VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA PODNIKATELSKÁ
FACULTY OF BUSINESS AND MANAGEMENT

ÚSTAV INFORMATIKY
INSTITUTE OF INFORMATICS

VYUŽITÍ DATA MININGU V PERSONÁLNÍ AGENTUŘE
UTILIZATION OF DATA MINING FOR PERSONNEL AGENCY

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — **Bc. Erik Ondruš**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. Jan Luhan, Ph.D., MSc**
SUPERVISOR

BRNO 2017

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA PODNIKATELSKÁ

Zadání diplomové práce

Ústav:	Ústav informatiky
Student:	**Bc. Erik Ondruš**
Studijní program:	Systémové inženýrství a informatika
Studijní obor:	Informační management
Vedoucí práce:	**Ing. Jan Luhan, Ph.D., MSc**
Akademický rok:	2016/17

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

Využití data miningu v personální agentuře

Charakteristika problematiky úkolu:

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

Cíle, kterých má být dosaženo:

Cílem práce je segmentace kandidátů na základě scoringu s využitím dolování dat z datového skladu pro účely zefektivnění firemních procesů v oblasti zpracování poptávek a zlepšení osobního přístupu ke kandidátům.
K dosažení tohoto cíle bude klíčové určit hodnotu kandidáta na základě expertních odhadů a vytvořit algoritmus pro automatizaci této úlohy. Následně bude provedena segmentace kandidátů a vybrán vhodný predikční model k získání skupiny potenciálně úspěšných kandidátů. Na závěr budou předloženy návrhy využití dolování dat v oblasti přímého marketingu.
Celá práce bude dle požadavků zadavatele navržena nad platformou Microsoft, s využitím databázových nástrojů MS SQL Server 2014 (Analysis Services).

Základní literární prameny:

CIOS, K., W. PEDRYCZ, R. W. SWINIARSKI a L. A. KURGAN. Data mining: a knowledge discovery approach. 1st ed. New York: Springer Sience Business Media, 2007. 606 p. ISBN 978-0-387-33333-5.
DE VILLE, B. Microsoft data mining: integrated business intelligence for e-Commerce and knowledge management. 1st ed. Boston: Digital Press, 2001. 315 p. ISBN 15-555-8242-7.

DELEN, D. Real-world data mining: applied business analytics and decision making. 1st ed. New Jersey: Pearson FT press, 2015. 288 p. ISBN 01-335-5107-5.

HARINATH, S., R. PIHLGREN, D. G. LEE, J. SIRMON a R. M. BRUCKNER. Professional Microsoft SQL Server 2012 Analysis Services with MDX and DAX. 1st ed. Indianapolis: Jogn Wiley and Sons, 2012. 1176 p. ISBN 978-1-118-10110-0.

LACKO, L. Mistrovství v SQL Server 2012. 1. vyd. Brno: Computer Press, 2013. 640 s. ISBN 978-8--251-3773-4.

Termín odevzdání diplomové práce je stanoven časovým plánem akademického roku 2016/17

V Brně dne 28.2.2017

L. S.

doc. RNDr. Bedřich Půža, CSc.
ředitel

doc. Ing. et Ing. Stanislav Škapa, Ph.D.
děkan

Abstrakt

Tato diplomová práce se zabývá využitím technik dolování dat v oblasti segmentace a predikce nástupu kandidátů personální agentury. Získané výsledky by měly zefektivnit firemní procesy v oblasti zpracování poptávek a podpořit osobní přístup ke kandidátům.

První kapitola obsahuje nezbytné teoretické minimum z oblasti Business Intelligence, datových skladů, data miningu a marketingu. Dále je představena analýza současného stavu s důrazem na zachycení klíčového procesu zpracování poptávky. Poslední kapitola je určena samotnému návrhu a implementaci vlastního řešení nad platformou Microsoft SQL Server 2014. Na závěr jsou předloženy návrhy využití dolování dat v oblasti přímého marketingu.

Abstract

This master's thesis will look into the use of data mining in the area of segmentation and the prediction of onboarding candidates of a recruitment agency. The obtained results should serve to make company processes more effective concerning the processing of orders, and should also facilitate a more personal approach to candidates.

The first chapter includes imperetive theoretical bases from the studies of Business Intelligence, data warehouses, data mining and marketing. Thereafter an analysis of the current state is presented with a focus on the capture of the key processes in processing and order. The last chapter looks at the proposed solution and implementation on the platform Microsoft SQL Server 2014. To conclude there are proposals of utilizing data mining in direct marketing.

Klíčová slova

Data Mining, Business Intelligence, Datový sklad, SQL server, Přímý marketing, CRISP-DM, Segmentace, Predikce

Key words

Data Mining, Business Intelligence, Data Warehouse, SQL server, Direct Marketing, CRISP-DM, Segmentation, Prediction

Bibliografická citace

ONDRUŠ, E. *Využití data miningu v personální agentuře*. Brno: Vysoké učení technické v Brně, Fakulta podnikatelská, 2017, 106 s. Vedoucí diplomové práce Ing. Jan Luhan, Ph.D., MSc

Čestné prohlášení

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

V Brně dne 17. 5. 2017

_____

Ondruš Erik

Poděkování

Rád bych tímto poděkoval především svému vedoucímu práce, panu Ing. Janovi Luhanovi, Ph.D., MSc, za vstřícný přístup, cenné rady a odborné vedení práce. Dále děkuji svému oponentovi, panu Ing. Michalovi Janatovi za jeho čas a přínosné připomínky k práci.

OBSAH

ÚVOD

V současné době působí na českém trhu přibližně 1800 personálních agentur, které mezi sebou tvoří velké konkurenční prostředí. Jeden z rozhodujících faktorů pro udržení konkurenceschopnosti je umět správně využít dostupná data, pro podporu rozhodování. Nestačí tedy pouze umět data získat a ukládat, ale je klíčové z nich vytěžit správné informace a znalosti. Pokud data nejsou zpracována na informace, které transformujeme ve znalosti a ty později přeměníme v moudrost, jejich velký potenciál přijde nazmar.

V reakci na technologický rozvoj a shromažďování ohromného množství dat v datových skladech vznikl data mining, kterým se tato práce zabývá. Data mining je schopen poskytnout komplexnější porozumění datům pomocí nalezení nových vzorů, které dosud nebyly objeveny a zároveň vytvořit prediktivní modely pomocí kombinace tradiční statistické analýzy, umělé inteligence a technik strojového učení.

Data mining je často využíván v oblasti přímého marketingu při výběru oslovených klientů, segmentaci zákazníků či predikci událostí. Proto jsem se jej rozhodl v této míře uplatit v prostředí personální agentury s cílem segmentace kandidátů pro zefektivnění firemních procesů v oblasti zpracování poptávek a zlepšení osobního přístupu ke kandidátům. Tato práce se však nezabývá porovnáním ani výběrem dostupných nástrojů pro data mining, nicméně se zaměří přímo na platformu společnosti Microsoft, pod kterou je již zavedena firemní databáze, včetně řešení Business Intelligece.

Tato diplomová práce je rozdělena na tři hlavní kapitoly. První z nich obsahuje nezbytné teoretické minimum z oblasti Business Intelligence, datových skladů, data miningu a marketingu. Druhá část představuje analýzu současného stavu včetně zachycení klíčových procesů společnosti, pro kterou bude řešení navrhováno. Zároveň je zde provedena analýza rizik implementace tohoto řešení. Poslední kapitola je určena samotnému návrhu a implementaci vlastního řešení nad platformou Microsoft SQL Server 2014. Na závěr jsou předloženy návrhy využití dolování dat v oblasti přímého marketingu. Konkrétní cíle této práce a metodika jejich zpracování jsou uvedeny v následující kapitole.

CÍLE PRÁCE, METODY A POSTUPY ZPRACOVÁNÍ

Cíle

Cílem práce je segmentace kandidátů na základě scoringu s využitím dolování dat z datového skladu pro účely zefektivnění firemních procesů v oblasti zpracování poptávek a zlepšení osobního přístupu ke kandidátům.

K dosažení tohoto cíle bude klíčové určení hodnoty kandidáta na základě expertních odhadů a vytvoření algoritmu pro automatizaci této úlohy. Následně bude provedena segmentace kandidátů a vybrán vhodný predikční model k získání skupiny potenciálně úspěšných kandidátů. Na závěr budou předloženy návrhy využití dolování dat v oblasti přímého marketingu.

Celá práce bude dle požadavků zadavatele navržena nad platformou Microsoft, s využitím databázových nástrojů MS SQL Server 2014 (Analysis Services).

Metody a postupy zpracování

V analýze současného stavu budou identifikovány klíčové procesy společnosti. Bude vypracována PEST analýza, Porterův model pěti sil, McKinseyho model 7S a závěrem SWOT analýza. Zároveň proběhne analýza rizik za využití metody RIPRAN.

Pro odhalení kritérií, definující úspěšného kandidáta, budou využity expertní odhady a analýza historických dat ve spolupráci se společností. Samotný data mining se bude opírat o metodiku CRISP-DM.

Na závěr bude vypracováno ekonomické zhodnocení této práce.

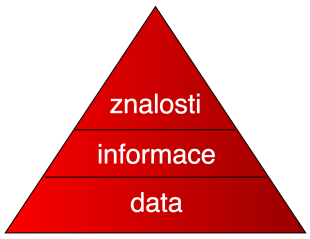
1 TEORETICKÁ VÝCHODISKA PRÁCE

V této kapitole jsou uvedeny základní teoretické znalosti a pojmy, ze kterých bude celá tato práce vycházet. Ve zkratce představím problematiku Business Intelligence, datových skladů, marketingu a dále se především zaměřím na samotnou oblast dolování dat.

1.1 Business Inteligence

Termín Business Intelligence (BI) poprvé definoval Howard Dresner takto: *„Business Intelligence je množina konceptů a metodik, která zlepší rozhodovací proces za použití metrik, nebo systémů založených na metrikách. Účelem procesu je konvertovat velké objemy dat na poznatky, které jsou potřebné pro koncové uživatele. Tyto poznatky potom můžeme efektivně použít například v procesu rozhodování a mohou tvořit velmi významnou konkurenční výhodu“* (2, str. 14).

Nejprve si musíme ujasnit, že data jsou jednoduchá, nezpracovaná fakta, která mají určitou důležitost pro jednotlivce nebo celou organizaci. Pokud tato data zpracujeme a získají určitou strukturu, která jim dává pro jednotlivce či celou organizaci význam, stávají se informací. (1, str. 36)



Obrázek 1 : Hierarchie informačních úrovní

Zdroj: (2, str. 15)

Business Intelligence jako proces transformace dat na informace a převod těchto informací na poznatky prostřednictvím objevování je jedním z nejdynamičtěji rostoucích odvětví na trhu informačních technologií. Nejvíce se využívá v podnikových informačních systémech, ale nachází si cestu i do zdravotnictví, výzkumu a vývoje. Hlavní přínosy řešení Business Intelligence spočívají v přechodu z intuitivního

rozhodování na rozhodování kvalifikované, realizované na základě kvalitních a operativně dostupných informací, které jsou dodány ve správný čas správným osobám. To umožňuje zlepšení obchodních a marketingových aktivit, možnost sledování a předvídání trendů a tedy v konečném důsledku hlavně zvýšení konkurenceschopnosti firmy. (2, str. 11)

Moderní databázové servery již obsahují rozsáhlou podporu pro budování datových skladů, OLAP a dolování dat, protože je často potřeba sledovat určité trendy či závislosti (například při obchodování s akciemi nebo detekci podvodů). (5, str. 354)

Výstup nebo samotná prezentace BI může mít poté různé formy, jako jsou například sestavy, dotazy, OLAP, ovládací panely či přehledy výsledků. Obecně známé sestavy jsou statické, obvykle předem plánované a spouštěné rutiny, které vytvářejí konkrétní přehledy. OLAP metoda je další formou dotazování, která doplňuje obvykle statické sestavy o dynamické procesy. Ovládací panely (dashboard) a přehledy výsledků (scorecard) představují další typ vykazování s důrazem na vizuální prezentaci, obvykle obsahují značně agregované klíčové indikátory výkonu, které informují o vývoji podnikových metrik a jejich aktuální hodnotě vzhledem k určitému předem určenému rozsahu. (4, str. 29-32)

Technologie business intelligence se v ideálním případě vyznačuje těmito vlastnostmi: (4, str. 26)

- Rozšíření možností – zajišťuje přímou použitelnost
- Rychlost – reaguje na požadavky
- Aktuálnost – je dostupná
- Přesnost – lze se spolehnout na kvalitu
- Užitečnost – poskytuje hodnotu

Nakonec můžeme stručně shrnout, že nástroj BI není nezávislým prvkem. Vyžaduje koordinaci se základní databází, architekturou a celkovým řešením. BI je tedy řešení, nikoliv pouze nástroj krychle či určitá sestava. (4, str. 33)

1.2 Datový sklad

Koncoví uživatelé se dotazují na data uložená v prostředí datového skladu a odpovědi na tyto dotazy jim poté pomáhají přijímat obchodní rozhodnutí. Dotazy mohou mít různou složitost od jednoduchých dotazů, analýz trendů, dolování dat pro asociativní analýzu, prediktivní analýzu budoucího vývoje až po kombinaci těchto a dalších postupů v závislosti na požadavcích podnikových uživatelů. (4, str. 36)

Pravděpodobně nejznámější definice datového skladu, jejímž autorem je Bill Inmon zní: *„Datový sklad je podnikově strukturovaný depozitář subjektově orientovaných, integrovaných, časově proměnných, historických dat použitých pro získávání informací a podporu rozhodování. V datovém skladu jsou uložena atomická a sumární data“* (2, str. 38).

Orientace na předmět – údaje se do datového skladu zapisují spíše podle předmětu zájmu než podle aplikace, ve které byly vytvořeny. (3, str. 360)

Integrovanost – datový sklad musí být jednotný a integrovaný. To znamená, že údaje týkající se jednoho předmětu se do datového skladu ukládají jen jednou. (3, str. 361)

Časová variabilita – údaje se ukládají do datového skladu jako série snímků, z nichž každý reprezentuje určitý časový úsek. (3, str. 361)

Neměnnost – Údaje v datovém skladu se obvykle nemění ani neodstraňují, jen se v pravidelných intervalech přidávají nové záznamy. To znamená, že transakční přístup a většina metod pro optimalizaci a normalizaci dat je nepotřebná. (3, str. 361)

1.2.1 Datový sklad jako jediný zdroj datové pravdy

U informačního systému, jehož součástí je datový sklad, musíme předpokládat, že nejlepší způsob jak dosáhnout odstranění redundance a s ní související nejednoznačností dat je, že datový sklad bude trochu nadsazeně řečeno jediným zdrojem datové pravdy v informačním systému. Uživatelé by tedy měli na všech úrovních kromě operační, kde data vznikají, vidět jen data z datového skladu. (3, str. 361)

1.2.2 Srovnání OLTP s datovými sklady

Tabulka 1: Srovnání OLTP s DW (1, str. 462)

OLTP systém	Datový sklad
Obsahuje aktuální data	Obsahuje historická data
Obsahuje podrobná data	Obsahuje podrobná, sumarizovaná data
Data jsou dynamická	Data jsou většinou statická
Vysoká průchodnost transakcí	Střední až nízká průchodnost transakcí
Předvídatelné vzorce použití	Nepředvídatelné vzorce chování
Řízení transakcemi	Řízení analýzou

OLTP systémy nejsou budovány tak, aby rychle zodpovídaly ad hoc dotazy, které zahrnují komplexní analýzu dat. Také obvykle neobsahují historická data, která jsou pro analýzu trendů nezbytná. (1, str. 462)

Údaje se z pravidla ukládají do operačních databází, které mohou být v různých odděleních firem, nebo dokonce i v jiných geografických lokalitách. Tyto data jsou v pravidelných intervalech sesbírány, předzpracovány a zavedeny do datového skladu. (3, str. 362)

1.2.3 Datové trhy

Datové trhy jsou v podstatě menší datové sklady, respektive jejich podmnožina, která může být vytvořena pro organizační jednotku společnosti na nižší úrovni hierarchie. Slouží například pro oddělení či geografickou lokaci, případně k ukládání a dalšímu zpracovávání dat pouze z některých vybraných oblastí podnikání. (3, str. 362)

1.2.4 Tabulky faktů a dimenzí

Tabulka faktů obsahuje numerické měrné jednotky obchodování kvalifikované podle dimenzí. Tabulka faktů je zpravidla největší tabulka v databázi a obsahuje velký objem dat. (3, str. 431)

Tabulky dimenzí jsou zpravidla menší než tabulky faktů a data v nich se nemění tak často. Zatímco dimenze ve všeobecnosti se stromovou hierarchickou strukturou obsahují relativně stabilní data, dimenze zákazníků, produktů a podobně se aktualizují častěji. Vysvětlují všechna „proč“ a „jak“, pokud se jedná o obchodování a transakce prvků. (3, str. 433)

Existuje ještě třetí typ, takzvané tabulky faktů bez faktů. Jsou to takové tabulky, které neobsahují žádné metriky. Tyto tabulky se vytvářejí kvůli sledování událostí. Dochází k tomu v případech, kdy tabulka faktů sdružuje dimenze a existence řádku ve faktové tabulce představuje výskyt dimenzí. (4, str. 162)

1.2.5 Budování datového skladu

Zásadním krokem při budování datového skladu je samotný výběr nejvhodnější metody. Musíme brát v úvahu nejen organizační strukturu a informační „kulturu“ firmy, ale také předvídat různé problémy, které se během budování datového skladu nevyhnutelně objeví. (2, str. 44)

Metoda velkého třesku

Jedinou výhodou této metody je skutečnost, že celý projekt lze kompletně vypracovat ještě před začátkem jeho realizace. Ale převažují zde však spíše rizika jako je například změna požadavků a také trvá velmi dlouho, než se projeví první výsledky obrovských investic do datového skladu.

Metoda velkého třesku se skládá ze tří etap (2, str. 44):

- Analýza požadavků podniku
- Vytvoření podnikového datového skladu
- Vytvoření přístupu buď přímo, nebo přes datové trhy

Přírůstková metoda

Budování datového skladu po jednotlivých etapách, tedy místo vybudování celého datového skladu postupně přibývají přírůstková řešení, která zapadají do celkové architektury datového skladu.

Přírůstková metoda se skládá z následujících kroků (2, str. 46):

- Strategie
- Definice
- Analýza
- Návrh
- Sestavení
- Produkce

Přírůstková metoda směrem „shora dolů“

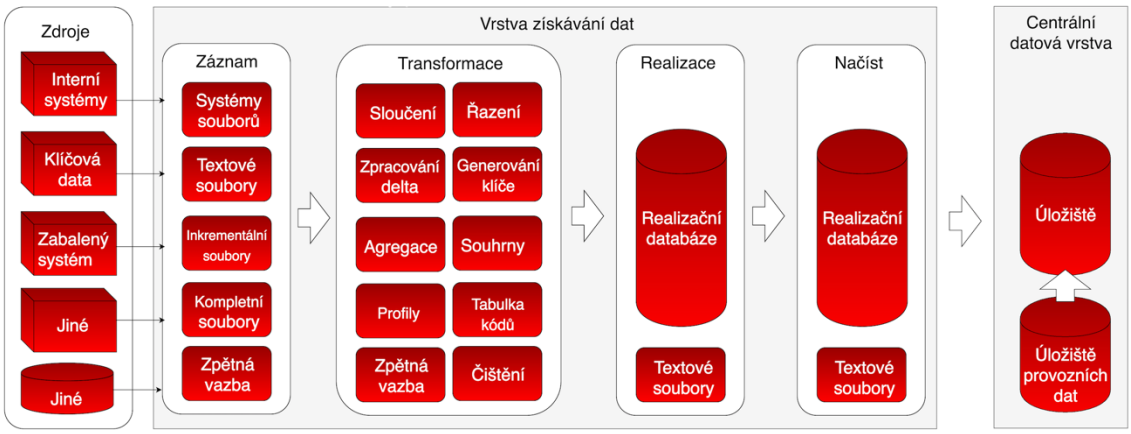
U této metody je nejprve na základě požadavků vytvořen konceptuální model datového skladu, přičemž důležitou roli hraje stanovení hierarchie předmětných oblastí. Následně jsou sestaveny konceptuální modely jednotlivých předmětných oblastí. Tato metoda poskytuje poměrně rychlou implementaci jednotlivých datových trhů, a tím i návratnost investic, je zatížena menším rizikem. (2, str. 46)

Přírůstková metoda směrem „zdola nahoru“

U této metody vystupuje do popředí IT oddělení podniku. Převažují zde spíše nevýhody. Protože se konceptuální model odvíjí od zdrojových systémů, je celková rozšiřitelnost v některých případech značně problematická. Navíc je IT oddělení zvyklé pracovat spíše s daty než s informacemi, proto není úloha hlavního realizátora projektu pro IT oddělení nejšťastnější řešení. (2, str. 46)

1.2.6 ETL a kvalita dat

Z obecného hlediska lze ETL popsat jako extrakci dat ze vstupního zdroje, transformaci těchto dat do příslušného formátu a následné nahrání dat do cílové databáze. Termín „transformace“ zahrnuje několik dílčích procesů a kroků, například: čištění, slučování, třídění, definování jedinečných identifikátorů, zajištění časových razítek, zpracování delta, vytváření dat, ověřování dat, zajištění referenční integrity, sumarizace a profilování dat. V podsystému ETL lze provádět libovolný počet uvedených kroků, což závisí na architektuře a na požadavcích řešení. Transformaci dat je možné provádět na neformátovaných souborech, v databázovém systému nebo v jejich kombinaci. (4, str. 242)



Obrázek 2: Vrstva získávání dat

Zdroj: (4, str. 257), Vlastní zpracování

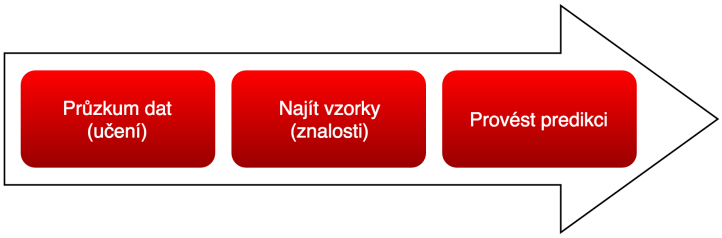
1.3 Data mining

„Data mining je proces analýzy dat z různých perspektiv a jejich přeměna na užitečné informace. Z matematického a statistického hlediska jde o hledání korelací, tedy vzájemných vztahů nebo vzorů v datech. Data mining je proces, jehož cílem je těžba informací v databázích. Využívá statistické metody a další metody hraničící s oblastí umělé inteligence“ (3, str. 572).

Využití data miningu, neboli dolování dat je úzce spjato s oblastí BI. Je principiálně založeno na heuristických algoritmech, neuronových sítích, umělé inteligenci a jiných pokročilých softwarových technologiích. Pomáhá nejen sledovat a analyzovat trendy, ale také předvídat určité události. Využívá se například v bankovnictví při analýze a predikci úrokového rizika, či ve zdravotnictví při analýze laboratorních vzorků. Je to ve své podstatě prostředek k získávání informací pro podporu rozhodování, nicméně samotné rozhodování musí provádět příslušný zodpovědný pracovník. (3, str. 572)

Typické problémy řešené pomocí data miningu: (3, str. 573)

- Segmentace zákazníků do skupin s podobnými vzory chování
- Profilování zákazníků pro řízení individuálních vztahů s nimi
- Identifikace zákazníků, kteří přinášejí největší zisk, a identifikace příčin a důvodů
- Zkoumání faktorů, které významně ovlivňují nákupní chování
- Predikce budoucího chování zákazníků na základě jejich historie a charakteristik
- Predikce neoprávněných pojistných událostí



Obrázek 3: Procesní schéma data miningu

Zdroj: (3, str. 573), Vlastní zpracování

1.3.1 Statistické metody využívané při dolování dat

Data miningové modely využívají některé statistické metody, mezi které patří hlavně korelace, lineární a logistická regrese, diskriminantní analýza a metody předpovídání. Složitější modely jsou také někdy realizovány pomocí neuronových sítí a genetický algoritmů. (3, str. 574)

Korelace

Jedná se o míru závislosti mezi dvěma proměnnými, která může být pozitivní či negativní. Pozitivní korelace udává, že vysoká úroveň jedné proměnné bude provázena vysokou úrovní korelační proměnné. Negativní korelace naopak udává, že vysoká úroveň jedné proměnné bude provázena nízkou úrovní korelační proměnné. (3, str. 575)

Lineární regrese

Je statistická metoda, která kvantifikuje závislost mezi dvěma spojitými proměnnými: a to závislou proměnnou, která je potřeba predikovat a nezávislou, tedy prediktivní proměnnou. Jejím cílem je nalezení parametrů přímky procházející mezi jednotlivými body, pro kterou platí, že součet druhých mocnin odchylek od každého bodu je minimální. (3, str. 575)

Logistická regrese

Logistická regrese je velmi podobná regresi lineární s tím rozdílem, že závislá proměnná není spojitá, ale diskrétní. Tuto regresi je možné použít k predikování výsledků dvou nebo více úrovní, například proč nastane jev nesplácení půjček. (3, str. 576)

Diskriminantní analýza

Měří důležitost faktorů určujících příslušnost do dané kategorie. (3, str. 576)

Předpovědi trendů

Jsou založeny na analýze údajů z minulosti, na základě kterých se definují určitá pravidla, pomocí nichž se predikuje budoucí trend dané proměnné. Používá se na příklad regrese časových úseků či neuronové sítě. Předpovědi trendů v sobě zahrnují i krátkodobé cyklické fluktuace. (3, str. 576)

Neuronové sítě

Neuronové sítě určitým způsobem simulují strukturu lidského mozku, kdy získávají poznatky z množiny vstupů a na jejich základě upřesňují své parametry modelu vzhledem k novým znalostem. Zpracování pomocí neuronových sítí nevychází ze žádného statistického rozdělení, ale pracuje podobně jako lidský mozek na principu rozpoznávání vzorů a minimalizace chyb. (3, str. 576)

Genetické algoritmy

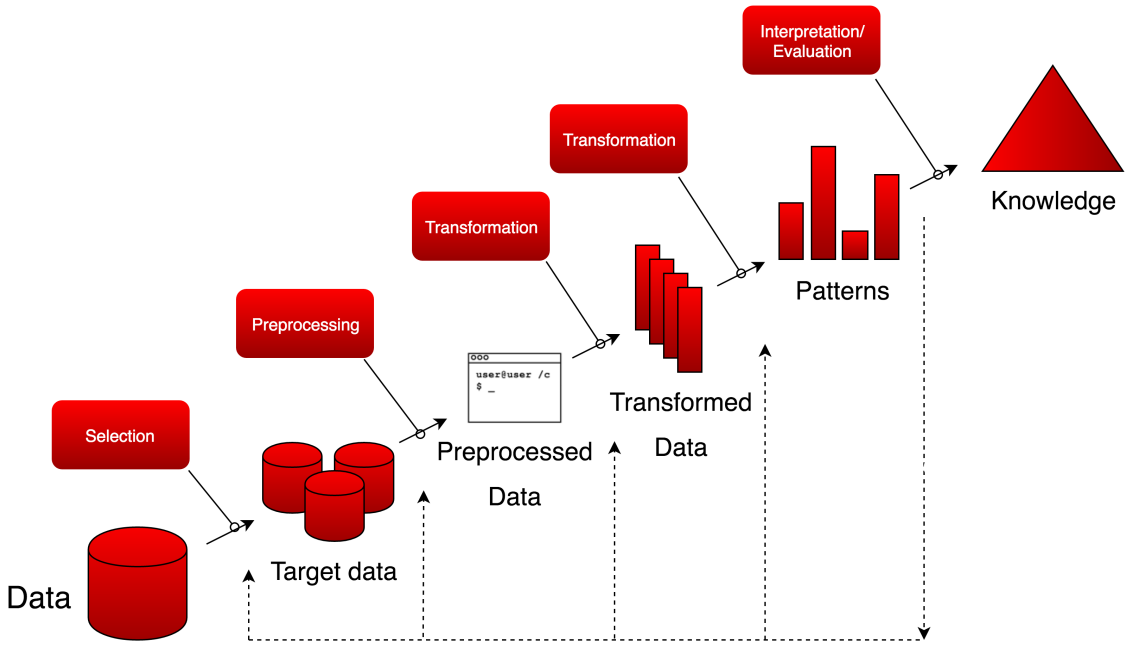
Genetické algoritmy předpokládají vliv evolučního procesu, kdy se porovnává více modelů, které se v jednotlivých krocích upravují křížením, mutací a klonováním, náhodnými výměnami hodnot, znamének a funkcí. Je to velmi náročná metoda na výpočetní kapacitu. (3, str. 576)

1.3.2 Metodologie data miningu

Existuje řada osvědčených postupů při aplikaci dolování dat ve firemních procesech. Mezi nejvíce rozšířené metodologie patří CRISP-DM (Cross-Industry Standard Process for Data Mining) a SEMMA (Sample, Explore, Modify, Model and Assess) obě tyto metodologie vychází z procesů KDD (Knowledge Discovery in Databases). (5, str. 37)

1.3.3 Knowledge Discovery in Database

Dobývání znalostí z databází (KDD) je proces, který využívá metody data miningu. Dle Usama Fayyada se jedná o proces netriviálního objevování implicitních, dopředu neznámých a potenciálních použitelných znalostí v datech. Nicméně nejde o samotný data mining, ten je pouze jedením z kroků tohoto procesu. Na níže uvedeném obrázku je znázorněn technologický proces KDD, který je složen z 5 fází: selekce, předzpracování, transformace, data mining, interpretace/vyhodnocení. (7, str. 183)

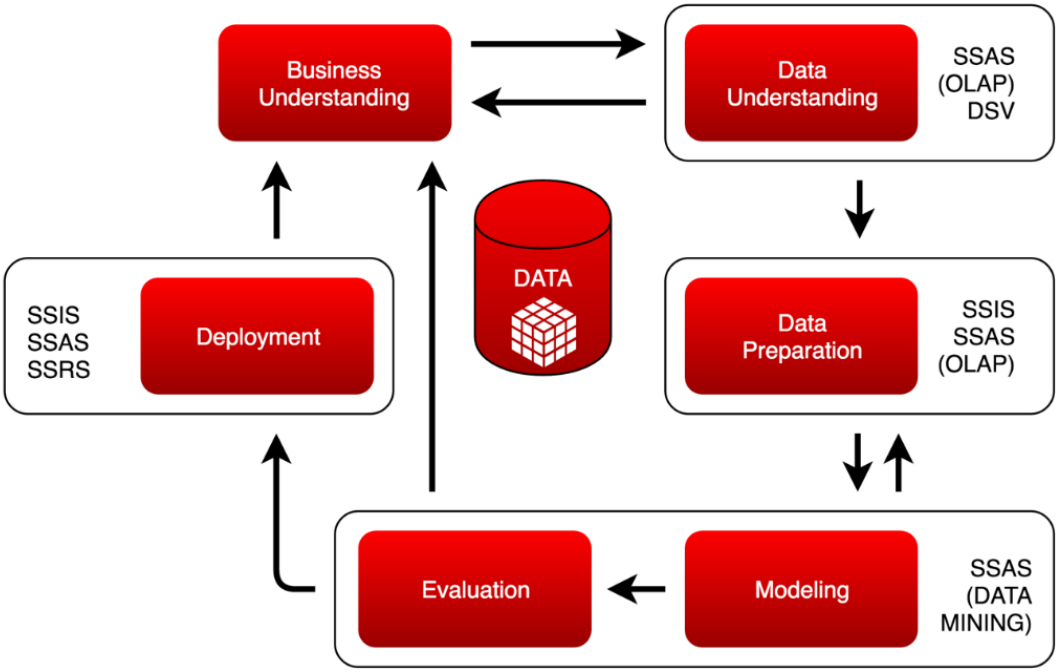


Obrázek 4: Proces KDD

Zdroj: (7, str. 183), Vlastní zpracování

1.3.4 Metodika CRISP-DM

Metodika CRISP-DM (CRoss-Industry Standard Process for Data Mining) vznikla v rámci výzkumného projektu Evropské komise. Cílem bylo navrhnout univerzální postup použitelný v různých komerčních aplikacích. Skládá se ze šesti fází, kdy jejich pořadí není přesně určeno. (8, str. 70)



Obrázek 5: Procesní schéma CRISP-DM

Zdroj: (3, str. 578), Vlastní zpracování

Výše uvedený obrázek popisuje základní fáze v procesu dolování dat a jejich významné vztahy. Základní fáze a činnosti dolování dat: (4, str. 18)

1. **Porozumění podnikání (Business Understanding)**

Představuje porozumění cílů a požadavků z obchodního hlediska, které mají být dolováním dat naplněny. Zahrnuje stanovení obchodních cílů, zhodnocení situace a plánování.

2. **Porozumění datům (Data Understanding)**

Je základním předpokladem úspěchu dolování dat a začíná počátečním sběrem a seznámením s daty. Umožňuje získat základní představu o datech, které máme k dispozici (popis dat, průzkum a posouzení kvality dat). Je to klíčový krok vedoucí ke správnému výběru techniky a algoritmům pro dolování.

3. **Příprava dat (Data Preparation)**

Tento krok se vztahuje na všechny činnosti potřebné k vytvoření cílové datové množiny, kterou budeme prostřednictvím technik data miningu zkoumat. Tato příprava zahrnuje selekci, čistění, transformaci, integraci a formátování dat.

4. **Modelování (Modeling)**

Je to fáze, kdy z velkého počtu modelů, technik a algoritmů vybíráme ty nejvhodnější kandidáty, kteří nejlépe popisují zkoumanou datovou množinu. Fáze modelování zahrnuje výběr modelovací techniky, generování testovacích vzorů, tvorba a vyhodnocení modelů.

5. **Vyhodnocení (Evaluation)**

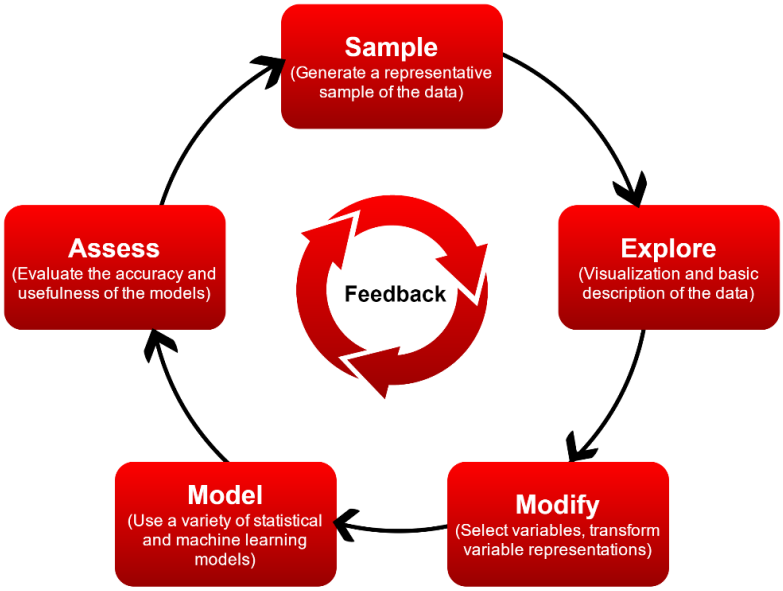
V tomto kroku ověřujeme kvalitu zjištěných informací, a zda jsou splněny všechny požadavky, které byly definovány na začátku procesu. Na konci této fáze by tedy mělo dojít k rozhodnutí z pohledu manažerů, zda budou výsledky přijaty, nebo bude nutné se vrátit až k prvnímu kroku.

6. **Nasazení (Deployment)**

V případě, že byly všechny požadavky naplněny, zjištěné informace odpovídají očekáváním a nebyla nalezena žádná významná nekonzistence, je možné přistoupit k nasazení finálního modelu do provozního prostředí. Tato fáze zahrnuje plán nasazení, monitorování a údržbu, závěrečnou zprávu.

1.3.5 Metodika SEMMA

Tato metodika byla vytvořena společností SAS Institute pro realizaci data miningového procesu. Jednoduše aplikuje explorační statistické a vizualizační techniky, vybere a transformuje nejvýznamnější predikční proměnné, modeluje proměnné k předpovědi výstupů a zhodnotí přesnost modelu. Jak již vyplývá z názvu, postup je tvořen tedy těmito pěti kroky: Sample, Explore, Modify, Model, Assess. (8, str. 78)



Obrázek 6: Procesní schéma SEMMA

Zdroj: (8, str. 78), Vlastní zpracování

1. Příprava vstupních dat (Sample)

Má za úkol vytvořit vhodný vzorek dat extrahováním z rozsáhlého objemu, který je dostatečně velký na to, aby obsahoval hledané informace a zároveň dostatečně malý, aby byla zaručena rychlá manipulace. Dolování z reprezentativních vzorků na rozdíl od celého souboru významně redukuje výpočetní čas. Data jsou dále rozděleny do následujících množin: (8, str. 79)

- Trénovací – pro vývoj modelu.
- Validační – pro vyhodnocení modelu a pro prevenci proti přeučení modelu.
- Testovací – pro finální vyhodnocení modelu.

2. **Průzkumová analýza dat (Explore)**

V tomto kroku získáváme základní představu o obsahu a kvalitě podkladových dat. Jestliže pomocí vizualizačních technik neodhalíme jasné trendy a závislosti v datech, můžeme je dále prohledat statistickými metodami v čele s faktorovou analýzou, korespondenční analýzou či shlukováním. (8, str. 80)

3. **Příprava dat pro analýzu (Modify)**

Tato fáze je založena na poznatcích z předchozího kroku, kdy modifikujeme vysvětlující proměnné v datové sadě. Pokud například v kroku Explore odhalíme lineární závislost mezi vstupními proměnnými, je vhodné některé proměnné odstranit, aby tato závislost zmizela. Dále můžeme odstranit extrémní hodnoty. Všechny tyto kroky pomáhají optimalizovat proces trénování. (8, str. 80)

4. **Výběr a odhad modelu (Model)**

Tento krok již realizuje vytvoření příslušného modelu. Mezi nejčastěji zde používané modelovací techniky patří např. neuronové sítě, rozhodovací stromy a další statistické modely, jako je analýza časových řad či regresní analýza. (8, str. 81)

5. **Interpretace a vyhodnocení výsledků (Assess)**

Představuje závěrečné zhodnocení úspěšnosti modelu a jeho případná implementace do praxe. (8, str. 81)

1.3.6 CRISP-DM versus SEMMA

Oba tyto přístupy jsou zcela kompatibilní a mají za cíl zjednodušit proces získávání znalostí. I když mají stejný cíl a jsou si v mnoha ohledech velmi podobné, mají několik rozdílů, které jsou uvedeny v následující tabulce. (8, str. 82)

Tabulka 2: Porovnání CRISP-DM a SEMMA (8, str. 82)

Task	CRISP-DM	SEMMA	Comments
Project Initiation	Business understanding	-	CRISP-DM includes activities like project initiation, problem definition, and goal setting. SEMMA does not have a step for this phase.
Data access	Data understanding	Sample	Both have steps access, sample and explore the data.
		Explore	
Data transformation	Data preparation	Modify	Both process the data to make it amenable to machine processing
Model building	Model building	Model	Both build and test various models
Project evaluation	Testing and evaluation	Assess	Both assess the findings against the project goals
Project finalization	Deployment	-	CRISP-DM prescribes deploying the results, while SEMMA does not explicitly have a step for this

1.3.7 Algoritmy pro data mining

V této části budou popsány nejvíce rozšířené algoritmy využívané pro dolování dat, jako jsou rozhodovací stromy, shlukování, asociační pravidla, časové řady, neuronové sítě a další. Je třeba podotknout, že neexistuje jednoznačné doporučení na výběr nejvhodnějšího algoritmu pro daný typ úlohy, ale výběr vždy závisí právě na konkrétním řešeném případu. (3, str. 579)

Rozhodovací stromy

Tento typ algoritmu odhaluje závislosti, vyhledává specifické vlastnosti a vzorce, které následně slouží k sestavení predikčního modelu rozhodování na jednotlivých úrovních hierarchické struktury. Jeho výhodou je rychlost, přehlednost, srozumitelnost a snadná interpretace. (3, str. 579)

Při využívání rozhodovacích stromů je nutné správně odhadnout optimální velikost množiny testovacích údajů. Pokud je příliš malá, strom nemusí být správně specifický a naopak pokud je příliš velká, strom může být přeučen. (3, str. 580)

Shlukování

Tento algoritmus se používá k odhalování shluků (clusterů) dat. Jedná se o proces organizování objektů do skupin, na základě jejich určité podobnosti, přičemž nejsou kritéria shlukování předem dána, ale odhalují se z přirozené struktury údajů. Tyto jednotlivé shluky se mohou a nemusí překrývat, to znamená, že jedna instance může patřit do jednoho čí více shluků. (3, str. 580)

Shlukování je vhodné použít například k identifikaci zákaznických segmentů, které jsou založeny na společných charakteristikách (demografické, sociální, profesní a podobně). (3, str. 581)

Sekvenční shlukování

Jedná se o specifický případ shlukování. Z matematického hlediska se jedná o aplikaci Markových procesů a modelů. (3, str. 581)

Asociační pravidla

Analýza na základě asociačních pravidel je zaměřena na odkrývání vztahů v datech a odhalování souvislostí přiřazování. Nejčastěji se pravděpodobně využívají při analýze nákupního košíku, respektive při hledání odpovědi na otázku, jaké zboží je nakupováno spolu. Z hlediska matematické statistiky se jedná o zkoumání korelace, ať už pozitivní či negativní. Pozitivní korelace udává, že vysoká úroveň jedné proměnné bude provázána vysokou úrovní korelační proměnné. Poznání této pozitivní korelace je důležité například při marketingových rozhodnutí, jaké zboží prodávat spolu, případně co nabídnout určitému zákazníkovi. (3, str. 582)

Časové řady

Tento algoritmus se zpravidla vztahuje k určité proměnné, například k obratu nákladům, zisku a podobně. Na základě analýzy údajů z minulosti a současnosti můžeme definovat určitá pravidla, pomocí nichž následně předpovídáme budoucí trend dané proměnné. Principiálně jde o regresi časových úseků, respektive předpovědi trendů v sobě zahrnující i krátkodobé cyklické fluktuace. (3, str. 582)

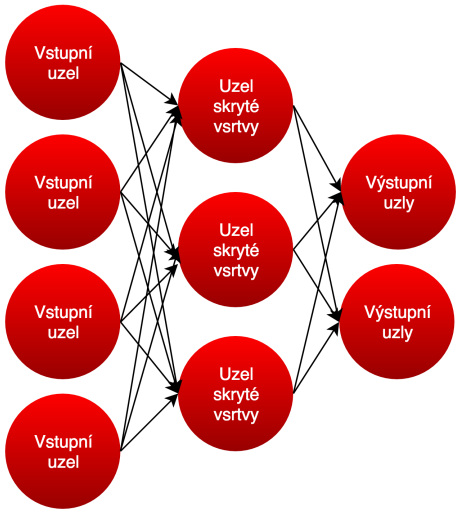
Neuronové sítě

Mezi základní přednosti neuronových sítí patří schopnost učit se, zobecňovat, identifikovat a reprezentovat závislosti ve vstupní datové množině, které nejsou zřejmé. Neuronové sítě tvoří uzly uspořádané do vrstev. Jejich základním principem je, že každý neuron obsahuje několik vstupů, které jsou ohodnoceny váhami a několik výstupů. Pokud celkový účinek všech vstupních podnětů překročí stanovený práh, dochází ke změně chování tohoto neuronu – aktivuje se a sám bude ovlivňovat následující neurony, bude tedy v roli vstupu pro následnou vrstvu neuronů. Aby se byly neuronové sítě schopny učit, musí být splněny následující předpoklady: dostatečné množství referenčních příkladů, tréninková, výběrová a testovací množina, výběr správného typu neuronové sítě a příslušného algoritmu. (6, str. 217)

Před začátkem vlastního procesu se údaje rozdělí do tréninkové a testovací množiny. Během každé iterace jsou vstupy zpracovávány systémem a jsou porovnávány se skutečnou hodnotou. Změří se chyba a odevzdává se ke zpracování systému, aby upravil

původní váhy. Tento proces končí zpravidla v okamžiku dosažení předem určené minimální chyby. Algoritmus sice podporuje predikci diskrétních i spojitých atributů, nicméně je nejvýhodnější použití pouze diskrétní a diskreditované atributy. (3, str. 584)

Neuronové sítě mají však také vlastnost, kterou lze označit jako negativní. S ohledem na způsob jejich chování a učení lze tento proces považovat za černou skříňku, jelikož z pohledu externího uživatele není zcela možné určit způsoby a postupy, na jejichž základě dospěla neuronová síť k výsledku. Vnitřní chování sítě je nám tedy neznámé. (6, str. 218)



Obrázek 7: Schéma neuronové sítě s jednou skrytou vrstvou

Zdroj: (3, str. 584), Vlastní zpracování

Na výše uvedeném schématu jednoduché neuronové sítě lze už na první pohled vidět velkou výhodu této architektury – údaje jsou totiž v jednotlivých vrstvách zpracovávány paralelně mnoha neurony. (3, str. 584)

Naive Bayes

Jedná se o velmi rychlý a přitom poměrně přesný algoritmus, založený na Bayesově větě, pomocí které je možné pracovat přímo s pravděpodobnostmi. Používá se na složitější analýzy, například k výpočtu paralelní korelace a kombinuje nová data s předchozími znalostmi. Využívá matematické pravidlo vysvětlující, jak se má změnit existující domněnka při konfrontaci s novými skutečnostmi. (3, str. 584)

1.4 Reporting

„Reporting představuje komplexní systém vnitropodnikových výkazů a zpráv, které syntetizují informace pro řízení podniku jako celku i jeho základních organizačních jednotek“ (10, str. 10).

Reporty lze rozdělit na dvě hlavní skupiny a to na statické a interaktivní. I přesto, že statické reporty mohou být v elektronické podobně, principiálně se nijak neliší od těch papírových. Je možné v nich především číst a listovat, nicméně není již možné je dále různě přizpůsobovat. Naproti reporty, které se řadí do skupiny interaktivních, je možné dále přizpůsobovat pomocí různých ovládacích prvků. Je možné tak získat informace, které právě potřebujeme, v takové formě, ve které je chceme prezentovat. (2, str. 324)

Dále je možné členit reporty dle oblasti a filozofie nasazení. První skupina je takzvaná Enteprise, která prezentuje data v podnikové informatice. Druhou skupinou je Embedded, kdy se generují reporty, jako integrální součást aplikací. Třetí skupinou je B2B, neboli Business To Business, kdy se jedná především o generování reportů pro obchodní partnery. (2, str. 324)

Uživatelů, kteří nahlíží do reportů bývá zpravidla mnoho a navíc mají velmi různící se požadavky. To klade na formální i obsahovou stránku reportingu velké nároky. Nedílnou součástí reportingu je i výběr, zpracování, formální úprava a distribuce informací, určených pro různé skupiny uživatelů. Přitom každý uživatel, respektive řídící pracovník by měl mít přístup pouze k těm informacím, které svou činností nějak ovlivňuje, a to v takové podobě, aby byla pro něj srozumitelná a přehledná. Naopak vrcholový management by měl mít zpravidla přístup ke komplexnímu systému informací. (10, str. 11)

1.5 Nástroje

V tomto oddíle jsou pouze ve zkratce představeny zásadní prostředky a nástroje, které jsou při tvorbě této práce využity. Jedná se o nástroje společnosti Microsoft, které společnost aktivně využívá.

1.5.1 Platforma Microsoft SQL Server 2014

Výčet veškerých možností, které tato platforma nabízí, by vydal na několik stovek stran. Základem je samozřejmě databázový server a sada komplexních nástrojů určených pro jeho konfiguraci, přístup, řízení, správu zpravidla s přehledným grafickým prostředím. Klíčovou vlastností jsou analytické služby a Business Intelligence. Data Miningová funkčnost je součástí analytických služeb SQL Serveru. Analytické služby (SSAS) musí být nainstalované v multidimenzionálním provedení. Tvorba data miningového řešení probíhá Visual Studio v rámci Multidimenzionálního projektu Analysis Services. (9)

1.5.2 SQL Server Management Studio a SQL Server Data Tools

SQL Server Management Studio je integrované prostředí na správu databázového serveru, jehož součástí je i prostředí pro zadávání a ladění SQL příkazů. Mimo jiné umožňuje vytvářet a spravovat nové databáze, uživatelské účty včetně oprávnění a podobně. SQL Server Data Tools je univerzální nástroj založený na rozhraní Visual Studia pro vytváření nejen aplikací Business Intelligence, ale také pro databázové vývojáře. Tento nástroj využiji především pro samotný data mining. (9)

1.5.3 Microsoft Excel

Jedná se o známý tabulkový procesor s dominantním postavením na trhu, který výborně zapadá do celkového ekosystému platformy MS SQL Server. Zajímavou možností je využít doplněk umožňující provádět data mining přímo v prostředí Excelu. Podmínkou je připojený MS SQL server ve verzi, která obsahuje Analysis services, jelikož samotné výpočty data miningu probíhají na straně SQL serveru. Nicméně pro méně zkušené uživatele jde o celkem přívětivé řešení v povědomém uživatelském prostředí. (9)

1.6 Marketing

V tomto oddíle bude popsáno nezbytné teoretické minimum, s důrazem na oblast direct marketingu.

Dnešní marketing je třeba chápat nejen jako schopnost přesvědčit a prodat, ale především jako uspokojování potřeb zákazníka. Marketing definujeme jako společenský a manažerský proces, jehož prostřednictvím uspokojují jednotlivci a skupiny své potřeby a přání v procesu výroby a směny produktů a hodnot. (11, str. 38)

1.6.1 Marketing management

Je to věda a umění zvolit cílové trhy a vybudovat s nimi ziskové vztahy. Marketing management zahrnuje řízení poptávky, které dále zahrnuje řízení vztahů se zákazníky. (11, str. 46)

1.6.2 Marketingový mix

Marketingový mix představuje soubor nástrojů, jejichž pomocí marketingový manažer utváří vlastnosti služeb nabízených zákazníkům. Jednotlivé prvky mixu může marketingový manažer namíchat v různé intenzitě i v různém pořadí. Slouží stejnému cíli: uspokojit potřeby zákazníků a přinést organizaci zisk. Původně obsahoval čtyři prvky (4P): **product**, **price**, **place**, **promotion**. Aplikace marketingové orientace v organizacích poskytujících služby ukázala, že tato čtyři P pro účinné vytváření marketingových plánů zcela nestačí. Výsledkem tedy bylo připojit další tři P. **physical evidence** - pomáhá zhmotnění služby, **people** - usnadňují vzájemnou interakci mezi poskytovatelem služeb a zákazníkem, **processes** - usnadňují a řídí poskytování služeb zákazníkům. (12, str. 22)

1.6.3 Marketingový komunikační mix

Jedná se o podsystém mixu marketingového. Je to specifická směs reklamy, osobního prodeje a public relations, kterou firma používá pro dosažení svých reklamních a marketingových cílů. Mezi pět hlavních komunikačních nástrojů patří reklama, osobní prodeje, podpora prodeje, public relations a přímý marketing. (11, str. 809)

1.6.4 Direct marketing

„Přímý marketing představuje přímou komunikaci s pečlivě vybranými individuálními zákazníky s cílem získat okamžitou odezvu a vybudovat dlouhodobé vztahy se zákazníky“ (11, str. 928).

I když existuje mnoho forem přímého marketingu jako je například direct mail, telemarketing, elektronický marketing a další, tak všechny sdílejí čtyři charakteristické rysy: (11, str. 837)

- Přímý marketing je okamžitý, protože sdělení lze připravovat velmi rychle.
- Je neveřejný, neboť sdělení je obvykle adresováno konkrétní osobě.
- Lze ho přizpůsobit tak, aby bylo sdělení přitažlivé pro konkrétní zákazníky.
- Je interaktivní: umožňuje dialog mezi komunikátorem a spotřebitelem a zprávy lze upravovat na základě reakce spotřebitele.

Z tohoto důvodu se přímý marketing výborně hodí pro přesně cílené marketingové snahy a budování individuálních vztahů se zákazníky. (11, str. 837)

Direct mail a E-mailing

Direct mail zahrnuje zaslání nabídky či oznámení konkrétní osobě. Rizikem tohoto oslovení je, že může být považováno za SPAM – tedy nevyžádanou poštu, pokud jsou osloveni nesprávní lidé. A stejně tak je tomu u elektronické pošty. (11, str. 837)

Rozesílání e-mailů patří k velmi účinným formám marketingové komunikace. Zároveň to však vyžaduje velké úsilí z hlediska přípravy obsahu, kvalitní databázi zákazníků a také překonání řady právních a technických překážek. Mezi výhody přímé komunikace se zákazníky prostřednictvím e-mailu patří například snadná personalizace, diferencovaný přístup, možnost okamžité reakce, nízké náklady a snadné vyhodnocování efektivity. Naopak mezi možné nevýhody patří nedoručitelnost (možnost „spadnutí“ do spamu), obtěžování zákazníka příliš častý zasíláním zpráv a technické problémy jako jsou špatná optimalizace pro e-mailové klienty či nezobrazení obrázků. (13, str. 204)

1.7 Analytické metody

V tomto oddíle jsou ve zkratce definovány metody, které budou využity v kapitole Analýza současného stavu.

1.7.1 Analýza vnějšího obecného okolí – PEST

Tento druh analýzy je považován za všeobecný a platí pro všechny organizace. Identifikuje klíčové trendy a vlivy, zajímá se o to, jaké vnější vlivy budou na různé organizace působit a jaké zde budou odlišnosti. Zkoumá: (14, str. 41)

- Politické a legislativní prostředí
- Ekonomické prostředí
- Sociální prostředí
- Technologické prostředí

1.7.2 Analýza oborového okolí – Porterův model pěti sil

Porterova analýza je zaměřena na analýzu struktury odvětví a je založena na předpokladu, že vývoj odvětví je funkcí jeho struktury. Předmětem hodnocení je těchto pět ukazatelů, které představují determinanty struktury odvětví: (14, str. 49)

- Vyjednávací síla zákazníků
- Vyjednávací síla dodavatelů
- Hrozba vstupu nových konkurentů
- Hrozba substitutů
- Konkurence existujících firem na daném trhu

1.7.3 Analýza vnitřních faktorů – Model 7S

Jak již vyplývá z názvu, McKinseyho model 7S, byl vytvořen firmou McKinsey a to již v sedmdesátých letech, aby pomohl manažerům lépe porozumět složitostem, které jsou spojeny s organizačními změnami v podniku. Model 7S je takto nazýván z důvodu, že obsahuje těchto sedm faktorů: (14, str. 73)

- Strategie (Strategy)
- Struktura (Structure)
- Systémy (Systems)
- Styl práce vedení (Style)
- Spolupracovníci (Staff)
- Schopnosti (Skills)
- Sdílené hodnoty (Shared values)

1.7.4 SWOT analýza

Cílem SWOT analýzy je sestavit reprezentativní seznamy pro silné a slabé stránky, příležitosti a hrozby. Tvoří základ strategické analýzy. Nejčastěji se využívá v předprojektových fázích, zejména ve studii příležitostí, může se však provést kdykoliv v průběhu řízení projektu. (15, str. 102)

1.7.5 Metoda RIPRAN

Metoda RIPRAN (Risk Project Analysis) je jednou z metod využívaných při řízení rizik. V její současné druhé verzi se skládá ze čtyř základních kroků: (15, str. 90)

- Identifikace nebezpečí projektu
- Kvantifikace rizik projektu
- Reakce na rizika projektu
- Celkové posouzení rizik projektu

V prvním kroku provádí projektový tým identifikaci nebezpečí sestavením seznamu nejlépe ve formě tabulky, která obsahuje číslo rizika, hrozbu, scénář a případně poznámku. Hrozbou se zde rozumí konkrétní projev nebezpečí a scénářem děj, který nastane v důsledku výskytu hrozby. Hrozba je příčinou scénáře. Ve druhém kroku se provádí kvantifikace rizika. Předchozí tabulka se rozšíří o pravděpodobnost výskytu scénáře, hodnotu dopadu na projekt a výslednou hodnotu rizika. Ve třetím kroku se sestavují opatření ke snížení hodnoty rizika. A v posledním se posoudí celková hodnota rizika a vyhodnotí se, zda je vhodná realizace projektu. (15, str. 91)

2 ANALÝZA SOUČASNÉHO STAVU

V této kapitole bude představen současný stav společnosti, včetně uvedení jejich klíčových či problémových procesů. Dále bude definována představa společnosti a její požadavky na řešení.

2.1 Představení společnosti

Vzhledem k povaze citlivých dat a informací uvedených v této práci jsme se s vedením dohodli, že společnost nebude jmenována.

Jedná se o společnost s ručením omezeným působící na trhu jako personální agentura již 15 let. Během její existence si vybudovala vedoucí postavení v poskytování recruitmentu a patří mezi 3 nejvýznamnější personální agentury v České republice.

2.2 Popis současné situace

Personální agentura v současné době disponuje nově vytvořeným informačním systémem, který byl navržen přímo pro její specifické požadavky a zaštiťuje tak většinu firemních procesů. Tento IS byl spuštěn v polovině roku 2015 s označením XIS4, kde číslo 4 vyjadřuje již čtvrtou generaci firemního IS. Zároveň bylo zavedeno Business Intelligence řešení, které zahrnuje firemní datový sklad sloužící jako centralizované, historizované úložiště a také automatický reporting zásadních informací pro podporu rozhodování. Databázová vrstva včetně BI běží pod platformou Microsoft SQL Server 2014.

Zaměstnanci společnosti dále využívají kancelářské aplikace z řady Microsoft Office 365, především však aplikaci Microsoft Excel, kde mají možnost spravovat výše zmíněné reporty. Společnost má za celou dobu svého působení nasazen ekonomický systém Pohoda. Komunikace mezi zaměstnanci a klienty probíhá často za využití aplikace Skype, zvláště pak pokud se jedná o videokonference. SMS zprávy adresované především kandidátům, jsou zpravidla zasílány přímo v prostředí informačního systému, který zároveň udržuje historii zaslaných zpráv, konkrétním uživatelem.

2.3 Analýza vnějšího a vnitřního prostředí

V této části jsou provedeny analýzy vnějšího obecného prostředí, oborového okolí a vnitřního prostředí.

2.3.1 Analýza vnějšího obecného prostředí PEST

Politické a legislativní prostředí

Mezi základní dokumenty, charakterizující zprostředkování zaměstnání patří Zákon o zaměstnanosti (Zákon č. 435/2004 Sb.). Zprostředkováváním zaměstnání agenturami práce se zabývají ustanovení § 58 až § 66 zákona o zaměstnanosti. V těchto ustanoveních jsou definovány podmínky, za kterých se fyzická nebo právnická osoba může stát personální agenturou. (18)

Ve dne 25. 5. 2018 vstoupí v platnost GDRP, neboli General Data Protection Regulation. Jedná se o novou legislativu EU, která má za úkol výrazně zvýšit ochranu osobních dat občanů. Týká se všech firem, institucí, ale i jednotlivců a online služeb, které zpracovávají data uživatelů. Toto nařízení zavádí celou řadu nových pravidel a tím přináší velkou administrativní zátěž. V případě porušení, nezavedení či nepřipravenosti na nové nařízení hrozí povinným subjektům vysoké pokuty, které mohou být v některých případech až likvidační. (19)

Ekonomické prostředí

Služby personálních agentur jsou pro kandidáty z pravidla zdarma, a proto nemá kupní síla obyvatelstva vliv na volbu konkrétní agentury. Ekonomické prostředí ovlivňuje především jejich klienty, kteří potřebují zaměstnance. Míra inflace, se tak značně projevuje například na zvýšení či omezení výroby společností a tím také na množství poptávaných zaměstnanců. Dalším významným faktorem je stav ekonomické krize, která vede k nezaměstnanosti, poklesu příjmů či nárůstu cen. Aktuálně se ale ekonomika nachází spíše ve fázi růstu. V roce 2016 ve 4. čtvrtletí nastal růst HDP o 1,9 %. Mezi roky 2015 a 2016 došlo ke změně průměrné roční inflace z 0,3 na 0,7. Spotřebitelské ceny meziročně vzrostly o 2,5 %, což bylo 0,3 procentního bodu více než v lednu 2017. (20)

Sociální prostředí

Dle Českého statistického úřadu je k 31. prosinci 2016 v České republice 10 578 820 obyvatel. Z toho v lednu 2017 bylo ekonomicky aktivních 76 % obyvatel, což znamená oproti lednu 2016 navýšení o 1,6 procentního bodu. (20)

Obecná míra nezaměstnanosti dosáhla 3,5 %, tato hodnota se meziročně snížila o 0,9 procentního bodu. V roce 2016 ve třetím čtvrtletí bylo v České republice 213 000 nezaměstnaných osob. Nejvyšší nezaměstnanost je v Moravskoslezském kraji s počtem 39 500 osob, dále v Jihomoravském kraji s počtem 26 300 osob, a také ve Středočeském kraji, kde je počet nezaměstnaných osob 21 200. Nejvyšší počet nezaměstnaných obyvatel dle jejich dosaženého vzdělání tvoří skupina se středoškolským vzděláním bez maturity a naopak skupina s nejmenším počtem nezaměstnaných je s vysokoškolským vzděláním. (20)

I přes dlouhodobě klesající nezaměstnanost, je zde stále řada lidí, kteří hledají nové pracovní příležitosti a tvoří tak potenciální kandidáty personální agentury.

Technologické prostředí

S rozvojem informačních technologií a nových možností rozšiřovat funkcionalitu podnikových informačních systémů, získává personální agentura příležitost neustále optimalizovat své firemní procesy a zvyšovat tak svou konkurenceschopnost. Jak již bylo zmíněno výše, agentura aktivně využívá svůj aktualizovaný informační systém, disponující řadou nástrojů Business Intelligence, zvláště pak reportovacími služby. Nicméně postrádá pokročilé analytické nástroje, které by uplatnila na svou rozsáhlou interní databázi a externí zdroje dat ve formě databází populárních pracovních serverů jako je LinkedIn, Jobs.cz, Profesia a podobně. I přesto, že Česká republika patří k zemím s nejvyššími cenami tarifů mobilních operátorů, dochází v současné době k mírnému tlaku na snížení cen jak jednotlivých tarifů tak zahraničního volání. Vzhledem k tomu, že personální agentura aktivně komunikuje právě především prostřednictvím mobilního telefonu, představuje případné snížení cen výrazné snížení nákladů na komunikaci s kandidáty a partnery. S neustále rostoucí popularitou sociálních sítí jako je například Facebook, LinkedIn či Twitter, vzniká společnosti velký prostor pro silnou inzerci za minimální náklady.

2.3.2 Analýza oborového okolí – Porter

Stávající konkurenti

Aktuálně působí na trhu přibližně 1800 personálních agentur, které mezi sebou tvoří velké konkurenční prostředí. Vzhledem k velkému počtu popíši jen část nejsilnějších konkurentů působících v Jihomoravském kraji, který je pro společnost nejvýznamnější.

Grafton Recruitment, s.r.o. – patří mezi přední poskytovatele řešení pro nábor, agenturní zaměstnávání, talent management, FDI projekty a lidské zdroje s více než 30 lety zkušeností. Působí v 6 zemích – Česká republika, Irsko, Maďarsko, Polsko, Severní Irsko a Slovensko – s 35 pobočkami a více než 500 zaměstnanci. Zaměřuje se na umístění profesionálů do soukromého i státního sektoru napříč spektrem oborů.

Advantage Consulting, s.r.o. – působí po celé České republice, se sítí poboček v Praze, Plzni, Brně, Ostravě, Olomouci, Hradci Králové a Ústí nad Labem. Jednotlivé divize se soustředí na obory: IT, Telco; Finance, Ekonomie, Administrativa; Marketing, Obchod; Logistika; Jakost ; Farmacie a zdravotnictví.

ANEX personální agentura – specializuje se na vyhledávání pracovníků, pracovní a psychologické poradenství, personální audity, Assessment/Development Centra. Pobočky má v Ostravě, Praze a Brně.

Trenkwalder, a.s. – nabízí komplexní služby v oblasti lidských zdrojů, od dočasného přidělení a výběru zaměstnanců, přes vzdělávání, assessment/develepment centra, testování až po personální poradenství a executive search. Má více než 65 trvalých zaměstnanců a pobočky v České republice, na Slovensku a v Bulharsku.

Potenciální konkurenti

Vstup nových konkurentů je ovlivněn několika bariérami. První z nich je kapitálová náročnost, jelikož založení personální agentury s sebou přináší počáteční investice ve formě pronájmů prostor, informačního systému, kancelářského vybavení a hardwaru, licencí, nákladů na propagaci a podobně. Nicméně takové počáteční investice s sebou nese téměř každá firma. Mezi podstatnější překážky se jeví to, že firmy jsou ochotné spolupracovat spíše s již zavedenými agenturami, které zároveň disponují obsáhlou

databází kandidátů. Nová agentura tedy musí budovat svou klientelu jak ze strany partnerů tak i ze strany kandidátů. Mezi formální požadavky patří například Zákon o zaměstnanosti určující podmínky, které musí být naplněny. Zároveň fyzická osoba musí dosáhnout věku 18 let, být způsobilá k právním úkonům, odborně způsobilá, bezúhonná a musí mít také bydliště na území České republiky.

Dodavatelé

Mezi klíčové dodavatele společnosti můžeme zařadit poskytovatele cloudových služeb, pod kterými běží celý informační systém. Další významný dodavatel je například telefonní operátor, jelikož konzultanti tráví většinu času právě telefonováním s kandidáty. V neposlední řadě je třeba zmínit inzertní servery, jako je například Jobs.cz a dodavatele reklamních předmětů, elektřiny či pracovních prostorů. Dodavatelské ceny tedy určují významnou složku podnikání a jakékoliv jejich zvýšení mohou být případnou hrozbou pro společnost.

Kupující

Mezi kupující služby personální agentury patří společnosti, které poptávají zaměstnance, ale z určitých důvodů jako je například personální či časová kapacita nemůžou obsadit pozici svépomocí. Tito zákazníci jsou hlavním zdrojem příjmu, jelikož za tuto využitou službu platí.

Na druhé straně stojí kandidáti, kteří aktivně hledají pracovní místo nebo jsou osloveni personální agenturou s danou pracovní nabídkou. Pro tyto kandidáty jsou služby zcela zdarma a netvoří tak přímo zdroj příjmu, ale klíčové aktivum společnosti.

Substituty

Společnosti, které nechtějí využít služby personální agentury mohou hledat zaměstnance vlastními silami. Nejčastěji vydávají inzeráty na vlastních webových stránkách, nebo pracovních portálech a uchazeči tak mohou reagovat na pracovní příležitost bez zprostředkovatele.

Dalším substitutem může být například úřad práce, který pomáhá hledat práci přímo jednotlivým uchazečům.

2.3.3 Analýza vnitřního prostředí 7S

Strategie

Společnost se snaží poskytovat nadstandardní služby oproti jiným personálním agenturám tak, že se soustředí na osobní a zejména individuální přístup ke svým partnerům. Konzultanti jsou vedeni k tomu, aby přistupovali ke každému kandidátovi osobitě a vyšli vstříc jeho nejrůznějším potřebám a požadavkům.

Dlouhodobým strategickým záměrem společnosti je trvale zvyšovat její hodnoty a vytvořit stabilizovaný tým zaměstnanců s potřebnou kvalifikací, který bude dále šířit její dobré jméno.

Společnost chce být solidním a spolehlivým obchodním partnerem i zaměstnavatelem, proto se snaží držet krok s moderními technologiemi, postupy a aktuálními nástroji v oblasti recruitmentu, které předává svým zaměstnancům v rámci pravidelných školení.

Sdílené hodnoty

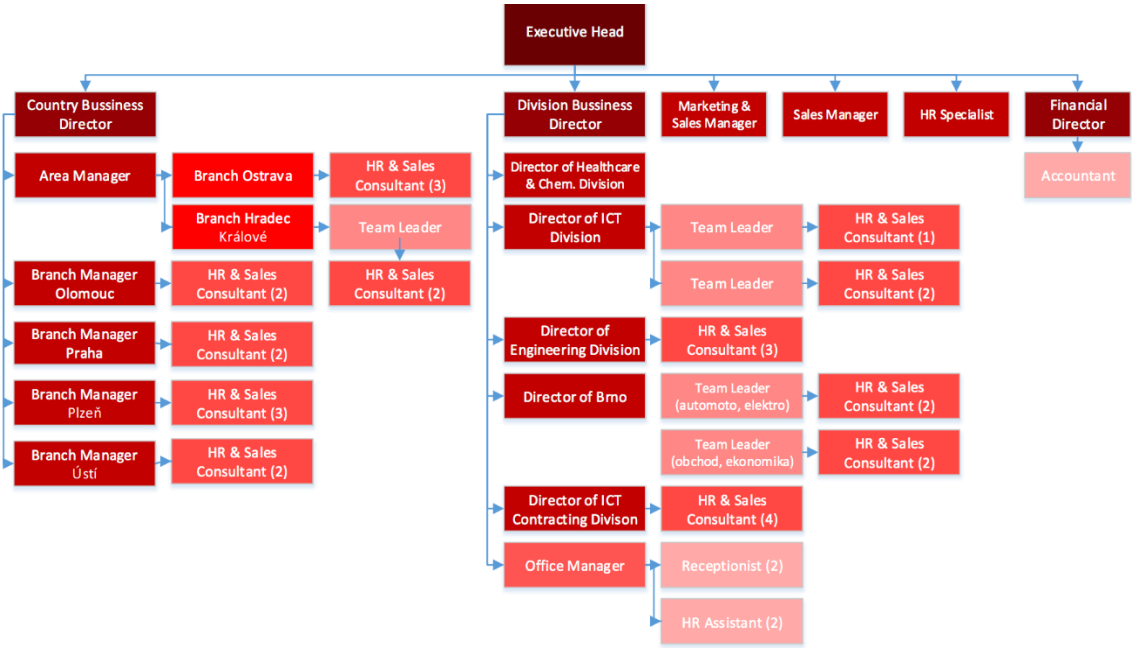
Společenská odpovědnost firmy a sociální politika jsou pro společnost velmi důležitá témata a prolínají se napříč její firemní kulturou. Za vnímání sociální odpovědnosti v oblasti zaměstnávání osob se znevýhodněním získali ocenění TOP Odpovědná firma 2014 a 2016 – cena veřejnosti a Stejná šance – zaměstnavatel 2014.

Hodnoty firmy:

- Respekt – chováme se k druhým tak, jak chceme, aby se druzí chovali k nám
- Spolupráce – naše společné úsilí je zaměřené na dosažení prospěchu všech, kteří se na něm podílejí
- Důvěra – věříme sami v sebe, věříme si navzájem, dodržujeme slovo a spoléháme jeden na druhého
- Odpovědnost – pracujeme tak, aby všechny naše činnosti měly pozitivní výsledek, nebojíme se být odpovědni za sebe a za svá rozhodnutí
- Prosperita – věříme v dlouhodobý růst a smysluplnost naší práce
- Profesionalita – jsme odborníky. Svoji práci vykonáváme efektivně, důsledně, s plným nasazením a maximální snahou o kvalitní výkon

Struktura

Společnost v současné době provozuje sedm poboček a sedm odborných divizí s více než osmdesáti konzultanty na divizi. Nejvyšší postavení zde zaujímá majitelka a zároveň jednatelka společnosti. Jednotlivé divize vedou divizní manažeři a ti mají pod sebou teamleadry, kteří vedou své konzultanty. Pozice konzultantů se dále člení na junior a senior pozice. Na níže uvedeném obrázku je znázorněna organizační struktura společnosti.



Obrázek 8: Organizační struktura společnosti

Zdroj: Vlastní zpracování

Styl vedení

Ve společnosti se postupně vytrácí direktivní styl vedení, delegování a přikazování, které se spíše transformuje do podpory a koučování jednotlivých konzultantů. Ve vedení je tedy uplatňován styl Laissez-faire a je založeno na kooperaci manažerů, teamleadrů a konzultantů. Manažeři se snaží o maximální předání zpětné vazby na pravidelných poradách. Teamleadři poté komunikují se svými podřízenými na denní bázi a jsou jim prakticky neustále k dispozici při jakékoliv otázce či problému.

Systémy

Systém řízení se dá označit za velmi flexibilní a přizpůsobující se situaci danému trhu, který je výrazně proměnlivý. Alespoň jednou měsíčně se konají strategické porady manažerů jednotlivých divizí, kde se projednává především cenová strategie a strategie vůči konkurenci, zaměstnancům a podobně. Dále každý týden se poté konají porady konzultantů a jejich teamleaderů, které vede manažer dané divize. Na těchto poradách se definují prioritní zakázky a přerozdělují úkoly.

Konzultanti jsou ohodnocováni fixní mzdou a také variabilní složkou, respektive bonusem, který je závislý na měsíčním obratu konzultanta. Takovým způsobem odměňování jsou zaměstnanci velmi motivováni k lepším výkonům, vzhledem k tomu, že bonus může přesáhnout dalekosáhle jejich fixní mzdu. Dále společnost pravidelně vyhlašuje soutěže o nejvýkonnějšího konzultanta, kde výhra představuje například luxusní zájezd na dovolenou.

Spolupracovníci

Mezi jednotlivými konzultanty panuje občasná rivalita a napětí, protože v nich obchodní prostředí vyvolává určitou míru soutěživosti, kde se snaží dosáhnout co nejlepších výsledků a získat tak patřičné benefity. Nicméně vedení, a především teamleadři se snaží udržovat spíše přátelské prostředí a vést zaměstnance k tomu, aby si byli navzájem nápomocní a dosahovali i dobrých kolektivních výsledků i přes mírně soutěživou atmosféru.

Schopnosti

Většina konzultantů disponuje vysokoškolským vzděláním v oblasti psychologie, personalistiky či ekonomie. Vysokoškolské vzdělání konzultantů však není striktní podmínkou pro přijetí. Důležitou vlastností jsou spíše jejich „soft skills“, jako je například komunikativnost, příjemné vystupování, asertivní jednání, flexibilita, zodpovědnost a v neposlední řadě chuť se stále vzdělávat. Jak jsem již zmínil dříve, konzultanti se se musí stále přizpůsobovat dynamickému trhu, proto jsou neustále školeni v nových trendech.

2.4 Analýza problému

V této části budou uvedeny jednotlivé procesy, jako jsou výběr vhodných kandidátů, inzerce poptávek či oslovení cílové skupiny.

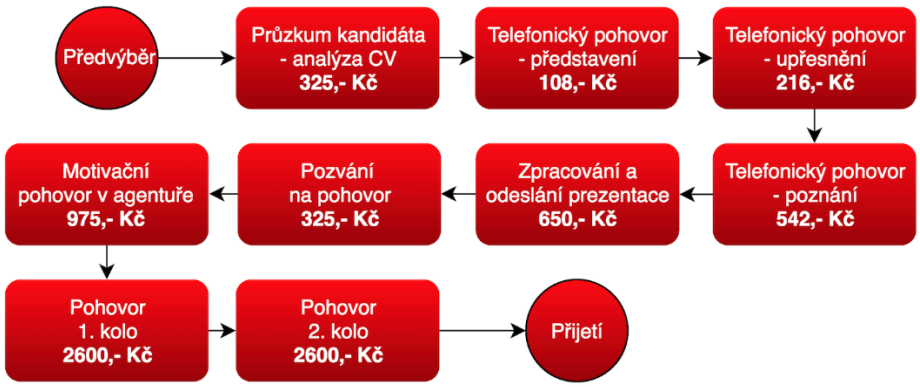
2.4.1 Proces výběru vhodného kandidáta

Každodenním opakujícím se procesem ve společnosti, je výběr vhodných kandidátů k zadaným poptávkám. Poptávka představuje otevřenou pozici u klienta, která má být do určitého časového období obsazena. Společnost si zakládá na detailní specifikaci otevřených pozic, tak aby měli konzultanti jasnou představu o hledaném profilu kandidáta. Konzultanti mají k dispozici seznam otázek, které musí s klientem během specifikace projít. Hlavní body v seznamu jsou délka praxe, vzdělání, znalosti, mzdové požadavky, lokalita a osobní předpoklady. Specifikace otevřené pozice probíhá mezi konzultantem a klientem telefonicky, emailem nebo během osobního setkání.

Proces specifikace pozice je nejdůležitější částí Recruitmentu, protože bez jasné představy o hledaném profilu, není možné správně cílit, a tak efektivně obsadit poptávku. Doba hledání vhodného kandidáta musí být co nejkratší. S časem hledáním vhodných profilů na poptávku rostou společnosti náklady a riziko, že poptávka bude obsazena konkurenční agenturou nebo si ji klient obsadí sám.

Společnost se neustále snaží optimalizovat proces Recruitmentu tak, aby konzultanti vybírali ty nejvhodnější kandidáty. S každým vybraným kandidátem provede konzultant telefonický pohovor, zpracuje prezentaci a pozve ho na osobní motivační pohovor do agentury. V případě, že se potvrdí vhodnost kandidáta, domluví s klientem termín výběrové řízení, kterého se účastní také odpovědný konzultant. Je tedy zřejmé, že výběr nevhodného kandidáta společnost stojí nemalou částku, protože je mu věnováno velké množství času.

Na níže uvedeném obrázku jsou uvedeny průměrné náklady na jednotlivé kroky procesu, kterým je provázen kandidát od jeho předvýběru až po jeho přijetí či zamítnutí. Tyto orientační náklady na jednoho kandidáta definovala společnost jako průměrné hodnoty získané za období jednoho roku od nasazení nového IS.



Obrázek 9: Náklady na provedení kandidáta procesem obsazení

Zdroj: Vlastní zpracování

Každým jednotlivým podstoupeným krokem se náklady na kandidáta kumulují, celková částka za takto provedeného kandidáta celým běžným procesem byla tedy odhadnuta na 8 341,- Kč. Jelikož každý jednotlivý krok může skončit neúspěšně a nejmenší úspěšnost provází právě první kroky, je pro společnost zásadní oslovovat co nejmenší počet kandidátů s nejlepšími předpoklady pro přijetí.

Předvýběr

Vzhledem k zavedení nového IS je proces výběru vhodného kandidáta k poptávce značně zjednodušen a do jisté míry automatizován. Každý registrovaný kandidát má v systému svou kartu, kde jsou jeho osobní informace, vzdělání, praxe a preferované místo výkonu práce. Veškeré tyto údaje jsou uloženy ve formě strukturovaných dat a zároveň jsou zde přiloženy další dokumenty, jako jsou životopis či motivační dopis. Naproti kandidátovi stojí poptávka, která má rovněž definované požadavky na pozici v obdobné formě. Prakticky se jedná o tabulky s atomickými atributy, které je možné mezi sebou velmi snadno porovnávat.

Proto je informační systém schopen na základě několika těchto atributů spárovat vyhovující kandidáty k poptávce a nabídnout tak pouze ty, kteří splňují všechny nebo vybrané podmínky. Nicméně systém takto může vyhodnotit desítky až stovky vhodných kandidátů, a proto není možné je vždy všechny přezkoumat. Konzultant obvykle vybere subjektivně či zcela náhodně menší vzorek o velikosti 30-50 kandidátech, které jde dále přezkoumat.

Průzkum kandidáta

V této fázi má konzultant možnost více analyzovat vybrané kandidáty. Především se jedná o podrobný průzkum jejich životopisů, profilů na sociální síti LinkedIn a podobně. Tímto krokem projde zhruba 75 % nejzajímavějších kandidátů, kteří jsou dále osloveni. Naneštěstí plná shoda hard-skills a pozitivní závěr průzkumu nezaručuje úspěšné nastoupení či alespoň jeho zájem o tuto pozici.

Ne vždy kandidát, který splňuje veškeré požadavky na praxi a vzdělání je zaměstnavatele vhodný či má o pracovní příležitost zájem. O tom rozhoduje řada další faktorů, které jsou objeveny postupně během telefonického pohovoru a motivačního či oficiálního pohovoru.

Telefonický pohovor

Telefonický pohovor má 3 klíčové části. V první části konzultant nastíní v co nejkratším možném čase, o jakou pozici se jedná a zda má kandidát zájem tuto pracovní nabídku dále probrat. Tato část zpravidla netrvá déle než 5 minut. Asi 50 % z oslovených kandidátů projeví zájem pokračovat v hovoru a chce se dozvědět více.

V další části je představena pracovní nabídka, konkrétně její popis, nabízená mzda, zajímavé benefity a hlavně požadavky. Konzultant se snaží ověřit klíčové znalosti uvedené v životopise. Například u pozice, kde je klíčová znalost cizího jazyka, může kandidáta přezkoušet, jelikož se jedná o jednu z nejvíce přeceňovanou schopností uváděnou v životopise. Tato část trvá zhruba 10 minut a projde jí přibližně 30 % kandidátů.

Pokud konzultant nenarazí na žádnou vážnou nesrovnalost a kandidát má stále zájem o pozici, přistoupí se k poslední části hovoru. Tato část je časově nejnáročnější, jedná se o upřesnění časových možností kandidáta. Naplánování motivačního pohovoru a získání upřesňujících informací k vytvoření prezentace kandidáta.

Zpracování prezentace kandidáta

Na základě případného upřesnění informací z telefonického pohovoru a životopisu kandidáta je zpracován strukturovaný dokument – interně nazýván „šablonka“.

Jedná se o prezentaci kandidáta, ve kterém je souhrn významných vlastností a informací, které jsou prezentovány klientovi. Je to klíčový dokument, rozhodující o pozvání kandidáta na pracovní pohovor.

Pozvání na pohovor

Pokud klient zamítne či naopak přistoupí na pozvání kandidáta k pracovnímu pohovoru, společnost ho informuje telefonicky. V případě pozvání spolu upřesní nejprve datum a čas motivačního pohovoru v agentuře a následně také pracovního pohovoru.

Motivační pohovor v agentuře

Motivační pohovor je součástí služby personální agentury. Trvá průměrně asi 30 minut a jedná se v podstatě o důkladné připravení kandidáta na skutečný pohovor a upřesnění jeho průběhu.

Pracovní pohovor

Pracovní pohovor má zpravidla 1 až 2 kola, kterých se účastní osobně také odpovědný konzultant, proto i tato fáze představuje pro společnost velké náklady. V případě neúspěchu, firma v podstatě vynaložila veškeré dosavadní náklady za proces zbytečně. Nicméně, pro další jednání se stejným kandidátem o jiné pracovní pozici nebude již nutné vynakládat všechny kroky.

Je nutné podotknout, že celý tento popsaný proces se týká pouze registrovaných kandidátů v systému a pro oslovení široké veřejnosti slouží další procesy, které tomuto ve své podstatě předchází.

2.4.2 Přímé oslovení pomocí komunikačních kanálů

Pak je zde skupina uživatelů, která svolila k zasílání nabídek pracovních příležitostí prostřednictvím e-mailů, zpráv na sociálních sítích či dokonce formou SMS. Nicméně i když uživatel svolí k takové formě oslovení, které je špatně zacíleno, stává se z něj z pravidla SPAM, neboli nevyžádaná pošta.

E-mail

Pokud budeme mluvit o e-mailovém marketingu, tedy o cíleném zasílání pracovních nabídek, je zřejmé, že jsou zde pouze nepatrné náklady a rozdíl mezi zasláním stovky či tisíce e-mailový zpráv bude zanedbatelný. Společnosti tedy z pravidla necítí tak velkou potřebu zvolit správnou cílovou skupinu a oslovit pouze relevantní osoby, jako je tomu například u telefonického hovoru, kdy jsou náklady přímo úměrné jejich objemu.

Zde je nutné se dívat na problematiku z jiného pohledu. Na jednu stranu zasláním nabídek všem kandidátům bez ohledu na jakékoliv zacílení, bude sice statisticky vzato znamenat určitou jistotu odezvy, nicméně i tak spíše při velmi nízké konverzi. Problém nastává ve fázi, kdy společnost takto zahltí všechny své kandidáty, kteří po čase začnou zprávy naprosto ignorovat a zároveň se velmi snižuje úroveň a důvěryhodnost firmy. Velmi snadno tak můžeme jedním kanálem poškodit celý komunikační mix. Společnost se tento problém snaží řešit alespoň hrubým kategorizováním pracovních nabídek na ty, které má předdefinován kandidát, ale při pohledu na statistiky reakcí zjistíme, že to pravděpodobně nestačí.

Stejný problém se týká typických newsletterů obsahující čerstvé informace o probíhajících seminářích, zajímavých novinkách, či odborných publikacích. Tyto zprávy jsou ale spíše více obecné a je tedy i vyšší pravděpodobnost zaujetí širší skupiny příjemců.

SMS

Společnost uvažuje o zasílání exkluzivních poptávek kandidátům i prostřednictvím SMS brány. Nicméně vzhledem k nákladům převyšující zprávy e-mailové a zároveň stejného, ne-li vyššího problému, který je popsán výše, je zásadní, aby cílila ještě lépe.

Sociální sítě LinkedIn a Facebook

Obě tyto sociální sítě nabízí přímé oslovení pomocí soukromé zprávy, která je však omezena určitou vzájemnou vazbou. Dále je možné využít propagované příspěvky, které je možné cílit na určitý okruh uživatelů dle zadaných parametrů. Volba těchto parametrů musí být pečlivě uvážena, jelikož platba probíhá za každého osloveného uživatele, nikoliv proklik.

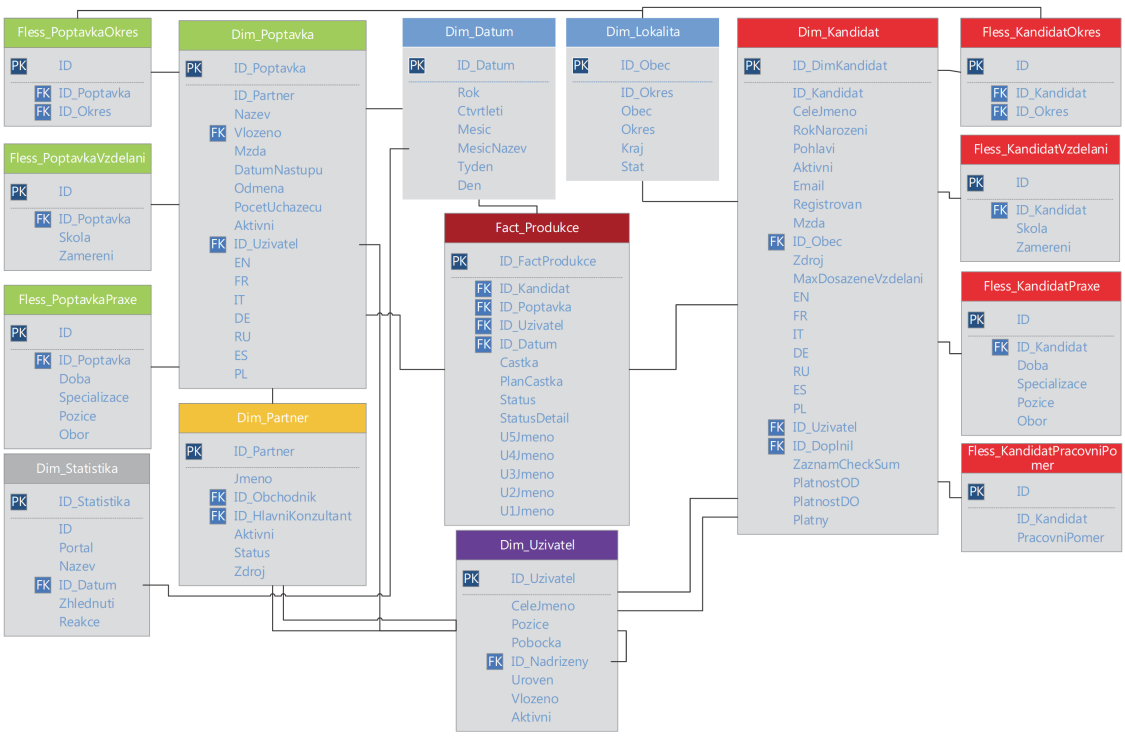
2.5 Analýza vstupních dat

Vzhledem k tomu, že se data mining prakticky obejde bez datových kostek, použiji jako primární zdroj dat přímo datový sklad. Rovněž zde integruji automatické hodnocení kandidáta. Datový sklad jsem vytvořil pro společnost již dříve, jako součást mé bakalářské práce a je aktivně využíván pro pravidelně generované reporty.

2.5.1 Datový sklad

Pro potřeby hodnocení kandidáta se zde nacházejí veškeré dříve definované informace. Samozřejmě bude ale nejprve nutné doplnit dimenzi *Dim_Kandidat* o atribut výsledek hodnocení.

Níže uvedené schéma zobrazuje podstatnou část současné podoby datového skladu. Pro mne je nejvýznamnější červeně označená část, která se týká přímo kandidátů.



Obrázek 10: Schéma datového skladu

Zdroj: Vlastní zpracování

Níže budou popsány atributy pro aplikaci data miningu, které jsou dostupné v datovém skladu.

Dim_Kandidat – Definuje základní údaje o kandidátovi.

ID_Kandidat – Unikátní identifikátor kandidáta.

CeleJmeno – Jméno Příjmení, Titul kandidáta.

RokNarozeni – Rok narození kandidáta.

Pohlaví – Pohlaví kandidáta (muž, žena, null).

Aktivni – Hodnota, která určuje, zda je kandidáta možné obsadit.

Registrovan – Datum registrace kandidáta do systému agentury.

Mzda – Minimální požadovaná mzda kandidáta.

ID_Obec – Identifikátor obce, ve které kandidát přebývá.

Zdroj – Zdroj získání kandidáta, například doporučení, pracovní portál, inzerát.

MaxDosazeneVzdelani – Maximální dosažené vzdělání kandidáta – úroveň školy.

EN, FR, IT, DE, RU, ES, PL – Jazyková úroveň jednotlivých jazyků (A1, A2, B1, B2, ...).

ID_Uzivatel – Uživatel, který vložil kandidáta do systému.

ID_Doplnil – Uživatel, který doplnil informace do karty kandidáta.

Fless_KandidatOkres – Definuje okres, ve kterém kandidát hledá pracovní místo.

ID_Okres – Identifikátor okresu.

Fless_KandidatVzdelani – Definuje vzdělání kandidáta. Školy, které absolvoval.

Skola – Úroveň vystudované školy (Střední škola, Vysoká škola, …).

Zamereni – Zaměření vystudovaného oboru (Ekonomie, Strojírenství, …).

Fless_KandidatPraxe – Definuje kandidátovu dosaženou praxi.

Doba – Doba jednoho zaměstnání, respektive pracovní pozice.

Specializace – Specializace zaměstnání (Java, C#, …).

Pozice – Pozice, kterou kandidát vykonával (Tester, Team leader, …).

Obor – Nadřazený obor pozice (Ekonomie, IT, …).

Fless_KandidatPracovniPomer – Definuje preferovaný pracovní poměr.

PracovniPomer – Typ pracovního poměru (HPP, Živnostenský list, …).

2.6 Výběr platformy

Již v teoretické části jsem popisoval systém Microsoft SQL Server 2014 a to z důvodu, že právě tuto platformu personální agentura dlouhodobě využívá. Společnost řešením od Microsoftu právem plně důvěřuje, a proto veškerá databázová vrstva, včetně datového skladu a řešení BI běží právě na této platformě.

I přestože tedy existuje velká řada jak Open-Source tak i dalších komerčních nástrojů, je vzhledem k situaci naprosto vhodné držet se této úspěšně zavedené platformy. Navíc s integrací nového doplňku dolování dat do uživatelům známého prostředí aplikace Microsoft Excel, se otvírají zkušenějším uživatelům nové příležitosti a možnosti data mining aktivněji využívat. Samozřejmě by obsluha tohoto doplňku vyžadovala nejprve vhodné školení vybraných uživatelů, což je možné díky široké nabídce kvalitních kurzů na trhu.

Obrázek 11: 2017 Magic Quadrant for Business Intelligence and Analytics Platforms

Zdroj: (21)

2.7 SWOT Analýza

V této části bude sestavena SWOT analýza popisující silné a slabé stránky podniku a také příležitosti a hrozby působící na podnik z vnějšího prostředí. Tato analýza využívá výstupy veškerých doposud vypracovaných analýz.

2.7.1 Silné stránky

Know-how – Vzhledem k tomu, že společnost působí na trhu již od roku 2002 disponuje sadou ověřených postupů a cenných informací v oblasti recruitmentu.

Zavedené jméno firmy – S délkou působení na trhu souvisí také silné jméno společnosti a velké množství dlouhodobých partnerů.

Kvalitní informační systém – Společnost disponuje moderním IS, který usnadňuje většinu firemních procesů.

Databáze kandidátů – Společnost rovněž disponuje obsáhlou, a především kvalitní databází uchazečů o zaměstnání, což je její největší předností.

Rozsáhlá síť poboček a divizí – Provozuje sedm poboček v různých krajích a sedm odborných divizí s více než osmdesáti konzultanty.

2.7.2 Slabé stránky

Nízká automatizace klíčových procesů – Ačkoliv se při zavedení BI řešení automatizovala řada náročných úloh jako je například reporting, stále zůstává několik procesů, jichž se automatizace zatím nedotkla. Například právě hodnocení kandidáta.

Nezajištěná pokročilá analýza dat – I přes existenci sady reportů firma dosud nevyužívá žádné pokročilé analytické nástroje. Vzhledem k zavedenému datovému skladu a obrovské databázi kandidátů se tak dobrovolně připravuje o nové možnosti.

Velké vytížení konzultanta – konzultant musí kromě své běžné funkce zastat také pozici researchera – musí vyhledávat a mezi sebou porovnávat vhodné uchazeče.

Nízký důraz na cílený marketing – Firma aktivně využívá nástroje přímého marketingu, nicméně neklade velký důraz na zacílení. To tak snadno může být považováno za SPAM.

2.7.3 Příležitosti

Pokročilá analýza dat – V případě nasazení pokročilých analytických nástrojů, může společnost ve svých datech objevit nové informace a získat tak další znalosti pro podporou rozhodování.

Rostoucí poptávka po zaměstnancích – S příchodem nových, či růstem stávajících firem se otevírají nové pracovní pozice, které může agentura obsadit.

Nové metody a technologie – Nástup nových technologií či zavedení moderních metod v oblasti recruitmentu může pomoci agentuře zefektivnit její procesy a snížit náklady.

Snížení konkurence – S případným odchodem konkurenta společnost může vyplnit jeho nepokrytou poptávku.

2.7.4 Hrozby

Silná konkurence – Vzhledem k vysokému počtu konkurenčních firem, není vyloučena ztráta klíčového partnera, který by nebyl zcela spokojen se službami či cenou služeb společnosti.

Legislativní změny (příchod GDRP) – Změna zákonů v České republice může způsobit personální agentuře různorodé problémy, jako jsou například: nutná úprava stávajících procesů, zvýšení nákladů na ochranu dat, omezení zpracování osobních údajů, vysoké sankce a podobně. Příkladem změny může být právě příchod legislativy GDRP.

Zvyšování nákladů na provoz – S případným zvyšováním cen tarifů mobilních operátorů či dodavatelů cloudových služeb, může v budoucnu dojít k celkovému zvýšení nákladů na provoz.

Únik citlivých dat – Pro personální agenturu je únik citlivých dat vážnou hrozbou, vzhledem k velkému množství osobních údajů, které zpracovává a uchovává. To však

hrozí nejen z okolí v podobě případných útoků, ale také zevnitř společnosti v podobě vynášení informací či zanedbání bezpečnosti.

2.7.5 Vyhodnocení

Tabulka 3: Vyhodnocení SWOT analýzy (Vlastní zpracování)

SWOT analýza	
Silné stránky	**Příležitosti**
Know-how Zavedené jméno firmy Kvalitní IS Databáze kandidátů Rozsáhlá síť poboček a divizí	Pokročilá analýza dat Rostoucí poptávka po zaměstnancích Nové metody a technologie Snížení konkurence
Slabé stránky	**Hrozby**
Nízká automatizace klíčových procesů Nezajištěná pokročilá analýza dat Velké vytížení konzultanta Nízký důraz na cílený marketing	Silná konkurence Legislativní změny (Příchod GDRP) Zvyšování nákladů na provoz Únik citlivých dat

Na základě vypracované analýzy současného stavu jsou odhaleny slabé stránky společnosti, které se týkají především problému nízké automatizace klíčových procesů, zvláště pak procesu výběru vhodného kandidáta. Je tedy možné usoudit, že současný stav je nevyhovující. Vzhledem k tomu, že společnost disponuje kvalitní a velmi obsáhlou databází kandidátů, jsem potvrdil závěr, že nasazení pokročilé analýzy dat pomocí data miningu a přispění k automatizaci procesu výběru vhodného kandidáta představuje vhodnou změnu. Tato změna by měla vést ke zvýšení konkurenceschopnosti, optimalizovat dosud zavedené procesy a podpořit manažerské rozhodování.

2.8 Analýza rizik

Jako u každého projektu, i zde existuje řada možných hrozeb související s jeho implementací, které je vhodné mít pod kontrolou. Pro analýzu rizik využiji metodu RIPRAN.

2.8.1 Identifikace rizik

Nekvalitní data

Pokud očekáváme kvalitní výsledky analýzy, je potřeba vždy analyzovat kvalitní data a v tomto případě to platí více než kdy jindy. Mezi nejčastější problémy patří například různě kódovaná data, neatomické hodnoty, chybějící hodnoty, duplicitní záznamy, konvence pojmů a objektů, kdy data z různých zdrojů popisující stejný jev mají odlišné názvy entit apod. Data, která budou vstupovat do modelu musí být řádně očištěna.

Organizační změny ve firmě

Jedním z nejdůležitějších kritérií pro úspěšné dokončení projektu je právě lidský faktor. Nicméně v průběhu projektu může dojít ke změnám v organizační struktuře, která může ovlivnit projektový tým. Účast na projektu by tedy měla být pro jednotlivé členy zavázána smlouvou či jinou dohodou.

Nekvalitní počáteční analýza

Pokud provedená analýza obsahuje chybné výstupy, nebo neobjevila klíčové vlastnosti je možné, že dojde ke špatně definovanému cíli, či jeho nenaplnění. Zároveň je hrozba nepokrytí klíčových procesů.

Neochota zaměstnanců pracovat s výsledky

Zaměstnanci nejsou z pravidla příliš ochotni měnit své pracovní návyky a zaběhnuté procesy. Proto je tyto zaměstnance nutné správně motivovat a ujistit, že úprava jejich stávajících procesů má za úkol podpořit a zefektivnit jejich práci.

Změna priorit ve firmě

V případě změny priorit ve firmě například k jinému projektu, je možné, že by na tento projekt nezbyly patřičné prostředky, a tak by se mohl pozastavit či dokonce zrušit.

Nedostatečné školení uživatelů

Uživatel, který nerozumí výsledkům projektu, nebo neví, jak s nimi správně naložit je další hrozbou, která se může objevit ihned po spuštění do ostrého provozu. Je tedy velmi důležité věnovat patřičnou pozornost školení všech budoucích uživatelů a ujistit se, že jsou si vědomi nové změny a rozumí, jak ji využít při své práci.

Únik a zneužití citlivých dat

Jelikož v tomto projektu bude nakládáno s velmi citlivými daty, jako jsou osobní údaje kandidátů je nutné zajistit, aby v žádném případě nedošlo k jejich úniku či zneužití. V nejhorším možném případě by takový únik mohl znamenat pokuty až do výše několika milionů korun.

Nedostatečná rychlost systému

Zpracování data miningových modelů lze považovat za výpočetně velmi náročnou operaci. Vzhledem k tomu, že konzultanti aktivně využívají informační systém během celé pracovní doby, je nutné, aby jeho rychlost nebyla nijak omezena.

Špatně definovaná kritéria pro scoringový algoritmus

Špatný návrh kritérií scoringového algoritmu může vést, k chybnému hodnocení kandidátů, kdy ve skutečnosti méně kvalitní kandidáti budou hodnoceni lépe, než lepší a naopak.

Neznalost spolehlivosti výsledku predikce

Pokud uživatel nebude vědět, s jakou pravděpodobností nabývá konkrétní záznam výsledku predikce, nemůže se na něj zcela spolehnout.

Změna legislativy

Změna legislativy s sebou může přinést nutnost nových školení, které budou muset zaměstnanci absolvovat. Z tohoto důvodu může nastat odklonění od plánovaných termínů jednotlivých činností či zvýšení nákladů.

Nenaplnění očekávání

V případě nízké komunikace mezi projektovým týmem, sponzorem změny a dalšími uživateli, může dojít k nejasnosti v představě, požadavcích a tím následně k nenaplnění očekávání na výstup projektu.

Zpoždění termínu ukončení projektu

Při zpoždění termínu dokončení projektu může dojít ke zvýšení vynaložených nákladů na projekt.

Všechna výše identifikovaná rizika je nutné po celou dobu projektu sledovat, a pokud nastanou, patřičně na ně reagovat. V další kapitole budou tato rizika posouzena z hlediska pravděpodobnosti a míry dopadu.

2.8.2 Kvantifikace rizik

Při hodnocení rizik posuzuji dopad rizika a pravděpodobnost jeho vzniku, při čemž součinem těchto hodnot získám celkový význam rizika. Podle této celkové hodnoty se dále rozhoduji, zda riziko přijmu či nikoliv. Pokud ne, je nutné navrhnout vhodná opatření, které sníží jeho hodnotu na přijatelnou úroveň, nebo ho zcela odstraní.

Pro ohodnocení rizik si vytvořím následující stupnice, ve které definuji intervaly a hodnoty pravděpodobnosti vzniku a velikosti dopadu rizika.

Tabulka 4: Klasifikační stupnice pravděpodobností vzniku (Vlastní zpracování)

Pravděpodobnost	Interval
(NP) Nízká	od 0 % - do 25 %
(SP) Střední	od 26 % - do 65 %
(VP) Vysoká	od 66 % - do 100 %

Tabulka 5: Klasifikační stupnice dopadu (Vlastní zpracování)

Dopad	Hodnota dopadu
Zanedbatelný	1
Mírný	2
Velký	3
Závažný	4

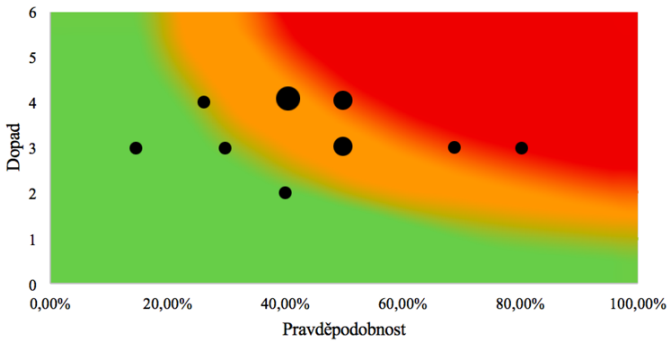
V následující tabulce je dále definována stupnice pro celkový význam rizika, která rozlišuje význam rizika. Běžná rizika budou přijata. Na závažná a kritická rizika budou navrženy vhodná opatření. Je nutné podotknout, že v tomto kontextu kritické riziko neznamená ohrožení, jehož potenciální ztráty mohou vyústit v bankrot společnosti, ale kriticky ovlivňují cíle projektu.

Tabulka 6: Klasifikační stupnice hodnoty rizika (Vlastní zpracování)

Hodnota rizika	Hodnota rizika
Běžné	< 1
Závažné	1 – 1,99
Kritické	>= 2

Tabulka 7: Hodnocení identifikovaných rizik (Vlastní zpracování)

Č.	Hrozba	Scénář	P	D	H
1	Nekvalitní data	Špatné výsledky, které ovlivní rozhodování	50	4	2
2	Organizační změny ve firmě	Zdržení dokončení projektu	15	3	0.45
3	Nekvalitní počáteční analýza	Nenaplnění potřeb organizace	40	4	1.6
4	Neochota zaměstnanců pracovat s výsledky	Neuplatnění výsledků projektů	80	3	2.4
5	Změna priorit ve firmě	Pozastavení či zrušení projektu	40	2	0.8
6	Nedostatečné školení uživatelů	Neefektivní práce s výsledky	70	3	2.1
7	Únik a zneužití citlivých dat	Pokuta a poškození jména společnosti	25	4	1
8	Nedostatečná rychlost systému	Nestabilita systému, omezení práce uživatelů	50	4	2
9	Špatně definovaná kritéria pro scoringový algoritmus	Špatné výsledky, které ovlivní rozhodování	40	4	1.6
10	Neznalost spolehlivosti výsledku predikce	Značná nejistota v rozhodování	50	3	1.5
11	Změna legislativy	Zvýšení nákladů na projekt, případné zpoždění	30	3	0.9
12	Nenaplnění očekávání	Vynaložení nákladů bez naplnění představy sponzora	40	4	1.6
13	Zpoždění termínu ukončení projektu	Zvýšení předpokládaných nákladů na projekt	50	3	1.5



Obrázek 12: Mapa rizik

Zdroj: Vlastní zpracování

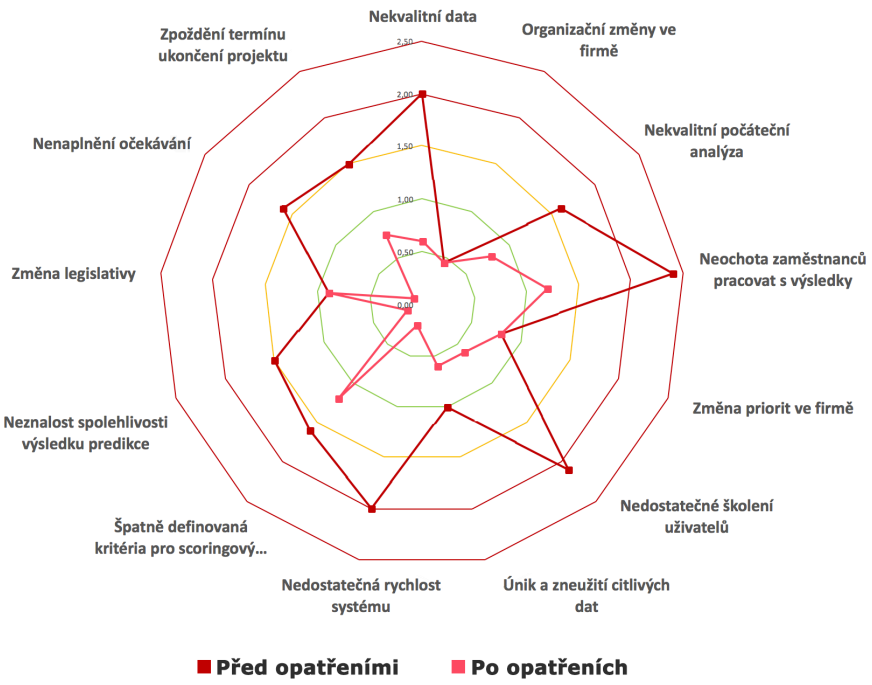
2.8.3 Metody snižování rizik

Z hodnocení rizik vyplývá, že tři rizika jsou běžná a není nutné se jimi dále zabývat, šest rizik je závažných a čtyři kritická.

Tabulka 8: Návrhy na opatření (Vlastní zpracování)

Č.	Opatření	Náklady / Zodpovědná osoba	Realizace	P*	D*	H*
1	Podrobná analýza vstupních dat a důraz na ETL fázi	10.000 Kč / Vývojář	Během	15	4	0.6
3	Vykonání interního, případně externího auditu	15.000 Kč / Projektový manažer	Před	20	4	0.8
4	Motivační meeting s důrazem na informování zaměstnanců o přínosech výstupů projektu	5.000 Kč / Projektový manažer	Během	40	3	1.2
6	Vícefázový workshop s důrazem na aktivitu uživatelů	5.000 Kč / Projektový manažer	Během	20	3	0.6
7	Šifrování přenášených materiálů, poučení zaměstnanců a smluvní ošetření	2.500 Kč / Projektový manažer	Před	20	3	0.6
8	Zpracování veškerých modelů mimo pracovní dobu	0 / Vývojář	Po	5	4	0.2
9	Důraz na expertní odhady kvalifikovaných konzultantů	0 / Projektový manažer	Před	30	4	1.2
10	Doplnění výsledku predikce, o hodnotu její pravděpodobnosti	0 / Vývojář	Během	5	3	0.15
12	Analýza požadavků, představení možností data miningu	2.500 Kč / Projektový manažer	Před	10	1	0.1
13	Důkladné zpracování síťové analýzy	3.500 Kč / Projektový manažer	Před	25	3	0.75

Po uplatnění opatření na snížení rizik, se většina z nich přesunula do kategorie běžných rizik. Dvě rizika jsou však stále klasifikována jako závažné, a proto je nutné, je dále pečlivě sledovat.



Obrázek 13: Pavučinový graf rizik před a po opatřeních

Zdroj: Vlastní zpracování

2.8.4 Zhodnocení rizikovosti projektu

Z provedené analýzy vyplývá, že se jedná o středně rizikovou změnu. Objevil jsem tři běžná, šest závažných a čtyři kritická rizika, nicméně po uplatnění navržených opatření byla hodnota těchto rizik snížena. Nyní tato změna představuje pouze dvě závažná rizika a lze tedy projekt doporučit k realizaci.

3 VLASTNÍ NÁVRHY ŘEŠENÍ

Na základě znalosti současného stavu, analýzy problému a definovaných požadavků jsem nyní schopen přistoupit k samotnému návrhu řešení. Postup mé práce se bude opírat o metodiku CRISP-DM, jejíž první dva kroky – porozumění podnikání a porozumění datům byly naplněny již v kapitole Analýza současného stavu.

3.1 Návrh procesu hodnocení kandidáta

Jak již bylo uvedeno v části analýza problému, konzultanti nejsou schopni bez bližšího průzkumu rychle definovat, jakou hodnotu pro společnost představuje konkrétní kandidát, respektive nejsou schopni mezi sebou kandidáty efektivně porovnat. I když v jeho profilu mohou uvést své subjektivní hodnocení ve formě textové poznámky, není možné tyto texty zohlednit v pořadí či třídění hledaných výsledků IS. Navíc se tyto hodnocení zpravidla udělují až po vynaloženém času stráveným nad profilem.

Z těchto důvodů jsem společnosti doporučil automatizované hodnocení kandidátů na základě definovaných parametrů. Hodnota kandidáta pro společnost bude počítána součinem váhy vybraných atributů a bodů jejich definovaných hodnot. Tyto váhy a body byly po konzultaci zajištěny společností pomocí expertních odhadů na základě zkušeností a analýzy historických dat. Se společností jsme se zároveň dohodli, že pro zkušební provoz bude toto řešení nasazeno na vrstvu datového skladu, aby prozatím žádným způsobem neomezovalo fungování IS. Vzhledem k tomu, že samotný algoritmus je velmi nenáročný je vhodné, aby fungoval zcela samostatně mimo oblast dolování dat. Tato hodnota tedy bude nezávislá na výstupu data miningu, nicméně bude dále uvažována jako jeho vstupní atribut.

Je zřejmé, že jednotlivé váhy a hodnoty atributů budou bodovány odlišně alespoň v závislosti na konkrétní divizi. Například při hledání ideálního právníka, nebudeme klást význam tomu, že má praxi jako programátor a podobně. Každá divize tedy bude muset mít nejméně jednu sadu takových tabulek, kde rozdělí váhy a body odpovídající jednotlivým oborům či pozicím.

V této práci bude pro ilustraci uvažována pouze divize IT, tedy pracovní trh s IT specialisty. Mezi hlavní důvody výběru právě této divize patří to, že většina nových projektů je implementována v první řadě zde a také sídlí v Brně, což umožňuje pravidelné osobní konzultace.

3.1.1 Postup

Nejprve je nutné zvolit správná kritéria, kterým je třeba definovat určitou váhu. Tato váha jistým způsobem vyjadřuje prioritu jednotlivých kritérií. Vzhledem k tomu, že hodnotící proces má fungovat automaticky bez dalších nezbytných zásahů uživatele, je nutné, aby tyto data byly přímo dostupné z firemní databáze, či dosažitelné jiným způsobem s ohledem na legislativu.

Následně budou ohodnoceny jednotlivé hodnoty, které tyto atributy nabývají. Vzhledem k tomu, že prakticky všechny podstatné hodnoty v databázi jsou tvořeny číselníky, není potíž s nimi dále pracovat. Spojité hodnoty jako je věk či požadovaná mzda budou dále diskretizovány pro vytvoření ideálních rozsahů.

Nakonec navrhnu vhodný algoritmus, který bude zpracován jako databázový T-SQL skript, kdy jako primární zdroj dat poslouží přímo datový sklad.

3.1.2 Výběr kritérií

Na základě zkušeností firemních expertů byly pečlivě vybrány kritéria, kterým byly přiřazeny určité váhy. Veškeré tyto atributy jsou dostupné jak v datovém skladu, tak v relační databázi, se kterou přímo pracuje informační systém.

Jak je možné vidět v níže uvedené tabulce, je zřejmé že například znalost anglického jazyka je pro firmu, respektive pro divizi IT velmi klíčovou schopností, a to dokonce více než samotné dosažené vzdělání či délka zaměstnání. Stejnou váhu má dále typ pracovní zkušenosti, respektive dosavadní pracovní pozice kandidáta. Minimální požadovaná mzda dále určuje potenciální výnos, protože odměna pro společnost bývá zpravidla přímo závislá na nástupním platu kandidáta. Nepatrný vliv má také preferovaný typ pracovního

poměru, kdy je pro společnost snazší například uplatnit kandidáta na HPP či zkrácený úvazek než pouze práci na živnostenský list.

Tabulka 9: Váhy jednotlivých kritérií (Vlastní zpracování)

Atribut	Váha
Znalost Jazyka - EN	8
Praxe Pozice	8
Obor vyšší odborné a vysoké školy	7
Minimální požadovaná mzda	7
Znalost Jazyka - DE	7
Znalost Jazyka - RU	7
Praxe obor	7
Maximální dosažené vzdělání	6
Obor střední školy	5
Znalost Jazyka - FR	5
Věk	5
Délka zaměstnání	5
Znalost Jazyka - IT	4
Znalost Jazyka - ES	4
Znalost Jazyka - PL	4
Délka posledního/současného zaměstnání	4
Preferovaný pracovní poměr	3

3.1.3 Bodování jednotlivých hodnot

Nyní stejným principem jsou ohodnoceny jednotlivé hodnoty, které tyto atributy nabývají. Většina hodnot spojitých atributů jako je například minimální požadovaná mzda či věk kandidáta jsou převedeny na potřebné rozsahy. V jiném případě jsou ohodnoceny všechny hodnoty, které atribut může nabývat.

Vzhledem k počtu jednotlivých hodnot, které vybrané atributy mohou nabývat, je níže na uvedené tabulce představena pro ilustraci pouze jejich velmi malá část. Širší ohodnocený seznam je uveden v příloze číslo 1 této práce.

Tabulka 10: Body jednotlivých hodnot (Vlastní zpracování)

Hodnota	Body
Maximální dosažené vzdělání	
Vysoká škola	100
Vyšší odborná škola	65
Střední škola	50
Vyučen	15
Minimální požadovaná mzda	
40.000 - 50.000	100
30.000 - 40.000	90
50.000 - 60.000	80
> 60.000	70
20.000 - 30.000	40
< 20.000	10
Znalost Jazyka	
C2	100
MA	85
C1	80
B2	60
B1	40
A2	15
A1	5
Věk	
30 - 39	100
20 - 29	90
40 - 50	80
< 20	40
> 50	40
Délka praxe v letech	
3-5	100
1-3	85
5-7	85
7 a více	85
0-1	65
zkušební doba	20
Praxe pozice	
Technická podpora / Administrátor/Správce OS a sítí	100
Programátor	100
ICT manager / ředitel IT/IS	85

3.2 Příprava dat

Aby bylo možné implementovat toto automatické hodnocení kandidáta a navrhnout vhodné data miningové modely, je třeba nejprve zajistit, aby se v datovém skladu nacházely všechna patřičná data. Jak již je vysvětleno v teoretické části, data, která vstupují do datového skladu, musí nejprve projít takzvanou fází ETL, pro zaručení patřičné kvality. Na tuto fázi využiji především nástroj SQL Server Data Tools a TSQL skripty.

3.2.1 Implementace hodnocení v datovém skladu

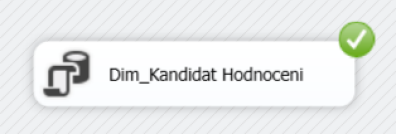
Samotná problematika hodnocení byla uvedena v předchozí kapitole. Jedno z možných řešení je nasazení bodových ohodnocení přímo do produkční databáze a vytvořit tak novou sadu tabulek s váhami a body s řádně zajištěnou referenční integritou. Vzhledem k situaci, kdy je počet jednotlivých divizí v podstatě konstantní a změna může nastat jen ve výjimečném případě, lze definovat váhy ve formě atributů. I tak bude tabulku kdykoliv možné snadno rozšířit o další atribut, představující hodnotu konkrétní divize či jinou specifickou oblast. Tyto hodnoty by byly dále zpracovány pomocí skriptu, který by aktualizoval výstup hodnocení.

Nicméně tento způsob není prozatím zcela možný. Jedná se o zásah do struktury produkční databáze, který si společnost prozatím nepřeje provést. V současnosti to není však v podstatě ani nutné, jelikož nasazujeme zkušební provoz pouze pro IT divizi, je elegantnější řešení využít pouze jednoduchý databázový skript, který bude mít stejné vlastnosti, jen neovlivní strukturu databáze.

Rozšířím tedy současný SSIS balík *MainIncrement* v projektu Integration Services, který je zodpovědný za integraci části dat z produkční databáze do datového skladu. Vytvořím zde nový Execute SQL Task *Dim_kandidat Hodnoceni* představující T-SQL skript, který je uveden níže.

Vzhledem k celkovému rozsahu skriptu, zde bude pouze pro ilustraci uvedena pouze jeho malá část a celý zdrojový kód naleznete v příloze číslo 2 tohoto dokumentu.

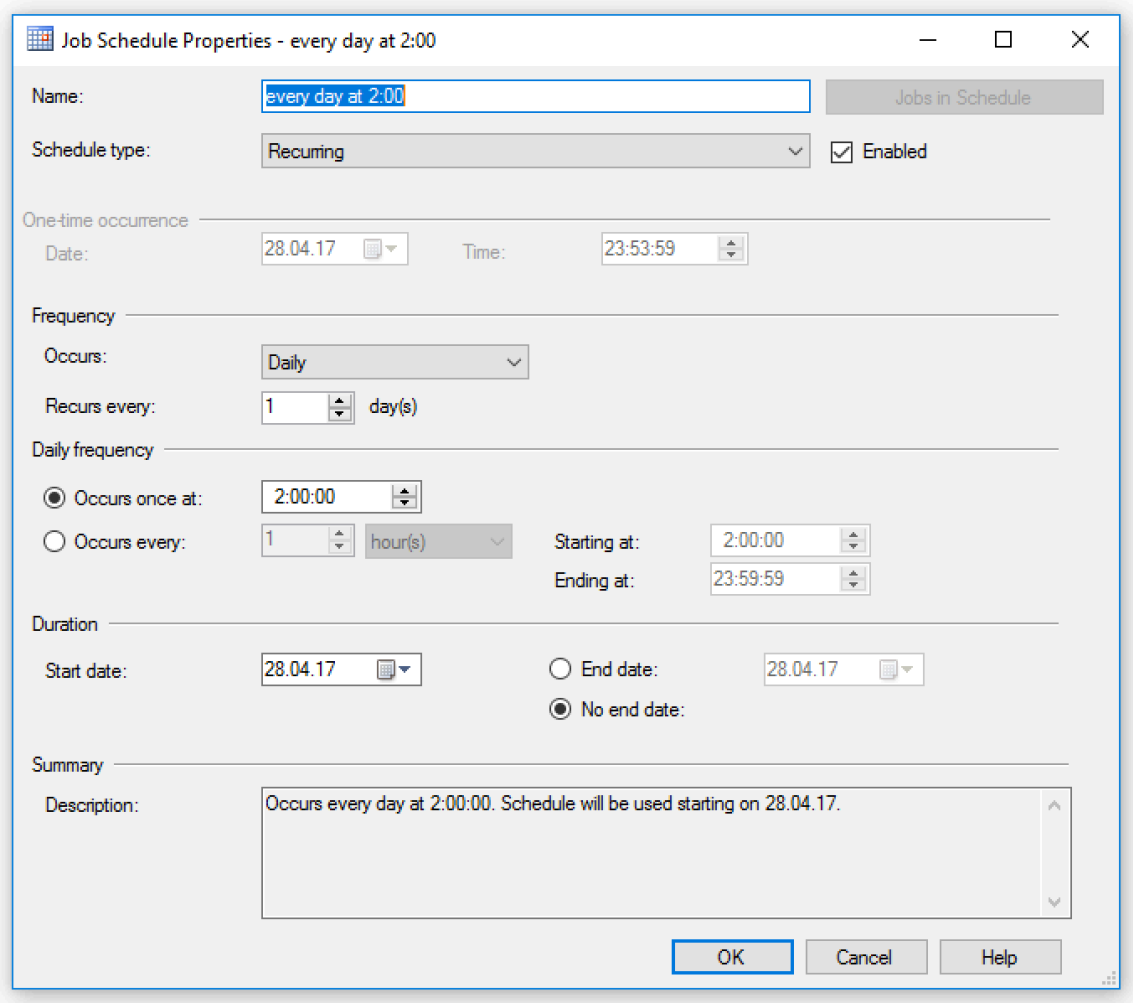
```
ALTER PROCEDURE hodnoceni AS
Declare
	--vahy
	@vek tinyint = 5,
	@PraxeDobaCela tinyint = 5,
	@PraxeObor tinyint = 7

select ID_Kandidat, sum(hodnoceni) as hodnoceni into #temp from (
	select
	--Vek
		case
		  when year(getdate())-RokNarozeni between 30 and 39 then 100 * @vek
		  when year(getdate())-RokNarozeni between 20 and 29 then 90 * @vek
		  when year(getdate())-RokNarozeni between 40 and 50 then 80 * @vek
		  when year(getdate())-RokNarozeni not between 20 and 50 then 40 * @vek
		  else 40 * @vek
		end as 'hodnoceni', ID_Kandidat
	from Dim_Kandidat dk
union all
	select
		--Delka praxe v letech
		sum(
			case
			  when dkp.doba = '0-1' then 65 * @praxeDobaCela
			  when dkp.doba = '1-3' then 85 * @praxeDobaCela
			  when dkp.doba = '3-5' then 100 * @praxeDobaCela
			  when dkp.doba = '5-7' then 85 * @praxeDobaCela
			  when dkp.doba = '7 a více' then 85 * @praxeDobaCela
			  when dkp.doba = 'zkušební doba' then 20 * @praxeDobaCela
			  else 0
			end
		) +
		--Praxe obor
		sum(
			case
			  when left(obor, 2) = 'IS' then 100 * @PraxeObor
			  when dkp.doba = 'Elektrotechnika' then 80 * @PraxeObor
			  -- etc.
			  else 0
			end
		) as 'hodnoceni', ID_Kandidat
	from Dim_KandidatPraxe dkp
	group by dkp.ID_Kandidat
)
x group by ID_Kandidat

MERGE INTO dim_Kandidat dk
   USING #temp t
      ON t.id_kandidat = dk.id_kandidat
   WHEN MATCHED THEN
      UPDATE
         SET hodnoceni = t.hodnoceni;
GO
```



Obrázek 14: Execute SQL Task Dim_kandidat Hodnoceni

Zdroj: Vlastní zpracování

Tímto krokem jsou zajištěny fáze extrakce a transformace. Fáze loading bude probíhat současně s ostatními balíčky v pravidelném intervalu, který zajišťuje SQL Server Agent. Toto nahrávání probíhá každý den ve 2 hodiny ráno proto, aby přenos dat jakkoliv neomezil databázový výkon v době, kdy ho využívají uživatelé systému.



Obrázek 15: Integrace - SQL Server Agent

Zdroj: Vlastní zpracování

3.2.2 Vytvoření pohledů pro Data Mining

Nyní pro aplikaci data miningu je nutné vytvořit dva pohledy. První z nich poslouží pro trénování modelu a bude obsahovat všechny kandidáty, kteří již prošli výběrovým řízením a byly tak přijati či zamítnuti. Druhý pohled slouží k testování modelu, respektive k samotné predikci a obsahuje již všechny kandidáty v databázi, které chceme zpracovat. Ukázka prvního pohledu je zobrazena níže, druhý pohled je již jednodušší modifikací toho prvního.

```
create view dm_obsazeni as
select
       dk.ID_Kandidat, Pohlavi, Aktivni, Email, Registrovan, Mzda, dl.Okres as
       'bydliste', Zdroj, MaxDosazeneVzdelani, en, fr, it, de, ru, es, pl, Hodnoceni,
       dkp.doba, dkp.Specializace, dkp.Pozice, dkp.Obor, dkv.Zamereni, dll.Okres,
       fp.PoptavkaStatus,
       case
              when year(getdate())-RokNarozeni between 15 and 20 then '15-20'
              when year(getdate())-RokNarozeni between 20 and 25 then '20-25'
              when year(getdate())-RokNarozeni between 25 and 30 then '25-30'
              when year(getdate())-RokNarozeni between 30 and 35 then '30-35'
              when year(getdate())-RokNarozeni between 35 and 45 then '35-45'
              when year(getdate())-RokNarozeni between 45 and 55 then '45-55'
              when year(getdate())-RokNarozeni > 55 then '> 55'
          else '< 15'
       end as 'vek',
       RokNarozeni
from Dim_Kandidat dk
       left join Fact_Produkce fp
              on fp.ID_Kandidat = dk.ID_Kandidat
       left join Dim_KandidatPraxe dkp
              on dkp.ID_Kandidat = dk.ID_Kandidat
       left join Dim_Lokalita dl
              on dl.ID_Okres = dk.ID_Okres
       left join Dim_KandidatVzdelani dkv
              on dkv.ID_Kandidat = dk.ID_Kandidat
       left join Dim_KandidatOkres dko
              on dko.ID_Kandidat = dk.ID_Kandidat
       left join Dim_Lokalita dll
              on dll.ID_Okres = dko.ID_Okres
where fp.scdAktualni = 1
```

Tímto krokem jsou již připraveny veškerá potřebná data a mohu tak přistoupit k fázi modelování.

3.3 Hodnocení výsledků automatického scoringu

Pro ověření správných výsledků algoritmu, jsem předložil třem konzultantům z IT divize deset náhodně vybraných kandidátů. Tyto kandidáty se pokusili seřadit dle zajímavosti pro společnost, na základě veškerých dostupných informací, se kterými běžně pracují. Číslo deset představuje nejvyšší prioritu, tedy nejzajímavějšího kandidáta, kterého by upřednostnili před jinými.

Tabulka 11: Porovnání výsledků algoritmu (Vlastní zpracování)

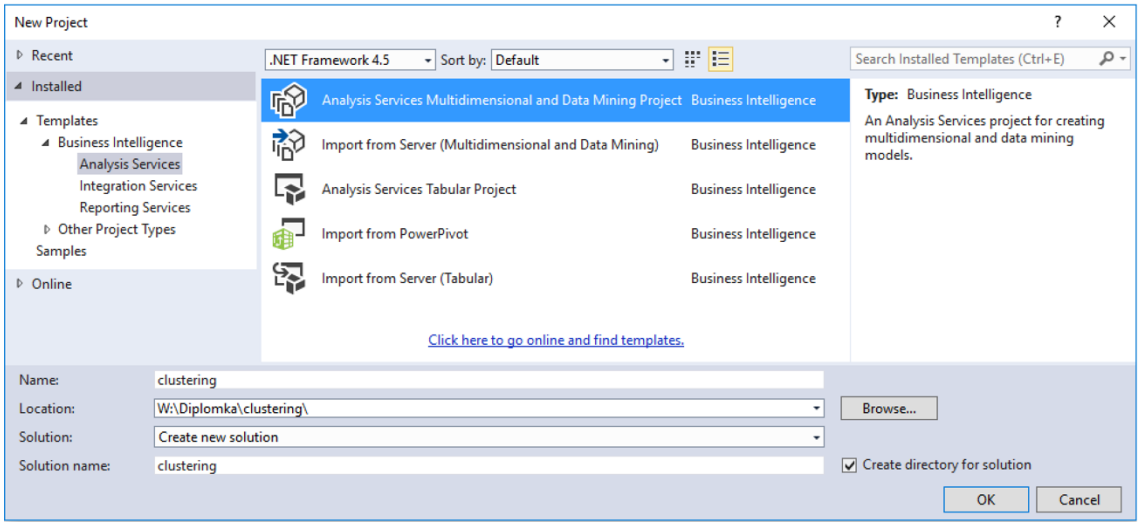
Kandidát	Automatické hodnocení		Hodnocení konzultanta 1	Hodnocení konzultanta 2	Hodnocení konzultanta 3
1	18630	10	10	**9**	10
2	16945	9	9	**10**	9
3	10555	8	**7**	8	8
4	9245	7	**8**	7	7
5	6765	6	6	6	5
6	6765	5	5	5	6
7	5955	4	4	4	**3**
8	4495	3	3	3	**4**
9	3375	2	2	2	2
10	1055	1	1	1	1

Po srovnání subjektivního hodnocení a hodnocení automatického scoringu na tomto malém vzorku, je shoda u každého konzultanta v 9 z 10 případů. Na tomto příkladu však není možné určit, kdy se zmýlil scoringový algoritmus, nebo konzultant. Nicméně se jedná o velmi uspokojivý výsledek, který demonstruje, že lze tímto algoritmem značně podpořit činnost konzultantů.

Je nutné podotknout, že konzultanti jsou vedeni k tomu, aby postupovali při hodnocení kandidátů obdobně. Kritéria jsou v tomto případě know-how společnosti a konzultanti jsou poučeni jak je brát na vědomí. Nicméně, zda jsou tyto kritéria navrženy opravdu správně, pomůže ověřit následující kapitola zabývající se samotným dolováním dat.

3.4 Modelování

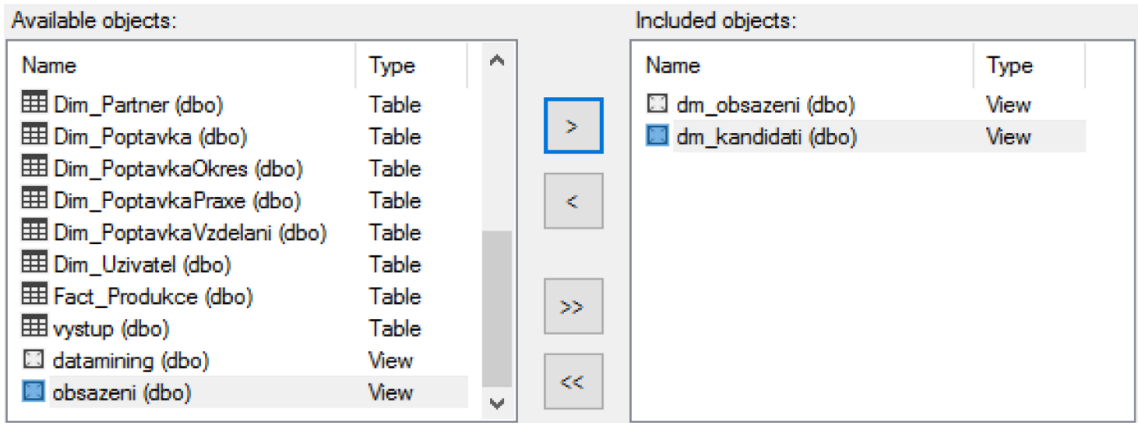
Tento oddíl se zabývá výběrem nejvhodnějšího modelu pro predikci nástupů a segmentaci kandidátů. Nejprve však musím vytvořit nový Analysis Services Multidimensional and Data Mining project v prostředí aplikace Server Data Tools. V tomto projektu vyberu nový Data Source, kde jako zdroj dat poslouží datový sklad společnosti.



Obrázek 16: Prostředí SQL Server Data Tools

Zdroj: Vlastní zpracování

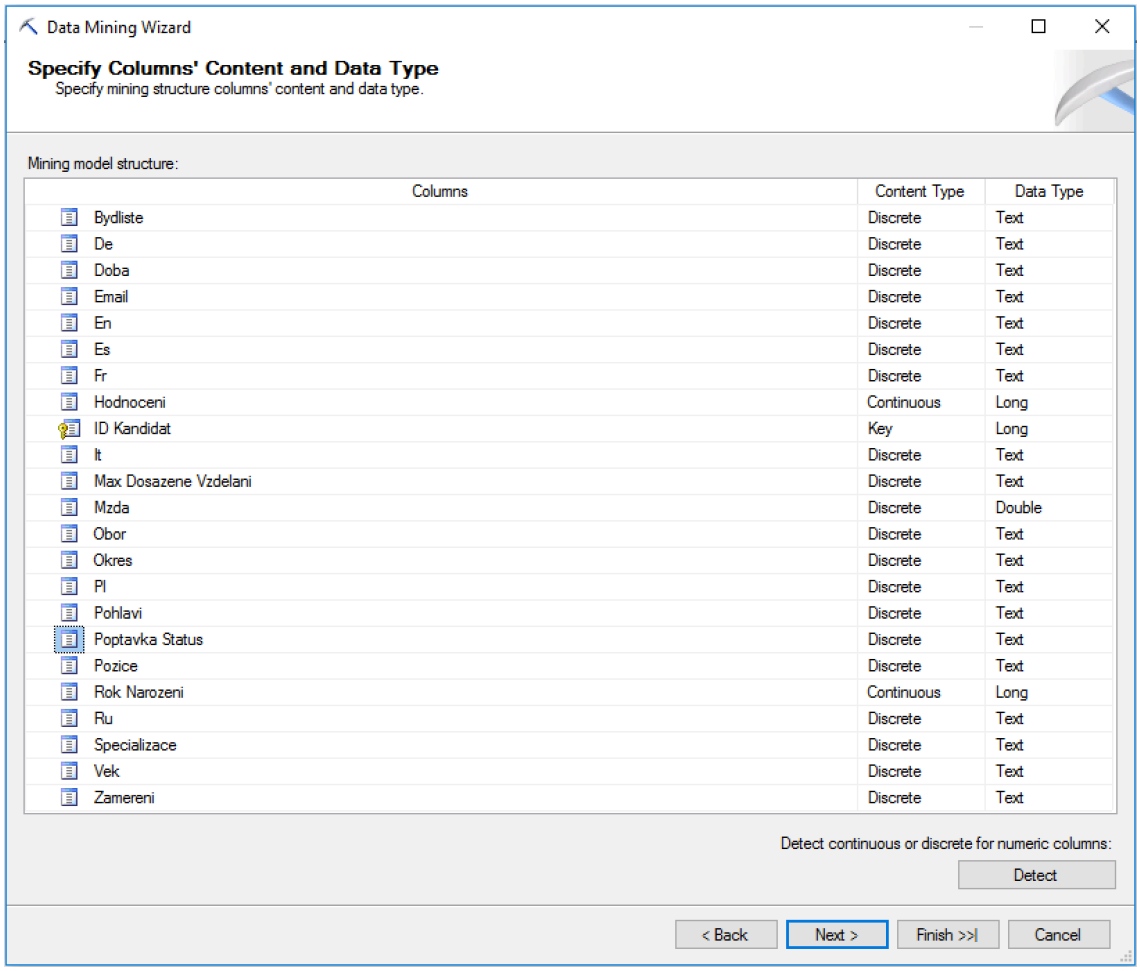
Následně definuji Data Source View, kde vyberu dříve vytvořené pohledy na kandidáty obsahující potřebné atributy. Pohled *dm_obsazeni* je určen k učení modelu a *dm_kandidati* k jeho vyhodnocení, respektive k predikci.



Obrázek 17: Data Source View

Zdroj: Vlastní zpracování

Dále vytvořím novou Mining Structure, kde vyberu potřebné input atributy. Jako klíč poslouží *ID Kandidat* a predikovaná hodnota bude *Poptavka Status*, která nabývá hodnoty *zamítnut/nástup*. Automatická detekce se v tomto případě postará o správné nastavení diskrétních a spojitých atributů. Nakonec rozdělím data na dvě množiny – 70 % dat poslouží pro trénování modelu a zbylých 30 % k jeho testování.



Obrázek 18: Nastavení diskrétních a spojitých atributů

Zdroj: Vlastní zpracování

3.4.1 Výběr ideální verze modelu a zhodnocení vlivu scoringu

Před samotným výběrem ideálního algoritmu se chci pozastavit nad samotnou důležitostí naprogramovaného scoringu a jeho vlivu na výsledky predikce. To v podstatě znamená, že musím vytvořit a porovnat tři různé modely s odlišnými vstupy, abych zjistil, zda scoring pomůže nejen konzultantům při preferenci kandidátů, ale také zda zlepší predikční model.

První model *All Attributes* tedy zahrnuje veškeré atributy v dříve vytvořeném pohledu *dm_obsazeni*. Do dalšího modelu *Just Scoring* vstupuje pouze hodnota scoringu - *Hodnoceni*. A do posledního modelu *Without Scoring* vstupují všechny vybrané atributy právě kromě hodnoty scoringu. Všechny tři popsané modely budou porovnány s využitím algoritmu rozhodovacích stromů, který se ukázal jako nejpřesnější.

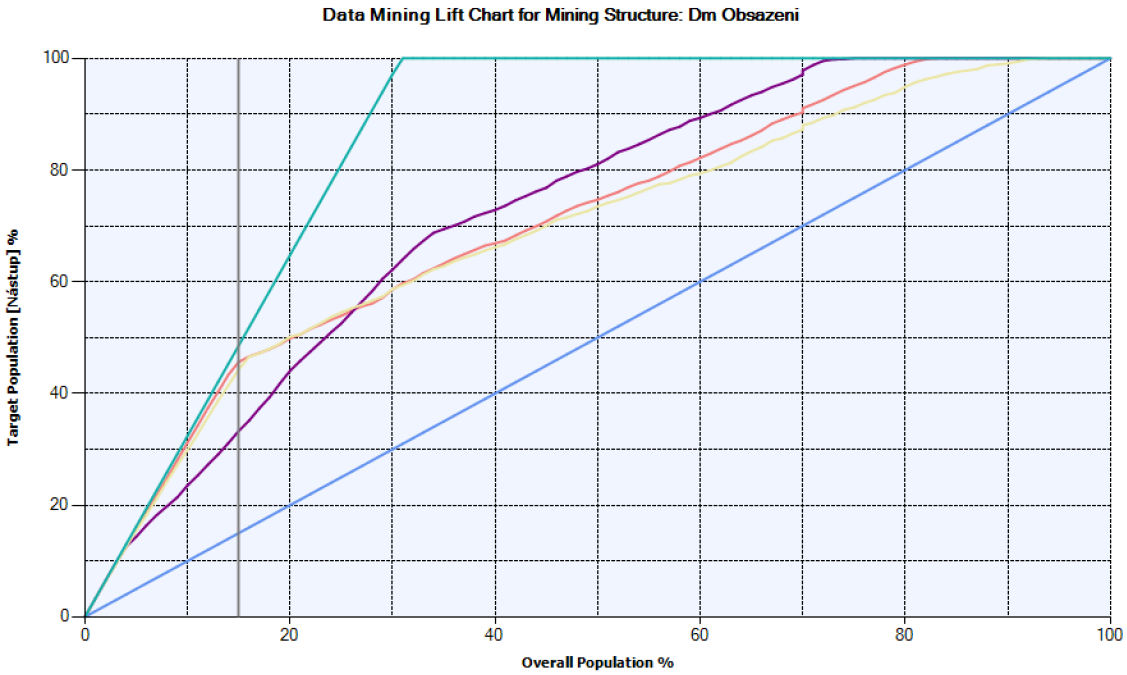
Structure	All attributes	Just Scoring	Without scoring
	Microsoft_Decision_Trees	Microsoft_Decision_Trees	Microsoft_Decision_Trees
Bydliste	Input	Ignore	Input
De	Input	Ignore	Input
Doba	Input	Ignore	Input
Email	Ignore	Ignore	Ignore
En	Input	Ignore	Input
Es	Input	Ignore	Input
Fr	Input	Ignore	Input
Hodnoceni	Input	Input	Ignore
ID Kandidat	Key	Key	Key
It	Input	Ignore	Input
Max Dosazene Vzdelani	Input	Ignore	Input
Mzda	Input	Ignore	Input
Obor	Input	Ignore	Input
Okres	Input	Ignore	Input
Pl	Input	Ignore	Input
Pohlavi	Input	Ignore	Input
Poptavka Status	PredictOnly	PredictOnly	PredictOnly
Pozice	Input	Ignore	Input
Ru	Input	Ignore	Input
Specializace	Input	Ignore	Input
Vek	Input	Ignore	Input
Zamereni	Input	Ignore	Input

Obrázek 19: Mining models – vstupující atributy

Zdroj: Vlastní zpracování

Tímto způsobem budu zároveň schopen odhalit případné odchylky ve scoringovém algoritmu. Pokud by se ukázalo, že se příliš vzdaluje ideálnímu modelu, nemusel by být vhodně navržen.

Pro porovnání přesnosti jednotlivých modelů slouží přehledný Lift Chart diagram, který je dostupný na záložce Mining Accuracy Chart. Křivka správně navrženého modelu by se měla co nejvíce blížit ke křivce ideálního modelu. V nabídce *input selection* je zvolena hodnota *nástup*, která je pro mne zásadní.



Series, Model	Score	Target population	Predict probability
All attributes	0,84	45,64%	37,54%
Just Scoring	0,86	33,28%	61,84%
Without scoring	0,82	44,31%	89,74%
Random Guess M...		15,00%	
Ideal Model for: Al...		48,58%	

Obrázek 20: Lift Chart

Zdroj: Vlastní zpracování

Na první pohled lze na výše uvedeném diagramu vidět, že model *All atributes* a *Without Scoring* mají velmi podobný průběh křivky, nicméně model se všemi atributy je přesnější. Je to tím, že scoringový algoritmus v sobě zahrnuje již většinu kritérií, které představují zbytek atributů. Model *Just Scoring*, dosahuje dle tabulky největšího skóre, nicméně pokud se podíváme na jeho průběh, začíná převyšovat přesnost ostatních modelů zhruba až ve 27 % populace, do té doby je model *All attributes* nejpřesnější.

Z tohoto důvodu je pro mne nejvýznamnější, jelikož je nutné si uvědomit, že 100 % populace v tomto případě znamená zhruba 200 000 kandidátů. Takto vysoké číslo nemají konzultanti běžně možnost zpracovat a oslovit. Významnou část populace představuje zhruba 15 % a do té doby je vyhodnocen jako nejvhodnější model *All Atributes*, proto bude dále uvažován pro následnou predikci.

Zároveň můžu na základě přesnosti modelu *Just Scoring* usoudit, že metoda scoringu je navržena pro potřeby preference správně a vybraná kritéria odpovídají skutečnosti.

3.4.2 Výběr vhodného algoritmu a porovnání přesnosti

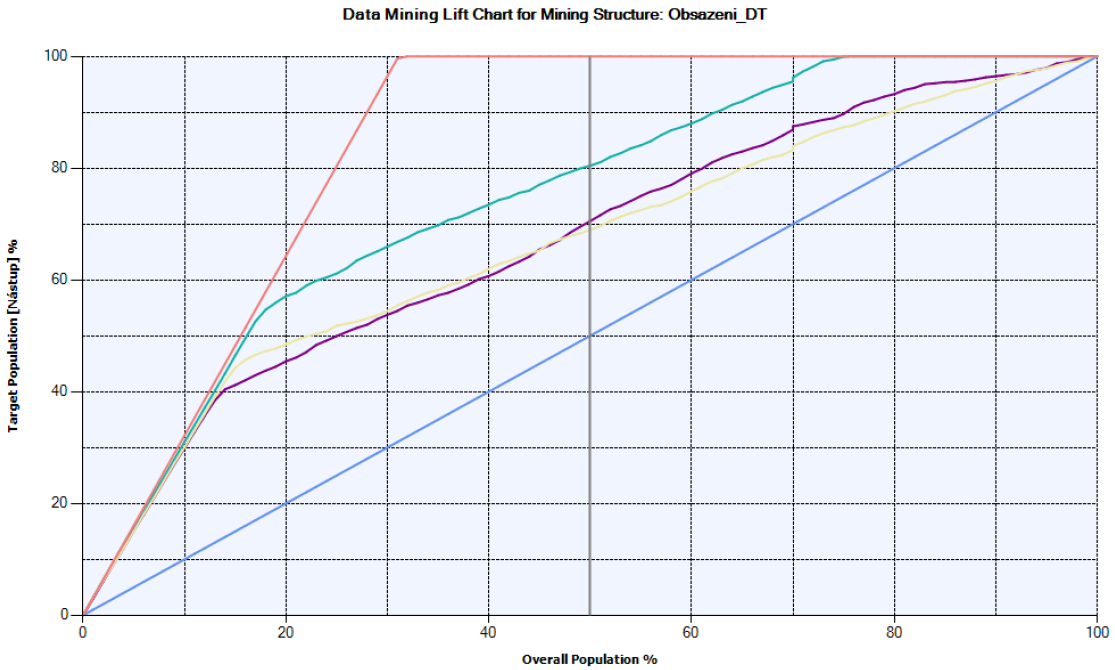
Definuji tedy další Mining Structures, kde znovu porovnám přesnost jednotlivých modelů obsahující všechny atributy, které se však již liší různými využitými algoritmy. Pro ilustraci zde srovnám pouze tři algoritmy: Rozhodovací stromy, Clustering a Neuronové sítě. Ostatní algoritmy měly horší výsledky, nebo například jako algoritmus Naive Bayes nepodporují spojité atributy, tudíž by ignorovaly atribut *Hodnoceni*.

Structure	Obsazeni_DT	Obsazeni_CL	Obsazeni_NN
	Microsoft_Decision_Trees	Microsoft_Clustering	Microsoft_Neural_Network
Bydliste	Input	Input	Input
De	Input	Input	Input
Doba	Input	Input	Input
Email	Ignore	Ignore	Ignore
En	Input	Input	Input
Es	Input	Input	Input
Fr	Input	Input	Input
Hodnoceni	Input	Input	Input
ID Kandidat	Key	Key	Key
It	Input	Input	Input
Max Dosazene Vzdelani	Input	Input	Input
Mzda	Input	Input	Input
Obor	Input	Input	Input
Okres	Input	Input	Input
Pl	Input	Input	Input
Pohlavi	Input	Input	Input
Poptavka Status	PredictOnly	PredictOnly	PredictOnly
Pozice	Input	Input	Input
Ru	Input	Input	Input
Specializace	Input	Input	Input
Vek	Input	Input	Input
Zamereni	Input	Input	Input

Obrázek 21: Mining models – algoritmy

Zdroj: Vlastní zpracování

Stejně jako v předchozím případě lze na níže uvedeném diagramu vidět přesnost jednotlivých algoritmů, respektive modelů. Nejlepších výsledků stále dosahuje algoritmus rozhodovacích stromů Obsazeni_DT, jelikož se jeho křivka nejvíce blíží k ideálnímu modelu. Rozhodovací stromy tedy využiji pro následnou predikci obsazení.



Series, Model	Score	Target population	Predict probability
Obsazeni_DT	0,88	80,45%	23,06%
Obsazeni_CL	0,79	70,52%	25,03%
Obsazeni_NN	0,79	68,95%	26,50%
Random Guess Model		50,00%	
Ideal Model for: Obsazeni_...		100,00%	

Obrázek 22: Lift Chart - Porovnání přesnosti využitých algoritmů

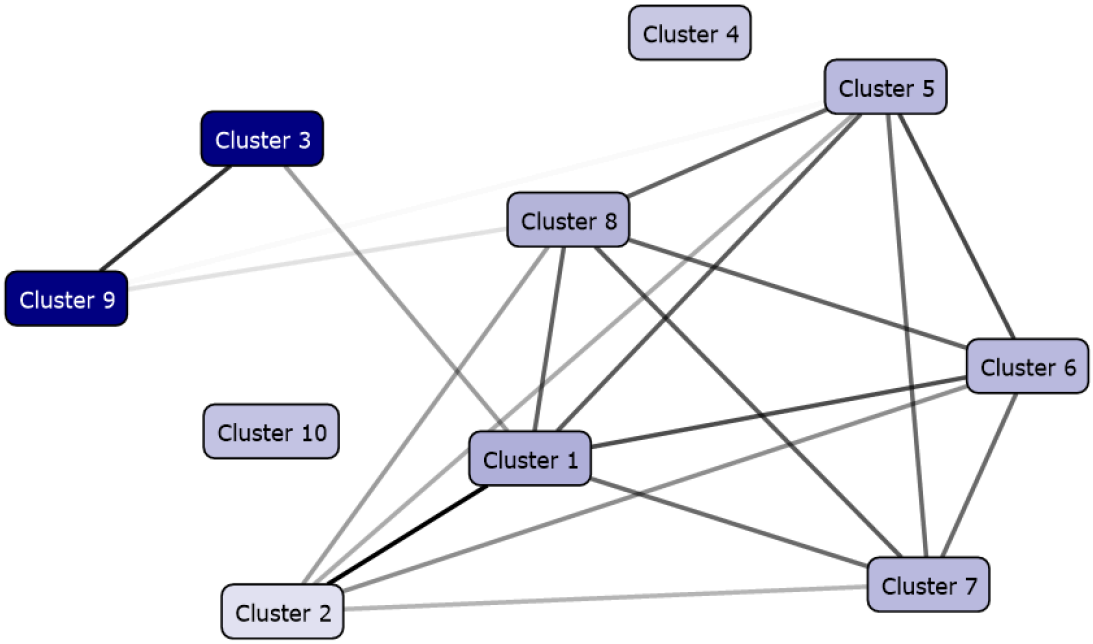
Zdroj: Vlastní zpracování

Model Obsazeni_CL s algoritmem shlukování sice nedosáhl tak vysoké přesnosti jako rozhodovací stromy, nicméně má stále velmi dobré výsledky a je zároveň výborným nástrojem pro definování segmentů. Využiji ho tedy pro rozpoznání a definici charakteristik jednotlivých skupin kandidátů.

3.4.3 Vyhodnocení modelů

Segmentace kandidátů pomocí Microsoft Clurstering Algorithm

Na níže uvedeném diagramu je znázorněna síť jednotlivých segmentů. Tmavě modré clustery představují segmenty kandidátů s největším nástupem. Cluster 3 a 9 obsahují až 89 % nástupů a zároveň si budou podobné díky silné vazbě mezi nimi. Naopak Cluster 2 obsahuje nejvyšší procento zamítnutí a to až 90 %.

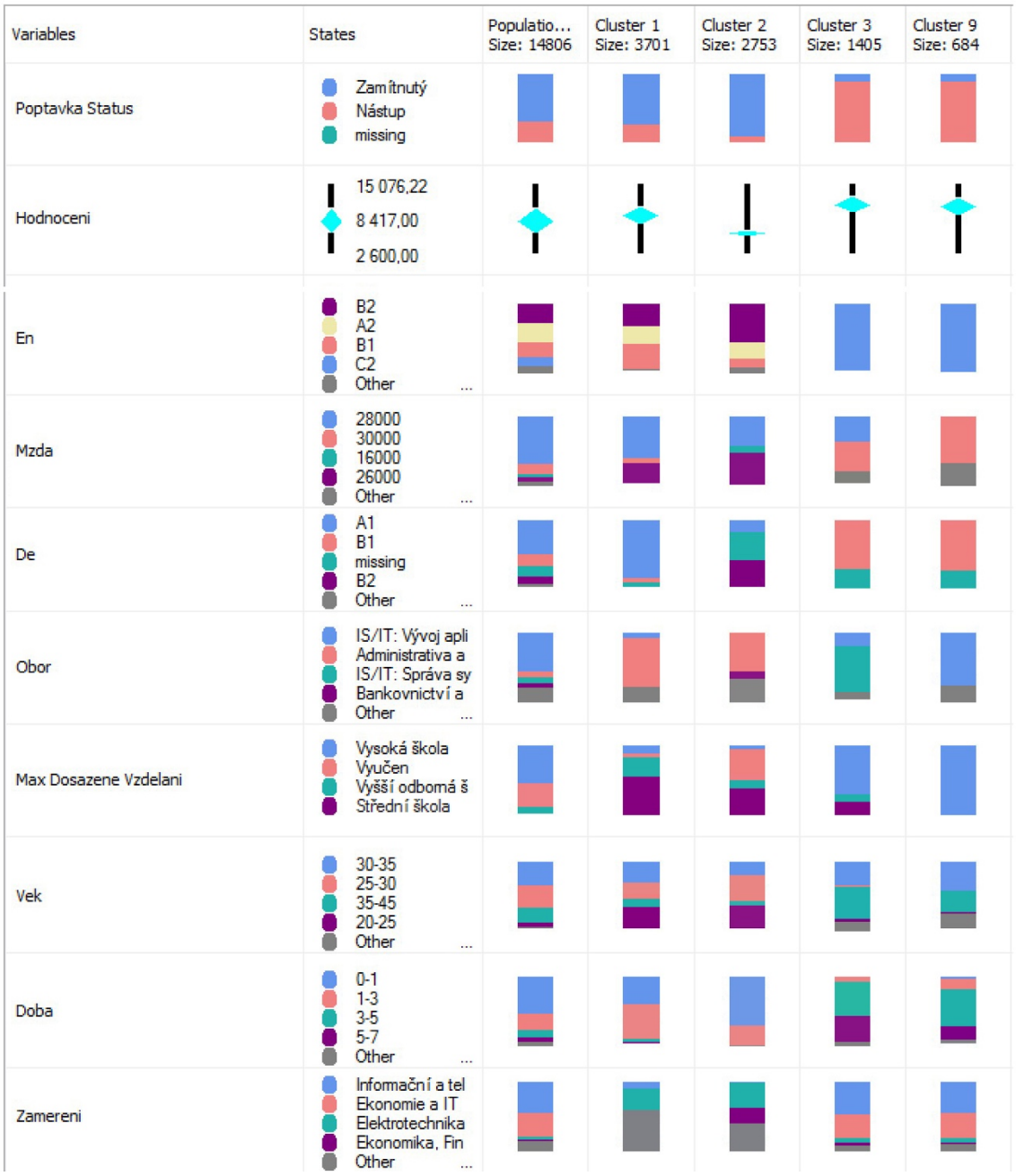


Obrázek 23: Diagram shlukovacího algoritmu

Zdroj: Vlastní zpracování

Na záložce Cluster Profiles si zobrazím diagram charakteristiky jednotlivých shluků, jehož část je uvedena níže. Pro ilustraci jsem záměrně vybral segmenty 3 a 9 s nejvyšším nástupem, segment 2 s nejvyšším zamítnutím a dále segment 1 který představuje průměrnou hodnotu nástupů. Kvůli rozsahu atributů jsou zde zobrazeny pouze ty, které mají největší vliv na určení segmentu. Celý diagram je připojen v příloze číslo 3 této práce.

Na diagramu je vždy uveden název atributu, jeho hodnoty a velikost zastoupení v jednotlivých segmentech.



Obrázek 24: Diagram charakteristiky jednotlivých segmentů

Zdroj: Vlastní zpracování

Cluster 2 s nejnižším nástupem

Kandidáti v tomto segmentu dosahují nejnižšího hodnocení ve velmi malém rozsahu 6.183 (+/- 586). Věk kandidátů se pohybuje převážně v rozsahu od 20 do 30 let. Vzdělání je zde zastoupeno nejvíce vyučením, v malé míře střední školou a nejméně vysokou a vyšší odbornou školou. Zhruba 40 % vzdělání je v různých oborech, 40 % v oboru

Elektrotechniky a zbylých 20 % v oboru Ekonomie a finance. Co se týče jazykových dovednostní, angličtina dosahuje úrovně B2 asi v 60 % případech, dále v menší míře úrovně A2 – B1. Vyšší úroveň C1 – C2 zde není vůbec zastoupena. Převážná část kandidátů nemá znalost německého jazyka, menší část dosahuje úrovně B2. Délka jejich praxe je zhruba v 67 % případech do jednoho roku, ve zbylém množství v rozsahu 1 – 3 let. Tato praxe je zastoupena převážně administrativní pozicí, ostatními činnostmi a v malé míře bankovnictvím a finančními službami. Zároveň tito kandidáti požadují poměrně vysokou mzdu v průměrné výši 27.000 korun českých.

Cluster 1 s nízkým nástupem

Kandidáti v tomto segmentu dosahují rovněž nižšího hodnocení v rozsahu 9.355 (+/- 1692). Jejich věk se pohybuje v jednotlivých rozsazích rovnoměrně, s menším zastoupením kandidátů starších 35 let. Vzdělání je zde zastoupeno nejvíce střední školou, v malé míře vyšší odbornou školou a poté vysokou školou. Zhruba 60 % vzdělání je v různých oborech, 30 % v oboru Elektrotechniky a zbylých 10 % v oboru Informační a telekomunikační technologie. Co se týče jazykových dovednostní, angličtina zde dosahuje spíše základní úrovně A2 až B2. Vyšší úroveň C1 – C2 zde není rovněž zastoupena. Převážná část kandidátů má znalost německého jazyka na úrovni A1. Délka jejich praxe se nejvíce pohybuje v rozsahu 1 – 3 let a v menší míře do jednoho roku. Tato praxe je zastoupena převážně administrativní pozicí. Mzda je zde rovněž v průměrné výši 27.000 korun českých.

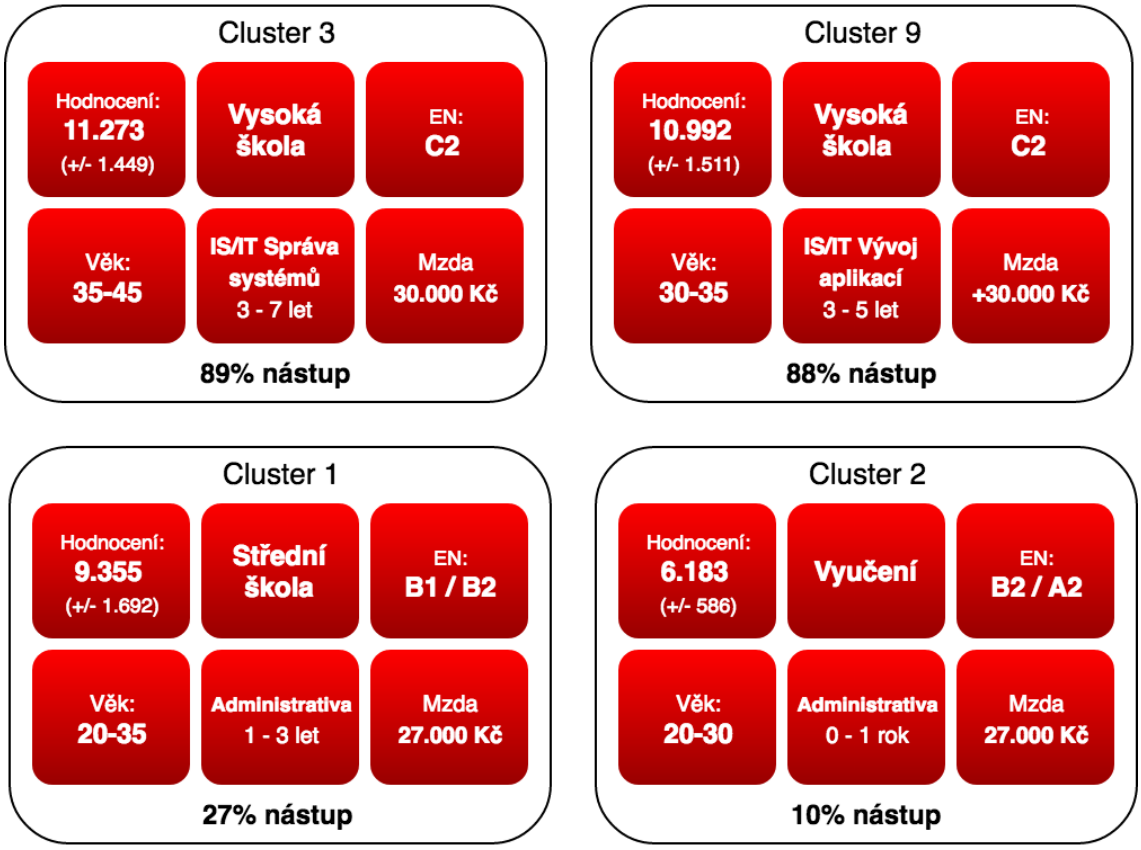
Cluster 3 s vysokým nástupem

Dosahují poměrně vysokého hodnocení v rozsahu 11.273 (+/- 1,449). Nejvíce jsou zde zastoupeni kandidáti ve věku 35 až 45 let a v menší míře ve věku 30 až 35 let s vysokoškolským vzděláním. Nicméně zhruba 15 % z nich má pouze středoškolské a 10 % vyšší odborné vzdělání. Obor těchto vzdělání je zhruba ve 48 % případech Informační a telekomunikační technologie, a ve 33 % Ekonomie a IT. Tento segment je tvořen čistě kandidáty s úrovní angličtiny C2. Znalost německého jazyka dosahuje převážně úrovně B1 a ve 30 % chybí. Délka jejich praxe se nejvíce pohybuje v rozsahu 3 – 5 let a v nepatrně menší míře v rozsahu 5 – 7 let. Tato praxe je zastoupena

převážně pozicí IS/IT: Správy systémů a poté méně IS/IT: Vývojem aplikací. Minimální požadovaná mzda dosahuje částky 30.000 korun českých.

Cluster 9 s vysokým nástupem

Tento segment dosahuje rovněž poměrně vysokého hodnocení v rozsahu 10.992 (+/- 1.511). Nejvíce jsou zde zastoupeni kandidáti ve věku 30 až 35 let a v menší míře ve věku 35 až 45 let čistě s vysokoškolským vzděláním. Obor těchto vzdělání je zastoupen téměř totožně jako v clusteru 3. Tento segment je rovněž tvořen čistě kandidáty s úrovní angličtiny C2. Znalost německého jazyka dosahuje převážně úrovně B1 a ve 28 % chybí. Délka jejich praxe je zastoupena v 55% případě v rozsahu 3 – 5 let, ve 20% v rozsahu 5 – 7 let a v 15% 1 – 3 let. Tato praxe je tvořena převážně pozicí IS/IT: Vývojem aplikací. Minimální požadovaná mzda dosahuje částky 30.000 korun českých a ve třetině případů tuto částku převyšuje.

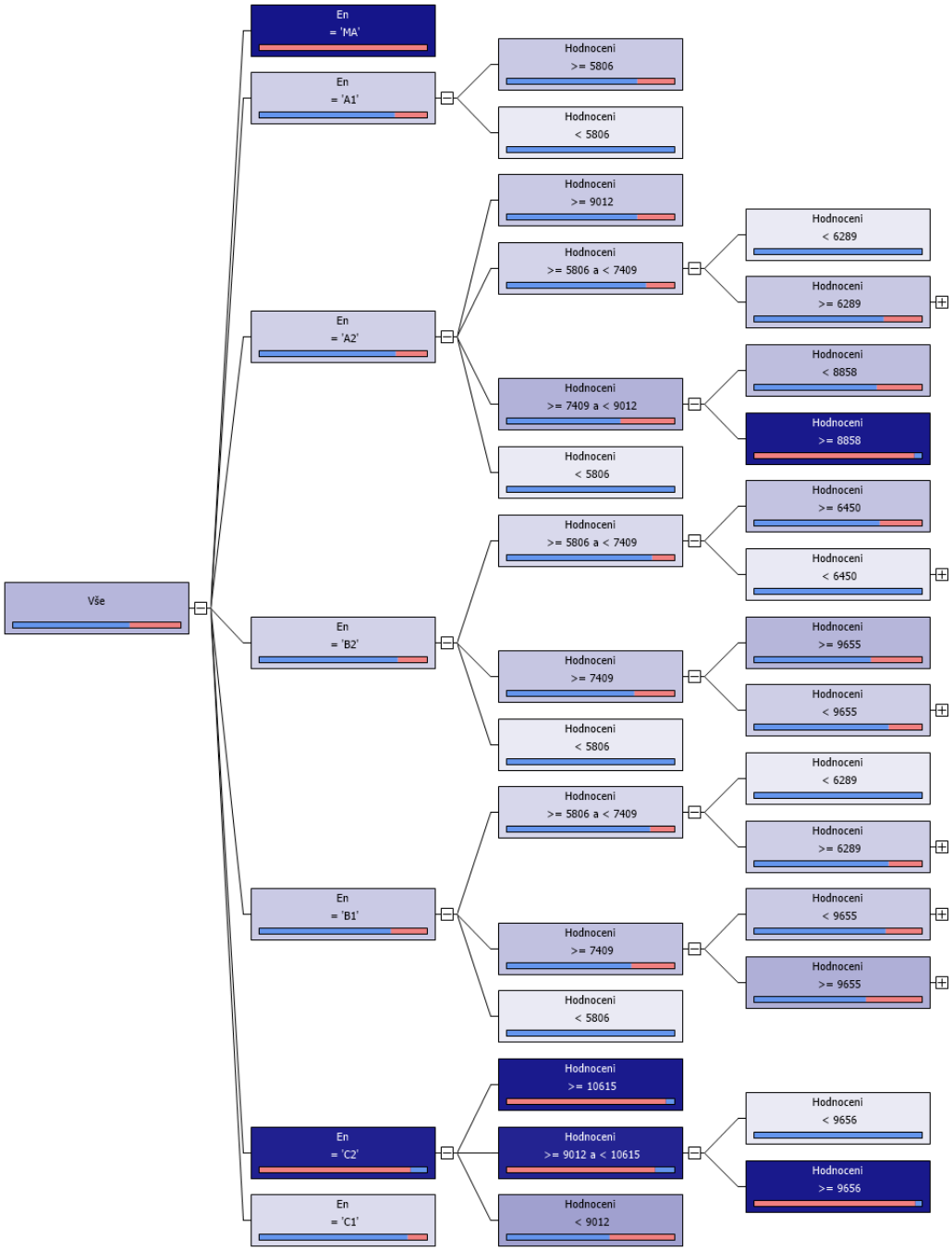


Obrázek 25: Významné vlastnosti jednotlivých clusterů

Zdroj: Vlastní zpracování

Predikce pomocí Microsoft Decision Trees Algorithm

Na následujícím diagramu je zobrazen diagram nevyváženého rozpadového stromu, na kterém mohu sledovat váhu jednotlivých atributů ovlivňujících predikovanou veličinu *Poptavka Status* s výrazněnou hodnotou *Nástup*.



Obrázek 26: Diagram rozhodovacích stromů

Zdroj: Vlastní zpracování

Jak lze vidět, predikovaná veličina je závislá ve čtyřech úrovních jen na úrovni znalosti angličtiny a hodnoty hodnocení, to potvrzuje také diagram závislosti na záložce Dependency Network.

Obrázek 27: Diagram závislosti predikovaného atributu na vstupech

Zdroj: Vlastní zpracování

Je to proto, že algoritmus hodnocení v sobě prakticky již zahrnuje všechny vstupující atributy. Z tohoto důvodu se také model *Just Scoring* příliš nelišil svou přesností od modelu *Without Scoring*. Nicméně právě jazyková úroveň angličtiny byl ten atribut, který mu lehce přidal na přesnosti. Z tohoto závěru lze navíc usoudit, že by ve scoringovém algoritmu mohlo být vhodné upravit váhu atributu EN či bodové ohodnocení nabývajících hodnot.

Červená část pruhu představuje nástup, modrá zamítnutí a tmavě zvýrazněné položky hodnoty s nejvyšší úspěšností.

Nejvyšší pravděpodobnost na nástup mají kandidáti disponující mateřským jazykem angličtina. Dále kandidáti s úrovní angličtiny C2 a hodnocením v rozsahu a hodnocením v rozsahu od 9656 do 10615. Poněkud menší šanci dosahují kandidáti s úrovní angličtiny A2 a hodnocením v rozsahu 8858 až 9012.

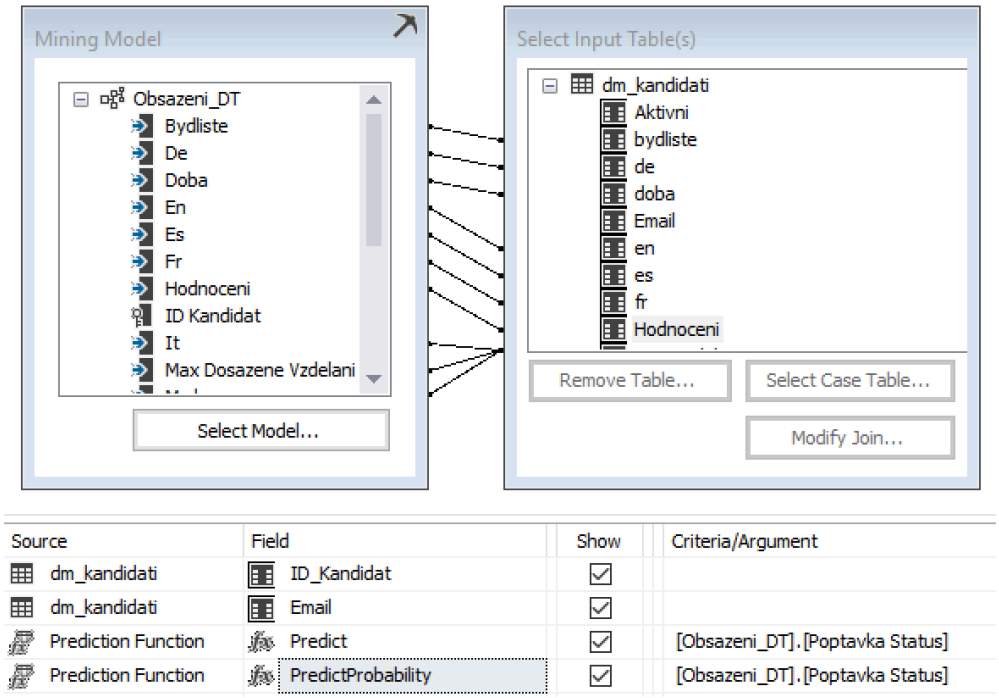
Naopak nejmenší šanci na úspěšný nástup mají kandidáti s hodnocením menším než 5806 a již příliš nezávisí na jazykové úrovni angličtiny.

Z uvedeného příkladu však nelze potvrdit, že vyšší hodnota scoringu než 10615 může představovat horší výsledky. Spíše se tak vysoké hodnoty nevyskytují v dostatečném množství ve zdrojových datech.

3.5 Predikce

Nyní po vyhodnocení jednotlivých modelů jsem schopen konečně přistoupit ke klíčové části této práce, a to k predikci nástupu kandidátů.

Předmětem predikce pro každý záznam je tedy určit, zda daný kandidát nastoupí či bude zamítnut. Predikční model se bude skládat ze dvou pohledů, respektive tabulek. V jedné budou atributy data miningového modelu a v druhé atributy tabulky, na základě které se bude predikovat. Tento model je vytvořen na záložce Mining Model Prediction.



Obrázek 28: Predikční model

Zdroj: Vlastní zpracování

Do výsledku predikce zahrnu identifikační číslo kandidáta a pro rychlý přehled také jeho e-mailovou adresu. Dále je samozřejmě nastavena predikční funkce pro atribut *Poptavka Status* a navíc pro důvěryhodnost jednotlivých výsledků jsem připojil také funkci na výpočet pravděpodobnosti této predikce. Na základě takto zadaných parametrů je vygenerován následující dotaz.

```
SELECT
     t.[ID_Kandidat], t.[Email],
     Predict([Obsazeni_DT].[Poptavka Status]),
     PredictProbability([Obsazeni_DT].[Poptavka Status])
From
     [Obsazeni_DT]
     PREDICTION JOIN
     OPENQUERY([Acis Dw],
          'SELECT
          [ID_Kandidat], [Email], [Mzda], [bydliste], [MaxDosazeneVzdelani], [en],
          [fr], [it], [de], [ru], [es], [pl], [Hodnoceni], [doba], [Pozice], [Obor],
          [Zamereni], [Okres], [vek]
          FROM [dbo].[dm_kandidati]
          ') AS t
          ON
          [Obsazeni_DT].[Mzda] = t.[Mzda] AND [Obsazeni_DT].[Obor] = t.[Obor] AND
          [Obsazeni_DT].[Bydliste] = t.[bydliste] AND
          [Obsazeni_DT].[Max Dosazene Vzdelani] = t.[MaxDosazeneVzdelani] AND
          [Obsazeni_DT].[En] = t.[en] AND [Obsazeni_DT].[Fr] = t.[fr] AND
          [Obsazeni_DT].[It] = t.[it] AND [Obsazeni_DT].[De] = t.[de] AND
          [Obsazeni_DT].[Ru] = t.[ru] AND [Obsazeni_DT].[Es] = t.[es] AND
          [Obsazeni_DT].[Pl] = t.[pl] AND [Obsazeni_DT].[Doba] = t.[doba] AND
          [Obsazeni_DT].[Hodnoceni] = t.[Hodnoceni] AND
          [Obsazeni_DT].[Pozice] = t.[Pozice] AND
          [Obsazeni_DT].[Zamereni] = t.[Zamereni] AND
          [Obsazeni_DT].[Okres] = t.[Okres] AND [Obsazeni_DT].[Vek] = t.[vek]
```

3.5.1 Vyhodnocení predikce

Výsledek predikce jsem si pro další zpracování uložil jako databázovou tabulku do datového skladu. Její náhled můžeme vidět níže. Pravděpodobnost výsledků dosahuje celkově velmi vysokých čísel, pohybujících se přes 95 %. Nicméně zde existuje i malé množství záznamů, které dosahují pravděpodobnosti nižší, a to okolo 66 %. Z tohoto důvodu bude ještě vhodné ověřit, ve kterých případech, respektive v kolika se predikce doopravdy zmýlila.

ID_Kandidat	Email	Obsazeni	PravdepodobnostVysledku
3611	[redacted]	Zamítnutý	0,999993523333175
5225	[redacted]	Zamítnutý	0,999993523333175
1182	[redacted]	Nástup	0,999990902640054
1399	[redacted]	Nástup	0,999990902640054
1770	[redacted]	Nástup	0,999990902640054

Obrázek 29: Výsledky predikce

Zdroj: Vlastní zpracování

Pro tyto účely vytvořím další pohled, ve kterém mohu porovnat predikci s realitou. Sloučím tedy tabulku s predikcí a tabulku nástupů kandidátů, abych byl schopen objevit všechny nesrovnalosti. Tento pohled bude obsahovat atributy *ID_Kandidat*, *Email*, *Obsazeni* – predikovaná hodnota, *PoptavkaStatus* – skutečná hodnota a *PravdepodobnostVysledku*. Nyní jsem schopen dle potřeby filtrovat veškeré chyby či chyby prvního a druhého řádu.

Vzhledem k počtu záznamů přikládám navíc jednoduchý dotaz, pro zjištění celkové úspěšnosti vyjádřené v procentech.

```
SELECT
       100 -
           (100 * CONVERT(float,
              (SELECT COUNT(*) FROM porovani WHERE obsazeni <> PoptavkaStatus)) /
              CONVERT(float, COUNT(*)))
       AS 'Uspesnost v %'

FROM Porovani
```

	Uspesnost v %
1	84.8181818181818

Obrázek 30: Celková úspěšnost predikce

Zdroj: Vlastní zpracování

Výsledek úspěšnosti predikce lze považovat za uspokojivý, jelikož na tomto vzorku dat dosáhl vysoké hodnoty a to téměř 85 %.

3.5.2 Nasazení

Obdobně jako automatický scoring, i zde bude probíhat výpočet modelu v pozdních nočních hodinách mimo pracovní dobu. Výsledek predikce bude generován ve formě reportu do přehledných XLS tabulek pro snadné uživatelské zpracování, a navíc ukládán rovněž do datového skladu. Celý tento proces bude mít na svědomí již dříve definovaný SQL Server Agent.

3.6 Data mining v e-mailové marketingové strategii

Běžně by bylo vhodné umístit tuto problematiku na začátek práce, avšak záměrně se jí zabývám až nyní, abych demonstroval možnosti neustálého rozšiřování oblasti Business Intelligence v podniku.

Úspěšně jsem dokázal určit pravděpodobnou hodnotu kandidáta pro společnost, a to ať již za pomocí automatického scoringu tak výsledků predikce. Tyto výsledky by bylo možné již nyní uplatnit při návrhu a cílení e-mailových kampaní, jelikož je zřejmé, že je pro společnost vhodné soustředit se výhradně na kvalitní kandidáty.

Nicméně samotná hodnota kandidáta již neurčuje, jakým způsobem bude na tyto kampaně reagovat. Proto se pokusím tuto problematiku posunout ještě o kus dále a zaměřit se na oslovení kandidátů, kteří mají nejen velkou cenu pro společnost, ale zároveň přinesou požadovanou konverzi. Konverzí v tomto případě myslím určitou odezvu, jako je například odpověď na sdělení či proklik na internetový odkaz.

Na neštěstí v současnosti však pro společnost není možné tyto uživatele definovat. Je to tím, že jí schází data, která by mohla tuto hypotézu vyvrátit či potvrdit, a proto nejsem schopen na tuto problematiku analyzovat. Společnost většinu hromadných e-mailových sdělení zasílá přímo z informačního systému. Ten však bohužel v současné době nenabízí uživatelům žádné pokročilé funkce jako je například trackování a statistiky.

3.6.1 Doporučení na změnu strategie

Tato cenná data lze velmi jednoduše získat pomocí trackování odeslaných e-mailů a následně je využít při určování dalšího směru strategie a výběru vhodné cílové skupiny. Společnosti proto důrazně doporučuji, aby aktivně začala využívat jeden z nástrojů pro e-mail marketing, který nejen, že umožňuje veškeré e-maily trackovat, ale také obsahuje další funkce jako jsou automatizované odesílání, správu kontaktů, A/B testování, přehledné statistiky a responsivní grafické šablony.

Výběr nástroje

Na trhu existuje velká řada online nástrojů pro správu emailových kampaní. Pro středně velké společnosti je vhodný například MailChimp, Ecomail.cz, MailerLite, SmartEmaling.cz a další. Jak potvrzuje následující tabulka, většina z nich nabízí prakticky stejné funkce, nicméně v odlišných cenových tarifech.

Tabulka 12: Srovnání e-mailingových nástrojů (Vlastní zpracování)

	Ecomail	Mailchimp	Mailer Lite	SmartEmailing
Zdarma do	100 kontakt.	2000 kontakt.	1000 kontakt.	200 kontakt.
Cena za 10 000 kontaktů (Kč bez DPH)	750 - 1 425	1850	860	2 400
E-maily v ceně	30 000	Bez limitu	Bez limitu	Bez limitu
Zasílání SMS	ANO	NE	NE	ANO
Drag&Drop Edit	ANO	ANO	ANO	ANO
HTML Editor	ANO	ANO	ANO	NE
Šablony	ANO	ANO	ANO	ANO
Facebook	ANO	ANO	ANO	NE
Automatizace	Zdarma	Placená verze	Zdarma	Zdarma
A/B testování	Zdarma	Zdarma	Zdarma	Zdarma
Podpora	Chat v reálném čase	Odpověď do 24 hodin	Chat v reálném čase	Reaguje v řádu minut
Obsluha	Snadná	Snadná	Snadná	Středně obtížná

Pro personální agenturou se jeví jako vhodná volba nástroj Ecomail, který je nejen že velmi dobře cenově dostupný a nabízí veškeré potřebné funkce, včetně zasílání SMS, ale je především kompletně v českém jazyce a to včetně uživatelské podpory.

3.6.2 Integrace do datového skladu

Základní údaje, které lze pomocí prostřednictvím tohoto nástroje získat o e-mailové kampani, či konkrétním uživateli jsou:

- Počet celkově doručených e-mailů
- Počet a doba otevřených e-mailů (jeden uživatel může otevřít e-mail vícekrát)
- Počet a doba unikátních/celkových prokliků na odkaz
- Demografické údaje
- Operační systém a e-mailový klient
- Odhlášení uživatele z odběru

Informace o přečtení, respektive otevření e-mailu uživatelem probíhá na základě trackovacího obrázku. Jedná se o unikátní obrázek pro každý jednotlivý e-mail, kdy ve chvíli jeho zobrazení, zaznamenává systém informaci o otevřeném e-mailu. V případě, kdy poštovní klient nestahuje automaticky obrázky, je možné zaznamenat otevření až po samotném prokliknutí na uvedený odkaz. Zaznamenání prokliku funguje velmi jednoduše, na bázi nahrazení původního odkazu. Ve chvíli, kdy uživatel klikne na odkaz uvedený v e-mailu, přechází na adresu nástroje, který proklik zaznamená a následně přesměruje uživatele na původní odkaz. Tato operace však trvá pouze zlomek vteřiny a tak uživatele nikterak neomezí.

Vybraný nástroj umožňuje samozřejmě veškeré reporty o odeslaných kampaní generovat v podobě csv souboru, a proto je snadné integrovat data do datového skladu. Postačí tedy vytvořit nový SSIS balík v projektu Integration Services, který bude zodpovědný za tuto integraci.

3.6.3 Aplikace data miningu

V kombinaci s dosavadními daty uvedených v datovém skladu, by tak společnost získala významný přehled o kandidátech a jejich reakcích na sdělení. Po aplikaci data miningu může znovu definovat jednotlivé segmenty, nyní s důrazem konverzní poměr či otevření sdělení. Obdobně jako o předchozího případu predikce, by byly vygenerovány kontaktní e-maily ve formě csv souboru, které by se importovaly přímo do e-mailingového nástroje.

4 EKONOMICKÉ ZHODNOCENÍ PRÁCE

Rozpočet na realizaci tohoto projektu byl stanoven na 65.000 Kč a konečné náklady dosáhly výše 53.500 Kč. Tyto náklady zahrnují veškeré nutné kroky, jako je analýza současného stavu, vývoj, konzultace, testování, opatření ke snížení rizik implementace, nasazení a následné školení uživatelů.

Samotná implementace této změny nevyžaduje žádné investice do nového hardwaru či dalšího softwaru, jelikož potřebná platforma Microsoft SQL Server 2014 je již úspěšně zavedena. Náklady na údržbu, které představují případnou úpravu modelů, nepřesahují částku 10.000 Kč ročně. Předpokládaná životnost tohoto řešení, je stejná jako u stávajícího informačního systému a to 5 let.

Případná investice do licence e-mailingového nástroje Ecomail, by pro společnost znamenala částku 1.725 Kč měsíčně. Společnost by tak však získala nový zdroj kvalitních dat a především možnost správně cíleného přímého marketingu.

Hlavním přínosem tohoto projektu je především zvýšení produktivity práce a ušetření nákladů v procesu Recruitmentu. Jelikož mezi jeho výstupy patří částečná automatizace procesu výběru vhodných kandidátů a zároveň zvýšení jeho úspěšnosti. Zbavení konzultantů do jisté míry odpovědnosti za ruční předvýběr a následný průzkum velkého množství kandidátů, znamená značnou časovou úsporu a eliminaci faktoru náhody při předvýběru.

Konzultanti mají nyní k dispozici číselnou hodnotu, která vyjadřuje význam kandidáta pro společnost. Na základě této hodnoty jim může informační systém nabídnout nejen pouze adekvátní kandidáty, ale zároveň je seřadit od takzvaně potenciálně nejúspěšnějších. Zároveň s výsledky data miningu, které definují jednotlivé segmenty kandidátů a současně predikují jejich pravděpodobnost k nástupu, jsou schopni lépe rozhodovat o tom, kterým kandidátům budou věnovat svůj čas.

Společnost definovala, že dříve konzultanti do předvýběru zařadili k jedné poptávce běžně i 30 až 50 kandidátů, které museli dále přezkoumat. Průzkum jednoho kandidáta stojí společnost průměrně 325 Kč. Což znamená, že jen samotný výběr vhodných kandidátů k jedné poptávce stojí zhruba 13.000 Kč. Po nasazení projektu je možné díky

jasnější představě o kvalitě kandidátů snížit jejich počet, a tedy i vynaložené náklady na tuto činnost až o polovinu.

Nicméně tato úspora se nedotkne pouze nákladů na činnost průzkumu, ale významně ovlivní náklady celého procesu obsazení. Každou jednotlivou činností v tomto procesu se náklady na kandidáta kumulují a celková částka byla odhadnuta společností na 8.341 Kč. Je tedy zřejmé, že po nasazení řešení, které sníží počet kandidátů vstupujících do procesu, významně sníží náklady na celý proces.

Nicméně vyčíslit celkovou konkrétní částku, o kterou se sníží náklady společnosti, bude možné až později při srovnání delšího časového období, ve kterém se výsledky aktivně využívaly.

Mezi nevyčíslitelné přínosy, lze zařadit například podporu při rozhodování i méně zkušených konzultantů, vyšší spokojenost pracovníků a spokojenost kandidátů, která je ovlivněna zlepšením osobního přístupu, na základě výsledků segmentace.

ZÁVĚR

Záměrem této práce bylo aplikovat nástroje data miningu do prostředí personální agentury s cílem zefektivnit firemní procesy v oblasti zpracování poptávek a zlepšení osobního přístupu ke kandidátům. K dosažení tohoto cíle bylo klíčové určit hodnotu kandidáta pro společnost na základě expertních odhadů a vytvořit algoritmus pro automatizaci této úlohy.

V teoretické části byla za pomoci odborné literatury čtenáři nastíněna problematika Business Intelligence, datových skladů, data miningu a marketingu.

V analýze současného stavu je představen klíčový proces zpracování poptávky, který tvoří základ pro vypracování vlastního návrhu řešení. Dále je zde analyzováno vnější a vnitřní prostředí společnosti, včetně vyhotovení analýzy rizik za využití metody RIPRAN. Veškeré tyto informace byly čerpány především prostřednictvím osobní konzultace se zástupci společnosti.

Po odhalení kritérií, definující úspěšného kandidáta, za využití expertních odhadů a analýzy historických dat ve spolupráci se společností bylo možné sestavit funkční scoringový algoritmus. Ten nejenže do značné míry automatizuje proces předvýběru, ale zároveň zvyšuje úspěšnost celého procesu obsazení.

Samotný postup dolování dat, byl podepřen metodikou CRISP-DM. Jeho výsledky lze považovat za velmi uspokojivé, jelikož se podařilo dosáhnout téměř 85% celkové úspěšnosti predikce obsazení. Současně byly odhaleny segmenty kandidátů, představující skupinu s nejvyšší a nejnižší úspěšností. A zároveň pomocí výsledků data miningu bylo možné ověřit, že scoringový model je navržen správně.

Na závěr jsem předložil návrhy využití dolování dat v oblasti přímého marketingu, konkrétně v oblasti direct e-mail.

V kapitole ekonomické zhodnocení práce, bylo zjištěno, že částečná automatizace předvýběru kandidátů a výsledky data miningu mají značný pozitivní vliv na náklady společnosti.

SEZNAM POUŽITÝCH ZDROJŮ

1) CONOLLY, T., C. E BEGG a R. HOLOWCZAK. *Mistrovství - databáze: profesionální průvodce tvorbou efektivních databází*. 1. vydání. Brno: Computer Press, 2009, 584 s. ISBN 978-80-251-2328-7.

2) LACKO, Ľ. *Business Intelligence v SQL Serveru 2008: reportovací, analytické a další datové služby*. 1. vydání. Brno: Computer Press, 2009, 456 s. ISBN 978-80-251-2887-9.

3) LACKO, Ľ. *Mistrovství v SQL Server 2012*. 1. vydání. Brno: Computer Press, 2013, 640 s. ISBN 978-80-251-3773-4.

4) CIOS, Krzysztof, Witold PEDRYCZ, Roman W. SWINIARSKI a Lukasz A. KURGAN. *Data mining: a knowledge discovery approach*. New York: Springer Sience Business Media, 2007. ISBN 978-0-387-33333-5.

5) DE VILLE, Barry. Microsoft data mining: integrated business intelligence for e-Commerce and knowledge management. Boston: Digital Press, c2001. ISBN 15-555-8242-7.

6) POUR, J., MARYŠKA, M. a NOVOTNÝ, O. *Business intelligence v podnikové praxi: reportovací, analytické a další datové služby*. Praha: Professional Publishing, 2012, 276 s. ISBN 978-80-7431-065-2.

7) AZEVEDO, A. – Santos Manuel Filipe. 2008. *KDD, SEMMA and CRISP-DM: A parallel overview*. Amsterdam : Iadis, 2008. ISBN: 978-972-8924-63-8.

8) DELEN, Dursun. *Real-world data mining: applied business analytics and decision making.* ISBN 01-335-5107-5.

9) Oficiální domovská stránka Microsoft [online]. [cit. 2017-23-01]. Dostupné z: http://www.microsoft.com/cs-cz/

10) ŠOLJAKOVÁ, L., J. FIBÍROVÁ. *Reporting.* 3., rozš. a aktualiz. vydání. Praha: Grada, 2010, 221 s. Finance (Grada). ISBN 978-80-247-2759-2.

11) KOTLER, P. *Moderní marketing*: 4. evropské vydání. Praha: Grada, 2007. ISBN 80-247-1545-7.

12) VAŠTÍKOVÁ, M. *Marketing služeb: efektivně a moderně*. 2., aktualiz. a rozš. vydání. Praha: Grada, 2014. Manažer. ISBN 978-80-247-5037-8.

13) JANOUCH, V. *Internetový marketing.* 2. vydání. V Brně: Computer Press, 2014. ISBN 978-80-251-4311-7.

14) FOTR, J. *Tvorba strategie a strategické plánování: teorie a praxe*. Praha: Grada, 2012. Expert (Grada). ISBN 978-80-247-3985-4.

15) DOLEŽAL, J., MÁCHAL, P. a LACKO, B. *Projektový management podle IPMA: teorie a praxe.* 1. vydání. Praha: Grada, 2012. Expert (Grada). ISBN 978-80-247-4275-5.

16) LABERGE, R. *Datové sklady: agilní metody a business intelligence.* 1. vydání. Brno: Computer Press, 2012, 350 s. ISBN 978-80-251-3729-1.

17) LIEBOWITZ, J. *Strategic intelligence: business intelligence, competitive intelligence, and knowledge management.* New York: Auerbach Publications, 2006, xviii, 223 s. ISBN 0-8493-9868-1.

18) *Měšec.cz: Zákony online* [online]. [cit. 2017-05-07]. Dostupné z: https://www.mesec.cz/zakony/zakon-o-zamestnanosti/f2610214/

19) *Obecné nařízení o ochraně osobních údajů*: GDRP [online]. [cit. 2017-03-27]. Dostupné z: https://www.gdpr.cz/

20) *Český statistický úřad: Veřejná databáze* [online]. 2017 [cit. 2017-03-27]. Dostupné z: https://www.czso.cz/

21) HATHI, K. *Microsoft breaks through in the Gartner Magic Quadrant for Business Intelligence and Analytics Platforms*. In: Official Microsoft Blog [online]. 2017 [cit. 2017-03-27]. Dostupné z: https://blogs.microsoft.com/blog/2017/02/16/microsoft-breaks-gartner-magic-quadrant-business-intelligence-analytics-platforms/

SEZNAM ZKRATEK

BI	Business Intelligence
CRISP-DM	Cross-Industry Standard Process for Data Mining
CSV	Comma-Separated Values
DM	Data Mining
DW	Data Warehouse
ETL	Extraction, Transformation, Loading
GDRP	General Data Protection Regulation
IS	Information System
IT	Information Technology
OLAP	On-Line Analytical Processing
OLE DB	Object Linking and Embedding Database
SQL	Structured Query Language

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

SEZNAM PŘÍLOH

Příloha 1 – Scoring – bodové hodnocení hodnot

Hodnota	Body
Maximální dosažené vzdělání	
Vysoká škola	100
Vyšší odborná škola	65
Střední škola	50
Vyučen	15
Minimální požadovaná mzda	
40.000 - 50.000	100
30.000 - 40.000	90
50.000 - 60.000	80
> 60.000	70
20.000 - 30.000	40
< 20.000	10
Znalost Jazyka	
C2	100
MA	85
C1	80
B2	60
B1	40
A2	15
A1	5
Věk	**Body**
30 - 39	50
20 - 29	45
40 - 50	40
< 20	20
> 50	20
Preferovaný pracovní poměr	**Column1**
HPP	100
Zkrácený poměr	50
Živnostenský list	50
Brigáda	20
Jiné	20
Praxe obor	
IS/IT: Projekce a konzultace	100
IS/IT: Správa systémů a HW	100
IS/IT: Vývoj aplikací a systémů	100
Elektrotechnika / Energetika	80

Délka praxe v letech	
3-5	100
1-3	85
5-7	85
7 a více	85
0-1	65
zkušební doba	20
Praxe pozice	
Technická podpora / Administrátor/Správce OS a sítí	100
Programátor	100
Team leader	100
Tester	100
ICT manager / ředitel IT/IS	85
Správce aplikací	85
Systémový inženýr správy aplikací	85
Technik HW	85
Product manager	80
Grafické	80
Project manager	80
Analytik IS/IT	75
Scrum master	75
Konzultant IS/IT	70
Architekt/ Projektant IS/IT	65
Problem manager	60
Auditor	40
Lektor/ Instruktor	40
Střední Škola Obor	
Informační a telekomunikační technologie	100
Elektro, automatizace	85
Průmyslová škola	85
Gymnázium	20
Vysoká škola a Vyšší odborná škola - obor	
Informační a telekomunikační technologie	100
Elektrotechnika	95
Ekonomie a IT	80
Ekonomika, Finance, Management	50
Administrativa	40
Humanitní obory	0
Marketing	0
Právo	0
Strojírenství	0

Příloha 2 – Scoring – zdrojový kód

```
alter PROCEDURE hodnoceni
AS
	--IT--
	--vahy
	declare
		@vek tinyint = 5, @mzda tinyint = 7, @maxVzdelani tinyint = 6,
		@EN tinyint = 8, @DE tinyint = 7, @RU tinyint = 7,
		@FR tinyint = 5, @IT tinyint = 4, @ES tinyint = 4,
		@PL tinyint = 4, @pomer tinyint = 3, @PraxeDobaCela tinyint = 5,
		@PraxeDobaSoucasna tinyint = 4, @PraxeObor tinyint = 7,
		@PraxePozice tinyint = 8, @skolaVysObor tinyint = 7,
		@skolaStrObor tinyint = 5

select ID_Kandidat, sum(hodnoceni) as hodnoceni into #temp from (
	select
	--Vek
	case
		when year(getdate())-RokNarozeni between 30 and 39 then 100 * @vek
		when year(getdate())-RokNarozeni between 20 and 29 then 90 * @vek
		when year(getdate())-RokNarozeni between 40 and 50 then 80 * @vek
		when year(getdate())-RokNarozeni < 20 then 40 * @vek
		when year(getdate())-RokNarozeni > 50 then 40 * @vek
		else 40 * @vek
	end  +
	--Mzda
	case
		when Mzda between 40000 and 50000 then 100 * @mzda
		when Mzda between 30000 and 40000 then 90  * @mzda
		when Mzda between 50000 and 60000 then 80  * @mzda
		when Mzda > 60000 then 70 * @mzda
		when Mzda between 20000 and 30000 then 40 * @mzda
		when Mzda < 20000 then 10  * @mzda
		else 10 * @mzda
	end  +
	--Vzdelani
	case
	when MaxDosazeneVzdelani = 'Vysoká škola' then 100 * @maxVzdelani
	when MaxDosazeneVzdelani = 'Vyšší odborná škola' then 65 * @maxVzdelani
	when MaxDosazeneVzdelani = 'Střední škola' then 50 * @maxVzdelani
	when MaxDosazeneVzdelani = 'Vyučen' then 15 * @maxVzdelani
	else 0
	end  +
	--Jazyky
	case
		when EN = 'C2' then 100 * @EN
		when EN = 'MA' then 85 * @EN
		when EN = 'C1' then 80 * @EN
		when EN = 'B2' then 60 * @EN
		when EN = 'B1' then 40 * @EN
		when EN = 'A2' then 15 * @EN
		when EN = 'A1' then 5 * @EN
		else 0
	end  +
```

```
case
       when DE = 'C2' then 100 * @DE
       when DE = 'MA' then 85 * @DE
       when DE = 'C1' then 80 * @DE
       when DE = 'B2' then 60 * @DE
       when DE = 'B1' then 40 * @DE
       when DE = 'A2' then 15 * @DE
       when DE = 'A1' then 5 * @DE
       else 0
end  +
case
       when DE = 'C2' then 100 * @DE
       when DE = 'MA' then 85 * @DE
       when DE = 'C1' then 80 * @DE
       when DE = 'B2' then 60 * @DE
       when DE = 'B1' then 40 * @DE
       when DE = 'A2' then 15 * @DE
       when DE = 'A1' then 5 * @DE
       else 0
end  +
case
       when RU = 'C2' then 100 * @RU
       when RU = 'MA' then 85 * @RU
       when RU = 'C1' then 80 * @RU
       when RU = 'B2' then 60 * @RU
       when RU = 'B1' then 40 * @RU
       when RU = 'A2' then 15 * @RU
       when RU = 'A1' then 5 * @RU
       else 0
end  +
case
       when FR = 'C2' then 100 * @FR
       when FR = 'MA' then 85 * @FR
       when FR = 'C1' then 80 * @FR
       when FR = 'B2' then 60 * @FR
       when FR = 'B1' then 40 * @FR
       when FR = 'A2' then 15 * @FR
       when FR = 'A1' then 5 * @FR
       else 0
end  +
case
       when IT = 'C2' then 100 * @IT
       when IT = 'MA' then 85 * @IT
       when IT = 'C1' then 80 * @IT
       when IT = 'B2' then 60 * @IT
       when IT = 'B1' then 40 * @IT
       when IT = 'A2' then 15 * @IT
       when IT = 'A1' then 5 * @IT
       else 0
end  +
case
       when ES = 'C2' then 100 * @ES
       when ES = 'MA' then 85 * @ES
       when ES = 'C1' then 80 * @ES
       when ES = 'B2' then 60 * @ES
       when ES = 'B1' then 40 * @ES
       when ES = 'A2' then 15 * @ES
       when ES = 'A1' then 5 * @ES
       else 0
end +
```

```
        case
                when PL = 'C2' then 100 * @PL
                when PL = 'MA' then 85 * @PL
                when PL = 'C1' then 80 * @PL
                when PL = 'B2' then 60 * @PL
                when PL = 'B1' then 40 * @PL
                when PL = 'A2' then 15 * @PL
                when PL = 'A1' then 5 * @PL
                else 0
        end as 'hodnoceni',
        ID_Kandidat

        from Dim_Kandidat dk

union all

        select
        --Delka praxe v letech
        sum(
        case
                when dkp.doba = '0-1' then 65 * @praxeDobaCela
                when dkp.doba = '1-3' then 85 * @praxeDobaCela
                when dkp.doba = '3-5' then 100 * @praxeDobaCela
                when dkp.doba = '5-7' then 85 * @praxeDobaCela
                when dkp.doba = '7 a více' then 85 * @praxeDobaCela
                when dkp.doba = 'zkušební doba' then 20 * @praxeDobaCela
                else 0
        end
        )+
        --Praxe obor
        sum(
        case
                when left(obor, 2) = 'IS' then 100 * @PraxeObor
                when dkp.doba = 'Elektrotechnika / Energetika' then 80 *@PraxeObor
                --atd
                else 0
        end
        )+
        --Praxe pozice
        sum(
        case
        when dkp.Pozice = 'Technická podpora then 100 * @praxePozice
        when dkp.Pozice = 'Programátor' then 100 * @praxePozice
        when dkp.Pozice = 'Team leader' then 100 * @praxePozice
        when dkp.Pozice = 'Tester' then 100 * @praxePozice
        when dkp.Pozice = 'ICT manager / ředitel IT/IS' then 85 * @praxePozice
        when dkp.Pozice = 'Správce aplikací' then 85 * @praxePozice
        when dkp.Pozice = 'Systémový inžený' then 85 * @praxePozice
        when dkp.Pozice = 'Technik HW' then 85 * @praxePozice
        when dkp.Pozice = 'Product manager' then 80 * @praxePozice
        when dkp.Pozice = 'Grafické' then 80 * @praxePozice
        when dkp.Pozice = 'Project manager' then 80 * @praxePozice
        when dkp.Pozice = 'Analytik IS/IT' then 75 * @praxePozice
        when dkp.Pozice = 'Scrum master' then 75 * @praxePozice
        when dkp.Pozice = 'Konzultant IS/IT' then 70 * @praxePozice
        when dkp.Pozice = 'Architekt/ Projektant IS/IT' then 65 * @praxePozice
        when dkp.Pozice = 'Problem manager' then 60 * @praxePozice
        when dkp.Pozice = 'Auditor' then 40 * @praxePozice
        when dkp.Pozice = 'Lektor/ Instruktor' then 40 * @praxePozice
        --atd
```

```
	else 0
	end
	) as 'hodnoceni',
	ID_Kandidat

	from Dim_KandidatPraxe dkp
	group by dkp.ID_Kandidat

union all

	--Delka posledni/soucasne praxe v letech
	select
		case
			when dkp.doba = '0-1' then 65 * @PraxeDobaSoucasna
			when dkp.doba = '1-3' then 85 * @PraxeDobaSoucasna
			when dkp.doba = '3-5' then 100 * @PraxeDobaSoucasna
			when dkp.doba = '5-7' then 85 * @PraxeDobaSoucasna
			when dkp.doba = '7 a více' then 85 * @PraxeDobaSoucasna
			when dkp.doba = 'zkušební doba' then 20 * @PraxeDobaSoucasna
			else 0
		end as 'hodnoceni',
		dkp.id_kandidat

	from Dim_KandidatPraxe dkp
		inner join(
			select max(id) as id from Dim_KandidatPraxe group by
ID_Kandidat
		) dkp2
		on dkp.id = dkp2.id
	group by ID_Kandidat, doba

union all

	--Obor vysoke a vyssi odborne skoly
	select
		sum(
		case
		when dkv.Zamereni = IT/ICT' then 100 * @skolaVysObor
		when dkv.Zamereni = 'Elektrotechnika' then 95 * @skolaVysObor
		when dkv.Zamereni = 'Ekonomie a IT' then 80 * @skolaVysObor
		when dkv.Zamereni = 'Ekonomika, Management' then 50 * @skolaVysObor
		when dkv.Zamereni = 'Administrativa' then 40 * @skolaVysObor
		else 0 * @skolaVysObor
		end
		) as 'hodnoceni',
		ID_Kandidat

	from Dim_KandidatVzdelani dkv
	group by dkv.ID_Kandidat, dkv.Skola
	having dkv.Skola = 'Vysoká škola' or dkv.Skola = 'Vyšší odborná škola'

union all
```

```
	--Obor stredni skoly
	select
		sum(
		case
		when dkv.Zamereni = 'IT/ICT' then 100 * @skolaStrObor
		when dkv.Zamereni = 'Elektro, automatizace' then 85 * @skolaStrObor
		when dkv.Zamereni = 'Průmyslová škola' then 85 * @skolaStrObor
		when dkv.Zamereni = 'Gymnázium' then 20 * @skolaStrObor
		else 0 * @skolaStrObor
		end
		) as 'hodnoceni',
		ID_Kandidat

	from Dim_KandidatVzdelani dkv
	group by dkv.ID_Kandidat, dkv.Skola
	having dkv.Skola = 'Střední škola'

union all

	--Preferovany pracovni pomer
	select
		sum(
		case
			when dkpp.Typ = 'HPP' then 100 * @pomer
			when dkpp.Typ = 'Zkrácený poměr' then 50 * @pomer
			when dkpp.Typ = 'Živnostenský list' then 50 * @pomer
			when dkpp.Typ = 'Brigáda' then 20 * @pomer
			else 20 * @pomer
		end
		) as 'hodnoceni',
		ID_Kandidat

	from Dim_KandidatPracovniPomer dkpp
	group by dkpp.ID_Kandidat
)
x group by ID_Kandidat

MERGE INTO dim_Kandidat dk
   USING #temp t
      ON t.id_kandidat = dk.id_kandidat
WHEN MATCHED THEN
   UPDATE
      SET hodnoceni = t.hodnoceni;
GO
```

Příloha 3 - Diagram charakteristiky jednotlivých segmentů