Vysoká škola ekonomická v Praze

Fakulta informatiky a statistiky

Studijní program: Kvantitativní metody v ekonomice

Studijní obor: Statistické metody v ekonomii



METODA CHAID A LOGISTICKÉ REGRESE

Bakalářská práce

Autor bakalářské práce: Jaroslav Novák

Vedoucí bakalářské práce: Ing. Adam Čabla

školní rok 2015/2016

Prohlášení

Prohlašuji, že jsem bakalářskou práci zpracoval samostatně a že jsem uvedl všechny použité prameny a literaturu, ze kterých jsem čerpal.

V Praze dne 18. 5. 2016

…….
podpis

Poděkování

Na tomto místě bych rád poděkoval vedoucímu práce, Ing. Adamu Čablovi za cenné rady, připomínky a vstřícný přístup.

Abstrakt:

Tato bakalářská práce si klade za cíl představit metodu pro tvorbu rozhodovacích stromů zvanou CHAID a metodu logistické regrese. Následně je porovnat z hlediska jejich interpretace. Toho bude dosaženo aplikací těchto metod na reálná data pomocí statistického software, interpretací výsledných modelů a vyvozením následných závěrů. Práce tedy představuje možnosti interpretace modelů sestrojených pomocí metody CHAID a logistické regrese a s tím spojené výhody a nevýhody jednotlivých metod.

Klíčová slova: CHAID, rozhodovací stromy, logistická regrese, klasifikace

Abstract:

The aim of this thesis is to introduce logistic regression and method dedicated to construction of decision trees called CHAID, compare these two methods with regard to interpretation of their outputs. In order to accomplish the set goals application of these methods will be presented on real dataset. Statistical software will be used to obtain outputs. The outputs will be interpreted and conclusions on their bases will be presented. This thesis will also introduce possibilities of interpretation of these outputs and pros and cons that are connected with them.

Keywords: CHAID, decision trees, logistic regression, classification

Obsah

Obsah ... 1
Úvod ... 3
1 Metoda CHAID ... 4
1.1 Rozhodovací stromy ... 4
1.2 Algoritmus metody CHAID ... 6
1.2.1 První fáze algoritmu ... 6
1.2.2 Druhá fáze algoritmu ... 9
1.2.3 Třetí fáze algoritmu ... 10
1.3 Výhody metody CHAID ... 11
1.4 Nevýhody metody CHAID ... 12
1.5 Další metody pro tvorbu rozhodovacích stromů ... 13
2 Metoda logistické regrese ... 14
2.1 Princip logistické regrese ... 15
2.2 Logistický regresní model ... 19
2.2.1 Model s kategoriální vysvětlující proměnnou ... 19
2.3 Bodový odhad parametrů modelu ... 20
2.4 Hodnocení modelu ... 22
2.4.1 Testování parametrů modelu ... 22
2.4.2 Hodnocení kvality modelu ... 24
2.5 Výběr vysvětlujících proměnných ... 27
2.6 Interpretace modelu ... 28
2.7 Výhody a nevýhody logistické regrese ... 30
3 Aplikace metod na reálných datech ... 31
3.1 Popis dat ... 31
3.2 Metoda CHAID ... 33
3.2.1 Nastavení kritérií algoritmu ... 33
3.2.2 Interpretace ... 34
3.3 Metoda logistické regrese ... 38
3.3.1 Výběr vhodného modelu ... 38
3.3.2 Interpretace ... 39
3.4 Porovnání klasifikačních schopností ... 42
Závěr ... 45
Přehled literatury a zdrojů ... 47

Seznam použitých obrázků 48
Seznam použitých tabulek 49
Přílohy 50

Úvod

Metoda CHAID a logistická regrese jsou vícerozměrnými statistickými metodami pro modelování závislosti vysvětlované kategoriální proměnné, které se začali využívat až během 80. let minulého století. Od té doby se jejich využití rozšířilo do mnoha odvětví lidské činnosti. Tato práce kompiluje současné poznání o těchto metodách a snaží se o jejich porovnání z hlediska možností a způsobu interpretace výsledků.

Práce si klade za cíl představit základy těchto metod a porovnat obě metody z hlediska jejich výstupů a možné interpretace. Cíle bude dosaženo teoretickým popisem a následnou aplikací metod na reálná data. Tato aplikace bude zahrnovat interpretaci výsledků, zhodnocení výsledných modelů a v konečném důsledku vyvození závěrů.

Práce je členěna do tří hlavních částí. První část se zabývá teoretickým popisem metody CHAID, tedy obecným představením rozhodovacích stromů, popisem algoritmu metody CHAID a popisem možných výhod a nevýhod. Dále jsou stručně představeny další metody pro tvorbu rozhodovacích stromů. Druhá část teoreticky popisuje logistickou regresi. V této části je představen princip logistické regrese, způsob tvorby logistického regresního modelu, jeho hodnocení a interpretace. V poslední části jsou skutečnosti popsané v předchozích částech aplikovány na reálná data. Na základě obou metod jsou vytvořeny výsledné modely, které jsou interpretovány a porovnány pomocí jejich klasifikační schopnosti. V závěru jsou shrnuty zjištěné poznatky.

1 Metoda CHAID

Název této metody je zkratkou pro Chi-square Automatic Interaction Detection, poprvé byla publikována v roce 1980 v článku G. V. Kasse [1]. Tato metoda je primárně určena pro analýzu kategoriálních dat v rámci rozsáhlých datových souborů s velkým počtem proměnných. Patří mezi další techniky používané k tvorbě rozhodovacích stromů [2].

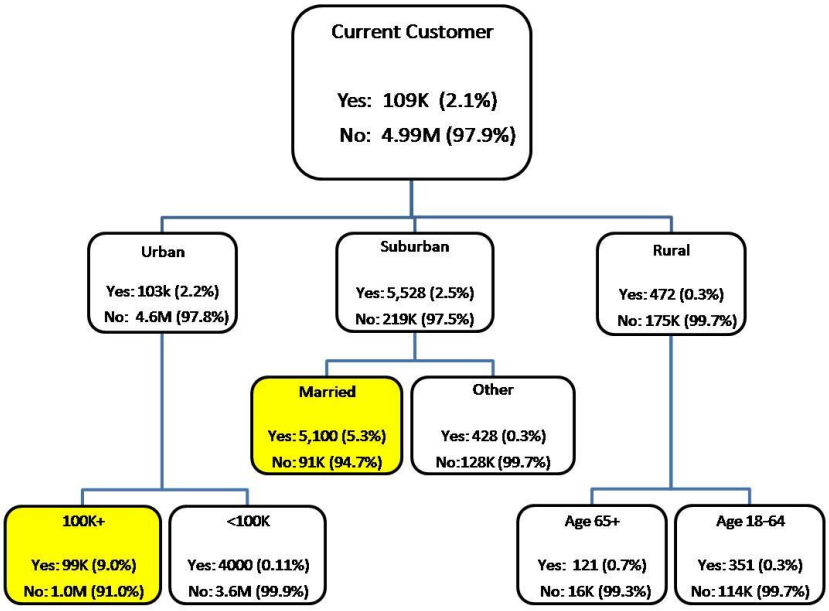
1.1 Rozhodovací stromy

Rozhodovací stromy je možné definovat jako strukturu tvarem podobnou převrácenému stromu, která zobrazuje hierarchickou skupinu vztahů mezi vysvětlovanou proměnou a dalšími vysvětlujícími proměnnými, takzvanými prediktory. Techniky sloužící k tvorbě těchto stromů, se řadí nejen mezi statistické ale i data miningové metody. Rozhodovací stromy jsou populární a poměrně hojně používané zejména pro jejich dvojí využitelnost. Jednak mohou sloužit ke klasifikaci, neboli roztřídění datového souboru. Při tomto roztřídění dochází k rozdělení heterogenní výchozí populace (např. všichni respondenti v příslušném průzkumu) na menší podskupiny, které jsou více homogenní (např. skupiny respondentů s podobnými vlastnostmi). Druhý způsob jakým mohou být rozhodovací stromy využity, je hledání závislostí mezi hodnotami vysvětlované proměnné a hodnotami již zmíněných prediktorů [2], [3], [4] a [5].

Celý strom je tvořen dvěma hlavními prvky a to uzly a větvemi. Uzly vznikají při dělení souboru na menší podskupiny podle hodnot vysvětlujících proměnných (např. podle pohlaví na muže a ženy). Takto vytvořené uzly obsahují hodnoty vysvětlované proměnné v příslušné podskupině (např. podíl mužů, eventuálně žen, kteří reagovali kladně na příslušnou otázku v dotazníkovém šetření). Větve mají za účel spojit jednotlivé hierarchické úrovně stromu [2].

U rozhodovacích stromů rozeznáváme několik typů uzlů. Samotný strom začíná nejvýše postaveným uzlem, který se nazývá kořen stromu, ten představuje kategorie vysvětlované proměnné pro celý soubor. Dále rozeznáváme takzvané koncové uzly. To jsou uzly, u kterých již nedochází k následnému dělení na další podskupiny. Při dělení příslušného uzlu na podskupiny, se rozlišují další dva typy uzlů. Výchozí uzel pro dělení je v anglické literatuře označen jako „parrent node“ (uzel rodič). Uzly, které tímto dělením vzniknou, se podle anglické literatury nazývají „child nodes“ (uzly potomci). Již zmíněné větve spojují právě rodičovské uzly se svými potomky. Rozhodovací stromy je možné dělit na binární a vícenásobné. Binární stromy mají u nekoncových uzlů vždy právě dvě větve, zatím co u vícenásobných stromů existuje alespoň jeden nekoncový uzel s více než dvěma větvemi. [2] a [5].

Na obrázku 1.1 je zobrazen příklad rozhodovacího stromu, vytvořeného pomocí metody CHAID. Strom byl sestrojen za účelem, aby odhalil segment osob, které mají větší tendenci reagovat kladně na přímou marketingovou kampaň. Jako základní soubor sloužil vzorek současných zákazníků, výsledky vyplývající ze sestrojeného stromu byly následně uvažovány na celou populaci. Strom odpovídá na otázku v jaké podskupině osob, bude nejvyšší podíl kladných odpovědí v rámci zmíněné marketingové kampaně. Z kořene stromu je patrné, že v celém souboru právě 2,1% zákazníků nejmenované společnosti reagovalo kladně. Kořen stromu se dále větví, podle proměnné představující místo bydliště. Podle této proměnné je příslušná populace rozdělena na osoby bydlící v městských oblastech, příměstských oblastech a na venkově. Tyto uzly se dále větví podle různých prediktorů. Z obrázku lze snadno vyčíst, že nejvyšší podíl zákazníků (9 %) je u osob s příjmem vyšším než 100 tisíc dolarů ročně, žijících v městských oblastech. Na základě těchto výsledků může příslušná společnost zacílit svou reklamní kampaň právě na tyto segmenty osob a učinit ji více efektivní.



Obrázek 1.1: Příklad rozhodovacího stromu [1].

S rozhodovacími stromy je možné se setkat v rámci různých oborů lidské činnosti. Vlastnosti rozhodovacích stromů se využívají například během analýzy marketingových kampaní, jak bylo popsáno výše. Tato problematika je podrobně popsána v [2]. Dále se objevují

[1] *Zdroj: SM Research*, dostupné online: http://www.smres.com/Images/CHAIDTree.jpg

v analýzách spojených se zákaznickými průzkumy, kde se s jejich pomocí zjišťuje, jaké vlastnosti zákazníků determinují jejich chování (např. zájem či nezájem o příslušný produkt). Dále se techniky konstrukce rozhodovacích stromů často používají v lékařství, kdy se zjišťuje jaké veličiny, vlastnosti pacientů a jejich chování mají zásadní vliv na vznik daného onemocnění (např. zjišťování, které faktory ovlivňují výskyt nádorových onemocnění). Rozhodovací stromy je také možné využít k získání základního přehledu o vztazích mezi vysvětlovanou proměnnou a jejími prediktory ve velkých datových souborech s mnoha proměnnými, a následně pomocí takto nabytých znalostí sestavit výsledný model pomocí jiné techniky zabývající se zjišťováním vzájemných vztahů mezi proměnnými. [2], [3] a [4].

1.2 Algoritmus metody CHAID

Algoritmus této metody byl poprvé představen G. V. Kassem v [1]. Algoritmus vede k výsledku v podobě vícenásobného rozhodovacího stromu. Jelikož je tato metoda určena k analýze kategoriálních dat, je nutné proměnné spojitého typu rozdělit na intervaly. Tím se příslušná spojitá proměnná převede na ordinální proměnnou s kategoriemi odpovídajícími zmíněným intervalům [2].

Celý algoritmus je možné rozdělit do tří jednotlivých fází. Pokud zkoumaný soubor dat neobsahuje žádnou spojitou proměnnou, nebo byla-li převedena na ordinální, je možné přistoupit k první fázi algoritmu, která se zabývá slučováním. V této fázi probíhá testování, zda jsou u každého jednotlivého prediktoru jeho katogerie vzájemně statisticky významě odlišné. K tomu se používá Pearsonův chí-kvadrát test o nezávislosti v kontingenční tabulce. Pokud statisticky významná odlišnost mezi kategoriemi není prokázána, příslušné kategorie se sloučí. Ve druhé fázi dochází k rozdělení rodičovského uzlu podle kategorií prediktoru, který nejlépe diferencuje hodnoty vysvětlované proměnné. Tento prediktor se určí také na základě Pearsonova chí-kvadrát testu o nezávislosti v kontingeční tabulce. V závěrečné třetí fázi se jen ověřují pravidla pro ukončení algoritu. Pokud jsou naplněna, algoritmus končí, v opačném případě se dále pokračuje ve větvení stromu [2], [3] a [5].

1.2.1 První fáze algoritmu

Jak již bylo řečeno, v tomto kroku dochází ke slučování kategorií příslušných prediktorů. Ke zjištění jaké kategorie sloučit a zda je vůbec sloučit se využívá Pearsonův chí-kvadrát test o nezávislosti v kontingenční tabulce, způsob jeho výpočtu je popsán podrobněji v [6]. K rozhodnutí o sloučení jednotlivých kategorií je postačující znát příslušnou p-hodnotu a

hladinu významnosti, která je stanovena již před zahájením celého algoritmu. P-hodnotu je možné vypočítat podle vztahu

$$p = P\left(\chi_k^2 \geq G\right). \tag{1.2.1}$$

P-hodnota tedy odpovídá pravděpodobnosti, že testová statistika s rozdělením chí-kvadrát a *k* stupni volnosti nabude hodnoty vyšší než *G*, což je empirická hodnota testové statistiky při platnosti nulové hypotézy. Pro *k* platí vztah

$$k = (d-1)(c-1), \tag{1.2.2}$$

kde *d* a *c* jsou počty kategorií vysvětlované a příslušné vysvětlující proměnné [6].

Nulová hypotéza u Pearsonova chí-kvadrát testu o nezávislosti v kontingenční tabulce předpokládá nezávislost hodnot vysvětlované proměnné na jejím prediktoru. Nezávislost znamená, že kategorie prediktoru statisticky významně nediferencují hodnoty vysvětlované proměnné. Naopak alternativní hypotéza hovoří o závislosti a tedy o statisticky významné odlišnosti hodnot vysvětlované proměnné v rámci kategorií prediktoru. Se zvyšující se hodnotou testové statistiky *G* klesá příslušná p-hodnota. Pokud p-hodnota nabyde hodnoty nižší než předem stanovená hladina významnosti α, nulová hypotéza bude zamítnuta a tím bude přijata alternativní hypotéza o závislosti na dané hladině významnosti závislost [7].

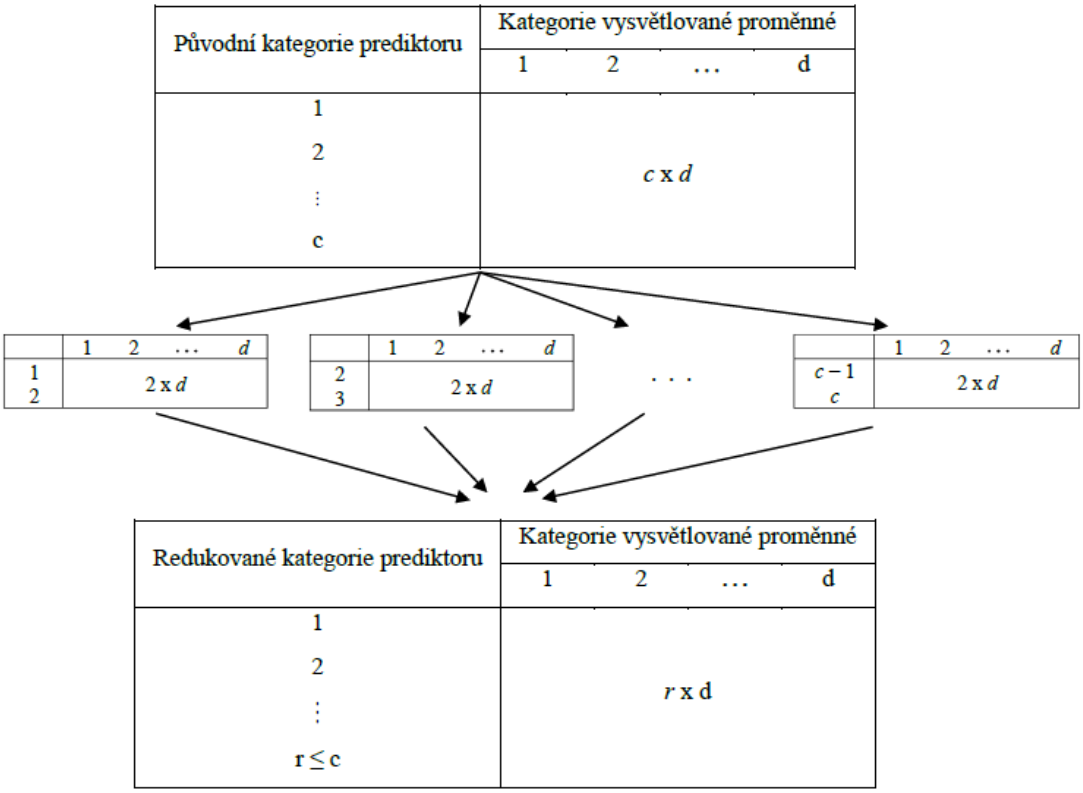
Pro první fázi algoritmu předpokládáme, že vysvětlovaná proměnná je kategoriální a má $d \geq 2$ kategorií. Dále právě analyzovanému prediktoru musí příslušet $c \geq 2$ kategorií (prediktor s pouze jednou kategorií má automaticky přiřazenou p-hodnotu rovnu jedné). Poté se sestaví kontingenční tabulka o rozměru *c* x *d*, která bude následně redukována tak, že se spojí ty kategorie, které se statisticky významně neliší s ohledem na vysvětlovanou proměnnou. O této odlišnosti bude rozhodnuto na základě již zmíněného testu [1], [2] a [6].

Postupně jsou testovány příslušné dvojice jednotlivých prediktorů, zda statisticky významně diferencují hodnoty vysvětlované proměnné. U nominálních vysvětlujících proměnných se testují všechny možné dvojice kategorií. Naproti tomu u ordinálních proměnných dojde k testování pouze dvojic kategorií, které po sobě bezprostředně následují. Jednotlivé testy tedy budou probíhat v rámci kontingenčních tabulek o rozměru 2 x *d*. Postupně se u každého prediktoru otestují jednotlivé dvojice kategorií [1] a [2].

Pokud na dané hladině významnosti nebude prokázáno, že se daná dvojice kategorií prediktoru statisticky významně liší, je možné poté tyto dvě kategorie sloučit. V případě, že je

během testování zjištěno více dvojic kategorií, které je možné sloučit, sloučí se ta dvojice kategorií, jejíž p-hodnota bude nejvyšší, jelikož vyšší p-hodnota a nižší testová statistika G značí méně významnou odlišnost. Jestliže testy prokáží na dané hladině významnosti u každé z dvojic kategorií příslušného prediktoru statisticky významnou odlišnost, k žádnému slučování v rámci tohoto prediktoru nedojde. Sloučené dvojice kategorií se nadále uvažují jako jedna samostatná kategorie příslušného prediktoru. [2], [6] a [8].

Tento proces slučování bude iterativně probíhat, dokud všechny kategorie příslušného prediktoru nebudou statisticky významně odlišovat hodnoty vysvětlované proměnné na zvolené hladině významnosti. Poté co bude tento proces ukončen, vznikne nový prediktor s redukovaným počtem kategorií r, kde platí $r \leq c$. Celý proces slučování je graficky znázorněn na obrázku 1.2.



Obrázek 1.2: Schéma slučování kategorií vysvětlující proměnné [2].

Jak je uvedeno v [1], v rámci této fáze algoritmu by kategorie vzniklé sloučením tří a více kategorií měly být binárně rozděleny, tedy do dvou následných kategorií. Je tedy nutné najít to nejlepší možné binární rozdělení opět pomocí Pearsonova chí-kvadrát testu.

[2] Zdroj: [2]

Znovurozdělení bude provedeno u té možnosti, u které byla p-hodnota nejnižší a zároveň byla nižší než stanovená hladina významnosti. V [6] je tato část algoritmu uvedena pouze jako volitelná a záleží na uživateli, zda ji do celého algoritmu zahrne.

1.2.2 Druhá fáze algoritmu

V dalším kroku algoritmu dochází k rozdělení rodičovských uzlů. Je tedy nutné nalézt vhodný prediktor (pokud takový existuje), podle jehož kategorií bude rodičovský uzel rozdělen. Ze skupiny prediktorů je tedy nutné vybrat ten, který nejlépe diferencuje hodnoty vysvětlované proměnné. K tomu, aby bylo možné takový prediktor jednoznačně určit, bude opět využit Pearsonův chí-kvadrát test o nezávislosti v kontingenční tabulce. Ovšem nyní test proběhne v rámci kontingenční tabulky o velikosti r x d [2].

Nejprve se vypočte standardní chí-kvadrát test, avšak následně bude nutné jeho výsledky upravit. Problémy s klasickým chí-kvadrát testem o nezávislosti se objevují u prediktorů, jejichž kategorie byly redukovány. To je způsobeno tím, že tento test nezohledňuje počet původních kategorií vysvětlující proměnné. Aby bylo dosaženo správného výsledku, používá se Bonferroniho úprava, která již bere ohled na počet původních kategorií. [1] a [2].

Tato úprava spočívá v použití takzvaného Bonferroniho multiplikátoru. Tento multiplikátor představuje počet možných způsobů, jak c původních kategorií sloučit do r redukovaných kategorií. Počet způsobů, jak je možné sloučit původní kategorie vysvětlující proměnné, závisí na typu proměnné. Proto se i výpočet Bonferroniho multiplikátoru pro jednotlivé typy proměnných liší. Vzorce výpočtu multiplikátorů pro dané typy proměnných jsou za podmínky $2 \leq r < c$ následující [1], [2] a [6].

Pro ordinální prediktor:

$$B_{ord} = \binom{c-1}{r-1}. \tag{1.2.3}$$

Pro nominální prediktor:

$$B_{nom} = \sum_{i=0}^{r-1} (-1) \frac{(r-i)^c}{i!(r-i)!}. \tag{1.2.4}$$

V [1] a [6] je dále uveden vzorec pro takzvaný „floating predictor“. V podstatě se jedná o ordinální prediktor zahrnující kategorii, u které nelze jednoznačně určit pořadí, s ohledem na

ostatní kategorie ordinálního prediktoru. Tato kategorie se tedy nalézá mimo pomyslnou stupnici hodnot ordinální proměnné. Konkrétně se může jednat například o chybějící hodnoty, odpovědi v dotazníku typu „nevím“. Pro tento typ prediktoru je vzorec Bonferroniho multiplikátoru následující:

$$B_{float} = \binom{c-2}{r-2} + r\binom{c-2}{r-1} = \frac{r-1+r(c-r)}{c-1} B_{ord} . \qquad (1.2.5)$$

Výsledky standardního Pearsonova testu o nezávislosti v kontingenční tabulce se upraví tak, že se výsledná p-hodnota testu vynásobí příslušným multiplikátorem. Tímto způsobem se získá upravená p-hodnota, na jejímž základě již bude možné rozhodnout, podle jakého prediktoru se bude větvit daný rodičovský uzel. Nejvhodnějším prediktorem pro následné větvení je ten, jemuž přísluší nejnižší upravená p-hodnota. Větvení proběhne pouze v případě, že tato upravená p-hodnota je nižší, než zvolená hladina významnosti pro větvení. Příslušný rodičovský uzel se rozdělí podle vybraného prediktoru. Redukované kategorie náležící tomuto prediktoru budou tvořit nové uzly rozhodovacího stromu. Tyto uzly se následně mohou dále větvit, pokud nebude splněno některé z kritérií pro zastavení algoritmu [2], [5] a [6].

1.2.3 Třetí fáze algoritmu

Tato fáze se zabývá případy, kdy dojde k ukončení celého algoritmu. Rozhodnutí o ukončení algoritmu, je založeno na pravidlech, po jejichž splnění se algoritmus zastaví. První a druhý krok algoritmu probíhá opakovaně do té doby, dokud není splněno alespoň jedno z pravidel pro ukončení algoritmu. Tři z těchto pravidel jsou uvedena v [2].

V prvním případě se jedná o zastavení algoritmu z toho důvodu, že již neexistuje žádný prediktor, podle kterého by bylo možné dále větvit výsledné uzly stromu. Tato situace nastane, pokud není možné nalézt další prediktor s upravenou p-hodnotou nižší než zvolená hladina významnosti pro větvení.

Druhé pravidlo hovoří o ukončení algoritmu v případě, že byl dosažen maximální počet úrovní stromu.

U třetího pravidla se jedná o situaci, kdy po dalším větvení by nebyla splněna podmínka o minimálním počtu pozorování v takto vzniklých uzlech. Z toho důvodu toto větvení již neproběhne.

Dále je nutné zmínit, že parametry, na kterých závisí ukončení algoritmu, se volí ještě před analýzou, tedy zahájením celého algoritmu. Mezi tyto parametry patří hladina významnosti pro větvení, maximální počet úrovní stromu a minimální počet pozorování pro jednotlivé uzly. Navzdory tomu, že tyto parametry výrazně ovlivňují vlastnosti výsledného stromu, pro jejich volbu neexistují pevně stanovená kritéria. Proto závisí na uživateli, aby byly zvoleny takové hodnoty parametrů, které zajistí výslednému rozhodovacímu stromu dostatečnou vypovídací hodnotu.

1.3 Výhody metody CHAID

V tomto oddíle budou shrnuty zásadní výhody použití metody CHAID pro analýzu interakcí mezi proměnnými uvedené v [2], [5] a [8].

Jednou z výhod metody CHAID je *rychlost*, s jakou je celý algoritmus vypočten. Samozřejmě za předpokladu použití moderního software, který umožňuje konstrukci rozhodovacích stromů. Tímto způsobem není problém zpracovat datové soubory s velkým počtem pozorování a proměnných v relativně krátkém čase.

Další nespornou výhodou je *nenáročnost*. Kromě velikosti datového souboru (dostatečný počet pozorování) neklade tato metoda žádné další předpoklady na data. To je zapříčiněno tím, že celý algoritmus závisí na chí-kvadrát testu, jehož jediným požadavkem je počet očekávaných pozorování. Dále by měl mít výsledný rozhodovací strom dostatečnou vypovídací hodnotu, ale pokud je k dispozici dostatečný počet pozorování, tato vypovídací hodnota závisí jen na správné volbě vstupních parametrů, které byly zmíněny v předchozím oddíle.

Mezi výhody se řadí také *přehlednost* výstupu celé metody. Tímto výstupem je grafické znázornění výsledného rozhodovacího stromu. Přehledné zobrazení ve formě rozhodovacího stromu umožňuje snadno porozumět výsledkům této metody i osobám, které nedisponují znalostí celého algoritmu.

Za výhodu je také možné považovat, že metoda CHAID vytváří vícenásobné rozhodovací stromy. Oproti jiným metodám, které umožňují jen binární větevní, metoda CHAID vytváří „širší“ stromy. To umožňuje lépe rozdělit jednotlivé podskupiny, které se liší hodnotou vysvětlované proměnné. Pokud by bylo možné jen binární dělení rodičovských uzlů,

některé z těchto podskupin by mohly být sloučeny a jejich rozdílnost v hodnotách vysvětlované proměnné by nebyla odhalena.

Další výhodou je, že tato metoda dokáže pracovat s chybějícími hodnotami a to dvěma způsoby. První možnost, je že chybějící hodnoty dané proměnné budou představovat samostatnou kategorii. Druhou možností je sloučení chybějících hodnot s jinou kategorií příslušné proměnné.

Poslední výhoda, kterou zmíním, je schopnost odhalit a vystihnout některé podmíněné vztahy mezi prediktory. Jako příklad lze uvést vztah pohlaví a věku. Z výsledného stromu, poté může být například patrné, že věk hraje roli pouze u mužů a u žen ne.

1.4 Nevýhody metody CHAID

Nyní budou popsány některé nevýhody spojené s metodou CHAID, které jsou zmíněny v [2], [5] a [8].

Jednou z nejvýznamnějších nevýhod, hovořících proti použití metody CHAID k analýze dat, je absence výstupu v podobě matematického modelu, který by zřetelně popisoval vztahy mezi proměnnými. Přesto je možné vyčíst směr vztahu z kontingenční tabulky. Chybějící matematický model je zásadním rozdílem, kterým se metoda CHAID liší od jiných metod používaných k analýze kategoriálních dat. Jako příklad takových metod je možné uvést logistickou regresi, které bude rozebrána v následující kapitole.

Další vlastností, která omezuje využití této metody, je nutnost velkého počtu pozorování v rámci datového souboru. Tím se tato metoda stává nepoužitelnou pro malé vzorky populací. V některých zdrojích se uvádí, že by minimální počet pozorování pro každý uzel měl být alespoň 30 – 50 pozorování, aby byla zajištěna dostatečná vypovídací hodnota rozhodovacího stromu. Například statistický výpočetní software SPSS v rámci svého výchozího nastavení pro metodu CHAID určuje minimálně 100 pozorování pro uzly typu rodič a alespoň 50 pozorování pro uzly typu potomek. Toto výchozí nastavení je možné libovolně upravit, aby odpovídalo velikosti vstupního datového souboru. Pokud by zvolená kritéria na minimální počet pozorování byla příliš přísná, výsledný strom se poté dostatečně nerozvětví. V opačném případě, kdy bude nastaven příliš nízký minimální počet pozorování, může dojít k velkému rozvětvení stromu, který se tím stane nepřehledným.

K nevýhodám patří také to, že je tato metoda určena primárně pro kategoriální data. I když je možné numerické proměnné převést na kategoriální pomocí rozdělení do jednotlivých intervalů, není toto řešení ideální. Při tomto převodu numerické proměnné dochází ke ztrátě části informace, která je v datech obsažena.

1.5 Další metody pro tvorbu rozhodovacích stromů

Algoritmus metody CHAID není jediným, který se pro tvorbu rozhodovacích stromů používá. Statistický software SPSS například umožňuje sestavení rozhodovacího stromu pomocí metod CART a QUEST. Metoda CART (Breiman, Friedman, Olshen, Stone, 1984) je zkratkou pro anglické pojmenování Classification And Regression Tree. Rozhodovací stromy vzniklé pomocí této metody jsou binární. Proto kategorie příslušných prediktorů musejí být redukovány tak, aby bylo možné rodičovský uzel rozdělit právě na dva uzly potomky a tak aby se dosáhlo nejlepší možné hodnoty štěpícího kritéria. Jako štěpící kritérium v rámci této metody slouží Giniho index, pomocí jehož hodnoty se určí správný prediktor, podle kterého bude rodičovský uzel rozdělen [6] a [8].

Metoda QUEST (Loh, Shih, 1997) neboli Quick, Unbiased, Effecient, Statistical Tree stejně jako metoda CART vytváří binární rozhodovací stromy. V rámci tohoto algoritmu existují dvě kritéria pro výběr vhodného prediktoru, podle kterého se rozdělí příslušný rodičovský uzel. Jaké kritérium bude použito, závisí na typu prediktoru. Mezi spojitými nezávislými proměnnými se hledá vhodný prediktor pro rozdělení na základě p-hodnoty Fisherova F-testu. U kategoriálních nezávislých proměnných se k tomuto účelu využívá p-hodnoty Pearsonova chí-kvadrát testu. Jelikož se opět jedná o binární strom, hodnoty příslušných prediktorů musejí být upraveny tak, aby se rodičovské uzly rozdělovaly právě na dvě větve a aby tyto prediktory co nejlépe diferencovaly hodnoty vysvětlované proměnné [6].

Posledním algoritmem pro tvorbu rozhodovacích stromů, který zmíním, je C5.0 (Quinlan, 1998). Rozhodovací strom vzniklý na základě této metody je vícenásobný. Algoritmus je založen na maximalizaci takzvaného informačního zisku, který je založen na konceptu informační entropie [8].

2 Metoda logistické regrese

Logistická regrese, podobně jako další regresní metody určené k analýze dat, slouží ke zkoumání jednostranného vztahu mezi vysvětlovanou proměnnou a jednou nebo i více vysvětlujícími proměnnými. Jedná se tedy o vícerozměrnou metodu. K popisu tohoto vztahu se využívá matematický model, který je výstupem algoritmu příslušné metody a je vyjádřen regresní funkcí. Pro takový model platí, že by měl být úsporný, spolehlivý, věcně smysluplný a co nejlépe by měl vystihovat skutečný vztah mezi proměnnými. Na rozdíl od ostatních regresních metod, z nichž je pravděpodobně nejznámější lineární regrese, logistická regrese umožňuje analyzovat závislost kategoriální vysvětlované proměnné na jejích prediktorech. Přičemž vysvětlující proměnné mohou být jakéhokoliv typu. Předmět analýzy je tedy v podstatě shodný jak u logistické regrese, tak u metody CHAID popsané v předchozí kapitole [9] a [10].

Ačkoliv principy, na kterých je metoda logistické regrese založena, byly známy již dříve, větší popularitě se této metodě dostalo až po roce 1970 především díky knize *Analysis of Binary*, kterou ve stejném roce publikoval britský statistik David Cox. K popularizaci této metody přispěly také nové výpočetní technologie, jejichž následný vývoj výrazně usnadnil proces výpočtu této metody. Navzdory tomu, že od 80. let minulého století byla logistická regrese již v zemích Západní Evropy a Severní Ameriky standardní metodou pro analýzu dat, v České republice ještě na přelomu tisíciletí byla tato metoda spíše méně známá [9] [11] a [12].

Metoda logistické regrese byla vyvinuta primárně pro analýzu vztahu závislosti binární proměnné na jejích prediktorech. Pro tento případ se užívá označení binární logistická regrese. V této práci se zaměřím, především na tento typ regrese. Je třeba zmínit, že tato situace není jediná, kdy je možné logistickou regresi k analýze použít. Postupem času byly vyvinuty i postupy, které umožnily zkoumat příslušnou vysvětlovanou proměnnou s více než dvěma kategoriemi a postupy, díky kterým bylo možné analyzovat i vysvětlované proměnné ordinální. V takových případech se poté hovoří o multinomické či ordinální logistické regresi [9] a [13].

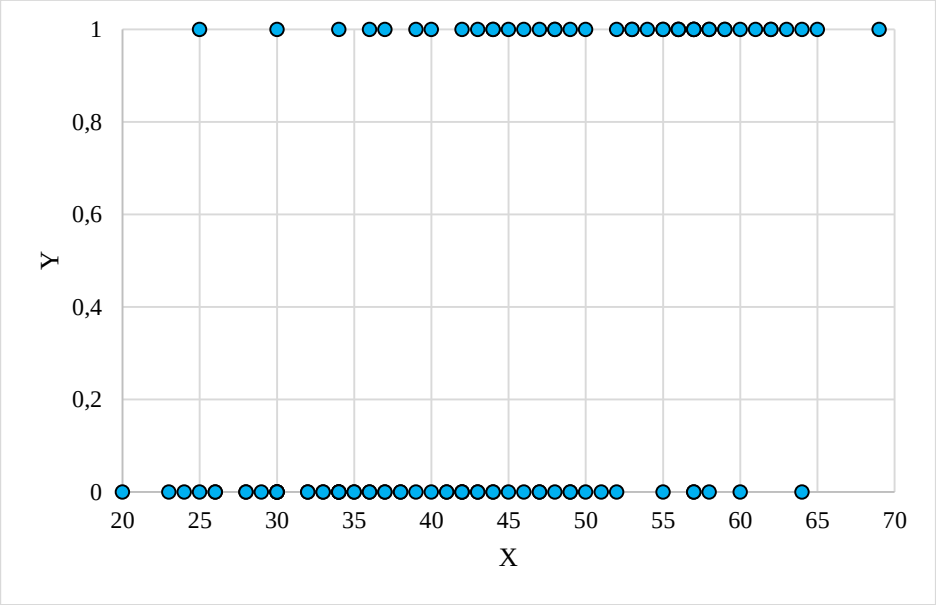
Logistická regrese byla nejprve využívána především v medicíně a epidemiologii, kde se s její pomocí předpovídala například pravděpodobnost rozšíření choroby nebo mortalita následkem určitého úrazu. Následně se začala využívat i v dalších oborech. V marketingu například při modelování pravděpodobnosti kladné reakce na reklamní kampaň, či zakoupení určitého produktu. Logistická regrese je také častým nástrojem ve finančnictví pro ověřování schopnosti potencionálního klienta splácet závazky v případě půjček, nebo hypoték. V rámci

sociologie je možné tuto metodu využít například pro vytvoření modelu, zda je respondent v závislosti na daných faktorech spokojen s určitým společenským jevem, takový jev může kupříkladu představovat politickou či ekonomickou situaci. V technických oborech se může používat pro modelování rizika poruchy určitého zařízení za daných podmínek. Dalšími obory, kde se využívá logistická regrese, jsou například genetika nebo biologie [9], [10] a [11].

Princip logistické regrese, tvorba modelu, jeho interpretace a hodnocení jeho kvality bude popsán primárně pro binomickou vysvětlovanou proměnnou.

2.1 Princip logistické regrese

Jak již bylo řečeno, v logistické regresi vysvětlovanou proměnnou představuje proměnná kategoriálního typu. V takovém případě není možné použití klasického lineárního modelu, kde kategoriální proměnné mohou vystupovat pouze na pozici vysvětlujících proměnných, neboť by nebyly splněny příslušné podmínky daného výpočetního postupu a došlo by i k dalším numerickým problémům. K odstranění těchto potíží využívá logistická regrese transformace, která vychází ze zobecněného lineárního modelu[3] [9] a [13].



Obrázek 2.1: Graf hodnot alternativní proměnné *Y* v závislosti na *X* [4].

V případě binární vysvětlované proměnné, kterou označíme *Y*, se zpravidla značí její hodnoty nulou a jedničkou. Přičemž $Y = 1$, pokud sledovaný jev nastal a $Y = 0$, jestliže daný jev nenastal. Za předpokladu, že se v modelu objeví pouze jedna spojitá vysvětlující proměnná

[3] O zobecněných lineárních modelech se pojednává například v [11].
[4] *Zdroj*: *vlastní zpracování, MS Excel, data z* [10] *viz příloha A*

X, je možné tuto situace znázornit pomocí grafu jako na obrázku 2.1. V konkrétní aplikaci může proměnná *X* představovat například věk účastníků příslušné studie a proměnná *Y* výskyt určitého onemocnění. Ze zmíněného obrázku je patrné rozdělení hodnot proměnné *Y*. Pro jednotlivé hodnoty proměnné *X* se vysvětlovaná proměnná *Y* řídí alternativním rozdělením, jedná se tedy o jednotlivá podmíněná alternativní rozdělení [9], [10] a [13].

Jelikož povaha vysvětlované proměnné způsobuje problémy, které brání použití lineární regrese, je nutná transformace jejích hodnot na hodnoty spojité. Toho se docílí tím způsobem, že nebudou odhadovány jednotlivé hodnoty proměnné *Y*, ale podmíněné pravděpodobnosti nastoupení příslušného jevu při dané hodnotě vysvětlující proměnné *X*, tedy $P(Y = 1 | x)$. Střední hodnoty jednotlivých podmíněných rozdělení proměnné *Y* odpovídají právě těmto pravděpodobnostem. To znamená, že platí následující vztah:

$$\pi(x) = E(Y \mid x) = P(Y = 1 \mid x) \ . \tag{2.1.1}$$

Při pokusu sestavit následující lineární regresní model:

$$\pi(x) = \beta_0 + \beta_1 X \tag{2.1.2}$$

kde β_0 a β_1 představují regresní parametry, se opět objeví další problém, jelikož se jedná o pravděpodobnost jevu. Ta je definována tak, že může nabývat pouze hodnot mezi nulou a jedničkou. Podstata problému spočívá v tom, že tento model není schopen zajistit zmíněné omezení pro hodnoty pravděpodobnosti, jelikož regresní funkce (2.1.2) může teoretický nabývat jakýchkoliv hodnot z intervalu $(-\infty ; +\infty)$. Proti použití tohoto modelu hovoří také fakt, že vztah mezi pravděpodobností jevu a vysvětlující proměnnou není lineární. Neboť pokud dojde k jednotkové změně vysvětlující proměnné, nedojde ke změně pravděpodobnosti takové, která by byla stejná pro každou další jednotkovou změnu vysvětlující proměnné. Takový model musí být tedy dále upraven. [8], [9], [13] a [14]

Další úprava se provede pomocí další transformace, tedy záměnou pravděpodobnosti za takzvanou šanci jevu. Šance nastoupení jevu neboli *šance (Y=1)*, je definována jako poměr pravděpodobnosti, že daný jev nastane $P(Y = 1)$ a pravděpodobnosti, že příslušný jev nenastane $P(Y \neq 1)$. Z této definice vyplývá následující vztah:

$$\check{s}ance(Y = 1 \mid x) = \frac{\pi(x)}{1 - \pi(x)} \ . \tag{2.1.3}$$

Díky této úpravě se sice odstraní problém, s hodnotami pravděpodobnosti omezenými na intervalu mezi jedničkou a nulou, ale hodnoty, kterých může nabývat šance jevu, jsou také omezené. Jelikož šance jevu nemůže nabývat záporných hodnot [9].

Šance jevu tedy musí být dále transformována tak, aby výsledek této transformace byl definovaný na celém intervalu (−∞ ; +∞). Toho se dosáhne pomocí přirozeného logaritmu šance. Veličina vzniklá touto transformací se nazývá logit. Vztah definující logit je tedy následující [9] a [13]:

$$\mathrm{logit}(\pi(x)) = \ln \frac{\pi(x)}{1-\pi(x)} \tag{2.1.4}$$

Úprava, která byla výše naznačena v jednotlivých krocích, se nazývá logitová transformace. Logit již může být použit jako vysvětlovaná proměnná v logistickém regresním modelu, aniž by docházelo k numerickým problémům. Logistický regresní model pro jednu spojitou vysvětlující proměnnou bude vypadat následovně:

$$g(x) = \ln \frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 X \ . \tag{2.1.5}$$

Kde *g(x)* značí název regresní funkce vzniklé logitovou transformací pravděpodobnosti *π(x)*. Tato funkce je již na rozdíl od *π(x)* definována na celém intervalu (−∞ ; +∞) a je lineární v parametrech [9], [10] a [13].

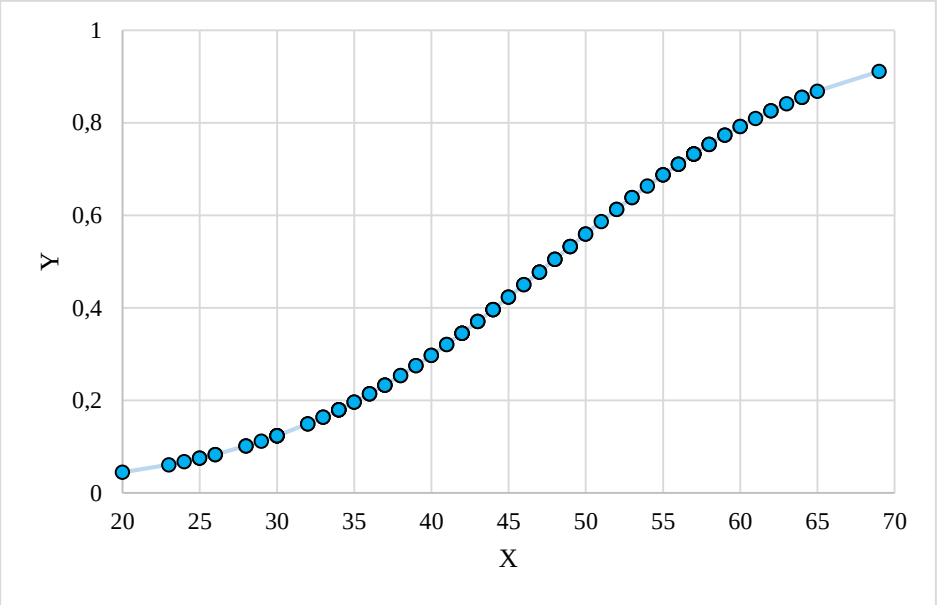
Logitová transformace může proběhnout i zpětně. Z modelu vyjádřeného vztahem (2.1.5) je tedy možné získat i hodnoty pravděpodobnosti nebo šance. Pro šanci jevu poté platí:

$$\check{s}ance(Y = 1 \mid x) = e^{\beta_0 + \beta_1 X} \ . \tag{2.1.6}$$

Podmíněnou pravděpodobnost *π(x)* lze zapsat následujícím způsobem:

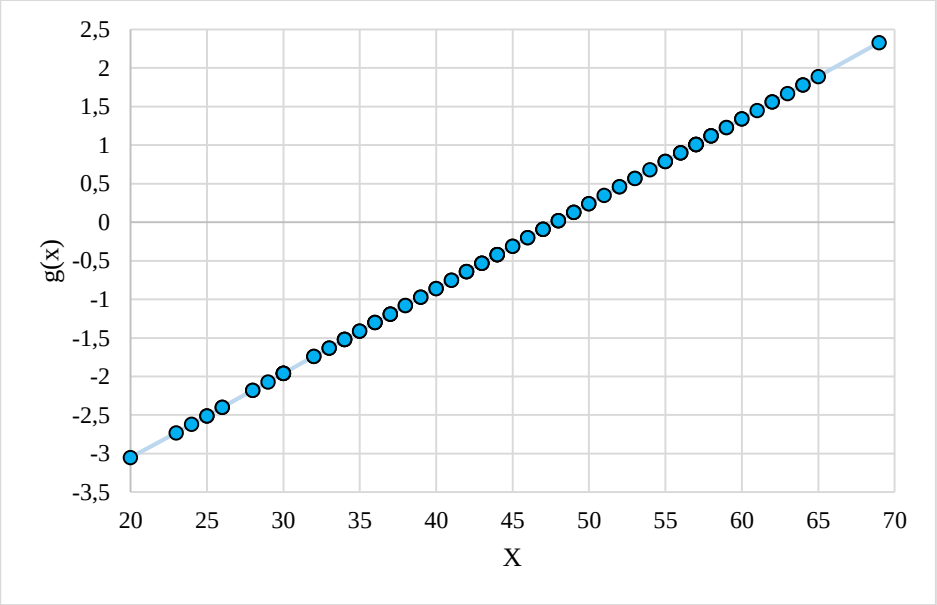
$$\pi(x) = \frac{e^{\beta_0 + \beta_1 X}}{1 - e^{\beta_0 + \beta_1 X}} = \left(1 + e^{-\beta_0 - \beta_1 X}\right)^{-1} \ . \tag{2.1.7}$$

Jak je z předcházejícího vztahu patrné, podmíněná pravděpodobnost *π(x)* je vyjádřena jako nelineární funkce vysvětlované proměnné. Tato funkce je zároveň předpisem pro logistické rozdělení. Podle tohoto rozdělení je metoda logistické regrese pojmenována. Pro grafickou představu je vztah (2.1.7) vyjádřen grafem na obrázku 2.2 [10] a [13].



Obrázek 2.2: Graf odhadnutých hodnot pravděpodobnosti π(x) v závislosti na X [5].

Pro úplnost je na obrázku 2.3 popsán vztah logitu a vysvětlující proměnné X. Na zmíněném obrázku je logit vyjádřen pomocí odhadnutých hodnot funkce $g(x)$. Je patrné, že tento vztah je lineární.



Obrázek 2.3: Graf odhadnutých hodnot logitu v závislosti na X [6].

Interpretace výsledného modelu je snadněji pochopitelná, pokud jsou výsledné hodnoty vyjádřeny pomocí šancí, nebo pravděpodobností. Proto se hodnoty logitu často zpětně převádí.

[5] *Zdroj: vlastní zpracování, MS Excel, data z [10] viz příloha A*
[6] *Zdroj: vlastní zpracování, MS Excel, data z* [10] *viz příloha A*

Při interpretaci modelu je tedy nutné dbát na to, aby nedocházelo k záměně těchto různých veličin.

2.2 Logistický regresní model

V minulém oddílu bylo objasněno použití logitu jako vysvětlované proměnné v rámci logistického regresního modelu. V předchozí části však byla uvažována pouze jedna vysvětlující proměnná spojitého typu. I když je logistickou regresi možno použít i v takovém případě, většinou se využívá její schopnosti pracovat s vícerozměrnými daty. Poté v modelu vystupuje více vysvětlujících proměnných a to jak proměnné spojité, tak i proměnné kategoriální, ty ovšem vyžadují odlišný způsob zpracování, aby mohly být do modelu zařazeny. Nyní tedy bude logistický regresní model zobecněn pro více vysvětlujících proměnných [10].

Za předpokladu, že do modelu vstupuje více vysvětlujících proměnných, je vhodné tyto proměnné vyjádřit pomocí vektoru. Pro soubor p nezávislých (vysvětlujících) proměnných bude tento vektor vypadat následovně:

$$\mathbf{x}' = (X_1, X_2, \ldots, X_p) \ . \tag{2.2.1}$$

Kde p značí počet vysvětlujících proměnných. Obecně zapsaný logistický regresní model se poté vyjádří následujícím způsobem:

$$g(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p = \beta_0 + \mathbf{x}'\mathbf{B} \ . \tag{2.2.2}$$

Kde ***B*** označuje vektor parametrů [10] a [13].

2.2.1 Model s kategoriální vysvětlující proměnnou

V případě nominálních vysvětlujících proměnných, které mají více než dvě kategorie, není možné, aby vstoupily do modelu stejným způsobem jako spojité proměnné. Číslice, pomocí kterých mohou být hodnoty kategoriálních proměnných vyjádřeny, nemají numerický význam. Mají význam pouze jako kódovací znaky jednotlivých hodnot příslušné proměnné. Aby bylo možné takové proměnné do modelu zahrnout, je nutné vytvořit takzvané indikátorové proměnné.

Jestliže má příslušná proměnná $k \geq 2$ kategorií, je jedním z nejpoužívanějších způsobů pro tvorbu indikátorových proměnných vytvoření $k - 1$ nových indikátorových proměnných. Přičemž je jedna kategorie zvolena jako referenční a všem ostatním kategoriím je přiřazena

právě jedna indikátorová proměnná. Taková indikátorová proměnná nabývá hodnot nula nebo jedna v závislosti na tom, zda sledovaná statistická jednotka (respondent) bude patřit do dané kategorie vysvětlující proměnné. Pokud do dané kategorie patřit bude, příslušná indikátorová proměnná nabyde hodnoty jedna a všechny ostatní indikátorové proměnné se budou rovnat nule. Takto vzniklé proměnné se nazývají *dummy* indikátorové proměnné. Tento způsob není jediný, například v [10] jsou popsány další postupy pro tvorbu indikátorových proměnných [9] [10] a [13].

Rovnici popisující logistický regresní model je možné upravit tak, aby bylo patrné, že obsahuje nominální vysvětlující proměnou. Jako nominální proměnná bude zvolena proměnná X_p. Příslušné indikátorové proměnné budou označeny D_{p1}, D_{p2}, …, D_{pk-1}. Logistický regresní model poté vypadá následovně:

$$g(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \sum_{j}^{k-1} \beta_{pj} D_{pj} \ . \qquad (2.2.3)$$

Z modelu zobrazeného v (2.2.3) je patrné, že nominální proměnné vyjádřené pomocí indikátorových proměnných nepřísluší pouze jeden regresní koeficient β jako v případě spojitých proměnných. Pokud se jedná o nominální proměnnou, je každé její indikátorové proměnné (každé kategorii, kromě referenční) přiřazen právě jeden regresní koeficient β. Jak bude dále zmíněno, takové koeficienty mají i odlišnou interpretaci než koeficienty náležící spojitým vysvětlujícím proměnným [9], [10] a [13].

Pokud mají do modelu vstoupit binomické vysvětlující proměnné, pro které se počet kategorií k rovná dvěma, je možné si vybrat, zda s takovou proměnnou pracovat jako se spojitou či vytvořit indikátorové proměnné. V obou případech bude v modelu vystupovat pouze jedna proměnná. V případě ordinální proměnné se pro zanesení do regresního modelu využívá primárně stejný způsob jako pro nominální proměnnou s více než dvěma kategoriemi. Pokud však vysvětlující ordinální proměnná nabývá dostatečného počtu kategorií je s ní možné pracovat jako se spojitou proměnnou. Jako dostatečný počet se například v [9] uvádí alespoň sedm kategorii [9] a [10].

2.3 Bodový odhad parametrů modelu

K získání výsledných hodnot logitu, šance či pravděpodobnosti je nejprve nutné odhadnout jednotlivé parametry, neboli regresní koeficienty, modelu. Je tedy nutné odhadnout $p + 1$ parametrů. K tomu se využívá takzvaná metoda maximální věrohodnosti. Pomocí této

metody se v podstatě hledají hodnoty neznámých parametrů, které maximalizují pravděpodobnost získání dat, která se budou shodovat s těmi pozorovanými [10].

Zásadní roli v rámci této metody hraje věrohodnostní funkce. Ta právě představuje pravděpodobnost získání pozorovaných dat jako funkci $p + 1$ parametrů. Cílem metody maximální věrohodnosti tedy je najít takové hodnoty zmíněných parametrů, které budou maximalizovat věrohodnostní funkci. Věrohodnostní funkce má za předpokladu nezávislosti jednotlivých pozorování následující tvar:

$$L(\beta) = \prod_{i}^{n} \pi(\mathbf{x}_i)^{y_i} \left[1 - \pi(\mathbf{x}_i)\right]^{1-y_i} \quad . \tag{2.3.1}$$

Kde β je vektorem parametrů, tedy $\beta' = (\beta_0, \beta_1, \ldots, \beta_p)$, n značí počet pozorování, $\boldsymbol{x}_i$ je vektor představující kombinaci vysvětlujících proměnných pro i-té pozorování a y_i vyjadřuje hodnoty binární vysvětlované proměnné, nabývá tedy hodnot nula nebo jedna, podle toho, zda sledovaný jev nastal či nenastal. Výraz $\pi(\boldsymbol{x}_i)$ udává podmíněnou pravděpodobnost nastání sledovaného jevu při dané kombinaci vysvětlujících proměnných $\boldsymbol{x}_i$ a platí pro něj následující vztah:

$$\pi(\mathbf{x}_i) = \frac{e^{\beta_0 \sum_j^p \beta_j \mathbf{x}_{ij}}}{1 - e^{\beta_0 \sum_j^p \beta_j \mathbf{x}_{ij}}} \quad . \tag{2.3.2}$$

V rovnici (2.3.1) podmíněná pravděpodobnost $\pi(\boldsymbol{x}_i)$ představuje příspěvek k věrohodnostní funkci, pokud sledovaný jev nastal. Pokud sledovaný jev nenastal, je příspěvkem k věrohodnostní funkci výraz $1 - \pi(\boldsymbol{x}_i)$. Aby se dalo s věrohodnostní funkcí z matematického hlediska lépe pracovat, využívá se logaritmus věrohodnostní funkce, který se značí $l(\beta)$ [8], [10] a [15].

Pro nalezení vhodných hodnot parametrů β je tedy nutné maximalizovat logaritmus věrohodnostní funkce $l(\beta)$. Za tímto účelem jsou nejprve vypočteny parciální derivace funkce $l(\beta)$ podle jednotlivých parametrů β. Tyto parciální derivace jsou následně postaveny rovny nule, tím vznikne soustava $p + 1$ takzvaných věrohodnostních rovnic. Každému parametru odpovídá jedna rovnice [10].

Takto vyniklé rovnice jsou nelineární v parametrech, proto je k vyřešení soustavy nutné použití iteračních metod. Jednou z nejpoužívanějších metod k nalezení řešení této soustavy je Newton-Raphsonova metoda. Aplikací této metody se získají maximálně věrohodné odhady parametrů β, které odpovídají řešením příslušných rovnic a značí se $\hat{\beta}$. Maximálně věrohodné

odhady podmíněných pravděpodobností je poté možné získat dosazením do vztahu (2.3.2) [8], [13] a [10].

2.4 Hodnocení modelu

Poté co je získán výsledný model pomocí odhadu jednotlivých parametrů, je před samotnou interpretací modelu nutné ověřit správnost modelu. Hodnocení modelu je možné chápat ve dvou rovinách. V prvním případě se jedná o testování parametrů modelu. Neboli o ověření, zda všechny vysvětlující proměnné mají statisticky významný vliv na vysvětlovanou proměnnou. Jestliže jsou všechny vysvětlující proměnné v modelu statisticky významné, je možné přistoupit k druhému způsobu hodnocení modelu, který ověřuje celkovou kvalitu modelu. Kvalita modelu se posuzuje podle toho, jak efektivně model popisuje závislost mezi vysvětlovanou a vysvětlujícími proměnnými [10].

K hodnocení logistického regresního modelu bylo vyvinuto více metod. Toto hodnocení probíhá na základě jednotlivých statistik. Zde budou zmíněny statistiky, které se objevují ve výstupu statistického výpočetního softwaru SPSS [6] ,[9].

2.4.1 Testování parametrů modelu

Testování statistické významnosti parametrů se provádí pomocí Waldovy statistiky a pomocí statistiky věrohodnostního poměru, která bývá také označována jako statistika chí-kvadrát modelu. U datových souborů s velkým počtem pozorování se na základě těchto dvou statistik dospěje k podobným výsledkům. Avšak u menších datových souborů se výsledky mohou lišit. V takových případech se dává přednost testu založenému na věrohodnostním poměru. Na Waldově statistice a věrohodnostním poměru jsou také založeny algoritmy pro automatický výběr vysvětlujících proměnných, které se využívají především u rozsáhlých datových souborů obsahujících velký počet potencionálních vysvětlujících proměnných. Tyto algoritmy budou popsány později [6], [9], [10] a [15].

Test parametrů pomocí věrohodnostního poměru testuje model jako celek a vychází z odhadu věrohodnostní funkce, ten se získá dosazením odhadnutých podmíněných pravděpodobností do vztahu (2.3.1). Tento odhad je dále nutné upravit, aby odpovídal známému statistickému rozdělení, tím se získá statistika $-2LL$ (log likehood), pro kterou tedy platí:

$$-2LL = -2\ln\left\lfloor L\left(\hat{\beta}\right)\right\rfloor . \tag{2.4.1}$$

Tato statistika má asymptoticky χ^2 rozdělení s počtem stupňů $J - (p + 1)$, kde J představuje počet různých kombinací hodnot vysvětlujících proměnných, jinými slovy počet různých hodnot vektoru $\boldsymbol{x}$ [6], [9], [10] a [15].

K testování statistické významnosti parametrů se využívá rozdílu statistiky $-2LL$ dvou různých modelů. A to modelu obsahující pouze parametr β_0 a modelu obsahujícího $p + 1$ parametrů β, který bere v potaz všechny vysvětlující proměnné. Tento rozdíl se nazývá statistika věrohodnostního poměru nebo také chí-kvadrát modelu a značí se G. Platí pro ni následující vztah:

$$G = -2\ln\left[L\left(\hat{\beta}_0\right)\right] + 2\ln\left[l\left(\hat{\beta}\right)\right] = -2\ln\left[\frac{L\left(\hat{\beta}_0\right)}{L\left(\hat{\beta}\right)}\right]. \tag{2.4.2}$$

Statistika G může nabývat hodnot z intervalu $(0; \infty)$ a má také χ^2 rozdělení se stupni volnosti odpovídajícím rozdílu počtu parametrů obou modelů. Využívá se k testování hypotézy, zda je model jako celek statisticky významný. Nulová hypotéza předpokládá, že se parametry $\beta_1, \beta_2, \ldots, \beta_{p+1}$ rovnají nule, podobně jako u F-testu v lineární regresi. V takovém případě je model jako celek nevýznamný, tato situace nastává, pokud je hodnota statistiky G blízká nule. Výsledek testu se určí na základě p-hodnoty. Pokud je výsledná p-hodnota testu menší než stanovená hladina významnosti α, nulová hypotéza se zamítá. To znamená, že alespoň jeden ze zmíněných parametrů se nerovná nule a že model je jako celek statisticky významný. Jinými slovy zamítnutí nulové hypotézy znamená, že zahrnutí příslušných vysvětlujících proměnných do modelu umožňuje lepší predikci hodnot závislé proměnné, než jaká by byla možná bez nich [6], [9], [10], [13] a [15].

Další způsob jak ověřit statistickou významnost vysvětlujících proměnných je pomocí Waldova testu. Ten na rozdíl od předešlého testu zkoumá statistickou významnost individuálně. Pomocí Waldovy statistiky se testuje nulová hypotéza, že některý z parametrů β_j je roven nule, podobně jako v případě t-testů v lineární regresi. Pro $j = 1, 2, \ldots, p+1$ má testová statistika následující tvar:

$$W_j = \frac{\hat{\beta}_j}{S\hat{E}\left(\hat{\beta}_j\right)}. \tag{2.4.3}$$

Waldova statistika představuje poměr odhadnuté hodnoty příslušného parametru a jeho směrodatné chyby. Jak je ze vztahu (2.4.3) patrné, Waldova statistika se počítá pro každý

parametr modelu j zvlášť. Tato statistika se řídí normovaným normálním rozdělením. Rozhodnutí o statistické významnosti daného parametru opět probíhá na základě p-hodnoty. Je-li tedy p-hodnota menší než příslušná hladina významnosti, nulová hypotéza se zamítá a parametr β_j je možné považovat za statisticky významný. Tudíž příslušná vysvětlující proměnná má statisticky významný vliv na hodnotu vysvětlované proměnné. Pokud nedojde k zamítnutí nulové hypotézy, příslušná proměnná by měla být z modelu odstraněna, pokud pro její zařazení do modelu neexistují jiné objektivní důvody [10], [13] a [15].

2.4.2 Hodnocení kvality modelu

Při hodnocení kvality modelu, je zkoumán model jako celek, ne jeho jednotlivé komponenty. Kvalitou je myšlena schopnost efektivně predikovat hodnoty závislé proměnné pomocí vysvětlujících proměnných na základě pozorovaných dat. K hodnocení kvality modelu mohou být použity různé metody, některé z nich budou zde popsány. Kvalita modelu může být popsána klasifikační tabulkou, ROC křivkou, nebo různými statistikami, například pomocí Cox-Snellova koeficientu determinace, Nagelkerkeova koeficientu determinace, anebo pomocí statistiky $-2LL$ [10].

Statistika $-2LL$ se spíše než k hodnocení konkrétního modelu používá k porovnání dvou různých modelů. Přičemž platí, že nižší hodnoty statistiky $-2LL$ značí kvalitnější model. Jelikož hodnota této statistiky je ovlivněna počtem parametrů v modelu, dochází k upřesňování složitějších modelů neboli modelů s více parametry. Proto se používají i modifikace této statistiky, které již berou počet parametrů v potaz. Například Akaikeho informační kritérium AIC, pro které platí následující vztah [9], [13] a [15]:

$$AIC = -2LL + 2(p+1). \tag{2.4.4}$$

Cox-Snellova statistika stejně jako Nagelkerkeova statistika je obdobou koeficientu determinace v lineární regresi. Hodnoty těchto statistik je možné chápat jako míru snížení neurčitosti v datech, kterého bylo dosaženo pomocí zkonstruovaného modelu. Vyšší hodnota tedy znamená větší míru snížení neurčitosti a zároveň kvalitnější model. Cox-Snellova statistika je definována takto:

$$R^2_{CS} = 1 - \left[\frac{l(\hat{\beta}_0)}{l(\hat{\beta})}\right]^{\frac{2}{n}}. \tag{2.4.5}$$

Kde *n* je počet pozorování a věrohodností funkce jsou stejné jako v rovnici (2.4.2)**Chyba! Nenalezen zdroj odkazů.**. Maximální hodnota, které může Cox-Snellova statistika dosáhnout je následující:

$$\max\left(R_{CS}^2\right) = 1 - \left[l\left(\hat{\beta}_0\right)\right]^{\frac{2}{n}}. \tag{2.4.6}$$

Nagelkerkeova statistika je modifikací předcházející statistiky a má následující tvar:

$$R_N^2 = \frac{R_{CS}^2}{\max(R_{CS}^2)}. \tag{2.4.7}$$

Modifikace spočívá v upravení intervalu hodnot, kterých může statistika nabývat. Nagelkerkeova statistika na rozdíl od Cox-Snellovy nabývá pouze hodnot mezi nulou a jedničkou, proto je lépe interpretovatelná [6], [8], [9] a [13].

Dalším nástrojem pro hodnocení modelu je klasifikační tabulka. Ta porovnává pozorované a modelem odhadnuté hodnoty vysvětlované proměnné. V následující tabulce je zobrazen příklad klasifikační tabulky.

Tabulka 2.1: Příklad klasifikační tabulky.

Pozorované hodnoty		Predikované hodnoty	
		Y	
		1	0
Y	1	a	b
	0	c	d

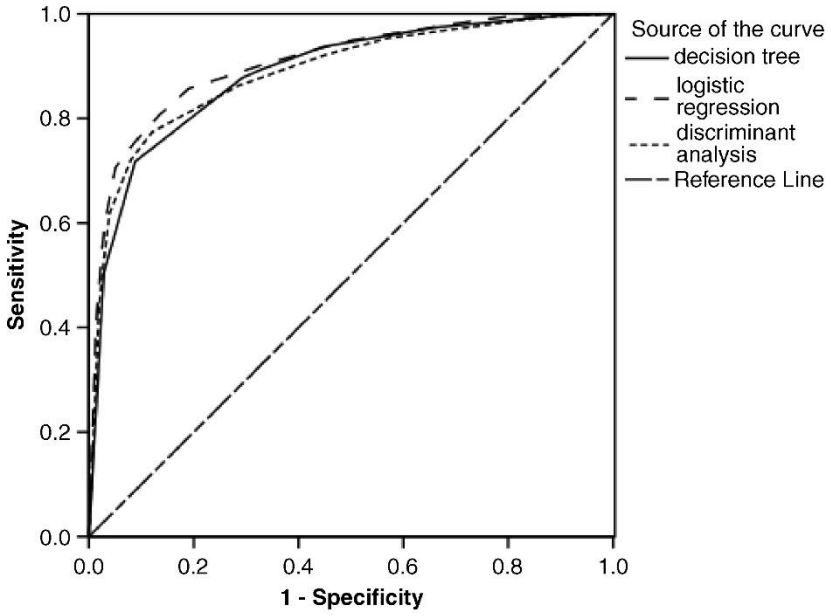
Zdroj: Vlastní zpracování, MS Excel.

Predikované hodnoty závislé proměnné jsou určeny na základě odhadnutých podmíněných pravděpodobností. Pro odhady podmíněných pravděpodobností je určena jistá hranice (většinou 0,5), podle které jsou do kategorií vysvětlované proměnné zařazeny jednotlivé případy. Pokud je zmíněná hranice zvolena jako 0,5, poté případy, jimž náleží odhadnuté podmíněné pravděpodobnosti vyšší než 0,5, budou zařazeny do kategorie odpovídající nastání jevu. Případy s podmíněnými pravděpodobnostmi nižšími než zmíněná hranice, budou naopak zařazeny do kategorie, která značí nenastání příslušného jevu [8] a [10].

Pro binární vysvětlovanou proměnnou má klasifikační tabulka rozměr 2x2. Tabulka zobrazuje počty případů, které odpovídají jednotlivým hodnotám vysvětlované proměnné. Ve sloupcích jsou zobrazeny predikované počty případů a v řádcích pozorované počty případů. Na

hlavní diagonále se nachází počty případů, které byly modelem klasifikovány správně. Na základě podílu správně zařazených případů do jednotlivých kategorií, se pomocí klasifikační tabulky určuje jak efektivně je model schopen klasifikovat jednotlivé případy. Tento podíl se nazývá celková úspěšnost a má následující tvar $(a + d) / (a + b + c + d)$. [6], [9] a [10].

Další nástroj pro hodnocení modelu souvisí s klasifikační tabulkou a je jím ROC (reciever operating characteristic) křivka. Jak název napovídá, tato metoda pochází původně z teorie detekce signálu. Graf ROC křivky je konstruován na základě takzvané senzitivity a specificity. Senzitivita udává pravděpodobnost správné predikce nastání jevu, který sleduje vysvětlovaná proměnná. V souvislosti s klasifikační tabulkou zobrazenou v tabulce 2.1 senzitivita odpovídá podílu $a / (a + c)$. Specificita odpovídá pravděpodobnosti správné detekce nenastání jevu příslušným modelem. Z příkladu klasifikační tabulky se vypočte jako $d / (b + d)$. Z toho vyplývá, že senzitivita i specificita jsou závislé na zvolené hranici pro odhady podmíněné pravděpodobnosti zmíněné u popisu klasifikační tabulky. V grafu ROC křivky je senzitivita zobrazena na ose *y* a osa *x* udává hodnoty odpovídající výrazu 1 – specificita. Příklad ROC křivky je uveden na obrázku 2.4.



Obrázek 2.4: Příklad ROC křivky[7]

Body ROC křivky se získají tak, že pro jednotlivé hodnoty hranice pravděpodobnosti (na základě které se získávají odhadnuté hodnoty závislé proměnné pro klasifikační tabulku, viz výše) z intervalu od nuly do jedné je vypočtena senzitivita a 1 – specificita. Hodnoty senzitivity

[7] Zdroj: [8]

a 1 – specificity poté představují souřadnice příslušného bodu. Další body ROC křivky se získají vypočtením senzitivity a 1 – specificity pro další hodnoty zmíněné hranice pravděpodobnosti [8] a [10].

Model se hodnotí na základě plochy pod ROC křivkou. Čím je plocha větší, tím lépe model klasifikuje jednotlivé případy. Maximální hodnota plochy pod ROC křivkou je rovna jedné [10].

Tyto dva způsoby pro měření kvality modelu, tedy klasifikační tabulka a ROC křivka, nejsou vlastní pouze logistické regresi, ale dají se aplikovat i na jiné klasifikační modely. Umožňují vlastně srovnání modelů vzniklých pomocí různých metod [8].

2.5 Výběr vysvětlujících proměnných

Správným výběrem vysvětlujících proměnných je nutné se zabývat v případě datových souborů obsahujících velký počet vysvětlujících proměnných, které by mohly být potencionálně zařazeny do modelu. Je potřeba vybrat vysvětlující proměnné tak, aby model nebyl příliš složitý a přesto aby efektivně predikoval hodnoty vysvětlované proměnné. Složitější modely obsahujících větší počet vysvětlujících proměnných bývají více závislé na příslušném vzorku dat a nemusí spolehlivě odrážet realitu [10].

Rozhodnutí o zahrnutí jednotlivých vysvětlujících proměnných do modelu může provést sám uživatel, a to buď intuitivně, nebo na základě analýzy vztahu závislé proměnné a jednotlivých potencionálních vysvětlujících proměnných. Avšak statistické softwary nabízejí procedury, které zajišťují automatický výběr vysvětlujících proměnných. Tyto procedury jsou označovány jako postupná regrese (*stepwise regression*) [6] a [10].

Statistický software SPSS nabízí dva způsoby automatického výběru vysvětlujících proměnných. Oba tyto způsoby jsou založeny na algoritmech používající stejné statistické testy. Zmíněné algoritmy se liší v podstatě pouze počáteční fází a pořadím dalších kroků. Jednotlivé fáze obou algoritmů jsou podrobněji rozebrány v [6].

V příslušných fázích algoritmů se pro posouzení, zda by proměnná měla vstoupit do modelu, se používá takzvaný *score test*, který je podrobněji popsán v [6]. Rozhodnutí je provedeno na základě p-hodnoty a zvolené hladiny významnosti α_E. V SPSS je standardně pro zařazení vysvětlující proměnné do modelu nastavena hodnota hladiny významnosti $\alpha_E = 0{,}05$. Navzdory tomu se v [10] uvádí, že tato hodnota je příliš striktní a je vhodné zvolit hodnotu

hladiny významnosti α_E vyšší, a to 0,15 nebo 0,20. Pokud bude příslušná p-hodnota nižší než hladina významnosti α_E, je možné uvažovat o zařazení dané proměnné do modelu [6] a [10].

Oba algoritmy dále obsahují test pro vyřazení potencionální proměnné z modelu. V SPSS má uživatel na výběr ze tří statistik na základě, kterých test bude v rámci algoritmu probíhat. A to na základě Waldovy statistiky nebo věrohodnostního poměru, které byly již popsány. Dále pak na základě takzvané podmíněné statistiky (*conditional statistic*), která je blíže popsána v [6]. K rozhodnutí o vyřazení potencionální proměnné z modelu dochází na základě příslušné p-hodnoty a zvolené hladiny významnosti α_R. Hladina významnosti pro vyřazení proměnné z modelu α_R musí vždy nabývat vyšší hodnoty než hladina významnosti α_E. Pokud je výsledná p-hodnota větší než hladina významnosti α_R, příslušná vysvětlující proměnná nebude do modelu zahrnuta [6] a [10].

První způsob označen jako *forward stepwise selection* v první fázi algoritmu pracuje s modelem obsahujícím pouze parametr β_0. Poté na základě score testu se postupně zařazují vhodné potencionální vysvětlující proměnné do modelu. Zároveň se postupně ověřuje, zda proměnné, které byly do modelu zařazeny, odpovídají i testu založenému na jedné ze tří zmíněných statistik. Vysvětlující proměnné, které splní kritéria obou testů, jsou zařazeny do výsledného modelu [6] a [10].

Druhý způsob je označený jako *backward stepwise selection*. Algoritmus začíná modelem obsahujícím všechny potencionální vysvětlující proměnné. Poté na základě jedné ze tří vybraných statistik jsou jednotlivé potencionální proměnné vyřazovány. Zároveň proměnné, které předešlým testem nebyly vyřazeny, musejí být ještě ověřeny score testem. Pokud příslušná vysvětlující proměnná nebyla z počátečního modelu vyřazena prvním testem a zároveň splnila kritérium score testu, může být zařazeno do výsledného modelu [6].

2.6 Interpretace modelu

Interpretce modelu spočívá v popsání významu hodnot odhadnutých parametrů β. Interpretace parametrů β příslušejících vysvětlujícím proměnným se mírně liší v závislosti na tom, zda se jedná o proměnnou spojitou či kategoriální. K samotné interpretaci modelu dochází poté, co byla vyhodnocena statistická významnost jednotlivých vysvětlujících proměnných, a došlo ke zhodnocení kvality celého modelu [10].

V logistickém regresním modelu představuje závislou proměnnou logit, jak je popsáno v oddílu 2.1. Proto parametr β_0 představuje hodnotu logitu v případě kdy se všechny spojité vysvětlující proměnné modelu rovnají nule a kategoriální proměnné nabývají svých referenčních kategorií. Z hlediska interpretace nemá parametr β_0 takový význam jako parametry $\beta_1, \beta_2, \ldots, \beta_{p+1}$ [9] a [10].

V případě spojité nezávislé proměnné je možné hodnotu parametru β_j pro $j = 1, 2, \ldots, p+1$ interpretovat jako změnu logitu odpovídající jednotkové změně příslušné vysvětlující proměnné X_j za předpokladu, že se hodnota ostatních vysvětlujících proměnných nezmění. Změna logitu představuje rozdíl logitu po jednotkové změně proměnné X_j a logitu před touto změnou. Pro model obsahující pouze jednu spojitou vysvětlující proměnou, je změna logitu definována následovně:

$$\beta_1 = \ln\left(\frac{\pi(x_1)}{1-\pi(x_1)}\right) - \ln\left(\frac{\pi(x_0)}{1-\pi(x_0)}\right). \qquad (2.6.1)$$

Kde x_1 je hodnota vysvětlující proměnné po jednotkové změně a x_0 je její hodnota před změnou. Model je samozřejmě výhodnější interpretovat pomocí šancí. Proto je nutné rovnici (2.6.1) upravit. Výsledkem úpravy je následující vztah:

$$e^{\beta_1} = \frac{\pi(x_1)/[1-\pi(x_1)]}{\pi(x_0)/[1-\pi(x_0)]}. \qquad (2.6.2)$$

Výraz na pravé straně rovnice (2.6.2) se nazývá poměr šancí. Jak vyplývá ze vztahu (2.1.6) poměr šancí udává, kolikrát se zvýší šance nastání jevu, při jednotkové změně spojité vysvětlující proměnné. Stejný způsob interpretace platí i pro modely obsahující více spojitých proměnných Pokud je odhadnutá hodnota parametru β_j záporná, šance nastání jevu se při jednotkové změně příslušné proměnné X_j sníží [9], [10], [11] a [14].

Interpretace parametrů β_{ij} se u kategoriální vysvětlující proměnné liší. Způsob interpretace navíc závisí i na způsobu jakým byly indikátorové proměnné vytvořeny. Zde bude uvažován způsob tvorby indikátorových proměnných, který je popsán v oddílu 2.2.1.V případě, že vysvětlující proměnná X_j na místo své referenční kategorie nabude kategorie i, kterou určuje daný parametr β_{ij}. Poté odhad tohoto parametru udává, jak se za této situace změnila hodnota logitu. Poměr šancí udává, kolikrát se změní šance nastání sledovaného jevu, pokud kategoriální vysvětlující proměnná nabyde dané kategorie na místo své kategorie referenční. Poměr šancí pro referenční kategorii je roven jedné. Poměr šancí je u kategoriální proměnné v

podstatě definován podobně jako ve vztahu (2.6.2) s tím rozdílem, že x_0 představuje referenční kategorii a x_1 jakoukoliv kategorii nereferenční [9] [10].

Odhadnuté hodnoty parametrů β, stejně tak odhady poměru šancí bývají zobrazeny ve výstupech většiny statistických softwarů [10].

2.7 Výhody a nevýhody logistické regrese

Nyní budou ve stručnosti popsány některé výhody a nevýhody použití metody logistické regrese.

Jednou z výhod logistické regrese je výstup v podobě matematického modelu. Ten umožňuje kvantifikaci vztahu závislé proměnné na ostatních vysvětlujících proměnných, který není obtížné srozumitelně interpretovat pomocí poměru šancí. Poměr šancí jasně popisuje přímý vliv jednotlivých vysvětlujících proměnných na závislou proměnnou. Z matematického modelu je možné také vyčíst statistickou významnost jednotlivých vysvětlujících proměnných. Logistická regrese navíc nabízí více možností pro hodnocení kvality modelu oproti rozhodovacím stromům [8].

Logistický regresní model dokáže pracovat s vysvětlujícími proměnnými jakéhokoliv typu. V případě spojité proměnné není nutné její hodnoty rozdělovat do intervalů jako v případě metody CHAID. Co se týče vysvětlované proměnné, binární logistický model je možné zobecnit pro proměnné kategoriální s větším počtem kategorií i pro proměnné ordinální [10].

Další výhodou je skutečnost, že logistická regrese nabízí možnost automatického postupného výběru vysvětlujících proměnných.

Výstup ve formě matematického modelu může představovat i jistou nevýhodu, v tom smyslu, že může být nesrozumitelný pro osoby s nedostatečnou znalostí dané problematiky. V tomto ohledu mají rozhodovací stromy výhodu.

Další nevýhoda souvisí se spojitými vysvětlujícími proměnnými. Logistická regrese nedokáže pracovat s chybějícími hodnotami spojité vysvětlující proměnné. Navíc je citlivá na extrémní hodnoty spojitých proměnných. Avšak těmto komplikacím se nechá předejít rozdělením spojité proměnné do intervalů a vytvoření zvláštní kategorie pro chybějící hodnoty. Poté, ale dochází ke ztrátě části informace obsažené v datech, jak již bylo zmíněno v kapitole týkající se metody CHAID [8].

3 Aplikace metod na reálných datech

V této části práce budou zmíněnými metodami sestaveny modely pro data získaná z Českého sociálněvědního datového archivu Sociologického ústavu Akademie věd ČR, v. v. i., dále jen ČSDA.

3.1 Popis dat

Datový soubor byl poskytnut ČSDA na základě registrace do systému datového katalogu ČSDA[8]. Data pocházejí z výzkumu Sociologického ústavu AV ČR „Naše společnost 2016“ a byla získána pomocí terénního sběru dat na základě kvótního výběru.

Datový soubor obsahuje celkem 1047 pozorování a 201 proměnných. Za vysvětlující byla zvolena proměnná s kódovým označením *ER_42*, která zjišťuje, zda si příslušný respondent, či jeho domácnost vytváří finanční rezervu. Tato proměnná byla následně překódována, závislá proměnná, která bude v modelech vystupovat, ponese označení *Y*. Během dotazníkového šetření 2 respondenti odmítli odpovědět a 33 respondentů zvolilo odpověď „*nevím*“. Těchto 35 zmíněných případů budou považovány za neplatné.

V případě, že nejsou uvažovány neplatné odpovědi, nabývá vysvětlovaná proměnná *Y* pouze hodnot *„ano“* a „*ne*“. Jedná se tedy o binární proměnnou. Hodnoty závislé proměnné představují odpovědi na otázku, zda si respondent, nebo jeho domácnost vytváří finanční rezervu. Kladná odpověď znamená nastání sledovaného jevu a v datovém souboru je kódována jedničkou, záporná odpověď představuje nenastoupení sledovaného jevu a je kódována nulou.

Tabulka 3.1: Absolutní a relativní četnosti vysvětlované proměnné.

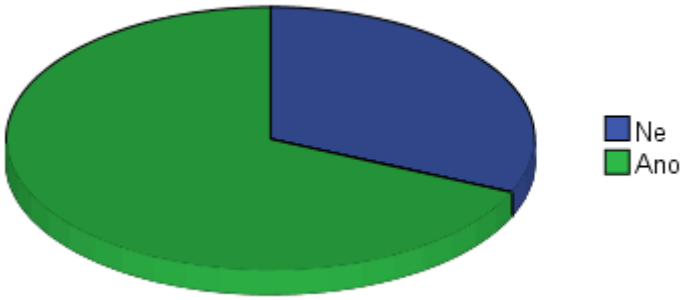
Y	Absolutní četnosti	Relativní četnosti [%]
ne (0)	320	31,6
ano (1)	692	68,4
Celkem	1012	100

Zdroj: Vlastní zpracování, MS Excel.

Absolutní a relativní četnosti jednotlivých kategorií vysvětlované proměnné jsou zobrazeny v tabulce 2.1. Relativní četnosti jsou zároveň vyjádřeny graficky na obrázku 3.1. Z tabulky i z grafu je patrné, že finanční rezervu si vytváří přibližně dvě třetiny respondentů. Zda se

[8] http://nesstar.soc.cas.cz/webview/

toto rozdělení četností mění v závislosti na vybraných vysvětlujících proměnných, bude zkoumáno v dalších oddílech této práce.



Obrázek 3.1: Graf relativních četností závislé proměnné *Y*[9]

Z datového souboru bylo dále intuitivně vybráno 15 proměnných, které by na zmíněnou závislou proměnnou mohly mít vliv, viz tabulka 3.2. Tyto proměnné budou uvažovány jako vysvětlující. Vytvořené modely následně odhalí, zda je vliv na závislou proměnnou statisticky významný a popřípadě i jak je silný. Při konstrukci modelů, byla u některých vysvětlujících proměnných problémová kategorie popisující odpověď „*nevím*“, jelikož do této kategorie spadalo velmi málo případů, většinou jen v řádu jednotek. Z toho důvodu u proměnných, kde absolutní četnost zmíněné kategorie byla nižší než 20 případů, byla odpověď „*nevím*“ označena jako neplatná. Přestože se tak snížil celkový počet pozorování na 937, u logistické regrese to zapříčinilo pokles Akaikeho informační kritéria, což značí zlepšení modelu. Kategorie „*nevím*“ byla označena za neplatnou odpověď u proměnných s následujícími kódovými označeními *EV*_10, *ER*_35*D*, *ER*_44*AA*, *ER*_44*AC*, *EU*_179*A*, *IDE*_3*A*, *t_ZIVUR*. U ostatních proměnných, kde se odpověď „*nevím“* objevuje, představuje samostatnou kategorii, jelikož je zastoupena dostatečným počtem pozorování.

Tabulka 3.2 zobrazuje výpis jednotlivých vysvětlujících proměnných, jejich kódové označení v rámci datového souboru, dále jejich typ a skutečnosti, které popisují. Hodnoty proměnných představují dopovědi na otázky získávané v rámci dotazníkového šetření. Hodnoty, kterých mohou jednotlivé vysvětlující kategoriální proměnné nabývat, jsou popsány v tabulce v rámci přílohy *B*.

[9] *Zdroj: vlastní zpracování, SPSS, data z* [16].

Tabulka 3.2: Výpis vysvětlujících proměnných.

Kód	Typ	Význam
EV_10	Ordinální	Hodnocení současné ekonomické situace v ČR
EV_11	Ordinální	Jak se změní ekonomická situace v ČR v tomto roce
ER_35D	Ordinální	Je půjčování peněz riziko
ER_35E	Ordinální	Vyplatí se mít dluhy
ER_38A	Ordinální	Hodnocení míry zadlužení občanů ČR
ER_39	Ordinální	Je míra zadlužení občanů ČR problém
ER_44AA	Binární	Splácí hypotéku
ER_44AC	Binární	Splácí zboží na úvěr
EU_179A	Binární	Byl nezaměstnaný déle než 6 měsíců
IDE_3A	Nominální	Rodinný stav
IDE_13	Spojitá	Počet členů domácnosti (v případě CHAID ordinální)
IDE_2	Spojitá	Věk (v případě logistické regrese)
t_VEK_5	Ordinální	Věk (v případě metody CHAID)
t_ZIVUR	Ordinální	Životní úroveň
t_EA_a_2	Binární	Ekonomická aktivita
t_VZD	Ordinální	Dosažená úroveň vzdělání

Zdroj: Vlastní zpracování, MS Excel.

3.2 Metoda CHAID

Nyní na datech z [16] bude předvedena analýza vztahu vysvětlované proměnné *Y* a vybraných vysvětlujících proměnných pomocí metody CHAID. Výstupem bude tedy rozhodovací strom sestrojený touto metodou v statistickém softwaru SPSS.

Jelikož metoda CHAID dokáže pracovat s chybějícími (neplatnými) hodnotami vysvětlujících proměnných, bude do modelu zahrnuto všech 1012 platných pozorování vysvětlované proměnné. Na druhou v rámci této metody nelze použít vysvětlující proměnné spojitého typu. Proto na místo proměnné *IDE_2* popisující věk bude použita proměnná *t_VEK*_5, která věk respondentů popisuje pomocí pěti intervalů. Jednotlivé intervaly jsou popsány v tabulce v příloze *B*. Dále proměnná *IDE*_13 popisující počet členů domácnosti bude uvažována jako ordinální, není nutné ji převádět na jednotlivé intervaly, neboť nabývá jen sedmi hodnot.

3.2.1 Nastavení kritérií algoritmu

Poté, co jsou určena vysvětlovaná a vysvětlující proměnné (prediktory), které budou vstupovat do algoritmu metody CHAID, je nutné nastavit vhodné hodnoty kritérií, podle nichž bude algoritmus probíhat.

Pro první fázi algoritmu je nutné zvolit hladinu významnosti Pearsonova chí-kvadrát testu o nezávislosti v kontingenční tabulce, na jehož základě se rozhodne o případném sloučení kategorií jednotlivých vysvětlujících proměnných, tak jak je popsáno v oddíle 1.2.1. V tomto případě byla zvolena hodnota hladiny významnosti 0,05. Jestliže bude p-hodnota Pearsonova chí-kvadrát testu vyšší než zmíněné hladina významnosti, bude možné příslušné kategorie, pro které je test prováděn, sloučit.

Aby mohla správně proběhnout druhá fáze algoritmu, je nutné zvolit hladinu významnosti pro Pearsonův chí-kvadrát test, jehož výsledky rozhodují o tom, podle kterého prediktoru bude rozdělen příslušný rodičovský uzel. Hladina významnosti byla zvolena shodně jako v předešlém případě, tedy 0,05. Dále je třeba nastavit, aby výsledné p-hodnoty Pearsonova chí-kvadrát testu byly upraveny pomocí Bonferroniho multiplikátorů. Poté výsledky testu nebudou ovlivněny původním počtem kategorií prediktorů. Příslušný rodičovský uzel poté bude rozdělen na základě prediktoru s nejnižší upravenou p-hodnotou, která zároveň musí být nižší než zvolená hladina významnosti.

Další kritéria se týkají třetí fáze algoritmu. Určují, za jakých podmínek se má algoritmus zastavit. Hodnoty těchto kritérií byly nastaveny v souladu se standardním nastavením výpočetního programu SPSS. Maximální hloubka stromu neboli maximální počet úrovní větvení byl nastaven na hodnotu 3. Dalšími kritérii jsou minimální počet pozorování pro rodičovský uzel a pro uzel typu potomek. U rodičovského uzlu byl nastaven minimální počet pozorování 100 a uzlu typu potomek 50. Pro nižší hodnoty těchto kritérií by výsledky zobrazené ve formě rozhodovacího stromu mohly být méně spolehlivé.

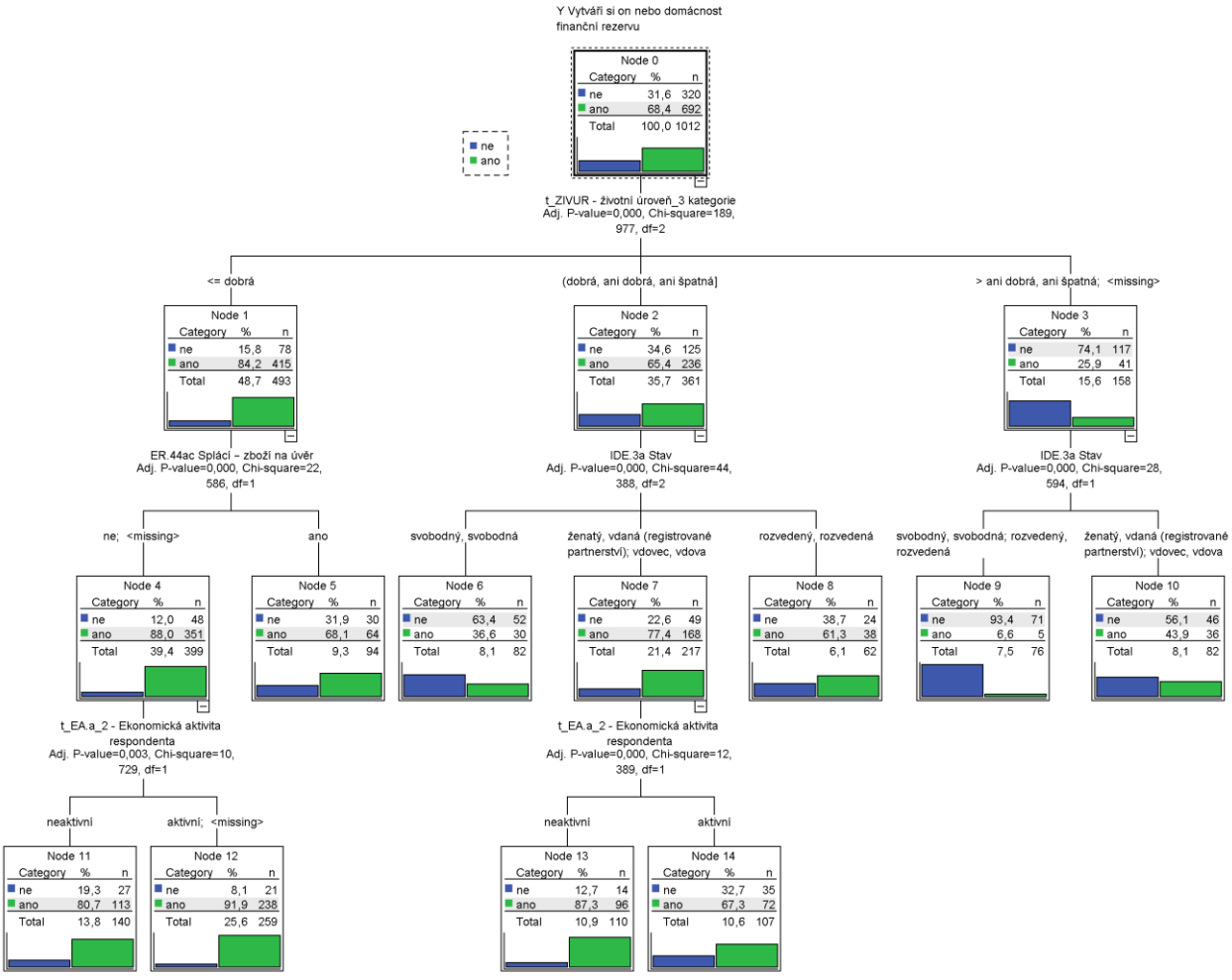
Poslední atribut, který lze před spuštěním algoritmu nastavit, je znovurozdělování kategorií, které vznikly sloučením tří a více původních kategorií. Tato možnost, ale do modelu konstruovaném v této práci, zahrnuta nebyla.

3.2.2 Interpretace

Pokud jsou správně zadány všechny zmíněné parametry, nic již nebrání spuštění algoritmu. Ten proběhne způsobem, který je popsán v oddílu 1.2. Statistický software SPSS po výpočtu algoritmu nabízí výstup v podobě rozhodovacího stromu zobrazeného na obrázku 3.2. Tento rozhodovací strom u každého uzlu zobrazuje absolutní i relativní četnosti hodnot

vysvětlované proměnné, a to i graficky. Pro lepší orientaci jsou jednotlivé uzly[10] očíslovány, kořen stromu je označen jako uzel č. 0.

Kořen stromu popisuje rozdělení hodnot závislé proměnné *Y* v celém datovém souboru. Jak je patrné, finanční rezervu si vytváří přibližně dvě třetiny dotázaných domácností. Kořen stromu se dále větví. To znamená, že v souboru vysvětlujících proměnných byla nalezena proměnná taková, která statisticky významně diferencuje hodnoty závislé proměnné. Hodnoty vysvětlované proměnné svými kategoriemi nejlépe diferencuje vysvětlující proměnná *t_ZIVUR* popisující životní úroveň dotázaných domácností. Rozhodovací strom udává i upravenou p-hodnotu a hodnotu statistiky *G*, na jejímž základě byla pomocí Pearsonova chí-kvadrát testu právě proměnná *t_ZIVUR* vybrána k rozvětvení kořene stromu.



Obrázek 3.2: Rozhodovací strom sestrojený metodou CHAID[11]

[10] V rozhodovacím stromu jsou uzly označeny anglickým výrazem „*node*“
[11] *Zdroj: Vlastní zpracování, SPSS, data z* [16]

Během větvení kořene stromu došlo ke sloučení některých kategorií proměnné *t_ZIVUR*. Konkrétně byla sloučena kategorie, která představovala špatnou životní úroveň s kategorií zastupující neplatné dopovědi. Neplatné odpovědi jsou v rozhodovacím stromu označeny jako „*missing*“. Skutečnost, že tyto kategorie byly sloučeny, znamená, že rozdělení hodnot vysvětlované proměnné v rámci těchto kategorií nejsou statisticky významně odlišná. Žádné další kategorie proměnné *t_ZIVUR* sloučeny nebyly. Kořen stromu se tedy větví na další tři uzly, podle kategorií představující životní úroveň dotázaných domácností. Jednotlivé uzly v první úrovni větvení představují domácnosti s dobrou životní úrovní, domácnosti s ani dobrou, ani špatnou životní úrovní a domácnosti kde životní úroveň nebyla zjištěna, nebo je špatná. Z obrázku 3.2 je patrné, že u domácností s dobrou životní úrovní je vyšší tendence vytvářet si finanční rezervu, vytváří si ji 84,2 % dotázaných. U domácností jejichž životní úroveň není hodnocena ani jako dobrá, ani jako špatná, je rozdělení četností závislé proměnné podobné tomuto rozdělení v kořenovém uzlu. V uzlu představujícím špatnou a nezjištěnou životní úroveň je relativní četnost domácností vytvářející si finanční rezervu velmi nízká oproti základnímu souboru. V této podskupině si finanční rezervu vytváří pouze necelých 26 % dotázaných.

Všechny uzly zmíněné v minulém odstavci se dále větví. Uzel představující domácnosti s dobrou životní úrovní je rozvětven podle proměnné *ER_44AC*, která zjišťuje, zda domácnost splácí nějaké zboží na úvěr. U této proměnné došlo ke sloučení kategorie neplatných odpovědí a kategorie „*ne*“. Uzel č. 1 je tedy rozvětven podle proměnné *ER_44AC* na domácnosti splácející zboží na úvěr a na domácnosti, které zboží na úvěr nesplácejí, nebo se u nich tuto skutečnost nepodařilo zjistit. U domácnosti s dobrou životní úrovní splácející zboží na úvěr je podíl domácností vytvářející si finanční rezervu výrazně nižší než u domácností, které zboží na úvěr nesplácejí, přibližně o 20 procentních bodů. K dalšímu větvení dochází jen u uzlu č. 4 představujícím domácnosti s dobrou životní úrovní, které nesplácejí žádné zboží na úvěr, nebo u nich nebyl tento údaj zjištěn. Uzel č. 4 je dále rozdělen na základě proměnné *t_EA_a_2* zjišťující ekonomickou aktivitu respondenta. Ekonomicky aktivní jsou sloučeni s kategorií pro nezjištěné případy. V rámci této větve stromu je podíl respondentů vytvářejících si finanční rezervu vysoký jak u ekonomicky aktivních, tak i u ekonomicky neaktivních. Uzly vzniklé posledním zmíněným větvením se již dále nevětví, neboť bylo dosaženo maximální možné úrovně větvení rozhodovacího stromu.

Pro uzel č. 2 představující domácnosti, u kterých není životní úroveň ani dobrá, ani špatná, nejlépe diferencuje rozdělení hodnot závislé proměnné *Y* vysvětlující proměnná

IDE_3_a. Tato proměnná sleduje rodinný stav respondenta. Zde opět došlo ke sloučení dvou kategorií, a to kategorie „*ženatý, vdaná (registrované partnerství)*“ s kategorií „*vdovec, vdova*“. Zajímavý je vysoký podíl respondentů nevytvářející si finanční rezervu u kategorie „*svobodný, svobodná*“ tento podíl má hodnotu 63,4 %, což je oproti podílu za celý datový soubor velký rozdíl. Na této úrovni stromu se dále větví jen uzel č. 7 popisující kategorii „*ženatý, vdaná (registrované partnerství)*“. Uzel č. 7 se větví podle ekonomické aktivity respondentů. Zde již k žádnému slučování nedochází. Z rozhodovacího stromu je patrné, že podíl respondentů vytvářejících si finanční rezervu, kteří jsou ekonomicky neaktivní, ženatí, vdaní, nebo v registrovaném partnerství a jejich životní úroveň není ani dobrá ani špatná je o 20 procentních bodů vyšší než podíl u ekonomicky aktivních respondentů se stejnými vlastnostmi.

K poslednímu větvení dochází u uzlu č. 3, který představuje respondenty z domácností se špatnou životní úrovní, nebo respondenty u nichž životní úroveň nebyla zjištěna. Zde dochází opět k větvení na základě rodinného stavu respondenta, ale v tomto případě byly kategorie sloučeny jiným způsobem. Po sloučení vznikly jen dvě nové kategorie. Jedna představuje svobodné a rozvedené respondenty a druhá respondenty patřící do kategorií „*ženatý, vdaná (registrované partnerství)*“ a „*vdovec vdova*“. Již v uzlu č. 3 byl vysoký podíl respondentů, kteří si nevytvářejí finanční rezervu, ale jak je z rozhodovacího stromu patrné, velmi k tomu přispívají svobodní a rozvedení respondenti z uzlu č. 9.

Rozhodovací strom nebyl již dále větven, jelikož byla splněna některá z podmínek pro zastavení algoritmu. Strom by se pravděpodobně rozvětvil více, pokud by byly zvoleny nižší minimální počty pozorování pro příslušné typy uzlů. To by však vzhledem k počtu pozorování v základním souboru vedlo k snížení vypovídací hodnoty stromu.

Za předpokladu dostatečně velkého základního souboru je možné uvažovat výsledky vyplývající z rozhodovacího stromu, které byly popsány výše uvažovat na celou populaci. Obecně lze říci, že proměnnou nejvíce ovlivňující hodnoty závisle proměnné je životní úroveň respondentů. Rozdělení hodnot vysvětlované proměnné dále závisí na rodinném stavu, na skutečnosti, zda v domácnostech dochází ke splácení zboží na úvěr a na ekonomické aktivitě. Na základě rozhodovacího stromu z obrázku 3.2 je možné předpokládat, že nejvyšší podíl osob, které si vytvářejí finanční rezervu, by byl mezi ekonomicky aktivními osobami nesplácející zboží na úvěr a žijící v domácnosti s dobrou životní úrovní. Naopak nejnižší podíl osob vytvářejících si finanční rezervu by byl mezi osobami svobodnými, nebo rozvedenými, které pocházejí z domácnosti se špatnou životní úrovní.

3.3 Metoda logistické regrese

K analýze vztahu vysvětlované proměnné *Y* a souboru vysvětlujících proměnných bude nyní použita metoda logistické regrese. Statistický software SPSS jako výstup pro tuto metodu nabízí tabulku s odhady regresních koeficientů.

Data, která budou vstupovat do modelu, se budou mírně lišit od těch použitých v předchozí metodě. Jelikož metoda logistické regrese nedokáže pracovat s chybějícími hodnotami, jsou z modelu vyloučena neplatná pozorování u příslušných vysvětlujících proměnných způsobem popsaným v oddílu 3.1. V tomto případě bude základní soubor obsahovat jen 937 pozorování. Dále bude využita vlastnost logistické regrese, díky které dokáže pracovat se spojitými proměnnými. Proto věk respondentů bude udávat proměnná *IDE_*2 a hodnoty proměnné *IDE_*13 představující počet členů domácnosti budou považovány za spojité.

3.3.1 Výběr vhodného modelu

V SPSS může být logistický regresní model sestrojen více způsoby. Proto budou zkonstruovány celkem tři modely, a to pomocí metod postupného automatického výběru proměnných *forward stepwise selection* a *backward stepwise selection*. Poslední model bude sestrojen ručně a bude obsahovat jen proměnné, které byly na základě Waldova testu statisticky významné. Interpretován bude pouze nejlepší model z této trojice. O tom, který z modelů to bude, se rozhodne na základě Akaikeho informační kritéria a Nagelkerkeova R^2.

U metod postupného výběru proměnných bude pro score test zvolena hladina významnosti α_E rovna 0,05. Jako statistika pro test o vyřazení potencionální proměnné bude sloužit věrohodnostní poměr. Tento test bude probíhat na hladině významnosti α_R rovné 0,1.

Model sestrojený způsobem postupného výběru proměnných *forward stepwise selection* je shodný s modelem zkonstruovaným manuálním vyřazením statisticky nevýznamných proměnných na základě Waldova testu. Výsledná tabulka odhadů parametrů pro tyto dva modely je zobrazena v příloze C. Akaikeho informační kritérium pro tyto modely je rovno 824,804 a Nagelkerkeova R^2 je rovno 0,438.

Logistický regresní model sestrojený pomocí metody postupného výběru proměnných *backward stepwise selection* má hodnotu Akaikeho informační kritéria rovnu 814,968. Nagelkerkeovo R^2 je pro tento rovno 0,448. Tento model obsahuje navíc vysvětlující proměnnou *ER_38A*. Tato proměnná je však Waldovým testem vyhodnocena jako statisticky

nevýznamná. Výsledky založené na Waldově testu a věrohodnostním poměru se mohou lišit, jak je uvedeno v oddílu 2.4.1. V takovém případě by měl být upřednostněn výsledek získaný pomocí věrohodnostního poměru. Proto bude pomocí věrohodnostního poměru ověřeno, zda přidání proměnné *ER_38A* do modelu má smysl. Tento postup je doporučen například v [13]. Testová statistika G se získá rozdílem statistiky $-2LL$ modelu z předchozího odstavce a statistiky $-2LL$ příslušející modelu s vysvětlující proměnnou *ER_38A*. V tomto případě má statistika G χ^2 rozdělení s jedním stupněm volnosti. Výsledná p-hodnota tohoto testu je nižší než zvolená hodnota hladina významnosti 0,05. Zamítá se tedy nulová hypotéza, že parametr β příslušející proměnné *ER_38A* se rovná nule. To znamená, že proměnná *ER_38A* by měla být do modelu zahrnuta.

3.3.2 Interpretace

Interpretován bude logistický regresní model sestrojený pomocí postupného výběru vysvětlujících proměnných metodou *backward stepwise selection*, jak bylo v předešlém oddílu objasněno. Statistický software SPSS nabízí výstup ve formě tabulky zobrazující hodnoty a další informace o maximálně věrohodných bodových odhadech parametrů β získané postupem popsaným v oddílu 2.3. Tento výstup je zobrazen v tabulce 3.4.

Ještě než se přistoupí k interpretaci samotné, bude určeno, zda je model jako celek statisticky významný. Toho bude dosaženo pomocí věrohodnostního poměru, který porovná model obsahující pouze konstantu β_0 a model zahrnující všechny parametry β zobrazené v tabulce 3.4. Dále bude zhodnocena hodnota Cox-Snellova a Nagelkerkeova R^2.

Hodnota statistiky $-2LL$ modelu obsahující pouze parametr β_0 je 1154,582, stejná statistika pro model obsahující všechny parametry β je zobrazena v tabulce 3.3. Statistika G představující jejich rozdíl je rovna 357,884 a má χ^2 rozdělení s 22 stupni volnosti. P-hodnota příslušející testu, který je založen na věrohodnostním poměru, je nulová. Výsledný logistický regresní model je tedy možné hodnotit jako celkově statisticky významný.

Tabulka 3.3: Charakteristiky modelu

-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
796,698[a]	,317	,448

Tabulka 3.3 dále obsahuje hodnoty Cox-Snellova a Nagelkerkeova R^2. V [10] je zmíněno, že hodnotu Nagelkerkeova R^2 je interpretovat podobně jako koeficient determinace v lineární

regresi, tedy jako procento variability závislé proměnné vysvětlené daným model. Pomocí popisovaného modelu se tedy podařilo objasnit 44,8 % celkové variability vysvětlované proměnné.

Tabulka 3.4: Výsledný logistický regresní model.

		β	S.E.	Wald	df	Sig.	Exp(β)
β_1	**IDE_2**	0,019	0,008	6,422	1	0,011	1,019
	t_VZD			13,971	3	0,003	
β_{21}	Střední bez maturity a vyučení	0,261	0,263	,990	1	0,320	1,299
β_{22}	Střední s maturitou	0,702	0,271	6,686	1	0,010	2,017
β_{23}	VOŠ, bakalářské, VŠ	1,099	0,346	10,086	1	0,001	3,001
	ER_35D			9,010	3	0,029	
β_{31}	Spíše souhlasí	-0,542	0,204	7,029	1	0,008	,582
β_{32}	Spíše nesouhlasí	-0,216	0,341	,402	1	0,526	,805
β_{33}	Rozhodně nesouhlasí	,581	0,644	,816	1	0,366	1,789
	ER_38A			10,268	5	0,068	
β_{41}	Velmi nízká	-0,728	0,869	,702	1	0,402	,483
β_{42}	Spíše nízká	,273	0,609	,201	1	0,654	1,314
β_{43}	Odpovídající	-0,001	0,527	,000	1	0,999	,999
β_{44}	Spíše vysoká	0,625	0,518	1,458	1	0,227	1,869
β_{45}	Velmi vysoká	0,794	0,548	2,097	1	0,148	2,211
	ER_39			11,357	4	0,023	
β_{51}	Velmi závažný	-0,194	0,522	,138	1	0,711	,824
β_{52}	Závažný	-0,047	0,480	,010	1	0,921	,954
β_{53}	Málo závažný	-0,0890	0,495	3,233	1	0,072	,410
β_{54}	není problémem	-0,569	0,637	,797	1	0,372	,566
	ER_44AC			37,713	1	0,000	
β_6	Ano	-1,257	0,205	37,713	1	0,000	,285
	IDE_3A			14,769	3	0,002	
β_{71}	Ženatý/vdaná (reg. partnerství)	0,843	0,266	10,037	1	0,002	2,323
β_{72}	Rozvedený/ná	0,099	0,325	,093	1	0,760	1,104
β_{73}	Vdovec/vdova	0,543	0,443	1,506	1	0,220	1,722
	t_ZIVUR			109,911	2	0,000	
β_{81}	Dobrá	3,107	0,299	108,130	1	0,000	22,355
β_{82}	Ani dobrá, ani špatná	1,843	0,273	45,715	1	0,000	6,314
β_0	Constant	-2,641	0,683	14,958	1	,000	,071

Zdroj: Vlastní zpracování, SPSS. Data z [16].

Předcházející tabulka již popisuje výsledný logistický regresní model. Udává hodnoty odhadnutých regresních koeficientů pro jednotlivé proměnné, popřípadě pro jejich nereferenční kategorie. Tabulka dále zobrazuje směrodatné chyby odhadů, hodnoty Waldovy statistiky, příslušné stupně volnosti, p-hodnoty Waldova testu a v neposlední řadě odhady poměrů šancí.

Poté co byla ověřena statistická významnost modelu, jsou z interpretačního hlediska nejzajímavější hodnoty jednotlivých parametrů a především hodnoty poměrů šancí.

Z důvodu větší přehlednosti jsou v tabulce 3.5 zobrazeny referenční kategorie jednotlivých kategoriálních proměnných.

Do logistického regresního modelu bylo ze souboru potencionálních vysvětlujících proměnných zařazeno celkem osm proměnných. To je výrazně více než modelu ve formě rozhodovacího stromu vytvořeného metodou CHAID. V logistickém regresním modelu navíc vystupují proměnné *IDE*_2, *t_VZD*, *ER*_35, *ER*_38*A* a *ER*_39. Oproti metodě CHAID nebyla do regresního modelu zahrnuta proměnná *t_EA_A*_2 sledující ekonomickou aktivitu respondentů.

Tabulka 3.5: Referenčních kategorie.

Proměnná	Referenční kategorie
ER_35D	Rozhodně souhlasí
ER_38A	Neví
ER_39	Neví
ER_44AC	Ne
IDE_3A	Svobodný
t_ZIVUR	špatná
t_VZD	neúplné/základní

Zdroj: Vlastní zpracování, MS Excel.

Stejně jako u metody CHAID závislou proměnnou nejvíce ovlivňuje proměnná *t_ZIVUR* sledující životní úroveň respondentů. Je patrné, že životní úroveň zvyšuje šanci vytváření si finanční rezervy. U respondentů s dobrou životní úrovní je tato šance přibližně 22krát vyšší než u respondentů se špatnou životní úrovní za předpokladu, že se hodnoty ostatních vysvětlujících proměnných nezmění, jak vyplývá s příslušného poměru šancí. Logistický regresní model ukazuje, že vliv proměnných *IDE*_3*A* a *ER*_44*AC* na závislou proměnnou je podobný tomu jaký byl zjištěn z rozhodovacího stromu u metody CHAID. U osob žijících v manželství je šance na vytváření finanční rezervy 2,3krát vyšší než u svobodných osob. U vdov a vdovců je tato šance 1,7krát vyšší. Mezi svobodnými a rozvedenými osobami s ohledem na vysvětlovanou proměnnou naopak velký rozdíl není. Model dále odhaduje nižší tendenci na vytváření finanční rezervy u osob splácejících nějaké zboží na úvěr.

Jediná spojitá proměnná zařazená do modelu je *IDE*_2, která představuje věk respondentů. Nízká hodnota poměru šancí, však značí, že příliš velký vliv na vysvětlovanou proměnnou nemá. Ačkoliv se proměnná *t_VZD* v rámci metody CHAID neobjevila, má poměrně silný vliv na vytváření si finanční rezervy vzdělání. Z poměrů šancí příslušejících kategoriím této proměnné je patrné, že vyšší vzdělání doprovází vyšší šance na vytvoření si finanční rezervy. U zbývajících proměnných se takto jasný trend již neobjevuje. Například u proměnné *ER*_35*D*, která zjišťuje, do jaké míry respondenti souhlasí s výrokem, že půjčování peněz je riziko. U osob, které s tímto výrokem spíše souhlasí, je šance na vytváření finanční rezervy téměř poloviční než u osob, které s tímto výrokem souhlasí. Zatímco pro osoby, které s tímto výrokem rozhodně nesouhlasí, je tato šance až 1,8krát vyšší. Vyšší hodnoty poměrů šancí se dále objevují u proměnné *ER*_38*A*, především u kategorií osob, které hodnotí míru zadlužení občanů ČR jako spíše vysokou a velmi vysokou. U osob, které hodnotí zadlužení občanů ČR jako velmi vážný či vážný problém se šance na vytváření finanční rezervy příliš neliší od osob, které na tuto věc nemají názor, jak dokládají poměry šancí blízké jedné u proměnné *ER*_39. U osob, které vnímají tuto problematiku jako málo vážný problém, nebo ji vůbec nevnímají jako problém, je šance na vytváření finanční rezervy přibližně poloviční oproti osobám bez názoru na tuto problematiku.

Obecně lze z modelu vyvodit, že finanční rezervu si budou vytvářet spíše osoby z domácností s dobrou životní úrovní, s vysokoškolským vzděláním, žijící v manželském svazku vnímající míru zadlužení občanů ČR jako velmi vysokou.

3.4 Porovnání klasifikačních schopností

Klasifikační schopnost logistické regrese a modelu ve formě rozhodovacího stromu je možné hodnotit na základě klasifikační tabulky, nebo ROC křivky, jak je zmíněno v oddílu 2.4.2. Ve statistickém softwaru SPSS je klasifikační tabulka standardním výstupem v rámci metody CHAID i logistické regrese. ROC křivku k modelům je nutné vytvořit zvlášť. Konstrukce křivky pro logistickou regresi je popsána v oddílu 2.4.2. Pro metodu CHAID je sestrojena na základě relativních četností nastání sledovaného jevu v koncových uzlech. Hodnoty hranice pravděpodobnosti, pro které je senzitivita a specificita vypočítána u logistické regrese, zde zastupují právě relativní četnosti v koncových uzlech. Jak klasifikační tabulka, tak i plocha pod ROC křivkou měří klasifikační sílu modelu.

V tabulce 3.6 je zobrazena klasifikační tabulka pro rozhodovací strom zobrazený na obrázku 3.2. Klasifikační tabulka porovnává pozorované a odhadnuté hodnoty závislé proměnné *Y*. Pro daný případ je odhadnuta kategorie „*ano*“ pokud jemu příslušející hodnota relativní četnosti v koncovém uzlu je vyšší než 0,5. Pokud je tato relativní četnost nižší než hranice 0,5, je daný případ zařazen do kategorie „*ne*“.

Z tabulky vyplývá, že pomocí modelu bylo pro kategorii „*ne*“ správně klasifikováno 169 případů, což představuje 52,8 % případů, u kterých tato kategorie byla pozorována. Do kategorie „ano“ bylo správně klasifikováno 621 případů, to odpovídá 89,7 %, u kterých byla tato kategorie pozorována. Z těchto hodnot se vypočte celková úspěšnost modelu, která nabývá hodnoty 78,1 %.

Tabulka 3.6: Klasifikační tabulka rozhodovacího stromu

	Predicted		
Observed	ne	ano	Percent Correct
Ne	169	151	52,8%
Ano	71	621	89,7%
Overall Percentage	23,7%	76,3%	78,1%

Zdroj: Vlastní zpracování, SPSS. Data z [16].

Tabulka 3.7 zobrazuje klasifikační tabulku pro logistický regresní model. Způsob jakým byla získána je popsán v oddílu 2.4.2. Z tabulky je patrné, že modelem bylo správně klasifikováno do kategorie „*ne*“ 159 případů, které představují 55,4 % případů, kde byla pozorována kategorie „*ne*“. Do kategorie „ano“ bylo správně klasifikováno celkem 597 případů, což představuje 91,8 % případů, u kterých byla tato kategorie pozorována. Celková úspěšnost logistického regresního modelu je rovna 80,7 %.

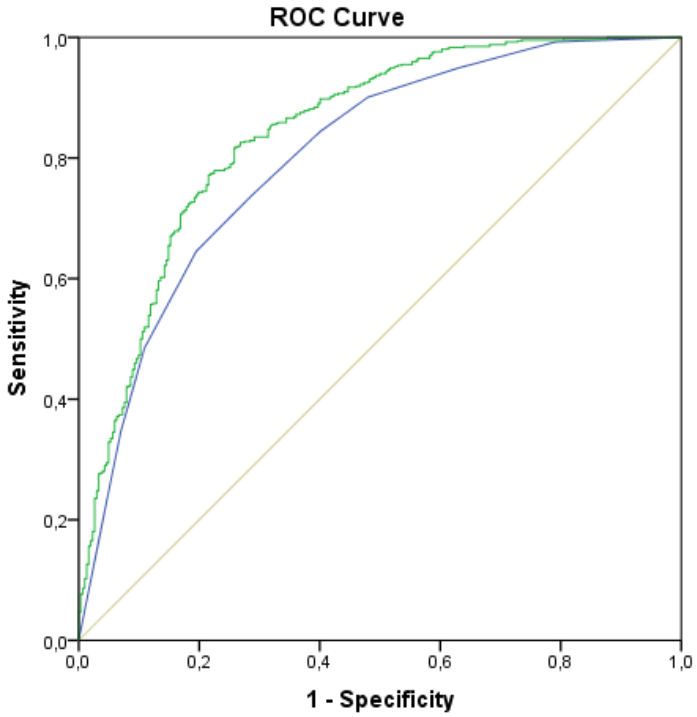
Tabulka 3.7: Klasifikační tabulka logistické regrese

	Predicted		
Observed	ne	ano	Percent Correct
Ne	159	128	55,4%
Ano	53	597	91,8%
Overall Percentage	22,6%	77,4%	80,7%

Zdroj: Vlastní zpracování, SPSS. Data z [16].

Na základě klasifikačních tabulek lze konstatovat, že logistická regrese klasifikovala hodnoty vysvětlované proměnné *Y* jen o něco málo lépe než rozhodovací strom sestrojený metodou CHAID

Na obrázku 3.3 je zobrazena ROC křivka pro model ve formě rozhodovacího stromu a pro logistickou regresi. ROC křivka metody CHAID je označena modrou barvou a křivka pro logistickou regresi zeleně. Diagonála představuje situaci, kdy by model klasifikoval hodnoty závislé proměnné *Y* zcela náhodně. V takovém případě by hodnota plochy pod křivkou měla hodnotu 0,5. Výpočetní program SPSS v rámci výstupu k ROC křivce nabízí i vypočtenou plochu pod křivkou. Ta je pro metodu CHAID rovna 0,805 a pro logistickou regresi 0,843. Podle [10] hodnota plochy pod ROC křivkou v intervalu od 0,8 do 0,9 představuje model s velmi dobrou klasifikační silou.



Obrázek 3.3: ROC křivka pro metodu CHAID a logistickou regresi[12]

Na základě hodnot celkové úspěšnosti a plochy pod ROC křivkou je možné konstatovat, že oba sestrojené model klasifikují hodnoty vysvětlované proměnné *Y* velice dobře, avšak logistická regrese o něco lépe.

[12] *Zdroj: Vlastní zpracování, SPSS, data z* [16]

Závěr

Tato práce se zabývala metodou CHAID a logistické regrese. V úvodní teoretické části byly představeny obě metody. Byly popsány principy a postupy pro aplikaci těchto metod společně s možnostmi hodnocení jejich výstupů. Dále byly v práci uvedeny oblasti využití, kde je možné se s těmito metodami setkat a v neposlední řadě došlo k popsání výhod a nevýhod, které s těmito metodami souvisejí.

V praktické části byly výše uvedené postupy aplikovány na reálná data poskytnutá Českým sociálněvědním datovým archivem Sociologického ústavu Akademie věd ČR, v. v. i. Pomocí zkoumaných metod byly ze zmíněných dat získány výstupy, které byly následně interpretovány a v závěru práce porovnány na základě jejich klasifikačních schopností.

Na základě reálné aplikace těchto metod se bylo možné přesvědčit, že jak metoda pro tvorbu rozhodovacích stromů CHAID, tak logistická regrese dokázali identifikovat vysvětlující proměnné s nejzásadnějším vlivem na závislou proměnnou. Dále bylo prokázáno, že klasifikační schopnost obou metoda se v podstatě výrazně neliší.

Logistická regrese oproti metodě CHAID nabídla podrobnější, avšak složitější model. Díky tomu bylo možné odhalit vztahy, které z rozhodovacího stromu nebyly patrné. Například vliv úrovně vzdělání na vysvětlovanou proměnnou. Vztahy popsané logistickým regresním modelem byly navíc jasně kvantifikované pomocí poměrů šancí. Logistická regrese navíc umožňuje hodnocení modelů na základě konkrétních statistik, to usnadňuje hledání nejlepšího možného modelu. Jak již bylo řečeno, logistický regresní model byl ve srovnání s modelem rozhodovacího stromu složitější a zároveň hůře interpretovatelný. K složitosti modelu přispívali především vysvětlující kategoriální proměnné s větším počtem kategorií.

Metoda CHAID naopak nabídla stručný, přehledný a velice snadno interpretovatelný model. Ovšem za cenu toho, že neodhalil další významné vztahy mezi závislou proměnnou a vysvětlujícími proměnnými. K větší úspornosti modelu ve formě rozhodovacího stromu pomohla především schopnost slučování jednotlivých kategorií vysvětlujících proměnných.

Na základě zjištěných skutečností popsaných v této práci nelze konstatovat, které metoda je lepší či která by měla být upřednostňována. Lze však doporučit kombinaci použití obou metod. Metodu CHAID k předběžné analýze vztahů v datovém souboru, která umožní základní vhled do zkoumané problematiky a následně pro podrobnější a jasně kvantifikované

výsledky použít logistickou regresi. Sloučené kategorie vysvětlujících proměnných, získané pomocí metody CHAID, by navíc bylo výhodné využít v následné logistické regresi. Umožnilo by to zjednodušení výsledného logistického regresního modelu.

Přehled literatury a zdrojů

[1] G. V. Kass, "An Exploratory Technique for Investigating Large Quantities of Categorical Data," *Journal of the Royal Statistical Society. Series C (Applied Statistics),* no. Vol. 29, s 119 – 127, South Africa: University of Witwatersrand, 1980.

[2] J. Vild, "Analýza CHAID jako nástroj vyhodnocování marketingových kampaní," in *Doucek - P., Nedomová, L. (ed). Sborník prací účastníků vědeckého semináře doktorského studia FIS VŠE*, Praha, 2012.

[3] M. J. A. Berry and G. S. Linoff, Data mining techniques for marketing, sales, and customer relationship management, 2. ed., Indianapolis: Wiley Publishing, Inc. , 2004.

[4] J. F. Hair, R. B. Bush and O. D. J., . Marketing research within a changing informational environment, 2. ed., New York: The McGraw-Hill Companies, Inc., 2003.

[5] R. Nisbet, J. Elder and G. Miner, Statistical analysis & data mining applications, Boston: Elsevier Academic Press, 2009.

[6] IBM Corp., "IBM SPSS Statistics 20 Algorithms," IBM Corp., Armonk, NY, 2011.

[7] D. B. P. V. V. Bílková, Pravděpodobnost a statistika, Plzeň: Aleš Čeněk, 2009.

[8] S. Tufféry, "Data Mining and Statistics for Decision Making," John Wiley & Sons, Ltd., Chichester, 2011.

[9] B. Řeháková, "Nebojte se logistické regrese," *Sociologický časopis,* vol. 4, no. 36, pp. 475-492, 2000.

[10] D. W. Hosmer and S. Lemeshow, Applied logistic regression, 2. ed., New York: John Wiley & Sons, 2000.

[11] A. Agresti, Categorical data analysis, 2. ed., New York: John Wiley, 2002.

[12] A. Paul, "http://statisticalhorizons.com," 1 Duben 2015. [Online]. Available: http://statisticalhorizons.com/whats-so-special-about-logit. [Accessed 20 Březen 2016].

[13] I. Pecáková, "Logistická regrese s vícekategoriální vysvětlovanou proměnnou," *Acta Oeconomica Pragensia*, vol. 1, pp. 86-96, 2007.

[14] J. Tvrdík, "Logistická regrese a vyheldávání modelů," in *Robust '98* , Praha, 1998.

[15] I. Stankovičová and M. Vojtková, Viacrozmerné štatistické metódy s aplikáciami, Bratislava: Edícia Ekonómia, 2007.

[16] Sociologický ústav (Akademie věd ČR). Centrum pro výzkum veřejného, *Naše společnost 2016 - leden* [*Datový soubor*], Praha: Český sociálněvědní datový archiv, 2016.

Seznam použitých obrázků

Obrázek 1.1: Příklad rozhodovacího stromu 5
Obrázek 1.2: Schéma slučování kategorií vysvětlující proměnné 8
Obrázek 2.1: Graf hodnot alternativní proměnné Y v závislosti na X 15
Obrázek 2.2: Graf odhadnutých hodnot pravděpodobnosti $\pi(x)$ v závislosti na X 18
Obrázek 2.3: Graf odhadnutých hodnot logitu v závislosti na X 18
Obrázek 2.4: Příklad ROC křivky 26
Obrázek 3.1: Graf relativních četností závislé proměnné Y 32
Obrázek 3.2: Rozhodovací strom sestrojený metodou CHAID 35
Obrázek 3.3: ROC křivka pro metodu CHAID a logistickou regresi 44

Seznam použitých tabulek

Tabulka 2.1: Příklad klasifikační tabulky. 25
Tabulka 3.1: Absolutní a relativní četnosti vysvětlované proměnné. 31
Tabulka 3.2: Výpis vysvětlujících proměnných. 33
Tabulka 3.3: Charakteristiky modelu 39
Tabulka 3.4: Výsledný logistický regresní model. 40
Tabulka 3.5: Tabulka referenčních kategorií. 41
Tabulka 3.6: Klasifikační tabulka rozhodovacího stromu 43
Tabulka 3.7: Klasifikační tabulka logistické regrese 43

Přílohy

Příloha A: Tabulka dat k obrázkům číslo 2.1, 2.2 a 2.3 pocházející z *[10]*.

ID	X	Y	π(x)	g(x)	ID	X	Y	π(x)	g(x)
1	20	0	0,045	-3,05	**51**	44	1	0,397	-0,42
2	23	0	0,062	-2,73	**52**	44	1	0,397	-0,42
3	24	0	0,068	-2,62	**53**	45	0	0,424	-0,31
4	25	0	0,075	-2,51	**54**	45	1	0,424	-0,31
5	25	1	0,075	-2,51	**55**	46	0	0,451	-0,2
6	26	0	0,084	-2,4	**56**	46	1	0,451	-0,2
7	26	0	0,084	-2,4	**57**	47	0	0,478	-0,09
8	28	0	0,102	-2,18	**58**	47	0	0,478	-0,09
9	28	0	0,102	-2,18	**59**	47	1	0,478	-0,09
10	29	0	0,112	-2,07	**60**	48	0	0,505	0,02
11	30	0	0,124	-1,96	**61**	48	1	0,505	0,02
12	30	0	0,124	-1,96	**62**	48	1	0,505	0,02
13	30	0	0,124	-1,96	**63**	49	0	0,533	0,13
14	30	0	0,124	-1,96	**64**	49	0	0,533	0,13
15	30	0	0,124	-1,96	**65**	49	1	0,533	0,13
16	30	1	0,124	-1,96	**66**	50	0	0,560	0,24
17	32	0	0,150	-1,74	**67**	50	1	0,560	0,24
18	32	0	0,150	-1,74	**68**	51	0	0,587	0,35
19	33	0	0,164	-1,63	**69**	52	0	0,613	0,46
20	33	0	0,164	-1,63	**70**	52	1	0,613	0,46
21	34	0	0,180	-1,52	**71**	53	1	0,639	0,57
22	34	0	0,180	-1,52	**72**	53	1	0,639	0,57
23	34	1	0,180	-1,52	**73**	54	1	0,664	0,68
24	34	0	0,180	-1,52	**74**	55	0	0,688	0,79
25	34	0	0,180	-1,52	**75**	55	1	0,688	0,79
26	35	0	0,197	-1,41	**76**	55	1	0,688	0,79
27	35	0	0,197	-1,41	**77**	56	1	0,711	0,9
28	36	0	0,215	-1,3	**78**	56	1	0,711	0,9
29	36	1	0,215	-1,3	**79**	56	1	0,711	0,9
30	36	0	0,215	-1,3	**80**	57	0	0,733	1,01
31	37	0	0,234	-1,19	**81**	57	0	0,733	1,01
32	37	1	0,234	-1,19	**82**	57	1	0,733	1,01
33	37	0	0,234	-1,19	**83**	57	1	0,733	1,01
34	38	0	0,254	-1,08	**84**	57	1	0,733	1,01
35	38	0	0,254	-1,08	**85**	57	1	0,733	1,01
36	39	0	0,275	-0,97	**86**	58	0	0,754	1,12
37	39	1	0,275	-0,97	**87**	58	1	0,754	1,12
38	40	0	0,298	-0,86	**88**	58	1	0,754	1,12
39	40	1	0,298	-0,86	**89**	59	1	0,774	1,23
40	41	0	0,321	-0,75	**90**	59	1	0,774	1,23
41	41	0	0,321	-0,75	**91**	60	0	0,792	1,34
42	42	0	0,346	-0,64	**92**	60	1	0,792	1,34
43	42	0	0,346	-0,64	**93**	61	1	0,810	1,45
44	42	0	0,346	-0,64	**94**	62	1	0,826	1,56
45	42	1	0,346	-0,64	**95**	62	1	0,826	1,56
46	43	0	0,371	-0,53	**96**	63	1	0,842	1,67
47	43	0	0,371	-0,53	**97**	64	0	0,856	1,78
48	43	1	0,371	-0,53	**98**	64	1	0,856	1,78
49	44	0	0,397	-0,42	**99**	65	1	0,869	1,89
50	44	0	0,397	-0,42	**100**	69	1	0,911	2,33

Příloha B: Tabulka s hodnotami kategoriálních proměnných z datového souboru *[16]*.

EV_10	Hodnocení současné ekonomické situace v ČR
1	Velmi dobrá
2	Dobrá
3	Ani dobrá ani špatná
4	Špatná
5	Velmi špatná
6	Neví
EV_11	Jak se změní ekonomická situace v ČR
1	Velmi zlepší
2	Trochu zlepší
3	Nezmění se
4	Trochu zhorší
5	Velmi zhorší
6	Neví
ER_35D	Je půjčování peněz riziko
1	Rozhodně souhlasí
2	Spíše souhlasí
3	Spíše nesouhlasí
4	Rozhodně nesouhlasí
ER_35E	Vyplatí se mít dluhy
1	Rozhodně souhlasí
2	Spíše souhlasí
3	Spíše nesouhlasí
4	Rozhodně nesouhlasí
5	Neví
ER_38A	Hodnocení míry zadlužení občanů ČR
1	Velmi nízká
2	Spíše nízká
3	Odpovídající (tak akorát)
4	Spíše vysoká
5	Velmi vysoká
6	Neví
ER_39	Je míra zadlužení občanů ČR problém
1	Velmi závažný
2	Závažný
3	Málo závažný
4	Není problémem
5	Neví

ER_44AA	Splácí hypotéku
1	Ano
2	Ne
ER_44AC	Splácí zboží na úvěr
1	Ano
2	Ne
EU_179A	Byl nezaměstnaný déle než 6 měsíců
1	Ano
2	Ne
IDE_3A	Rodinný stav
1	Svobodný/ná,
2	Ženatý/vdaná, registrované partnerství
3	Rozvedený/ná
4	Vdovec/vdova
t_VEK_5	Věk
1	15-19
2	20-29
3	30-44
4	45-59
5	60+
t_ZIVUR	Životní úroveň
1	Dobrá,
2	Ani dobrá, ani špatná
3	Špatná
t_EA_a_2	Ekonomická aktivita
1	Aktivní
2	Neaktivní
t_VZD	Dosažená úroveň vzdělání
1	Neúplné nebo základní
2	Střední bez maturity
3	Střední s maturitou
4	VOŠ, bakalářské a VŠ

Příloha C: Regresní model sestrojený manuálně a pomocí *forward stepwise selection*

Variables in the Equation						
	B	S.E.	Wald	df	Sig.	Exp(B)
IDE_2	,020	,007	7,175	1	,007	1,020
t_VZD			14,184	3	,003	
t_VZD(1)	,317	,259	1,495	1	,221	1,373
t_VZD(2)	,697	,266	6,860	1	,009	2,007
t_VZD(3)	1,137	,341	11,119	1	,001	3,117
ER_35D			10,935	3	,012	
ER_35D(1)	-,589	,201	8,554	1	,003	,555
ER_35D(2)	-,233	,335	,483	1	,487	,792
ER_35D(3)	,632	,636	,988	1	,320	1,882
ER_39			24,208	4	,000	
ER_39(1)	,274	,450	,372	1	,542	1,316
ER_39(2)	,294	,421	,485	1	,486	1,341
ER_39(3)	-,799	,452	3,127	1	,077	,450
ER_39(4)	-,641	,602	1,136	1	,287	,527
ER_44AC(1)	-1,199	,202	35,418	1	,000	,301
IDE_3A			13,672	3	,003	
IDE_3A(1)	,801	,262	9,331	1	,002	2,229
IDE_3A(2)	,095	,324	,086	1	,769	1,100
IDE_3A(3)	,547	,441	1,539	1	,215	1,727
t_ZIVUR			107,692	2	,000	
t_ZIVUR(1)	2,996	,291	105,991	1	,000	20,010
t_ZIVUR(2)	1,797	,269	44,780	1	,000	6,033
Constant	-2,431	,612	15,778	1	,000	,088