových vysvětlujících proměnných – REES46

Pro bližší pohled na zákaznické shluky je využito lokální interpretovatelnosti, tj. pro každý shluk určíme pozorování nejblíže příslušnému těžišti. Na vybrané pozorování pak je nahlíženo prostřednictvím významných SHAP hodnot i vysvětlujících proměnných. Pro demonstraci přístupu využíváme protiklad zákazníků, které je ekonomicky výhodné (čtvrtý shluk) nebo nevýhodné (první shluk) do retenční kampaně řadit. Pozorování reprezentativní pro první shluk zobrazuje Obr. 57, kde pro zákazníka user_id 600479056 předpokládáme inkrementální dopad zahrnutí do retenční kampaně ve výši −11.19 CU. Ztráta je oproti očekávané základní inkrementální ekonomické hodnotě vyšší, především díky uživatelskému chování i nižší zákaznické hodnotě.



Obr. 57 Dopady významných vysvětlujících proměnných na predikci inkrementálního zisku retenční kampaně, který je těžištěm shluku zákazníků, které je nevýhodné uvažovat v retenční kampani – REES46

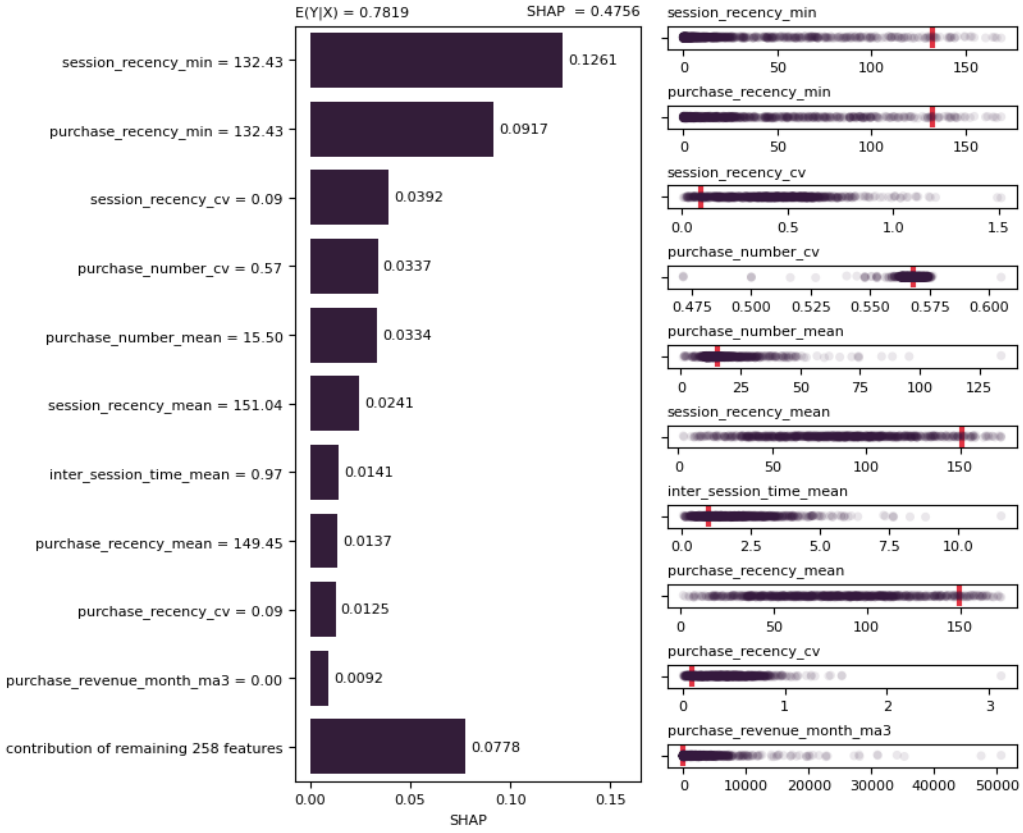
Pozorování reprezentativní pro čtvrtý shluk je vyneseno na Obr. 58, kde pro zákazníka user_id 537404834 přepokládáme inkrementální dopad zahrnutí do retenční kampaně ve výši 9.04 CU. Zisk je oproti očekávané základní inkrementální ekonomické hodnotě vyšší, především díky vysokému stáří uživatelské relace a vysoké aktuální hodnoty zákazníka, což souvisí s počtem a objemem realizovaných nákupů.



Obr. 58 Dopady významných vysvětlujících proměnných na inkrementálního zisku retenční kampaně, který je těžištěm shluku zákazníků, které je vhodné oslovit v rámci retenční kampaně – REES46

Představené srovnání odhaluje u zákazníků, u kterých je zařazení do retenční aktivity nevýhodné, především nedávnou uživatelskou relaci, průměrný počet transakcí a nižší hodnotu zákazníka. Naproti tomu u zákazníků, které je vhodné oslovit, pozorujeme vysokou i zákaznickou hodnotu a počet transakcí; současně také značné stáří poslední uživatelské relace.

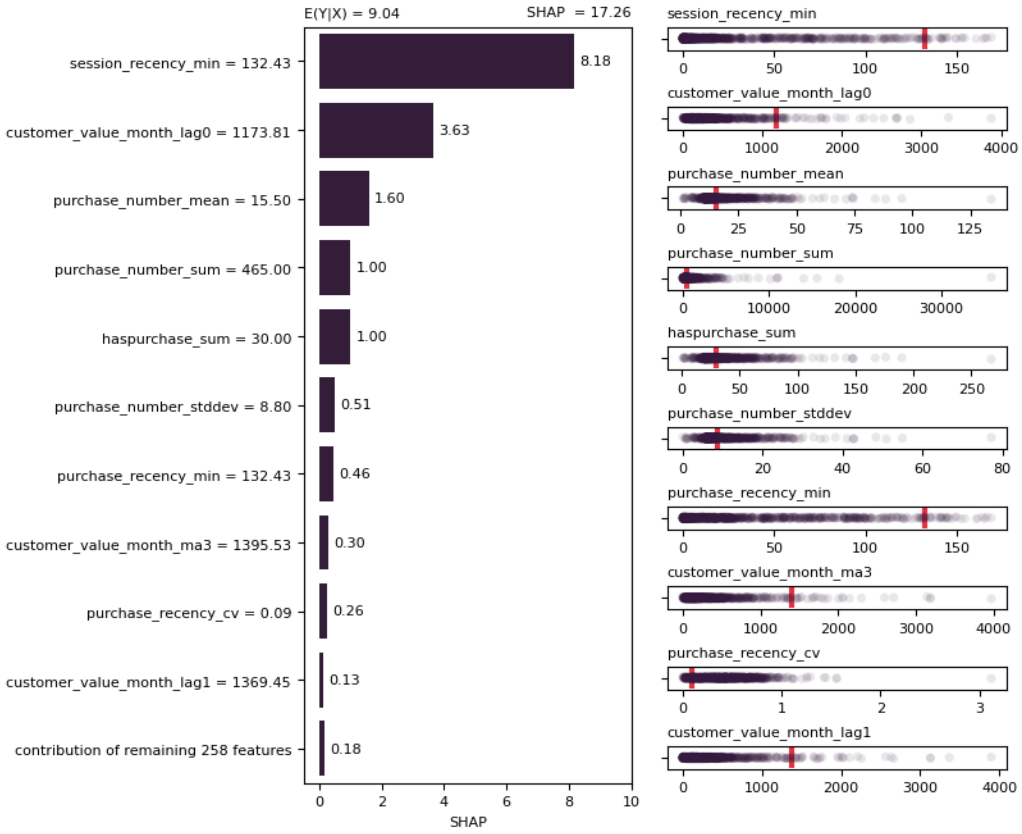
Úspěšná retenční kampaň směřuje na zákazníky, u nichž lze předpokládat kladný inkrementální ekonomický výsledek aktivity. Proto je vhodné podrobněji nahlédnout na těžiště posledního shluku, tj. zákazníka s user_id 537404834 a to perspektivou modelu predikujícího pravděpodobnost ztráty i modelu predikujícího očekávaný inkrementální dopad zařazení daného zákazníka. Pohledem klasifikačního gradient boosting modelu je zřejmé že zákazník velmi dlouho s platformou prodejce neinteragoval, ani nenakupoval. Současně jsou však variační koeficienty obou veličin nízké, tj. v minulém období docházelo k návštěvám i nákupům krátce po sobě. Významným se zdají být i různé reprezentace počtu nákupů, v této dimenzi lze dané pozorování považovat za zákazníka s běžným počtem transakcí, viz Obr. 59.



Obr. 59 Detail těžiště shluku zákazníků, které je výhodné zařadit do retenčních aktivit pohledem klasifikačního modelu – REES46

Struktura významných faktorů je u regresního rozhodovacího stromu odlišná, klíčové je ale i v tomto případě stáří návštěv, nákupů, i počet realizovaných nákupů. Velmi významným faktorem je však také aktuální i historická hodnota zákazníka. Model tedy cílí na zákazníky, kteří dlouho neinteragovali s platformou, mají nižší počet nákupů a vysokou zákaznickou hodnotu, která klesá, viz Obr. 60.

Interpretace prediktivních řešení nastiňuje možné směřování retenční kampaně. Vysoké stáří zákaznických návštěv omezuje výběr vhodných kontaktních kanálů, tj. není možné kontaktovat zákazníka během probíhající seance. Zákaznické shluky odkrývají některé další faktory, především s ohledem na pozorovanou hodnotu zákazníka, zde pozorujeme potenciál pro diferenciaci retenční kampaně pro třetí a čtvrtý shluk, především z hlediska výše a formy incentivy. Vysvětlující proměnné bohužel reflektují především zákaznické chování, tj. nepodařilo se identifikovat skutečné příčiny odlivu zákazníků.



Obr. 60 Detail těžiště shluku zákazníků, které je výhodné zařadit do retenčních aktivit pohledem regresního modelu – REES46

6 Shrnutí a diskuse

6.1 Realizace výzkumu

Sekundární výzkum

Úvodní kapitolu, popisující teoretická východiska práce, autor zaměřuje na vymezení pojmů a specifik elektronického obchodování, řízení vztahů se zákazníky a strojového učení. Význam zvoleného odvětví dobře ilustruje překotný růst elektronického maloobchodu na úkor maloobchodu tradičního. Výrazným zástupcem je společnost Amazon.com Inc., za jejímž úspěchem stojí orientace na technologické inovace, strojové učení a zákaznickou zkušenost, respektive spokojenost (Mackenzie et al., 2013; Morgan, 2018; Terdiman, 2018).

V části věnované řízení vztahů se zákazníky práce popisuje postupující příklon k zákaznické orientaci firem, tj. k vnímání zákazníka jako středobodu podnikových aktivit, včetně některých trendů v oblastech chování spotřebitele, trhu a marketingových funkcí. Významný pohled na způsob tvorby hodnoty pro zákazníka prostřednictvím zákaznické orientace a moderních technologií předkládají Kumar & Reinartz (2018), kteří se zaměřují na ekonomické dopady marketingových aktivit než zákaznickou spokojenost nebo loajalitu. Dawkins & Reichheld (1990), Gupta et al. (2004) a Buttle & Maklan (2019) poukazují na ekonomický význam marketingových aktivit jejichž cílem je udržení souvislého vztahu se zákazníky, tj. retenčního managementu. Za určující pro směřování disertační práce považuje autor aspekty individuální retenční kampaně, které směřují za rámec identifikace zákazníků, kteří se chystají vztah se společností rozvázat. Ascarza et al. (2018) uvádějí snahu o odhalení příčin zamýšleného odchodu, výběr, na jaké ze zákazníků cílit, načasování retenční kampaně, návrh incentivy, a v neposlední řadě celkové vyhodnocení retenčního úsilí.

Teoretická východiska uzavírá sekce věnovaná strojovému učení, tj. prostředku učení počítačů na základě dat nebo interakcí. Géron (2019) spatřuje příležitosti k využití strojového učení především v úlohách, které nelze řešit soustavou pevně stanovených pravidel, které se vyvíjí v čase, nebo jsou příliš rozsáhlé. Kapitola se zabývá přístupy k procesu učení, vztahem k novým instancím datového souboru, nebo zobecněním zachycených znalostí. Významná část je věnována selekci a hodnocení systémů, konkrétně pojetím experimentu, ukazateli úspěšnosti, ale i různými přístupy k interpretaci. Text dále předkládá popis a vlastnosti oblíbených klasifikačních a regresních algoritmů.

Následující kapitolou je literární rešerše, která popisuje aktuální stav vědeckého poznání v oblasti modelování ztráty zákazníka, a to nejprve v celé šíři, a následně s ohledem na specifika elektronického obchodování. Cílem je identifikovat výzkumné mezery, na jejichž základě bude možné formulovat další směřování disertační práce. Nejprve autor prostředky výpočetní lingvistiky analyzuje obsah široké škály vědeckých textů. Za tímto účelem využívá strukturálních modelů témat (Roberts et al., 2016), které umožňují vysvětlit výskyt témat pomocí dalších nezávislých faktorů. Mezí pozoruhodná zjištění je lze zařadit nesoulad hojně citovaných a méně častých témat, která postihují počítačové experimenty, hodnocení modelů, nebo rozpory mezi řešenou úlohou a potřebami podniku. V přehledových článcích Britto & Gobinath (2020), Jain et al., (2021), Ngai et al. (2009) se podobné skutečnosti identifikovat nepodařilo. Zevrubné shrnutí závěrů počítačově asistované analýzy literatury je obsahem sekce 2.1.4.

Prostředky tradiční rešerše pak autor využívá ke zkoumání textů zaměřených na predikci odchodu zákazníka v prostředí e-commerce, analýza literatury je strukturována s využitím metodického rámce CRISP-DM (Chapman et al., 2000). Ukazuje se, že úvahy o ekonomickém dopadu retenčních aktivit jsou podceňovány. Mezi výjimky náleží Coussement & De Bock (2013), Castro & Zsuzuki (2015), Tamaddoni et al. (2014) a Lee et al. (2020), kteří nahlíží na úspěšnost vybraných klasifikačních modelů prostřednictvím dynamiky retenčních nákladů a výnosů představené Neslin et al. (2006). Z přehledových článků reflektuje potřebu hodnocení ekonomických dopadů pouze Ahn et al. (2020). Nastíněné hledisko autoři užívají výhradně pro hodnocení modelů, dílčí kroky jako výběr nezávislých proměnných nebo optimalizace vnějších parametrů tento aspekt nereflektují. Za podceňovaný aspekt považuje autor i snahu o bližší porozumění modelovanému fenoménu. Interpretace zákaznických motivům případně identifikace společných vlastnosti ohrožených zákazníků mohou vést k vyšší úspěšnosti retenčních kampaní podniku. Existujícím projevem takové snahy je identifikace významných nezávislých proměnných, kde však nedochází k silnému konsensu, což lze zčásti vysvětlit povahou jednotlivých odvětví a dostupností dat. Zkoumané práce zpravidla řadí proměnné dle významu, nedochází ale k analýze směru, síly a charakteru vztahu. Vybočují práce Song et al. (2004) a Kim et al. (2005), které porovnávají sousedící shluky setrvávajících a ohrožených uživatelů, čímž je možné odhalit problematické faktory a na ty následně příslušnými nástroji reagovat. V rámci přehledových článků se podobné aspekty identifikovat nepodařilo. Zevrubné shrnutí závěrů tradiční rešerše je obsahem kapitoly 2.2.7.

Primární výzkum

Další směřování disertační reflektuje slepá místa, jež se podařilo identifikovat v rámci sekundárního výzkumu. Hlavním cílem primárního výzkumu je tvorba prediktivního řešení reflektujícího ekonomické dopady retenčních aktivit i porozumění modelovanému fenoménu. Návrh a implementaci systému strojového učení autor strukturuje, s využitím referenčního modelu CRISP-DM (Chapman et al., 2000), do vymezení problému, porozumění datovému souboru, zpracování datového souboru, modelování, vyhodnocení a interpretace, a aplikace řešení. Vymezení problému je úzce spjato s cílem práce, podsekce se zabývá popisem původního přístup k výpočtu zákaznické hodnoty, i odhadu ekonomického dopadu retenční kampaně vycházejícího z Tamaddoni et al. (2014).

V rámci porozumění datovému souboru se autor věnuje akvizici dostupných datových souborů, simulaci produktové marže, konstrukci a exploraci modelu zákazníka. Pro potřeby primárního výzkumu je využito datových sad Retail Rocket (2017) a REES46 (2020), které reprezentují historii interakcí uživatele s nabízeným produktem a související atributy. Původní vlastnosti navíc jsou rozšířeny o počítačově simulovanou úroveň marže.

Datovou reprezentaci problému, tzv. model zákazníka, vymezuje disertační práce jak s pomocí tradiční závislé binární proměnné, určující zda zákazník v budoucím období nakoupí či nikoliv, tak originální spojitou proměnou, která reflektuje inkrementální ekonomický dopad zařazení zákazníka do retenční kampaně. Nastíněný přístup umožňuje překlenout nesoulad mezi procesem konstrukce prediktivního řešení a jeho hodnocením, na který poukazujeme v rámci sekundárního výzkumu. Skupiny nezávislých proměnných vycházejí z transakčního i netransakčního zákaznického chování, na které je nahlíženo prizmatem stáří, frekvence a peněžní hodnoty, čímž autor spojuje přístup obvyklý v sektoru her a služeb, ale i maloobchodu. Původní je zavedení zákaznických preferencí, prostřednictvím latentních faktorů doporučovacích systémů, což vede k vyšší variabilitě proměnných. Explorativní analýza datové reprezentace odkrývá další společné rysy, kde u závislé proměnné klasifikační úlohy je obvyklá dominance jedné z tříd. Regresní proměnné vykazují zápornou střední hodnotu a asymetrické rozdělení, tj. pokud bychom zákazníky do uvažované retenční kampaně vybírali náhodně, pak bude retenční kampaň generovat ztrátu. U nezávislých proměnných také je možné pozorovat asymetrii, nízkou hustotu ale i vnitřní korelaci napříč proměnnými.

Navazující zpracování datového souboru reflektuje uvedená zjištění, autor využívá škálování, eliminaci proměnných s nízkou variabilitou, a výběr proměnných. Pro klasifikační úlohy

je zvažováno vzorkování instancí datového souboru, u regresních úloh transformace závislé proměnné. K výběru a nastavení jednotlivých kroků dochází během optimalizace vnějších parametrů prediktivního systému; pojetí je inspirováno prací Feurer et al. (2015).

Modelování sestává z návrhu, přístupu k hodnocení a konstrukce systému strojového učení. Pro dílčí úlohu, algoritmus a časový řez je určen skelet modelu, vymezen prostor vnějších parametrů, který je prohledáván za účelem nalezení přijatelného nastavení systému. Výsledný model je kalibrován a využit ke konstrukci predikcí. Na schopnosti modelu nahlíží autor s pomocí přirozených ukazatelů úspěšnosti, doby potřebné k sestavení modelu, ale také s pomocí odhadovaného ekonomického dopadu retenční aktivity informované příslušným řešením.

Významným aspektem hodnocení modelu je pojetí experimentu, skutečné podmínky využití systému jsou inspirovány křížovou validací časových řad (Hyndman & Athanasopoulos, 2013). Pro konstrukci trénovací množiny dat je využito seskupení historických výřezů (Gattermann-Itschert & Thonemann, 2021). Na schopnosti řešení nahlíží autor přirozenými ukazateli a dobou potřebnou pro konstrukci řešení. V rámci klasifikační úlohy vychází především z matice záměn (ACC, F1), využívá i pravděpodobnost příslušnosti k dané třídě (AUCROC). U regresní úlohy se autor soustředí na podíl vysvětlené variability (R^2), případně odchylky predikce od pozorovaných hodnot závislé proměnné (MAE, MSE). Perspektivu ekonomického dopadu retenční kampaně reflektuje originálním přístupem, který umožňuje odhadnout dopad kampaně s ohledem na zařazení individuálního zákazníka do retenční aktivity.

Jádro systému strojového učení tvoří populární klasifikační a regresní algoritmy, kam lze řadit zobecněné lineární modely, podpůrné vektory, umělé neuronové sítě, rozhodovací stromy a meta-algoritmy. K vhodnému nastavení vnějších parametrů je přistupováno s pomocí Bayesovské optimalizace (Bergstra et al., 2013).

Pro posouzení přirozených ukazatelů a potřebného výpočetního času je využito prostého srovnání odhadů středních hodnot a souvisejících intervalů spolehlivosti spočtených na testovací části dat. Přirozené hledisko je doplněno analýzou kompromisu mezi vychýlením a rozptylem predikcí. Na ekonomický dopad prediktivních řešení je nahlíženo s pomocí očekávaného a skutečného zisku uvažované retenční kampaně, přesněji řečeno náležitými odhady středních hodnot a intervalů spolehlivosti. Statistický význam rozdílu mezi tradičními klasifikačními přístupy a novými regresními přístupy je hodnocen s pomocí Wilcoxonova testu. Schopnost

řadit zákazníky dle očekávaného zisku je dále analyzována prostřednictvím křivek očekávaného a skutečného kumulativního zisku zařazení daného zákazníka do retenční aktivity.

Pro dosažení lepší srozumitelnosti prediktivního řešení je využito agnostického přístupu SHAP (Lundberg & Lee, 2017), jenž umožňuje sjednotit pohled na globální a lokální interpretaci systému. Vlastní postup staví na distribuovaném výpočtu Shapleyho hodnot, na které je nahlíženo skrze celkové vztahy mezi veličinami, i prostřednictvím vybraných datových instancí. Použité vizuální prvky rozšiřují původní nástroje o vlastnosti datového souboru a přístup ke shlukování instancí.

Globální perspektiva se zaměřuje na reflexi vztahů skrze atributy, ale i skupiny instancí datového souboru. Nejprve jsou identifikovány význačné nezávislé proměnné, nejvlivnější z nich jsou zkoumány s cílem porozumět síle, směru a charakteristice vztahu, ale i robustnosti odhadu. Pro identifikaci zákaznických skupin, ke kterým model přistupuje podobným způsobem, je aplikováno shlukování napříč SHAP hodnotami. Navržený přístup umožňuje nahlédnout na rozložení SHAP hodnot, ale i identifikaci směru a síly význačných atributů pro každý shluk. Lokální perspektiva se soustřední na instance, které dobře zastupují zákaznické shluky. Vizuální prvky slouží k identifikaci směru a síly působení vysvětlující proměnné, i polohy pozorování v rámci datového souboru.

V rámci aplikace řešení je kladen důraz na praktické aspekty výzkumu, jako jsou technologická koncepce řešení a odhad nákladů na provoz systému. Reference na datové soubory zákaznických modelů i programový kód aplikace je možné nalézt v příloze B1.

6.2 Výzkumné otázky

Následující odstavce shrnují závěry, kterých bylo v rámci uvažovaných výzkumných otázek dosaženo. Výstupy jsou dále diskutovány v kontextu relevantní vědecké literatury.

VO1: Jaké jsou výzkumné mezery současného poznání v oblasti predikce ztráty zákazníka v daném kontextu?

Otázka je adresována prostřednictvím literární rešerše, kde je nejprve, s využitím výpočetní lingvistiky, analyzován obsah široké škály textů zaměřených na modelování ztráty zákazníka. Následuje tradiční rešerše podmnožiny textů relevantních pro elektronické obchodování. Srovnáním obou větví literární rešerše jsou identifikovány oblasti vhodné pro další výzkum, které jsou využity jak pro formulaci odpovědi na kýženou výzkumnou otázku, tak i k určení dalšího směřování disertační práce.

S pomocí strukturálních modelů témat se podařilo odkrýt nesoulad hojně citovaných a nepříliš prevalentních témat „classification performance“ a „economic performance“, která se zabývají experimenty, hodnocením modelů, ale i rozpor mezi řešeným problémem a potřebami podniku. V přehledových článcích Britto & Gobinath (2020), Jain et al., (2021), Ngai et al. (2009) autoři na podobné skutečnosti neupozorňují. Shrnutí počítačově asistované analýzy literatury je obsahem kapitoly 2.1.4.

Tradiční rešerší se podařilo obnažit opomíjené aspekty retenčního managementu, které mířící za rámec identifikace rizikových zákazníků. I přes výjimky jako jsou Coussement & De Bock (2013), Castro & Zsuzuki (2015), Tamaddoni et al. (2014) nebo Lee et al. (2020) se zdá, že ekonomickému dopadu retenčních aktivit není věnována přílišná pozornost. Uvedené práce využívají ekonomické hledisko výhradně k hodnocení prediktivních systémů, dílčí kroky konstrukce řešení jako výběr nezávislých proměnných, optimalizace vnějších parametrů nebo konstrukce modelu tuto perspektivu nereflektují. Další podceňovanou oblastí je snaha o bližší porozumění modelovanému fenoménu, což je jedním z význačných teoretických východisek řízení vztahu se zákazníky. Pozorovaným projevem snahy je identifikace nezávislých proměnných, na které vybrané modely spoléhají, bohužel nedochází ke zkoumání směru, síly nebo

charakteru vztahů. Z hlediska skupin významných proměnných nebyl pozorován silný konsenzus. Závěrečný přehled klasické rešerše podmnožiny zkoumaných prací je obsahem kapitoly 2.2.7.

Zásadní výzkumné mezery současného poznání tak autor spatřuje především v nedostatečné pozornosti věnované podnikového kontextu predikce ztráty zákazníka, tj. další otázky retenčního managementu. Za oblasti zájmu považuje ekonomický dopad retenčních aktivit a bližší porozumění modelovanému jevu. Tyto aspekty jsou reflektovány při formulaci cílů a dalších výzkumných otázkách práce.

VO2: Jaké třídy modelů vedou k lepším predikčním schopnostem řešení?

V disertační práci uvažuje autor dva přístupy k modelování ztráty zákazníka, kde první pojetí vymezuje ztrátu zákazníka tradičním způsobem, jako absenci transakcí v budoucím období, kterou chápe jako úlohu klasifikační. Druhé pojetí zavádí ekonomický inkrementální dopad zařazení zákazníka do retenční aktivity, jako spojitou závislou proměnnou, jedná se tedy o úlohu regresní. K hodnocení dochází s pomocí časově odlišené křížové validace, napříč časovými řezy. Pokud uvažujeme o klasifikační úloze, pak v rámci obou datových souborů vyčnívají především umělé neuronové sítě, náhodné lesy a gradient boosting. Uvedené přístupy dosahují nejlepších výsledků napříč ukazateli ACC, F1, i AUCROC. Náhodné lesy a gradient boosting dominují i v rámci úlohy regresní. Regresní metody dosahují nejlepších, nebo srovnatelných výsledků napříč ukazateli R^2, MAE, i MSE. Pro praktické použití v rámci obou úloh se zdají být vhodné zejména meta-algoritmy, a to jak s ohledem na predikční schopnosti, tak i časem potřebným ke konstrukci řešení. Zevrubná vyhodnocení predikčních schopností modelů v přirozených ukazatelích jsou, pro datový soubor Retail Rocket náplní podkapitoly 5.1.1, respektive podkapitoly 5.2.1 pro datový soubor REES46.

Podobně i Wang et al. (2019), Venkatesh & Jeyakarthic (2020) a Almuqren et al. (2021) prokazují schopnosti umělých neuronových sítí. Význam meta-algoritmů naproti tomu podporují výsledky experimentů zahrnujících metody „bootstrap aggregating“ (Coussement & De Bock, 2013; Rachid et al., 2018; Rothmeier et al., 2021), nebo „gradient boosting“ (Tamaddoni et al., 2014; Milosevic et al., 2017). Relevanci algoritmů potvrzují i odpovídající témata

detekovaná metodami výpočetní lingvistiky. Lze tedy tvrdit, že v perspektivě přirozených ukazatelů predikčních schopností jsou dosažené výsledky v souladu s existující literaturou.

VO3: Jaké třídy modelů vedou k lepším ekonomickým výsledkům retenční kampaně?

V rámci ekonomických výsledků se autor soustředí především na dosažený zisk kampaně, kde zahrnutí zákazníka do retenční aktivity chápe jako funkci prediktivního systému. Při pohledu na pořadí úspěšných klasifikačních modelů v přirozených a ekonomických ukazatelích lze pozorovat téměř perfektní shodu. Mezi regresními modely takový vztah neplatí, vynikají především řešení využívajících rozhodovacích stromů, tj. prosté rozhodovací stromy, náhodné lesy a gradient boosting. Šíře intervalů spolehlivosti, které v prvním datovém souboru obsahují i záporné hodnoty, dobře ilustruje potřebu pečlivého výběru zákazníků do zamýšlené kampaně.

Ze srovnání nejlepších regresních a klasifikačních přístupů vyplývá, že využití regresního přístupu v datovém souboru Retail Rocket vede ke zlepšení dosaženého zisku v průměru o ~ 13.6 %, což odpovídá ~ 11415.4 CU, podobně i v datovém souboru REES46 vede využití regresního přístupu ke zlepšení dosaženého zisku v průměru o ~ 6.1 %, což odpovídá ~ 798.2 CU. Statistický význam rozdílů mezi středními hodnotami zisků napříč časovými řezy je dále porovnán s pomocí párového Wilcoxonova testu. U prvního datového souboru se nepodařilo prokázat, že pozorovaný kladný rozdíl není nahodilý, na vině je nízký počet časových řezů. U druhého rozsáhlejšího datového souboru naopak autor alternativní hypotézu přijímá, tj. využití regresního přístupu zde vedlo na zvolené hladině významnosti k prokazatelnému zlepšení ekonomických výsledků. Pro porozumění řazení zákazníků dle očekávaného inkrementálního zisku jsou zkoumány kumulativní křivky očekávaného a dosaženého zisku kampaně pro nejlepší klasifikační a regresní přístupy. Dosažené výsledky naznačují užitečnost nového, regresního pojetí úlohy. Další dopady představeného přístupu mohou vést ke zlepšením řízení kampaně, kde je možné uvažovat o odhadech ekonomického výsledku, rozpočtu kampaně, případně optimálního počtu a složení cílové skupiny zákazníků. Obšírná zhodnocení ekonomických dopadů uvažovaných retenčních kapaní jsou pro uvažované soubory dat obsahem podkapitol 5.1.2, respektive 5.2.2.

Podobně rozsáhlé srovnání se v rámci zkoumané literatury nevyskytuje, a to ani v rámci klasifikačního pojetí úlohy. Za relevantní lze považovat výsledky Coussement & De Bock (2013), Tamaddoni et al. (2014) a Lee et al. (2020), které potvrzují užitečnost meta-algoritmů v rámci vlastních pohledů na ekonomické aspekty retenčního řízení.

VO4: Jaké vysvětlující proměnné jsou klíčové pro predikci modelů?

Pro bližší interpretaci je pro každé z uvažovaných pojetí a datových souborů vybrán nejúspěšnější z prediktivních systémů. Modely jsou interpretovány na aktuálním časovém řezu s pomocí SHAP nástrojů a navržených rozšíření. V rámci klasifikační úlohy pozoruje autor význam jak tradičních skupin proměnných popisujících stáří a frekvenci uživatelských interakcí, tak i skupin proměnných popisujících chování uvnitř uživatelské seance. Ke shodě na konkrétních nezávislých proměnných dochází pouze u proměnné reprezentující stáří poslední relace, což je do značné míry způsobeno odlišnostmi mezi soubory dat a konstrukcí řešení. Zajímavý je proto obdobný charakter vztahu mezi pravděpodobností ztráty zákazníka a stářím poslední uživatelské relace. Ukazuje se, že s rostoucím stářím relace roste i pravděpodobnost odchodu zákazníka, která se po v určitém bodě ustálí. V obou případech je tento bod dosažen po přibližně třech týdnech, což naznačuje asociaci s časovým vymezením závislé proměnné. Dále lze pozorovat exploataci minimálního množství transakcí pro výběr zákazníků, jev se však projevuje prostřednictvím odlišných nezávislých proměnných. Podrobnosti, včetně vizuální ztvárnění, je možné pro datový soubor Retail Rocket nalézt v podkapitole 5.1.3; pro datový soubor REES46 v podkapitole 5.2.3.

Význam faktorů popisujících transakční a netransakční chování uživatele, včetně interakcí uvnitř relace, naznačují v podobném kontextu i Abbasi et al. (2015) a Rachid et al. (2018). Množiny zákaznických preferencí, a data a času se naopak ukazují jako méně významné, což je v rozporu s úvahami Gordini & Veglio (2017) a Li & Li (2019). Zdá se tedy, že dobrým společným základem modelu zákazníka pro klasifikační pojetí predikce ztráty zákazníka v daném odvětví jsou stáří, frekvence a interakce uvnitř uživatelské seance. Ostatní proměnné popisující peněžní aspekty chování, preference, nebo datum a čas se ukazují jako méně významné.

V rámci regresní úlohy je třeba upozornit na důležitost nezávislých proměnných charakterizujících stáří interakcí, ale i skupiny proměnných popisující nákupní chování uživatele, včetně hodnoty zákazníka. Ke shodě na konkrétních nezávislých proměnných dochází v případě stáří poslední uživatelské relace, počtu nákupů a zákaznické hodnoty. Vyčnívá především stáří poslední relace. Ukazuje se, že s rostoucím stářím relace roste i inkrementální zisk zařazení zákazníka do retenční aktivity, který se po v určitém bodě ustálí. Uvedený aspekt zákaznického chování odráží vztah mezi závislými proměnnými klasifikačního a regresního pojetí. Počet nákupů napomáhá k určení hodnotných zákazníků, současně také exploatuje proces sestavení datových souborů, podobně jako u tradičního klasifikačního pojetí problému. Zákazníky, které je ekonomicky výhodné do kampaně zařadit, je možné přesněji popsat s pomocí zákaznické hodnoty. Je zřejmé, že faktory významné pro oba datové soubory, těsně reflektují představený způsob konstrukce inkrementálního zisku retenční kampaně.

Původní přístup k problému omezuje přímé srovnání s existující literaturou. Pokud je ovšem regresní pojetí zasazeno do kontextu významných faktorů pojetí klasifikačního, pak lze pozorovat shodu na důležitosti transakčního a netransakčního chování uživatele. Nad rámec těchto faktorů dochází k růstu významu peněžní hodnoty transakčních interakcí, což lze s ohledem na definici inkrementálního zisku retenční kampaně očekávat.

VO5: Jaké společné znaky vykazují zákazníci, na které je vhodné retenční aktivity cílit?

Autor se soustředí, podobně jako v předchozí výzkumné otázce, na interpretaci nejlepších řešení. Modely zkoumá na aktuálním časovém řezu s pomocí SHAP nástrojů a navržených rozšíření. Zákazníci jsou nejprve rozřazeni do shluků, dle SHAP hodnot napříč vysvětlujícími proměnnými. Shluk tak sdružuje pozorování, u kterých prediktivní systém uvažuje obdobnou povahu asociací mezi vysvětlovanou a vysvětlovanými proměnnými. Pro bližší porozumění je využito lokální interpretovatelnosti, kde pro zkoumaný shluk určíme pozorování nejblíže odpovídajícímu těžišti. Na pozorování je pak nahlíženo prostřednictvím SHAP hodnot i vysvětlujících proměnných.

Pozornost je nejprve věnována ohroženým zákaznickým shlukům, vyplývajících z klasifikačního pojetí problému. Napříč datovými soubory se ukazuje jako klíčové stáří poslední

relace. V rámci Retail Rocket navíc vyčnívá chování uvnitř uživatelské relace, kde vysoké prodlevy mezi transakčními interakcemi vedou k vyššímu předpokládanému riziku ztráty. V datovém souboru REES46 zase identifikují rizikové zákazníky vysoké hodnoty stáří poslední transakce a vysoký počet transakcí. Ohrožené zákazníky tedy lze odhalit prostřednictvím některých aspektů uživatelských relací a transakční historie. Nedostatkem je však omezená využitelnost předestřených skutečností pro porozumění zákaznickému rozhodování.

S ohledem na ekonomické výsledky retenčních aktivit se zdá být užitečný především regresní přístup k úloze. Cílové shluky v tomto pojetí sdružují zákazníky, u kterých lze očekávat kladný inkrementální výsledek zařazení do kampaně. Napříč datovými soubory vyčnívá stáří poslední relace; nově na významu nabývají reprezentace frekvence a peněžní hodnoty transakcí, což reflektuje způsob konstrukce inkrementálního zisku retenční kampaně. Rozdíly pozorujeme v pořadí a struktuře proměnných, kde v Retail Rocket patří první příčky transakčním aspektům chování, což odráží odlišnosti v chování zákazníků.

Nastíněná interpretace představuje možné směřování retenčního úsilí. V obou datových souborech se jako významné ukazuje stáří poslední návštěvy, což omezuje výběr kontaktních kanálů. Charakter vztahu je možné využít jako podklad pro automatizaci dalších marketingových aktivit, eg. doporučení relevantního obsahu nebo služby. Užitečné je také odlišení zákaznických shluků dle očekávané střední hodnoty inkrementálního výsledku retenční kampaně, otevírající prostor pro diferenciaci retenčního úsilí především s ohledem na výši a formu incentivy. Významné vysvětlující proměnné reflektují spíše zákaznické chování, tj. nepřispívají k identifikaci skutečných příčin odlivu zákazníků. Detailní rozbor problematiky je možné dohledat v podkapitolách 5.1.3 a 5.2.3.

Obdobnou snahu o interpretaci cílové skupiny retenční aktivity je možné pozorovat v Song et al. (2004) a Kim et al. (2005), kteří srovnávají blízké skupiny setrvávajících a ohrožených uživatelů. Autorem představený přístup k interpretaci však umožňuje těsnější reflexi vztahu k cílové proměnné. Užití SHAP hodnot vede ke zmírnění problémů s vícerozměrnými prostory vysvětlujících proměnných, kterými řešení Song et al. (2004) a Kim et al. (2005) trpí. Značnou výhodou je také reflexe ekonomické perspektivy problému v rámci interpretace modelů.

6.3 Limity a budoucí směřování výzkumu

Datové soubory

Určitá omezení vyplývají z povahy použitých datových souborů, které zachycují interakce v prostředí elektronického maloobchodu, tj. v jiných vztazích nebo odvětvích nemusí být dosažené poznatky platné. Dostupná úroveň detailu odpovídá interakcím uživatele s nabízeným produktem. Průnik vlastností souborů tak limitoval možné směry, kterými se bylo možné vydat, především s ohledem na další vlastnosti zákazníka (mezilidské vztahy, sociálně-ekonomické nebo místní odlišnosti), podniků (mikroprostředí, makroprostředí) atp. Další omezení plynou z dílčích rozhodnutí při sestavení reprezentace zákaznického modelu, limitujícím faktorem je i časový rozsah.

V budoucím výzkumu by tak bylo vhodné ověřit prezentované přístupy napříč různými typy vztahů, případně v odvětvích, která využívají prvky elektronického obchodování a současně reflektují význam retenčního řízení např. telekomunikace, bankovnictví, zábavní průmysl, nebo pohostinství. Dále by bylo dobré získat subjektivní vysvětlující proměnné, nebo uvažovat o takové reprezentaci, která povede ke zlepšení schopnosti daného systému jak s ohledem na prediktivních modelování, tak i interpretaci řešení. Rozšíření časového horizontu pro sběr dat by mohlo být využito pro modelování úspěšnosti aktivit u kterých předpokládáme trvání delší než čtyři týdny.

Strojové učení

Limity vychází z částí z návrhu experimentu, jenž staví na konceptech křížové validace časových řad a seskupení časových výřezů. Pojetí sice vede ke zlepšení prediktivních schopností systémů a dobře reflektuje jejich skutečné nasazení; spojení časových výřezů však vede k porušení nezávislosti datových instancí. Doba, po kterou byla data sbírána, omezuje dostupný počet časových řezů, což následně vede i k vyšší variabilitě zkoumaných ukazatelů.

Řada omezení plyne z návrhu a implementace systému strojového učení, kde autor uvažuje dílčí komponenty, pořadí a vzájemné vazby, vnější parametry aj. Nad rámec selekce některých komponent a optimalizace vnějších parametrů by bylo možné zvážit další přístupy k automatizaci konstrukce prediktivního systému. Inspirací by mohly sloužit práce Olson & Moore (2019), ve které autoři přistupují k sestavení grafu komponent pomocí genetického programování; případně práce Feurer et al., (2020), která využívá předchozích znalostí o chování prediktivního systému a obecných vlastností datového souboru k výběru komponent, prostoru

vnějších parametrů. S ohledem na dosažené výsledky, se zdá výhodné důkladně prozkoumat moderní algoritmy strojového učení, jako jsou meta-learning (Chen & Guestring, 2016; Ke et al., 2017; Dorogush et al., 2018), případně komplexní architektury umělých neuronových sítí (Zai & Brown, 2020; Ferlitsch, 2021; Raff & Borne, 2022), pro které by však bylo třeba zajistit vhodný hardware, tj. grafické procesory (GPU).

Ekonomické dopady retenční kampaně

Omezujícím faktorem je absence atributu, který by reprezentoval marže produktu. Autor atribut doplňuje prostřednictvím počítačové simulace, jejímž cílem je reflexe vývoje veličiny v čase. Zvolený přístup nemusí ukazovat skutečnou úroveň marže produktu, ve svém důsledku tedy nemusí odrážet ekonomickou realitu daných podnikatelských subjektů. Bylo by tedy dobré ověřit dosažené závěry s využitím takového souboru dat, který obsahuje i informaci o maržích podniku.

Dalším limitujícím faktorem jsou dílčí komponenty pro výpočet inkrementálního zisku zahrnutí zákazníka do retenční kampaně, potažmo maximálního zisku retenční aktivity. Navržený výpočet zákaznické hodnoty je úzce svázán s časovým období realizace retenční aktivity skrze délku časového okna cílové proměnné. Hodnotu zákazníka tak nejde dost dobře využít nad rámec časového okna závislé proměnné, pro zevrubnou reflexi životního cyklu zákazníka nebo spočtení zákaznického kapitálu. Do budoucna by bylo možné rozšířit tento přístup tak, aby lépe reflektoval vztah mezi ziskem plynoucím ze zákaznických transakcí a délkou vzájemného vztahu. Další komponenty výpočtu jsou odhadnuty prostřednictvím počítačové simulace, tj. nejsou brány v potaz vlastnosti jednotlivých zákazníků. Zajímavá je především pravděpodobnost přijetí retenční nabídky ohrožených zákazníků, u které předpokládáme rozdílné individuální chování a současně pozitivní korelaci s výší peněžní incentivy. Autor má za to, že explorace a modelování nastíněné části problému povede k dalšímu zdokonalení retenčních aktivit.

Interpretace systémů strojového učení

K porozumění je přistupováno skrze detailní analýzu chování prediktivního systému, což v případě méně spolehlivých modelů může vést k zavádějícím zjištěním, což je závažným konceptuálním limitem. Podobný neduh představuje i náchylnost SHAP přístupu k manipulaci reflektující předsudky autora. Oba problémy jsou zmírněny využitím alespoň dvou datových souborů. V budoucnu by bylo dobré uvažovat více různorodých přístupů k porozumění prediktivnímu systému, což by vedlo k vyšší objektivitě závěrů.

Interpretaci dále limituje komplexní zákaznický model, a to především s ohledem na několik úrovní transformací proměnných a jejich vzájemné vztahy. Možným řešením je implementace výchozího zákaznického modelu jako prvku systému strojového učení, případně oddělení interpretace od původního zákaznického modelu s využitím metod jako LIME (Ribeiro et al., 2016). První navržený přístup by vedl k zásadnímu růstu výpočetní náročnosti, druhý zase k nižší věrnosti interpretace.

Praktickým limitem, je podobně jako v části strojového učení, výpočetní náročnost zvoleného přístupu. Problém je zmírněn vzorkováním pozorování a implementací na distribuované výpočetním systému Apache Spark. V budoucnu by bylo možné zvážit interpretaci pouze vybraných zákaznických shluků, případně přenesení výpočetně náročných částí systému na odpovídající hardware, tj. grafické procesory (GPU).

7 Přínosy práce

7.1 Přínosy pro vědu a výzkum

Klíčovým výstupem disertační práce je posun ve formulaci úlohy predikce odchodu zákazníka od předpovědi absence transakce k inkrementálnímu ekonomickému dopadu retenční aktivity v budoucím období. Nové pojetí zasahuje do konstrukce, hodnocení, i interpretace systému strojového učení; rozšiřuje tak obzor pro nová zjištění a poznatky v oblasti aplikací strojového učení při řízení vztahů se zákazníky.

Mezi pozitivní dopady snahy o interpretaci prediktivního řešení lze řadit užší porozumění modelovanému jevu, zhodnocení dopadu vysvětlujících proměnných na závislých proměnných, a to nejen ve smyslu řazení, ale také ve smyslu směru působení a charakteru zachyceného vztahu, validaci vstupních dat, ale i celkového přístupu k řešení úlohy. Na prezentovaný výzkum je možné navázat v oblastech výběru a konstrukce vysvětlujících proměnných, prezentace a komunikace zachycených znalostí, odhalení podjatosti řešení, případně zkoumáním navazujícího rozhodování v retenčním řízení.

Srovnání komplexních systémů strojového učení je také přínosné, a to jak z hlediska jejich prediktivních schopností, tak i ekonomických dopadů retenčních aktivit. Disertační práce čtenáře informuje o klíčových charakteristikách řešení, ale i možných problémech a omezeních. Benchmark tak může inspirovat vědecké pracovníky k tvorbě nových řešení, metod nebo formulaci navazujících perspektiv. V intencích zkoumaného jevu se podobně rozsáhlou komparaci moderních systémů nalézt nepodařilo.

Dalším z výstupů je zpracování rozsáhlé obsahové rešerše relevantních vědeckých článků, jenž byly zkoumány s pomocí metod zpracování přirozeného jazyka. Analýza je zaměřena na prevalenci skrytých témat, změny prevalence v čase a závislost prevalence na počtu citací. Rešerši literatury lze využít jako podklad pro směřování navazujícího výzkumu. Metodický postup a aplikaci pro počítačově asistovanou analýzu vědecké literatury je možné aplikovat i v dalších oblastech vědeckého zkoumání.

V neposlední řadě disertační práce otevírá možnosti navazujícího vědeckého úsilí poskytnutím podkladových datových souborů a programového kódu aplikace, díky čemuž je možné předestřený výzkum reprodukovat, nezávisle zhodnotit, případně využít jako jeden ze stavebních kamenů dalšího úsilí.

7.2 Přínosy pro podnikatelskou praxi

Výstupem pro podnikatelskou praxi je demonstrace zasazení předpovědi odchodu zákazníka do kontextu retenčního řízení. Mezi dopady lze řadit identifikaci ohrožených zákazníků, na které je výhodné cílit, prioritizaci retenčního úsilí, užší porozumění zákaznickému chování, a zlepšení ekonomických výsledků retenčních aktivit. Kladný vliv retenčního managementu na ekonomické výsledky podniku dokládají texty Dawkins & Reichheld (1990), Gupta et al. (2004), nebo Buttle & Maklan (2019).

Disertační práce může sloužit jako podrobný recept na sestavení systému strojového učení pro predikci ztráty zákazníka, od pojetí k modelovaného jevu, přes shromažďování dat, vytváření datové reprezentace, sestavení a hodnocení prediktivních modelů, až po zajištění srozumitelnosti a vlastní nasazení. Text nabízí nejen návod k vytvoření systému strojového učení, demonstruje i jeho praktickou implementaci.

Možnými příjemci výzkumu jsou dvě skupiny podnikatelských subjektů. První skupinou mohou být úspěšné maloobchodní společnosti, využívající elektronické obchodování jako jeden z prodejních kanálů. Zde by bylo možné rozšířit soubor nástrojů pro prevence odlivu zákazníků. Druhou skupinou jsou podniky zabývající se vývojem IT systémů jako jsou e-commerce platformy nebo systémy pro správu zákaznických vztahů. Zde by bylo výhodné začlenit některé z popsaných přístupů mezi prostředky pro podporu rozhodování.

7.3 Přínosy pro vzdělávání

Za pedagogický přínos disertační práce lze považovat shrnutí vybraných teoretických partií řízení vztahů se zákazníky a strojového učení, případně hlubší vhled do problematiky predikce odchodu zákazníků. Text rovněž nabízí ukázku obsahové analýzy literatury zpracovanou s využitím výpočetní lingvistiky, respektive metod pro zpracování přirozeného jazyka. Studenti, výzkumníci, i odborníci z praxe mohou těžit z přehledu relevantních technik a metod používaných k řešení úlohy, včetně nejnovějších přístupů a jejich silných a slabých stránek. Disertační práce poskytuje ucelený přehled o procesu realizace výzkumu, včetně vymezení modelovaného problému, sběru dat a konstrukce datové reprezentace, sestavení a hodnocení prediktivních modelů, interpretace, nebo aplikace. Text čtenářům přibližuje i některá omezení, se kterými se bylo třeba vypořádat, ať už se jedná o dostupnost dat, srozumitelnost nebo škálovatelnost navrženého systému, aj. V neposlední řadě ilustruje disertační práce potřebu mezioborového

přístupu k řešení podnikových problémů, může tak poskytnout solidní základ pro navazující výzkumnou nebo aplikovanou práci, která se zaobírá průnikem oblastí řízení vztahů se zákazníky a strojového učení. Další využití je podpořeno již zmiňovanou ukázkou implementace systému, včetně datové reprezentace zákaznických modelů.

Závěr

Posun ve vnímání individuálního zákazníka jako těžiště podnikových aktivit je přirozeným hybatelem snah o správu vzájemných vztahů, potažmo úsilí směřovaného k prevenci odchodu a udržení stávajících zákazníků. Gupta et al. (2004), Kumar et al. (2018), Umashanjar et al (2017) aj., dokládají vazbu mezi realizací takového úsilí a ekonomickými výsledky podniku. Podpora retenčních aktivit je tak přirozenou prioritou. Daunis & Iwan (2014) a Handley (2013) však upozorňují na skutečnost, že ani vrcholový management, ani zákazníci nejsou s úrovní těchto snah příliš spokojení. Aby byly retenční snahy úspěšné, je nutné předvídat, který zákazník bude chtít vztah ukončit, a na to reagovat pomocí vhodné pobídky nebo intervence. Prvotním krokem je tedy předpověď odchodu zákazníka, ke které bývá přistupováno s pomocí strojového učení. Nedostatečná predikční schopnost mnoha přístupů naznačuje potenciál využití velkých dat a nových přístupů strojového učení. Nicméně, jak upozorňují Ascarza et al. (2018), úspěšná retenční kampaň zahrnuje i některé opomíjené aspekty, jako jsou porozumění zákaznickému chování a výběr cílové skupiny. Technologický pokrok v oblastech komunikačních a informačních technologií umožňuje podnikům orientaci na individuálního zákazníka, ohledání fenoménu tak probíhá v prostředí elektronického maloobchodu, jenž je produktem těchto změn (Chaffey, 2015).

Disertační práce je uvedena vybranými teoretickými aspekty elektronického obchodování, řízení zákaznických vztahů a strojového učení. Kapitola představuje základní pojmy a východiska nezbytná pro uchopení problematiky. Autor se následně věnuje zevrubné rešerši vědeckých článků zaměřených na predikci odchodu zákazníka. Úsilí je rozděleno do dvou větví, kde v první z větví analyzuje širokou škálu textů prostřednictvím metod zpracování přirozeného jazyka, druhá z větví pak tradičním způsobem zkoumá podmnožinu textů zaměřenou na elektronické obchodování. Slepé skvrny současného poznání byly identifikovány v absenci reflexe podnikového kontextu daného problému, kam náleží ekonomický dopad retenčních aktivit, porozumění modelovanému jevu. Hlavním cílem disertační práce je tak návrh, implementace a zhodnocení systému strojového učení pro predikci odchodu zákazníka v prostředí elektronického maloobchodu, který tyto výzkumné mezery reflektuje.

Návrh a implementace vlastního řešení jsou strukturovány do částí vymezení problému, porozumění a zpracování dat, modelování, ale i vyhodnocení, interpretace a produkční nasazení systému. Za účelem užší reflexe podnikového kontextu je nad rámec závislé proměnné popisující absenci transakce v budoucím období zavedena závislá proměnné charakterizující

inkrementální ekonomický dopad retenčních aktivit. Výzkum využívá dvou datových souborů, Retail Rocket (2017) a REES46 (2020), které sdružují informace o interakcích uživatelů s nabízenými produkty. Po exploraci dat následuje jejich zpracování, zahrnující škálování, eliminaci proměnných s nízkou variabilitou a výběr relevantních proměnných. Pro modelování je využito oblíbených tříd klasifikačních a regresních algoritmů, jako jsou zobecněné lineární modely, podpůrné vektory, umělé neuronové sítě, rozhodovací stromy a meta-algoritmy. Vnější parametry modelů jsou určeny s pomocí Bayesovské optimalizace. Posouzení schopností modelů je provedeno jak s využitím přirozených ukazatelů, tak prizmatem předpokládaného ekonomického dopadu zamýšlených retenčních opatření. Pro porozumění je využito přístupů SHAP (Lundberg & Lee, 2017), která jsou vhodně rozšířeny jak v oblasti implementace, tak v oblasti vizuálních nástrojů. Pozornost je věnována i některým praktickým aspektům využití systému jako jsou technologická koncepce nebo odhad části provozních nákladů.

Autor užívá dvě pojetí modelovaného jevu, tradiční klasifikační a původní regresní, obě nejprve hodnotí s využitím přirozených ukazatelů. V regresní úloze se ukázaly jako význačné náhodné lesy a gradient boosting; u klasifikační úlohy dominovaly navíc i umělé neuronové sítě. Lze tedy doporučit využití meta-algoritmů, zejména s ohledem na prokázanou všestrannost, úroveň predikčních schopností a nízkou časovou náročností. Na řešení je dále nahlíženo perspektivou ekonomického dopadu zamýšlené retenční kampaně, kde se podařilo demonstrovat užitečnost nového regresního pojetí úlohy, zejména v kombinaci s rozhodovacími stromy a meta-algoritmy. I přes pozitivní výsledky originálního přístupu je třeba upozornit na možná omezení ve smyslu přenositelnosti do dalších podniků nebo odvětví.

Významné pro obě pojetí se ukazují vysvětlující proměnné popisující stáří a frekvenci uživatelských interakcí, případně chování uvnitř uživatelské relace. Pro regresní přístup je navíc významná i hodnota zákazníka. Ostatní proměnné charakterizující peněžní aspekty chování, preference, nebo datum a čas se zdají být méně podstatné. Prakticky využitelný pro návrh a automatizaci retenčních aktivit je především popis vzájemných vztahů, směru a síly působení. Neduhem je však skutečnost, že zvolené proměnné odrážejí spíše zákaznické chování než příčiny odlivu zákazníků. Související perspektivou je určení zákazníků, na které je vhodné retenční aktivity cílit. S ohledem na ekonomický dopad jsou významné především výstupy regresního pojetí problému, které u zákaznických shluků s kladnou střední hodnotou inkrementálního výsledku retenční kampaně potvrzují význam zákaznické hodnoty a stáří poslední uživatelské relace. Nad rámec užitečnosti popsané prostřednictvím klíčových proměnných umožňuje tento

pohled bližší porozumění individuálnímu zákazníku, což může vést k diferenciaci pobídek, případně další úpravě zamýšlené kampaně.

Disertační práce představuje nové pojetí modelování odchodu zákazníků prostřednictvím závislé proměnné charakterizující inkrementální ekonomický dopad retenčních aktivit v budoucím období. Součástí je i moderní přístup k interpretaci systémů, doplněný o rozsáhlé srovnání prediktivních schopností a jejich ekonomických dopadů. Pozoruhodná je i obsahovaná analýza literatury, realizovaná metodami zpracování přirozeného jazyka.

V kontextu podnikové praxe lze uvažovat o pozitivním dopadu na ekonomické výsledky retenčních aktivit prostřednictvím cílení na vhodné zákazníky, prioritizace retenčního úsilí, porozumění zákaznickému chování. Mezi uživateli představených nástrojů je možné uvažovat o společnostech zaměřených na elektronický maloobchod, vývoj e-commerce platforem, případně systémů pro řízení zákaznických vztahů.

Za přínosné pro vzdělání a pedagogickou praxi lze považovat shrnutí teoretického úvodu řízení zákaznických vztahů a strojového učení, s důrazem na úlohu predikce odchodu zákazníka. Pozoruhodná je i obsahovaná analýza literatury, která byla zpracována s pomocí metod zpracování přirozeného jazyka. Text ilustruje proces výzkumu, včetně definice problému, sběru dat, vytvoření datové reprezentace, hodnocení a interpretace modelů. Čtenáři jsou seznámeni s omezeními, ale i přínosy takového řešení. Disertační práce tak představuje solidní základ pro navazující výzkum a aplikovanou práci v oblastech řízení vztahů se zákazníky a strojového učení, včetně ukázky implementace systému a datové reprezentace zákaznických modelů.

Literární zdroje

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems: Large-Scale Machine Learning on Heterogeneous Systems*. https://www.tensorflow.org/

Abbasi, A., Lau, R. Y. K., & Brown, D. E. (2015). Predicting behavior. *IEEE Intelligent Systems*, *30*(3), 35-43. https://doi.org/10.1109/MIS.2015.19

Aggarwal, C. C. (2016). *Recommender systems: The Textbook*. Springer Science+Business Media.

Ahmed, A. A. Q., & Maheswari, D. (2019). An enhanced ensemble classifier for telecom churn prediction using cost based uplift modelling. *International journal of information technology (Singapore. Online)*, *11*(2), 381-391. https://doi.org/10.1007/s41870-018-0248-3

Ahn, J., Hwang, J., Kim, D., Choi, H., & Kang, S. (2020). A Survey on Churn Analysis in Various Business Domains. *IEEE Access*, *8*, 220816-220839. https://doi.org/10.1109/ACCESS.2020.3042657

Almuqren, L., Alrayes, F. S., & Cristea, A. I. (2021). An Empirical Study on Customer Churn Behaviours Prediction Using Arabic Twitter Mining Approach. *Future Internet, 13*(7). https://doi.org/10.3390/fi13070175

Alpaydin, E. (2020). *Introduction to Machine Learning: Adaptive Computation and Machine Learning* (4th). MIT Press. https://books.google.cz/books?id=uZnSDwAAQBAJ

Anderson, E. W., & Mittal, V. (2000). Strengthening the Satisfaction-Profit Chain. *Journal of Service Research*, *3*(2), 107-120. https://doi.org/10.1177/109467050032001

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, *4*(none), 40-79. https://doi.org/10.1214/09-SS054

Ascarza, E. (2018). Retention Futility: Targeting High-Risk Customers Might Be Ineffective. *Journal of marketing research*, *55*(1), 80-98. https://doi.org/10.1509/jmr.16.0163

Ascarza, E., Fader, P. S., & Hardie, B. G. S. (2017). Marketing Models for the Customer-Centric Firm. In *Handbook of Marketing Decision Models* (pp. 297-329). Springer International Publishing. https://doi.org/10.1007/978-3-319-56941-3_10

Ascarza, E., Iyengar, R., & Schleicher, M. (2016). The Perils of Proactive Churn Prevention Using Plan Recommendations: Evidence from a Field Experiment. *Journal of Marketing Research*, *53*(1), 46-60. https://doi.org/10.1509/jmr.13.0483

Ascarza, E., Neslin, S. A., Netzer, O., Anderson, Z., Fader, P. S., Gupta, S., Hardie, B. G. S., Lemmens, A., Libai, B., Neal, D., Provost, F., & Schrift, R. (2018). In Pursuit of Enhanced Customer Retention Management: Review, Key Issues, and Future Directions. *Customer Needs and Solutions*, *5*(1-2), 65-81. https://doi.org/10.1007/s40547-017-0080-0

Bergstra, J., Yamins, D., & Cox, D. (2013). *Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures* (Vol. 28, p. -123). PMLR. https://proceedings.mlr.press/v28/bergstra13.html

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bischof, J. M., & Airoldi, E. M. (2012). Summarizing topical content with word frequency and exclusivity. In *ICML'12 Proceedings of the 29th International Coference on International Conference on Machine Learning* (pp. 9-19). Omnipress.

Blattberg, R. C., & Deighton, J. (1996). Manage marketing by the customer equity test. *Harvard business review*, *74*(4), 136-144.

Blattberg, R. C., Kim, B. D., & Neslin, S. A. (2008). *Database Marketing: Analyzing and Managing Customers: International Series in Quantitative Marketing*. Springer New York. https://books.google.cz/books?id=-JwptfFItaoC

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84. https://doi.org/10.1145/2133806.2133826

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal Of Machine Learning Research*, *3*(4-5), 993-1022.

Bohr, J., & Dunlap, R. E. (2017). Key Topics in environmental sociology, 1990–2014: results from a computational text analysis. *Environmental Sociology*, *4*(2), 181-195. https://doi.org/10.1080/23251042.2017.1393863

Botev, Z., & Ridder, A. (2017). Variance Reduction. In *Wiley StatsRef: Statistics Reference Online* (pp. 1-6). Wiley.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145-1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Braun, M., & Schweidel, D. A. (2011). Modeling Customer Lifetimes with Multiple Causes of Churn. *Marketing science (Providence, R.I.)*, *30*(5), 881-902. https://doi.org/10.1287/mksc.1110.0665

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, *24*(2), 123-140. https://doi.org/10.1023/A:1018054314350

Britto, J., & Gobinath, R. (2020). A Detailed Review For Marketing Decision Making Support System In A Customer Churn Prediction. *International Journal of Scientific & Technology Research*, *9*(4), 3698-3703.

Bryman, A. (2012). *Social research methods* (4th ed.). Oxford University Press.

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research*, *164*(1), 252-268. https://doi.org/10.1016/j.ejor.2003.12.010

Buttle, F., & Maklan, S. (2019). *Customer relationship management: concepts and technologies* (4th). Routledge.

Castro, E. G., & Tsuzuki, M. S. G. (2015). Churn Prediction in Online Games Using Players' Login Records: A Frequency Analysis Approach. *IEEE Transactions on Computational Intelligence and AI in Games*, *7*(3), 255-265. https://doi.org/10.1109/TCIAIG.2015.2401979

Commission, E., Eurostat, Corselli-Nordblad, L., & Strandell, H. (2020). *Ageing Europe: looking at the lives of older people in the EU: 2020 edition: looking at the lives of older people in the EU*. Publications Office. https://doi.org/doi/10.2785/628105

Coussement, K., & De Bock, K. W. (2013). Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. *Journal of Business Research*, *66*(9), 1629-1636. https://doi.org/10.1016/j.jbusres.2012.12.008

Daunis, L., & Iwan, E. (2014). *Companies Struggling To Win Customers For Life, Says New Study By Forbes Insights And Sitecore*. Forbes Insights. Retrieved May 3, 2020, from https://www.forbes.com/sites/forbespr/2014/09/10/companies-struggling-to-win-customers-for-life-says-new-study-by-forbes-insights-and-sitecore/

Delgosha, M. S., Hajiheydari, N., & Saadatmanesh, H. (2020). Semantic structures of business analytics research: applying text mining methods. *Information Research*, *25*(2).

Devriendt, F., Van Belle, J., Guns, T., & Verbeke, W. (2022). Learning to Rank for Uplift Modeling. *IEEE Transactions on Knowledge and Data Engineering*, *34*(10), 4888-4904. https://doi.org/10.1109/TKDE.2020.3048510

Ding, J., Gao, D., & Chen, X. (2015). Alone in the game: Dynamic Spread of Churn Behavior in a Large Social Network a Longitudinal Study in MMORPG. *International Journal of Smart Home*, *9*(3), 35-44. https://doi.org/10.14257/ijsh.2015.9.3.04

Dodson, J. A., Tybout, A. M., & Sternthal, B. (1978). Impact of Deals and Deal Retraction on Brand Switching. *Journal of Marketing Research*, *15*(1). https://doi.org/10.2307/3150402

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. https://doi.org/10.48550/arxiv.1810.11363

Etaati, L. (2019). Azure Databricks. In *Machine Learning with Microsoft Technologies* (pp. 159-171). Apress. https://doi.org/10.1007/978-1-4842-3658-1_10

Fader, P. (2012). *Customer Centricity Focus on the Right Customers for Strategic Advantage*. Wharton School Press. https://doi.org/10.2307/j.ctv2hdrfj0

Ferlitsch, A. (2021). *Deep Learning Patterns and Practices*. Manning.

Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M., & Hutter, F. (2020). Auto-Sklearn 2.0: Hands-free AutoML via Meta-Learning. https://doi.org/10.48550/arxiv.2007.04074

Feurer, M., & Hutter, F. (2019). Hyperparameter Optimization. In F. Hutter, L. Kotthoff, & J. Vanschoren (Eds.), *Automated Machine Learning* (pp. 3-33). Springer International Publishing. https://doi.org/10.1007/978-3-030-05318-5_1

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., & Hutter, F. (2015). Efficient and Robust Automated Machine Learning. In *28th Conference on Neural Information Processing Systems*. Neural Information Processing Systems. https://papers.nips.cc/paper_files/paper/2016/file/5680522b8e2bb01943234bce7bf84534-Paper.pdf https://proceedings.neurips.cc/paper_files/paper/2015/file/11d0e6287202fced83f79975ec59a3a6-Paper.pdf

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of machine learning research*, *20*.

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, *55*(1), 119-139. https://doi.org/10.1006/jcss.1997.1504

Fridrich, M. (2019). Explanatory variable selection with balanced clustering in customer churn prediction. *Ad Alta: Journal of Interdisciplinary Research*, *9*(1), 56-66. http://www.magnanimitas.cz/ADALTA/0901/papers/A_fridrich.pdf

Fridrich, M. (2020). Understanding Customer Churn Prediction Research with Structural Topic Models. *ECONOMIC COMPUTATION AND ECONOMIC CYBERNETICS STUDIES AND RESEARCH*, *54*(4/2020), 301-317. https://doi.org/10.24818/18423264/54.4.20.19

Gattermann-Itschert, T., & Thonemann, U. W. (2021). How training on multiple time slices improves performance in churn prediction. *European Journal of Operational Research*, *295*(2), 664-674. https://doi.org/10.1016/j.ejor.2021.05.035

Geeter, D. (2018). Here's what it's like to shop at Toys R Us for the last time. *CNBC*. https://www.cnbc.com/2018/06/29/toys-r-us-closing-sales-stores-retail-toys.html

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems* (2nd). O'Reilly.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Journal of machine learning research*, *9*, 249-256.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning: Adaptive Computation and Machine Learning series*. MIT Press. https://books.google.cz/books?id=Np9SDQAAQBAJ

Gordini, N., & Veglio, V. (2017). Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. *Industrial Marketing Management*, *62*, 100-107. https://doi.org/10.1016/j.indmarman.2016.08.003

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(1), 5228-5235. https://doi.org/10.1073/pnas.0307752101

Gronwald, K. D. (2017). *Integrated Business Information Systems: A Holistic View of the Linked Business Process Chain ERP-SCM-CRM-BI-Big Data: A Holistic View of the Linked Business Process Chain ERP-SCM-CRM-BI-Big Data*. Springer Berlin Heidelberg. https://books.google.cz/books?id=mSYmDwAAQBAJ

Gupta, S., Lehmann, D. R., & Stuart, J. A. (2004). Valuing Customers. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.459595

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, *77*(1), 103-123. https://doi.org/10.1007/s10994-009-5119-5

Handley, L. (2013). *Customer retention: brave new world of consumer dynamics*. Marketing Week Online Ed 21. Retrieved December 8, 2019, from https://www.marketingweek.com/customer-retention-brave-new-world-of-consumer-dynamics/

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd). Springer.

Hengliang, W., & Weiwei, Z. (2012). A Customer Churn Analysis Model in E-business Environment. *International Journal of Digital Content Technology and its Applications*, *6*(9), 296-302. https://doi.org/10.4156/jdcta.vol6.issue9.37

Hinton, G., Srivastava, N., & Swersky, K. (2018). *Neural Networks for Machine Learning: Lecture 6a Overview of mini-batch gradient descent*. https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf

Hockenmaier, J. (2020). *CS447: Natural Language Processing*. University of Illinois. Retrieved April 17, 2023, from https://courses.engr.illinois.edu/cs447/fa2020/

Hodson, H. (2019). *DeepMind and Google: the battle to control artificial intelligence*. The Economist. Retrieved April 16, 2023, from https://www.economist.com/1843/2019/03/01/deepmind-and-google-the-battle-to-control-artificial-intelligence

Holčík, J., & Komenda, M. (2015). *Matematická biologie: e-learningová učebnice [online].: e-learningová učebnice [online]*. Masarykova univerzita. http://portal.matematickabiologie.cz/

Honnibal, M., & Montani, I. (2017). *SpaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd). OTexts.

Chaffey, D. (2015). *E-business and e-commerce management: strategy, implementation and practice* (6th). FT Prentice Hall.

Chapman, P., Clinton, J., KERBER, R., KHABAZA, T., REINARTZ, T., SHEARER, C., & WIRTH, R. (2000). *CRISP-DM 1.0 Step-by-step data mining guide*.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *ArXiv.org*. https://doi.org/10.1145/2939672.2939785

Chollet, F., & et al. (2015). *Keras*. Retrieved May 13, 2023, from https://github.com/fchollet/keras

Chou, Y. -C., & Chuang, H. H. -C. (2018). A predictive investigation of first-time customer retention in online reservation services. *Service Business*, *12*(4), 685-699. https://doi.org/10.1007/s11628-018-0371-z

Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. https://doi.org/10.48550/arxiv.1502.03167

Jain, H., Khunteta, A., & Srivastava, S. (2021). Telecom churn prediction and used techniques, datasets and performance measures: a review. *Telecommunication Systems*, *76*(4), 613-630. https://doi.org/10.1007/s11235-020-00727-0

Kay, A. (2007). Tesseract: An Open-Source Optical Character Recognition Engine: An Open-Source Optical Character Recognition Engine. *Linux J*, *2007*(159), 2.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. -Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *31st Conference on Neural Information Processing Systems*. Neural Information Processing Systems.

Kim, B., Khanna, R., & Koyejo, O. (2016). Examples are not Enough, Learn to Criticize!: Criticism for Interpretability. In *30th Conference on Neural Information Processing Systems*. Neural Information Processing Systems. https://papers.nips.cc/paper_files/paper/2016/file/5680522b8e2bb01943234bce7bf84534-Paper.pdf

Kim, S., Shin, K. -shik, & Park, K. (2005). An Application of Support Vector Machines for Customer Churn Analysis: Credit Card Case. *Advances in Natural Computation*, 636-647. https://doi.org/10.1007/11539117_91

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. https://doi.org/10.48550/arxiv.1412.6980

Koshiyama, A., Firoozye, N., & Treleaven, P. (2020). Algorithms in future capital markets. In *Proceedings of the First ACM International Conference on AI in Finance* (pp. 1-8). ACM. https://doi.org/10.1145/3383455.3422539

Kumar, A., Adlakaha, A., & Mukherjee, K. (2018). The effect of perceived security and grievance redressal on continuance intention to use M-wallets in a developing country.

International Journal of Bank Marketing, *36*(7), 1170-1189. https://doi.org/10.1108/IJBM-04-2017-0077

Kumar, R. (2019). *Research Methodology* (4 ed.). SAGE Publications.

Kumar, V., & Reinartz, W. (2018). *Customer Relationship Management: Concept, Strategy, and Tools: Springer Texts in Business and Economics* (3rd). Springer. https://books.google.cz/books?id=wBLYtNotoE0C

Kumar, V., & Reinartz, W. (2016). Creating Enduring Customer Value. *Journal of Marketing*, *80*(6), 36-68. https://doi.org/10.1509/jm.15.0414

Kumar, V., Aksoy, L., Donkers, B., Venkatesan, R., Wiesel, T., & Tillmanns, S. (2010). Undervalued or Overvalued Customers: Capturing Total Customer Engagement Value: Capturing Total Customer Engagement Value. *Journal of Service Research*, *13*(3), 297-310. https://doi.org/10.1177/1094670510375602

Kumar, V., Leone, R. P., Aaker, D. A., & Day, G. S. (2018). *Marketing Research* (13th). Wiley. https://books.google.cz/books?id=c-dKuAEACAAJ

Lee, D., & Seung, H. S. (2000). Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, *13*.

Lee, E. -B., Kim, J., & Lee, S. -G. (2017). Predicting customer churn in mobile industry using data mining technology. *Industrial Management & Data Systems*, *117*(1), 90-109. https://doi.org/10.1108/IMDS-12-2015-0509

Lee, E., Kim, B., Kang, S., Kang, B., Jang, Y., & Kim, H. K. (2020). Profit Optimizing Churn Prediction for Long-Term Loyal Customers in Online Games. *IEEE Transactions on Games*, *12*(1), 41-53. https://doi.org/10.1109/TG.2018.2871215

Lee, L. (2001). *On the effectiveness of the skew divergence for statistical language analysis* (R3). PMLR. https://proceedings.mlr.press/r3/lee01a.html

Liu, D. -ren, & Shih, Y. -yueh. (2005). Integrating AHP and data mining for product recommendation based on customer lifetime value. *Information & Management*, *42*(3), 387-400. https://doi.org/10.1016/j.im.2004.01.008

Li, X. -S., Zhang, H. -L., Zhu, Z. -X., Xiang, Z. -B., Chen, Z. -X., & Shi, Y. (2013). An Intelligent Transformation Knowledge Mining Method Based On Extenics. *Journal of Internet Technology*, *14*(2).

Li, X., & Li, Z. (2019). A Hybrid Prediction Model for E-Commerce Customer Churn Based on Logistic Regression and Extreme Gradient Boosting Algorithm. *Ingénierie des systèmes d information*, *24*(5), 525-530. https://doi.org/10.18280/isi.240510

Llave Montiel, M. A., & López, F. (2020). Spatial models for online retail churn: Evidence from an online grocery delivery service in Madrid. *Papers in Regional Science*, *99*(6), 1643-1665. https://doi.org/10.1111/pirs.12552

Lundberg, S., & Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. In Advances in *Neural Information Processing Systems*. Curran Associates. https://doi.org/10.48550/arxiv.1705.07874

MacKenzie, I., Meyer, C., & Noble, S. (2013). How retailers can keep up with consumers. *McKInsey & Company Insights*. https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers

Masís, S. (2021). *Interpretable Machine Learning with Python: Learn to Build Interpretable High-performance Models with Hands-on Real-world Examples: Learn to Build Interpretable High-performance Models with Hands-on Real-world Examples*. Packt Publishing. https://books.google.cz/books?id=eWQmzgEACAAJ

Mathai, N., Chen, Y., & Kirchmair, J. (2020). Validation strategies for target prediction methods. *Briefings in Bioinformatics*, *21*(3), 791-802. https://doi.org/10.1093/bib/bbz026

Metz, C. (2017). *An Improved AlphaGo Wins Its First Game Against the World's Top Go Player*. Wired. Retrieved April 16, 2023, from https://www.wired.com/2017/05/revamped-alphago-wins-first-game-chinese-go-grandmaster/

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1-38. https://doi.org/10.1016/j.artint.2018.07.007

Milošević, M., Živić, N., & Andjelković, I. (2017). Early churn prediction with personalized targeting in mobile social games. *Expert Systems with Applications*, *83*, 326-332. https://doi.org/10.1016/j.eswa.2017.04.056

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing Semantic Coherence in Topic Models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 262-272). Association for Computational Linguistics.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine*, *6*(7). https://doi.org/10.1371/journal.pmed.1000097

Molnar, C. (2022). *Interpretable Machine Learning* (2st). Lulu.

Molnár, Z. (2020). *Úvod do základů vědecké práce*. Fakulta Stavební - ČVUT. Retrieved February 1, 2020, from https://people.fsv.cvut.cz/~k126/predmety/d26mvp/mvp_sylabus-mvp.pdf

Molnár, Z. (2012). *Pokročilé metody vědecké práce*. Profess Consulting.

Morgan, B. (2018). How Amazon Has Reorganized Around Artificial Intelligence And Machine Learning. *Forbes*. https://www.forbes.com/sites/blakemorgan/2018/07/16/how-amazon-has-re-organized-around-artificial-intelligence-and-machine-learning/

Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., & Kaushansky, H. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on Neural Networks*, *11*(3), 690-696. https://doi.org/10.1109/72.846740

Neslin, S., Gupta, S., Kamakura, W., Lu, J., & Mason, C. (2006). Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. *Journal of Marketing Research (JMR)*, *43*(2), 204-211. http://web.a.ebsco-host.com.ezproxy.lib.vutbr.cz/ehost/detail/detail?sid=d9c07cd1-3964-40ea-9802-538946b2ec64%40session-mgr4010&vid=0&hid=4106&bdata=Jmxhbmc9Y3Mmc2l0ZT1laG9zdC1saXZl#AN=20949381&db=bth

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36*(2), 2592-2602. https://doi.org/10.1016/j.eswa.2008.02.021

Olson, R. S., & Moore, J. H. (2019). TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning. *Automated Machine Learning*, 151-160. https://doi.org/10.1007/978-3-030-05318-5_8

Ovide, S. (2011). Bookstore Chain Borders Is Dead. *The Wallstreet Journal*. https://blogs.wsj.com/deals/2011/07/18/its-almost-official-borders-is-dead/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825-2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

Perisic, A., & Pahor, M. RFM-LIR feature framework for churn prediction in the mobile games market. *IEEE Transactions on Games*, 1-1. https://doi.org/10.1109/TG.2021.3067114

Perrott, B. (2005). Towards a manager's model for e-business strategy decisions. *Journal of General Management*, *30*(4), 73-90. https://doi.org/10.1177/030630700503000405

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, *10*(3), 61-74.

Poornappriya, T. S., & Durairaj, M. (2019). High relevancy low redundancy vague set based feature selection method for telecom dataset. *Journal of Intelligent & Fuzzy Systems*, *37*(5), 6743-6760. https://doi.org/10.3233/JIFS-190242

Raff, E., & Borne, K. (2022). *Inside deep learning: math, algorithms, models*. Manning.

Rachid, A. D., Abdellah, A., Belaid, B., & Rachid, L. (2018). Clustering Prediction Techniques in Defining and Predicting Customers Defection: The Case of E-Commerce Context. *International Journal of Electrical and Computer Engineering (IJECE)*, *8*(4), 2367-2383. https://doi.org/10.11591/ijece.v8i4.pp2367-2383

Raschka, S., Liu, Y., & Mirjalili, V. (2022). *Machine learning with PyTorch and scikit-learn: Develop machine learning and deep learning models with Python*. Packt.

Reichheld, F. F., & Dawkings, P. M. (1990). Customer Retention as a Competitive Weapon. *Directors Broads*, *14*(1), 42-47.

Reichheld, F. F., & Sasser, J. R. (1990). Zero defections: quality comes to services. *Harvard business review*, *68*(5), 105-105. https://hbr.org/1990/09/zero-defections-quality-comes-to-services

Reinartz, W., Thomas, J. S., & Kumar, V. (2005). Balancing Acquisition and Retention Resources to Maximize Customer Profitability. *Journal of marketing*, *69*(1), 63-79. https://doi.org/10.1509/jmkg.69.1.63.55511

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-Agnostic Interpretability of Machine Learning. https://doi.org/10.48550/arxiv.1606.05386

Ricci, F., Rokach, L., Shapira, B., & Kantor, P. B. (Eds.). (2011). *Recommender Systems Handbook*. Springer US. https://doi.org/10.1007/978-0-387-85820-3

Roberts, M. e, Stewart, B. m, & Airoldi, E. m. (2016). A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association*, *111*(515), 988-1003. https://doi.org/10.1080/01621459.2016.1141684

Roberts, M., Steward, B., & Tingley, D. (2019). Stm: An R Package for Structural Topic Models. *Journal of Statistical Software*, *91*(2), 1-40. https://doi.org/10.18637/jss.v091.i02

Roberts, M., Stewart, B., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, *58*(4), 1064-1082. https://doi.org/10.1111/ajps.12103

Rothmeier, K., Pflanzl, N., Hullmann, J. A., & Preuss, M. (2021). Prediction of Player Churn and Disengagement Based on User Activity Data of a Freemium Online Strategy Game. *IEEE Transactions on Games*, *13*(1), 78-88. https://doi.org/10.1109/TG.2020.2992282

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

Russell, S. J., & Norvig, P. (2022). *Artificial intelligence: A modern approach* (4th). Pearson.

Sammut, C., & Webb, G. I. (2017). *Encyclopedia of Machine Learning and Data Mining* (2nd). Springer.

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*(3), 210-229. https://doi.org/10.1147/rd.33.0210

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & de Freitas, N. (2016). Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE*, *104*(1), 148-175. https://doi.org/10.1109/JPROC.2015.2494218

Shapley, L. S. (1953). *A value for n-person games*. Contributions to the Theory of Games.

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526-534. https://doi.org/10.1016/j.procs.2021.01.199

Singh, N., Singh, P., & Gupta, M. (2020). An inclusive survey on machine learning for CRM: A paradigm shift. *Decision*, *47*(4), 447-457. https://doi.org/10.1007/s40622-020-00261-7

Solomon, M. (2015). *The Year Of The Millennial Customer: Is Your Customer Experience Ready?*. Forbes. Retrieved April 15, 2023, from https://www.forbes.com/sites/micahsolomon/2015/11/14/2016-is-the-year-of-the-millennial-customer-heres-how-to-be-ready/#2784a68f6e72

Song, H. S., Kim, J. K., Cho, Y. B., & Kim, S. H. (2004). A Personalized Defection Detection and Prevention Procedure based on the Self-Organizing Map and Association Rule Mining: Applied to Online Game Site. *Artificial Intelligence Review*, *21*(2), 161-184. https://doi.org/10.1023/B:AIRE.0000021067.66616.b0

Stahl, F., Heitmann, M., Lehmann, D. R., & Neslin, S. A. (2012). The Impact of Brand Equity on Customer Acquisition, Retention, and Profit Margin. *Journal of marketing*, *76*(4), 44-63. https://doi.org/10.1509/jm.10.0522

Tamaddoni Jahromi, A., Stakhovych, S., & Ewing, M. (2014). Managing B2B customer churn, retention and profitability. *Industrial Marketing Management*, *43*(7), 1258-1268. https://doi.org/10.1016/j.indmarman.2014.06.016

Teixeira, T. (2014). *The Rising Cost of Consumer Attention: Why You Should Care, and What You Can Do about It* [Working paper]. Harvard Business School.

Terdiman, D. (2018). How AI is helping Amazon become a trillion-dollar company. *Fast company*. https://www.fastcompany.com/90246028/how-ai-is-helping-amazon-become-a-trillion-dollar-company

Thampi, A. (2022). *Interpretable AI: Building explainable machine learning systems*. Manning Publications Co.

Tsai, C. -F., & Chen, M. -Y. (2010). Variable selection by association rules for customer churn prediction of multimedia on demand. *Expert Systems with Applications*, *37*(3), 2006-2015. https://doi.org/10.1016/j.eswa.2009.06.076

Umashankar, N., Bhagwat, Y., & Kumar, V. (2017). Do loyal customers really pay more for services?. *Journal of the Academy of Marketing Science*, *45*(6), 807-826. https://doi.org/10.1007/s11747-016-0491-8

Valinsky, J. (2019). Macy's is closing 28 stores and a Bloomingdale's store. *CNN Business*. https://amp.cnn.com/cnn/2020/01/08/business/macys-store-closures/index.html

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, *7*(1), 91-91. https://doi.org/10.1186/1471-2105-7-91

Venkatesh, S., & Jeyakarthic, D. M. (2020). Adagrad Optimizer with Elephant Herding Optimization based Hyper Parameter Tuned Bidirectional LSTM for Customer Churn Prediction in IoT Enabled Cloud Environment. *Webology*, *17*(2), 631-651. https://doi.org/10.14704/WEB/V17I2/WEB17057

Wang, C., & Blei, D. M. (2011). *Collaborative Topic Modeling for Recommending Scientific Articles*. Association for Computing Machinery. https://doi.org/10.1145/2020408.2020480

Wang, C., Han, D., Fan, W., & Liu, Q. (2019). Customer Churn Prediction with Feature Embedded Convolutional Neural Network: An Empirical Study in the Internet Funds Industry.

International Journal of Computational Intelligence and Applications, *18*(01). https://doi.org/10.1142/S1469026819500032

Wang, S., Chaovalitwongse, W., & Babuska, R. (2012). Machine Learning Algorithms in Bipedal Robot Control. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(5), 728-743. https://doi.org/10.1109/TSMCC.2012.2186565

Williams, C., & Seeger, M. (2000). Using the nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13* (pp. 682-688). MIT Press. https://infoscience.epfl.ch/record/161322?ln=en

Yeo, I. K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*(4), 954-959. https://doi.org/10.1093/biomet/87.4.954

Yu, J., Seo, J., & Hyun, S. S. (2021). Perceived hygiene attributes in the hotel industry: customer retention amid the COVID-19 crisis. *International Journal of Hospitality Management*, *93*. https://doi.org/10.1016/j.ijhm.2020.102768

Yu, X., Guo, S., Guo, J., & Huang, X. (2011). An extended support vector machine forecasting framework for customer churn in e-commerce. *Expert Systems with Applications*, *38*(3), 1425-1430. https://doi.org/10.1016/j.eswa.2010.07.049

Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 694-699). ACM. https://doi.org/10.1145/775047.775151

Zaharia, M., Xin, R., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M., Ghodsi, A., Gonzalez, J., Shenker, S., & Stoica, I. (2016). Apache Spark: A unified engine for big data processing. *Communications of the ACM*, *59*(11), 56-65. https://doi.org/10.1145/2934664

Zai, A., & Brown, B. (2020). *Deep Reinforcement Learning in Action*. Manning.

Zhu, B., Baesens, B., Backiel, A., & vanden Broucke, S. K. L. M. (2018). Benchmarking sampling techniques for imbalance learning in churn prediction. *Journal of the Operational Research Society*, *69*(1), 49-65. https://doi.org/10.1057/s41274-016-0176-1

Investopedia. (2020). *Markets today: Amazon.com, inc*. Investopedia. Retrieved April 15, 2020, from https://www.investopedia.com/markets/quote?tvwidgetsymbol=AMZN

Insider Intelligence. (2020). *Global Ecommerce 2020: Ecommerce Decelerates amid Global Retail Contraction but Remains a Bright Spot*. Insider Intelligence. Retrieved April 15, 2021, from https://www.insiderintelligence.com/content/global-ecommerce-2020#page-report

Amazon Inc. (2020). *Annual Report 2020*. Annual reports, proxies and shareholder letters. Retrieved April 15, 2023, from https://s2.q4cdn.com/299287126/files/doc_financials/2021/ar/Amazon-2020-Annual-Report.pdf

Microsoft. (2022). *Azure Databricks pricing*. Microsoft. https://azure.microsoft.com/en-us/pricing/details/databricks/

Retail Rocket. (2017). *Retailrocket recommender system dataset*. Kaggle.com. https://www.kaggle.com/datasets/retailrocket/ecommerce-dataset/versions/4

Cabinet Office. (1999). *E-commerce: A Performance and Innovation Unit report*. UK Cabinet Office. Retrieved January 22, 2020, from www.cabinet-office.gov.uk/ innovation/1999/ecommerce/ec.body.pdf

REES46. (2020). *ECommerce behavior data from multi category store*. Kaggle.com. https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store

Seznam tabulek

Seznam obrázků

Seznam zkratek

ACC	angl. Accuracy, ukazatel přesnosti klasifikátoru
ALS	angl. Alternating Least Squares, metoda pro rozklad nezáporných matic
ANN	angl. Artificial Neural Networks, umělé neuronové sítě
AraBERT	angl. Arabic Bidirectional Encoder Representations from Transformers, velký jazykový model
AUCROC	angl. Area Under the Curve – Receiver Operating Characteristics, ukazatel přesnosti klasifikátoru
AWS	angl. Amazon Web Services
Bi-GRU	angl. Bi-Directional Gated Recurrent Units, komponenta/architektura rekurentní neuronové sítě
BO	angl. Bayesian Optimization, optimalizační strategie
C4.5	algoritmus pro konstrukci rozhodovacího stromu
C5.0	algoritmus pro konstrukci rozhodovacího stromu
CART	angl. Classification And Regression Tree, algoritmus pro konstrukci rozhodovacího stromu
CHAID	angl. Chi-Square Automatic Interaction Detection, algoritmus pro konstrukci rozhodovacího stromu
CLV	angl. Customer Lifetime Value, celoživotní hodnota zákazníka
CNN	angl. Convolutional Neural Network, konvoluční neuronová síť
CRISP-DM	angl. Cross-Industry Standard Process for Data Mining, metodický rámec pro implementaci projektů založených na dobývání znalostí nebo strojovém učení
CRM	angl. Customer Relationship Management, řízení vztahů se zákazníky
CU	angl. Currency Units, peněžní jednotka
DT	angl. Decision Tree, rozhodovací strom
EHO	angl. Elephant Herd Optimization, optimalizační strategie
EM	angl. Expectation Maximization, optimalizační strategie
F1	angl. F1 Score, ukazatel přesnosti klasifikátoru
FNR	angl. False Negative Rate, poměr chybně klasifikovaných pozorování – negativní třída
FPR	angl. False Positive Rate, poměr chybně klasifikovaných pozorování – pozitivní třída
FREX	angl. Frequency and Exclusivity, ukazatel kvality modelu témat
GAM	angl. Generalized Additive Models, zobecněné aditivní modely
GBM	angl. Gradient Boosted Machines, metoda boosting
GLM	angl. Generalized Linear Models, zobecněné lineární modely
GPU	angl. Graphics Processing Unit, grafický procesor
KNN	angl. K-Nearest Neighbors, metoda nejbližších sousedů
LDA	angl. Latent Dirichlet Allocation, metoda modelování témat
LIME	angl. Local Interpretable Model-Agnostic Explanations
LR	angl. Logistic Regression, logistická regrese
LSTM	angl. Long-Short Term Memory Networks, architektura rekurentních neuronových sítí
MAE	angl. Mean Absolute Error, průměrná absolutní chyba
MDA	angl. Multiple Discriminant Analysis, metoda pro snížení počtu dimenzí
MLP	angl. Multi-Layer Perceptron, architektura dopředných neuronových sítí
MSE	angl. Mean Squared Error, průměrná kvadratická chyba

NB	angl. Naive Bayes, klasifikační metoda
NBD	angl. Negative Binomial Distribution, diskrétní rozdělení pravděpodobnosti
NLP	angl. Natural Language Processing, zpracování přirozeného jazyka
OGA	angl. Optimal Genetic Algorithm, optimalizační strategie
PCA	angl. Principal Component Analysis, metoda dekompozice matic
PDF	angl. Portable Document Format, formát souborů
PRE	angl. Precision, podíl správně identifikovaných pozitivních pozorování vzhledem ke všem pozitivním predikcím
PRISMA	angl. Preferred Reporting Items for Systematic Reviews and Meta-Analyses, metodický rámec pro zpracování meta-analýz
REC	angl. Recall, podíl správně identifikovaných pozitivních pozorování vzhledem ke všem skutečným pozitivním pozorováním
RF	angl. Random Forests, metoda náhodných lesů
RFM	angl. Recency-Frequency-Monetary, přístup k marketingové analýze zákaznické báze
SEM-PLS	angl. Structural Equation Modeling – Partial Least Squares
SHAP	angl. Shapley Additive Explanations, Shapleyho aditivní vysvětlení
SSD	angl. Solid State Drive, diskové úložiště bez pohyblivých částí
STM	angl. Structural Topic Model, strukturální model témat
SVM	angl. Support Vector Machines, metoda podpůrných vektorů
TDL	angl. Top Decile Lift, zlepšení přesnosti predikce klasifikátoru v horním decilu pravděpodobností příslušnosti k pozitivní třídě
TPR	angl. True Positive Rate, poměr správně klasifikovaných pozorování – pozitivní třída
UK	angl. United Kingdom, Spojené Království
USD	angl. US Dollar, Americký dolar
YOY	angl. Year-On-Year, meziroční srovnání

Seznam příloh

A Modelování témat

Tab. A1 Ukázka profilace témat a reprezentativních slov

topic label	prevalence	scoring strategy	tokens
survey studies	1.3%	proba	limit, store, system, group, subject
		frex	limit, permission, subject, store, green
		lift	submission, permission, green, limit, subject
		score	permission, green, store, document, participant
customer satisfaction and loyalty	2.2%	proba	satisfaction, customer, effect, relationship, intention
		frex	satisfaction, repurchase, trust, moderate, intention
		lift	satisfaction, faction, experienced, moderator, incident
		score	satisfaction, trust, intention, repurchase, perceive
data mining	4.0%	proba	datum, mining, business, data, information
		frex	mining, analytic, big, tool, data
		lift	analytic, warehouse, big, tool, automation
		score	mining, analytic, datum, warehouse, crm
attitudes and behavior	0.9%	proba	attitude, behavior, use, survey, intention
		frex	attitude, respondent, transportation, mode, survey
		lift	transportation, attitude, transport, mode, bus
		score	attitude, respondent, intention, transportation, travel
statistical methods	2.3%	proba	variable, model, factor, regression, analysis
		frex	regression, variable, correlation, logistic, factor
		lift	probable, push, bar, box, dissertation
		score	regression, variable, model, logistic, factor
economic performance	2.0%	proba	profit, customer, churn, lift, target
		frex	lift, campaign, profit, decile, target
		lift	decile, campaign, lift, fraction, intervention
		score	lift, decile, churn, fraction, campaign
clustering techniques	2.4%	proba	cluster, clustering, datum, algorithm, mean
		frex	cluster, clustering, outlier, distance, fuzzy
		lift	cluster, clustering, outlier, unsupervised, distance
		score	clustering, cluster, algorithm, outlier, object
datasets	2.0%	proba	set, datum, sample, test, number
		frex	set, sample, training, test, datum
		lift	disease, gene, set, convert, cancer
		score	set, disease, training, gene, sample
customer relationship management	2.7%	proba	marketing, relationship, trust, customer, management
		frex	relationship, buyer, trust, marketing, supplier
		lift	buyer, side, referral, dependency, facet
		score	trust, buyer, relational, marketing, relationship
employee engagement	1.6%	proba	brand, employee, behavior, community, commitment
		frex	employee, engagement, brand, community, equity
		lift	engagement, employee, bond, equity, worker
		score	brand, employee, equity, orientation, engagement
classification performance	2.7%	proba	prediction, performance, datum, churner, technique
		frex	churner, auc, prediction, performance, classifier
		lift	chumer, benchmarke, baesen, mue, comprehensible
		score	churner, classifier, auc, prediction, classification
structural equation modeling	2.8%	proba	loyalty, construct, structural, effect, analysis
		frex	loyalty, structural, validity, construct, variance
		lift	gfi, cfi, agfi, larcker, rmsea
		score	loyalty, hair, respondent, mediate, structural
comparing performance	2.9%	proba	model, base, type, performance, propose
		frex	model, type, base, modeling, build
		lift	model, sente, scoring, type, modeling
		score	model, performance, scoring, predictive, type
social network analysis	1.7%	proba	network, social, influence, relational, number
		frex	network, social, relational, entity, graph

topic label	prevalence	scoring strategy	tokens
		lift	maximal, entity, spread, collective, neighborhood
		score	network, relational, node, neighbor, social
sales forecasting	1.1%	proba	sale, salesperson, customer, medium, forecast
		frex	salesperson, sale, forecast, student, selling
		lift	salesperson, efficacy, selling, trait, student
		score	salesperson, sale, personality, selling, relational
rule-based learning	2.6%	proba	rule, tree, decision, mining, classification
		frex	rule, tree, mining, rough, classification
		lift	rule, induction, exhaustive, tree, dissatisfied
		score	tree, mining, rule, algorithm, node
service quality	2.7%	proba	service, quality, provider, influence, recovery
		frex	quality, service, recovery, complaint, delivery
		lift	servqual, empathy, tangible, assurance, parasuraman
		score	recovery, service, quality, servqual, parasuraman
survival modeling	2.0%	proba	time, risk, model, customer, datum
		frex	risk, hazard, survival, covariate, event
		lift	hazard, covariate, survival, mixture, median
		score	covariate, hazard, risk, posterior, survival
ensemble learning	2.8%	proba	classifier, ensemble, forest, boost, base
		frex	ensemble, boost, classifier, forest, bagging
		lift	ensemble, bagging, voting, boosting, boost
		score	ensemble, classifier, forest, svm, learner
decision modeling	1.2%	proba	decision, state, problem, agent, step
		frex	agent, situation, objective, state, problem
		lift	agent, awareness, team, situation, conversion
		score	agent, awareness, object, team, simulation
acquisition and retention value	1.9%	proba	firm, model, acquisition, retention, marketing
		frex	acquisition, diffusion, firm, innovation, acquire
		lift	diffusion, acquisition, valuation, prospect, budget
		score	diffusion, firm, equity, acquisition, adoption
probabilistic modeling	1.2%	proba	variable, bayesian, probability, state, model
		frex	bayesian, spatial, causal, probit, explanatory
		lift	spatial, probit, independence, bayesian, parent
		score	bayesian, spatial, probit, node, causal
natural language processing	1.2%	proba	topic, sentiment, word, text, document
		frex	sentiment, topic, text, document, word
		lift	sentiment, topic, formal, text, neutral
		score	sentiment, topic, document, node, friend
attributes	1.4%	proba	attribute, value, information, system, step
		frex	attribute, miss, contain, step, subset
		lift	attribute, attri, missing, miss, red
		score	attribute, dataset, subset, value, object
gaming industry	1.6%	proba	user, game, time, feature, day
		frex	game, session, player, user, day
		lift	game, session, video, player, daily
		score	game, session, user, player, video
tourism and hospitality	1.8%	proba	tourism, hotel, perceive, restaurant, intention
		frex	tourism, hotel, restaurant, tourist, fairness
		lift	tourist, tourism, hotel, destination, restaurant
		score	tourist, tourism, hotel, restaurant, hospitality
pricing strategy	1.3%	proba	price, usage, effect, renewal, contract
		frex	renewal, price, usage, pricing, fee
		lift	renewal, uncertain, renew, pricing, price
		score	renewal, price, renew, contract, uncertain
e-commerce industry	1.9%	proba	online, consumer, product, shopping, information
		frex	online, shopping, website, commerce, site
		lift	shopper, functionality, website, shopping, offline
		score	shopping, online, shopper, commerce, retailer
customer segmentation	1.8%	proba	customer, segmentation, segment, group, analysis
		frex	segmentation, segment, enterprise, crm, group

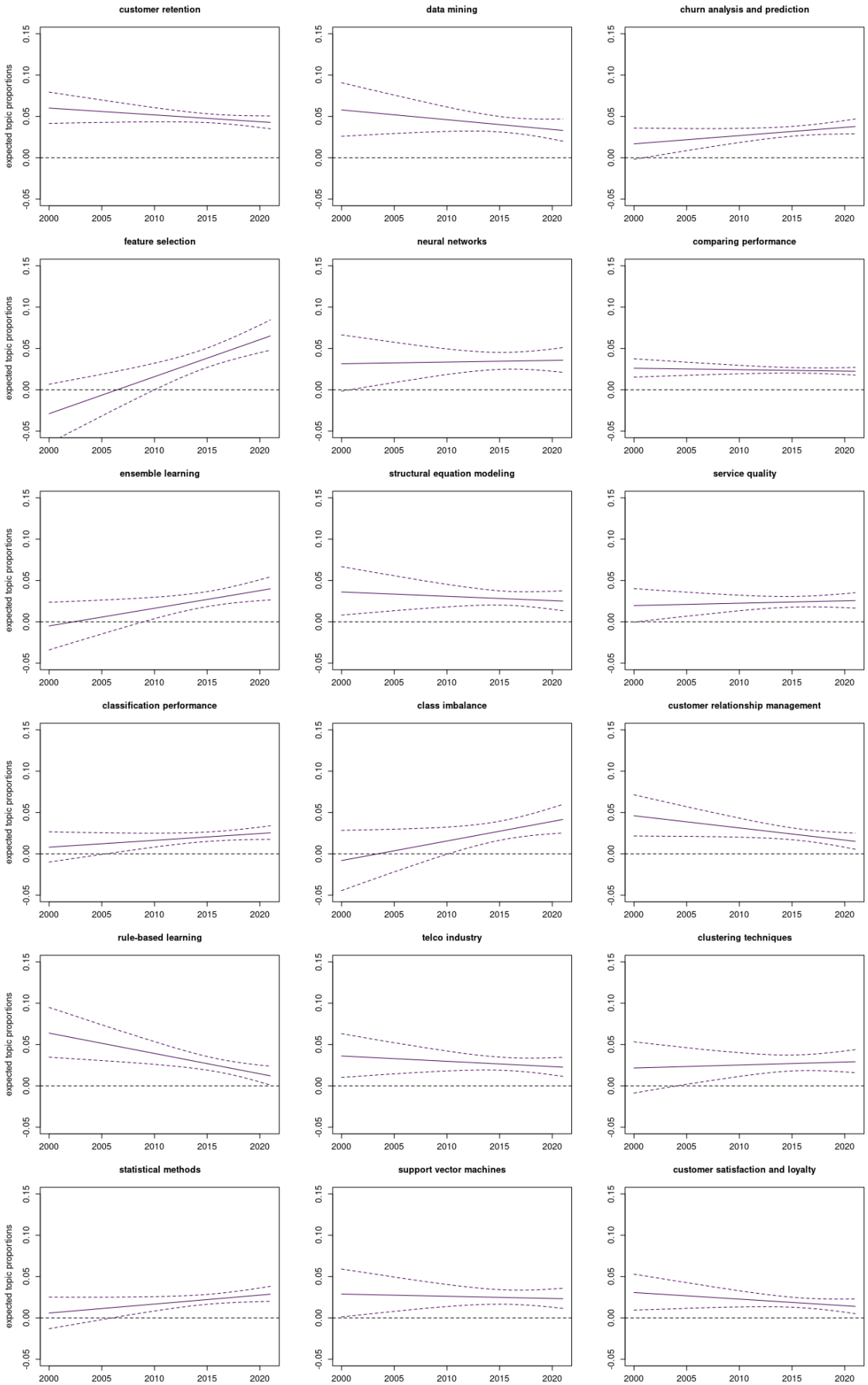
topic label	prevalence	scoring strategy	tokens
		Lift	segmentation, neighbour, verification, enterprise, segment
		Score	segmentation, enterprise, segment, neighbour, rfm
image and brand	1.2%	Proba	experience, image, corporate, market, consumer
		Frex	corporate, image, reputation, experience, energy
		Lift	corporate, reputation, responsibility, image, stakeholder
		Score	corporate, image, reputation, responsibility, energy
value creation	1.5%	Proba	value, creation, use, create, measure
		Frex	creation, value, create, resource, establish
		Lift	creation, value, absence, proposal, logic
		Score	value, creation, proposition, absence, resource
cloud infrastructure	1.4%	Proba	business, cost, system, cloud, profit
		Frex	cloud, recovery, infrastructure, scheme, open
		Lift	cloud, damage, infrastructure, upgrade, server
		Score	recovery, cloud, server, infrastructure, layer
strategic orientation	2.0%	Proba	performance, firm, management, market, business
		Frex	strategic, capability, firm, orientation, organizational
		Lift	manufacturing, strategic, capability, orientation, organizational
		Score	firm, orientation, organizational, performance, strategic
feature selection	3.6%	proba	churn, feature, dataset, prediction, customer
		frex	particle, recall, dataset, precision, swarm
		lift	particle, swarm, redundancy, night, recall
		score	churn, dataset, particle, classifier, algorithm
loyalty program	1.6%	proba	consumer, program, loyalty, brand, intention
		frex	program, food, consumer, app, reward
		lift	food, app, program, habit, reward
		score	intention, food, consumer, app, perceive
supply-chain	1.0%	proba	chain, demand, supply, system, product
		frex	supply, chain, demand, stock, sharing
		lift	sharing, supply, stock, chain, demand
		score	node, stock, respondent, supply, sharing
customer retention	5.8%	proba	customer, retention, company, relationship, service
		frex	customer, company, retention, retain, strategy
		lift	customer, company, retain, try, retention
		score	customer, company, retention, loyalty, retain
purchasing behavior	2.0%	proba	customer, purchase, state, model, period
		frex	defection, purchase, lifetime, category, recency
		lift	recency, fader, click, transition, defect
		score	purchase, recency, fader, shopping, contractual
mobile services	1.5%	proba	mobile, banking, service, user, use
		frex	banking, mobile, adoption, acceptance, internet
		lift	wallet, acceptance, banking, continuance, instant
		score	banking, continuance, mobile, trust, intention
support vector machines	2.3%	proba	feature, svm, vector, selection, machine
		frex	svm, feature, kernel, vector, selection
		lift	hyper, kernel, svm, static, gaussian
		score	svm, feature, vector, kernel, machine
neural networks	3.3%	proba	network, neural, layer, prediction, datum
		frex	layer, neural, output, neuron, input
		lift	neuron, som, mlp, layer, multilayer
		score	neural, layer, neuron, node, som
data and modeling	2.0%	proba	variable, client, company, product, regression
		frex	client, partial, coussement, length, expert
		lift	den, buckinx, client, fax, poel
		score	forest, client, coussement, regression, variable
telco industry	2.4%	proba	service, mobile, telecom, subscriber, operator
		frex	subscriber, operator, telecom, telecommunication, voice
		lift	subscriber, bill, handset, voice, billing
		score	subscriber, telecom, mobile, operator, contract
switching behavior	1.9%	proba	switch, cost, service, commitment, provider
		frex	switch, switching, commitment, cost, barrier

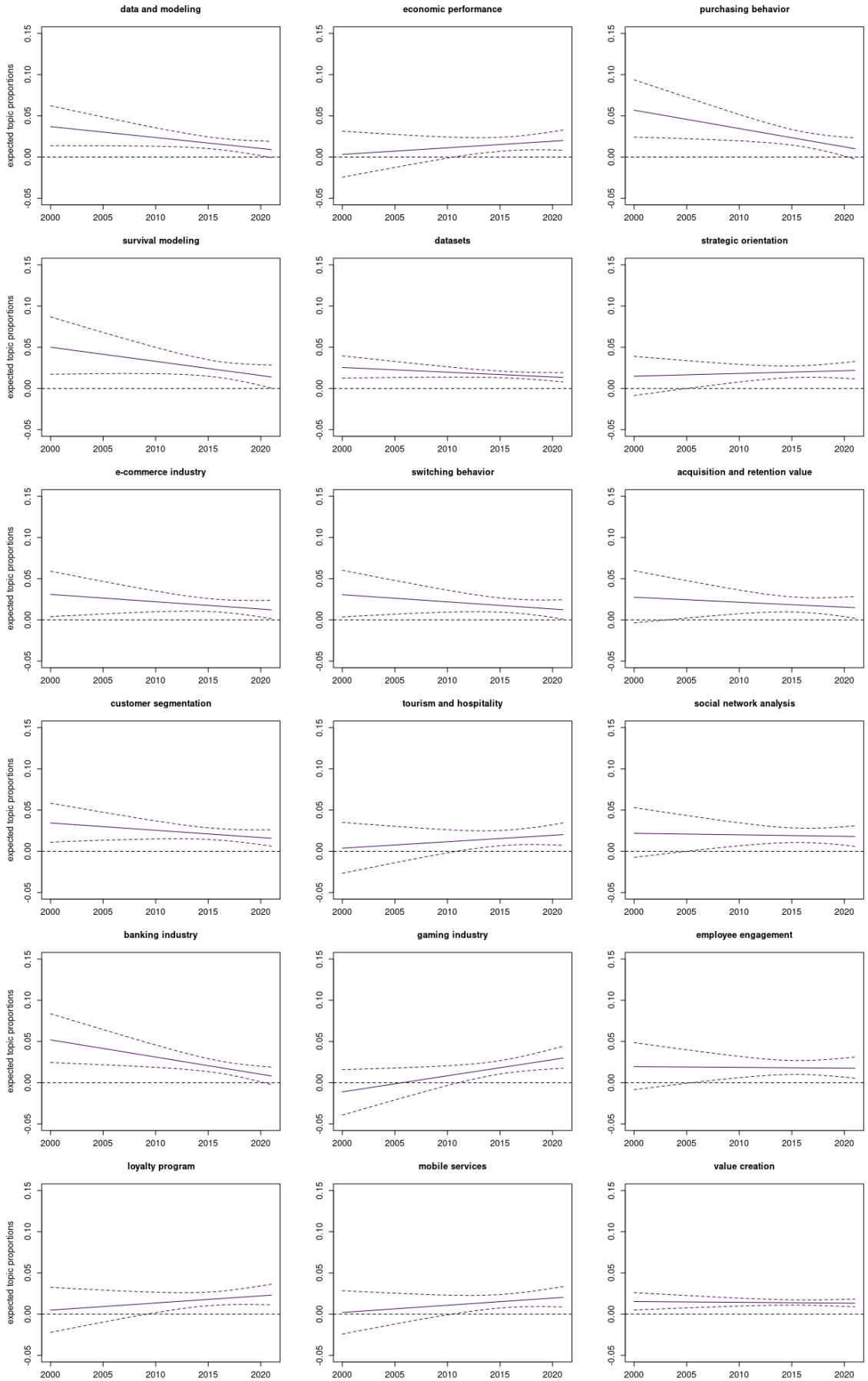
topic label	prevalence	scoring strategy	tokens
		lift	switching, lock, inertia, switch, barrier
		score	commitment, inertia, switching, switch, affective
class imbalance	2.7%	proba	class, datum, sample, method, imbalanced
		frex	imbalanced, minority, class, smite, imbalance
		lift	oversampling, smite, minority, oversample, imbalanced
		score	imbalanced, smite, minority, oversampling, class
churn analysis and prediction	3.8%	proba	churn, prediction, predict, churner, analysis
		frex	churn, prediction, churner, predict, expert
		lift	churn, chum, pany, syst, prediction
		score	churn, churner, tree, prediction, mining
banking industry	1.7%	proba	bank, financial, credit, banking, insurance
		frex	bank, insurance, credit, card, financial
		lift	loan, holder, bank, insurance, premium
		score	bank, credit, attrition, banking, card

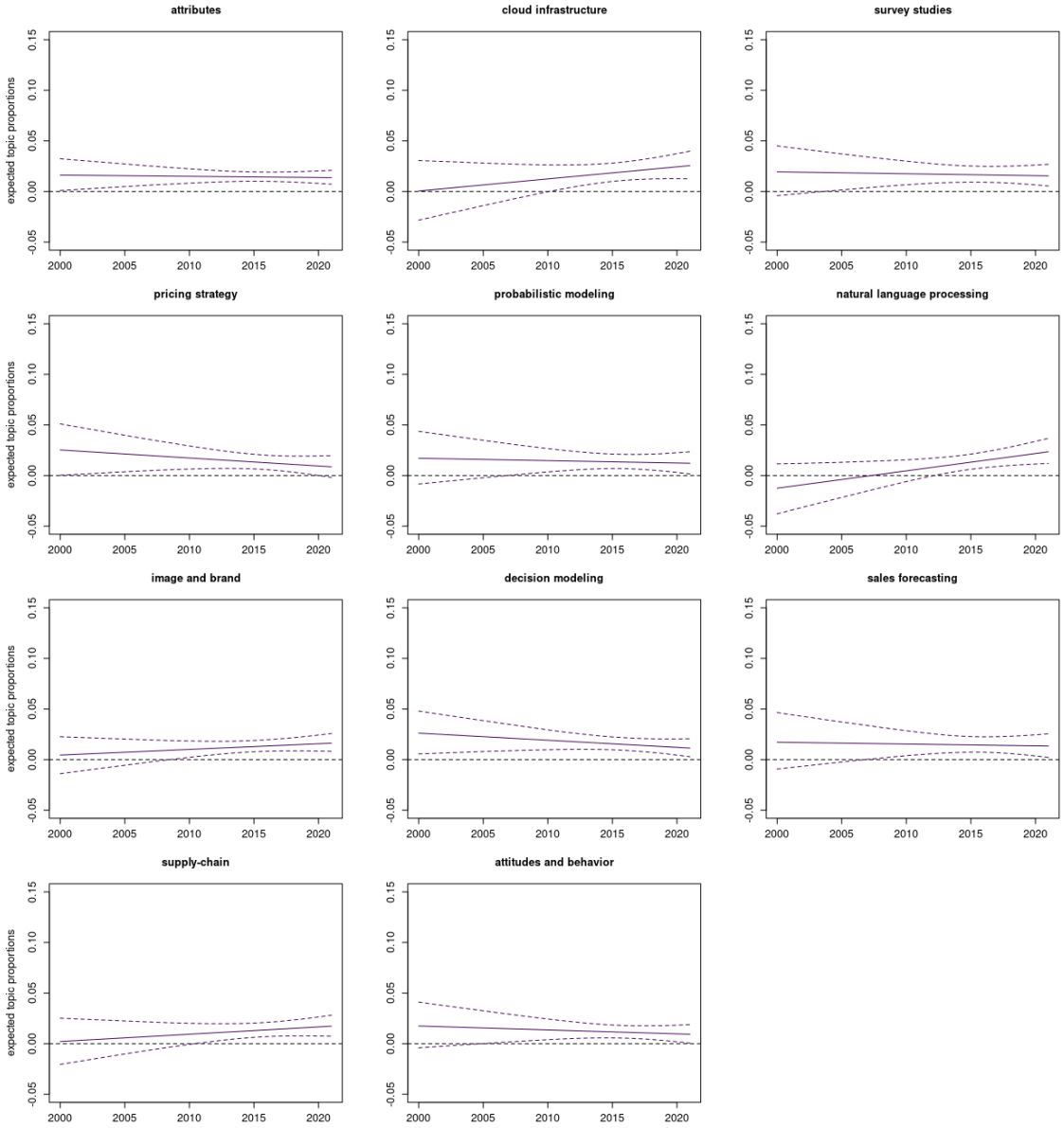
Tab. A2 Ukázka profilace témat a reprezentativních dokumentů

topic label	authors	title	year	avg cit. per year	preval.
survey studies	Kang JYM & Kim J	Online customer relationship marketing tactics through social media and perceived customer retention orientation of the green retailer	2017	3.58	68.0%
customer satisfaction and loyalty	Teichert T & Rost K	Trust, involvement profile and customer retention - Modelling, effects and implications	2003	1.17	69.8%
data mining	Ranjan J & Bhatnagar V	A holistic framework for mCRM - Data mining perspective	2009	1.57	70.5%
attitudes and behavior	Carrel AL & Li M	Survey-based measurement of transit customer loyalty: Evaluation of measures and systematic biases	2019	1.45	87.8%
statistical methods	Bhargava M et al.	Prediction model for telecom postpaid customer churn using Six-Sigma methodology	2017	0.21	49.9%
economic performance	Lemmens A & Gupta S	Managing Churn to Maximize Profits	2020	1.14	73.4%
clustering techniques	Sangeetha T & Mary GA	A Study on Different Methods of Outlier Detection Algorithms in Data Mining	2020	0.00	73.0%
datasets	Dong G et al.	Mining disease state converters for medical intervention of diseases	2010	0.26	49.3%
customer relationship management	Doney PM et al.	Trust determinants and outcomes in global B2B services	2007	9.42	56.8%
employee engagement	Cho SH & Johanson MM	Organizational citizenship behavior and employee performance: A moderating effect of work status in restaurant employees	2008	3.13	72.9%
classification performance	Ahmed AAQ & Maheswari D	An enhanced ensemble classifier for telecom churn prediction using cost based uplift modelling	2019	1.09	28.3%
structural equation modeling	Matzler K et al.	Switching experience, customer satisfaction, and switching costs in the ICT industry	2015	4.59	43.3%
comparing performance	Lee JS & Lee JC	Customer churn prediction by hybrid model	2006	0.25	35.3%
social network analysis	Fang X & Hu PJH	Top persuader prediction for social networks	2018	4.53	78.5%
sales forecasting	Yang B	How does personality link to customer retention? A sequential mediating model of self-efficacy and relational investment	2015	0.15	76.7%
rule-based learning	Li X et al.	An intelligent transformation knowledge mining method based on extenics	2013	2.29	51.0%
service quality	Mostafa R et al.	The CURE scale: a multidimensional measure of service recovery strategy	2014	2.71	47.7%
survival modeling	Galloway M et al.	Time-to-default analysis of mortgage portfolios	2017	0.21	78.1%
ensemble learning	Adhikary DD & Gupta D	Applying over 100 classifiers for churn prediction in telecom companies	2020	0.57	57.6%
decision modeling	Ma J et al.	Team situation awareness measure using semantic utility functions for supporting dynamic decision-making	2010	0.34	83.5%
acquisition and retention value	Fazıl Paç M et al.	When to adopt a service innovation: Nash equilibria in a competitive diffusion framework	2018	0.27	85.1%
probabilistic modeling	Cox LA	Data mining and causal modeling of customer behaviors	2002	0.35	64.7%
natural language processing	Ravi K et al.	Fuzzy formal concept analysis based opinion mining for CRM in financial services	2017	6.95	81.7%
attributes	Pinheiro P & Cavique L	Regular sports services: Dataset of demographic, frequency and service level agreement	2021	0.00	49.2%

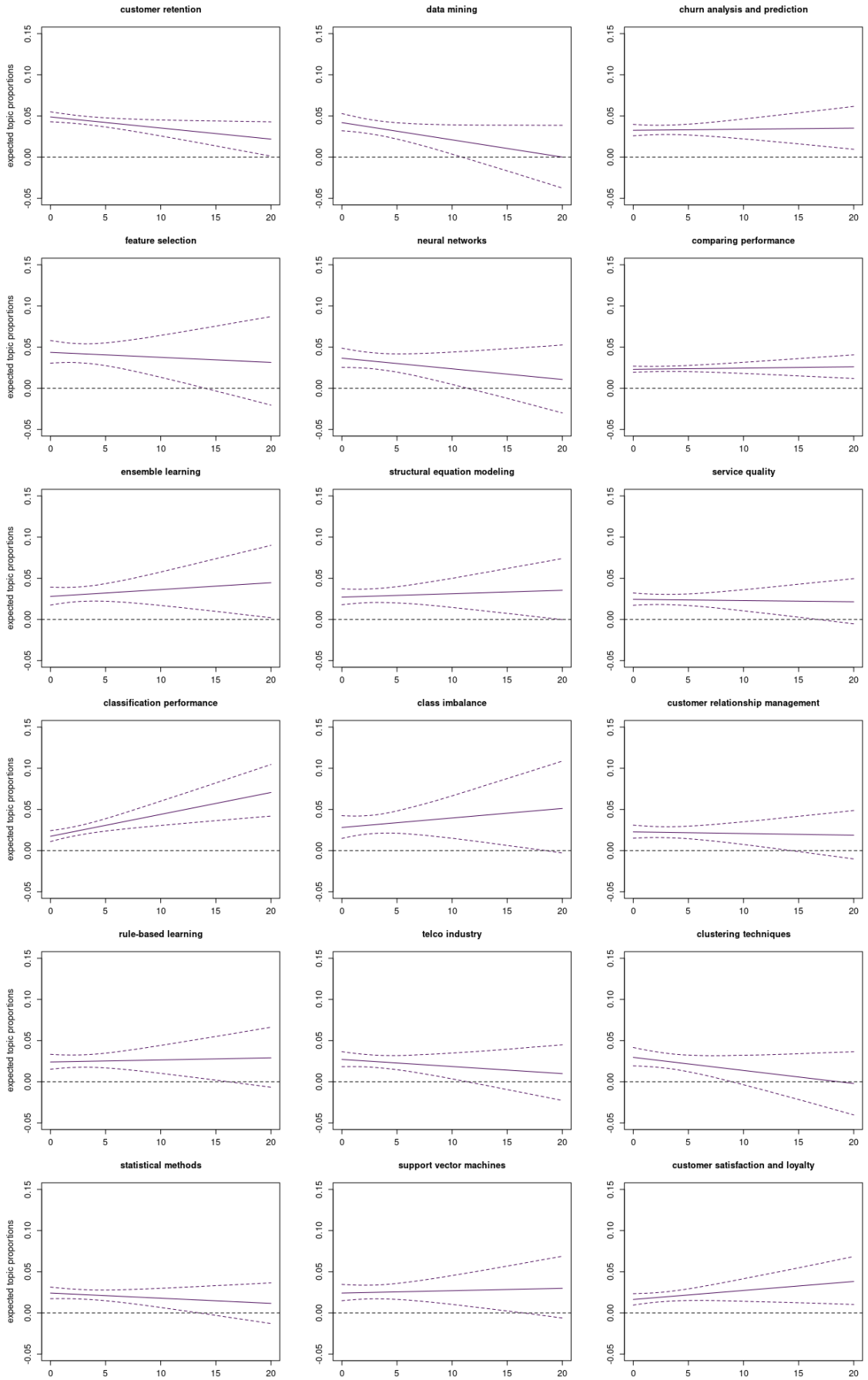
topic label	authors	title	year	avg cit. per year	preval.
gaming industry	Castro EG & Tsuzuki MSG	Churn Prediction in Online Games Using Players' Login Records: A Frequency Analysis Approach	2015	4.15	70.7%
tourism and hospitality	Han H et al.	Nature based solutions and customer retention strategy: Eliciting customer well-being experiences and self-rated mental health	2020	8.57	72.2%
pricing strategy	Shen LX et al.	The Fun and Function of Uncertainty: Uncertain Incentives Reinforce Repetition Decisions	2019	4.36	88.1%
e-commerce industry	Zhang T et al.	The value of it-enabled retailer learning: Personalized product recommendations and customer store loyalty in electronic markets	2011	11.35	61.8%
customer segmentation	Tanaka M & Kurahashi S	An analysis of customer retention rates by time series data mining	2015	0.44	68.7%
image and brand	Dudek D et al.	Changing energy supplier on the market with a strong position of incumbent suppliers—polish example	2021	0.00	36.8%
value creation	Dal Bo G et al.	Proposal and validation of a theoretical model of customer retention determinants in a service environment	2018	0.53	35.3%
cloud infrastructure	Malleswari M et al.	Comparative analysis of machine learning techniques to Identify churn for telecom data	2018	0.27	88.3%
strategic orientation	Shin H et al.	Strategic agility of Korean small and medium enterprises and its influence on operational and firm performance	2015	8.74	64.7%
feature selection	Poornappriya TS& Durairaj M	High relevancy low redundancy vague set based feature selection method for telecom dataset	2019	0.73	69.8%
loyalty program	Loh Z& Hassan SH	Consumers' attitudes, perceived risks and perceived benefits towards repurchase intention of food truck products	2021	0.00	75.8%
supply-chain	Esmaeili N et al.	A scenario-based optimization model for planning and redesigning the sale and after-sales services closed-loop supply chain	2021	0.00	92.6%
customer retention	Min H et al.	A data mining approach to developing the profiles of hotel customers	2002	2.58	36.5%
purchasing behavior	Park CHet al.	A multi-category customer base analysis	2014	2.06	88.8%
mobile services	Ghobakhloo M & Fathi M	Modeling the Success of Application-Based Mobile Banking	2019	1.09	75.1%
support vector machines	Huang B et al.	Multi-objective feature selection by using NSGA-II for customer churn prediction in telecommunications	2010	8.60	56.9%
neural networks	Adwan O et al.	Predicting customer churn in telecom industry using multilayer preceptron neural networks: Modeling and analysis	2014	3.48	57.9%
data and modeling	Larivière B & Van Den Poel D	Predicting customer retention and profitability by using random forests and regression forests techniques	2005	11.28	67.8%
telco industry	Chu BH et al.	Toward a hybrid data mining model for customer retention	2007	5.49	42.8%
switching behavior	Bansal HS et al.	A three-component model of customer commitment to service providers	2004	20.00	78.1%
class imbalance	Toor AA & Usman M	Adaptive telecom churn prediction for concept-sensitive imbalance data streams	2021	0.00	76.0%
churn analysis and prediction	Ahn J et al.	A Survey on Churn Analysis in Various Business Domains	2020	0.00	41.5%
banking industry	Prinzie A & Van den Poel D	Incorporating sequential information into traditional classification models by using an element/position-sensitive SAM	2006	2.48	50.3%

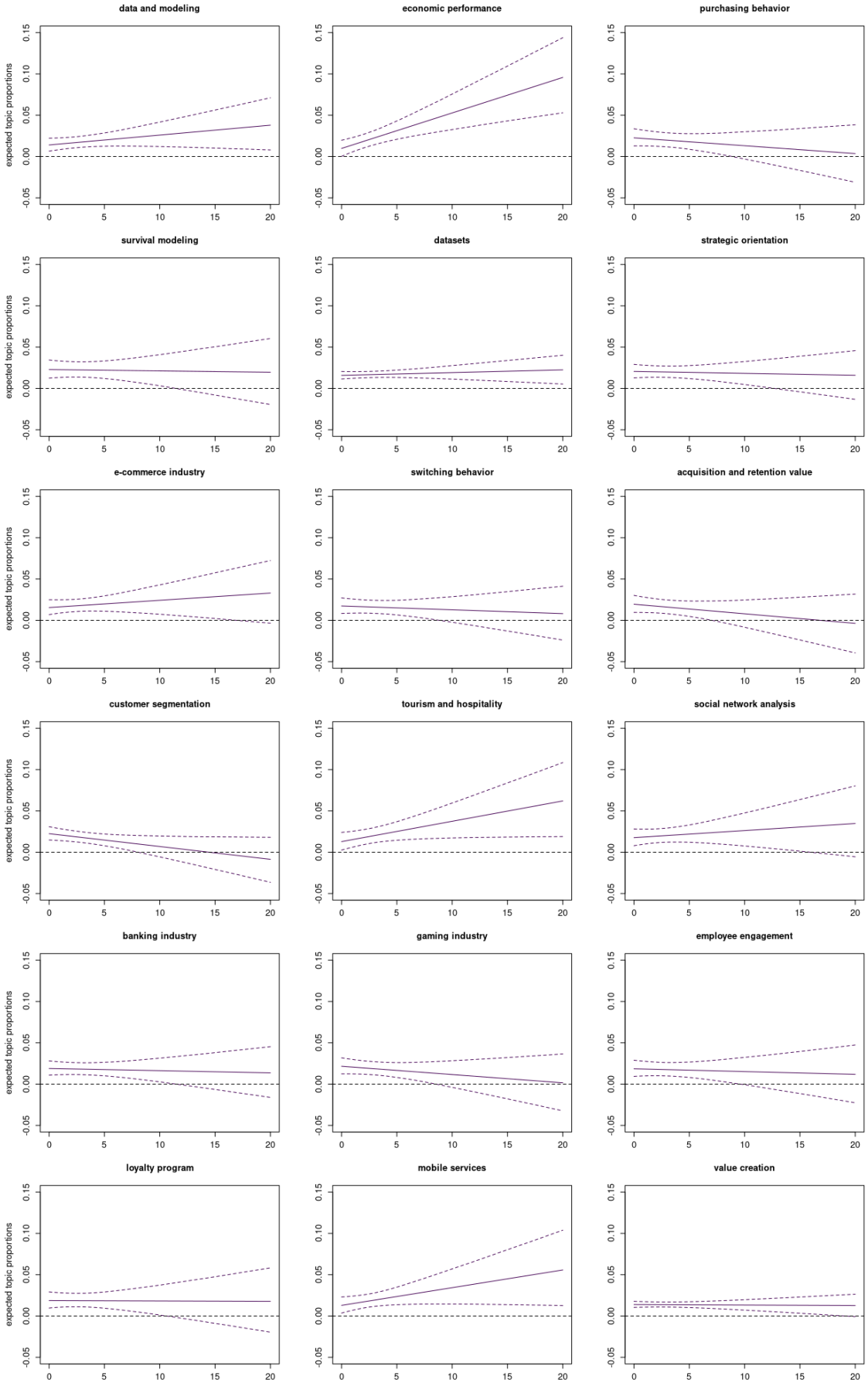

customer retention
data mining
churn analysis and prediction
feature selection
neural networks
comparing performance
ensemble learning
structural equation modeling
service quality
classification performance
class imbalance
customer relationship management
rule-based learning
telco industry
clustering techniques
statistical methods
support vector machines
customer satisfaction and loyalty
expected topic proportions
-0.05
0.00
0.05
0.10
0.15
2000
2005
2010
2015
2020

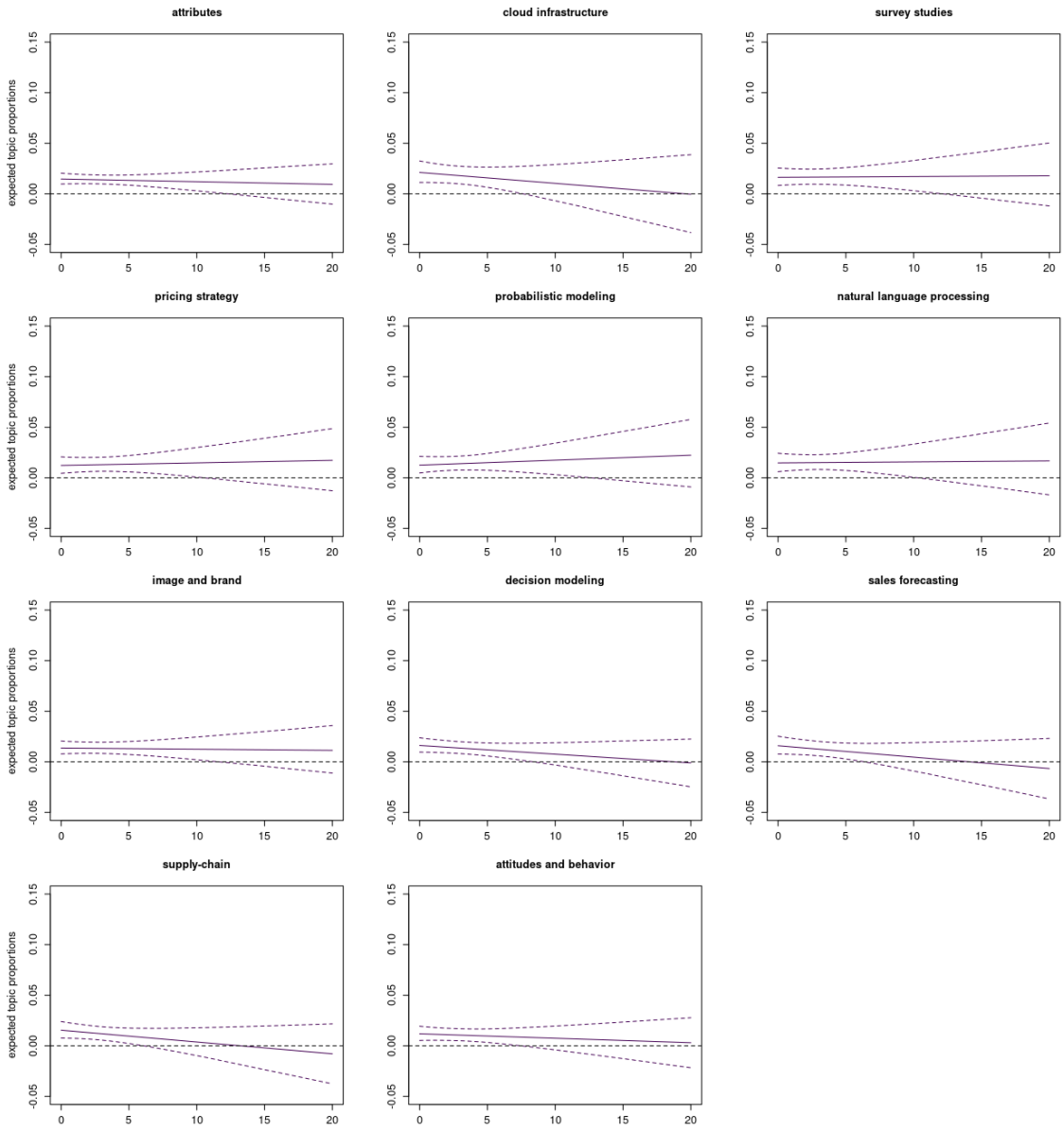

data and modeling
economic performance
purchasing behavior
survival modeling
datasets
strategic orientation
e-commerce industry
switching behavior
acquisition and retention value
customer segmentation
tourism and hospitality
social network analysis
banking industry
gaming industry
employee engagement
loyalty program
mobile services
value creation
expected topic proportions
0.15
0.10
0.05
0.00
-0.05
2000
2005
2010
2015
2020



Obr. A3 Změna očekávané míry výskytu témat v čase


customer retention
data mining
churn analysis and prediction
feature selection
neural networks
comparing performance
ensemble learning
structural equation modeling
service quality
classification performance
class imbalance
customer relationship management
rule-based learning
telco industry
clustering techniques
statistical methods
support vector machines
customer satisfaction and loyalty
expected topic proportions
-0.05
0.00
0.05
0.10
0.15
0
5
10
15
20


data and modeling
economic performance
purchasing behavior
survival modeling
datasets
strategic orientation
e-commerce industry
switching behavior
acquisition and retention value
customer segmentation
tourism and hospitality
social network analysis
banking industry
gaming industry
employee engagement
loyalty program
mobile services
value creation
expected topic proportions
-0.05
0.00
0.05
0.10
0.15
0
5
10
15
20



Obr. A4 Očekávaná míra výskytu témat jako funkce průměrného počtu citací za kalendářní rok

B Modelování odchodu zákazníka

Tab. B1 Datové soubory a programový kód

supplementary material	hyperlink
Retail Rocket – customer model	https://www.kaggle.com/datasets/fridrichmrtn/e-commerce-churn-dataset-retail-rocket
REES46 – customer model	https://www.kaggle.com/datasets/fridrichmrtn/e-commerce-churn-dataset-rees46
Code	https://github.com/fridrichmrtn/churn-modeling

Tab. B2 Vnější parametry systému strojového učení

pipeline step	parameter description	parameter value	type
data scaler	scaler object	power transformer	instance
		quantile transformer	instance
		robust scaler	instance
variance filter	threshold	<0.001, 0.1>, uniform	float
feature selector	number of features	<5, 50>, uniform	integer
data sampler	sampler object	random under-sampler	instance
		random over-sampler	instance
		passthrough	string
logistic regression	inverse regularization strength	<0.01, 10>, uniform	float
	L1 regularization weight, complementary to L2	<0, 1>, uniform	float
elastic net regression	inverse regularization strength	<0.01, 10>, uniform	float
	L1 regularization weight, complementary to L2	<0, 1>, uniform	float
support vector machine with linear kernel	inverse regularization strength	<0.01, 10>, uniform	float
support vector machine with approx. radial basis kernel	number of components for Nystroem projection	<10, 100>, uniform	integer
	inverse regularization strength	<0.01, 10>, uniform	float
multi-layer perceptron	batch size	<4, 16>, uniform	integer
	number of epochs	<50, 500>, uniform	integer
	number of hidden layers	<1, 5>, uniform	integer
	number of units in a hidden layer	<8, 64>, uniform	integer
	activation function	exponential linear unit	string
		leaky rectified linear unit	string
	learning rate	<10e-5, 10e-2>, uniform	float
	optimization algorithm	stochastic gradient boosting	string
		adam	string
		rmsprop	string
decision tree	maximum tree depth	<2, 25>, uniform	integer
	minimum samples in a split	<10, 50>, uniform	integer
	minimum samples in a leaf	<2, 50>, uniform	integer
	minimal impurity decrease	<0.05, 0.1>, uniform	float
	minimal weight fraction in a leaf	<0, 0.1>, uniform	float
random forest	number of trees	<100, 750>, uniform	integer
	ratio of sampled features	<0.2, 0.6>, uniform	float
	ratio of sampled observations	<0.2, 0.6>, uniform	float

pipeline step	parameter description	parameter value	type
	minimum samples in a split	<10, 200>, uniform	integer
	minimum samples in a leaf	<2, 50>, uniform	integer
	minimal impurity decrease	<0.05, 0.1>, uniform	float
	minimal weight fraction in a leaf	<0, 0.1>, uniform	float
	number of boosting iterations	<100, 750>, uniform	integer
	learning rate	<0.01, 0.15>, uniform	float
	number of leaves	<300, 625>, uniform	integer
	maximum tree depth	<5, 25>, uniform	integer
gradient boosting machine	minimum samples in a split	<5, 50>, uniform	integer
	sampling frequency	<1, 5>, uniform	integer
	ratio of sampled features	<0.2, 0.6>, uniform	float
	ratio of sampled observations	<0.2, 0.6>, uniform	float
	L2 regularization on tree weights	<0.01, 100>, uniform	float

C Životopis autora

Martin Fridrich

fridrichmartin@yahoo.com | in

Vybrané pracovní zkušenosti

Období	2023 až dosud
Společnost	TD Synnex
Pozice	Senior Manager – Data Science

Období	2015 až dosud
Společnost	Martin Fridrich
Pozice	Independent Researcher
Náplň	Výzkum a aplikace strojového učení v různých podnikových kontextech, realizace přednášek a workshopů zaměřených na datovou vědu, strojové učení a řízení datových projektů.

Období	2017–2021
Společnost	Alza.cz
Pozice	Head of Data Science and Analytics
Náplň	Vedení týmů datové vědy a business intelligence, odpovědnost za spoluvytváření strategie datových inovací, jejich realizaci, metodiku projektů, architekturu řešení, a návratnost investic.

Období	2014, 2015–2016
Společnost	DSV Road
Pozice	Head of Tender Management and Analytics
Náplň	Vedení analytického týmu, odpovědnost za cenovou politiku, proces zpracování výběrových řízení, návrh a hodnocení dopravních řešení, interní školení a reporting.

Vzdělání

Období	2016 až dosud
Stupeň	Doktorské studium (Ph.D.)
Program	Ekonomika a management
Instituce	Vysoké učení technické v Brně

Období	2013–2015
Stupeň	Magisterské studium (M.Sc.), dokončeno s vyznamenáním
Program	Business and Informatics
Instituce	Nottingham Trent University, Vysoké učení technické v Brně

Období	2006–2011
Stupeň	Magisterské a bakalářské studium (Ing., Bc.)
Program	Dopravní inženýrství a spoje
Instituce	Univerzita Pardubice

Další aktivity

Období	2019 až dosud
Časopisy	User Modeling and User-Adapted Interaction, International Journal of Engineering Business Management, Proceedings of Digital Transformation of Corporate Business
Role	Recenzent příspěvků zaměřených na umělou inteligenci, strojové učení, velká data, případně modelování zákaznického chování.

Období	2017–2018
Instituce	CHEDTEB, FH Bielefeld, Vysoké učení technické v Brně
Role	Vedení workshopů a prezentace zaměřené na využití velkých dat, strojového učení a řízení datových projektů.

Období	2017–2018
Organizace	Czechitas
Role	Mentoring v oblasti návrhu a vývoje datových produktů.

Certifikace

Období	2018–2022 (220 hodin)
Kurz	Data Scientist in R, Data Scientist in Python, Machine Learning Scientist with R, Machine Learning Scientist with Python, Statistician with R
Instituce	Datacamp.com

Období	2015–2018 (170 hodin)
Kurz	Data Science Specialization, Executive Data Science Specialization
Instituce	Coursera, Johns Hopkins University

Období	2016 (110 hodin)
Kurz	Machine Learning
Instituce	Coursera, Stanford University

Jazykové dovednosti

Český jazyk – rodilý mluvčí, Anglický jazyk – C1

D Přehled publikací

Články indexované v databázi Web of Science (IF)

Kvasničková Stanislavská, L., Pilař, L., Fridrich, M., Kvasnička, R., Pilařová, L., Asfar, B., Gorton, M., (2023). Sustainability reports: The difference between developing and developed countries. Frontiers in Environmental Science, 11(1).

Fridrich, M. (2020). Understanding Customer Churn Prediction Research with Structural Topic Models. Economic Computation and Economic Cybernetics Studies and Research, 54(4/2020), 301-317.

Články indexované v databázi Scopus

Fridrich, M., & Dostál, P. (2022). User Churn Model in E-Commerce Retail. Scientific Papers of the University of Pardubice, Series D: Faculty of Economics and Administration, 30(1).

Ostatní články

Fridrich, M. (2019). Explanatory variable selection with balanced clustering in customer churn prediction. Ad Alta: Journal of Interdisciplinary Research, 9(1), 56-66.

Fridrich, M. (2017). Experimental Parameter Tuning of Artificial Neural Network in Customer Churn Prediction. Trends Economics and Management, 11(28), 9-21.

Příspěvky na konferencích

Fridrich, M. (2018). Cost-benefit metrics in customer churn prediction: A review. In MMK 2018: International Masaryk konference for Ph.D. students and young researchers (pp. 178-185). MAGNANIMITAS.