Univerzita Pardubice

Fakulta ekonomicko-správní

Analýza výstupů informačního systému pro podporu rozhodování managementu

Ivona Barvová

Bakalářská práce
2011

Univerzita Pardubice
Fakulta ekonomicko-správní
Akademický rok: 2010/2011

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Ivona BARVOVÁ**
Osobní číslo:	**E08929**
Studijní program:	**B6209 Systémové inženýrství a informatika**
Studijní obor:	**Regionální a informační management**
Název tématu:	**Analýza výstupů informačního systému pro podporu rozhodování managementu.**
Zadávající katedra:	**Ústav systémového inženýrství a informatiky**

Zásady pro vypracování:

Základní pojmy z oblasti data miningu a informačních systémů.
Shromáždění a příprava dat pro modelování a popis procesů s využitím data miningu.
Modelování dat.
Vyhodnocení výsledků a doporučení pro rozhodování managementu.

Rozsah grafických prací:

Rozsah pracovní zprávy:

Forma zpracování bakalářské práce: **tištěná/elektronická**

Seznam odborné literatury:

Basl, Josef; Blažíček, Roman. Podnikové informační systémy. Praha : Grada, 2008. 283 s. ISBN 978-80-247-2279-5.

Berka, Petr. Dobývání znalostí z databází. Praha : Academia, 2003. 366 s. ISBN 80-200-1062-9.

Dostál, Petr; Rais, Karel; Sojka, Zdeněk. Pokročilé metody manažerského rozhodování. Praha : Grada, 2005. 166 s. ISBN 80-247-1338-1.

Rud, Olivia Parr. Data Mining. Praha : Computer Press, 2001. 329 s. ISBN 80-7226-577-6.

Vedoucí bakalářské práce: **Ing. Jana Filipová**
Ústav systémového inženýrství a informatiky

Datum zadání bakalářské práce: **4. října 2010**

Termín odevzdání bakalářské práce: **6. května 2011**

L.S.

doc. Ing. Renáta Myšková, Ph.D.
děkanka

doc. Ing. Jiří Křupka, Ph.D.
vedoucí ústavu

V Pardubicích dne 4. října 2010

Prohlášení

Tuto práci jsem vypracovala samostatně. Veškeré literární prameny a informace, které jsem v práci využila, jsou uvedeny v seznamu použité literatury.

Byla jsem seznámena s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorský zákon, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

Souhlasím se zveřejněním své práce v Univerzitní knihovně.

V Pardubicích dne 6.5.2011

Ivona Barvová

Poděkování

Ráda bych touto cestou poděkovala vedoucí své bakalářské práce Ing. Janě Filipové za odborné vedení, cenné připomínky, náměty, rady a čas věnovaný konzultacím.

Zároveň také děkuji mým nejbližším, Pavlovi, Davidovi a Janovi. Bez jejich nesmírné pomoci, trpělivosti, podpory, pochopení a tolerance by tato práce nikdy nevznikla.

Anotace

Bakalářská práce je zaměřena na využití metodiky CRISP-DM v konkrétní firmě k získání nových poznatků a informací s využitím dat uložených v databázi v podnikovém informačním systému. Práce popisuje jednotlivé dílčí etapy metodiky, jejich uplatnění v praxi při realizaci konkrétního dataminingového projektu. Závěrečné vyhodnocení může využít management firmy ke stanovení nových obchodních strategií a cílů firmy.

Annotation

The Bachelor thesis is focused on the use of the methodology CRISP-DM in a particular company to acquire new knowledge and information with data stored in a database in an enterprise information system. This work describes separate partial stages of methodology - their practical application during a specific data mining project. The final evaluation can be used by company's management to set new business strategies and company objectives.

Klíčová slova

Data Mining, dolování z dat, CRISP-DM, modelování dat, podnikový informační systém.

Key words

Data Mining, CRISP-DM, Data Modelling, Enterprise Information System.

Obsah

Úvod

Bakalářská práce na téma Analýza výstupů informačního systému pro podporu rozhodování managementu se zabývá využitím metodiky CRISP-DM, jejích nástrojů a postupů při hledání zajímavých skutečností ve velkých datových souborech uložených v informačním systému podniku.

Cílem práce je při využití konkrétních dat skutečně existující firmy ukázat, jaké poznatky, informace, vztahy, vzájemné závislosti lze z podnikového informačního systému získat za pomoci této metodologie a využít je v praxi.

U mnoha firem management při řízení společností při svém rozhodování zdaleka nevyužívá všech informací, které se skrývají v útrobách jejich informačního systému, ale jedná na základě momentálního rozhodnutí. Převládá stále názor, že oddělení IT nemá pro řízení společnosti žádné zajímavé poznatky, že se jedná o útvar, který stále jen požaduje finanční prostředky na rozvoj a inovace v oblasti IT. Ale je tomu právě naopak. Všechny podstatné informace o firmě a jejím vztahu k dodavatelům, zákazníkům (i potencionálním), veškerá historie jednotlivých nabídek, komunikace, realizace a následně i fakturace, jsou v informačním systému firmy.

A právě na toto je bakalářská práce zaměřena. Provedená analýza dat ukazuje, jaké důležité a zajímavé informace podnikový informační systém skrývá. Zjištěné vazby a vztahy mezi daty by mohly sloužit managementu firmy při rozhodování o dalších cílech a směrech vývoje firmy, při obchodních aktivitách společnosti.

1 Základní pojmy z oblasti Data Miningu a informačních systémů

1.1 Základní definice

Nejzákladnější pojmy z podnikové informatiky jsou definovány takto:

„**Informace** je zpráva o nastalém jevu, který u nás (příjemců) snižuje míru neznalosti o tomto jevu.“ [8, s.23]

„**Systém** je účelově definovaná neprázdná množina prvků a množina vazeb mezi nimi, přičemž vlastnosti prvků a vazeb mezi nimi určují vlastnosti (chování) celku.“ [8, s.23]

„**Informační systém (IS)** představuje konzistentní uspořádanou množinu komponent spolupracujících za účelem tvorby, shromažďování, zpracování, přenášení a rozšiřování informací. Prvky informačního systému tvoří lidé, respektive uživatelé informací, a informatické zdroje. Komponenta je tvořena jedním prvkem nebo více prvky.“ [8, s.25]

„**Proces** je definován jako soubor vzájemně souvisejících nebo vzájemně působících činností, který přeměňuje vstupy na výstupy. Činnosti využívají zdrojů (lidí, nástrojů, materiálů apod.). Proces může mít více vstupů a také více výstupů.“ [8, s.25]

„Protože podniková informatika se soustřeďuje na podnik, pak tento systém označujeme jako informační systém v podniku anebo **podnikový informační systém (PIS)**. Jeho účel, respektive cílové chování je dáno základním požadavkem podniku na soulad ICT a podnikových procesů, resp. na adekvátní podporu podnikových procesů informačními a komunikačními technologiemi (business-IT alignment). V současné době je podnikový IS často i nositelem nových obchodních příležitostí, nové podoby podnikání nebo zvyšování celkové efektivity podniku. Prvky podnikového informačního systému jsou lidé, ICT a data.“ [8, s.28]

„**Data** (podniková data), jakožto prvek podnikového informačního systému představují zaznamenaná fakta o všech podstatných skutečnostech, které souvisejí s aktivitami podniku.“ [8, s.29]

„**Data Mining (DM)** lze charakterizovat jako proces extrakce relevantních, předem neznámých nebo nedefinovaných informací z velmi rozsáhlých databází. Důležitou vlastností dolování z dat je, že se jedná o analýzy odvozené z obsahu dat, nikoli analýzy předem specifikované uživatelem, a jedná se především o odvozování prediktivních informací, nikoli pouze deskriptivních.“ [8, s.230]

Pojmy obsažené v předchozích definicích hýbou v dnešní době společností. Aniž si to v reálném životě uvědomujeme, IS se významně podílejí na našem každodenním životě, nejen

v podnicích, ale i v domácnostech. Rychle a podstatně se mění výrobní i nevýrobní technologie, výrobky, služby, a tím i pracovní postupy, metody a přístup k podnikovým procesům. Tyto změny se musí nutně projevit i v potřebě získávat data a informace, uchovávat je, následně je vhodně využívat a tím podpořit efektivní a dynamický rozvoj firmy.

Snahou všech firem, zabývajících se vývojem a prodejem hardware či software je, aby svým zákazníkům poskytovaly co možná nejlepší technické a programové vybavení. Aby svá data a důležité informace měli stále vhodně dostupné a to ve formě, v jaké jsou pro ně přijatelné, srozumitelné a využitelné.

Dnes bychom již opravdu hledali těžce firmu, která s nějakým informačním systémem nepracuje. Pokud se nejedná zrovna o některé z manažerských či rozhodovacích IS, tak rozhodně firma využívá některý z velkého množství na trhu dostupných ekonomických či účetních informačních systémů.

Současná doba přímo nahrává změnám a rozvoji informačních systémů v podnicích. Finanční prostředky na obnovu a inovace informačních a komunikačních technologií lze získat i z dotačních rozvojových a operačních programů fondů Evropské unie či s podílem státního rozpočtu. O jednotlivé dotace se stará Ministerstvo průmyslu a obchodu ve spolupráci s firmou CzechInvest. Příjem a posuzování žádostí o finanční prostředky je stále aktuální a je využíván nejen soukromými malými či středními podniky, ale i státní sférou, pro kterou je to velká příležitost pomoci si k lepšímu technickému vybavení v oblasti ICT.

1.2 Podnikový informační systém

Podnikový informační systém je jeden z mnoha možných typů informačního systému. Je to systém konkrétního podniku, jehož cílem je správa informací mapující podnikové procesy za použití informačních a komunikačních technologií. Mezi PIS můžeme zařadit např. systémy řízení vztahu se zákazníky, účetní systémy, řídicí systémy podniku, manažerské systémy a další. Bývá přizpůsoben požadavkům konkrétního podniku. Zohledňuje charakter jeho podnikatelské činnosti, odvětví, typ výroby, jeho velikost a zároveň i procesní a organizační způsoby řízení. [1]

Důvodů, proč se již v praxi víceméně skoro nikde nepoužívá papírová kartotéka, účetní deník ve formě knihy či psací stroj, je celá řada. Rychlost, přesnost, jednoduchost, efektivnost, to jsou charakteristické vlastnosti, proč firmy neustále se snaží zlepšovat své PIS. Kdyby tak nečinily, jejich konkurenceschopnost, vztahy se zákazníky, řízení výrobních technologií a tím i jejich výrobky a služby by nenašly uplatnění na trhu a jeden z hlavních cílů každého podniku, tvorba zisku, by nemohl být naplněn.

Z výše uvedeného je zřejmé, že data, databáze, informace v nich uložené, získávání informací, zpracování informací, to jsou v současném konkurenčním prostředí velmi důležité pojmy. Trvale vzrůstá objem jednotlivých firemních obchodních či průmyslových databází. Většina podniků, zejména jejich manažeři, si ani pořádně neuvědomuje, jaká všechna data ve svých informačních systémech skrývají a jak je vhodně využít ke stanovení dalších cílů, strategií, k odhalení slabých míst podnikových procesů. Z různých dat je možné vyčíst velké množství informací, jejich zpracování má při dostupnosti dnešních informačních technologií nejrůznější formy. [15]

A právě Data Mining je možno vnímat jako jeden z možných analytických nástrojů pro transformaci datových zdrojů na informace, které ovlivní obchodní strategii, cíle, organizaci řízení i rozhodování jednotlivých firem. Nalézání souvislostí v datech, které nejsou na první pohled zřejmé, můžeme nazvat dolováním z dat. [7]

1.3 Data Mining

Data Mining je analytická metoda získávání skrytých, ale přesto užitečných, informací z dat. Je to jistý proces výběru, prohledávání a modelování velkých objemů dat, k vyhledání skrytých závislostí a k odhalení dříve neznámých vztahů mezi daty. [2]

Již od svého vzniku na konci osmdesátých a počátku devadesátých let dvacátého století, kdy byla nejvíce využívána bankovním sektorem, se rychle rozšiřovala mezi další obory. V současné době se uplatňuje nejvíce v pojišťovnictví, i nadále v bankovnictví, ve veřejných službách, telekomunikacích, energetice, maloobchodě, zásilkových službách, farmaceutickém průmyslu či cestovním ruchu. [2, 15]

Pro zjednodušení práce uživatelů při řešení různých úloh v oblasti získávání dat z databází začaly vznikat metodologie, jejichž cílem bylo poskytnout návod, průvodce, zkušenosti z již realizovaných projektů. Některé metodologie byly vytvořeny softwarovými výrobci, např. metodologie SEMMA od firmy SAS nebo 5A firmy SPSS. Zároveň ale vznikaly i metodologie nezávislých výrobců, jako např. CRISP-DM (zkratka CRISP-DM znamená Cross-Industry Standard proces for Data Mining). [6]

Metodologie CRISP-DM se snaží nalézt univerzální postup pro dobývání z dat. Krok za krokem prochází postupně celý proces projektu a již vyzkoušenými postupy se snaží efektivně a co nejrychleji řešit zadaný problém. [10]

Tato metodologie dle [2] rozděluje proces dobývání znalostí z databází do šesti základních etap:

- Porozumění problematice
- Porozumění datům
- Příprava dat
- Modelování
- Vyhodnocení výsledků
- Využití výsledků

V první etapě, porozumění problematice, se jedná o pochopení cílů úlohy, stanovení požadavků z manažerského hlediska a jejich transformace do úlohy vhodné k řešení z pohledu dobývání z dat v databázích. Již v této fázi je vhodné stanovit předběžný plán a soupis prací. [2, 13]

Druhá etapa je porozumění datům. Samotný název je dostatečně výstižný. Data je třeba nejprve získat a následně si o nich utvořit představu z pohledu jejich kvality, případně zjistit i jejich deskriptivní charakteristiky. Velmi vhodné je i nahlédnutí do dat či pouhého vzorku dat některým z běžných nástrojů, jako jsou např. dnes velmi rozšířené programy Microsoft Office Excel či Access. [7]

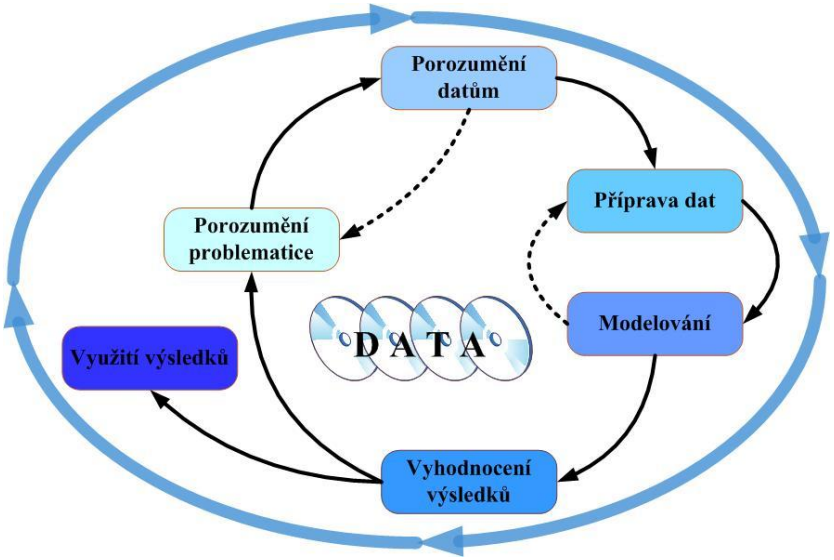
Samotná příprava dat, třetí z etap metodologie CRISP-DM, již zahrnuje činnosti vedoucí k vytvoření takového datového souboru, který bude připraven pro zpracování některou z modelovacích technik. Z dat je vhodné selekcí vybrat pouze ta, která se jeví jako potřebná, provést jejich čištění, transformaci, integraci, vytvořit nové odvozené datové proměnné a následně celý datový soubor zformátovat do využitelného tvaru. Tato fáze je časově nejnáročnější ze všech. Některé z úkonů je třeba provádět i opakovaně. [15]

Existují různé modelovací metody v oblasti Data Miningu. Pro každý konkrétní typ úlohy je třeba vybrat tu nejvhodnější, případně použít více metod a výsledky pak porovnávat, analyzovat, kombinovat. Mezi nejčastěji používané analytické metody patří např. rozhodovací stromy, asociační pravidla, rozhodovací pravidla, neuronové sítě, či některá ze statistických metod, jako např. regresní analýza, shluková analýza či diskriminační analýza. Modelování je čtvrtou etapou metodologie CRISP-DM. [10]

V páté etapě, vyhodnocení výsledků, je prováděno hodnocení dosažených výsledků dobývání z dat na základě cílů, které byly stanoveny v první etapě. Pokud cíle nebylo uspokojujícím způsobem z pohledu manažerského dosaženo, celý proces se opakuje od etapy jedna, porozumění problematice. Pokud bylo dosaženo cíle, mělo by být rozhodnuto o využití výsledků. [2, 13]

Na toto rozhodnutí navazuje i celá šestá etapa metodologie CRISP-DM. Proces využití výsledků znamená, že je sepsána projektová zpráva a celý projekt je zhodnocen. Podoba zprávy by měla být čitelná pro zadavatele úlohy, aby bylo na první pohled jasné, jaké kroky je třeba učinit, aby dosažené výsledky byly využity co možná nejefektivněji. Sepsání projektové, závěrečné zprávy patří k sice nepopulárním činnostem v reálném životě, ale její význam je následně oceněn při řešení další problematiky z oblasti dobývání z dat. [2, 13]

Propojení a návaznosti jednotlivých fází metodologie CRISP-DM jsou znázorněny na následujícím obrázku 1.



Obrázek 1 - Fáze CRISP-DM (zdroj: upraveno podle [2])

Data Mining najde vhodné uplatnění zejména tam, kde je třeba vyhledat zajímavé charakteristiky mezi daty. Při velkých objemech dat s velkým množstvím atributů, či při tvorbě modelu, kdy na základě již existujících dat jsou předem hodnocena data teprve vznikající. [14]

Jednou z takovýchto oblastí informačního systému firmy můžeme být modul řízení vztahů se zákazníky neboli modul pro řízení obchodních procesů ve firmě. Bývají zde zachyceny veškeré obchodní aktivity, stavy zakázky či projektu, komunikace se zákazníkem i finální plnění zakázek. Praktickým příkladem je třeba analýza prodeje, vývoj nového produktu, marketingová kampaň, analýza rizik, nebezpečí zneužití či předpoklad možné ztráty zákazníka. Největším přínosem samozřejmě je, že veškeré informace jsou soustředěny na jednom místě. [15]

Trh softwarových nástrojů v oblasti DM se neustále vyvíjí, je to dynamicky se rozvíjející oblast. Pro svůj praktický příklad jsem si vybrala jeden z produktů společnosti

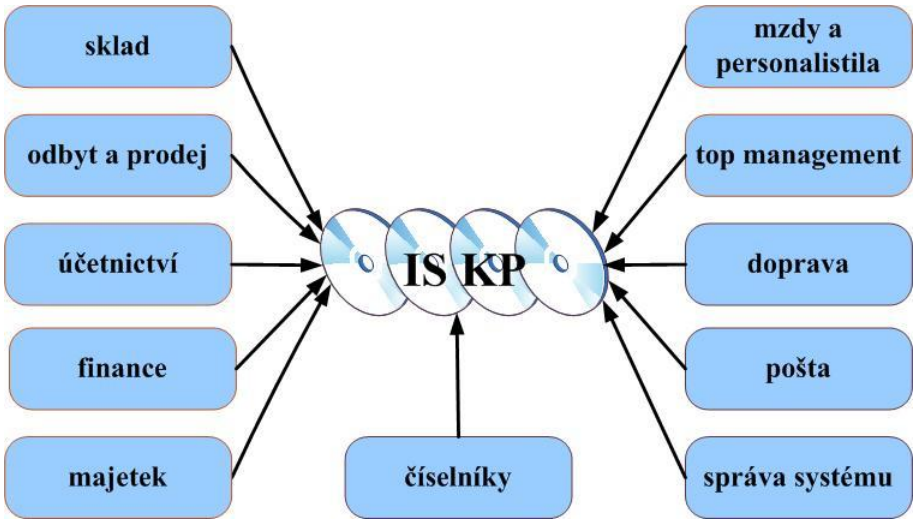
SPSS Inc. se zastoupením pro Českou republiku SPSS ČR, spol. s r.o., a to analytický nástroj Clementine, který je dostupný na počítačových učebnách Fakulty ekonomicko-správní Pardubice. V současné době, po té, co byla společnost SPSS odkoupena firmou IBM, byl produkt Clementine nahrazen softwarem IBM SPSS Modeler. Jedná se o moderní intuitivní program pro Data Mining poskytující popis a vizualizaci datových vztahů, analýzu příčin a pro rozhodování. [10]

1.4 Informační systém KP

Konkrétním IS, jehož data byla využita pro praktickou realizaci užití metodiky CRISP-DM, je ekonomický informační systém KP. Jedná se o zakázkově řešený parametrizovaný IS pro řízení malých a středních firem.

Jak uvádějí internetové stránky firmy, která IS KP vyvinula a dodává, jednotlivé moduly informačního systému KP jsou zobrazeny na obrázku 2 a jejich dílčí rozpad na podmoduly je následují:

- skladové hospodářství – poptávky, objednávky, dodavatelské faktury, skladové karty, skladové uzávěrky, nákupní ceníky,
- odbyt a prodej – nabídky, zakázky, odběratelské faktury, dodací listy, prodejní ceny,
- účetnictví – příprava a účtování jednotlivých dokladů, účetní výkazy, roční závěrka,
- finance – pokladní a bankovní transakce, příkazy k úhradě, neuhrazenost,
- mzdy a personalistika – mzdové a personální údaje, zpracování mezd včetně zákonem stanovených povinností, roční zúčtování daně,
- majetek – evidence hmotného, nehmotného, drobného i finančního majetku, daňové a účetní odpisy,
- doprava – evidence vozidel a jízd, rezervace,
- pošta – evidence došlé a odeslané pošty,
- top management – náklady a výnosy, manažerské analýzy,
- správa systému – nastavení jednotlivých uživatelů a přístupových práv, jazykové varianty, slovník textů,
- číselníky – adresář partnerů, dokladové řady, poštovní směrovací čísla, střediska, atd.



Obrázek 2 - Informační systém KP (zdroj: vlastní)

Informační systém KP pracuje na architektuře klient/server založené na datové základně Microsoft SQL Server umožňující zpracování údajů v reálném čase.

Nastavení přístupových práv jednotlivým pracovníkům firmy zajišťuje, že uživatel může nahlížet, pořizovat či opravovat pouze určitou skupinu dokladů, stejně jako je mu umožněno spouštět pouze některé činnosti. Již samotné názvy modulů naznačují, které z nich budou mít pouze velmi omezený počet uživatelů. Modul mzdy a personalistika či top management jistě nebudou patřit mezi moduly určené běžným uživatelům.

Data jsou v informačním systému KP uložena dle různého typu použitelnosti následovně:

- parametry – obsahují hodnoty určené k fungování celého IS,
- číselníky – obsahují údaje využitelné ve více modulech,
- základní data z jednotlivých modulů – nemají konkrétní vazbu na zakázku,
- konkrétní data vztahující se k jednomu zakázkovému případu,
- archivní data – z již realizovaných a uzavřených zakázek.

1.5 Proces rozhodování managementu

Vztahy mezi IS firmy a podnikovými procesy jsou velmi těsné. Jakákoliv změna firemního procesu vyvolává změny v informačním systému, stejně jako naopak, je-li změněna funkčnost IS, nutně musí dojít i k procesní změně. Provázanost jednotlivých procesů je následně zohledněna i v provázanosti jednotlivých modulů IS. [5]

V běhu posledních let se mění i způsob procesu rozhodování a požadavků na něj. Stále obtížněji se odhaduje chování zákazníků, vývoj trhu, roste nabídka počtu výrobků a služeb, požadavky zákazníků jsou specifičtější.

Chce-li management vést svoji firmu tak, aby uspěla na trhu, musí si jasně definovat svoje hlavní úkoly, dát všem činnostem jasný cíl a strategii. Strategie se týká i efektivního využívání IS. Aby veškeré údaje z IS managementu sloužily právě k naplňování těchto cílů, je potřebné mít všechny informace o stavu a potřebách podniku, o finančních ukazatelích, o aktuálním stavu a průběhu zakázek v přijatelné formě. Management má nejraději grafické vyjádření, a pokud možno v reálném čase. Jen tak může IS napomoci k strategickým rozhodnutím, která směřují ke zvýšení konkurenceschopnosti, výkonnosti, kvality podnikového řízení a v neposlední řadě k dobrému jménu firmy či ke zlepšení ekonomických ukazatelů firmy.

2 Shromáždění a příprava dat pro modelování

2.1 Porozumění problematice

První fáze metodiky CRISP-DM se zaměřuje na pochopení manažerských požadavků a cílů úlohy v rámci projektu a jejich transformaci do takového zadání, kterému porozumí specialisté – experti na dataminingové zpracování dat.

A právě o vzájemné komunikaci těchto dvou zúčastněných osob v rámci projektu je hlavně tato fáze. Manažer vidí potřebu získat informace o svých obchodních aktivitách, ale nemá reálnou představu o tom, z jakých dat je to realizovatelné. Proto i stanovení cíle je ve slovníku manažerské terminologie. Naopak dataminingový specialista, který se nevěnuje obchodním strategiím, ale ovládá analytické metody a ví, jak z dostupných dat na požadavky manažera odpovědět, které analýzy vhodně použít. Je schopen manažerskou terminologii převést do terminologie pro dataminingové zpracování. Definovaný cíl projektu musí být společným dílem obou hlavních zúčastněných. [2, 15]

Dle [2] se zároveň vyhodnotí předpokládané náklady a přínosy projektu, potřebnost jednotlivých zdrojů, ať již lidských nebo technických, datových či v neposlední řadě finančních, a připraví se předběžný časový plán projektu a upozorní na případná možná rizika.

Pro praktický příklad využití metodiky v konkrétní firmě byly po konzultaci s managementem firmy navrženy pro zpracování následující požadavky – úkoly – zadání.

2.2 Stanovení jednotlivých cílů

Hlavní náplní firmy je poskytování služeb v oblasti informačních technologií a s tím související obchodní aktivity. Její dodávky jsou určeny zákazníkům zejména v těchto oblastech činností, rozdělených do kategorií:

- finanční instituce,
- průmysl,
- služby,
- státní správa,
- stavebnictví,
- školství,
- zdravotnictví.

V této práci jsou jednotliví zákazníci zařazeni do kategorií s názvem kateg1 – kateg7, zákazník nezařazený v žádné kategorii je označen dvěma pomlčkami (--). Pořadí jednotlivých kategorií neodpovídá výše uvedenému seznamu.

Struktura nabízených služeb se dá rozdělit do následujících podskupin, sortimentů:

- dodávky aktivních prvků,
- dodávky hardware,
- dodávky řídicích systémů,
- dodávky software,
- návrhy, výstavba a komplexní dodávky sítí LAN a WAN,
- prodej materiálu,
- služby – servisní a systémová podpora.

Pro možnost prezentace této práce jsou jednotlivé sortimenty pojmenovány sort1 až sort7, pořadí neodpovídá výše uvedenému seznamu.

Management firmy se samozřejmě zajímá o to, v jakém odvětví jsou realizovány tržby za poskytované činnosti. Informační systém firmy je schopen takovéto údaje vyhodnotit. Ale již určení přímého vztahu mezi dodávaným sortimentem služeb a odvětvím na straně zákazníka, to již informační systém vyhodnotit neumí. A přesto je tato znalost pro obchodní aktivity firmy jednou z důležitých. Vědět, o jaké služby je v které oblasti života zájem a nabídky těchto služeb směřovat právě těmto vybraným zákazníkům.

Cílem je tedy zjistit vzájemné vazby – vztahy – souvislosti mezi kategorií zákazníků a sortimentem dodávaných služeb. Podpůrným hlediskem by měla být i období, ve kterém jsou služby poskytovány, a samozřejmě výše fakturované částky.

Důležitým atributem při pohledu na zákazníka je jeho platební morálka, schopnost hradit včas své finanční závazky vůči firmě.

Informace o výši finančních prostředků firmy patří vždy k nejsledovanějším a pro management rozhodně i nejdůležitějším. V praktickém životě to znamená, aby firma měla dostatek finančních prostředků na úhradu svých závazků i v okamžiku, kdy nejsou v řádných termínech splatnosti plněny pohledávky za jednotlivá poskytnutá plnění zákazníkům.

Zjištění souvislostí mezi fakturačním obdobím, datem následné úhrady faktury, případně i kategorií zákazníka, u kterých dochází nejčastěji ke zpožďování plateb, jistě přispějí k finanční stabilitě firmy. Již při uzavírání nových obchodních smluv je možné stanovit s konkrétním zákazníkem fakturační podmínky tak, aby nedocházelo k pozdním úhradám.

Vzhledem k okolnostem, které v ekonomice nejen České republiky proběhly v posledních měsících až roce, kdy chování zákazníků na trhu je ovlivněno ekonomickou

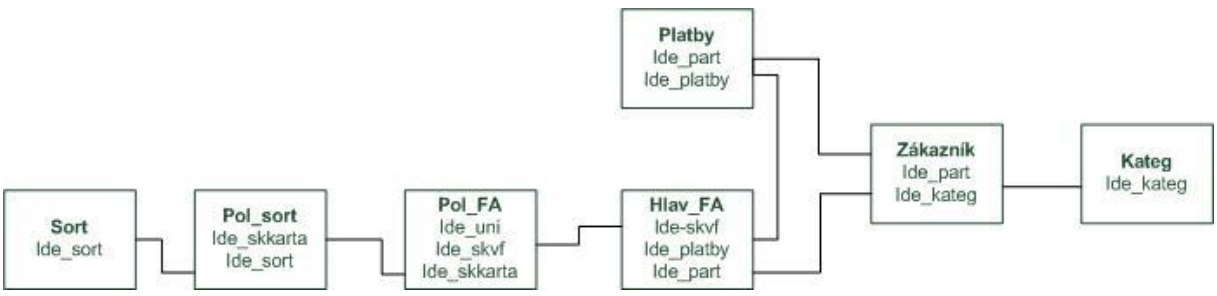
situací celé společnosti, musí i firma pružně reagovat na měnící se potřeby jednotlivých kategorií zákazníků. I proto je datová analýza zaměřena na srovnání období let 2008-2010.

2.3 Shromáždění dat

Data jsou v IS KP uložena v SQL databázi. Pro výběr vhodných dat je potřebné v tabulkách příslušné databáze vybrat ta, které obsahují informace vhodné pro zpracování metodikou CRISP-DM. Zároveň musí mezi tabulkami existovat vzájemné propojení, provázanost.

2.3.1 Výběr vhodných dat

Z obrázku 3 je patrné, že základní tabulkou je Zákazník a s ním jsou svázány další tabulky. Každý zákazník je zařazen v některé z kategorií podle odvětví svého působení. Jednotlivé kategorie jsou obsahem tabulky Kateg. Zákazník je zařazen pouze do jedné kategorie. Pakliže působí ve více odvětvích, je vybráno to, které je u něj uváděno jako hlavní podnikatelská činnost. V rámci obchodních aktivit firmy jsou zákazníkům vystavovány za poskytnuté zboží či služby faktury. Hlavičky jednotlivých faktur jsou v tabulce Hlav_FA. Zákazník ve sledovaném období nemusí mít žádnou fakturu, může mít pouze jednu či více faktur. Každá faktura obsahuje položky za jednotlivá poskytnutá plnění. Tyto údaje jsou obsaženy v tabulce Pol_FA. Hlav_FA musí obsahovat alespoň jednu položku. Položka faktury může dle svého typu patřit mezi sortimentní položky. Pak je obsahem tabulky Pol_sort a její konkrétní zařazení do sortimentu poskytovaných činností a služeb je přiřazeno pomocí tabulky Sort. Každá sortimentní položka musí být zařazena v některém sortimentu. Dále je na obrázku 3 ještě znázorněna tabulka Platby. Informace o platebním styku se zákazníkem jsou shromažďovány jak přímo v návaznosti na tabulku Zákazník, tak musí být propojeny s Hlav_FA jako informace o úhradě konkrétní faktury. Datové struktury jednotlivých tabulek jsou uvedeny v příloze A.



Obrázek 3 - Propojenost tabulek IS KP (zdroj: vlastní)

2.3.2 Porozumění datům

Základní datový soubor pro jednotlivé tabulky byl vždy získán exportem datového obsahu tabulek pomocí Query Analyzeru přímo z SQL serveru, na kterém se data IS KP nacházejí.

Před přípravou dat, která budou vstupovat do modelovacích technik v programu Clementine, je nutné data správně pochopit hlavně s cílem a zaměřením pouze na atributy a jejich vlastnosti, které napomohou k vytvoření modelů.

Pro první představu o podobě, tvaru, formátu a rozsahu dat byla takto získaná data naimportována do prostředí Microsoft Office Excel. Zde již bylo možné prvotně rozlišit, které atributy jednotlivých tabulek jsou zde nadbytečné pro další zpracování, neboť se jedná o systémové atributy informačního systému KP. Tyto atributy, které jsou pro dataminingový projekt nevyužitelné, je možné odstranit.

Vytvoření jednotlivých souborů dat patří k nejnáročnější činnosti metodologie CRISP-DM. Jejich kvalita podstatně napomůže k naplnění a dosažení cílů projektu.

2.3.3 Tabulka Zákazník

Tabulka Zákazník obsahuje široké spektrum údajů, které jsou potřebné u zákazníků evidovat. Jedná se zejména o údaje napomáhající splnění základních zákonných povinností firmy při vystavování daňových dokladů, jako například:

- IČO a DIČ,
- přesný název a adresa firmy.

Další údaje už jsou využitelné v rámci funkčnosti IS a tudíž nezbytné pro jeho správný provoz. Sem patří údaje typu:

- označení dealera, tj. pracovníka, který se o konkrétního zákazníka stará,
- typ úhrady,
- standardní délka splatnosti pro zákazníka,
- je-li zákazník kvalifikován v rámci ISO.

Tabulka obsahovala další údaje, které byly shledány jako nepotřebné pro tento dataminingový projekt a proto byly z tabulky vyřazeny. Jedná se např. o atribut:

- poznámka,
- rabat,
- cenová skupina.

Tabulka Zákazník ke dni exportu dat z IS obsahovala 4256 záznamů.

2.3.4 Tabulka Kategorie

Tabulka Kategorie (Kateg) je velmi jednoduchá. Jsou zde pouze atributy Označení kategorie a Název kategorie. Jedná se o rozčlenění, ze kterého odvětví či oblasti poskytovaných služeb zákazník pochází.

Ke dni exportu obsahovala 8 záznamů, nezařazený zákazník je označen dvěma pomlčkami (--).

2.3.5 Tabulka Platby

Tabulka Platby obsahuje údaje o tom, ve kterém období byla faktura na poskytnutá plnění vystavena, jaký byl termín její splatnosti a datum, kdy byla zákazníkem skutečně uhrazena. Nechybí samozřejmě atributy jako:

- částka
- identifikační číslo zákazníka,
- identifikační číslo faktury.

Naopak opět některé nepotřebné atributy byly odstraněny, jako například:

- interní číslo dokladu, kterým byla provedena platba,
- kód banky a bankovní účet, na který byla platba provedena.

Z datové struktury uvedené v příloze A je patrné, že nechybí provázanost přes ide_zak00000 na zákazníka a ide_platba00000 na konkrétní vystavenou fakturu.

Ke dni exportu tabulka obsahovala 12893 záznamů.

2.3.6 Tabulka Hlavička faktury

Tabulka Hlavička faktury (Hlav_FA) obsahuje údaje rozhodné pro vystavení faktury. Nejdůležitější je samozřejmě provázanost na tabulku Zákazník, ze které jsou přebírány potřebné fakturační údaje. Neméně důležitým je datum uskutečnění zdanitelného plnění a datum splatnosti. Další ponechané důležité atributy jsou patrné z datové struktury uvedené v příloze A

Ke dni exportu obsahovala tabulka 12351 záznamů.

2.3.7 Tabulka Položky faktury

Tabulka Položky faktury (Pol_FA) je shledána jako nejobtížnější na orientace v jednotlivých atributech. Jedná se o tabulku, ve které jsou obsaženy položky ze všech výstupních dokladů vznikajících v rámci IS, a na základě propojení přes ID konkrétního výstupního dokladu je blíže určeno, o kterou položku faktury, výdejky, dodacího listu či jiného výstupního dokladu jde. Na první pohled je tedy patrné, že tato tabulka musí být

předzpracována přímo selektivním výběrem na SQL serveru, aby došlo k podstatnému zmenšení počtu záznamů a ke zlepšení pochopení jednotlivých atributů.

Nejdůležitějšími údaji, které jednotlivé záznamy obsahují, jsou údaje o tom, ke které faktuře uvedená položka patří, k jaké sortimentní položce se vztahuje. Neopominutelný je samozřejmě atribut Cena.

Velmi důležitým je atribut Typ_řádku. Určuje, jedná-li se pouze o textový řádek, řádek s vedením textu a ceny či o řádek návazný na skladový sortiment.

Ke dni exportu tabulka obsahovala 45792 záznamů.

2.3.8 Tabulka Položka sortimentu

Jedná se o tabulku (Pol_sort) blíže určující položky z tabulky Položka faktury. Je-li některá z položek faktury typu skladový sortiment, jedná se o položky mající návaznost na skladovou položku nebo na položky blíže specifikující.

Ke dni exportu obsahovala 9199 záznamů.

2.3.9 Tabulka Sortiment

Tabulka Sortiment (Sort) je stejně jako tabulka Kategorie velmi jednoduchá. Jsou zde pouze atributy Označení sortimentu a Název sortimentu. Jedná se o rozčlenění sortimentu poskytovaných služeb.

Ke dni exportu obsahovala 7 záznamů.

2.4 Příprava dat

Z jednotlivých analýz datových struktur vyplývá, jak je potřebné se získanými daty dále pracovat a které atributy budou vybrány jako klíčové pro modelování k naplnění stanovených cílů.

Před samotným importem do Clementine je třeba ještě upravit některé datové formáty. Například veškeré datumové typy jsou v IS KP evidovány s přesným časem na jednotky vteřin. Pro dataminingové zkoumání toto není potřebné, postačí datumový typ v podobě dd.mm.rrrr.

Vzhledem k tomu, že se jedná o reálná, skutečná data IS, která prošla zaúčtováním a auditováním v rámci účetní závěrky firmy, není předpoklad chybějících dat u rozhodujících atributů. Přesto byla kontrola kvality dat provedena.

V textových polích byl nalezen výskyt znaku středník, což by dělalo velké problémy při importu do Clementine vzhledem ke skutečnosti, že středník je použit jako oddělovač.

Všechny výskyty znaku středník, jednalo se o umístění v textových, názvových či poznámkových polích, byly nahrazeny mezerou. Toto nijak neovlivnilo kvalitu dat.

Na základě údajů o splatnosti faktury a datu skutečné úhrady byl přidán nový atribut s názvem Neuhrazenost. Představuje údaj, o kolik dní po splatnosti byla faktura uhrazena. V případě záporného čísla se jedná o fakt, že faktura byla uhrazena před termínem splatnosti.

Všechny datové soubory vhodné pro import byly v prostředí Microsoft Office Excel uloženy do datového formátu souboru *.csv, který je vhodným vstupním souborem do modelování v prostředí Clementine.

3 Modelování

Jádrem celého procesu dobývání znalostí z databází je použití analytických metod [2]. Vstupem do modelování jsou předzpracované datové soubory ve formátu csv, z nichž je patrné, které atributy budou voleny jako rozhodující pro jednotlivé modelovací techniky. Rozdíly mezi jednotlivými metodami spočívají zejména dle typu dat.

Jednotlivými atributy, které mohou nejvíce napomoci v odpovědi na otázky stanovené managementem v cílech projektu, jsou:

- Sortiment – představuje jednotlivé druhy sortimentu poskytovaných plnění,
- Kategorie – jedná se o oblasti působení zákazníků,
- Částka – výše jednotlivých plnění v Kč,
- Období – jedná se o měsíce, ve kterých je plnění poskytnuto,
- Neuhrazenost – počet dnů, o které byla faktura za poskytnutá plnění uhrazena zákazníkem později.

Proces modelování je implementován do prostředí Clementine za pomoci [3, 4] tak, aby bylo možné co nejlépe nalézt odpovědi na otázky managementu, na stanovené cíle v kroku jedna metodologie CRISP-DM.

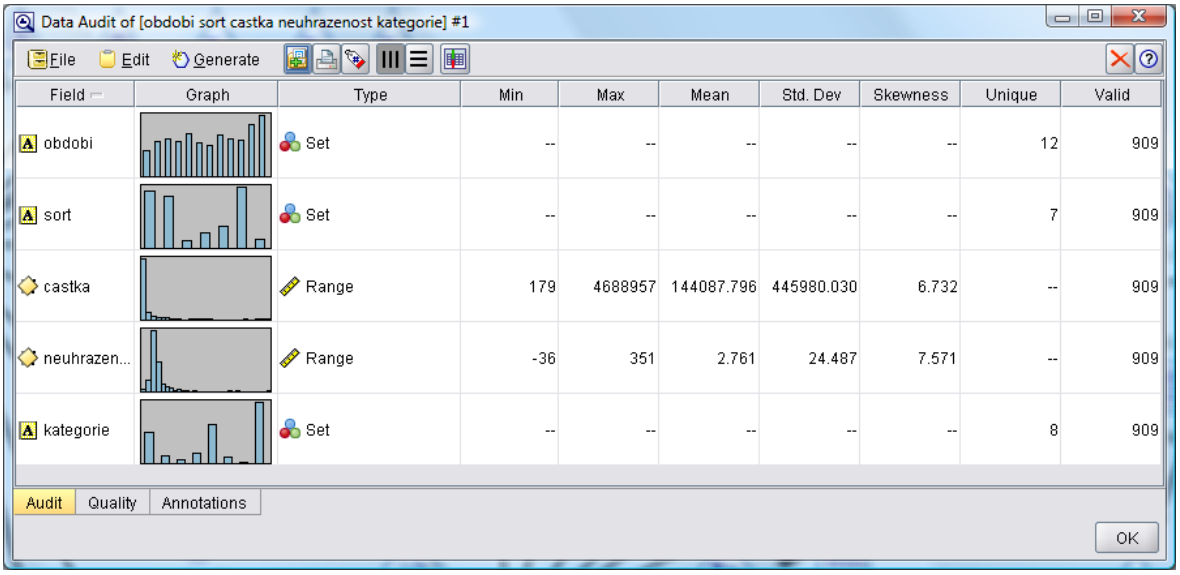
Vybranými metodami jsou:

- Grafické znázornění vazby mezi kategorií, sortimentem a obdobím pavučinovým grafem,
- Grafické znázornění vazby mezi kategorií, sortimentem a obdobím sloupcovým grafem,
- Zohlednění výše fakturace zákazníkům podle sortimentu pomocí agregace,
- Grafické znázornění vztahu mezi neuhrazeností a výší fakturované částky v návaznosti na kategorii zákazníka a období pomocí,
- Grafické znázornění neuhrazenosti,
- Modelování rozhodovacího stromu pomocí algoritmu C5.0,
- Shlukování pomocí algoritmu K-Means a KSOM.

Jednotlivé vstupní datové soubory je nutné nejprve v prostředí Clementine ještě upravit. Jedná se o výběr dat za jednotlivé porovnávané roky z tabulek Platby, Hlav_FA, Pol_FA. Některé atributy, které byly ponechány z původních datových souborů pro lepší přehlednost v datech, ale pro modelování nemají význam, byly potlačeny.

Postupně došlo ke sloučení vstupních datových souborů v jeden, který obsahuje všechny důležité atributy a je vhodným vstupem do modelovacích metod.

V průběhu vytváření tohoto finálního datového souboru je nezbytné průběžně ověřovat, zda zůstává zachována kvalita dat, jak je znázorněno na obrázku 4.



Obrázek 4 - Kvalita vybraných dat roku 2010 (zdroj: vlastní)

3.1 Grafické znázornění vazby pavučinovým grafem

Díky upravenému datovému souboru je nyní možné modelovat a graficky znázornit následující závislosti:

- v jakém období se nejlépe prodává který sortiment,
- která kategorie zákazníků nejvíce nakupuje v kterém období,
- jaký sortiment kupuje která kategorie.

Vstupními parametry jsou voleny vždy právě dvě zkoumané proměnné jednotlivých vztahů, tj. vždy kombinace dvou atributů z následujících tří: Sortiment, Kategorie a Období.

Výstupem je pak výstižný pavučinový graf vyjadřující sílu vzájemné vazby vstupních proměnných, nebo tabulkové vyjádření. Toto znázorňuje obrázek 5 a 7 pro vyjádření vztahu mezi sortimentem a kategorií. Obrázek 6 a 8 je modelován na základě proměnné období a kategorie.

V obou případech se jedná o znázornění silné vazby buď mezi sortimentem a kategorií zákazníka, nebo obdobím a kategorií zákazníka. V nastavení modelu byla tato silná vazba přednastavena na hodnotu počtu vzájemných výskytů vyšší než 35. Jednotlivé tloušťky čar pavučinového grafu odpovídají kvantitativnímu vyjádření.

Jak je patrné z níže uvedených obrázků, v roce 2010 dominovaly nákupy sort1 kategorií zákazníka kateg1 a zákazník z kateg2 nejvíce nakupoval sort4. Obě tyto kombinace překročily více než 100 výskytů v daném roce.

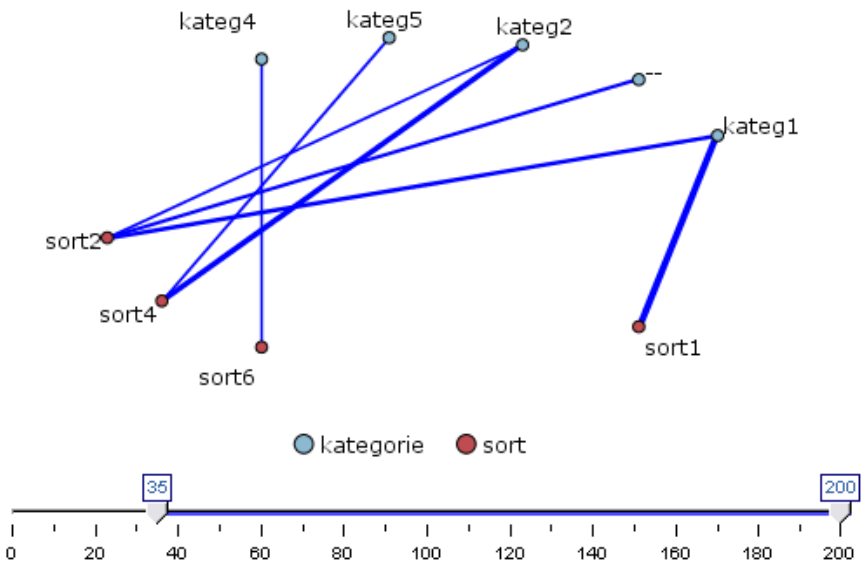
Co se týče nejvyššího počtu poskytnutých plnění v jednom měsíci pro jednu kategorii v roce 2009, pak nejvíce nakupovala kateg1 v měsíci listopadu a květnu, v prosinci pak převládaly nákupy zákazníkem z kateg2.

Links	Field 1	Field 2
191	sort = "sort1"	kategorie = "kateg1"
145	sort = "sort4"	kategorie = "kateg2"
93	sort = "sort2"	kategorie = "kateg1"
70	sort = "sort2"	kategorie = "--"
40	sort = "sort6"	kategorie = "kateg4"
38	sort = "sort4"	kategorie = "kateg5"
36	sort = "sort2"	kategorie = "kateg2"

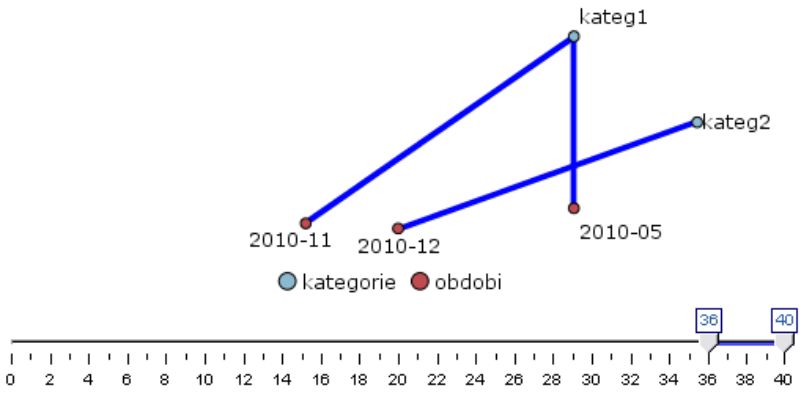
Obrázek 5 - Tabulkové znázornění výstupu – rok 2010 (zdroj: vlastní)

Links	Field 1	Field 2
38	obdobi = "2010-11"	kategorie = "kateg1"
37	obdobi = "2010-05"	kategorie = "kateg1"
36	obdobi = "2010-12"	kategorie = "kateg2"

Obrázek 6 - Tabulkové znázornění výstupu - rok 2009 (zdroj: vlastní)



Obrázek 7 - Pavučinový graf – rok 2010 (zdroj: vlastní)

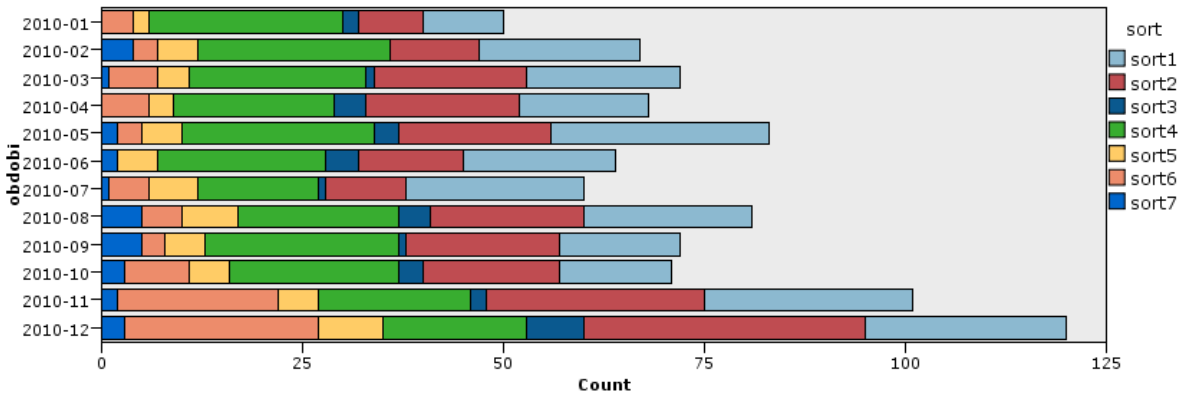


Obrázek 8 - Pavučinový graf - rok 2010 (zdroj: vlastní)

3.2 Grafické znázornění vazby sloupcovým grafem

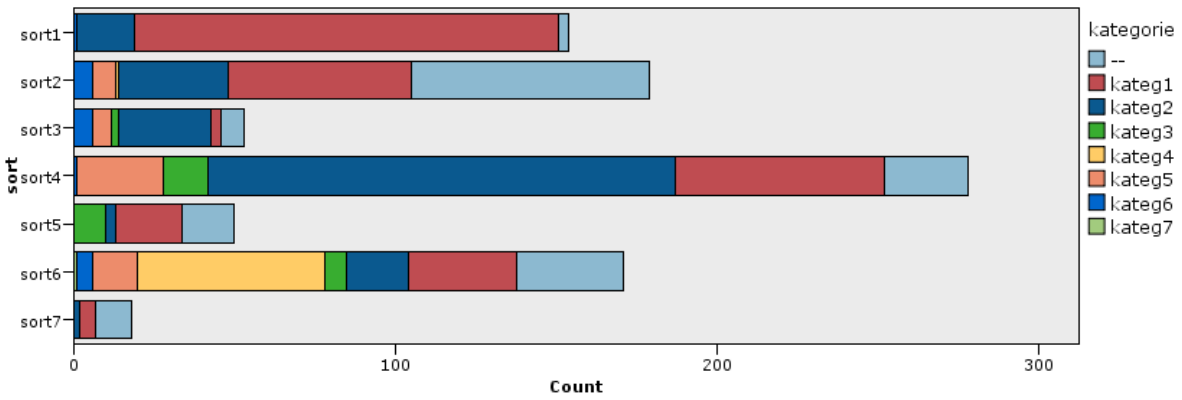
Jiná podoba grafického znázornění stejných závislostí jako v předchozím typu modelování je znázorněna pomocí sloupcového grafu.

Vstupními parametry jsou opět postupně voleny kombinace dvojic atributů Sortiment, Kategorie a Období. Výstupem z modelování je sloupcový graf zachycený na obrázku 9. Z něj je jasně patrné, ve kterých obdobích je největší prodej daného sortimentu. Například v listopadu a prosinci 2010 výrazně vzrost prodej sort6. Naproti tomu sort7 se vůbec neprodával v lednu a dubnu. Obdobným výstupem byl i graf vyjadřující, kteří zákazníci v určitém období nejvíce nakupují. Tyto souvislosti by se však při výborné znalosti IS KP a za podpory např. Microsoft Office Excel daly vysledovat.



Obrázek 9 - Vztah mezi obdobím a sortimentem - rok 2010 (zdroj: vlastní)

Získat informace, jaký sortiment kupuje který za zákazníků, je z PIS naprosto nemožné. Pro Clementine z připraveného datového souboru je to již velmi jednoduché, jak ukazuje obrázek 10.



Obrázek 10 - Vztah mezi sortimentem a kategorií - rok 2009 (zdroj: vlastní)

V roce 2009 zde dominuje sortiment 1, který je kupován především pouze zákazníkem z kateg1. Nejširší spektrum zákazníků má sort6. Ten nakupuje i zákazník z kateg4, který jiné sortimenty nenakupuje. Kategorie 2 se nejvíce zaměřuje na sort4, který je dle počtu výskytů nejprodávanějším v daném roce.

3.3 Zohlednění výše fakturace zákazníkům dle sortimentu

Všechny výše provedené analýzy se zabývají pouze porovnáním počtu plnění bez zohlednění výše plnění, tj. atributu Částka. Pro srovnání finančních objemů jednotlivých modelů se použije agregace.

V nastavení modelování jsou vstupními parametry voleny atributy Sortiment a Kategorie, agregovaným polem je potom atribut Částka, u kterého se zvolí součtování dle proměnných vstupních parametrů. Je možné zvolit i další parametry, v konkrétním případě bylo vybráno zobrazení minimálních a maximálních hodnot pro příslušné kombinace vstupních proměnných.

Výstup z modelování je ve formě tabulky zobrazené na obrázku 11, názorně a přehledně vyjadřující, která kategorie si v jaké výši zakoupila který sortiment.

Sortiment1 nejvíce nakupoval zákazník z kateg1. Nejvyšší je jak celková částka za poskytnutá plnění, tak i počet uskutečněných obchodních případů. Sort7 patří mezi nejméně obchodovatelný sortiment. Je oblíben pouze u kategorie nezařazených zákazníků a velmi drobné obchody má s kategorií kateg2. Zajímavým zjištěním v roce 2010 je i fakt, že žádný ze sortimentů není nakupován všemi kategoriemi zákazníků.

	sort	castka_Sum	castka_Min	castka_Max	kategorie	pocet_vyskytu
1	sort1	533441	179	327701	--	10
2	sort1	39276570	748	4423782	kateg1	191
3	sort1	6722010	582	2342657	kateg2	27
4	sort1	387401	7978	245492	kateg3	5
5	sort1	14474	14474	14474	kateg5	1
6	sort2	479926	300	39996	--	70
7	sort2	2867231	348	551111	kateg1	93
8	sort2	2311587	1433	438624	kateg2	36
9	sort2	4356	4356	4356	kateg4	1
10	sort2	234998	1188	156650	kateg5	9
11	sort2	1473816	3336	800617	kateg6	6
12	sort2	27732	27732	27732	kateg7	1
13	sort3	30840	1356	14424	--	9
14	sort3	124422	24792	36240	kateg1	4
15	sort3	6463274	6312	3781836	kateg2	13
16	sort3	16188	3216	7728	kateg3	3
17	sort3	432307	432307	432307	kateg4	1
18	sort3	23460	10560	12900	kateg5	2
19	sort4	124420	179	31251	--	19
20	sort4	583857	1363	156000	kateg1	34
21	sort4	15320497	179	668850	kateg2	145
22	sort4	102626	179	57360	kateg3	9
23	sort4	740662	1260	149920	kateg5	38
24	sort4	24444	3480	6348	kateg6	6
25	sort4	120000	120000	120000	kateg7	1
26	sort5	4901789	2616	1079996	--	28
27	sort5	10777138	4368	4658246	kateg1	20
28	sort5	2501944	5400	2376304	kateg3	12
29	sort6	6868341	11436	1762198	--	15
30	sort6	5066879	10091	1178042	kateg1	17
31	sort6	110621	7662	31822	kateg2	6
32	sort6	9520024	115330	4688957	kateg3	6
33	sort6	11581578	5244	1272000	kateg4	40
34	sort6	45478	45478	45478	kateg5	1
35	sort6	20299	5516	14783	kateg6	2
36	sort7	1097039	1337	445681	--	25
37	sort7	44138	2064	35810	kateg2	3

Obrázek 11 - Agregate - fakturace podle sortimentu a kategorií - rok 2010 (zdroj: vlastní)

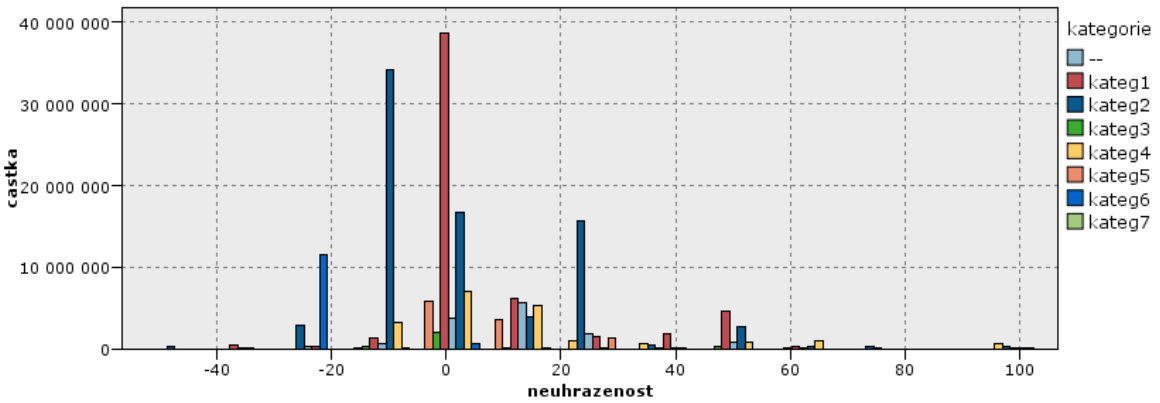
3.4 Grafické znázornění vztahu mezi neuhrazeností a výší fakturované částky v návaznosti na kategorii zákazníka a období

Další typ grafického výstupu ukazuje, jaká je platební morálka jednotlivých zákazníků v návaznosti na výši fakturovaných částek. Znalost informace o období, kdy dochází k největšímu zpoždění plateb, patří rovněž k těm, které firmy potřebují využívat k plánování svých finančních toků.

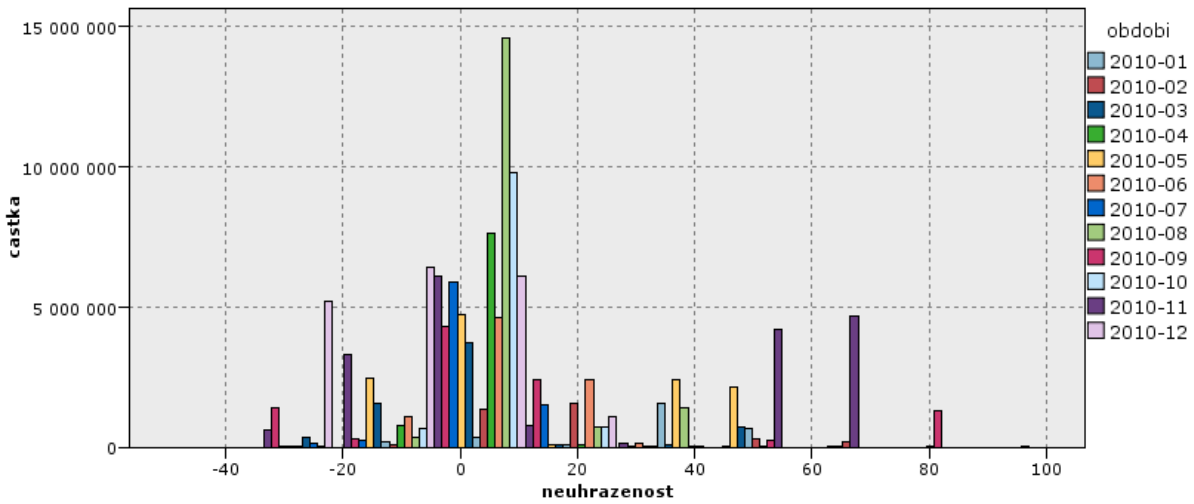
V nastavení modelování jsou jako vstupní proměnné zvoleny kvantifikovatelné atributy Částka a Neuhrazenost. Pro názorný grafický výstup je zvoleno barevné rozlišení

podle kategorie, jak je znázorněno na obrázku 12, a výstup v návaznosti na období je na obrázku 13. Grafy jsou velmi jednoduché, přehledné, na první pohled vypovídající.

Pro zvýšení názornosti byl pro neuhrazenost zvolen interval hodnot, ze kterého byly vyloučeny odlehlé hodnoty. V roce 2009 kateg1 hradila své faktury téměř ve většině případů přesně ve splatnosti, naproti tomu kateg2 hradí některá svá poskytnutá plnění před splatností, a některá i více než 20 dní po splatnosti. V roce 2010 byly s úhradou nejvíce opožděny faktury vystavené v měsíci listopadu, kdy neuhrazenost se pohybovala až kolem 60 dní.



Obrázek 12 - Neuhrazenost dle kategorie a částky - rok 2009 (zdroj: vlastní)



Obrázek 13 - Neuhrazenost dle období a částky - rok 2010 (zdroj: vlastní)

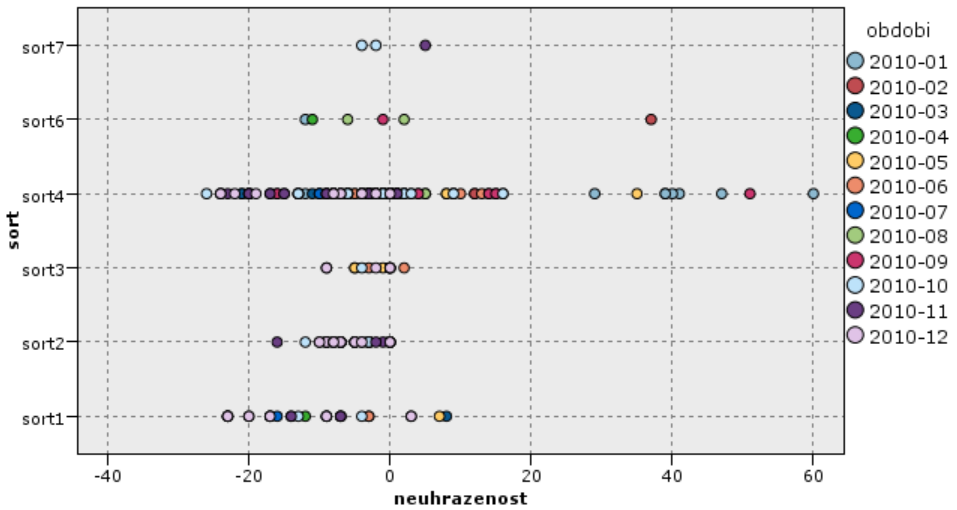
3.5 Grafické znázornění neuhrazenosti pro jednotlivé kategorie

Tento grafický výstup umožňuje ukázat, jaká je neuhrazenost jednotlivých kategorií zákazníků za jednotlivé typy sortimentu v členění po obdobích. Aby bylo možné zobrazit tyto závislosti, je potřebné rozšířit předzpracování vstupního datového souboru o výběr jednotlivých kategorií.

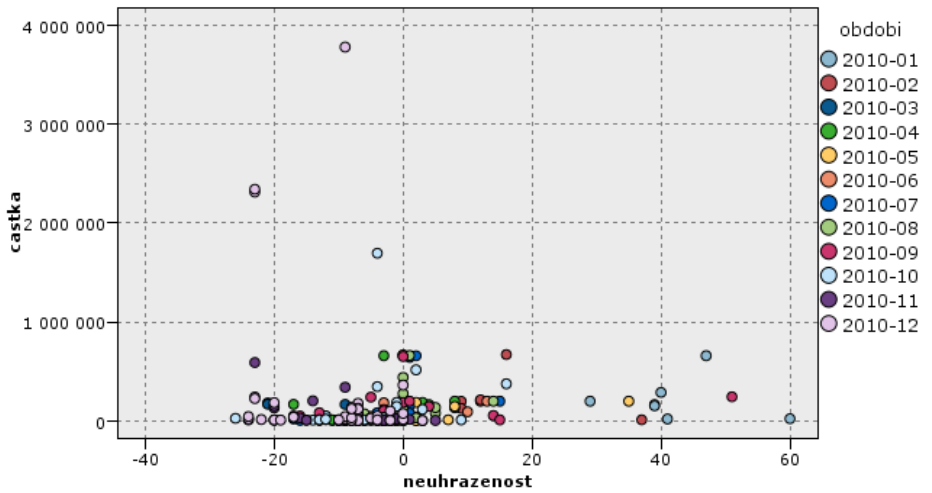
Vstupní nastavení je voleno tak, aby barevné rozlišení znázorňovalo jednotlivá období. Vstupními parametry jsou pro srovnání v prvém případě voleny Neuhrazenost a Sortiment, pro druhý výstup potom Neuhrazenost a Částka.

Grafický výstup je krásně přehledný, čitelný, jak znázorňuje obrázek 14. Opět je zde zachycena skutečnost, kterou IS firmy není schopen v žádné podobě poskytnout.

Obrázek 14 znázorňuje, jaká je neuhrazenost dle sortimentu a obrázek 15 pak neuhrazenost dle výše fakturované částky v rozdělení po obdobích. V obou případech se jedná o analýzu zákazníka z oblasti kateg2. Na první pohled je patrné, že tento zákazník vůbec nekupuje sort5. Nejvyšší neuhrazenost se pak pohybuje u sort4 a to řádově v částkách do 300 tisíc Kč za jednotlivé obchodní případy. U sort2 a u sort3 jsou faktury hrazeny většinou před splatností. Jedná se o údaje roku 2010.



Obrázek 14 - Platební morálka zákazníka z kateg2 dle sortimentu - rok 2010 (zdroj: vlastní)



Obrázek 15 - Platební morálka zákazníka z kateg2 dle částky - rok 2010 (zdroj: vlastní)

3.6 Modelování rozhodovacího stromu pomocí algoritmu C5.0

Jak uvádí [9], modelování pomocí rozhodovacího stromu rozděluje datový soubor podle vstupních proměnných do skupin, na základě kterých je tvořena predikce nebo klasifikace. Snahou je naleznout v datovém souboru pravidla, podle kterých by byl systematicky rozdělen.

Prvky stromové struktury, která se graficky zobrazuje jako schéma, jsou větve a uzly. Uzly jsou uspořádány do různých úrovní, z nichž uzel na nejvyšší úrovni je nazýván kořen. Uzly mohou být dvojího typu, nelistové, odkazují na nižší úrovně, a listové, které se dále již nedělí. Jednotlivé větve vedou od kořene k listům. [19]

„Obecné schéma algoritmu TDIDT pro tvorbu rozhodovacích stromů:

1. Zvol jeden atribut jako kořen dílčího stromu.
2. Rozděl data v tomto uzlu na podmnožiny podle hodnot zvoleného atributu a přidej uzel pro každou podmnožinu.
3. Existuje-li uzel, pro který nepatří všechna data do téže třídy, pro tento uzel opakuj postup od bodu 1, jinak ukonči.“ [2, s.86]

Mezi nejznámější vyvinuté a běžně užívané algoritmy pro tvorbu rozhodovacích stromů patří např. C5.0, CHAID, QUEST, C&RT a základní charakteristika jejich algoritmu dle [9, 19] je:

- C5.0 – pracuje podle obecného algoritmu TDIDT, standard pro tvorbu rozhodovacích stromů,
- CHAID – statistický algoritmus založený na optimální hodnotě chí-kvadrát testu závislosti nebo F-testu,
- QUEST – statistický algoritmus vybírající proměnné rychle a nevychýleně, efektivně vytváří přesné binární stromy,
- C&RT – klasifikační a regresní algoritmus, jedná se o binární strom dělící se maximálně na 2 větve.

Pro samotné modelování je vybrán algoritmus C5.0, který je založen na větvení na základě největšího informačního zisku. Pracuje jak s kategorizovanými, tak i s číselnými hodnotami na vstupu, na výstupu jsou data kategorizovaná. [2, 9]

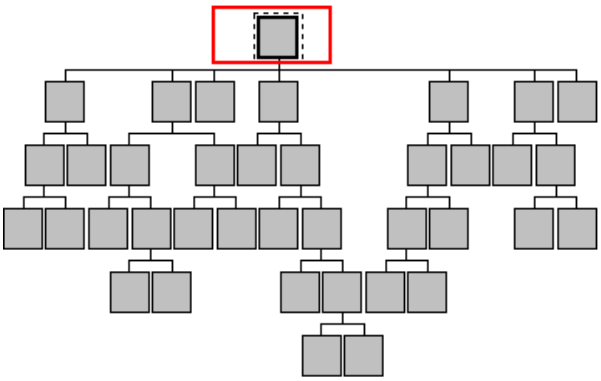
Rozhodovací strom lze převést pro lepší orientaci na rozhodovací pravidla. To se provádí tak, že z každého průchodu stromem od kořene k listu se vytvoří jedno pravidlo. Nelistové uzly jsou předpoklady, listový uzel pak závěr pravidla. [2, 9]

Tato modelovací technika provedla rozdělení datového soubor na základě vstupních parametrů, které byly zvoleny takto: za kořen stromu byla zvolena Kategorie a vstupními

proměnnými Neuhrazenost a Sortiment. Není vhodné zadávat více vstupních parametrů, rozhodovací strom je pak velmi rozsáhlý a pro orientaci značně nepřehledný.

Na obrázku 16 je rozhodovací strom pro výše zvolené atributy. Jde zde pouze o grafickou podobu náhledu složitosti vzhledu celého rozhodovacího stromu.

Podrobný rozhodovací strom pro výše uvedené proměnné z dat roku 2010 je uveden v příloze B.



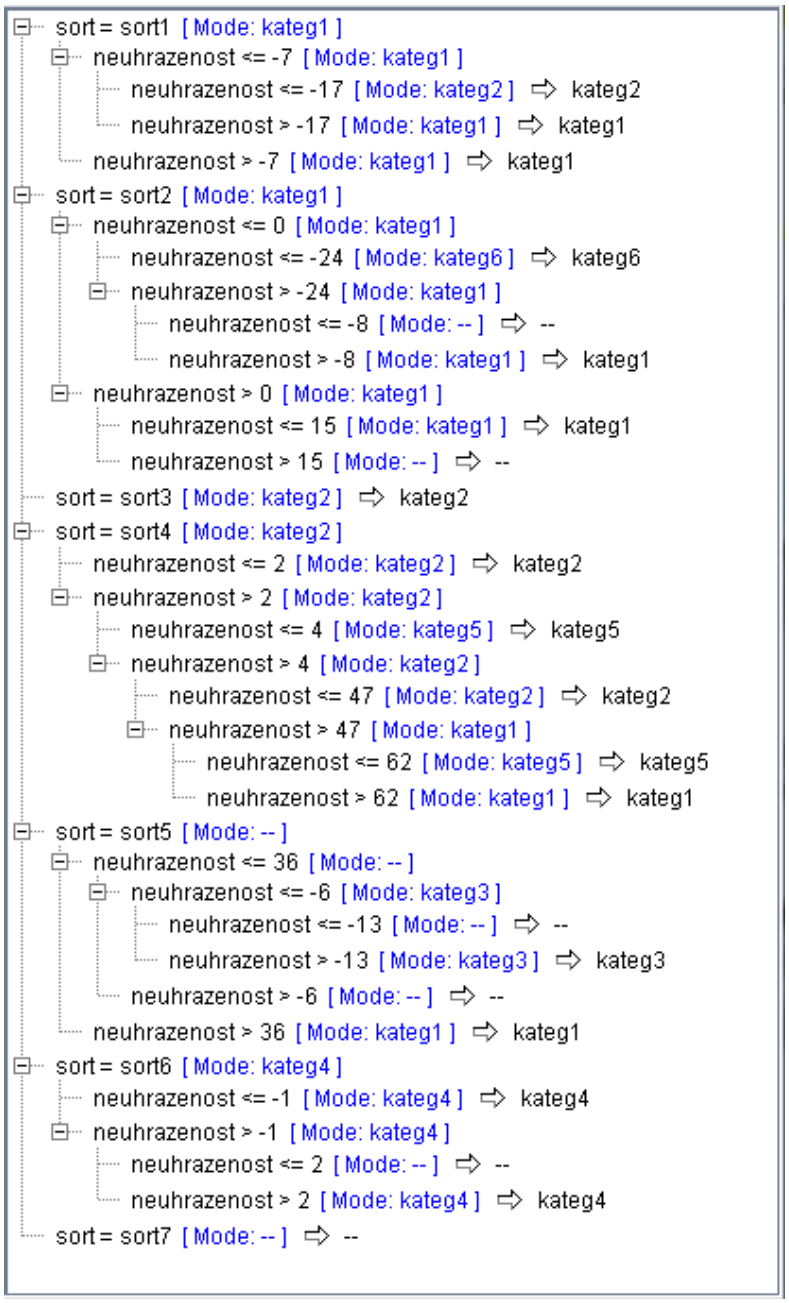
Obrázek 16 - Rozhodovací strom (zdroj: vlastní)

V případě přidání kvantifikované vstupní proměnné Částka do modelování se podstatně změní vzhled rozhodovacího stromu. V tomto konkrétním případě došlo k rozdělení záznamů do 132 uzlů a rozhodovací strom se tím stal nepřehledným a složitým.

Popisný způsob výstupu z modelování pomocí rozhodovacích pravidel je uveden na obrázku 17. V případě jednodušší podoby rozhodovacího stromu je i toto popisné vyjádření lehce čitelné, přehledné, snadno orientovatelné. V závislosti na zadaných vstupních proměnných ukazuje postupný rozpad stromu až na jednotlivé listy.

Příkladem z dat roku 2010 je sortiment sort1, jak ukazuje obrázek 17, tak i příloha B. Sort1 ve sledovaném roce nejvíce nakupoval zákazník z kateg1, ale jeho platební morálku již bylo možné dále větvit. Pokud byla neuhrazenost ≤ -7, pak se jednalo nejspíše o zákazníka z kateg1. Byla-li neuhrazenost > -7, pak byl rovněž z kateg1. Ale další větvení ukazuje, že dosahovala-li neuhrazenost výše až ≤ -17, pak mohl být zákazník i z kateg2. Obdobně by šel popsat celý rozhodovací strom.

Stejně tak je možné za kořen stromu zvolit Sortiment a vstupními proměnnými Neuhrazenost a Kategorii. Tento rozhodovací strom pak ukazuje, že je-li zákazník např. z kateg6, bude nejspíše odebírat sortiment sort2 s neuhrazeností ≤ -19, pokud bude neuhrazenost > -19, bude jeho poptávanou činností sort4.



Obrázek 17 - Rozhodovací pravidla (zdroj: vlastní)

3.7 Shlukování

Jedná se o modelovací analýzy pomocí shluků, což jsou metody, která ze vstupního datového souboru na základě vzájemné podobnosti vytvoří seskupováním jedinců skupiny, tak zvané shluky. Tyto skupiny záznamů mají co možná nejvíce obdobné vlastnosti a naproti tomu od záznamů patřících do jiného shluku se co možná nejvíce liší. Obě metody modelování, jak K-Means, tak i KSOM (Kohonen Self-Organizing Map), jsou založeny na principu vytváření shluků. [16, 18]

3.7.1 Kohonenovy samoorganizující se mapy

Kohonenovy samoorganizující se mapy jsou samoorganizující se a samoučící se neuronové sítě vhodné pro analýzu dat. Jejich základ vznikl počátkem sedmdesátých let minulého století a na ně navázal v osmdesátých letech Teuvo Kohonen. Základní myšlenku a princip popsal ve své knize [11].

Ve volné a zkrácené podobě si lze Kohonenův algoritmus představit dle [20] takto:

1. Inicializace sítě
2. Předložení vstupního vektoru
3. Výpočet vzdáleností
4. Výběr minimální vzdálenosti
5. Úprava vah
6. Přestup k bodu 2.

Aby bylo dosaženo co možná nejlepšího uspořádání do shluků, volí se na začátku velké rozsah okolí a velký vliv učeného vzoru na změny ve vahách neuronů. Síť v průběhu učení vytvoří reprezentanty shluků a téže vah. K učení sítě postačí velké skupiny vstupních vektorů, není nutné mít k dispozici ideální vzory. Jedná se tedy o proces, kdy si KSOM samy určí, jak bude výstup klasifikován. Mají tedy schopnost vytvářet shluky objektů s podobnými vlastnostmi a roztřídit je do skupin. [12, 16, 17]

V prostředí Clementine se u modelování KSOM nenastavuje počet shluků na výstupu, tato metoda, jak již bylo zmíněno výše, si počet shluků vytvoří sama. Vstupními parametry jsou voleny Částka a Neuhrazenost, vždy se zde musí jednat o kvantifikovatelné proměnné.

Obrázek 19 pak ukazuje vytvořené shluky touto metodou. Došlo k rozdělení dat do 12 shluků v souřadnicích x a y. Výstup z modelu dále ukazuje, kolik výskytů je ve kterém shluku a jaká je průměrná hodnota atributu Částka a Neuhrazenost v jednotlivých shlucích.

3.7.2 K-Means

Metoda K-Means, nebo-li metoda K-průměrů, někdy je v literatuře nazývána i metodou K-centroidů, je oproti KSOM optimalizační metodou, kde je počet shluků jedním ze vstupních parametrů, je zadán ručně v počátečním nastavení. Vytváří se tedy předem stanovený počet shluků a to tak, že se určí počáteční centroidy jako středy shluků a poté se postupně zkoumají vzdálenosti každého objektu od každého centroidu pomocí výpočtu euklidovské vzdálenosti. Objekt je přiřazen k nejbližšímu centroidu. Pro každý shluk se počítá nový centroid a následně se opět zkoumá vzdálenost objektu od každého centroidu. [16, 18]

Jednotlivé kroky algoritmu jsou dle [16, 18] popsány takto:

1. Vyber počet shluků k, vygeneruj nebo zadej těžiště shluků.
2. Pro každý objekt vyber shluk, jehož těžišti je nejblíže, a pokud se vybraný shluk nerovná původnímu shluku, přemísti do něj objekt a přepočítej těžiště.
3. Pokud nedošlo ke změně žádného shluku, skonči, jinak pokračuj krokem 2.

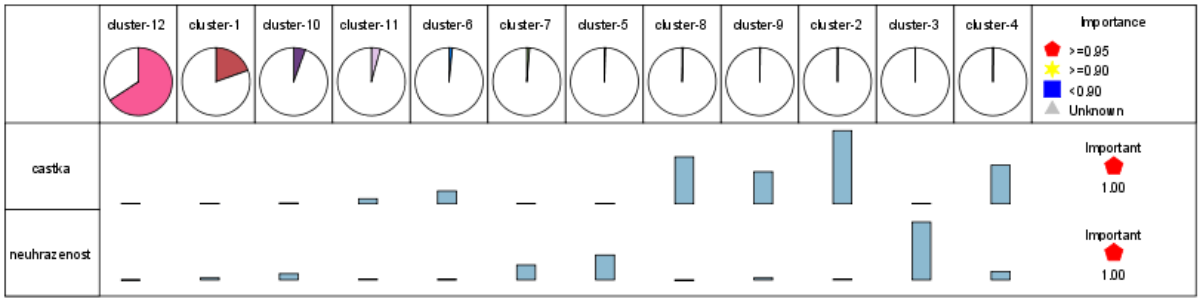
Více informací a přesný postup je uveden například v [16, 18].

Aby bylo dobře patrné, jak která metoda podle algoritmu svého výpočtu vytváří shluky, bylo pro metodu K-Means v úvodním nastavení modelování požadováno vytvoření stejného počtu shluků, jako jich vytvořil u modelování KSOM. Vstupní atributy byly zachovány stejné jako u KSOM.

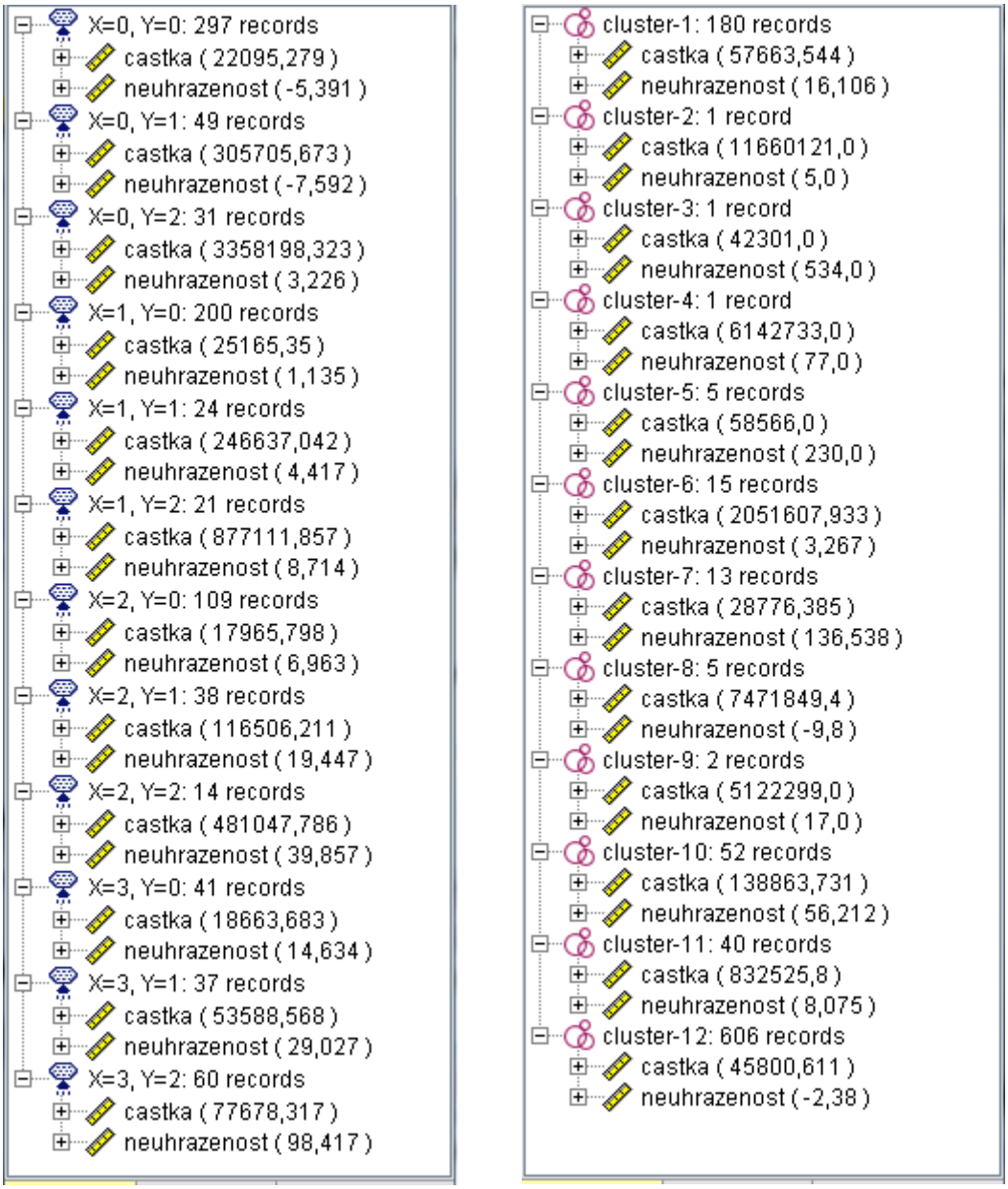
Výstupem je buďto vytvořený model v popisném tvaru, jak ukazuje obrázek 19, kde soupis jednotlivých shluků přesně určuje i počet výskytů ve shluku, nebo je výstup v grafické podobě, jak znázorňuje obrázek 18.

V případě dat roku 2008 je vidět velké zastoupení ve shluku č. 12 a naproti tomu polovina shluků obsahuje pouze do 5 výskytů. Jedná se o extrémní neuhrazenost či vysoké fakturované částky, jak ukazují při modelování z dat roku 2008 shluky č. 2, 3, 4, 5, 8 a 9, které v případě zjištěných vzdáleností od ostatních centroidů nebylo možné při požadovaném počtu shluků 12 zařadit do některého jiného shluku.

Největší shluk, shluk č. 12, je vytvořen ze záznamů, kde průměrná fakturovaná částka je ve výši 45800 Kč při průměrné neuhrazenosti -2 dny. Druhým shlukem s nejvíce záznamů je shluk 1, kde průměrná částka je ve výši 57663 Kč a průměrná neuhrazenost potom 16 dní.



Obrázek 18 - Metoda K-Means - rok 2008 (zdroj: vlastní)



Obrázek 19 - Vytvořené shluky - KSOM a K-Means - rok 2008 (zdroj: vlastní)

4 Vyhodnocení výsledků, jejich využití a doporučení pro rozhodování managementu

Zhodnocení výsledků z modelování je směřováno ke srovnání jednotlivých let 2008 až 2010 s cílem najít, pokud nastávají, nové obchodní tendence zákazníků konkrétní firmy se zaměřením na kategorie zákazníků a obchodovaný sortiment. Jedním z cílů bylo analyzovat závislosti právě mezi těmito dvěma rozhodujícími skupinami. Důležitými pomocnými prvky pro vyhodnocení jsou vztahy mezi výší fakturované částky, obdobím, kdy fakturace proběhla a následně, kdy došlo k uhrazení fakturované částky zákazníkem.

Datová analýza pomocí pavučinového grafu ukázala, že se rozložení nejčastěji uskutečněných prodejů počtem výskytů u obchodovaných sortimentů a skupin zákazníků nemění. Jednoznačně převládá zastoupení zákazníků z oblasti kateg1, kteří nejčastěji nakupují sort1 pro komunikační infrastrukturu. Naproti tomu kateg2 nejčastěji nakupuje sort4. Tento trend se v posledních třech letech nezměnil.

Jednotlivé obchody z pohledu období, ve kterém jsou uskutečněny, lze prohlásit za stálé. Na počátku roku převládají většinou smluvně dlouhodobé kontrakty. Naopak poslední dva měsíce v roce narůstá objem i počet uskutečněných obchodů. Podíl na tomto mají i drobní odběratelé nakupující vánoční dárky v podobě sortimentu HW či SW.

Co se však podstatně změnilo, je výše finančního objemu těchto kontraktů u jednotlivých zákazníků.

U zákazníků z kateg1 se dodávky sortimentu1 podstatně neliší, ale velmi vzrůstající tendenci mají naopak dodávky sort5. Tato oblast trhu může být i nadále velmi atraktivní, neboť zákazníci z kategorie1 byli před dvěma lety v situaci, kdy mnoho neinvestovali, a nyní se již výrazně začínají projevovat snahy o nové investice do technologií pro obnovu vlastních technických zařízení a zákazníci z této oblasti se stávají velmi vyhledávanými.

S tímto nepochybně souvisí i skupina zákazníků zařazených v kateg3. V roce 2008 se objem uskutečněných obchodů pohyboval řádově ve stovkách tisíc, rok 2009 přinesl první větší kontrakty a následně rok 2010 ukazuje, že po době útlumu tato oblast začíná být velmi zajímavá z pohledu zákazníků. Tento obor nejvíce poptával činností jako budování komunikačních infrastruktur a dodávky sort5.

Kategorie2 je naopak příkladem, jak ještě v roce 2008 investovala i do výstavby sítí LAN a WAN, s tím souvisejících nákupů HW a SW. V roce 2009 došlo k prvním státním finančním restrikcím a obchodní vztahy s kateg2 se zúžily téměř jen na poskytování služeb nezbytných pro fungování této skupiny zákazníků. Stejný trend pokračoval i v roce 2010

částečným omezením i v této činnosti. Objem obchodů s kateg2 ještě i v roce 2009 představoval skoro 40% objemu tržeb firmy. Naproti tomu rok 2010 přinesl zásadní změnu. Propad fakturovaných objemů na pouhých 20% z celkových tržeb. Tento stav jasně ukazuje na skutečnost, že firma se nemůže klíčově orientovat pouze na jednu kategorii zákazníků, ale musí se pohybovat v širokém spektru. Jednostranná orientace by mohla vést až k samotným existenčním problémům.

Obdobný propad je patrný i u zákazníků z kategorie kateg5. V roce 2010 jsou i zde zaznamenány pouze minimální kontrakty, které se omezily pouze na nezbytné služby a drobnou obnovu hardware. Zejména v této skupině zákazníků jsou určitě mnohé příležitosti pro nové obchodní kontrakty, zvláště s ohledem na skutečnost, že kategorie těchto zákazníků patří k těm s lepší platební morálkou.

Stabilní oblastí se jeví kategorie zákazníků kateg4. V uplynulých třech letech zde nenastaly podstatné rozdíly v počtu či objemů obchodů, ani v sortimentním zaměření. Trvale převládají dodávky a montáž sort6. Tento stav zřejmě souvisí s trvalými investicemi zákazníka do obnov jednotlivých poboček těchto institucí. Platební morálku lze prohlásit za uspokojivou, pouze výjimečně dochází ke zpoždění úhrad faktur za poskytnuté činnosti.

Srovnání platební morálky jednotlivých zákazníků v letech 2008 až 2010 ukazuje, jak např. zákazníci zařazení v kateg2 si i přes změnu ve výši obchodních transakcí udržují stejný trend. Za dodávky HW, SW, či sort1 platí v předstihu, naproti tomu za sort4 i po splatnosti. Vzhledem ke stavu, že tato oblast se v současné době zaměřuje téměř výhradně na nákup poskytovaných služeb, stává se tento zákazník lehce rizikový. Na druhou stranu je však patrná skutečnost, že výše fakturovaných částek za jednotlivá plnění není až tak vysoká a skluz neuhrazenosti se pohybuje v těchto případech kolem 10 až 15 dní, pouze výjimečně až kolem 40 dní.

Dříve příjemné zjištění o rozvoji obchodní činnosti se zákazníky z kateg3 kazí při analyzování neuhrazenosti fakt, že platební morálka této kategorie je velmi špatná, pravidelně dochází ke zpožďování plateb a to řádově v desítkách dnů. To již je podstatný zásah do finanční situace firmy. Nejkritičtější je to u sortimentu sort6, což je oblast úzce související se stavebními pracemi. Dodávky dalších technologií, ať již sort5, či sort1, zdaleka tak vysokou neuhrazenost nemají.

Zákazníci z kateg1 v současné době až na drobné výjimky hradí své faktury v termínech splatnosti či těsně před nebo po splatnosti, ale pouze v jednotkách dní. Ještě v roce 2008 bylo možné kategorii kateg1 prohlásit za kategorii ohrožující finanční toky firmy, zejména u sortimentu sort5. Po útlumu obchodů v roce 2009 se však v roce následujícím

podstatně zpět navýšily obchodní kontrakty a s tím se zlepšila i platební morálka u však poskytovaných sortimentů.

Vybrat období, kdy dochází k největší neuhrazenosti faktur a které by se pravidelně opakovalo v letech po sobě, není možné. Výstupní modely neukázaly žádný typický případ měsíce, kdy by tato situace nastávala. Jednotlivé případy spíše souvisí s konkrétním zákazníkem. Přesto je toto zjištění cenné. Je možné konstatovat, že nenastává období roku, kdy by bylo potřebné předem alokovat finanční prostředky pro běžný chod firmy z důvodu předpokládané špatné platební morálky zákazníků.

Cílem modelování pomocí rozhodovacích stromů je získat představu o typickém zařazení zákazníka do skupiny, jak by se mohla vyvíjet neuhrazenost u budoucích zákazníků na základě doposud uskutečněných kontraktů, na základě dosavadního chování stávajících zákazníků.

Rozhodovací strom, ať již v popisném, či grafickém vyjádření, je lehce čitelný. Příkladem může být např. sort3, který se dále nevětví. Tuto skupinu sortimentu nakupují převážně zákazníci z kategorie kateg2 a jejich výše neuhrazenosti se nijak dramaticky neliší. Pro management firmy dobrá informace, komu nejvíce tento sortiment nabízet. Další příklad rozhodovacího pravidla již byl popsán v kapitole 4.6.

V roce 2009 a 2010 mají rozhodovací stromy obdobný charakter. V roce 2009 se větví do 46 uzlů, jeden sortiment, sort1, se nevětví vůbec. Rozhodovací strom má 5 úrovní. V roce 2010 jsou úrovně dokonce jen 4, uzlů je 38 a nevětví se dokonce dva sortimenty, sort3 a sort7. V roce 2008 měl rozhodovací strom 49 uzlů, což je ještě srovnatelné, nevětvily se dva sortimenty, sort5 a sotr7, které byly dokonce spojeny v jeden uzel. Velmi rozdílné je však větvení u sort6, které je rozpadnuto až do 11 úrovní, do 30 uzlů. Tento stav je dán velmi rozdílnou výši neuhrazenosti u tohoto sortimentu v roce 2008.

Modelování pomocí shluků dobře ukázalo skutečnost, jak široký rozsah činností firma má, že nedochází ke specializaci na drobné či naopak jen velké zákazníky, ani k úzkému sortimentnímu zaměření. Výše fakturovaných částek je v řádech stokorun až milionů korun. Vytvoření menšího počtu shluků při vstupních atributech Částka a Neuhrazenost je pak víceméně nemožné. Kohonenovy samoorganizující se mapy ve všech třech sledovaných letech vytvořily 12 shluků. Žádný ze shluků v porovnání s ostatními neobsahoval extrémní počet záznamů. Naopak vždy se vyskytl shluk s počtem záznamů do deseti, v roce 2008 dokonce až se 60 záznamy, který obsahoval záznamy s extrémně vysokou neuhrazeností.

V případě metody K-Means bylo při spouštění analýzy nastaveno požadované vytvoření 12 shluků pro možnost porovnání výstupů z obou metod. Výsledný model

se sestavoval z jednoho velkého shluku obsahujícího více než polovinu záznamů. Naopak dalších 5 shluků vždy obsahovalo pouze 1 až 5 záznamů. Jednalo se o extrémně vysoké hodnoty atributu Neuhrazenost nebo Částka. Tato metoda velmi extrémní hodnoty vhledem ke způsobu výpočtu nepřiřazuje k jiným hodnotám a ponechává je jako osamocené. Tento stav by samozřejmě nenastal, pokud by bylo zvoleno např. jen 5 shluků v úvodním nastavení. Pak i nadále převládá jeden velký shluk čítající kolem 85 – 90% záznamů, ale již se nevyskytuje žádný shluk s pouze jedním záznamem.

Z obou výše popsaných metod modelování je tedy patrný jejich zásadní rozdíl. Metoda K-Means je velmi citlivá na odlehlé hodnoty, ale na druhou stranu si vlastním odhadem volíme počet shluků. Kdežto KSOM si na základě svého postupu samy určí, kolik shluků bude vytvořeno.

Vyhodnocením a rozborem modelovaných situací je skončena pátá etapa metodologie CRISP-DM.

Podstata šesté etapy metodologie CRISP-DM spočívá ve využití výsledků. V případě tohoto projektu je samotnou podstatou a řešením podle metodologie popis výstupů, jak je uvedeno v celé této kapitole. Samotné vyhodnocení jednotlivých modelovaných situací, tj. srovnání let 2008 – 2010 z pohledu měnící se ekonomické situace na trhu, znamená pro management firmy lepší orientaci na tomto trhu. Na kterou oblast zákazníků se zaměřit, kde v případě sjednávání nových obchodních kontraktů dbát zvýšené pozornosti ve vztahu k fakturaci a neuhrazenosti, který sortiment je na trhu žádanější a nebylo by od věci jeho rozšíření či spojení s případnými nově nabízenými službami.

Určitý druh implementace nějakých softwarových nástrojů v tomto případě nebude proveden ani doporučen.

Změny v návaznosti na tuto rozsáhlou analýzu by mohli směřovat buď do organizačních opatření, nebo do obchodních strategií firmy.

Z organizačního hlediska je možná managementu doporučit rozšíření kapacity obchodního oddělení zejména se zaměřením na zákazníky z kategorie kateg1 a s tím spojený i tým realizující zakázky ze sortimentu sort5.

Jako vedlejší činnost se z pohledu obchodů jeví prodej HW a SW. Tento sortiment na trhu prochází neustálým vývojem, je zde sice i velká konkurence, ale zákazníci přesto rádi reagují na novinky na trhu. V tomto směru firma nenabízí žádné velké možnosti. Jedním z řešení by mohlo být upozorňovat současné zákazníky odebírající jiné produkty a služby na vlastní e-shop, jehož obchody jsou dost managementem firmy opomíjeny, stejně jako celé fungování, vzhled a kvalita internetového obchodu firmy.

V rámci obchodních strategií je třeba se soustředit na platební morálku zákazníků. Pokud již v dřívější době byly se zákazníkem problémy při úhradách faktur, bylo by vhodné smluvně sjednávat dílčí plnění jednotlivých zakázek, aby se v případě nepříznivého vývoje předešlo velkých výpadkům předpokládaných finančních prostředků. Jaké možnosti či naopak úskalí přinášejí a mohou přinášet obchodní kontrakty s jednotlivými kategoriemi zákazníků i v návaznosti na poskytovaný sortiment činností, již bylo popsáno výše.

Cíl celého projektu, tak jak byl ve fázi 1 metodologie CRISP-DM stanoven, spočíval v popisném pojetí závěrečné zprávy a ta je v rámci této kapitoly ve své podstatě sepsána. Z jednotlivých modelovacích technik či grafických výstupů byly získány zajímavé informace a skutečnosti, byla popsána možnost jejich využití a na základě celého projektu byly dány i doporučení pro management.

Závěr

V závěru této bakalářské práce je na prvním místě třeba připomenout, že modelování všech dat proběhlo na skutečně existujících reálných datech firmy a výsledky modelování a následná vyhodnocení mohou sloužit pro podporu rozhodování managementu a být vstupním klíčem při stanovení dalších strategických a obchodních cílů firmy.

Stanovené cíle práce byly naplněny. Bylo provedeno zkoumání datových vazeb, které nejsou na první pohled patrné. Informační systém firmy takovéto údaje neposkytuje, a přesto jsou velmi důležitou skutečností a potřebnou informací v životě firmy. Bylo postupováno podle metodologie CRISP-DM, jsou popsány všechny její fáze. Byť to není na první pohled třeba až tak patrné, ale tak, jak uvádí veškerá literatura týkající se této metodologie, fáze přípravy dat a vytvoření vstupních datových souborů pro modelování byla opravdu časově nejnáročnějším úsekem této bakalářské práce. Přesto se jednalo o velmi zajímavou část.

Stejně tak seznámení se a následná práce se softwarovým produktem Clementine byla velmi zajímavá. Možnosti grafického výstupu ze SW jsou z manažerského pohledu velmi příznivě přijímány. Mají lepší vypovídací schopnost než soubory číselných údajů. Proto byly ve velké míře využity při modelování a prezentovány při vyhodnocení.

Provedené vyhodnocení jednotlivých metod modelování je jednou z fází metodologie CRISP-DM a je popsáno podrobně výše.

Srovnáním let 2008 – 2010 je možné vidět, k jakým změnám na trhu dochází při poskytování činností v oblasti informačních technologií, jak se mění struktura zákazníků a tím i sortiment poskytovaných služeb. Je jasně patrný rozdíl mezi zákazníky ze soukromé sféry a zákazníky financovanými státním rozpočtem. Soukromá sféra byla první oblastí, kde docházelo k finančním úsporám. Státní restrikční opatření se výrazněji projevily až v roce 2010, kdy naopak soukromý sektor již zvolna začíná s prvními investicemi do sortimentu, které omezil nejvíce, neboť další setrvání na současném stavu by se těžce projevilo v jeho konkurenceschopnosti.

U všech zákazníků se projevil jasný trend, kdy jako první byly omezeny obchodní aktivity se sortimenty, které souvisejí s pořízením nových informačních technologií či komponent. Naproti tomu oblast služeb souvisejících s IT svým způsobem omezit nejde, proto finanční objemy ani počty obchodů v tomto poskytovaném sortimentu neprošly v minulých letech výraznými výkyvy.

Vzhledem k tomu, že není předpoklad velkých změn na trhu ani v roce 2011, jsou tyto závěry o to cennější.

Seznam použité literatury

[1] BASL, Josef; BLAŽÍČEK, Roman. *Podnikové informační systémy*. Praha : Grada, 2008. 283 s. ISBN 978-80-247-2279-5.

[2] BERKA, Petr. *Dobývání znalostí z databází*. Praha : Academia, 2003. 366 s. ISBN 80-200-1062-9.

[3] *Clementine® 11.0 Desktop User's Guide* [online]. 2006 [cit. 2011-01-25]. Clementine® 11.0 Desktop User's Guide. Dostupné z WWW: <http://www.forms.manchester.ac.uk/applications-media/document/clementine/11.0/ClementineUsersGuide_11.0.pdf>.

[4] *Clementine® 10.0 Desktop User's Guide* [online]. 2005 [cit. 2011-01-25]. Clementine® 10.0 Desktop User's Guide. Dostupné z WWW: <http://www.forms.manchester.ac.uk/applications-media/document/clementine/10.0/ClementineUsersGuide_10.0.pdf>.

[5] DOSTÁL, Petr; RAIS, Karel; SOJKA, Zdeněk. *Pokročilé metody manažerského rozhodování*. Praha : Grada, 2005. 166 s. ISBN 80-247-1338-1.

[6] *EuroMISE* [online]. 24.10.2002 [cit. 2011-02-16]. Proces dobývání znalostí. Dostupné z WWW: <euromise.vse.cz/kdd/index.php?page=proceskdd>.

[7] *EuroMISE* [online]. 24.10.2002 [cit. 2011-02-16]. Proces dobývání znalostí. Dostupné z WWW: <euromise.vse.cz/kdd/index.php?page=metody>.

[8] GÁLA, Libor; POUR, Jan; ŠEDIVÁ, Zuzana. *Podniková informatika*. Praha : Grada, 2009. 496 s. ISBN 978-80-247-2615-1.

[9] *IBM SPSS Decision Trees* [online]. 2010 [cit. 2011-03-19]. IBM SPSS Decision Trees. Dostupné z WWW: <http://www.spss.cz/sw_mcla.htm>.

[10] *IBM SPSS, Kurzy, Software, Data, Statistika, Data mining - SPSS CR* [online]. 2011 [cit. 2011-02-09]. IBM SPSS, Kurzy, Software, Data, Statistika, Data mining - SPSS CR. Dostupné z WWW: <www.spss.cz>.

[11] KOHONEN, Teuvo. *Self-Organizing maps*. Berlín : Springer, 2001. 501 s. ISBN 3-540-67921-9.

[12] MAŘÍK, Vladimír; LAŽANSKÝ, Jiří ; ŠTĚPÁNKOVÁ, Olga. *Umělá inteligence, sv.1*. Praha : Academia, 1993. 254 s. ISBN 80-200-0496-3.

[13] PETR, Pavel. *Data Mining. Díl 1*. Pardubice : Univerzita Pardubice, 2008. 139 s. ISBN 978-80-7395-098-9.

[14] POŽÁR, Josef. *Manažerská informatika*. Plzeň : Vydavatelství a nakladatelství Aleš Čeněk, 2010. 357 s. ISBN 978-80-7380-276-9.

[15] RUD, Olivia Parr. *Data Mining*. Praha : Computer Press, 2001. 329 s. ISBN 80-7226-577-6.

[16] ŘEZANKOVÁ, Hana; HÚSEK, Dušan; SNÁŠEL, Václav. *Shluková analýza dat*. Praha : Professional Publishing, 2007. 196 s. ISBN 978-80-86946-26-9.

[17] *Samoorganizace* [online]. srpen 2008 [cit. 2011-03-19]. Automatizace. Dostupné z WWW: <http://www.automatizace.cz/article.php?a=2174>.

[18] *Shluková analýza* [online]. 2004 [cit. 2011-04-02]. Shluková analýza. Dostupné z WWW: <http://staff.utia.cas.cz/nagy/skola/Projekty/Classification/ShlukovaAnalyza.pdf>.

[19] SKALSKÁ, Hana. *Data Mining a klasifikační metody*. Hradec Králové : Gaudeamus, 2010. 154 s. ISBN 978-80-7435-088-7.

[20] VONDRÁK, Ivo. *Umělá inteligence a neuronové sítě*. Ostrava : VŠB - Technická univerzita Ostrava, 2009. 139 s. ISBN 978-80-248-1981-5.

Seznam obrázků

Seznam použitých zkratek

CRISP-DM	Cross-Industry Standard proces for Data Mining (metodologie)
csv	Comma-separated values (hodnoty oddělené čárkami)
C&RT	Classification & Regression Trees (klasifikační a regresní algoritmus)
DIČ	Daňové identifikační číslo
DM	Data Mining
HW	Hardware
CHAID	Chi-squared Automatic Interaction Detection
ICT	Informační a komunikační technologie
IČO	Identifikační číslo organizace
IS	Informační systém
IS KP	Konkrétní informační systém analyzované firmy
ISO	Mezinárodní organizace pro normalizaci
IT	Informační technologie
KP	Vlastní zkratka pro název reálného existujícího informačního systému
KSOM	Kohonen Self-Organizing Map (Kohonenovy samoorganizující se mapy)
LAN	Local Area Network (lokální – místní síť)
PIS	Podnikový informační systém
QUEST	Quick, Unbiased, Efficient Statistical Tree
SOM	Self-Organizing Map (samoorganizující se mapa)
SQL	Structured Query Language (strukturovaný dotazovací jazyk)
SW	Software
TDIDT	Top-Down Induction of Decision Trees (indukce rozhodovacích stromů)
WAN	Wide Area Network (rozlehlá síť)

Seznam příloh

Příloha A – Datové struktury tabulek

TABULKA ZÁKAZNÍK				
NAZEV ATRIBUTU	**TYP**	**VELIKOST**	**POPIS**	**ROZSAH**
ide_part00000	***cislo***	***20***	***ID - číslo zákazníka***	***<48587564÷998756704386>***
ic	text	20	IČO zákazníka	
nazev_firmy	text	50	název firmy	
adresa2	text	50	adresa sídla firmy	
adresa3	text	50	adresa sídla firmy	
místo	text	50	město sídla firmy	
ide_psc00000	cislo	20	ID - poštovní směrovací číslo	<10÷996857776073>
ide_stat00000	cislo	20	ID - stát	<0÷5448088779144>
ide_pravf00000	cislo	20	ID - právní forma	<0÷612975214375>
ide_kateg00000	cislo	20	ID - kategorie zákazníka	<0÷26540511986>
dic	text	25	DIČ zákazníka	
certif_iso	text	25	certifikáty ISO	
poznamka	text	250	poznámka	
ide_platp00000	cislo	20	ID - platební podmínka	<0÷977494ž32781>
ide_dealer00000	cislo	20	ID - dealer	<0÷980694836471>
www	text	100	www stránky	

TABULKA PLATBY				
NAZEV ATRIBUTU	**TYP**	**VELIKOST**	**POPIS**	**ROZSAH**
ide_platby00000	***cislo***	***20***	***ID - platba***	***<6245407508÷988987828372>***
ide_skvf00000	cislo	20	ID - vystavená faktura	<1724655041÷998734157532>
obdobi	text	10	období vystavení faktury	
text	text	100	text	
castka	cislo	8	částka	<1÷18213782>
ide_part00000	cislo	20	ID - zákazník	<48587564÷998756704386>
datum_platby	datum	8	datum úhrady	<04071998÷15022011>
ide_dokluhr00000	cislo	20	ID - doklad úhrady interní	<5726931151÷999389891978>
neuhrazenost	cislo	5	dny po splatnosti	<-36÷351>

TABULKA KATEG				
NAZEV ATRIBUTU	**TYP**	**VELIKOST**	**POPIS**	**ROZSAH**
ide_kateg00000	***cislo***	***20***	***ID - číslo kategorie***	***<0÷26540511986>***
kategorie	text	50	název kategorie	
oznaceni	text	6	zkratka kategorie	

TABULKA SORT				
NAZEV ATRIBUTU	**TYP**	**VELIKOST**	**POPIS**	**ROZSAH**
ide_sort00000	***cislo***	***20***	***ID - číslo sortimentu***	***<0÷78505598474>***
sort	text	50	název sortimentu	
oznaceni	text	6	zkratka sortimentu	

TABULKA HLAV_FA				
NAZEV ATRIBUTU	**TYP**	**VELIKOST**	**POPIS**	**ROZSAH**
ide_skvf00000	***cislo***	***20***	***ID - vystavená faktura***	***<1724655041÷99873457532>***
ide_drady00000	cislo	20	ID - dokladová řada	<77539104269÷9703882359350>
rok_operace	cislo	4	rok	<1998÷2011.
cislo_operace	cislo	4	číslo operace - interní	<1-734>
doklad	text	25	doklad zákazníka	
ide_part00000	cislo	20	ID - číslo zákazníka	<48587564÷998756704386>
ide_dealer00000	cislo	20	ID - dealer	<0÷980694836471>
ide_stred00000	cislo	20	ID - středisko	<16494751328÷97307905675>
ide_czak00000	cislo	20	ID - číslo zakázky	<602779677÷996456764871>
var_sym_	cislo	10	variabilní symbol	<4080000001÷5090001452>
ide_dopr00000	cislo	20	ID - doprava	<0-482705880189>
dat_vystavení	datum	8	datum vystavení	<01071998÷15022011>
dat_zd_pln_	datum	8	datum zdanitelného plnění	<01071998÷15022011>
dat_splatnosti	datum	8	datum splatnosti	<12041998÷15042011>
castka	cislo	8	částka	<1÷18213782>
text	text	100	text	
ide_platby00000	cislo	20	ID - platba	<6245407508÷98898728372>
zakl_0	cislo	8	částka - základ dph 0	<1÷900426>
zakl_1	cislo	8	částka - základ dph 1	<0÷17346459>
zakl_2	cislo	8	částka - základ dph 2	<60÷800000>
dph0	cislo	8	částka - dph 0	<1÷180085>
dph1	cislo	8	částka - dph 1	<1÷867323>
dph2	cislo	8	částka - dph 2	<0÷72000>
ide_dph0_00000	cislo	20	ID - dph 0	<0÷949228017>
ide_dph1_00000	cislo	20	ID - dph 1	<0÷949228017>
ide_dph2_00000	cislo	20	ID - dph 2	<0÷949228017>
zaokr_	cislo	8	částka - zaokrouhlení	<0÷1>
bez_dph_celkem	cislo	8	částka celkem - bez dph	<1÷17514342>
dph_celkem	cislo	8	částka celkem - dph	<0÷1843081>

TABULKA POL_FA				
NAZEV ATRIBUTU	**TYP**	**VELIKOST**	**POPIS**	**ROZSAH**
ide_uni00000	***cislo***	***20***	***ID - položka faktury***	***<745886471÷999877428868>***
ide_skvf00000	cislo	20	ID - vystavená faktura	<1724655041÷998734157532>
ide_skkarta00000	cislo	20	ID - skladová karta	<51710754÷999959183391>
ide_mj00000	cislo	20	ID - měrná jednotka	<0÷79030193155>
ide_dph00000	cislo	20	ID - dph	<0÷949228017>
text	text	50	text	
cena_příj__vyd_	cislo	8	cena	<0÷17346459>
mnozstvi	cislo	8	množství	<1÷15000>
typ_radku	cislo	4	typ řádku	<1÷9>
poradi	cislo	4	pořadí řádku	<0001÷0078>

TABULKA POL_SORT				
NAZEV ATRIBUTU	**TYP**	**VELIKOST**	**POPIS**	**ROZSAH**
ide_skkarta00000	***cislo***	***20***	***ID - skladová položka***	***<51710754÷999959183391>***
cislo_polozky	cislo	8	interní číslo skladové položky	<12108005÷90090000>
nazev_skl_karty	text	100	název	
alt_cislo	text	50	alt.číslo výrobce-dodavatele	
ide_dph00000	cislo	20	ID - dph	<0÷949228017>
ide_mj00000	cislo	20	ID - měrná jednotka	<0÷79030193155>
ide_sort00000	cislo	20	ID - sortiment	<0÷78505598474>

Příloha B – Rozhodovací strom (zdroj: vlastní)