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

TEZE DISERTAČNÍ PRÁCE
DOCTORAL THESIS (ABRIDGED VERSION)

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

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

BRNO 2023

Klíčová slova

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

Keywords

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

Místo uložení práce

Vysoké učení technické v Brně

Fakulta podnikatelská

Oddělení pro vědu a výzkum

Kolejní 2906/4

61200 Brno

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.

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.

Obsah

Úvod 7

1 Teoretická východiska 10

1.1 E-commerce retail 10

1.2 Řízení vztahů se zákazníky 10

1.2.1 Koncept zákaznické hodnoty 11

1.2.2 Retenční management 11

1.3 Strojové učení 12

2 Literární rešerše 13

2.1 Predikce ztráty zákazníka 13

2.2 Ztráta zákazníka v e-commerce 14

3 Cíle práce a užité metody 17

3.1 Cíle práce 17

3.2 Výzkumné otázky 17

3.3 Užité metody 18

3.3.1 Metody vědeckého zkoumání 18

3.3.2 Strojové učení 20

4 Návrh a implementace řešení 23

4.1 Vymezení problému 23

4.2 Porozumění datovému souboru 24

4.3 Zpracování datového souboru 25

4.4 Modelování 25

4.5 Vyhodnocení a interpretace 26

4.6 Aplikace řešení 27

5 Shrnutí a diskuse dosažených výsledků 28

5.1 Výzkumné otázky 28

5.2 Limity a budoucí směřování výzkumu 36
6 Přínosy práce 38
6.1 Přínosy pro vědu a výzkum 38
6.2 Přínosy pro podnikatelskou praxi 38
6.3 Přínosy pro vzdělávání 38
Závěr 39
Literární zdroje 42
Seznam tabulek 48
Seznam obrázků 48
Seznam zkratek 48
Životopis autora 49
Přehled publikací 51

Úvod

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).

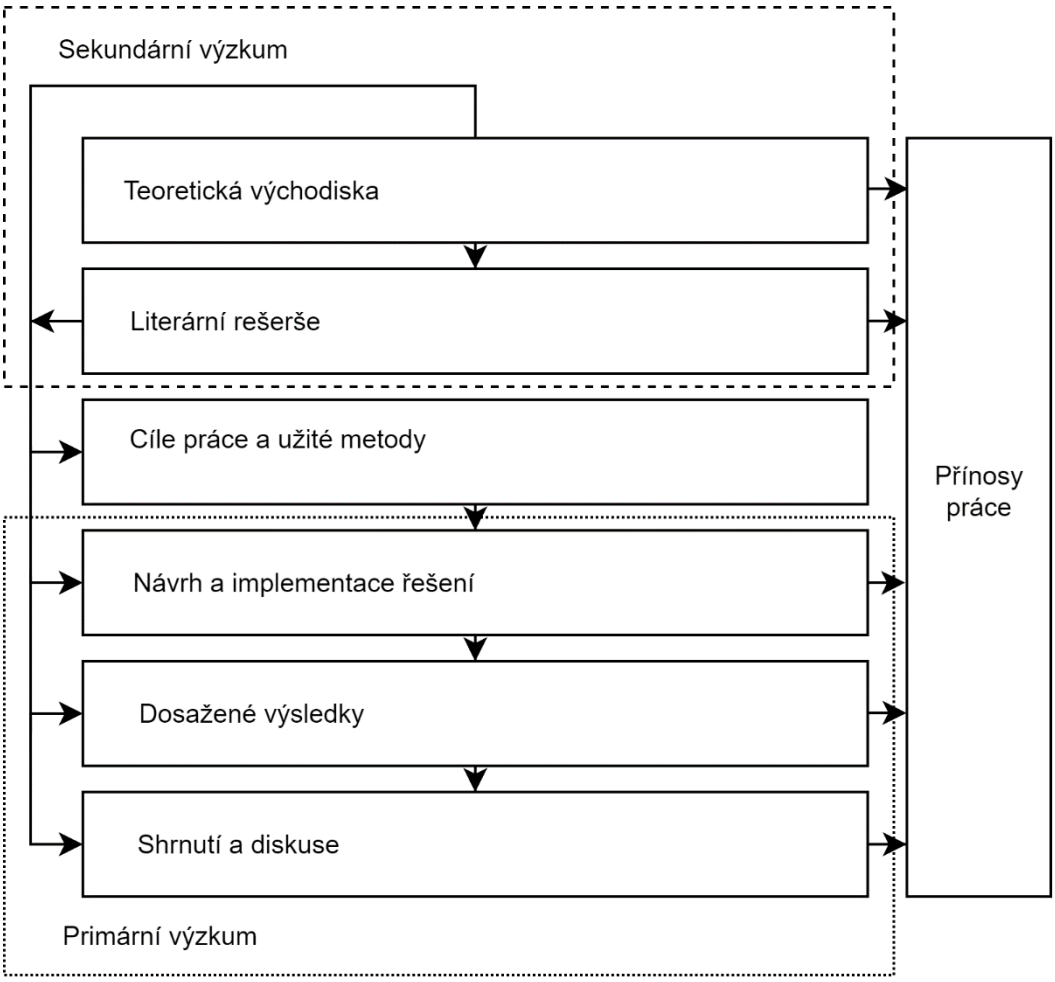
Struktura disertační práce

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.

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.



Obr. 1 Koncepce disertační práce

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.

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, konkurenční, tržní výhody a přidanou hodnotu. 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).

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“. Největší maloobchodní společností současnosti je právě e-commerce gigant Amazon.com Inc., s tržní kapitalizací 1.515 bilionů USD (Investopedia, 2022). 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, společnost 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.

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). Pojetí vlastního 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.

1.2.2 Retenční management

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 %. Gupta et al. (2004) odhadují 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.

1.3 Strojové učení

Strojové učení sdružuje přístupy, 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).

Z hlediska ukotvení bývá část strojového učení pojímána jako část umělé inteligence, tedy oboru počítačové vědy zabývajícím se řešením komplikovaných úloh, v kterých vynikají lidé. Systémy strojového učení však využívají i metod z oblasti matematické statistiky nebo matematické analýzy, hranice mezi obory jsou v tomto případě nejasné. 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).

2 Literární rešerše

Jako 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.

Množina textů popisující modelování odchodu zákazníka je zkoumána 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 odkrývají výzkumné mezery adresované cílem disertační práce.

2.1 Predikce ztráty zákazníka

Napříč změnami prevalence v čase lze 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 lze konstatovat, že se autorovi nepodařilo identifikovat třídu metod založených na fuzzy logice, její nízkou prevalenci naznačuje i práce Britto & Gobinath (2020). Následující sekce zevrubně analyzuje relevantní podmnožinu dokumentů popisující předpověď ztráty zákazníka ve vybrané vertikále elektronického obchodování.

2.2 Ztráta zákazníka v e-commerce

V rámci studované literatury lze rozlišit 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 se 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 lze pozorovat 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

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. 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. 1.

Tab. 1 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?

3.3 Užité metody

3.3.1 Metody vědeckého zkoumání

Empirické metody

Molnár et al. (2012) vymezuje jako empirické metody založené na přímém pozorování a měření skutečností. 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 Strojové učení

Dělení datového souboru

Za přístupem k dělení datového souboru je snaha napodobit podmínky aplikace prediktivního systému v reálném prostředí. Návrh experimentu vychází z křížové validace časových řad (Hyndman & Athanasopoulos, 2013) a seskupení časových výřezů (Gattermann-Itschert & Thonemann, 2021), tj, zabezpečuje časové odlišení trénovací a testovací množiny dat.

Ukazatele prediktivních schopností

Pro zhodnocení prediktivních schopností klasifikačních systémů jsou využity následující ukazatele:

- *Accuracy (ACC)* vystihuje podíl správně klasifikovaných pozorování,
- *F1 Score (F1)* sdružuje do jediného ukazatele podíl správně klasifikovaných predikovaných instancí pozitivní třídy a podíl správně klasifikovaných pozorovaných instancí pozitivní třídy s pomocí harmonického průměru,
- *Area Under Curve of the Receiver Operating Characteristic (AUCROC)* lze interpretovat jako pravděpodobnost, že pozorování náhodně vybrané z pozitivní třídy přiřadí klasifikační model vyšší skóre příslušnosti k pozitivní třídě než u náhodně vybraného pozorování negativní třídy.

Pro zhodnocení prediktivních schopností regresních systémů jsou uvažovány následující ukazatele:

- *Coeficient of Determination (R^2)* určuje podíl variability závislé proměnné vysvětlený regresním modelem,

- *Mean Absolute Error (MAE)* popisuje průměr absolutních odchylek pozorovaných a předpovídaných hodnot,
- *Mean Squared Error (MSE)* vystihuje průměr kvadratických odchylek pozorovaných a předpovídaných hodnot.

Srozumitelnost modelu

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, včetně konceptuálního 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. K odhadu dochází s pomocí simulace, která uvažuje pouze část koaličních proměnných. S výhodou je možné spočíst odpovídající vlastnosti s pomocí lineárního modelu. Mezi přednosti lze řadit pevné teoretické základy a spravedlivé rozdělení příspěvků vysvětlujících proměnných, dále také možnost jednotným způsobem interpretovat globální i lokální chování prediktivního systému, případně efektivní implementace pro vybrané algoritmy. Mezi nedostatky lze uvažovat výpočetní komplexitu, tvorbu nerealistických datových instancí u univerzálních metod, dále také náročnost interpretace nebo manipulaci výsledků (Masís,2021; Molnar, 2022).

Vybrané algoritmy

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 technik využívající více modelů. Metoda je postavená na sekvenční konstrukci 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.

4 Návrh a implementace řešení

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.

4.1 Vymezení problému

Vymezení problému je úzce spjato s cílem práce, podsekce se zabývá popisem původního přístupu k výpočtu zákaznické hodnoty, i odhadu ekonomického dopadu retenční kampaně vycházejícího z Tamaddoni et al. (2014).

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

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

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.

Maximální dosažený zisk

Ekonomický dopad je odhadnut 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 jako

$$\pi_i^{actual} = y_i[\gamma_i(CV_i - \delta)] + (1 - y_i)[-\psi_i \delta], \tag{2}$$

kde y_i označuje binární závislou proměnnou, jenž nabývá hodnoty 0 pokud zákazník setrvá a 1 pokud bude ztracen, 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. Druhý sčítanec koresponduje se zacílením retenční aktivity na zákazníka, u kterého odchod nehrozí. Výše incentivy δ je zvolena arbitrárně. Skutečný dosažený zisk určíme jako součet individuálních zisků zákazníků zahrnutých do kampaně, tj.

$$\Pi^{actual} = \sum \pi_i^{actual}, \tag{3}$$

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. 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 lze řadit především nutnost stanovení dílčích parametrů kampaně.

4.2 Porozumění datovému souboru

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 autor poukazuje 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.

4.3 Zpracování datového souboru

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).

4.4 Modelování

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).

4.5 Vyhodnocení a interpretace

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.

4.6 Aplikace řešení

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 následující Tab. 2.

Tab. 2 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

5 Shrnutí a diskuse dosažených výsledků

5.1 Výzkumné otázky

V následujících odstavcích shrnuje autor 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.

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 nepozoruje autor silný konsenzus.

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.

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. Výsledky jsou vyneseny v tabulkách Tab. 3 a Tab. 4. 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.

Tab. 3 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)

Tab. 4 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)

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.

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.

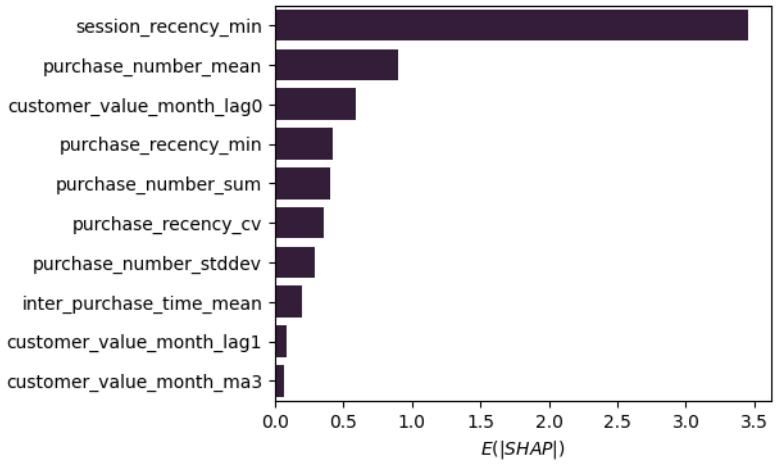
VO4: Jaké vysvětlující proměnné jsou klíčové pro predikci modelů?

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

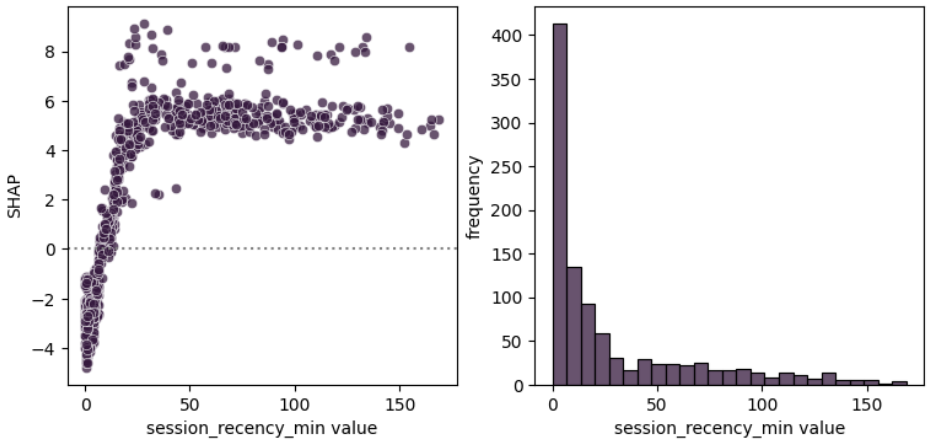
zákazníka, která se po v určitém bodě ustálí. V obou případech je tento bod dosažen po přibližně třech týdnech, což naznačuje asociaci s časovým vymezením závislé proměnné. Dále lze pozorovat exploataci minimálního množství transakcí pro výběr zákazníků, jev se však projevuje prostřednictvím odlišných nezávislých proměnných.

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. Popsané vlastnosti regresního rozhodovacího stromu pro datový soubor REES46 ilustrují Obr. 2 a Obr. 3.



Obr. 2 Proměnné významné pro predikci inkrementálního zisku retenční kampaně, s využitím rozhodovacího stromu – REES46



Obr. 3 SHAP hodnoty proměnných, významných pro predikci inkrementálního zisku retenční kampaně, včetně pozorovaného rozdělení – REES46

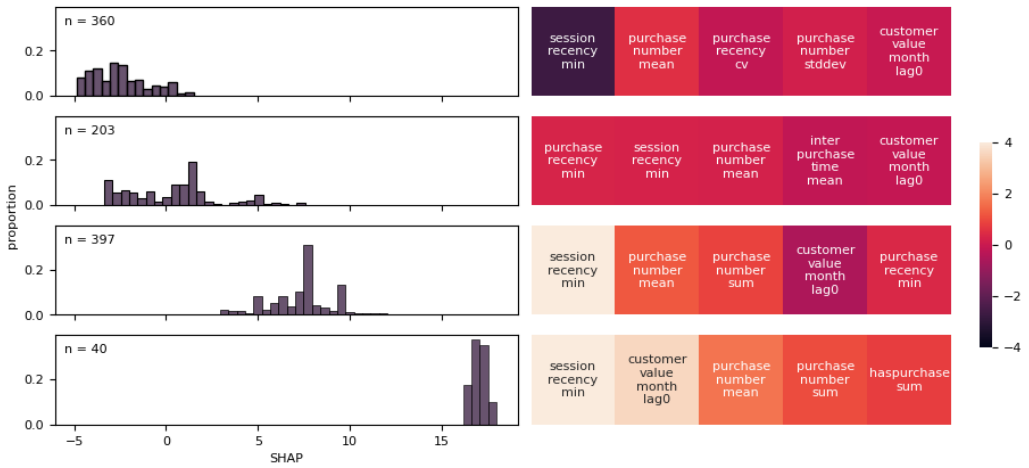
Původní přístup k problému omezuje přímé srovnání s existující literaturou. Pokud je ovšem regresní pojetí zasazeno do kontextu významných faktorů pojetí klasifikačního, pak lze pozorovat shodu na důležitosti transakčního a netransakčního chování uživatele. Nad rámec těchto faktorů dochází k růstu významu 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?

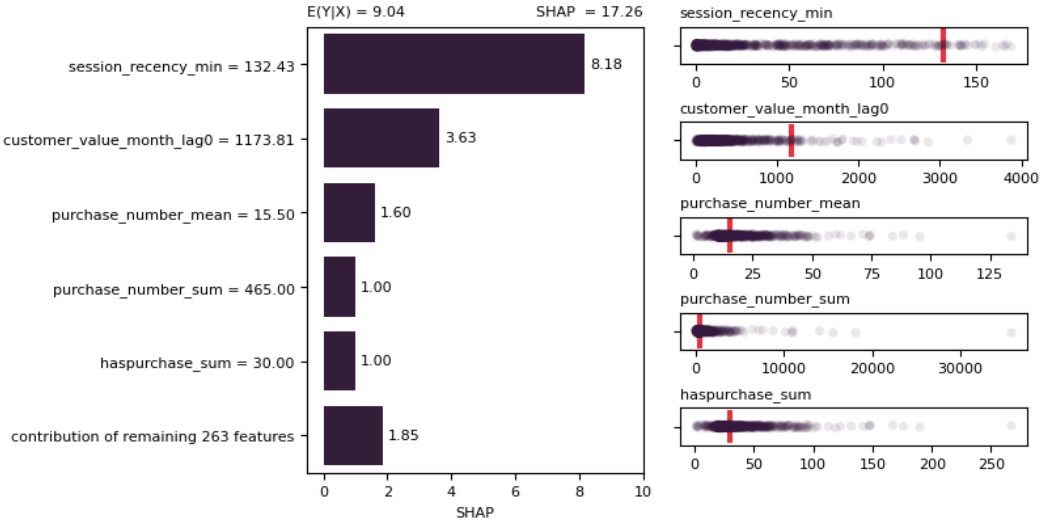
Autor se soustředí, podobně jako v předchozí výzkumné otázce, na interpretaci nejlepších řešení. Modely zkoumá na aktuálním časovém řezu s pomocí SHAP nástrojů a navržených rozšíření. Zákazníci jsou nejprve rozřazeni do shluků, dle SHAP hodnot napříč vysvětlujícími proměnnými. Shluk tak sdružuje pozorování, 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. Uvedené vlastnosti zákaznických shluků pro datový soubor REES46 dokumentují Obr. 4 a Obr. 5.



Obr. 4 Zákaznické shluky SHAP hodnot v predikci inkrementálního zisku retenční kampaně, včetně klíčových vysvětlujících proměnných – REES46



Obr. 5 Dopady významných vysvětlujících proměnných na inkrementálního zisku retenční kampaně, který je těžištěm shluku zákazníků, které je vhodné oslovit v rámci retenční kampaně – REES46

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.

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.

5.2 Limity a budoucí směřování výzkumu

Datové soubory

Určitá omezení vyplývají z povahy a detailu datových souborů, které zachycují interakce v prostředí elektronického maloobchodu, tj. v jiných vztazích nebo odvětvích nemusí být dosažené poznatky platné. Mezi další omezení náleží dostupná úroveň detailu, která nereflektuje další vlastnosti zákazníka a podniků; případně dílčí rozhodnutí při sestavení datové reprezentace. V budoucím výzkumu by tak bylo vhodné ověřit prezentované přístupy na datových souborů z dalších domén, ve větší úrovni detailu, případně v různých časových horizontech.

Strojové učení

Limity strojového učení vychází z části z návrhu experimentu, které může porušovat předpoklady některých modelů s ohledem na požadavek nezávislosti datových instancí. Další 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 mohl sloužit práce Olson & Moore (2019), případně Feurer et al., (2020). 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).

Ekonomické dopady retenční kampaně

Jedním z omezujících faktorů je absence atributu marží produktu, které jsou doplněny počítačovou simulací a nemusí odrážet ekonomickou realitu daných podnikatelských subjektů, bylo by tedy vhodné ověřit přístupy na kompletním datovém souboru. Dalším limitujícím faktorem jsou dílčí komponenty výpočtu inkrementálního zisku zahrnutí zákazníka do retenční kampaně. 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 k manipulaci reflektující předsudků autora. Oba problémy 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. Mezi další omezení lze řadit velmi komplexní zákaznický model, případně výpočetní náročnost zvoleného přístupu.

6 Přínosy práce

6.1 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. Vymezení jevu mění konstrukci, přístup k hodnocení, i interpretaci systému strojového učení. Mezi další přínosy pro vědu a výzkum lze řadit právě interpretaci modelovaného fenoménu, benchmark schopností prediktivních systémů, obsahovou analýzu vědecké literatury, nebo otevřený přístup k datové reprezentaci a programovému kódu aplikace. Autor tak připravuje prostor pro další zjištění a poznatky v oblasti aplikací strojového učení při řízení vztahů se zákazníky.

6.2 Přínosy pro podnikatelskou praxi

Výstupem pro podnikatelskou praxi je demonstrace zasazení předpovědi odchodu zákazníka do kontextu potřeb 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. Jako možné příjemce výzkumu předpokládá autor úspěšné maloobchodní společnosti, případně podniky zabývající se vývojem IT systémů jako jsou e-commerce platformy nebo systémy pro správu zákaznických vztahů.

6.3 Přínosy pro vzdělávání

Za přínosné pro pedagogiku 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.

Závěr

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 popisuje 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 původní 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), které jsou vhodně rozšířeny jak v oblasti implementace, tak v oblasti vizuálních nástrojů. V neposlední řadě se disertační práce zabývá některými praktickými aspekty aplikace systému, jako jsou technologická koncepce nebo odhad 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

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

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

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.

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.

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

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).

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

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

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

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

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

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

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

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

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

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/

Handley, L. (2013). *Customer retention: brave new world of consumer dynamics*. Marketing Week. Retrieved May 3, 2020, 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.

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

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

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

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, 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

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

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

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

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

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

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

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

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

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/

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

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

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

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

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

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

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

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

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

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

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

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., 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

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

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

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

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

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

Seznam tabulek

Tab. 1 Struktura cílů a výzkumných otázek disertační práce 18
Tab. 2 Datové soubory a programový kód 27
Tab. 3 Ukazatele ekonomického dopadu retenční kampaně – Retail Rocket 30
Tab. 4 Ukazatele ekonomického dopadu retenční kampaně – REES46 30

Seznam obrázků

Obr. 1 Koncepce disertační práce 8
Obr. 2 Proměnné významné pro predikci inkrementálního zisku retenční kampaně, s využitím rozhodovacího stromu – REES46 33
Obr. 3 SHAP hodnoty proměnných, významných pro predikci inkrementálního zisku retenční kampaně, včetně pozorovaného rozdělení – REES46 33
Obr. 4 Zákaznické shluky SHAP hodnot v predikci inkrementálního zisku retenční kampaně, včetně klíčových vysvětlujících proměnných – REES46 35
Obr. 5 Dopady významných vysvětlujících proměnných na inkrementálního zisku retenční kampaně, který je těžištěm shluku zákazníků, které je vhodné oslovit v rámci retenční kampaně – REES46 35

Seznam zkratek

ACC	angl. Accuracy, ukazatel přesnosti klasifikátoru
ANN	angl. Artificial Neural Networks, umělé neuronové sítě
AUC ROC	angl. Area Under the Curve – Receiver Operating Characteristics, ukazatel přesnosti klasifikátoru
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
F1	angl. F1 Score, ukazatel přesnosti klasifikátoru
GLM	angl. Generalized Linear Models, zobecněné lineární modely
LDA	angl. Latent Dirichlet Allocation, metoda modelování témat
LIME	angl. Local Interpretable Model-Agnostic Explanations
MAE	angl. Mean Absolute Error, průměrná absolutní chyba
MSE	angl. Mean Squared Error, průměrná kvadratická chyba
SHAP	angl. Shapley Additive Explanations, Shapleyho aditivní vysvětlení
SVM	angl. Support Vector Machines, metoda podpůrných vektorů
USD	angl. US Dollar, Americký dolar

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

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.