Česká zemědělská univerzita v Praze

Provozně ekonomická fakulta

Katedra statistiky

Bakalářská práce

Prediktivní modelování opětovného objednání služby

Vojtěch Dušek

ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

Provozně ekonomická fakulta

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Vojtěch Dušek

Informatika

Název práce

Prediktivní modelování opětovného objednání služby

Název anglicky

Predictive modeling of service re-ordering

Cíle práce

Cílem bakalářské práce bude vyhodnocení vybrané zákaznické databáze statistickými metodami užívanými v Big Data. Analýza bude založena na datech vybrané společnosti. Snahou bude identifikovat faktory, které mají vliv na objednávání vybraných služeb.

Metodika

Těžiště práce bude spočívat v analýze a vyhodnocení rozsáhlejší databáze. K řešení budou využity statistické metody mající uplatnění v Big Data, tj. metody vícerozměrné statistiky a metody prediktivního modelování.

Doporučený rozsah práce

30 – 40 stran

Klíčová slova

Data mining, prediktivní modelování, služba, zákazník, chování

Doporučené zdroje informací

ABBOTT, D. Applied Predictive Analytics : Principles and Techniques for the Professional Data Analyst. Praha: John Wiley & Sons, Incorporated, 2014. ISBN 9781118727935.

HEBÁK, P. *Statistické myšlení a nástroje analýzy dat.* Praha: Informatorium, 2015. ISBN 978-80-7333-118-4.

HENDL, J. *Přehled statistických metod : analýza a metaanalýza dat.* Praha: Portál, 2015. ISBN 978-80-262-0981-2.

KOTLER, P. *Moderní marketing : 4. evropské vydání.* Praha: Grada, 2007. ISBN 978-80-247-1545-2.

MAYER-SCHOENBERGER, V., CUKIER, K. Big Data: Revoluce, která změní způsob, jak žijeme, pracujeme a myslíme. Brno: Computer Press, 2014. ISBN 978-80-251-4119-9.

NISBET, R. – MINER, G. – ELDER, J. Handbook of statistical analysis and data mining applications. Amsterdam: Amsterdam, 2009. ISBN 978-0-12-374765-5.

RUD, O P. *Data mining : praktický průvodce dolováním dat pro efektivní prodej, cílený marketing a podporu zákazníků (CRM).* Praha: Computer Press, 2001. ISBN 80-7226-577-6.

SCHIFFMAN, L G. – KANUK, L L. *Nákupní chování.* Brno: Computer Press, 2004. ISBN 80-251-0094-4.

Předběžný termín obhajoby

2017/18 LS – PEF

Vedoucí práce

Ing. Tomáš Hlavsa, Ph.D.

Garantující pracoviště

Katedra statistiky

Elektronicky schváleno dne 15. 1. 2018

prof. Ing. Libuše Svatošová, CSc.

Vedoucí katedry

Elektronicky schváleno dne 15. 1. 2018

Ing. Martin Pelikán, Ph.D.

Děkan

V Praze dne 13. 03. 2018

Čestné prohlášení

Prohlašuji, že svou bakalářskou práci "Prediktivní modelování opětovného objednání služby" jsem vypracoval samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou citovány v práci a uvedeny v seznamu použitých zdrojů na konci práce. Jako autor uvedené bakalářské práce dále prohlašuji, že jsem v souvislosti s jejím vytvořením neporušil autorská práva třetích osob.

V Praze dne 15.3.2018 _____

Poděkování

Rád bych touto cestou poděkoval vedoucímu bakalářské práce panu Ing. Tomášovi Hlavsovi, Ph.D. za odborné rady a připomínky a také za vstřícný přístup. Také bych chtěl poděkovat vedení úklidové společnosti za zpřístupnění databáze.

Prediktivní modelování opětovného objednání služby

Souhrn

Cílem bakalářské práce bylo vyhodnotit hlavní faktory, které ovlivňují opětovně objednat služby u vybrané společnosti. Vybraná společnost poskytla svou zákaznickou databázi, nad kterou byla provedena statistická analýza.

Teoretická část pojednává o základech marketingu a nastiňuje důležitost udržení si svých stávajících zákazníků. Vybraná společnost zprostředkovává své služby na internetu, a proto se teoretická část soustředí na marketing služeb a marketing v internetovém věku.

Druhá část teoretické části pojednává o porozumění datům, data miningových metodách, strukturách firemních databází a prediktivním modelování.

V praktické části byla provedena popisná statistika všech atributů získaných z databáze vybrané společnosti. Popisná statistika identifikovala strukturu jednotlivých atributů a podnítila k očištění databáze od odlehlých a extrémních hodnot, které by mohly zásadně ovlivnit prediktivní modelování. Byly vytvořeny prediktivní modely pomocí metody rozhodovacích stromů a logistické regrese. Obě metody určily totožné faktory. Prodloužení doby služby do 21 minut, cenu služby v rozmezí 279 Kč a 1105 Kč a použití maximálně 2 promo kódů.

V závěru práce byly popsány hlavní faktory ovlivňující zákazníky při opětovném objednání služby. Vybrané společnosti bylo doporučeno, na které zákazníky se zaměřit, aby se zvýšil počet zákazníků, kteří službu opětovně objednají.

Klíčová slova: Data mining, prediktivní modelování, služba, zákazník, chování

Predictive modelling of service re-ordering

Summary

The main goal of thesis was focused on evaluating the main factors influencing the re-ordering of the services of selected company. The statistical analysis was performed on database provided by selected company.

The theoretical part described basics of marketing and importance of retaining existing customers. Theoretical part was focused on internet marketing because selected company is based on internet.

The second part of theoretical part described data mining methods, structure of company databases and predictive modelling.

The practical part was based on descriptive statistics of all attributes in database and predictive modelling. The database was cleared from remote and extreme values based on descriptive statistics. Predictive models were based on decision threes method and logistic regression method. Identical factors were identified by both methods. Extending service time to 21 minutes, service cost between 279 Kč and 1105 Kč and maximal 2 promo code per user. At the end of the thesis were described the main factors influencing re-ordering the service. Selected company have been advised how increase the number of re-ordering customers based on identified factors by analysis.

Keywords: Data mining, predictive modeling, service, customer, behavior

Obsah

1 Úvod

V dnešní době se stále více společností přesouvá na internet a poskytuje svoje služby online. Prostředí internetu se stává domovem čím dál tím více lidí v produktivním věku, a proto zde můžou prosperovat i služby zaměřené nejen na internetovou mládež.

Celkově na internetu probíhá velký konkurenční boj, který se nezakládá pouze na cenách. Velice významnými faktory jsou zákaznická podpora, spokojenost se službou nebo důvěra ve službu.

Všechny tyto faktory lze zaznamenávat ve firemních databázích, ze kterých lze dále těžit. Například spokojenost zákazníka se službou lze zaznamenat pomocí hodnocení služby zákazníkem. Zákazník hodnotí služby v rozsahu 1 až 5 hvězdiček. Ovšem hvězdičky nejsou jediný způsob, jak zákazníkovu spokojenost zjistit. Informace o spokojenosti jsou skryty v datech a za pomoc statistické analýzy lze tyto informace objevit.

Díky big data analýzám, data miningu a prediktivnímu modelování je možné z těchto dat vytěžit maximu a zvýšit efektivnost firmy na trhu. Z dat lze zjisti spoustu zajímavých faktů, na které by se bez analýzy nikdy nepřišlo. Příkladem zjištěných faktů můžou být faktory, které zákazníky nejvíce ovlivňují při rozhodování opětovného objednání služby. Díky analýze se firma může zaměřit na důležité faktory a zvýšit svou úspěšnost na trhu.

2 Cíl práce a metodika

2.1. Cíl práce

Cílem bakalářské práce je vyhodnocení vybrané zákaznické databáze statistickými metodami užívanými v Big Data. Analýza bude založena na datech společnosti, která online zprostředkovává úklidy v domácnostech a firmách. Snahou bude identifikovat faktory, které mají vliv na opětovné objednávání služeb.

2.2. Metodika

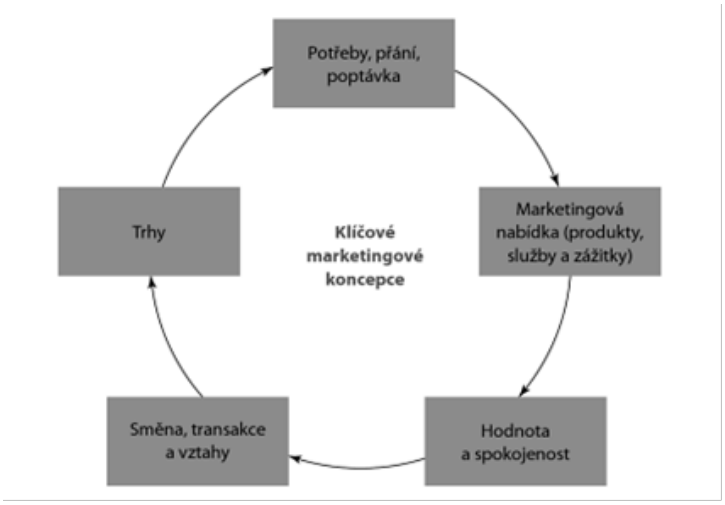
Těžiště práce bude spočívat v analýze a vyhodnocení rozsáhlejší databáze. K řešení budou využity statistické metody mající uplatnění v Big Data, tj. metody vícerozměrné statistiky a metody prediktivního modelování - rozhodovací stromy, logistické regrese, charakteristiky míry polohy, charakteristiky míry rozptýlenosti a box splot graf pro odlehlé a extrémní hodnoty. Výpočty metod rozhodovacích stromů a logistické regrese byly provedeny v programu Rapid Miner. Pro charakteristiky míry polohy, rozptýlenosti a odlehlých či extrémních hodnot MS Excel 2016.

3 Teoretická východiska

3.1. Úvod do marketingu

S marketingem jsme konfrontování dnes a denně. Každá firma ví, pokud má efektivní marketing, že podíl na trhu a zisky se zvýší. Základem marketingu je tvorba hodnot u zákazníka a jeho uspokojení. Jednoduše řečeno, marketing je uspokojení zákazníka se ziskem. V dnešní době se o marketing nezajímají pouze velké firmy, ale i subjekty, které na první pohled marketing nepotřebují. Například právníci, lékaři, neziskové organizace (školy, muzea, policejní oddělení). [1]

Marketing není pouze umístění reklamy na Facebook nebo do televize. Marketing je chápán jako celek, jako produkt nebo dokonce jak celou firmu vnímáme. Není cílem pouze přilákat nové zákazníky, ale udržet spokojenost a věrnost těch stávajících. Dokonalým příkladem jsou takzvané „žhavé výrobky“. Například telefon iPhone, herní stanice PlayStation, vozy Smart. Na tyto výrobky se při spuštění prodeje stály dlouhé fronty a kvůli návalu zákazníků jich byl nedostatek. Mají jedno podobné, a to že se odlišovaly od ostatních konkurenčních výrobků. Tento úspěch je výsledkem práce specialistů, kteří identifikovali potřeby zákazníka, vytvořili produkt nabízející vysokou hodnotu a nakonec ho dobře propagují. Prodej a reklama jsou důležitou součástí marketingu, ale také jsou součástí důležitého marketingového mixu. [1]



Obrázek 1 Klíčová marketingová koncepce [1]

3.1.1. Potřeby, produkty, prožitky

Marketing je založen na lidských potřebách a pociťovaných nedostatcích. Pokud zákazníkova potřeba není uspokojena, zákazník může buď vyhledat předmět, který jeho potřebu uspokojí, nebo se pokusí potřebu omezit. Zákazníkovy potřeby jsou uspokojovány hodnotovou propozicí, což je výčet vlastností, které zákazníkovu potřebu splní. Marketingová propozice je naplněna pomocí marketingové nabídky, což je kombinace produktů, služeb, informací nebo prožitků. [1]

Důležité je zákazníkům ukázat výhody a prožitky. Pokud se marketing soustředí na produkt a jeho vlastnosti, je to chybný postup. Například výrobce vrtáků si myslí, že zákazníci potřebují vrták. Opak je pravdou. Zákazník potřebuje vyvrtat díru, anebo upevnit něco za pomocí vrtáku. Je důležité si uvědomit, co jsou skutečné potřeby zákazníků a upřednostnit je před okamžitým přáním. [1]

Úplně nejlepším způsobem marketingu je vytvořit poselství a prožitek ze značky. Tyto značky se ve svém odvětví stávají ikonou a její zákazníci jsou ji věrní, i když jejich produkty občas nejsou nejlepší na trhu. Takzvaný prožitek ze značky je silnější, než lepší vlastnosti konkurenčního produktu. [1]

3.1.2. Výhodné vztahy se zákazníky

Poptávka po produktech pochází od dvou skupin zákazníků. Nových a opakovaně nakupujících zákazníků. V dnešní době internetu zákazníci nejsou vázání demograficky, takže získat nového zákazníka je velmi těžké. Náklady na nového zákazníka pětkrát převyšují náklady na udržení stávajícího zákazníka. Právě z tohoto důvodu se pozornost přesouvá na udržení zákazníků a vybudování pevných vztahů. Ztráta zákazníka neznamená pouze jednorázovou ztrátu, ale ztrátu jeho budoucích, ne-li celoživotních útrat. Proto je klíčová hodnota a uspokojení zákazníka. [9]

3.2. Marketingový proces

V dnešní době firmy potřebují zákazníkovi nabídnout přidanou hodnotu oproti konkurenci, aby ho přesvědčily k nákupům u nich. Základem je pochopit jejích přání a potřeby. K pochopení těchto dvou základních faktorů je zapotřebí provést analýzu zákazníkova chování. Analýza se nezajímá jen o klíčovou transakci, ale o proces před ní a po ní.

- Společenské postavení spotřebitele
- Životní styl
- Preference dávno předtím, než vznikne potřeba produktu
- Uvědomění si produktu
- Zájem o určitý produkt
- Touha po konkrétním způsobu uspokojení potřeby
- Samotný proces nákupu produktu

Marketing je členitý a různorodý právě díky základním dvěma složkám. Lidský mozek a společnost, ve které žijeme. Snaha porozumět spotřebiteli je prováděna pomocí různých metod. Počínaje u psychologie a konče u fyziky. Právě kvůli obrovskému množství vlivů působících na zákazníkovo myšlení neexistuje ucelený postup, jak dosáhnout úspěšného marketingu. [1]
Důležité je si uvědomit, že provedením transakce marketing rozhodně nekončí. Právě naopak v tento moment nastává další fáze. Fáze vytváření dlouhodobého vztahu. Postoj a životní styl jsou ovlivněny předchozími nákupy. Každý nákup a zkušenost s ním ovlivňuje budoucí nákupy. [1]

3.3. Strategický marketing – segmentace trhu

Společnosti vědí, že nemohou cílit na celou společnost, a proto si segmentují trh. Na jednotlivé segmenty trhu použijí odlišné marketingové strategie, aby jednotlivé segmenty mohly ziskově obsluhovat.

- Segmentace trhu: Odlišení skupin s odlišnými potřebami.
- Segment trhu: Sdružuje lidi, kteří reagují na určitý marketing podobně.
- Targeting: Vyhodnocení atraktivity segmentů a výběr nejlepších segmentů.
- Positioning: Zákazníci výrazně vnímají produkt ve srovnání s konkurencí.
- Pozice trhu: Odlišné zachycení produktu v paměti zákazníků ve srovnání s konkurencí. [1]

3.3.1. Marketingový mix

Po jasně definované marketingové strategii je potřeba definovat marketingový mix. Zahrnuje všechny možnosti, jak firma může ovlivnit poptávku po svém produktu. Zahrnuje čtyři části. Produktová politika, cenová politika, konkurenční politika a distribuční politika. [1]

Obrázek 2 Marketingový mix [2]

3.4. Marketingové prostředí

Tempo dnešní doby se stále zrychluje a právě marketing na to musí být schopný reagovat. Čím dál tím více je dnešní svět propojený a společnost musí reagovat nejen na zájmy svých nejbližší zákazníků, ale i na dodavatele, konkurenci i veřejnost. Důležité je pochopit kontext prostředí, ve kterém marketing probíhá. Faktory ovlivňující prostředí jsou demografické, ekonomické, přírodní, technologické, politické a kulturní. [1]

3.4.1. Demografické prostředí

Zajímá se o to, kdo vlastně trh tvoří. Vychází ze statistických údajů o složení populace z hlediska pohlaví, rasy a zaměstnání.
Například změny ve věkové struktuře obyvatel, odlišné chování generace X a Y, migrační tlaky, stálý růst vzdělanosti, rostoucí různorodost všech členů společnosti. [1]

3.4.2. Ekonomické prostředí

Jednotlivé státy se liší v trendu utrácení a marketéři musí tyto trendy pečlivě sledovat. Mohou se právě lišit podle uvedených rozdílných generací X a Y, ale zároveň je ovlivňuje ekonomické prostředí, ve kterém se pohybují. [1]

3.4.3. Přírodní prostředí

Zahrnuje přírodní zdroje, které nejsou nekonečné. Jak se přírodní zdroje spotřebovávají, prostředí, ve kterém žijeme, se mění. Například nedostatek surovin, rostoucí ceny energií, růst znečištění a další. [1]

3.4.4. Technologické prostředí

Životnost technologií je dnes stále kratší, protože jsou vytlačovány novými a lepšími technologiemi. Proto se zvyšují rozpočty na výzkum a vývoj, aby se produkt mohl delší dobu udržet na trhu. Je zapotřebí stále uvádět na trh produkty s drobnými zlepšeními, aby zákazníci zase mohli nakupovat lepší a lepší produkty. [1]

3.4.5. Politické a kulturní prostředí

Politické a kulturní prostředí se v každém státu liší. Rozdílné zákony, důraz na etiku a společenskou zodpovědnost, stálost kulturních hodnot. Marketingová strategie musí se všemi těmito odlišnostmi počítat a chytře je využívat ve vlastní prospěch. [1]

3.5. Marketing služeb

„Služba je aktivita nebo výhoda, kterou může jedna strana nabídnout druhé, je v zásadě nehmotná a nepřenáší vlastnictví.“ [1] Fenomenální růst služeb je zapříčiněn větším bohatstvím ve společnosti a tím pádem způsobuje přesouvání nezábavných či složitých úkonů na třetí stranu. Úplně jiný rozměr to dostává v době internetu.
Základními vlastnostmi služeb jsou nehmotnost, neoddělitelnost, proměnlivost, pomíjivost a absence vlastnictví. [1]

3.5.1. Proměnlivost

Proměnlivost můžeme chápat, jako když dvě firmy provádějí stejnou službu, ale liší se svým přístupem. Díky tomuto přístupu má zákazník úplně jiný dojem z totožné služby a vždy se raději vrátí k té s lepším přístupem. [1]

3.5.2. Pomíjivost

Hodnota služby existuje jen v určitý moment. Pokud tento moment uplyne, služba už nemusí mít žádnou hodnotu. Logicky z toho vychází strategie různých cen v různou dobu. Část poptávky se přesune mimo špičku a část poptávky, která si může dovolit připlatit, zaplatí vyšší cenu. [1]

3.6. Marketing v internetovém věku

Internet změnil vše a marketing firem působících na internetu nevyjímaje. Dokonce i firmy, které odmítaly mít cokoli společného s internetem, byly nepřímo ovlivněny. V internetovém věku se potlačuje mainstreamové cílení na potencionální klienty. Zvyšuje se intenzita na personalizaci, budování vztahů a přesnější cílení na ještě menší segmenty trhu. [1]

3.6.1. Customizace a customerizace

Hlavní trend internetové doby se točí kolem informací. Díky informacím o zákazníkovi můžeme snadno rozlišit cílové skupiny, personalizovat a individualizovat jejich potřeby a reagovat na tyto potřeby téměř okamžitě přes moderní komunikační kanály nebo přímo upravit nabízený produkt či službu.

Velice důležité odvětví podporující marketing je sbírání dat o svých zákaznících, dodavatelích a vlastně všech vstupech a výstupech při podnikání. A ještě důležitější než samotné sbírání dat je následné vyhodnocení dat a ziskové použití získaných informací v marketingu. [1]

3.6.2. Online spotřebitelé

Uživatelé internetu již dávno nejsou jen „hackeři“ či „geekové“. Internet je dnes domovem široké veřejnost, a proto sbírání dat je o to důležitější.

Internetoví uživatelé reagují mnohem více na negativní marketing firmy a nemají problém takřka okamžitě přejít ke konkurenci. [1]

3.7. Data mining

Data mining je soubor technik, které firmy využívají ke zlepšení marketingových cílů. Díky těmto technikám můžeme z datových skladů získat velmi důležité informace, které nejsou na první pohled vůbec jasné. Právě proto data mining hraje klíčovou roli, jak si firmy udržují stabilní pozici na trhu vůči konkurenci. Metody data miningu se rozšířily v nejrůznějších odvětvích v poměrně nedávné době, ale první zárodky data miningu se objevily již v 60. letech. Úspěšné využití přišlo v oblasti získávání nových zákazníků až v druhé polovině 80. let. V oblasti budování dlouhodobého vztahu se zákazníky se data mining rozšířil až později. Hlavní obory, které využívají data mining, jsou bankovnictví, pojišťovnictví a další velká odvětví, které mají spoustu dat od svých zdrojů. Na druhou stranu, v dnešní době se data miningem nezbývají pouze velké korporátní společnosti, ale i podniky menších rozměrů. [8]

3.7.1. Datové sklady

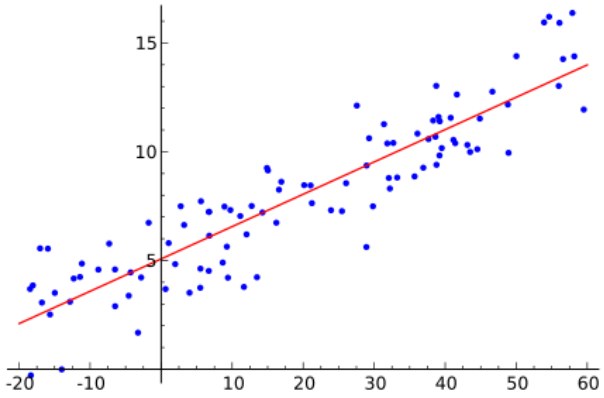
Datový sklad je rozsáhlé centrální místo pro ukládání dat z různých zdrojů. Obsahují pouze vymezené části dat z datových zdrojů, které jsou relevantní pro další zkoumání data miningovými technikami. [3]

3.7.2. Data miningové metody

Data miningové metody využívají matematické, statistické znalosti a metody umělé inteligence. [4]

3.7.2.1. Regresní analýza

Regresní analýza zkoumá závislost mezi dvěma spojitými proměnnými. Jednodušeji řečeno popisuje vztah mezi proměnou X a Y. Závislou proměnnou Y se snažíme predikovat na základě vztahu s nezávislou proměnnou X (prediktivní či vysvětlující proměnnou). Regresní funkce říká, že závislé proměnné jsou spojeny s nezávislými na základě neznámých parametrů. Cílem regresní analýzy je zjistit závislost mezi proměnnými a vyjádřit jí matematickou funkcí. [4]



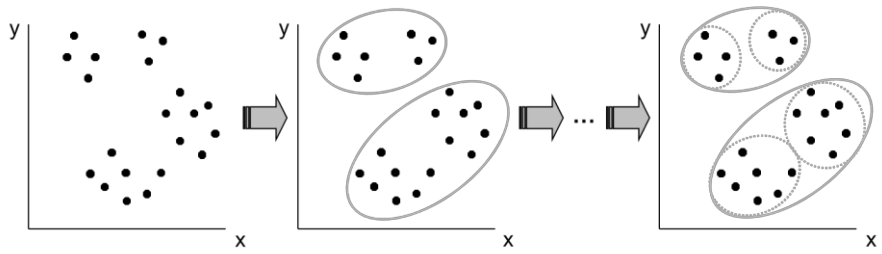
Obrázek 3 Lineární regresní model [5]

3.7.2.2. Diskriminační analýza

Diskriminační analýza je používána k rozdělení prvků do skupin podle třídícího kritéria. Data jsou tříděna do odlišných skupin podle kritéria diskriminačních funkcí. Známe deskriptivní a predikční diskriminační analýzu. [4]

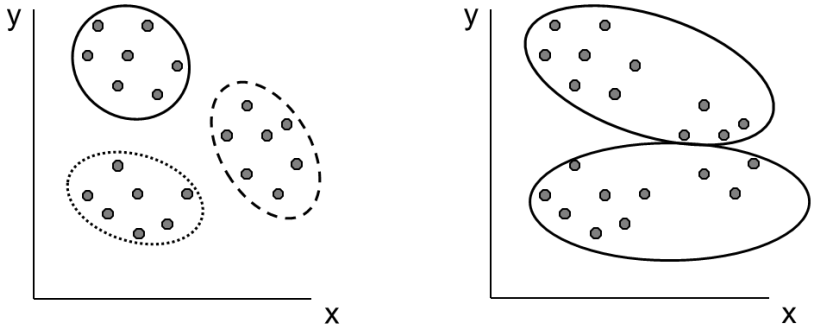
3.7.2.3. Shluková analýza

Shluková analýza třídí data do určitých shluků neboli skupin. Jednotky, které se nacházejí ve stejné skupině, si budou mezi sebou více podobné než jednotky nacházející se v odlišných skupinách. Shlukovou analýzu můžeme dělit na hierarchickou a nehierarchickou. Hierarchické metody vytvářejí novou podmnožinu již existující skupiny, neboli detailněji rozdělí skupinu do dalších podskupin. Hierarchické metody se dále dělí na aglomerativní a divizní přístup. [6]



Obrázek 4 Divizní hierarchická shluková analýza [7]

Když data nevykazují hiearchitickou strukturu, je potřeba použít nehierarchitický přístup. Data ve skupinách jsou co nejvíce homogenní a skupiny mezi sebou co nejvíce odlišné. [7]



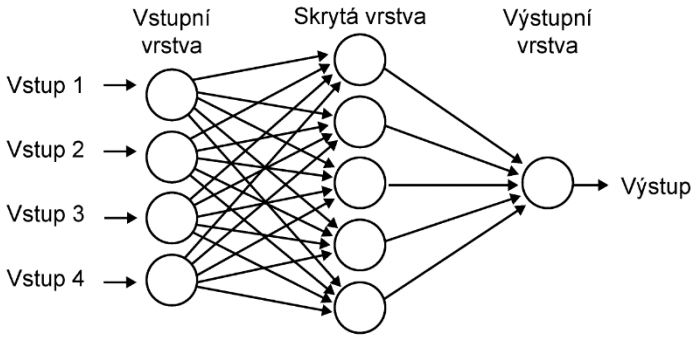
Obrázek 5 K-průměr nehierarchická shluková analýza [7]

3.7.2.4. Rozhodovací stromy

Rozhodovací stromy rozdělují data do odlišných skupin či větví tak, aby se maximalizovaly rozdíly hodnot závislé proměnné. Zobrazí data ve formě stromu a na každém uzlu je určeno kritérium pro rozdělení hodnot do větví. V každé větvi se vyskytují hodnoty s podobnými vlastnostmi. Tato metoda je velmi oblíbená pro její přehlednost a možnost rychlé interpretace výsledků. [4]

3.7.2.5. Neuronové sítě

Neuronové sítě nevychází z žádné statistické metody. Jsou modelovány podle funkcí organizace nebo lidského mozku. Jsou velice citlivé na kvalitu vstupů. Na vstupních datech se postupně učí každou iterací. [4]



Obrázek 6 Uspořádání neuronů v neuronové síti [7]

Obrázek zobrazuje jednoduchou neuronovou síť s jednou skrytou vrstvou. Vstupní data se rozdělí na trénovací a testovací množinu. Potom se přiřadí váhy ke každému uzlu ve vstupní vrstvě a po každé iteraci jsou vstupy porovnány se skutečnou hodnotou. Pokud se vyskytne chyba, systém upraví původní váhy. Iterace pokračují, doku se nedosáhne minimální chyby. [4]

3.8. Porozumění datům

Data jsou informace získané pozorováním objektů. Nehledě na původ, data můžeme dělit do tří základních skupin. [4]

3.8.1. Demografická data

Obecně můžeme říct, že popisují charakteristiky osob nebo domácností. Jako příklad lze uvést pohlaví, věk, jméno, příjmení, vzdělání, bydliště, národnost apod. Výhodou je jejich stabilita, a proto se dají dobře použít v prediktivních modelech. Jsou velice levná na pořízení. Na druhou stranu jsou obtížně získatelná s detailem na jednoho jedince. [4]

3.8.2. Behaviorální data

Tato data představují míru akce nebo chování. Příkladem může být prodané množství výrobků, datum nákupu, míra utracené částky za nákup nebo chování zákazníka na webu. Například kam kliknul, zachycení cesty proklikávání webem, naplnění košíku. Hlavní výhodou je, že tato data poskytují nejsilnější predikční sílu. Také jsou velice obtížně získatelná nebo velice drahá. [4]

3.8.3. Psychologická data

Představují názory, osobní hodnoty či životní styl lidí. Lze je získat výzkumem mínění u názorových skupin, nebo vyvodit z nákupního chování zákazníků. Tato data můžeme považovat za doplňková, jelikož nemají takovou predikční sílu, jako předešlé kategorie. Nelze je považovat za 100% pravdivé, protože se jedná o názorová data, která jen odhadujeme. Dost často mohou jen lehce korelovat se skutečností. Obvykle se využijí v momentě, kdy firma získala maximum z demografických a behaviorálních dat. Psychologická data mohou posloužit jako další rozměr pro datovou analýzu. [4]

Zdroj: [4]

	Schopnost predikce	Stabilita	Cena
Demografická	Střední	Vysoká	Nízká
Behaviorální	Vysoká	Nízká	Vysoká
Psychologická	Střední	Střední	Vysoká

Tabulka 1 Vlastnosti typů dat [4]

3.8.4. Zdroje dat

3.8.4.1. Interní zdroje dat

Představují hlavní firemní datové zdroje, které firmě přímo náleží. Většinou právě tato data mají nejvyšší predikční sílu, protože se přímo týkají prodávaných výrobků a oboru firmy. [4]

3.8.4.1.1. Zákaznická báze

Pokud firma má pouze jedinou databázi, tak v ní budou obsaženy všechny informace. Mnohem častější a profesionálnější přístup je, že v zákaznické databázi jeden identifikační záznam představuje jednoho zákazníka. Přes tento identifikační záznam se propojuje s ostatními firemními databázemi, které obsahují podrobnější informace o nákupech a chování zákazníka. Příkladem údajů v zákaznické databázi může být: ID, jméno, příjmení, adresa, telefon, číslo nakoupeného výrobku a podobně. [4]

3.8.4.1.2. Transakční databáze

Tento typ databáze obsahuje veškerou transakční aktivitu zákazníka. Každý řádek v databází představuje jednu transakci mezi zákazníkem a firmou. Počet řádků jednotlivého zákazníka se může lišit podle jeho aktivity. Právě tato data mají vysokou predikční hodnotu, ale pouze v případě, že je firma schopna tato data zužitkovat. Obvykle obsahuje tyto položky: ID, číslo účtu, velikost transakce a datum aktivity. [4]

3.8.4.1.3. Databáze historie nabídek

Tato databáze se skládá ze všech nabídek a podrobnostech o nabídkách, které byly vytvořeny pro stávající i potencionální zákazníky. Proměnné z této databáze mají obvykle tu nejvyšší predikční schopnost u cílených modelů odpovědí. Tato databáze často velmi rychle roste a je potřeba ji pravidelně udržovat. Běžné položky jsou: ID, jméno, příjmení, adresa, telefon, podrobnosti o nabídce, souhrn nabídky, hodnocení modelu a prediktivní údaje. [4]

3.8.4.2. Externí zdroje dat

Externí zdroje dat jsou poskytovány subjekty, které se na tuto činnost přímo specializují. Prodávají celé databáze nasbíraných jmen, telefonních čísel, adres a podobně. Často je obohacují o demografické, behaviorální a psychologické údaje. Databáze mohou být prodávány v surové formě. Některé subjekty databází neprodávají pouze v surové formě, ale ještě s ní dále pracují a databáze takzvaně pročišťují. Tento proces zvyšuje tržní hodnotu externího zdroje dat. [4]

3.9. Prediktivní modelování

Prediktivní modelování je proces, pomocí kterého hledáme v datech vzory chování a na základě těchto vzorů předpovídáme budoucí chování. Prediktivní modelování je založeno na proměnlivých faktorech, což jsou nezávislé proměnné, které definují chování vysvětlované proměnné. V procesu odvození vysvětlované proměnně je obsažena identifikace nezávislých proměnných, vybrání charakteristik definujících model a definování koeficientů a váhy pro vybrané proměnné. [10]

Algoritmy prediktivního modelování automatizují proces hledání vzoru chování v datech. Některé algoritmy nedefinují pouze koeficienty a váhy z dat, ale jdou ještě dál. Například rozhodovací stromy určí, které vstupní prvky mají nejlepší hodnoty nezávislých proměnných pro predikční schopnost modelu. Zároveň identifikují, které hodnoty nezávislých proměnných jsou charakteristické pro vzory chování. [10]

3.10. Logistická regrese

Logistická regrese se liší od lineární regrese v jednom zásadním bodě. Logistická regrese je vhodná pro data, kde závislá proměnná není kvantitativní, ale kvalitativní. Závislá proměnná může nabývat dvou a více úrovní, ovšem nejčastěji bývá pouze dvouúrovňová. Například 0 a 1. Pro každou hodnotu se spočítá pravděpodobnost pro výskyt pozitivní predikce a pravděpodobnost pro případ negativní predikce. [4]

Obecný vzorec logistické regrese

$$\text{logit(Y)} = ln\ \frac{\pi}{1-\pi} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_k X_k$$

kde $\beta_0, \beta_1 \ldots \beta_k$ jsou regresní koeficienty a π je podmíněná střední hodnota vysvětlované proměnné. Střední hodnotu lze vyjádřit ze vztahu $\pi = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k}}$

Výsledkem je logit, který je špatně interpretovatelný a může nabývat hodnot od -∞ do +∞. Proto se logistická regrese dá vyjádřit ve formě šancí nebo pravděpodobností. Vždy je zapotřebí definovat, pro jaký stav závislé proměnné šanci nebo pravděpodobnost počítáme. Šance se odvodí z původního obecného vzorce pomocí exponenciální funkce. Nabývá hodnot od 0 do +∞ a přesně hodnotí, kolikrát je pravděpodobnost vyšší, že predikovaná hodnota bude 1, než že predikovaná hodnota bude 0. [11]

$$\text{šance}(Y = 1) = \exp[logit(Y)] = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_k X_k)$$

Pravděpodobnost nabývá hodnot mezi 0 a 1 a představuje, o kolik se zvýší pravděpodobnost predikce predikované proměnné na 0 nebo 1, pokud se hodnota zkoumané proměnné změní o jednotku, ve které je měřená. Například koruny, hodiny apod. Zároveň hodnoty ostatní znaků musí zůstat nezměněny. V uvedeném vzorci je výchozí pravděpodobnost pro predikovanou hodnotu 1.

$$\check{s}ance(Y = 1) = \exp[logit(Y)] = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_k X_k)}$$

[11]

3.11. Rozhodovací stromy

Existují odlišné typy rozhodovacích stromů. Typy se liší algoritmem, který rozhodovací stromy vytváří. Struktury stromů se vytvářejí dělením dat do odlišných skupin na základě závislé proměnné. Na každém uzlu se vyhodnotí, který atribut data nejlépe rozděluje do podmnožin. Pokud bude splněno kritérium pro zastavení větvení, strom se již dál nerozvětví. Kritérium může být definováno maximální hloubkou stromu, maximálním počtem listů nebo stupněm homogennosti dat v listu.

3.11.1. Algoritmus ID3

ID3 je založen na principu vytváření stromu od shora dolů. Lze ho použít v případě, kdy jsou všechny atributy kvalitativní. Pro každý uzel je vybrán atribut, podle kterého se listy budou dělit. Vybrání atributu je založeno na minimální homogenitě dat neboli entropii. Entropie je vypočítána pro každý atribut následujícím vzorcem. [13]

$$E(b) = \sum_e (-\frac{n_{bc}}{n_b}) \log_2 (\frac{n_{bc}}{n_b})$$

„Kde b je vznikající větev, c je třída závislé proměnnou, n_b *je počet objektů ve větvi b,* n_{bc} *představuje počet objektů třídy c ve větvi b.“* [13] Na daném uzlu se pro všechny atributy spočítá průměrná entropie, která je spočítána následujícím vzorcem.

$$E = \sum \left(\frac{n_{bc}}{n_t}\right) E(b)$$

„Kde n_t *je celkový počet objektů ve všech větvích.“* [13] Pokud entropie vyjde s hodnotou 0, dělení se v daném listu ukončí.

3.11.2. Algoritmus C4.5

C4.5 je vylepšená verze algoritmu ID3. Zásadní vylepšení spočívají ve využití kvantitativních atributů a možnosti chybějících dat. V algoritmu C4.5 se navíc počítá podmíněná entropie, která je vyjádřena následujícím vzorcem. [13]

$$E(x|T) = \frac{n_{bx}}{n_b}\log_2(\frac{n_{bx}}{n_b})$$

Pomocí obecné entropie a podmíněné entropie je spočten zisk. Zisk by měl být maximalizován skrze dělení atributů podle hodnoty x. [13]

$$Zisk(A, x) = E_A - E(x|A)$$

„Kde A je vybraný atribut, E_A jeho obecná entropie, x hodnota.“ [13]

3.12. ROC křivka

ROC křivka graficky vyjadřuje kvalitu modelu. Rozřadí výsledky do dvou skupin. Skupina s predikovanou hodnotou 1 a skupina s predikovanou hodnotou 0. Rozřadí vzorky na základě shody predikované hodnoty a reálné hodnoty z výběrové databáze.

- Skupina s predikovanou hodnotou 1 a reálnou hodnotou 1
- Skupina s predikovanou hodnotou 0 a reálnou hodnota 0

Ideální model by měl všechny vzorky zařazený v obou skupinách a ROC křivka by měla tvar lomené čáry. Reálný model nezařadí do dvou skupiny vzorky, u kterých predikce nebyla správná a reální hodnota není totožná s predikovanou hodnotou. Křivka začne měnit tvar. Kvalita modelu se hodnotí podle velikosti plochy pod ROC křivkou. Čím je plocha pod křivkou větší, tím je model kvalitnější. [12]

4 Vlastní práce

4.1. Vybraná společnost

Vybraná společnost působí v oblasti úklidu domácností a firem. Na trhu se před pár lety etablovala díky modernímu online marketingu a online platformě pro objednávání úklidů. Povedlo se jí prosadit již na zavedeném trhu díky inovativnímu přístupu. Především se orientuje na klientelu ve velkých českých městech, jako je Praha nebo Brno. Poskytuje službu pravidelných úklidů, ale lze si ji objednat i na jednorázový úklid. Zakládají si na osobním přístupu. Pokud je zákazník spokojen s úklidem, je možné si v příští objednávce objednat zase stejnou paní na úklid. Každá objednávka je ukončena hodnocením klienta v rozsahu 1 až 5 hvězdiček.

Tato společnost pro tuto bakalářskou práci poskytla svou celkovou anonymizovanou databázi, ve které lze najít informace o všech provedených objednávkách úklidu v rozmezí 2 let.

4.2. Popisná statistika

Byl poskytnut základní soubor, který obsahoval spoustu tabulek a relací mezi nimi. Z tohoto základního souboru byl díky data miningu vytvořen výběrový soubor založený na určitých podmínkách. Každý záznam představuje unikátního uživatele, který si objednal službu a následně za ni zaplatil. Jelikož hodnocení služby není povinné, bylo zapotřebí odstranit všechny uživatele, kteří ani jednou neuvedli hodnocení za objednávku. Hodnocení spokojenosti zákazníkem je velice důležitý faktor pro predikování opětovného objednání služby. Proto nebylo možné do výběrového souboru uvést uživatele, kteří nikdy nehodnotili. Mohlo by to zásadně ovlivnit výsledek.

4.2.1. Statistické znaky

Statistickou jednotkou je unikátní uživatel.

Statistické znaky jsou: user_id, re_order, saved_card, avg_order_price, count_promo_code, avg_promo_price, avg_base_length, avg_margin_length, order_subscription, last_order_active_subscription, subject, city, zipcode, avg_customer_rating

Predikovaná proměnná je znak re_order.

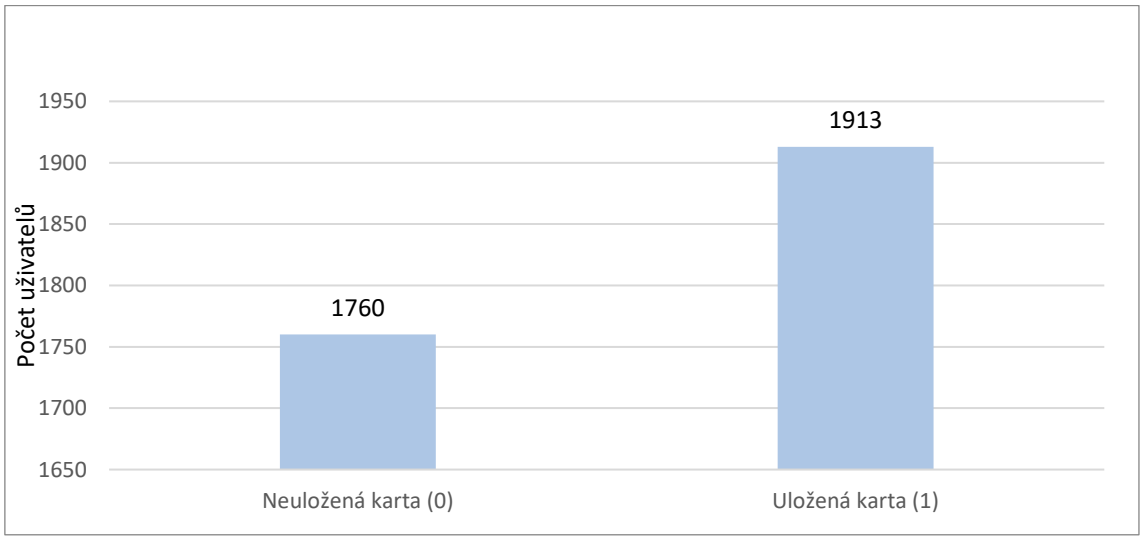
4.2.1.1. User_id

Každý uživatel má v databázi své unikátní číslo. Tento údaj nebyl nijak zahrnut do prediktivních modelů. Zůstal pouze pro jednoduchou identifikaci.

4.2.1.2. Saved_card

Jedná se o kvalitativní alternativní statistický znak, který je představován hodnotou 0 nebo 1. Tento znak představuje, zda uživatel projevil k službě takovou důvěru, že si u nich v platformě uložil svou platební kartu.

Zdroj: Vlastní práce



Graf 1 Saved_card histogram

Počet uživatelů s uloženou a neuloženou kartou v platformě se relativně moc neliší.

4.2.1.3. Count_order

Kvantitativní nespojitý statistický znak představující počet objednávek, které si uživatel objednal v časovém v rozmezí dvou let. Maximální hodnota je 151 objednávek a minimální je 1 objednávka.

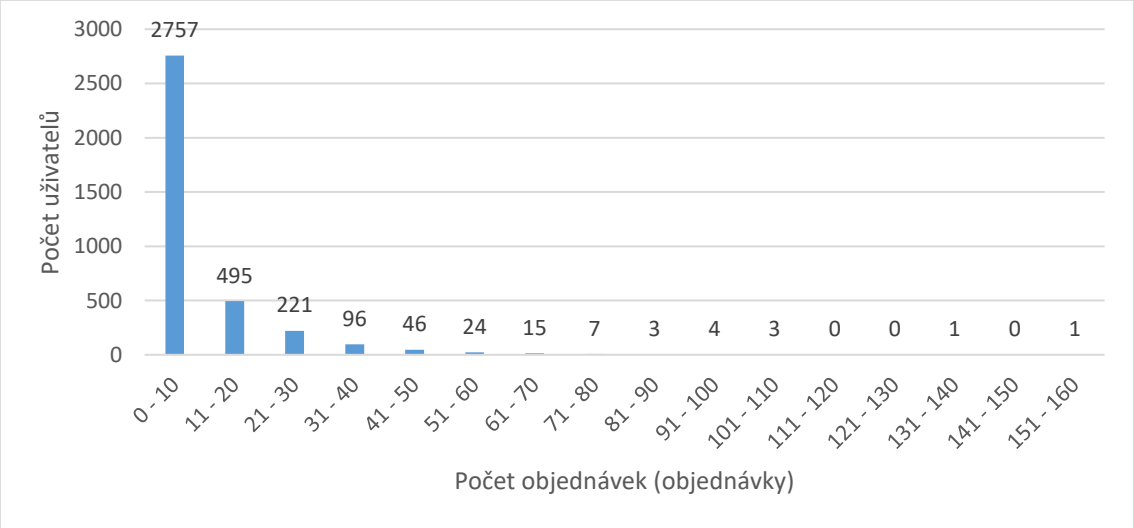
Zdroj: Vlastní práce

význam	Počet objednávek
aritmetický průměr	8,24 objednávek
medián	3,00 objednávek
variační rozpětí	150,00 objednávek
rozptyl	151,85
směrodatná odchylka	12,32
variační koeficient	149,6 %
První kvartil	1 objednávka
Třetí kvartil	10 objednávek
Spodní fous box plot	1
Horní fous box plot	23

Tabulka 2 Count_order popisné charakteristiky

Z uvedené tabulky na první pohled můžeme vyčíst, že má vysokou variabilitu. Z následujícího grafu je na první pohled zřejmé, že největší počet objednávek na uživatele je v intervalu od 1 do 10. Existují zde však extrémy až k hodnotám v intervalu 151 – 160 objednávek.

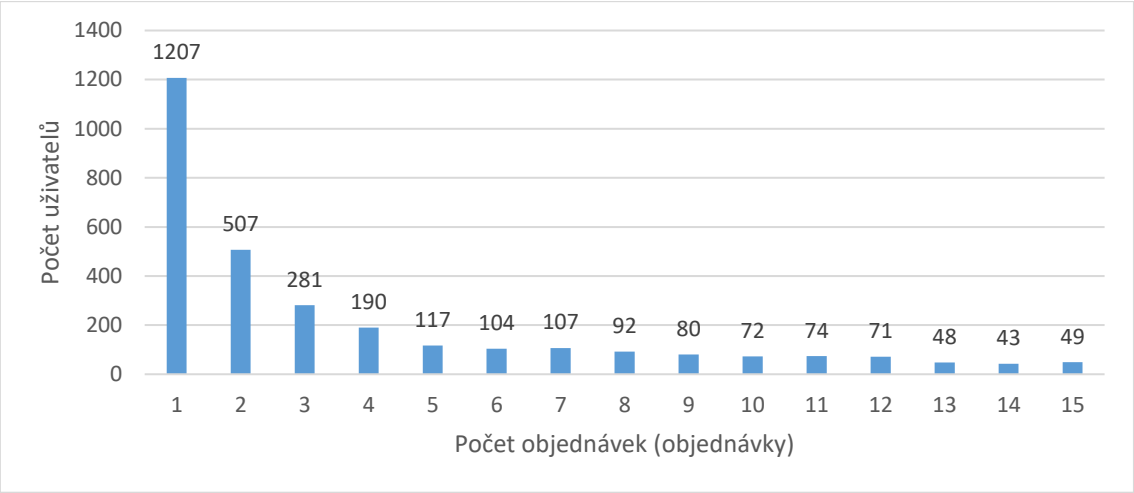
Zdroj: Vlastní práce



Graf 2 Count_order všechny hodnoty

Protože interval 10 nepřesně zobrazuje reálnou četnost počtu objednávek, vytvořil jsem detailnější graf zaměřený na uživatele s počtem objednávek od 1 do 15. Lze si povšimnout, že největší zastoupení má jedna objednávka, což je vlastně skupina uživatelů, kteří si objednali pouze jednou.

Zdroj: Vlastní práce



Graf 3 Count_order detailnější histogram

4.2.1.4. Avg_order_price

Kvantitativní spojitý znak představující aritmetický průměr ceny objednávky za všechny realizované objednávky uživatele v korunách.

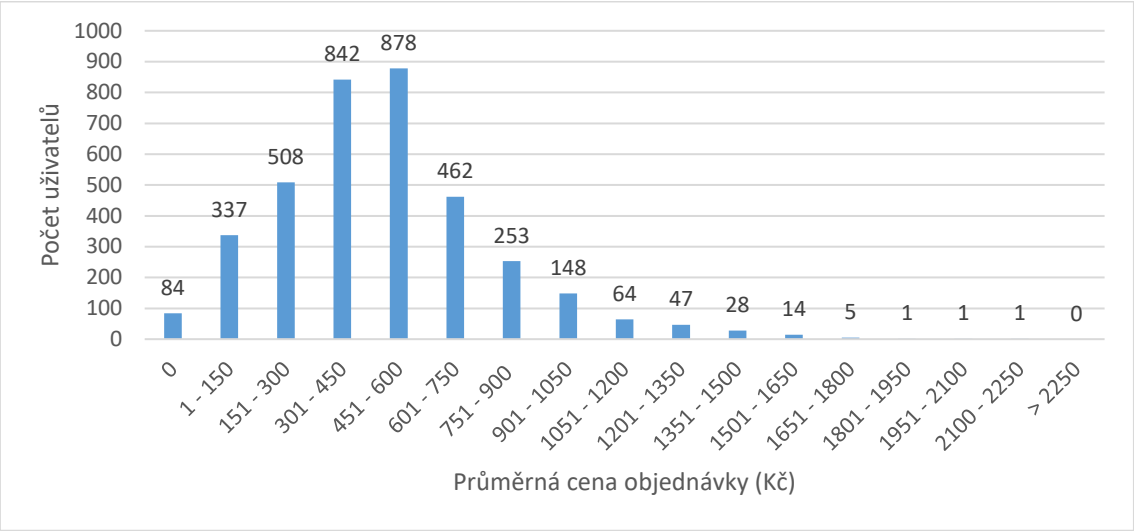
Zdroj: Vlastní práce

význam	Průměrná cena objednávky
aritmetický průměr	487,25 Kč
medián	462,00 Kč
variační rozpětí	2168,00 Kč
rozptyl	83883,91
směrodatná odchylka	289,63
variační koeficient	59,4 %
První kvartil	299 Kč
Třetí kvartil	626 Kč
Spodní fous box plot	0
Horní fous box plot	1113

Tabulka 3 Avg_order_price popisné charakteristiky

Uživatelé si tedy průměrně objednávali službu za 487,25 Kč. Tento průměr může být ovlivněn extrémními hodnotami, které se pohybují nad 1200 Kč. Spolehlivější hodnota než průměr bude medián, která není ovlivněna extrémy. Tento znak má nižší variabilitu, což hezky znázorňuje následující graf. Většina hodnot se pohybuje kolem 150 až 900 Kč.

Zdroj: Vlastní práce



Graf 4 Avg_order_price histogram

4.2.1.5. Count_promo_code

Kvantitativní nespojitý statistický znak, který představuje počet použitých promo kódů za všechny provedené objednávky uživatele. Maximální hodnota je 59 promo kódů a minimální hodnota je 0.

Z tabulky lze vyčíst, že každý uživatel použil 0,76 promo kódu. To je na první pohled nesmysl, ale po prozkoumání grafu je to hned jasnější. Většina uživatelů použila 0 promo kódů, tedy žádný. Díky tomu průměr může vycházet takto na první pohled nelogicky. Medián je v tento moment zbytečný, protože více než polovina hodnot je 0, tak i medián vychází 0. Variabilita tohoto znaku je extrémně vysoká.

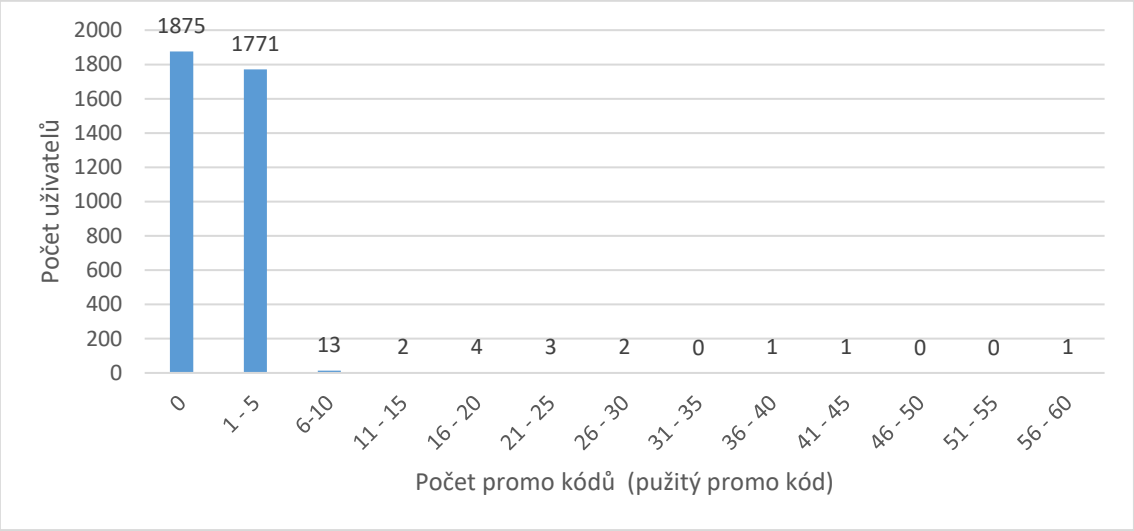
Zdroj: Vlastní práce

význam	Počet použitých promo kódů
aritmetický průměr	0,76 kódů
medián	0,00 kódů
variační rozpětí	59,00 kódů
rozptyl	3,78
směrodatná odchylka	1,95
variační koeficient	255 %
První kvartil	0 kódů
Třetí kvartil	1 kód
Spodní fous box plot	0
Horní fous box plot	2

Tabulka 4 Count_promo_code popisné charakteristiky

V grafu je vidět, že pokud už uživatel použil promo kód, tak ho použil maximálně 5krát. Vyskytly se extrémy. Našel se uživatel, který spadal do intervalu 56 – 60 použitých promo kódů.

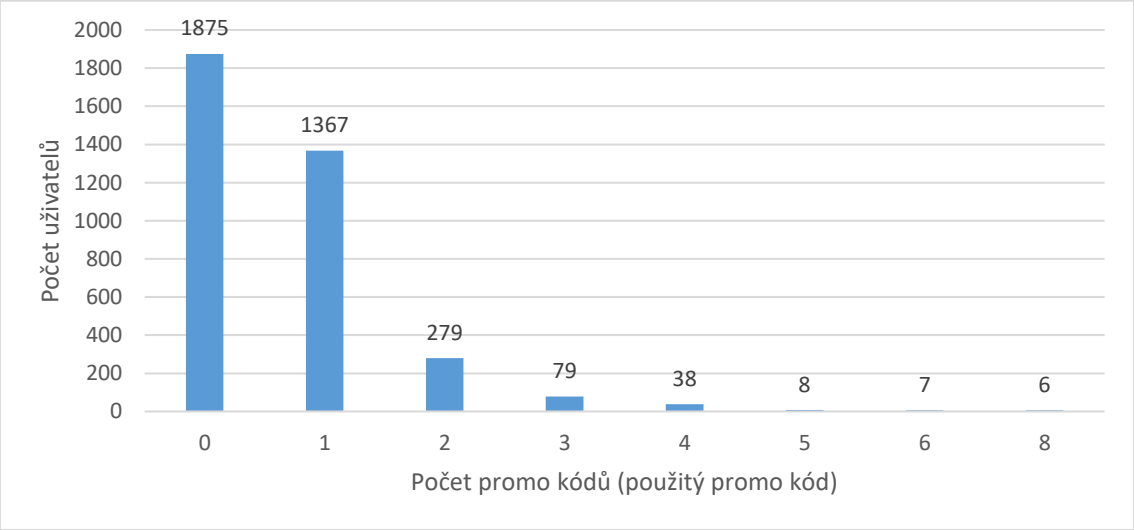
Zdroj: Vlastní práce



Graf 5 Count_promo_code všechny hodnoty

První graf může být trochu zavádějící. Detailnější graf zobrazuje uživatele do 8 použitých promo kódů. Je evidentní, že většina uživatelů, kteří tedy již použili promo kód, ho znovu nepoužívali.

Zdroj: Vlastní práce



Graf 6 Count_promo_code detailnější histogram (do 8 promo kódů)

4.2.1.6. Avg_promo_price

Kvantitativní spojitý znak představující aritmetický průměr za hodnotu promo kódu v korunách, který uživatel použil. Maximální průměrná sleva, kterou uživatel dostal, byla 1901 Kč. Minimální sleva byla 0 Kč.

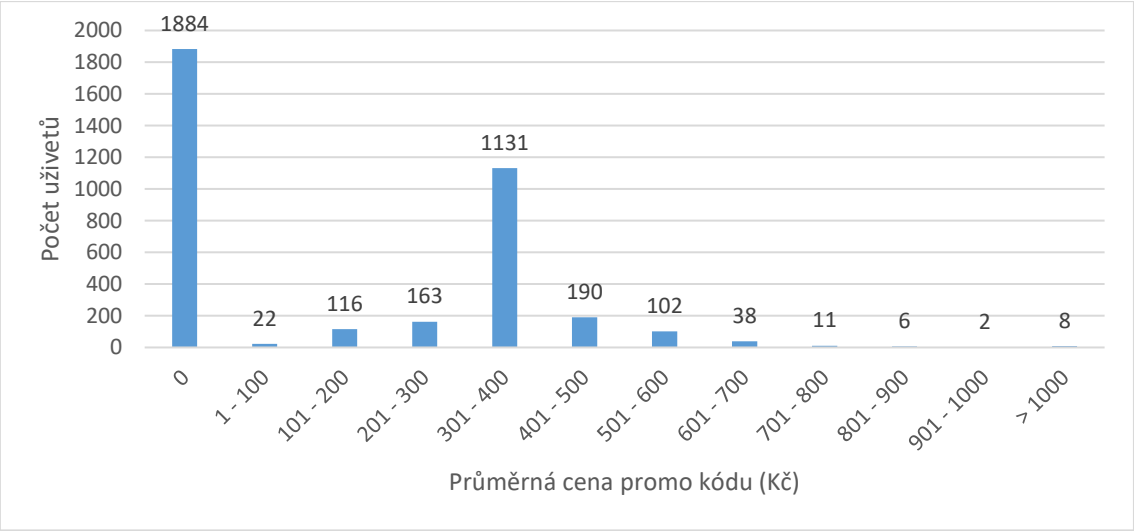
Zdroj: Vlastní práce

význam	Průměrná cena promo kódu
aritmetický průměr	186,95 Kč
medián	0,00 Kč
variační rozpětí	1901,00 Kč
rozptyl	44195,68
směrodatná odchylka	210,23
variační koeficient	112,4 %
První kvartil	0 Kč
Třetí kvartil	400 Kč
Spodní fous box plot	0
Horní fous box plot	993

Tabulka 5 Avg_promo_price popisné charakteristiky

Průměrná cena promo kódu vychází aritmetickým průměrem na 186,95 Kč. Bohužel tato hodnota je ovlivněna všemi uživateli, kteří nepoužily žádný promo kód. Medián taky nevychází reálně. Pokud by se nezapočítaly nulové hodnoty, kterých je podle grafu většina, průměrná cena by se mohla pohybovat okolo 350 Kč. Tento údaj lze jednoduše odvodit z grafu.

Zdroj: Vlastní práce



Graf 7 Avg_promo_price histogram

4.2.1.7. Avg_base_length

Kvantitativní spojitý statistický znak představující aritmetický průměr doby úklidu v hodinách za všechny objednávky uživatele. Maximální průměrná doba úklidu je 11 hodin. Minimální doba úklidu je 1 hodina.

Tento statistický znak má malou variabilitu, což lze vnímat kladně. To se odráží i na aritmetickém průměru a mediánu, jelikož se jejich hodnoty skoro neliší. Dá se předpokládat, že se zde nevyskytují žádné extrémy.

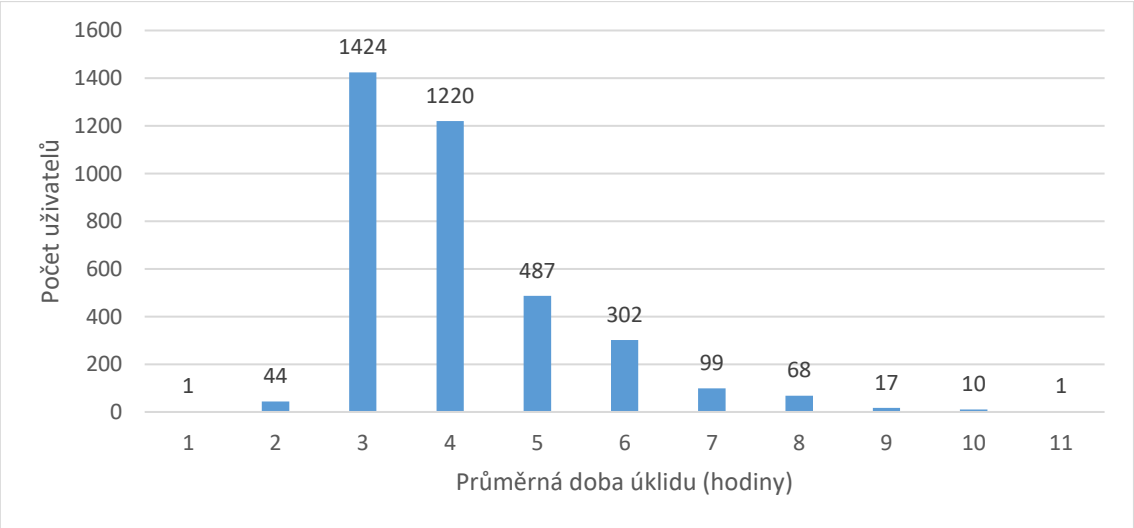
Zdroj: Vlastní práce

význam	Průměrná doba služby
aritmetický průměr	4,08 h
medián	4,00 h
variační rozpětí	10,00 h
rozptyl	1,66
směrodatná odchylka	1,29
variační koeficient	31,561
První kvartil	3 h
Třetí kvartil	5 h
Spodní fous box plot	1
Horní fous box plot	8

Tabulka 6 Avg_base_length popisné charakteristiky

Z grafu je na první pohled jasné, že se služba objednává na větší a náročnější úklidy od 3 a více hodin. Krátké úklidy do 2 hodin nebo extrémně dlouhé úklidy se vyskytují minimálně.

Zdroj: Vlastní práce



Graf 8 Avg_base_length histogram

4.2.1.8. Avg_margin_length

Kvantitativní spojitý statistický znak představující aritmetický průměr prodloužení doby úklidu oproti předpokládané době. Hodnoty jsou uvedeny v hodinách. Maximální prodloužení je 1 hodina a minimální prodloužení je 0.2 hodiny. Tento znak jsem použil, protože mění předpokládanou cenu objednávky až na místě. Tento faktor může v zákazníkovi vyvolat negativní emoci, způsobenou pocitem, že dáma na úklid pracuje příliš pomalu a zákazník za to musí připlácet.

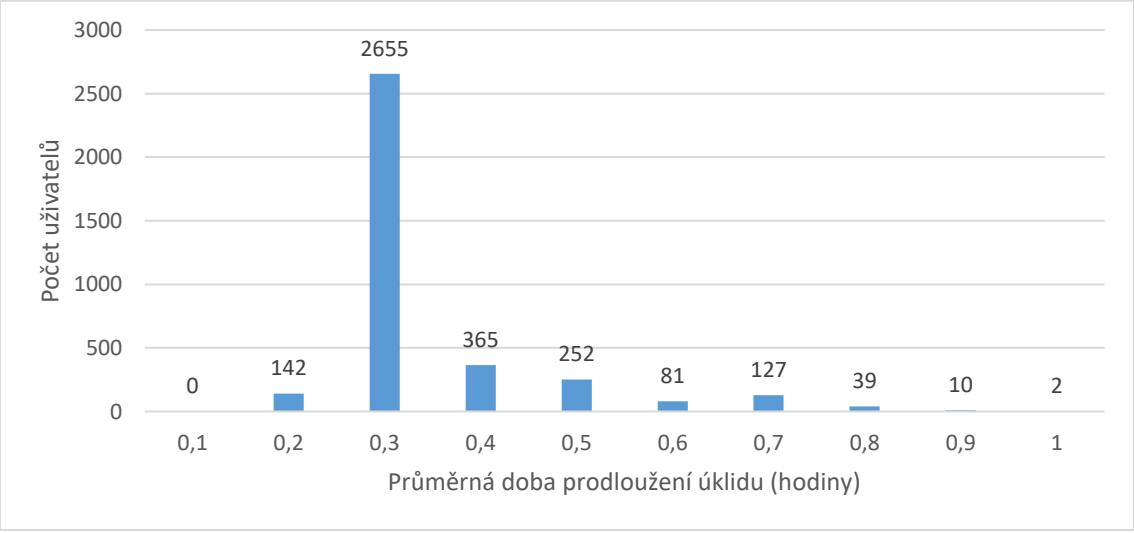
Zdroj: Vlastní práce

význam	Průměrná délka prodloužení objednávky
aritmetický průměr	0,35 h
medián	0,30 h
variační rozpětí	0,80 h
rozptyl	0,01
směrodatná odchylka	0,12
variační koeficient	33,2 %

Tabulka 7 Avg_margin_length popisné charakteristiky

Znak o prodlužování úklidů má podobné vlastnosti, jako znak o standartní délce úklidů. Lze si povšimnout, že ani jeden uživatel nemá nulovou hodnotu. Znamená to tedy, že vždy se úklid o trochu protáhne a většinou to bývá 0,3 hodiny.

Zdroj: Vlastní práce



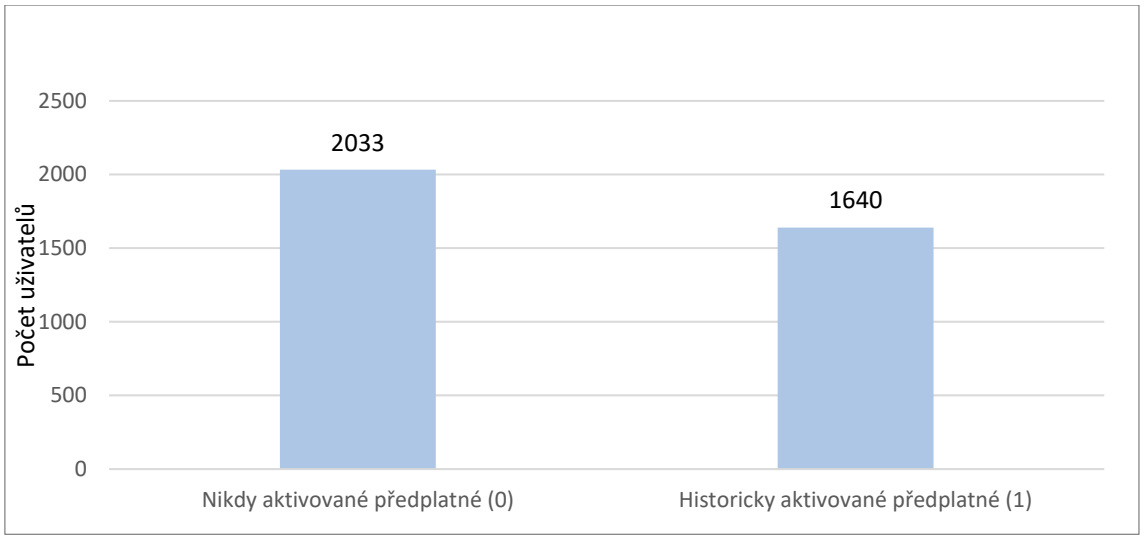
Graf 9 Avg_margin_length histogram

4.2.1.9. Order_subscription

Je to kvalitativní alternativní statistický znak, který představuje historicky, zda uživatel měl někdy aktivní předplatné. Předplatné funguje tak, že v určitých intervalech se uklízení opakuje. Hodnota 1 znamená, že někdy již měl aktivní předplatné, ale nemusí ho mít stále aktivní. Hodnota 0 znamená, že nikdy si předplatné neaktivoval.

Z grafu je jasné, že méně uživatelů využívá předplatného.

Zdroj: Vlastní práce

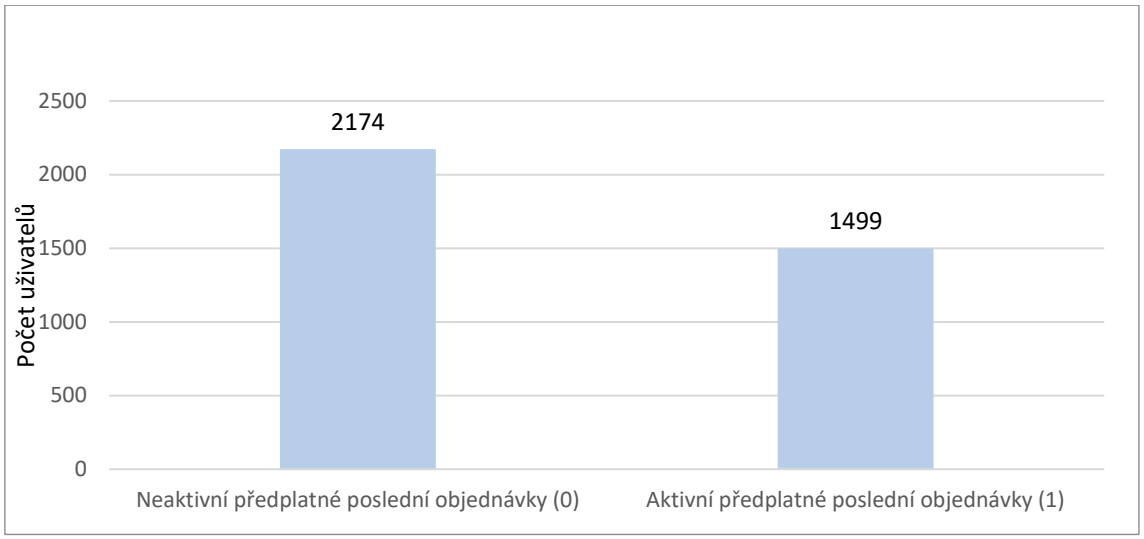


Graf 10 Order_subscription histogram

4.2.1.10. Last_order_active_subsription

Tato hodnota je také kvalitativní alternativní statistický znak, který na rozdíl od order_subscription nepředstavuje historický údaj, ale aktuální stav předplatného k provedené objednávce úklidu. Právě z hodnot order_subscription a last_order_active_subcription se jednoduše dá vyvodit, že si předplatné zrušil. V případě že hodnota obsahuje 0, znamená to, že poslední objednávka byla objednána skrz předplatné. V opačném případě poslední objednávka uživatele nebyla objednána skrz předplatné.

Zdroj: Vlastní práce



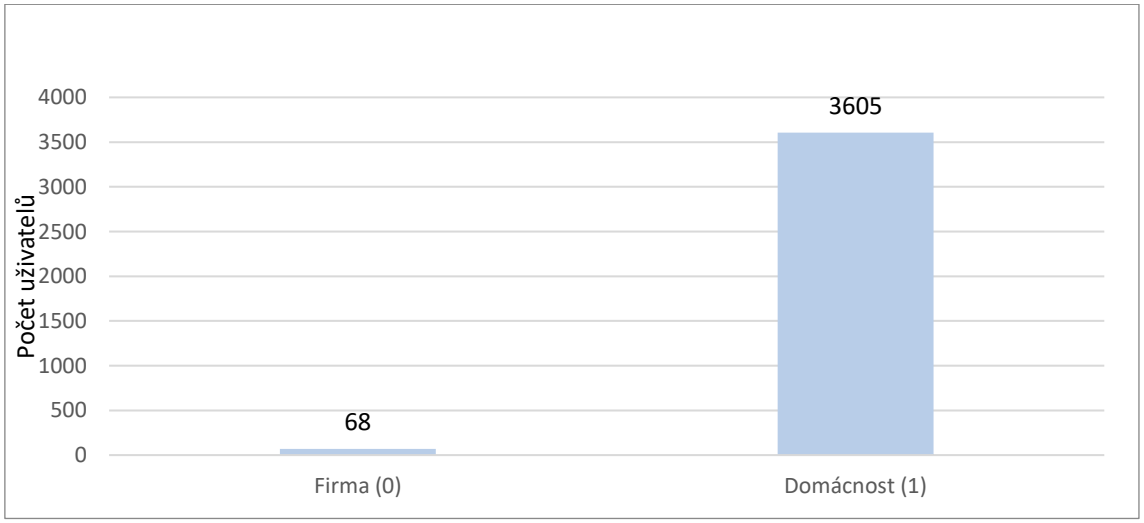
Graf 11 Last_order_active_subscription histogram

Z grafu lze vyčíst podobné hodnoty jako u znaku order_subscription. Aktivní předplatné u poslední objednávky je o trochu nižší než historická hodnota.

4.2.1.11. Subject

Subject je kvalitativní alternativní statistický znak. Reprezentuje typ uživatele, jestli je to klasická domácnost, nebo se jedná o firmu. Hodnota 1 je domácnost, tedy fyzická osoba. Hodnota 0 je právnická osoba. Chování mezi těmito subjekty se liší. Předpokládal jsem, že firmy si budou objednávat úklid častěji a budou platit větší částky.
Graf jednoznačně ukazuje, že drtivá většina uživatelů jsou soukromé osoby poptávající úklid v domácnosti. Pouze pár subjektů jsou firmy.

Zdroj: Vlastní práce

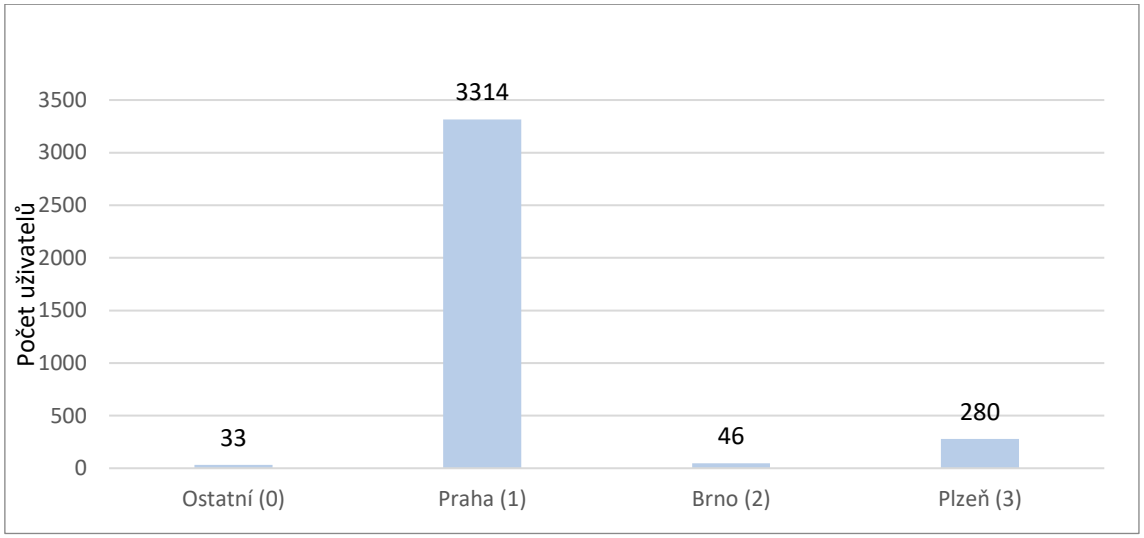


Graf 12 Subject histogram

4.2.1.12. City

Kvalitativní množinný statistický znak reprezentující město, ve kterém objednávka proběhla. Hodnota může nabývat hodnot 1,2,3 a 0, což odpovídá městům Praha, Brno, Plzeň a ostatní. Odpovídá to velkým městům, na které se firma soustředí. Hodnota 0 (Ostatní) může představovat objednávky realizované v okolí jmenovaných měst, nebo i jiné lokace.

Zdroj: Vlastní práce



Graf 13 City histogram

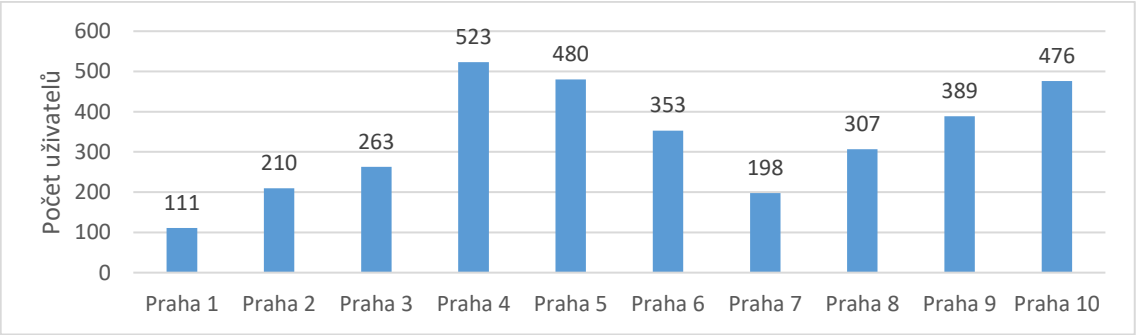
Jednoznačně se úklidy objednávají v Praze oproti ostatním městům.

4.2.1.13. Zipcode

Kvalitativní množinný statistický znak je doplňkový údaj k znaku City. V rámci měst lze odlišit jednotlivé čtvrti. U hodnoty 0 u znaku City lze díky Zipcode dohledat přesněji, o kterou lokaci se jedná.

Jelikož se služba úklidu objednává hlavně v Praze, použil jsem zipcode pro detailnější zmapování Prahy. Zipcode jsem sjednotil na městské části Praha 1 až Praha 10.

Zdroj: Vlastní práce



Graf 14 Zipcode histogram (pouze Praha)

Rozhodně nelze říct, že by Praha objednávala úklid rovnoměrně. Nejvíce se objednává ve čtvrtích s vyšším počtem obyvatel. Tedy Praha 4, Praha 5 a Praha 10. Naopak méně často se úklid objednává do čtvrtí v centru Prahy, kde jsou spíše kanceláře a lidé tolik nebydlí.

4.2.1.14. Avg_customer_rating

Kvantitativní spojitý statistický znak. Představuje aritmetický průměr hodnocení za všechny uživatelovy objednávky. Může nabývat hodnot od 1 až 5. Uživatelé hodnotí pomocí hvězdiček. Takže hodnota 1 je nejhorší a hodnota 5 je nejlepší, tedy naprostá spokojenost se službou.

Zdroj: Vlastní práce

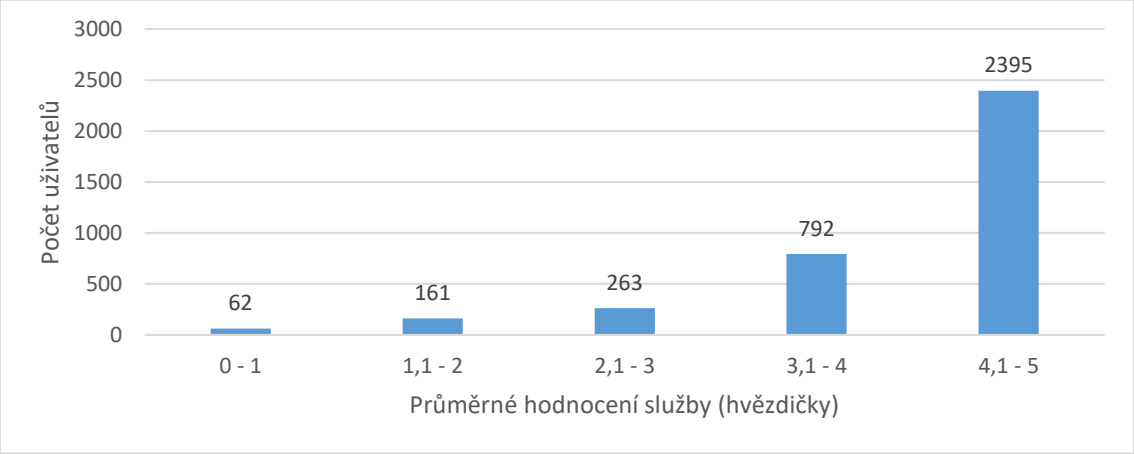
význam	Počet udělených hvězdiček
aritmetický průměr	4,33 hvězdiček
medián	4,82 hvězdiček
variační rozpětí	4,00 hvězdiček
rozptyl	0,85
směrodatná odchylka	0,92
variační koeficient	21,3 %

Tabulka 8 Avg_customer_rating popisné charakteristiky

Variabilita průměrného hodnocení uživatele je nízká. Aritmetický průměr a medián se nijak zásadně neliší, protože se nevyskytují žádné extrémy. Vychází to z podstaty znaku, který může nabývat pouze 5 hodnot.

Z grafu lze vyčíst, že jsou uživatelé většinou spokojeni a nespokojenost má sestupný trend.

Zdroj: Vlastní práce



Graf 15 Avg_customer_rating histogram

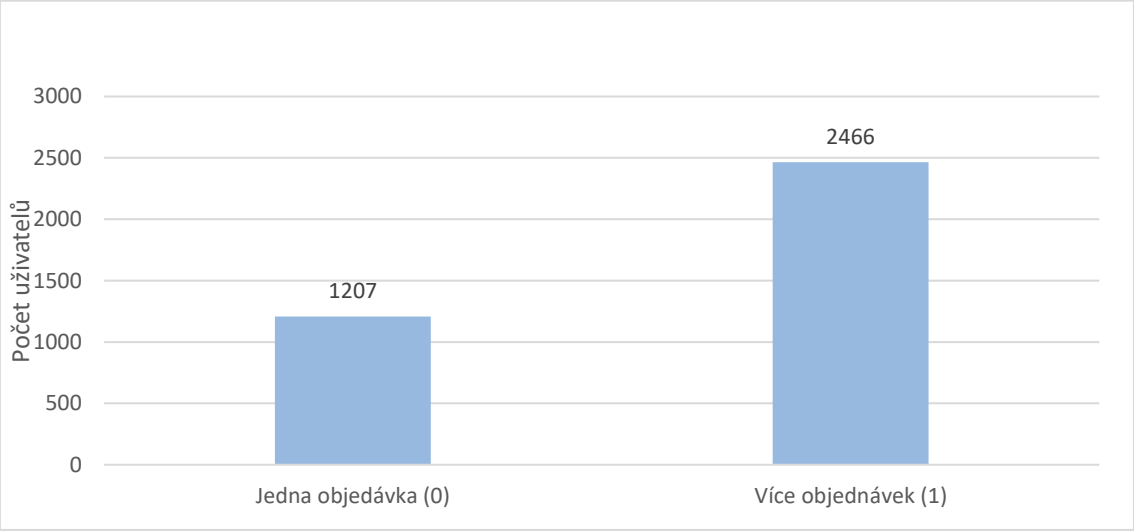
4.2.2. Charakteristika modelované proměnné

4.2.2.1. Re_order

Je to kvalitativní alternativní statistický znak a zároveň prediktivní proměnná pro celý model. Je představován pouze hodnotou 0 nebo 1. Představuje informaci, zda si uživatel v minulosti objednal službu pouze jednou (0) nebo víckrát (1). Pokud službu objednal víckrát, spadá do skupiny uživatelů, kteří provedli opětovné objednání služby.

Tento znak vznikl z počtu objednávek uživatele. Pokud uživatel měl pouze jednu objednávku, ve znaku re_order mu byla přiřazena 0. Pokud měl historicky objednáno více jak jednu objednávku, byla mu u znaku re_order přiřazena 1.

Zdroj: Vlastní práce



Graf 16 Re_order histogram

V grafu je jasně vidět dvojnásobný počet uživatelů, kteří provedli opětovné objednání služby

4.2.3. Znaky bez extrémů

U některých statistických znaků se vyskytuje mnoho odlehlých až extrémních hodnot. Lze to jednoduše poznat v rozdílných hodnotách průměru a mediánu. Tyto hodnoty by mohli zásadně ovlivnit výpočet prediktivních modelů. Bylo zapotřebí výběrovou databázi očistit u kvalitativních znaků od odlehlých či extrémních hodnot. U každého znaku, který byl očištěn od odlehlých nebo extrémních hodnot, se zmenšil rozdíl mezi průměrem a mediánem. Dále se snížil variační koeficient daného statistického znaku.

Původní výběrový statistický soubor obsahoval záznam o 3673 uživatelích. Po očištění všech odlehlých a extrémních hodnot se výběrový statistický soubor zmenšil o 621 uživatelů na 3052 uživatelů. To je přibližně 17% záznamů.

4.2.3.1. Count_order

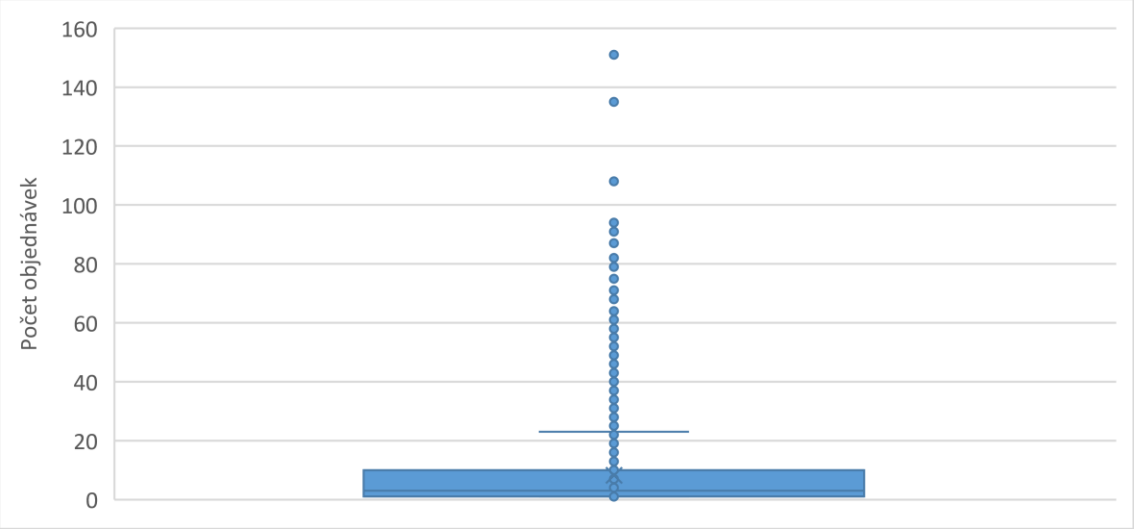
Podle box-plot grafu všechny hodnoty větší než konec horního fousu jsou odlehlé nebo extrémní. Celkem to představuje 333 uživatelů, kteří mají počet objednávek větší než 23.

Zdroj: Vlastní práce

význam	Počet objednávek
aritmetický průměr	5,15 objednávek
medián	2,00 objednávky
variační rozpětí	22,00 objednávek
rozptyl	31,91
směrodatná odchylka	5,65
variační koeficient	109,6 %

Tabulka 9 Count_order popisné charakteristiky bez odlehlých a extrémních hodnot

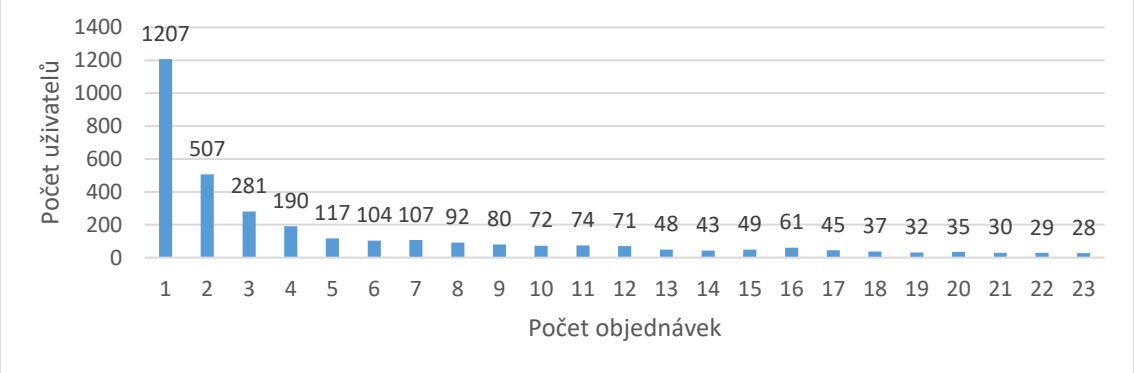
Zdroj: Vlastní práce



Graf 17 Count_order box-plot

Výběrová databáze byla očištěna o 333 hodnot s hodnotou větší než 23. Výsledný histogram znaku count_order vypadá následovně.

Zdroj: Vlastní práce



Graf 18 Count_order histogram bez odlehlých a extrémních hodnot

4.2.3.2. Avg_order_price

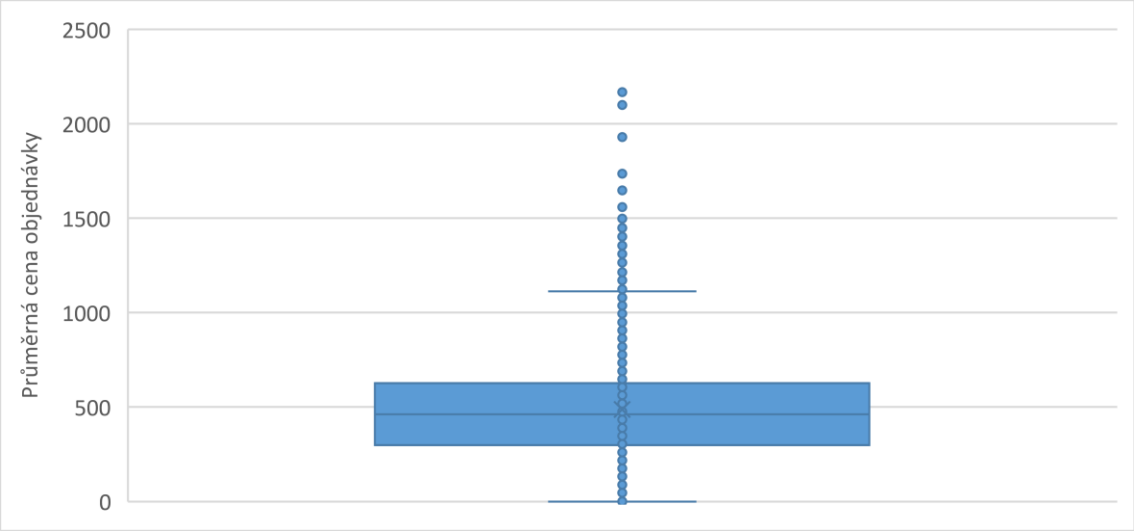
Podle grafu box plot existuje 126 uživatelů, kteří mají průměrnou objednávku větší než 1113 Kč, což lze považovat za odlehlou nebo extrémní hodnotu.

Zdroj: Vlastní práce

význam	Průměrná cena objednávky
aritmetický průměr	456,88 Kč
medián	451,00 Kč
variační rozpětí	1113,00 Kč
rozptyl	58915,49
směrodatná odchylka	242,73
variační koeficient	53,1 %

Tabulka 10 Avg_order_price popisné charakteristiky bez odlehlých a extrémních hodnot

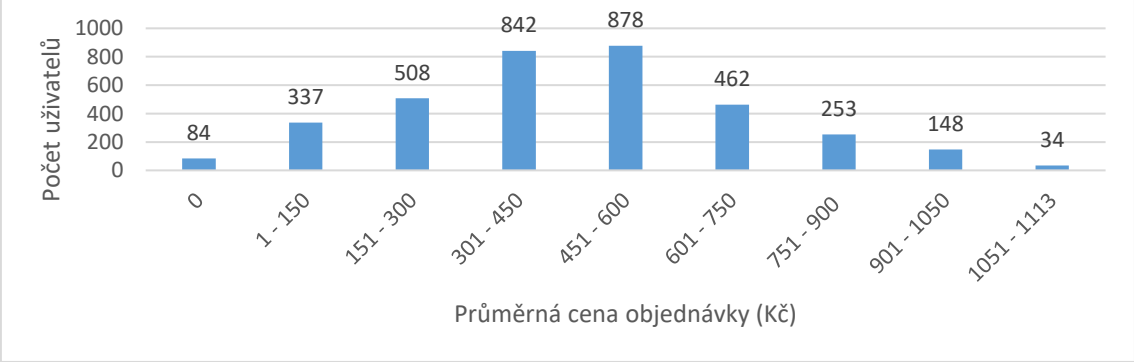
Zdroj: Vlastní práce



Graf 19 Avg_order_price box-plot

Výběrová databáze byla očištěna o 126 uživatelů a histogram statistického znaku avg_order_price vypadá následovně.

Zdroj: Vlastní práce



Graf 20 Avg_order_price histogram bez odlehlých a extrémních hodnot

4.2.3.3. Count_promo_code

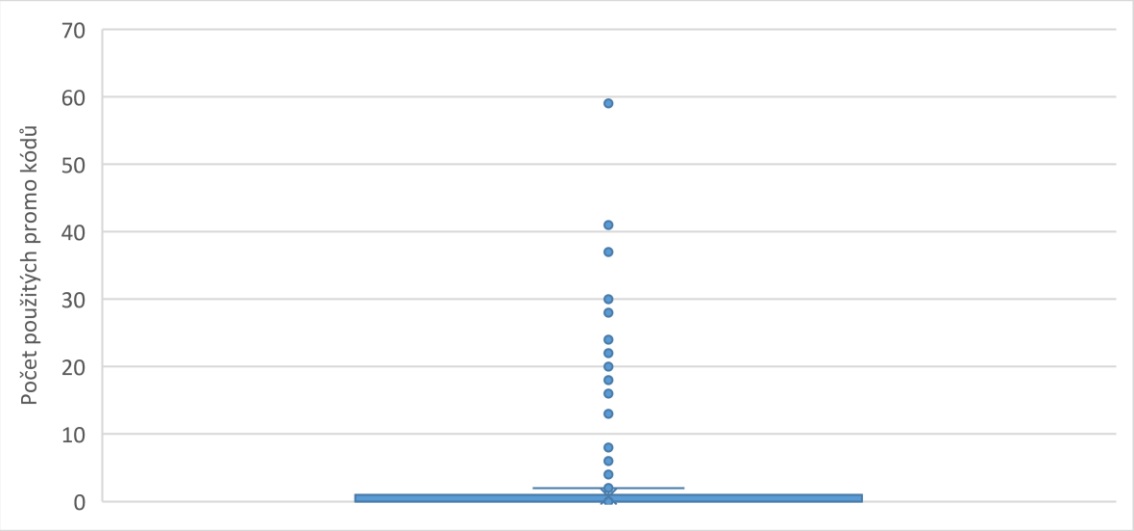
Podle box plot grafu je 151 uživatelů, kteří mají počet použitých promo kódů větší než 2. Všechny hodnoty větší než 2 lze považovat za odlehlé nebo extrémní.

Zdroj: Vlastní práce

význam	Počet použitých promo kódů
aritmetický průměr	0,55 kódu
medián	0,00 kódů
variační rozpětí	2,00 kódy
rozptyl	0,41
směrodatná odchylka	0,64
variační koeficient	116,6 %

Tabulka 11 Count_promo_code popisné charakteristiky bez odlehlých a extrémních hodnot

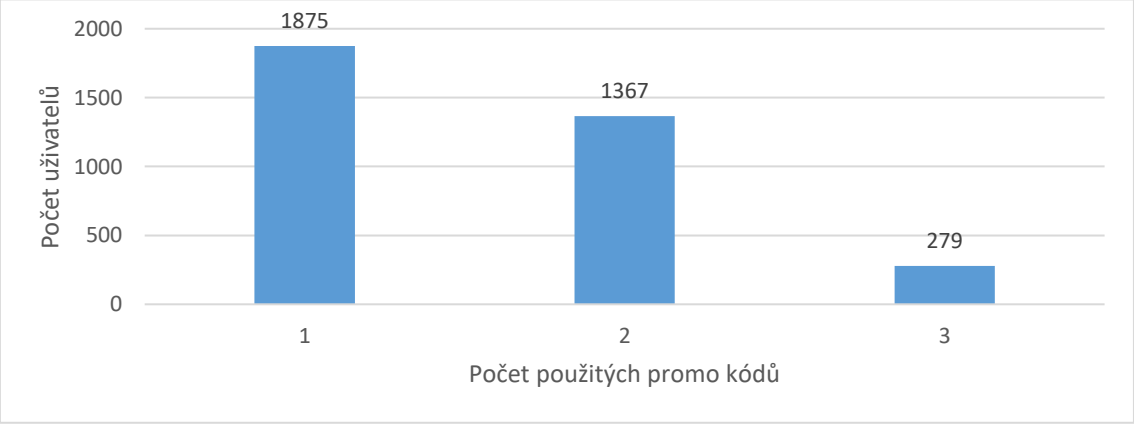
Zdroj: Vlastní práce



Graf 21 Count_promo_code box-plot

Výběrová databáze byla očištěna o 151 uživatelů, kteří měli statistický znak count_promo_code větší než 2. Histogram tohoto statistického znaku vypadá následovně.

Zdroj: Vlastní práce



Graf 22 Count_promo_code histogram bez odlehlých a extrémních hodnot

4.2.3.4. Avg_promo_price

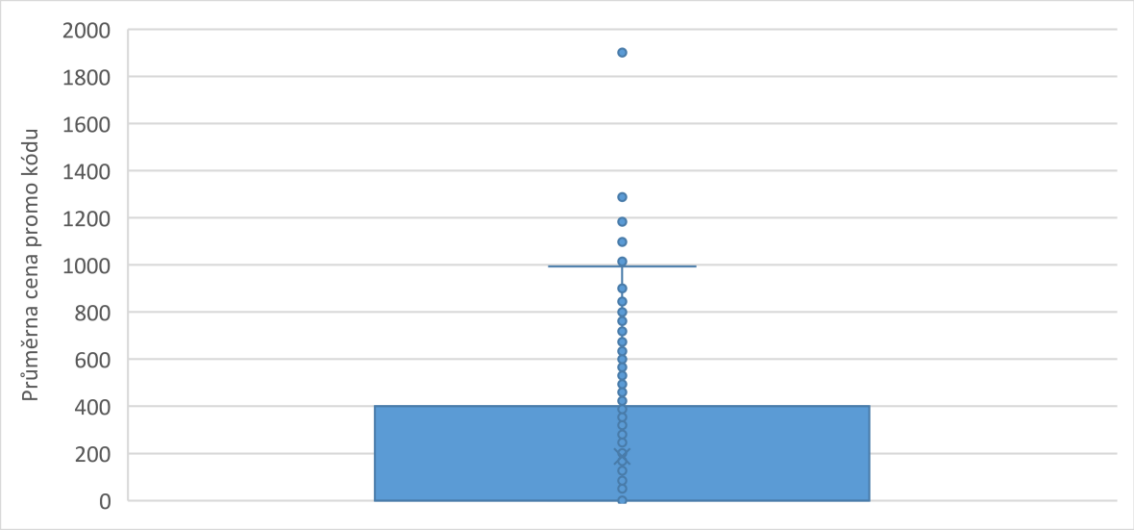
Podle box plot grafu má 8 uživatelů průměrnou cenu promo kódu větší než 933 Kč. Tyto hodnoty jsou považovány jako odlehlé nebo extrémní.

Zdroj: Vlastní práce

význam	Průměrná cena promo kódu
aritmetický průměr	184,57 Kč
medián	0,00 Kč
variační rozpětí	993,00 Kč
rozptyl	41551,12
směrodatná odchylka	203,84
variační koeficient	110,4 %

Tabulka 12 Avg_promo_price popisné charakteristiky bez odlehlých a extrémních hodnot

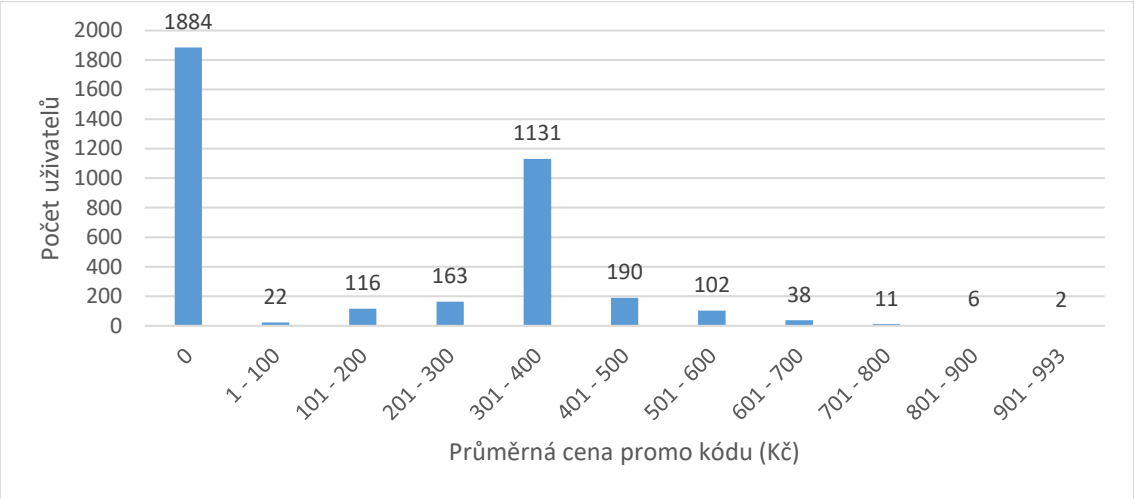
Zdroj: Vlastní práce



Graf 23 Avg_promo_price box-plot

Výběrová databáze byla očištěna o 8 uživatelů, u kterých statistický znak Avg_promo_price vykazoval známky odlehlých nebo extrémních hodnot. Na histogramu se očištění výrazně neprojevilo.

Zdroj: Vlastní práce



Graf 24 Avg_promo_price histogram bez odlehlých a extrémních hodnot

4.2.3.5. Avg_base_length

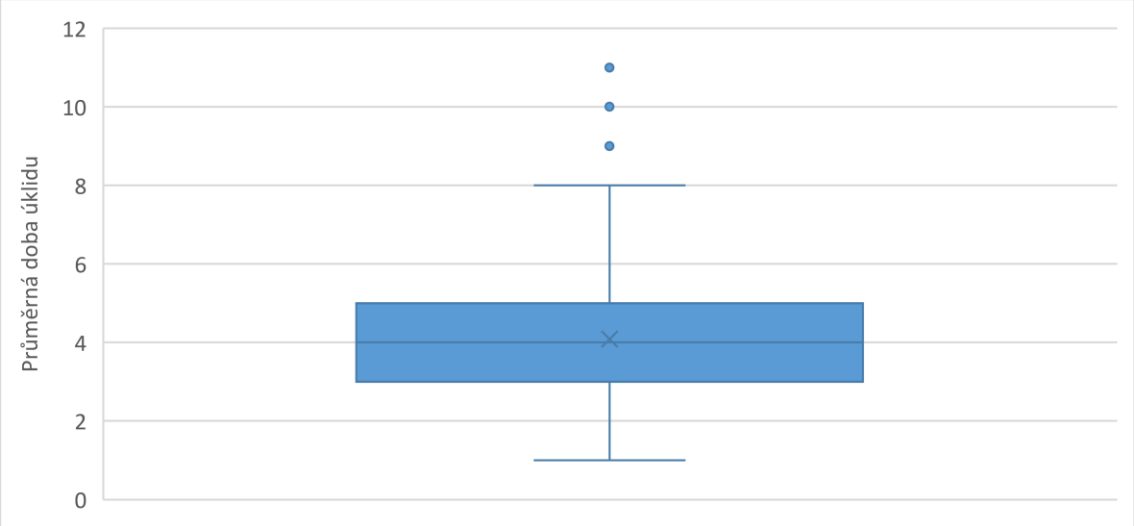
Podle box plot grafu vyšla u 27 uživatelů průměrná doba úklidu jako odlehlá nebo extrémní hodnota. Odlehlé nebo extrémní hodnoty jsou menší než 1 nebo větší než 8.

Zdroj: Vlastní práce

význam	Průměrná doba služby
aritmetický průměr	4,04 h
medián	4,00 h
variační rozpětí	7,00 h
rozptyl	1,45
směrodatná odchylka	1,20
variační koeficient	29,8 %

Tabulka 13 Avg_base_length popisné charakteristiky bez odlehlých a extrémních hodnot

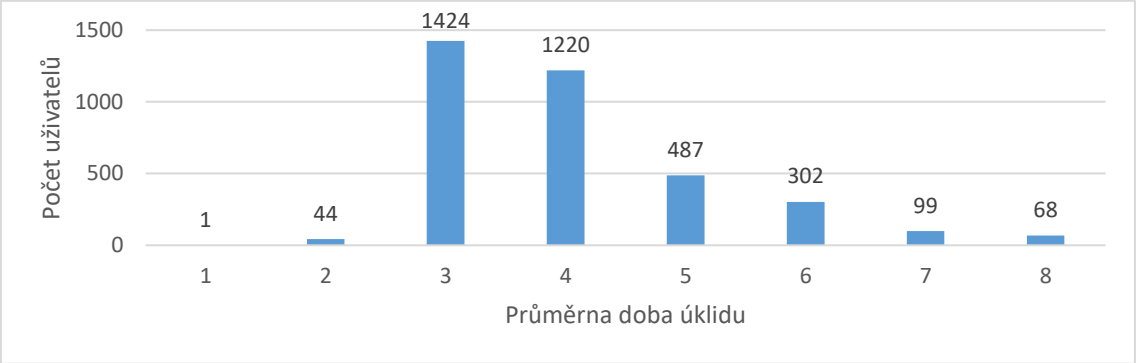
Zdroj: Vlastní práce



Graf 25 Avg_base_length box-plot

Výběrová databáze byla očištěna o 27 uživatelů, kteří měli statistický znak Avg_base_length menší než 1 nebo větší než 8. Opět se to na histogramu průměrné délky úklidu výrazně neprojevilo.

Zdroj: Vlastní práce



Graf 26 Avg_base_length histogram bez odlehlých a extrémních hodnot

4.3. Prediktivní modely

Pro prediktivní modely byl použit nástroj Rapid Miner, ve kterém jsem nastavil predikování proměnné Re_order pomocí rozhodovacích stromů a logistické regrese.

Oba modely mi vytvořily 3 nové sloupce.

- Prediction (Re_order)
- Confidence (1)
- Confidence (0)

4.3.1.1. Prediction (Re_order)

Prediktivní model spočítal pro každého uživatele v databázi hodnotu nabývající hodnot 0 nebo 1. U většiny uživatelů se tato hodnota rovnala hodnotě ve sloupci Re_order. Úspěšnost predikce je rozebrána u jednotlivých prediktivních modelů rozhodovacích stromů a logistické regrese.

4.3.1.2. Confidence (1)

Prediktivní model spočítal pravděpodobnost opětovného objednání služby uživatelem. Confidence (1) nabývá hodnot mezi 0 a 1. Čím vyšší číslo, tím větší pravděpodobnost opětovného objednání služby. Hodnota 1 znamená 100% pravděpodobnost. Pokud je pravděpodobnost menší než 0,5, tak se většinou hodnota Prediction Re_order změní na 0.

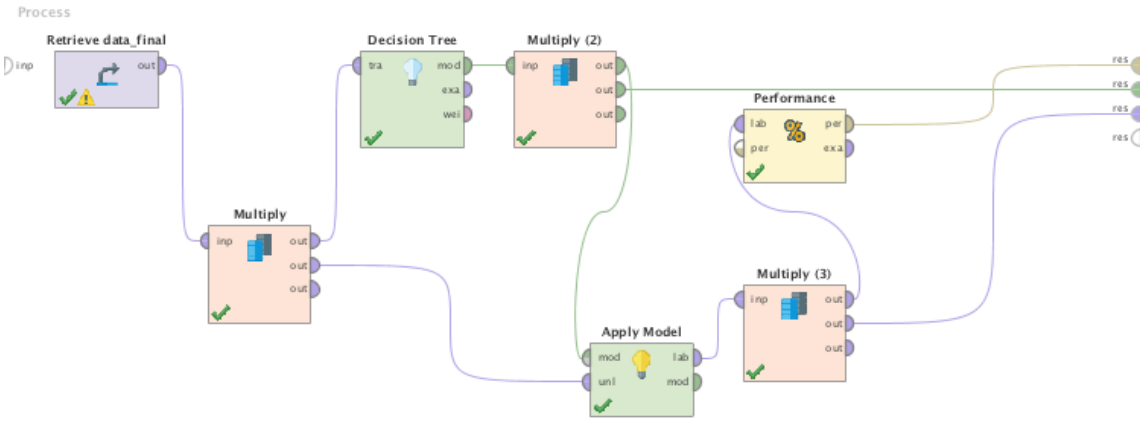
4.3.1.3. Confidence (0)

Prediktivní model spočítal pravděpodobnost opětovného neobjednání služby uživatelem. Confidence (0) také nabývá hodnot mezi 0 a 1. Vždy se tedy součet Confidence (0) a Confidence (1) rovná 1, tedy 100% pravděpodobnosti. Když hodnota Confidence(0) je větší než 0,5, hodnota Prediction Re_order vychází 0. Trend Prediction Re_order se mění z 0 na 1 přibližně tehdy, když Confidence (0) začíná být větší než Confidence (1).

4.3.2. Rozhodovací stromy

V programu Rapid Miner byl nastaven prediktivní model rozhodovacího stromu s C4.5 algoritmem.

Zdroj: Vlastní práce

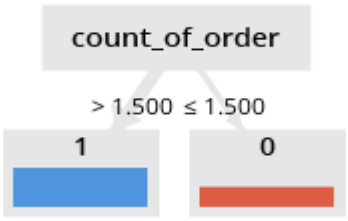


Obrázek 7 Prediktivní proces Rozhodovací strom

- První funkce Retrieve data_final načetla data z nahrané databáze
- Funkce Multiply poslala data oběma výstupy v nezměněné podobě
- Funkce Decision Tree vytvořila rozhodovací strom
- Funkce Apply Model na vstupu získal původní data a prediktivní model rozhodovacího stromu, model použil na původní data.
- Funkce Performance vytvořila tabulku úspěšnosti prediktivního modelu
- Proces vrátil 3 výstupy
    - Tabulku s predikovanou hodnotou Prediction Re_order, Confidence 0 a Confidence 1
    - Grafické znázornění rozhodovacího stromu
    - Tabulku úspěšnosti prediktivního modelu

První výstup rozhodovacího stromu se nerozvětvil a měl 100% úspěšnost. Nastalo to proto, že predikovaná proměnná Re_order vycházela ze statistického znaku Count_order.

Zdroj: Vlastní práce



Graf 27 Rozhodovací strom s Count_order

Z tabulky lze snadno vyčíst, že se model nemýlil ani v jednom případě. Všude, kde byl počet objednávek roven 1, predikoval opětovné neobjednání služby. Tam, kde byl počet objednávek větší než 1, vždy predikoval znovu objednání služby.

Zdroj: Vlastní práce

accuracy: 100.00%

	true 1	true 0	class precision
pred. 1	2466	0	100.00%
pred. 0	0	1207	100.00%
class recall	100.00%	100.00%	

Tabulka 14 Úspěšnost rozhodovacího stromu s Count_order

Z toho důvodu jsem hodnotu Count_order vyřadil a dále jsem s ní nepočítal. Rozhodovací strom bez znaku Count_order vyšel již detailněji a rozvětvil se. Modrá barva představuje pozitivní výsledek Re_order = 1 a červená barva negativní výsledek Re_order = 0. Výška barevného sloupce představuje procentuální zastoupení počtu vyhodnocených uživatelů v daném listu stromu.

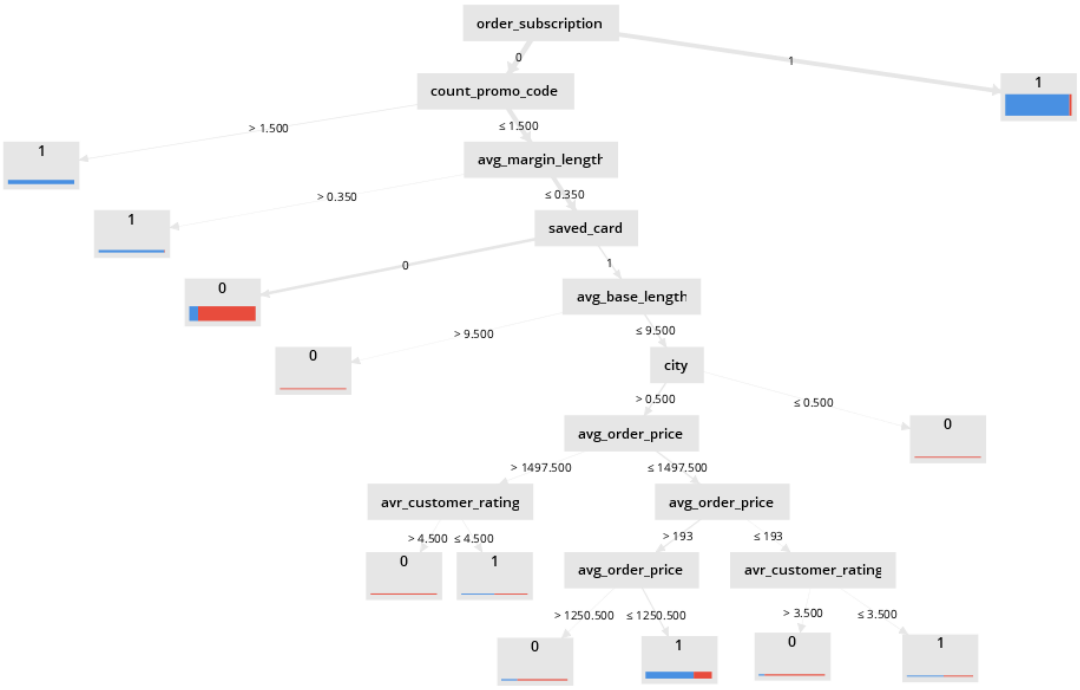
4.3.2.1. Rozhodovací strom s odlehlými a extrémními hodnotami

Na první pohled je vidět, že nejlepší prediktor je Order_subscription. Když uživatel má aktivní předplatné, tak je velice pravděpodobné, že provede znovu objednávku. Celkově 44,65% všech záznamů skončilo v tomto listu stromu. 1556 uživatelů s kladnou hodnotou (1, modrá) a pouze 54 uživatelů se zápornou hodnotou (0, červená).

Naopak uživatelé, kteří nemají uloženou kartu (Saved_card), často neudělají opětovné objednání služby. V tomto listu stromu skončilo 29,76 % uživatelů. 948 uživatelů s neuloženou kartou již neprovedlo opětovné objednání služby. Pouze 145 uživatelů s neuloženou kartou opětovně službu objednalo.

Uživatelé, kteří mají hodnotu průměrné ceny objednávky mezi 193 Kč a 1250 Kč, mají uloženou kartu, průměrně se jim úklid neprodlužuje o víc jak 0.35 hodiny a použili méně než 1,5 promo kódu, tak velice často opakují objednávku. V tomto listu skončilo 12,74% uživatelů. Celkem 340 uživatelů provedlo opětovné objednání služby. Pouze 120 uživatelů si službu znovu již neobjednalo.

Zdroj: Vlastní práce



Graf 28 Rozhodovací strom s extrémy

Správně predikovaných proměnných je 90,63 %. Přesněji 2310 uživatelů bylo správně predikováno, že znovu službu objednají. Pouze 156 uživatelů bylo chybně predikováno, že službu znovu objednají.

Naopak 1019 uživatelů bylo správně predikováno, že službu již neobjednají a pouze 188 uživatelů bylo chybně predikováno, že službu neobjednají znovu.

Zdroj: Vlastní práce

accuracy: 90.63%

	true 1	true 0	class precision
pred. 1	2310	188	92.47%
pred. 0	156	1019	86.72%
class recall	93.67%	84.42%	

Tabulka 15 Úspěšnost rozhodovacího stromu s extrémy

Úspěšnost 90,63 % je podezřele vysoká. Bylo zapotřebí odfiltrovat znaky, které model příliš ovlivňují.

4.3.2.2. Rozhodovací strom bez odlehlých a extrémních hodnot

Data bez odlehlých a externích hodnot a bez znaků order_subscription, last_order_active_subscription a saved_card příliš ovlivňovaly model. Červená barva znamená opětovné objednání služby (re_order = 1) a modrá barva, že uživatel službu znovu již neobjednal (re_order = 0). Nyní je rozhodovací strom mnohem méně rozvětvený, ale přesněji identifikuje významné znaky.

Celkem 57,47 % uživatelů si službu objednalo znovu, pokud měli průměrnou cenu objednávky mezi 279 Kč a 1105 Kč a zároveň počet použitých promo kódů nepřesahuje 1,5 a prodloužení úklidu nebylo delší než 0,35 hodiny.

Zdroj: Vlastní práce

Graf 29 Rozhodovací strom bez extrémů

Zdroj: Vlastní práce

accuracy: 76.25%

	true 0	true 1	class precision
pred. 0	439	54	89.05%
pred. 1	671	1888	73.78%
class recall	39.55%	97.22%	

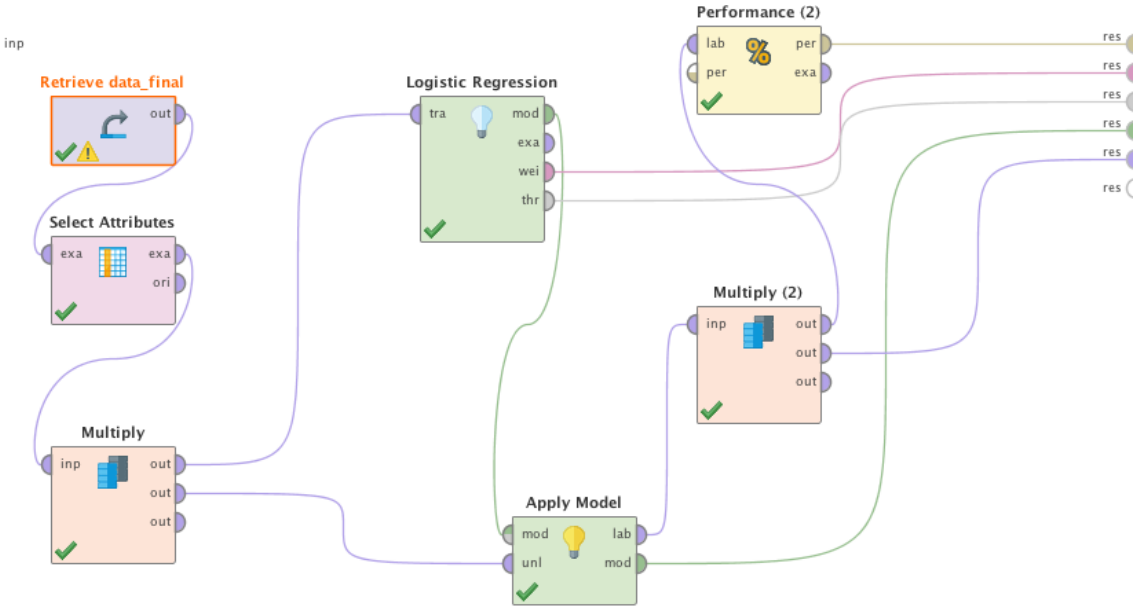
Tabulka 16 Úspěšnost rozhodovacího stromu bez extrémů

Nyní je úspěšnost výrazně nižší, ale reálnější. Není ovlivněna proměnnými, které byly vyřazeny. Celková úspěšnost modelu je 76.25 %.

4.3.3. Logistická regrese

Nastavení prediktivního modelu v programu Rapid Miner vypadá následovně.

Zdroj: Vlastní práce



Obrázek 8 Prediktivní proces Logistická regrese

- První funkce Retriveve data_final načetla výběrovou databázi
- Funkce Multiply poslala data oběma výstupy v nezměněné podobě
- Logistic Regression proved výpočet Logistické regrese
- Funkce Apply Model na vstupu získala původní data a prediktivní model Logistické regrese, použil na původní data
- Funkce Performance vytvoří tabulku úspěšnosti prediktivního modelu
- Proces má 5 výstupů

- Samotnou výstupní databázi rozšířenou o Prediction Re_order, Confidence(0) a Confidence(1)
- Prahovou hodnotu prediktivního modelu pro nově vytvořené Confidence 0 a 1
- Tabulku vah jednotlivých statistických znaků
- Tabulku Odhadu regresních parametrů s p-hodnotou
- Tabulku úspěšnosti prediktivního modelu

Prahová hodnota definuje, kdy se predikovaná proměnná Prediction Re_order mění z 0 na 1 v závislosti hodnot Confidence(0) a Confidence(1). Jakmile Confidence(1) je menší než 0,44, tak predikovaná proměnná už nemá vysokou pravděpodobnost správné predikce.

Zdroj: Vlastní práce

Threshold

```
Threshold: 0.44061708563847923
first class: 0
second class: 1
if confidence(1) > 0.44061708563847923 then 1
else 0
```

Tabulka 17 Logistická regrese Prahová hodnota

Další výstup Logistické regrese je tabulka koeficientů s p-hodnotou. Podle p-hodnoty v posledním sloupci, která testuje významnost regresního modelu, by se měl vyřadit z modelu znak subject a city.

Nejdůležitější hodnotou je sloupec Coefficient. Koeficient vypovídá o statistické významnosti ovlivnění prediktivního modelu.

Zdroj: Vlastní práce

Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value ↓
subject.0	0.773	0.773	0.531	1.454	0.146
city	-0.117	-0.066	0.074	-1.566	0.117
avr_customer_rating	-0.085	-0.081	0.047	-1.818	0.069
avg_base_length	-0.406	-0.439	0.062	-6.523	0.000
avg_order_price	0.002	0.598	0.000	7.520	0.000
count_promo_code	2.283	1.470	0.194	11.739	0
avg_promo_price	-0.006	-1.295	0.001	-10.725	0
avg_margin_length	19.105	2.249	1.355	14.102	0
Intercept	-4.552	1.173	-0.287	15.880	0

Tabulka 18 Logistická regrese koeficienty s p-hodnotou

Po odstranění znaků Subject a City se prahová hodnota změnila na 0.49 pro Confidence(1). Pokud Confidence(1) bude menší než 0.49, tak prediktivní model vyhodnotí Prediction Re_order s hodnotou 0.

U všech znaků se nyní p-hodnota rovná nule a není potřeba další znaky z prediktivního modelu vyřadit. Podle koeficientu jsou statisticky významné znaky působící pozitivně ve prospěch predikované proměnné s hodnotou 1 následující: průměrná doba prodloužení, počet promo kódů a průměrná cena objednávky.

Zdroj: Vlastní práce

Attribute	Coefficient ↓
avg_margin_length	19.193
count_promo_code	2.289
avg_order_price	0.002
avg_promo_price	-0.006
avg_base_length	-0.399
Intercept	-5.071

Tabulka 19 Logistická regrese koeficienty

Dle vzorců logistické regrese, lze interpretovat významnost koeficientů. Pokud se zvýší hodnota koeficientu avg_margin_length o 0.1 hodiny, pravděpodobnost určení predikované hodnoty 1 je o 99% větší. U count_promo_code při zvýšení počtu promo kódů o 1 se zvýší pravděpodobnost o 90% a u avg_order_price při zvýšení o 1 Kč se zvýší pravděpodobnost 50% za předpokladu, že se hodnoty ostatních znaků nezmění. Znaky se zápornou hodnotou působí ve prospěch neopětovného objednání služby.

Podle tabulky úspěšnosti je Logistická regrese v porovnání s Rozhodovacími stromy o pár setin procenta méně úspěšná. Úspěšnost je 76,34 %. 1715 hodnot správně predikuje Predicition Re_order = 1 a 615 hodnot správně jako Prediction Re_order = 0.

Zdroj: Vlastní práce

accuracy: 76.34%

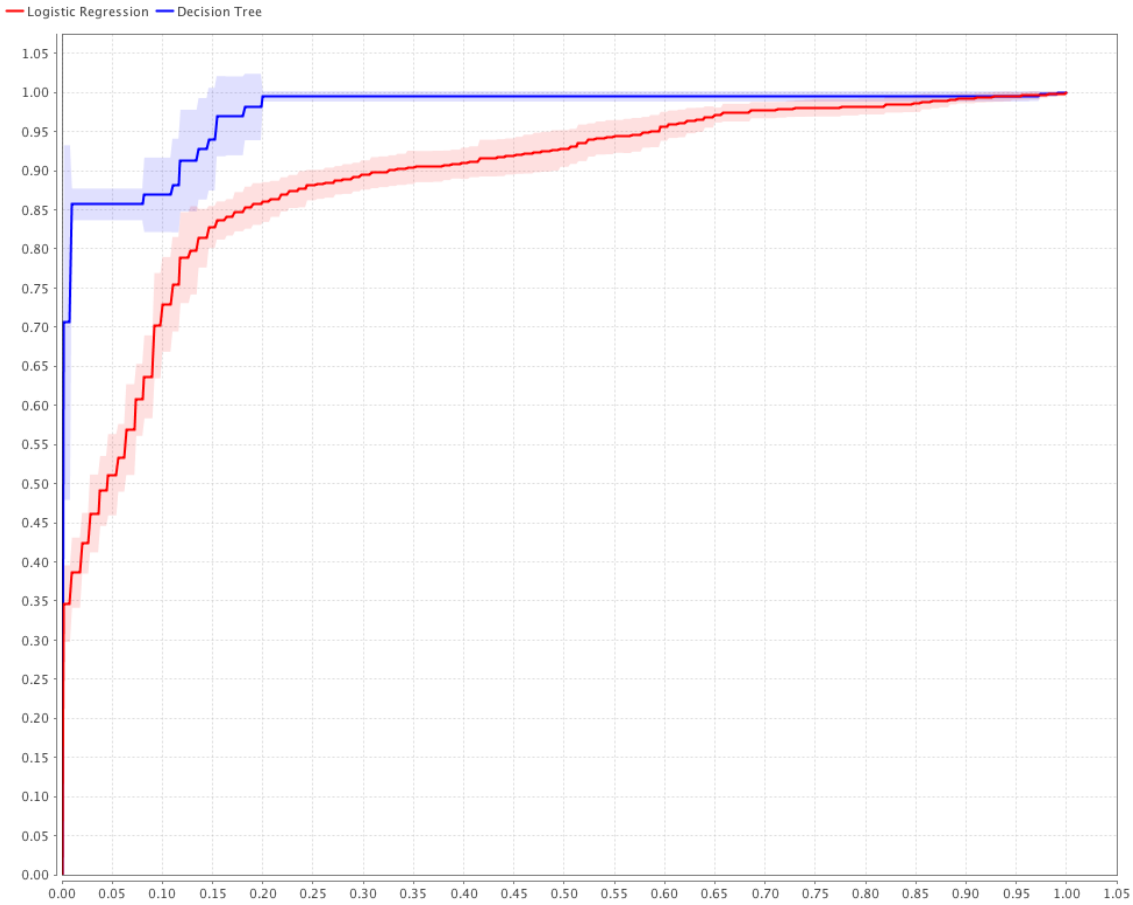
	true 0	true 1	class precision
pred. 0	615	227	73.04%
pred. 1	495	1715	77.60%
class recall	55.41%	88.31%	

Tabulka 20 Logistická regrese tabulka úspěšnosti

4.4. Porovnání prediktivních modelů

Pomocí ROC grafu byla porovnána kvalitu modelu Rozhodovacích stromy (modrá) a Logistické regrese (červená). Na základě grafu lze říct, že model Rozhodovacích stromů je kvalitnější než model Logistické regrese. Rozhodovací stromy mají velikost plochy pod křivkou 97 % a logistická regrese 88 %. Oba modely lze považovat za úspěšné s tím, že Rozhodovací stromy mají v tomto případě kvalitnější predikční schopnost.

Zdroj: Vlastní práce



Graf 30 Roc graf

5 Závěr

Cílem práce bylo vyhodnotit faktory ovlivňující opětovné objednání služby z databáze obsahující 3673 uživatelů. Byly použity metody rozhodovacích stromů a logistická regrese. Rozhodovací stromy měly vysokou predikční schopnost dosahující 97 % kvality. Logistická regrese měla dobrou predikční schopnost a dosáhla 88 % kvality.

Bylo zapotřebí z predikčních modelů odfiltrovat znaky počtu objednávek, předplatného, uložené karty, subjekt a město, které velmi ovlivňovaly modely.

Zároveň z původní databáze byly odfiltrovány odlehlé a extrémní hodnoty, které zvýšily úspěšnost prediktivních modelů. Bylo odfiltrováno celkem 621 uživatelů, což představuje 17% záznamů.

Metoda rozhodovacího stromu identifikovala, jako statisticky nejvýznamnější faktory působící pozitivně ve prospěch opětovného objednání služby, následující znaky. Průměrnou cenu objednávky pohybující se mezi 279 Kč a 1105 Kč, počet použitých promo kódů méně než 1,5 a dobu prodloužení úklidu méně než 0,35 hodiny.

Metoda logistické regrese identifikovala, jako statisticky nejvýznamnější faktory působící pozitivně ve prospěch opětovného objednání služby, následující znaky. Průměrnou dobu prodloužení úklidu, počet použitých promo kódů a průměrnou cenu objednávky.

V identifikovaných faktorech se obě metody shodly.

Z analýzy vyplývá, že by se vybraná společnost mohla zaměřit na uživatele, u kterých průměrná doba prodloužení úklidu nepřekročí 21 minut a zároveň nevyužili více než 2 promo kódy. Dále se mohou zaměřit na uživatele s objednávkou mezi 279 Kč a 1105 Kč. U uživatelů splňující tyto faktory je vysoká pravděpodobnost, že si službu objednají znovu, a proto má smysl se na tyto zákazníky marketingově zaměřit.

6 Seznam použitých zdrojů

[1] KOTLER, Philip. *Moderní marketing: 4. evropské vydání*. 1. vyd. Praha: Grada, 2007. ISBN 978-80-247-1545-2.

[2] Marketingový mix. In: *Sun Marketing* [online]. Praha: Sun Marketing s.r.o., 2006 [cit. 2018-02-09]. Dostupné z: http://www.sunmarketing.cz/nastroje/navody-pro-klienty/marketingovy-mix

[3] *Systemonline.cz: Datové sklady a jejich optimalizace* [online]. Praha: Jan Vrána, 2001 [cit. 2018-01-28]. Dostupné z: https://www.systemonline.cz/clanky/datove-sklady-a-jejich-optimalizace.htm

[4] RUD, Olivia. *Data mining: Praktický průvodce dolováním dat pro efektivní prodej, cílený marketing a podporu zákazníkům (CRM)*. 1. Praha: Computer Press, 2001. ISBN 80-7226-577-6.

[5] Korelační a regresní analýza. In: *Wikisofia* [online]. Praha: Wikisofia, 2010 [cit. 2018-02-09]. Dostupné z: http://wikisofia.cz/wiki/Korela%C4%8Dn%C3%AD_a_regresn%C3%AD_anal%C3%BDza

[6] *Muni.cz: Shluková analýza* [online]. Jiří Kučera, b.r. [cit. 2018-01-29]. Dostupné z: https://is.muni.cz/th/172767/fi_b/5739129/web/web/main.html

[7] *Matematickabiologie.cz: Shlukoval nehierarchická analýza* [online]. b.r. [cit. 2018-01-29]. Dostupné z: http://portal.matematickabiologie.cz/index.php?pg=analyza-a-hodnoceni-biologickych-dat--vicerozmerne-metody-pro-analyzu-dat--shlukova-analyza--shlukova-nehierarchicka-analyza--uvod

[8] PETR, Pavel. *Data Mining*. Vyd. 2. Pardubice: Univerzita Pardubice, 2008. ISBN 978-80-7395-098-9.

[9] VYSEKALOVÁ, Jitka. *Chování zákazníka: jak odkrýt tajemství "černé skříňky"*. Praha: Grada, 2011. ISBN 978-80-247-3528-3.

[10] ABBOTT, Dean. *Applied predictive analytics: principles and techniques for the professional data analyst*. Indianapolis, Indiana, 2014. ISBN 978-111-8727-690.

[11] ŘEHÁKOVÁ, Blanka. Nebojte se logistické regrese. *Sociologický časopis* [online]. 2000, 2000(4), 475-492 [cit. 2018-03-10]. Dostupné z: http://sreview.soc.cas.cz/cs/issue/64-sociologicky-casopis-4-2000/1149

[12] ZLÁMAL, Filip. *Logistická regrese v R*. Brno, 2013. Bakalářská práce. Masarykova univerzita, Přírodovědecká fakulta, Ústav matematiky a statistiky.

[13] ŽAMBOCHOVÁ, RNDr. Marta. *Jak na rozhodovací stromy*. Ústní nad Labem, 2007. Univerzita J. E. Purkyně v Ústní n.L., Fakulta sociálně ekonomická, Katedra matematiky a statistiky.

7 Seznam obrázků

8 Seznam tabulek

9 Seznam grafů