VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA PODNIKATELSKÁ

FACULTY OF BUSINESS AND MANAGEMENT

ÚSTAV INFORMATIKY

INSTITUTE OF INFORMATICS

VYUŽITÍ DATA MININGU VE FIREMNÍCH PROCESECH

USE OF DATA MINING IN BUSINESS PROCESSES

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

AUTOR PRÁCE — **Vendula Procházková**

AUTHOR

VEDOUCÍ PRÁCE — **Ing. Jiří Kříž, Ph.D.**

SUPERVISOR

BRNO 2022

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA PODNIKATELSKÁ

Zadání bakalářské práce

Ústav:	Ústav informatiky
Studentka:	**Vendula Procházková**
Vedoucí práce:	**Ing. Jiří Kříž, Ph.D.**
Akademický rok:	2021/22
Studijní program:	Manažerská informatika

Garant studijního programu Vám v souladu se zákonem č. 111/1998 Sb., o vysokých školách ve znění pozdějších předpisů a se Studijním a zkušebním řádem VUT v Brně zadává bakalářskou práci s názvem:

Využití data miningu ve firemních procesech

Charakteristika problematiky úkolu:

Úvod
Cíle práce, metody a postupy zpracování
Teoretická východiska práce
Analýza současného stavu
Vlastní návrhy řešení
Závěr
Seznam použité literatury
Přílohy

Cíle, kterých má být dosaženo:

Cílem práce je navrhnout efektivní data mining – modely pro podporu firemních procesů.

Základní literární prameny:

ALASADI, Suad A. a Wesam S. BHAYA. Review of data preprocessing techniques in data mining. Journal of Engineering and Applied Sciences, 2017. str. 4102-4107.

GUPTA, Manoj Kumar a Chandra PRAVIN. A comprehensive survey of data mining. International Journal of Information Technology. 6. Únor 2020, str. 1243-1257.

CHEN, Hsinchun, Roger H. L. CHIANG and Veda C. STOREY. Business intelligence and analytics: from big data to big impact. MIS Quarterly. 2012, pp. 1165-1188.

NGUYEN, Giang, et al. Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. Artificial Intelligence Review. Leden 19, 2019. str. 77-124.

NOVOTNÝ, Ota, Jan POUR a David SLÁNSKÝ. Business intelligence: jak využít bohatství ve vašich datech. Praha: Grada, 2005. Management v informační společnosti. ISBN 80-247-1094-3.

Termín odevzdání bakalářské práce je stanoven časovým plánem akademického roku 2021/22

V Brně dne 28.2.2022

L. S.

Ing. Jiří Kříž, Ph.D. garant	doc. Ing. Vojtěch Bartoš, Ph.D. děkan

Abstrakt

Tato bakalářská práce je zaměřena na analýzu prodejů produktů dané značky s orientací na výběr TOP produktů. Cílem práce je tento proces automatizovat a umožnit přenositelnost procesu výběru produktů na další značky z portfolia společnosti. Pro realizaci stanovených cílů v práci používám mnou vytvořenou metriku výběru TOP produktů. Výsledky získané aplikací mého řešení porovnávám s konkurencí pro určení kvality mnou navrženého procesu výběru TOP produktů. Dodatečně své řešení porovnávám s vybranými technikami data miningu a ukazuji jeho vhodnost pro tento typ analýzy.

Abstract

This bachelor thesis focuses on the sales analysis of the products of a specific brand, focusing on the selection of TOP products. The thesis aims to automate this process and make the product selection process transferable to other brands in the company's portfolio. In order to realize the set objectives, I use a TOP product selection metric created by me. The results obtained by applying my solution are compared with competitors to determine the quality of my proposed TOP product selection process. Additionally, I compare my solution with selected data mining techniques and show its suitability for this type of analysis.

Klíčová slova

Business Intelligence, data mining, analýza prodejů, RapidMiner, obchodní rozhodování

Keywords

Business Intelligence, data mining, sales analysis, RapidMiner, business decision making

Bibliografická citace

PROCHÁZKOVÁ, Vendula. Využití data miningu ve firemních procesech [online]. Brno, 2022 [cit. 2022-05-08]. Dostupné z: https://www.vutbr.cz/studenti/zav-prace/detail/143736. Bakalářská práce. Vysoké učení technické v Brně, Fakulta podnikatelská, Ústav informatiky. Vedoucí práce Jiří Kříž.

Čestné prohlášení

Prohlašuji, že předložená bakalářská práce je původní a zpracovala jsem ji samostatně. Prohlašuji, že citace použitých pramenů je úplná, že jsem ve své práci neporušila autorská práva (ve smyslu Zákona č. 121/2000 Sb., o právu autorském a o právech souvisejících s právem autorským).

V Brně dne 9. května 2022

……

podpis studenta

Poděkování

Ráda bych poděkovala Ing. Jiřímu Křížovi, Ph.D. za odborné vedení, jeho čas a rady při zpracování této práce. Dále chci poděkovat zaměstnancům společnosti, ve které jsem práci zpracovala, za poskytnutá data a cenné rady při jejich analýze. Poděkování patří i mé rodině a příteli, kteří mě vždy během studia podpořili.

Obsah

Úvod

Správné nakládání s informacemi se v dnešním konkurenčním prostředí stalo významnou výhodou v obchodních i dalších procesech. Business Intelligence využívá software a služby k přeměně dat na užitečné informace, které slouží jako podklad pro obchodní rozhodnutí organizace. Business Intelligence se stalo důležitým aspektem, který by si měli uvědomovat jak podnikoví manažeři, tak manažeři IT, a využívat jej pro proměnu informací na konkurenční výhody.

Data mining (dolování z dat) je jeden z nástrojů Business Intelligence. Je to analytická metoda, která umožňuje získat z velkého množství dat užitečné informace, které jsou významné při obchodním rozhodování společnosti. Vzhledem k stále většímu objemu a složitosti ukládaných dat již není v lidských silách tato data ručně zpracovávat, a proto je role data miningu v podniku stále důležitější.

Důležitost Business Intelligence si uvědomuje i společnost, u které práci vypracovávám, a proto se rozhodli učinit první kroky k zavedení data miningových technik do dalších firemních procesů. Konkrétně se v práci věnuji výběru produktů pro promování na hlavních stránkách e-shopů vybrané značky. Momentálně se tímto výběrem zabývá několik analytiků a expertů na dané značky. Díky zavedení data miningových technik do tohoto procesu bude výběr zautomatizován, což ušetří lidské zdroje. Zároveň bude možné tyto techniky aplikovat i na další značky z portfolia společnosti.

V návaznosti na uvedenou motivaci prezentuji řešení pro výběr TOP produktů konkrétní značky na základě analýzy dat prodejů za posledních 24 měsíců. Pro tento výběr jsem navrhla vlastní metriku zohledňující průběžné prodeje za celé období a prodeje za poslední čtvrtletí. Pro určení vhodnosti mého řešení jsem porovnávala výsledky získané aplikací metriky s výsledky různých data miningových technik. Mnou navržená metrika se ukázala jako nejvhodnější. Toto zjištění bylo potvrzeno i porovnáním TOP produktů s konkurencí.

Za hlavní přínosy práce můžeme považovat:

- Vytvoření metriky pro výběr TOP produktů na základě informací o prodejích za posledních 24 měsíců.

- Porovnání různých technik výběru TOP produktů a prokazování jejich vhodnosti pro tento typ analýzy
- Automatizace procesu výběru TOP produktů a jeho rozšiřitelnost pro další značky

V kapitole 1 jsou definovány problémy a cíle práce. Kapitola 2 popisuje teoretická východiska práce, zejména Business Intelligence a data mining. V kapitole 3 se zabývám analýzou současného stavu společnosti, za pomoci které je práce zpracována. Kapitola 4 obsahuje popis a realizaci vlastního návrhu řešení. Závěr pojednává o dosažených výsledcích práce.

1 Vymezení problému a cíle práce

Tato kapitola definuje řešený problém a cíle, kterých má být v práci dosaženo. Zároveň pojednává o zvolené metodice vypracování.

1.1 Vymezení problému

Společnost XYZ poskytuje B2C (Business to customer) e-shopy (brandpage) značkám, které v České republice nemají zastoupení. Výběr TOP produktů zobrazovaných na těchto e-shopech není zautomatizovaný a provádí ho tým analytiků a expertů na dané značky. To je pro společnost zbytečně časově i finančně náročné.

1.2 Cíle práce

Hlavním cílem této bakalářské práce je zautomatizovat proces výběru TOP produktů.

Zautomatizováním tohoto procesu bude možné analyzovat produkty častěji s nižšími požadavky na lidské a finanční zdroje. Zároveň tím bude usnadněno porovnávání výsledků s konkurencí.

V bakalářské práci budou popsána teoretická východiska v oblasti Business Intelligence a data miningu, analýza současného stavu společnosti a návrh řešení s jeho realizací.

1.3 Metodika práce

Pro zpracování výše popsaného úkolu jsem si zvolila nástroj RapidMiner Studio. Tento nástroj jsem vybrala, protože je volně přístupný studentům fakulty a má přívětivé uživatelské rozhraní.

Výběr produktů zobrazovaných na brandpage bude založen na analýze jejich prodejů za posledních 24 měsíců, ceně a kategorii. Výsledkem bude výběr TOP produktů (nejprodávanější, nejzajímavější pro zákazníky atd.) rozdělených podle cenových hladin a kategorií.

Ověření správnosti zautomatizovaného procesu bude probíhat porovnáním s výsledky konkurenčního portálu Heureka.

2 Teoretická východiska práce

V této kapitole se věnuji teorii spojené s Business Intelligence a data miningem. Nejdříve poskytuji úvod do problematiky Business Intelligence a transakčních a analytických systémů. V další části se věnuji úvodu do data miningu, jeho úlohám, technikám a algoritmům.

2.1 Business Intelligence

Business Intelligence (BI) jsou techniky, technologie, systémy, praktiky, metodiky a aplikace využívané pro analýzu kritických obchodních dat tak, aby pomohly podniku lépe pochopit jeho činnost a chování trhu a na základě těchto znalostí podpořit rozhodovací procesy. Kromě základních technologií pro zpracování dat a analýzu zahrnuje BI také postupy a metodiky zaměřené na podnikání, které lze aplikovat na různé důležité oblasti jako jsou e-commerce, průzkum trhu, e-goverment, zdravotnictví a bezpečnost (1).

Termín Intelligence používají výzkumníci v oblasti umělé inteligence již od 50. let 20. století. Termín Business Intelligence zavedl v roce 1989 Howard J. Dresner, analytik ve společnosti Gartner Group a pojem BI se stal populárním v podnikatelské a IT komunitě v 90. letech 20. století. Koncem roku 2000 byl zaveden pojem Business Analytics, který představuje klíčovou analytickou složku v BI (1; 2).

Mezi nástroje a aplikace Business Intelligence patří (2):

- Produkční, zdrojové systémy
- Dočasná uložiště dat (DSA – Data Staging Area)
- Operativní uložiště dat (ODS – Operational Data Store)
- Transformační nástroje (ETL – Extraction Transformation Loading)
- Integrační zdroje (EAI – Enterprise Application Integration)
- Datové sklady (DWH – Data Warehouses)
- Datová tržiště (DMA – Data Marts)
- OLAP

- Reporting
- Manažerské aplikace (EIS – Executive Information Systems)
- Dolování dat (Data Mining)
- Nástroje pro zajištění kvality dat
- Nástroje pro správu metadat
- Ostatní

2.2 Rozdíl mezi OLTP a analytickými systémy

Tato podkapitola obsahuje základní informace o OLTP a analytických systémech a přehledné porovnání jejich vlastností.

2.2.1 OLTP (On Line Transaction Processing) systémy

OLTP je systém pro online zpracování transakcí. Hlavním cílem systému OLTP je zaznamenávat probíhající aktualizace, vkládání a mazání při transakcích. Dotazy OLTP jsou jednodušší a krátké, a proto vyžadují méně času na zpracování a také méně místa. Systém OLTP se stává zdrojem dat pro OLAP (2; 3).

2.2.2 Analytické systémy – OLAP (On Line Analytical Processing)

OLAP je systém online analytického zpracování dat pro podporu dotazování. Databáze OLAP uchovává historická data, která byla vložena pomocí OLTP. Umožňuje uživateli zobrazit různé souhrny vícerozměrných dat. Pomocí OLAP lze z rozsáhlé databáze získávat informace a analyzovat je pro účely rozhodování. OLAP také umožňuje uživateli provádět složité dotazy k získání vícerozměrných dat. Transakce v OLAP jsou dlouhé, a proto jejich zpracování trvá relativně déle a vyžadují velký prostor (2; 3).

Detailněji jsou rozdíly OLTP a OLAP popsány v Tabulce 1.

Tabulka 1 : Rozdíly mezi technologiemi OLTP a OLAP. (Zdroj: Vlastní zpracování dle *(2; 3; 4; 5)*)

Vlastnost	*OLTP*	*OLAP*
Zaměření	Vkládání, aktualizace a mazání informací z databáze	Výběr dat pro analýzu, která pomáhá při rozhodování
Data	Současná a detailní	Historická a sloučená
Zdroje dat	Operační, Interní	Operační, Interní a Externí
Typy dotazů	Jednoduché standardizované dotazy	Složité dotazy
Základní operace	Založeno na INSERT, UPDATE, DELETE příkazech	Založeno na příkazu SELECT
Integrita dat	Vysoký důraz na integritu dat	Data se normálně nemění – není potřeba řešit integritu dat
Normalizace dat	Ano	Ne
Doba odezvy	Milisekundy	Sekundy, minuty nebo hodiny v závislosti na množství zpracovaných dat
Činnosti	Procesy	Analýza
Požadavky na uložiště (objem dat)	Malé	Vysoké
Pohled na data	Seznam transakcí	Multidimenzionální pohled na podniková data
Příklady uživatelů	Zaměstnanci ve styku se zákazníky, prodavači, online nakupující	Datoví analytici, obchodní analytici nebo vedoucí pracovníci.

2.3 Data Mining

Data mining (dolování dat) je základní fází procesu objevování znalostí, jejímž cílem je získat z dat zajímavé a potenciálně užitečné informace. Data mining se skládá z různých funkcí, technik a algoritmů, které se používají k objevování a získávání zajímavých vzorů z velkého úložiště dat. Ačkoli se často pojem "dolování dat" orientuje především na dolování dat velkého rozsahu, mnoho technik, které dobře fungují u rozsáhlých souborů dat, lze efektivně aplikovat i na menší soubory. Vzhledem k významu při rozhodování se v posledních dvou desetiletích data mining dostal do širokého povědomí a stal se základním nástrojem při provádění nejrůznějších operací organizací (6; 7).

Do data miningu je začleněno mnoho dalších technik, jako je statistika, databázové systémy/datové sklady, strojové učení, algoritmy, rozpoznávání vzorů, vizualizace, vyhledávání informací, vysoce výkonná výpočetní technika atd. První tři uvedené techniky jsou hlavními přispěvateli data miningu (6; 7).

2.3.1 Příprava dat

Dolování dat v zásadě závisí na kvalitě dat. Surová data jsou obvykle náchylná k chybějícím hodnotám, zašuměným datům, neúplným datům, nekonzistentním datům a datům s odlehlými hodnotami. Je tedy důležité, aby tato data byla před dolováním zpracována. Předběžné zpracování dat je nezbytným krokem ke zvýšení kvality dat. Zpracování dat je jedním z nejdůležitějších kroků data miningu, který se zabývá přípravou a transformací datového souboru a zároveň se snaží zefektivnit objevování znalostí. Předzpracování zahrnuje několik technik, jako je čištění, integrace, transformace a redukce (8).

2.3.2 Úlohy data miningu

Funkce nebo úlohy data miningu lze použít k určení typů vzorů nebo znalostí, které mají být během procesu dolování objeveny. Mezi hlavní funkce data miningu patří sumarizace, charakterizace a diskriminace, asociace, shlukování, klasifikace, analýza odlehlých hodnot, regrese, analýza trendů. Vybrané úlohy jsou pospány dále v této podkapitole (6).

Sumarizace

Sumarizace vede k rozdělení výsledků do menšího souboru a představuje shrnutí podrobných dat na základě hierarchie pojmů. Obvykle se sumarizace provádí pomocí agregace, kterou lze rozšířit na různé úrovně abstrakce a nahlížet na ni z různých úhlů (6).

Charakterizace a diskriminace

Charakterizace je v podstatě shrnutí dat na základě hierarchie pojmů a generuje charakterizační pravidla. Na druhé straně diskriminace slouží k identifikaci odlišností mezi různými soubory dat. Výstup diskriminace je generován ve formě diskriminačních pravidel (6).

Klasifikace

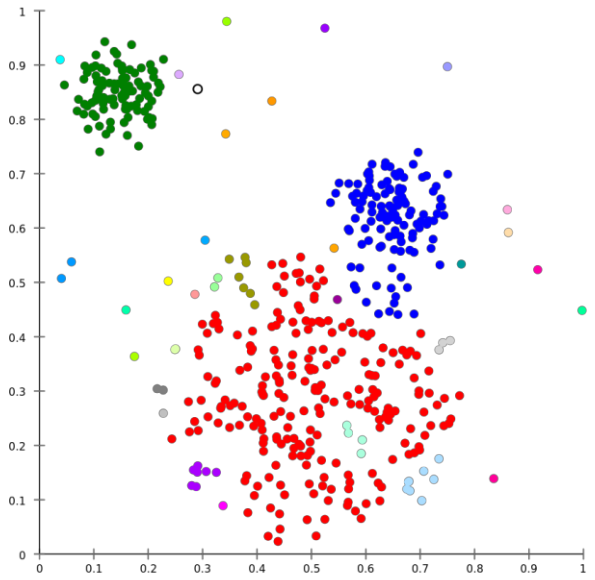
Klasifikace je proces, při kterém zařazujeme objekty do předem definovaných tříd. Třídou se rozumí například: *zdravý/nemocný*, *SPAM/ne SPAM*. Pro trénink klasifikačního modelu je nutné trénovacím datům přiřadit správné třídy. Jedná se o takzvané „učení s učitelem“ (6; 9).

Shlukování (Clustering)

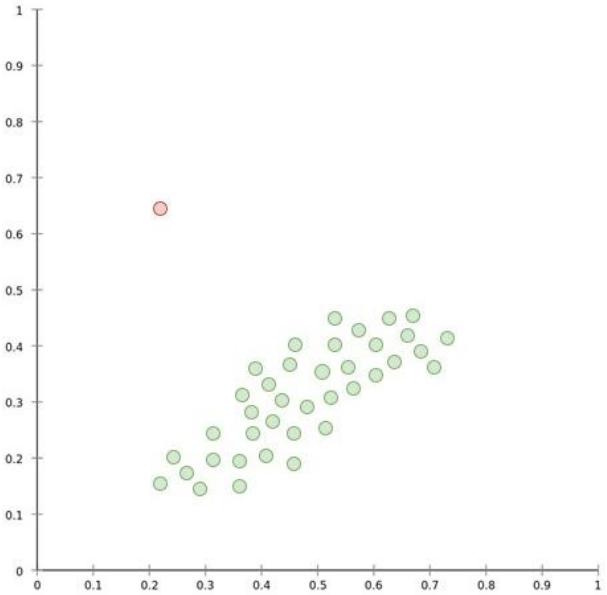
Shlukování se používá k rozdělení nebo segmentaci datových objektů do podmnožin nazývaných skupiny nebo shluky, jak ukazuje Obrázek 1. Objekty, které jsou si navzájem blízké, jsou zařazeny do stejné skupiny. Stejně jako klasifikace i shlukování klasifikuje podobné datové objekty, ale na rozdíl od klasifikace jsou značky tříd neznámé. Shluková analýza je jednou z nejoblíbenějších technik, která se používá nejen při dolování dat, ale také v dalších oblastech, jako je statistika, segmentace obrazu, bioinformatika, a další (6).

Analýza odlehlých hodnot

Datové objekty, které se liší v běžném chování dat, se nazývají odlehlé hodnoty (viz Obrázek 2). Většina metod dolování dat tyto odlehlé hodnoty zpravidla vyřazuje jako šum nebo výjimky. Někdy mohou mít odlehlé hodnoty ve srovnání s jinými datovými objekty více informací. Proto je analýza odlehlých hodnot důležitá pro některé oblasti použití, jako je detekce narušení, detekce podvodů, detekce anomálií a další (6).



Obrázek 1: Příklad grafu shlukování. (Zdroj: *(23)*)



Obrázek 2: Příklad grafu analýzy odlehlých hodnot. Červený bod značí odlehlou hodnotu. (Zdroj: Vlastní zpracování)

Asociační analýza

Asociační analýza spočívá v objevování asociačních pravidel (vazeb) ve velkých souborech dat. Asociační pravidla se vytvářejí tak, že se v datech hledají časté vzory typu *jestliže-pak* a pomocí kritérií podpory a důvěryhodnosti se identifikují nejdůležitější vztahy. Asociační pravidla jsou tvrzení typu *jestliže-pak*, která pomáhají zobrazit pravděpodobnost vztahů mezi datovými položkami v rámci rozsáhlých datových souborů v různých typech databází. Používá se například při analýze nákupního košíku, kdy se zjišťuje, jaké produkty zákazníci často nakupují společně. Pro asociační analýzu se široce používá algoritmus Apriori (6; 10; 11).

Regresní analýza a analýza trendů (nebo analýza vývoje)

Regrese je funkce pro dolování dat, která předpovídá číslo. Regresní model lze například použít k předpovědi hodnoty domu na základě polohy, počtu místností, velikosti pozemku a dalších faktorů (6; 12).

Analýza trendů (nazývaná také analýza vývoje) odhaluje zajímavé zákonitosti v historii vývoje objektů. Identifikace vzorců ve vývoji objektu a porovnávání měnících se trendů objektů jsou dva hlavní aspekty analýzy trendů (6).

2.3.3 Techniky data miningu

Úlohy dolování dat se provádějí na základě řady technik nebo přístupů. Například strojové učení, statistika, neuronové sítě, databázové systémy a datové sklady, genetické algoritmy, fuzzy množiny, vizualizace a další. V této podkapitole jsou dále popsány vybrané techniky (2; 6).

Statistické metody

Většina statistických modelů se obvykle vytváří na základě trénovacího souboru dat. Z modelu se pak vyvozují různá pravidla a vzory. Většina úloh dolování dat se provádí pomocí jednoho nebo více statistických přístupů.

Statistické metody běžně používané při dolování dat (6):

- Bayesovská síť – jedná se o pravděpodobnostní model, který využívá grafovou reprezentaci k zobrazení pravděpodobnostních vztahů mezi jednotlivými jevy. Reprezentuje znalosti o neurčité oblasti, kde každý uzel odpovídá náhodné veličině a každá hrana představuje podmíněnou pravděpodobnost pro příslušné náhodné veličiny (13; 14).
- Korelace – lineární závislost dvou veličin. Míru korelace lze vyjádřit korelačním koeficientem, který může nabývat hodnot -1 až 1. Hodnota 0 značí lineární nezávislost, hodnota blízká -1 značí nepřímou závislost a hodnota blízká 1 přímou závislost (15).
- Regresní analýza – zkoumá vztah mezi dvěma proměnnými – nezávisle proměnnou X a závisle proměnnou Y. Regresní analýza nám pomáhá pochopit, jak se změní hodnota závisle proměnné v návaznosti na změnu jedné z nezávisle proměnných (ostatní nezávisle proměnné zůstávají konstantní) (16).
- Shluková analýza – umožňuje klasifikovat objekty na základně podobnosti. Cílem je vytvořit shluky objektů tak, aby si objekty patřící do stejné skupiny byly podobnější než objekty z různých skupin (17).
- Diskriminační analýza – slouží k rozlišení objektů pocházejících z konečného počtu tříd na základě objektů z trénovací množiny sestavením rozhodovacího pravidla a následným zařazením zůstávajících objektů do jejich odhadovaných tříd (18).

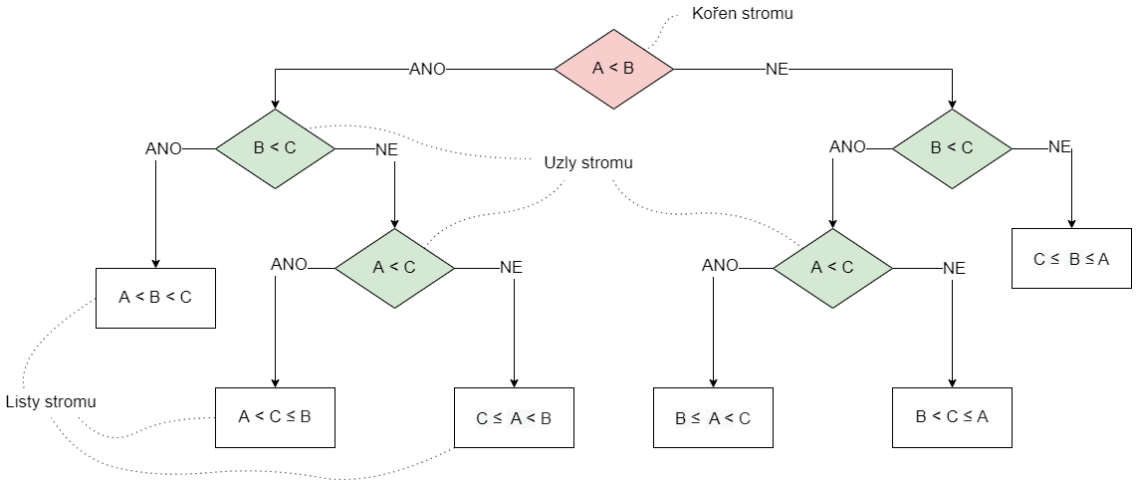
Rozhodovací stromy

Rozhodovací stromy jsou grafický nástroj pro podporu rozhodování (viz Obrázek 3). Rozdělují zkoumaná data podle určitých rozhodovacích kritérií. Kořen stromu zastupuje celý soubor a postupně probíhá větvení do dalších uzlů. Rozhodovací stromy jsou složeny z uzlů, hran a listů. Každý z uzlů rozhodovacího stromu představuje určitou vlastnost objektu a hrany představují možné hodnoty vlastností, listy určují predikovanou třídu (6).

Strojové učení (Machine Learning)

Strojové učení je studium výpočetních metod pro automatizaci procesu získávání znalostí z příkladů. Běžně používanou strategií je objevování vzorů v trénovací množině dat. Tento vzor je pak použit ke klasifikaci a/nebo předpovědi chování nových příkladů.

Strojové učení zvyšuje úroveň automatizace procesu zjišťování znalostí v databázích s cílem zvýšit přesnost a efektivitu (6; 19).



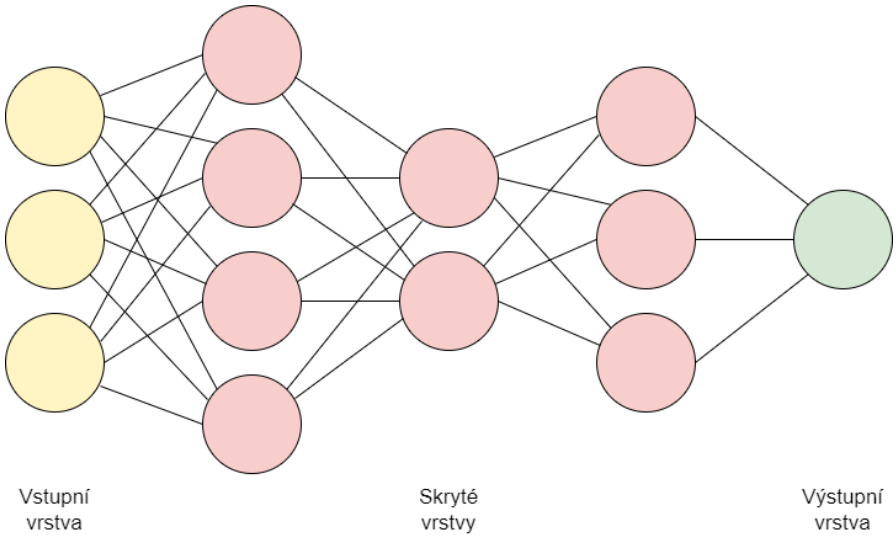
Obrázek 3: Příklad grafu větvení rozhodovacího stromu. (Zdroj: Vlastní zpracování)

Neuronové sítě

Umělá neuronová síť (artificial neural network - ANN), často nazývaná jen „neuronová síť“ (neural network – NN), je matematický model nebo výpočetní model založený na biologických neuronových sítích, jinými slovy je to emulace biologického neuronového systému. Příklad architektury neuronové sítě je znázorněn na Obrázku 4. Skládá se z propojené skupiny umělých neuronů a zpracovává informace pomocí spojového přístupu k výpočtu. Ve většině případů je ANN adaptivní systém, který mění svou strukturu na základě vnějších nebo vnitřních informací, které proudí (prochází) sítí během fáze učení (6; 20).

Databázové systémy a datové sklady

Pro dosažení škálovatelnosti a velké efektivity úloh dolování dat, které potřebují zpracovávat velké soubory dat, lze pro dolování dat využít databázové technologie. Hlavními metodami tohoto přístupu jsou iterativní prohledávání databáze za účelem zaměření na atributy, atributově orientovaná indukce a množiny častých položek. Vícerozměrná povaha struktury dat v datovém skladu podporuje také vícerozměrné dolování dat (6).



Obrázek 4: Příklad architektury neuronové sítě. (Zdroj: Vlastní zpracování)

Fuzzy množiny

Fuzzy množina definuje stupeň příslušnosti na základě hodnoty pravděpodobnosti vypočtené pomocí funkce příslušnosti. Je široce používána při klasifikaci a shlukové analýze (6).

Vizualizace

Technika dolování dat, která umožňuje identifikovat a reprezentovat vzory v datových souborech. Při vizualizaci se data převádějí na objekty, jako jsou body, čáry, plochy atd., které se zobrazují ve dvourozměrném nebo trojrozměrném prostoru (6).

2.3.4 Algoritmy data miningu (metody)

Algoritmy, také známé jako metody, slouží k provádění úloh data miningu založených na technikách data minigu. Například algoritmus Apriori, Naïve Bayesian, k-Nearest Neighbour, k-Means, CLIQUE, STING a další (6).

2.3.5 Vztah mezi úlohami data miningu a technikami data miningu

Úlohy data miningu se provádí pomocí jedné nebo více technik data miningu. V technice dolování dat lze použít jednu nebo více metod data miningu (6).

Tabulka 2 představuje úlohy dolování dat, které se provádějí na základě hlavních technik.

Tabulka 2: Úlohy a techniky data miningu. (Zdroj: Vlastní zpracování dle *(6)*)

		Úlohy data miningu						
Techniky data miningu		*Sumarizace*	*Charakterizace a diskriminace*	*Klasifikace*	*Shlukování*	*Asociace*	*Analýza odlehlých hodnot*	*Regresní analýza*
	Statistika	X	X	X	X	X	X	X
	Strojové učení		X	X	X	X	X	X
	Neuronové sítě		X	X	X	X	X	X
	Databázové systémy	X	X			X	X	
	Fuzzy množiny		X	X	X	X	X	
	Vizualizace		X	X	X		X	X

2.3.6 Domény data miningu

Data mining lze využít v řadě oblastí, např. data mining časových řad, web mining, temporal data mining, spatial data mining, tempo-spatial data mining, data mining ve vzdělávání, obchodu, medicíně, vědě, technice atd. Každá doména může mít jednu nebo více aplikací dolování dat (6).

2.3.7 Aplikace data miningu

Jedná se o soubor aplikačních oblastí, ve kterých lze použít jednu nebo více funkcí data miningu. Například analýza finančních dat, analýza nákupního košíku, detekce narušení, detekce podvodů, doporučovací systémy, detekce rakoviny a další (6).

2.3.8 Proces dolování dat

Data mining je iterativní proces zahrnující několik kroků, počínaje pochopením a definicí problému a konče analýzou výsledků společně s vytvořením strategie pro využití výsledků k získání výhod. Následující sekce popisuje základní kroky procesu dolování dat, tak jak jsou popsány v publikacích (2; 21; 22).

Definice problému (cíle)

Základním a zároveň prvním krokem v procesu dolování dat je definice problému. Je tedy potřeba jasně definovat s jakým cílem je analýza vykonávaná a na co budou její výsledky použity. Bez jasného pochopení problému může být výsledek bezcenný. Problém by tedy měl být dostatečně specifický, aby byl řešitelný a aby byly výsledky měřitelné.

Výběr dat

Po definici problému musíme definovat zdroje dat. Zdroje dat mohou být:

- Interní – data vznikající uvnitř firmy (DB zákazníků, zaměstnanců, produktů, …)
- Externí – z otevřených zdrojů (velikost sídla zákazníka, …)
- Produkční – data z operačních databází podniku (transakce, …)
- Archivní – uložená historická data (vyřazené produkty, …)

Ve většině případů jsou data získána ze stávající provozní databáze nebo datového skladu, který byl původně vytvořen pro různé analytické potřeby. Vybraná data jsou obvykle extrahována ze zdrojové databáze na zvláštní server, kde je realizován data mining, a uložena ve formátu vhodném pro data miningové algoritmy.

Příprava dat

Cílem je připravit data na aplikaci dolovacího algoritmu a zajistit co nejvyšší kvalitu vstupních dat. Mezi typické aktivity tohoto kroku patří:

- Čištění dat – zvýšení kvality dat (ošetřit chybějící hodnoty, odstranění šumu, …)
- Integrace dat – spojení/agregace dat z několika zdrojů
- Úprava datové sady – redukce (dimenzionality, počtu záznamů, počtu hodnot), řešení nevyváženosti
- Transformace dat – upravení hodnot do podoby vhodné pro dolování (agregace, normalizace, diskretizace numerických hodnot, …)

Příprava dat je časově nejnáročnější krok každého projektu data miningu a zároveň nejkritičtější – výsledné modely jsou pouze tak dobré, jak jsou dobrá data, která byla použita pro jejich vytvoření.

Dolování

Aplikace zvoleného algoritmu na předzpracovaná data dle typu znalosti a dat. Mezi typy znalostí řadíme:

- Asociační pravidla – hledání vazeb mezi objekty (např.: Analýza nákupního košíku využívaná k vyhledávání skupin a prvku, které mají tendenci vyskytovat se pospolu)
- Shlukování – seskupování podobných objektů
- Klasifikace – rozdělování objektů do předem známých tříd
- Predikce – přiřazovat datům hodnoty, které mají obecně spojitý charakter
- Detekce odchylek

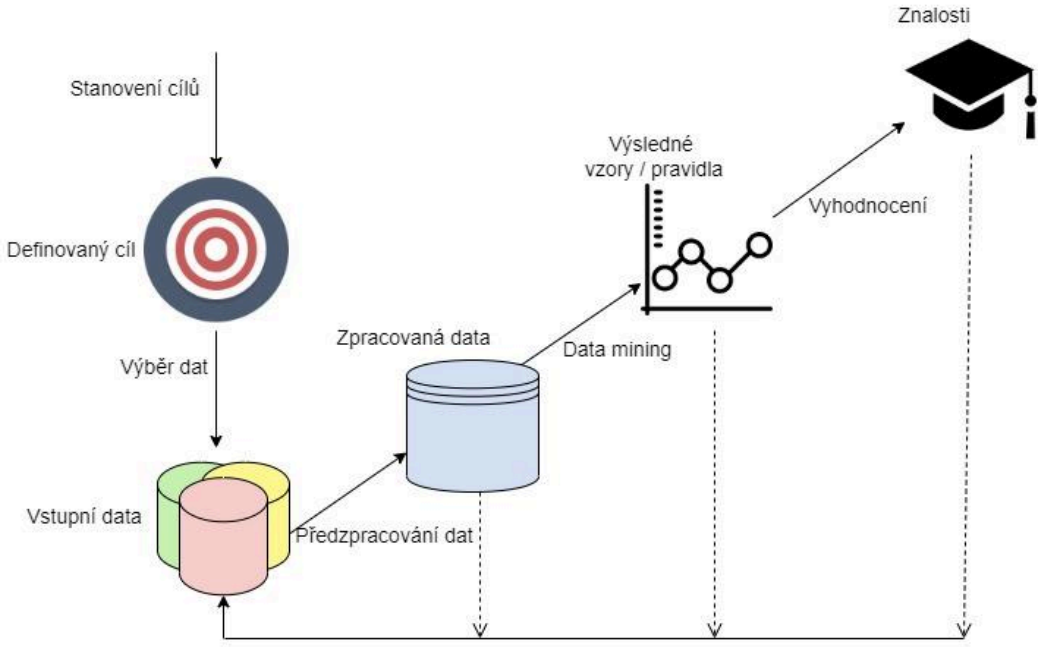
Vyhodnocení

Identifikace zajímavých vzorů (pravidel) a zhodnocení výsledků z pohledu zadání. Často návrat zpět na začátek celého procesu a stanovení nových cílů (úprava zadání).

Získání znalostí

Pochopení a prezentace výsledků data miningu použitím vizualizačních technik.

Proces data miningu je zobrazen na Obrázku 5.



Obrázek 5: Proces data miningu. (Zdroj: Vlastní zpracování)

2.4 RapidMiner

RapidMiner je univerzální softwarová platforma pro přípravu dat, Machine Learning, Deep Learning, text mining a prediktivní analýzu. RapidMiner (dříve YALE, Yet Another Learning Environment) byl vyvíjen od roku 2001 na oddělení umělé inteligence Technické univerzity v Dortmundu (7).

Jedná se o multiplatformní framework vyvinutý na základě otevřeného modelu jádra napsaného v jazyce Java. RapidMiner podporuje interaktivní režim (GUI), rozhraní příkazového řádku (CLI) a rozhraní Java API. Od verze 6.0 se jedná především o proprietární komerční produkt. Jeho architektura je založena na modelu klient/server, přičemž server je nabízen buď jako on-premise, nebo ve veřejných či soukromých cloudových infrastrukturách (Amazon AWS a Microsoft Azure). Pro rozsáhlou analýzu dat podporuje RapidMiner učení bez dohledu v prostředí Hadoop (Radoop), učení s dohledem v paměti se skórováním na clusteru (SparkRM) a skórování s nativními algoritmy na clusteru. V tomto případě je pokrytí algoritmů zúženo na Naive Bayes, lineární regresi, logistickou regresi, SVM, rozhodovací strom, náhodný les a shlukování pomocí k-means a fuzzy k-means (7).

3 Analýza současného stavu

V této kapitole se budu zabývat analýzou současného stavu společnosti, ve které práci zpracovávám. Jelikož si společnost nepřeje být v práci jmenována budu dále používat "XYZ" jako pracovní označení společnosti.

3.1 Základní informace o společnosti

Společnost XYZ patří mezi největší distributory výpočetní techniky v České republice. Mimo to se spolu se sesterskými společnostmi výrazně uplatňuje v celém regionu Česko-Slovensko-Polsko. Společnost vznikla v 90. letech 20. století zápisem do obchodního rejstříku.

Portfolio nabízených produktů zahrnuje úplný sortiment ICT trhu, tedy více než 300 značek nejvýznamnějších světových výrobců a přes 135 000 produktů. Společnost disponuje skladem o rozloze 20 000 m2 a špičkovým logistickým zázemím s více než 16 000 druhy produktů k okamžitému dodání.

Společnost je zároveň vlastníkem značky počítačů, EET pokladen a PC příslušenství. Jedná se o jednu z nejprodávanějších značek české výpočetní techniky s momentálním prodejní základnou více než 150 prodejců.

3.1.1 Hlavní poskytované služby

Mezi hlavní poskytované služby společnosti XYZ patří:

- Pomoc při dostávání zboží menších značek na trh
- Marketingová podpora distribuce produktů k prodejcům
- Zajištění skladové dostupnosti pro koncový trh
- Představení produktů prodejcům prostřednictvím obchodních eventů
- Zaškolení prodejců
- Zalistování produktů do vlastního katalogu

3.2 Vybavení společnosti

Tato sekce popisuje hardwarové a softwarové vybavení společnosti.

3.2.1 Hardware

Při nástupu do společnosti dostane každý zaměstnanec vlastní notebook značky HP, který používá v kanceláři nebo na setkáních se zákazníky/dodavateli. Zároveň jsou pracovní notebooky vhodné pro využívání v rámci home office a obchodních jednání, kdy se může zaměstnanec připojit pomocí VPN. Mimo to je každému zaměstnanci přiděleno pracovní místo v kanceláři s minimálně jedním monitorem o úhlopříčce 24", klávesnicí a myší. Na pobočkách nechybí ani multifunkční tiskárny.

3.2.2 Software

Pro správný a efektivní chod společnosti je využíváno více programů. Na počítačích je nainstalován operační systém Windows 10 společnosti Microsoft. Každý zaměstnanec má vlastní uživatelské jméno a heslo, které používá pro přístup do počítače, mailu, k VPN i informačního systému. Mezi další programy používané na denní bázi se řadí zejména:

Microsoft Excel

Program Microsoft Excel je jeden z nejpoužívanějších nástrojů ve společnosti. Je používán například pro zpracování většího množství dat do tabulek, zalistování nových produktů zákazníkům, uchování logistických informací apod. V rámci tabulek jsou následně využívány jednodušší funkce pro zpracování požadovaných výsledků.

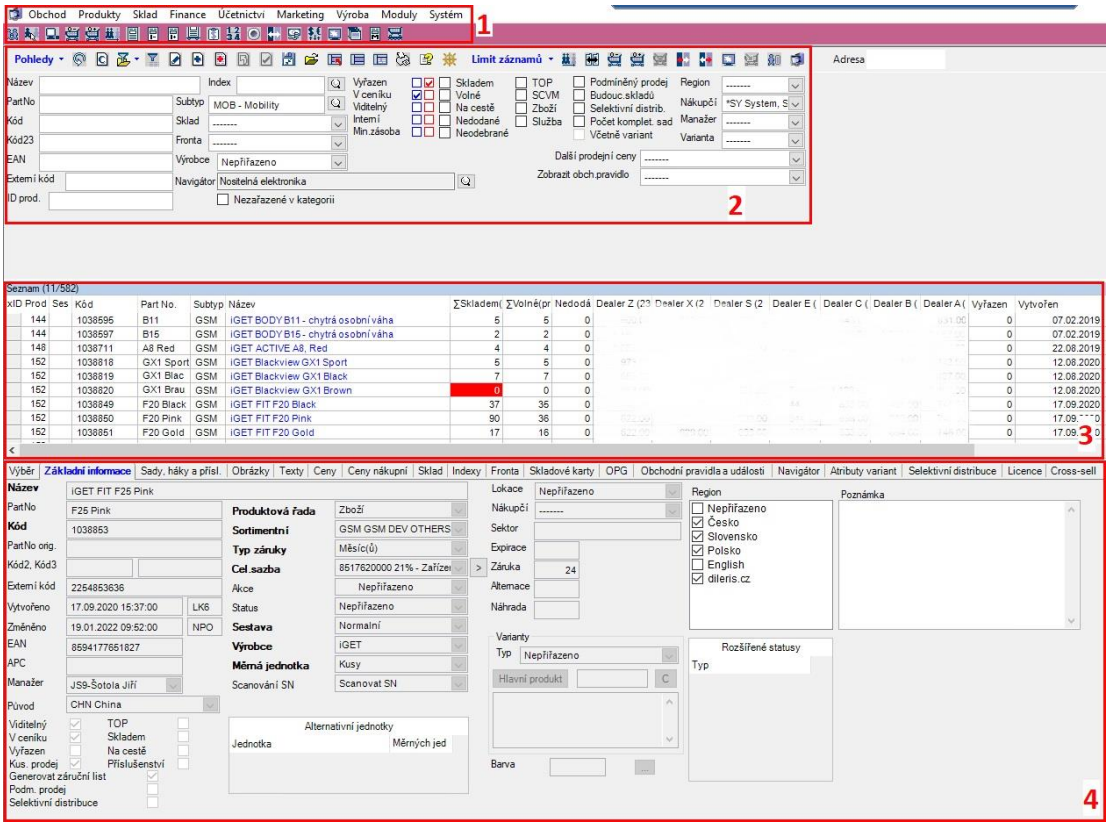
Komunikační kanály

Mezi komunikační kanály uvnitř společnosti patří elektronická pošta řešená za pomoci aplikace Outlook, pro textovou komunikaci, hovory a videokonference je využíván nástroj Microsoft Teams, osobní setkání a telefonní komunikace.

Na komunikaci s dodavateli a zákazníky využívá společnost hlavně elektronickou poštu a telefonní komunikaci. Pro komunikaci s veřejností využívá společnost sociální sítě LinkedIn a Facebook především pro propagaci společnosti, oslovení potenciálních zákazníků a sdílení nových informací.

Informační systém ESYCO

Společnost využívá vlastní ERP (plánování podnikových zdrojů) systém ESYCO, který pokrývá základní ekonomické, obchodní, účetní a evidenční funkce. Skládá se z jednotlivých modulů, které jsou navzájem propojeny a poskytují tak ucelené informace o chodu společnosti. Předností tohoto systému je také možnost řídit e-shop a skladové zásoby z jednoho místa. Snímek obrazovky z informačního systému je zobrazen na Obrázku 6.



Obrázek 6: Ukázka informačního systému. 1 – moduly IS, 2 – filtr produktů, 3 – seznam vyfiltrovaných produktů, 4 – detail vybraného produktu. (Zdroj: Vlastní zpracování)

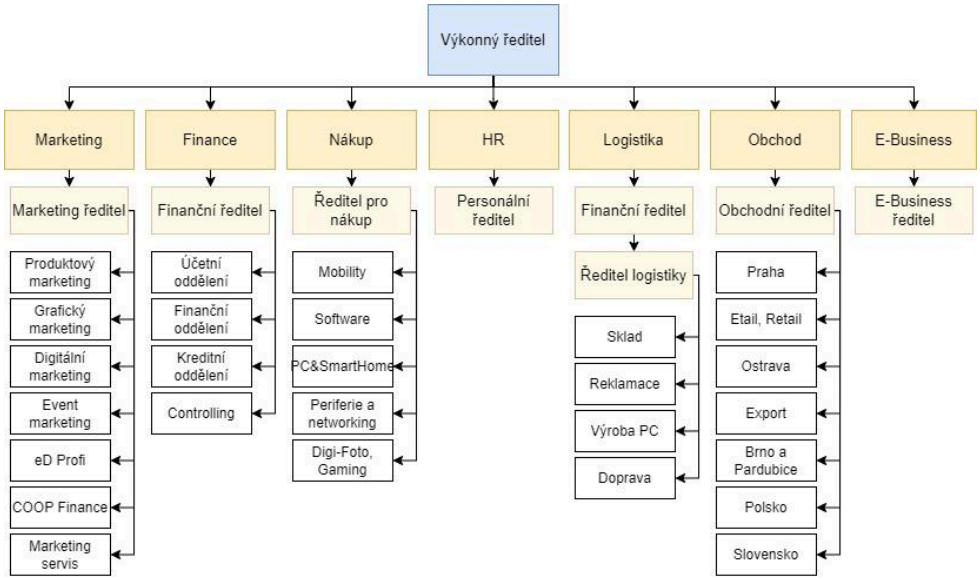
Moduly informačního systému a jejich části (viz Obrázek 6 sekce 1):

- Obchod – informace o klientech (dodavatelé i odběratelé), prodejní i nákupní doklady (poptávky, nabídky, objednávky, zápůjčky, faktury, …), dodací listy, dodávky (expedice), záruky
- Produkty – přehled produktů (produktové karty), navigátor (nastavení hierarchie produktů pro e-shop)

- Sklad – reklamace a servis, skladové přesuny, stáří skladu, inventura, montáž / demontáž, centrální kurz
- Finance – evidence plateb, kurzovní lístek, zápočty, platební příkazy, evidence smluv, platby bankou a pokladnou
- Účetnictví – účetní doklady, hlavní kniha, obraty účtů, předvaha, rozvaha, výsledovka, deník majetku
- Marketing – marketingové kampaně, věrnostní kampaně, úkoly, projekty, příležitosti, aktivity
- Výroba – položky výrobních příkazů, plánování výroby, odvádění výroby
- Další – statistiky prodeje a nákupů, docházka, evidence slev, potvrzování cen objednávek, zboží na cestě, sestavy a BI

3.3 Organizační struktura

Ve společnosti pracuje přes 330 stálých zaměstnanců. Výkonný ředitel zajišťuje komplexní vedení a řízení celé společnosti. Na každém oddělení je ředitel, který zpracovává a řídí hlavní procesy na svém oddělení a kontroluje ekonomickou bilanci svého střediska. Organizační struktura se následně dělí na jednotlivé oblasti, z nichž každá disponuje manažerem. Pro větší přehlednost je organizační struktura společnosti zobrazena na Obrázku 7.



Obrázek 7: Organizační struktura společnosti. (Zdroj: Vlastní zpracování)

3.3.1 Hlavní činnosti jednotlivých oddělení

Společnost má vícero oddělení a každé zastává odlišnou funkci. V této podkapitole jsou pospány základní činnosti každého oddělení:

- Marketing – příprava marketingových aktivit, eventů a promoakcí, výroba vizitek, bannerů, grafických materiálů, dárky zákazníkům, reklamní předměty, příprava prezentačních materiálů, katalogů, eD web, digitální marketing
- Nákup – komunikace s dodavateli, zalistování nových produktů do systému, objednávání a nákup zboží, management skladu a komodit, schvalování cen a cenotvorba, smluvní dokumentace
- Obchod – strategické řízení společnosti, rozvoj obchodních vztahů, řízení obchodních aktivit, fakturace, monitoring trhu, zákaznické průzkumy, smluvní dokumentace
- E-Business – řízení e-commerce projektů, správa eD shopu pro B2B, správa aplikací, elektronických datových výměn, změnové procedury v IS platformách, implementace automatizace procesů
- Finance – účetnictví, pokladna, účetní uzávěrky, reportování výsledků, archivace dokladů, kompletní finanční řízení, platby faktur, nákup deviz, zajišťování proti kurzovému riziku, řízení všech finančních rizik, řízení cash flow, odsouhlasování bilancí s dodavateli, odsouhlasování finančních podmínek obchodních smluv
- Logistika – řízení nákupu, dodávek, dopravy a distribuce zboží, controlling, zlepšování procesů logistický služeb podle nejmodernějších trendů
- HR – nábor zaměstnanců a adaptace, jejich vzdělávání a motivace, teambuildingové aktivity, talent management, odměňovací a bonusový systém, hodnocení zaměstnanců, benefitní systém, HR marketing, správa interního blogu, správa kariérních stránek, studentský program, ISO certifikace, EU projekty

3.4 Business Intelligence ve společnosti

Tato podkapitola popisuje jeden z případů využití business intelligence ve společnosti – zprostředkování B2C e-shopů. Tento proces zahrnuje analýzy prodeje a oblíbenosti produktů pro daný e-shop. Další podkapitoly zahrnují detailnější popis tohoto procesu a zadání úlohy pro zvýšení efektivity analýz vykonávaných v rámci tohoto procesu.

3.4.1 Analýza prodeje produktů

Společnost XYZ se kromě B2B (Business to business) podnikání zabývá i zprostředkováním B2C (Business to customer) e-shopů značkám, které nemají v České republice zastoupení. Jedná se o takzvané *brandpage*. Zajišťují tak lepší dostupnost produktů těchto značek pro český trh. Důležitým aspektem tvorby a provozu takového e-shopu je výběr a promování produktů různých kategorií, o který je aktuálně na trhu největší zájem a poptávka. Tyto produkty jsou vybírány na základě analýzy dosavadních prodejů, oblíbenosti a popularity. To celé je podloženo znalostmi expertů na dané značky, kteří mají dlouholeté zkušenosti z prodeje produktů přímo koncovým zákazníkům.

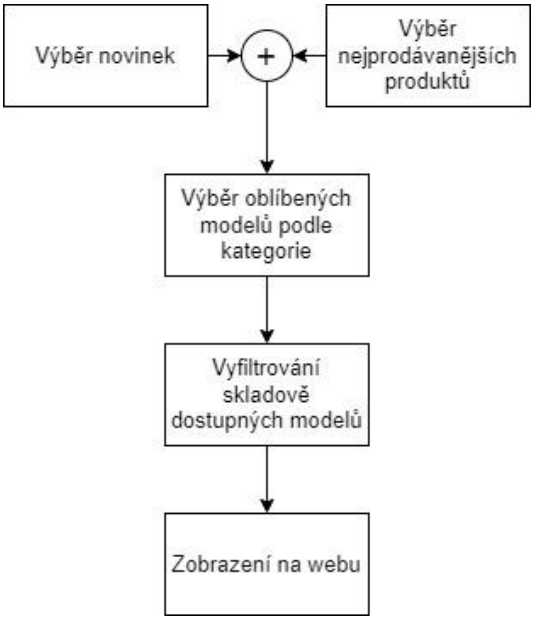
Výběr produktů pro zobrazení

Proces výběru produktů pro zobrazení na webu se skládá z následujících kroků:

1. Výběr novinek – výběr čerstvě vydaných modelů
2. Výběr nejzajímavějších produktů – jedná se o produkty, které jsou nejvíce inovativní, designově zajímavé, cenově zajímavé a nejprodávanější[1]
3. Výběr oblíbených modelů a konfigurací podle kategorie – z 1. a 2. kroku jsou vybrány modely oblíbených řad a jejich konfigurace pro každou kategorii, do které je produkt zařazený
4. Vyfiltrování skladově dostupných modelů – porovnání seznamu vybraných produktů se skladovými zásobami a úprava tak, aby nedošlo k promování nedostupných produktů
5. Zobrazení na webu – nahrání vybraných a vyfiltrovaných produktů na brandpage

[1] V posledních dvou letech jsou nejprodávanější produkty výrazně zkresleny tím, co je právě skladem. Obecný nedostatek různých komponentů omezil dodávky především dražších a novějších modelů.

Proces výběru produktů je vizualizovaný na Obrázku 8.



Obrázek 8: Diagram procesu výběru top produktů pro zobrazení na brandpage. (Zdroj: Vlastní zpracování)

3.4.2 Zvýšení efektivity analýzy

Proces analýzy, popsaný v předcházející podkapitole, chce společnost zautomatizovat a zjistit, zda ho lze vykonávat bez expertních znalostí produktů dané značky. Společnost tím ušetří čas, lidské zdroje a analýzu bude možné vykonávat častěji. Mnou provedené analýzy budou následně porovnány s konkurencí pro porovnání správnosti výsledků. Dále bude analýza sloužit jako vzor pro další nabízené značky a určí přesnost aktuálního provedení.

4 Vlastní návrh řešení

Tato kapitola obsahuje návrh a realizaci mého vlastního řešení. Prvně je uvedeno zadání řešeného problému a návrh jeho řešení. Následně popisuji realizaci tohoto návrhu pomocí mnou navržených metrik a data miningových technik. Závěrem porovnávám výsledky mého řešení s konkurencí.

4.1 Specifikace zadání

Zadáním je analyzovat prodeje produktů značky ABC za poslední dva roky. Obsahem analýzy budou 4 kategorie produktů – notebooky, monitory, tiskárny a PC sestavy. Cílem je napříč všemi kategoriemi vybrat TOP 100 produktů pro další zobrazení na brandpage dané značky. Dalším požadavkem je, aby tato analýza probíhala automatizovaně a její výsledky bylo možné porovnávat s konkurencí.

4.1.1 Postup analýzy

Pro vytvoření procesu výběru TOP produktů budu používat nástroj RapidMiner, protože poskytuje licence pro studenty a umožňuje vytvoření automatizovaných procesů. Data, která jsem od společnosti dostala, bude nejprve nutné očistit a upravit. Bude potřeba odstranit duplicity, které reprezentují produkty prodané v rámci akcí a bazaru. Dále pak rozdělit produkty podle kategorií a v každé kategorii určit cenové hladiny produktů. Pro určení cenových hladin bude vyzkoušeno více přístupů, aby bylo zjištěno, který funguje nejlépe. Jako jedna možnost se nabízí využít cenové hladiny nastavené konkurencí (Alza, Heureka), nebo pomocí různých metod diskretizace v rámci nástroje RapidMiner. Následně v každé kategorii a cenové hladině analyzovat prodeje produktů a najít nejprodávanější produkty, prozkoumat trendy prodejů a případně získat jiná zajímavá data o prodejích.

Vybrané nejprodávanější produkty budou porovnány s portálem Heureka a jejich seznamem TOP produktů pro danou značku. Dále je potřeba zaznamenat shodu/neshodu v seznamu produktů. V případě neshody (produkt není v seznamu společnosti, ale na portálu Heureka ano) poznačím odkaz, hlavní cenu na Heurece a prodeje, které společnost eviduje. V případě shody zapíšu odkaz a cenu. Důvodem tohoto kroku je zjistit, jaké produkty se liší a proč se tyto produkty liší.

4.1.2 Vstupy analýzy

Společnost mi poskytla výpisy z databáze prodejů dané značky za posledních 24 měsíců. Struktura dat je znázorněna v Tabulce 3.

Tabulka 3: Struktura zdrojových dat. (Zdroj: Vlastní zpracování)

Atribut	Datový typ
Kód produktu podniku	Číslo
Název produktu	Textový řetězec
Produktové číslo	Textový řetězec
Výrobce	Textový řetězec
Kategorie produktu	Textový řetězec
Doporučená cena (B2C)	Číslo
Cena pro zákazníky (B2B)	Číslo
Nákupní cena	Číslo
Počet prodejů v jednotlivých měsících	Číslo

4.1.3 Výstup analýzy

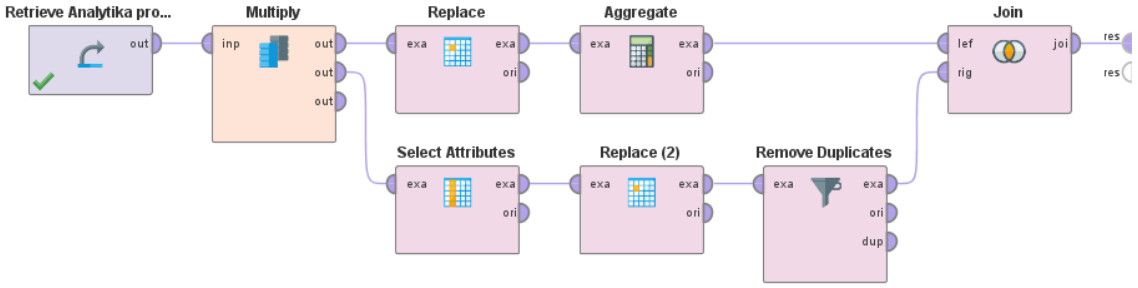
Výstupem bude rozdělení produktů do kategorií podle cenových hladin a případně v podkategoriích (kancelářské, herní, …) a v každé cenové hladině nalezení TOP produktů. Následně se výsledky porovnají s konkurencí. Vybrané produkty budou poté zobrazeny na brandpage analyzované značky.

4.2 Úprava datové sady

Vstupní datová sada obsahuje duplicitní položky. Jedná se většinou o bazarové a promo produkty. Tyto produkty je možné rozlišit pomocí jejich produktového čísla, které v případě bazarových, promo a podobných produktů obsahují extra řetězec začínající znaky „//“ a pak informaci, o jaký druh produktu se jedná (např.: 0000000#AAA//Bazar).

Z produktového čísla jsem prvně odstranila extra řetězce, čímž vznikla duplicitní produktová čísla. Duplicity jsem odstranila pomocí agregace: seskupení podle produktového čísla a suma prodejů za jednotlivé měsíce. Agregované mohly být pouze počty prodejů, a proto bylo potřeba přidat chybějící atributy pomocí operátoru *Join*.

Druhý vstup do tohoto operátoru byl očištěn, aby neobsahoval duplicitní hodnoty. Proces úpravy datové sady je zobrazen na Obrázku 9.



Obrázek 9: Proces úpravy datové sady. (Zdroj: Vlastní zpracování)

4.3 Rozdělení produktů

Produkty je potřeba rozdělit do kategorií podle typu produktu (monitory, tiskárny, notebooky a PC sestavy) a následně tyto produkty rozdělit do cenových hladin. Postup a výsledky jsou popsány v následujících podkapitolách.

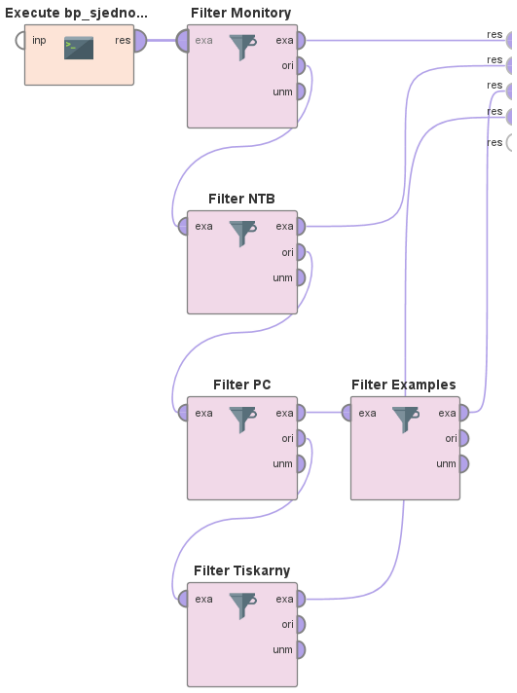
4.3.1 Rozdělení produktů podle kategorie

Očištěná datová sada obsahuje produkty všech výše zmíněných kategorií dohromady. Pro efektivní práci s daty je nutné kategorie rozdělit do samostatných sad.

Očištěná data jsem pomocí filtrů rozdělila do jednotlivých kategorií a výsledky exportovala jako samostatné výstupy procesu pro další použití. Rozdělení do podkategorií herní, kancelářské atd. nebylo pouze s využitím vstupních dat možné, a proto jsem se tím dále nezabývala. U kategorie PC sestav se vyskytují i špatně zařazené produkty z kategorie notebooků. Proto jsem do procesu přidala ještě jeden filtr, který odstraňuje produkty s řetězcem NTB v názvu produktu. Proces filtrace a exportu je znázorněný na Obrázku 10.

4.3.2 Rozdělení produktů do cenových hladin

V každé kategorii produktů je potřeba určit cenové hladiny. Cenová hladina je definována intervalem *<min;max)*, kde *min* a *max* reprezentuje cenu v Kč. K určení cenových hladin jsem na doporučení společnosti porovnávala dva přístupy:



Obrázek 10: Rozdělení produktů do kategorií. (Zdroj: Vlastní zpracování)

1. Určení cenových hladin podle konkurence (Heureka, Alza)
2. Určení cenových hladin pomocí diskretizace

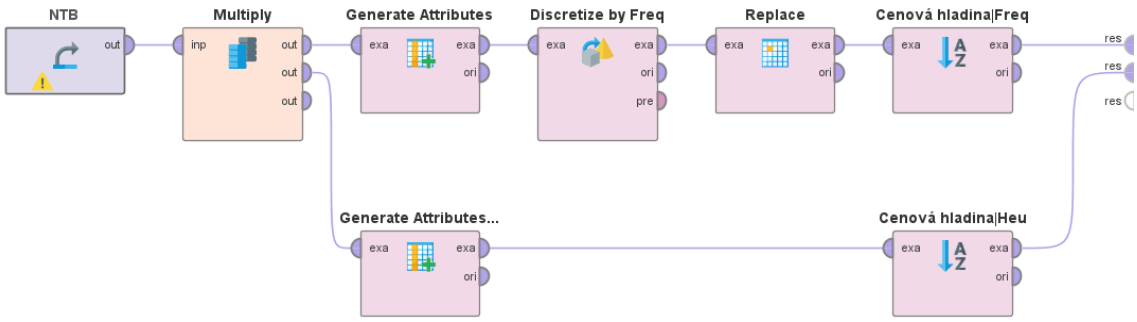
Oba přístupy určení cenových hladin probíhaly současně v jednom procesu. Nejdříve bylo třeba nakopírovat atribut ceny, protože by se při diskretizaci ztratil. Z dostupných cen jsem zvolila doporučenou cenu (B2C), protože tuto cenu vidí koncový zákazník.

Určování cenových hladin podle konkurence probíhalo pouze podle Heureky, protože Alza na svém webu nemá toto rozdělení definováno. Pro každou kategorii jsem definovala cenové hladiny podle rozdělení na Heurece, jak znázorňuje Tabulka 4. Pomocí operátoru *Generate Attributes* jsem vytvořila nový atribut „Cenová hladina“ a s použitím série podmínek přiřadila každému produktu cenovou hladinu. Pro lepší vizualizaci je nutné produkty seřadit dle atributu Cenová hladina.

Tabulka 4: Cenové hladiny Heureka. (Zdroj: Vlastní zpracování)

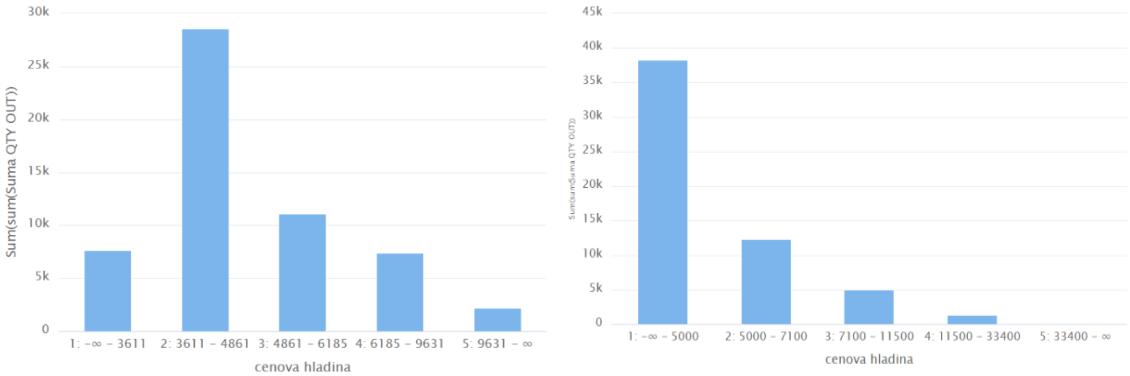
	Kategorie	*NTB*	*Monitory*	*Tiskárny*	*PC sestavy*
Cen. hlad.	1:	-∞ – 16000	-∞ – 5000	-∞ – 5500	-∞ – 14000
	2:	16000 – 21000	5000 – 7200	5500 – 11000	14000 – 18000
	3:	21000 – 28000	7200 – 11700	11000 – 22000	18000 – 25000
	4:	28000 – 41000	11700 – 33700	22000 – 52500	25000 – 41000
	5:	41000 – ∞	33700 – ∞	52500 – ∞	41000 – ∞

Určování cenové hladiny pomocí diskretizace jsem prováděla na základě frekvence produktů. Každá cenová hladina tedy obsahuje stejný počet produktů. Výsledkem bylo rozdělení do 5 cenových hladin stejně jako v předcházejícím přístupu pro jejich jednoduché porovnání. Pomocí operátoru *Replace* byly přejmenovány hodnoty atributu Cenová hladina tak, aby měly stejný formát jako v předchozím přístupu. Proces tohoto rozdělení je znázorněn na Obrázku 11.



Obrázek 11: Rozdělení do cenových hladin. (Zdroj: Vlastní zpracování)

Na základě analýzy výsledků obou přístupů se jeví jako vhodnější použití diskretizace na základě frekvence. Při použití tohoto přístupu dochází k jemnějšímu rozdělení cenových hladin a díky tomu nevznikají prázdné cenové hladiny. Rozdíl mezi použitými přístupy je nejlépe viditelný na analýze monitorů (viz Obrázek 12).



Obrázek 12: Porovnání rozložení produktů do cenových hladin při jejich určení pomocí diskretizace (vlevo) a portálu Heureka (vpravo). (Zdroj: Vlastní zpracování)

4.4 Výběr top produktů dle prodejů

V předcházejících podkapitolách byl popsán proces rozdělení produktů do kategorií a proces určení cenových hladin v nich. Tato podkapitola popisuje proces výběru TOP produktů dle prodejů za poslední dva roky. Dle zadání je cílem vybrat TOP 100 produktů napříč všemi kategoriemi a cenovými hladinami.

4.4.1 Určení počtu vybraných produktů podle kategorie a cenové hladiny

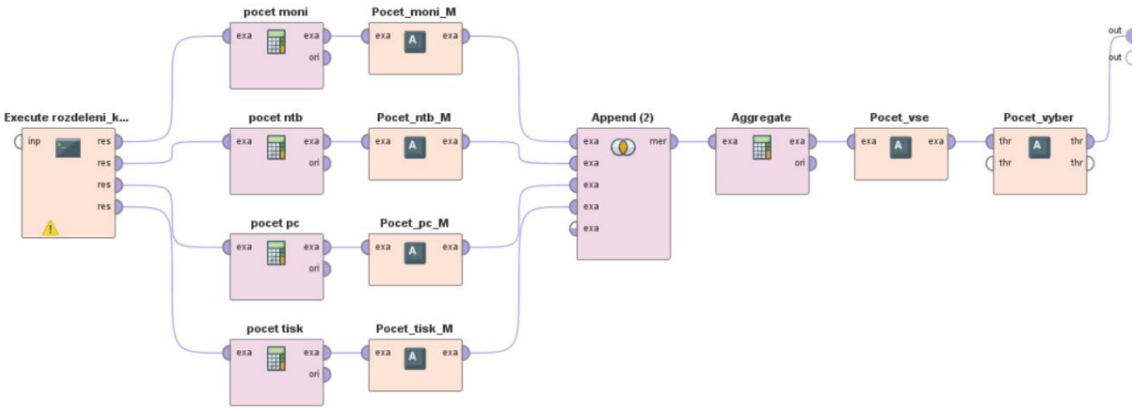
Počet vybraných produktů z každé kategorie a cenové hladiny závisí na celkovém počtu prodejů v daných kategoriích a cenových hladinách. Při rozdělení pouze podle procentuálního zastoupení prodejů každé třídy docházelo k nevhodnému rozložení, kdy například v nejvyšší cenové hladině u tiskáren nebo monitorů nebyly vybrány žádné top produkty.

Pro získání vhodnějšího rozložení jsem se rozhodla z každé cenové hladiny vybrat vždy minimálně dva produkty. Pro splnění této podmínky tedy musí být z každé kategorie vybráno alespoň 10 produktů. Bez ohledu na procentuální zastoupení tříd je vždy vybráno TOP 40 produktů. Zbylých 60 produktů je vybráno na základě procentuálního zastoupení prodejů každé třídy.

Výběr počtu produktů z kategorie je navržený jako podproces, který se pro výběr produktů v každé kategorii spouští zvlášť. Vstupem jsou data rozdělená do kategorií. Pro každou kategorii je pomocí operátoru *Aggregate* vypočítaný celkový počet prodejů a zapsán do makra. Operátor *Append* znovu sjednotí všechny produkty do jedné tabulky, z které je pomocí operátoru *Aggregate* vypočítána suma všech prodejů a zapsána do makra. Posledním krokem je výpočet, kolik ze zbylých 60 produktů bude vybráno právě z dané kategorie (viz Rovnice 1) a následné uložení tohoto počtu do makra. Celý podproces je znázorněný na Obrázku 13.

$$Počet\ výběr \doteq \left(\frac{Počet\ prodeje\ kategorie}{Počet\ prodeje\ celkem}\right) * 60$$

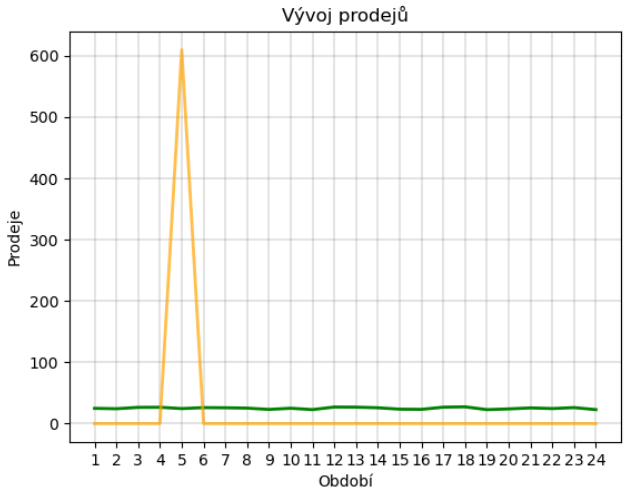
Rovnice 1: Výpočet počtu vybraných produktů z jedné kategorie. (Zdroj: Vlastní zpracování)



Obrázek 13: Určení počtu vybraných produktů podle kategorie a cenové hladiny. (Zdroj: Vlastní zpracování)

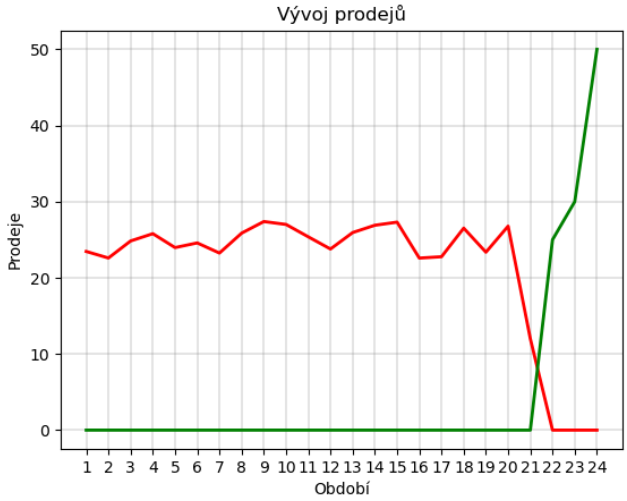
4.4.2 Vytvoření metriky

Pro výběr top produktů je nutné zvolit vhodnou metriku. Prvním kritériem jsou stabilní prodeje produktu za analyzované období. Pro výsledný výběr je hodnotnější produkt, který se ve sledovaném období prodává neustále než produkt, který má ve výsledku vyšší počet celkových prodejů, ale tyto prodeje vznikly v jednom nebo dvou měsících. Rozdíl mezi prodeji takových produktů je znázorněn na Obrázku 14.



Obrázek 14: Motivace pro vznik prvního kritéria: Rozdíl v prodejích dvou produktů. Cílem je upřednostnit produkt znázorněný zelenou křivkou. (Zdroj: Vlastní zpracování)

První kritérium však nezohledňuje nové produkty, ani zvýšení (např.: nový produkt) nebo snížení (např.: v důsledku vyřazení produktu z nabídky) prodejů za poslední měsíce. Pro odstranění těchto nedostatků jsem přidala druhé kritérium, kterým je počet prodaných



Obrázek 15: Motivace pro vznik druhého kritéria: Rozdíl v prodejích dvou produktů v posledních měsících. Cílem je upřednostnit produkt znázorněný zelenou křivkou. (Zdroj: Vlastní zpracování)

produktů za poslední tři měsíce. Motivace pro vznik druhého kritéria je znázorněna na Obrázku 15.

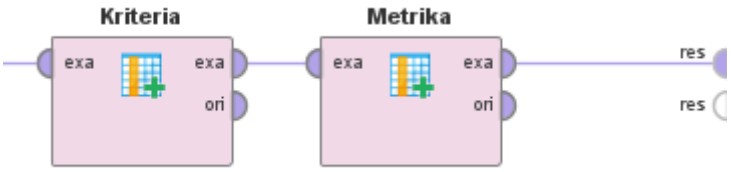
Pro číselné ohodnocení prvního kritéria jsem zvolila geometrický průměr, který je schopný vyrovnat nárazové skoky v prodejích. Zároveň jsem přidala každému čtvrtletí váhy tak, aby prodeje v posledním čtvrtletí byly vyhodnoceny jako nejdůležitější a váha předcházejících období se postupně snižovala. Vzhledem k vlastnostem geometrického průměru a vyskytujícím se nulovým prodejům v některých měsících u mnoha produktů bylo nutné při výpočtu připočítat k prodejům za každé období nenulovou kladnou hodnotu. Tato hodnota byla nastavena na 1.

Číselné ohodnocení druhého kritéria jsem provedla pomocí průměru prodejů za poslední tři měsíce. Kombinací těchto dvou kritérií vznikla výsledná metrika pro ohodnocení produktů:

$$Metrika = kritérium\ 1 * kritérium\ 2$$

Rovnice 2: Vzorec výpočtu metriky. (Zdroj: Vlastní zpracování)

Zapojení procesu pro výběr metriky je ukázáno na Obrázku 16. Tímto výpočtem metriky jsou zvýhodněny právě ty produkty, které se ve sledovaném období stabilně prodávají, produkty s vysokými prodeji v posledním období a produkty nové.



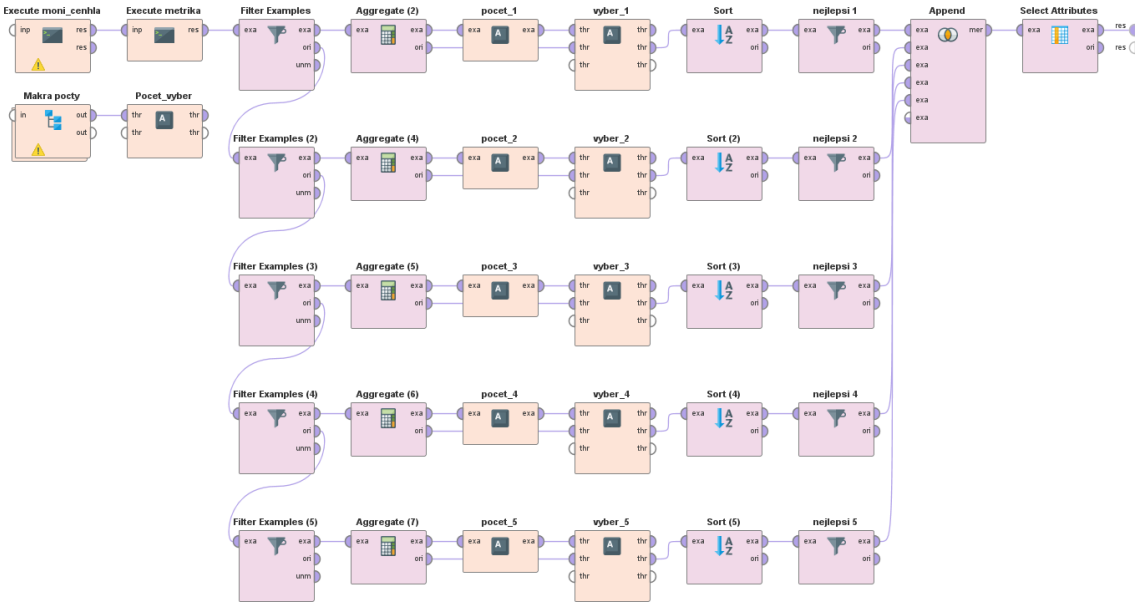
Obrázek 16: Zapojení procesu pro výpočet metriky. (Zdroj: Vlastní zpracování)

4.4.3 Proces výběru TOP produktů

Pro každou kategorii probíhá výběr produktů v samostatném procesu. Vstupem procesu jsou data s přiřazenými cenovými hladinami. Následně jsou pro každý produkt vytvořeny nové atributy *Kritérium 1* a *Kritérium 2* a z nich vypočítán atribut *Metrika*. Paralelně s tím se spouští i podproces (viz Obrázek 13), který nastaví do maker celkový počet prodejů produktů v každé kategorii a celkový součet těchto prodejů (viz podkapitola 4.4.1). Pomocí těchto maker je vypočítáno, kolik produktů se má z dané kategorie ze zbylých 60 produktů (viz podkapitola 4.4.1) vybrat.

Vstupní data jsou rozdělena operátorem *Filter* do cenových hladin. V každé cenové hladině se pomocí operátoru *Aggregate* opět sečtou celkové prodeje a uloží do makra. Následně se pomocí dříve uložených maker vypočítá počet vybraných produktů z dané cenové hladiny (2 + *dle procentuálního zastoupení*) a vypočítaná hodnota se opět nastaví jako makro. Všechny produkty v cenové hladině jsou sestupně seřazeny pomocí operátoru S*ort* podle metriky. Pomocí operátoru *Filter* je vybrán výsledný počet produktů vycházející z dříve vypočtených maker. Operátorem *Append* jsou výsledky z každé cenové hladiny sloučeny do jedné tabulky a jako poslední jsou pomocí operátoru *Select Atributes* vybrány jen důležité atributy (název, PN, metrika, kategorie, celkové prodeje, cena, …). Celý proces výběru top produktů je zobrazen na Obrázku 17.

Výstupem tohoto procesu jsou 4 tabulky, které dohromady zobrazují TOP 100 vybraných produktů.



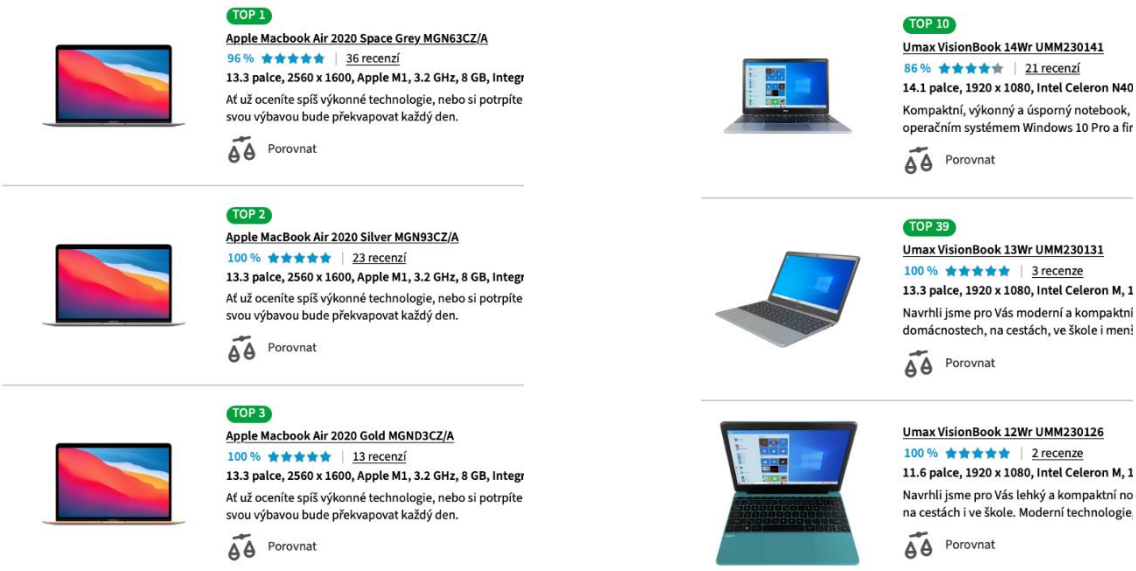
Obrázek 17: Proces výběru TOP produktů. (Zdroj: Vlastní zpracování)

4.5 Porovnání s konkurencí

Po výběru TOP produktů bylo potřeba porovnat výběr s portálem Heureka. Cílem bylo ověřit, zda mají v top produktech stejné produkty, a v případě neshod zjistit příčinu. Toto porovnání proběhlo v půlce března, kdy jsem měla k dispozici firemní data za období březen 2020 - únor 2022.

Portál Heureka pro každou kategorii zařízení nabízí seznam TOP 50 produktů napříč různými značkami. Po vyfiltrování produktů podle vybrané značky a seřazení dle oblíbenosti se prvně zobrazují produkty vybrané značky ze seznamu TOP 50 a následně ostatní produkty bez štítku TOP. Toto chování je znázorněno na Obrázku 18.

Pro porovnání mého řešení s portálem Heureka využívám dva přístupy. V prvním přístupu porovnávám jen produkty se štítkem TOP a v druhém vybírám z každé kategorie na Heurece stejný počet nejoblíbenějších produktů jako je ve výstupu z mé analýzy (např.: 40 notebooků, 17 tiskáren, …).



Obrázek 18: Řazení TOP produktů na portálu Heureka. (Zdroj: Vlastní zpracování)

4.5.1 První přístup (produkty se štítkem TOP)

Tabulka 5 zobrazuje počty shodných produktů z Heureky se štítkem TOP s mnou vybranými TOP 100 produkty. V kategorii tiskáren a monitorů je shodných vysoké procento produktů, toto procento klesá u notebooků a u PC sestav, kde je procento velmi nízké.

Tabulka 5: Výsledky porovnání prvním přístupem. (Zdroj: Vlastní zpracování)

Kategorie	*Počet Heureka*	*Počet stejných*	*Úspěšnost (%)*
NTB	13	8	61,6
Tiskárny	12	10	83
PC sestavy	6	1	16,7
Monitory	3	3	100

Při další analýze výsledků jsem zjišťovala příčiny neshod. Nejčastější příčinou, proč se produkt vyskytující se na portálu Heureka v TOP a v mém výstupu nikoliv, byly nízké počty prodejů za celé sledované období anebo rapidní pokles prodejů v posledních měsících.

Nízké procento shody u PC sestav je pravděpodobně způsobeno širokou variabilitou této kategorie. Existuje velké množství v podstatě stejných produktů, které se liší například pouze jedním komponentem, což je z pohledu analýzy bráno jako odlišný produkt. Tato

skutečnost se odráží i na kategorii notebooků. Tato hypotéza momentálně není podložena bližším výzkumem a do budoucna může být zajímavým předmětem zkoumání.

4.5.2 Druhý přístup (100 produktů z Heureky)

Tabulka 6 zobrazuje počty shodných produktů z Heureky dané značky a kategorie seřazených dle oblíbenosti a mnou vybranými TOP 100 produkty. Stejně jako u předchozího přístupu je procento shodných PC sestav nejnižší a u ostatních kategorií procento kleslo přibližně na 50 %.

Tabulka 6: Výsledky porovnání druhým přístupem. (Zdroj: Vlastní zpracování)

Kategorie	*Počet celkem*	*Počet stejných*	*Úspěšnost (%)*
NTB	40	16	40
Tiskárny	17	10	58,9
PC sestavy	17	3	17,6
Monitory	25	14	56

Následně jsem stejně jako v předcházejícím přístupu hledala příčiny neshod mezi mými výsledky a výsledky z Heureky. Závěr je podobný jako u prvního přístupu, a to nízké počty prodejů za celé sledované období nebo rapidní pokles prodejů v posledních měsících.

Při analýze PC sestav jsem zjistila, že přes polovinu produktů, které má Heureka v nejoblíbenějších, nemáme vůbec v datech. Chybějící produkty se objevily i v ostatních kategoriích, ale jejich množství bylo oproti PC sestavám zanedbatelné.

4.5.3 Závěr porovnání s konkurencí

Jak ukazují výsledky, úspěšnost při výběru TOP produktů je napříč všemi kategoriemi velmi rozdílná. Hlavním faktorem, proč se top produkty z Heureky neobjevují v TOP produktech vybraných mojí analýzou, je nízký počet celkových prodejů daného produktu. Tyto prodeje jsou často pouze v desítkách prodaných kusů, což je z hlediska délky sledovaného období nevýznamný objem. Dalším faktorem je pokles prodejů za poslední měsíce, kdy se prodeje silně prodávaných produktů propadly často až na nulu.

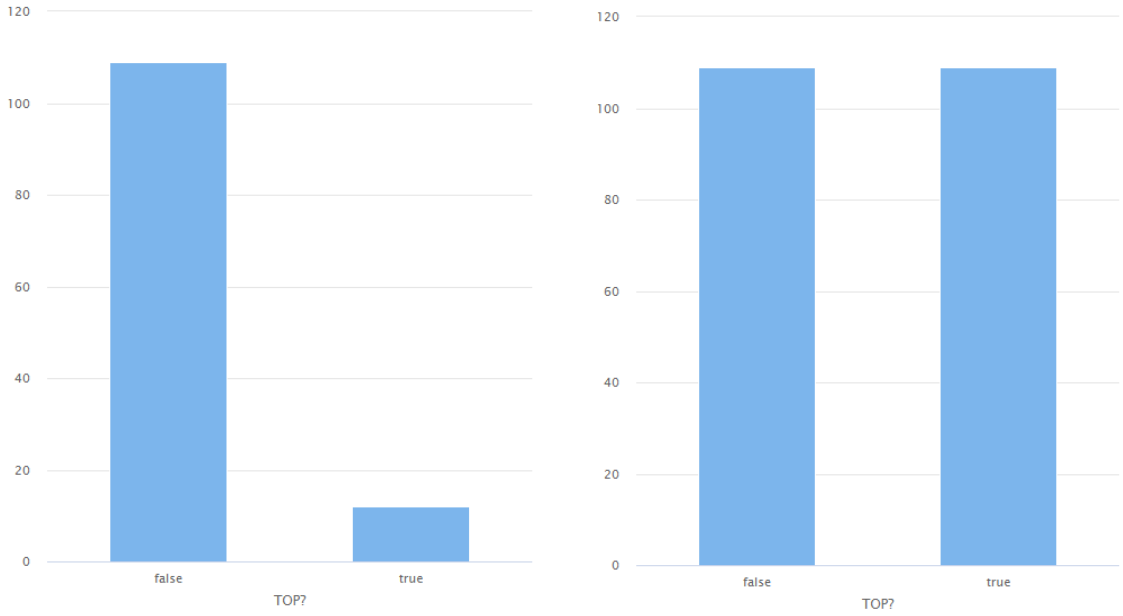
Porovnání s výsledky na portálu Heureka jednoznačně neodráží kvalitu výsledků mé analýzy primárně z důvodu rozdílných vstupních dat. Data o prodejích, kterými disponuji, jsou pouze částí trhu. Zároveň cílem analýzy není poskytnout stejný výběr top produktů, jako má Heureka, ale přizpůsobit tento výběr oblasti, která je reprezentována dodanými daty. Tohoto cíle se mi podařilo dosáhnout, jak značí průnik mnou vybraných produktů a produktů z Heureky, a zároveň neshody mezi těmito produkty jsou zapříčiněny obsahem vstupních dat, nikoliv procesem výběru.

4.6 Využití technik data miningu

Tato podkapitola diskutuje použití různých technik pro výběr TOP produktů pomocí data miningu. Konkrétně rozhodovací strom, Naive Bayes a neuronová síť. Pro každou z těchto technik je popsané její použití, výsledky a diskutovaná vhodnost jejího dalšího využití.

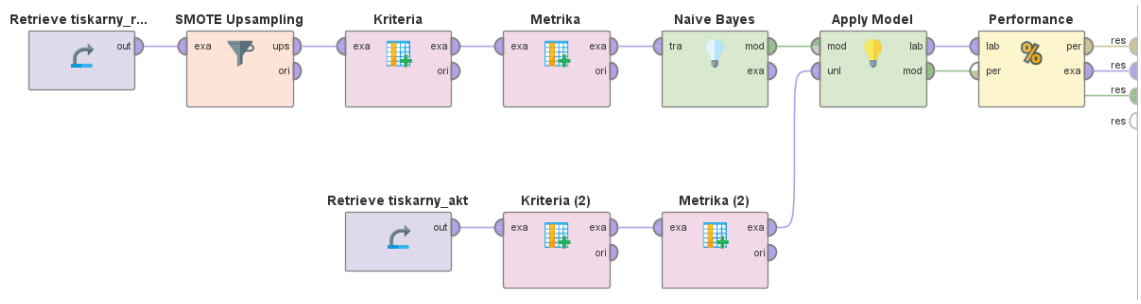
4.6.1 Příprava pro použití data miningových technik

Pro využití výše zmiňovaných technik bylo nutné data upravit do vhodné podoby. V každé kategorii jsem pracovala s daty rozdělenými do cenových hladin (viz podkapitola 4.3.2). Nejdříve bylo potřeba data anotovat. Ke každému produktu jsem poznačila, zda se nacházejí v TOP produktech na portálu Heureka. Vzhledem k nízkému počtu top produktů v porovnání k celkovému počtu produktů docházelo k výraznému nepoměru zastoupení obou tříd (je v TOP / není v TOP). Proto jsem se rozhodla použít nadvzorkování minoritní třídy pomocí operátoru *SMOTE Upsampling* z rozšíření RapidMineru Operator Toolbox. Nadvzorkování pomocí uměle vytvořených dat není vždy ideálním řešením, ale v tomto případě je nejvhodnějším přístupem, jak dosáhnout vyváženosti tříd a zároveň zachovat rozumnou velikost datové sady. Rozdíl mezi vyvážeností tříd před a po nadvzorkování dat je znázorněn na Obrázku 19. Takto upravená data byla použita pro trénink modelů a pro jejich vyhodnocení byla použita aktualizovaná neupravená data. Zároveň byl mezi trénovacími a testovacími daty rozdíl tří měsíců, trénovací data byla za období prosinec 2019–listopad 2021 a testovací data



Obrázek 19: Rozdíl zastoupení produktů před (vlevo) a po (vpravo) nadvzorkování. (Zdroj: Vlastní zpracování)

březen 2020–únor 2022. Celé schéma zapojení procesu pro trénink a testování technik je ukázáno na Obrázku 20.



Obrázek 20: Proces pro trénink a testování technik data miningu. (Zdroj: Vlastní zpracování)

4.6.2 Rozhodovací strom

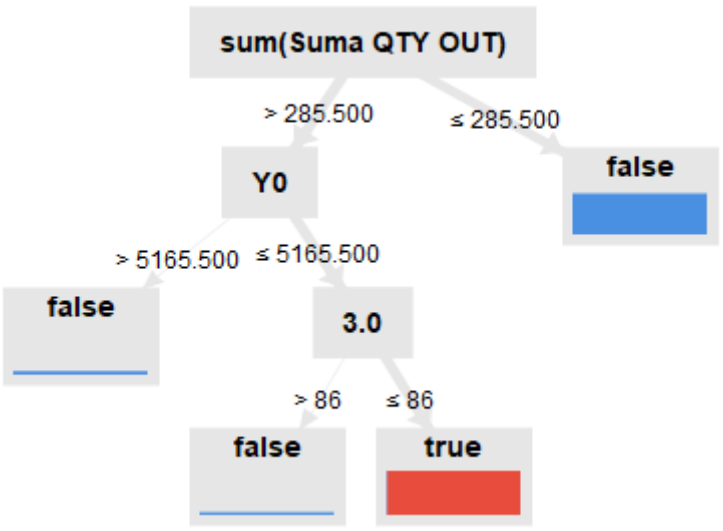
Rozhodovací strom je v RapidMineru dostupný jako operátor *Decision Tree*. Operátor nabízí možnost výběru kritéria pro výběr nejdůležitějších atributů při sestavování stromu.

Prvním krokem bylo porovnat nabízená kritéria pro výběr uzlů při sestavování rozhodovacího stromu a posoudit vliv dříve navržené metriky (viz podkapitola 4.4.2) na úspěšnost predikce. Toto porovnání je znázorněno v Tabulce 7. Všechny ostatní parametry jsem ponechala ve výchozích hodnotách.

Tabulka 7: Porovnání vlivu kritéria a metriky na přesnost rozhodovacího stromu. (Zdroj: Vlastní zpracování)

Kritérium	*Kategorie*	*Přesnost (%)*	*Přesnost bez metriky*
Gain_ratio	Tiskárny	61,54	61,54
	Notebooky	7,69	7,69
	PC sestavy	20	20
	Monitory	0	0
Information_gain	Tiskárny	61,54	61,54
	Notebooky	7,69	7,69
	PC sestavy	20	20
	Monitory	0	0
Gini_index	Tiskárny	61,54	61,54
	Notebooky	7,69	7,69
	PC sestavy	20	20
	Monitory	0	0
Accuracy	Tiskárny	61,54	61,54
	Notebooky	7,69	7,69
	PC sestavy	0	0
	Monitory	0	0

Z tabulky je zřejmé, že změna kritéria ani přítomnost mnou navržených metrik nemají vliv na výsledky predikce. Samotná přesnost predikce je velmi nízká. Jedním z hlavních důvodů bude pravděpodobně nevhodná struktura rozhodovacího stromu. Pro ilustraci je struktura rozhodovacího stromu pro kategorii tiskáren znázorněna na Obrázku 21. Z této struktury je zřejmé, že se strom rozhoduje na základě malého množství atributů. Zároveň zvolené atributy neodpovídají dříve odhaleným vzorům v prodeji TOP produktů jako je například počet prodejů za poslední tři měsíce.



Obrázek 21: Struktura rozhodovacího stromu pro kategorii tiskáren. (Zdroj: Vlastní zpracování)

4.6.3 Naive Bayes

Naivní Bayesův klasifikátor je v RapidMineru dostupný pod operátorem *Naive Bayes*. Jako vstup jsem opět zkoušela jak data s mnou vytvořenou metrikou, tak data bez metriky.

Z Tabulky 8 lze vyčíst, že naivní Bayesův klasifikátor dosahuje lepších výsledků než rozhodovací strom a zároveň přidání metriky opět nemělo na klasifikátor žádný vliv.

Tabulka 8: Vliv metriky na naivní Bayesův klasifikátor. (Zdroj: Vlastní zpracování)

Kategorie	*S metrikou*	*Bez metriky*
Tiskárny	92,31	92,31
Notebooky	15,38	15,38
PC sestavy	60	60
Monitory	0	0

U tiskáren se podařilo dosáhnout velmi vysoké úspěšnosti při predikci top produktů, kdežto u kategorie monitorů se klasifikátoru nepodařilo správně predikovat ani jeden monitor.

4.6.4 Neuronová síť

Pomocí operátoru *Neural Net* je v RapidMineru možné využívat neuronové sítě. Při jeho použití jsem nechala parametry ve výchozích hodnotách a znovu porovnávala vstupy bez metriky a s jejím přidáním.

V Tabulce 9 jsou znázorněny výsledky porovnání. Neuronová síť jako jediný z použitých modelů dokázala využít přítomnost metriky pro zvýšení přesnosti. Toto zlepšení bylo na úkor kategorie PC sestav, kdy neuronová síť nebyla schopna správně klasifikovat žádný počítač. Zároveň jako jediný model dokázala neuronová síť správně klasifikovat aspoň jeden monitor.

Celkově neuronová síť dosahuje srovnatelných výsledků jako naivní Bayesův klasifikátor.

Tabulka 9: Vliv metriky na neuronovou síť. (Zdroj: Vlastní zpracování)

Kategorie	*S metrikou*	*Bez metriky*
Tiskárny	84,62	76,92
Notebooky	38,46	23,08
PC sestavy	0	20
Monitory	33,33	33,33

4.6.5 Závěr použití technik data miningu

Žádná z použitých technik nedosahovala uspokojivých výsledků ve všech kategoriích. Nejhůře predikoval rozhodovací strom. Naivní Bayesův klasifikátor a neuronová síť dosahovali podobných výsledků.

Porovnáním přesnosti bez metriky a s jejím přidáním stejně jako prozkoumáním struktury modelu rozhodovacího stromu a naivního Bayesovského klasifikátoru jsem zjistila, že vytvořené modely přidanou metriku nepoužívají. Na základě tohoto zjištění jsem upravila výběr atributů pro trénink modelu tak, aby obsahoval pouze ceny, kritéria, metriku a celkovou sumu prodejů. Tato úprava výrazně zlepšila predikci top produktů (*true positive*), ale zároveň rapidně stoupla predikce falešně pozitivních výsledků (*false positive*). Tento přístup dodával nejhorší výsledky, protože vytvářel nejvíce chyb v obou třídách (TOP i obyčejných produktech).

Z dosažených výsledků je zřejmé, že tyto data miningové techniky v mnou použitých případech nejsou pro výběr TOP produktů vhodné. Pro další vyhodnocení, zda jsou tyto techniky vhodné pro tento typ analýzy, bude potřeba hlubší zkoumání a experimenty, což už je mimo rozsah této práce.

Závěr

Při analýze prodejů produktů jsem zjistila, že nejdůležitějšími faktory při výběru TOP produktů jsou za prvé stabilní prodeje za celé analyzované období a za druhé prodeje v posledním čtvrtletí. Tyto dva faktory jsem zohlednila při tvorbě metriky pro výběr TOP produktů. Své výsledky z mnou navrženého řešení jsem porovnávala s produkty vybranými konkurencí. Toto srovnání potvrdilo správnou funkčnost mého řešení. K případným rozdílům ve vybraných produktech docházelo zejména kvůli obsahu poskytnutých dat. Hlavním důvodem, proč se TOP produkty vybrané konkurencí neshodují s TOP produkty vybranými mou analýzou, je nízký počet celkových prodejů daného produktu. Tyto prodeje jsou často pouze v desítkách prodaných kusů, což je z hlediska délky sledovaného období nevýznamný objem.

Porovnání mého řešení a data miningových technik pro výběr TOP produktů potvrdilo vhodnost mnou navrženého řešení. Data miningové techniky se v použitém formátu ukázaly jako nevhodné pro tento typ analýzy a zároveň by jejich úspěšné nasazení vyžadovalo hlubší přípravu dat a úpravu modelů.

Prvním krokem pro vylepšení procesu výběru TOP produktů je konzultace vzniklých řešení a výsledků s experty dané značky. Na základě toho může být řešení dále upraveno nebo rozšířeno pro dosahovaní kvalitnějších výsledků. Dalším krokem může být detailnější prozkoumání a použití data miningových technik pro proces výběru TOP produktů.

Seznam použité literatury

1. **Chen, Hsinchun, Chiang, Roger H. L. and Storey, Veda C.** Business intelligence and analytics: from big data to big impact. *MIS Quarterly.* 2012, pp. 1165-1188.

2. **Novotný, Ota, Pour, Jan a Slánský, David.** *Business Intelligence: Jak využít bohatství ve vašich datech.* Praha : Grada Publishing, 2005. 80-247-1094-3.

3. **TechDifferences.** Difference Between OLTP and OLAP. *techdifferences.com.* [Online] [Dátum: 4. Prosinec 2021.] https://techdifferences.com/difference-between-oltp-and-olap.html.

4. **Giceva, Jana a Sadoghi, Mohammad.** Hybrid OLTP and OLAP. *Encyclopedia of Big Data Technologies.* 19. Únor 2018.

5. **stitchdata.com.** OLTP and OLAP: a practical comparison. *stitchdata.com.* [Online] 2021. [Dátum: 9. Leden 2022.] https://www.stitchdata.com/resources/oltp-vs-olap/.

6. **Gupta, Manoj Kumar a Pravin, Chandra.** A comprehensive survey of data mining. *International Journal of Information Technology.* 6. Únor 2020, s. 1243-1257.

7. **Nguyen, Giang, et al.** Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. *Artificial Intelligence Review.* Leden 19, 2019, pp. 77-124.

8. **Alasadi, Suad A. a Bhaya, Wesam S.** Review of data preprocessing techniques in data mining. *Journal of Engineering and Applied Sciences.* 2017, s. 4102-4107.

9. **StatSoft.** Úvod do data miningu. *StatSoft.* [Online] 26. Únor 2014. [Dátum: 20. Únor 2022.] http://www.statsoft.cz/file1/PDF/newsletter/2014_02_26_StatSoft_Uvod_do_data_miningu.pdf.

10. **Lutkevich, Ben.** DEFINITION association rules. *searchbusinessanalytics.techtarget.com.* [Online] Září 2020. [Dátum: 10. Leden 2022.] https://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining.

11. **Raval, Kalyani M.** Data mining techniques. *International Journal of Advanced Research in Computer Science and Software Engineering.* 2012, Zv. 2, 10, s. 439-442.

12. **Oracle.** Oracle® Data Mining. [Online] 2010. [Dátum: 12. Leden 2022.] https://docs.oracle.com/cd/E18283_01/datamine.112/e16808.pdf.

13. **Yang, Xin-She.** 2 - Mathematical foundations. *Introduction to Algorithms for Data Mining and Machine Learning.* : Academic Press, 2019, s. 19-43.

14. **Matoušek, Kamil.** Pravděpodobnostní reprezentace neurčitosti, bayesovské sítě. *cw.fel.cvut.cz.* [Online] [Dátum: 4. Březen 2022.] https://cw.fel.cvut.cz/old/_media/courses/a7b33sui/bayesovske_site.pdf.

15. **Fiřtová, Lenka.** Korelace – co to je korelace a co znamená korelační koeficient. *ExcelTown.com.* [Online] 2020. [Dátum: 4. Březen 2022.] https://exceltown.com/navody/pokrocila-analyza-regrese-korelace/korelace-co-to-vlastne-je/.

16. **wikisofia.cz.** Korelační a regresní analýza. *wikisofia.cz.* [Online] 2013. [Dátum: 4. Březen 2022.] https://wikisofia.cz/wiki/Korela%C4%8Dn%C3%AD_a_regresn%C3%AD_anal%C3%BDza. 2336-5897.

17. **Kelbel, Jan a Šilhán, David.** Shluková analýza. [Online] Květen 2002. [Dátum: 4. Březen 2022.] http://cmp.felk.cvut.cz/cmp/courses/recognition/zapis_prednasky/zapis_02/13/shlukovani.pdf.

18. **Klecka, William.** *Discriminant Analysis.* Thousand Oaks, California : Sage Publications, 1980. 978-0-8039-1491-9.

19. **Bose, Indranil a Radha, K. Mahapatra.** Business data mining—a machine learning perspective. *Information & management.* 2001, Zv. 39, 3, s. 211-225.

20. **Singh, Yashpal a Chauhan, Alok Singh.** NEURAL NETWORKS IN DATA MINING. *Journal of Theoretical & Applied Information Technology.* 2009, Zv. 5, 1, s. 37-42.

21. **George Harrison, John.** *Enhancements to the data mining process.* Ann Arbor : Stanford University, 1997. ISBN 978-0-591-31808-1.

22. **Zendulka, Jaroslav, a iní.** *Získávání znalostí z databází. Studijní opora.* Brno : FIT VUT, 2009.

23. **Chire.** Cluster analysis. *wikipedia.* [Online] 22. Říjen 2011. [Dátum: 27. Prosinec 2021.] https://en.wikipedia.org/wiki/Cluster_analysis#/media/File:SLINK-Gaussian-data.svg.

Seznam obrázků

Seznam tabulek