

EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA PODNIKOVÉHO MANAŽMENTU

Evidenčné číslo: 104006/B/2021/36122176490522372

APLIKÁCIA DATAMININGU V PODNIKOVEJ PRAXI

Bakalárska práca

2021

Patrik Valach

EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA PODNIKOVÉHO MANAŽMENTU

APLIKÁCIA DATAMININGU V PODNIKOVEJ PRAXI

Bakalárska práca

Študijný program: ekonomika a manažment podniku

Študijný odbor: ekonómia a manažment

Školiace pracovisko: Katedra informačného manažmentu

Vedúci záverečnej práce: doc. Ing. Anna Hamranová, PhD.

Bratislava 2021

Patrik Valach

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Patrik Valach
Študijný program: ekonomika a manažment podniku (Jednoodborové štúdium, bakalársky I. st., denná forma)
Študijný odbor: ekonómia a manažment
Typ záverečnej práce: Bakalárska záverečná práca
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Aplikácia data miningu v podnikovej praxi

Anotácia: Cieľom práce je analyzovať metódy hĺbkovej analýzy údajov (Data Mining), charakterizovať jednotlivé Data Miningové techniky a aplikovať ich na konkrétne podnikové údaje. Súčasťou práce je vyhodnotenie možností aplikácie Data Miningu a ich praktického použitia z pohľadu samotných metód, ako aj z pohľadu podnikových procesov.

Vedúci: doc. Ing. Anna Hamranová, PhD.
Oponent: Ing. Peter Zahradník
Katedra: KIM FPM - Kat. informacneho manazmentu FPM
Vedúci katedry: doc. Ing. Anita Romanová, PhD.
Dátum zadania: 31.08.2020

Dátum schválenia: 06.02.2020
doc. Ing. Mojmír Kokles, PhD.
vedúci katedry

Čestné vyhlásenie

Čestne vyhlasujem, že bakalársku prácu som vypracoval samostatne a uviedol som všetku použitú literatúru.

Dátum:

.....

Patrik Valach

Pod'akovanie

V prvom rade by som sa chcel veľmi pekne pod'akovať vedúcej mojej bakalárskej práce doc. Ing. Anne Hamranovej, PhD. za ochotu a pomoc pri usmerňovaní a všetky konzultácie počas prípravy a písania práce. Ďakujem tiež mojej rodine a priateľom za neustálu podporu.

ABSTRAKT

VALACH, Patrik: Aplikácia dataminingu v podnikovej praxi. – Ekonomická univerzita v Bratislave. Fakulta podnikového manažmentu; Katedra informačného manažmentu. – Vedúci záverečnej práce: doc. Ing. Anna Hamranová, PhD. – Bratislava: FPM, 2021, 54 strán.

Cieľom záverečnej práce je definovať súčasné poznatky v oblasti dataminingu a ich následná aplikácia pri tvorbe dataminingových modelov na základe dát vybranej spoločnosti. Práca zahŕňa charakteristiku pojmu „datamining“, charakteristiku vybraných metód analýzy dát, ako aj jednotlivé kroky, ktoré proces dataminingu zahŕňa. Aplikačná časť je riešená vo vybranom podniku, kde na základe podnikových dát je realizovaná hĺbková analýza s cieľom navrhnúť zlepšenie v oblasti tvorby portfólia klientov a umiestnených reklamných spotov. Práca je rozdelená do troch kapitol. Obsahuje tri grafy a tri obrázky. Prvá kapitola je venovaná súčasnému stavu riešenej problematiky doma a v zahraničí. V ďalšej časti je charakterizovaný hlavný cieľ, čiastkové ciele práce a metódy využité pre ich dosiahnutie. Tretia kapitola sa zaoberá aplikáciou dataminingových metód na dáta vybranej spoločnosti a vyhodnotením splnenia vytýčených cieľov využitými metódami. Výsledkom riešenia danej problematiky je analýza momentálneho portfólia klientov spoločnosti a návrhy pre vylepšenie tohto portfólia za účelom zlepšenia jeho rentability a ziskovosti tejto spoločnosti.

Kľúčové slová:

datamining, postupy a metódy dataminingu, regresná analýza, dataset, vizualizácia dát

ABSTRACT

VALACH, Patrik: Application of datamining in business practice. - University of Economics in Bratislava. Faculty of business management; Department of Information Management. - Thesis supervisor: doc. Ing. Anna Hamranová, PhD. - Bratislava: FPM, 2021, 54 pages.

The aim of the bachelor thesis work is to define current knowledge in the field of datamining and its subsequent application in the creation of data mining models based on data from a selected company. The work includes the characteristics of the term "datamining", the characteristics of selected methods of data analysis, as well as the individual steps that the process of datamining includes. The application part is solved in a selected company, where an in-depth analysis is performed on the basis of company data in order to propose improvements in the creation of a client-portfolio and allocated commercials. The thesis is divided into three chapters. It contains three graphs and three pictures. The first chapter is devoted to the current state of the problem at home and abroad. The next part characterizes the main goal and partial goals of the work and the methods used to achieve them. The third chapter deals with the application of data mining methods to the data of the selected company and the evaluation of the fulfillment of the set goals considering the methods used. The result of solving the problem is an analysis of the current portfolio of the company's clients and subsequent suggestions for potential improvements regarding this portfolio in order to improve its rentability and the profitability of the company.

Key words:

datamining, approaches and methods of datamining, regression analysis, dataset, data visualization

Obsah

Úvod	9
1 Súčasný stav riešenej problematiky doma a v zahraničí	11
1.1 VYUŽITIE DATAMININGU.....	14
1.2 DATAMINING A JEHO PROCESNÉ PRVKY	15
1.3 SPÔSOBY ULOŽENIA DÁT.....	18
1.4 PRÍSTUPY DATAMININGU.....	20
1.5 CHARAKTERIZÁCIA METÓD DATAMININGU	21
1.6 REGRESIA.....	24
1.6.1 Lineárny regresný model	25
1.6.2 Viacrozmerný regresný model	26
1.7 ASOCIÁCIA.....	27
2 Cieľ práce a metódy skúmania.....	29
3 Výsledky práce a diskusia	33
3.1 PREDSTAVENIE SPOLOČNOSTI	33
3.2 ANALÝZA AKTUÁLNEHO STAVU V PODNIKU V OBLASTI REKLAMNÝCH SPOTOV ZA ROK 2019	34
VYŠPECIFIKOVANIE NEDOSTATKOV	35
3.3 PROBLEMATIKA RENTABILITY REKLAMNÝCH SPOTOV A ZISKOVOSTI V JEDNOTLIVÝCH KVARTÁLOCH ROKU 2019	36
3.3.1 Riešenie problematiky sledovanosti a rentability v rámci demografickej skupiny A 12-54 pomocou lineárnej regresie	37
3.3.2 Riešenie distribúcie ziskov z reklám v jednotlivých kvartáloch fiškálneho roka pomocou Power BI	44
3.4 POROVNANIE A HODNOTENIE PROGRAMU KNIME A POWER BI.....	48
Záver	51
Zoznam bibliografických zdrojov	53

Úvod

Život v 21. storočí je výrazne poznačený rapídnyim pokrokom v informatizácii, výpočtových technológiách a štatistikách, ako aj v sfére umelých inteligencií. „Doba dát“ je veľmi presným opisom prítomnosti, ako aj nadchádzajúcich rokov. Takýto pokrok so sebou ruka v ruku denne prináša aj vznik enormného počtu dát a informácií, ktoré sa následne zaznamenávajú, klasifikujú, zakladajú a uchovávajú pre rôzne potreby ako štátnych, tak aj súkromných inštitúcií. Spolu tieto informácie vytvárajú ucelený obraz o skutočnosti a o chode každodenného života z rôznych uhlov pohľadov, ktoré predstavujú hodnotné výstupy pre spracovávateľa takýchto informácií. Bližší pohľad na veľké množstvo každodenne vyskytujúcich sa úkonov odhalí, že v úzadí zanechávame stopu v podobe dát, ktoré sú výsledkom práve spomenutej automatizácie systému spoločnosti, ako aj vývinu nástrojov, ktoré tieto zhluky dát dokážu zbierať a kategorizovať. Kostra globálnej telekomunikačnej siete denne distribuuje desiatky petabytov dát, farmaceutický priemysel obsahuje skrz databázy pacientov, záznamy a iné dokumenty taktiež obrovské objemy dát. Banky, úrady a iné inštitúcie sú schopné archivovať a operujú s rôznymi detailne spísanými dátami podrobne popisujúcimi ich klientelu a jej správanie. V neposlednom rade taktiež sociálne siete, miesto kde sa efektívne ukladajú dáta najrozličnejšieho charakteru a dôležitosti, sú platformou, ktorá vďaka veľkému počtu používateľov slúži ako jeden z hlavných zdrojov takýchto datasetov. Takpovediac v každej sfére vieme identifikovať zdroje a práve v dobe, kedy sú naprieč celým svetom každou minútou zbierané veľké množstvá dát je dôležité a potrebné tieto vzorky analyzovať a výstupy využívať v náš prospech. Cieľom tejto bakalárskej práce je práve tento prístup aplikovať za účelom zlepšenia situácie a ziskovosti podniku. Tieto rapídne rozrastajúce sa a voľne dostupné súbory dát vôkol nás plynúce z takmer úplne každého aspektu nášho života volajú po systémoch, ktoré by v ich rámci identifikovali relevantné informácie a datamining je prostriedok, ktorý ponúka na túto potrebu efektívne riešenie.

Bakalárska práca pozostáva z troch častí. V prvej kapitole sme sa venovali priblíženiu pojmu datamining, spomenuli sme podstatné výrazy a opísali proces ako taký. Venovali sme sa taktiež problémom spojeným s dataminingom a popísali využitie dataminingu v rôznych sférach. Priblížili sme taktiež, čo mu predchádza formou charakterizácie prípravy dát. Taktiež sme definovali, aké informácie sa snažíme extrahovať a určili sme z akých zdrojov môžeme čerpať dáta, ktoré môžu byť „dolované“.

V neposlednom rade sme sa venovali rozdeleniu predikčných a deskriptívnych typov modelov a ich metódam. Ako ďalší bod sme stručne charakterizovali jednotlivé metódy a napokon sa bližšie venovali regresii a asociácii. V druhej kapitole sme podrobne popísali hlavný cieľ a taktiež čiastkové ciele tejto práce a dôvody, prečo sú výstupy dosiahnuté pomocou týchto cieľov hodnotné pre konkrétnu spoločnosť. V poslednej tretej časti tejto bakalárskej práce sme pristúpili k skúmaniu praktického využitia dataminingu v podniku. Aplikovali sme metódy za účelom dosiahnutia poznatkov, ktoré sme v ďalšom kroku mohli využiť na uspokojenie potrieb podniku. V závere sú metódy a výsledky z aplikačnej časti vyhodnotené vzhľadom k jednotlivým cieľom.

1 Súčasný stav riešenej problematiky doma a v zahraničí

Podľa Jiawei Hana¹ ktorý vo svojej knihe hovorí o procese dolovania dát definuje datamining ako niečo, čo predstavuje nástroj a stratégiu ktorou sa takéto veľké súbory dát dajú hĺbkovo analyzovať. Ďalší pohľad prinášajú vo svojej knihe autori Ian H. Witten a Eibe Frank², ktorí tento proces opisujú ako nástroj, ktorý analýzou dát obsiahnutých v súboroch rieši problémy, respektíve svojím výsledným produktom v podobe modelov pomôže identifikovať možné stratégie a tak výraznou mierou prispieva k ich riešeniu. V tomto smere môže byť označený ako prirodzená evolúcia informačných technológií. V prvej fáze vývoja vznikla možnosť dáta zatriediť do databáz a systémov, čo položilo základy pre dátový manažment, t. j. získavanie dát na základe vytvorenia dotazu ktorý pracuje s uloženými dátami. Oba kroky vydláždili cestu vzniku dataminingu, ktorý vytvorené súbory analyzoval. V čase, kedy takmer každý vlastní zariadenie, ktoré takéto dáta vie zaznamenávať je proces dataminingu prísľubom efektívneho vypracovania korektného a presného modelu, ktorého výsledkom sú potrebné dáta znamenajúce vedomosti určené pre účely analyzujúceho. Hĺbková analýza dát je veľmi komplexný pojem, ktorý ako proces nechápeme ako úkon, ktorým by sa súbory dát jednoducho spracovali a zároveň v jednom kroku zanalyzovali. Samotný proces pozostáva z niekoľkých úkonov a v konečnom dôsledku sú informácie a vedomosti z jednej analýzy posudzované a opäť analyzované inými metódami za účelom nadobudnutia viacerých perspektív a v konečnom dôsledku vytvorenia správneho záveru. Tento proces je využívaný podnikmi na domácom, ale aj zahraničnom trhu a to najmä za účelom zlepšenia svojich produktov a služieb alebo k zlepšeniu vnútropodnikovej štruktúry. Výstupy ktoré sú podnikom pomocou dataminingu poskytované slúžia k vyhotovovaniu reportov, lepšiemu poznaniu správania konzumentov, zefektívneniu marketingových metód, predikcie správania zákazníkov a vo všeobecnosti poskytujú podniku informácie, ktoré napomáhajú lepšej orientácii, vyhodnocovaniu situácii a následnému prijatiu opatrení alebo prípadnej úprave biznis modelu pre lepšiu reakciu na možné zmeny trhu. Už z uvedených funkcií hĺbkovej analýzy dát je jasné, že ako primárny prínos tohto procesu spracovania dát

¹ HAN, Jiawei – PEI, Jian – KAMBER, Micheline. Data mining: concepts and techniques. 3. vydanie. San Francisco: Elsevier, 2011. 560 s. ISBN 9780123814791

² WITTEN, Ian H. – FRANK, Eibe. Data mining: practical machine learning tools and techniques with Java implementations. 2. vydanie. San Francisco: Elsevier, 2005. 525 s. ISBN 1-55860-552-5

ako takého sú smerodajné ukazovatele predikčného charakteru, ktoré vytvoria projekciu budúceho vývoja udalostí, trhu a správania zákazníkov na základe doposiaľ poskytnutých informácií. Hĺbková analýza dát je dnes širokospektrálne využívaná mnohými spoločnosťami z rôznych odvetví od podnikov, ktorých cieľom je inovácia produktového mixu, spoločnosťami v reštrukturalizácii až po podniky, ktorých obchodná činnosť stagnuje a potrebuje stratégiu pre reštart. Metódy dataminingu sú rozdelené do dvoch hlavných typov modelov. Identifikujeme predikčné a deskriptívne modely/prístupy, na základe ktorých sú produkované relevantné informácie. Oba typy svojou variáciou príslušných metód ponúkajú jedinečnú perspektívu na skúmanú oblasť, a preto sa podnety ktoré vedú k využitiu hĺbkovej analýzy dát, ako aj následná voľba metódy líšia na individuálnej báze. Dáta samotné vieme taktiež rozpoznávať a distingoovať na základe rôznych faktorov. Zaznamenané dáta môžu byť skúmané ako časové alebo dátumové údaje, ale rovnako aj ako textové či číselné hodnoty.

V užšom slova zmysle je dolovanie dát súborom nástrojov a techník a taktiež je jednou z niekoľkých technológií potrebných na podporu riadenia podniku zameraného na zákazníka. V širšom zmysle je dolovanie dát prístup, ktorým by mali byť obchodné akcie na základe poznatku, že rozhodnutia prijaté v stave plnej informovanosti sú lepšie ako neinformované rozhodnutia a že kvantitatívna alebo kvalitatívna analýza dát je pre podnik prospešná.

Keďže datamining je sám o sebe iba postup, ktorý nástroje a metódy uplatňuje je potrebné pre efektívnu ťažbu zabezpečiť, aby boli splnené aj ďalšie požiadavky na analytický CRM (customer relationship management) systém, respektíve manažment klientskych vzťahov. Michael J.A. Berry a Gordon S. Linoff³ určili, že za účelom vytvorenia vzdelávacieho vzťahu so svojimi zákazníkmi, firma musí byť schopná:

- Všímať si, čo robia jej zákazníci.
- Pamätať si, čo všetko samotná firma a jej zákazníci v priebehu času urobili
- Poučiť sa z toho, čo bolo zapamätané.
- Konat' podľa toho, čo sa firma naučila, aby sa zvýšila ziskovosť

³ BERRY, Michael J.A. – LINOFF, Gordon S. Data mining techniques: for marketing, sales, and customer relationship management. 2. vydanie. Indianapolis: John Wiley & Sons, 2004. 643 s. ISBN 9780471470649

V podniku musia taktiež existovať systémy na spracovanie transakcií a na zachytenie interakcií so zákazníkmi, ako aj údajové sklady na uskladnenie historických informácií o správaní zákazníkov. Z poskytnutých dát minulých období následne hĺbková analýza vyvodí výsledné potrebné poznatky a tie v konečnom dôsledku podniku pomôžu do plánov a budúcich opatrení, ako aj stratégie vzťahov so zákazníkmi implementovať zmeny na základe týchto dát tak, aby spoločnosť vďaka nim vykonávala svoju podnikateľskú činnosť lepšie a vedela uspokojiť dopyt efektívnejšie. Výhodou modernej doby je fakt, že mnoho odvetví dokáže automatizovať a digitalizovať procesy, ktoré bolo nutné bez technologicky pokročilých zariadení spracovávať manuálne. Každé odvetvie podnikania sa snaží tomuto trendu priblížiť a uvoľniť čo najviac personálnej kapacity pre čo najefektívnejšie využitie zdrojov, nakoľko s novodobou technikou je možné nahrádzať mnoho pozícií, respektíve prenechať mnoho manuálne náročných úloh strojom. V priemyselnej výrobe sa stretávame pomerne často s robotickými komponentami, ktoré sú naprogramované na opakovanie singulárnej aktivity počas žiadanej doby a s potrebnými materiálmi. Pre využitie hĺbkovej analýzy dát je však najpodstatnejší pokrok v oblasti informatiky, nakoľko moderné počítače sú schopné uniesť veľké množstvá dát, a preto veľa podnikov s pomerne dostupnou technológiou je zameraných na analýzu mnohých súborov interných databáz, ktoré obsahujú rôzne typy dát pre svoje komparatívne výhody. Vníname to ako obrovský pokrok, nakoľko sa všetky úpravy a analýzy dát museli v predošlých desaťročiach vypracovávať ručne pomocou tímov štatistov a analytikov, ktorí sa pri každom výskume v rámci dát museli spoliehať na vlastné vedomosti a v čo najväčšom množstve eliminovať prípadné ľudské pochybenia, respektíve „human error“. Podľa Fostera Provosta a Toma Fawcetta⁴, sú metódy dataminingu v najväčšom množstve využívané v odvetviach marketingu, online reklamy, a odporúčaní pre krízový predaj. V tomto tvrdení je potrebné s autormi súhlasiť, nakoľko je toto konštatovanie postavené na fakte, že menované segmenty patria k jedným z najmladších. Zároveň ako ďalší utvrdzujúci aspekt slúži, že v podnikovej praxi firiem v tomto obore sú dáta najelementárnejšou zložkou informácií, ktoré za účelom vytvorenia produktu na uspokojenie potrieb svojej klientely môžu využiť vo forme vstupov pre svoje interné procesy.

⁴ PROVOST, Foster – FAWCETT, Tom. Data Science for Business: What you need to know about data mining and data-analytic thinking. Sebastopol: O'Reilly Media, Inc., 2013. 414 s. ISBN 9781449374280

1.1 Využitie dataminingu

Dobre vedené malé podniky prirodzene vytvárajú vzťahy so svojimi zákazníkmi. Postupom času sa dozvedajú viac a viac o svojich zákazníkoch a tieto vedomosti využívajú na to, aby im lepšie slúžili. Výsledkom sú spokojní, verní zákazníci a ziskové podniky. Väčšie spoločnosti so státisícmi alebo miliónmi zákazníkov si neužívajú luxus takto prirodzene budujúcich sa vzťahov, kde správanie oboch strán je plne transparentné a jednoducho zaznamenateľné či dokonca štandardizované. Tieto väčšie firmy sa musia spoliehať na iné prostriedky a musia manuálne formovať vzťahy so svojimi zákazníkmi. Musia sa predovšetkým učiť naplno využívať výhody niečoho, čo majú v hojnosti - dáta vyprodukované takmer pri každej interakcii so zákazníkom. Využitie dataminingu v prípade modelovania vzťahov a predikcie chodu trhu je ale len jedným z mála možných využití, pre ktoré je takáto analýza vhodná. Mimo podnikovej sféry je dolovanie dát využiteľné aj v rôznych aplikačných doménach. Svoje využitie dátová analýza našla v bankovníctve a poisťovníctve, taktiež v biofarmaceutickom priemysle na prieskum efektivity nových liekov ale aj v telekomunikačnom priemysle. Datamining je využívaný v rámci systémov Business Intelligence pri odhaľovaní finančných alebo poisťovacích podvodov, hodnotení rizík či dolovaní dát z textu, ale je rovnako použiteľný aj pri úkonoch ako je stanovenie diagnózy v medicíne. Ďalšími oblasťami kde je analýza dát využiteľná je školstvo; existuje nová rozvíjajúca sa oblasť s názvom edukačný datamining, ktorá sa zaoberá vývojom metód, ktoré objavujú poznatky v súboroch údajov pochádzajúcich zo vzdelávacích prostredí. Ciele takéhoto dataminingu sú vyšpecifikované ako predpovedanie budúceho správania študentov pri učení, štúdium účinkov vzdelávania a rozširovanie vedeckých poznatkov o učení. Datamining môže inštitúcia využiť na prijímanie presných rozhodnutí a tiež na predpovedanie výsledkov študenta. S výsledkami sa inštitúcia môže zamerať na to, čo učiť a ako učiť. Rovnako v sfére vyšetrovania protizákonných činností je datamining rozšírenou praktikou. Kriminológia je proces zameraný na identifikáciu charakteristík kriminality. Analýza kriminality zahŕňa skúmanie a odhaľovanie trestných činov a ich vzťahov s zločincami. Vysoký objem uskladnených údajov o trestnej činnosti a tiež zložitosť vzťahov, ktoré medzi týmito druhmi údajov existujú spôsobili, že kriminalistika je vhodnou oblasťou na uplatnenie techník ťažby údajov. Tieto informácie možno použiť na vykonanie procesu porovnávania trestných činov. Ďalším príkladom pre využitie dátovej analýzy sú agentúry v mediálnom priemysle, ktoré na dennej báze vytvárajú a následne taktiež pracujú s databázami obsahujúcimi enormné počty záznamov. Primárne využitie tieto datasety

nachádzajú v reportingových aktivitách ako aj v bežnom vyhodnocovaní uplynulých období a teda za účelom modifikácie predajnej stratégie a iných komponentov podnikania. Je však už z uvedeného počtu a charakteru príkladov využitia dataminingu jasné, že je možné ho aplikovať takmer v každom smere ak to situácia umožňuje a vyžaduje.

Faktom je, že vo všetkých druhoch priemyslu inklinujú spoločnosti k cieľu porozumieť každému zákazníkovi individuálne a využiť ho na to aby uľahčili zákazníkovi obchodovanie s nimi a nie s konkurenciou. Rovnako sa firmy snažia a postupným procesom učia pozeráť na hodnotu každého zákazníka aby vedeli, do akej klientely sa oplatí investovať peniaze a úsilie, aby obchodný vzťah udržali a s ktorou by obchodovanie nebolo ziskové. Tvrdenie, že datamining je multidisciplinárny proces podporuje vo svojej knihe Jiawei Han, Micheline Kamber a Jian Pei⁵ ktorí ho rozdelili do niekoľkých úkonov.

1.2 Datamining a jeho procesné prvky

Datamining je často považovaný za elementárny proces a mnoho ľudí ho vníma ako synonymum pre ďalší populárny pojem, získavanie poznatkov z dát, zatiaľ čo iní považujú dolovanie dát iba za zásadný krok v procese zisťovania znalostí. Ako však vo svojom článku uviedli Festim Halili a Avni Rustemi⁶, je to súbor komplexných činností, ktoré majú za následok získanie potrebných relácií medzi údajmi datasetu.

- 1) Očistenie dát – krok v ktorom sa z kompletného súboru dát odstraňujú nekorektné a irelevantné prvky
- 2) Integrácia dát – vo fáze integrácie sa dáta získané z rôznych zdrojov a oblastí vložia do jednotného formátu, ktorý zabráni rozličnej reprezentácii vstupných formátov dát
- 3) Selekcia/výber dát – počas selekcie sa identifikujú najrelevantnejšie vzorky pre stanovený výskum a preto je často nutná znalosť v oblasti skúmaných dát
- 4) Transformácia dát – aby bolo určité metódy dataminingu možné uplatňovať je potrebné skúmané dáta transformovať a skonsolidovať do náležitej formy použitím

⁵ HAN, Jiawei – PEI, Jian – KAMBER, Micheline. Data mining: concepts and techniques. 3. vydanie. San Francisco: Elsevier, 2011. 560 s. ISBN 9780123814791

⁶ HALILI, Festim – RUSTEMI, Avni. In Predictive modeling: data mining regression technique applied in a prototype [online]. Telangana: IJCSMC, 2016, roč. 5, č. 8, : 207-215 s. ISSN 2320-088X. Dostupné na:

<https://www.ijcsmc.com/docs/papers/August2016/V5I8201642.pdf>

sumarizačných alebo agregátnych operácii. Transformácia sa môže taktiež vykonať za účelom odľahčenia súboru o nadbytočné atribúty.

- 5) Ťažba dát – esenciálny proces uplatňovania algoritmov pre extrahovanie znalostí
- 6) Ohodnotenie výsledných hodnôt – fáza počas ktorej sa vydolované znalosti ohodnocujú na základe stanovenej miery zaujímavosti (interestingness measures). Znalosti sú individuálne posudzované na základe ich pragmatického využitia, presnosti odhadovania budúceho stavu a iných parametrov
- 7) Prezentácia poznatkov/znalostí – Poznatky ktoré sú po ohodnotení identifikované ako najužitočnejšie sú v konečnom štádiu prezentované pomocou vizualizácie a/alebo poznatkovej reprezentácie užívateľom pre pochopenie perspektívy ktorú nám o skúmanom súbore ponúkajú.

Autor Charu C. Aggarwal⁷ vo svojom diele Data Mining : The text book tento proces taktiež vysvetľuje ako celok pozostávajúci z niekoľkých úkonov. Uvádza tri kroky:

- 1) Zber dát
- 2) Extrakcia črtou a čistenie dát
- 3) Analytické procesovanie a algoritmy

V prvom rozdelení sú kroky 1 – 4 sú chápané ako rôzne druhy úprav, kedy sa súbor dát formátuje pre následné dolovanie. Samotný krok dolovania môže zahŕňať interakciu s užívateľom alebo vedomostnou základňou (knowledge base). Relevantné a hodnotné poznatky sú zdieľané s a prezentované užívateľovi a môžu byť následne založené ako nové poznatky vo vedomostnej základni.

Pod pojmom knowledge base alebo vedomostná základňa vieme nájsť dva rôzne spôsoby interpretácie jeho definície. Matthew Petterson⁸ uvádza, že o vedomostnej základni môžeme hovoriť ako o konkrétnom mieste, respektíve publikovanej platforme, ktorá slúži ako úložný priestor všetkých relevantných informácií, dokumentov, náučných manuálov, odpovedí na často sa vyskytujúce otázky a údajov súvisiacich s podnikom alebo taktiež ako

⁷ AGGARWAL, Charu C. Data mining: the textbook. Springer, 2015. 734 s. ISBN 978-3-319-14142-8

⁸ PATTERSON, Mathew. In Knowledge base 101: Definition, Benefits, Examples and Tips [online]. Dostupné na: <https://www.helpscout.com/playlists/knowledge-base/>

o softvéri, ktorý je zodpovedný za administráciu a taktiež prezentáciu takýchto dát. Je to teda výsledný produkt zbierania a organizovania vyššie spomenutých dát do užitočnej formy a následne využitie softvéru, ktorý konkrétnu vyžiadajú informáciu z tohto zdroja vytvorí a poskytne užívateľovi.

Vyššie uvedené rozloženie krokov kompletného procesu nám ukazuje, že datamining je jedným z niekoľkých procesných úkonov v rámci celku. Avšak v iných odvetviach, ako v médiách alebo priemysle, je datamining uvádzaný ako proces pokrývajúci všetky úkony; pravdepodobne za účelom odstránenia nevyhnutnosti nazývať tento proces komplexne, a to „Získavanie poznatkov z dát“. Ukazuje sa nám preto širšia škála chápania funkcionality dataminingu – je to proces odhaľovania hodnotných informácií a poznatkov uložených vo veľkých súboroch dát.

Popri uvedení pozitívnych aspektov a krokov, ktoré ucelene tvoria proces hĺbkovej analýzy dát, je taktiež potrebné uviesť aj možné problémy a ťažkosti, s ktorými sa užívateľ alebo dataminingové softvéry môžu stretnúť počas jednotlivých fáz. Pri objemoch dát, ktoré sú každým dňom zaznamenávané je nemožné vyhnúť sa takýmto stavom a skutočnostiam, nakoľko charakter a forma jednotlivých vzoriek je v každom prípade iná a líši sa na základe zdroja aj spôsobu zachytávania. Uvedieme teda najčastejšie problémy ktorým je datamining vystavený.⁹

1. Zlá kvalita dát - ako napríklad údaje obsahujúce šum, údaje s neaktuálnymi alebo chýbajúcimi hodnotami, nepresné respektíve nesprávne hodnoty, neprimeraná veľkosť dát a nedostatočné zastúpenie pri vzorkovaní údajov.
2. Integrácia konfliktných alebo zbytočných údajov z rôznych zdrojov a foriem: multimediálne súbory (audio, video a obrázky), geografické údaje, textové, sociálne, číselné údaje atď.
3. Čoraz aktuálnejšia je aj téma čím ďalej tým viac sa zväčšujúcich obáv o bezpečnosť a súkromie spoločnosti na ktoré je konštantne poukazované jednotlivcami, organizáciami a vládami.

⁹ KATIYAR, Pranchal. Difference Between Descriptive and Predictive Data Mining [online]. Uverejnené: 17.09.2019 [cit. 2021-01-05] Dostupné na: <https://www.geeksforgeeks.org/difference-between-descriptive-and-predictive-data-mining/>

4. Nemožný alebo sťažený prístup k údajom.
5. Škálovateľnosť algoritmov využitých na dolovanie pre efektívne získavanie informácií z veľkého množstva dát v databázach.
6. Práca s obrovskými setmi údajov, ktoré vyžadujú viac ako jeden distingvovaný prístup.
7. Procesovanie nestatických, nevyvážených a citlivých údajov.
8. Ťažba informácií z heterogénnych databáz a globálnych informačných systémov.
9. Konštantná aktualizácia modelov pre zachovanie schopnosti zachytávať prichádzajúce údaje a udržať funkčnosť úmerne s ich rýchlosťou.
10. Vysoké náklady spojené s nákupom a údržbou dostatočne výkonného softvéru, serverov a hardvéru na ukladanie dát, ktoré majú kapacitu spracúvať veľké množstvá dát.
11. Úprava veľkých, zložitých a neštruktúrovaných údajov do štruktúrovaného formátu a jeho následné spracovanie.
12. Problémovým faktorom môže v konečnom dôsledku byť aj kvantita výstupov z množstva metód dolovania údajov.

1.3 Spôsoby uloženia dát

Ako všeobecná technológia sa dá ťažba dát použiť na akýkoľvek druh údajov, pokiaľ tieto údaje majú pre cieľovú aplikáciu zmysel. Z tohto hľadiska vieme preto možné dátové zdroje rozdeliť do viacerých kategórií.

Základné formy dát pre ťažbu sú:

- transakčné údaje
- databázové údaje
- údaje dátového skladu
- iné druhy dát

Transakčné údaje

Všeobecne každý záznam v transakčnej databáze zachytáva transakciu, napríklad nákup zákazníka, rezervácia letu alebo návšteva webovej stránky používateľom. Transakcia spravidla obsahuje jedinečné identifikačné číslo transakcie (transaction ID) a zoznam položiek, ktoré transakcia obsahuje, napríklad položky zakúpené v tejto transakcii. Transakčnú databázu môžu tvoriť ďalšie tabuľky, ktoré obsahujú podrobnejšie súvisiace

informácie k transakciám, ako je napríklad popis položky, informácie o predajcovi alebo pobočka.

Databázové údaje

Databázový systém, ktorý sa tiež nazýva systém manažmentu databázy, pozostáva zo súboru vzájomne prepojených údajov, známych aj ako databáza, a taktiež sadu softvérových programov, ktorých účelom je spravovať a umožňovať prístup k údajom. Softvérové programy poskytujú mechanizmy na definovanie databázovej štruktúry a ukladanie údajov; na špecifikáciu prístupu k súbežným, zdieľaným alebo distribuovaným dátam a na zabezpečenie konzistencie a bezpečnosti uložených informácií pre ochranu pred poruchami systému alebo pokusmi o neoprávnený prístup.

Údaje dátového skladu

Dátový sklad je úložisko informácií zhromaždených z viacerých zdrojov, ktoré sú následne uložené v jednotnej schéme a formáte a obvykle sú lokalizované na jednom mieste. Dátové sklady sú vytvorené prostredníctvom procesu čistenia údajov, integrácie údajov, transformácie údajov, načítania údajov a periodického obnovovania údajov. Na uľahčenie rozhodovania sú dáta v dátovom sklade usporiadané na základe hlavných predmetov (napr. zákazník, položka, dodávateľ a činnosť). Údaje sa ukladajú, aby poskytli informácie z retrospektívneho hľadiska, napríklad za posledných 6 až 12 mesiacov, a sú zvyčajne zhrnuté. Napríklad namiesto ukladania podrobností o každej predajnej transakcii, môže dátový sklad uložiť súhrn transakcií podľa typu položky pre každý obchod alebo sumárne na vyššej úrovni vedenia podniku, prípadne taktiež pre každý región predaja.

Dátový sklad je zvyčajne modelovaný multidimenziálnou dátovou štruktúrou, ktorá sa nazýva dátová kocka, v ktorej každá dimenzia zodpovedá atribútu alebo množine atribútov v schéme a každá bunka ukladá hodnotu nejakej agregovanej miery, napríklad súčtu alebo súčinu.

Iné druhy dát

Okrem databázových a transakčných údajov a údajov dátového skladu existuje mnoho ďalších druhov dát rôznorodých foriem a štruktúr, ako aj všeobecne rozdielneho sémantického významu. Takýto druh dát môžeme vidieť v širokom spektre aplikácií.

Napríklad časovo súvisiace alebo sekvenčné údaje (napríklad historické záznamy, údaje o burze cenných papierov a údaje o časových radoch a biologických sekvenciách), dátové toky (napríklad údaje z video monitorovania a senzorov, ktoré sú nepretržite prenášané), priestorové údaje (napríklad mapy), údaje o konštrukčnom návrhu (napríklad budovy, komponenty systému alebo integrované obvody), hypertextové a multimediálne údaje (vrátane textových, obrazových, obrazových a zvukových údajov), grafy a údaje v sieti (napríklad sociálne a informačné siete) a web (obrovské a široko distribuované úložisko informácií prístupné internetom). Pri vzniku týchto dát vyvstáva nová výzva, a to akým spôsobom zaobchádzať s dátami rôznych štruktúr a sémantických významov a rovnako aj ako dolovať poznatky, ktoré nesú rozmanité štruktúry a sémantiku. Je dôležité si uvedomiť, že v tomto prípade sa v aplikáciách môžu vyskytovať skupiny dát z rôznych zdrojov disponujúcich veľkou variáciou odlišných atribútov, ktoré na jednej strane nesú potenciál vytvoriť podklad pre analýzu dát bohatú na konečné poznatky vďaka vzájomnému zdokonaľovaniu a konsolidácii týchto zdrojov, avšak na druhej strane ich komplexná náтура taktiež vytvára pre tento proces nepriaznivé podmienky na základe vzniknutých ťažkostí pri čistení a integrácii údajov, ako aj zložitých interakcií medzi viacerými zdrojmi dát.

1.4 Prístupy dataminingu

Po klasifikácii rôznych typov úložísk údajov a informácií, na ktorých je možné vykonávať ťažbu je taktiež potrebné uviesť druhy vzorov, ktoré je možné vytážiť. Zaujímavé vzory predstavujú vedomosti a existuje niekoľko dataminingových techník na dolovanie údajov, ktorých koncovým produktom sú práve údaje a poznatky, ktoré nie sú explicitne pozorovateľné respektíve uložené v dátových súboroch. V rámci týchto metód poznáme charakterizáciu a diskrimináciu, ťažbu častých vzorov, združení a koreláciu, klasifikáciu a regresiu, klastrovú analýzu a analýzu odľahlých hodnôt. Spomenuté metódy dolovania dát sú využívané na špecifikáciu druhov vzorov, ktoré sa nachádzajú v úlohách dolovania dát. Vo všeobecnosti vieme tieto úlohy rozdeliť na základe nimi vytvorenej perspektívy do dvoch skupín : **deskriptívne** a **prediktívne**.¹⁰

¹⁰ Difference Between Descriptive and Predictive Data Mining [online]. [cit. 2021-01-07] Dostupné na: <https://techdifferences.com/difference-between-descriptive-and-predictive-datamining.html#:~:text=Key%20Differences%20Between%20Descriptive%20and,that%20predictions%20can%20be%20made>

Deskriptívna úloha ťažby kladie dôraz na sumarizáciu a transformáciu údajov na zmysluplné informácie pre monitorovanie a ich prezentáciu alebo vizualizáciu. Vytvára korelácie a používa sa na určenie podobnosti údajov a na nájdenie existujúcich vzorcov. Táto úloha je sústredená okolo opisu charakteru údajov v datasete.

Prediktívna úloha na druhej strane umožňuje vykonať indukciu aktuálnych údajov, z ktorej sú po spracovaní výstupom predpokladané údaje budúcich období. V porovnaní s deskriptívnou úlohou, pri prediktívnom prístupe vyvstáva z aktuálnej situácie aj možnosť správnej reakcie na základe predpokladaného diania v budúcom vývoji. Dátové modelovanie je nevyhnutnou súčasťou prediktívnej analýzy, ktorá využíva niektoré premenné na predikciu neznámych budúcich dátových hodnôt pre iné premenné. Prediktívny prístup je využívaný na identifikovanie tej najoptimálnejšej dataminingovej metódy, ktorá čo najefektívnejšie vydoluje požadované vzory.¹¹

1.5 Charakterizácia metód dataminingu

V tejto kapitole sa budeme venovať opisu metód hĺbkovej analýzy dát, ich využitiu a taktiež konečnému výstupu, ktorý počas dolovania vzniká, a teda poznatkom a ich reprezentácii a taktiež podrobnému rozboru metód regresie a asociácie.

Klasifikácia a predikcia

Klasifikácia je proces hľadania modelu, ktorý popisuje a rozlišuje triedy a pojmy údajov. Táto metóda analyzuje nové relevantné informácie v skúmanom súbore a na základe vopred stanovených vlastností prvkov, ktoré slúžia k následnému správne zaradeniu dát, identifikuje do ktorého zo súboru kategórií (subpopulácií) patrí nové pozorovanie, využívajúc súbor údajov obsahujúci pozorovania, ktorých členstvo v kategóriách je známe.

¹¹ KATIYAR, Pranchal. Difference Between Descriptive and Predictive Data Mining [online]. Uverejnené: 17.09.2019 [cit. 2021-01-05] Dostupné na: <https://www.geeksforgeeks.org/difference-between-descriptive-and-predictive-data-mining/>

Je teda využívaný softvér, ktorý prijaté dáta analyzuje a následne triedi do kategórii na základe bodov a atribútov, ktoré jednotlivé prvky súboru vykazovali.¹²

Predikcia funguje na takmer rovnakom princípe, ako klasifikácia s rozdielom, že výstupom nie je model obsahujúci klasifikované informácie, ale funkcia ktorá predpovedá hodnoty atribútu, ktorý bol skúmaný.

Zhluková analýza

Zhluková analýza alebo klastrovanie je úlohou hĺbkovej analýzy dát zoskupiť skupinu vzoriek takým spôsobom, že vzorky v tej istej skupine (nazývanej klaster) sú si navzájom viac podobné v zmysle spoločných atribútov, a kategorizačných znakov, ako v iných skupinách. Táto metóda teda na základe vopred stanovených prvkov rozdeľuje dataset do viacerých skupín, ktoré obsahujú dáta, ktoré sú si práve tak podobné, a teda nie sú zhodné s dátami v iných skupinách.¹³ Pri prvotnom zbere údajov môže byť súbor neusporiadaný a neštruktúrovaný, čo môže v neskoršej fáze a najmä pri snahe využitia takýchto dát spôsobiť problémy. Aby sa takýmto stavom predišlo je využívaná zhluková analýza, ktorá okrem svojej hlavnej funkcie vytvorenia štruktúry v rámci datasetu je taktiež možnosťou objavenia nových skupín v rámci zaznamenaných vzoriek. Je to primárna úloha prieskumnej ťažby údajov a často využívaná technika štatistickej analýzy údajov, ktorá sa používa v mnohých oblastiach, vrátane rozpoznávania vzorov, analýzy obrázkov, získavania informácií, bioinformatiky, kompresie údajov, počítačovej grafiky a strojového učenia.

¹² Basic Concept of Classification (Data Mining) [online]. Uverejnené: 12.12.2019 [cit. 2021-01-15] Dostupné na: <https://www.geeksforgeeks.org/difference-between-descriptive-and-predictive-data-mining/>

¹³ SHARMA, Rohit. Cluster Analysis in Data Mining: Applications, Methods & Requirements [online]. Uverejnené: 20.01.2020 [cit. 2021-02-01] Dostupné na: <https://www.upgrad.com/blog/cluster-analysis-data-mining/#:~:text=Conclusion-,What%20is%20Clustering%20in%20Data%20Mining%3F,the%20similarity%20of%20the%20data.>

Detekcia anomálii

Detekcia anomálii, taktiež často nazývaná „analýza odľahlých hodnôt“ je metóda v dolovaní dát, ktorá identifikuje dátové body, udalosti alebo pozorovania, ktoré sa líšia od bežného správania množiny údajov. V dobe, kedy vďaka dostupným analytickým programom a rôznorodým softvérom na administráciu je pre spoločnosti jednoduchšie než kedykoľvek predtým efektívne merať každý aspekt ekonomickej činnosti a s možnosťou použitia miliónov rôznych metrík, ktoré je možné merať, majú spoločnosti tendenciu skončiť s celkom pôsobivým súborom údajov na preskúmanie ich podnikovej výkonnosti. Tieto konštantne zaznamenávané dáta obsahujú informácie, ktoré reprezentujú podnik pri bežnom chode a každá neočakávaná zmena v týchto údajových vzorkách, respektíve udalosť, ktorá nezodpovedá očakávanému vzorcu údajov, je vykazovaná a považovaná za anomáliu. Inými slovami, anomália je odchýlka od bežného chodu podniku. Anomálne údaje môžu naznačovať kritické udalosti, ako je technická chyba alebo potenciálne príležitosti, akou je napríklad zmena spotrebiteľského správania. Na automatizáciu detekcie anomálii sa postupne používa strojové učenie.

Rozhodovací strom

Tvorba rozhodovacieho stromu je jednou z metód prediktívneho prístupu a modelovania používaného v dataminingu. Hlavným objektom tohto prediktívneho konceptu je práve rozhodovací strom používaný na prechod od pozorovaní položiek v datasete, ktoré sú v tomto strome zastúpené formou vetiev, k záverom o cieľových hodnotách znázornených na listoch. Strom sa zostaví rozdelením datasetu, ktorý je vizualizovaný ako koreňový uzol stromu, na podkategórie, ktoré tvoria podklad pre ďalšie rozdelenia. Toto delenie, inak nazývané aj rekúzia, je vykonávané na základe pravidiel odvodených z klasifikačných znakov a je opakované pre každú odvodenú podmnožinu. Rekúzia je dokončená, keď sú v rámci podmnožiny kategorizované všetky hodnoty cieľovej premennej, alebo keď rozdelenie už nemá ďalšiu pridanú hodnotu pre predpovede. Rozhodovacie stromy delíme na modely, kde cieľová premenná môže mať samostatnú množinu hodnôt. Tento model sa nazýva **klasifikačný strom**; v tejto štruktúre predstavujú listy názvy tried a vetvy konektor prvkov, ktoré vedú k týmto triedam. Rozhodovacie stromy, kde cieľová premenná môže nadobúdať spojité hodnoty (zvyčajne reálne čísla), sú nazývané **regresné stromy**. Rozhodovacie stromy patria vďaka svojej priamočiarosti a zrozumiteľnosti medzi

najobľúbenejšie algoritmy dataminingu a strojového učenia. Nevyžadujú taktiež žiadnu znalosť domény a je jednoduché naučiť sa takéto stromy zhotovovať.

V ďalšej časti práce sme bližšie definovali a opísali metódu regresie a asociácie. Uviedli sme, na akom princípe tieto algoritmy fungujú, ako dáta predpripraviť aby ich použitie bolo efektívne, aké sú výstupy a taktiež príklady ktoré budú ilustrovať vhodné využitie týchto metód.

1.6 Regresia

Regresia je funkcia ťažby dát, ktorá predpovedá numerické hodnoty. Zisk, predaj, hypotekárne úroky, hodnota domu, rozmery plochy v metroch štvorcových, teplota alebo vzdialenosť sú všetko možné výstupy predpovedané pomocou regresných techník. Takáto numerická hodnota je stanovená procesom zhodnotenia charakteristických vlastností skúmaných vzoriek a teda regresný model by sa napríklad mohol použiť na predpovedanie hodnoty domu na základe polohy, počtu izieb, veľkosti pozemku a ďalších faktorov. Regresná metóda sa začína tvorbou a prípravou súboru údajov, v ktorom sú známe cieľové hodnoty. Napríklad na základe pozorovaných údajov mnohých podnikov za určité časové obdobie je možné vytvoriť regresný model, ktorý predpovedá zisk alebo iné ukazovatele ekonomickej činnosti. Okrem zisku môžu údaje sledovať počet zákazníkov, ceny produktov, počet pobočiek, odvedené dane, mestskú časť kde je podnik situovaný, atď.¹⁴

Predpovedaná hodnota je označovaná ako cieľová hodnota, atribúty, na základe ktorých je zostavovaná regresná funkcia, sú nazývané prediktory a individuálne údaje pre každú hodnotu sú uvádzané ako prípady. V procese zostavovania výsledného prediktívneho modelu regresný algoritmus odhaduje hodnotu cieľa ako funkciu prediktorov pre každý prípad v údajoch o zostavení. Tieto spojenia medzi prediktormi a cieľom sú zakomponované v modeli, ktorý je potom možné použiť na iný súbor údajov, v ktorom nie sú cieľové hodnoty známe. Regresné modely sa testujú výpočtom rôznych štatistík, ktoré merali rozdiel medzi predpovedanými hodnotami a želanými hodnotami, respektíve hodnotami, ktoré

¹⁴ GUPTA, Swati. A Regression Modeling Technique on Data Mining. In International Journal of Computer Applications [online]. Foundation of Computer Science, 2015, roč. 9. č. 116, s. 27-29 [cit. 2020-11-20]. ISBN: 973-93-80886-20-6. Dostupné na: <https://research.ijcaonline.org/volume116/number9/pxc3902570.pdf>

očakávame. Regresné modelovanie má veľa aplikácií v analýze trendov, obchodnom plánovaní, marketingu, finančných prognózach, predikciách časových radov, biomedicínskom a farmaceutickom modelovaní, ako aj environmentálnom modelovaní.

Na vývoj a použitie regresných modelov za účelom dolovania údajov nie je potrebné rozumieť matematickým postupom použitým v regresnej analýze, no je však užitočné pochopiť základné pojmy. Princíp metódy regresnej analýzy spočíva v jej snahe zistiť hodnoty parametrov pre funkciu, ktoré spôsobia, že táto funkcia najlepšie zapadne do súboru pozorovaných údajov, ktoré poskytnete. Nižšie vyobrazená rovnica vyjadruje tieto vzťahy v symboloch. Na jej základe zistíme, že regresia je proces odhadu hodnoty spojitého cieľa (y) ako funkcie (F) jedného alebo viacerých atribútov respektíve prediktorov ($x_1, x_2, \dots, x_n$), množiny parametrov ($\theta_1, \theta_2, \dots, \theta_n$) a miery odchýlky (e).¹⁵

$$y = F(x, \theta) + e$$

Pod prediktormi môžeme rozumieť nezávislé premenné a cieľ chápeme ako závislú premennú. Chyba, ktorá je taktiež nazývaná reziduálna, je rozdielom medzi očakávanou a predpokladanou hodnotou závislej premennej. Regresné parametre sú tiež známe ako regresné koeficienty. Proces vytvárania regresného modelu zahŕňa nájdenie hodnôt parametrov, ktoré minimalizujú mieru odchýlky, napríklad reziduálny súčet kvadrátov.

Poznáme rôzne typy regresných funkcií a rôzne spôsoby merania miery odchýlky:

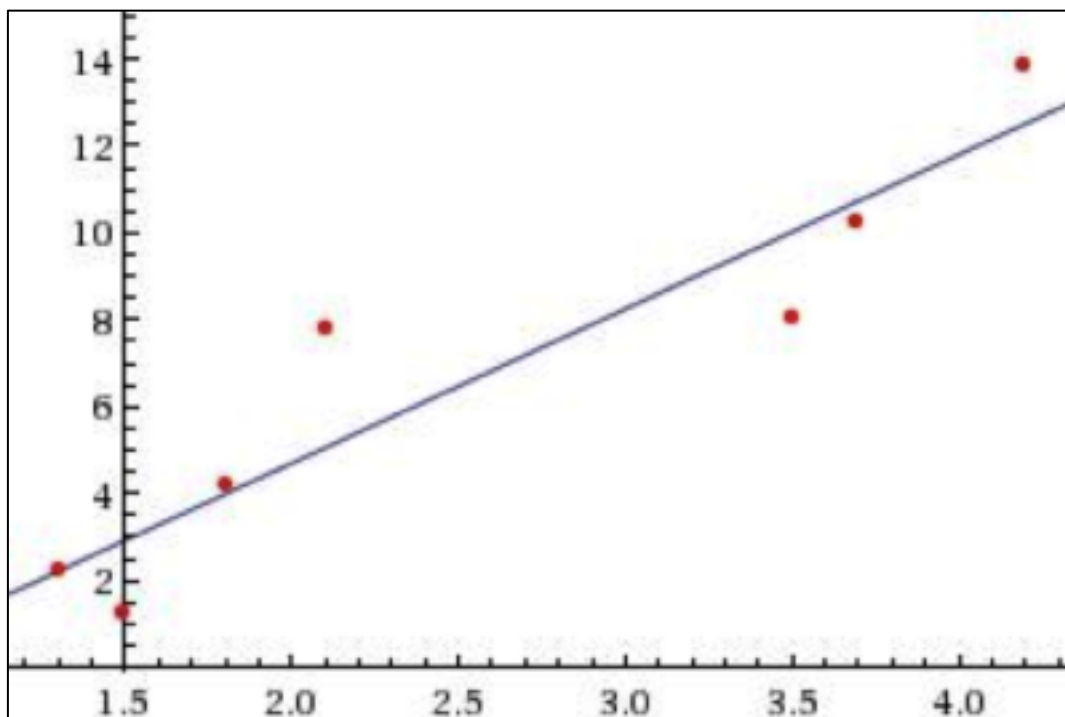
1.6.1 Lineárny regresný model

Najjednoduchšia forma regresie, lineárna regresia, využíva vzorec priamky ($y = mx + b$) a určuje príslušnú hodnotu pre m a b na predpovedanie hodnoty y na základe danej hodnoty x . V zásade lineárne regresné modely sa používajú na zobrazenie alebo predvídanie vzťahu medzi dvoma stanovenými premennými alebo faktormi. Na grafickej prezentácii takéhoto modelu by sme mohli pozorovať dve možné udalosti. Ak by bol tvar čiary zaoblený smerom nahor, znamená to že nezávislá premenná má pozitívny efekt na závislú premennú. Naopak pri zaoblení čiary smerom nadol by bol tento efekt negatívny.

¹⁵ ORACLE Help Center: Regression [online]. In Data Mining Concepts. [cit. 2021-02-22] Dostupné na:

https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/regress.htm#DMCON005

Faktor, ktorý sa predpovedá (pre ktorý rovnica rieši hodnoty) je v tejto rovnici zaznamenaný ako závislá premenná. Faktory, ktoré sa používajú na predpovedanie hodnoty faktora ako závislej premennej sa nazývajú nezávislé premenné.



Obrázok 1 Lineárny regresný model

Zdroj: SVOBODA, Lukáš. Data mining, Posterus, 2003

1.6.2 Viacrozmerný regresný model

Termín viacrozmerná lineárna regresia sa týka lineárnej regresie s dvoma alebo viacerými prediktormi ($x_1, x_2, \dots, x_n$). Ak je do procesu zapojených viac prediktorov, nie je možné regresnú priamku vizualizovať v dvojrozmernom priestore. Priamku je však možné vypočítať jednoduchým rozšírením rovnice pre lineárnu regresiu s jedným prediktorom tak, aby obsahovala parametre každého z prediktorov.

Funkcia tohto typu regresného modelu znie:

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n + e$$

Pri viacrozmernej forme lineárnej regresie sa regresné parametre často označujú ako koeficienty. Keď takýto model zostavujeme, algoritmus počíta koeficient pre každý

z prediktorov v ňom použitých. Koeficient je mierou vplyvu prediktora x na cieľ y . K dispozícii je veľa štatistík na analýzu regresných koeficientov, slúžiacich na vyhodnotenie toho, ako dobre sa regresná čiara hodí k údajom.

1.7 Asociácia

Asociácia, ktorú vo svojom článku analyzoval Festim Halili a Avni Rustemi¹⁶, je zvyčajne aplikovaná pre nachádzanie frekventovanej množiny položiek medzi veľkými súbormi údajov. Tento typ nálezu pomáha podnikom prijímať určité rozhodnutia, ako napríklad návrh katalógov, krížový marketing a analýza správania zákazníkov pri nakupovaní. Algoritmy asociačného pravidla musia byť schopné generovať pravidlá s hodnotami spoľahlivosti menšími ako jedna. Počet možných asociačných pravidiel pre daný súbor údajov je však vo všeobecnosti veľmi vysoký a vysoký podiel pravidiel má zvyčajne malú respektíve nulovú hodnotu. Typy asociačného pravidla:

- Pravidlo viacúrovňovej asociácie
- Multidimenzionálne asociačné pravidlo
- Kvantitatívne asociačné pravidlo

Na základe koncepcie striktných pravidiel zaviedla skupina počítačových expertov asociačné pravidlá pre objavovanie zákonitostí medzi produktmi v rozsiahlych transakčných dátach zaznamenávaných systémami point-of-sale (POS), v preklade bod nákupu, v supermarketoch. Napríklad pravidlo {cibuľa, zemiaky} $\rightarrow$ {burger} nájdené v údajoch o predaji v supermarkete naznačuje, že ak si zákazník kúpi cibuľu a zemiaky spoločne, pravdepodobne si kúpi aj hamburgerové mäso. Tieto informácie môžu byť použité ako základ pre rozhodovanie o marketingových aktivitách, ako sú napríklad propagačné ceny alebo umiestňovanie produktov. Okrem vyššie uvedeného príkladu z analýzy trhových košov sa dnes asociačné pravidlá používajú v mnohých aplikačných oblastiach vrátane dolovania webového využívania, detekcie narušenia, nepretržitej výroby a bioinformatiky.

¹⁶ HALILI, Festim – RUSTEMI, Avni. In Predictive modeling: data mining regression technique applied in a prototype [online]. Telangana: IJCSMC, 2016, roč. 5, č. 8, : 207-215 s. ISSN 2320-088X. Dostupné na: <https://www.ijcsmc.com/docs/papers/August2016/V5I8201642.pdf>

Na rozdiel od sekvenčného dolovania, asociačných pravidiel zvyčajne nezohľadňujú poradie položiek v rámci transakcie ani medzi transakciami.

Výber z uvedených metód sme použili v aplikačnej časti pri riešení podnikovej problematiky.

2 Cieľ práce a metódy skúmania

Hlavným cieľom tejto práce je definovať súčasné poznatky v oblasti hĺbkovej analýzy dát dataminingu a ich následná aplikácia pri tvorbe dataminingových modelov na základe dát vybranej spoločnosti. Následne pomocou analýzy skutočných dát navrhnúť zlepšenie v oblasti tvorby portfólia klientov a umiestnených reklamných spotov spoločnosti Mindshare.

Na dosiahnutie cieľa bakalárskej práce bolo nutné zakomponovať nižšie uvedené čiastkové ciele, ktoré boli v ďalšej časti podrobnejšie opísané:

- definovanie súčasných vedomostí v oblasti dataminingu
- uvedenie rôznych pojmov a ucelených definícií autorov na tému dataminingu
- stanovenie procesných prvkov a ich nevyhnutnosti v kontexte analýzy dát
- uvedenie možných problémov spojených s dolovaním dát
- charakteristika viacerých dataminingových metód
- analýza aktuálnej situácie v podniku
- realizácia hĺbkovej analýzy dát riešením lineárnej regresie a vizualizáciou dát
- kategorizácia kľúčových a nevýnosných zmluvných vzťahov spoločnosti Mindshare
- vyhodnotenie ziskov počas kvartálov fiškálneho roka
- vyhodnotenie analýzy s prihliadnutím na momentálnu situáciu

Jedným z čiastkových cieľov tejto bakalárskej práce bolo definovať súčasné vedomosti prostredia a využívania dataminingu a taktiež zvýšiť povedomie o možnostiach vyplývajúcich z využívania dolovania dát za účelom dosiahnutia výstupov obsahujúcich vedomosti. Tieto vedomosti je ďalej možno využiť pre prijímanie rozhodnutí v rámci podnikateľskej sféry. Základné pojmy a všeobecné informácie o krokoch ktoré predchádzajú už konkrétnemu využitiu rôznych dataminingových metód, sú prostriedkom ako si vytvoriť predstavu o funkcionalite takejto platformy ktorá je čoraz častejšie využívaná naprieč širokým spektrom podnikateľských subjektov. Bližšia definícia daných metód umožňuje na základe vopred vytýčených potrieb a požiadaviek na získanie vedomostí z dátových súborov správne určiť potrebnú stratégiu pre riešenie problému. Ako ďalší čiastkový cieľ, ktorý je pre túto bakalársku prácu vytýčený, je poukázať na fakt, že hĺbkovú analýzu dát je potrebné chápať v jej ucelenej forme ako komplexný proces ktorý sa každým riešeným problémom

líši a nie je teda možné aplikovať už vyhotovené výstupy skúmaní z minulých období na novo vzniknuté požiadavky na analýzu dát. Dáta vnímame ako dynamický element a je preto nutné ku každému skúmanému datasetu pristupovať individuálne a prikladať každému kroku predprípravy dát, ako aj aplikovaniu konkrétnej dataminingovej metódy v rôznych analytických aplikáciách vysokú dôležitosť. V neposlednom rade je potrebné spomenúť, že čiastkovým cieľom je taktiež poukázať na možné problémy, ktoré sa vyskytujú pri analýze dát takých rozmerov, akými dnešné spoločnosti akéhokoľvek veľkostného typu disponujú. V každom prípade ide o problémy najmä technického charakteru. Ďalším čiastkovým cieľom bakalárskej práce bolo charakterizovať niekoľko vybraných metód hĺbkovej analýzy dát za účelom rozšíriť vedomosti a poukázať na možnosti ich využitia pre rôzne účely. Analýza aktuálnej situácie vo vybranom podniku, konkrétne v oblasti reklamných spotov za rok 2019, bola stanovená ako čiastkový cieľ, nakoľko bolo potrebné poukázať na nedostatky, ktoré sme analýzou dát riešili. Vytýčili sme si taktiež čiastkové ciele, ktoré sa priamo týkali riešenia problematiky ktorú sme si pre bakalársku prácu stanovili, a to vytvorenie modelu lineárnej regresnej analýzy a vizualizáciu dát. Pre lepšie pochopenie navrhovaných zmien sme taktiež museli v rámci čiastkových cieľov definovať dve kategórie klientov, do ktorých je potrebné v ďalšom navrhovanom postupe existujúce zmluvné vzťahy so svojimi klientami selektovať. Konkrétne sú to kľúčoví a nevýnosní klienti. Vyhodnotením ziskov v jednotlivých kvartáloch fiškálneho roka 2019 a vyhodnotením vytvorenej analýzy sme zosumarizovali aplikačnú časť, ktorá nám umožnila formulovať potenciálne zlepšenia pre portfóliovú stratégiu vybranej spoločnosti.

Čiastkové ciele bakalárskej práce môžeme kolektívne definovať ako snahu o to, vytvoriť dokument, ktorý čo najefektívnejším spôsobom zdieľa informácie počnúc všeobecným porozumením pojmu „dolovanie dát“, ako aj jeho následné vloženie do kontextu funkcionality naprieč rôznymi segmentami dnešnej spoločnosti, a to najmä na úrovni potrieb podniku. Praktická časť práce je vďaka postupnému rozboru vypracovanej analýzy na reálnych dátach ideálnym príkladom pre potenciálne budúce dataminingové procesy, ktoré by mohli byť vypracovávané v spoločnostiach s rovnakým zameraním a obdobnými požiadavkami na výsledné poznatky nadobudnuté z metadát. Je preto na uvážení kompetentných osôb v rámci manažmentu spoločnosti, či, a ak áno, aké úpravy bude nutné implementovať v rámci momentálne veľmi výkonného portfólia.

2.1. Metódy

Bakalárska práca obsahuje údaje ktoré boli nadobudnuté týmito metódami:

- Literárna rešerš
- Analýza z vybranej domácej a zahraničnej literatúry
- Komparácia a syntéza pri analýze literatúry
- Výber dataminingového softvéru
- Vyhodenie dataminingových modelov pomocou dataminingového softvéru Knime
- Dedukcia na základe dostupných dát spoločnosti Mindshare
- Vytvorenie vizualizácii v programe Power BI

Metódy, ktoré boli využité za účelom dosiahnutia vyššie uvedených cieľov sa mimo štúdia domácej i zahraničnej odbornej literatúry, ktorá zahŕňala nielen poznatky o všeobecnej teoretickej stránke analýzy dát, ale taktiež tvorila základ k porozumeniu strojového učenia a procesom, ktoré sa dejú v pozadí využívaných programov mimo nášho vnímania, vyžadovali taktiež v rámci praktickej časti získať potrebné dataminingové softvéry Knime a Power BI. Tieto programy umožnili aplikovať rôzne metódy analýzy dát, ako aj vytvoriť vizualizácie výsledných modelov, čím bolo možné v praxi pozorovať ako deskriptívny, tak prediktívny prístup dataminingu. Pre stanovenie problémov ktoré sú riešené v praktickej časti bolo taktiež nutné prioritizovať informácie obsiahnuté v databáze, ktoré boli v programoch stanovené ako kľúčové atribúty pri riešení regresného modelu a rovnako aj vizualizácie dát. Podkladom pre údaje obsiahnuté v prvej časti bakalárskej práce boli relevantné domáce, ako aj zahraničné odborné literárne diela, ktoré využitím komparatívneho prístupu vytvorili dvojstranný ucelený obraz o momentálnom poznaní prostredia strojového učenia. Neopomenuteľným zdrojom boli taktiež zahraničné odborné články, ktoré detailne rozoberali rôzne elementy dataminingového procesu. Nakoľko je pohľad na problematiku konkrétnych fáz dátovej analýzy v individuálnom podaní autorov rozdielny, je tu opäť predpoklad pre mnohostrannú charakteristiku utvorenú z diferencovaných názorov týchto autorov. Pre čo najrelevantnejší výstup boli stanovené problémy, ktoré by vypracovanou analýzou mali byť riešené vo vybranej spoločnosti za účelom prijímania lepších rozhodnutí na úrovni stredného manažmentu, ako aj za účelom reportingu. Predikčné modely a dátové vizualizácie boli vyhotovené v softvéroch, ktoré by mali odzrkadľovať čo najideálnejšiu kombináciu atribútov, ktoré sú uvedené a zhodnotené v poslednej časti bakalárskej práce a to najmä z dôvodu čo najefektívnejšej práce a čo

najideálnejšieho možného využitia takejto analýzy za rovnakých podmienok. Nakoľko oba programy fungujú na rozdielnom rozhraní, spĺňajú iné funkcie za iným účelom a oba obsahujú odlišné komponenty, bolo taktiež nutné nadobudnúť vedomosti v oblasti ovládania a zaobchádzania s nimi.

V ďalšej kapitole sme aplikovali metódy dataminingu na skutočné dáta a následne sme ohodnotili mieru splnenia vytýčených cieľov.

3 Výsledky práce a diskusia

V časti Výsledky práce a diskusia sa zaoberáme celým procesom výberu a špecifikácie problému ktorý sme riešili na dátach poskytnutých spoločnosťou, ktorú v prvej časti tejto kapitoly patrične predstavíme a uvedieme všetky potrebné informácie ktoré predchádzali procesu mnou vypracovanej analýzy dát, problémov s ktorými sme sa pri tejto analýze stretli a taktiež v grafickej podobe znázorníme výsledky mojej práce. Všetko za podporného textu v ktorom zdôrazníme potrebu dolovania týchto dát, softvér ktorý sme si pre túto analýzu vybrali a taktiež k výstupom dolovania týchto dát poskytneme komentár.

3.1 Predstavenie spoločnosti

Bakalársku prácu riešime v spoločnosti Mindshare. Organizačná štruktúra tejto spoločnosti, konkrétne jej pobočky na Slovensku s lokalitou v Bratislave, pozostáva z jej riaditeľa, ktorému podliehajú manažéri jednotlivých oddelení a následne členovia jednotlivých tímov. V Bratislave sa nachádzajú dva druhy tímov, ktoré majú na starosti rôzne skupiny klientov. Sú to Client Leadership tímy, ktoré zaobstarávajú kompletnú komunikáciu s klientom a Digital tímy ktoré na základe komunikovaných objednávok realizujú požiadavky klienta komunikáciou s médiami, ktoré potenciálne umiestnia reklamy týchto klientov. Mediálna agentúra Mindshare je globálna spoločnosť s pobočkami vo viac ako osemdesiatich krajinách sveta. Ich cieľom je ponúknuť tie najlepšie riešenia v oblasti marketingovej komunikácie týkajúcej sa reklám a umiestňovať reklamné kampane svojich klientov v mediálnom priestore. Klientami tejto mediálnej agentúry sú spoločnosti z rôznych odvetví. Vieme túto spoločnosť taktiež vnímať aj ako sprostredkovateľa kontaktu medzi klientom požadujúcim mediálny priestor a konkrétnym médiom, v ktorom bude reklama umiestnená. Produktom takejto firmy je konečná stratégia komunikácie v médiách ktorá zahŕňa podrobnú expertízu v podobe odporúčania alokácie rozpočtu a spôsobu rozloženia reklamných prostriedkov, pre ktoré si klient žiada vyhradiť tento priestor v médiách. Pri globálnych spoločnostiach podnikajúcich v tomto segmente sú interné procesy zväčša rozdielne, avšak využívanie rôznych súborov dát a ich analýza je základom pre všetky. Ide o využívanie veľkých súborov dát, ktoré obsahujú nielen databázy so zaznamenanými údajmi klientov, s ktorými spoločnosť v minulosti nadviazala alebo v blízkej budúcnosti nadviaže dohodu o spolupráci, ale taktiež sú to aj dáta zaznamenané pri konkrétnych transakciách vykonaných po úspešnom ukončení negociácii, dáta o dostupných mediálnych priestoroch, interné údaje spoločnosti spojené s operačnou stránkou spoločnosti

a v neposlednom rade taktiež databázy, ktoré obsahujú detaily uskutočnených obchodov a umiestnených reklamných spotov. Tieto pre mnohých irelevantné záznamy sú pre agentúru Mindshare základným zdrojom informácií a dát, ktoré sú v rámci spoločnosti mnohostranne využívané. Pre spoločnosť Mindshare je dôležité pristupovať ku každej objednávke tak, aby proces podliehal najvyšším štandardom tejto agentúry. Preto musia okrem iných aspektov rovnako aj všetky databázy ako súčasť interne spracovaného a využívaného portfólia údajov a vedomostí byť za každých podmienok prístupné a najmä vyobrazujúce ako aktuálne, tak aj pravdivé informácie. Tieto údaje sú vo viacerých fázach procesu spracovávania jednotlivých objednávok využívané mnohými spôsobmi naprieč rôznymi oddeleniami, ktoré sa na tomto procesnom celku podieľajú svojím vstupom.

3.2 Analýza aktuálneho stavu v podniku v oblasti reklamných spotov za rok 2019

V podniku, ktorý sa zaoberá problematikou riešenia stratégie umiestňovania reklám v mediálnom priestore, a preto teda poskytovania abstraktných výstupov je dôležité brať ohľad na viacero ukazovateľov, ktoré môžu v zásadnej miere zasiahnuť do výberu a následného poskytnutia správnej a optimálnej stratégie. Mimo iného môžu taktiež ovplyvniť okolnosti a detaily zmluvného vzťahu s klientom. Sú to práve trendy minulých období a predikcia ich opätovného výskytu, respektíve predpoveď budúcich trendov a anomálii, ktoré sú dôležitými ukazovateľmi rentability projektov a zmlúv. Spoločnosťou Mindshare nám bola poskytnutá databáza, ktorá zaznamenáva podrobné detaily týkajúce sa medializovaných reklamných spotov z roku 2019. Túto databázu bližšie predstavíme v ďalšej kapitole. Poznatky z prediktívneho prístupu dataminingu sú teda v prípade tejto databázy podnetmi k prípadnej zmene a prijatiu rozhodnutí za účelom exitujúce portfólio optimalizovať.

Popri možnostiach výstupy a vedomosti poskytnuté aplikovaním dataminingu využiť na iniciovanie zmeny akéhokoľvek charakteru, je taktiež možné využiť ich pre posúdenie ďalšieho postupu spoločnosti v zmysle expanzie, respektíve prehodnotenia poskytovaných produktov. Takáto zmena by mohla nastať za účelom vylepšenia dostupnosti, dosahu a v konečnom dôsledku budúcej ziskovosti podniku. Využitím regresných modelov, ktoré by na základe jedného z ukazovateľov pomohli identifikovať silné a slabé odvetvia a medzery v momentálnej stratégii podniku, by sa mohli doceliť práve takéto úpravy, ktoré by danému podniku mimoriadne napomohli v jeho podnikateľskej činnosti. Ako podporný

aspekt je možné využiť vizualizované dáta, ktorých primárnou úlohou je výsledky ukázať vo forme, ktorá nám umožní lepšie pochopiť momentálnu situáciu a sledovanosť v rámci detailnejšie rozdelených demografických skupín, ktorá v konečnom dôsledku ovplyvňuje, akým spôsobom je distribuovaný relatívny podiel na výnosoch agentúry v tomto segmente. Je taktiež podstatné disponovať možnosťou spätnej kontroly pre účtovné účely, a práve v tomto smere je opätovné využitie vizualizácii ako aj vypracovaných modelov, ktoré vedia porovnať cieľové atribúty za sledované obdobie merané v kvartáloch alebo inom požadovanom časovom rozmedzí veľmi užitočné.

Jedným z dôležitých ukazovateľov ziskovosti tejto spoločnosti je rentabilita reklamných spotov v rámci rôznych demografických skupín. Na základe tohto ukazovateľa vieme presne určiť aj ďalšie parametre, ako napríklad kedy reklamu vysielat' alebo v akom poradí usporiadať reklamy v daných blokoch. Jedným z riešených problémov preto bude analýza rentability reklamných spotov v rámci vybranej demografickej skupiny A 12-54. Písmenko A ktoré v tomto kontexte predstavuje anglické „all“, teda v preklade „všetci“, sleduje veľmi širokú demografickú skupinu vo vekovom rozmedzí 12 až 54 rokov a považovali sme preto za podstatné určiť, aké zastúpenie a pri akých reklamných spotoch je najrentabilnejšie a na základe vzťahov medzi týmito ukazovateľmi sledovať trend vývoja týchto aspektov. Ako ďalší problém ktorý je riešený v tejto bakalárskej práci je následná analýza ziskov rozdelená do kvartálov fiškálneho roka. Považovali sme za potrebné tieto dáta vizualizovať, nakoľko sú tieto údaje podnetom pre ďalšie rozhodnutia pri tvorení budgetu a finančného, ako aj projektového plánovania. Vo forme grafov vieme tento krok efektívnejšie zrealizovať vo forme prezentácie ktorú by sme mohli poskytnúť výkonnému manažmentu spoločnosti Mindshare.

Vyšpecifikovanie nedostatkov

Dôvodom riešenia práve problematiky rentability reklamných spotov v rámci konkrétnej demografickej skupiny v spojitosti s vizualizáciou ziskov naprieč kvartálmi fiškálneho roka je absencia analýzy uplynulých období pre účely úpravy stratégie a portfólia klientov. Podniky, ktoré sú v ich činnosti závislé na správaní a rozhodnutiach konzumentov, disponujú veľkými počtami reportov a databáz, ktoré je možné premeniť na poznatky. V tejto súvislosti je tvorených nespočetne veľa analýz, avšak nie vždy je jednou z nich analýza rentability medializovaných reklamných spotov. Tento nedostatok je potrebné pokryť práve takou analýzou, akú sme vypracovali v tejto záverečnej práci. Jej výstupom

sú okrem konkrétnych údajov, o ktorých budeme diskutovať v rozboroch konkrétnych analytických modelov, najmä možnosti ktoré sa naskytnú kompetentným osobám. Sú to možnosti zdokonaľiť existujúci mix klientov zmenami ktoré pomôžu spoločnosti mimo iného najmä po ekonomickej stránke. Dostupné dáta je potrebné za každých okolností využívať čo najefektívnejšie vo svoj prospech. Mnohokrát je práve táto činnosť kľúčovým rozdielom medzi lídrom a druhotnými spoločnosťami podnikajúcimi v príslušnej podnikateľskej sfére.

3.3 Problematika rentability reklamných spotov a ziskovosti v jednotlivých kvartáloch roku 2019

Spoločnosťou Mindshare nám bol poskytnutý súbor formátu Excel ktorý zaznamenával rôzne údaje týkajúce sa reklamných spotov, ktoré počas fiškálneho roku 2019 táto spoločnosť pre svojich klientov umiestnila do mediálneho priestoru. Tento súbor sa skladá z dvadsiaticich stĺpcov, v ktorých vieme nájsť údaje od tých najelementárnejších, ktorými sú názvy klientov, cez názvy konkrétnych spotov, názvy televíznych staníc na ktorých boli tieto reklamné spoty vysielané, cez detailnejšie a zložitejšie údaje akými sú formáty reklamných blokov, respektíve ich dĺžka. Ďalšie z týchto detailných údajov sú programy, ktoré týmto blokom predchádzali až po zastúpenia sledovanosti v rámci rôznych demografických skupín a ich podiel na celkových výnosoch reklamných spotov umiestňovaných v roku 2019 touto mediálnou agentúrou. Obsahuje tridsať tri tisíc riadkov vyššie spomenutých údajov. Pre účely aplikácie dataminingu boli využité iba stĺpce obsahujúce údaje: Názov brandu, Kvartál a TRP bloku A 12-54.

	B	G	O
11			TRP bloku
12	Názov brandu	Kvartál	A 12-54
13	Σ		23572,3
14	CARTE D'OR	II	0,2
15	CARTE D'OR	II	0,7
16	CARTE D'OR	II	0
17	CARTE D'OR	II	0
18	CARTE D'OR	II	0,5
19	CARTE D'OR	II	0,3
20	CARTE D'OR	II	0

Obrázok 2 Údaje využité na účely dataminingu

Zdroj: Vlastné spracovanie, snímka obrazovky

Dáta, ktoré sú v rámci celku uložené v dátovom sklade spoločnosti Mindshare sú využívané nielen na tvorbu reportov, ktoré vedenie spoločnosti informujú o údajoch z predošlých fiškálnych alebo iných sledovaných období a ich komparácii, ale aj na priame podieľanie sa na finálnom produkte ktorý táto spoločnosť pri svojom fungovaní vytvára. Vieme teda s určitosťou povedať, že v dobe kedy sú dáta všetkých formátov generované v rámci mnohých aspektov nášho každodenného života, sú to práve tieto údaje ktoré boli v minulosti získané a v prítomnosti dopĺňané respektíve aktualizované, ktoré sú platformou a hybnou silou pre relatívne tak mladý trh, ako sú médiá a konkrétne mediálne agentúry.

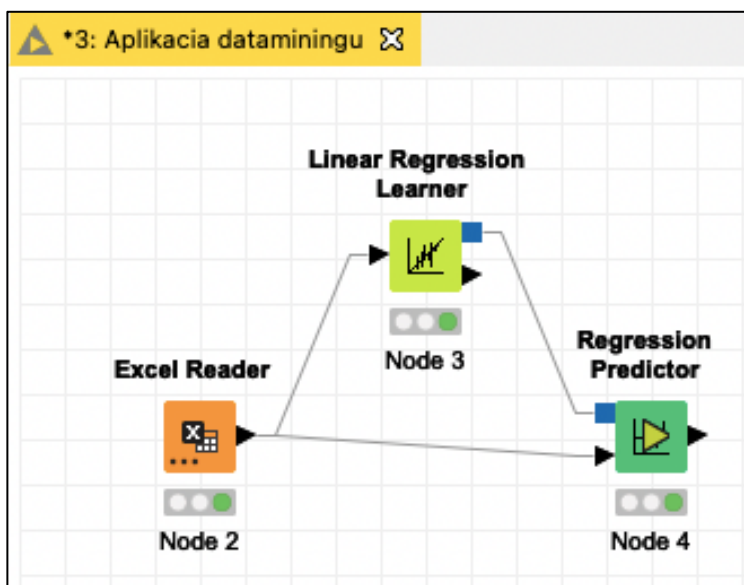
3.3.1 Riešenie problematiky sledovanosti a rentability v rámci demografickej skupiny A 12-54 pomocou lineárnej regresie

V tejto časti je objasnené využitie získaných dát na vytvorenie regresného modelu. Pri riešení tejto problematiky sme sa rozhodli využiť lineárny regresný model, nakoľko je najčastejšie využívaný práve pri analýze vzťahov medzi premennými, čo bola v prípade analýzy sledovanosti a rentability v rámci danej demografickej skupiny tá najvýhodnejšia metóda analýzy. Na úvod je nutné uviesť kroky, ktoré bolo potrebné podstúpiť pred samotným dolovaním dát, ako aj rozhodovací proces a schému, ktorá bola na základe požiadaviek pre výsledný výstup vyhotovená za účelom konformity s potrebami podniku. Bola priložená taktiež grafická podoba tohto modelu, ako aj podporný komentár ku každému z týchto krokov.

Príprava pre vytvorenie tohto predikčného modelu zahŕňala získanie dát od spoločnosti, ktorá obdobné analýzy využíva pri jej každodenných činnostiach a interných rozhodnutiach. Bolo teda potrebné stanoviť si predpoklad a požiadavky, ktoré pomohli následnou dedukciou z užšieho okruhu vybrať spoločnosť Mindshare. Výhodou pri tomto rozhodovaní bol najmä fakt, že hĺbková analýza dát je širokospektrálne využívaná naprieč takmer všetkými odvetvami. Skutočnosť, že dáta na ktorých sú v časti Výsledky práce vypracované dataminingové modely pochádzajú od spoločnosti, ktorej podnikateľská činnosť je vykonávaná v rámci jednej z najmladších oblastí biznisu nasvedčujú, že analýza a jej detailná schéma v rámci dataminingových softvérov v kapitolách nižšie je aplikovateľná na dáta podobného charakteru iných firiem. Je teda možné v týchto programoch jednoducho obmieňať používané súbory, či už zmenou kmeňového podniku alebo zmenou databázy vytvorenej za iné fiškálne roky. Dôvod a charakter potreby, ktorá by

bola podnetom pre vytvorenie takýchto modelov sa pre každú spoločnosť, jej stratégiu ale aj interné procesy líšia, avšak výhodou celkového procesu dataminingu je, že vydolované dáta vieme následne vizualizovať a ich výpovedná hodnota môže s každou vizualizáciou meniť svoje využitie pre potreby firmy. Ako ďalší krok prípravy tejto databázy a samotného prostredia v ktorom sa analýza vykonávala, bolo potrebné okrem vyhodnotenia aktuality a relevantnosti jednotlivých stĺpcov obsahujúcich dáta, vyriešiť operačné požiadavky ktoré vyvstali z tohto úkonu vo zvolenom softvéri. Je totiž nutné zdôrazniť, že prostredie, ako aj jazyk a funkcionality sa pri každom jednom programe využívanom na dolovanie dát, nech už ide o platené alebo open source verzie, líši v rôznych stupňoch na základe rôznych komponentov ktoré tieto programy obsahujú. Ide o odlišné funkcionality a jednou z prvých, a zároveň pravdepodobne najpodstatnejšou je kapacita, ktorou tieto softvéry disponujú – tento aspekt taktiež pôsobí ako jedno z hlavných kritérií pri výbere programu na hĺbkovú analýzu dát, nakoľko je pravdepodobné, že súbory ktoré budú do tohto procesu vstupovať budú obsahovať niekoľko desiatok tisíc riadkov rôznych informácií a pri nedostatočnej kapacite softvéru by bolo nemožné takýto súbor analyzovať.

Po výbere adekvátneho softvéru bolo potrebné získať znalosti v oblasti ovládania a navigácie v priestore softvéru Knime. Je potrebné pochopiť, že každý program funguje na inej báze a príkazy pre jednotlivé analýzy sú zadávané iným spôsobom. Úspešne splniť aj tento krok prípravy bolo možné štúdiom integrovaného manuálu a inštrukcií, ktoré tento program ponúka pri prvotnom otvorení aplikácie. Aby však bolo možné nadobudnúť vedomosti z už úspešne vyhotovených procesov, využili sme taktiež platformu YouTube na sledovanie rôznych analýz a postupných vytváraní schém, ktorými je v programe Knime možné naprogramovať celý proces, a teda preniesť požiadavky formou rôznych dataminingových metód aplikovaných na dataset do konečného výstupu – vedomostí.



Obrázok 3: Pracovné prostredie a schéma v programe Knime

Zdroj: Vlastné spracovanie dát spoločnosti Mindshare v programe Knime

Zostavenie samotnej schémy procesu pre vytvorenie lineárneho regresného modelu sa odohrávalo na pracovnej ploche softvéru Knime ktorá slúži ako interaktívna tabuľa na ktorú môže užívateľ z ponuky rôznych komponentov načítat variáciu prvkov, ktoré krok po kroku dotvárajú po jednotlivých fázach kompletný proces dolovania dát. Aby sme mohli lepšie pochopiť prácu v tomto programe, je nutné upresniť, že všetky jednotlivé komponenty predstavujú uzly a vzťahy sú medzi nimi tvorené manuálne používateľom ktorý vie podľa potrieb pomocou spojnic vytvoriť už spomenutú schému, respektíve ak to situácia a potreby podniku vyžadujú, je možné na tejto pracovnej ploche vytvoriť viacero schém ktoré budú navzájom nezávisle vypracované zo zvoleného zdroja dát, ktorý sme prvotne načítali v potrebnom formáte do programu Knime. Individuálne tieto komponenty taktiež potrebujú programovanie, nakoľko si tento proces chceme do čo najväčšej miery prispôbiť aby výstupy plne zodpovedali našim požiadavkám. Samostatne stojace komponenty/uzly/funkcie ktoré sme načítali na pracovnú plochu nevykonávajú funkciu, ktorá by v tejto forme bola prínosom pre celkový proces dolovania dát, a preto je nutné sa pri každom komponente postarať o nevyhnutnú individuálnu úpravu fáz v rámci týchto komponentov. Každý z načítaných komponentov prechádza troma fázami: načítanie, nastavenie, vykonanie. Pre kompletný proces je nutné, aby všetky zakomponované prvky prešli všetkými troma fázami a pod ich ikonou svietila zelená farba ktorá signalizuje, že daný komponent má všetky interné operatívne úlohy splnené a môžeme ho ďalej začleniť do našej schémy. Prvá fáza je splnená automaticky v momente, kedy užívateľ načíta konkrétny

komponent na pracovnú plochu, druhá fáza vyžaduje vstup užívateľa, ktorý modifikuje potrebné prednastavenia, ktoré budeme bližšie špecifikovať v ďalších riadkoch. Po splnení tejto fázy zvolíme príkaz „Vykonaj“ a následne tento komponent prejde do tretej fázy, ktorá, ako sme vyššie konštatovali, má za úlohu naznačiť skompletizovanie všetkých potrebných krokov prípravy daného prvku. Program Knime nám po otvorení okna možností jednotlivých komponentov a následného výberu nastavení z ponuky možností, umožňuje modifikovať rôzne nastavenia týchto uzlov. Jedná sa o pomerne široké spektrum nastavení od výberu plochy dát, ktorá bude do analýzy zapojená v prvotnom komponente „Excel Reader“, do ktorého sme súbor nahrali a následne špecifikovali, ktoré bunky budú zapojené do procesu hĺbkovej analýzy, cez nastavenia konkrétnych kategórií jednotlivých stĺpcov aby mohol program jednoduchšie rozpoznať či sa bude jednať o číselné, alebo slovné hodnoty, až po nastavenia želanej formy výstupu ktorú sme mohli zobrazit' v grafickej podobe. Program Knime taktiež umožňuje zobrazit' tabuľku súboru Excel obsahujúcu tieto analyzované dáta už po aplikovaní nami zvolenej metódy. Je taktiež možné použiť metódy ako rozhodovací strom, analýzu diskriminantov, klasifikáciu a mnoho ďalších.

Faktom však zostáva, že priestor na pochybenie je nepopierateľne prítomný a súčasne je pomerne veľký. Je to najmä z dôvodu, že samotný softvér je schopný detegovať nesprávne nastavenia iba do úrovne, kedy mu neznemožňujú vykonať samotnú hĺbkovú analýzu dát. Užívateľ však môže po preskúmaní výsledného modelu schému poopraviť bez akejkoľvek straty dát. Je to taktiež jedna z pozitívnych vlastností tohto programu aj keď patrí do kategórie korekcií a nemusí k nej prísť pri každom skúmaní dát. Keďže finálny proces skladá samotný používateľ z niekoľkých fáz, ktoré zohrávajú v tomto ucelenom procese veľmi dôležité úlohy, je možné identifikovať, v ktorom uzle nastala chyba a následne upraviť iba potrebný komponent bez toho aby bol narušený tok dát alebo funkcia iných komponentov. Akonáhle je chyba odstránená, môže sa celý proces spustiť nanovo a môže sa posúdiť či v tomto prípade bola analýza úspešná. Pri príprave dát poskytnutých spoločnosťou Mindshare nebolo nutné využiť žiadnu formu čistenia dát, nakoľko nám súbor bol poslaný s očistenými a zaručene relevantnými dátami. Jednou z možných funkcionalít, ktoré program Knime ponúka, je na pracovnej ploche do schémy zapojiť aj komponenty, ktoré majú na starosti úpravu dát a dátových súborov. Vedeli by sme pomocou týchto uzlov napríklad databázu rozdeliť na základe rôznych kritérií, vedeli by sme v nej vyhľadať určité hodnoty a nahradiť ich inými ale rovnako by sme vedeli databázu prispôbiť požiadavkám jednotlivých metód ktoré by sme sa rozhodli aplikovať. Nakoľko bol tento krok irelevantný,

rozhodli sme sa žiadnym spôsobom nezasahovať do formy dát v ktorej boli do programu načítané, aby sme predišli akejkol'vek strate alebo možnému poškodeniu dát. Naopak pre čo najpresnejšie a najspoľahlivejšie výsledky sme analýzu metódou lineárnej regresie na databáze spoločnosti Mindshare spustili tri krát, aby sme dokázali anulovať prípadné nepresnosti a problémy. Zároveň sme si týmto spôsobom vedeli overiť, že sú skutočne údaje pri každej analýze rovnaké a to aj za cenu dlhšieho celkového času a nižšej časovej efektívnosti. Faktor času však v tomto prípade nezohrával dôležitú rolu. Uviedli sme už, že Knime funguje na princípe pracovnej plochy na ktorú sa spôsobom drag-and-drop/potiahni a pusti dajú načítať rôzne komponenty, ktoré v ucelenom procese pôsobia ako príkazy. Doplnková funkcionálna softvéru Knime je taktiež možnosť analyzovať dáta zo štatistickej perspektívy. Využili sme preto možnosť vizualizácie skúmaných dát aj v tejto rovine, aby sme podporili výstupy ktoré sú výsledkom tejto práce. Pridať takýto analytický prvok bolo veľmi jednoduché, nakoľko je štatistika zahrnutá v jedinom komponente, ktorý bolo potrebné načítať na pracovnú plochu programu a vytvoriť vzťah medzi ním a načítanou databázou.

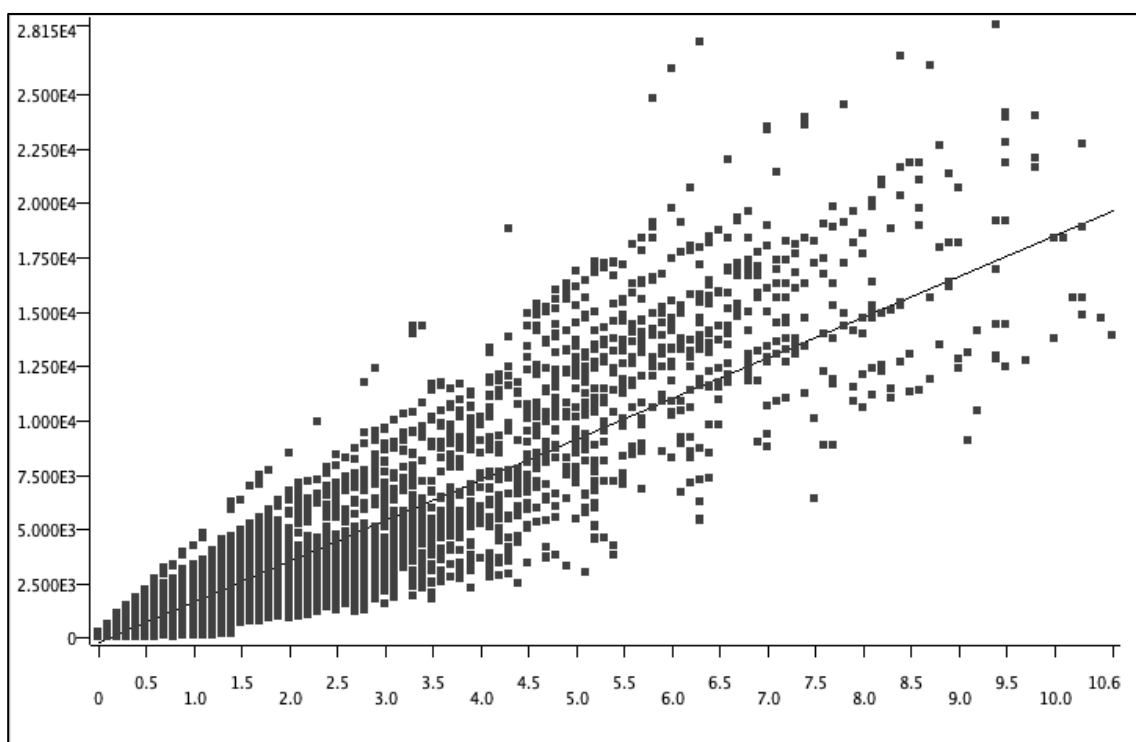
Na sledovanie vzťahu dvoch premenných a ovplyvňovania zmeny hodnôt jednej veličiny pri rôznych hodnotách a vstupoch druhej je ideálne využiť lineárnu regresnú analýzu. Táto metóda nám umožnila sledovať prepojenie a vytvoriť predikčný model a predpovedať, aké kauzálne vzťahy existujú medzi dvomi, nami zvolenými veličinami ktoré v tomto prípade budú kvartály fiškálneho roka v ktorých boli tieto spoty vysielané v médiách ,cena reklamy ktorá nám v konečnom dôsledku taktiež relatívne prezradí, ktorý kvartál akým podielom prispel do celkovej sumy, ktorá sa vo forme zárobku akumulovala počas celého roka a taktiež veková skupina A 12-54, ktorá nám v grafickej podobe ukáže, aký podiel mala táto demografická skupina na zárobkoch spoločnosti Mindshare v oblasti reklamných spotov. Budeme teda skúmať, ako sa vyvíja relatívny príjem individuálnych reklamných spotov pri zmene zastúpenia vybranej cieľovej demografickej skupiny. Pri vytváraní lineárneho regresného modelu v dataminingovom programe Knime bolo potrebné súbor dostať do správnej formy a to z dôvodu, že si tento softvér vyžadoval prácu s csv. (comma separated values) formátom. Nakoľko bola poskytnutá databáza vo svojej prvotnej forme očistená od irelevantných dát, bolo možné upustiť od prípravných procesov a bolo teda možné pristúpiť priamo k procesu aplikovania metód dolovania dát. Je dôležité poznamenať, že sledovanosť v rámci jednotlivých demografických blokov je zaznamenávaná v podobe TRP bodov. Cieľový hodnotiaci bod (TRP), ktorý je taktiež nazývaný televízny hodnotiaci

bod, je metrika používaná v marketingu a reklame na označenie percenta cieľového publika osloveného kampanou alebo reklamného spotu prostredníctvom komunikačného média. Tieto čísla sa považujú za vzorku od celkových vlastníkov televízorov v rôznych geografických a demografických sektoroch. Skúmaná vzorka v poskytnutej databáze má teda pre každý reklamný spot priradenú práve jednu hodnotu ktorá pre nás bude relatívnym ukazovateľom z celkovej sledovanosti. Ďalším podstatným aspektom ktorý nemôže byť opomenutý je fakt, že ak chce užívateľ použiť regresiu na predpovedanie alebo na odvodenie kauzálnych vzťahov, musí starostlivo zdôvodniť, prečo majú existujúce vzťahy prediktívnu silu pre kontext analýzy alebo prečo má vzťah medzi dvoma premennými kauzálnu interpretáciu. Táto zmienka je obzvlášť dôležitá, nakoľko chceme odhadnúť kauzálne vzťahy pomocou analýzy pozorovacích údajov.

Po analyzovaní dát a vytvorení lineárneho regresného modelu sme mohli pozorovať výsledný trend, ktorý naznačuje hneď niekoľko faktov a podporuje niekoľko tvrdení. Výstup ukázal, že pri nižšom zastúpení skúmanej demografickej skupiny, v pomyslenej skupine 0 – 4 TRP, sú príjmy za vysielané reklamy v komunikačných médiách pomerne stabilné a hýbu sa vo veľmi malej odchýlke od regresnej priamky, ktorá je naznačená v grafe nižšie. Body ktoré sú viditeľné na tomto grafe značia priesek určitej hodnoty zastúpenia v rámci demografickej skupiny a jej relatívnych príjmov ktoré generujú naprieč portfóliom. Čím sa posúvame ďalej a postupujeme k vyššiemu zastúpeniu začína sa prejavovať pomerne kolísavý, ale taktiež pozitívne vyvíjajúci sa trend. Výstup v tomto prípade, konkrétne pri zastúpení v pomere od cirka 4.0 do 7.0 televíznych hodnotiacich bodov, značí, že trend sledovaný v rámci vzorky s nižšou hodnotou TRP je aplikovateľný aj na túto kategóriu a je teda jasné, že i keď niektoré reklamné spoty produkujú nižšie zárobky, drvivá väčšina bodov zastupujúcich spoty, ktoré sú sledované touto kategóriou demografickej skupiny A 12-54, sa nachádza nad regresnou priamkou a vo všeobecnosti je teda táto skupina nadmieru zároboková.

Je zreteľné, že body s podpriemernými hodnotami sa vyskytujú najmä v oblasti prvej sledovanej kategórie 0 – 4-0 TRP, kedy síce máme možnosť sledovať prvotné náznaky ďalej pozitívne smerujúcej krivky, avšak nakoľko tieto hodnoty značia pomerne nízke zastúpenie, je taktiež pravdepodobné, že kumulovaný zisk v tomto sektore nebude dosť vysoký na to, aby naprieč touto kategóriou boli priesečky TRP a ziskov z reklám na dostatočne vysokej úrovni. Avšak v kategórii od 4.0 do 7.0 TRP, kedy je zastúpenie relatívne vyššie ako v prvej

skupine a dostávame sa k číslam, ktoré naznačujú pomerne dobrú sledovanosť, vidíme vzor ktorý potvrdzuje naše predchádzajúce tvrdenie, nakoľko sa body vyskytujú v prevažnej väčšine nad regresnou priamkou, a teda značia nadpriemerný alebo očakávaný výnos. Rovnaký trend môžeme sledovať aj v poslednej kategórii rozdelenia hodnôt osy x, a teda môžeme označiť reklamné spoty ktoré sa vyskytujú v prvej kategórii za potrebné, avšak minimálne výnosné.



Graf 1 Lineárny regresný model

Zdroj: Vlastné spracovanie dát spoločnosti Mindshare v programe Knime

Nakoniec detailne skúmame aj skupinu, ktorá je v zastúpení od hodnoty 7.0 až po konečnú hodnotu stupnice a vidíme prvky kontinuity z predošle komentovaných kategórii. Aj napriek tomu, že takúto vysokú koncentráciu sledovanosti nemá veľké množstvo reklamných spotov v portfóliu spoločnosti Mindshare, je stále prevažná väčšina z nich nad regresnou priamkou. Vieme opäť konštatovať, že v momentálnom nastavení sú všetky nami stanovené kategórie v rámci tejto demografickej skupiny veľmi balansované a vyhlíadky do budúcich období signalizujú pozitívne výsledky. Je nutné dodať, že tieto skupiny sú ovplyvňované mnohými externými faktormi a situácia, ktorá sa naskytla s vypuknutím celosvetovej pandemickej situácie je viac než pravdepodobné, že i bez akéhokoľvek upravenia portfólia a zmien v rámci zmluvných podmienok, budú dáta reflektujúce rok 2020 vykazovať pri rovnakej analýze úmerne lepšie výsledky nakoľko sú televízie

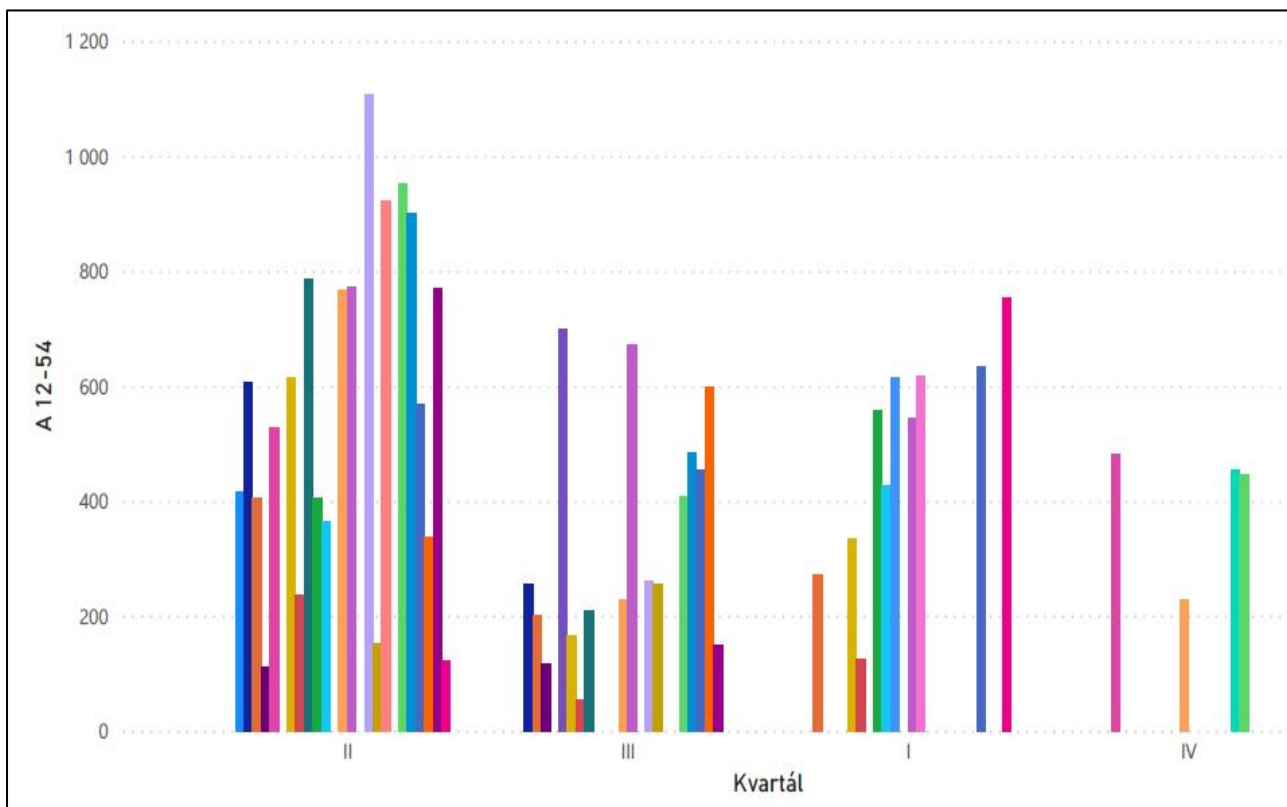
v domácnostiach momentálne využívané viac, než tomu bolo pred touto situáciou, kedy široká spoločnosť nemala rovnaké nadpriemerné množstvo dostupného času na sledovanie televízie.

3.3.2 Riešenie distribúcie ziskov z reklám v jednotlivých kvartáloch fiškálneho roka pomocou Power BI

V ďalšej časti sme sa venovali skúmaniu a analýze dát v programe Power BI. Pre Power BI sme sa rozhodli najmä kvôli tomu, že v najväčšej miere spĺňal nároky, ktoré sme kládli na všetky programy pri výbere adekvátnej platformy pre vykonanie analýzy dát. Pre potreby vizualizácie dát pri tejto problematike boli výstupy vytvorené týmto softvérom vo výbornej kvalite a obsahovali všetky potrebné náležitosti.

Na úvod je potrebné predstaviť tento program v praktickej rovine a vykresliť v akom prostredí sa užívateľ pri tvorbe vizualizácií a analýze dát pohybuje. Prostredie Power BI podobne ako v prípade dataminingového softvéru Knime, obsahuje hlavnú pracovnú plochu ktorá slúži ako priestor pre prvotné načítanie dát a následné prednastavenie konkrétnej spracovávanej oblasti ktorú budeme analyzovať v prípade, ak súbor obsahuje riadky respektíve stĺpce nepotrebných dát. Keď sú dáta načítané, je možné pristúpiť priamo k tvorbe vizualizácii vo forme rôznych typov grafov. Z ponuky si užívateľ jednoducho vyberie potrebný graf a na lište, ktorá obsahuje všetky názvy stĺpcov z načítaného súboru dát si zvolí premenné, ktorých vzťahy a zmeny hodnôt chceme pomocou zvolených vizualizácií analyzovať.

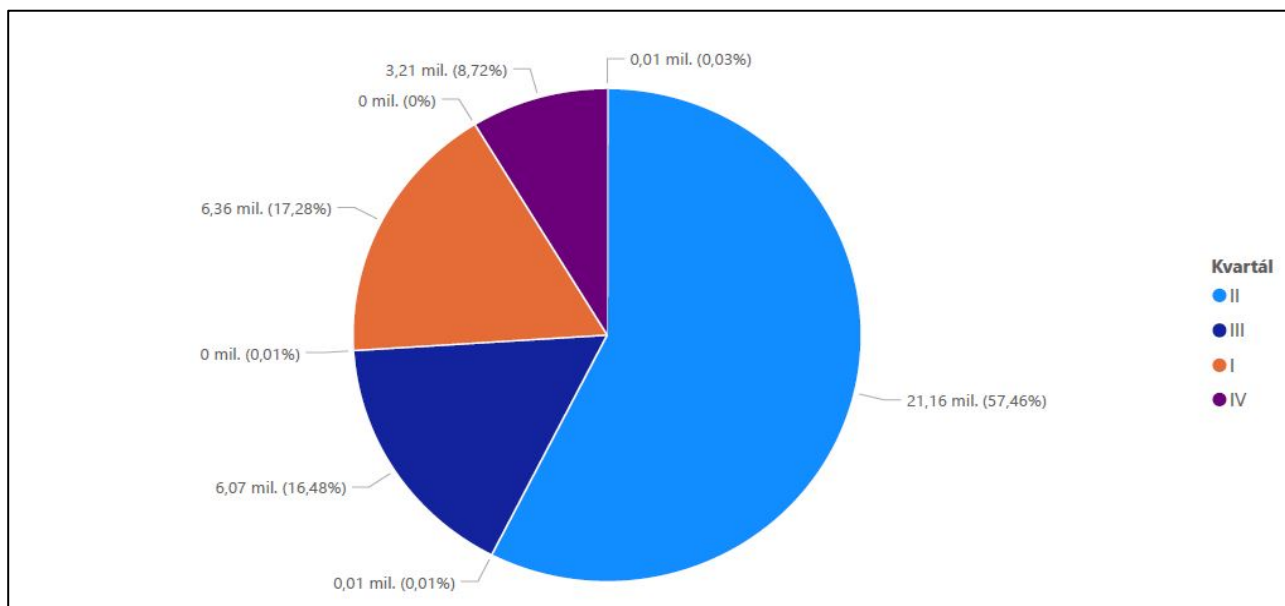
V prípade tejto analýzy slúžia tieto grafy ako podporné vizualizácie pre regresný model ktorý bol v predošlej časti vytvorený v prostredí softvéru Knime. Vieme však poznatky z týchto vizualizácií využiť ako deskriptívne prvky pre bližšie poznanie výkonnosti portfólia v rámci kvartálov. Skúmané veličiny teda boli podobné ako v prípade lineárneho regresného modelu. Do grafu bolo potrebné načítať údaje demografickej skupiny A 12-54, ktoré boli načítanými údajmi z priradených kvartálov klasifikované práve do týchto štyroch kapitol. Posledným vstupom do vizualizácie skupinovým stĺpcovým grafom boli názvy konkrétnych reklamných spotov pre lepšiu identifikáciu tých výnosovo najprínosnejších.



Graf 2 Vizualizácia dát v skupinovom stĺpcovom grafe v programe Power BI

Zdroj: Vlastné spracovanie dát spoločnosti Mindshare v programe Power BI

Výsledná vizualizácia poukázala na fakt, že druhý kvartál bol všeobecne najvyťaženejší a v perspektíve profitov generoval najväčšie zárobky. Vďaka tomuto grafu bolo taktiež evidentné, spot ktorého klienta bol najsledovanejší. Za účelom zmeny zamerania portfólia by bolo možné prejsť aj k detailnejšej analýze, do ktorej by spadali čas odvysielania spotu, spot ktorý jej predchádzal alebo formát bloku, v ktorom sa reklama počas konkrétného dňa nachádzala. Na základe výsledných informácií by sme vedeli predpovedať, v ktorom čase sú reklamy výnosné bez ohľadu na obsah. My sa však zameriavame na výnosy generované týmito spotmi a preto podstatná informácia, ktorú nám tieto vizualizácie skrz skupinový stĺpcový a koláčový graf ukazujú, je fakt, že táto demografická skupina v najväčšom počte sleduje reklamné spoty najmä počas druhého kvartálu. Konkrétne počas mesiacov apríl, máj a jún. Najväčší dosah majú spoty spoločnosti Carte D'or, čo môže byť odôvodnené viacerými argumentami. Najpodstatnejší z nich by istotne bol, že sa v týchto mesiacoch vylepšuje počasie a konzumenti budú čoraz viac do svojich nákupov zahŕňať aj zmrzlinové produkty. Veľkú rolu v tomto prípade zohráva aj už vytvorené povedomie týkajúce sa značky Carte D'or v rámci širokej verejnosti.



Graf 3 Vizualizácia dát v koláčovom grafe v programe Power BI

Zdroj: Zdroj: Vlastné spracovanie dát spoločnosti Mindshare v programe Power BI

Koláčový graf je špecifickým ukazovateľom distribúcie výnosov naprieč celým fiškálnym rokom. Druhý kvartál bol podľa očakávaní najsilnejším a tvoril takmer 58% všetkých profitov spoločnosti Mindshare v oblasti televíznych reklamných spotov. Na druhom mieste je prvý kvartál s takmer 18 percentami. Pomerne malé zastúpenie, štvrtého kvartálu, v ktorom evidujeme iba necelých 9% celkových výnosov. Konštatovali sme už, že portfólio spoločnosti Mindshare v rámci demografickej skupiny A 12-54 je pomerne vyvážené v oblasti výnosov, avšak ak sa sústredíme na distribúciu naprieč fiškálnym rokom, nie je možné opätovne konštatovať tento záver. Je podstatné diskutovať o týchto výsledkoch aby bolo možné vytvoriť kompletný obraz o skutočnostiach ktoré nám hĺbková analýza dát spoločnosti Mindshare poskytla. Výsledky, ktoré boli vizualizované v lineárnom regresnom modeli a taktiež v grafoch majú veľkú hodnotu pre operatívne zložky tejto firmy nakoľko táto analýza bližšie identifikovala silné a profitujúce a zároveň slabšie položky v rámci portfólia spoločnosti Mindshare. Táto diskusia preto prinesie tieto poznatky do kontextu a bude sa taktiež venovať ich možnej implementácii pri prijímaní budúcich rozhodnutí tejto firmy.

Hlavný cieľ bakalárskej práce, ktorým bolo na základe analýzy navrhnúť zmeny v klientskom portfóliu a zložení umiestňovaných reklám, sme plnili pomocou vytvorenia regresného modelu a dátovej vizualizácie. Komentármi k jednotlivým krokom aplikačnej časti a rovnako aj touto diskusiou sme navrhli možné zmeny, ktoré spoločnosť Mindshare

môže implementovať za účelom dosiahnutia stabilnejšieho a rentabilnejšieho portfólia. Konkrétne sme analyzovali dáta reklamných spotov, ktoré bolo vďaka tejto mediálnej agentúre možné sledovať na televíznych obrazovkách. Ide o jedny z najpodstatnejších rozhodnutí v rámci podniku, nakoľko sa jedná o samotnú podstatu fungovania spoločnosti, ktorá je založená na ponúkaní služieb tohto charakteru. Je nutné regulovať, akých klientov bude Mindshare naďalej považovať za kľúčových, avšak ide aj o špecifickejšie rozhodnutia na úrovni jednotlivých reklamných spotov ktoré sa firma rozhodne umiestniť. Výsledky tejto analýzy sú časťou celkového riešenia problému nakoľko ďalším krokom, ktorý bude nutné podstúpiť je pri obnove zmluvného vzťahu tieto projektované zmeny implementovať. Jedným z najpodstatnejších elementov pri skúmaní databázy reklamných spotov bolo určenie optimálnej metódy, ktorá by čo najlepšie zhodnotila, ako sa vyvíjajú ceny reklám v rámci vekovej skupiny A 12-54. Ide teda o dvojcestnú analýzu, ktorá manažmentu Mindshare umožní urobiť rozhodnutie založené jednak na výstupe, ktorý poukáže na najviac profitujúce reklamy a zároveň klientov. Taktiež ukáže, v akej miere sú konkrétne spoty populárne v tejto demografickej skupine, ktorá je síce veľmi všeobecná, avšak predchádzame jej výberom skresleným údajom nakoľko špecifické demografické skupiny obsahujúce výlučne jedno z pohlaví by sme vedeli zahrnúť pri analýze konkrétnych kategórií reklám, pre ktoré bola daná skupina identifikovaná ako cieľová. Vizualizáciami v programe Power BI sme si poznatky doplnili o informácie o najaktívnejších kvartáloch a o relatívnych podieloch týchto kvartálov na celkových príjmoch spoločnosti. Táto informácia je taktiež kategorizovaná ako kľúčová, nakoľko bude mať manažment spoločnosti Mindshare možnosť zhodnotiť, ktoré zmeny je a zároveň nie je možné implementovať pri momentálnom stave akvizície kapitálu. Pre príklad teda nie je možné meniť portfólio klientely v štvrtom kvartáli nakoľko je to jeden z najmenej aktívnych častí fiškálneho roka a drasticky by to ovplyvnilo akumulovanú výnosnosť firmy.

Pri detailnejšom preskúmaní výstupov z programu Knime je zreteľné, že spoty, ktoré majú nižšiu sledovanosť sú kritické body, ktoré by mali byť v rámci predajného oddelenia podrobnejšie preskúmané a ich umiestnenie by malo byť prehodnotené. Je totiž pravdepodobné, že pri bližšej analýze zistíme, že reálna hodnota výkonnosti týchto spotov nenaplnila očakávania a je teda ďaleko za projektovanou výkonnosťou. V takom prípade je možné, že i keď produkujú určité percento celkových ziskov, mohli by tieto spoty byť nahradené výnosnejšou spoluprácou s iným klientom, a iným reklamným spotom, ktorého umiestnenie by mohlo priniesť priaznivejšie výsledky.

Taktiež je potrebné konfrontovať skutočnosť, že v štvrtom kvartáli je pomerne veľa priestoru, ktorý by mohol byť pri získavaní novej klientely avizovaný ako jeden z možných bodov na ktorých by negociácie boli založené. Je totiž pomocou rôznych marketingových nástrojov možné tento priestor predať za výhodnejších, avšak stále rentabilných podmienok čo by spôsobilo prínos ďalšieho kapitálu a možnosť dlhodobej spolupráce. Z opačnej perspektívy sa taktiež môžeme detailnejšie pozrieť na najlepšie vynášajúce spoty, ktoré sú umiestňované pre klienta Carte D'or, ktorý je dlhoročným klientom spoločnosti Mindshare. Pri pohľade na jeho osamotenú pozíciu je toto prvenstvo možným príslubom prehodnotenia faktu, že sú spoty tohto klienta umiestňované výlučne počas jedného kvartálu. Je tu totiž potenciál využiť popularitu a osvedčenú kvalitu tohto klienta na vysielanie ďalších reklamných spotov počas iných mesiacov fiškálneho roka, avšak s prihliadnutím na relevantnosť obsahu týchto reklám nakoľko sa jedná o zmrzlinové výrobky. Ide teda o distingvované produkty ponúkané prevažne počas letných mesiacov.

Je však na ďalšom posúdení manažmentu spoločnosti Mindshare, či budú obdobné kroky podstúpené alebo budú aplikované ešte podrobnejšie analytické kritériá, ktoré prinesú nielen vyššie spomenuté výstupy, ale budú taktiež možné analyzovať časové úseky, ktoré v rámci reklamných blokov boli tie najprofitujúcejšie. Na základe týchto pridaných informácií budú ďalej postupovať v rozhodovacom procese. Vizualizácie a regresný model, ktoré sme vytvorili pomocou spomenutých programov jasne ukázali výkonné a pomerne všestranné portfólio, ktoré je však možné na základe nami vytýčených bodov vylepšiť a priniesť tým proporčne stúpajúcu efektivitu podnikateľskej činnosti spoločnosti Mindshare.

3.4 Porovnanie a hodnotenie programu Knime a Power BI

Ak sa firma rozhodne na podporu svojich rozhodovacích procesov využiť parametre a výstupy nadobudnuté dolovaním dát, treba taktiež zobrať do úvahy rôzne faktory, ktoré pri voľbe optimálneho analytického softvéru zohrávajú svoju roľu. Medzi najzákladnejšie aspekty by sme zaradili náročnosť práce s daným softvérom, jazykovú dostupnosť ktorá zaistí, že pri práci s analytickým prostriedkom sa nevyskytnú komplikácie s jazykovou bariérou, kompatibilita s operačným systémom a prvkami vhodnými pre zariadenia, ktorými podnikateľ, respektíve firma disponuje. Ako ďalší aspekt vnímame cenovú dostupnosť nakoľko nie všetky softvéry sú vo svojej primárnej/alpha verzii dostupné v bezplatnej forme,

respektíve ako open source programy. Dôležitým kritériom je aj prístupnosť podpory pre užívateľa, ak by prišlo k technickým chybám vážnejšieho charakteru. Teda takým, ktoré nie je možné riešiť operatívne skrz príkazy zadávané do zdrojového kódu.

Power Business Intelligence považujem z hľadiska jednoduchosti používania, rýchlosti a iným aspektom spojených s funkcionalitou za jednu z najlepších platforiem pre analýzu dát. Rovnako tento program nezaostáva ani po estetickej stránke. Spomenuli sme už, že program je čo sa jednoduchosti ovládania týka špičkový a odporučil by som tento program každému podniku, ktorý chce k relevantným dátam vždy dostupné interpretácie. Vďaka rýchlosti a dynamickosti tohto softvéru je to správna voľba pre každú firmu, ktorej rozhodnutia sú prijímané a založené na dátach. Ako jednu z menej pozitívnych skúseností, ktorá pre mnoho spoločností môže tvoriť pomerne závažný problém a prekážku pri využívaní tohto softvéru je skutočnosť, že nakoľko bol Power BI produktom vývojárov zo spoločnosti Microsoft, nie je tento softvér dostupný pre žiadne zo zariadení ktoré funguje na operačnom systéme MacOS od spoločnosti Apple, nakoľko sa jedná o konkurenčnú firmu. I keď tvrdenie, že ide o stratégiu ako zabrániť konkurenčnej spoločnosti využívať a eventuálne prísť na trh s obdobnou aplikáciou môže byť iba jedným z možných dôvodov nedostupnosti tohto programu na operačnom systéme tejto spoločnosti, je to obštrukcia ktorú nemožno prehliadnuť. Riešenie pre takúto situáciu spadá do kategórie nekomplexných, avšak nákladných nakoľko je možné tento problém, v prípade že spoločnosť disponuje iba zariadeniami spoločnosti Apple, riešiť dvoma spôsobmi. Prvý, rentabilnejší a dostupnejší spôsob riešenia je zakúpiť licenciu na operačný systém spoločnosti Microsoft a na zariadeniach, ktorým to softvérové vybavenie dovoľuje, môžu fungovať s duálnym operačným systémom. Druhé, nákladnejšie riešenie by v tomto prípade bolo investovať do nového, vyhovujúceho zariadenia inej značky ktoré disponuje operačným systémom Windows . Toto riešenie by malo byť uprednostnené iba s situácií, kedy sú dáta a ich analýza nevyhnutnou a permanentnou súčasťou interných procesov podniku, a to najmä z dôvodu, že k programu Power BI je možné dodatočne stiahnuť aj doplnkové funkcionality, ktoré môžu byť využité k ešte detailnejšej a na mieru prednastavenej analýze.

Čo sa softvéru Knime týka, práve pre mnohostrannosť a pomerne jednoduché ovládanie ktoré bolo možné pochopiť na základe štúdie relevantných materiálov bola voľba tohto programu veľmi jednoduchá. Ako ďalší pozitívny aspekt ktorý nasvedčuje o kvalite tohto softvéru je možnosť veľmi jednoducho načítať ďalšie funkcie do pracovného priestoru

a týmto spôsobom pozmeniť formu alebo oblasť vstupných dát. Pri nastavovaní a vyhotovovaní schémy je potrebné byť čo najpozornejší a to najmä z dôvodu, že s veľkým počtom možných modifikácií vzniká taktiež riziko neželanej voľby niektorého z ponúkaných nastavení čo sa v konečnom dôsledku môže prejaviť ako fatálne pri presnosti výsledného modelu. Odporučili by sme preto pozorne prejsť každým dialógovým oknom a rovnako každou zobrazenou ponukou nastavení, nakoľko je možné, že program nenačítal všetky prednastavenia v absolútnej korektnosti. Pre čo najviac efektnú prácu s týmto softvérom by sa takáto kontrola nemala vykonávať v konečnej fáze ale operatívne pri postupnom výbere a zhotovovaní schémy, keďže túto činnosť vykonávame s najaktuálnejšími inštrukciami ktoré očakávame, že budú odzrkadlené vo výstupe vyhotovenom analýzou dát. Záverečná kontrola nie je v žiadnom prípade zavrňovaná avšak sme toho názoru, že pri takejto kontrole môže byť používateľ pod návalom všetkých informácií uvedený do mylného pocitu správnosti alebo naopak nesprávnosti niektorého z nastavení.

Záver

Hlavnou témou práce bolo aplikovať metódy dataminingu v podnikovej praxi existujúcej spoločnosti, pričom hlavným cieľom bolo pomocou analýzy skutočných dát navrhnúť zlepšenie v oblasti tvorby portfólia klientov a umiestnených reklamných spotov spoločnosti Mindshare. Táto úloha bola plnená rôznymi metódami, ktoré okrem iného v aplikačnej časti zahŕňali vyhotovenie regresného modelu a viacerých dátových vizualizácií, ktoré mali za úlohu vyvodit' riešenia na špecifikované problémy. Analýzou, komparáciou a sumarizáciou odbornej literatúry, ako aj uvedením niekoľkých pohľadov a definícií rôznych zahraničných i domácich autorov, sme v prvej časti tejto práce plnili čiastkové ciele definovania súčasných vedomostí a uvedenia viacerých podstatných pojmov z oblasti dataminingu. Definovaním pojmov a opísaním súčasných vedomostí z oblasti analýzy dát sme prispeli k pochopeniu tejto analýzy ako viacstupňovej a dynamickej činnosti ktorá si mimo samotnej časti dolovania vyžaduje pozornú a detailnú prípravu dát. Jasne sme stanovili fundamentálne princípy hĺbkovej analýzy dát a definovali, že sa tento ucelený proces skladá z niekoľkých krokov ktoré svojou dôležitosťou rovnakým dielom prispievajú k efektívnemu získavaniu poznatkov z databáz rôznych veľkostí. Týmto bol splnený čiastkový cieľ stanovenia procesných prvkov prípravy dát a samotnej hĺbkovej analýzy. Pri mnohých funkcionalitách a využitíach analýzy dát je takmer prirodzené, že so sebou takýto komplexný súbor frakčných úkonov prináša aj problémy ktoré môžu používateľa dataminingových softvérov postretnúť v rôznych podobách. Pri jasnom definovaní postupu, ktorý bude dodržaný počas celého procesu hĺbkovej analýzy dát, je však možné sa týmto problémom z veľkej časti vyhnúť. Charakterizáciou týchto problémov bol plnený ďalší z čiastkových cieľov tejto práce. Ako ďalší bol plnený čiastkový cieľ analýzy aktuálnej situácie v podniku. Detailnou analýzou dát poskytnutých spoločnosťou Mindshare, ako aj analýzou dostupných informácií a následnou dedukciou, sme vytvorili podrobný obraz o aktuálnom stave a nedostatkoch v portfóliovej stratégii tejto spoločnosti. Získaním potrebných softvérov, nadobudnutím znalostí pre ich ovládanie a následne samotnou analýzou dát vybranej spoločnosti pomocou týchto softvérov boli splnené čiastkové ciele týkajúce sa vyhotovenia modelu lineárnej regresie a dátových vizualizácií. Po vytvorení analytických modelov sme pristúpili k diskusii a formulácii možných zlepšení. Za účelom zrozumiteľnejšieho výstupu bolo nutné vytvoriť dve kategórie, podľa ktorých budú klienti hodnotený. Vytýčili sme si túto skutočnosť ako čiastkový cieľ a plnili sme ho kategorizáciou klientov do skupiny kľúčových a nevýnosných zmluvných vzťahov. V

neposlednom rade sme podrobne vyhodnotili úroveň ziskov počas jednotlivých kvartálov roka 2019 a zároveň sme vyhodnotili vyhotovené analýzy za účelom formulácie odporúčaní a možných zlepšení, ktoré je možné implementovať s cieľom vylepšiť rentabilitu portfólia a výnosnosť samotnej spoločnosti Mindshare. Prvá časť tejto bakalárskej práce nám, ako sme už spomenuli vytvorila hrubé obrisy ktoré sme vyplnili vedomosťami obsiahnutými v praktickej časti, ktorá zahŕňala získanie reálnych dát spoločnosti Mindshare. Aplikovanie dataminingu na tieto dáta poukázalo na skutočne všestranné využitie výstupov a viac než hodnotné závery, ktoré bolo možné na základe vytvorených modelov a vizualizácii vyvodiť. Cieľom praktickej časti tejto práce bolo vytvoriť funkčný model, ktorý by informáciami v ňom obsiahnutými prispel k chodu tejto spoločnosti a ktorý by bolo možné aplikovať do momentálnej schémy procesov v rámci relevantných oddelení tejto spoločnosti. Poznatkami, ktoré sme vydolovali pomocou použitých metód sme teda úspešne aplikovali datamining do dát tak kľúčového sektora, ako je samotný predaj a komunikácia s klientom. V spoločnostiach ktoré sú client-driven a teda sú ich kroky a rozhodovacie procesy založené výhradne na správaní spotrebiteľov, je nutné prediktívnym prístupom v čo najväčšej miere prispievať k rozhodnutiam, ktoré budú spoločnosti viesť k čoraz úspešnejšiemu účinkovaniu v podnikateľskej sfére. Datamining je preto jednou z neoddeliteľných súčastí chodu spoločnosti a je potrebné naše vedomosti v tejto oblasti zdokonaľovať, nakoľko možnosti využitia v oblasti podnikania, ale rovnako aj spoločnosti ako takej, ktoré nám tento analytický prostriedok ponúka, majú enormný potenciál. Je iba na nás, ako dokážeme tento dynamický a učentlivý mechanizmus zužitkovať.

Zoznam bibliografických zdrojov

Knižné zdroje

HAN, Jiawei – PEI, Jian – KAMBER, Micheline. Data mining: concepts and techniques. 3. vydanie. San Francisco: Elsevier, 2011. 560 s. ISBN 9780123814791

WITTEN, Ian H. – FRANK, Eibe. Data mining: practical machine learning tools and techniques with Java implementations. 2. vydanie. San Francisco: Elsevier, 2005. 525 s. ISBN 1558605525

BERRY, Michael J.A. – LINOFF, Gordon S. Data mining techniques: for marketing, sales, and customer relationship management. 2. vydanie. Indianapolis: John Wiley & Sons, 2004. 643 s. ISBN 9780471470649

PROVOST, Foster – FAWCETT, Tom. Data Science for Business: What you need to know about data mining and data-analytic thinking. Sebastopol: O'Reilly Media, Inc., 2013. 414 s. ISBN 9781449374280

AGGARWAL, Charu C. Data mining: the textbook. Springer, 2015. 734 s. ISBN 9783319141428

Články v elektronických časopisoch

HALILI, Festim – RUSTEMI, Avni. Predictive modeling: data mining regression technique applied in a prototype, In International Journal of Computer Science and Mobile Computing [online]. Telangana: IJCSMC, 2016, roč. 5, č. 8, s. 207-215 [cit. 2020-11-20] ISSN 2320-088X. Dostupné na: <https://www.ijcsmc.com/docs/papers/August2016/V5I8201642.pdf>

GUPTA, Swati. A Regression Modeling Technique on Data Mining. In International Journal of Computer Applications [online]. Foundation of Computer Science, 2015, roč. 9. č. 116, s. 27-29 [cit. 2020-11-20]. ISBN: 973-93-80886-20-6. Dostupné na: <https://research.ijcaonline.org/volume116/number9/pxc3902570.pdf>

Internetové zdroje

PATTERSON, Mathew. Knowledge base 101: Definition, Benefits, Examples and Tips [online]. [cit. 2020-12-03]. Dostupné na: <https://www.helpscout.com/playlists/knowledge-base/>

KATIYAR, Pranchal. Difference Between Descriptive and Predictive Data Mining [online]. Uveřejněné: 17.09.2019 [cit. 2021-01-05] Dostupné na: <https://www.geeksforgeeks.org/difference-between-descriptive-and-predictive-data-mining/>

12 common problems in Data Mining [online]. Uveřejněné: 03.02.2015 [cit. 2021-01-05] Dostupné na: <https://bigdata-madesimple.com/12-common-problems-in-data-mining/>

Difference Between Descriptive and Predictive Data Mining [online]. [cit. 2021-01-07] Dostupné na: <https://techdifferences.com/difference-between-descriptive-and-predictive-datamining.html#:~:text=Key%20Differences%20Between%20Descriptive%20and,that%20predictions%20can%20be%20made>

Basic Concept of Classification (Data Mining) [online]. Uveřejněné: 12.12.2019 [cit. 2021-01-15] Dostupné na: <https://www.geeksforgeeks.org/difference-between-descriptive-and-predictive-data-mining/>

SHARMA, Rohit. Cluster Analysis in Data Mining: Applications, Methods & Requirements [online]. Uveřejněné: 20.01.2020 [cit. 2021-02-01] Dostupné na: <https://www.upgrad.com/blog/cluster-analysis-data-mining/#:~:text=Conclusion-,What%20is%20Clustering%20in%20Data%20Mining%3F,the%20similarity%20of%20the%20data.>

ORACLE Help Center: Regression [online]. In Data Mining Concepts. [cit. 2021-02-22] Dostupné na: https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/regress.htm#DMCON005