Univerzita Pardubice

Fakulta ekonomicko-správní

Model pro ohodnocení ojetého vozidla

Bc. Ivo Brett

Diplomová práce

2008

Univerzita Pardubice
Fakulta ekonomicko-správní
Ústav systémového inženýrství a informatiky
Akademický rok: **2007/2008**

ZADÁNÍ DIPLOMOVÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení: **Bc. Ivo BRETT**

Studijní program: **N6209 Systémové inženýrství a informatika**

Studijní obor: **Informatika ve veřejné správě**

Název tématu: **Model pro odhad ceny ojetého vozidla**

Zásady pro vypracování:

Popis současného stavu stanovení ceny ojetých vozodel.
Návrh modelu pro odhad ceny ojetého vozidla.
Implementace navrženého modelu.
Ohodnocení navrženého modelu.

Rozsah grafických prací:

Rozsah pracovní zprávy:

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

Seznam odborné literatury:

BERKA, P.: Expertní sytémy.1.vyd. V Praze. 1998. 160 s. ISBN 80-7079-873-4

DUŠEK, F.: Matlab a Simulink, úvod do používání.1. vyd. Pardubice, 2005. 172 s. ISBN 80-7194-776-8

KLIR, G.: Facest of Systéme Science.1.vyd. London: Plenum, 1991. 664 s. ISBN 0-306-43959-X

Vedoucí diplomové práce: **Ing. Miloslav Hub, Ph.D.**
Ústav systémového inženýrství a informatiky

Datum zadání diplomové práce: **30. října 2007**

Termín odevzdání diplomové práce: **26. května 2008**

prof. Ing. Jan Čapek, CSc.
děkan

L.S.

doc. Ing. Pavel Petr, Ph.D.
vedoucí ústavu

V Pardubicích dne 30. ledna 2007

SOUHRN

Diplomová práce se zabývá problematikou stanovení ceny ojetých vozidel. Věnuje pozornost současnému stavu ohodnocování použitých automobilů. Pojednává o důležitosti přípravy vhodných dat pro další zpracování. Zaměřuje se zejména na návrh modelu hodnotící ojetá vozidla s využitím neuronových sítí a rozhodovacích stromů. Navržené modely jsou implementovány a testovány pomocí připravených dat. Na základě výsledků testů je vybrán model dosahující nejnižší chybovosti.

KLÍČOVÁ SLOVA

Předzpracovaní, CRISP-DM, rozhodovací stromy, neuronové sítě, modelování

TITLE

Rating model of used cars

SUMARY

The main objective of the study is to design, develop and determine system for pricing used cars via neural network systems and decision tree analysis. Thesis consists of: current evaluating model, data and data preparation, system model design, results tests and describtion of chosen model with the lowest error rate.

KEY WORDS

Data preparation, CRISP-DM, decision trees, neural network, modelling

OBSAH

ÚVOD

Velká část lidí jednoho dne bude řešit problém, zda koupit nový nebo ojetý automobil. Přitom vznikne otázka: „Neplatím za tento automobil zbytečně moc?“ Automobily se dají porovnávat podle mnoha parametrů a kritérií, proto je velmi složité pro jednotlivce, který se prodejem a nákupem automobilů nezabývá, říci: „Ano, toto je odpovídající cena tomuto druhu vozidla“. Na českém trhu jsou velké autobazary, které velice dobře prosperují. Prosperují buď z důvodu, že levně nakupují a prodávají za tržní cenu, nebo nakupují za tržní cenu a prodávají za cenu vyšší. Dále jsou na českém trhu i menší autobazary a soukromí dopravci, kteří vydělávají na silné koruně a automobily do České republiky dovážejí ze zahraničí. Není žádným tajemstvím, že použitá vozidla v zahraničí bývají zpravidla prodávána za mnohem nižší ceny než v České republice.

Cílem diplomové práce je modelovat a navrhnout systém, který bude provádět ohodnocení vozidel. Modely budou jednotlivým automobilům přidělovat adekvátní prodejní cenu. Dále budou zhodnoceny výsledky jednotlivých modelů a vybrán nejvhodnější model. K této práci budou použita data získaná od různých autobazarů podnikajících v České republice. Jako vzorová data pro naučení modelu budou sloužit data o automobilech, u nichž je známa i prodejní cena nabízená v autobazarech.

Diplomová práce má následující členění: V první části bude popsán současný stav stanovení cen ojetých automobilů. Druhá část bude věnována návrhu modelu na odhad ojetého vozidla. Závěrečná část se zaměří na implementaci a ohodnocení navrženého modelu. Zároveň bude provedeno porovnání funkčnosti modelů.

K této práci byla k dispozici vstupní data, bylo přihlédnuto k odborné literatuře vztahující se k danému tématu.

1. STÁVAJÍCÍ SITUACE

Jak již bylo předesláno v úvodu, mnoho z nás bude nebo již řešilo otázku koupě vozidla. Někteří se spoléhali a spoléhají na své subjektivní rozhodnutí, jiní důvěřují známým a někteří využívají například ohodnocení od soudních znalců.

1.1 Stávající možnosti ohodnocování ojetých vozidel

1.1.1. Společnost IBS

IBS expert je společností s mezinárodní působností poskytující svým klientů poradenství a informace v automobilové a dopravní problematice. Díky orientaci na moderní komunikační technologie a vlastnímu vývoji v oblasti informační techniky se z odborně zaměřené firmy postupem času vyprofilovala v poskytovatele informačních služeb v sektoru automobilového průmyslu a obchodu. Jedním z produktů společnosti IBS expert je systém pro ohodnocování ojetých vozidel TAXexpert. Na webových stránkách této společnosti je možnost stáhnutí demoverze programu TAXexpert [7].

Po nastavení vstupních informací o vozu do programu TAXexpert, např. druh vozu, výrobce, model, rok výroby, výbava, poškození různých částí vozu, atd., provede program propočet a nabídne dvě hodnoty: prodejní a nákupní cenu.

Jak vyplynulo z konzultace s pracovníky jednoho středně velkého autobazaru není shora uvedený systém mezi prodejci v praxi používán. Systém slouží poskytovatelům úvěrů (bankovním institucím). Poskytovatel se na základě tohoto systému rozhodne zda kupujícímu poskytne na kupovaný vůz úvěr a v jaké výši. Z rozhovoru byly zjištěny skutečnosti, proč prodejci tento systém nepoužívají. Autoprodejci rozhodují podle svých zkušeností a velmi pěkné automobily ohodnotí vyšší, než průměrnou cenou. Systém TAXexpert nemá možnost tohoto subjekviního ohodnocení, z toho plyne, že velice pěkný vůz podhodnotí. Zákazník následně nedostane na tento vůz úvěr v potřebné výši. Zbývající částku pak musí kupující složit u poskytovatele úvěru ve formě akontace.

1.1.2. Ohodnocení na základě odhadu majitele či prodejce

Menší autobazary fungují odlišně. Majitel autobazaru vlastní pouze plochu, na které je autobazar provozován a účtuje si určitou částku za den a místo. Cena, za kterou je automobil v autobazaru vystavován, závisí pouze na majiteli prodávaného vozu. Provozovatel autobazaru dostává peníze, i když auto neprodá. Automobil je ohodnocen pouze subjektivně, a to v závislosti na značce vozu, opotřebení různých částí podvozku, jízdních vlastností atd. Důležitou roli u těchto menších autobazarů hraje i lukrativnost místa tohoto autobazaru.

Pokud provozovatel autobazaru je současně i majitelem nabízených vozů, prodává je také, jako středně velké autobazary na základě svých zkušeností a subjektivního odhadu.

1.1.3. Využití soudního znalce

Ohodnocení ojetého vozidla je podrobně popsán ve znaleckých standardech určených k oceňování motorových vozidel vydaných v roce 2005[11]. V těchto pokynech jsou naprosto striktně nastavené stupnice, v jakých může znalec dané vozidlo ohodnotit. Na základě jeho uvážení udává do vzorců popsaných ve standardech různou výši opotřebení. Pomocí těchto vzorců následně dopočítá srážku z původní ceny a tím ohodnotí sledovaný automobil. Tento způsob není často využíván, neboť se za znalecký posudek vynakládají značné finanční prostředky.

2. MOŽNÉ METODIKY

Bylo zjištěno že existují metodiky, které si kladou za cíl poskytnout uživatelům jednotný rámec pro řešení různých úloh z oblasti dobývání znalostí z dat. Tyto metodiky umožňují sdílet a přenášet zkušenosti z úspěšných projektů. Za některými stojí producenti programových systémů (metodika 5A firmy SPSS nebo metodika SEMMA firmy SAS). Jiné vznikají ve spolupráci výzkumných a komerčních institucí jako „softwarově nezávislé" CRISP-DM [1].

2.1 Metodika 5A

Metodiku 5A nabízí firma SPSS jako svůj pohled na proces dobývání znalostí. Název metodiky je akronymem pro jednotlivé prováděné kroky.

- assess – posouzení potřeb projektu,
- access – shromáždění potřebných dat,
- analyze – provedení analýz,
- akt – přeměna znalostí na akční znalosti,
- automate – převedení výsledků analýzy do praxe.

Žádná data nemají význam, jestliže jsou oddělena od kontextu. Prvním krokem v analytickém procesu je tedy stanovení kontextu – cílů, strategií a procesů. Materiál SPSS k tomu říká[1]:

- určete data, jejich sběr, pořízení a skladování je nutné zajistit k provedení takových analýz, které chcete realizovat.
- připravte se na své projekty a obory, v nichž rozhodujete – jejich porozuměním zabezpečte ty analytické nástroje, které potřebujete.
- vzdělávejte a trénujte všechny lidi, kteří myslí analyticky a používají efektivně software jako součást přemýšlení nad problémy a analýzu dat jako nedílnou složku rozhodovacího procesu.

Druhým krokem v metodice 5A je sběr a příprava dat. Je třeba získat vhodné soubory z podnikových datových skladů, datových bází, odkazových systémů a jiných interních

zdrojů, Lze využít i data týkající se daného problému, která jsou nabízena veřejně. Data lze rovněž získat vlastními průzkumy nebo od výzkumné firmy.

Třetím krokem je používání různých analytických postupu k tomu, abychom nalezli odpovědi na otázky stanovené v prvním kroku. V tomto kroku se data přeměňují na informace a znalosti. Firma SPSS doporučuje širokou škálu nástrojů pro zkoumání a porozumění datům počínaje deskriptivní statistikou, přes metodu OLAP až po metody strojového učení (rozhodovací stromy, neuronové sítě). Doporučení je zřejmé: „Použijte více metod a porovnejte jejich výsledky a vhodnost, abyste získali nejlepší řešení a navíc rychle a jednoduše“.

Čtvrtý krok procesu obsahuje doporučení, řadu dodatečných otázek a následné rozhodnutí. Znalosti nalezené v předcházejícím kroku je zde mění na znalosti akční. Nalezené výsledky by měly být předkládány v jasné a srozumitelné podobě.

Pátým krokem je převedení výsledků analýzy do praxe. Tento krok obsahuje všechny činnosti, kterými lze zajistit aplikaci učiněných rozhodnutí. Sem patří například to, že vytvoříme praktické rozhraní, abychom rozvinuli nalezené modely do takového formátu, který je snadný pro užívání a porozumění v běžné a opakované praxi organizace a pro monitorování výsledků (a důsledků) prováděných rozhodnutí. Další z doporučení zní: „Automatizujte své analýzy tak, aby opakující se úlohy nezabíraly čas a abyste mohli snadno aktualizovat své modely s tím, jak přicházejí nové výsledky“ [1].

2.2 Metodika SEMMA

Enterprise Miner, softwarový produkt firmy SAS, vychází z vlastní metodiky pro dobývání znalostí z databází. Název SEMMA opět charakterizuje jednotlivé prováděné kroky:

- sample (vybírání náhodných objektů),
- explore (vizuální explorace a redukce dat),
- modify (seskupování objektů a hodnot atributů, datové transformace),
- model (analýza dat: neuronové sítě, rozhodovací stromy, statistikcé techniky, asociace a shlukování)
- assess (porovnání modelů a interpretace).

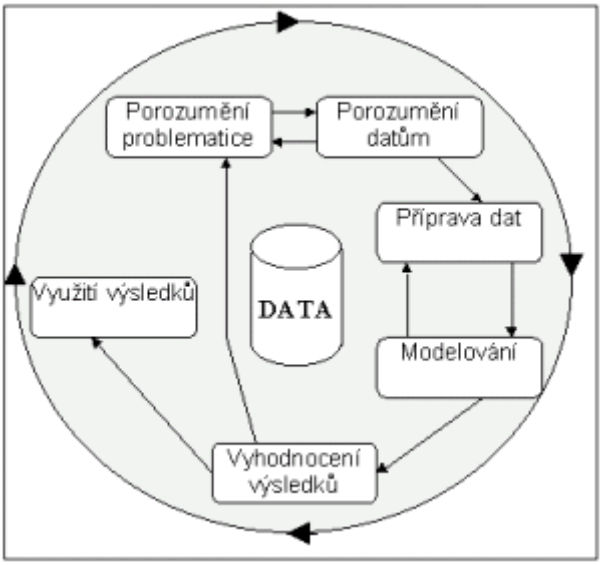
Důraz se klade na snadnou interpretaci výstupů ve formě srozumitelné obchodnímu uživateli [1].

2.3 Metodologie CRISP-DM

Metodologie CRISP-DM (CRoss-Industry Standard Process for Data Mining). Vývoj metodologie CRISP-DM byl zahájen jako projekt evropské komise definující model standardního postupu při vytváření data miningových projektů. Tato metodologie je majetkem partnerů CRISP-DM konsorcia: NCR systems Engineering Kopenhagen, Daimler Chrysler atd [6]. Firma SPSS se dolováním z dat zabývá a vytvořila programové prostředí Clementine, které je přímo specializované na tuto problematiku.

Metodologie CRISP-DM rozděluje celý proces data miningového projektu do šesti základních etap, v rámci nichž dále rozlišuje další kroky. Těmito etapami jsou (obr.1):

- Definování cílů,
- Porozumění datům,
- Předzpracování dat,
- Modelování,
- Hodnocení výsledků,
- Implementace vytvořeného modelu [6].



Obr.1: Schéma etap metodologie Crisp-dm (zdroj: [14])

Metodiky 5A, SEMMA a Crisp-dm se principiálně v postupu řešení projektu v mnohém shodují. V práci byla zvolena metodologie Crisp-dm. Crisp-dm je výsledkem dohody více firem: NCR Systems Engineering Copenhagen (USA a Dánsko), DaimlerChrysler AG (Německo), SPSS Inc. (USA) and OHRA Verzekeringen en Bank Groep B.V (Holandsko). Metodiky 5A a Semma jsou produktem jedné firmy (SPSS, resp. SAS)[13].

Dalším důvodem pro zvolení metodologie CRISP-DM byl fakt, že Crisp-dm je popsána v mnoha dostupných zdrojích [1,2,5,6,10,13,14], a že byl k dispozici softwarový produkt firmy SPSS program Clementine, ve kterém je práce z velké části tvořena.

3. DEFINOVÁNÍ CÍLŮ

Tato úvodní fáze je zaměřena na pochopení cílů úlohy a požadavků na řešení formulovaných z manažerského hlediska. Manažerská formulace musí být následně převedena do základní úlohy pro dobývání znalostí z databází. V této fázi se rovněž provádí inventura zdrojů, hodnotí se možná rizika, náklady a přínos použití metod dobývání znalostí z databází a stanovuje se předběžný plán prací[1].

Hlavním cílem celé práce je navrhnutí takového modelu, který dokáže odhadnout cenu ojetého vozidla. S využitím výpočetní techniky a postupů budou tvořeny modely. Následně proběhne test vytvořených modelů. Po vyhodnocení bude vybrán nejúspěšnější model. Aby bylo možné k těmto cílům dospět, bude nutné data k možnému použití připravit.

4. POROZUMĚNÍ DATŮM

Fáze porozumění datům začíná prvotním sběrem dat. Následují činnosti které umožní získat základní představu o datech, která jsou k dispozici (posouzení kvality dat, první „vhled“ do dat, vytipování zajímavých podmnožin záznamů v databázi…). Obvykle se zjišťují různé deskriptivní charakteristiky dat (četnosti hodnot různých atributů, průměrné hodnoty, minima, maxima apod.), s výhodou se využívají i různé vizualizační techniky [1].

Data musí popisovat daný problém, nesmí být chybná a zavádějící, musí mít nějakou informační hodnotu a je dobré, když je dat dostatek.

Vstupní data popisující ojetá vozidla (Tabulka 1) se týkají výhradně českého výrobce automobilů značky Škoda.

Tabulka 1: Prvotní datový slovník (zdroj: vlastní)

par.	jméno	typ proměnné	hodnoty, příklady	popisek
p1	typ	více kategoriální	„Octavia, 120“	názvy vozů
p2	obsah	intervalová	„987,2800“	objem válců v cm^3
p3	rok výroby	intervalová	„1968,2007“	rok výroby vozidla
p4	výkon	intervalová	„33,147“	výkon motoru v kW
p5	stav tachometru	intervalová	„150,842134“	najeté km
p6	prodejní cena	intervalová	„2900,729000“	nabízená cena v bazarech
p7	barva	více kategoriální	„bílá, červená..“	barva vozidla
p8	palivo	kategoriální	„benzín, nafta..“	druh paliva
p9	druh vozidla	kategoriální	„osobní, užit...“	použití vozu
p10	typ karoserie	kategoriální	„sedan, kupé...“	typ karoserie

p11	kraj	více kategoriální	„pardubický,...“	místo kde je auto prodáváno
p12	počet dveří	ordinální	„2,5“	
p13	počet míst	ordinální	„2,5“	počet sedadel ve vozidle
p14	výbava	vícekategoriální	„airbag, ABS, CD..	výčet výbavy ve vozidle

Po prvotním náhledu na data bylo zjištěno, že chybí parametry jako je například síla koroze, hlučnost motoru, spotřeba pohonných hmot, spotřeba oleje, velikost zavazadlového prostoru a další, které podstatnou měrou určují cenu ojetého vozidla. Dle shledaných nedostatků musela být některá data odstraněna.

5. PŘEDZPRACOVÁNÍ DAT

Předzpracování dat zahrnuje činnosti, jež vedou k vytvoření datového souboru, který bude zpracováván jednotlivými analytickými metodami. Data by tedy měla obsahovat údaje relevantní k dané úloze, a mít podobu, která je vyžadována vlastními analytickými algoritmy. Příprava dat tedy zahrnuje selekci dat, čištění dat, transformaci dat, vytváření dat, integrování dat a formátování dat. Tato fáze je obvykle nejpracnější částí řešení celé úlohy. Jednotlivé úkony jsou obvykle prováděny opakovaně, v nejrůznějším pořadí[1].

V dataminigových projektech se až 80 % času alokuje na přípravu dat. Zároveň je to patrně nejproblémovější fáze celého cyklu. V přípravě dat se střetává realita pochopení problému a jeho relevantnost vzhledem k dostupným datům. Dobrým předpokladem data miningu je kvalitní datový sklad, který integruje a agreguje všechny dostupné informace[13].

Jednotlivé kroky procesu dobývání znalostí z dat jsou různě časově náročné a mají i různou důležitost pro úspěšné vyřešení dané úlohy. Praktici v oboru uvádějí, že nejdůležitější je fáze porozumění problému a časově nejnáročnější je fáze přípravy dat [1].

Prvotní porozumění datům bylo provedeno v MS Excel. Vstupní data byla velice různorodá, byly zjištěny chyby a nesmyslné údaje. Nedostatky byly způsobeny zřejmě tím, že data nezapisuje jeden zadavatel a že jednotliví zadavatelé dat používají jiný způsob zapisování a jsou různě pečliví. Z tohoto důvodu bylo provedeno čištění dat.

5.1 Čistění dat

Četnosti hodnot různých atributů, průměrné hodnoty, minima, maxima apod. provedené v MS Excel na vstupních datech ukázaly, že data nelze k našemu modelování použít anebo je lze použít pouze ve velmi omezené míře. Proto bylo přistoupeno k předzpracování dat:

- vstupní data charakteru typ automobilu byla vícekategoriální, tudíž bylo nutné vozidla téhož typu pojmenovat stejně. Záznamy byly abecedně seřazeny. Na základě podobnosti názvu typu vozu a shodnosti zdvihového objemu a výkonu, byly tyto podobné záznamy pojmenovány stejně.
- vstupní data typu obsah byly ošetřeny v závislosti na četnosti obsahů jednotlivých typů vozidel. Tzn. nejvíce opakovaná hodnota byla vzata jako

vzor a použita k nastavení vstupních dat typu obsah. Na základě této operace bylo mnoho záznamů doplněno, popřípadě upraveno,

- vstupní data typu rok výroby byla ošetřena pouze funkcí maxima a minima. Pokud nějaká hodnota dosahovala nesmyslných mezí, byla následně upravena. Několikrát se u těchto dat vyskytlo překlepnutí zadavatele, a tak byl velice často zaměňován rok 1998 s 1989,
- vstupní data typu výkon byla upravována na základě zdvihového objemu, roku výroby, typu vozidla a pohonných hmot,
- vstupní data stav tachometru byly upraveny podobně jako typ rok výroby, s rozdílem, že u těchto dat nebyla žádná možnost, jak zjistit, zda jsou data opravdu zadána správně,
- u vstupních dat typu prodejní cena byly odstraněny záznamy, kdy prodejní cena nebyla vůbec zadaná. Pro další zpracování tyto záznamy představovaly problém,
- u vstupních dat typu barva byly opravovány hlavně pravopisné chyby a překlepy,
- vstupní data typu palivo (benzín, nafta a LPG) byly upraveny tak, že pokud tento záznam chyběl, byl dohledán v závislosti na zdvihovém objemu nebo výkonu,
- vstupní data typu druh vozidla byly pouze dvě hodnoty vůz užitkový a osobní. Mnoho typických osobních automobilů mělo jako druh vozidla zaznamenáno užitkový automobil. Příčina je obvykle na straně prodejce, který může specifikovat vůz jako užitkový, pokud je automobil vybaven mřížkou mezi zavazadlovým prostorem a prostorem pro cestující,
- vstupní data typu karoserie byla doplněna o chybějící záznamy v závislosti na názvu typu vozidla,
- vstupní data typu kraj, zde byla provedena úprava sjednocení názvů na konečných 14 krajů,
- vstupní data typu počet dveří byla upravena tak, aby uvedený počet byl od dvou do pěti,

- vstupní data typu počet míst byla podobně upravena jako data typu počet dveří,
- vstupní data výbava byla nepoužitelná pro výpočet, a proto z těchto dat bylo potřeba odvodit nové proměnné.

5.2 Odvození nových proměnných

Velký problém byl se vstupními daty výbava, jelikož někteří autoprodejci do výbavy vozu nepoznamenali nic, jiní prodejci popsali výbavu vozidla velice podrobně. Následně byl vytvořen dotaz a vypsány dichotomické proměnné např. ABS, Airbag, ESP atd. V tabulce 2 jsou uvedeny odvozené proměnné.

Tabulka 2: Odvozené vstupní proměnné (zdroj: vlastní)

p14	airbag	ditochomická	„0,1“	
p15	ABS	ditochomická	„0,1“	antiblokovací systém
p16	klimatizace	ditochomická	„0,1“	
p17	centrál dálkový	ditochomická	„0,1“	centrální zamykání na DO
p18	CD	ditochomická	„0,1“	rádio na CD
p19	imobilizér	ditochomická	„0,1“	zabezpečení proti odcizení
p20	hliníková kola	ditochomická	„0,1“	
p21	palubní počítač	ditochomická	„0,1“	
p22	ASR	ditochomická	„0,1“	Sys. regulace prokluzu kol
p23	ESP	ditochomická	„0,1“	Elektronický stabilizační program
p24	tažné zařízení	ditochomická	„0,1“	možnost přípojného vozidla
p25	střešní lyžiny	ditochomická	„0,1“	možnost nákladu na střeše

5.3 Korelační analýza

Cílem zkoumání je číselné charakterizování závislosti. Každá korelační závislost kvantitativních znaků je předmětem měření a analýz.

Těsnost (síla, intenzita) závislosti, tj. míra vzájemného vztahu mezi proměnnými. To je důležité jednak pro nalezení resp. rozlišení nejdůležitějších faktorů, jednak pro posouzení kvality vystižení průběhu závislosti. Toto hledisko zkoumání je označováno jako korelační analýza [3].

Analýza nám ukázala, že některé parametry jako je např. počet dveří nebo počet míst k sezení nejsou s parametrem prodejní cenou téměř korelovány. Tudíž tyto vstupní parametry můžeme vynechat, aniž by se změnila kvalita výpočtů.

Korelační analýza vypočítala, že většina spojitých vstupních parametrů koreluje s výstupním parametrem prodejní ceny ve větší než zanedbatelné síle. V zanedbatelné síle koreluje s prodejní cenou vstupní parametr počet dveří a počet míst (tabulka 3). Proto se s těmito dvěma vstupními parametry dále nebude počítat. Korelační analýza byla prováděna v MS Excel (funkce CORREL) a následně v prostředí Clementine 9.0 (uzel Statistic).

Tabulka 3: Síla korelace vstupních parametrů s prodejní cenou (zdroj: vlastní)

Typ vstupních dat	Síla korelace
obsah	0,5757
výkon	0,6942
rok Výroby	0,7666
stav Tachometru	-0,4121
počet Dveří	0,0264
počet Míst	-0,0009
airbag	0,6273
ABS	0,6626
klimatizace	0,6222
centrál dálkový	0,6078
CD	0,5402
imobilizér	0,4677
hliníková kola	0,4607
palubní počítač	0,7009
ASR	0,6339
ESP	0,5347
tažné zařízení	-0,1556
střešní nosič	0,1690

6. POUŽITÉ MODELOVACÍ TECHNIKY

Důležitou vlastností živých organismů je schopnost přizpůsobovat se měnícím se podmínkám, eventuálně se učit na základě vlastních zkušeností. Schopnost se učit je považována za definici inteligence. Cílem umělé inteligence je tedy vybavit touto schopností technické systémy. Strojové učení může tedy být realizováno dvěma způsoby a to učením bez učitele a učením s učitelem.

Učení s učitelem se týká učení, ke kterému máme informaci o tom, do jaké třídy je příklad zařazen, které se má systém naučit. Učitel poskytuje systému explicitní informaci o požadovaném chování. Jinými slovy se dá říci, že trénovací množina obsahuje hodnotící vektor. Strojové učení bez učitele užívá systém, pokud postrádá doplňkovou informaci.

Obvykle existuje řada různých metod pro řešení dané úlohy, je tedy třeba vybrat ty nejvhodnější (doporučuje se použít více různých metod a jejich výsledky kombinovat) a vhodně nastavit jejich parametry. Jde tedy opět o iterativní činnost (opakovaná aplikace algoritmů s různými parametry). Použití analytických algoritmů může navíc vést k potřebě modifikovat data a tedy k návratu k datovým transformacím z předcházející fáze [1].

Součástí této fáze je rovněž ověřování nalezených znalostí z pohledu metod dobývání znalostí. To může představovat např. testování klasifikačních znalostí na nezávislých datech.

Protože byla k dispozici data včetně ohodnocovacího vektoru (prodejní cena), bude užito metod vykazující učení s učitelem. Budou použity neinferenční techniky, rozhodovací stromy a neuronové sítě. Bylo rozhodnuto pro tyto techniky, protože v práci jsou modelována data a nemáme předpoklady o rozdělení dat, dále byly tyto techniky vybrány proto, že vstupní data jsou spojitá, ale i kategorizovaná. Přitom by měla být výstupní veličina spojitá (odhadovaná prodejní cena).

6.1 Rozhodovací stromy

Rozhodovací stromy (RS) jsou analytické nástroje sloužící k nalezení pravidel a vztahů v datovém souboru pomocí systematického rozdělování a větvení na nižší úrovně. Rozhodovací stromy jsou také prediktivní modely, které zobrazují data v podobě stromu, každý uzel určuje kritérium pro následné rozdělení dat do jednotlivých větví. Strom se skládá z uzlů. Uzel na nejvyšší úrovni je označován pojmem kořenový. Vnitřní uzle přestavují testy

jednotlivých atributů (kořenový uzel je rovněž testem). Větví nazýváme možný výsledek testu. Externí uzly označované jako listy reprezentují jednotlivé třídy [10].

Strom tak rozděluje veškerá zdrojová data do segmentů, kde každý list odpovídá určitému segmentu definovanému předchozími uzly. Data, která jsou zařazena do určitého segmentu se vyznačují shodnými vlastnostmi [2].

Rozhodovací stromy jsou vhodné pro úlohy, ve kterých má být provedena klasifikace nebo předpověď. Užitečné jsou v oblastech, ve kterých můžeme hodnoty proměnných rozdělit do relativně malého počtu skupin. Na druhou stranu nejsou vhodné pro případy, kdy je úkolem předpovězení kvantitativních hodnot.

6.1.1. Algoritmus TDIDT

Algoritmus TDIDT (top down induction of decision trees) je jeden z hlavních algoritmů rozkladu dat. Na počátku tvoří celá trénovací data jednu množinu, na konci máme podmnožiny tvořené příklady téže třídy.

1. zvol jeden atribut jako kořen dílčího stromu,
2. rozděl data v tomto uzlu na podmnožiny podle hodnot zvoleného atributu a přidej uzel pro každou podmnožinu,
3. existuje-li uzel, pro který nepatří všechna data do téže třídy, pro tento uzel opakuj postup od bodu 1, jinak skonči [1].

6.1.2. Regresní stromy

Výše byly popsány stromy pro klasifikaci objektů do tříd. Takovým stromům se obvykle říká klasifikační stromy. Existují ale i stromy regresní, které umožňují odhadovat hodnotu nějakého numerického atributu. V listových uzlech mají takové stromy místo názvu třídy například konkrétní hodnotu, která odpovídá průměrné hodnotě cílového atributu.

Algoritmus pro tvorbu regresního stromu odpovídá algoritmu TDIDT. Rozdíl je ve způsobu volby atributu pro větvení. Místo entropie se vychází ze směrodatné odchylky hodnot cílového atributu. Větvení skončí, pokud se hodnota cílového atributu pro příklady v uvažovaném uzlu jen málo liší, nebo pokud je v uvažovaném uzlu jen málo příkladů [1].

6.2 Neuronové sítě

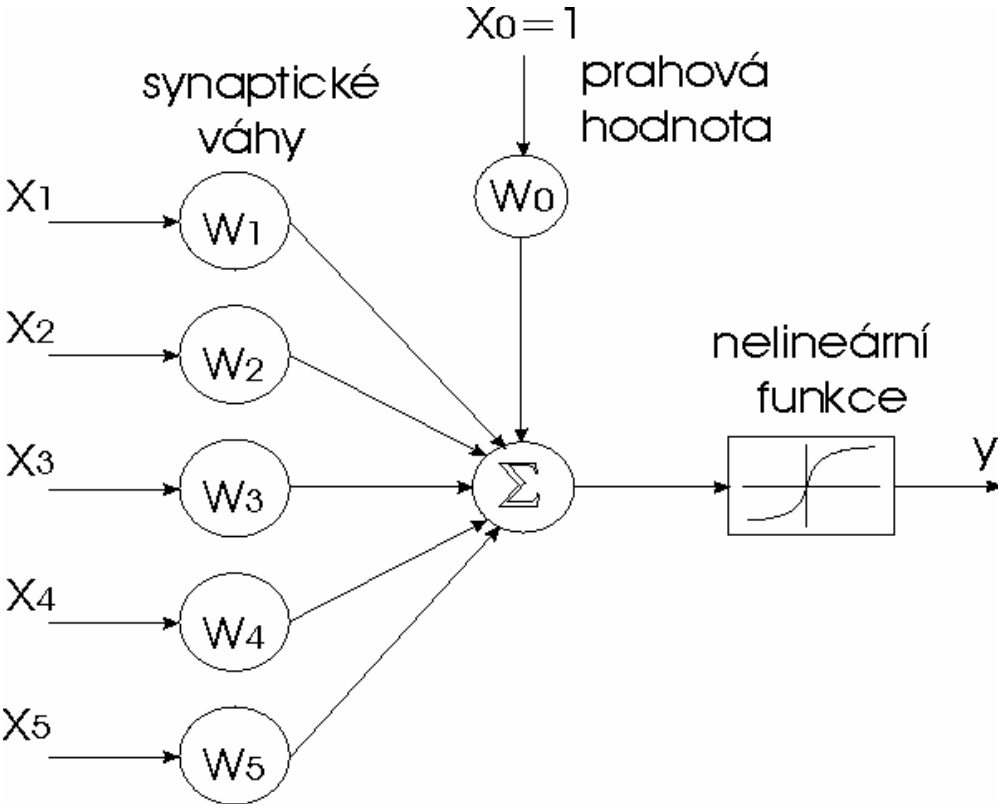
Neuronové sítě (NS) jsou využívány pro tvorbu prediktivních modelů. Jsou založeny na principech napodobujících organizaci nebo chování lidského mozku.

Jednoduchá neuronová síť, ať už jednovrstvá nebo vícevrstvá, je schopná korektně řešit jen omezenou třídu problémů. Lineární neuronové sítě mohou být jednoduše zevšeobecněné tak, že aktivity výstupních neuronů jsou určené pomocí nelineární přechodové funkce (která v nejjednodušším případě odpovídá tzv. tvrdé nelinearitě, např. skokové nebo znaménkové funkci). Avšak takovéto zevšeobecnění lineární neuronové sítě je též schopné klasifikovat jen lineárně separovatelné problémy. Toto omezení bylo považováno za vážný nedostatek neuronových sítí. Teoreticky se uvažovalo o možnosti zavedení dalších (skrytých) vrstev nelineárních neuronů do sítí. Bohužel, nebylo jasné, jak adaptovat váhové koeficienty, které jsou přiřazené neuronům ze skryté vrstvy. Až byl navrhnut jednoduchý gradientní algoritmus (nazvaný metoda zpětného šíření – angl. Back-propagation) adaptace vícevrstvých neuronových sítí s dopředným šířením. Tímto se vícevrstvé neuronové sítě staly velmi populární a patří mezi univerzální přístupy teorie neuronových sítí se širokou paletou aplikací v různých oblastech informatiky a přírodních věd. Navíc bylo dokázáno, že neuronové sítě tohoto typu jsou univerzálním aproximátorem, tj. jsou schopné aproximovat s požadovanou přesností libovolnou spojitou funkci.

Umělé neuronové sítě byly vytvořeny na základě jednoduchých modelů neuronů - funkčních buněk nervového systému živých organizmů. Většina současných aplikací umělých neuronových sítí využívá selektivní a generalizační vlastnosti těchto struktur. Některé novější struktury jsou navíc schopné řešit i úlohy složitějšího typu, jako jsou např. optimalizační úlohy[12].

6.2.1. Matematický model neuronové sítě

První matematický model neuronu vytvořili McCulloch a Pitts v roce 1943 a tento model se dodnes používá pro běžné aplikace. Tento matematický model se skládá ze tří hlavních částí. Obsahuje vstupní, výstupní a funkční část. Vstupní část se skládá ze vstupů a z přirazených, nastavitelných vah (synaptické váhy). Na základě váhových koeficientů mohou být jednotlivé vstupy zvýhodňovány či potlačeny. Následující částí je výkonná jednotka, která zpracuje informace ze vstupu a vygeneruje výstupní odezvu. Třetí část je výstupní jednotka, která přivádí výstupní informace na vstup jiných neuronů. Z toho je patrná podoba mezi klasickými výpočetními systémy a umělými neurony. Oba systémy obsahují vstupní část, paměť, výkonnou jednotku a výstupní část. Velké rozdíly jsou ovšem v uspořádání těchto částí. Paměť umělého neuronu není samostatná jednotka, ale je rozprostřená ve vstupní části formou váhových koeficientů. Pomocí těchto koeficientů je systém schopný zapamatovat si informace. Jak je vidět na obrázku obr.1, výkonná jednotka umělého neuronu je mnohem jednodušší než výkonná jednotka výpočetních systémů a je tvořena jednoduchou nelineární funkcí [12].



Obr.2: Jednoduchý model neuronu (zdroj: [12])

Z obrázku obr. 2 je jasná funkce jednoho neuronu. Vstupní hodnoty jsou vynásobeny příslušnými váhovými koeficienty a sečtou se. Na výsledek součtu se aplikuje funkce (obecně

nelineární) a výsledná hodnota funkce je přivedena na vstup jiných neuronů pomocí výstupní části. Na obr. 1 je navíc vidět, že neuron má jeden zvláštní vstup, který není připojený k výstupu žádného neuronu, ale přivádí konstantní veličinu do neuronu. Tato veličina funguje jako prahová hodnota při aktivování výstupu. Když suma váženého součtu vstupů nepřesahuje prahovou hodnotu, tak se neuron neaktivuje a jeho výstup zůstane nezměněný.

6.2.2. Matematický popis neuronu

Matematicky lze funkci neuronu popsat následovně:

$$y = F\left(\sum_{i=1}^{n} x_i w_i + Q\right) \tag{1}$$

kde:
xi - je hodnota na i-tém vstupu,
wi - je váha i-tého vstupu,
Q - je prahová hodnota,
n - je celkový počet vstupů,
F - je obecná nelineární funkce,
y - je hodnota výstupu [12].

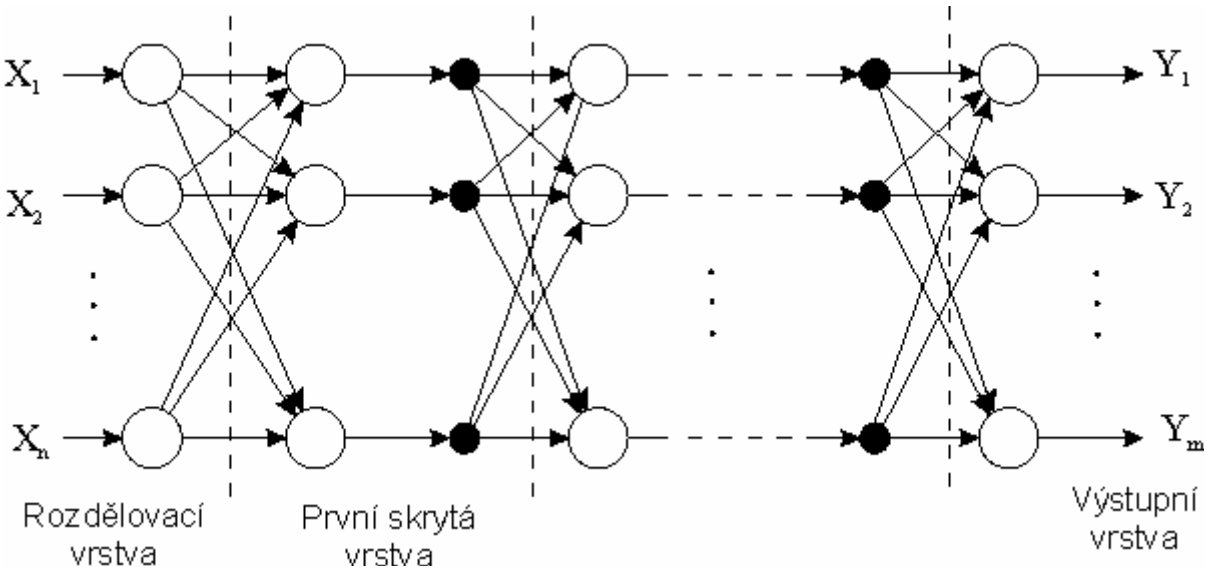
6.2.3. Umělá neuronová síť

Je asi zřejmé, že jediný neuron není schopen vykonat příliš složitou funkci. Síla systému, využívající umělé neurony, je ve struktuře, v síti velkého počtu neuronů. Umělá neuronová síť je pole jednoduchých výkonných prvků - neuronů. Takovéto uspořádání má velkou flexibilitu a spolehlivost . Umožňuje různě propojovat vstupy a výstupy neuronů, zvýhodnit či potlačit některé vstupy a minimalizovat vliv nesprávně fungujícího neuronu na celkový výsledek.

Samozřejmě i tento systém má nevýhody. Největší problémy se vyskytují při realizaci velmi složitých struktur, kde velký počet propojení mezi neurony se realizuje velmi obtížně. Dalším problémem je, že neexistuje jednoznačný postup při syntéze složitějších struktur.

6.2.4. Vrstvová struktura umělé neuronové sítě

Neurony jsou většinou sdružovány do vrstev, jak to ukazuje obr. 3. Výstupy z n-té vrstvy jsou přivedeny na vstup obecně každého neuronu ve vrstvě n + 1. První vrstva se nazývá vstupní či rozdělovací vrstva a má za úkol přijímat hodnoty z okolí pro zpracování a přivést je na vstup každého neuronu následující vrstvy. Poslední vrstva nese název výstupní a hodnoty na jejím výstupu jsou odezvou celého systému na vstupní vzorky. Vnitřní vrstvy se nazývají skryté vrstvy. Jejich počet závisí na složitosti funkce, kterou má síť vykonat a na zvoleném typu sítě.

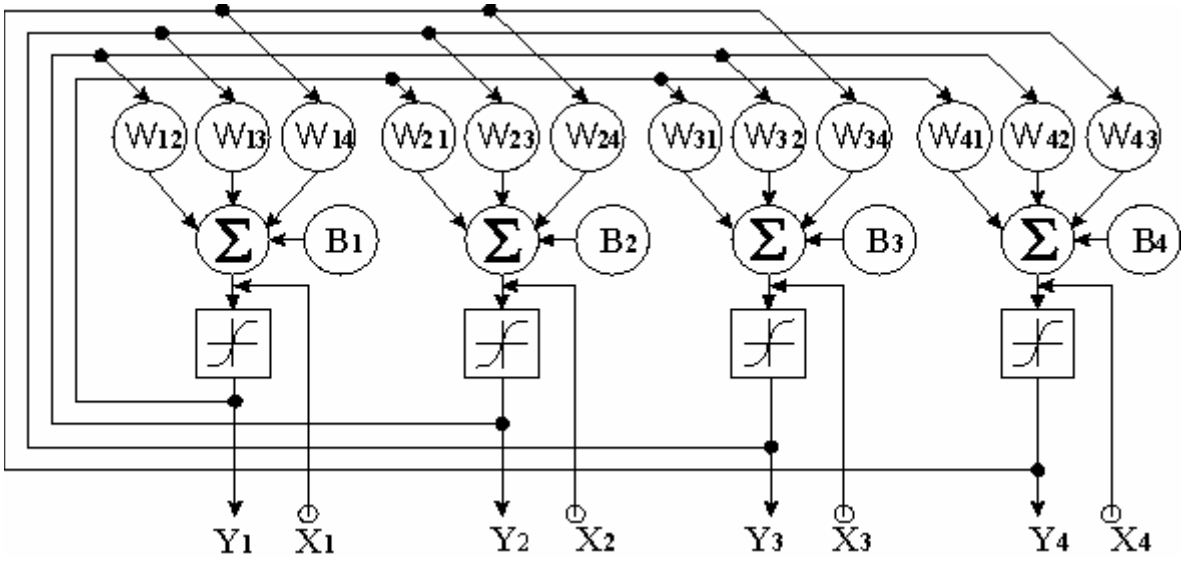


Obr.3: Vrstvová struktura umělé neuronové sítě (zdroj: [12])

6.2.5. Typy neuronových sítí

Neuronové sítě lze rozdělit do dvou hlavních skupin podle struktury: na sítě s dopředním šířením signálu a na sítě se zpětnou vazbou. V současnosti se nejčasněji používají struktury s dopředním šířením signálu, kde výstupy z jedné vrstvy jsou vedeny na vstup následující vrstvy, jak to ukazuje obr.3. Výstupy z poslední, výstupní vrstvy jsou výstupy z celé sítě.

Struktura sítí se zpětnou vazbou se liší od předchozích v tom, že výstupy z vrstvy jsou vedeny zpět na vstup dané vrstvy. Taková struktura umožňuje realizovat výpočty založené na iteračním procesu a tak řešit např. optimalizační úlohy. Příklad struktury takové sítě je na obr. 4, který znázorňuje Hopfieldovu síť se čtyřmi neurony [12].



Obr.4: Hopfieldova síť (zdroj [12])

Neuronové sítě s dopředním šířením signálu lze rozdělit do dvou skupin podle funkce kterou realizují, a to na lineární a nelineární. Tato funkce samozřejmě není totožná s výstupní funkcí jednoho neuronu. Sítě lineární jsou schopné realizovat lineární matematické funkce, tj. funkce skládající se ze součtů a z násobení.

Charakteristickou vlastností nelineárních neuronových sítí s dopředním šířením signálu je schopnost učení. Fáze učení předchází fázi vlastní práce a slouží k určení váhových koeficientů a tak vlastně k uložení informací do paměti systému. Učení se může probíhat dvěma způsoby, s učitelem a bez učitele. Při prvním způsobu je síť trénována pomocí dvojic vstupní vzorek a příslušný, očekávaný výstupní vzorek. Trénovací vstupní vzorky jsou vybrány z celkové množiny vstupních vzorků tak, aby plně popsaly všechny vlastnosti množiny důležité pro danou úlohu. V této fázi nenatrénované síti přiložíme vstupní vzorek. Na základě skutečné odezvy a očekávané odezvy se upravují váhové koeficienty. Během trénování se na vstupy sítě přivedou všechny trénovací vzorky, obecně vícekrát a navíc v náhodném pořadí. Po natrénování síť musí správně reagovat na všechny trénovací vzorky a dále má pracovat dobře i pro ostatní vzorky množiny. Aby síť pracovala dobře potřebujeme velký počet trénovacích vzorků. Obecně platí, že čím větší je počet trénovacích vzorků, tím přesněji bude síť pracovat. Příkladem takové sítě je síť "back-propagation", která je pravděpodobně nejčastěji používaným typem.

Při učení bez učitele máme jenom trénovací vzorky, ale neexistují očekávané výstupní vzorky. Tyto výstupní vzorky, příslušející k jednotlivým vstupním vzorkům se určí během procesu učení. Váhové koeficienty se postupně nakonfigurují tak, aby pro každý vstupní

trénovací vzorek existoval jediný aktivní výstup. Tak na konci trénování dosáhneme toho, že přivedením trénovacího vzorku se aktivuje vždy jediný, jednoznačně určený výstup[8].

6.3 Definice neuronové sítě

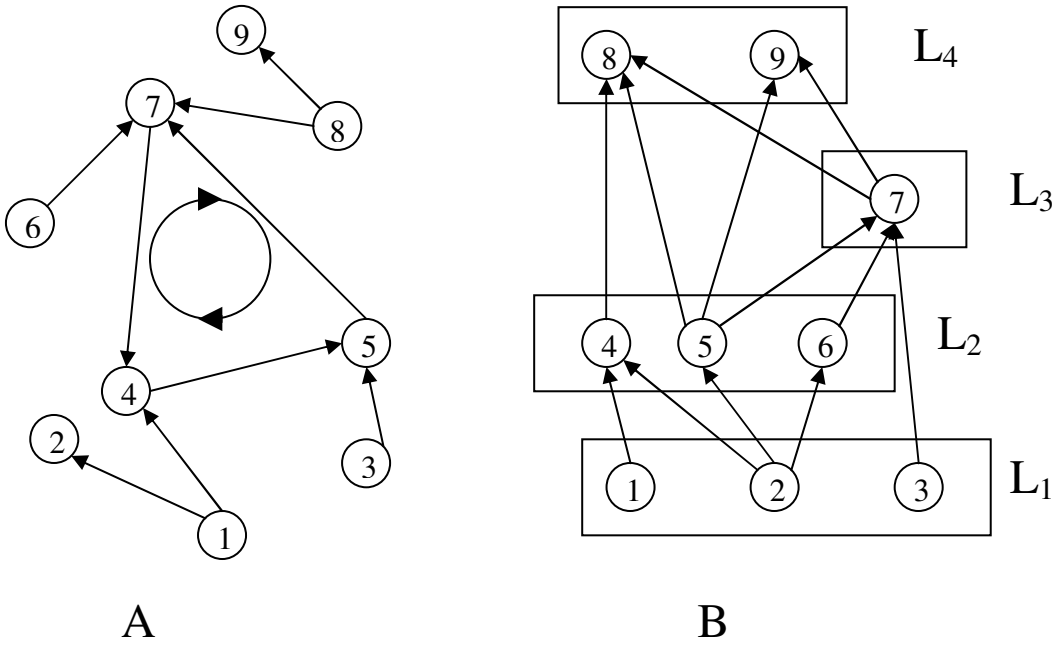
Vývoj neuronové sítě bude formulováno pomocí grafově-teoretického přístupu. Přitom se vychází z analogie s lidským mozkem a koncept neuronové sítě bude použitý na konstrukci modelové funkce $G(x,w)$. Formálně je neuronová síť určená jako orientovaný graf $G = (V,E)$. Výrazy $V = \{v_1,v_2, \dots ,v_N\}$ a $E = \{e_1,e_2, \dots ,e_M\}$ označují neprázdnou vrcholovou množinu, resp. hranovou množinu grafu G obsahujícího N vrcholů (neuronů) a M hran (synapsí). Každý spoj $e \in E$ se interpretuje jako uspořádaná dvojice dvou neuronů z množiny V, $e = (v,v')$. Spoj e začíná v neuronu v a končí v neuronu v'. Množina V je rozložená na disjunktní podmnožiny následujícím způsobem

$$V = V_I \cup V_H \cup V_O \tag{2}$$

kde:

V - je neprázdná vrcholová množina,

kde V_I obsahuje N_I vstupních neuronů, které sousedí jen s vycházejícími hranami V_H, obsahuje N_H skrytých neuronů, které sousedí současně s vycházejícími jako s vcházejícími hranami a V_O obsahuje N_O výstupních neuronů, které sousedí jen s vcházejícími hranami. V následujících úvahách se bude vždy předpokládat, že množiny V_I a V_O jsou neprázdné, tj. neuronová síť obsahuje vždy alespoň jeden vstupní a jeden výstupní neuron. Pro acyklické neuronové sítě (které neobsahují orientované cykly (obr. 5 A) dále mohou být neurony uspořádány do vrstev (obr. 5 B).



Obr.5: Neuronová síť definovaná jako orientovaný souvislý graf (zdroj: [4])

Na obr. 5 A je zobrazen orientovaný graf s jedním cyklem a tedy nemůže být použitý pro definici neuronové sítě s dopředným šířením. Na obr. 5 B je znázorněna možnost rozkladu vrcholů (neuronů) acyklického orientovaného grafu na vrstvy L_1, … ,L_4.

$$V = L_1 \cup L_2 \cup L_3 \cup ... \cup L_t \tag{3}$$

kde:

V - je neprázdná vrcholová množina,

L – vrstva neuronové sítě

kde $L_1 = V_I$ je *vstupní vrstva* (obsahuje pouze vstupní neurony), L_2,L_3, … ,L_{t-1} jsou *skryté vrstvy* a L_t je *výstupní vrstva*. Vrstva L_i (pro $1 \le i \le t$) je určená následujícím jednoduchým způsobem

$$L_i = \{v \in V; d(v) = i - 1\} \tag{4}$$

kde:

d(v) – vzdálenost rovnající se délce max. délce spojující daný neuron se vstupním neuronem,

L – vrstva neuronové sítě

kde vzdálenost $d(v)$ se rovná délce maximální cesty, která spojuje daný neuron se vstupním neuronem. Potom musí platit $d(v) = 0,\ v \in V_I$. Neuronová síť určená acyklickým grafem je obvykle zvolená tak, že neurony ze dvou sousedících vrstev jsou pospojované všemi možnými spoji. Bohužel, takovýto rozklad množiny neuronů na vrstvy je možný jen pro neuronové sítě reprezentované acyklickými grafy, pro cyklické grafy vzdálenost $d(v)$ může nabývat libovolnou kladnou celočíselnou hodnotu[4].

7. MODELOVÁNÍ V SYSTÉMU CLEMENTINE

V každé fázi data miningového procesu podporuje Clementine standardní metodiku oboru CRISP-DM (CRoss-Industry Standard Process for Data Mining). To znamená, že se analytici mohou plně soustředit na řešení věcných (marketingových, prodejních, personálních, atd.) problémů postupem data miningu a nemusí se v každém projektu zabývat hledáním a definicí nových procesů a novou metodikou postupu. Individuální projekty jsou v Clementine efektivně organizovány pomocí správce projektů podle CRISP-DM [5].

Metodika CRISP-DM pomáhá analytikům data miningu efektivně implementovat data miningové projekty, které končí měřitelnými obchodními výsledky.

K samotnému učení neuronové sítě bylo použito učení s učitelem, přičemž prodejní cena představovala ohodnocovací vektor. Po prvotním převzetí a předzpracování dat se práce ubírá k samotným analýzám.

7.1 Načtení dat

V této fázi již předzpracovaná data z MS Excel ve formátu csv (data oddělená středníkem) byla načtena pomocí uzlu Var. file. Systém Clementine upravil datový typ vstupních dat (tabulka 4). Následná práce se bude odvíjet z tohoto výchozího bodu.

Tabulka 4: Vstupní data do systému Clementine (zdroj: vlastní)

par.	jméno	typ proměnné	hodnoty, příklady	typ v systému Clementine
p1	typ	kategoriální	„Octavia, 120“	množina (set)
p2	obsah	numerická	„987,2800“	rozsah (range)
p3	rok výroby	numerická	„1968,2007“	rozsah (range)
p4	výkon	numerická	„33,147“	rozsah (range)
p5	stav tachometru	numerická	„150,842134“	rozsah (range)
p6	prodejní cena	numerická	„2900,729000“	rozsah (range)

p7	barva	kategoriální	„bílá, červená..“	množina (set)
p8	palivo	kategoriální	„benzín, nafta..“	množina (set)
p9	druh vozidla	kategoriální	„osobní, užit...“	příznak (flag)
p10	typ karoserie	kategoriální	„sedan, kupé...“	množina (set)
p11	kraj	kategoriální	„pardubický,...“	množina (set)
p12	počet dveří	numerická	„2,5“	příznak (flag)
p13	počet míst	numerická	„2,5“	příznak (flag)
p14	airbag	numerická	„0,1“	příznak (flag)
p15	ABS	numerická	„0,1“	příznak (flag)
p16	klimatizace	numerická	„0,1“	příznak (flag)
p17	centrál dálkový	numerická	„0,1“	příznak (flag)
p18	CD	numerická	„0,1“	příznak (flag)
p19	imobilizér	numerická	„0,1“	příznak (flag)
p20	hliníková kola	numerická	„0,1“	příznak (flag)
p21	palubní počítač	numerická	„0,1“	příznak (flag)
p22	ASR	numerická	„0,1“	příznak (flag)
p23	ESP	numerická	„0,1“	příznak (flag)
p24	tažné zařízení	numerická	„0,1“	příznak (flag)
p25	střešní lyžiny	numerická	„0,1“	příznak (flag)

Po načtení dat byla provedena vstupní analýza pomocí uzlu statistics včetně výše popsané korelace. Na základě výsledků korelační analýzy byly vstupy o slabé korelační síle s prodejní cenou odfiltrovány pomocí uzlu filter. Jednalo se o vstupní proměnné „počet Dveří“ a „počet Míst“

7.2 Tvorba trénovací a testovací množiny

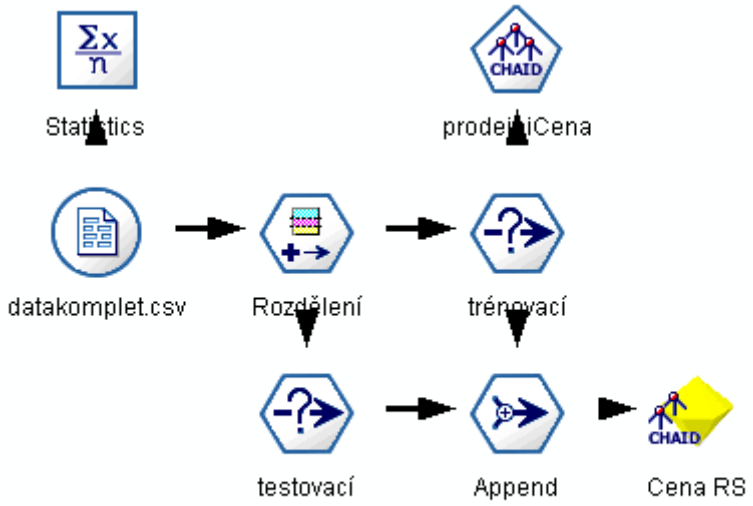
Pod tvorbou trénovaní, resp. testovací množiny chápeme sběr potřebných vstupně/výstupních dat analyzované soustavy a jejich následné uspořádání do tvaru, který je potřebný pro proces trénování, resp. testování. Tento sběr dat může být plánovaný, kdy se na vstup soustavy přivádí definovaná posloupnost hodnot a odečítají se odpovídající hodnoty ustálené výstupní veličiny. Nebo lze snímat přímo „provozní“ data, která pak mají náhodný charakter. Tyto varianty sběru lze mezi trénovací a testovací množinou podle potřeby kombinovat. Přitom, jak je uvedeno [12] a jinde, počet vzorů v testovací množině by se měl rovnat minimálně 1/3 až 1/2 počtu vzorů množiny trénovaní. Rozklad vstupní množiny na trénovací množinu a na množinu testovací byl proveden z důvodu zpětné vazby. Bylo důležité mít možnost porovnat, zda naučený systém pracuje správně a jeho výsledky jsou odpovídající. Rozdělení dat proběhlo v poměru 75 % trénovací množina a 25 % testovací množina viz. tabulka 5.

Tabulka 5: Rozdělení dat na množinu trénovací a testovací (zdroj: vlastní)

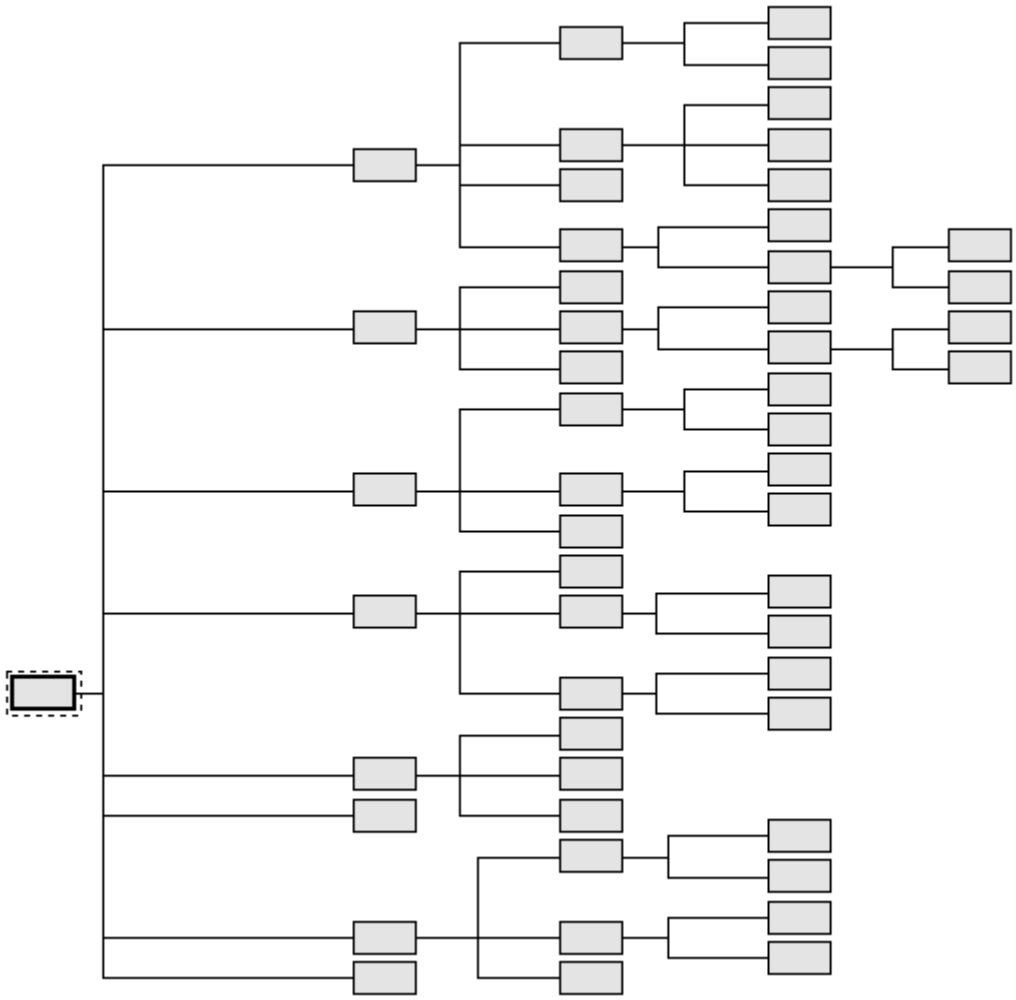
	Rozdělení v %	Počet záznamů
Trénovací	75	2328
Testovací	25	844
Celkem	100	3172

7.3 Rozhodovací regresní stromy

Modelování rozhodovacích stromů v systému Clementine bylo provedeno přes uzel CHAID Tree. Na obr.6 je naznačen stream, kde je znázorněno načtení a rozdělení dat a dále návrh rozhodovacího regresního stromu, který je naučen na trénovacích datech a implementován na všech datech. Strom byl vytvořen načtením spojitých i kategoriálních vstupních proměnných a jako cílová proměnná byl zvolen vstupní parametr prodejní cena.



Obr.6: Stream tvorby modelu rozhodovacího stromu (zdroj: vlastní)

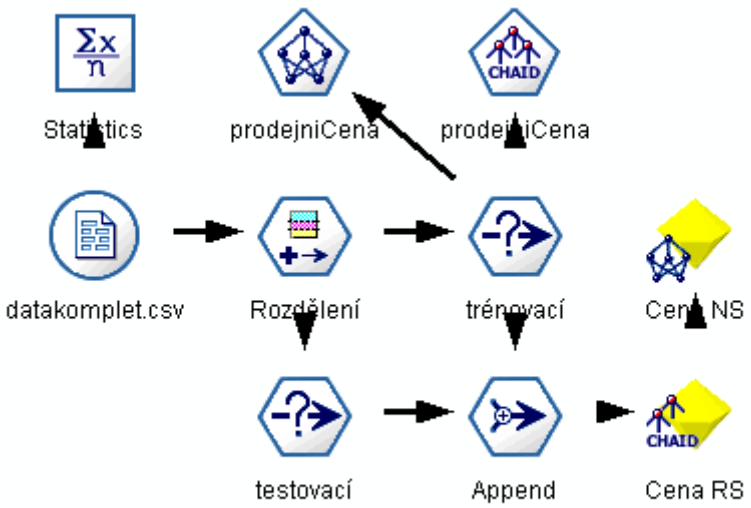


Obr.7: Struktura rozhodovacího stromu (zdroj: vlastní)

Struktura rozhodovacího stromu je znázorněna na obr.7., kde je patrné rozdělení dat na podmnožiny. V každým uzlu je jasně zapsáno pravidlo určující rozdělení na další uzly. Uzel charakterizují údaje o počtu položek, jež splňují podmínky uzlu, o procentuálním podílu těchto položek z celkového počtu, a o průměrné hodnotě položek uzlu. Průměrná hodnota uzlů je předpovídaná prodejní cena. Dá se tedy říci, že čím více by byl strom rozvětvený, tím by podával lepší výsledky.

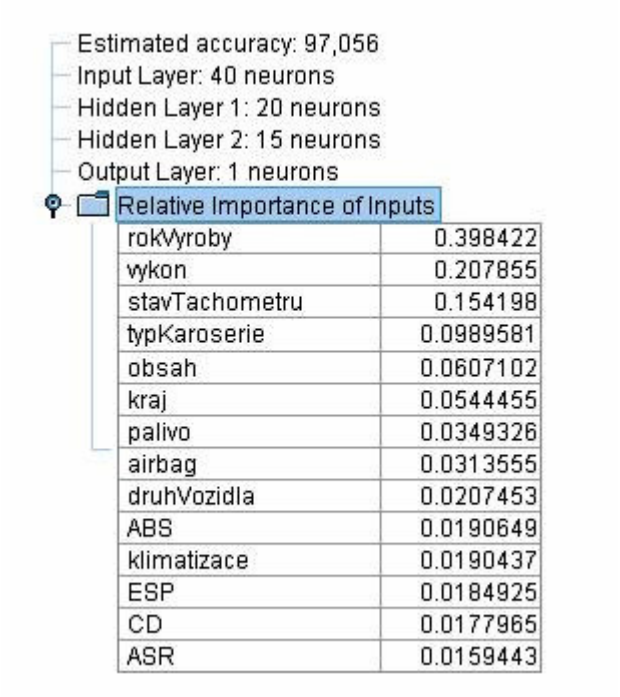
7.4 Neuronové sítě

Podobně jako regresní rozhodovací stromy byl vytvořen model neuronových sítí na trénovacích datech obr.8. Jako nastavení neuronové sítě byla zvolena metoda „Quick“a počet neuronů na skrytých vrstvách, kde bylo nastaveno 20 neuronů na první skryté vrstvě a 15 neuronů na druhé skryté vrstvě. Výstupní vrstva byla pouze jedna, protože je pouze jeden výstup. Neuronová síť je na rozdíl od rozhodovacích stromů pro uživatele taková tzv. „černá skříňka“ to znamená, že uživatel neví jakým způsobem jsou nastaveny váhy na synapsích a jak mají jednotlivé neurony nastavenu hodnotu.



Obr.8: Stream tvorby neuronové sítě (zdroj: vlastní)

Na obr.9 je znázorněna charakteristika vytvořeného modelu neuronové sítě, předpokládaná přesnost modelu počet neuronů na skrytých vrstvách a hodnoty relativních důležitostí vstupních vektorů.



rokVyroby	0.398422
vykon	0.207855
stavTachometru	0.154198
typKaroserie	0.0989581
obsah	0.0607102
kraj	0.0544455
palivo	0.0349326
airbag	0.0313555
druhVozidla	0.0207453
ABS	0.0190649
klimatizace	0.0190437
ESP	0.0184925
CD	0.0177965
ASR	0.0159443

Obr.9: Informace o modelu neuronových sítí (zdroj vlastní)

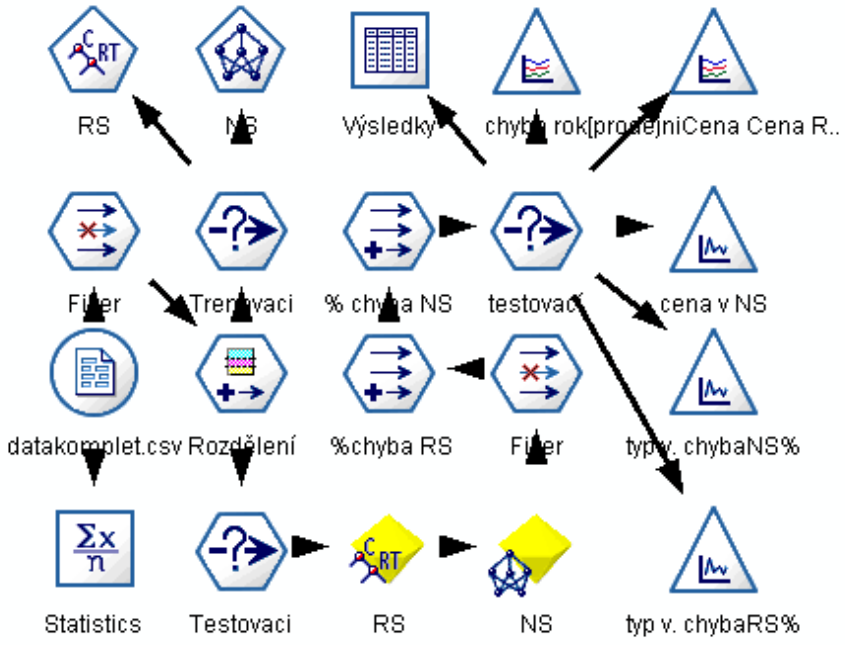
8. HODNOCENÍ

V této fázi jsme se dopracovali do stavu, kdy jsme nalezli znalosti, které se zdají být v pořádku z hlediska metod dobývání znalostí. Dosažené výsledky je ale ještě třeba vyhodnotit z pohledu manažerů, zda byly splněny cíle formulované při zadání úlohy (zdroj: [1]).

8.1.1. Navržené analýzy

Na obr.10 jsou znázorněny různé analýzy, které byly na datech provedeny. Byly to tyto:

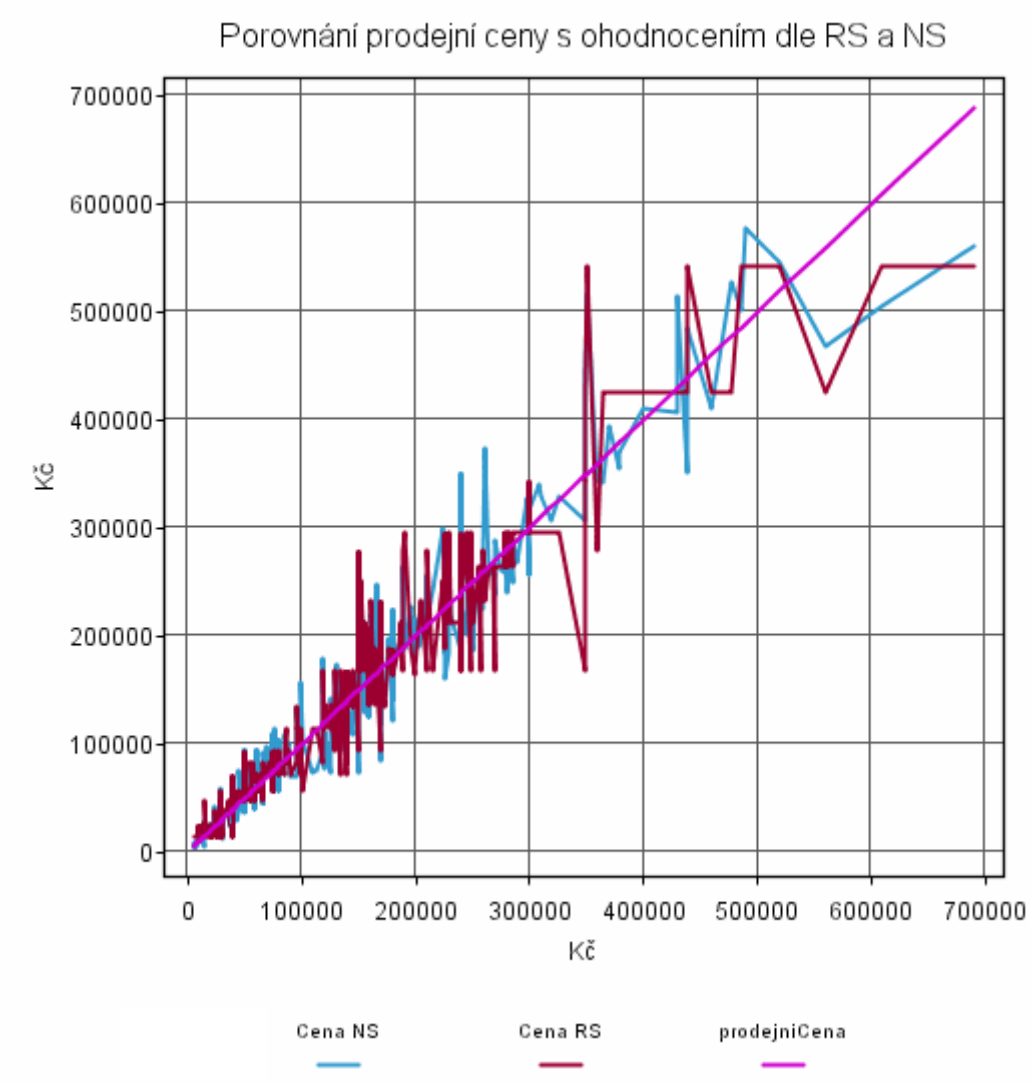
- porovnání prodejní ceny s odhadnutou cenou podle NS a RS
- procentuální chyba NS a RS při odhadu prodejní ceny v závislosti na roku výroby vozidla
- porovnání prodejní ceny a předpovězené ceny podle RS
- porovnání prodejní ceny a předpovězené ceny podle NS
- procentuální chyba RS v závislosti na typu automobilu
- procentuální chyba NS v závislosti na typu automobilu



Obr.10: Navržené analýzy (zdroj: vlastní)

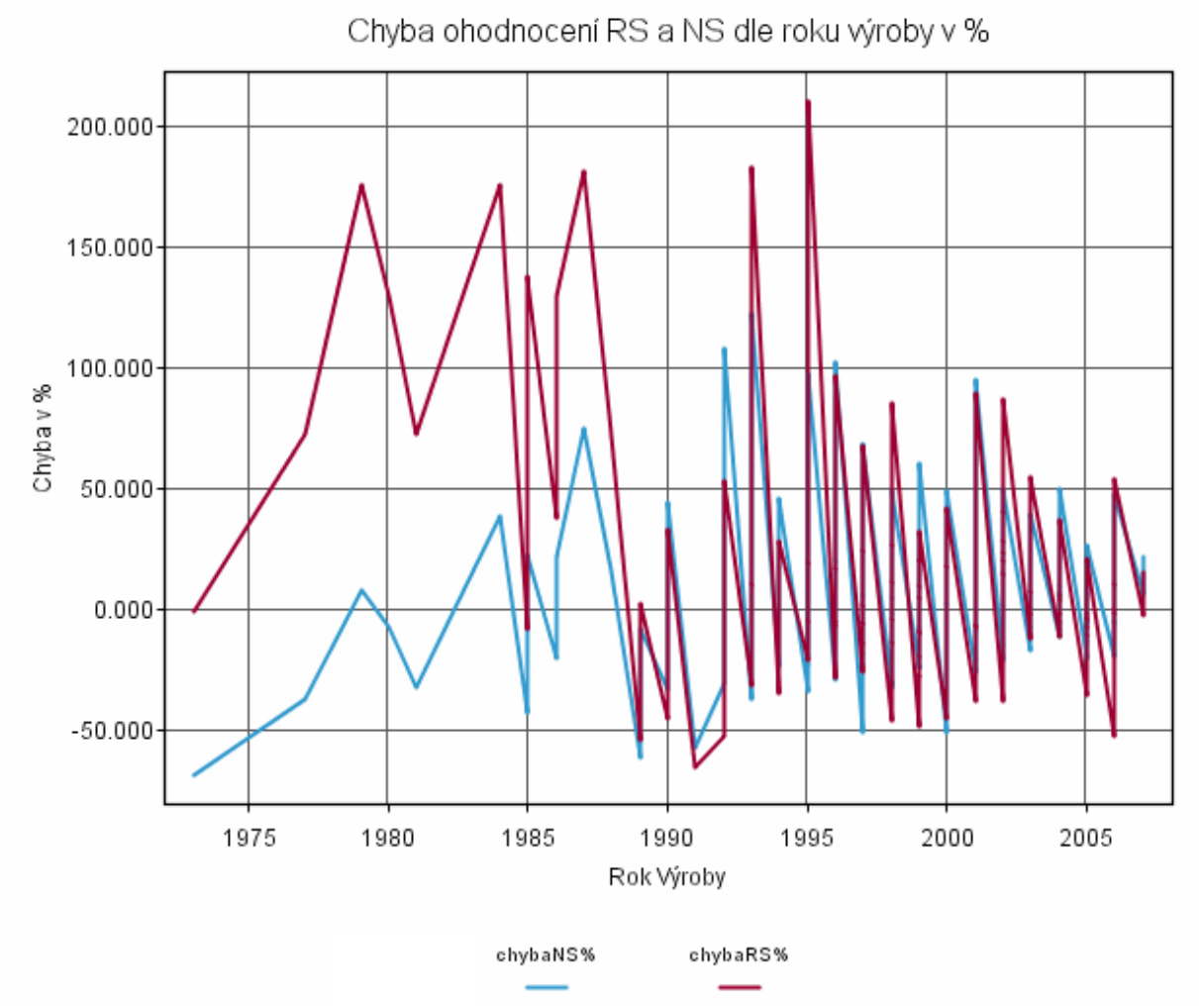
8.1.2. Výsledky analýz

Z výsledků vyplývá, že neuronové sítě a rozhodovací stromy jsou možným nástrojem k ohodnocení vozidla. Na grafech je znázorněno, jakým způsobem si jednotlivé metody vedly při ohodnocování. Na grafu 1. je porovnání odhadované ceny pomocí neuronových sítí a rozhodovacích stromů. V ideálním případě by tyto křivky kopírovali linii odpovídající prodejní ceně. Křivky jak RS (rozhodovací stromy)i NS (neuronové sítě) podávají velice dobré výsledky od prodejní ceny od 0 až po 300000 Kč. Od 300000 Kč zhruba do 400000 Kč se lépe chová křivka neuronových sítí.



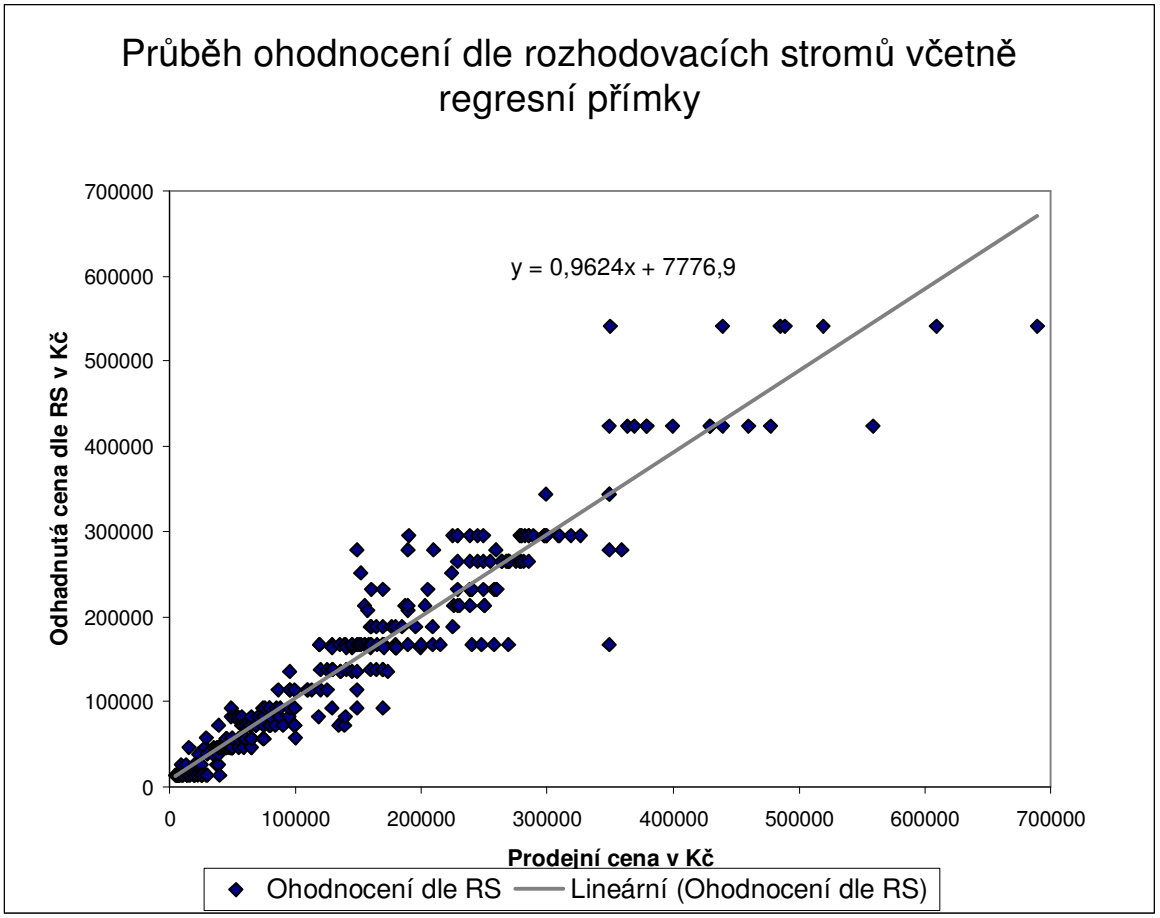
Graf 1: Porovnání prodejní ceny s cenou odhadnutou pomocí neuronových sítí a rozhodovacích stromů (zdroj: vlastní)

Analýza procentuální chyby NS a RS v závislosti na roku výroby vozidla (Graf.2). Vykazuje také velice zajímavé výsledky. Ukazuje, jak velké výkyvy při odhadu ceny ojetého vozidla vznikly. V případě starších automobilů je to v celku logické, protože starší vozidla mohou být jak „vraky“ tak „krásní veteráni“. Proto nelze starší automobily klasifikovat podle technických vlastností, které byly k dispozici v této práci, ale spíše dle vizuálního smyslu.



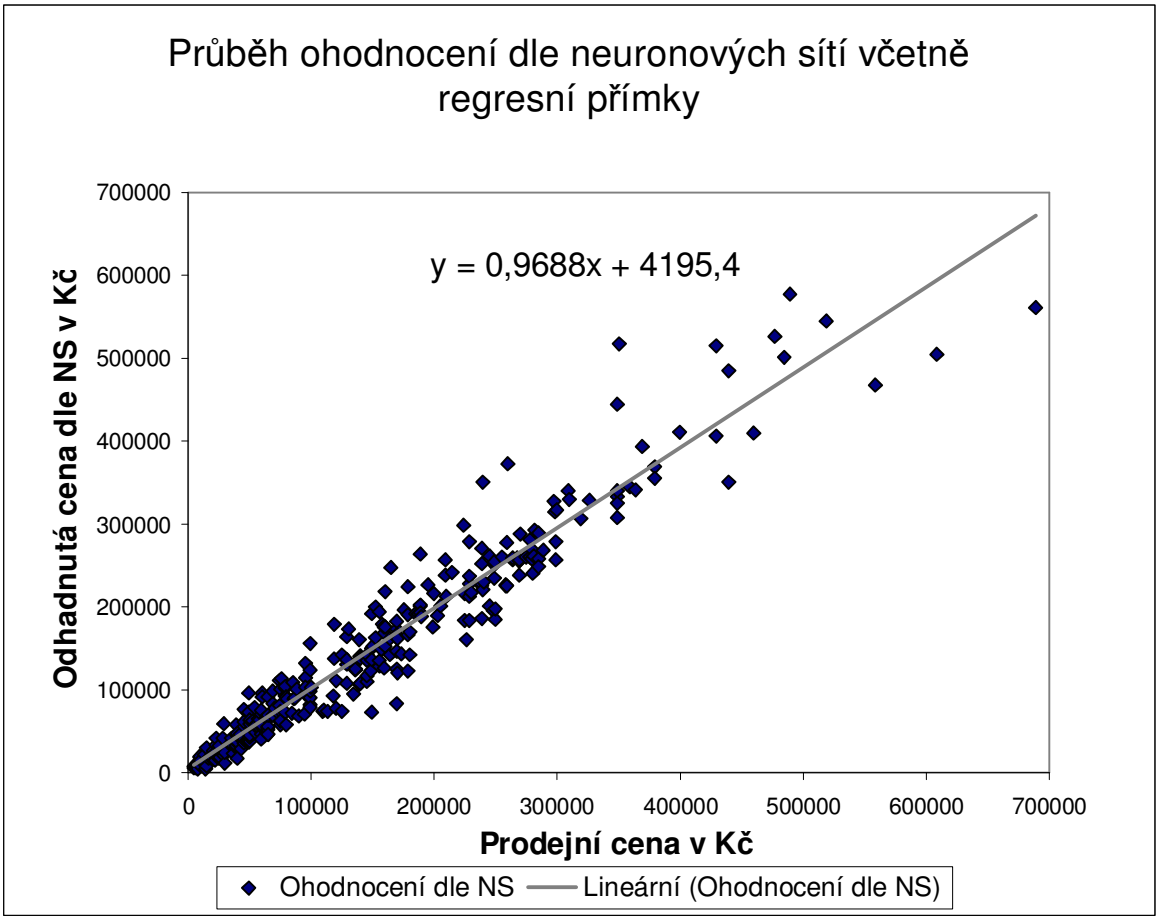
Graf 2: Porovnání procentuálních chyb ve výpočtech neuronových sítí a rozhodovacích v závislosti na roce výroby automobilu (zdroj: vlastní)

Hodnocení dle rozhodovacích stromů v závislosti na prodejní ceně lze okomentovat tak, že rozhodovací stromy nejsou úplně vhodným nástrojem, neboť mají určitý a konečný počet možných výstupů. Směrnice křivky aproximující rozložení výsledků má však hodnotu velice blízkou hodnotě 1, jenž naznačuje dobrou výslednost modelu. Hodnota směrnice je menší než 1, což ve výsledku znamená mírné průměrné podhodnocení oproti prodejní ceně.



Graf 3: Porovnání nezávisle proměnné prodejní ceny s závisle proměnou cenou rozhodavacích stromů (zdroj: vlastní)

Na druhé straně porovnání prodejní ceny a navrhované ceny dle neuronových sítí naznačuje větší použitelnost, protože neuronové sítě vykazují výstupy spojité. Směrnice přímky aproximující výstupy se více blíží k hodnotě 1, což znamená nepatrně lepší výsledky než u rozhodovacích stromů. V ideálním případě by hodnota směrnice přímky byla rovna 1. To by znamenalo, že modely hodnotí vozidla s nulovou průměrnou chybou.



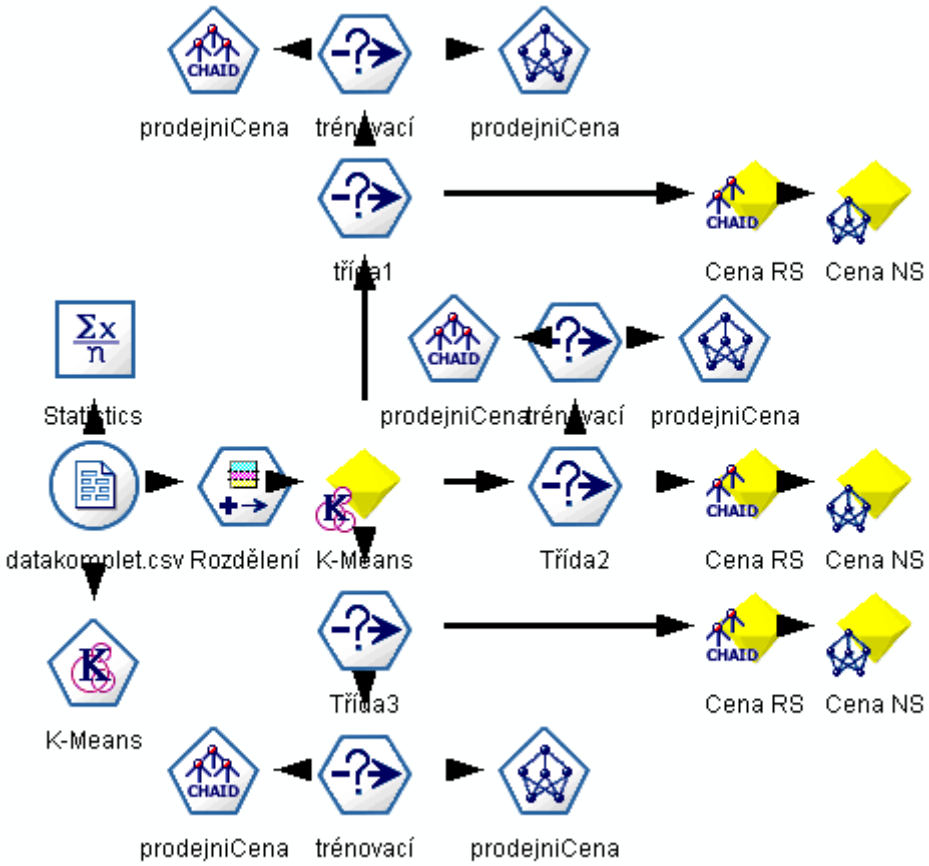
Graf 4: Porovnání nezávisle proměnné prodejní ceny s závisle proměnou cenou neuronových sítí (zdroj: vlastní)

8.1.3. Algoritmus K-means

Automobily byly rozděleny na tři homogenní skupiny, pro lepší výslednost modelů, bylo použito modelovacího uzlu K-means. Algoritmus je tvořen těmito kroky:

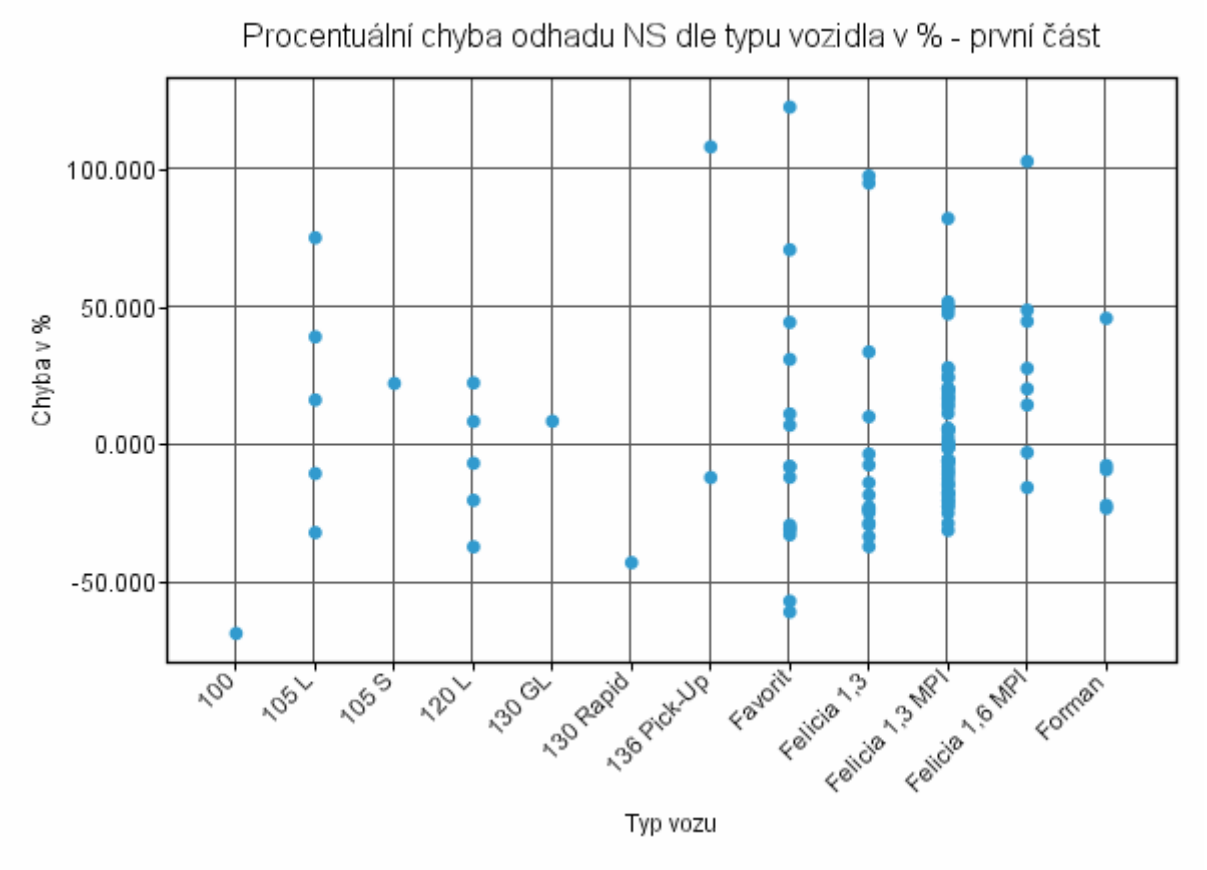
1) stanovte požadovaný počet tříd a jejich vzory
2) přiřaďte každý vektor do té třídy, od jejíhož vzoru má nejmenší vzdálenost
3) vypočtěte nové vzory tříd jako střední hodnoty vektorů příslušné třídy
4) opakujte kroky 2 a 3, dokud se mění vzory tříd.

Pomocí tohoto algoritmu znázorněném na obr.11 bylo nastaveno, aby uzel K-means vytvořil tři třídy (clustery) a dále bylo nastaveno, aby rozdělení provedl podle typu, roku výroby, objemu válců a výkonu. Výsledkem měly být tři navzájem odlišné homogenní skupiny automobilů pro další přesnější výsledky.



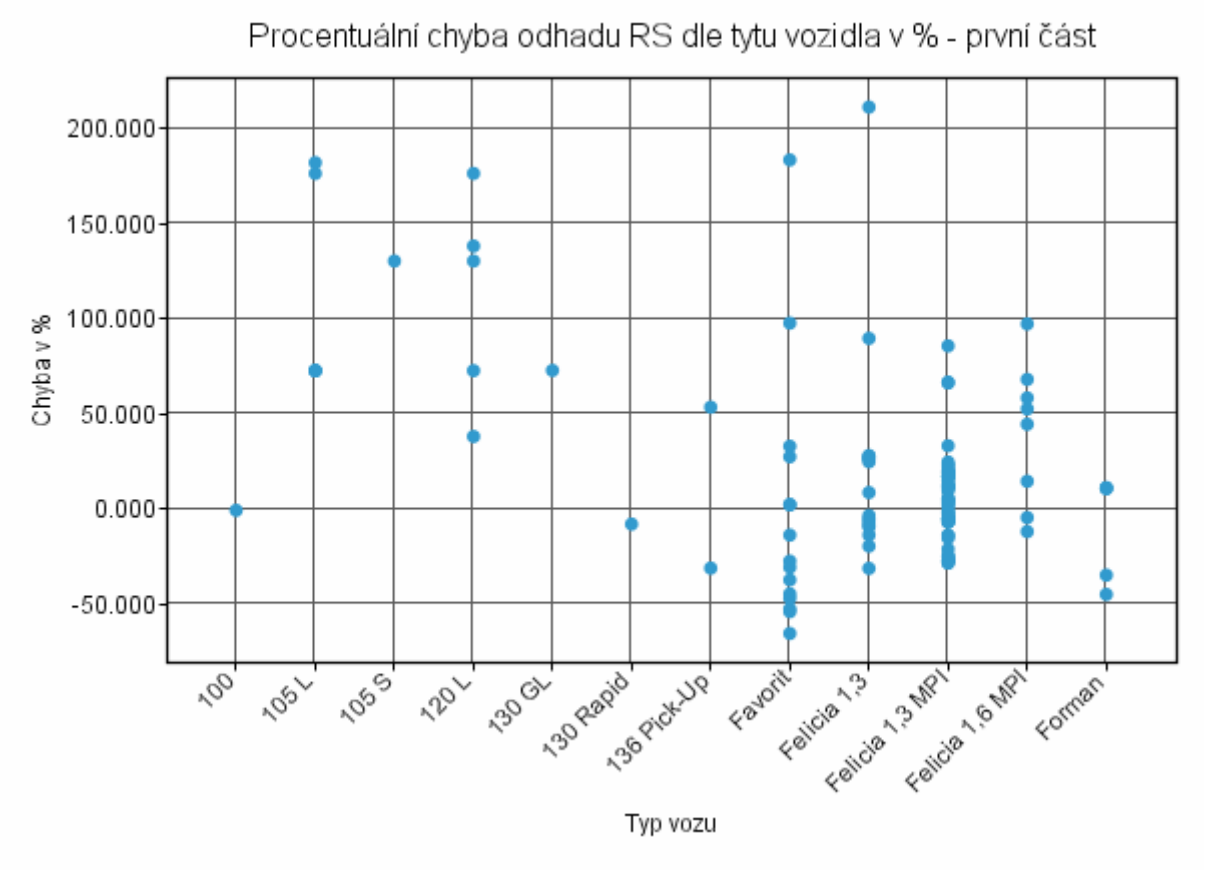
Obr.11: Rozdělení kvůli přesnosti modelů na tři podskupiny (zdroj vlastní)

V první třídě byly vybrány starší automobily o menším objemu a nižší prodejní ceně. Na grafu 5 a grafu 6 je vidět, jakým způsobem si u jednotlivých typů vozidel vedly neuronové sítě a jakým způsobem rozhodovací stromy. Grafy shodně ukazují největší procentuální chybu u vozu typu favorit. Možná příčina je v údržbě a užití vozu.



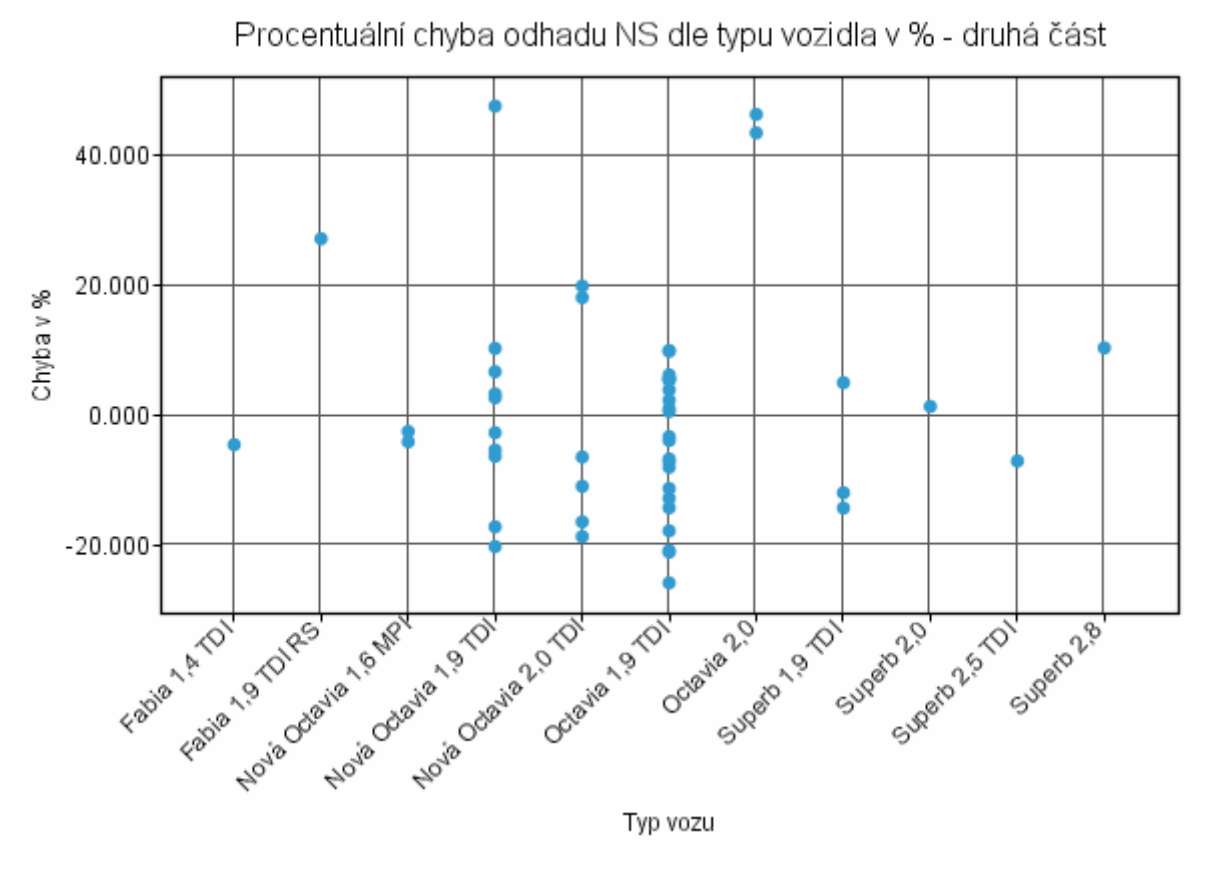
Graf 5: Procentuelní chyba neuronových sítí v závislosti na typu automobilu 1. část (zdroj: vlastní)

Z grafů 5 a 6 vyplývá, že jak neuronové sítě tak rozhodovací stromy starší typy vozidel nehodnotily zcela správně. I malé rozdíly mezi odhadnutou a prodejní cenou starších vozů mohou v závislosti na nízké prodejní ceně znamenat velkou procentuální chybu. Velké výkyvy jsou zapříčiněny stářím vozidla, nízkou cenou a velmi rozdílným stavem vozidel.



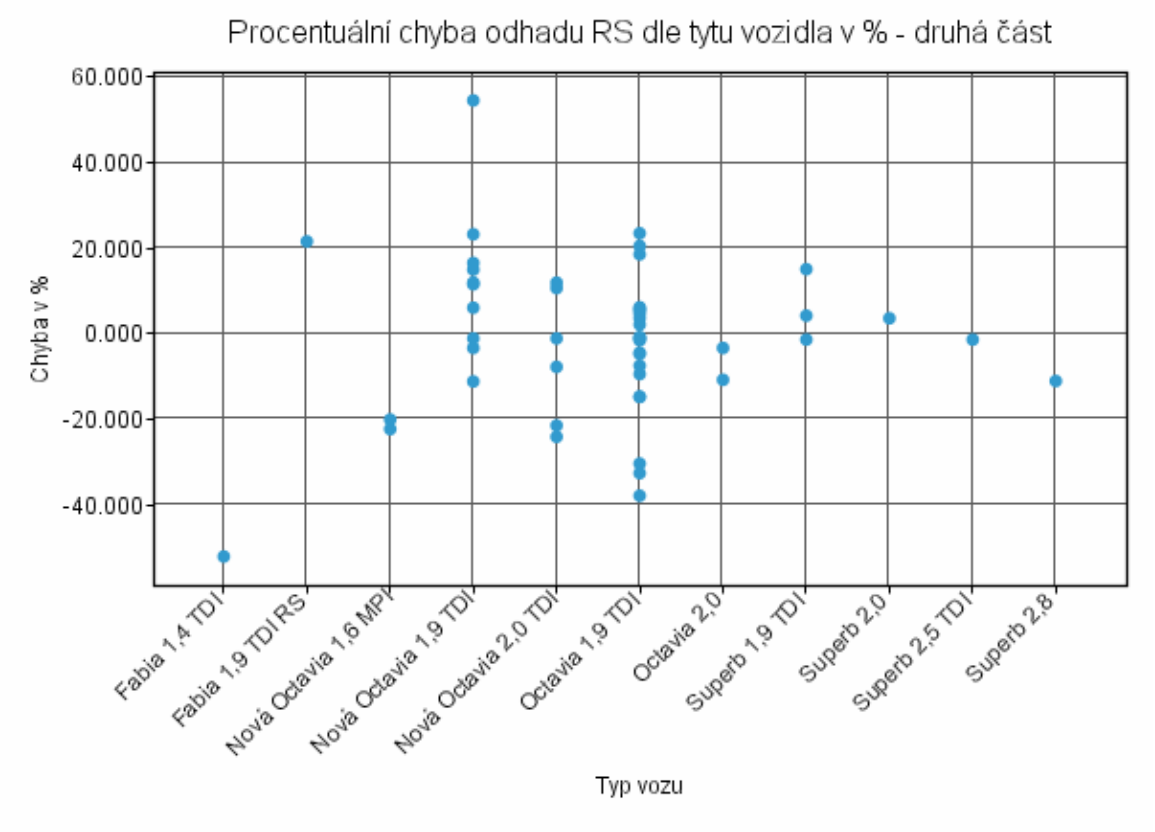
Graf 6: Procentuelní chyba rozhodovacích stromů v závislosti na typu automobilu 1. část (zdroj: vlastní)

Na grafech 7 a 8 jsou automobily vyššího roku výroby, většího zdvihového objemu a vysoké prodejní ceny. Jak graf neuronových sítí, tak graf podle rozhodovacích stromů dosahuje neporovnatelně lepších výsledků než graf 6 a 7, protože automobily těchto ročníků neničí v takové míře koroze a v povědomí prodejců jsou ceny, za které se tyto automobily prodávají.



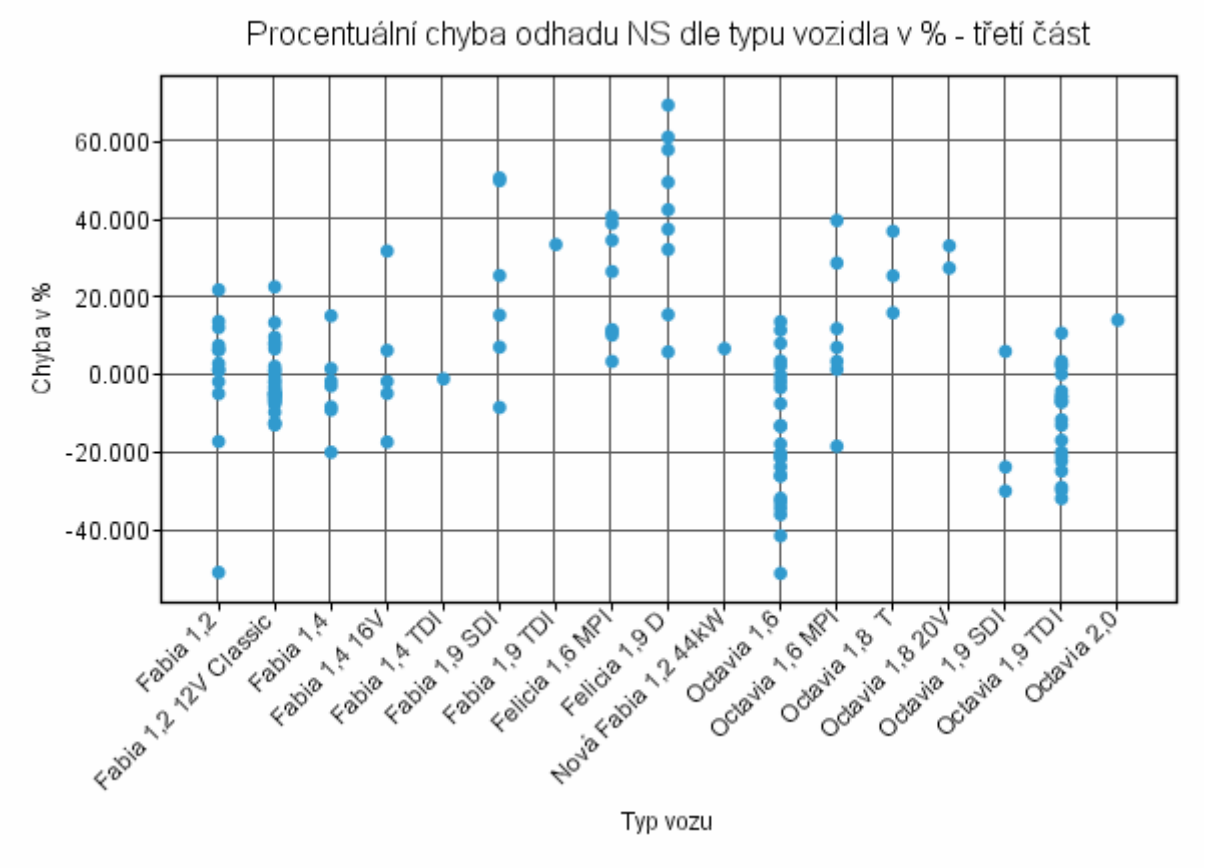
Graf 7: Procentuelní chyba neuronových sítí v závislosti na typu automobilu 2. část (zdroj: vlastní)

Za zmínku stojí snad jen typ vozu Octavia 2,0. Neuronové sítě tento automobil nadhodnotily, naopak rozhodovací stromy si vedly s ohodnocením velice dobře. V opačném případě se jednalo o vůz Fabia 1,4 TDI, u tohoto vozu vykazoval lepší výsledek model neuronových sítí.



Graf 8: Procentuelní chyba rozhodovacích stromů v závislosti na typu automobilu 2. část (zdroj: vlastní)

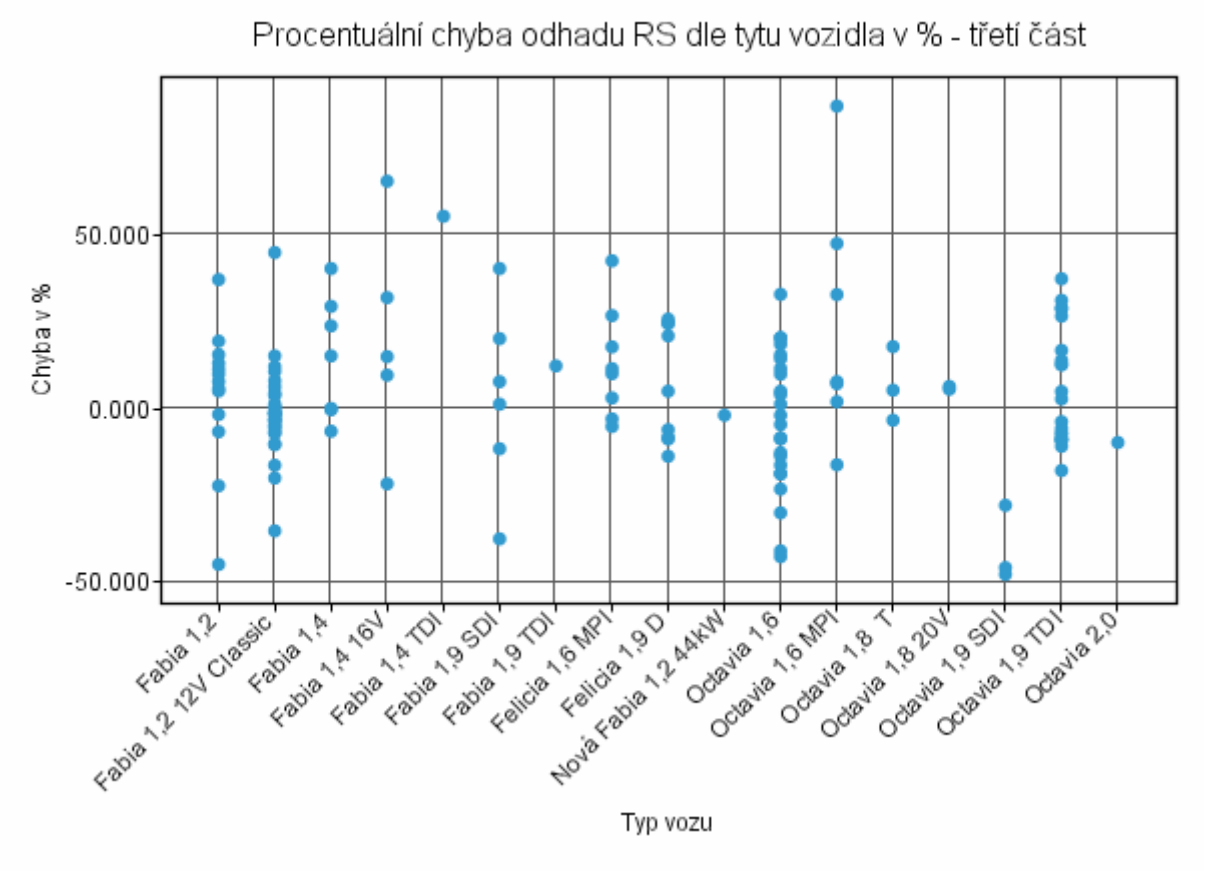
Ve třetí části jsou zobrazeny automobily středního stáří, středního zdvihového objemu a střední ceny. Těchto automobilů bylo největší množství. Na grafech 9 a 10 je vidět, že s těmito daty si obě modelovací techniky poradily velice dobře, neboť valná většina ohodnocení osciluje kolem nulové procentuální chyby.



Graf 9: Procentuelní chyba neuronových sítí v závislosti na typu automobilu 3. část (zdroj: vlastní)

Za zmínku stojí že rozhodovací stromy u vozu Felicia 1,9 měly lepší výsledky než neuronové sítě. Neuronové sítě vůz typu Felicia 1,9 nadhonocovaly a to až s 60 procentní chybovostí. Naopak rozhodovací stromy odhadly tento vůz s chybou do 30 procent.

.



Graf 10: Procentuelní chyba rozhodovacích stromů v závislosti na typu automobilu 3. část (zdroj: vlastní)

8.1.4. Porovnání výsledků

Pro lepší orientaci ve výsledcích jsou výsledky chyb zapsány v tabulce č. 3. Odhady cen ojetých vozidel se v případě vzoru (clusteru) č.1 vymykaly. Jak již bylo předesláno, starší automobil lze ohodnotit velice vysokou cenou, za kterou si tento vůz nikdo nekoupí. Anebo bude-li starý vůz velice zachovalý a vzhledově přitažlivý, systém bude naučen na vozech „ošklivých – používaných“, proto tento vůz podhodnotí. U vzoru č.2 a u vzoru č.3 jak rozhodovací stromy, tak neuronové sítě dosahovaly velmi dobrých výsledků. U těchto vzorů se systém dostal pod 20% procentní hodnotu chybovosti.

Tabulka 6: Porovnání celkových průměrných procentuálních chyb odhadů rozhodovacích stromů a neuronových sítí (zdroj: vlastní)

typ dat	celková % chyba RS	celková % chyba NS	záporná % průměrná chyba RS	kladná % průměrná chyba RS	záporná % průměrná chyba NS	kladná % průměrná chyba NS
vzor 1 (cluster-1)	19%	4%	-18%	46%	-19%	32%
vzor 2 (cluster-2)	-1%	0%	-13%	13%	-11%	13%
vzor 3 (cluster-3)	4%	2%	-13%	18%	-14%	18%
všechny testovací data	9%	2%	-15%	29%	-15%	23%

9. VYUŽITÍ V PRAXI

Vytvořením vhodného modelu řešení úlohy obecně nekončí. Dokonce i v případě, že řešenou úlohou byl pouze popis dat, je třeba získané znalosti upravit do podoby použitelné pro zákazníka. Podle typu úlohy tedy využití výsledků může na jedné straně znamenat prosté sepsání závěrečné zprávy, na straně druhé pak zavedení systému pro automatickou klasifikaci nových případů.

Ve většině případů je to zákazník a nikoliv analytik, kdo provádí kroky vedoucí k využívání výsledků analýzy. Z tohoto důvodu je důležité, aby pochopil, co je nezbytné učinit pro efektivní využívání výsledků.[1].

Aby zákazník mohl využívat výsledků analýz, musí mít přístup k poměrně drahému dataminingovému nástroji Clementine. Dále musí mít k dispozici vytvořenou analýzu, ve které jsou nastaveny a uloženy stromová struktura rozhodovacích stromů, rozvržení neuronové sítě a váhy na synapsích jednotlivých neuronů. Má-li zákazník dataminingový nástroj Clementine a příslušnou analýzu, je další zpracování velice jednoduché. Stačí pouze načíst data a „poklepat“ na výstupní ikonu tabulky. Další problém nastává v aktualizacích systému, protože postupem času automobily ztrácejí na hodnotě. Data je nutné obnovovat a systém učit na nových vstupních datech. Pro zákazníka, který tento systém má k dispozici a zvládá jednoduchou operaci nastavení vstupního souboru, pak tento systém vykoná velice zajímavou práci.

ZÁVĚR PRÁCE

V předložené práci byl navrhnut systém, který ohodnocuje automobily s využitím neuronových sítí a rozhodovacích stromů. Systém fungoval s výslednou průměrnou chybou, která měla u rozhodovacích stromů hodnotu 9% a u neuronových sítí byla hodnota průměrné chyby pouhé 2%. Uvedená chyba je způsobena pravděpodobně tím, že data, která byla k dispozici, nedávala stoprocentní obraz o tom, jak automobil vypadá resp. jaká je jeho hodnota.

Z této práce tedy vyplývá, že typ vozu, obsah válců, rok výroby, výkon, stav tachometru, barva, palivo, druh vozidla, typ karoserie, kraj, počet dveří, počet míst k sezení, airbag, ABS, klimatizace, centrál na dálkové ovládání, rádio na CD, imobilizér, hliníková kola, palubní počítač, ASR, ESP a tažné zařízení, nejsou nejpodstatnějšími parametry k odhadu ceny ojetého vozidla. Chybí zde např. celková koroze, hlučnost motoru, zda je nebo není vozidlo zakoupeno v České republice. Záleží také na tom, zda bylo vozidlo dodatečně upraveno, případně jinak vylepšeno. Ke zlepšení výslednosti modelů by došlo v tom případě, pokud by byly jasně nastaveny amortizační koeficienty, kterých se musí prodejci aut držet. Kdyby tedy byla mezi vstupními parametry, krom parametrů se kterými bylo již pracováno, například i cena nového vozidla, spotřeba pohonných látek, cena servisových hodin, již zmiňovaná koroze, zašlost laku, zašlost polstrování uvnitř vozu, zda automobil nadměrně nekouří, atd.

Práce prokázala, že použité techniky jako neuronové sítě a rozhodovací stromy jsou vhodné pro řešení této problematiky. Dosahují velmi dobrých výsledků. Srovnáním obou použitých modelů se ukázalo, že neuronové sítě ohodnocují s menší chybovostí než rozhodovací stromy. Výhoda spočívá ve spojitosti výstupů této metody. Potvrdilo se, že s rostoucí kvalitou dat přímo úměrně roste i kvalita výstupů.

POUŽITÁ LITERATURA

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

[2] PETR, Pavel. *Data Mining*. Pardubice : [s.n.], 2006. 144 s. ISBN 80-7194-886-1.

[3] SOUČEK, Eduard. *Základy pravděpodobnosti a statistiky*. Pardubice : [s.n.], 2003. 170 s. ISBN 80-7194-611-7.

[4] KVASNIČKA, V. a kol. *Úvod do teórie neurónových sietí*. Bratislava : IRIS, 1997. 285 s. ISBN 80- 88778-30-1.

[5] *Clementine* [online]. 2007 [cit. 2008-05-05]. Dostupný z WWW: <http://www.spss.cz/sw_clementine.htm>.

[6] *Wikipedie - otevřená encyklopedie* [online]. 2008 , 14.2.2008 [cit. 2008-05-06]. Dostupný z WWW: <http://cs.wikipedia.org/wiki/CRISP-DM>.

[7] *IBS expert* [online]. 2004 [cit. 2008-05-04]. Dostupný z WWW: <http://www.ibs-expert.cz/cz/page.php>.

[8] NOVÁK, M. a kol.: *Umělé neuronové sítě, teorie a aplikace*, C. H. Beck, Praha, 1998

[9] Kotek Z., Vysoký P., Zdráhal Z. *Kybernetika*. SNTL, Praha, 1990.

[10] Rozhodovací stromy. *Solutions* [online]. 2002 [cit. 2008-05-01]. Dostupný z WWW: <http://datamining.xf.cz/view.php?cisloclanku=2002102802>.

[11] KREJČÍŘ, Pavel, BRADÁČ, Albert. *Znalecký standard I/2005. Oceňování motorových vozidel.*. CERM. [s.l.] : [s.n.], 2005. 104 s. ISBN 8072043706.

[12] NOVÁK, M. a kol.: *Umělé neuronové sítě, teorie a aplikace*, C. H. Beck, Praha, 1998. 382 s. ISBN 80-7179-132-6.

[13] Prevence odchodu zákazníka pomocí data miningu. *IT systems* [online]. 2007 [cit. 2008-05-05]. Dostupný z WWW: <http://www.systemonline.cz/business-intelligence/prevence-odchodu-zakaznika-pomoci-data-miningu-1.htm>. ISSN 1802-615X.

[14] *Proces dobývání znalostí* [online]. 2002 [cit. 2008-05-06]. Dostupný z WWW: <http://euromise.vse.cz/kdd/index.php?page=proceskdd>.

POUŽITÉ ZKRATKY

CRISP-DM – standardní postup při vytváření dataminingových projektů (cross-industry standard process for data mining)
RS – rozhodovací stromy
NS – neuronové sítě
OLAP – technologie uložení dat v databázi (online analytical processing)
ABS – antiblokovací systém
ASR – systém regulace prokluzu kol
ESP – elektronický stabilizační systém
TDIDT – algoritmus pro tvorbu stromů (top down induction of desicion trees)

SEZNAM OBRÁZKŮ

SEZNAM GRAFŮ

SEZNAM TABULEK

SEZNAM POUŽITÝCH SYMBOLŮ

xi - hodnota na i-tém vstupu,

wi - váha i-tého vstupu,

Q - prahová hodnota,

n - celkový počet vstupů,

F - obecná nelineární funkce,

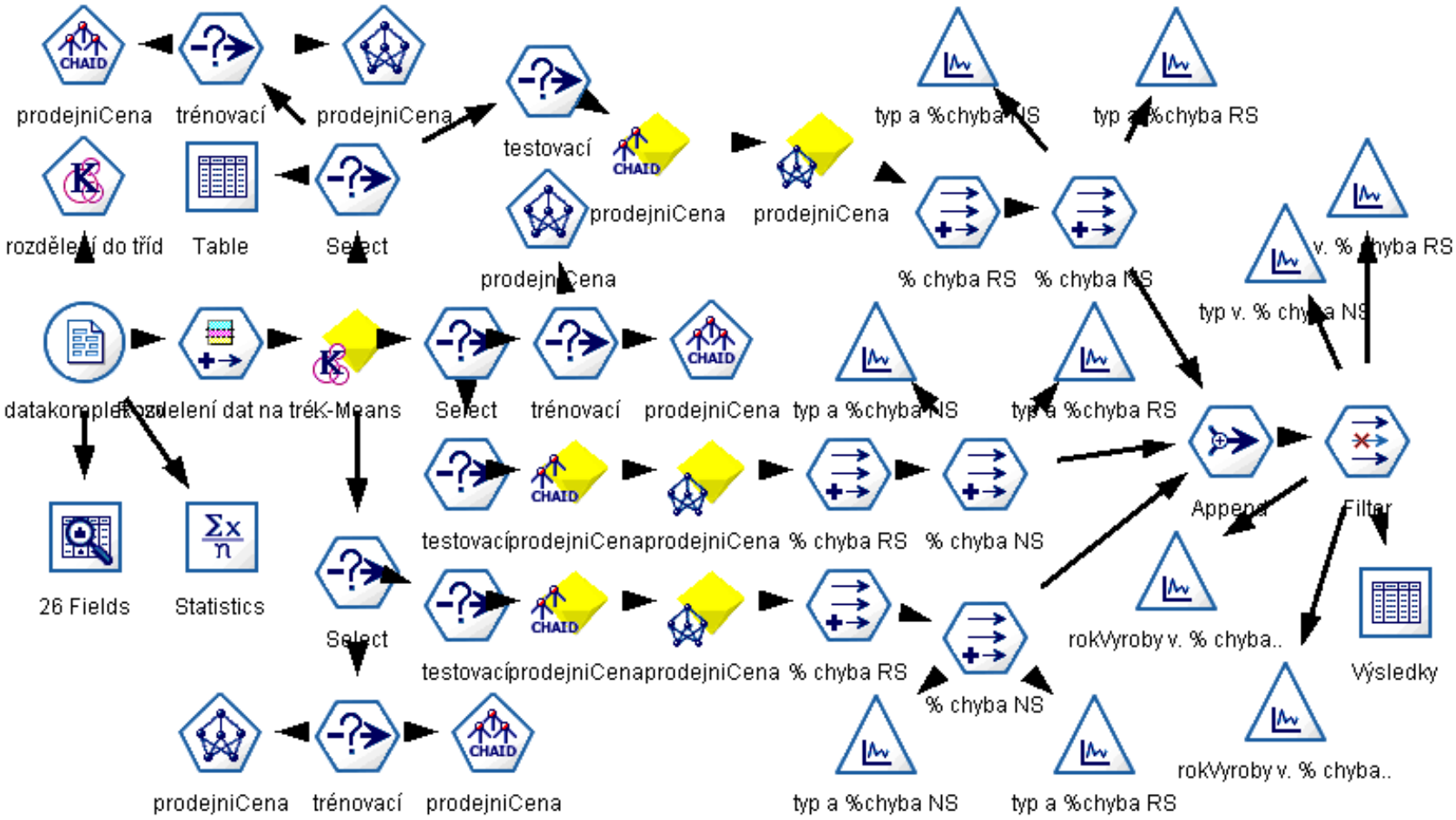
y - hodnota výstupu [12].

V - neprázdná vrcholová množina,

L – vrstva neuronové sítě

SEZNAM PŘÍLOH

PŘÍLOHA 1: CELKOVÝ STREAM PRÁCE



PŘÍLOHA 2: UKÁZKA VÝSLEDKŮ V TABULCE

ID	typ	obsah	vykon	rokVyroby	prodejniCena	RS-prodejniCena	NS-prodejniCena	% chyba RS
228	Fabia 1,2 12V Classic	1198	47	2005	209900	249043	232112	19%
316	Fabia 1,2 12V Classic	1198	47	2006	264000	262772	258740	0%
370	Fabia 1,2 12V Classic	1198	47	2006	278000	262772	255965	-5%
490	Fabia 1,4	1397	50	2003	146900	151846	154668	3%
556	Fabia 1,4 16V	1390	74	2000	139900	167475	162272	20%
612	Fabia 1,4 16V	1390	55	2004	215000	199434	225646	-7%
1744	Felicia 1,3 MPI	1289	50	2000	89990	86227	107556	-4%
1885	Felicia 1,6 MPI	1598	55	1998	69000	71650	84235	4%
2002	Felicia 1,9 D	1896	47	1997	59444	65181	65577	10%
2605	Octavia 1,6	1598	55	1997	119900	119876	70543	0%
2777	Nová Octavia 1,6 MPI	1595	75	2005	369000	393979	397076	7%
2225	Nová Octavia 1,9 TDI	1896	77	2005	470050	432940	354329	-8%
2755	Octavia 1,6 MPI	1595	75	2002	249000	216245	236148	-13%
2824	Octavia 1,8 20V	1781	92	1998	84000	124825	140362	49%
2381	Octavia 1,9 TDI	1897	66	1999	179000	168890	176796	-6%
2972	Octavia 1,9 TDI	1896	81	2002	268000	259350	289271	-3%
2537	Octavia 1,9 TDI	1896	66	2004	299000	288066	337215	-4%
3084	Octavia 2,0 TFSI	1984	147	2006	669000	599807	576237	-10%
34	105 L	1046	33	1986	11000	6545	11618	-41%
838	Favorit	1289	43	1992	13000	20547	22014	58%
931	Favorit	1289	43	1993	22000	20547	38423	-7%
976	Felicia 1,3	1289	40	1995	15000	43008	38178	187%
1035	Felicia 1,3	1289	40	1997	45000	59671	56524	33%
1179	Felicia 1,3 MPI	1289	50	1996	32000	47500	50018	48%
1316	Felicia 1,3 MPI	1289	50	1996	45000	39835	42622	-11%
1394	Felicia 1,3 MPI	1289	50	1996	49900	47500	48447	-5%
1498	Felicia 1,3 MPI	1289	50	1997	59000	59671	57418	1%
1641	Felicia 1,3 MPI	1289	50	1997	74000	59671	84197	-19%

PŘÍLOHA 3: PROCENTUÁLNÍ CHYBOVOST