Česká zemědělská univerzita v Praze

Provozně ekonomická fakulta

Katedra Statistiky

Diplomová práce

Big data a prediktivní modelování v oblasti skladového hospodářství vybrané firmy

Jakub Kukla

ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

Provozně ekonomická fakulta

ZADÁNÍ DIPLOMOVÉ PRÁCE

Bc. Jakub Kukla

Provoz a ekonomika

Název práce

Big Data a prediktivní modelování v oblasti skladového hospodářství vybrané firmy

Název anglicky

Possibilities of Big Data and predictive analysis in the warehouse management of a selected company

---

Cíle práce

Diplomová práce se zabývá hledáním a analýzou faktorů, které ovlivňují četnost skladových pohybů artiklů (výrobků) společnosti EYELEVEL s.r.o. Hlavním cílem této práce je předvídání chování jednotlivých výrobků z hlediska naskladňování a vyskladňování tak, aby bylo dopředu možné určit, jak moc bude výrobek pohyblivý a na základě toho vybrat vhodnou lokaci k jeho uskladnění. Tento krok společnosti umožní vybrat pro uskladnění takové skladové pozice, které budou efektivní z hlediska výdeje do oblasti montáže a balení. Toto opatření společnosti zkrátí časové prostoje při výdeji materiálu a bude tak zefektivněn celý proces vyskladňování. Ke splnění hlavního cíle práce bude třeba analyzovat databázi současných skladových pohybů společně s databází jednotlivých artiklů a jejich atributů, což povede k odhalení faktorů, které mají na četnost skladových pohybů vliv.

Metodika

Těžiště vlastní práce bude postaveno na využití statistických postupů data mining. Dle povahy dat se počítá se zapojením klasifikačních metod (např. shluková analýza), dále pak rozhodovacích stromů a především regresních modelů (klasických i logistických).

Doporučený rozsah práce

60 – 80 stran

Klíčová slova

Skladové hospodářství, výrobek, skladová pozice, prediktivní model

---

Doporučené zdroje informací

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

AGRESTI, A. *Categorical data analysis.* Hoboken: John Wiley & Sons, 2013. ISBN 978-0-470-46363-5.

JAIN V. K. Big Data and Hadoop. Indie: Khanna Publishing, 2017. ISBN 9789382609131

RAVINDRA, S. Understanding the relationship between IoT and Big Data. Jaxenter.com [online]. USA: Jaxenter, 2017. Dostupné z: https://jaxenter.com/relationship-between-iot-big-data-138220.html

TUFFÉRY, S. Data Mining and Statistics for Decision Making. UK, West Sussex: Wiley, 2011. ISBN 978-0-470-68829-8.

Předběžný termín obhajoby

2018/19 LS – PEF

Vedoucí práce

Ing. Tomáš Hlavsa, Ph.D.

Garantující pracoviště

Katedra statistiky

Elektronicky schváleno dne 24. 1. 2019

prof. Ing. Libuše Svatošová, CSc.

Vedoucí katedry

Elektronicky schváleno dne 5. 2. 2019

Ing. Martin Pelikán, Ph.D.

Děkan

V Praze dne 14. 03. 2019

Čestné prohlášení

Prohlašuji, že svou diplomovou práci " Big data a prediktivní modelování v oblasti skladového hospodářství vybrané firmy " jsem vypracoval samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou citovány v práci a uvedeny v seznamu použitých zdrojů na konci práce. Jako autor uvedené diplomové práce dále prohlašuji, že jsem v souvislosti s jejím vytvořením neporušil autorská práva třetích osob.

V Praze dne 31.3.2019 _____

Poděkování

Rád bych touto cestou poděkoval Ing. Tomáši Hlavsovi, Ph.D. za jeho odborný dohled, konzultace, rady, doporučení a ochotu při zpracování této diplomové práce. Dále bych chtěl poděkovat společnosti EYELEVEL s.r.o., která mi prostřednictvím vedoucích jednotlivých oddělení poskytla veškeré informace, data a podklady potřebné k realizaci praktické části práce.

Big Data a prediktivní modelování v oblasti skladového hospodářství vybrané firmy

Abstrakt

Tato diplomová práce je zaměřena na demonstraci možností využití data miningu a prediktivního modelování ve skladovém hospodářství vybrané společnosti. V teoretické části této práce jsou popsány principy a způsoby využití velkých objemů dat v obchodním prostředí. Praktická část práce se potom věnuje vlastnímu výzkumu, který zkoumá vybranou databázi produktů společnosti, jejich atributů a četnosti skladových pohybů. S využitím této databáze byly zpracovány tři prediktivní modely, jejichž cílem je naučení souvislostí mezi atributy jednotlivých produktů a výsledným počtem jejich skladových pohybů. Díky tomu bude u nových produktů společnosti možné určit, zda budou rychloobrátkové, či nikoliv a na základě toho vybrat pro jejich uskladnění vhodnou skladovou lokaci, čímž dojde ke zkrácení dráhy toku materiálu, ušetření času pro vyskladnění a ušetření nákladů. Po optimalizaci modelů a vzájemném porovnání jejich výkonů byl vybrán model rozhodovacího stromu, který dokáže s téměř devadesátiprocentní jistotou určit výslednou obrátkovost produktů, adekvátní kategorii pro jejich zařazení a nejvýznamnější faktory, které tento proces ovlivňují. Rozhodovací strom byl následně v praktickém příkladu hypoteticky aplikován do struktur informačního systému společnosti, přičemž byly demonstrovány přínosy jeho možného využívání. V rámci tohoto příkladu bylo zjištěno, že pokud bude využita skladová lokace pro rychloobrátkové zboží, do které budou produkty zařazovány na základě výstupu prediktivního modelu, bude možné ušetřit v rámci jedné zakázky přibližně 23 minut, což v kontextu ročního objemu zakázek umožní společnosti ušetřit náklady ve výši přibližně 240 000 Kč.

Klíčová slova: big data, prediktivní analýza, skladové hospodářství, business intelligence, zlepšování procesů

Possibilities of Big Data and predictive analysis in the warehouse management of a selected company

Abstract

This diploma thesis is focused on application of data mining and predictive analytics in the warehouse management of a company. The theoretical part of this thesis describes the basic principles of using big data in business. Second part of the diploma thesis is focusing on the experiment which analyses selected database of the company products, their attributes and stock transaction quantities. The database is then used for the construction of three predictive models including logistic regression, decision tree and neural network. The main goal of those models is to learn the predictive algorithm based on the products attributes and classify newly created products as those with high or low quantity of stock transactions through the year. Based on this prediction, the most effective container can be selected for the product storage concerning the distance from the assembly or packaging area. Using this strategy, the material flow distance, time required for the product handling and overall costs for warehouse operations will be reduced. The model of decision tree was chosen after all model parameter optimization and accuracy comparison. This model also analysed the most important factors influencing the stock transaction quantities. The decision tree was also hypothetically applied to the ERP system of the company to demonstrate the benefits of its usage. This model showed that if the company creates the stock location for fast – tracked products (which will be picked by the predictive model) it will be possible to spare approximately 23 minutes of operation time for one customer order. This spare will lead to the cost reduction of roughly 240 000 CZK per year.

Keywords: big data, predictive analytics, warehouse management, business intelligence, process improvement

Obsah

1 Úvod 12

2 Cíl práce a metodika 14

2.1 Cíl práce 14

2.2 Metodika 14

3 Teoretická východiska 16

3.1 Big data – pojem a realita 16

3.1.1 3 V a jejich význam 16

3.1.2 Původ dat a internet věcí 18

3.1.3 Historie vývoje 19

3.1.4 Využití big data 20

3.2 Business intelligence a zlepšování procesů 24

3.2.1 Business intelligence 24

3.2.2 Zlepšování procesů 26

3.3 Prediktivní analýza 27

3.3.1 Úvod do prediktivní analýzy 27

3.3.2 Komparace prediktivní analýzy a příbuzných statistických disciplín 28

3.3.3 Překážky pro aplikaci prediktivní analýzy v podniku 31

3.3.4 Procesní metodika prediktivní analýzy – CRISP-DM 33

3.4 Warehouse Management (skladové hospodářství) 46

4 Vlastní práce 49

4.1 Popis společnosti 49

4.2 Porozumění businessové stránce 50

4.3 Porozumění datům 53

4.3.1 MITTRA data 53

4.3.2 Item Basics 54

4.3.3 Metadata 54

4.4 Příprava dat 54

4.4.1 Agregování databází 55

4.4.2 Základní informace o souboru 56

4.4.3 Testování a výběr proměnných 63

4.5 Modelování 67

4.5.1 Logistická regrese 67

4.5.2 Rozhodovací Strom 72

4.5.3 Neuronové sítě 79

4.6 Srovnání a vyhodnocení výsledků jednotlivých modelů 82

4.7 Aplikace rozhodovacího stromu 83
4.7.1 Action steps 83
4.7.2 Časový plán zavedení modelu do provozu 84
4.7.3 Komunikace řešení koncovým uživatelům 85
4.7.4 Měření výkonu modelu a opatření do budoucna 86
4.7.5 Příklad využití modelu v praxi a vyčíslení možných úspor 86

5 Výsledky a diskuse 91

6 Závěr 94

7 Seznam použitých zdrojů 96

8 Přílohy 99

Seznam obrázků

Obrázek 1 3 V Big Dat. 18
Obrázek 2 Schéma systému BI.. 26
Obrázek 3 Schéma procesu prediktivní analýzy. 27
Obrázek 4 Fáze referenčního modelu metodiky CRISP-DM. 34
Obrázek 5 Porovnání lineární a logistické regrese. 41
Obrázek 6 Příklad struktury rozhodovacího stromu. 42
Obrázek 7 Neuron transformující vstupní hodnoty. 43
Obrázek 8 Jednotlivé vrstvy modelu neuronové sítě. 44
Obrázek 9 Typické funkce skladování a související toky materiálu. 47
Obrázek 10 Organizační struktura pražské pobočky společnosti EYELEVEL s.r.o. 49
Obrázek 11 Rozdělení souboru dle klientů. 56
Obrázek 12 Rozdělení souboru dle klientů v porovnání s četností skladových pohybů. 57
Obrázek 13 Struktura jednotlivých klientů a pod ně spadajících projektů. 58
Obrázek 14 Kategorizace souboru dle jednotlivých projektů. 59
Obrázek 15 Rozdělení produktů dle typu a vlastnictví. 60
Obrázek 16 Histogram rozdělení Manufactured itemů dle počtu komponentů. 61
Obrázek 17 Scatter diagram rozdělení nákupních cen produktů dle jejich typu. 62
Obrázek 18 Wordcloudová kategorizace produktů dle dodavatele. 62
Obrázek 19 Schéma procesu generace korelační matice.. 64
Obrázek 20 Ukazatel vhodnosti atributu pro modelování. 65
Obrázek 21 Výsledný proces logistické regrese. 67
Obrázek 22 Výsledková tabulka pravděpodobnosti výsledků modelu logistické regrese. 69
Obrázek 23 Klasifikační matice modelu logistické regrese.. 70
Obrázek 24 Upravený proces logistické regrese. 72
Obrázek 25 Klasifikační matice upravené logistické regrese. 72
Obrázek 26 Výsledný proces rozhodovacího stromu. 74
Obrázek 27 Výřez modelu rozhodovacího stromu. 75
Obrázek 28 Klasifikační matice prvního rozhodovacího stromu. 75
Obrázek 29 Klasifikační matice prvního rozhodovacího stromu. 76
Obrázek 30 Proces optimalizace parametrů rozhodovacího stromu. 77
Obrázek 31 Ideální kombinace hodnot parametrů rozhodovacího stromu. 77

Obrázek 32 Výřez výstupu optimalizovaného rozhodovacího stromu. 78
Obrázek 33 Klasifikační matice optimalizovaného rozhodovacího stromu. 79
Obrázek 34 Proces tvorby umělé neuronové sítě. 80
Obrázek 35 Subproces operátoru Validation. 80
Obrázek 36 Klasifikační matice neuronové sítě (200 tréninkových cyklů, 1 skrytá vrstva). 81
Obrázek 37 Klasifikační matice neuronové sítě (500 tréninkových cyklů, 2 skryté vrstvy). 81
Obrázek 38 Průtok materiálu skladem dle standardního postupu. 87
Obrázek 39 Průtok materiálu skladem s využitím lokace Fast Pick. 89

Seznam Tabulek

Tabulka 1 Porovnání datových základen privátních společností v čase. 19
Tabulka 2 Přehled korelovaných proměnných a jejich korelační koeficientů. 64
Tabulka 3 Vhodnost jednotlivých atributů pro modelování. 66
Tabulka 4 Seznam vyřazených proměnných a důvod jejich vyřazení. 66
Tabulka 5 Porovnání celkového výkonu jednotlivých modelů. 82
Tabulka 6 Časová osa aplikace prediktivního modelu. 85
Tabulka 7 Přehled manipulačních časů a nákladů standardního postupu. 88
Tabulka 8 Přehled manipulačních časů a nákladů s využitím Fast Pick lokace 90
Tabulka 9 Rozdíl v manipulačních vzdálenostech, dobách a nákladech standardního a navrhovaného postupu. 90

Seznam Grafů

Graf 1 Časová osa vývoje datových základen privátních společností v čase. 20

1 Úvod

Doba, ve které žijeme je plná rychlých změn, kterým se musíme neustále přizpůsobovat. Toto tvrzení, s jehož důsledky se setkáváme v rámci studia, zaměstnání i běžného života platí v rámci obchodního prostředí několikanásobně silněji. Vezmeme-li v potaz kupní sílu vyspělých trhů, technologické možnosti, kterými v jednadvacátém století disponujeme a neustálý růst konkurence z pozice mezinárodních korporátů, či začínajících start-upů, je třeba k udržení tempa neustále toto prostředí monitorovat a adekvátně upravovat strategii řízení společností téměř na každodenní bázi. K tomu, aby byly takovéto flexibilní úpravy možné, je nejprve nutné vnímat a pochopit veškeré relevantní vlivy vnějšího prostředí, na jejichž základě je možné podniknout konkrétní kroky vedoucí k úpravám vnitřního prostředí společnosti. K poznání a zaznamenání veškerých těchto vlivů a skutečností slouží data, jejichž význam je v posledním desetiletí neustále vyzdvihován a hovoří se o tom, že jsou často vedle půdy, práce a kapitálu považována za jeden ze základních výrobních faktorů. V praxi lze prostřednictvím dat zaznamenat veškeré události a skutečnosti reálného světa od těch základních, jakými jsou například počty obyvatel hlavních měst, až po ty zcela irelevantní jako například průměrný odstín zelené barvy jablek pěstovaných v jihovýchodní Austrálii. Jednou z nejzajímavějších a nejrozšířenějších metod získávání cenných dat je v současné době zaznamenávání údajů pomocí informačních technologií. Do této kategorie spadají nejrůznější způsoby zaznamenávání dat. Ať už se jedná o sbírání dat prostřednictvím chytrých (smart) přístrojů připojených k internetové síti (tzv. internet věcí), zaznamenávání dat z používání webových stránek a sociálních sítí, či ukládání údajů do informačních systémů společností, hlavním znakem tohoto fenoménu je objem a množství neustále generovaných informací. Tyto velké objemy dat, které jsou vytvářeny závratnou, neustále se zvyšující rychlostí bývají označovány jako big data a pro mnoho společností se jedná o klíč k úspěchu, porozumění své oblasti působení, či nekonečný zdroj podnětů pro zlepšování pozice v rámci konkurence. Big data jsou ve své podstatě záznamy o fungování a průběhu určitých procesů, či skutečností na jejichž základě lze tyto události efektivně monitorovat, analyzovat proběhlé situace a hledat souvislosti mezi procesními vstupy a jejich výstupem. Z tohoto důvodu jsou big data využívána v nejrůznějších oblastech od lékařství, kde je možné na základě záznamů o průběhu nemoci pacienta určit například jeho šance na uzdravení až po marketingový výzkum, kde lze s využitím údajů o

charakteristikách zákazníka (věk, pohlaví, záliby atd.) s určitou pravděpodobností předpovědět, zda daný produkt koupí, či nikoliv. Díky neustálému zdokonalování informačních systémů společností a zaznamenávání údajů o podnikových procesech tato metoda v posledních letech prostupuje téměř do všech odvětví. Obecně lze říci, že s vyšší, či nižší přidanou hodnotou lze big data a jejich analýzu využít ve své podstatě všude, kde existuje dostatečný přísun dat, na jejichž základě lze situace modelovat a naučit algoritmus předpovídat předem neznámý výstup s využitím známých atributů. Ačkoliv se jedná o silný nástroj pro pochopení obchodního prostředí a interních procesů společnosti, nejedná se o univerzální řešení jakéhokoliv problému. Prediktivní modelování a big data poskytují pouze rámec a podklady pro zlepšení efektivity a celkové konkurenceschopnosti firmy, které musí být podloženo především intenzivním zájmem zaměstnanců a silnou podporou managementu.

2 Cíl práce a metodika

2.1 Cíl práce

Hlavním cílem této práce je předvídání chování jednotlivých výrobků společnosti EYELEVEL s.r.o. z hlediska naskladňování a vyskladňování tak, aby bylo dopředu možné určit, jak moc bude výrobek pohyblivý a na základě toho vybrat vhodnou lokaci k jeho uskladnění. Tento krok společnosti umožní vybrat pro uskladnění takové skladové pozice, které budou efektivní z hlediska výdeje do oblasti montáže a balení. Toto opatření společnosti zkrátí časové prostoje při výdeji materiálu a bude tak zefektivněn celý proces vyskladňování. Ke splnění hlavního cíle práce bude třeba analyzovat databázi současných skladových pohybů společně s databází jednotlivých artiklů a jejich atributů, což povede k odhalení faktorů, které mají na četnost skladových pohybů vliv.

K dosažení hlavního cíle této práce jsou stanoveny tři cíle dílčí. Prvním z dílčích cílů je agregace dat z různých databází a jejich příprava do podoby vhodné k modelování. Druhým dílčím cílem je sestavení tří prediktivních modelů, které budou co nejpřesněji předpovídat, zda bude zboží rychloobrátkové, či nikoliv. Třetím dílčím cílem je potom vypracování plánu zavedení vybraného modelu do provozu společnosti a zpracování krátké případová studie, která vyčíslí možné roční úspory vyplývající z aplikace modelu.

2.2 Metodika

Teoretická část této práce je zpracována na základě studia literatury. Vlastní práce je potom založena na analýze databází z informačního systému společnosti a jejich následné agregaci do základního souboru čítajícího 8909 záznamů o produktech společnosti a jejich 29 atributech. V rámci vlastní práce je postupováno dle metodiky Cross-Industry Standard Process for Data Mining (CRISP-DM), která je blíže popsaná v teoretických východiscích. Postup zpracování projektu lze shrnout následujícími body:

- Porozumění obchodnímu případu, problematice a specifikám odvětví.
- Porozumění datům, se kterými bude dále pracováno, jejich popis a vizualizace.
- Příprava dat do podoby, ve které je možné je zpracovat.
- Vypracování tří prediktivních modelů a jejich optimalizace.

- Měření přesnosti modelů a jejich vzájemné porovnání.
- Výběr nejvhodnějšího modelu pro daný obchodní případ.

Model, který byl v průběhu vlastní práce označený jako nejvhodnější je následně hypoteticky aplikován do provozu společnosti, přičemž je vypracována vzorová případová studie popisující proces skladové manipulace v rámci nově vzniklých skladových lokací pro rychloobrátkové zboží. Na základě této případové studie jsou následně vyčísleny časové i finanční úspory, které bude možné tímto řešením dosáhnout.

3 Teoretická východiska

3.1 Big data – pojem a realita

3.1.1 3 V a jejich význam

Samotný český překlad slovního spojení „big data“, tedy „velká data“ naznačuje, že se pod tímto pojmem skrývá označení rozsáhlého souboru digitálních dat, popisující určité skutečnosti nebo jevy reálného světa. První otázka, kterou je nutno si klást je to, jak rozsáhlý soubor dat musí být, aby bylo možné ho označovat jako big data.

Formální, obecně uznávaná jednotná definice pojmu big data nebyla zatím stanovena, nicméně nejčastěji se setkáváme s popisem, zakládajícím se na takzvaných třech V. Dle tohoto popisu jsou big data taková data, jejichž objem (anglicky volume), rychlost růstu (velocity) a různorodost (variety) neumožňují zpracování pomocí doposud známých a ověřených technologií v rozumném čase (Holubová, Kosek, Minařík, Novák 2015).

První ze zmiňovaných "V", tedy volume, či v českém překladu objem je nejzákladnější charakteristikou popisovaných dat. V souvislosti s pojmem big data lze tento objem chápat jako rozsah záznamů v určitém souboru dat. Tento rozsah nabývá často těžko uchopitelné rozměry. Je-li zkoumána například datová základna sociální sítě Facebook, kde každý uživatel (v současné chvíli tuto síť používá okolo 2,2 miliardy aktivních uživatelů) sdílí a uloží do databáze během svého působení několik fotografií (v současné chvíli databáze obsahuje přibližně 250 miliard fotografií), lze hovořit o opravdu nepředstavitelně velkém objemu dat (Gewirtz 2018).

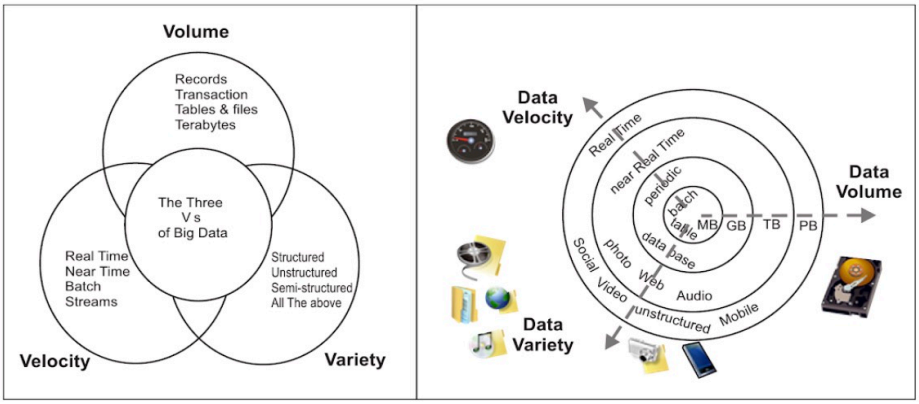
Databáze sociálních sítí jsou samy o sobě extrémní případ. Lépe uchopitelná pro představu je například databáze aplikace Todoist, která uživatelům pomáhá tvořit jednoduché to-do listy a propojovat je s ostatními aplikacemi, jako například s emailovým klientem, či kalendářem. Tato aplikace má v současné chvíli zhruba 10 milionů aktivních uživatelů, přičemž každý má své seznamy úkolů, které se do databáze propisují. Kromě takto rozsáhlých databází existuje samozřejmě nespočet databází, obsahující záznamy o

podnikových procesech, lékařství, či jiné oblasti, které svým objemem spadají do kategorie Big Data (Gewirtz 2018).

Za druhým "V" se skrývá pojem velocity, česky překládaný jako rychlost růstu. Tento pojem lze chápat jako rychlost, s jakou záznamy v databázi přibývají. Výborným příkladem pro demonstraci nepředstavitelné rychlosti růstu množství dat je opět sociální síť Facebook a její databáze fotografií. Dle statistik z prvního čtvrtletí roku 2018 je každý den nahráno na servery Facebooku okolo 300 milionů fotografií a databáze tedy roste každým okamžikem (Gewirtz 2018).

Třetí a poslední "V", zastupující slovo variety, které lze přeložit jako různorodost, či rozmanitost, vypovídá o nepřeberném množství nestrukturovaných dat v databázi, přičemž je každý záznam unikátní. U výše zmiňované databáze fotografií sociální sítě Facebook lze například hovořit o tom, že každá nahraná fotografie má unikátní složení barev, velikost na serveru, či datum a čas nahrání. Pokud by byla zkoumána například emailová komunikace celého světa poslaná během jednoho dne, bylo by zjištěno, že každý email má unikátní kombinace adres odesílatele a příjemce, čas odeslání, obsah samotné zprávy a přílohy. Kromě těchto příkladů lze obecně tvrdit, že každá akce generuje nestrukturovaná data, která mohou být neustále ukládána do rozsáhlých databází, které jsou však kvůli různorodosti dat velice komplexní a náročné na zpracování (Gewirtz 2018). Obrázek číslo 1 znázorňuje tři V big data a jejich součásti.

Výše zmíněné tři základní vlastnosti dat, bývají běžně označovány jako tzv. 3 V a postupem času k nim přibývají další "V", například věrohodnost (veracity), hodnota pro firmu (value), limitovaná doba platnosti (validity), či přechodná doba jejich nutného ukládání (volatility) (Holubová, Kosek, Minařík, Novák 2015).



Obrázek 1 3 V Big Dat. Zdroj: V. K. Jain 2017.

3.1.2 Původ dat a internet věcí

Hovoří-li se o tématu big data je třeba si také klást otázku, odkud se vlastně všechna ta rozsáhlá data berou. Zdrojů nestrukturovaných informací je po celém světě nespočet. V souvislosti se sbíráním dat bývá v současné době často zmiňováno slovní spojení „internet věcí“ (internet of things). Koncept internetu věcí spočívá v tom, že je v současné době stále více běžně používaných zařízení vybavováno nejrůznějšími senzory a operačním systémem a přeměňováno na tzv. chytrá zařízení (smart objects), která jsou schopna zaznamenávat data o svém provozu, či okolí. Ve své podstatě se může jednat o jakékoliv zařízení. V dnešní době jsou běžně vybaveny operačním systémem hodinky, mapující například životní funkce nositele, auta s palubním počítačem, zaznamenávající polohové údaje v reálném čase, či lednice objednávající automaticky dodávky potravin (Ravindra, 2017).

Každé zařízení, spadající do internetu věcí je, jak již název napovídá, připojené k internetu prostřednictvím datové sítě, či GPS. Pomocí těchto propojení je možné v reálném čase zaznamenávat nepřeberné množství uživatelských dat, a sledovat a analyzovat jejich vývoj minutu po minutě. V tomto bodě nastupují nástroje big data analýzy, které jsou schopny třídit, zpracovávat a vyhodnocovat konstantní příliv nestrukturovaných informací pocházejících ze zařízení připojených do internetu věcí (Ravindra, 2017).

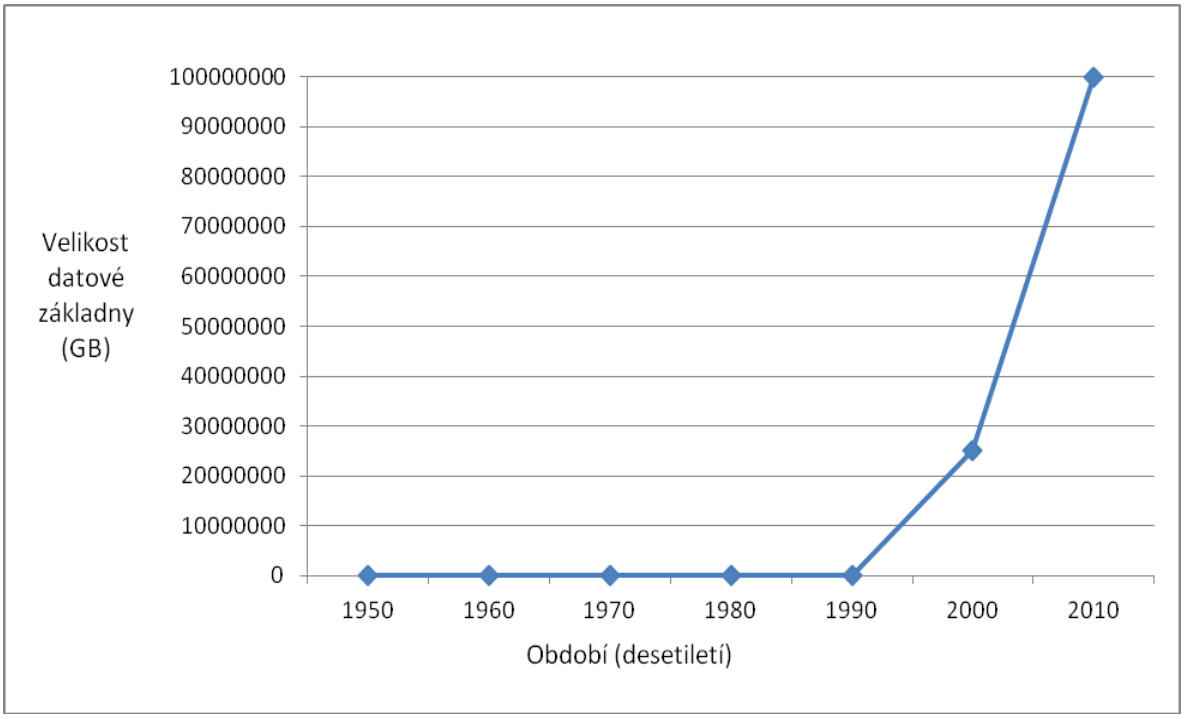
3.1.3 Historie vývoje

Potřeba sbírat, uchovávat a určitým způsobem využívat data vznikla již na konci třicátých let minulého století v USA. Po schválení zákona zvaného Social Security Act, který zajišťoval právo na sociální a zdravotní pojištění bylo potřeba zaznamenat a kontrolovat příspěvky od více než šestadvaceti milionů lidí a třech milionů zaměstnavatelů. Na základě této potřeby přišla americká společnost IBM s řešením, které spočívalo ve vytvoření zařízení, které umělo vytvářet a číst děrované štítky. Jednalo se v podstatě o základ výpočetní techniky (Van Rijmenam 2015).

V průběhu následujících let se důležitost a v první řadě množství dat neustále zvyšovala, čehož si začaly všímat i soukromé firmy. Na konci třicátých let tento proces započal s vynálezem čtečky děrovaných štítků, pokračoval v padesátých letech s vývojem prvních aplikací ve spolupráci s firmou IBM a nyní nám dává možnost uchovávat téměř nekonečné množství nejrůznějších dat. Tabulka číslo 1, zobrazující největší databáze firemních dat od padesátých let názorně demonstruje aktuální trend v rychlosti rozvoje datové základny firem. Tento trend je vyobrazen i na grafu číslo 1 (Rosenbush 2013).

Období (desetiletí)	Jméno společnosti	Obor působení	Velikost datové základny
1950	John Hancock Mutual Life Insurance Co.	Pojišťovnictví	600 Megabajtů
1960	American Airlines	Letecký průmysl	800 Megabajtů
1970	Federal Express Cosmos	Logistika	80 Gigabajtů
1980	CitiCorp's NAIB	Bankovnictví	450 Gigabajtů
1990	Walmart	Obchod	180 Terabajtů
2000	Google	Informační technologie	25 Petabajtů
2010	Facebook	Informační technologie	100 Petabajtů

Tabulka 1 Porovnání datových základen privátních společností v čase. Zdroj: Rosenbush 2013.



Graf 1 Časová osa vývoje datových základen privátních společností v čase. Zdroj: Rosenbush 2013.

Z grafu číslo jedna je patrné, že zlom v kvantitě dat přichází s nástupem internetových společností okolo roku 2005, kdy vzniká i takzvaný Web 2.0. na jehož principech jsou největší internetové společnosti (Facebook, Youtube atd.) založeny. Tento princip spočívá v tom, že obsah webu už není publikovaný jeho tvůrci, či provozovateli, ale hlavně uživateli samotnými. Díky rozmachu takovýchto sítí, zařízení a komunikačních technologií neustále rapidně narůstá množství dat, která jsou produkována. Od počátku informačních technologií ve třicátých letech do roku 2003 bylo vyprodukováno zhruba 5 miliard gigabajtů dat. V roce 2011 bylo toto množství produkováno každé 2 dny, v roce 2012 každých deset minut a tato rychlost neustále narůstá (Marr, 2015) (VK Jain, 2017).

3.1.4 Využití big data

Jak již bylo zmíněno dříve v této diplomové práci, množství dat zachycující skutečnosti světa, ve kterém žijeme neustále roste. Dle výzkumu společnosti McKinsey schopnost tato data správně zachytit, analyzovat a reálně využít v praxi se nyní stává nezbytným základem pro inovaci, konkurenceschopnost a produktivitu každé společnosti.

Dalo by se obecně tvrdit, že data se stávají vedle půdy, práce a kapitálu dalším nezbytným výrobním faktorem (Manyika, 2011).

Sama o sobě by big data by jakožto výrobní faktor nebyla společnostem k ničemu. Je tudíž třeba klást si otázku, jak lze data využít a jakou přidanou hodnotu dané společnosti přináší. V současné době je možné vymezit pět relativně širokých oblastí, kde je možné tuto hodnotu nalézt. V první řadě lze hovořit o zvýšení transparentnosti informací a mnohem snazším přístupu k nim. V řadě druhé vzniká přidaná hodnota využití big data tím, že umožňují organizacím vytvářet a shromažďovat záznamy o veškerých uskutečněných transakcích, pomocí čehož je možné získat detailní přehled o výkonech veškerých podnikových procesů – od externích jako například prodej zboží do určitých regionů, po interní v podobě využití nemocenské dovolené jednotlivými zaměstnanci. Pokud je za pomocí dat proces detailně zmapován, je možné odhalit jeho slabá místa a odstranit, či je zefektivnit patřičnými opatřeními. Nejpokročilejší společnosti využívají data také k provádění kontrolovaných experimentů, které mohou pomoci při provádění manažerských rozhodnutí, či při vytvoření predikce výkonu firmy na další období. Do třetí oblasti lze zahrnout využití datové základny pro úzkou segmentaci zákazníků společnosti, na základě které je možné optimalizovat produkt, či poskytovanou službu. Čtvrtou oblastí je podpora a zlepšení rozhodovacího procesu. Poslední oblastí, kterou je třeba zmínit je ovlivnění další generace výrobků, nebo služeb. V praxi se může jednat například o využití dat získaných ze senzorů umístěných na výrobku k proaktivnímu nabízení poprodejních služeb, údržby a preventivních opatření (oprava výrobku je nabídnuta předtím, než zákazník postřehne závadu) (Manyika, 2011).

Vedle společností, které se v poslední době stále více orientují na datovou analýzu a rozhodování na základě dat existují samozřejmě společnosti, které na data tolik nehledí, a i tak velice dobře prosperují. Naskýtá se tedy otázka, zda má využití dat nějaký reálný dopad na výkon firem jako takových. Na tuto otázku hledala odpověď skupina výzkumníků americké univerzity MIT a společnosti McKinsey & Company vedená informatikem Andrewem McAfeem. Tato výzkumná skupina ve svém projektu testovala hypotézu, která tvrdí, že firmy využívající pro rozhodování podklady vyplývající z datové analýzy budou mít lepší výsledky než firmy řízené tradičně. Do studie se zapojilo více než 330

severoamerických společností, jejichž zástupci uváděli při strukturovaných rozhovorech detaily o struktuře a řízení firmy. Informace z těchto interview zkombinované s ročními výsledky firmy a dalšími nezávislými zdroji daly dohromady široké spektrum nejrůznějších přístupů napříč mnoha obory působení firem, zahrnující různé úrovně využití dat. Po srovnání všech údajů, zahrnující například i rozdíly v technickém zázemí, velikosti společnosti, či investice do IT bylo zjištěno, že firmy, které se popisují jako datově orientované dosahují v průměru o 5 procent větší produktivity a o 6 procent vyšší profitability než jejich konkurenti řízení tradičním způsobem (McAfee, Brynjolfsson, 2012).

Pro lepší představu reálného využití Big Data v praxi lze zmínit například případ společnosti Passur Aerospace popsaný v časopise Harvard Business Review v článku výše zmíněného Andrewa McAfeeho. V popsané případové studii tato společnost optimalizovala zobrazování příletových časů na letištích USA. Je obecně známo, že na letištích závisí vše na přesném načasování a mnohdy jde doslova o minuty. Pokud letadlo přistane dříve, než je na něj personál připraven, musí cestující společně s posádkou čekat uvnitř. Pokud přistane později, musí čekat naopak personál letiště, jehož kapacity by v danou chvíli bylo možné využít efektivně jinak. Z průzkumu jedné z největších leteckých společností v USA vyplynulo, že zhruba 10 % všech letů ve skutečnosti přistává o více než 10 minut dříve, či později, než je předpokládaný čas příletu a přibližně 30 % všech letů má tento rozdíl více než pětiminutový. V době, kdy tato studie probíhala byly veškeré odhady prováděny přímo piloty blížících se letadel, na které jsou však při přistávání kladeny mnohé jiné požadavky a nemají tak kapacitu pro přesné počítání času do přistání, který potom mnohdy odhadovali jen na základě své osobní zkušenosti. V rámci řešení tohoto problému se letecká společnost obrátila na firmu Passur Aerospace, která poskytuje implementaci business inteligence se specializací právě v leteckém průmyslu. Na základě tohoto požadavku přichází společnost Passur Aerospace s novou technologií odhadu, zvanou RightETA, která nahrazuje odhady pilotů přesným výpočtem, který zahrnuje veřejně dostupné údaje o počasí nad konkrétním letištěm, údaje o odletech a příletech, údaje z radarových stanic v okolí letiště a další zdroje. Na základě tohoto výpočtu, dokáže program RightETA přesně predikovat čas přistání všech letadel, která se zrovna pohybují v okolí letiště a chystají se přistát. V současné době program sbírá již více než 15 let

každých přibližně 5 sekund údaje o všech přibližujících se letadlech na více než 155 letištích, což dohromady dává obrovské objemy dat, které je možné dále analyzovat a využívat pro další odhady. Ve své podstatě program RightETA nyní funguje tak, že když „vidí“ blížící se letadlo, porovná údaje o současných podmínkách v okolí letiště se svou databází, vyhledá nejpodobnější případ z historie a na základně něj předvídá s vysokou přesností budoucí čas přistání. Díky tomuto řešení se podařilo téměř odstranit rozdíl, mezi odhadovaným a skutečným časem přistání a optimalizovat tak letištní procesy. Autor článku shrnuje tento postup jednoduchým vzorcem – použití big data vede k lepším předpovědím, lepší předpovědi vedou k lepšímu rozhodování.

Výborným příkladem je také případová studie společnosti Sears (USA, síť obchodních domů), která pojednává o zefektivnění marketingové strategie při ušetření nákladů. Celá iniciativa začala rozhodnutím, že by společnost Sears ráda využila množství zákaznických dat, která měla k dispozici, pro vytvoření personalizovaných nabídek pro zákazníky určitých značek, které pod společnost patří. Pro vytvoření těchto nabídek potřebovalo marketingové a analytické oddělení společnosti zhruba osm týdnů, což v konečném důsledku znamenalo, že značná část nabídek nebyla už v době své publikace aktuální a celé počínání nebylo příliš efektivní. Dlouhá doba, potřebná pro vytvoření personalizovaných nabídek byla způsobena především roztříštěnou datovou základnou obsahující zákaznické údaje, jejíž části byly uloženy v datových skladech mnoha společností, poskytující tuto službu. Ve snaze o zefektivnění tohoto neflexibilního procesu sáhla společnost Sears po řešení z oblasti Big Data. V první fázi se jednalo především o zavedení distribuovaného zpracování velkých objemů dat pomocí metody Hadoop, která propojila veškerá již uložená data z existujících datových skladů s daty nově vytvořenými. Ve fázi druhé bylo zavedeno provádění analýz s využitím všech dat přímo v systému Hadoop, což odstranilo jeden celý procesní krok, spočívající v extrakci dat z různých zdrojů a jejich přípravy pro samotnou analýzu. Tato změna vedla ke značnému zrychlení všech procesů a dala společnosti Sears možnost dynamicky vytvářet kvalitní personalizované nabídky. Dle tehdejšího technického ředitele Phila Shelleyho poklesl čas, potřebný k přípravě nabídek z dvou měsíců na pouhý jeden týden a tato doba neustále klesá. Distribuovaný systém Hadoop společnosti Sears denně zprocesuje několik petabajtů dat, při nesrovnatelně nižších provozních nákladech, než by bylo potřeba vynaložit

k uskladnění tohoto objemu v klasických datových skladech. Celková investice do transformace tradiční správy dat na moderní, cloudové řešení užívající principy big data se díky těmto nižším provozním nákladům svým investorům během doby zcela navrátí (McAfee, Brynjolfsson, 2012).

Příklady využití principů big data společností Passur Aerospace a Sears názorně demonstrují sílu a možnosti, které v současné době využití dat poskytuje. Tyto principy umožňují vytváření stále přesnějších predikcí, lepší podporu rozhodování, zavedení nejrůznější preventivních opatření dříve, než jsou nezbytně nutné a mnoho dalšího. V dnešní době je možné se setkat s využitím datové analýzy v nejrůznějších oborech napříč celým soukromým i státním sektorem. S tímto přístupem se lze setkat v oborech, které svou podstatou k využití dat přímo vybízí (online marketing, či automobilový průmysl) i v oborech, kde využití není zcela nasnadě (zdravotnictví, či lidské zdroje) a ve své podstatě tento přístup stále silněji ovlivňuje celou ekonomiku (McAfee, Brynjolfsson, 2012).

3.2 Business intelligence a zlepšování procesů

3.2.1 Business intelligence

V současném velice konkurenčním prostředí se podnikoví vedoucí pracovníci a analytici musí rozhodovat pod neustálým časovým tlakem bez prostoru pro chybné kroky. K tomu, aby takovéto rozhodování mohlo být efektivní, je potřeba velké množství kvalitních, aktuálních a relevantních informací, které musí být dostupné bez velkých překážek. Termín Business intelligence (zkráceně BI) zastřešuje souhrn znalostí, procesů, technologií, aplikací a postupů, které mají za cíl usnadnit podniková rozhodování. Technologie BI využívá pro práci historická data v určitém kontextu, s jejichž pomocí vytváří podklady pro podniková rozhodnutí ovlivňující budoucí události (Laberge, 2012).

Hodnota systémů Business intelligence spočívá především v tom, že umožňuje uživatelům získávat podklady pro rozhodování bez nutnosti zásahu třetích osob. Na konci 70. let, v začátcích rozvoje informačních technologií, byla pro získání jakýchkoliv dat nezbytná spolupráce několika lidí napříč různými podnikovými odděleními. Uživatel, který v dané době datový výstup vyžadoval, musel nejprve poslat žádost s přesnými

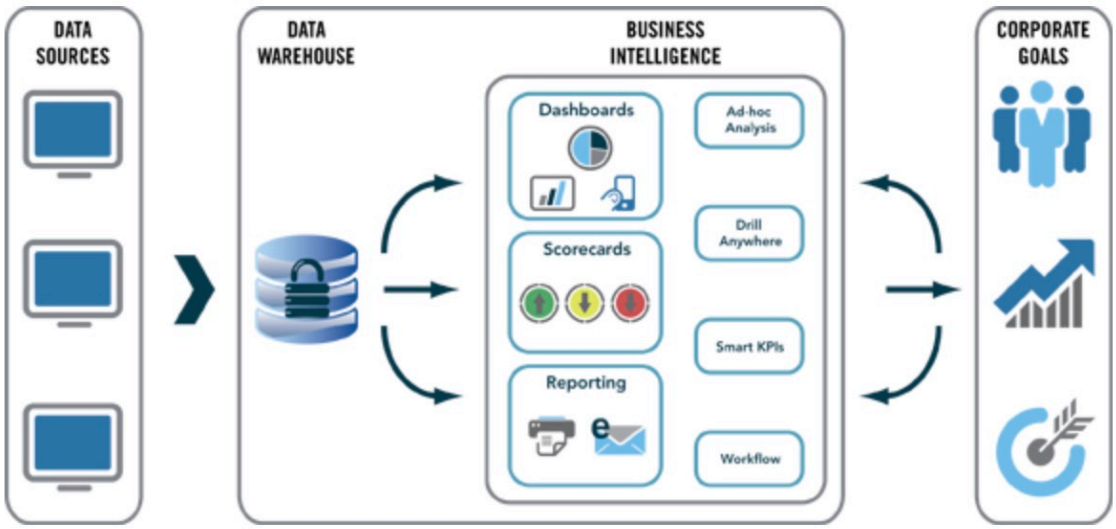
podrobnostmi výstupu na IT oddělení, které po dostatečném upřesnění napsalo program, který vytvořil požadovanou sestavu. V tomto procesu docházelo zpravidla k potížím v komunikaci, a tak vydání informace trvalo zpravidla několik dní až týdnů, což data činilo neaktuálními a ztrácela tak svoji hodnotu (Laberge, 2012).

Současné systémy BI svým uživatelům umožňují, aby měli možnost se k požadovaným datům dostat sami, bez zainteresování dalších osob, během několika okamžiků. Díky tomuto zefektivnění přístupu k informacím je možné zkrátit rozhodovací proces z dnů až týdnů na pouhých pár minut. Business intelligence dramaticky zvyšuje možnosti podnikových uživatelů při zpracovávání informací a také jim umožňuje lépe pochopit kontext, ve kterém se tyto informace vyskytují. IT oddělení, které bylo dříve využívané k samotnému přístupu do databází nyní celý systém monitoruje a ladí. Po porozumění chování uživatelů systému je zpravidla soustředěna pozornost IT oddělení na maximální přizpůsobení výstupů BI tak, aby uživatel našel data, která zrovna vyžaduje ve strukturované, aktuální a srozumitelné formě, která lze rovnou využít jako podklad manažerského rozhodnutí (Laberge, 2012).

Systémy Business Intelligence se dle Labergeho zpravidla vyznačují těmito znaky:

- Rozšíření možností – přímá použitelnost dat jako podklad rozhodnutí
- Rychlost – okamžitá reakce na požadavky
- Aktuálnost – data jsou pravidelně aktualizována
- Přesnost – na kvalitu dat se lze spolehnout
- Užitečnost – systém poskytuje uživatelům hodnotu

Na obrázku č. 2 je názorně vidět proměna nestrukturovaných dat z mnoha zdrojů (ERP systém, HR databáze atd.) na uživatelsky přívětivější, snadno pochopitelnou formu pomocí systému BI.



Obrázek 2 Schéma systému BI. Zdroj: Motivity 2018.

3.2.2 Zlepšování procesů

Alena Svozilová definuje ve své publikaci Zlepšování podnikových procesů (2011) proces jako *„Sled činností, při nichž je aplikováno aktivní působení obsluhujícího personálu, a to jak intelektuální, tak manuální, na postupně vznikající předmět nebo službu, která má přinést nějakou hodnotu pro zamýšleného uživatele – zákazníka procesu.“* Zjednodušeně jako *„Sérii logicky souvisejících činností nebo úkolů, jejichž prostřednictvím – jsou-li postupně vykonávány – má být vytvořen předem definovaný soubor výsledků.“*

Procesy jako takové jsou součástí každodenního života každého člověka do takové míry, že jsou již brány jako samozřejmost. Pokud však výsledky procesu pro jeho zákazníka nejsou uspokojující, nebo jsou naopak překvapivě dobré vstupuje do hry mapování procesů, při kterém jsou jednotlivé činnosti procesu analyzovány a popisovány. Pokud je proces zmapován, je potom možné jeho zlepšení (v případě překvapivě dobrých výsledků zjištění příčiny a její následná aplikace na podobné procesy), či narovnání. Mezi nejčastější úpravy podnikových procesů patří bezpochyby jejich zefektivňování v podobě

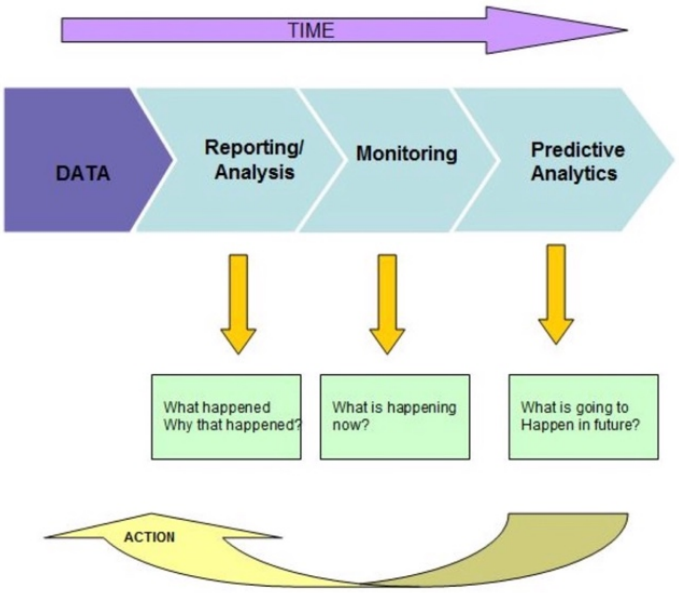
například zvyšování kvality produktů, snižování chybovosti, snižování nákladů a podobně (Svozilová, 2011).

3.3 Prediktivní analýza

3.3.1 Úvod do prediktivní analýzy

Termín prediktivní analýza zahrnuje soubor analytických nástrojů, pomocí kterých lze objevit vzorce vzájemné závislosti jednotlivých veličin v datových souborech. Na základě těchto vzorců lze potom s určitou pravděpodobností předvídat budoucí vývoj událostí a přizpůsobit tomuto vývoji manažerská rozhodnutí.

Prediktivní analýza používá nejrůznější techniky dolování dat (data mining), statistiky, modelování, strojového učení a umělé inteligence k analyzování současných dat a vytváření predikcí budoucího vývoje. Tyto techniky kombinují nástroje managementu, informačních technologií a modelování business procesů k předvídání budoucích událostí. Vzorce vzájemné závislosti, nalezené napříč analyzovanými daty lze využít k nalezení budoucích riskantních situací a příležitostí, které mohou pomoci k porozumění zákaznických dat, zlepšení podnikových procesů, či identifikaci nových obchodních příležitostí. Technika dolování dat společně se statistikou a textovou analýzou umožňují analytikům hledat vzorce závislosti a vytvářet predikce na základě strukturovaných (data připravená k analýze – například věk, výška, váha, pohlaví, rodinný status atd.) i nestrukturovaných (textová pole, data ze sociálních médií, zvukové záznamy atd.) dat. Prediktivní analýza umožňuje společnostem být proaktivní, předvídat budoucí vývoj situace na základě dat a plánovat konkrétní akce a rozhodnutí. Na obrázku č. 3 níže je znázorněn postup při využití nástrojů prediktivní analýzy v praxi. V datech, která jsou k dispozici lze dohledat proč došlo k určitým závěrům (what happened / why that happened), zjistit, co přesně se děje v současné chvíli (what is happening now) a odvodit co se stane v budoucnosti (what is going to happen in the future). Na základě této predikce jsou potom naplánovány konkrétní akce, jejichž výsledné údaje opět vstupují do dat, která jsou dále analyzována (Abbot, 2014).



Obrázek 3 Schéma procesu prediktivní analýzy. Zdroj: PAT Research, 2014.

3.3.2 Komparace prediktivní analýzy a příbuzných statistických disciplín

Prediktivní analýza X business intelligence

Jak již bylo v této práci popsáno, business intelligence i prediktivní analýza slouží jako podpora manažerských rozhodování, která ovlivní chod společnosti v budoucnosti. Obě metody také využívají historická data, hlavní rozdíl však tví ve způsobu, jakým jsou generovány výstupy. Business intelligence v první řadě shrnuje komplexní soubory dat do uživatelsky přívětivých, přehledných výstupů, zvaných KPI – key performance indicators (klíčové ukazatele výkonnosti). Ukazatele KPI jsou předem definovány osobou (user driven), která systém business intelligence spravuje a ví, jaké exaktní výstupy z programu požaduje. Typickými otázkami, na které systém business intelligence hledá odpovědi v praxi, můžou být například následující:

- Jaká je procentuální úspěšnost vymáhání klientských pohledávek?
- Kolik uživatelů otevřelo reklamní email s nabídkou?
- Kolik zákazníků se zapsalo do věrnostního programu?

Na rozdíl od business intelligence, prediktivní analýza hledá v datech předem neznámé souvislosti a vzorce. Na základě těchto vzorců potom určuje, jak se jednotlivé veličiny ovlivňují a jak případně ovlivní budoucí výsledky. Otázky, na které se lze ptát v praxi při prediktivní analýze mohou vypadat například takto:

- S jakou pravděpodobností klient zaplatí pohledávku?
- Jaká je pravděpodobnost toho, že zákazník otevře reklamní email?
- Kolik reklamních emailů je nutné zákazníkovi poslat, aby došlo ke konverzi (nákupu)?

Ve výše zmíněných otázkách se často vyskytuje slovní spojení „jaká je pravděpodobnost“, které značí, že je počítána pravděpodobnost výskytu určitého vzorce a veličiny, které ovlivňují výsledky jednotlivých úkonů (Abbot, 2014).

Prediktivní analýza X statistika

Prediktivní analýza má se statistikou mnoho společného a někteří odborníci ji dokonce považují za její odnož. Těmito společnými prvky jsou například algoritmy a statistické testy, které jsou při prediktivní analýze běžně používány. Rozdíl mezi prediktivní analýzou a statistikou tkví především v pojetí řešeného problému a míře kontroly, kterou nad modelem a daty analytik má. Statistické nástroje jsou často používány k vytvoření analýzy, potvrzující, či vyvracející určitou předem danou hypotézu o předem známém vztahu. Tato analýza má za cíl potvrzený, či vyvrácený vztah kvantifikovat a určit sílu závislosti. V praxi se může jednat například o potvrzení účinnosti určitého léku na základě potlačení výskytu určité nemoci v organismu testovaného jedince. U statistických analýz je třeba zajistit, aby vstupy a výstupy modelu odpovídaly jeho předpokladům a nedošlo tak k jeho chybné interpretaci. Tato chybná interpretace může být způsobena například zahrnutím extrémních hodnot do modelu, což značně ovlivní celkový výsledek a model tudíž reflektuje realitu nedokonale (Abbot, 2014).

Oproti výše zmíněným statistickým modelům se analytici, zabývající se prediktivní analýzou soustředí především na vytvoření modelu, který s co nejvyšší přesností dokáže predikovat budoucí vývoj a může ovlivnit rozhodování. Samotné parametry, které jsou do

modelu zahrnuty nehrají příliš velkou roli. Rozdíl mezi oběma disciplínami spočívá také v strukturovanosti problému. Zatímco statistika se zabývá především konfirmační analýzou strukturovaných problémů, prediktivní analýza se pokouší rozklíčovat nestrukturované problémy s použitím dat, která nebyla shromážděna za účelem využití v modelu a byla v patřičnou chvíli k dispozici. Jak již bylo zmíněno výše, prediktivní analýza bývá využívána především k řešení nestrukturovaných problémů v kontextu, který nelze žádným způsobem kontrolovat. V praxi se může jednat například o rozhodování, kterou marketingovou kampaň aplikovat na zákazníka elektronického obchodu. Zákazník může dostat jednu z deseti variant reklamního emailu, přičemž rozhodnutí bude vycházet z jeho demografických dat, předchozího chování na dané webové stránce a předchozí reakce na reklamní emaily (ignorace zprávy, otevření bez akce, či provedení konverze – nákupu produktu). Prediktivní modely dokáží na základě těchto dat předvídat s určitou pravděpodobností budoucí chování zákazníka, nicméně nedokáží postihnout veškeré možné nepředvídatelné události (zákazník je ovlivněn ostatními reklamami v bannerech po stranách webu, návrhy vyhledávání Google, či vyrušen v průběhu nákupu). Vzhledem k nekonečnému množství nepředvídatelných událostí, které mohou v průběhu zkoumání nastat, nelze chování zákazníka stoprocentně předvídat. Co však možné je, je zahrnutí veškerých proběhlých situací (i těch nepředvídaných) do prediktivního modelu, což v konečném důsledku zpřesní všechny budoucí předpovědi. Zatímco statistika je orientována především na model jako takový, prediktivní analýza se primárně soustředí na samotná data (Abbot, 2014).

Prediktivní analýza X data mining

Poslední oblastí, se kterou bývá prediktivní analýza srovnávána je dolování dat, tzv. data mining. Pojem data mining, označující proces hledání spojitostí, vzorců a trendů v rozsáhlých datových souborech má s prediktivní analýzou mnoho společného, nicméně není možné tyto dva pojmy zaměňovat. I přes mnoho společných elementů, kterými jsou například použité statistické metody ke zpracování dat, jejich následná segmentace, či využití principu strojového učení, je třeba chápat tyto dvě oblasti odlišně. Hlavní rozdíl tkví především ve finálním využití získaných poznatků. Zatímco data mining pomáhá zainteresované osobě lépe pochopit sesbíraná data, prediktivní analýza na jejich základě

určuje s danou pravděpodobností jejich budoucí vývoj a pomáhá tak s rozhodovacím procesem (Abbot, 2014).

V praxi se může jednat například o analýzu nákupního chování zákazníka. S pomocí metod data miningu je možné rozčlenit stávající zákazníky do určitých segmentů, či odhalit vzorec jejich nákupního chování v čase. S pomocí prediktivní analýzy lze expandovat tyto poznatky za hranice dostupných informací. S konkrétní pravděpodobností je kupříkladu možné určit, jaký výrobek zákazník koupí příště, kolik zákazníků zareaguje na výhodnou nabídku, nebo jak velký objem zboží bude možné příští měsíc prodat (Abbot,2014).

Klíčový rozdíl spočívá také v okruhu osob, který musí být v procesu data miningu a prediktivní analýzy zahrnut. Zatímco první jmenovaný přístup vyžaduje především znalosti algoritmů strojového učení, statistiky a matematiky pro vytvoření modelu, následná prediktivní analýza vyžaduje aplikaci znalostí specifického obchodního prostředí. Díky této znalosti je možné přiřadit nalezeným vzorcům a vztahům konkrétní hodnoty pro daný podnik a transformovat tyto poznatky do rozhodnutí ovlivňující budoucí vývoj (EduCBA, 2018).

3.3.3 Překážky pro aplikaci prediktivní analýzy v podniku

Prediktivní analýza může podniku výrazně pomoci se zvýšením celkové efektivity procesů, návratnosti investic, či jako podklad manažerských rozhodnutí. I přes všechny benefity, které může podniku přinést se však značná část vytvořených modelů nikdy do provozu neaplikuje. Dean Abbot (2014) ve své publikaci kategorizuje překážky, bránící nasazení prediktivní analýzy do čtyř hlavních skupin – překážky v managementu, překážky spojené se samotnými daty, překážky spojené s modelem a překážky spojené s nasazením.

Překážky v managementu

Aby mohl být vytvořený model aplikován, je nutné pro tuto akci alokovat značné množství zdrojů tak, aby bylo možné ho transformovat z výzkumné a testovací fáze do „ostrého“ provozu. Takováto zdrojově náročná činnost vyžaduje v praxi silnou podporu managementu s dostatečnou politickou mocí, pomocí které je možné prosadit vynaložení

zdrojů právě do této oblasti, která může na první pohled pro nezainteresované osoby působit nedůležitě a v porovnání se standardními výdaji neatraktivně. Management musí v případě nasazení modelu do praxe také zajistit aplikaci a dodržování nově vzniklých procesů a důkladné proškolení všech osob, kterých se budou tyto nové procesy týkat (Abbot, 2014).

Překážky spojené s daty

Druhá kategorie možných komplikací, se kterými je možné se při aplikaci prediktivní analýzy setkat, spočívá v nesprávné struktuře podkladových dat. Pro účely tohoto typu analýzy je nezbytně nutné mít podkladová data v jedné dvojrozměrné tabulce, obsahující řádky a sloupce. Pokud jsou data obsažena ve více různých zdrojích, je třeba je pomocí určitého klíče propojit do jednotného zobrazení, odpovídající výše zmíněnému formátu (Abbot, 2014).

Data, která jsou pro analýzu využívána musí také být konzistentní, úplná a aktuální vzhledem k prováděné analýze. Je-li například prováděna analýza dopadu marketingové kampaně, využívající pravidelně aktualizovanou databázi zákazníků a jejich demografických údajů (věk, příjem, bydliště atd.), musí být vždy využity tyto údaje z doby dané kampaně, a ne jejich pozdější aktualizace (Abbot, 2014).

Překážky spojené s modelem

Překážky spojené s nasazením prediktivní analýzy v podniku se mohou týkat také samotného modelu. Jedna z hlavních komplikací spočívá ve snaze analytiků zahrnout do modelu zbytečně moc proměnných ze stávajícího datového souboru, což vede k přílišné komplikovanosti modelu a jeho zkreslené vypovídací hodnotě. V takovémto případě může nastat situace, kdy jsou vynaloženy prostředky k aplikaci modelu, který právě kvůli zbytečné komplikaci a fixaci na původní datový soubor začne po nějakém čase selhávat. Obecné doporučení pro tvorbu prediktivních modelů říká, že je lepší aplikovat nejprve jednoduché modely, které mohou být postupně zlepšovány a rozšiřovány. Pokud takový jednoduchý model v prvotní fázi selže, není to pro podnik taková zdrojová zátěž a model může být snadno nahrazen (Abbot, 2014).

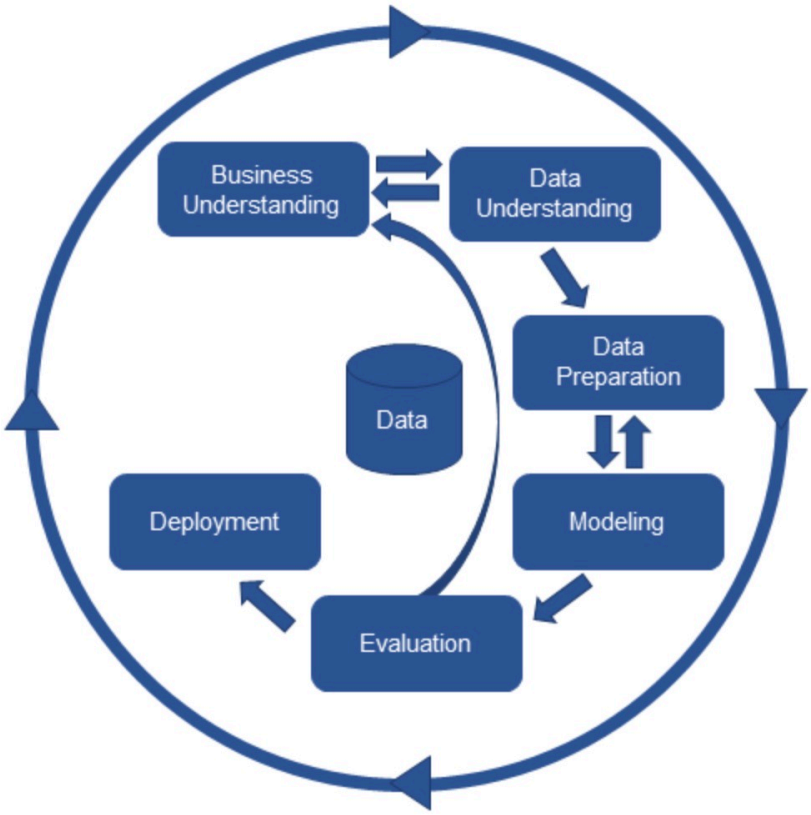
Překážky spojené s nasazením

Poslední kategorií překážek, kterou Abbot (2018) definuje jsou překážky spojené s nasazením modelu do praxe (deployment). Tato kategorie zahrnuje situace, kdy prediktivní model není čitelný operačním systémem stroje, který jej zpracovává. Modely jsou zpravidla zpracovávány komplexními programovacími jazyky (SQL, C++, Java) a pokud nelze model těmito jazyky přečíst, není možné je správně aplikovat.

3.3.4 Procesní metodika prediktivní analýzy – CRISP-DM

Pokud má být přístup data mining a vlastní prediktivní analýza aplikována, je nutné provést sérii činností, bez níž by nebylo možné žádné výsledky získat. Veškeré tyto činnosti a procesy pokrývá metodika zvaná Cross-Industry Standard Process for Data Mining (zkráceně CRISP-DM), která vznikla s podporou Evropské Unie na konci devadesátých let v reakci na masový rozvoj trendu sbírání dat a snaze jim lépe porozumět. Výhodou této metodiky je zejména její nezávislost na konkrétním odvětví, ve kterém má být aplikována, nezávislost na použitých softwarových nástrojích a konkrétních algoritmech data miningu (Chapman, 2000).

Metodika CRISP-DM popisovaná Petem Chapmanem (2000) představuje volně dostupný, obecně platný postup, který lze využít k řešení komplexních data miningových problémů. Dle metodiky CRISP-DM obsahuje každý data miningový projekt šest na sebe navazujících fází. Těmito fázemi jsou business understanding (porozumění bussinesové stránce věci, data understanding (porozumění datům), data preparation (příprava dat), modeling (modelování), evaluation (zhodnocení), deployment (aplikace do praxe) – viz. obrázek č. 7, přičemž pozdější fáze vždy navazuje na výstupy fáze předchozí. Mezi některými fázemi je možný i zpětný postup. V některých případech je potřeba vrátit se v postupu o krok zpět a provést potřebné změny tak, aby byly ovlivněny výstupy postupující do další fáze. Na obrázku číslo 7 je také znázorněn cyklický charakter této metodiky, značící její opakovatelnost v rámci postupného zlepšování celého procesu. Po úspěšném zavedení modelu je tudíž vhodné vrátit se zpět na začátek a překonfigurovat vstupy tak, aby došlu k co nejmarkantnější optimalizaci.



Obrázek 4 Fáze referenčního modelu metodiky CRISP-DM. Zdroj: Chapman, 2000.

Business understanding – porozumění problematice

První fáze metodiky CRISP-DM spočívá ve správném porozumění oblasti, do které řešený problém a jeho podkladová data spadají. V této fázi je potřeba především definovat veškeré požadavky a cíle celého projektu z manažerského pohledu včetně hodnocení rizik, potřebných prostředků, očekávaných přínosů a metriky, která určí, zda byl projekt úspěšný, či nikoliv. Po stanovení těchto požadavků je třeba vypracovat předběžný plán průběhu dílčích prací projektu a způsobů, jak chtěných výstupů dosáhnout. V této fázi se řešený problém transformuje do data miningové úlohy (Chapman, 2000).

Data understanding – porozumění datům

Fáze porozumění datům spočívá především jejich prvotním sběru, základním porozumění řešeného problému v jejich kontextu a možném odhalení na první pohled zřejmých poznatků. Chapman (2000) definuje v rámci této fáze několik základních kroků, které je nutné provést.

Sběr vstupních dat

V kroku, který se týká sběru dat je nutné získat data, či zajistit přístup k datům, na jejichž základě se bude daná analýza provádět. Zdroj těchto dat je definován v první fázi metodiky CRISP-DM. Výstupem tohoto kroku je seznam získaných datových souborů, jejich umístění (vlastnictví), metod s jejichž pomocí byla data získána a seznam komplikací, které při sběru nastaly, nebo potencionálně mohou nastat (Chapman, 2000).

Popis dat

Popisem dat je v tomto případě myšleno jejich povrchové zhodnocení a základní výčet jejich vlastností. V tomto kroku je datový soubor popsán z hlediska získání dat, jejich typu, množství záznamů, hodnot, kterých nabývají jednotlivá pole atd. V kroku popisující data je také důležité zhodnotit, do jaké míry jsou k danému projektu relevantní, jakou mají hodnotu a jaké jsou možnosti jejich využití (Chapman, 2000).

Poznávání dat

V rámci kroku poznávání dat popisuje Chapman (2000) úlohy spojené s jejich vizualizací, zkoumáním a vyhodnocováním. Spadá sem například určování četnosti výskytu nejdůležitějších atributů, krajních hodnot, průměrů atd.

Ověřování kvality dat

Ve čtvrtém kroku je dle Chapmana (2000) klíčové zodpovědět následující otázky:

- Pokrývají získaná data zkoumanou oblast v dostatečné míře?
- Je datový soubor v pořádku, nebo obsahuje chyby?
- Pokud obsahuje chyby, v jaké míře se vyskytují?
- Obsahuje datový soubor chybějící data?

- Pokud obsahuje chybějící data, v jaké míře se vyskytují a jak jsou reprezentovány?

Výstup z tohoto kroku by měl obsahovat odpovědi na výše zmíněné otázky včetně návrhů přípustných řešení vedoucích k odstranění nedostatků. Řešení kvalitativních problémů v datech je velice silně závislé na znalostech konkrétní problematiky (business understanding).

Data preparation – Příprava dat a výběr proměnných

Třetí fáze CRISP-DM má za cíl transformovat dostupná data do tvaru, který bude možné co nejlépe použít v rámci algoritmů prediktivního modelování. V rámci fáze přípravy dat je každý datový soubor unikátní a nelze určit obecně platný návod, jak data připravit a očistit z důvodu možnosti výskytu datových chyb, které jsou specifické pouze pro zkoumané odvětví. Hovoří-li se o fázi přípravy dat, mnoho autorů se shoduje, že se jedná o časově nejnáročnější fázi metodiky CRISP-DM, která může zabrat cca 60-90 % celkového času, který je na projekt potřeba. Hlavní problém tkví především v tom, že téměř veškerá data reálného světa obsahují určité chyby, které musí analytik odhalit a nahradit tak, aby zůstala vypovídací hodnota datového souboru zachována (Abbot, 2014).

Ve fázi přípravy dat je vhodné začít s čištěním hodnot proměnných (variable cleaning) od chybných hodnot, chybějících hodnot, či hodnot, které jsou spojené s nesprávným naprogramováním databáze. Proměnné, které by měly projít čištěním by měly být definovány už v předchozí fázi metodiky CRISP-DM – data understanding (porozumění datům) (Abbot, 2014).

Chybné hodnoty proměnné, kterou databáze obsahuje mohou být odhaleny například pomocí vizualizace četnosti výskytu hodnot. Vymyká-li se například hodnota ostatním a zároveň se v daném souboru vyskytuje pouze jednou, je zde velká pravděpodobnost, že se bude jednat o chybu. K určení toho, které hodnoty jsou akceptovatelné a které už lze považovat za chybné je nejvhodnější přizvat k analýze experta se zkušenostmi z konkrétního prostředí řešeného problému (Abbot, 2014).

Abbot (2014) popisuje kromě chyb proměnných také chyby, plynoucí z nejednotných formátů dat, či nesprávného naprogramování databáze. Chyby spojené s formáty dat plynou především z kombinace jejich zdrojů, což vede k zahrnutí například data ve formátu /dd/mm/yy a zároveň /mm/dd/yyyy. Takováto kombinace v modelu nebude fungovat a výsledky budou vykazovat chyby. Chyby spočívající v nesprávném naprogramování databáze si lze představit jako hodnoty, které se plošně vymykají očekávání. Pokud například v databázi, která obsahuje věk zákazníků bude u všech záznamů číslo trojciferné, začínající devítkou, je zde velká pravděpodobnost, že se jedná o chybné naprogramování a správně jsou pouze dvě čísla následující za devítkou.

Poslední a zřejmě nejčastější a nejvíce časově náročnou kategorií chyb je výskyt chybějících hodnot v datovém souboru. Pokud je takovýto datový soubor nutné analyzovat, naskýtá se několik řešení této situace. V první řadě je možné aplikovat postup odmazání záznamů (řádků), které obsahují chybějící hodnoty z datového souboru a pokračovat pouze se záznamy, které mají hodnoty vyplněné u všech proměnných (sloupců). Pokud se jedná o celou proměnnou, která vykazuje chybějící hodnoty, je žádoucí tuto proměnnou do modelu nezahrnovat a z datového souboru vymazat. Pokud se chybějící hodnoty vyskytují v souboru jen výjimečně, je tento postup velice efektivní. Pokud však soubor vykazuje mnoho chybějících hodnot a odmazání takovýchto záznamů a proměnných by znamenalo omezení vypovídací funkce modelu, nelze tento postup aplikovat (Abbot,2014).

V případě, že nelze aplikovat výše zmíněný postup odmazávání záznamů, či proměnných, doporučuje Abbot (2014) několik možných řešení. V prvním případě se jedná o nahrazení chybějících hodnot konstantou (například nulou), což ale v mnoha případech nedává v kontextu dat smysl (například u věku dospělých osob). Druhým možným řešením problému chybějících dat je jejich nahrazení průměrnou hodnotou, což ale při následném zkoumání rozptylu a standardních odchylek značně tyto údaje ovlivňuje (odchylka se snižuje směrem k průměrné hodnotě). Třetím, zřejmě nejefektivnějším způsobem je nahrazení chybějících hodnot náhodnými hodnotami vycházejícími z normálního rozdělení hodnot v daném souboru, průměru a standardní odchylky. Tímto zůstane vypovídající hodnota zachována a model je možné použít v plném rozsahu.

Na závěr výše zmíněné praxe odmazávání a nahrazování chybějících proměnných se Abbot (2014) zamýšlí nad tím, v jakém případě má a v jakém případě už nemá smysl proměnné nahrazovat a celkově jejich hodnoty do modelu zahrnovat. Odpověď na tuto otázku není dle Abbota jasně daná a vždy záleží na konkrétním modelu, a především na citu a zkušenostech analytika, který se jím zabývá.

Výběr proměnných

Zatímco Chapman v publikaci CRISP-DM 1.0 (2000) doporučuje po fázi přípravy dat přikročit k samotnému modelování, Dean Abbot (2014) doporučuje zahrnout krátký mezikrok zúžení výběru proměnných, které do modelu vstoupí. V tomto mezikroku Abbot doporučuje projít znovu výčet všech proměnných, které by měly být do modelu zahrnuty, otestovat je a vyřadit všechny, které se projeví jako nerelevantní, či redundantní.

Jako nerelevantní lze označit takové proměnné, jejichž hodnoty nebude možné z různých důvodů v budoucnu do modelu zahrnout. Jedná se například o regulované atributy klientů (náboženské vyznání, rasa, místo bydliště), či taková data, jejichž pravidelné získávání je příliš nákladné a není s ním možné pokračovat do budoucna. Nerelevantní proměnná může být například také strojově generované identifikační číslo zákazníka, které nemá žádnou vypovídací hodnotu a není nutné ho do modelu zohledňovat (Abbot, 2014).

Redundantní proměnné jsou ty, které vykazují velice silnou korelaci s jinými proměnnými. Pokud se v databázi objevují dvě, či více proměnných, jejichž korelace je vyšší než 0,9, nebo nižší než -0,9 stačí pro získání informace do modelu zahrnout pouze jednu z nich (Abbot, 2014).

Modeling – modelování

Další fáze metodiky CRISP-DM spočívá ve výběru a aplikaci nejrůznějších modelovacích technik a kalibrace jejich parametrů pro získání maximálně optimálních výsledků. V praxi je běžně k řešení jednoho problému aplikováno více modelů a jejich výstupy jsou následně zkombinovány. Některé techniky vyžadují specifický formát dat,

která do modelu vstupují, a proto je často nutné vrátit se v metodice o krok zpět a data upravit (Chapman, 2000).

V rámci modelování je nutné rozlišit, zda bude prováděno modelování tzv. deskriptivní, či prediktivní. V rámci deskriptivního modelování jsou vytvářeny algoritmy, které hledají souvislosti pouze v rámci vstupních dat a určují jejich přirozenou segmentaci. V rámci modelování prediktivního jsou potom vytvářeny algoritmy s cílem určit hodnotu cílové proměnné na základě hodnot ostatních vstupů (Abbot, 2014).

Regresní analýza

Regresní analýza je v současné době jednou z nejvyužívanějších technik analýzy více faktorových dat. Její hojné využívání pramení zejména z logického procesu, ve kterém je využito rovnic k vyjádření vztahu mezi konkrétní proměnnou a množinou souvisejících, prediktorových proměnných. Tento vztah vyjadřuje, jak se změní vysvětlovaná proměnná (závislá) při změně proměnné vysvětlující (nezávislé) za jinak nezměněných podmínek (ceteris paribus) (Montgomery, 2011).

Principy regresní analýzy jsou využívány v nejrůznějších regresních modelech, například v modelech lineární regrese, logistické regrese, rozhodovacích stromů, či neuronových sítí. Základním a zároveň nejméně komplexním z těchto modelů je model jednoduché lineární regrese, který předpokládá lineární závislost mezi dvěma veličinami. Rovnice regresní přímky může být zapsána v následujícím tvaru:

$$y=\beta 0 + \beta 1x1 + e_i$$

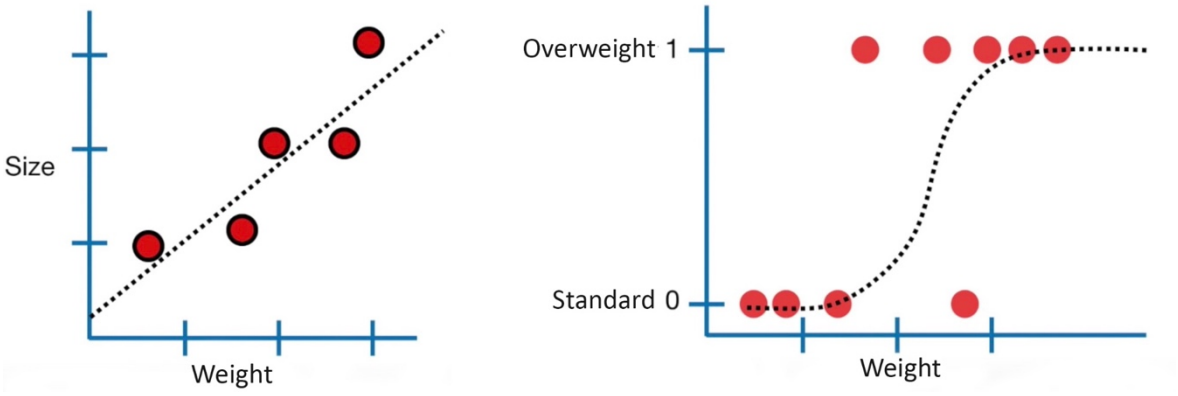
Přičemž Y reprezentuje závislou proměnnou, regresní koeficienty (β0 a β1x1) reprezentují deterministickou složku a chybový člen e_i reprezentuje stochastickou složku. U chybového členu c platí obecný předpoklad, že je jeho rozptyl konstantní, průměr roven nule a rozdělení normální. Cílem regresní analýzy je určení toho, jak se změní Y v závislosti na změně X neboli vyvození závěrů o závislé proměnné v využitím informací o proměnné nezávislé (Freund, 2006).

V souvislosti s lineární regresí bývá často také určován tzv. korelační koeficient, který nabývá hodnot v intervalu -1 až 1 včetně a vyjadřuje sílu vzájemné závislosti proměnných. Čím blíže je korelační koeficient ke krajním hodnotám intervalu, tím silnější je lineární vztah mezi proměnnými. Naopak čím blíže je daný koeficient nule, tím slabší existující vztah je (Montgomery, 2011).

Lineární regresní model může obsahovat i více než jednu nezávislou proměnnou. V takovém případě je model označován jako vícenásobný a jeho podstata a smysl je stejná, jako u modelu jednoduchého s tím rozdílem, že je náročnější na výpočet a je nutné brát v úvahu i vzájemné vztahy mezi nezávislými proměnnými (Freund, 2006).

V případě, že není z nějakého důvodu možné použít lineární regresní model (například kvůli výskytu binomických, či polynomických kvalitativních proměnných), je nutné zvolit pro analýzu model logistické regrese. Tento model je oproti výše zmíněnému lineárnímu modelu křivočarý a spočívá v analýze vzájemného působení jednotlivých proměnných (Freund, 2006).

Na rozdíl od lineární regrese, ve které je možné určit na základě nezávislých proměnných hodnotu proměnné závislé, logistická regrese je využívána především pro kategorizaci výstupů. Závislá proměnná tedy nabývá pouze hodnot 1 (true) pro výsledek potvrzující hypotézu a 0 (false) pro výsledek vyvracející hypotézu. Na obrázku číslo 5 je vidět porovnání grafických znázornění lineární (vlevo) a logistické (vpravo) regrese. V tomto konkrétním příkladu logistické regrese se jedná o zařazování jedinců do váhových kategorií (osa Y) dle tělesné hmotnosti (osa X). Křivka potom určuje jistotu v intervalu od nuly do jedné, s jakou jsou jednotlivé tělesné hmotnosti zařazovány do kategorií standardní váhy, či nadváhy (Starmer, 2018).

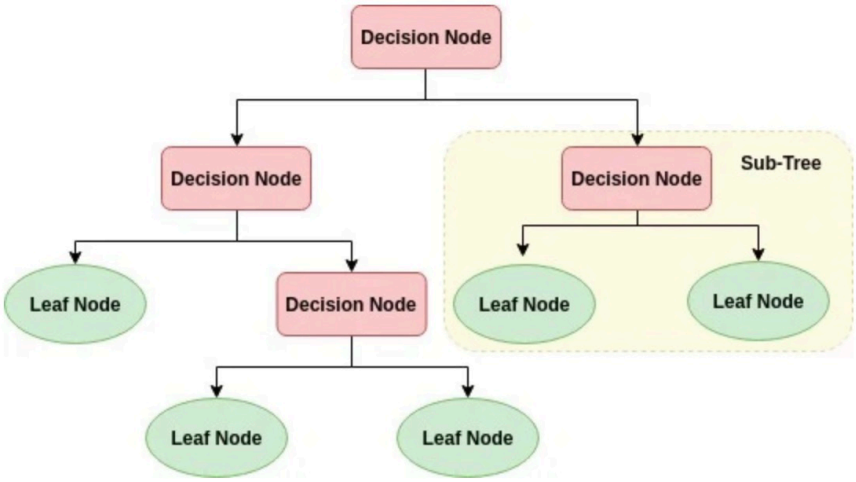


Obrázek 5 Porovnání lineární a logistické regrese. Zdroj: Vlastní zpracování.

Rozhodovací stromy

Dean Abbot (2014) ve své publikaci popisuje modelovací techniku rozhodovacích stromů jako jednu z nejpopulárnějších. Za svou popularitu vděčí tato technika především své srozumitelnosti i pro uživatele, kteří nemají v analytice, či matematice příliš silné zázemí a potřebují přehledný podklad pro svá manažerská rozhodnutí. Ve své podstatě algoritmus rozhodovacího stromu vytváří pomocí série pravidel „if – then – else“ grafy vzájemné závislosti proměnných. Tyto grafy jsou podobné rozvětvené struktuře stromu a identifikují možnosti, jak data rozdělit do určitých segmentů, do kterých následně jednotlivé vstupy spadají. Výhoda rozhodovacích stromů tkví také v jejich snadné aplikovatelnosti do nejrůznějších systémů (například SQL databází), možnosti kombinovat číselné i kategorické vstupy, či automatické schopnosti zpracování chybějících dat. Podle Hassouny et al. (2015) jsou v rozhodovacích stromech nejčastěji využívány algoritmy CART (Classification and Regression Trees), C 5.0 a CHAID (Chi-square Automatic Interaction Detector). Tyto algoritmy při vytváření rozhodovacích stromů procházejí dvě základní fáze – růst a rozvětvování, či pučení. Strom začíná růst z kořenového uzlu, který je paradoxně zobrazen na vrcholu stromu. Tento kořenový uzel obsahuje objekt analýzy, tedy proměnnou, jejímž modelováním se získá nejvíce přínosných informací (informací, které mají pro společnost největší hodnotu). Po stanovení kořenového uzlu dochází k postupnému růstu stromu a stanovování dalších dílčích uzlů na základě algoritmu „rozděl a panuj“, který spočívá v rozdělení komplexní úlohy na menší úlohy dílčí. Tato metoda spojuje cílové pole v datech (závislou proměnnou) s ostatními datovými poli (nezávislými proměnnými), které slouží jako podklad pro vytvoření

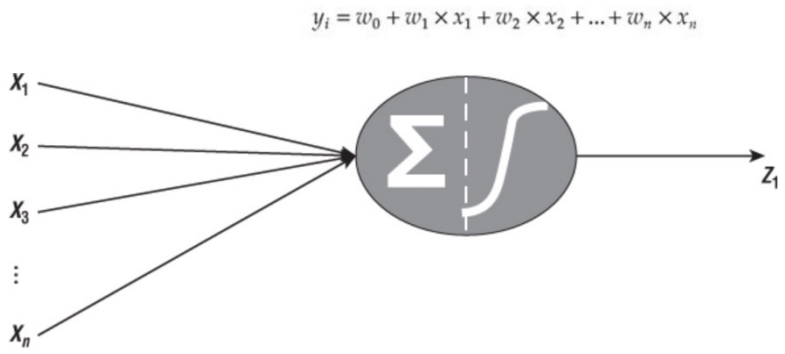
segmentů a větví. Po odhalení vztahu, který mezi těmito proměnnými existuje je možné odhalit i konkrétní pravidla, která vztah mezi vstupními a cílovými hodnotami ovlivňují. Výsledný rozhodovací strom pomáhá vizualizovat jednotlivé vztahy a analyzovat jejich síť. Na základě těchto vypozorovaných vztahů lze předvídat cílovou hodnotu nových záznamů (pozorování), u kterých jsou známy pouze vstupní hodnoty (SAS, 2008) (Hassouna, 2015).



Obrázek 6 Příklad struktury rozhodovacího stromu. Zdroj: Navlani, 2018.

Neuronové sítě

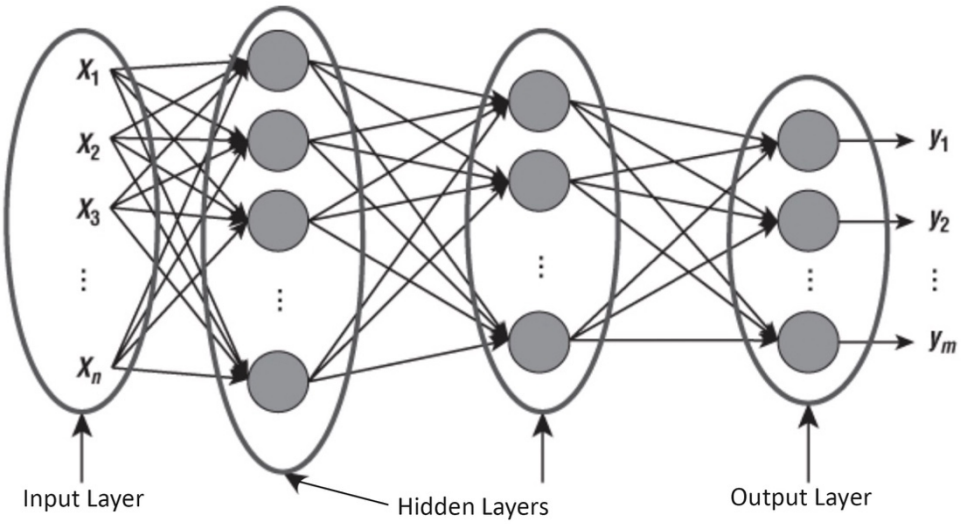
Třetím možným způsobem tvoření prediktivního modelu je využití principu neuronových sítí, inspirovaného biologickou funkcí a strukturou sítí v mozku živočichů, který hledá vzory v datech a dává jim konkrétní význam. Tento princip spadá do kategorie tzv. hlubokého učení (deep learning) a v porovnání s regresními modely, či rozhodovacími stromy je o dost komplexnější. Využívá se tedy kromě kategorizačních úloh také k řešení složitějších problémů, spočívajících například ve strojovém učení rozpoznávání objektů na obrázcích. Základní stavební jednotkou neuronových sítí jsou jednotlivé umělé neurony srovnané do vzájemně provázaných vrstev s postupující tendencí, které jsou schopny vzájemně přenášet signály. Každý neuron v sobě obsahuje rovnici transformující vstupy a jejich váhy na výstup, který je potom algoritmem dále zpracován (viz. obrázek číslo 7).



Obrázek 7 Neuron transformující vstupní hodnoty. Zdroj: Abbot, 2014.

Váhy jednotlivých kritérií můžou být v průběhu modelování algoritmem měněny v závislosti na pokroku strojového učení. Pokud je například některý ze vstupů na začátku nastaven jako důležitý, ale v průběhu modelování algoritmus zjistí, že je jeho přínos zanedbatelný, nastaví mu odpovídající váhu automaticky (Abbot, 2014).

Jak již bylo zmíněno výše, tyto neurony se skládají v rámci neuronové sítě do jednotlivých provázaných vrstev, kterých může být libovolný počet dle potřeby konkrétního modelu. Pro analytika jsou však viditelné pouze dvě vrstvy modelu. První (input layer) obsahuje veškeré vstupy z vnějšího prostředí a druhá (output layer) výstupy algoritmu. Ostatní vrstvy jsou označovány jako tzv. skryté vrstvy (hidden layers) a slouží pouze pro výpočetní účely modelu. Na obrázku číslo 8 je znázorněno vrstvení modelu neuronových sítí (Abbot, 2014).



Obrázek 8 Jednotlivé vrstvy modelu neuronové sítě. Zdroj: Abbot, 2014.

Abbot (2014) ve své publikaci také zdůrazňuje zásadní rozdíl modelu neuronových sítí oproti modelům ostatním. Zatímco v případě logistické regrese, či rozhodovacího stromu je model schopný naučit se určovat koeficienty a váhy jednotlivých proměnných na základě jednoho proběhlého procesního cyklu, u neuronových sítí probíhá učení postupně (iterativně). Model v prvním cyklu vezme například pouze jeden záznam, přiřadí náhodné váhy jednotlivým kritériím, předpoví výsledek a pak ho porovná s cílovou proměnnou a změří chybu. V dalších cyklech (může jich být desítky, stovky i tisíce) model prochází postupně záznamy, měří chyby a upravuje váhy kritérií, dokud nedosáhne uspokojivého výkonu s minimální chybovostí. Tento proces je velice náročný na výpočetní kapacity systému a oproti ostatním modelům trvá jeho zpracování několikanásobně déle (Abbot, 2014).

Evaluation – zhodnocení

Ve fázi evaluace, tedy vyhodnocení výsledků proběhlého modelování dochází k celkovému zhodnocení získaných znalostí a jejich hodnoty pro manažery. V této fázi také dochází ke hledání možných nedostatků, které by mohly bránit jeho nasazení a případnému testování v praxi. Pokud bylo ve fázi modelování vytvořeno více modelů řešící daný obchodní případ, je nutné je v této fázi vzájemně porovnat a vybrat nejvhodnější. Výstupem této fáze by dle Chapmana (2000) mělo být mimo jiné vyhodnocení, zda byly

splněny cíle zmíněné v první fázi metodiky, či doporučení pro příští data-miningové projekty na základě nově získaných poznatků.

Pro finální hodnocení binárních klasifikačních modelů lze využít nejrůznějších způsobů, které je však vždy vztáhnout ke konkrétnímu obchodnímu prostředí. Mezi nejběžnější užívané metody patří tzv. křížová validace, při které je daný model rozdělen na deset podmnožin, z nichž devět slouží k učení souvislostí a jedna potom k měření výkonu. Tento proces se následně mnohokrát opakuje pro nejrůznější kombinace učících a aplikačních podmnožin souboru. Při tomto procesu model predikuje na základě atributů zvolený výstup a porovnává svou predikci se skutečností. Výstupem křížové validace je tzv. klasifikační matice, která udává poměr správných a chybných předpovědí daného modelu. V této matici lze najít tři hlavní ukazatele, nabývající hodnot v intervalu od nuly do jedné (včetně krajních hodnot), které bývají zpravidla vyjadřovány v procentech. Prvním z nich je citlivost modelu, udávající poměr správně předpovězených případů s pozitivním výstupem a všech záznamů s pozitivním výstupem, které se v databázi vyskytují. Druhým porovnávaným ukazatelem je specifičnost modelu, udávající stejný poměr jako citlivost s tím rozdílem, že se týká pouze záznamů s negativním výstupem. Třetím a často nejdůležitějším ukazatelem je potom celková přesnost modelu, která značí souhrnný poměr správných a chybných předpovědí modelu (Eibe, 2005).

Deployment – nasazení modelu

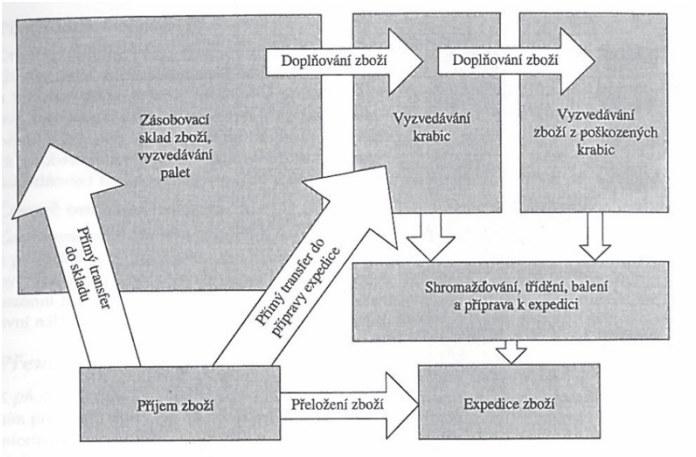
Poslední fáze metodiky CRISP-DM spočívá v konečném nasazení modelu do provozu. Tento úkol vyžaduje například stanovení časového plánu, podle kterého se bude daný prediktivní model v organizaci zavádět, stanovení konkrétních kroků, které je třeba k nasazení vykonat, stanovení způsobu, jakým výsledky modelu komunikovat koncovým uživatelům, či stanovení vhodného způsobu monitorování a měření jeho výsledků a využití.

Pro správnou funkci modelu v praxi je klíčová právě poslední jmenovaná činnost, tedy měření výsledků a kontrola využívání modelu. Stanovením správné strategie údržby a kontroly modelu lze předcházet chybám, které můžou v dlouhodobém horizontu výsledky modelu zkreslovat a snižovat tak jeho hodnotu pro organizaci (Chapman, 2000).

Výše zmíněný postup jednotlivých fází metodiky vychází z Chapmanovy publikace CRISP-DM 1.0, ve které doporučuje pečlivě dodržovat pořadí jednotlivých fází. Oproti tomuto přístupu, Dean Abbot ve své publikaci Applied Predictive Analytics (2014) popisuje alternativu, při které není nutné pořadí těchto fází brát v potaz. Dle Abbota je možné a v některých případech výhodné pořadí fází přeskupit, což v konečném důsledku urychlí celý proces. Abbot navrhuje například začít nejprve s fází vytváření modelu a jeho analýzy pomocí statistického softwaru. Pokud je tato fáze aplikována „naslepo“ před samotnou přípravou dat a porozumění problematice, může dát analytikovi základní přehled a může odhalit určité poznatky na první pohled bez manuálního hodnocení a procházení datového souboru (např. které proměnné budou mít v prediktivním modelu dobrou vypovídací hodnotu). Druhým nestandardním přístupem, který Abbot navrhuje je částečná aplikace požadavků modelu před jeho dokončením. Pokud je například v počáteční fázi zjištěno, jaké proměnné do modelu budou zahrnovány a v jaké podobě musí do modelu vstupovat není nutné čekat na dokončení jeho zbytku, které může trvat další týdny až měsíce. V takovém případě je například vhodné zajistit potřebnou změnu standardu vkládání těchto proměnných do informačních systémů tak aby byly v době aplikace modelu už ve správné podobě.

3.4 Warehouse Management (skladové hospodářství)

Warehouse management neboli skladové hospodářství je nedílnou součástí každého logistického řetězce, který ve své podstatě spojuje výrobce a koncového zákazníka daného produktu. Pojem warehouse management v sobě zahrnuje soubor skladů, jejich rozmístění, kapacitní využití, udržování optimální úrovně zásob, využívané technologie a mnoho dalšího (Lambert,2000).



Obrázek 9 Typické funkce skladování a související toky materiálu. Zdroj: Lambert, 2000.

Skladování může v rámci logistického řetězce plnit několik funkcí, které je možné roztřídit do tří základních kategorií. Těmito kategoriemi jsou přesun produktů, uskladnění produktů a přenos informací (Lambert, 2000).

- Přesun produktů
  - Příjem zboží – jakákoliv činnost spojená s nově naskladňovaným zbožím. Spadá sem vykládání a vybalování zboží, vstupní kontrola kvality a kontrola správnosti dodacích listů oproti skutečnosti.
  - Ukládání zboží – jakékoliv přesuny zboží do skladu, jeho uskladnění na příslušnou pozici, interní přesuny zboží v rámci skladu, či přesun zboží na expedici.
  - Kompletace zboží – třídění, sdružování a montáž zboží z komponentů dle objednávky zákazníka.
  - Překládka zboží (cross-docking) – překlad zboží z místa příjmu přímo do místa expedice (vynechání funkce uskladnění, ukládání i kompletace zboží).

- Uskladnění produktů
  - Přechodné uskladnění – jedná se o podpůrnou funkci k funkci přesun produktů zmiňované výše a zahrnuje pouze uskladnění produktů na dobu nezbytnou pro doplňování základních zásob.
  - Časově omezené uskladnění zásob – týká se nadměrných neboli pojistných, či nárazníkových zásob. Tyto zásoby jsou vytvářeny

k pokrytí např. sezónního výkyvu poptávky, nutné úpravy produktů (dozrávání ovoce), či spekulativních nákupů.

- Přenos informací – třetí hlavní funkcí skladování je přenos informací. Tyto informace jsou klíčové pro management firmy, kterému slouží jako podklad pro rozhodování. Jedná se například o informace o tom, jaký je současný stav zásob, jak zboží prochází skladem, v jakých kvantitách jím prochází, informace o dodávkách, o personálu skladu a mnoho dalších.

V současné době hrají v přenosu skladových informací zásadní roli počítačové systémy. Tyto systémy ve spojení s přenosem informací založeném na elektronické výměně dat (tzv. EDI – Electronic Data Interchange), systémem čtení čárových kódů a dalším technologiím napomáhají ke kontinuálnímu zefektivňování celého procesu. Společně s tímto zefektivňováním přenosu informací by se dle Lamberta (2000) měl management skladu soustředit na minimalizaci, či automatizaci administrativního zatížení pracovníků, které může vést ke vzniku chyb.

Lambert (2000) také poukazuje na neefektivity, které mohou ve skladu vznikat v rámci přesunu produktů, jejich uskladnění, či přenosu informací. Kontrole toho, zda neefektivity ve skladu nevznikají by měl management věnovat zvýšenou pozornost, neboť mohou vést ke vzniku značných nákladů. Konkurenční prostředí, které dnes vyžaduje stále preciznější systémy manipulace, uskladňování, expedice i kontroly kvality zboží takovéto chyby nepřipouští.

Příklady neefektivit ve skladování

- Přebytečná nebo nadměrná manipulace.
- Nízké využití skladovací plochy a prostoru.
- Nadměrné náklady na údržbu v důsledku zastaralosti zařízení.
- Zastaralé způsoby příjmu/expedice zboží.
- Zastaralé informační systémy.

4 Vlastní práce

4.1 Popis společnosti

Společnost EYELEVEL s.r.o., jejíž databáze bude pro účely této diplomové práce využita působí na českém i mezinárodním trhu od roku 2011. Z původně malé, rodinné firmy o cca desíti zaměstnancích s jednou kanceláří v Praze se během několika let rozrostla do rozměrů globálního hráče, který působí prostřednictvím zahraničních poboček po celém světě. Tato firma se zabývá návrhem, výrobou a distribucí nábytku, marketingových kampaní, digitálních technologií a vybavení pro značkové prodejny, které vyžadují jednolitý, unifikovaný koncept a styl nehledě na své umístění. Úspěšné poskytování této služby se EYELEVELU daří především díky rozsáhlé síti dodavatelů (truhlářství, kovovýroba, tiskařství, digitální technologie atd.), síti distribučních center, ve kterých probíhá montáž komponentů včetně jejich následného balení a expedice a síti regionálních poboček, ve kterých projektoví manažeři zajišťují zprostředkování služeb klientům.

Mezi klienty společnosti lze nalézt především globálně působící značky orientující se například na módní (oděvní) průmysl, potravinářství nebo spotřební elektroniku. Od roku 2015 patří EYELEVEL do portfolia americké korporace Innerworkings, která zaštiťuje podobnou škálu služeb jako právě EYELEVEL v ještě rozsáhlejším měřítku.

Diplomová práce se zabývá databází týkající se distribučního centra pražské pobočky společnosti, které vzniklo v roce 2017 v Modleticích u Prahy.

Obrázek 10 Organizační struktura pražské pobočky společnosti EYELEVEL s.r.o. Zdroj: Interní materiály společnosti.

4.2 Porozumění businessové stránce

Data využitá pro tuto diplomovou práci reprezentují vzorek databáze skladových pohybů produktů (v rámci názvosloví společnosti označovaných jako „item“), se kterými společnost operuje. Každý z těchto produktů je vyroben v rámci dodavatelské sítě společnosti, přijat na sklad, zaevidován, uskladněn a v případě potřeby vyskladněn. Vyskladnění itemu probíhá buď do oblasti montáže, kde je spotřebován jako součástka, či do oblasti balení a expedice odkud je odeslán k uspokojení objednávky klienta. Pohyb itemu v rámci skladu je evidován s využitím systému čárových kódů, jejichž informace jsou pomocí čteček načteny a odeslány do informačního systému. Informační systém Infor M3, který EYELEVEL využívá, spadá do kategorie ERP (Enterprise Resource Planning) a pomáhá společnosti efektivně řídit veškeré její aktivity.

Databáze obsahuje kromě záznamu o pohybu itemu také jeho evidenční číslo, typ transakce (zda se jedná pouze o přemístění v rámci skladu bez spotřeby, či vyskladnění do montáže/expedice se spotřebou), čas a místo vyskladnění, či identifikační číslo klientské objednávky, pro kterou byl daný pohyb iniciován. Tato skladová databáze byla pro účely diplomové práce zkombinována s produkční databází, obsahující veškeré atributy daného itemu, tedy jeho identifikační číslo, materiálové složení, cenu, identifikaci projektu, informace o klientovi, pro kterého je item určený, rozměry a další.

V současné době dochází při výrobě a naskladňování nového typu produktu k jeho uložení do aktuálně volné pozice v rámci rozdělení skladu dle oblastí určených pro jednotlivé klienty. Data miningová úloha a prediktivní model zpracovaný v této diplomové práci by měli odhalit, zda existuje spojitost mezi atributy produktu, které jsou známé už při jeho výrobě u dodavatele a počtem skladových pohybů, kterými bude následně daný výrobek procházet. Ve své podstatě bude prediktivní model na základě atributů produktu určovat, jak často bude tento nový výrobek vyskladňován do oblasti montáže, či expedice. Na základě této informace bude možné zvolit pro daný výrobek skladovou pozici tak, aby byl blíže, či naopak dále od těchto míst. Tento krok pomůže společnosti ušetřit čas, který pracovníci stráví převážením produktů mezi lokacemi, ušetřit energii spotřebovanou manipulační technikou a zefektivnit celý proces řízení zásob.

Cílem této data miningové úlohy je zpracování tří prediktivních modelů, které budou schopné určit, zda daný výrobek bude, či nebude rychloobrátkový a na základě toho vybrat vhodnou lokaci pro jeho uskladnění. K tomuto účelu budou sestaveny tři prediktivní modely vhodné k binomické klasifikaci (logistická regrese, rozhodovací strom, neuronová síť).

V tomto projektu bude postupováno dle metodiky CRISP-DM, která byla představena v rámci literární rešerše. V první fázi prvního kroku metodiky je třeba definovat obchodní cíle modelu z manažerského pohledu. Obchodní cíl tohoto projektu spočívá v analýze dostupných dat, na jejímž základě bude možné kategorizovat nově vyráběné produkty tak, aby bylo možné vybrat efektivní skladovou lokaci pro jejich uložení, což povede ke snížení nákladů společnosti a ušetření času při skladové manipulaci. V rámci tohoto projektu budou také určeny faktory, které mají z hlediska kategorizace produktů nejvyšší váhu při rozhodování, do které kategorie bude nový výrobek zařazen.

Druhý krok první fáze metodiky spočívá v definici potřebných a dostupných prostředků k realizaci projektu a rizik, která mohou nastat.

Zdroje, které jsou k realizaci projektu dostupné jsou následující:

- Databáze skladových pohybů produktů.
- Doprovodný soubor identifikující proměnné obsažené v databázi skladových pohybů.
- Databáze produktů a jejich atributů.
- Informační systém společnosti EYELEVEL (Infor M3).
- MS Excel.
- PC Microsoft Windows 7 64bit, AMD A8 3,30 GHz, 16 GB RAM
- Rapid Miner Studio 9.2

Při zpracovávání této data miningové úlohy může nastat několik rizik. V první řadě zde existuje riziko, že dostupná data nebudou mít pro model dostatečnou vypovídací

hodnotu. Tato situace může plynout například z toho, že kombinované databáze budou obsahovat příliš mnoho chybějících záznamů a výsledek tudíž bude zkreslený a nepoužitelný pro nově naskladňované itemy. Druhým rizikem může být fakt, že ani jeden ze zpracovaných modelů nebude dostatečně přesný. Pro daný obchodní případ byla stanovena hranice přesnosti modelu alespoň 80 %. Třetí možné riziko spočívá v nutnosti zachovat současnou podobu sbíraných dat. Pokud se společnost například v blízké době rozhodne, že změní ERP systém (Infor M3) na nějaký jiný, který pracuje na odlišném principu (např. SAP), nebude model možné využívat. Poslední možné riziko spočívá možnosti nedostatečné profitability projektu. Pokud bude například zjištěno, že náklady na reorganizaci skladu předčí možné úspory, ukáže se model jako nepotřebný a nebude tudíž vhodné ho implementovat.

Ve třetím kroku první fáze metodiky CRISP-DM je třeba stanovit dílčí úkoly, vedoucí ke splnění hlavního obchodního cíle projektu. V tomto kroku je také nutné určit metriku, která stanoví, zda byl daný projekt úspěšný, či nikoliv.

Jak již bylo zmíněno výše, hlavním obchodním cílem tohoto projektu je co nejpřesnější určení počtu budoucích skladových pohybů daného produktu, na základě analýzy dostupných dat.

Dílčí úkoly, které povedou ke splnění hlavního obchodního cíle projektu jsou následující:

- Získání databáze skladových pohybů.
- Získání databáze produktů a jejich atributů.
- Příprava dat do vhodné formy.
- Sestavení modelu č. 1 – logistická regrese.
- Ověření a vyhodnocení modelu č. 1 – logistická regrese.
- Sestavení modelu č. 2 – rozhodovací strom.
- Ověření a vyhodnocení modelu č. 2 – rozhodovací strom.
- Sestavení modelu č. 3 – neuronová síť.
- Ověření a vyhodnocení modelu č. 2 – neuronová síť.
- Výběr nejlepšího modelu a jeho hypotetická aplikace.

- Analýza výsledků a návrh opatření.
- Kalkulace možných nákladů a úspor daného opatření.

4.3 Porozumění datům

Druhou fází metodiky CRISP-DM je fáze porozumění datům, kterou lze standardně rozdělit na několik dílčích kroků. V první řadě je nutné získat samotná data, která budou pro následnou analýzu použita. Veškerá data použita v této práci vychází ze tří databází, které byly poskytnuty společností EYELEVEL. Těmito databázemi jsou MITTRA data.csv a ItemBasics.xlsx a metadata.xlsx. Po domluvě s vedením společnosti byla kvůli zachování obchodního tajemství některá data zašifrována a jména klientů a dodavatelů nahrazena generickým identifikátorem. Vzhledem ke značné rozsáhlosti databází, omezení softwaru MS Excel pouze pro 1 048 576 záznamů a omezení možností vizualizací softwaru Rapid Miner pro 1 000 000 záznamů budou některé činnosti, které by byly za normálních okolností provedeny ve fázi porozumění datům provedeny až ve fázi přípravy dat (po manuálním vytvoření agregované finální databáze, která bude pro samotné modelování využita).

4.3.1 MITTRA data

První databáze obsahuje záznamy o veškerých skladových pohybech jednotlivých itemů v distribučním centru Modletice od 1.1.2018 do 19.12.2018. Tato databáze byla získána jako .csv soubor čítající 1 171 335 záznamů a 81 proměnných. Veškeré proměnné obsažené v této databázi jsou pojmenované šestimístným systémovým kódem, pro jehož rozklíčování slouží soubor metadata.xlsx popsaný níže. Tento soubor obsahuje značné množství proměnných i záznamů které pro tento projekt nejsou relevantní a budou muset být v další fázi metodiky odstraněny (například systémové identifikátory transakcí, identifikátory společnosti, které nabývají pouze jedné hodnoty, či nevyužité proměnné obsahující pouze prázdné hodnoty). Tato databáze naopak postrádá proměnné týkající se atributů jednotlivých itemů, které budou na základě jejich identifikačního čísla doplněny v další fázi metodiky. Pro získání agregovaného počtu skladových pohybů daného itemu bude nutné převést současnou databázi do podoby kontingenční tabulky.

4.3.2 Item Basics

Druhá využitá databáze obsahuje záznamy o vlastnostech všech produktů, které společnost EYELEVEL za dobu svého působení na trhu nakoupila, prodala, či zpracovala. Tato databáze byla získána v podobě reportu z informačního systému Infor M3 a obsahuje 93599 záznamů a 14 proměnných. Tato databáze, stejně jako databáze předchozí obsahuje některé proměnné, které nejsou pro tento projekt relevantní, či vykazují na první pohled zjevnou multikolinearitu (například číselné označení výrobce a jeho jméno vyjádřené slovy) a budou z ní odmazány. Vlastnosti produktů, které jsou v této databázi obsaženy budou na základě jejich identifikačních čísel přiřazeny příslušným položkám v databázi MITTRA.

4.3.3 Metadata

Posledním zdrojem vstupů do tohoto projektu je pomocná databáze obsahující identifikační pole odpovídající kódům proměnných obsažených v záhlaví souboru MITTRA.csv. Tato databáze byla získána ze SQL databáze společnosti a obsahuje pouze 81 textových identifikátorů proměnných, údaje o datovém typu dané proměnné, počet znaků, jímž je proměnná vyjádřená, počet desetinných míst dané numerické proměnné atd. Identifikační popis proměnných bude na základě systémových kódů přiřazen do finální databáze, kde nahradí kódové označení pro snadnější orientaci čtenáře.

4.4 Příprava dat

Třetí fází metodiky CRISP-DM je fáze přípravy dat, která je pro úspěch celého projektu klíčová. Jak již bylo zmíněno ve fázi předchozí, je nyní třeba agregovat tři zdroje dat do jedné dvojrozměrné tabulky, která bude obsahovat pouze proměnné relevantní pro samotné modelování. Po promazání proměnných, které nejsou pro projekt důležité dojde také k otestování zbylých proměnných na výskyt multikolinearity, výskyt chybějících hodnot, chyb, či jiných nedostatků.

4.4.1 Agregování databází

Pro účely této diplomové práce je třeba v první řadě agregovat tři výše zmíněné databáze. Jako základní soubor slouží databáze MITTRA data.csv, který je převeden do dvojrozměrné tabulky .xlsx. Kvůli omezenému rozsahu sešitu softwaru MS Excel, který bude pro úpravu této databáze využit, bude původních 1 171 335 záznamů zredukováno na 1 048 576, tedy maximální možný rozsah, který tento formát tabulky dovoluje. V druhém kroku bude k systémovým označením jednotlivých sloupců (proměnných) přiřazena jejich slovní identifikace ze souboru metadata.xlsx, což napomůže ke snadnější čitelnosti databáze.

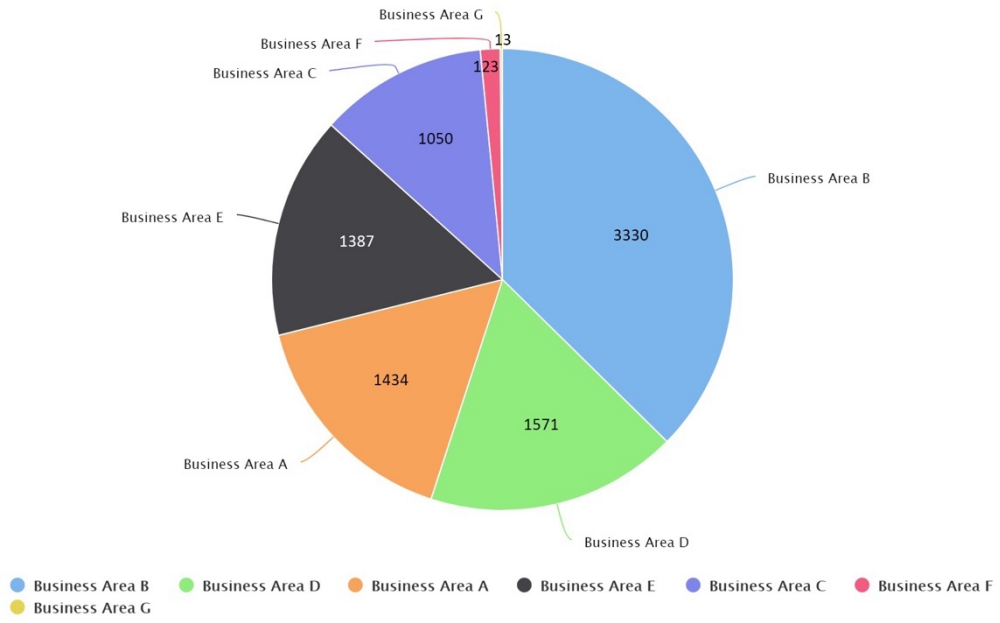
Následující krok spočívá v přiřazení atributů ze souboru Item basics.xlsx k jednotlivým záznamům databáze MITTRA data. Tohoto lze snadno dosáhnout s využitím vyhledávacích funkcí softwaru MS Excel (VLOOKUP, či kombinace funkcí INDEX a MATCH). Jako společný prvek obou databází, který bude sloužit pro přiřazování bude využita proměnná Item Number, reprezentující osmimístné číselné označení daného produktu. Před samotným přiřazováním bylo kvůli funkcionalitě vyhledávacích funkcí nutné ze všech polí proměnné Item Number databáze Item basics odmazat čtyři prázdné znaky. Tato chyba vznikla v důsledku nesprávného naprogramování reportu (pole Item Number je v informačním systému nastaveno na 12 znaků, nicméně vyplněno je pouze 8. Kvůli tomuto rozdílu systém nahradí nevyužité znaky mezerami (prázdnými znaky).

Po samotném sloučení databází je k dispozici dvojrozměrná tabulka, které však obsahuje pro každý produkt značné množství záznamů. Tyto nadbytečné záznamy označují i systémové operace, které se daného itemu týkají, a nejen fyzické skladové pohyby. V důsledku toho byla databáze pomocí proměnné „Order Type“ vyfiltrována na 77109 záznamů, které se týkají pouze vyskladňování zboží do konkrétních zakázek. Z této databáze byla následně zhotovena kontingenční tabulka čítající 8909 záznamů, přiřazující každému produktu jeho historickou četnost skladových pohybů, což bude v tomto projektu predikovaná proměnná. Tento soubor bude pro další práci považován za základní.

4.4.2 Základní informace o souboru

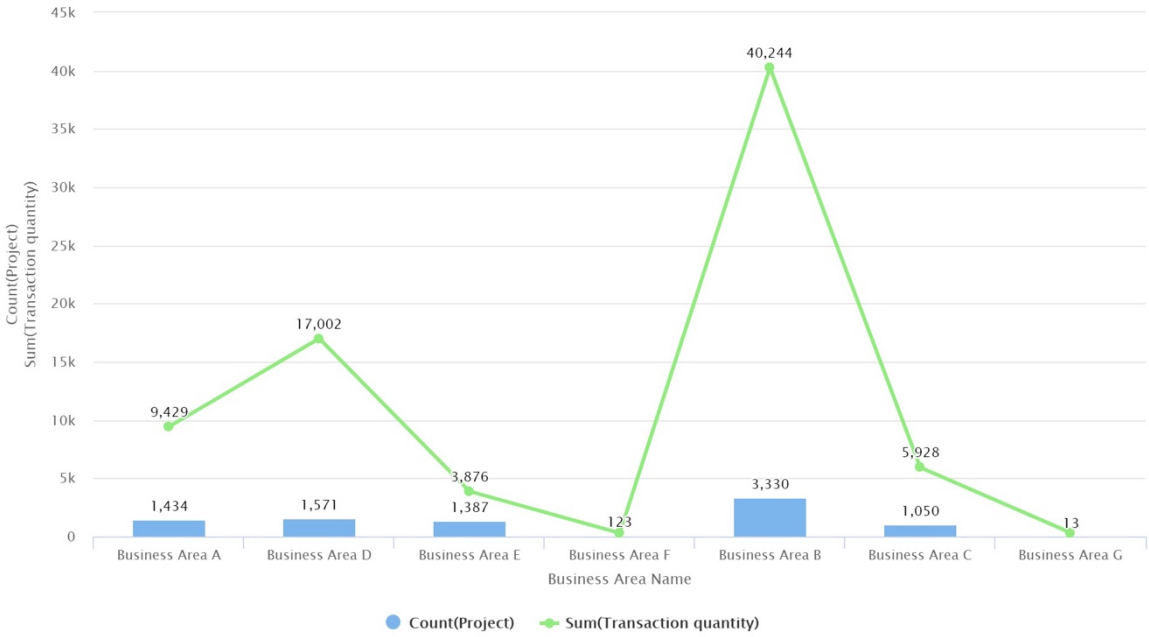
Finální datový soubor, využitý pro tento projekt obsahuje informace o vzorku 8909 produktů společnosti a jejich 29 atributech. Tento soubor je zpracován statistickým softwarem Rapid Miner Studio 9.2, který nabízí přehledný souhrn všech proměnných, jejich základních charakteristik (krajní hodnoty, průměr atd.) a možnost vizualizace. Pro následující popis souboru byly vybrány následující proměnné, které se z hlediska porozumění obchodní stránky projektu jeví jako nejdůležitější.

- Business Area – jméno klienta, pod kterého výrobek spadá.
- Transaction Quantity – počet skladových pohybů.
- Project – projekt, pod který daný výrobek spadá.
- Item Type – popis produktu z hlediska kompletace a vlastnictví.
- Number of Materials – počet komponentů, ze kterých je výrobek složený.
- Item Price – nákupní cena výrobku.
- Supplier Name – jméno dodavatele daného produktu.



Obrázek 11 Rozdělení souboru dle klientů. Zdroj: Vlastní zpracování.

Každý produkt, který společnost EYELEVEL zařadí do svého systému je vytvořen na základě přání, či objednávky svých klientů. Tito klienti jsou v informačním systému vedeni pod kódovým označením v rámci proměnné Business Area. Na obrázku číslo 11 je názorně vidět rozložení počtu produktů, které skladem procházejí dle jednotlivých klientů. Jména těchto klientů byla dle dohody s vedením společnosti skryta pomocí generické kategorizace. Z obrázku je vidět, že největší počet produktů spadá pod klienta Business Area B. Business Area A, D, E a C zaujímají přibližně stejný podíl a klienti Business Area G a F jsou pro společnost z hlediska počtu produktů téměř nevýznamní.

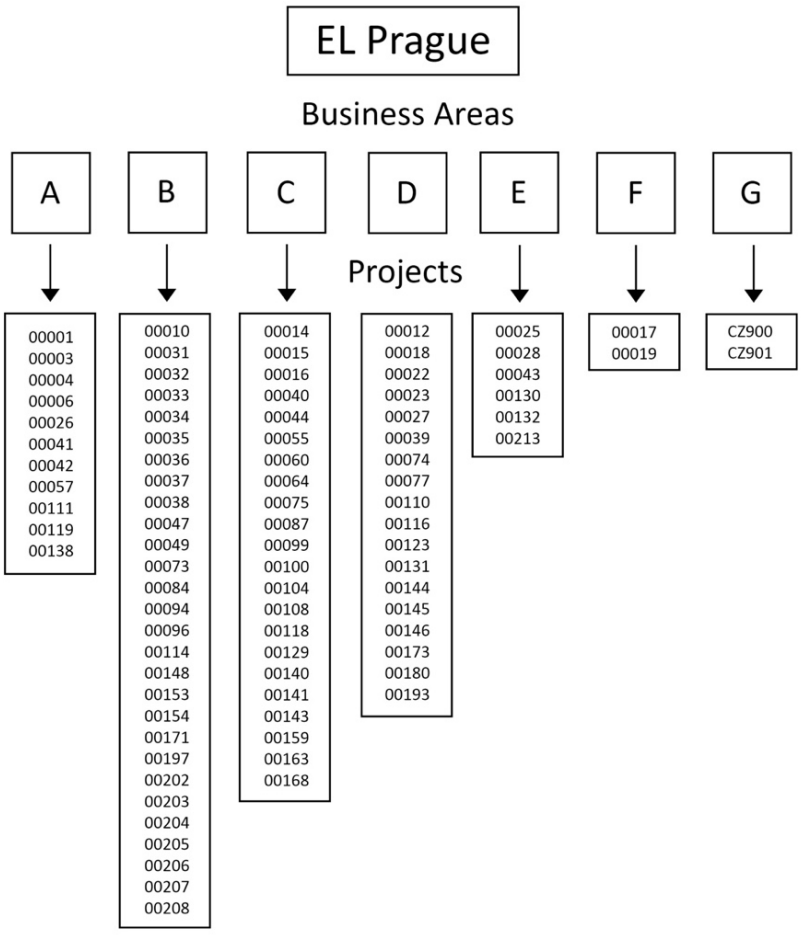


Obrázek 12 Rozdělení souboru dle klientů v porovnání s četností skladových pohybů. Zdroj: Vlastní zpracování.

Obrázek č. 12 znázorňuje rozdělení základního souboru dle klientů v porovnání s četností skladových pohybů produktů, které do jejich kategorie spadají. ose X jsou zobrazeni jednotliví klienti společnosti a osa Y slouží pro znázornění počtu produktů a zároveň počtu jejich skladových pohybů. Z obrázku je patrné, že sekce klienta Business Area B značně převyšuje ostatní klienty nejen v počtu produktů, ale také v počtu jejich skladových pohybů. Další v pořadí je potom Business Area D, která však svými hodnotami dosahuje zhruba do poloviny hodnot nejsilnějšího klienta.

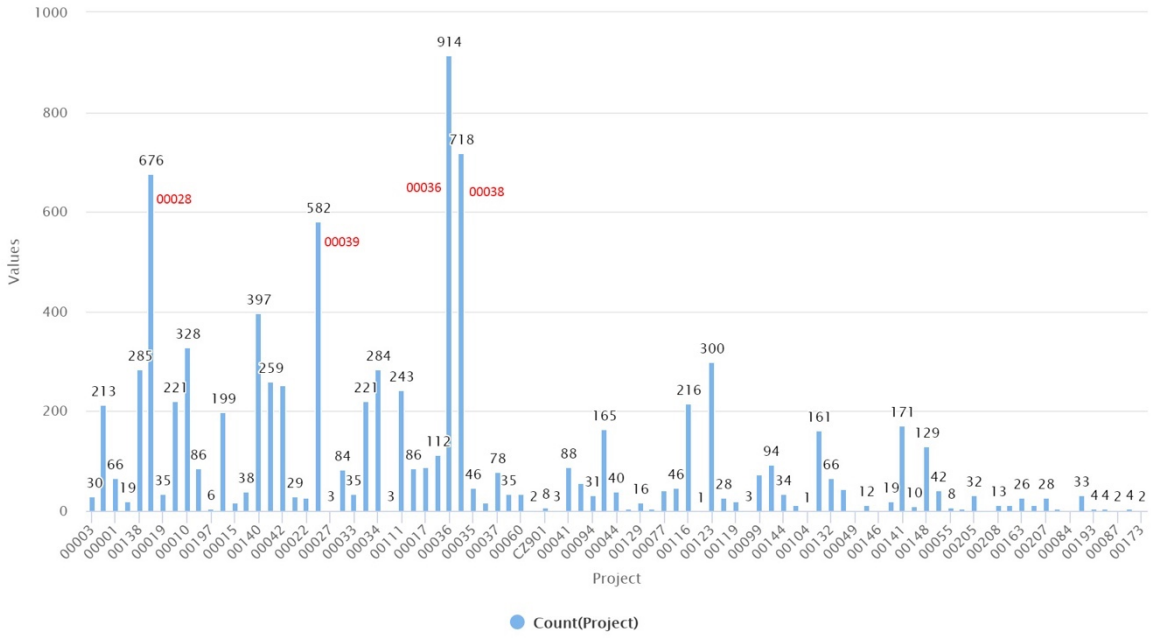
Tento trend zcela odpovídá očekáváním. V této fázi je nutné podotknout, že do Business Area B spadá klient, pro kterého jsou kromě standardní výroby realizovány i marketingové kampaně. Tyto kampaně se z hlediska skladového hospodářství vyznačují především tím, že je pro každou z nich vyrobeno značné množství jednorázových produktů, které do skladu přicházejí i odcházejí v poměrně krátkém čase. Kampaně se v různě dlouhých intervalech opakují a signifikantně tak navyšují počet produktů v evidenci, čemuž následně odpovídá i počet skladových pohybů.

Jako další kategorie třídění dat slouží proměnná Project, která udává číselné označení projektu, pod kterým byl daný produkt založen. V rámci systémové struktury společnosti označuje pojem „projekt“ podsložku klienta, založenou pro zpřehlednění a odlišení různých druhů výrobků. V praxi se může jednat například o odlišení odběratelských regionů, či různých konceptů nábytku, který je ale určen pro jednoho zákazníka. Struktura a přehled aktivních projektů spadající pod jednotlivé Business Areas je znázorněna na obrázku číslo 13.

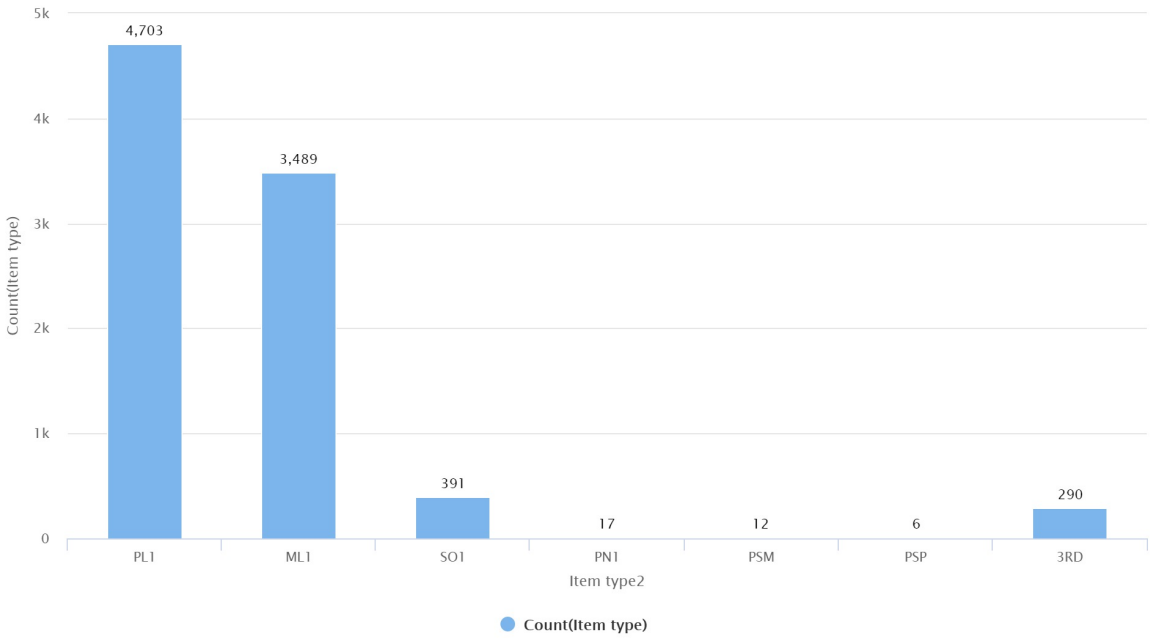


Obrázek 13 Struktura jednotlivých klientů a pod ně spadajících projektů. Zdroj: Vlastní zpracování.

Obrázek číslo 14 znázorňuje rozdělení základního souboru dle jednotlivých projektů. Na ose X je znázorněn počet záznamů (produktů) v databázi a na ose Y čísla jednotlivých projektů. Na obrázku lze vidět, že struktura rozdělení odpovídá rozdělení souboru dle jednotlivých klientů. Vysoká čísla jsou zde vidět zejména u projektů 00036, 00038, které spadají do Business Area B a oba se týkají výše zmíněných marketingových kampaní. Poměrně vysoko dosahují také projekty 00028 a 00039, které odpovídají standardní produkci.



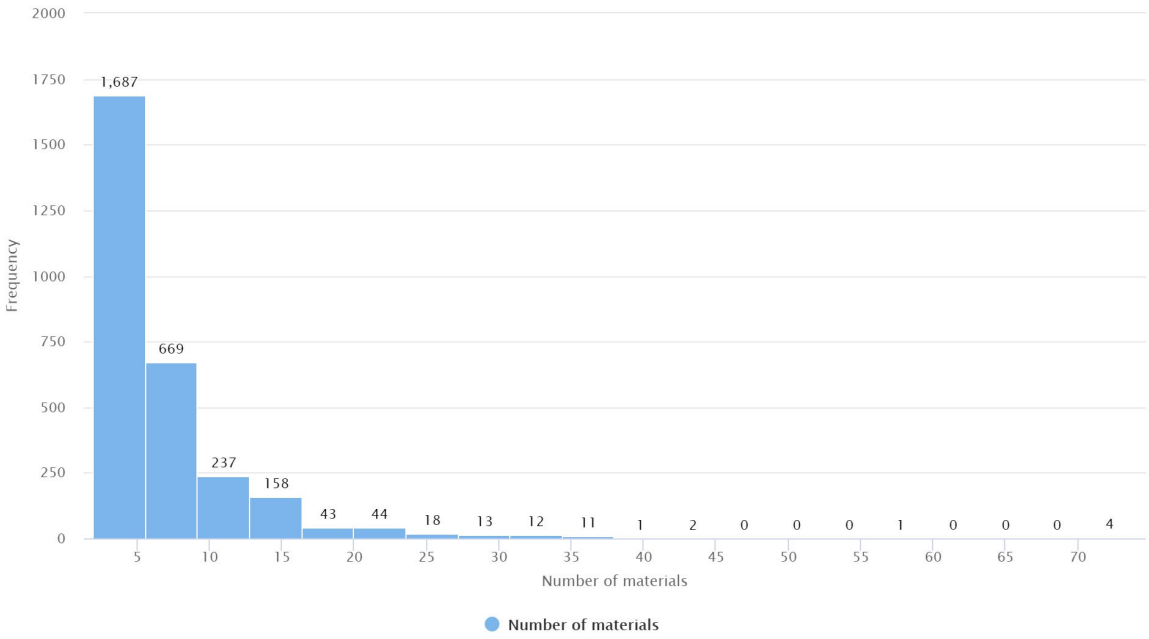
Obrázek 14 Kategorizace souboru dle jednotlivých projektů. Zdroj: Vlastní zpracování.



Obrázek 15 Rozdělení produktů dle typu a vlastnictví. Zdroj: Vlastní zpracování.

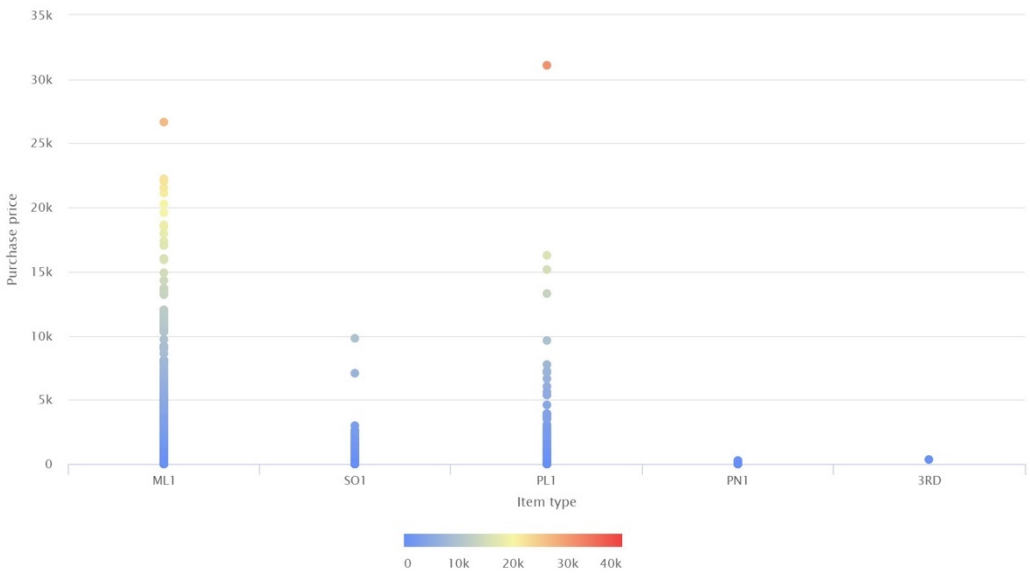
Další možností kategorizace produktů je jejich rozdělení dle typu. Systém společnosti rozlišuje čtyři hlavní kategorie produktů. První kategorií jsou itemy označované jako Purchased Lot (PL1). Tyto produkty jsou nakupovány od dodavatelů a nevyžadují žádnou další kompletaci. Je možné je využít dvěma způsoby. Buď jsou zabaleny a prodány ve stavu v jakém byly poskytnuty od dodavatele, nebo jsou spotřebovány jakožto komponenty do produktů označených jako ML1. Tyto produkty (tzv. Manufactured Lot – ML1) na rozdíl od prvních jmenovaných vyžadují kompletaci ze dvou a více komponentů. Třetí kategorií jsou produkty označované jako Subcontracting Order (SO1), které musí projít určitou úpravou u externího dodavatele (například potisk, či protipožární opatření). Poslední důležitou kategorií jsou tzv. Third-Party Items (3RD), což jsou produkty, které jsou ve valné většině ve vlastnictví klienta, kterému jsou za poplatek distribuovány společně se standardními výrobky. Zbylé kategorie zobrazené na obrázku č. 15 slouží např. pro systémovou klasifikaci, či interní spotřebu a jejich výskyt je minimální. Jak je vidět na obrázku číslo 15, značné množství produktů v databázi je označeno jako PL1 – tedy produkty, které je možné rovnou prodat, nebo spotřebovat pro výrobu pro výrobu složitějších manufactured itemů.

V Souvislosti s kategorizací produktů dle jejich typu je možné vizualizovat rozdělení produktů z hlediska počtu komponentů, ze kterých se kompletují. Pro zhotovení následující vizualizace byly z datového souboru vybrány pouze produkty typu ML1, které mají dva a více komponentů. U ostatních itemů se automaticky předpokládá, že mají komponent pouze jeden.



Obrázek 16 Histogram rozdělení Manufactured itemů dle počtu komponentů. Zdroj: Vlastní zpracování.

Na histogramu (obrázek č. 16) je znázorněn počet komponentů (osa X) daného produktu a četnost výskytu tohoto počtu v datovém souboru (osa Y). Z obrázku je vidět, že valná většina produktů se skládá z šesti a méně komponentů a více než 80 % všech produktů se skládá z méně než desíti komponentů. Maximální počet komponentů dávajících dohromady jeden manufactured produkt je 74, minimální počet jsou 2 (dle kritérií filtrace databáze) a průměrná hodnota je 6,55 komponentu na jeden složený produkt.



Obrázek 17 Scatter diagram rozdělení nákupních cen produktů dle jejich typu. Zdroj: Vlastní zpracování.

Na scatter diagramu (obrázek č. 17) je znázorněno rozdělení nákupních cen jednotlivých itemů dle jejich typu. Z obrázku je názorně vidět, že většina produktů, které společnost zpracovává má nákupní hodnotu do 5000 Kč. U itemů ML1, které se skládají z více komponentů jsou ceny zpravidla vyšší. Naopak u 3RD party itemů je vždy hodnota vždy vyjádřena pouze symbolicky (např. 1 Kč), protože její původní, nákupní cenu platil klient. Minimální cena za produkt v datovém souboru je tedy 1 Kč, maximální je 31 094,45 Kč a průměrná 1076,55 Kč. Pro tuto proměnnou je nutné zdůraznit, že soubor obsahuje 4087 chybějících hodnot a 1022 záznamů obsahuje nulovou hodnotu. Diagram vyobrazený výše znázorňuje zbylých 3799 záznamů.

Obrázek 18 Wordcloudová kategorizace produktů dle dodavatele. Zdroj: Vlastní zpracování.

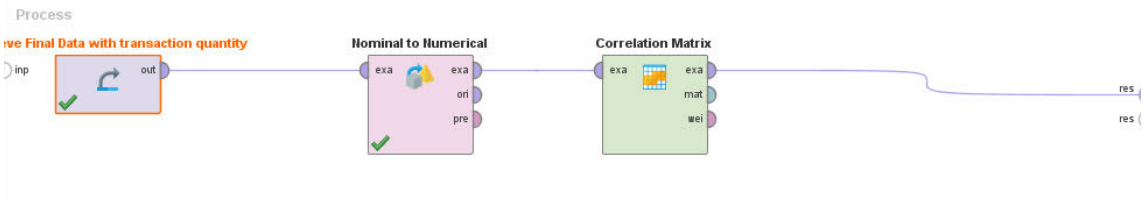
Poslední důležitá kategorizace produktů spočívá v jejich rozdělení dle dodavatele. Vzhledem ke značnému množství záznamů a unikátních výrobců v databázi je pro získání představy o souboru nejvhodnější vizualizovat data pomocí tzv. wordcloudu. Tato vizualizace analyzuje všechny záznamy databáze a následně vytvoří shluk slov, která svou velikostí odpovídají své četnosti výskytu v databázi. Z celkových 8909 záznamů obsahuje databáze 3211 záznamů označení výrobce „EYELEVEL“ – což znamená, že je výrobek montován z komponentů různých výrobců přímo ve skladu. Soubor také obsahuje 199 chybějících hodnot, které se do databáze dostaly zřejmě kvůli nepozornosti pracovníků, kteří daný produkt systémově zakládali.

Jak je vidět na obrázku č. 18, z hlediska kvantity itemů v databázi jsou největšími dodavateli společnosti firmy Dakacom.cz s.r.o, Astron Print s.r.o. a PrintPoint s.r.o. Tento fakt je opět v souladu s očekáváním. Všechny 3 jmenované firmy se zabývají tiskem a výrobou POS (Point of Sale) materiálů, které jsou prodávány společností EYELEVEL v rámci marketingových kampaní.

4.4.3 Testování a výběr proměnných

V dalším kroku této fáze metodiky CRISP-DM je třeba datový soubor připravit do konečné podoby pro modelování. Dle Abbota (2014) je v tomto kroku vhodné opakovaně projít všechny proměnné, zhodnotit jejich relevantnost k projektu a otestovat je na výskyt multikolinearity, chybějících hodnot, či jiných nežádoucích vlivů.

Jako první je třeba otestovat proměnné na výskyt multikolinearity, která je u některých atributů očekávaná. Odstraňování atributů, které vykazují vzájemnou korelaci (zpravidla více než 0,9, či méně než – 0,9) je žádoucí kvůli zjednodušení modelu a snížení výpočetního času. Pokud se v souboru vyskytují hodnoty s vysokým korelačním koeficientem, není vhodné je všechny do modelu zahrnovat. Tyto korelované atributy nepřinášejí do modelu žádnou další vypovídací hodnotu a zpravidla se jedná pouze o jiné vyjádření dané proměnné. Korelované proměnné byly odhaleny pomocí procesu software Rapid Miner, jehož schéma je zobrazeno na obrázku č. 20. Kroky tohoto procesu obsahují načtení souboru, převedení polynominálních proměnných na numerické a vytvoření rozsáhlé korelační matice, ze které byly výsledky vyčteny.



Obrázek 19 Schéma procesu generace korelační matice. Zdroj: Rapid Miner.

Ve zkoumaném souboru bylo vzájemně korelujících proměnných objeveno 11 a jsou vypsány v tabulce č. 2. Výskyt takovýchto proměnných je ve zkoumaném případě způsoben kombinací několika zdrojů informací, do kterých data vstupují z informačního systému společnosti. Do tohoto systému je například proměnná Responsible vyplňována pouze do jednoho pole, které se ale v rámci procesů propisuje do dalších oblastí systému a vystupuje tak v reportech samostatně.

First Attribute	Second Attribute	Correlation
Responsible	RESP	1,00
Item Type	Item Type2	1,00
Supplier ID	Supplier Name	1,00
Supplier ID	Supplier No	1,00
Supplier No	Supplier Name	1,00
Number of Materials	Count of Components	1,00
Business Area	Business Area Name	1,00
Item Type	Structure	0,93
Item Type2	Structure	0,93

Tabulka 2 Přehled korelovaných proměnných a jejich korelační koeficientů. Zdroj: Vlastní zpracování.

V rámci redukce počtu atributů budou tedy před vytvářením každého modelu odstraněny korelované atributy. Výběr těchto atributů bude probíhat na základě znalosti prostředí a snahy o maximálně uživatelsky přívětivý model. Pokud bude rozhodováno například o odstranění proměnných Supplier ID nebo Supplier Name, bude pro budoucí model vhodnější odstranit proměnnou Supplier Name a ponechat pouze tří až pětimístné číselné označení dodavatele. Pokud by byla odstraněna proměnná Supplier ID a ponechána druhá jmenovaná, bylo by komplikované do modelu vždy vypisovat slovní označení dodavatele, které by muselo stoprocentně souhlasit se záznamy v databázi (tzv. exact match – stejný počet znaků, pozice mezer i speciálních symbolů).

Z hlediska vzájemné korelace tedy budou z modelu odstraněny následující proměnné.

- RESP (korelační koeficient 1,0).
- Item Type2 (korelační koeficient 1,0).
- Supplier Name (korelační koeficient 1,0).
- Supplier No (korelační koeficient 1,0).
- Count of components (korelační koeficient 1,0).
- Business Area Name (korelační koeficient 1,0).
- Structure (korelační koeficient 0,93).

Z databáze budou dále odstraněny proměnné, jejichž kvalita není dostatečná pro vytváření modelu. Pro otestování všech proměnných bude využit nástroj softwaru Rapid Miner, který je součástí jeho funkce automatického modelování (Auto Model). Tento nástroj projde všechny proměnné v databázi a vyhodnotí jejich kvalitu na základě následujících čtyř kritérií.

- Correlation (C) – korelace s vysvětlovanou proměnnou.
- ID-Ness (I) – počet unikátních hodnot.
- Stability (S) – počet stejných hodnot.
- Missing (M) – počet chybějících hodnot.
- Text-ness (T) – výskyt textových polí místo hodnot.

Všechna tato kritéria software Rapid Miner vyhodnotí a určí, zda je atribut vhodný pro zahrnutí do modelu. Podoba, v jaké Rapid Miner výsledky zobrazuje je znázorněna na obrázku č. 20. Jednotlivé sloupce indikují míru výše zmíněných kritérií u daného atributu a barevný ukazatel potom kvalitu dat pro modelování (zelená = vhodný atribut, žlutá = vhodný atribut s výhradami, červená = nevhodný atribut).

Obrázek 20 Ukazatel vhodnosti atributu pro modelování. Zdroj: Rapid Miner.

Vhodnost jednotlivých atributů zdrojové databáze z hlediska kritérií zmíněných výše je znázorněna v tabulce číslo 3. Tento test slouží pouze k obecnému zhodnocení atributů a je možné, že se jejich vhodnost, či nevhodnost pro konkrétní model změní.

Quality	Attribute Name	Correlation	ID-ness	Stability	Missing	Text-ness
	ITEM NUMBER	0,06%	100,00%	0,01%	0,00%	36,89%
	Extended code	0,06%	99,99%	0,02%	0,00%	52,61%
	Status	0,00%	0,06%	94,86%	0,00%	0,91%
	Description	0,06%	95,84%	0,08%	0,04%	79,78%
	Project	0,16%	1,00%	10,26%	0,00%	2,56%
	Responsible	0,49%	0,53%	12,97%	0,00%	4,62%
	Item type	0,59%	0,08%	52,80%	0,00%	1,36%
	Salable	0,06%	0,02%	99,20%	0,00%	1,34%
	Supplier ID	0,07%	1,99%	38,28%	0,00%	2,54%
	Supplier No 2	0,00%	0,10%	100,00%	0,00%	0,45%
	Gross Weight	0,30%	0,00%	21,49%	0,15%	0,00%
	Net Weight	0,46%	0,00%	6,79%	0,00%	0,00%
	Length	0,25%	0,00%	31,63%	38,00%	0,00%
	Height	0,03%	0,00%	31,70%	38,00%	0,00%
	Width	0,08%	0,00%	31,54%	38,00%	0,00%
	Structure status	0,52%	0,01%	100,00%	60,87%	0,90%
	Number of materials	0,05%	0,47%	66,30%	0,00%	0,00%
	Number of operations	0,00%	0,04%	99,67%	0,00%	0,00%
	Purchase price	0,13%	0,00%	21,20%	45,88%	0,00%
	Business Area	0,17%	0,08%	37,38%	0,00%	1,36%
	Runtime	0,01%	0,00%	50,89%	74,01%	0,00%

Tabulka 3 Vhodnost jednotlivých atributů pro modelování. Zdroj: Vlastní zpracování.

Přehled vyřazených proměnných s krátkým zdůvodněním je znázorněn v tabulce číslo 4. Zbylé proměnné se zdají být v pořádku a mohou být pro vytváření modelu použity.

Quality	Attribute Name	Důvod vyřazení
	ITEM NUMBER	Každý záznam je unikátní
	Extended code	Každý záznam je unikátní
	Status	Identická hodnota ve všech záznamech
	Description	Každý záznam je unikátní
	Salable	Identická hodnota ve všech záznamech
	Supplier No 2	Identická hodnota ve všech záznamech
	Structure status	Identická hodnota ve všech záznamech
	Number of operations	Identická hodnota ve všech záznamech
	Runtime	Příliš mnoho chybějících hodnot

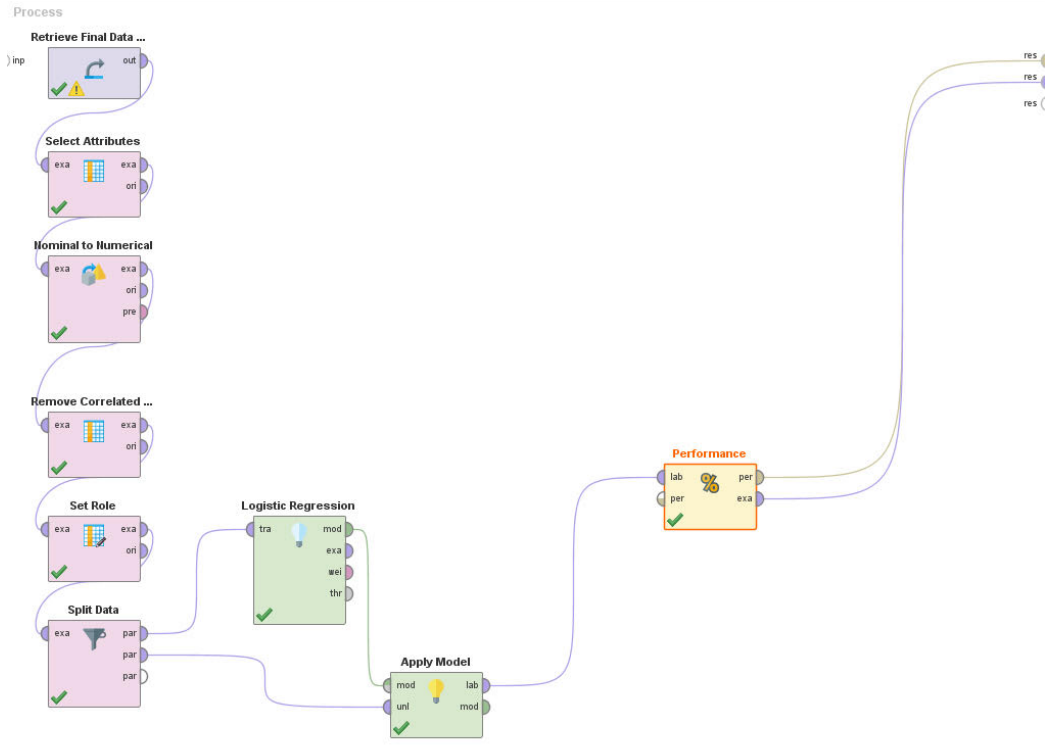
Tabulka 4 Seznam vyřazených proměnných a důvod jejich vyřazení. Zdroj: Vlastní zpracování.

4.5 Modelování

Další fáze metodiky CRISP-DM spočívá ve vytváření jednotlivých prediktivních modelů, které budou následně vzájemně porovnány z hlediska přesnosti a užitečnosti v daném obchodním prostředí. V této fázi metodiky budou vytvořeny tři prediktivní modely, a to logistická regrese, rozhodovací strom a umělá neuronová síť. K modelování bude využit software Rapid Miner, který pro vytváření jednotlivých modelů nabízí všechny potřebné moduly (operators). Tyto moduly je třeba zkombinovat s nejrůznějšími dalšími operátory tak, aby společně vytvořili sled kroků (proces), který postupně finální model vytvoří.

4.5.1 Logistická regrese

Jako první byl pro modelování zvolen model logistické regrese, který je hojně využívaný pro predikci binominální kategorizace vstupu na základě hodnot jeho atributů. Cílem tohoto modelu je vytvořit proces (viz. obrázek č. 21), který vybere vhodné atributy, vypočítá logistickou regresi a otestuje kvalitu a přesnost jejích výstupů.



Obrázek 21 Výsledný proces logistické regrese. Zdroj: Rapid Miner.

Po založení nového procesu je třeba do Rapid Mineru importovat data a upravit jejich typ tak, aby odpovídal potřebám modelu. Tuto úpravu lze provést v rámci nástroje pro import, který uživatele provádí výběrem souboru, odpovídajících sloupců, formátováním a stanovováním datových typů. Ve zkoumaném souboru byly v tomto kroku provedeny následující úpravy.

- Fast Pick Location – tato proměnná byla softwarem vyhodnocena jako polynominální, nicméně obsahuje pouze dvě hodnoty. Z tohoto důvodu byl datový typ změněn na binominální.
- Project – proměnná byla vyhodnocena jako integer (celé číslo), ale číslo projektu v tomto případě odpovídá jeho kategorii. Proto byla proměnná změněna na polynominální.

Po importu vstupních dat je třeba ze souboru odstranit proměnné, jejichž hodnoty nejsou pro modelování vhodné. Pomocí modulu Select Attributes byly zvoleny proměnné Item Number a Transaction quantity z modelu vyloučeny. Proměnná Item Number není k modelování vhodná kvůli tomu, že se jedná o unikátní systémový identifikátor a je pro každý záznam odlišná. Proměnná Transaction quantity udávající počet skladových pohybů byla vyloučena kvůli tomu, že se jedná o původně predikovanou proměnnou a nebude možné v budoucnosti do modelu použít v podobě vstupu. Při zakládání vkládání nového produktu do modelu nebude ještě znám počet jeho skladových pohybů.

Logistická regrese dokáže pracovat pouze s číselnými proměnnými, a proto do modelu nelze zahrnout polynominální proměnné v jejich původní podobě. V následujícím kroku je tudíž třeba pomocí operátoru Nominal to Numerical nominální datové vstupy transformovat na numerické a vytvořit tak umělé (dummy) proměnné. Tato konverze značně rozšíří množství proměnných (z původních 13 na 335), se kterými je v modelu pracováno a je zde vysoká pravděpodobnost výskytu multikolinearity. Z tohoto důvodu je do procesu přidán operátor Remove Correlated Attributes, který odstraní korelované proměnné a zredukuje tak celkové množství proměnných na 328. S takto vysokým počtem proměnných není kvůli přehlednosti vhodné vypisovat celkovou výslednou rovnici logistické regrese a testovat významnost jednotlivých parametrů.

Po odstranění korelovaných proměnných je třeba stanovit tzv. role proměnných, tedy určit tu, která má být modelem predikována. Operátorem Set Role je tedy třeba vybrat proměnnou Fast Pick Location a nastavit jí roli Label (predikovaná proměnná).

Nyní je vhodné do procesu přidat model pro testování validity modelu. Nejprve je třeba data rozdělit na dvě části. Jedna část bude sloužit pro samotné modelování a druhá potom pro aplikaci a ověření modelu. K tomuto rozdělení nejlépe slouží operátor Split Data, který rozdělí testovaný soubor v poměru 7 ku 3. 70 % dat tedy bude sloužit pro výpočet modelu a zbylých 30 % pro jeho validaci.

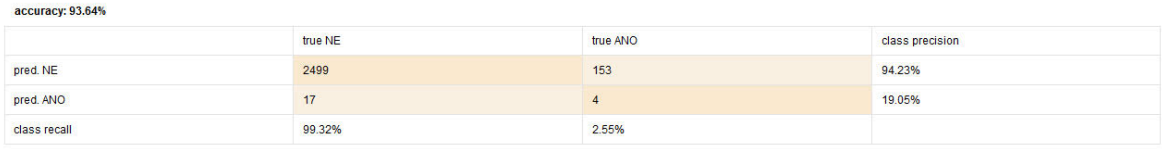
V této fázi jsou data připravena k modelování a je možné do procesu zapojit modul Logistic Regression do kterého se napojí 70 % datového souboru z operátoru Split Data. Zbylých 30 % dat je napojeno do modulu Apply Model, který model na připojená data aplikuje a pomocí operátoru Performance (Binominal Classification) vyhodnotí výsledky. Po správném napojení všech procesů je možné proces spustit a vyhodnotit výstupy, jimiž jsou v tomto případě výsledková tabulka predikce proměnných (obrázek č. 22) a tzv. Performance Vector, (obrázek č. 23) který měří přesnost modelu.

Row No.	Fast Pick Location	prediction(Fast Pick Location)	confidence(NE)	confidence(ANO)
1	NE	NE	0.99	0.01
2	NE	NE	0.99	0.01
3	NE	NE	0.99	0.01
4	NE	NE	0.97	0.03
5	NE	NE	0.98	0.02

Obrázek 22 Výsledková tabulka pravděpodobnosti výsledků modelu logistické regrese. Zdroj: Rapid Miner.

Na obrázku číslo 22 je znázorněn výstup programu Rapid Miner v podobě pravděpodobnostní tabulky výsledků. Na rozdíl od standardní lineární regrese, jejímž výstupem je hodnota závislé proměnné, logistická regrese počítá pouze pravděpodobnost toho, že nastane jeden, či druhý předem stanovený výsledek. Ve výše znázorněné tabulce výsledků jsou důležité čtyři výsledné sloupce. Hodnoty v řádcích prvního sloupce Fast Pick Location, odpovídají hodnotám predikované proměnné ze základního souboru. Druhý

sloupec prediction (Fast Pick Location) zobrazuje předpověď vypočítanou modelem logistické regrese pro daný řádek. Hodnoty v řádcích třetího sloupce confidence (NE) ukazují s jakou pravděpodobností, či jistotou nastane predikovaná kategorická proměnná na základě hodnot ostatních atributů. Čtvrtý sloupec confidence (ANO) znázorňuje pravděpodobnost, s jakou potom nastane druhá kategorická proměnná. Součet pravděpodobností nastání obou kategorií je vždy 1. Hodnoty, které vyjadřují pravděpodobnost správné predikce jednotlivých řádků se pohybují od 0 do 1 s tím, že defaultně nastavený práh pro predikci jedné, či druhé proměnné je hodnota jistoty 0,5 (pokud je ve zkoumaném případě hodnota jistoty pro možnost NE vyšší než 0,5, model předpoví hodnotu NE, pokud je nižší, předpoví model hodnotu ANO).



accuracy: 93.64%

	true NE	true ANO	class precision
pred. NE	2499	153	94.23%
pred. ANO	17	4	19.05%
class recall	99.32%	2.55%	

Obrázek 23 Klasifikační matice modelu logistické regrese. Zdroj: Rapid Miner.

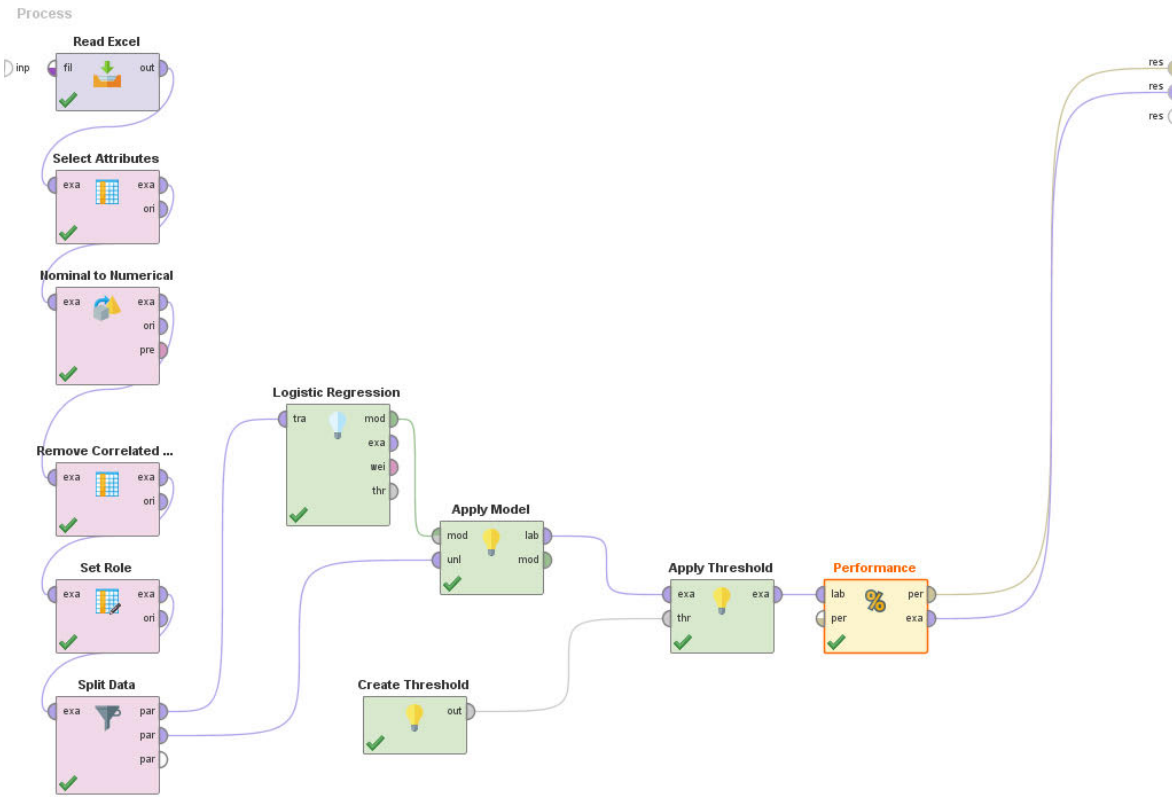
Na obrázku číslo 23 je znázorněn výsledek operátoru Performance, který měří přesnost modelu na základě využití oddělených 30 % dat, na které byl model aplikován. Jedná se o klasifikační matici 2 x 2 kterou lze interpretovat jako porovnání skutečnosti a modelem predikovaných hodnot. Z této matice lze vyčíst, že na základě hodnot atributů vysvětlovaných proměnných model předpověděl správně 2499 z 2516 případů výskytu hodnoty predikované proměnné NE a pouze 4 z 157 případů výskytu predikované proměnné ANO. Lze také vyčíst, že model dokáže předpovědět správně výslednou kategorii NE s pravděpodobností 99,32 % a správně výslednou kategorii ANO s pravděpodobností pouze 2,55 %. Celkovou přesnost modelu lze potom vypočítat jako poměr správných předpovědí a celkového počtu případů. Ve zkoumaném modelu to tudíž bude 2503 správných předpovědí ku 2673 všech případů, z čehož vyplývá celková přesnost modelu 93,64 %.

Přestože celková přesnost modelu je více než 90 %, je zde pozorovatelný značný rozdíl mezi přesností predikce první a druhé kategorie. Model je tudíž schopný velice přesně predikovat fakt, že produkt nebude rychloobrátkový, nicméně není schopný určit

druhý výsledek. Tento fenomén je způsoben tím, že v základní databázi bylo přítomno mnohem více případů, které spadají do kategorie NE. Jelikož byla hranice předělu kategorií ANO a NE stanovena na 30 pohybů, byl soubor rozdělen na 8387 záznamů spadajících do kategorie NE a pouze 522 záznamů spadajících do kategorie ANO.

Pokud by bylo cílem modelu predikovat, že výrobek nebude rychloobrátkový byl by tento model uspokojivý. Pokud by však byl obchodním prostředím požadován model, který přesněji predikuje zařazení nového produktu do kategorie ANO, bylo by nutné model upravit. Pro zpřesnění modelu pro danou kategorii lze ve zkoumaném případě aplikovat dvě opatření, které sice o něco sníží celkovou přesnost modelu, což ale může být vyváženo jeho přínosností pro daný obchodní případ. V první řadě se jedná o úpravu procesu, při které je pomocí operátorů Create Threshold a Apply Threshold změněna hranice jistoty se kterou model vyhodnotí výslednou kategorii jako ANO. Defaultně je tato hranice nastavena na 50 %, tedy 0,5. Pokud je tato hranice snížena například na 0,4 (modelu stačí pouze čtyřicetiprocentní jistota, aby určil výsledek jako ANO), dochází ke zpřesnění predikce modelu pro jednu z kategorií, za cenu poklesu přesnosti celkové.

Jak bylo zmíněno výše, nevyváženost predikce pro jednu z kategorií je způsoben především nevyvážeností poměru výskytu těchto kategorií v základním souboru. Pokud je tedy žádoucí zpřesnění modelu pro jednu z kategorií, je třeba vyvážit data v základním souboru. Po vyhodnocení základního souboru pomocí kontingenční tabulky bylo zjištěno, že hodnoty spadající do kategorie NE, (tedy méně než 30 skladových pohybů) mají průměrný počet záznamů 289 pro každou hodnotu. Tento fakt hodně ovlivňovala skutečnost, že se v souboru vyskytuje více než 4000 produktů, které vykazují pouze 1 skladový pohyb. Oproti tomu záznamů spadajících do kategorie ANO (tedy 30 a více skladových pohybů) je v souboru průměrně 5 pro každou hodnotu. K vyvážení je tedy třeba zredukovat počet záznamů pro každou hodnotu, která spadá do kategorie NE tak, aby byly průměrné počty záznamů v obou kategoriích přibližně stejné. Po zredukování vstupuje do modelu základní soubor čítající 591 záznamů s vyrovnanými průměrnými počty obě kategorie. Výsledný proces a klasifikační matice jsou znázorněny na obrázku č. 24 a 25.



Obrázek 24 Upravený proces logistické regrese. Zdroj: Rapid Miner.

accuracy: 80.68%

	true NE	true ANO	class precision
pred. NE	16	7	69.57%
pred. ANO	27	126	82.35%
class recall	37.21%	94.74%	

Obrázek 25 Klasifikační matice upravené logistické regrese. Zdroj: Rapid Miner.

Z klasifikační matice lze vyčíst, že bylo dosaženo značného zpřesnění kategorie ANO z původních 2,55 % na nynějších 94,74%. V důsledku úpravy modelu poklesla jeho celková přesnost o cca 13 %, což ale díky jeho vyšší přidané hodnotě pro obchodní prostředí nebrání jeho případné aplikaci.

4.5.2 Rozhodovací Strom

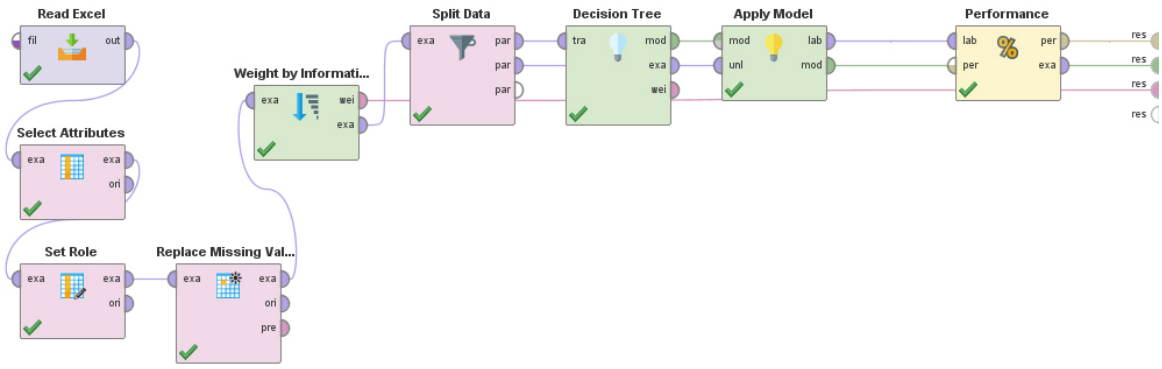
Druhý model, který bude na daný případ aplikován je model rozhodovacího stromu. Tento model bude aplikován na již upravený soubor, ve kterém je pro každý počet skladových pohybů přibližně stejný počet záznamů. Tímto krokem se předejde prvotní nevyváženosti modelu a bude možné se soustředit na jiné způsoby zpřesňování jeho předpovědí. Výhoda modelu založeného na tvoření rozhodovacího stromu spočívá v jeho

schopnosti pracovat i s polynominálními proměnnými, které není nutné převádět na numerické, což usnadňuje následnou srozumitelnost a přehlednost modelu.

Prvotní nastavení proměnných spočívá opět v určení cílové proměnné a vyloučení proměnných nevhodných pro modelování. Stejně jako v případě logistické regrese dojde k označení role proměnné Fast Pick Location kategorií Label a vyřazení proměnných Item Number a Transaction Quantity z modelu. Po omezení proměnných pouze na ty, které do modelu vstupují je třeba nahradit chybějící hodnoty. Ve vybraném datovém souboru, čítajícím 590 záznamů nejsou u některých atributů vyplněné všechny hodnoty, či obsahují pouze nuly. Těmito atributy jsou zejména ty, které značí převážně rozměry a váhy produktu Gross Weight (39 chybějících hodnot), Net Weight (2), Lenght (229), Height (228), Width (228), Purchase Price (312). V takovémto případě nastává situace, kdy je třeba rozhodnout, zda záznamy s chybějícími hodnotami odmazat, hodnoty nahradit, či zvolit jiný postup. V případě tohoto obchodního případu bude postupováno dle doporučení Deana Abotta (2014), který v této situaci nechává řešení čistě na posouzení a zkušenostech analytika, jenž modelování provádí. Jelikož se jedná o atributy, které se na rozhodování v rámci rozhodovacího stromu podílí pouze malou měrou (viz. obrázek č. 28 – váhy jednotlivých kritérií vstupujících do modelu), budou pro další postup pomocí operátoru Replace Missing Values nahrazeny průměrnými hodnotami.

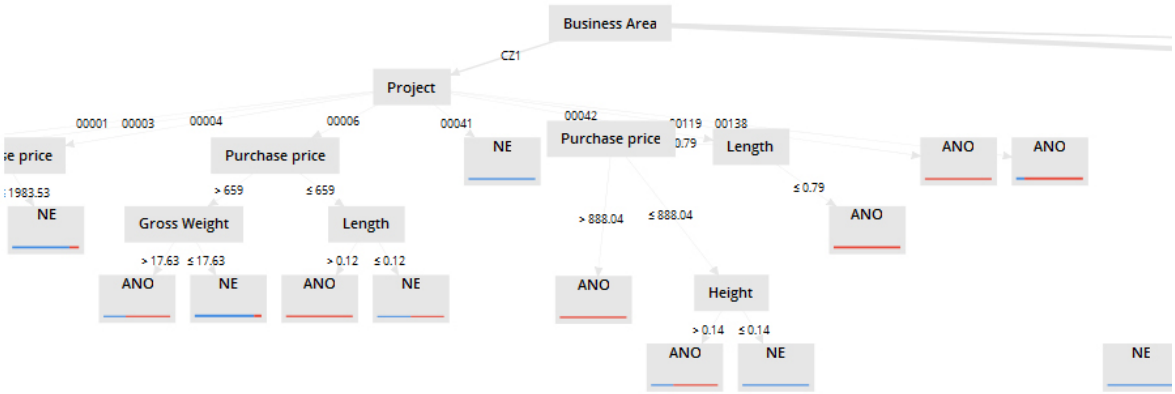
Po přípravě všech proměnných je třeba datový soubor opět rozdělit na dvě části v poměru 7 ku 3, aby bylo možné výsledky modelu otestovat a změřit jeho výkon. Jakmile jsou data rozdělena, je možné aplikovat samotný operátor Decision Tree, na kterém tento model stojí. Výsledný tvar rozhodovacího stromu lze v tomto kroku ovlivnit několika faktory v nastavení operátoru. V první řadě se jedná o kritérium, podle kterého algoritmus hodnotí a rozhoduje, jak nejlépe strom větvit. Ve zkoumaném případě bude postupováno dle kritéria Information_Gain, které řadí atributy s malou možností rozvětvování (převážně proměnné, jejichž hodnota spočívá v rozdělení do kategorií) na vrchol stromu. Dále je možné ovlivnit výsledný tvar stromu pomocí nastavení maximální hloubky, či počtu uzlů do které může strom růst. Obecně lze tvrdit, že čím vyšší toto číslo je, tím přesnější předpovídající schopnost daný strom má, nicméně se ale také stává nepřehlednějším. Pro tento model bude v prvním kole aplikováno 10 uzlů. Další parametry modelu spočívají

v nastavení vyřazování nepotřebných proměnných (prunning a preprunning), které jsou vyhodnocovány na základě své hladiny významnosti. Pro účely prvního modelu bude postupováno s přednastavenou hladinou významnosti 0,1. Po kalkulaci modelu stačí do procesu přidat operátor měřící výkon a pospojovat správně výstupy. Schéma celého procesu je znázorněno na obrázku číslo 26.



Obrázek 26 Výsledný proces rozhodovacího stromu. Zdroj: Rapid Miner.

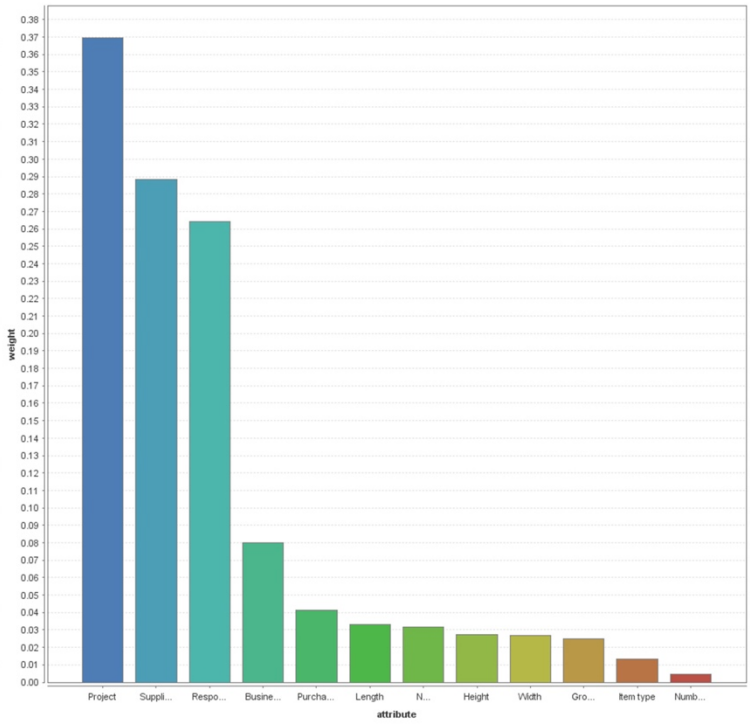
Po spuštění procesu je vytvořeno několik výstupů. V první řadě je pomocí operátoru Decision Tree vytvořeno schéma samotného rozhodovacího stromu. Vzhledem k relativně široké škále proměnných, které mohou jednotlivé proměnné nabývat je výsledný strom značně rozsáhlý. Na výřezu z tohoto stromu (obrázek číslo 27) je vidět, že algoritmus postupoval dle kritéria Information Gain a i přes to, že je dle proměnná Business Area na žebříčku vah kritérií relativně nízko (viz. obrázek číslo 28), umístil ji do pozice kořenového uzlu. Tento jev nastal proto, že daná proměnná nabývá pouze 7 různých hodnot a nedává tak moc prostoru ke složitému větvení. Dále je vzhledem ke své vysoké váze kritéria relativně vysoko proměnná Project, ze které se potom větví další atributy, jako například nákupní cena, rozměry, jméno výrobce atd. Celé schéma stromu je součástí přílohy číslo 1.



Obrázek 27 Výřez modelu rozhodovacího stromu. Zdroj: Rapid Miner.

Součástí softwarového výstupu je také tabulka a graf shrnující všechny vstupní atributy modelu a jejich váhy, hodnocené na základě získávání pro model přínosných informací. Na obrázku číslo 29 níže je vidět, že nejdůležitější proměnnými pro správné zařazení produktu do rychloobrátkové, či opačné kategorie jsou proměnné označující projekt, do kterého výrobek spadá (váha 0,37), dodavatele daného výrobku (váha 0,29), či označení produkčního zodpovědného za daný produkt (0,26).

attribute	weight
Project	0.37
Supplier ID	0.29
Responsible	0.26
Business Area	0.08
Purchase price	0.04
Length	0.03
Net Weight	0.03
Height	0.03
Width	0.03
Gross Weight	0.02
Item type	0.01
Number of materials	0.00



Obrázek 28 Klasifikační matice prvního rozhodovacího stromu. Zdroj: Rapid Miner.

Jednou z nejdůležitějších složek výstupu tohoto procesu je stejně jako v případě logistické regrese tzv. klasifikační matice, která měří přesnost předpovědí tohoto modelu. Pomocí této matice lze model evaluovat, porovnávat s ostatními a sledovat změny při úpravách modelu. Na obrázku číslo 29 je tato matice znázorněna pro výše zpracovaný rozhodovací strom.

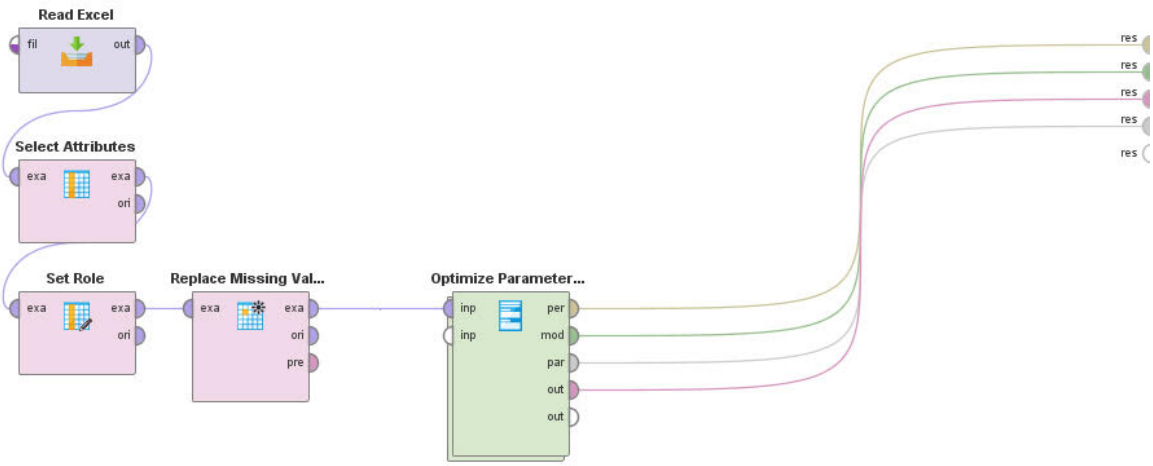
accuracy: 80.11%

	true NE	true ANO	class precision
pred. NE	19	11	63.33%
pred. ANO	24	122	83.56%
class recall	44.19%	91.73%	

Obrázek 29 Klasifikační matice prvního rozhodovacího stromu. Zdroj: Rapid Miner.

Z klasifikační matice lze vyčíst, že celková přesnost prvního zpracovaného rozhodovacího stromu je 80,11 % s tím, že model dokáže na 44,19 % určit, že se nejedná o rychloobrátkové zboží a naopak s 91,73 % pravděpodobností že se o toto zboží jedná. Model tedy vykazuje zhruba pětinovou chybovost.

Aby bylo dosaženo zpřesnění modelu, bude třeba postupně upravit jednotlivé parametry rozhodovacího stromu a sledovat, jak se přesnost předpovědi pro jednotlivé kategorie změní. Pro tento účel nabízí Rapid Miner operátor jménem Optimize Parameters (optimalizace parametrů), který dokáže vytvořit nejrůznější kombinace parametrů modelu a vybrat z nich tu, která vede k jeho nejvyšší přesnosti. Nejprve je nutné v tomto operátoru vytvořit sub proces, který zahrnuje veškeré optimalizované operátory. Následně je třeba zadat, které vzájemné kombinace parametrů má software testovat, přičemž je nutné vybrat pouze relevantní a postupně model adekvátně upravovat. Pokud by byly zahrnuty veškeré parametry sub procesu najednou, vzniklo by několik desítek milionů možných kombinací, což by značně zatěžovalo paměť systému a nebylo by tak možné model efektivně používat. Výsledný proces je znázorněn na obrázku číslo 30. Váhy kritérií zůstávají pro obě varianty modelu i při úpravách parametrů stejné.



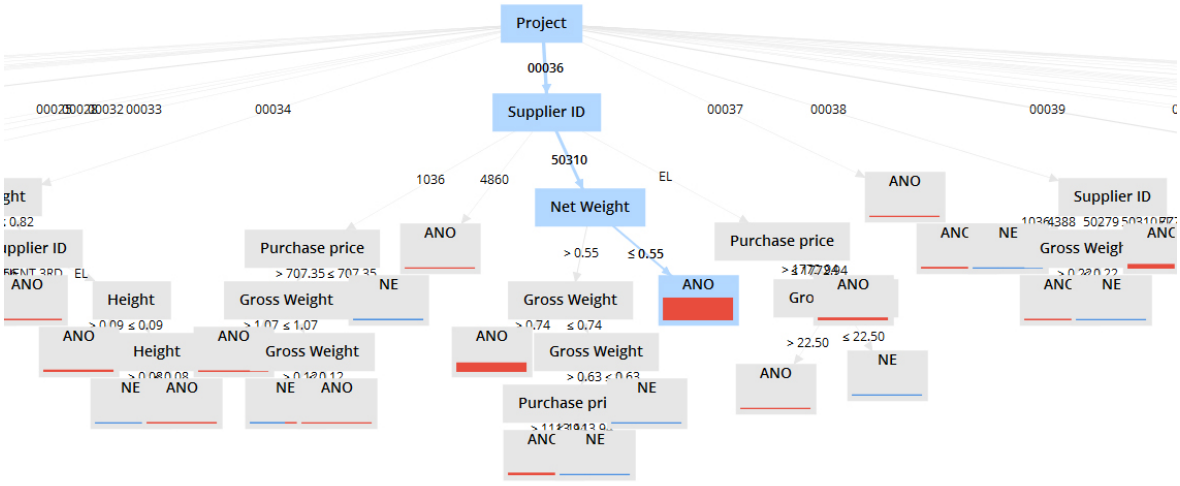
Obrázek 30 Proces optimalizace parametrů rozhodovacího stromu. Zdroj: Rapid Miner.

V rámci optimalizace tohoto rozhodovacího stromu je tedy vyhledána nejlepší kombinace 5 parametrů rozhodovacího stromu. Hodnoty parametrů, které vedou k vytvoření nejpřesnějšího modelu jsou znázorněny na obrázku číslo 31 níže.

```
Decision Tree (2).criterion     = information_gain
Decision Tree (2).maximal_depth = 19
Decision Tree (2).apply_pruning = true
Decision Tree (2).apply_prepruning      = false
Decision Tree (2).confidence    = 0.25
```

Obrázek 31 Ideální kombinace hodnot parametrů rozhodovacího stromu. Zdroj: Rapid Miner.

Z hodnot na obrázku výše vyplývá, že nejpřesnějším způsobem větvení stromu je větvení dle informačního zisku, který řadí do vyšších pozic atributy s menší možností rozvětvování. Maximální počet uzlů stromu (hloubka) je stanoven na 19 s tím, že model by měl vyřazovat některé nadbytečné proměnné. Hladina významnosti pro větvení je nastavena na 0,25.



Obrázek 32 Výřez výstupu optimalizovaného rozhodovacího stromu. Zdroj: Rapid Miner.

Na obrázku číslo 32 je znázorněn výřez optimalizovaného rozhodovacího stromu. Plná struktura je součástí přílohy číslo 2. Na obrázku je vidět, že je nyní strom více rozvětvený a složitější, nicméně nyní je jako kořenový uzel umístěna proměnná Project, následovaná proměnnou Supplier ID. Tyto proměnné mají v rámci modelu největší váhy, a tudíž jejich umístění na první dvě příčky modelu nyní odpovídá. Na výřezu rozhodovacího stromu si také lze povšimnout, že modře zvýrazněná cesta (čítající 82 záznamů) do hodnoty ANO vede přes projekt číslo 00036 a číslo dodavatele 50310. Tento projekt dle obrázku číslo 14 vykazuje v rámci celého zkoumaného souboru vůbec nejvíce skladových pohybů. Číselné označení dodavatele 50310 odpovídá výrobci Dakacom.cz s.r.o., který je dle word cloudového porovnání výrobců na obrázku číslo 19 dodavatelem, jehož produkty vykazují v rámci základního souboru také nejvíce pohybů. Na základě tohoto rozhodovacího stromu by se tudíž dalo předběžně určit jasné pravidlo, že pokud výrobek spadá pod výše zmíněný projekt a dodavatele a zároveň váží méně než 0,55 Kg bude automaticky rychloobrátkový a měl by být naskladněn do Fast Pick lokace. Toto zcela odpovídá realitě. Do této kategorie budou s největší pravděpodobností spadat jednoduché tisky, výměnné grafiky a sezónní materiály na papírové bázi, které jsou málokdy skladovány déle než 2 týdny.

accuracy: 89.20%

	true NE	true ANO	class precision
pred. NE	30	6	83.33%
pred. ANO	13	127	90.71%
class recall	69.77%	95.49%	

Obrázek 33 Klasifikační matice optimalizovaného rozhodovacího stromu.

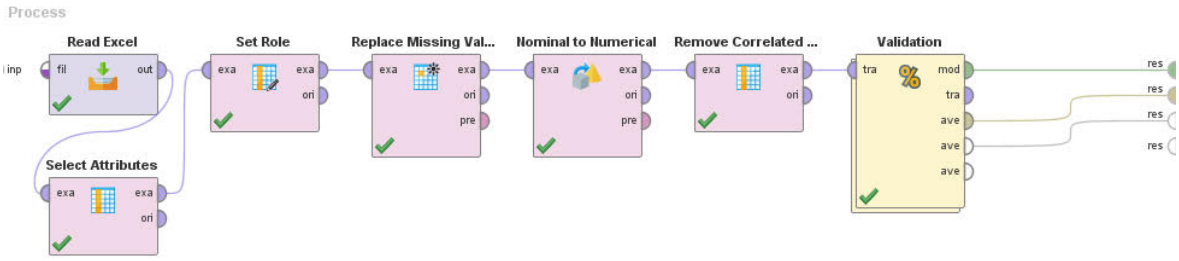
Na klasifikační matici znázorněné na obrázku číslo 33 je vidět, že se prediktivní schopnost modelu zvýšila ve všech směrech. Pro kategorii NE vzrostla z původních 44,19 % na nynějších 69,77 %, pro kategorii ANO potom z 91,73 % na 95,49 %. Celková spolehlivost předpovědi modelu vzrostla o 9,09 % na 89,20 %. Pokud jsou tedy známy veškeré atributy produktu, dokáže ho model v 9 z 10 případů správně zařadit do příslušné skladové lokace. Nevýhodou zpřesnění modelu je především jeho nynější komplikovanost a několikanásobně větší náročnost na výpočetní kapacitu systému.

4.5.3 Neuronové sítě

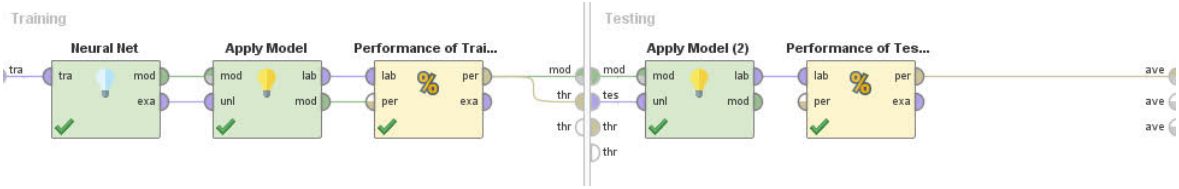
Třetím modelem aplikovaným na tento obchodní případ je model neuronové sítě. Jak již bylo zmíněno v teoretické části této práce tento model spočívá ve tvorbě vzájemně provázaných vrstev neuronů, které vzájemně přenáší signály a hledají spojitost mezi vstupy a cílovou proměnnou. Jako základní soubor pro modelování bude sloužit upravená databáze skladových pohybů, stejná jako v případě logistické regrese, či rozhodovacího stromu. Stejně jako v předchozích případech, cílová proměnná tohoto modelu je binomická proměnná Fast Pick Location, nabývající hodnot ANO, či NE. Proměnné Item Number a Transaction Quantity byly z modelu vyřazeny. První jmenovaná kvůli tomu, že se jedná pouze o systémové označení výrobku a druhá není v době zařazování nového produktu do skladu známá.

Po načtení databáze do softwaru je třeba nahradit chybějící hodnoty. Stejně jako při modelování pomocí rozhodovacího stromu jsou chybějící hodnoty proměnných značící fyzické vlastnosti produktů (rozměry, váha) nahrazeny průměry. V dalším kroku je potřeba převést nominální proměnné na numerické. Tato operace tvoří umělé (tzv. dummy) proměnné a rozšiřuje počet proměnných z původních 13 na 335. Při vytváření umělých proměnných je zde velká pravděpodobnost vzniku korelace, a proto je třeba zařadit

operátor Remove Correlated Values, který ve zkoumaném případě odstraní 7 korelovaných proměnných. V této fázi jsou veškerá data připravena pro aplikaci operátoru Validation, který rozdělí základní soubor na dvě části, z čehož jedna slouží k učení a druhá k ověřování výsledků. Aby bylo předejito ovlivnění modelu, je využita funkce tzv. Stratified Sampling, která zajistí, aby obě části souboru obsahovaly vyrovnaný počet vzorků pro obě varianty cílové proměnné. Po odpovídajícím rozdělení souboru je tedy možné do sub procesu operátoru Validation aplikovat samotný model Neural Net a měření jeho výkonu. V prvotním nastavení procesu je ponecháno generické nastavení modelu, tedy 200 učících cyklů, jedna skrytá vrstva a jemnost určování váhy kritérií 0,1. Učící neboli tréninkové cykly spočívají v opakování procesu, ve kterém jsou predikované hodnoty porovnávány s reálnými a na základě toho upravovány váhy jednotlivých kritérií tak, aby bylo dosaženo co nejmenší chybovosti. Finální podoba hlavního procesu a sub procesu modelování je znázorněna na obrázku 34 a 35.



Obrázek 34 Proces tvorby umělé neuronové sítě. Zdroj: Rapid Miner.



Obrázek 35 Subproces operátoru Validation. Zdroj: Rapid Miner.

Výše znázorněný proces generuje dva výstupy. Prvním výstupem je schéma výsledné neuronové sítě s jednou skrytou vrstvou. Toto schéma je kvůli převádění nominálních proměnných na numerické značně nepřehledné. Vrstva vstupů obsahuje 328 neuronů, skrytá vrstva 70 neuronů a výstup 2 neurony, které odpovídají hodnotám cílové proměnné. Celé schéma neuronové sítě je součástí přílohy číslo 3.

Druhým výstupem je stejně jako u předchozích modelů klasifikační matice, znázorňující poměr správných a nesprávných předpovědí a celkovou přesnost modelu.

accuracy: 84.66%

	true NE	true ANO	class precision
pred. NE	27	11	71.05%
pred. ANO	16	122	88.41%
class recall	62.79%	91.73%	

Obrázek 36 Klasifikační matice neuronové sítě (200 tréninkových cyklů, 1 skrytá vrstva). Zdroj: Rapid Miner.

Na obrázku číslo 36 je znázorněna klasifikační matice modelu neuronové sítě s 200 tréninkovými cykly a jednou skrytou vrstvou. Tento model dokázal správně předpovědět 62,79 % výsledků s hodnotou cílové proměnné NE a 91,73 % výsledků s hodnotou cílové proměnné ANO. Celková přesnost modelu je 84,66 %.

Aby došlo ke zpřesnění celkové prediktivní schopnosti modelu, je možné upravit parametry operátoru Neural Net. Pro druhé měření je tedy zvýšen počet tréninkových cyklů na 500 a zahrnuty 2 skryté vrstvy o 70 a 35 neuronech. Výsledné schéma nové neuronové sítě je součástí přílohy číslo 4, klasifikační matice je potom znázorněna na obrázku číslo 37.

accuracy: 85.23%

	true NE	true ANO	class precision
pred. NE	30	13	69.77%
pred. ANO	13	120	90.23%
class recall	69.77%	90.23%	

Obrázek 37 Klasifikační matice neuronové sítě (500 tréninkových cyklů, 2 skryté vrstvy). Zdroj: Rapid Miner.

Z výsledné klasifikační matice upraveného modelu neuronové sítě lze vyčíst, že se prediktivní schopnost pro cílovou proměnnou NE zvýšil o 6, 98 %, nicméně prediktivní schopnost pro opačnou proměnnou o 1,5 % poklesla. Celková přesnost modelu potom vzrostla pouze o 0,57 % na finálních 85,23 %. Oproti předchozímu nastavení neuronové sítě trval výpočet s těmito parametry zhruba trojnásobnou dobu, nicméně přesnost modelu vzrostla pouze nepatrně.

4.6 Srovnání a vyhodnocení výsledků jednotlivých modelů

Po fázi modelování přichází v rámci metodiky CRISP-DM na řadu fáze Evaluation, tedy srovnání a vyhodnocení výsledků jednotlivých modelů a výběr toho nejvhodnějšího pro daný obchodní případ. Pro srovnání výkonu jednotlivých modelů bude použita metoda křížové validace, která aplikuje daný model na celý základní soubor a vyhodnotí přesnost predikcí. Modely budou porovnávány z hlediska své celkové přesnosti, specifičnosti a citlivosti.

Porovnávány budou vždy modely s upravenými parametry pro lepší výkon. Jedná se tedy o logistickou regresi s upraveným základním souborem, rozhodovací strom s optimalizovanými parametry a umělou neuronovou síť s 500 tréninkovými cykly a dvěma skrytými vrstvami. V tabulce číslo 5 jsou porovnány výsledky křížové validace tří výše zmíněných modelů.

Model	Klasifikační matice			Přesnost	Specifičnost	Citlivost
		True NE	True ANO			
Logistická regrese	Pred NE	92	34	85,25%	88,58%	73,02%
	Pred ANO	53	411			
Rozhodovací strom	Pred NE	100	32	86,95%	90,17%	75,76%
	Pred ANO	45	413			
Neuronová síť	Pred NE	94	36	85,25%	88,91%	72,31%
	Pred ANO	51	409			

Tabulka 5 Porovnání celkového výkonu jednotlivých modelů. Zdroj: Vlastní zpracování.

Z tabulky číslo 5 je patrné, že v křížové validaci, ve které byly modely aplikovány na všech 591 záznamů základního souboru vykazovaly všechny 3 modely přibližně stejný výkon. Z hlediska přesnosti, specifičnosti i citlivosti dopadl nejlépe optimalizovaný model rozhodovacího stromu, který dokáže predikovat s přesností 86,95 % při specifičnosti 90,17 % a citlivosti 75,76 %. Jako druhý v pořadí by bylo možné určit jak logistickou regresi, tak neuronovou síť. Oba tyto modely vykazují identickou celkovou přesnost předpovědí a záleží pouze na daném obchodním případu, zda preferuje vyšší přesnost predikce pozitiv (citlivost), či negativ (specifičnost). Pokud by byla žádoucí vyšší specifičnost modelu, bylo by možné na druhé místo zařadit neuronovou síť, pokud by byla potřeba naopak vyšší citlivost modelu, zaujala by druhé místo logistická regrese.

Ve zkoumaném obchodním případě společnosti EYELEVEL vychází nejlépe model rozhodovacího stromu jak z hlediska ukazatelů klasifikační matice, tak z hlediska praktického. Rozhodovací stromy pracují na bázi série pravidel If – Then – Else a jejich algoritmus může být aplikován do systémů pracujících na bázi SQL, což je právě informační systém Infor M3, který společnost využívá.

4.7 Aplikace rozhodovacího stromu

Poslední fází metodiky CRISP-DM je tzv. Deployment, tedy nasazení vybraného modelu do provozu společnosti. Aplikace prediktivního modelu do struktur informačního systému Infor M3 pomůže společnosti v určování vhodné skladové lokace pro nově vytvářené a skladem přijímané produkty. V praxi bude toto řešení fungovat tak, že v momentě, kdy produkční manažer založí na základě podnětu klienta nový výrobek do informačního systému, systém automaticky podhodí jednu ze dvou oblastí (rychloobrátková lokace / ostatní) pro jeho uskladnění. Toto automatické navrhování zařazení produktu bude probíhat přímo v grafickém uživatelském prostředí informačního systému, kde bude k tomuto účelu zpřístupněno nové textové pole. Pole bude algoritmem vyplňováno pouze v případě, že má produkt vyplněné veškeré hodnoty svých atributů.

V této fázi projektu je třeba stanovit konkrétní kroky, které je pro zavedení modelu nutné vykonat (tzv. action steps), vypracovat přibližný časový plán zavedení, stanovit způsob, jakým bude inovace komunikována uživatelům a definovat metriku, s jejíž pomocí bude výkon modelu sledován a případně operativně upravován. Jelikož se jedná pouze o teoretický příklad aplikace modelu a není možné změřit reálné výsledky, je součástí této fáze také ukázka praktického využití modelu v situaci, kdy je plně zaveden a skladová lokace Fast Pick existuje. Na základě tohoto příkladu jsou také vyčísleny případné provozní úspory nákladů, které může tento model společnosti přinést.

4.7.1 Action steps

Ve fázi nasazování modelu do provozu je třeba nejprve stanovit konkrétní kroky (tzv. Action steps), které je nutné v rámci realizace projektu podniknout. Tyto kroky lze

definovat jako seznam dílčích činností, které je potřeba postupně vykonat, aby se projekt hýbal kupředu a bylo dosaženo stanoveného cíle. Každý krok má přiřazenu odpovědnou osobu, či oddělení, které je za činnosti zodpovědné a které dohlíží na jejich realizaci. V rámci projektu aplikace algoritmu rozhodovacího stromu do informačního systému Infor M3 jsou zahrnuty 4 oddělení společnosti EYELEVEL (zkráceně EL). Těmito odděleními jsou oddělení informačních technologií (IT) zodpovědné za programování a správu informačního systému, oddělení skladu (WH), kterého se bude nové řešení přímo týkat, oddělení Business Intelligence (BI) spravující data, reporty a uživatelskou stránku informačního systému a finanční oddělení (FIN), které určuje finanční rámec projektu. Ze strany poskytovatele informačního systému (Infor) bude zahrnut account manager, či projektový manager spravující společnost EYELEVEL coby klienta. Jednotlivé kroky včetně rozdělení odpovědností jsou vypsány v odrážkách níže.

- Tvorba zadání pro úpravu modulu informačního systému (EL IT, EL WH, EL BI).
- Tvorba projektového týmu (EL IT, EL WH, EL BI).
- Předběžná cenová kalkulace úpravy (Infor).
- Stanovení rozpočtu pro projekt (EL FIN, EL IT, EL BI).
- Programování modulu (EL IT, EL BI, Infor).
- Úpravy, konzultace, diskuze (EL IT, EL BI, Infor).
- Finální cenová kalkulace modulu (Infor).
- Představení nového řešení (Infor, EL IT, EL WH, EL BI).
- Školení úzké skupiny uživatelů (EL IT, EL WH, EL BI).
- Testovací provoz (EL WH, EL BI).
- Vyhodnocení testovacího provozu (EL WH, EL BI).
- Zavedení modelu do plného provozu (EL BI, EL IT).
- Školení uživatelů (EL WH, EL BI).
- Monitorování provozu, operativní úpravy (EL BI, Infor).

4.7.2 Časový plán zavedení modelu do provozu

Předběžný časový plán tohoto projektu je stanoven na 25 týdnů, přičemž časová náročnost jednotlivých úkolů je znázorněna na časové ose v tabulce číslo 6. Rozpočet na

realizaci projektu je omezen 100 000 Kč, které počítají i s neplánovanými úpravami. Některé úkoly, jako například tvorba projektového týmu a tvorba zadání, či usměrňování probíhají současně a jsou na sobě závislé. Nejdůležitější část tohoto projektu, tedy samotné programování algoritmu prediktivního modelu do struktury informačního systému zabere zhruba třetinu celého doby realizace. Jedním z bodů tohoto akčního plánu je také testovací provoz, který ovšem zahrnuje pouze procesní stránku věci a sledování funkcionality nového algoritmu. Pro vyhodnocení přínosů modelu z hlediska výsledků a ušetřených prostředků je třeba provést analýzu rozsáhlejšího souboru dat, tedy údajů alespoň za jeden kvartál. Pro dlouhodobý úspěch projektu je klíčové také jeho sledování, měření a vyhodnocování po zavedení do provozu. Tento proces již není v časové ose znázorněný, nicméně bude probíhat po celou dobu aktivního působení modelu a bude spadat do kompetence oddělení Business Intelligence.

Úkol	Odpovědnost	Časová osa (týdny)																							
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Tvorba zadání pro úpravu modulu informačního systému	IT, WH, BI																									
Tvorba projektového týmu	IT, WH, BI																									
Předběžná cenová kalkulace úpravy	Infor																									
Stanovení rozpočtu pro projekt	IT, BI, FIN																									
Programování modulu	IT, BI, Infor																									
Úpravy, konzultace, diskuze	IT, BI, Infor																									
Finální cenová kalkulace modulu	Infor																									
Představení nového řešení	IT, WH, BI, Infor																									
Školení úzké skupiny uživatelů	IT, WH, BI																									
Testovací provoz	WH, BI																									
Vyhodnocení testovacího provozu	WH, BI																									
Zavedení modelu do plného provozu	IT, BI																									
Školení uživatelů	WH, BI																									
Monitorování provozu, operativní úpravy	BI, Infor																									

Tabulka 6 Časová osa aplikace prediktivního modelu. Zdroj: Vlastní zpracování.

4.7.3 Komunikace řešení koncovým uživatelům

Po zakomponování algoritmu rozhodovacího stromu do modulu informačního systému je třeba inovaci komunikovat koncovým uživatelům. V rámci společnosti EYELEVEL k tomuto účelu slouží úzká skupina zaměstnanců (key users), kteří se pohybují na pomezí jednotlivých oddělení (klientské služby, produkce, logistika, sklad atd.) a Business Intelligence, či IT. Tito zaměstnanci rozumí dokonale strukturám informačního systému, hledají možnosti zlepšení z hlediska svého oddělení a komunikují nové postupy dále do firmy. Nově zakomponovaný algoritmus a související proces budou tedy touto formou komunikovány do oddělení, kterých se týká. Inovace v rámci společnosti ovlivní činnost produkčních manažerů, na které bude nově vyvíjen tlak k důslednému vyplňování veškerých atributů nově zakládaných produktů. Doposavad je v rámci atributů povinných polí pouze několik, což ve spojení s nedůsledností některých

zaměstnanců vede k neúplné databázi údajů o produktech, která nedosahuje svého plného potenciálu z hlediska vypovídací hodnoty pro možné budoucí analýzy. Druhou skupinou zaměstnanců, které se nově zavedený proces přímo týká je oblast skladu. Těmto zaměstnancům se nově při příjmu zboží na sklad zobrazí přímo na displeji čtecího zařízení čárových kódů informace, do jaké kategorie zboží spadá a do jaké skladové oblasti ho tedy uložit.

4.7.4 Měření výkonu modelu a opatření do budoucna

Součástí této fáze metodiky je také stanovení postupu, jakým výkon nově aplikovaného modelu měřit, monitorovat a upravovat. Měření výkonu nového modelu bude probíhat pravidelně s pomocí analýzy reportu ze SQL databáze. Tento report bude obsahovat výsledky veškerých predikcí modelu, tedy seznam itemů a jejich zařazení do kategorie rychloobrátkových, či ostatních položek. Výstupy tohoto reportu budou následně porovnávány s reálným počtem skladových pohybů daného produktu, což ověří, do jaké míry model predikuje přesně. Na základě těchto údajů bude možné sestavit klasifikační matici, kterou je následně možné porovnat s maticí využitou pro prvotní hodnocení modelu, či s klasifikační maticí jiného období. Výstupem této komparace bude graf mapující výkon modelu v průběhu času (např. po měsících).

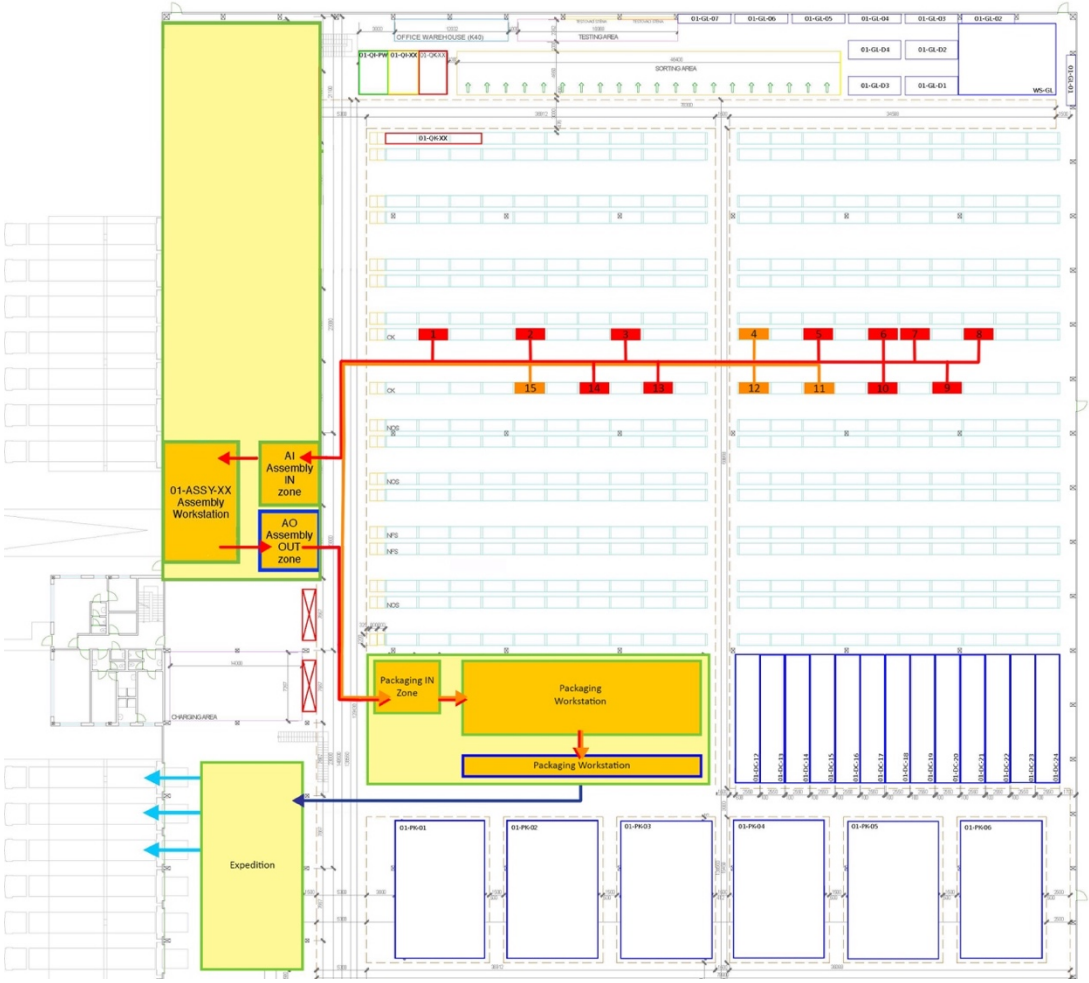
Databáze, na jejímž základě byl model vytvořen, obsahuje v současné chvíli údaje veškerých dodavatelů, klientů, projektů i ostatních atributů, které se v informačním systému společnosti vyskytují. Do budoucna je však samozřejmě možné, že se tyto údaje budou rozrůstat a mohlo by tak dojít v omezení prediktivní schopnosti modelu (do modelu budou vstupovat neznámé hodnoty polynomických proměnných, které nebude umět model správně zařadit). Z tohoto důvodu je třeba pravidelně upravovat tréninkovou databázi modelu a rozšiřovat ji o další záznamy se známými i neznámými hodnotami atributů.

4.7.5 Příklad využití modelu v praxi a vyčíslení možných úspor

Z informací, vyplývajících z rozhodovacího stromu je možné ve skladu vytvořit speciální lokaci Fast Pick, které bude využívána pro uskladnění zboží s více skladovými pohyby během roku. Tato lokace bude umístěna v blízkosti montáže a expedice tak, aby se

při jejím využití minimalizoval čas potřebný na vyskladňování produktů, což vede k ušetření finančních prostředků vynaložených na lidskou práci a činnost skladové techniky.

Pro ukázku využití této lokace byla sestavena vzorová zakázka, jejíž průtok skladem bude mapován jak standardním postupem, tak postupem využívajícím lokaci Fast Pick, do které budou produkty zařazovány dle výstupů prediktivního modelu. Tato zakázka obsahuje 15 produktů, z nichž 11 vyžaduje montáž a 4 nikoliv. Každý z těchto produktů je kvůli svým rozměrům a váze uskladněn na standardní paletě (velikost EUR – 1200 x 800 mm) a musí být manipulován individuálně. V rámci zjednodušení schématického znázornění procesu není uvažováno vertikální rozdělení skladových pozic a komponentní struktura produktů. Pro vyčíslení manipulačních časů a nákladů bylo počítáno s průměrnou dobou manipulace s jedním kontejnerem (13 minut) a hodinovými náklady na práci skladové techniky včetně operátora (300 Kč).



Obrázek 38 Průtok materiálu skladem dle standardního postupu. Zdroj: Vlastní zpracování.

Na obrázku číslo 38 je schematicky znázorněn průtok materiálu skladem bez využití lokace pro zrychlené vyskladnění. Jednotlivé produkty, označené v modelové situaci čísly 1–15 jsou uskladněny v regálech přibližně uprostřed skladu, odkud jsou postupně převáženy k příslušnému zpracování. Červeně vyznačené produkty vyžadují montáž a musí tudíž projít skrz oblast kompletace produktů (Assembly). Odtud jsou dále převezeny ve zkompletované podobě do oblasti balení (Packaging), kde jsou zabaleny a převezeny na expedici. Oranžově označené produkty montáž nevyžadují a je třeba je převézt přímo k zabalení a odeslání zákazníkovi.

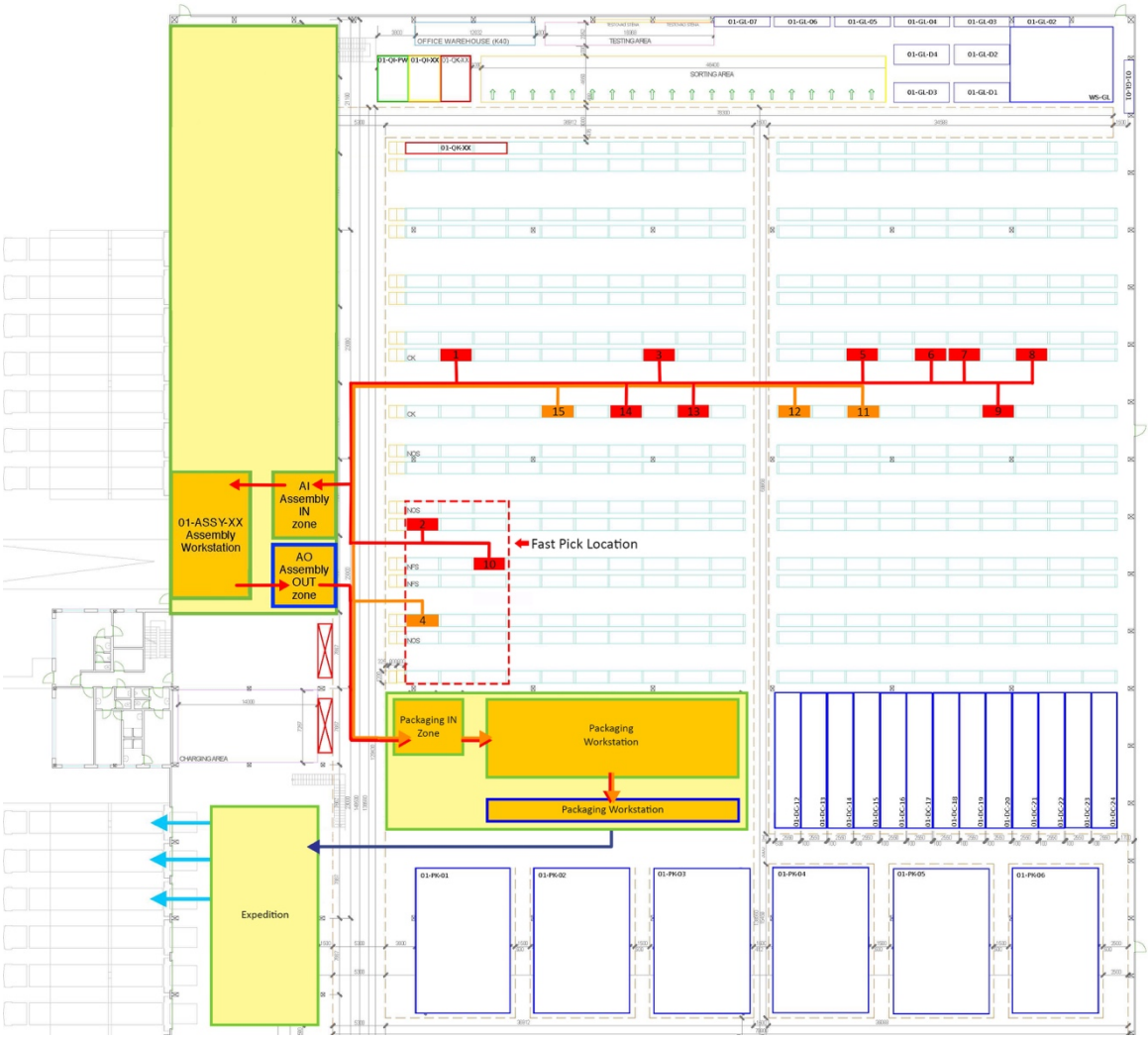
V tabulce číslo 7 níže jsou vypsány vzdálenosti, které při současném nastavení skladových pozic musí jednotlivé produkty urazit mezi místem svého uskladnění a oblastí balení. V případě kompletovaných itemů je počítána vzdálenost mezi skladovou pozicí (Kontejnerem), montáží a oblastí balení. Pokud kompletaci nevyžaduje, je brána v potaz pouze vzdálenost mezi kontejnerem a balením. Tyto vzdálenosti jsou dále přepočítány na časovou náročnost manipulace, která je dále převedena na finanční náklady na provoz skladové techniky a její obsluhy. Vzdálenost mezi oblastí balení a expedicí není v porovnání nákladů zahrnována, protože není případným využitím Fast Pick lokace ovlivněna a bude pro obě varianty stejná.

Stndard		Manipulační vzdálenost (m)				
Produkt	Vyžaduje montáž?	Kontejner - Montáž - Balení	Kontejner - Balení	Celkem	Manipulační doba (min)	Náklady na manipulaci (Kč)
1	ANO	36,14	0,00	36,14	6,54	32,68
2	ANO	46,68	0,00	46,68	8,44	42,21
3	ANO	57,22	0,00	57,22	10,35	51,75
4	NE	0,00	81,31	81,31	14,71	73,53
5	ANO	76,79	0,00	76,79	13,89	69,45
6	ANO	84,32	0,00	84,32	15,25	76,26
7	ANO	87,33	0,00	87,33	15,80	78,98
8	ANO	93,36	0,00	93,36	16,89	84,43
9	ANO	90,35	0,00	90,35	16,34	81,70
10	ANO	84,32	0,00	84,32	15,25	76,26
11	NE	0,00	87,33	87,33	15,80	78,98
12	NE	0,00	81,31	81,31	14,71	73,53
13	ANO	60,23	0,00	60,23	10,89	54,47
14	ANO	54,21	0,00	54,21	9,80	49,02
15	NE	0,00	57,22	57,22	10,35	51,75
	Celkem	770,95	307,18	1078,13	195,00	**975,00**

Tabulka 7 Přehled manipulačních časů a nákladů standardního postupu. Zdroj: Vlastní zpracování.

Z tabulky číslo 7 vyplývá, že produkty této vzorové zakázky musí být při současném nastavení skladového hospodářství převezeny celkem o 1078,13 metrů, což pracovníkům zabere 195 minut. Pro tuto manipulační dobu je možné na základě průměrné hodinové sazby vyčíslit náklady ve výši 975 Kč, které zahrnují lidskou práci, energie i opotřebení skladové techniky.

Na obrázku číslo 39 je znázorněn průtok stejné zakázky skladem při využití lokace Fast Pick, do které budou uskladněny všechny současné produkty s více než 30 skladovými pohyby ročně a zároveň všechny produkty, které prediktivní model vyhodnotí jako rychloobrátkové.



Obrázek 39 Průtok materiálu skladem s využitím lokace Fast Pick. Zdroj: Vlastní zpracování.

Pro ilustraci byly 3 produkty ze vzorové zakázky vyhodnoceny jakožto rychloobrátkové. Z těchto 3 produktů 2 vyžadují kompletaci (produkty číslo 2 a 10) a 1 je možné vyskladnit přímo do oblasti balení (produkt číslo 4). Pro lokaci Fast Pick bylo v rámci půdorysu skladu vybráno umístění v levém dolním rohu, které je nejvýhodnější z hlediska vzdálenosti mezi jednotlivými kontejnery a oblastí montáže, či balení. Tato lokace by svým objemem měla svým účelům postačovat, nicméně vždy je možné její velikost adekvátně upravit.

Stejně jako v případě standardního uskladnění produktů byly vypočítány vzdálenosti jednotlivých kontejnerů a oblastí montáže a balení. Přehled vzdáleností, manipulačních časů a nákladů je vypsán v tabulce číslo 8.

Fast Pick		Manipulační vzdálenost (m)				
Produkt	Vyžaduje montáž?	Kontejner - Montáž - Balení	Kontejner - Balení	Celkem	Manipulační doba (min)	Náklady na manipulaci (Kč)
1	ANO	36,14	0,00	36,14	6,54	32,68
2	**ANO**	**34,63**	**0,00**	**34,63**	**6,26**	**31,32**
3	ANO	57,22	0,00	57,22	10,35	51,75
4	**NE**	**0,00**	**22,59**	**22,59**	**4,09**	**20,43**
5	ANO	76,79	0,00	76,79	13,89	69,45
6	ANO	84,32	0,00	84,32	15,25	76,26
7	ANO	87,33	0,00	87,33	15,80	78,98
8	ANO	93,36	0,00	93,36	16,89	84,43
9	ANO	90,35	0,00	90,35	16,34	81,70
10	**ANO**	**28,61**	**0,00**	**28,61**	**5,17**	**25,87**
11	NE	0,00	87,33	87,33	15,80	78,98
12	NE	0,00	81,31	81,31	14,71	73,53
13	ANO	60,23	0,00	60,23	10,89	54,47
14	ANO	54,21	0,00	54,21	9,80	49,02
15	NE	0,00	57,22	57,22	10,35	51,75
	Celkem	703,19	248,45	951,65	172,12	**860,61**

Tabulka 8 Přehled manipulačních časů a nákladů s využitím Fast Pick lokace. Zdroj: Vlastní zpracování

V tabulce číslo 9 je znázorněn rozdíl jednotlivých ukazatelů v případě využití lokace Fast Pick pro uskladnění rychloobrátkových produktů číslo 2, 4 a 10. Z tabulky je patrné, že aplikací tohoto postupu do skladového hospodářství společnosti je možné docílit na takovéto zakázce zredukovat vzdálenost manipulace produktů o 126,48 metrů, zkrátit manipulační dobu o 22 minut 53 sekund a tím pádem ušetřit 114,39 Kč.

Produkt	Vyžaduje montáž?	Rozdíl Manipulační vzdálenosti (m)	Rozdíl Manipulační doby (min)	Rozdíl nákladů na manipulaci (Kč)
2	ANO	12,05	2,18	10,89
4	NE	58,73	10,62	53,11
10	ANO	55,71	10,08	50,38
	Celkem	**126,48**	**22,88**	**114,39**

Tabulka 9 Rozdíl v manipulačních vzdálenostech, dobách a nákladech standardního a navrhovaného postupu. Zdroj: Vlastní zpracování.

Vezme-li se v potaz referenční počet zakázek realizovaných v roce 2018 (10548) a fakt, že zhruba 20 % všech zakázek obsahuje produkty vhodné k uskladnění v lokaci Fast Pick, je možné vyčíslit možnou roční úsporu ve výši 241 317,14 Kč.

5 Výsledky a diskuse

V rámci praktické části této práce byla na konkrétní obchodní případ společnosti EYELEVEL aplikována metodika CRISP-DM, využívaná k rozdělení data miningových úkolů do dílčích fází, vedoucích ke konkrétnímu výsledku a aplikaci prediktivního modelu do praxe. V rámci této metodiky bylo v prvním kroku nutné porozumět businessové stránce věci, tedy fungování ovlivňovaných procesů ve společnosti, zdrojům dat, požadovaným přínosům celého projektu a možným rizikům, která mohou případně nastat. Dále bylo nutné analyzovat vstupní data a přetransformovat je do podoby, ve které je možné je využít pro samotné prediktivní modelování. V rámci této transformace bylo nutné agregovat informace získané z několika databází informačního systému společnosti, otestovat soubor z hlediska vhodnosti jednotlivých proměnných, nežádoucí proměnné ze souboru vyřadit a určit kategorickou proměnnou, jejíž hodnoty budou následně predikovány. V rámci zkoumaného obchodního případu se jednalo o proměnnou nazvanou Fast Pick Location, nabývající hodnot ANO/NE, která určovala, zda je daný produkt rychloobrátkový (vykazuje více než 30 skladových pohybů ročně), či nikoliv.

S využitím připraveného základního souboru byly následně s pomocí statistického softwaru Rapid Miner sestaveny 3 prediktivní modely, které by byly teoreticky vhodné pro aplikaci na daný obchodní případ. Těmito modely byla logistická regrese, model rozhodovacího stromu a model umělé neuronové sítě. Všechny tyto modely byly po svém zhotovení otestovány na 30 % základního souboru z hlediska celkového výkonu a přesnosti predikcí pro jednotlivé kategorie. Logistická regrese, rozhodovací strom i umělá neuronová síť byly následně parametricky optimalizovány tak, aby došlo ke zlepšení výkonu predikcí a bylo je možno případně aplikovat do praxe.

Po aplikaci prvního modelu na daný soubor bylo zjištěno, že datový soubor ve stávající podobě není pro modelování příliš vhodný, protože vykazuje značný nepoměr mezi počtem záznamů spadajících do kategorie ANO (522 záznamů) a NE (8387

záznamů). Model se tak dokázal velice přesně naučit předpověď pro kategorii NE, kterou s tímto základním souborem dokázal předpovědět s více než 99 % pravděpodobností, nicméně kategorii ano zvládal předpovědět pouze s 2,55 % pravděpodobností, což je pro zkoumaný obchodní případ nevhodné. Na základě této zkušenosti byl základní soubor upraven tak, aby pro všechny počty skladových pohybů spadající do kategorie ANO i NE existoval průměrně stejný počet záznamů, což určitým způsobem vyrovná přesnosti predikcí, či alespoň sníží rozdíl mezi přesnostmi predikcí jednotlivých kategorií. Logistická regrese byla tedy na nový základní soubor aplikována ještě jednou se sníženým prahem jistoty ve prospěch kategorie ANO. Tento upravený model již vykazoval přesnost 37,21 % pro kategorii NE a 94,74 % pro kategorii ANO, což ve výsledku znamená celkovou přesnost modelu 80,68 %.

Druhým modelem aplikovaným na tento upravený základní soubor byl model rozhodovacího stromu, který před optimalizací vykazoval celkovou přesnost 80,11 % (ANO 91, 73 %, NE 44, 19 %). V rámci tohoto modelu byly také zjištěny váhy jednotlivých kritérií, které mají na rozhodování při kategorizaci vliv. Tato kritéria také stojí v rozhodovacích stromech na vyšších pozicích, což znamená, že jsou pro rozhodování významná. To, zda bude produkt rychloobrátkový nejvíce ovlivňuje číslo projektu, do kterého patří (váha 0,37), číslo dodavatele (váha 0,29) a systémové označení produkčního, který je za výrobek zodpovědný (váha 0,26). V rámci optimalizace parametrů rozhodovacího stromu byla rozšířena jeho hloubka na 19 uzlů a kritérium větvení změněno na větvení dle váhy informace, což vedlo ke zpřesnění predikcí pro kategorii ANO o 3,76 %, pro kategorii NE o 25,58 % a celkové přesnosti o 9,09 %.

Posledním modelem, který byl v rámci praktické části práce aplikován je model umělých neuronových sítí, využívající princip hlubokého učení a skrytých vrstev. Tento model po prvotní aplikaci předpověděl výsledky pro kategorii ANO s přesností 91, 73 % a pro kategorii NE s přesností 62,79 %. Celková přesnost predikcí modelu byla potom, 84,66 %. Optimalizace tohoto modelu spočívala v úpravě počtu tréninkových cyklů na 500 a zahrnutí dvou skrytých vrstev o 70 a 35 neuronech. Tento krok drobně snížil přesnost predikce kategorie ANO na výsledných 90,23 %, nicméně přesnost kategorie NE byla

navýšena na výsledných 69,77 %. Tyto úpravy v konečném důsledku zvyšují celkovou přesnost modelu na 85,23 %.

Pro všechny výše zmíněné modely byla následně provedena tzv. cross validation, tedy jejich aplikace na celý základní soubor. Na základě výstupu tohoto procesu byly všechny tři modely porovnány z hlediska celkové přesnosti, specifičnosti a citlivosti, z čehož vyšel nejlépe optimalizovaný model rozhodovacího stromu s přesností. Tento model také určil faktory, které nejvyšší mírou přispívají ke kategorizaci produktů do jednotlivých kategorií.

Poslední součástí praktické části této práce je příklad aplikace vybraného modelu do reálného provozu firmy. V rámci této části byly stanoveny všechny dílčí činnosti, které je třeba pro realizaci vykonat. Byl také stanoven rozpočtový rámec projektu, časový plán zachycující veškeré činnosti, jejich odhadovanou časovou dotaci a oddělení společnosti zodpovědné jejich realizaci. Následně byl stanoven plán komunikace aplikovaného modelu do struktur informačního systému směrem ke koncovým uživatelům a plán řízení a kontroly modelu do budoucnosti.

Po hypotetické aplikaci modelu do struktur informačního systému společnosti byl vypracován praktický příklad průtoku materiálu skladem s využitím skladové lokace Fast Pick, do které budou nové produkty uskladňovány na základě výstupů prediktivního modelu. Na základě tohoto příkladu byly vyčísleny možné úspory, které zavedení modelu může společnosti přinést. Z této případové studie vyplynulo, že pokud bude tato lokace vybudována bude možné zkrátit tok materiálu v rámci jedné zakázky o 126,48 metrů, což pracovníkům ušetří 22,88 minut práce a společnosti 114,39 Kč v nákladech. Vezme-li se v potaz celkový roční objem zakázek (pro výpočet byl využit počet zakázek z roku 2018), je možné za 12 měsíců ušetřit 241 317,14 Kč. Pokud je tato částka ponížena o rozpočtový rámec 100 000 Kč, který bylo nutné vynaložit na tvorbu modelu a jeho programování do struktur SQL informačního systému, vychází roční úspora za první rok případného provozu na 141 317,14 Kč. Pro vedení společnosti je aplikace modelu velice výhodná, jelikož vysoké náklady na jeho aplikaci jsou pouze jednorázové. Následné náklady na dílčí úpravy modelu, rozšiřování tréninkové databáze, či průběžné měření výkonu budou bohatě

vyváženy jeho přidanou hodnotou vedoucí ke stále vyšším přesnostem předpovědí a tím pádem vyšším úsporám.

Kromě výše popsaného projektu by v rámci skladového hospodářství společnosti bylo možné realizovat nejen model předpovídající rozdělení produktů do dvou kategorií, ale i komplexnější model s polynomickými hodnotami predikované proměnné. Tímto modelem by bylo možné rozdělit oblast skladu do mnoha různých zón, které by reflektovaly obrátkovost produktů a dle tohoto údaje byly adekvátně rozmístěny v prostoru. Dosáhlo by se tak ještě markantnější optimalizace toku produktů a celkové efektivity procesu vyskladňování.

6 Závěr

Hlavním cílem této práce bylo předvídání chování nově zakládaných produktů společnosti EYELEVEL s.r.o. z hlediska počtu skladových pohybů tak, aby bylo možné dopředu určit vhodnou skladovou lokaci pro jejich umístění. S využitím prediktivního modelu rozhodovacího stromu s optimalizovanými parametry je nyní možné s 86,95 % pravděpodobností na základě atributů jednotlivých produktů určit, zda výrobek bude spadat do kategorie rychloobrátkového zboží, či do kategorie ostatního zboží.

V rámci tohoto projektu byly splněny tři dílčí cíle. Prvním dílčím cílem bylo sestavení základního souboru, který je vhodný pro prediktivní modelování. Tohoto cíle bylo dosaženo po agregaci několika databází získaných z informačního systému společnosti a jejich očištění od proměnných a záznamů, které vykazovaly multikolinearitu, či nebyly pro modelování důležité. Výsledný základní soubor měl podobu dvojrozměrné tabulky obsahující všechny produkty společnosti a jejich atributy, jako například jméno výrobce, číslo klienta, či počet skladových pohybů během roku. Na základě tohoto ročního počtu skladových pohybů byla stanovena hranice, podle které byl tréninkový soubor rozdělen na dvě kategorie, přičemž produkty vykazující více než 30 skladových pohybů byly kategorizovány jako rychloobrátkové a ostatní jako nerychloobrátkové (standardní).

Druhým dílčím cílem této práce bylo sestavení tří prediktivních modelů, jejich optimalizace a následné srovnání. V rámci plnění tohoto cíle byly sestaveny modely

logistické regrese, rozhodovacího stromu a neuronové sítě. Vzhledem k výkonu a aplikovatelnosti byl jako nejvhodnější model pro tento obchodní případ vyhodnocen model rozhodovacího stromu, vykazující celkovou přesnost 86,95 % při specifičnosti 90,17 % a citlivosti 75,76 %. Tento výsledek znamená, že model je schopen na základě atributů správně zařadit téměř devět nových produktů z deseti. V rámci zpracovávání rozhodovacího stromu byly také analyzovány faktory, které zařazení produktu ovlivňují nejsilněji. Tyto faktory byly v rámci stromu algoritmem umístěny na nejvyšší pozice a byly jimi zejména číslo projektu, do kterého produkt patří, číselné označení jeho dodavatele a iniciály produkčního manažera, který je za tento produkt zodpovědný.

Posledním dílčím cílem, který přispěl ke splnění cíle hlavního bylo v rámci tohoto projektu vypracování plánu praktického zavedení vybraného modelu do procesů firmy a jeho následná analýza z hlediska možnosti ušetření nákladů. Jak již bylo zmíněno výše v této práci, jako nejvhodnější model pro praktickou aplikaci byl vyhodnocen model rozhodovacího stromu, který by bylo možno zakomponovat do struktur SQL informačního systému společnosti. Pokud by k tomuto kroku došlo a byla zavedena nová skladová lokace sloužící pro uskladnění produktů, které prediktivní model vyhodnotí jako rychloobrátkové, bylo by možné zkrátit tok materiálu mezi jejich skladovacím kontejnerem a oblastí montáže, či balení. Na základě analýzy současných toků materiálů, průměrných manipulačních časů připadajících na jeden kontejner a průměrných hodinových nákladů na provoz skladové techniky a operátorů bylo vyhodnoceno, že zavedením nové skladové lokace může být ušetřeno průměrně 114,39 Kč v rámci jedné zakázky. Vezme-li se v potaz referenční objem zakázek roku 2018, bylo by možné ušetřit až 241 317,14 Kč ročně a tato částka může díky neustálému zdokonalování přesnosti predikcí dále růst.

Díky takovýmto drobným inovacím, které sofistikovaně zdokonalují interní proces firmy je možné neustále ladit celkovou efektivitu fungování společnosti. Dalším krokem může být aplikace komplexnějších modelů, které zvládnou rozdělit skladovací prostory na více než dvě oblasti a dokonale tak optimalizují tok materiálu mezi jednotlivými zónami. Prediktivní modelování může být také rozšířeno z prostředí skladového hospodářství dále do firmy, kde může sloužit k řešení nejrůznějších kategorizačních úloh a ušetření nákladů, což povede ke zlepšení celkové pozice společnosti v neustále dynamičtějším konkurenčním prostředí.

7 Seznam použitých zdrojů

HOLUBOVÁ, Irena, Jiří KOSEK, Karel MINAŘÍK a David NOVÁK. Big Data a NoSQL databáze. Praha: Grada, 2015. Profesionál. ISBN 978-80-247-5466-6.

GEWIRTZ, David. Volume, velocity, and variety: Understanding the three V's of big data. In: ZDNET [online]. USA: ZDNET, 2018, 2018 [cit. 2018-07-22]. Dostupné z: https://www.zdnet.com/article/volume-velocity-and-variety-understanding-the-three-vs-of-big-data/

VAN RIJMENAM, Mark. A Short History Of Big Data. Datafloq.com [online]. Netherlands: Datafloq, 2015, 7.1.2015 [cit. 2018-08-04]. Dostupné z: https://datafloq.com/read/big-data-history/239

ROSENBUSH, Steven. How Big Data Is Changing the Whole Equation for Business. In: Wall Street Journal online [online]. USA: Wall Street Journal, 2013, 2013 [cit. 2018-08-04]. Dostupné z: https://whatsthebigdata.files.wordpress.com/2013/03/bigdata_wsj.jpg

RAVINDRA, Savaram. Understanding the relationship between IoT and Big Data. Jaxenter.com [online]. USA: jaxenter, 2017, October 18, 2017 [cit. 2018-08-10]. Dostupné z: https://jaxenter.com/relationship-between-iot-big-data-138220.html

MARR, Bernard. A brief history of big data everyone should read. World Economic Forum [online]. Switzerland: World Economic Forum, 2015, February 25, 2015 [cit. 2018-08-10]. Dostupné z: https://www.weforum.org/agenda/2015/02/a-brief-history-of-big-data-everyone-should-read/

V.K. JAIN. Big Data and Hadoop. Indie: Khanna Publishing, 2017. ISBN 9789382609131.

LESLIE, Alexandra, 2018. What is Hadoop Good For?. HostingAdvice.com [online]. USA: HostingAdvice, 1.6.2018 [cit. 2019-02-23]. Dostupné z: https://www.hostingadvice.com/how-to/what-is-hadoop/

MANYIKA, James. Big Data: The next frontier for innovation, competition and productivity. McKinsey.com [online]. USA: McKinsey & Company, 2011, May 2015 [cit. 2018-08-19]. Dostupné z: https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

MCAFFEE, Andrew a Erik BRYNJOLFSSON. Big Data: The Management Revolution. Harvard Business Review [online]. 2012, **2012**(10), 10 [cit. 2018-08-19]. Dostupné z: https://pdfs.semanticscholar.org/02c7/740af5540f23a2da23d1769e64a8042ec62e.pdf

ABBOTT, Dean. Applied predictive analytics: principles and techniques for the professional data analyst. Indianapolis, IN: Wiley, [2014]. ISBN 1118727967.

SVOZILOVÁ, Alena. Zlepšování podnikových procesů. Praha: Grada, 2011. Expert (Grada). ISBN 978-80-247-3938-0.

MCCARTY, Tom. The Six Sigma black belt handbook. New York: McGraw-Hill, c2005, xiv, 588 p. ISBN 0071443290.

LABERGE, Robert. Datové sklady: agilní metody a business intelligence. Brno: Computer Press, 2012. ISBN 978-80-251-3729-1.

Motivity Business Intelligence: Gain Valuable Insight to Improve Business Goals [online], 2018. USA: Motivity Solutions [cit. 2019-02-23]. Dostupné z: https://motivitysolutions.com/business-intelligence/

STŘELEC, Jiří. DMAIC metoda. Vlastní Cesta [online]. Brno: Vlastnicesta.cz, 2012, 2012 [cit. 2018-09-03]. Dostupné z: https://www.vlastnicesta.cz/metody/dmaic-metoda-1/

FOGARTY, David John. Lean Six Sigma and Big Data: Continuing to Innovate and Optimize Business Processes. Journal of Management and Innovation [online]. New York City: University of Phoenix, 2015, 2015 [cit. 2018-09-03]. Dostupné z: http://jmi.mercy.edu/index.php/JMI/article/view/8

Predictive Analytics vs Data Mining - Which One Is More Useful: Head to Head Comparison between Predictive Analytics vs Data Mining. In: Www.educba.com [online]. Mumbai: eduCBA, 2018, 2018 [cit. 2019-01-24]. Dostupné z: https://www.educba.com/predictive-analytics-vs-data-mining/

CHAPMAN, Pete. CRISP-DM 1.0: Step-by-step data mining guide. The Modeling Agency [online]. Pittsburgh: One Oxford Centre, 2000, 2000 [cit. 2019-01-26]. Dostupné z: https://www.the-modeling-agency.com/crisp-dm.pdf

MONTGOMERY, Douglas C. Introduction to linear regression analysis. 5. ed. Oxford: Wiley-Blackwell, 2011. ISBN 9780470542811.

FREUND Rudolph, William J WILSON a Ping SA. Regression Analysis. 2nd ed. Burlington: Elsevier, 2006. ISBN 9780080522975.

HASSOUNA, Mohammed, Ali TARHINI, Tariq ELYAS a Mohammad Saeed ABOU TRAB, 2015. Customer Churn in Mobile Markets: A Comparison of Techniques. International Business Research [online]. 8(6), - [cit. 2018-01-22]. DOI: 10.5539/ibr.v8n6p224. ISSN 1913-9012. Dostupné z: http://www.ccsenet.org/journal/index.php/ibr/article/view/47593

SAS. Decision Trees for Business Intelligence and Data Mining, 2008. In: SAS Support [online]. SAS [cit. 2017-12-11]. Dostupné z: https://support.sas.com/publishing/pubcat/chaps/57587.pdf

LAMBERT, Douglas M. a Lisa M. ELLRAM. Logistika: příkladové studie, řízení zásob, přeprava a skladování, balení zboží. Praha: Computer Press, 2000. Business books (Computer Press). ISBN 80-7226-221-1.

STARMER, Josh, 2018. StatQuest.org. In: StatQuest.org: An epic journey through computational methods for molecular genetics [online]. USA: StatQuest [cit. 2019-02-26]. Dostupné z: https://statquest.org/video-index/

NAVLANI, Avinash, 2018. Decision Tree Classification in Python. In: DataCamp [online]. USA: DataCamp, 2018 [cit. 2019-02-26]. Dostupné z: https://www.datacamp.com/community/tutorials/decision-tree-classification-python

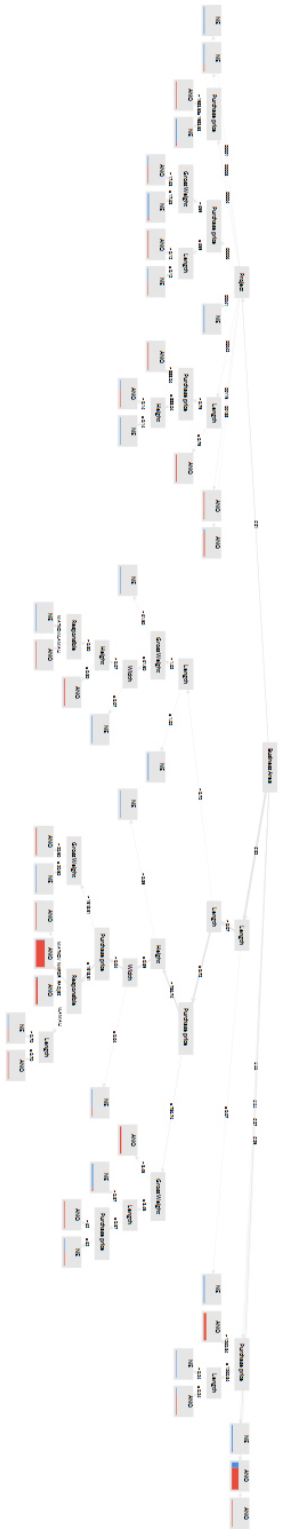
EIBE, Frank a Ian H. WITTEN, 2005. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition: Morgan Kaufmann Series in Data Management Systems. 2. USA: Elsevier. ISBN 978-0120884070.

Motivity Business Intelligence: Gain Valuable Insight to Improve Business Goals [online], 2018. USA: Motivity Solutions [cit. 2019-02-23]. Dostupné z: https://motivitysolutions.com/business-intelligence/

What is predictive analytics [online], 2014. Toronto, Canada: PAT Research [cit. 2019-03-14]. Dostupné z: https://www.predictiveanalyticstoday.com/what-is-predictive-analytics/

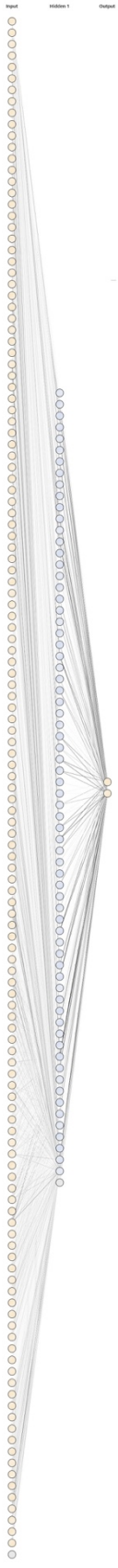
8 Přílohy

Příloha číslo 1 – schéma modelu prvního rozhodovacího stromu.



Příloha číslo 2 – schéma modelu optimalizovaného rozhodovacího stromu.

Příloha číslo 3 – schéma neuronové sítě (200 tréninkových cyklů, 1 skrytá vrstva).



Příloha číslo 4 – schéma neuronové sítě (500 tréninkových cyklů, 2 skryté vrstvy).