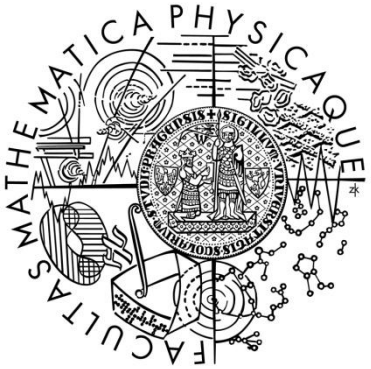
Univerzita Karlova v Praze

Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Jan Strnad

Analýza storna pojistných smluv

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: Mgr. Ing. Jakub Mertl

Studijní program: Matematika

Studijní obor: Finanční a pojistná matematika

Praha 2013

Rád bych poděkoval panu Ing. Mgr. Jakubovi Mertlovi za poskytnutí možnosti zabývat se reálným problémem na skutečných datech a také za trpělivost a cenné připomínky při vedení práce.

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona.

V Praze dne podpis

Název práce: Analýza storna pojistných smluv

Autor: Bc. Jan Strnad

Katedra / Ústav: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: Mgr. Ing. Jakub Mertl

Abstrakt: Cílem této práce je vyvinout na reálných datech nástroj k identifikaci smluv pojištění odpovědnosti z provozu motorového vozidla ohrožených stornem. Jsou zde představeny prostředky pro explorativní analýzu dat, výstavbu modelu logistické regrese, porovnání různých modelů a jejich validaci a kalibraci. Pomocí popsaných metod je na skutečných datech sestaveno několik modelů a z nich vybrán jeden finální. Vlastnosti tohoto modelu jsou poté ověřeny validací na vzorku z odlišného období. Posledním krokem je kalibrace modelu na očekávanou budoucí stornovost portfolia.

Klíčová slova: Pravděpodobnost storna, logistická regrese, vývoj a validace modelu

Title: Lapse Analysis of Insurance Contracts

Author: Bc. Jan Strnad

Department: Department of Probability and Mathematical Statistics

Supervisor: Mgr. Ing. Jakub Mertl

Abstract: The aim of the present work is to develop a tool for identification of Motor Third Party Liability insurance contracts which are at risk of cancellation. Methods for explorative data analysis, building a logistic regression model, comparing models and their validation and calibration are presented. Several models are developed on the real dataset using mentioned methods and then the final one is chosen. Behavior of the final model is verified by the validation on the out-of-time sample. Last step is calibration of the model to the expected value of the future portfolio cancellation rate.

Keywords: Probability of cancellation, logistic regression, model development and validation

Obsah

Seznam tabulek

Úvod

Na trhu s pojištěním odpovědnosti z provozu motorového vozidla pozorujeme v posledních letech velmi tvrdý konkurenční boj a z toho vyplývající vysokou fluktuaci klientů. Udržet si v tomto prostředí klienty je čím dál obtížnější, proto by bylo přínosné, kdyby pojišťovna dokázala s předstihem identifikovat klienty, kteří uvažují o odchodu ke konkurenci.

Možnost zabývat se tímto problémem na skutečných datech jedné z pojišťoven je pro mě velmi zajímavá jednak z hlediska obsahu, ale i pro to, že se jedná o řešení konkrétního problému, jehož cílem je implementace do reálného provozu pojišťovny.

Ve své práci nejprve popíši samotné pojištění odpovědnosti z provozu motorového vozidla, legislativní úpravu tohoto pojistného produktu a také situaci na trhu. Dále se budu věnovat teoretickému aparátu, který později využiji v praktické části práce a krátce představím základní funkce a procedury statistického softwaru, ve kterém budu pracovat. V praktické části poté pomocí modelu logistické regrese popíši, na čem závisí pravděpodobnost, zda klient na výročí smlouvy odejde od pojišťovny. Představím různé možnosti analýzy proměnných a výstavby modelu, porovnám jejich výsledky a vyberu finální model. Schopnost tohoto modelu odhadovat výše zmíněnou pravděpodobnost i v budoucnosti ověřím jeho validací na odlišném vzorku smluv. Na závěr provedu kalibraci modelu tak, aby dokázal nejen identifikovat klienty se zvýšeným rizikem storna, ale i co nejpřesněji odhadnout konkrétní pravděpodobnosti.

Cílem práce je analyzovat situaci okolo storna povinného ručení při obnově smlouvy a nastudovat problematiku explorativní analýzy dat a výstavby modelu logistické regrese. Dále pak na skutečných datech vyvinout model, který by dokázal s co největší přesností odhalit klienty, u kterých je pravděpodobné, že při obnově smlouvu stornují. Zároveň budu klást důraz na to, aby tento model byl co nejstabilnější v čase a tedy jej bylo možné využívat i v budoucnu a v neposlední řadě aby byl dobře interpretovatelný.

1 Úvod do problematiky

1.1 Pojištění odpovědnosti z provozu motorového vozidla

Pojištění odpovědnosti z provozu motorového vozidla, tzv. povinné ručení, upravuje zákon 168/1999 Sb. [15]. Jde o povinně smluvní pojištění, ve kterém se pojistitel zavazuje uhradit škody na majetku, zdraví, životě nebo škody, které mají formu ušlého zisku, jež pojištěný způsobí třetí osobě či osobám při provozu vozidla. Nevztahuje se však na škodu vzniklou na vozidle pojištěného, kterou způsobí vlastním zaviněním.

Na základě výše uvedeného zákona může vozidlo na dálnici, silnici, místní komunikaci či veřejně přístupnou účelovou komunikaci pouze pokud je pojištěno. K 1. 1. 2013 bylo v ČR podle [10] registrováno přes 7,5 milionu vozidel. Za rok 2012 uvádí Česká asociace pojišťoven v [4] 6 699 426 smluv povinného ručení s celkovým předepsaným smluvním pojistným ve výši 19,2 miliardy Kč, což tvoří 15,7 % z celkového předepsaného pojistného pro celý trh s životním i neživotním pojištěním za daný rok. Jde tedy o rozsáhlý pojistný kmen.

Specifikem tohoto odvětví v posledních letech je velký nárůst škodních úhrnů, který je způsoben především vysokou škodní inflací u škod na zdraví, doprovázený poklesem průměrného pojistného. Zatímco počty škod v letech 2002 až 2010 stagnují, průměrná výše škody na majetku podle ČKP stoupla o 31 % a průměrná výše škody na zdraví o 167 %. Index spotřebitelských cen ve stejném období vzrostl o 20 %. Roční objem škod pak stoupl z 9,7 mld. Kč v roce 2002 na 13,9 mld. Kč v roce 2010. Podle [8].

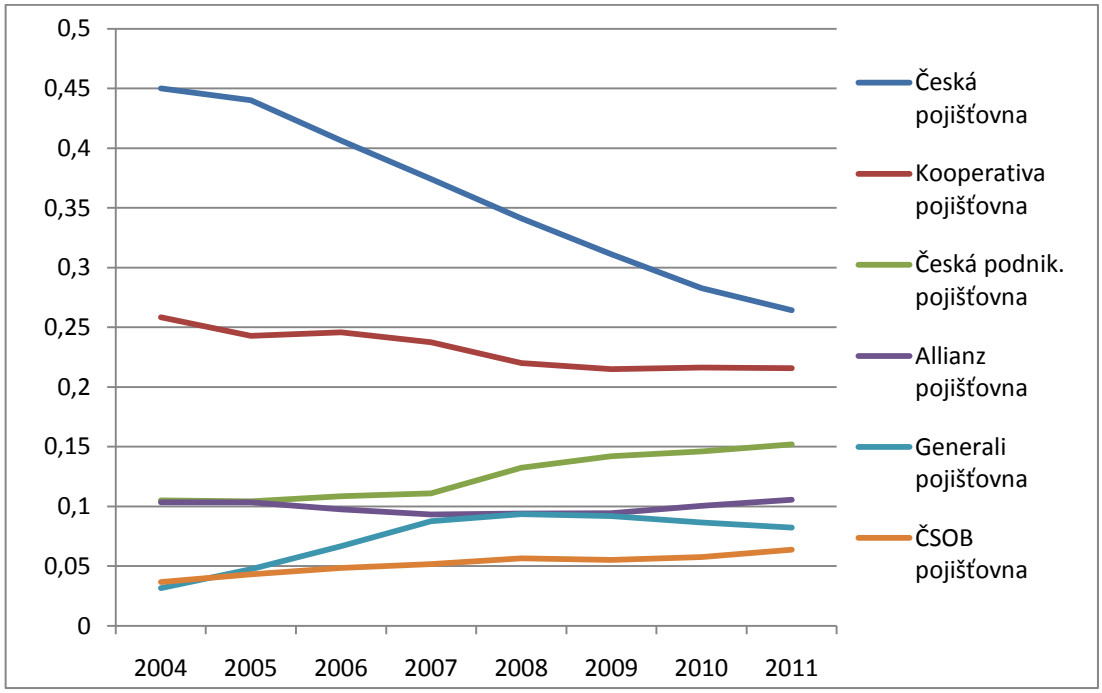
Povinné ručení bylo do konce roku 1999 zákonným pojištěním a provozovala jej pouze Česká pojišťovna. Od roku 2000 došlo ke změně na pojištění povinně smluvní a licence na jeho provozování byla udělena 11 pojišťovnám. Dnes má právo provozovat povinné ručení 14 subjektů. Otevření trhu zapříčinilo vznik konkurenčního prostředí, rozšíření nabídky produktů, různé výše pojistného plnění i sazeb pojistného a vznik vedlejších služeb a produktů. Čerpáno z [3].

1.2 Situace na trhu

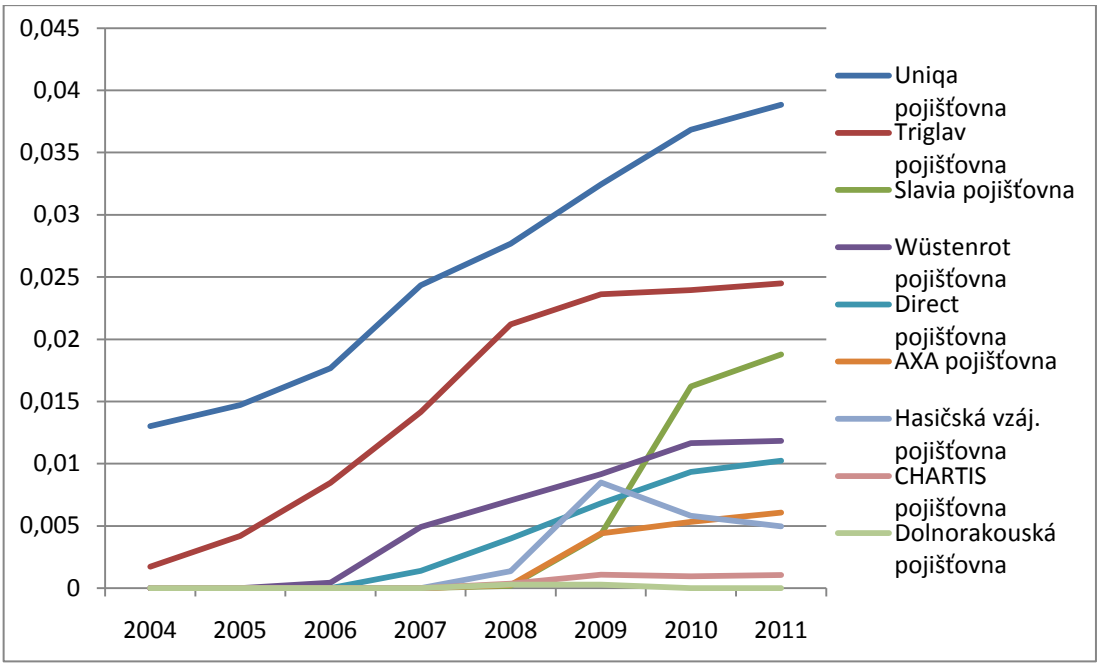
Protože se jedná o povinné pojištění, které se týká velké části populace, pro pojišťovny představuje dobrou příležitost k získání nových klientů, kterým je poté možné cíleně nabídnout další produkty. I proto jde o oblast

pojišťovnictví s pravděpodobně nejtvrdší konkurencí. To má za následek vysokou fluktuaci klientů a krátkou životnost smluv.

V posledních letech trh zaznamenal přechod klientů od největších pojišťoven ke středním a menším. Především podíl České pojišťovny, která až do roku 1999 poskytovala povinné ručení jako jediná, do roku 2011 klesl na 26,5 %, což umožnilo nárůst jak pojišťovnám pohybujícím se mezi 5 a 10 %, tak pojišťovnám s podílem na trhu v jednotkách procent. Detailnější pohled na situaci představují grafy č. 1 a 2. Podkladová data převzata z [3].



Graf 1: Pojišťovny s podílem na trhu větším než 5%. (Na vodorovné ose je období a na svislé podíl na trhu.)



Graf 2: Pojišťovny s podílem na trhu menším než 5%. (Na vodorovné ose je období a na svislé podíl na trhu.)

1.3 Legislativní rámec

Zánik pojištění odpovědnosti z provozu motorového vozidla jako soukromé pojištění upravují paragrafy 19 až 25 zákona č. 37/2004 Sb. o pojistné smlouvě [16]. Některé případy pak konkrétněji popisuje paragraf 12 zákona 168/1999 Sb. o pojištění odpovědnosti za škodu způsobenou provozem motorového vozidla [15]. K zániku může dojít z důvodu uplynutí doby, nezaplacení pojistného, dohodou, výpovědí, odstoupením, odmítnutím pojistného plnění, zánikem pojistného rizika či pojištěné věci, smrtí pojištěné osoby či zánikem pojištěné právnické osoby bez právního nástupce, nebo jiným způsobem uvedeným ve smlouvě. Jednotlivé možnosti nyní podrobněji popíši a pro přehlednost znázorním na obrázku 1.

- Uplynutí doby

Soukromé pojištění zaniká uplynutím pojistné doby. U pojištění na dobu určitou lze ve smlouvě stanovit, že se pojištění automaticky prodlouží (za stejných podmínek a na stejnou dobu na jakou bylo původně sjednáno), pokud pojistitel nebo pojistník nejméně 6 týdnů před vypršením pojistné doby nesdělí druhé straně, že o prodloužení nemá zájem.

Povinné ručení se sjednává na dobu určitou, obvykle s automatickým prodloužením. K jeho zániku uplynutím doby tedy dojde, pokud pojistník minimálně 6 týdnů před vypršením pojistné doby sdělí pojistiteli, že nemá zájem o prodloužení.

- Nezaplacení pojistného

K zániku soukromého pojištění z důvodu neplacení pojistného dojde tehdy, nezaplatí-li pojistník dlužné pojistné ve lhůtě stanovené v upomínce. Tato lhůta nesmí být kratší než jeden měsíc.

- Dohoda

Pojistitel a pojistník se mohou dohodnout na zániku soukromého pojištění. V tomto případě je okamžik zániku a způsob vzájemného vyrovnání závazků stanoven oboustrannou dohodou.

- Výpověď

Pojistitel nebo pojistník mohou soukromé pojištění vypovědět:

  - S osmidenní výpovědní lhůtou během dvou měsíců od uzavření smlouvy.
  - S měsíční výpovědní lhůtou během tří měsíců ode dne oznámení vzniku pojistné události.
  - Ke konci pojistného období (viz zánik uplynutím doby) pokud jde o pojištění s běžným pojistným.

Pojistník navíc může soukromé pojištění vypovědět v případě převodu pojistného kmene, nebo po odnětí povolení k provozování pojišťovací činnosti pojistitele. V obou případech tak musí učinit během jednoho měsíce a s osmidenní výpovědní lhůtou.

- Odstoupení

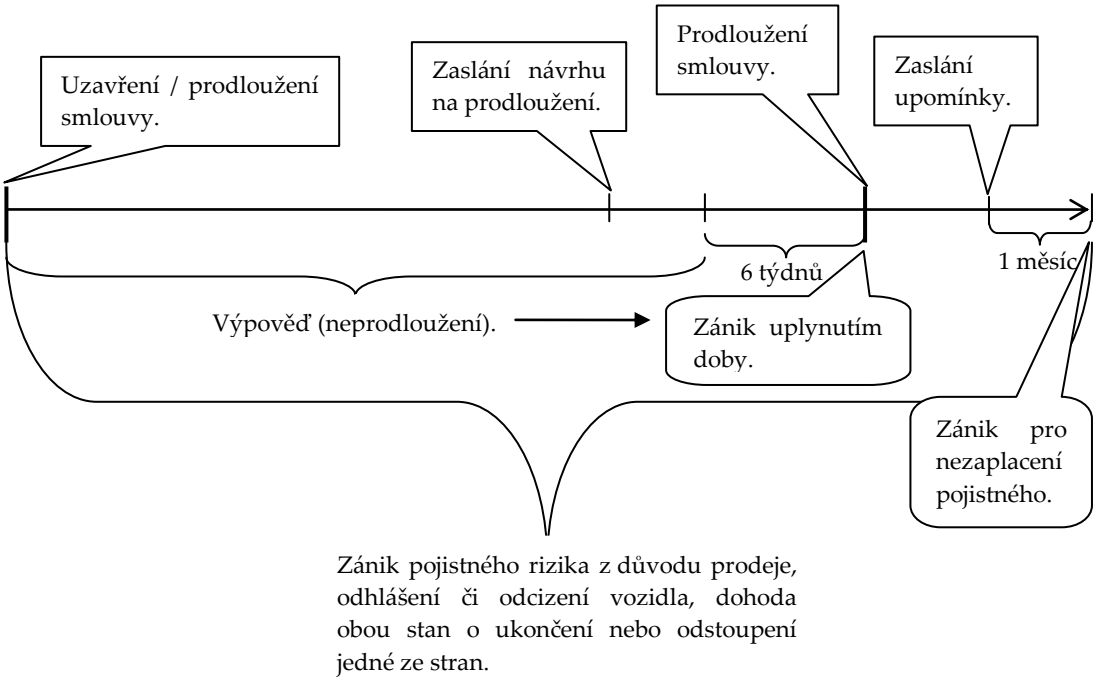
Pojistitel i pojistník mají právo od pojistné smlouvy odstoupit v případě, že jim druhá strana při sjednávání úmyslně či z nedbalosti nepravdivě nebo neúplně zodpověděla písemné dotazy týkající se pojištění, pokud by při pravdivém a úplném zodpovězení smlouvu neuzavřel. Toto právo lze uplatnit do dvou měsíců od zjištění popsané skutečnosti. Odstoupením se smlouva od počátku ruší.

- Jiné důvody

Soukromé pojištění zaniká dnem, kdy zaniklo pojistné riziko, pojištěná věc nebo jiná majetková hodnota, nebo dnem, kdy došlo ke smrti pojištěné fyzické osoby nebo zániku pojištěné právnické osoby bez právního nástupce, nestanoví-li tento zákon nebo pojistná smlouva jinak.

Podle zákona 168/1999 Sb. [15] mezi tyto důvody patří:

  - změna vlastníka tuzemského vozidla
  - zánik vozidla, které podléhá evidenci vozidel - vozidlo zanikne okamžikem, kdy nastane nevratná změna znemožňující jeho provoz
  - vyřazení tuzemského vozidla z evidence vozidel
  - odcizení vozidla



Obrázek 1: Znázornění života smlouvy

1.4 Předmět práce

Cílem této práce je vyvinout model, pomocí kterého by pojišťovna mohla identifikovat smlouvy ohrožené stornem během obnovy. Na základě informace o pravděpodobnosti storna jednotlivých smluv by se poté pojišťovna mohla efektivněji pokoušet těmto stornům předcházet například cíleným oslovováním rizikových klientů, či zohledněním pravděpodobnosti storna při návrhu nové smlouvy. Z důvodů, pro které může dojít ke stornu, zmíněných v předešlé sekci je tedy třeba vybrat ty, které se vztahují k obnově smlouvy a kterým je smysluplné pokusit se předcházet. Nejprve uvedu přehled nejčastějších příčin storna pojistné smlouvy. Jsou jimi:

- **Změna majitele vozidla** - Doložením smlouvy o prodeji vozidla dokládá pojistník zánik pojistného rizika a pojištění ke dni prodeje zaniká.
- **Odhlášení vozidla z evidence** - Po odhlášení vozidla z evidence již není možné vozidlo používat, a proto dochází k zániku pojistného rizika.
- **Nezaplacení pojistného**
- **Odcizení vozidla** - Odcizením vozidla dochází k zániku pojistného rizika a tím i pojištění.
- **Neprodloužení smlouvy** - Pojistník minimálně 6 týdnů před vypršením pojistné doby sdělí pojistiteli, že nemá zájem smlouvu prodlužovat. Pojištění poté zanikne z důvodu uplynutí doby.

Změna majitele vozidla, odhlášení vozidla z evidence i odcizení vozidla jsou situace, které by neměly souviset se spokojeností klienta s podmínkami smlouvy a s jeho případným přechodem ke konkurenci. Přestože by jistě bylo možné najít vztah například mezi typem, stářím a hodnotou vozidla a pravděpodobností, že dojde k jeho odcizení, neměla by taková informace z hlediska udržitelnosti klienta moc velkou hodnotu. Všechny zmíněné případy jsou storna z důvodu zániku pojistného rizika a těm pojišťovna zabránit nemůže.

Pro účely této práce mě budou zajímat storna, která se zpravidla vážou na přechod klienta k jiné pojišťovně. Těmi jsou neprodloužení pojistné smlouvy provedené řádným způsobem, tedy nejméně šest týdnů před vypršením smlouvy a nezaplacení pojistného na nové smlouvě.

2 Teoretický základ

V této kapitole popíši teoretický aparát užitý při vývoji modelu. Čerpat přitom budu převážně z [2], [6] a [17]. Cílem modelu je odhadnout pravděpodobnost, s jakou na jednotlivých smlouvách dojde ke stornu během obnovy. Pro vývoj mám k dispozici vzorek smluv a ke každé z nich informaci, zda u ní došlo ke stornu, či ne. Mohu je tedy rozdělit na smlouvy „dobré" (G) a smlouvy „špatné" (B). Existence storna na smlouvě pro mě tedy bude závislou proměnnou. Informace o smlouvě, vozidlu, kterého se týká a o klientovi pak budou vstupovat do modelu jako nezávislé proměnné. Jednotlivé hodnoty kategoriálních proměnných, případně jejich skupiny, budu označovat termínem *kategorie*. Stejně tak intervaly spojitých proměnných.

2.1 Weight of Evidence (WOE)

Metoda *Weight of Evidence* porovnává poměr „dobrých" a „špatných" smluv v jednotlivých kategoriích příslušné proměnné s poměrem „dobrých" a „špatných" smluv v celém vzorku. Vyjadřuje tak relativní rizikovost smluv dané kategorie. Pro spojité proměnné je nutné provést rozdělení na kategorie (intervaly). Dalším předpokladem je přítomnost alespoň jedné „dobré" a jedné „špatné" smlouvy v každé kategorii. Hodnota *WOE* pro *i*-tou kategorii je dána vztahem:

$$WOE_i = \ln\left(\frac{\frac{G_i}{\sum_{i=1}^{n} G_i}}{\frac{B_i}{\sum_{i=1}^{n} B_i}}\right),$$

kde n je celkový počet kategorií proměnné.

Pokud je tedy v příslušné kategorii stejná stornovost jako v celém vzorku, *WOE* této kategorie je rovno 0. Je-li stornovost v kategorii nižší, *WOE* nabývá záporných hodnot. Analogicky je-li stornovost smluv z dané kategorie vyšší než stornovost vzorku, nabývá *WOE* kladných hodnot.

Využití této metody je jednak ve dvourozměrné analýze, k porovnání rozdílů stornovosti v jednotlivých kategoriích proměnné a dále pak k transformaci kategoriálních proměnných na spojité.

Metoda *Weight of Evidence* vyjadřuje relativní rizikovost jednotlivých kategorií, nereflektuje však, jaké je v těchto kategoriích zastoupení smluv. Tuto informaci zohledňuje následující statistika zvaná *Informační hodnota*.

2.2 Informační hodnota

Informační hodnota vyjadřuje predikční sílu proměnné v tom smyslu, jak dobře lze na základě dané proměnné rozdělit smlouvy na „dobré" a „špatné". Nabývá vždy nezáporných hodnot a je tím vyšší, čím více se liší stornovost v jednotlivých kategoriích dané proměnné od průměrné stornovosti celého vzorku a také čím více smluv je v kategoriích lišících se od průměru. Proměnná, která má ve všech kategoriích stejnou stornovost má *Informační hodnotu* rovnou nule. Příspěvek i-té kategorie k *Informační hodnotě* proměnné je:

$$IV_i = \left(\frac{G_i}{\sum_{i=1}^{n} G_i} - \frac{B_i}{\sum_{i=1}^{n} B_i} \right) * WOE_i$$

a celková *Informační hodnota* pak je součet příspěvků jednotlivých kategorií:

$$IV = \sum_{i=1}^{n} \left[\left(\frac{G_i}{\sum_{i=1}^{n} G_i} - \frac{B_i}{\sum_{i=1}^{n} B_i} \right) * WOE_i \right],$$

kde *n* je celkový počet kategorií proměnné.

Pomocí této statistiky můžeme porovnávat proměnné mezi sebou a také různé varianty rozdělení jedné proměnné do kategorií. Může nám tedy pomoci s výběrem proměnných. Výhodou *Informační hodnoty* je také fakt, že na rozdíl od *Giniho koeficientu*, který popíši v následující sekci, nevyžaduje seřazení kategorií podle stornovosti. Dobře se tedy hodí pro prvotní analýzu dat.

2.3 Giniho koeficient

Giniho koeficient je charakteristika vyjadřující diverzifikační sílu modelu nebo proměnné. Vychází z *Lorenzovy křivky*, která je definovaná pomocí empirických kumulativních distribučních funkcí (*CDF*) skóre „dobrých" a „špatných" smluv. Skóre je zde zastoupeno pravděpodobností storna odhadnutou modelem, anebo hodnotou sledované proměnné.

$$F_G(a) = \frac{1}{n} * \sum_{i=1}^{m+n} I(s_i \leq a \wedge D_i = 1)$$

$$\text{a } F_B(a) = \frac{1}{m} * \sum_{i=1}^{m+n} I(s_i \leq a \wedge D_i = 0)$$

jsou CDF skóre „dobrých" (G) a „špatných" (B) smluv, přičemž *n* je celkový počet „dobrých" smluv, *m* je celkový počet „špatných" smluv, s_i je skóre *i-té* smlouvy, $D_i = 1$ je-li *i-tá* smlouvy „dobrá" a 0 v případě, že je „špatná" a

I(výrok) = 1 pokud je výrok pravdivý a 0 opačně. *Lorenzova křivka* je parametricky dána:

$$x = F_B(a)\,,$$
$$y = F_G(a)\,, \qquad a \in [s_{min}, s_{max}]\,,$$

kde s_{min} je minimální hodnota skóre a s_{max} maximální hodnota skóre. Křivka tedy pro každou hodnotu skóre zobrazuje, jak velká část „dobrých" a „špatných" smluv má skóre menší nebo rovné této hodnotě. V případě, že by stornovost byla ve všech kategoriích sledované proměnné stejná, nebo že by model přiřazoval smlouvám pravděpodobnost storna zcela náhodně, *Lorenzova křivka* by měla tvar úsečky spojující body [0,0] a [1,1].

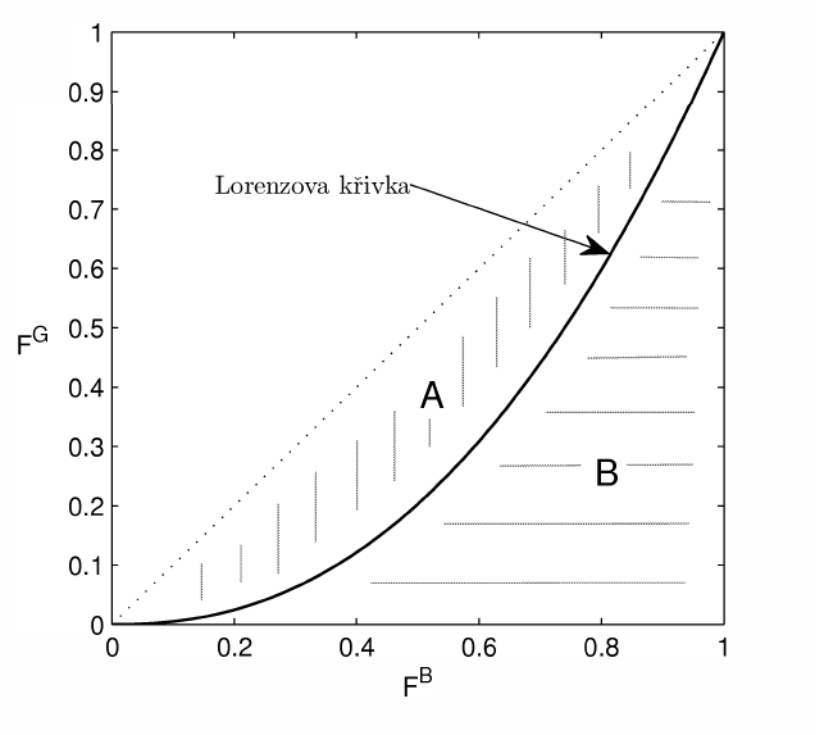
Giniho koeficient je definován jako podíl velikosti plochy mezi *Lorenzovou křivkou* a právě úsečkou spojující body [0,0] a [1,1] (plocha A) a celkovou velikostí plochy pod touto úsečkou (plochy A + B):

$$GC = \frac{A}{A + B}$$

A protože velikost plochy pod diagonálou je $\frac{1}{2}$, platí:

$$GC = 2A\,.$$

Giniho koeficient nabývá hodnot z intervalu [-1,1]. Hodnota 1 vyjadřuje ideální diverzifikační schopnost. Neboli libovolná „dobrá" smlouva bude v takovém případě mít nižší odhadnutou pravděpodobnost storna než libovolná „špatná" smlouva. Hodnota 0 značí stav, kdy po seřazení smluv podle odhadnuté pravděpodobnosti budou „dobré" a „špatné" smlouvy rovnoměrně promíchané. Záporné hodnoty pak znamenají, že proměnná nebo model schopnost diverzifikace mají, ale v opačném směru než jsme očekávali. Čerpáno z [12].



Obrázek 2: ***Lorenzova křivka*** **(zdroj: http://cs.wikipedia.org/wiki/Soubor:Giniho_koeficient.png)**

2.4 Population stability index

Population stability index slouží k vyjádření velikosti změn v rozdělení sledované veličiny ve dvou různých vzorcích. Danou veličinu je potřeba rozdělit do kategorií, nejčastěji decilů jednoho ze vzorků, či logických skupin. Pro každou kategorii se poté porovná zastoupení v jednotlivých vzorcích pomocí vztahu:

$$\mathrm{PSI_i} = \left(d_1^i - d_2^i\right) * \ln\left(\frac{d_1^i}{d_2^i}\right),$$

kde d_1^i je poměrné zastoupení *i-té* kategorie v prvním vzorku.

Celkovou hodnotu *PSI* pro veličinu s *k* kategoriemi pak spočteme součtem přes všechny kategorie:

$$\mathrm{PSI} = \sum_{i=1}^{k} \mathrm{PSI}_i \,.$$

Obdobně jako v případě *Informační hodnoty*, jedná se totiž o identický výpočet pouze jinak použitý, je nutné, aby každá kategorie byla zastoupena v obou výběrech. Za stabilní populaci se obecně považují vzorky s *PSI* menší než 0,1. Poznamenejme ještě, že hodnota *PSI* nezáleží na velikostech vzorků. Čerpáno z [11].

2.5 Logistická regrese

Pro modelování binomické veličiny Y_i, tedy veličiny s alternativním rozdělením s parametrem μ_i, střední hodnotou μ_i a rozptylem $\mu_i * (1 - \mu_i)$, pomocí vektoru vysvětlujících proměnných x_i použiji metodu logistické regrese. Ta vyjadřuje střední hodnotu rozdělení náhodné veličiny Y_i, neboli $P(Y_i=1)$, jako funkci nezávisle proměnných nazývanou logistická funkce. Tato funkce je definovaná:

$$\mu_i(\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{\beta}'\boldsymbol{x_i})}{1 + \exp(\boldsymbol{\beta}'\boldsymbol{x_i})},$$

kde $\boldsymbol{\beta}$ je vektor neznámých parametrů a $\boldsymbol{\beta}'$ jeho odhad. Úpravou

$$\ln\left(\frac{\mu_i(\boldsymbol{\beta})}{1 - \mu_i(\boldsymbol{\beta})}\right) = \boldsymbol{\beta}'\mathbf{x_i}$$

lze přejít k lineárnímu modelu. Zlomku v závorce se říká šance a funkce, která binomické veličině přiřazuje logaritmus šance, se označuje jako logit. Protože levá strana rovnosti může nabývat libovolné hodnoty z $\mathbb{R}$, nemusíme na parametry β klást žádné omezující podmínky.

2.5.1 Interpretace regresních parametrů

Uvažujme nyní dva vektory vysvětlujících proměnných $\boldsymbol{x}_i$ a $\boldsymbol{x}_j$, pro které platí:

$$x_{jk} = x_{ik} + 1 \text{ pro } k = m \text{ a } x_{jk} = x_{ik} \text{ pro } k \neq m\,.$$

Vliv parametru β_m pak je:

$$\operatorname{logit}\left(\mu_j(\boldsymbol{\beta})\right) - \operatorname{logit}\left(\mu_i(\boldsymbol{\beta})\right) = \ln\left(\frac{\frac{\mu_j(\boldsymbol{\beta})}{1 - \mu_j(\boldsymbol{\beta})}}{\frac{\mu_i(\boldsymbol{\beta})}{1 - \mu_i(\boldsymbol{\beta})}}\right) = \boldsymbol{\beta}\boldsymbol{x_j} - \boldsymbol{\beta}\boldsymbol{x_i} = \beta_m\,,$$

rovná se tedy logaritmu poměru šancí vysvětlovaných proměnných příslušejících $\boldsymbol{x}_j$ a $\boldsymbol{x}_i$.

2.5.2 Odhad regresních parametrů

Pravděpodobnost, že náhodná veličina Y_i nabude hodnoty 1 nebo 0, můžeme vyjádřit:

$$P(Y_i = y_i) = \mu_i(\boldsymbol{\beta})^{y_i}(1 - \mu_i(\boldsymbol{\beta}))^{1-y_i}\,, \qquad y_i = 0{,}1\,.$$

Logaritmická věrohodnostní funkce pak má tvar:

$$\ln L = \ln\left(\prod_{i=1}^{n} \mu_i(\boldsymbol{\beta})^{y_i}(1 - \mu_i(\boldsymbol{\beta}))^{1-y_i}\right) =$$

$$= \sum_{i=1}^{n} [y_i \ln(\mu_i(\boldsymbol{\beta})) + (1 - y_i)\ln(1 - \mu_i(\boldsymbol{\beta}))]\,,$$

kde n je počet pozorování a její parciální derivace podle vektoru parametrů $\boldsymbol{\beta}$ jsou rovny:

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial \ln L}{\partial \mu_i} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} (y_i - \mu_i(\boldsymbol{\beta}))\, \boldsymbol{x}_i \,.$$

Maximálně věrohodný odhad parametrů $\boldsymbol{\beta}$ získáme položením těchto rovnic rovných nule,

$$\boldsymbol{X}'\big(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})\big) = 0 \,,$$

a vyřešením takto vzniklé soustavy. To lze provést například pomocí Newton-Raphsonovy metody (viz [14]).

2.6 Deviance

Deviance je statistika, která porovnává pozorované hodnoty závislé proměnné s hodnotami odhadnutými modelem. Využívá k tomu věrohodnostní funkce a vyjádření pozorovaných hodnot jako výstupu ze saturovaného modelu, tedy modelu, který obsahuje stejný počet parametrů jako je pozorování.

$$D = -2 \ln\left(\frac{L_F}{L_S}\right),$$

kde L_F je věrohodnostní funkce vybraného modelu a L_S je věrohodnostní funkce saturovaného modelu. Zlomku v závorce se říká věrohodnostní poměr. Dosazením získáme

$$D = -2 * \sum_{i=1}^{n} \left[y_i \ln\left(\frac{\mu_i(\boldsymbol{\beta})}{y_i}\right) + (1 - y_i) \ln\left(\frac{1 - \mu_i(\boldsymbol{\beta})}{1 - y_i}\right)\right]. \qquad (2.1)$$

Tato statistika se používá jako míra těsnosti modelu (viz část 2.11.1) a také k porovnávání modelů mezi sebou.

2.7 Test věrohodnostního poměru

Tento test slouží k porovnávání modelů, nejčastěji modelu s a bez proměnné, jejíž vliv je předmětem našeho zájmu, nebo plného modelu a modelu pouze s pevným členem (interceptem), jako test významnosti celého modelu. Testová statistika vychází z porovnání statistiky *Deviance* plného modelu a podmodelu, který vznikl z plného modelu odebráním jedné nebo více nezávislých proměnných. Sleduje rozdíl v této statistice vzniklý zařazením příslušných proměnných do modelu. Tedy:

$$G = D(podmodelu) - D(plného\ modelu)$$

Protože saturovaný model je v obou případech totožný, lze testovou statistiku vyjádřit:

$$G = -2 \ln\left(\frac{L_1}{L_0}\right), \tag{2.2}$$

kde L_1 je věrohodnostní funkce podmodelu a L_0 věrohodnostní funkce plného modelu. Za platnosti hypotézy, že příslušné nezávislé proměnné jsou nevýznamné a tedy jejich regresní parametry β jsou rovny 0, má testová statistika Chí-kvadrát rozdělení s p stupni volnosti, kde je p je rozdíl počtu proměnných v plném modelu a v podmodelu.

2.8 Waldův test

Jinou možností testování významnosti konkrétní nezávislé proměnné, nebo celého modelu, je Waldův test, který porovnává maximálně věrohodné odhady regresních koeficientů daných proměnných s odhadem jejich směrodatných odchylek:

$$W = \hat{\beta}'(\widehat{\mathrm{var}}(\hat{\beta}))^{-1}\hat{\beta}$$

Opět se testuje hypotéza, že dané nezávislé proměnné (jedna v případě testu významnosti proměnné, všechny v případě testu významnosti celého modelu) jsou z hlediska vysvětlení závislé proměnné nevýznamné, tedy že jejich regresní koeficienty β jsou rovny nule. Za platnosti této hypotézy má testová statistika normální rozdělení, případně její druhá mocnina Chí-kvadrát rozdělení s p stupni volnosti, kde p je počet parametrů, které v hypotéze pokládáme rovné nule.

2.9 Score test

Score test je další test významnosti nezávislé proměnné či významnosti celého modelu, který vychází z parciálních derivací logaritmické věrohodnostní funkce podle vektoru parametrů $\boldsymbol{\beta}$:

$$U(\boldsymbol{\beta}) = \frac{\partial \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$$

Ta je porovnávána s

$$I(\boldsymbol{\beta}) = -\frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}.$$

Testová statistika má tedy tvar:

$$S = U'(\boldsymbol{\beta}) * I^{-1}(\boldsymbol{\beta}) * U(\boldsymbol{\beta})$$

a má za platnosti hypotézy $\boldsymbol{\beta} = \boldsymbol{\beta_0}$ Chí-kvadrát rozdělení s p stupni volnosti, kde p je rovno počtu omezení daných hypotézou.

2.10 Výběr modelu

2.10.1 Stepwise selection

Metoda *Stepwise selection*, neboli metoda postupné výstavby modelu, je založena na přidávání a odebírání proměnných na základě statistických kritérií. Jde o kombinaci metod *Forward selection*, která začíná modelem bez proměnných a postupně přidává vždy proměnnou, s největším příspěvkem k vysvětlení závislé proměnné a *Backward elimination*, která začne s plným modelem a postupně odebírá proměnné, které k vysvětlení závislé proměnné přispívají nejméně.

Jako kritéria pro výběr proměnných, které budou přidány, či odebrány z modelu, se nejčastěji používají výše popsané testy. Statistický software SAS, jenž budu užívat v praktické části práce, přidává proměnné na základě *Score testu* a odebírá proměnné dle výsledků *Waldova testu*.

Na začátku celého procesu se odhadne hodnota absolutního členu (interceptu). Poté se pro každou nezávislou proměnnou z množiny proměnných, které necháme vstoupit do procedury, spočítá testová statistika *Score testu*. Následně se vybere proměnná s nejvyšší hodnotou této statistiky. Je-li tato hodnota větší než předem stanovená minimální hodnota požadovaná pro vstup do modelu, je tato proměnná zařazena do modelu. Tento postup se pak zopakuje pro proměnné, které nebyly vybrány v prvním kroku. Pokud je výsledkem zařazení druhé proměnné do modelu, odhadne se nový model s absolutním členem a dvěma nezávislými proměnnými a poté se přistoupí k ověření významnosti obou těchto proměnných. Pro každou z nich se spočítá testová statistika *Waldova testu* a vybere se ta s nižší hodnotou této statistiky. Pokud je tato hodnota menší než předem nastavená hranice nutná pro zachování proměnné v modelu, je tato proměnná z modelu vyřazena. Do množiny proměnných, ze kterých se vybírají kandidáti na vstup do modelu, se tato proměnná ale již nevrací. Obdobně se pak postupuje až do doby, kdy buď jsou v modelu všechny proměnné, které byly k dispozici, nebo není žádná proměnná, která by splnila podmínky pro přidání do modelu a zároveň v modelu není žádná proměnná, která by splnila podmínky pro vyřazení.

Jak vyplývá z popisu procedury *Stepwise selection*, pro výsledek tohoto postupu je zásadní nastavení hraničních hodnot pro zařazení a vyřazení proměnné. Konkrétně v programu SAS se nastavují p-hodnoty pro jednotlivé testy. Protože *Score test* i *Waldův test* testují hypotézu, že příslušná proměnná je statisticky nevýznamná, pro vstup do modelu je potřeba p-hodnota nižší než nastavená hranice a pro vyřazení z modelu naopak p-hodnota vyšší než nastavená hranice.

Samotné nastavení pak záleží na strategii výstavby modelu. Nižší p-hodnoty vedou k modelům s méně proměnnými, u kterých lze předpokládat nižší sílu, ale větší stabilitu. Naopak vyšší p-hodnoty vedou k bohatším modelům, ze kterých se pak často odstraňují proměnné na základě dodatečných analýz, nebo u kterých se pouze využijí testy z jednotlivých kroků k manuálnímu sestavení modelu. Na základě práce založené na Monte Carlo simulacích navrhují Hosmer a Lemeshow v [6] maximální p-hodnoty pro vstup do modelu v rozmezí 0,15 – 0,20. Minimální p-hodnota pro vyřazení z modelu by pak měla být stejná, nebo vyšší.

2.10.2 Best subsets

Alternativní metodou pro výběr modelu je procedura *Best subsets*. Ta ze zvolené množiny pro každou možnou velikost modelu, tedy od modelu s jednou proměnnou až po model obsahující všechny proměnné ze zvolené množiny, vybere určený počet nejlepších variant. Statistický software SAS pro tuto proceduru užívá *Branch and Bound* algoritmus autorů Furnival a Wilson [5]. Algoritmus byl původně navržen pro modely lineární regrese, ovšem autoři Hosmer, Jovanovic a Lemeshow v [7] ukázali, že problém výpočtů odhadů koeficientů logistické regrese pro velké množství modelů lze modifikovat tak, aby bylo možné jej řešit metodami pro lineární regresi.

Pro výběr nejlepších modelů jednotlivých velikostí je nutné zvolit nějaké kritérium kvality modelu, podle kterého by bylo možné jednotlivé varianty mezi sebou porovnávat. Implementace *Best subsets* procedury v SASu k tomuto účelu užívá *Score test*.

I přes to, že tato procedura množinu potenciálních modelů výrazně zredukuje, jejím výstupem jsou ve většině případů desítky variant a je tedy nutné dokázat mezi sebou porovnat modely různých velikostí a vybrat z nich jeden finální. Pro modely lineární regrese se pro tyto účely často používá statistika C_q, viz [9], která porovnává reziduální součet čtverců daného modelu, SSE_q, se střední kvadratickou chybou modelu s maximálním počtem proměnných, MSE_p,:

$$C_q = \frac{SSE_q}{MSE_p} - n + 2q \ ,$$

kde n je počet pozorování, q počet proměnných příslušného modelu a p celkový počet proměnných, z nichž vybíráme kandidáty do modelu.

Hosmer, Jovanovic a Lemeshow v [7] nejprve ukázali, že pokud se pro výstavbu modelu logistické regrese pomocí *Best subsets* metody užívá metod pro lineární regresi, lze využít obdobu C_q statistiky s *Pearsonovým tvarem rezidua*:

$$C_q = \frac{\frac{X^2 + \lambda}{X^2}}{n - p - 1} + 2(q + 1) - n\,, \tag{2.3}$$

kde X^2 je *Pearsonova Chí-kvadrát statistika* pro model s p proměnnými a λ je testová statistika *Waldova testu* hypotézy, že koeficienty *(p-q)* proměnných, které nejsou zařazeny do modelu, jsou rovny 0. Za předpokladu, že je model správný, očekáváme, že hodnoty X^2 a λ jsou rovny *(n-p-1)* a *(p-q)*. Dosazením těchto hodnot do (2.3) dostáváme $C_q = q + 1$. Modely, jejichž C_q statistika se blíží této hodnotě, tedy lze považovat za nejlepší kandidáty.

Hosmer a Lemeshow v [6] poté představili způsob, kterým je možné pro logistickou regresi aproximovat hodnotu C_q statistiky pomocí výsledků *Score testu*. Opět předpokládejme, že správný model bude mít X^2 statistiku rovnou *(n-p-1)* a dále, že λ pro *(p-q)* proměnných nezařazených do modelu lze odhadnout rozdílem testové statistiky *Score testu* pro p respektive q proměnných. Tyto předpoklady vedou ke vztahu:

$$\begin{aligned} C_q &= \frac{\frac{X^2 + \lambda}{X^2}}{n - p - 1} + 2(q + 1) - n \approx \\ &\approx \frac{(n - p - 1) + (S_p - S_q)}{1} + 2(q + 1) - n = \\ &= S_p - S_q + 2q - p + 1\,, \end{aligned}$$

kde S_p je testová statistika *Score testu* nejbohatšího modelu s p proměnnými.

2.11 Míra těsnosti modelu

Zatímco v lineární regresi se jako míry těsnosti modelu využívají funkce reziduí, definovaných jako rozdíly pozorovaných a odhadnutých hodnot závislé proměnné, v logistické regresi je možností jak měřit schodu skutečných a odhadnutých hodnot více. Využívá se skutečnosti, že se odhady pravděpodobnosti přiřazují jednotlivým kombinacím hodnot vysvětlujících proměnných. Pozorováním se stejnou kombinací těchto hodnot je tedy přiřazen stejný odhad pravděpodobnosti. Pro jednotlivé kombinace je pak možné spočítat relativní četnost pozorovaných výskytů sledovaného jevu a tu porovnat s očekávanou, modelem odhadnutou, četností pro příslušnou kombinaci.

Nechť n je celkový počet pozorování, p počet nezávislých proměnných a J počet různých kombinací hodnot nezávislých proměnných. Pak m_j bude značit počet pozorování s *j-tou* kombinací proměnných, y_j počet pozorování s *j-tou* kombinací proměnných, pro která je závislá proměnná y = 1 a pro odhadnutou četnost výskytů sledovaného jevu mezi pozorováními s *j-tou* kombinací nezávislých proměnných bude platit

$$\hat{y}_j = m_j\hat{\pi}_j = m_j\mu_j(\boldsymbol{\beta}) = m_j\frac{\exp(\boldsymbol{\beta}'\boldsymbol{x_j})}{1+\exp(\boldsymbol{\beta}'\boldsymbol{x_j})}.$$

Na základě tohoto vztahu nyní definuji dva typy reziduí a od nich odvozené míry těsnosti.

2.11.1 Pearsonova Chí-kvadrát statistika a Deviance

Pearsonovo reziduum je dáno předpisem

$$r(y_j,\hat{\pi}_j) = \frac{(y_j - m_j\hat{\pi}_j)}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}}$$

a příslušná míra těsnosti, takzvaná *Pearsonova Chí-kvadrát statistika*, potom

$$\mathrm{X}^2 = \sum_{\mathrm{j}=1}^{\mathrm{J}} r(y_j,\hat{\pi}_j)^2 . \tag{2.4}$$

Devianční reziduum je definováno předpisem

$$d(y_j,\hat{\pi}_j) = \pm\left\{2\left[y_j\ln\left(\frac{y_j}{m_j\hat{\pi}_j}\right) + (m_j - y_j)\ln\left(\frac{(m_j-y_j)}{m_j(1-\hat{\pi}_j)}\right)\right]\right\}^{1/2},$$

kde znaménko odpovídá znaménku výrazu $(y_j - m_j\hat{\pi}_j)$ a odpovídající míra těsnosti má potom tvar

$$D = \sum_{j=1}^{J} d(y_j,\hat{\pi}_j)^2 . \tag{2.5}$$

V případě že $J = n$, tedy pokud se žádná kombinace hodnot vysvětlujících proměnných neopakuje, má tato statistika stejnou hodnotu jako v rovnici (2.1). Takový případ může nastat například tehdy, pokud je jedna z proměnných v modelu spojitá. Za předpokladu správnosti modelu, tedy že odhadnuté četnosti odpovídají skutečným hodnotám podmíněné střední hodnoty náhodné veličiny Y_i, mají oba dva výše zmíněné typy reziduí asymptoticky normované normální rozdělení a statistiky X^2 a D potom chí-kvadrát rozdělení s $J-(p+1)$ stupni volnosti. Pokud ovšem $J \approx n$, tak pro jednotlivé kombinace hodnot závislých proměnných máme příliš málo pozorování na to, abychom mohli použít toto asymptotické rozdělení. Možným řešením tohoto problému je umělé seskupení pozorování do méně skupin tak, aby se zastoupení v jednotlivých skupinách navýšilo. Takovýmto seskupováním se zabývali autoři Hosmer a Lemeshow v [6].

2.11.2 Hosmer – Lemeshow test

Autoři navrhují dva způsoby seskupení pozorování. V jednom případě podle percentilů rozdělení odhadnutých pravděpodobností na g stejně početných skupin. V druhém případě pak rozdělení škály odhadnutých pravděpodobností na g stejně velkých intervalů a následné rozdělení do

skupin podle toho, do kterého intervalu odhadnutá pravděpodobnost příslušného pozorování padla. Pro obě navržené možnosti testová statistika *Homser-Lemeshow testu těsnosti modelu*, $\hat{C}$, odpovídá *Pearsonově Chí-kvadrát statistice* aplikované na g dvojic odhadnutých a pozorovaných četností v jednotlivých skupinách. Označím-li n'_k celkový počet pozorování v k-té skupině, c_k počet různých kombinací hodnot nezávislých proměnných v k-té skupině,

$$o_k = \sum_{j=1}^{c_k} y_j$$

počet výskytů sledovaného jevu v těchto c_k kombinacích a

$$\bar{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j \hat{\pi}_j}{n'_k}$$

průměrnou odhadnutou pravděpodobnost v k-té skupině, pak testová statistika má tvar

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n'_k \bar{\pi}_k)^2}{n'_k \bar{\pi}_k (1 - \bar{\pi}_k)} \ .$$

Ve své práci Hosmer a Lemeshow pomocí simulací ukázali, že pokud $J = n$ a model logistické regrese je správný, rozdělení této testové statistiky lze aproximovat Chí-kvadrát rozdělením s $g - 2$ stupni volnosti, přičemž nejčastěji se g volí rovno deseti. Stejní autoři dále ukázali, že blíže k Chí-kvadrát rozdělní má metoda dělení založená na percentilech. Zvláště v případech, kdy jsou odhadnuté pravděpodobnosti nerovnoměrně rozloženy a metoda intervalů vede k nestejně velkým skupinám. V případě, že $J < n$, je nutno rozhodnout, jak se vypořádat s případy, kdy pozorování se stejnou odhadnutou pravděpodobností leží v okolí hraničního percentilu. Pokud zařazení pozorování se stejnou hodnotou odhadu do stejné skupiny nevede k výrazně nepoměrnému zastoupení v jednotlivých skupinách, výsledky testu nejsou příliš ovlivněny. Pokud by ovšem tato metoda vedla k situaci, kdy by v jedné nebo více skupinách nebyl dostatečný počet pozorování (uvádí se (např. v [1]), že všechny očekávané četnosti by měly být větší než 5), nebylo by možné využít předpokladu o asymptotickém rozdělní reziduí v jednotlivých skupinách. V případě, že bychom skupin získali příliš málo, by zase míra těsnosti modelu ztrácela na citlivosti. Hosmer a Lemeshow uvádí, že pro méně než 6 skupin, je téměř každý model hodnocen testem jako správný. Pokud pozorování se stejnými odhady rozdělíme do skupin tak, abychom těmto problémům předešli, hodnota testové statistiky bude záležet na tom, jakou metodu pro toto rozdělení zvolíme. Čím více případů se stejnými odhady budeme mít, tím více bude výsledek testu ovlivněn zvolenou metodou.

2.12 Kalibrace

Pokud je model logistické regrese vyvíjen za cílem budoucího odhadování pravděpodobností, což je i případ této práce, může se stát, že mezi vývojovým vzorkem a vzorkem, na kterém je model posléze aplikován, dojde ke změně poměrného zastoupení sledovaného jevu. V našem případě stornovosti. Pokud by se jednalo o obecnou změnu chování klientů a ne pouze o změnu složení vzorku, nebudou odhadnuté pravděpodobnosti storna odpovídat skutečnosti. Přestože hlavním úkolem modelu je diverzifikace klientů na základě jejich rizikovosti a na diverzifikační sílu nemá celkový posun stornovosti vliv, přesnost konkrétních odhadů rozšiřuje možnosti využití modelu. Proto se v takovém případě často přistupuje k dodatečné úpravě pravděpodobností odhadnutých modelem.

Prvním krokem při kalibraci je určení cílové hodnoty, ke které chceme průměrnou hodnotu pravděpodobnosti daného vzorku posunout. Může to být skutečné pozorované zastoupení sledovaného jevu na daném vzorku, pokud chceme provést validaci modelu bez vlivu změny, ke které mezi vývojovým a validačním vzorkem došlo. Nebo můžeme jako cílovou hodnotu určit očekávané zastoupení sledovaného jevu pro období, ve kterém plánujeme model užívat.

Poté je třeba upravit jednotlivé odhady pravděpodobnosti $\hat{\pi}_j$ tak, aby se průměrný dohad $\bar{\pi}$ přiblížil k cílovému průměrnému odhadu $\bar{\pi}_T$. Kalibrované hodnoty odhadů označím $\hat{\pi}_{j_C}$. Jako nejsnazší řešení by se nabízelo prosté vynásobení odhadů konstantou. V takovém případě by se ale mohlo stát, že bychom pro nějaké pozorování dostali pravděpodobnost větší než 1. Využiji tedy vyjádření odhadnuté pravděpodobnosti výskytu sledovaného jevu pro skupinu klientů s *j-tou* kombinací hodnot vysvětlujících proměnných pomocí poměru odhadovaného počtu pozorování s výskytem sledovaného jevu a součtu odhadovaných počtů pozorování s i bez výskytu sledovaného jevu:

$$\hat{\pi}_j = \frac{m_j \hat{\pi}_j}{m_j(1-\hat{\pi}_j) + m_j \hat{\pi}_j}, \tag{2.6}$$

kde m_j je počet pozorování s *j-tou* kombinací hodnot vysvětlujících proměnných. Pokud konstantou vynásobím pouze odhadovaný počet pozorování bez výskytu sledovaného jevu, docílím posunu odhadů pravděpodobnosti žádaným směrem a zároveň zachovám rozsah odhadů v intervalu [0,1]. Z (2.6) lze počet těchto pozorování vyjádřit:

$$m_j(1-\hat{\pi}_j) = \frac{(1-\hat{\pi}_j)}{\hat{\pi}_j} m_j \hat{\pi}_j \,. \tag{2.7}$$

Kalibrovaný odhad pak má tvar:

$$\hat{\pi}_{j_C}(\alpha) = \frac{m_j \hat{\pi}_j}{\alpha m_j (1 - \hat{\pi}_j) + m_j \hat{\pi}_j},$$

což lze pomocí (2.7) zapsat:

$$\hat{\pi}_{j_C}(\alpha) = \frac{m_j \hat{\pi}_j}{\alpha \frac{(1 - \hat{\pi}_j)}{\hat{\pi}_j} m_j \hat{\pi}_j + m_j \hat{\pi}_j} = \frac{1}{1 + \alpha \frac{(1 - \hat{\pi}_j)}{\hat{\pi}_j}}. \tag{2.8}$$

Hodnotu parametru α dopočtu aplikací vztahu (2.8) na celý vzorek, kdy kalibrovaný odhad $\hat{\pi}_{j_C}$ nahradím cílovou hodnotou průměru odhadů $\bar{\pi}_T$ a odhad $\hat{\pi}_j$ průměrem odhadů $\bar{\pi}$. Dostanu:

$$\bar{\pi}_T = \frac{1}{1 + \alpha \frac{(1 - \bar{\pi})}{\bar{\pi}}}$$

a odtud

$$\alpha = \frac{\bar{\pi}}{1 - \bar{\pi}} \frac{1 - \bar{\pi}_T}{\bar{\pi}_T}. \tag{2.9}$$

Finální vztah pro posun odhadnutých pravděpodobností získám dosazením (2.9) do (2.8):

$$\hat{\pi}_{j_C} = \frac{1}{1 + \frac{\bar{\pi}}{1 - \bar{\pi}} \frac{1 - \bar{\pi}_T}{\bar{\pi}_T} \frac{(1 - \hat{\pi}_j)}{\hat{\pi}_j}} = \frac{\hat{\pi}_j (1 - \bar{\pi}) \bar{\pi}_T}{(1 - \hat{\pi}_j) \bar{\pi} (1 - \bar{\pi}_T) + \hat{\pi}_j (1 - \bar{\pi}) \bar{\pi}_T}. \tag{2.10}$$

Na závěr je nutno podotknout, že tímto způsobem průměrnou odhadovanou pravděpodobnost k té cílové pouze přiblížíme, i když často velice blízko. Aby nastala rovnost těchto hodnot, musely by všechny kombinace hodnot vysvětlujících proměnných být stejně zastoupené. V opačném případě rozdílnost vah jednotlivých kombinací způsobí, že výsledný průměr kalibrovaných odhadů nedosáhne přesně průměru cílového. Pokud by rozdíl mezi těmito hodnotami nebyl zanedbatelný, je možné celý postup opakovat a kalibrované odhady opětovně posunovat tak, aby se jejich průměr dostatečně přiblížil k požadované cílové hodnotě.

3 Výpočetní prostředí

Pro úpravu dat, analýzu proměnných, výpočet regresních koeficientů, testování, validaci i kalibraci modelu budu používat statistický software SAS 9.1, konkrétně jeho základní modul, se kterým se pracuje pomocí programovacího jazyku SAS. Protože budu pracovat se skutečnými daty, jsem při jejich zpracování omezen pouze na tento statistický program. V této kapitole stručně popíši nejdůležitější procedury, které budu využívat. Čerpat při tom budu z uživatelské dokumentace programu [13].

DATA step – Základní nástroj pro přípravu datových souborů. Jedná se o sadu příkazů, pomocí nichž lze vytvářet nové, či upravovat stávající datové soubory.

PROC SQL – Implementace jazyku SQL pro SAS. SQL je standardizovaný dotazovací jazyk, který umožňuje snadnou manipulaci s daty ve formě relačních databází.

PROC SURVEYSELECT – Procedura pro různé varianty vytváření náhodných výběrů z datových souborů.

PROC MEANS – Počítá základní popisné statistiky jako je počet nechybějících pozorování, průměr, výběrová směrodatná odchylka, výběrový rozptyl, maximum, minimum, percentily a další.

PROC FREQ - Jedno a vícerozměrné tabulky četností a procentuálních zastoupení. Počítá míry závislosti a shody a nejrůznější statistické testy (např. pomocí příkazu SMDCR *Somerovo D*, čili *Giniho statistiku*).

PROC CORR – Procedura pro výpočet míry korelace. Nabízí jednu parametrickou metodu (*Pearsonův korelační koeficient*) a tři neparametrické (*Spearmenův koeficient pořadové korelace*, *Kendallův koeficient* a *Hoeffdingovu míru závislosti*). Pro *Pearsonův* a *Spearmenův koeficient* dále procedura pomocí *Fisherovy Z transformace* odvodí věrohodnostní interval a p-hodnotu testu hypotézy, že daný koeficient je roven nule.

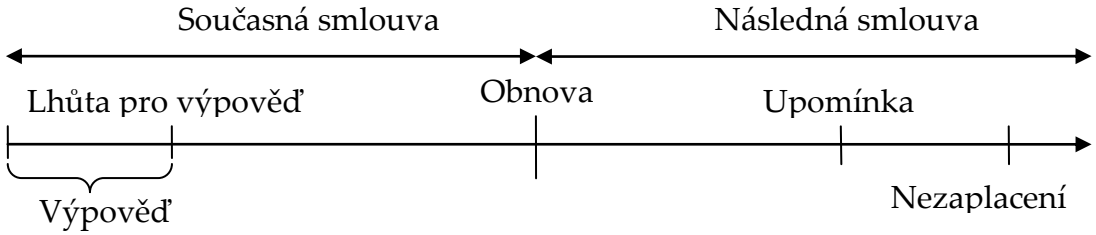
PROC LOGISTIC – Procedura, která metodou maximální věrohodnosti odhaduje koeficienty modelu logistické regrese. Pro numerický výpočet odhadů užívá volitelně buď *Newton-Raphsonův algoritmus* či *Fisherův skórovací algoritmus*. Umožňuje vstup spojitých i kategoriálních vysvětlujících proměnných a nastavení jedné ze čtyř metod výstavby modelu (*Forward, Backward, Stepwise* a *Best Subsets*). Jako výstup poskytuje kromě hodnot

odhadů, také testy významnosti jednotlivých parametrů, tři testy statistické významnosti celého modelu (*test věrohodnostního poměru*, *Waldův test* a *Score test*), několik různých statistik vyjadřujících prediktivní sílu modelu a *Hosmer-Lemeshow test těsnosti modelu*. Dále také pomocí této procedury lze na základě vytvořeného modelu přiřadit odhady pravděpodobností k pozorováním z dalších datových souborů.

4 Vývoj modelu

Jak bylo řečeno ve druhé kapitole, smlouvy povinného ručení jsou uzavírány na jeden rok. Dva měsíce před vypršením tohoto období dostane klient návrh nové smlouvy, a pokud nedojde k výpovědi, kterou je možné provést nejpozději šest týdnů před vypršením smlouvy původní, nová smlouva se stane aktivní.

Cílem této kapitoly je vyvinout model, který by pro každou smlouvu povinného ručení na základě informací, kterými pojišťovna disponuje, odhadl pravděpodobnost, že u dané smlouvy dojde ke stornu během obnovy. U každé smlouvy tedy budu pozorovat konec pojistného období a začátek pojistného období smlouvy následné a zkoumat zda v uvedených obdobích došlo k zániku výpovědí či pro nezaplacení pojistného. Sledovat tedy budu období mezi zasláním návrhu na novou smlouvu a lhůtou pro výpověď, která je 6 týdnů před výročím smlouvy a nejzazší termín pro zaplacení pojistného, který je jeden měsíc od zaslání upomínky (Viz obrázek 3).



Obrázek 3: Chronologické znázornění událostí okolo obnovy smlouvy

Jako závislá veličina tedy do modelu vstupuje binomická proměnná nabývající hodnoty 1, pakliže ve sledovaném období došlo ke stornu a hodnoty 0 v opačném případě. Jako nezávislé proměnné do modelu vstupují informace o klientovi (věk, pohlaví, bydliště), informace o vozidle, jehož provozu se pojištění týká (typ, původ, technické parametry, stáří, cena, atd.) a informace o nové i původní smlouvě (výše pojistného, bonus a malus, datum vzniku, atd.).

4.1 Popis dat

Pro účely této práce mám k dispozici datový soubor popisující smlouvy povinného ručení uzavřené v období ledna 2009 až června 2011, tedy

smlouvy, u kterých došlo k obnově v období ledna 2010 až června 2012. U každé smlouvy mám navíc informaci, zda po obnově bylo zaplaceno pojistné, neboli zda byla obnovená smlouva stornována z důvodu nezaplacení pojistného, nebo ne.

Během přípravy dat jsem zjistil, že případů včasné výpovědi je ve vývojovém vzorku velmi málo a jejich četnost je navíc v období po zaslání návrhu na prodloužení smlouvy obdobná jako ve zbytku roku. Pravděpodobně tedy nesouvisí s obnovou. Budu se tedy věnovat pouze stornům z důvodu nezaplacení pojistného.

Z datového souboru odstraním všechny smlouvy, u kterých během obnovy došlo k jinému než mnou sledovanému stornu a dále ty, které byly stornovány mimo období obnovy.

Pro vývoj modelu použiji smlouvy s obnovami v letech 2010 a 2011, které dále náhodně rozdělím v poměru 70:30 na vývojový a testovací vzorek. Smlouvy s obnovami v období ledna až března 2012 použiji k ověření validity modelu v průběhu času a smlouvy s obnovami v dubnu až červnu 2012 jako aktuální portfolio pro analýzu reprezentativnosti. V následující tabulce uvádím informace o jednotlivých vzorcích.

Vzorek	Celkový počet smluv	Počet stornovaných smluv	Stornovost
Vývojový	45 381	2 546	5,61 %
Testovací	19 402	1 098	5,66 %
Validační	9 525	452	4,75 %
Aktuální	10 944	---	---

Tabulka 1: Datové vzorky

4.2 Reprezentativnost vzorku

Pro použitelnost a funkčnost modelu je velice důležité, aby vzorek, na kterém je vyvíjen, co nejvíce odpovídal vzorku, na který bude aplikován. Protože vývoj bude probíhat na celém portfoliu, jediným rizikem zůstává změna portfolia v průběhu času. Ta může být způsobena fluktuací klientů a změnami u stávajících klientů.

Provedu proto porovnání vývojového vzorku smluv s obnovami v letech 2010 a 2011 s nejaktuálnějším vzorkem, který mám k dispozici, tedy se vzorkem smluv s obnovami v období dubna až června 2012. Pro porovnání použiji několik základních parametrů smlouvy: region, věk a pohlaví řidiče, stáří automobilu a výši pojistného. Každou z těchto proměnných rozdělím do několika kategorií a pomocí *Population stability indexu* porovnám změny zastoupení v jednotlivých kategoriích. Podrobnější

analýzu zaměřenou na konkrétní proměnné vybrané do modelu potom provedu při validaci modelu.

REGION	Vývojový vzorek	Procent. podíl	Aktuální portfolio	Procent. podíl	PSI
Hlavní město Praha	7 677	16,9 %	1 708	15,6 %	0,0011
Jihočeský	2 537	5,6 %	642	5,9 %	0,0001
Jihomoravský	3 498	7,7 %	790	7,2 %	0,0003
Karlovarský	1 398	3,1 %	384	3,5 %	0,0006
Královéhradecký	2 195	4,8 %	527	4,8 %	0,0000
Liberecký	2 358	5,2 %	651	5,9 %	0,0010
Moravskoslezský	3 768	8,3 %	864	7,9 %	0,0002
Olomoucký	1 815	4,0 %	447	4,1 %	0,0000
Pardubický	1 797	4,0 %	416	3,8 %	0,0001
Plzeňský	2 812	6,2 %	666	6,1 %	0,0000
Středočeský	7 022	15,5 %	1 738	15,9 %	0,0001
Vysočina	1 519	3,3 %	365	3,3 %	0,0000
Zlínský	1 741	3,8 %	467	4,3 %	0,0005
Ústecký	5 244	11,6 %	1 279	11,7 %	0,0000
Celkem	**45 381**	**100,0 %**	**10 944**	**100,0 %**	**0,0040**

Tabulka 2: Analýza reprezentativnosti, Region

VĚK ŘIDIČE	Vývojový vzorek	Procent. podíl	Aktuální portfolio	Procent. podíl	PSI
< 25	1 235	2,7 %	238	2,2 %	0,0012
26 - 35	14 118	31,1 %	2 914	26,6 %	0,0070
36 - 45	12 516	27,6 %	3 194	29,2 %	0,0009
46 - 55	7 579	16,7 %	1 882	17,2 %	0,0001
56 - 65	7 120	15,7 %	1 939	17,7 %	0,0025
66 - 75	2 385	5,3 %	670	6,1 %	0,0013
> 75	428	0,9 %	107	1,0 %	0,0000
Celkem	**45 381**	**100,0 %**	**10 944**	**100,0 %**	**0,0131**

Tabulka 3: Analýza reprezentativnosti, Věk řidiče

POHLAVÍ ŘIDIČE	Vývojový vzorek	Procent. podíl	Aktuální portfolio	Procent. podíl	PSI
Žena	11 810	26,0 %	2 729	24,9 %	0,0005
Muž	33 571	74,0 %	8 215	75,1 %	0,0002
Celkem	**45 381**	**100,0 %**	**10 944**	**100,0 %**	**0,0006**

Tabulka 4: Analýza reprezentativnosti, Pohlaví řidiče

STÁŘÍ VOZIDLA	Vývojový vzorek	Procent. podíl	Aktuální portfolio	Procent. podíl	PSI
<= 5 let	9 282	20,5 %	2 247	20,5 %	0,0000
6 - 10 let	14 165	31,2 %	3 289	30,1 %	0,0004
11 - 15 let	14 324	31,6 %	3 695	33,8 %	0,0015
16 - 20 let	4 594	10,1 %	1 119	10,2 %	0,0000
> 20 let	3 016	6,6 %	594	5,4 %	0,0025
Celkem	**45 381**	**100,0 %**	**10 944**	**100,0 %**	**0,0044**

Tabulka 5: Analýza reprezentativnosti, Stáří vozidla

POJISTNÉ (v Kč)	Vývojový vzorek	Procent. podíl	Aktuální portfolio	Procent. podíl	PSI
<= 2 000	4 785	10,5 %	917	8,4 %	0,0050
(2 000 - 3 000]	7 373	16,2 %	1 637	15,0 %	0,0011
(3 000 - 4 000]	7 580	16,7 %	1 868	17,1 %	0,0001
(4 000 - 5 000]	5 803	12,8 %	1 539	14,1 %	0,0012
(5 000 - 6 000]	5 066	11,2 %	1 188	10,9 %	0,0001
(6 000 - 7 000]	3 679	8,1 %	915	8,4 %	0,0001
(7 000 - 8 000]	2 837	6,3 %	696	6,4 %	0,0000
(8 000 - 9 000]	1 981	4,4 %	478	4,4 %	0,0000
(9 000 - 10 000]	1 353	3,0 %	372	3,4 %	0,0005
> 10 000	4 924	10,9 %	1 334	12,2 %	0,0016
Celkem	**45 381**	**100,0 %**	**10 944**	**100,0 %**	**0,0096**

Tabulka 6: Analýza reprezentativnosti, Výše pojistného

Z tabulek 2 až 6 vyplývá, že celkové hodnoty *PSI* jsou pro všechny vybrané ukazatele výrazně menší než 0,1. Lze tedy prohlásit, že portfolio je velmi stabilní v čase. Vývojový vzorek proto není třeba nijak upravovat.

4.3 Analýza proměnných

Ke každé smlouvě mám k dispozici hodnoty 52 proměnných, které se týkají smlouvy samotné, její obnovy, vozidla, ke kterému se smlouva vztahuje a řidičů tohoto vozidla. Pro tyto proměnné jsem provedl dvourozměrnou analýzu, na jejímž základě jsem sestavil dva výběry proměnných, širší a užší, které jsem poté používal jako vstup při výběru modelu.

4.3.1 Širší výběr

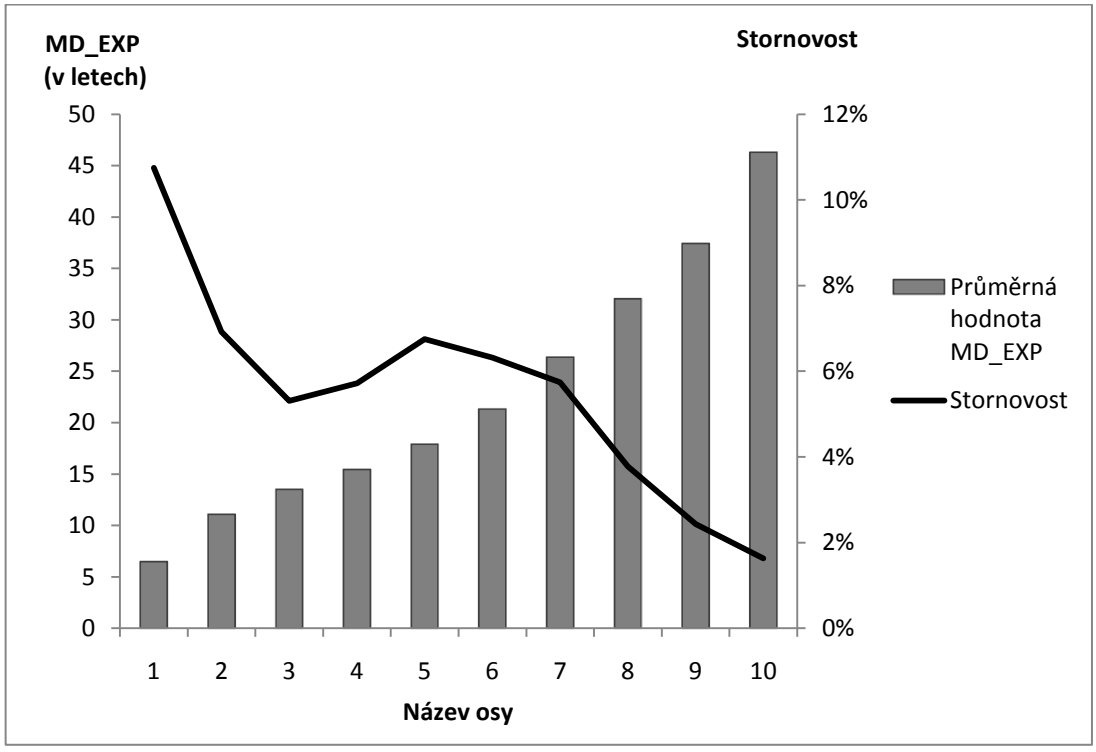
Hodnoty jednotlivých proměnných jsem nejprve rozdělil do skupin. Kvantitativní proměnné jsem rozdělil do maximálně deseti intervalů, tak aby v každém bylo pokud možno přibližně stejné množství pozorování. Kvalitativní proměnné jsem neupravoval. V jednotlivých kategoriích jsem pak spočítal stornovost.

Do širšího výběru jsem zařadil 17 proměnných, u kterých jsem na základě grafického znázornění usoudil, že by mohly mít statisticky významný a logicky vysvětlitelný vliv na míru stornovosti. Dále jsem do širšího výběru přidal 3 nově vytvořené proměnné – rozdíl bonusu/malusu na původní smlouvě a její obnově, rozdíl pojistného na původní smlouvě a její obnově a relativní změnu mezi pojistným na původní smlouvě a její obnově.

Proměnné z širšího výběru jsem poté přerozdělil do nových kategorií, tentokrát tak, abych jejich počet minimalizoval, ale současně nedošlo k výrazné ztrátě *Informační hodnoty*. Jako pomocné kritérium jsem použil hodnoty *Weight of Evidence* a sloučil jsem kategorie s podobnou hodnotou této statistiky.

Pro finální kategorie jednotlivých proměnných jsem znovu spočítal hodnoty *Weight of Evidence*, čímž jsem získal nové proměnné (transformace původních), které mají monotónní vztah ke stornovosti (s rostoucí hodnotou klesá stornovost), nabývají omezeného počtu hodnot (maximálně pěti), jejichž hodnoty odpovídají výši stornovosti (je vidět, jak moc se jednotlivé kategorie od sebe liší) a které jsou všechny ve stejném měřítku a proto snadno vzájemně porovnatelné.

Celý postup prezentuji na příkladu proměnné MD_EXP, neboli počtu let, po které má hlavní řidič vozidla, na které se povinné ručení vztahuje, řidičský průkaz. Proměnnou nejprve rozdělím do 10 kategorií, a protože trend stornovosti odpovídá předpokladu, že s rostoucí hodnotou proměnné bude stornovost klesat, zařadím ji do širšího výběru.



Graf 3: Vztah proměnné MD_EXP a stornovosti

Předpoklad vychází z úvahy, že řidiči mající řidičský průkaz kratší dobu jsou mladší a tudíž jednak méně konzervativní a zároveň méně zodpovědní. Méně konzervativní klienti budu s větší pravděpodobností chtít změnit pojišťovnu a méně zodpovědní nebudou dbát na fakt, že již došlo k obnově smlouvy a jsou povinni po celý rok platit pojistné. Současně se také dá předpokládat, že méně zkušení řidiči budou častěji způsobovat nehody, tím pádem budou mít vyšší pojistné a z toho důvodu budou uvažovat o změně pojistitele. Všechny zmíněné vlivy tedy působí stejným směrem. Kromě toho je nutné poznamenat, že z těchto úvah vyplývá, že proměnná zcela jistě bude silně korelovaná s věkem řidiče a v nějaké míře také pravděpodobně s počtem nehod a výší pojistného. Vztahy proměnných se budu zabývat v další části.

Proměnná je tedy zařazena do širšího výběru. Pro jednotlivé kategorie spočítám hodnoty *Weight of Evidence* a *Informační hodnotu* a pokusím se jejich počet zredukovat.

Kategorie	Min	Max	Pozorování	Storen	Stornovost	WOE	IV
1	0	9	4 994	537	10,75 %	-70,66	0,075513
2	10	12	4 912	340	6,92 %	-22,41	0,006007
3	13	14	3 864	205	5,31 %	5,91	0,000290
4	15	16	4 199	240	5,72 %	-1,972	0,000036
5	17	19	5 020	339	6,75 %	-19,76	0,004716
6	20	23	4 270	270	6,32 %	-12,72	0,001611
7	24	29	4 671	268	5,74 %	-2,378	0,000059
8	30	34	4 158	157	3,78 %	41,522	0,013179
9	35	40	4 823	117	2,43 %	87,159	0,055702
10	41	99	4 470	73	1,63 %	127,54	0,094350
Celkem	**0**	**99**	**45 381**	**2 546**	**5,61 %**	**…**	**0,251463**

Tabulka 7: Analýza proměnné MD_EXP

Sloučím kategorie 2 – 7, kde stornovost kolísá okolo 6 %. Získám tak proměnnou s pěti kategoriemi, jejíž celková *Informační hodnota* je jen nepatrně nižší než u původní proměnné s deseti kategoriemi.

Kategorie	Min	Max	Pozorování	Storen	Stornovost	WOE	IV
1	0	9	4 994	537	10,75 %	-70,66	0,075513
2	10	29	26 936	1 662	6,17 %	-10,108	0,006343
8	30	34	4 158	157	3,78 %	41,522	0,013179
9	35	40	4 823	117	2,43 %	87,159	0,055702
10	41	99	4 470	73	1,63 %	127,54	0,09435
Celkem	**0**	**99**	**45 381**	**2 546**	**5,61 %**	**…**	**0,245087**

Tabulka 8: Kategorizace proměnné MD_EXP

Následuje přehled širšího výběru proměnných s vypočítanou *Informační hodnotou*, *Giniho koeficientem* a testovou statistikou *Score testu* modelu obsahujícího pouze příslušnou proměnnou.

Název proměnné	Giniho koef.	IV	Score test	Počet kat.	Popis
gr_bm_tpl	39,49 %	0,59	1149,6	4	Bonus/malus.
gr_bm_tpl2	41,30 %	0,61	1284,6	5	Bonus/malus na obnovené smlouvě.
gr_bm_tpl_chng	12,49 %	0,18	255,5	4	Změna bonusu/malusu.
gr_car_age	28,87 %	0,31	620,6	4	Stáří vozidla.
gr_car_km	28,03 %	0,31	581,6	4	Najeté kilometry vozidla.
gr_car_value	24,44 %	0,21	460,3	4	Hodnota vozidla.
gr_freq	16,78 %	0,11	271,8	2	Frekvence plateb pojistného.
gr_md_age	20,73 %	0,19	361,2	4	Věk řidiče vozidla.
gr_md_exp	22,66 %	0,21	458,8	5	Počet roků, po které má řidič vozidla ŘP.
gr_parking	12,77 %	0,07	157,5	2	Parkování v garáži / mimo garáž.
gr_pay_method	5,38 %	0,02	45,7	2	Způsob úhrady pojistného.
gr_pocet_skod	4,35 %	0,02	68,3	2	Počet pojistných událostí na smlouvě.
gr_premium	21,07 %	0,18	331,5	5	Pojistné.
gr_premium2	23,15 %	0,21	406,2	5	Nové pojistné na obnovené smlouvě.
gr_premium_chng	22,89 %	0,18	454,7	4	Změna pojistného.
gr_premium_rel_chng	22,41 %	0,18	414,1	4	Relativní změna pojistného.
gr_product	10,34 %	0,13	202,1	2	Typ pojištění.
gr_riders	8,24 %	0,02	55,4	3	Počet připojištění.
gr_seller_team	13,01 %	0,07	149,6	3	Prodejní tým.
md_gender	2,95 %	0	10,9	2	Pohlaví řidiče vozidla.
rnw	19,67 %	0,16	334,1	4	Údaj o kolikátou obnovu smlouvy jde.

Tabulka 9: Širší výběr proměnných

4.3.2 Užší výběr

Do užšího výběru zařadím proměnné na základě *Giniho koeficientu*, *Informační hodnoty* a korelační analýzy.

Nejprve z širšího výběru odeberu proměnné GR_PARKING_WOE, GR_PAY_METHOD_WOE, GR_POCET_SKOD_WOE, GR_RIDERS_WOE a MD_GENDER_WOE, které mají velmi nízkou sílu diverzifikační schopnosti i informační hodnotu.

Dále budu zkoumat vztahy mezi proměnnými. Na základě korelační analýzy lze proměnné rozdělit do 10 skupin, uvnitř kterých existují velmi silné závislosti. Silnou závislost lze dále sledovat mezi skupinami 1a a 1b, 4a a 4b a také 5a, 5b a 5c. Zvýšené korelace se pak vyskytují ještě mezi proměnnými skupin 1, 2 a 3. Přehled absolutních hodnot *Pearsonových korelačních koeficientů* znázorňuje obrázek 4.

	Skupina	gr_bm_tpl	gr_bm_tpl2	gr_bm_tpl_chng	gr_md_age	gr_md_exp	gr_premium	gr_premium2	gr_freq	gr_car_value	gr_car_age	gr_product	gr_car_km	gr_premium_chng	gr_premium_rel_chng	RNW
Skupina		1a	1a	1b	2	2	3	3	4	5a	5a	5b	5b	6	6	7
gr_bm_tpl	1a	1,00	0,89	0,53	0,30	0,31	0,32	0,31	0,23	0,08	0,07	0,09	0,10	0,14	0,14	0,09
gr_bm_tpl2	1a	0,89	1,00	0,49	0,29	0,31	0,30	0,31	0,22	0,09	0,08	0,10	0,10	0,18	0,18	0,10
gr_bm_tpl_c	1b	0,53	0,49	1,00	0,21	0,22	0,22	0,18	0,16	0,03	0,03	0,09	0,07	0,01	0,00	0,01
gr_md_age	2	0,30	0,29	0,21	1,00	0,81	0,25	0,25	0,20	0,03	0,00	0,02	0,15	0,05	0,04	0,06
gr_md_exp	2	0,31	0,31	0,22	0,81	1,00	0,23	0,24	0,22	0,01	0,02	0,00	0,13	0,05	0,05	0,06
gr_premium	3	0,32	0,30	0,22	0,25	0,23	1,00	0,83	0,29	0,18	0,12	0,06	0,19	0,04	0,02	0,02
gr_premium2	3	0,31	0,31	0,18	0,25	0,24	0,83	1,00	0,29	0,27	0,20	0,19	0,18	0,12	0,09	0,09
gr_freq	4	0,23	0,22	0,16	0,20	0,22	0,29	0,29	1,00	0,06	0,12	0,02	0,19	0,02	0,02	0,10
gr_car_value	5a	0,08	0,09	0,03	0,03	0,01	0,18	0,27	0,06	1,00	0,75	0,44	0,31	0,01	0,07	0,12
gr_car_age	5a	0,07	0,08	0,03	0,00	0,02	0,12	0,20	0,12	0,75	1,00	0,45	0,52	0,00	0,06	0,15
gr_product	5b	0,09	0,10	0,09	0,02	0,00	0,06	0,19	0,02	0,44	0,45	1,00	0,26	0,05	0,13	0,10
gr_car_km	5c	0,10	0,10	0,07	0,15	0,13	0,19	0,18	0,19	0,31	0,52	0,26	1,00	0,04	0,06	0,04
gr_prem_c	6	0,14	0,18	0,01	0,05	0,05	0,04	0,12	0,02	0,01	0,00	0,05	0,04	1,00	0,74	0,20
gr_prem_rel_c	6	0,14	0,18	0,00	0,04	0,05	0,02	0,09	0,02	0,07	0,06	0,13	0,06	0,74	1,00	0,19
RNW	7	0,09	0,10	0,01	0,06	0,06	0,02	0,09	0,10	0,12	0,15	0,10	0,04	0,20	0,19	1,00

Obrázek 4: Korelační analýza

Do užšího seznamu proměnných vyberu z každé z 10 skupin proměnnou s nejvyššími hodnotami *Giniho koeficientu*, *Informační hodnoty* a testové statistiky *Score testu*.

Užší výběr zobrazuje následující tabulka:

Název proměnné	Sk.	Giniho koef.	IV	Score test	Počet kat.	Popis
gr_bm_tpl2	1a	41,30 %	0,61	1284,6	5	Bonus/malus na obnovené smlouvě.
gr_bm_tpl_chng	1b	12,49 %	0,18	255,5	4	Změna bonusu/malusu.
gr_md_exp	2	22,66 %	0,21	458,8	5	Počet roků, po které má řidič vozidla ŘP.
gr_premium2	3	23,15 %	0,21	406,2	5	Nové pojistné na obnovené smlouvě.
gr_freq	4	16,78 %	0,11	271,8	2	Frekvence placení pojistného.
gr_car_age	5a	28,87 %	0,31	620,6	4	Stáří vozidla.
gr_product	5b	10,34 %	0,13	202,1	2	Typ pojistění.
gr_car_km	5c	28,03 %	0,31	581,6	4	Najeté kilometry vozidla.
gr_premium_chng	6	22,89 %	0,18	454,7	4	Změna pojistného.
rnw	7	19,67 %	0,16	334,1	4	Údaj o kolikátou obnovu smlouvy jde.

Tabulka 10: Užší výběr proměnných

4.4 Logistická regrese

Model logistické regrese zkonstruuji několika různými způsoby. Pro výběr proměnných použiji metody *Best Subsets*, která na základě testové statistiky *Score testu* vybere pro všechny možné velikosti modelu tu nejlepší kombinaci proměnných a metodu *Stepwise selection*, která začne s modelem obsahujícím pouze absolutní člen, poté v každém kroku přidá nejvýznamnější proměnou a následně provede zpětnou kontrolu významnosti proměnných, které již jsou v modelu obsaženy. U metody *Stepwise selection* navíc vyzkouším různá nastavení hraničních hodnot pro vstup a vyřazování proměnných. Jako počáteční množiny kandidátů pro obě výše zmíněné metody použiji jednak širší výběr vzniklý pouze na základě grafické analýzy závislosti stornovosti na hodnotách jednotlivých proměnných a také užší výběr vytvořený na základě analýz *Giniho statistiky*, *Informační hodnoty* a korelační analýzy.

Vždy budu navíc konstruovat jednak plný logistický model, ve kterém každé kategorii každé proměnné odhadnu vlastní koeficient a pak *Weight of Evidence* model, kde jednotlivé kategorie každé proměnné mají přiřazenou *WOE* hodnotu, proměnné pak vstupují do modelu jako spojité a pro každou je odhadnut jen jeden regresní koeficient.

Nejlepší modely vybrané pomocí automatických procedur poté budu podrobněji zkoumat, případně je manuálně upravovat a na závěr z nich vyberu finální logistický model. Finální model by měl být co nejstabilnější a zároveň by měl mít co největší prediktivní a diverzifikační sílu. Mým cílem tedy bude najít model co možná nejjednodušší, s proměnnými, které budou vystihovat obecně platné a ekonomicky interpretovatelné jevy, model, který bude stabilní v čase a který bude dobře rozlišovat klienty s vysokým rizikem storna během obnovy smlouvy.

4.4.1 Best Subsets

Pomocí procedury PROC LOGISTIC s volbou SELECTION = SCORE a BEST = 1 najdu pro všechny možné velikosti modelu sadu proměnných s nejvyšší hodnotou testové statistiky *Score testu*. Na základě aproximace C_q statistiky pomocí hodnot testové statistiky *Score testu* poté vyberu model. Jak bylo řečeno v 2.10.2, nejlepším kandidátem je model, který má hodnotu této statistiky nejblíže *(q+1)*, kde *q* je počet proměnných v modelu. Protože touto procedurou nelze porovnávat modely s kategoriálními proměnnými, použiji pouze *WOE* varianty proměnných a teprve pro vybrané sady proměnných poté zkonstruuji jak *WOE* model, tak plný logistický model.

Počet proměnných v modelu	**Testová statistika *Score testu***	**C_q statistika založená na *Score testu***
1	1284,61	1472,97
2	1777,47	982,11
3	2124,49	637,08
4	2331,65	431,93
5	2455,27	310,30
6	2524,25	243,33
7	2577,64	191,94
8	2617,48	154,10
9	2655,40	118,18
10	2688,60	86,97
11	2720,74	56,83
12	2734,89	44,68
13	2747,98	33,60
14	2759,36	24,22
15	2765,53	20,05
16	**2771,42**	**16,15**
17	2773,75	15,83
18	2774,72	16,86
19	2775,57	18,00
20	2775,58	20,00

21	2775,58	22,00

Tabulka 11: Procedura Best subsets pro širší výběr

Počet proměnných v modelu	**Testová statistika** *Score testu*	**C_q statistika založená na** *Score testu*
1	1284,61	1276,35
2	1777,47	785,49
3	2124,49	440,47
4	2331,65	235,31
5	2455,27	113,69
6	2524,25	46,71
7	2544,90	28,06
8	2558,46	16,50
9	2565,94	11,02
10	**2567,96**	**11**

Tabulka 12: Procedura Best subsets pro užší výběr

4.4.2 Stepwise selection

Metodu postupného výběru proměnných aplikuji pomocí procedury PROC LOGISTIC s parametry SELECTION = STEPWISE a SLENTRY = α a SLSTAY = β pro nastavení hraničních p-hodnot potřebných pro vstup (α) a vyřazení (β). Pro vstup do procedury použiji postupně širší i užší výběr. Pomocí různých variant proměnných budu konstruovat *WOE* model, plný logistický model a také kombinovaný model. Pro posledně jmenovaný případ nechám do procedury vstoupit všechny proměnné jak v původní kategoriální variantě tak ve *WOE* verzi. V případě, že procedura vybere do modelu nějakou proměnnou v obou variantách, spustím proceduru znovu a nechám do ní vstoupit pouze tu variantu proměnné, která byla vybrána jako první.

Tabulka 13 zobrazuje přehled postupného výběru proměnných pro různá nastavení hranic pro vstup a vyřazení z modelu. Čísla vyjadřují pořadí vstupu do modelu. U kombinovaného modelu je navíc v závorce varianta vybrané proměnné (W – *WOE*, K – kategoriální). Jednou hvězdičkou jsou označeny proměnné, které byly vybrány při použití defaultního (středně přísného) nastavení navíc oproti nejpřísnějšímu nastavení. Dvě hvězdičky pak analogicky označují proměnné, o které se modely rozrostly při použití nejmírnějšího nastavení.

		Výběr: Typ modelu:	Širší WOE	Širší Plný	Širší Komb	Užší WOE	Užší Plný	Užší Komb
Širší výběr	Užší výběr	gr_bm_tpl2	1	1	1(K)	1	1	1 (K)
		gr_bm_tpl_chng	**19	16	*16(K)	*10	**10	*10(W)
		gr_md_exp	6	11	12(W)	6	6	6 (K)
		gr_premium2	4	3	3(K)	4	3	3 (K)
		gr_freq	*15	**20		9	9	*9 (K)
		gr_car_age	2	2	2(K)	2	2	2 (K)
		gr_product	14	6	5(K)	8	5	5 (W)
		gr_car_km	7	9	8(W)	7	8	8 (W)
		gr_premium_chng	5	7	7(W)	5	7	7 (W)
		rnw	3	4	4(W)	3	4	4 (W)
		gr_bm_tpl	9	14	14(W)			
		gr_car_value	12	12	10(W)			
		gr_md_age	**16	5	6 (K)			
		gr_parking	**18	**19	**18(K)			
		gr_pay_method	13	10	11(K)			
		gr_pocet_skod	8	8	9(W)			
		gr_premium		**18	**17(K)			
		gr_premium_rel_chng	**17	*17				
		gr_riders	11	15	15(W)			
		gr_seller_team						
		gr_md_gender	10	13	13(K)			
Celkový počet prom.		(α = 0,01 ; β = 0,01)	14	16	15	9	9	8
		(α = 0,05 ; β = 0,05)	15	17	16	10	9	10
		(α = 0,15 ; β = 0,15)	19	20	18	10	10	10

Tabulka 13: Přehled výstupů procedury *Stepwise selection*

Při mírnějším nastavení parametrů, $\alpha = \beta = 0{,}15$, jsou jednotlivé modely o několik proměnných bohatší. Pro užší výběr nastane změna pouze u plného modelu, do kterého je zařazena i poslední desátá proměnná GR_BM_TPL_CHNG, zbylé dva modely obsahovaly všech 10 proměnných již pro přísněji nastavené parametry α a β. Pro úplnost uvedu, že pro širší výběr se WOE model, plný model a kombinovaný model rozrostly o 4, 3 a 2 proměnné, nicméně tyto modely byly až příliš bohaté již pro původní nastavení parametrů, nebudu se jim proto věnovat detailněji.

Naopak při přísnějším nastavení parametrů, $\alpha = 0{,}01$, $\beta = 0{,}01$, se všechny tři modely vzešlé z širšího výběru o jednu proměnnou zjednodušily. Pro plný model z užšího výběru nenastala žádná změna, do *WOE* a kombinovaného modelu nevstoupila desátá proměnná GR_BM_TPL_CHNG a do kombinovaného modelu navíc ani devátá proměnná GR_FREQ.

4.4.3 Porovnání modelů

Pro všechny modely sestavené v předešlých sekcích spočítám *Giniho koeficient*, jako vyjádření diverzifikační síly, a p-hodnotu *Hosmer-Lemeshow testu* pro určení míry těsnosti modelu. Oboje nejprve pro *vývojový vzorek* (45 381 případů, stornovost 5,61 %), na základě kterého byly odhadnuty konkrétní regresní koeficienty. Následně pomocí jednotlivých modelů odhadnu pravděpodobnosti storna u smluv v *testovacím vzorku* (19 402 případů, stornovost 5,66 %) a výše zmíněné statistiky spočtu i pro tento vzorek. Vše shrnuje následující tabulka.

Č.	Množina prom.	Typ modelu	Procedura výběru	Počet prom.	DF	Vývoj		Test	
						Gini	HL	Gini	HL
1	Širší	WOE	Best Subsets	16	16	59,45%	99%	59,52%	15%
2	Širší	Plný	Best Subsets	16	38	60,40%	77%	60,44%	18%
3	Užší	WOE	Best Subsets	10	10	58,05%	82%	58,17%	14%
4	Užší	Plný	Best Subsets	10	27	59,16%	5%	59,41%	13%
5	Širší	WOE	Stpw. (0,05)	15	15	59,42%	95%	59,49%	13%
6	Širší	Plný	Stpw. (0,05)	17	43	60,62%	83%	60,70%	5%
7	Širší	Komb	Stpw. (0,05)	16	28	60,49%	56%	60,75%	5%
8	Užší	WOE	Stpw. (0,05)	10	10	58,05%	82%	58,17%	14%
9	Užší	Plný	Stpw. (0,05)	9	24	59,16%	5%	59,41%	13%
10	Užší	Komb	Stpw. (0,05)	10	19	59,19%	31%	59,51%	33%
12	Širší	WOE	Stpw. (0,01)	14	14	59,34%	99%	59,38%	12%
13	Širší	Plný	Stpw. (0,01)	16	40	60,57%	87%	60,60%	3%
14	Širší	Komb	Stpw. (0,01)	15	25	60,37%	62%	60,66%	21%
15	**Užší**	**WOE**	**Stpw. (0,01)**	**9**	**9**	**58,00%**	**43%**	**58,19%**	**12%**
16	Užší	Plný	Stpw. (0,01)	9	24	59,16%	5%	59,41%	13%
17	**Užší**	**Komb**	**Stpw. (0,01)**	**8**	**17**	**59,03%**	**0%**	**59,43%**	**30%**

Tabulka 14: Modely 1 až 17 - přehled

Z tabulky je patrné, že modely vzniklé ze širšího výběru jsou výrazně bohatší, ale nepřinášejí příliš velké zlepšení v žádném ze sledovaných kritérií. Na základě minimalizace počtu proměnných a stupňů volnosti (*DF*) a současné maximalizace *Giniho koeficientu* a p-hodnoty *H.-L. testu* pro testovací vzorek bych za nejlepší označil jeden z modelů 15 a 17.

4.4.4 Manuální výstavba modelu

V této sekci se pokusím s využitím výsledků automatických procedur výběru proměnných sestavit model manuálně. Jako základ použiji 4 proměnné, které byly ve všech případech procedurou *Stepwise selection* vybrány v prvních čtyřech krocích a současně procedurou *Best subsets* jako nejlepší model se čtyřmi proměnnými. Jsou to proměnné bonus/malus na nové smlouvě (GR_BM_TPL2), stáří vozidla (GR_CAR_AGE), pořadí obnovy (RNW) a výše pojistného na nové smlouvě (GR_PREMIUM2). Varianty

proměnných použiji podle výběru *Stepwise selection* procedury pro kombinovaný model. V obou případech (pro širší i užší výběr) procedura zařadila do modelu *WOE* variantu proměnné RNW a kategoriální varianty zbývajících tří proměnných. Model logistické regrese pro tyto 4 proměnné odhadnutý opět pomocí procedury PROC LOGISTIC má následující „výkonnostní parametry“:

Č.	Výběr	Typ modelu	Procedura výběru	Počet prom.	DF	Vývoj		Test	
						Gini	HL	Gini	HL
18	Širší	Komb	Manuální	4	12	56,24%	56%	56,82%	93%

Tabulka 15: Model 18 - přehled

Model má jen mírně slabší diverzifikační sílu (o 1,37 % a 2,61 % oproti modelům 15 a 17) a vykazuje velmi dobré výsledky ohledně míry těsnosti. Naopak další zjednodušování modelu již vede k poměrně výrazným poklesům diverzifikační síly (Tabulka 16).

Odstraněná proměnná	Pokles Gini
gr_bm_tpl2	-6,14 %
gr_car age	-9,24 %
gr_premium2	-3,64 %
RNW_WOE	-2,64 %

Tabulka 16: Změna diverzifikační síly modelu při odebrání jednotlivých proměnných – model 18

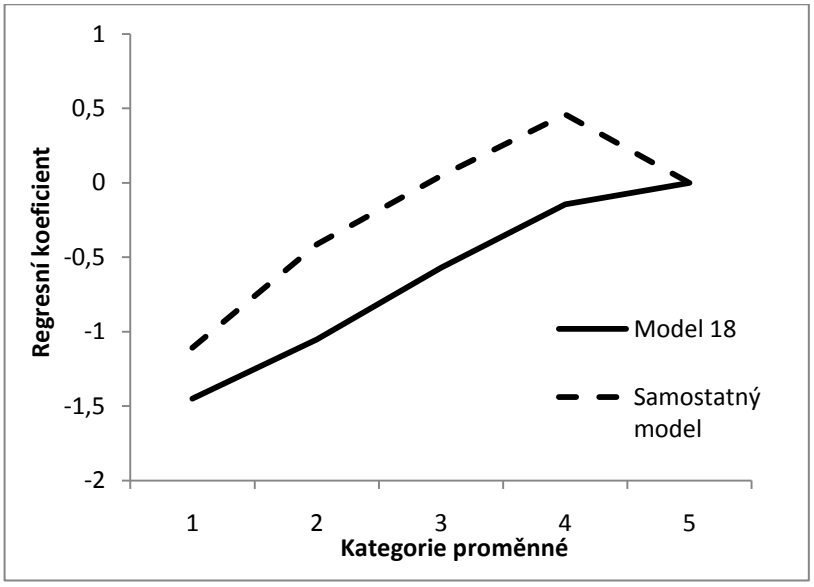
Nyní se na tento základní model podívám detailněji. Následující tabulka je jedním z výstupů SAS procedury PROC LOGISTIC a obsahuje odhady regresních koeficientů a *Waldův test* hypotézy, že příslušný koeficient je roven nule. Pro proměnnou RNW vstupuje do modelu její *WOE* varianta, je na ní tedy nahlíženo jako na spojitou a proto má pouze jeden koeficient. Ostatní proměnné mají koeficient pro každou svou kategorii až na jednu, která je obsažena v pevném členu (interceptu).

Proměnná	Kategorie	DF	Odhad koeficientu	SE	Wald Chi-kv.	p-hondota
Intercept		1	-2,5459	0,1086	549,179	<,0001
gr_premium2	1	1	-1,4512	0,1155	157,945	<,0001
gr_premium2	2	1	-1,0530	0,0930	128,275	<,0001
gr_premium2	3	1	-0,5702	0,0775	54,188	<,0001
gr_premium2	4	1	-0,1445	0,0751	3,702	0,0543
gr_bm_tpl2	1	1	1,6613	0,0847	384,888	<,0001
gr_bm_tpl2	2	1	1,4521	0,0839	299,477	<,0001
gr_bm_tpl2	3	1	0,9622	0,0942	104,388	<,0001
gr_bm_tpl2	4	1	0,7890	0,0914	74,575	<,0001
gr_car_age	1	1	-2,4567	0,1315	348,929	<,0001

gr_car_age	2	1	-1,5435	0,0660	546,383	<,0001
gr_car_age	3	1	-0,8268	0,0489	285,639	<,0001
RNW_WOE		1	-0,0102	0,0006	252,895	<,0001

Tabulka 17: Odhad regresních koeficientů – model 18

Z tabulky 17 vyplývá, že čtvrtá kategorie proměnné GR_PREMIUM2 je na hranici statistické významnosti. Z grafického porovnání koeficientů v modelu 18 a v modelu obsahujícím pouze tuto proměnnou je navíc vidět, že spojnice jednotlivých hodnot nemají stejný sklon. Zatímco v modelu s dalšími třemi proměnnými má GR_PREMIUM2 na stornovost monotónní vliv a vyššímu pojistnému jsou přiřazeny vyšší koeficienty, v modelu, který žádné další proměnné neobsahuje a vystihuje tedy pouze přímý vliv této samostatné proměnné, má křivka koeficientů maximum ve čtvrté kategorii a směrem k páté pak klesá.



Graf 4: Porovnání regresních koeficientů proměnné GR_PREMIUM2

Problém se pokusím vyřešit novým přerozdělením proměnné do kategorií. Původní rozdělení bylo následující.

Kat.	N	Storna	Stornovost
< 2 500	7 134	140	1,96%
< 3 500	7 726	298	3,86%
< 6 000	14 703	878	5,97%
< 10 000	10 679	936	8,76%
>=10 000	5 139	294	5,72%

Tabulka 18: Rozdělení proměnné Gr_premium2

Rostoucí stornovost s rostoucím pojistným je logický a očekávaný fakt. To, že se od určité hranice tento trend obrací, je možné interpretovat tak, že klienti s nejvyššími hodnotami pojistného jsou pravděpodobně movitější a cenu již nevnímají tak citlivě. Tuto hypotézu se však pomocí logistické

regrese nepodařilo prokázat a tak dvě nejvyšší kategorie sloučím dohromady.

Kat.	N	Storna	Stornovost
< 2 500	7 134	140	1,96%
< 3 500	7 726	298	3,86%
< 6 000	14 703	878	5,97%
>= 6 000	15 818	1 230	7,78%

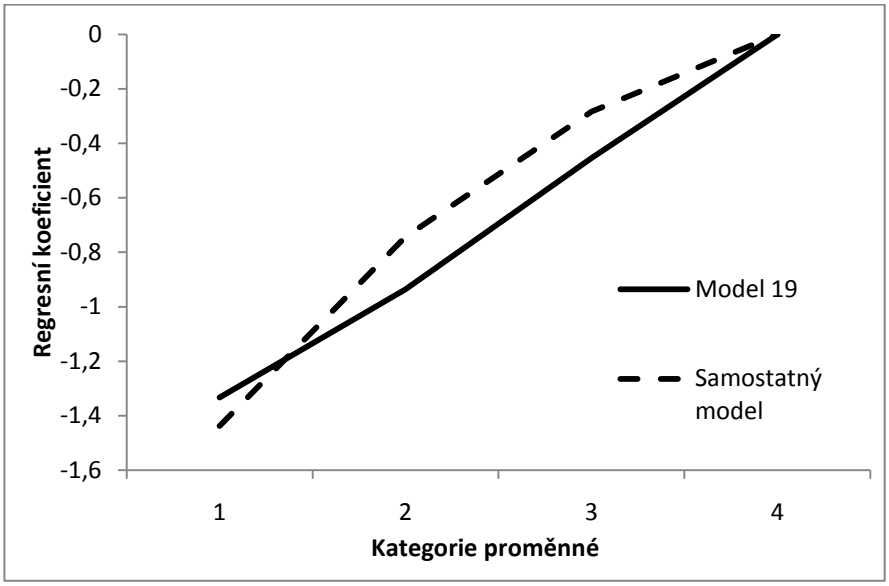
Tabulka 19: Rozdělení proměnné Gr_premium2_new

Model s nově rozdělenou proměnnou označím číslem 19 a porovnám jej s předchozí verzí.

Č.	Výběr	Typ modelu	Procedura výběru	Počet prom.	DF	Vývoj		Test	
						Gini	HL	Gini	HL
18	Širší	Komb	Manuální	4	12	56,24%	56%	56,82%	93%
19	Širší	Komb	Manuální	4	11	56,18%	56%	56,70%	87%

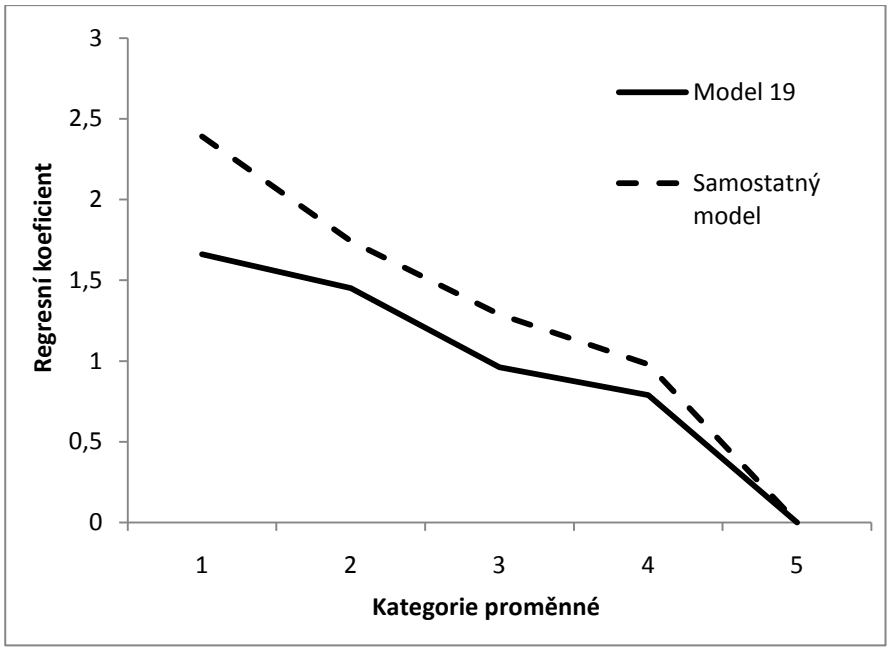
Tabulka 20: Modely 18 a 19 - přehled

Došlo k mírnému zhoršení diverzifikační síly a míry těsnosti, ovšem na druhou stranu se také snížil počet stupňů volnosti. Nekonzistence vlivu samostatné proměnné a proměnné jako součásti širšího modelu byla odstraněna.

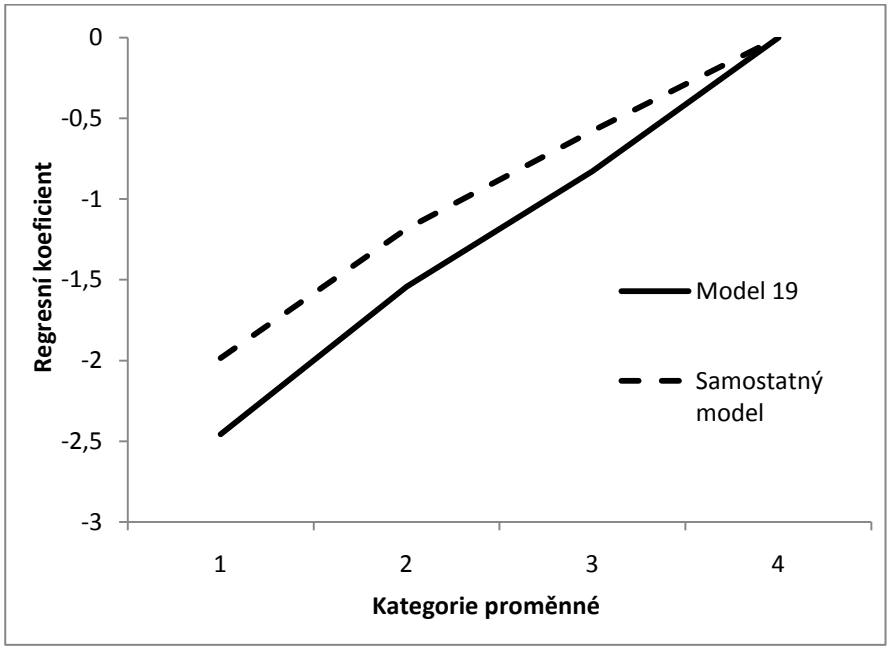


Graf 5: Porovnání regresních koeficientů proměnné GR_PREMIUM2_NEW

U zbylých dvou kategoriálních proměnných podobný problém nenastává.



Graf 6: Porovnání regresních koeficientů proměnné GR_BM_TPL2



Graf 7: Porovnání regresních koeficientů proměnné GR_CAR_AGE

Nyní se pokusím model rozšířit obdobným postupem, na kterém je založena metoda *Forward selection*, s tím rozdílem, že za kritérium pro vstup do modelu použiji přínos k hodnotě *Giniho statistiky* modelu. Jako potenciální kandidáty pro vstup do modelu vyberu proměnné, které byly v jednotlivých modelech vybrány procedurou *Stepwise selection* jako páté až deváté v pořadí (na dalších pořadích se již spektrum proměnných příliš rozšiřuje). Z těchto kandidátů vyřadím proměnnou bonus/malus na původní smlouvě (GR_BM_TPL), která je velice silně korelovaná s analogickou proměnnou pro novou smlouvu (GR_BM_TPL2) a ta již je v modelu zastoupena. V každém kroku nejprve spočítám, o kolik by vzrostla diverzifikační síla modelu po

rozšíření o jednotlivé proměnné a poté tu nejsilnější v tomto směru přidám do modelu.

Proměnná	Varianta proměnné	
	Kategoriální	WOE
gr_product	0,6	0,6
gr_car_km	0,7	0,7
gr_premium_chng	0,8	0,8
gr_md_exp	0,8	0,8
gr_md_age	0,9	0,7
gr_pocet_skod	0,4	0,4
gr_freq	0,2	0,2

Tabulka 21: Přehled změn *Giniho statistiky* modelu po přidání jednotlivých proměnných – krok 1

Model rozšířím o kategoriální variantu proměnné věk řidiče (GR_MD_AGE), která zvýší diverzifikační sílu o 0,9 procentního bodu na celkových 57,1% na vývojovém vzorku. Z množiny kandidátů poté navíc odstraním i proměnnou zkušenosti řidiče, která udává počet let, po které má řidič řidičský průkaz (GR_MD_EXP), protože je s věkem řidiče, který právě vstoupil do modelu, vysoce korelovaná. Během analýzy proměnných jsem do užšího výběru zařadil právě proměnnou zkušenosti řidiče, která samostatně vykazovala lepší výsledky. Nicméně procedura *Stepwise selection* dala ve dvou ze tří případů výběru ze širšího modelu přednost proměnné věk řidiče. A nyní je skutečně vidět, že v kombinaci s ostatními proměnnými, o trochu lépe pomůže vytřídit „špatné" klienty právě tato proměnná. Následně zopakuji předchozí krok.

Proměnná	Varianta proměnné	
	Kategoriální	WOE
gr_product	0,6	0,6
gr_car_km	0,6	0,6
gr_premium_chng	0,7	**0,7**
gr_pocet_skod	0,3	0,3
gr_freq	0,1	0,1

Tabulka 22: Přehled změn *Giniho statistiky* modelu po přidání jednotlivých proměnných – krok 2

Jako další nechám do modelu vstoupit proměnnou udávající rozdíl mezi předepsaným pojistným na nové a staré smlouvě (GR_PREMIUM_CHNG). Protože obě dvě varianty proměnné zvýší *Giniho statistiku* shodně o 0,7 procentního bodu na celkových 57,8 %, zařadím do modelu WOE variantu, která vyžaduje odhad pouze jednoho koeficientu. V množině kandidátů mi zůstanou 4 proměnné.

Proměnná	Varianta proměnné	
	Kategoriální	WOE
gr_product	0,5	0,5
gr_car_km	0,6	0,6
gr_pocet_skod	0,2	0,2
gr_freq	0,1	0,1

Tabulka 23: Přehled změn *Giniho statistiky* modelu po přidání jednotlivých proměnných – krok 3

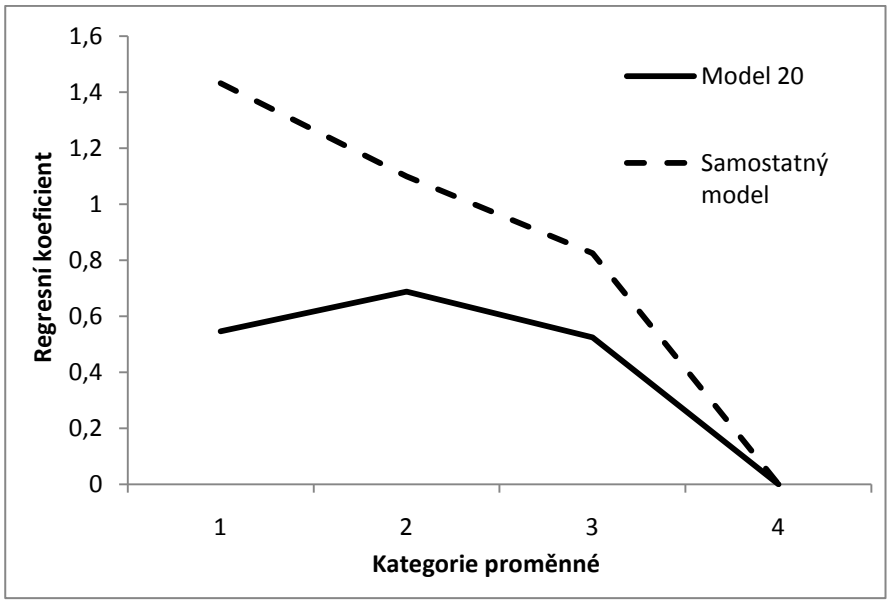
Tentokrát nevyberu proměnnou s nejvyšším příspěvkem k diverzifikační síle, kterou jsou najeté kilometry vozidla (GR_CAR_KM), ale druhou nejlepší proměnnou - typ produktu (GR_PRODUCT). Obě dvě tyto proměnné jsou poměrně silně korelované se stářím vozidla a to už je v modelu zahrnuto. Typ produktu má se stářím vozidla nižší korelační koeficient (0,45 vůči 0,52) a dle mého názoru navíc přináší jiný typ informace. Oproti tomu stáří vozidla a najeté kilometry lze považovat za informaci velmi podobného druhu. I tentokrát zvolím *WOE* variantu proměnné.

Dostávám se tedy k modelu se sedmi proměnnými s *Giniho koeficientem* 58,3%. Další proměnné již přinášejí zlepšení maximálně 0,2 procentního bodu, a proto rozšiřování modelu v tomto kroku ukončím a opět se na model podívám detailněji. Nejprve na konkrétní hodnoty odhadů koeficientů a jejich statistickou významnost.

Proměnná	Kategorie	DF	Odhad koeficientu	SE	Wald Chi-kv.	p-hondota
Intercept		1	-3,1186	0,1060	866,266	<,0001
gr_premium2_new	1	1	-1,2203	0,1002	148,392	<,0001
gr_premium2_new	2	1	-0,8525	0,0725	138,100	<,0001
gr_premium2_new	3	1	-0,4352	0,0507	73,822	<,0001
gr_bm_tpl2	1	1	1,4520	0,0866	281,016	<,0001
gr_bm_tpl2	2	1	1,2203	0,0860	201,438	<,0001
gr_bm_tpl2	3	1	0,7773	0,0955	66,315	<,0001
gr_bm_tpl2	4	1	0,6927	0,0921	56,614	<,0001
gr_car_age	1	1	-2,0653	0,1342	236,897	<,0001
gr_car_age	2	1	-1,3833	0,0659	440,836	<,0001
gr_car_age	3	1	-0,8252	0,0492	281,195	<,0001
RNW_WOE		1	-0,0091	0,0007	190,726	<,0001
gr_md_age	1	1	0,5459	0,0865	39,824	<,0001
gr_md_age	2	1	0,6873	0,0730	88,555	<,0001
gr_md_age	3	1	0,5239	0,0850	37,957	<,0001
gr_premium_chng_WOE		1	-0,0043	0,0005	66,068	<,0001
gr_product_WOE		1	-0,0056	0,0008	44,049	<,0001

Tabulka 24: Odhady regresních koeficientů – model 20

Všechny odhady regresních koeficientů jsou statisticky významné. Pro proměnnou GR_MD_AGE si ale neodpovídají tvary křivek koeficientů modelu 20 a modelu obsahujícího pouze tuto proměnnou. Hodnoty daných odhadů jsou navíc v modelu 20 relativně blízko sebe (vic Graf 8).



Graf 8: Porovnání regresních koeficientů proměnné GR_MD_AGE

S rostoucím věkem klesá stornovost. Koeficient pro nejmladší skupinu je ovšem nižší než bychom očekávali. Možným vysvětlením pro tento jev je skutečnost, že rizikovost této skupiny se již do modelu promítá skrze nějakou jinou proměnnou. Jednou z možností je výše pojistného (GR_PREMIUM2_NEW). Nejmladší klienti mají obecně vyšší pojistné, a protože s výší pojistného roste i pravděpodobnost storna, je možné, že výrazné zastoupení klientů nejmladší věkové kategorie ve vyšších kategoriích pojistného zohledňuje riziko této věkové skupiny dostatečným způsobem a samotné věkové kategorii pak je odhadnut nižší koeficient. Analýzy vztahu dvou zmíněných proměnných nabízí následující tabulky.

GR_PREMIUM2	GR_MD_AGE			
_NEW	< 30	< 45	< 55	>= 55
< 2500	4,8 %	10,9 %	17,1 %	29,6 %
< 3500	12,8 %	15,8 %	18,6 %	20,3 %
< 6000	36,2 %	33,1 %	32,9 %	28,8 %
>= 6000	46,2 %	40,2 %	31,3 %	21,2 %

Tabulka 25: Rozdělení klientů jednotlivých věkových skupin podle výše pojistného na nové smlouvě

GR_PREMIUM2	GR_MD_AGE			
_NEW	< 30	< 45	< 55	>= 55
< 2500	4,7 %	2,5 %	2,7 %	1,1 %
< 3500	4,0 %	5,1 %	3,7 %	2,0 %

< 6000	8,5 %	6,9 %	5,1 %	3,1 %
>= 6000	11,3 %	8,2 %	7,3 %	3,3 %

Tabulka 26: Stornovost, GR_PREMIUM_NEW vs. GR_MD_AGE

Vidíme, že nejmladší klienti opravdu jsou více zastoupeni ve skupinách s vyšším pojistným. Hodnoty stornovosti však rostou jak vertikálně, což odpovídá skutečnosti, že v rámci každé jednotlivé věkové kategorie roste s pojistným i stornovost, tak horizontálně což znamená, že i v rámci jednotlivých kategorií pojistného s klesajícím věkem roste pravděpodobnost storna. Zvýšené zastoupení mladších klientů ve vyšších kategoriích pojistného tedy k dostatečnému zohlednění vyšší stornovosti této kategorie nepostačuje.

Druhou možnou proměnnou se stejnou hypotézou představuje výše bonusu na nové smlouvě (GR_BM_TPL2). I zde jsou mladší klienti více zastoupeni v rizikovějších skupinách, tedy těch s nižším bonusem na pojistném. Následuje analogická analýza jako v předchozím případě.

	GR_MD_AGE			
GR_BM_TPL2	**< 30**	**< 45**	**< 55**	**>= 55**
<= 5 %	27,2 %	14,4 %	11,5 %	7,3 %
<= 15 %	34,9 %	25,5 %	20,6 %	13,2 %
<= 25 %	16,2 %	15,2 %	12,8 %	9,6 %
<= 40 %	14,9 %	21,1 %	18,8 %	19,8 %
> 40 %	27,2 %	14,4 %	11,5 %	7,3 %

Tabulka 27: Rozdělení klientů jednotlivých věkových skupin podle výše bonusu

	GR_MD_AGE			
GR_BM_TPL2	**< 30**	**< 45**	**< 55**	**>= 55**
<= 5 %	14,8 %	15,0 %	14,0 %	9,0 %
<= 15 %	8,6 %	8,3 %	7,8 %	5,4 %
<= 25 %	6,2 %	5,5 %	5,3 %	3,0 %
<= 40 %	5,9 %	4,6 %	3,3 %	1,9 %
> 40 %	1,5 %	2,3 %	1,7 %	0,6 %

Tabulka 28: Stornovost, GR_BM_TPL2 vs. GR_MD_AGE

Opět skutečně pozorujeme tentokrát dokonce výrazně vyšší zastoupení nejmladších klientů v rizikovějších skupinách s nižším bonusem. Oproti předchozímu případu ovšem v těchto rizikovějších skupinách („<= 5 %“, „<= 15 %“ a „<= 25 %“) s rostoucím věkem neklesá stornovost. Až pro nejvyšší věkovou skupinu dojde na všech úrovních bonusu k významnému poklesu.

Protože závislost stornovosti na výši bonusu je u prvních tří kategorií proměnné GR_MD_AGE stejná, výrazně vyšší zastoupení nejmladších klientů v nejrizikovějších kategoriích proměnné GR_BM_TPL2 je

dostatečným projevem vyšší rizikovosti této skupiny a proto je odhad příslušného koeficientu obdobný jako odhady koeficientů ve skupinách 30 až 44 let a 45 až 54 let.

Jestliže se vyšší stornovost ve skupinách „< 30“ a „< 45“ proměnné GR_MD_AGE dostatečně promítá skrze výši bonusu, mohu tyto dvě skupiny sloučit s kategorií „< 55“, tak abych z hlediska věku využil jen dodatečnou informaci o nižší stornovosti u klientů nejvyšší věkové kategorie.

Nyní pro model s upravenou proměnnou GR_MD_AGE_NEW, označím jej číslem 21, znovu odhadnu regresní koeficienty.

Proměnná	Kategorie	DF	Odhad koeficientu	SE	Wald Chi-kv.	p-hondota
Intercept		1	-3,1222	0,1059	869,660	<,0001
gr_premium2_new	1	1	-1,2229	0,1000	149,408	<,0001
gr_premium2_new	2	1	-0,8528	0,0724	138,771	<,0001
gr_premium2_new	3	1	-0,4381	0,0506	74,940	<,0001
gr_bm_tpl2	1	1	1,4520	0,0861	284,111	<,0001
gr_bm_tpl2	2	1	1,2207	0,0855	203,671	<,0001
gr_bm_tpl2	3	1	0,7816	0,0953	67,314	<,0001
gr_bm_tpl2	4	1	0,6997	0,0919	57,934	<,0001
gr_car_age	1	1	-2,0545	0,1341	234,852	<,0001
gr_car_age	2	1	-1,3740	0,0658	436,045	<,0001
gr_car_age	3	1	-0,8189	0,0491	277,821	<,0001
RNW_WOE		1	-0,0090	0,0007	188,905	<,0001
gr_md_age_new	1	1	0,6292	0,0710	78,600	<,0001
gr_premium _chng_WOE		1	-0,0043	0,0005	65,382	<,0001
gr_product_WOE		1	-0,0056	0,0008	43,758	<,0001

Tabulka 29: Odhady regresních koeficientů – model 21

Sloučení kategorií proměnné GR_MD_AGE nezpůsobilo pokles diverzifikační síly, *Giniho koeficient* je stále roven 58,3% a naopak došlo ke snížení počtu stupňů volnosti, změna tedy byla správná. Odhady koeficientů nyní pro všechny proměnné odpovídají stornovosti v jednotlivých kategoriích, mají tedy v modelu předpokládaný vliv, který jsem schopen logicky interpretovat. Model 21 budu považovat za finální model vzniklý manuálním výběrem proměnných a jejich dodatečnými úpravami. Spočtu pro něj všechny ukazatele jako pro předchozí verze.

Č.	Výběr	Typ modelu	Procedura výběru	Počet prom.	DF	Vývoj		Test	
						Gini	HL	Gini	HL
17	Užší	Komb	Stpw. (0,01)	8	17	59,03%	0%	59,43%	30%
21	Širší	Komb	Manuální	7	14	58,32%	4%	58,76%	15%

Tabulka 30: Modely 17 a 21 - přehled

Model 20 je velmi podobný jednomu ze dvou nejlepších modelů vzniklých automatickým výběrem proměnných, konkrétně kombinovanému modelu číslo 17 sestavenému procedurou *Stepwise selection* s nejpřísnější variantou kritérií pro zařazení proměnných do modelu nebo jejich vyřazení z modelu s výběrem z užší množiny proměnných. Oproti modelu 17 jsem do manuálně sestaveného modelu nezařadil proměnnou najeté kilometry (GR_CAR_KM), místo zkušeností řidiče (GR_MD_EXP) jsem použil věk řidiče (GR_MD_AGE) a proměnné GR_PREMIUM2 a GR_MD_AGE jsem nově přerozdělil do kategorií tak, aby jejich vliv v modelu odpovídal stornovostem v jednotlivých kategoriích. Model 20 má oproti modelu mírně slabší diverzifikační sílu, ale na druhou stranu je jednodušší, má méně stupňů volnosti a je lépe interpretovatelný.

4.5 Finální model a jeho vlastnosti

Jako finální model tedy vyberu model číslo 20, který považuji za dobrý kompromis mezi jednoduchostí, interpretovatelností, stabilitou, mírou těsnosti a diverzifikační silou. Následující tabulka zobrazuje přehled proměnných finálního modelu a jejich rozdělení do kategorií.

Proměnná	Kategorie	Celkem	Storen	Stornovost
gr_premium2_new	< 2 500	7 134	140	2,0 %
	< 3 500	7 726	298	3,9 %
	< 6 000	14 703	878	6,0 %
	>= 6 000	15 818	1 230	7,8 %
gr_bm_tpl2	<= 5 %	6 179	869	14,1 %
	<= 15 %	10 326	814	7,9 %
	<= 25 %	6 159	317	5,1 %
	<= 40 %	8 953	343	3,8 %
	> 40 %	13 764	203	1,5 %
gr_car_age	<= 3	4 819	70	1,5 %
	<= 8	12 271	390	3,2 %
	<= 13	16 325	927	5,7 %
	> 13	11 966	1 159	9,7 %
rnw	1	23 253	1 744	7,5 %
	2	14 071	576	4,1 %
	3	6 717	195	2,9 %
	4	1 340	31	2,3 %
gr_md_age_new	< 55	34 697	2 299	6,6 %
	>= 55	10 684	247	2,3 %

gr_premium_chng	< - 50	12 753	415	3,3 %
	< 50	8 190	575	7,0 %
	< 1 000	18 230	899	4,9 %
	>= 1 000	6 208	657	10,6 %
gr_product	žádný, nebo MINI Casco	38 600	2 414	6,3 %
	Ostatní	6 781	132	1,9 %

Tabulka 31: Přehled proměnných finálního modelu

Jak je vidět z tabulky 32, další zjednodušování modelu by již mělo velký vliv na *Giniho koeficient*.

Odstraněná proměnná	Změna Gini
gr_bm_tpl2	-3,7 %
gr_car age	-6,2 %
gr_premium2_new	-2,7 %
RNW_WOE	-1,9 %
gr_md_age_new	-0,8 %
gr_premium_chng_WOE	-0,6 %
gr_product_WOE	-0,6 %

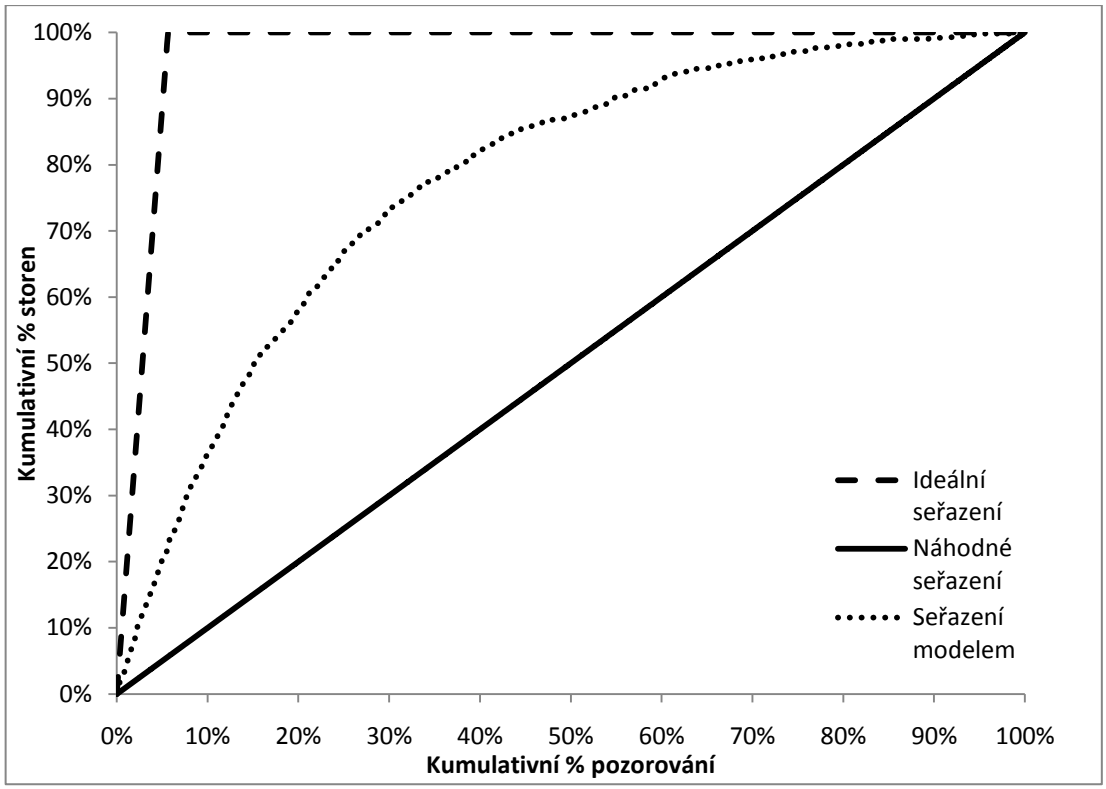
Tabulka 32: Změna diverzifikační síly modelu při odebrání jednotlivých proměnných – finální model

Pearsonovy korelační koeficienty pro *WOE* verze proměnných jsou vesměs nízké. Nejvyšší hodnoty dosahuje dvojice GR_PRODUCT a GR_CAR_AGE, konkrétně 0,45. Protože si myslím, že každá z těchto proměnných přináší jinou informaci a v předešlých analýzách jsem ukázal, že obě dvě přispívají k vysvětlení sledované proměnné, nechám je v modelu i přes zvýšenou korelaci obě dvě.

Proměnná	Č.	1	2	3	4	5	6	7
gr_premium2_new	1	1,00	0,31	-0,25	0,24	0,10	0,12	-0,32
gr_bm_tpl2	2	0,31	1,00	0,07	0,25	0,09	0,18	0,09
gr_car_age	3	-0,25	0,07	1,00	-0,01	-0,14	0,01	**0,45**
gr_md_age_new	4	0,24	0,25	-0,01	1,00	0,05	0,05	-0,01
RNW	5	0,10	0,09	-0,14	0,05	1,00	0,21	-0,09
gr_premium_chng	6	0,12	0,18	0,01	0,05	0,21	1,00	0,07
gr_product	7	-0,32	0,09	**0,45**	-0,01	-0,09	0,07	1,00

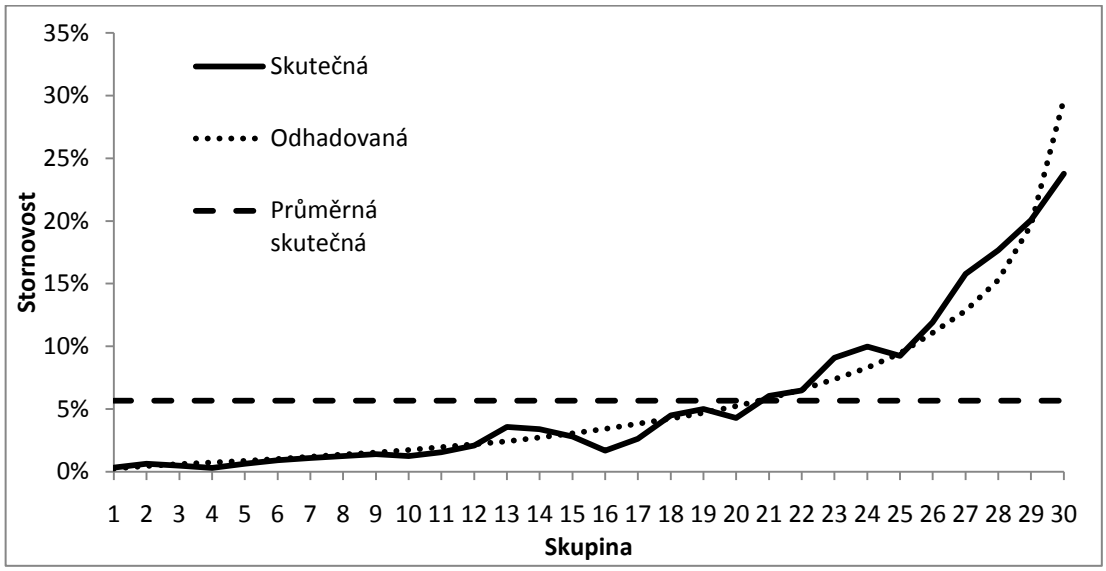
Tabulka 33: Korelační analýza finálních proměnných

Diverzifikační síla modelu na testovacím vzorku je 58,76 %. Grafické znázornění v grafu 9.



Graf 9: Znázornění diverzifikační síly modelu

Hosmer-Lemeshow test na testovacím vzorku dosahuje testové statistiky 11,7 a při p-hodnotě 0,165 nezamítá hypotézu, že pozorované a odhadnuté stornovosti v jednotlivých decilech se od sebe významně neliší. Lze tedy prohlásit, že model je dobře kalibrován. Grafické znázornění kalibrace modelu představuje následující graf.



Graf 10: Skutečná versus odhadovaná stornovost na Testovacím vzorku

Podobně jako při výpočtu testové statistiky *Hosmer_lemeshow testu* jsem pozorování testovacího vzorku seřadil podle odhadnutých pravděpodobností a poté je rozdělil do skupin. Pro vyšší citlivost jsem

tentokrát použil 30 skupin. V každé skupině jsem potom spočítal skutečnou a průměrnou odhadnutou stornovost. Do grafu jsem navíc vykreslil průměrnou stornovost celého vzorku, ke které by se blížila skutečná stornovost v jednotlivých skupinách, kdyby počáteční seřazení pozorování bylo náhodné.

5 Validace modelu

Validaci modelu provedu na vzorku 9 525 obnov smluv z období leden až březen 2012. Vzorek bude, stejně jako vývojový a testovací, obsahovat pouze smlouvy nevypovězené a smlouvy stornované z důvodu nezaplacení pojistného. Ostatní typy storen vyřadím. Stornovost validačního vzorku je 4,75 %. Pro ověření relevantnosti nejprve vzorek srovnám s nejnovějšími obnovami smluv, které mám k dispozici. Tedy s obnovami z období duben až červen 2012. Tento vzorek označím jako portfolio. Dále pak porovnám rozdělení jednotlivých proměnných, jejich diverzifikační sílu, sílu celého modelu a jeho míru těsnosti na testovacím a validačním vzorku.

Vzorek	Období	Celkový počet smluv	Počet stornovaných smluv	Stornovost
Testovací	2010 – 2011	19 402	1 098	5,66 %
Validační	leden – březen 2012	9 525	452	4,75 %
Portfolio	duben – červen 2012	10 944	---	---

Tabulka 34: Přehled datových vzorků – Testovací, Validační, Portfolio

5.1 Reprezentativnost

Obdobně jako v části 4.2, kde jsem porovnával rozdělení hodnot vybraných ukazatelů (region, věk a pohlaví řidiče, stáří automobilu a výši pojistného), nyní porovnám rozložení hodnot proměnných vybraných do modelu.

Výše zmíněná analýza reprezentativnosti na začátku vývoje modelu sestávala z obecného porovnání vývojového vzorku a aktuálního portfolia. V žádné ze sledovaných proměnných nebyl zaznamenán výrazný posun. Nyní se zaměřím konkrétně na proměnné z modelu a budu zkoumat, zdali mají obdobné rozdělení jako při vývoji a v případě, že tomu tak nebude, tak jaký má změna v rozdělení na model vliv. Použiji rozdělení proměnných do kategorií, které je užito v modelu.

Porovnám nejprve validační vzorek se současným portfoliem, abych zjistil, jakou vypovídací schopnost validace v tomto směru má. Poté provedu obdobné srovnání testovacího vzorku, na jehož základě jsem finální model vybral, s validačním vzorkem. K porovnávání vzorků opět použiji *Population*

stability index, přičemž za významný posun v rozdělení se považují hodnoty větší než 0,1.

Proměnná	Počet kategorií	PSI
gr_premium2_new	4	0,000
gr_bm_tpl2	5	0,009
gr_car_age	4	0,005
gr_md_age_new	2	0,000
rnw	4	0,004
gr_premium_chng	4	0,002
gr_product	2	0,000

Tabulka 35: Analýza reprezentativnosti, Validační vzorek vs. Portfolio

Proměnná	Počet kategorií	PSI
gr_premium2_new	4	0,007
gr_bm_tpl2	5	0,144
gr_car_age	4	0,014
gr_md_age_new	2	0,003
rnw	4	0,133
gr_premium_chng	4	0,043
gr_product	2	0,008

Tabulka 36: Analýza reprezentativnosti, Testovací vs. Validační vzorek

PSI mezi validačním vzorkem a portfoliem nepřesáhne pro žádnou proměnnou hodnotu 0,1 (viz Tabulka 35), validační vzorek tedy svým rozdělením odpovídá portfoliu. V porovnání testovacího a validačního vzorku v tabulce 36 vidíme dvě proměnné s hodnotou *PSI* nad 0,1. Jsou to bonus/malus na nové smlouvě (GR_BM_TPL2) a pořadí obnovy (RNW). Na jejich rozdělení se podívám detailněji.

Bonus	Testovací vzorek	Procent. Podíl	Validační vzorek	Procent. Podíl
<= 5 %	2 638	13,6 %	484	5,1 %
<= 15 %	4 491	23,1 %	1 605	16,9 %
<= 25 %	2 585	13,3 %	1 519	15,9 %
<= 40 %	3 903	20,1 %	2 022	21,2 %
> 40 %	5 785	29,8 %	3 895	40,9 %

Tabulka 37: Počty smluv a procentuální zastoupení v jednotlivých kategoriích proměnné GR_BM_TPL2

Obnova	Testovací vzorek	Procent. Podíl	Validační vzorek	Procent. Podíl
1	9 873	50,9 %	3 855	40,5 %
2	6 074	31,3 %	3 049	32,0 %

3	2 884	14,9 %	1 570	16,5 %
4	571	2,9 %	1 051	11,0 %

Tabulka 38: Počty smluv a procentuální zastoupení v jednotlivých kategoriích proměnné RNW

V tabulce 37 můžeme pozorovat přesun klientů do kategorií s vyšším bonusem. Konkrétně úbytek ve dvou nejnižších a nárůst ve třech zbylých kategoriích. Celkově tedy lze říci, že průměrná hodnota bonusu roste. To je částečně způsobeno změnami v rozdělení proměnné RNW (Tabulka 38), kde se snížilo zastoupení smluv s první obnovou a narostlo zastoupení smluv se čtvrtou obnovou. Věrní klienti, kteří jsou u pojišťovny déle, mají lepší bonus než ti nově příchozí. Tyto změny ovšem nejsou dostatečně výrazné, aby bezezbytku vysvětlily nárůst bonusu. Pravděpodobně tedy ještě sehrává roli silná konkurence na trhu a s ní spojené zvyšování bonusů klientům.

Pokud by podobný trend pokračoval i nadále, nejnižší kategorie s bonusem do 5 % by mohla úplně ztratit význam. S tím jak by se klienti přesouvali do vyšších kategorií a rozdělení této proměnné by bylo čím dál tím více nerovnoměrné, klesala by její síla. U proměnné RNW by při zachování trendu mělo docházet k opačnému efektu. Rozdělení by se stávalo rovnoměrnějším a s tím by bylo možné očekávat nárůst diverzifikační síly této proměnné. To by mělo být vidět již v porovnání testovacího a validačního vzorku. Ani u jedné proměnné ale změny zatím nejsou natolik velké, aby došlo k výraznějšímu ovlivnění modelu.

5.2 Diverzifikační síla

V této části porovnám pomocí *Giniho koeficientu* diverzifikační sílu jednotlivých proměnných, modelu jako celku a také sílu modelu pro některé vybrané skupiny klientů.

Proměnná	Test		Validace		Změna
	Gini	SE	Gini	SE	
gr_premium2_new	20,96 %	1,85 %	21,17 %	2,88 %	0,21 %
gr_bm_tpl2	40,64 %	1,80 %	38,10 %	2,82 %	-2,54 %
gr_car_age	28,51 %	1,85 %	29,11 %	2,87 %	0,60 %
gr_md_age_new	15,20 %	1,85 %	14,43 %	2,86 %	-0,77 %
rnw	20,66 %	1,85 %	26,46 %	2,87 %	5,80 %
gr_premium_chng	21,07 %	1,85 %	29,30 %	2,86 %	8,23 %
gr_product	11,11 %	1,84 %	13,15 %	2,86 %	2,04 %

Tabulka 39: Diverzifikační síla proměnných, Testovací vs Validační vzorek

Tabulka 39 uvádí hodnoty *Giniho koeficientu* pro jednotlivé proměnné jak na testovacím tak na validačním vzorku. Změny u proměnných BM_TPL2 a RNW odpovídají zjištěním učiněným v analýze

reprezentativnosti. Výrazný nárůst u proměnné PREMIUM_CHNG si vyžádá dodatečnou analýzu.

Kategorie	Testovací vzorek			Validační vzorek		
	N	Procent. podíl	Stornovost	N	Procent. podíl	Stornovost
<-50	5 551	28,6 %	3,3 %	2 174	22,8 %	1,9 %
<50	3 501	18,0 %	7,2 %	1 405	14,8 %	4,5 %
<1000	7 770	40,0 %	5,3 %	4 124	43,3 %	4,0 %
>=1000	2 580	13,3 %	9,9 %	1 822	19,1 %	10,0 %

Tabulka 40: PREMIUM_CHNG, Testovací vs. Validační vzorek

V tabulce 40 si lze všimnout, že stornovost v nejrizikovější kategorii klientů, u kterých pojistné na nové smlouvě vzrostlo o 1000 Kč a více, zůstala na hranici 10 %, zatímco ve zbylých kategoriích stornovost výrazně klesla. Tato kategorie se tedy od ostatních ve validačním vzorku odlišuje více než ve vzorku testovacím. Zároveň došlo k nárůstu zastoupení této poslední kategorie ze 13 % na 19 %. Obě dvě tyto skutečnosti dohromady významně posilují danou kategorii a tím i celou proměnnou. Opačně pak působí změna stornovosti druhé kategorie, která se výrazně přiblížila ke kategorii třetí a tím schopnost diverzifikace klientů oslabila. Silně ovšem převládl pozitivní efekt poslední nejrizikovější kategorie.

Celková diverzifikační síla modelu zůstala na necelých 59 %.

Vzorek	N	Stornovost	Gini	SE
Testovací	19 402	5,7 %	58,74 %	1,1 %
Validační	9 525	4,7 %	58,88 %	1,7 %

Tabulka 41: Diverzifikační síla modelu, Testovací vs. Validační vzorek

Na závěr této části otestuji diverzifikační sílu modelu na vybraných skupinách klientů se zvýšeným rizikem. Jde o mladé klienty do 30 let, nové klienty, kterým se bude smlouva obnovovat poprvé a klienty, kteří mají pouze povinné ručení a žádný dodatečný produkt (v tabulce 42 označeni jako „MTPL“). Pro žádnou z těchto skupin nedošlo k výrazné změně.

	Testovací vzorek			Validační vzorek			
Skupina	N	Gini	SE	N	Gini	SE	Změna
Do 30 let	2 685	48,74%	2,74%	1 048	47,82%	5,15%	-0,9%
Noví klienti	9 873	55,55%	1,41%	3 855	56,09%	2,27%	0,5%
MTPL	15 746	56,81%	1,18%	7 346	55,74%	1,89%	-1,1%

Tabulka 42: Diverzifikační síla modelu na vybraných skupinách klientů, Testovací vs. Validační vzorek

5.3 Kalibrace

Dalším pohledem na model, který nás bude zajímat je to, jak dobře odhaduje pravděpodobnost storna v různých částech spektra této sledované veličiny. Kvantitativně to ověřím pomocí testu míry těsnosti modelu, *Hosmer-Lemeshow testu.* Protože model neobsahuje žádnou spojitou proměnnou, existuje pouze omezený počet různých kombinací hodnot nezávislých proměnných a tím pádem i odhadů hodnot závislé proměnné. Konkrétně je takových kombinací 5 120. V testovacím vzorku, který má 19 402 pozorování, je různých kombinací vysvětlujících proměnných 1 918 a ve validačním vzorku s 9 525 pozorováními pak 1 529. Odhadnuté hodnoty vysvětlované proměnné se tedy často opakují, a proto budu muset, jak bylo řečeno v 2.11.2, věnovat zvýšenou pozornost jejich rozdělení do skupin. Skupin použiji 10, jak je obvyklé pro tento test a pro rozdělení pozorování na základě odhadnuté pravděpodobnosti storna do decilů použiji proceduru PROC RANK. Ta přiřazuje stejné hodnoty vždy do stejné skupiny a rozdělení do skupin optimalizuje tak, aby zastoupení v jednotlivých skupinách bylo co možná nejrovnoměrnější. Nepůjde tedy úplně přesně o decily.

Skup.	N	Očekávaná storna	Skutečná storna	Očekávaná stornovost	Skutečná stornovost	Testová statistika
1	1 932	8,27	9	0,43 %	0,47 %	0,07
2	1 948	14,94	12	0,77 %	0,62 %	0,58
3	1 940	25,41	24	1,31 %	1,24 %	0,08
4	1 911	37,29	31	1,95 %	1,62 %	1,08
5	1 968	56,73	64	2,88 %	3,25 %	0,96
6	1 926	69,41	56	3,60 %	2,91 %	2,69
7	1 953	102,44	100	5,25 %	5,12 %	0,06
8	1 930	144,73	164	7,50 %	8,50 %	2,77
9	1 955	219,44	241	11,22 %	12,33 %	2,39
10	1 939	415,15	397	21,41 %	20,47 %	1,01

Tabulka 43: Hosmer-Lemeshow test, Testovací vzorek

Pro testovací vzorek jsou si počty pozorování v jednotlivých skupinách velmi podobné, opakující se hodnoty odhadů tedy nezpůsobují žádný problém. Ve všech skupinách je také počet očekávaných storen vyšší než 5, lze tedy využít předpokladu o asymptotickém rozdělení reziduí. Testová statistika *Hosmer-Lemeshow testu* těsnosti modelu dosahuje hodnoty 11,7. Pravděpodobnost, že bychom za platnosti hypotézy o správnosti odhadů, napozorovali data, která by více svědčila proti této hypotéze, neboli p-hodnota, je při osmi stupních volnosti 16,5 %. Hypotézu tedy nezamítám.

Rozdělení do skupin pro *Hosmer-Lemeshow test* na validačním vzorku zobrazuje tabulka 44. Skupiny jsou opět rovnoměrně zastoupené, ale ve

skupině s nejnižšími hodnotami odhadů mám jen 3,19 očekávaných storen. Při interpretaci výsledků testu je tedy nutné vzít v potaz, že aproximace rozdělení testové statistiky pomocí Chí-kvadrát rozdělení nemusí v tomto případě být tak přesná.

Skup.	N	Očekávaná storna	Skutečná storna	Očekávaná stornovost	Skutečná stornovost	Testová statistika
1	949	3,19	5	0,34 %	0,53 %	1,03
2	955	6,62	6	0,69 %	0,63 %	0,06
3	942	10,10	10	1,07 %	1,06 %	0,00
4	964	14,99	9	1,56 %	0,93 %	2,43
5	950	20,74	17	2,18 %	1,79 %	0,69
6	966	30,26	31	3,13 %	3,21 %	0,02
7	949	40,60	41	4,28 %	4,32 %	0,00
8	946	59,75	73	6,32 %	7,72 %	3,14
9	951	94,12	94	9,90 %	9,88 %	0,00
10	953	193,47	166	20,30 %	17,42 %	4,89

Tabulka 44: Hosmer-Lemeshow test, Validační vzorek

Testová statistika *H.-L. testu* pro validační vzorek se rovná 12,26 a p-hodnota testu je při osmi stupních volnosti 15 %. Pokud bych chtěl dosáhnout splnění předpokladu o minimálních hodnotách očekávaných storen ve všech skupinách, mohl bych za cenu snížení citlivosti testu sloučit skupiny 1 a 2. Poté bych již měl ve všech skupinách více než 5 očekávaných storen. Testová statistika by pak měla hodnotu 11,32 a p-hodnota testu by při 7 stupních volnosti byla rovna 13 %. Ani v jednom případě tedy hypotézu o správnosti modelu nezamítám.

Z hlediska celého vzorku je třeba zmínit, že pokles pozorované stornovosti o 0,91 procentního bodu je pravděpodobně převážně způsoben změnami ve složení vzorku, které jsme pozorovali v reprezentativní analýze v části 5.1. Přesun klientů do nejméně rizikové skupiny s nejvyšším bonusem i vyšší zastoupení klientů se čtvrtou nebo vyšší obnovou smlouvy snižují riziko stornovosti vzorku, což dokazuje pokles odhadované stornovosti. Rozdíl mezi pozorovanou a odhadovanou stornovostí u validačního vzorku je ovšem způsoben ještě nějakou další změnou, kterou model nepostihuje. Může jít o nepřesnost modelu, o změnu složení klientů, která se ovšem nepromítá skrze proměnné v modelu, nebo o obecný trend mírného poklesu stornovosti. Blíže se na tuto skutečnost zaměřím v kapitole 6.

Vzorek	Pozorovaná stornovost	Odhadovaná stornovost	Rozdíl
Testovací	5,66 %	5,64 %	- 0,02 %
Validační	4,75 %	4,97 %	0,22 %

Tabulka 45: Pozorované a odhadované stornovosti, Testovací vs. Validační vzorek

5.4 Zhodnocení modelu

Mezi obdobími leden 2010 – prosinec 2011 a leden – březen 2012 došlo k mírným změnám ve složení klientů. Narostlo zastoupení klientů s vyšším bonusem a také klientů, kteří jsou u pojišťovny již delší dobu. Obě dvě tyto změny měly vliv také na diverzifikační sílu příslušných proměnných. Zatímco rozdělení proměnné výše bonusu na nové smlouvě, BM_TPL2, se vychýlilo směrem k nejvyšší kategorii a síla proměnné tak klesla, jednotlivé kategorie proměnné pořadí obnovy, RNW, jsou nyní zastoupeny rovnoměrněji a tato proměnná tak dokáže lépe třídit klienty. Další výraznější změnu diverzifikační síly jsem pak již zaznamenal pouze u proměnné změna výše pojistného, PREMIUM_CHNG. V tomto případě došlo k nárůstu rozdílu mezi stornovostí nejrizikovější kategorie a kategorií ostatních, což znamená, že přítomnost klienta v této skupině je silnějším příznakem jeho rizikovosti a proměnná je tak celkově silnější. Diverzifikační síla ostatních proměnných, modelu na celém vzorku i na vybraných skupinách zůstala stabilní.

U dalšího parametru modelu, jeho míry těsnosti, došlo k mírnému zhoršení. Na testovacím vzorku byla p-hodnota testu hypotézy o správnosti modelu ve smyslu rovnosti odhadnutých pravděpodobností s hodnotami podmíněné střední hodnoty pozorované veličiny 16,5 %. Pro validační vzorek stejný test dosahuje p-hodnoty 15 %, respektive 13 % při rozdělení do 9 skupin. V obou případech však hypotézu i nadále nezamítáme. Dále jsem pozoroval, že mezi testovacím a validačním vzorkem došlo ke snížení stornovosti o téměř celý jeden procentní bod. Z velké části tento pokles model postihnul, ale přesto na validačním vzorku model pravděpodobnosti storna lehce nadsazuje. V průměru o 0,22 % procentního bodu.

Celkově lze říci, že model je stabilní v čase. Pravděpodobnosti storna dobře odhaduje v celém spektru této veličiny a díky tomu dobře třídí rizikové klienty. Možným rizikem je slábnutí proměnné BM_TPL2, pokud by i nadále rostlo zastoupení klientů s nejvyšším bonusem.

6 Kalibrace

V části 5.3 jsem konstatoval, že mezi testovacím a validačním vzorkem došlo ke změně průměrné stornovosti. Tato změna byla z větší části způsobena změnami ve složení vzorků, ale částečně zůstala nevysvětlena. Konkrétně se jedná o rozdíl mezi průměrnou pozorovanou stornovostí na validačním vzorku, která činí 4,75 % a průměrnou odhadovanou stornovostí na stejném vzorku, která se rovná 4,97 %. Pokud by šlo o dlouhodobější trend, mohl by tento nesoulad v průběhu užívání modelu postupně narůstat. Analyzuji tedy vývoj stornovosti a poté se případně pokusím odhadnuté hodnoty upravit tak, aby lépe vystihovaly skutečnou pravděpodobnost, že na obnově dojde ke stornu pro nezaplacení pojistného.

K analýze využiji faktu, že vývojový a testovací vzorek, které oba pochází ze stejného období a pouze byly náhodně rozděleny na dvě části, obsahují obnovy smluv za celé dva roky 2010 a 2011. Společně s validačním vzorkem mám tedy k dispozici poměrně rozsáhlé časové období, ve kterém mohu zkoumat vývoj stornovosti. Toto období rozdělím na čtvrtletí, pro každé čtvrtletí spočítám průměrnou pozorovanou stornovost a získanou časovou řadu extrapoluji pomocí několika základních trendových funkcí pro následující čtyři čtvrtletí.

Pro extrapolaci použiji lineární trendovou funkci danou předpisem

$$y = \alpha x + \beta ,$$

mocninnou trendovou funkci danou předpisem

$$y = \alpha x^{\beta} ,$$

exponenciální trendovou funkci danou předpisem

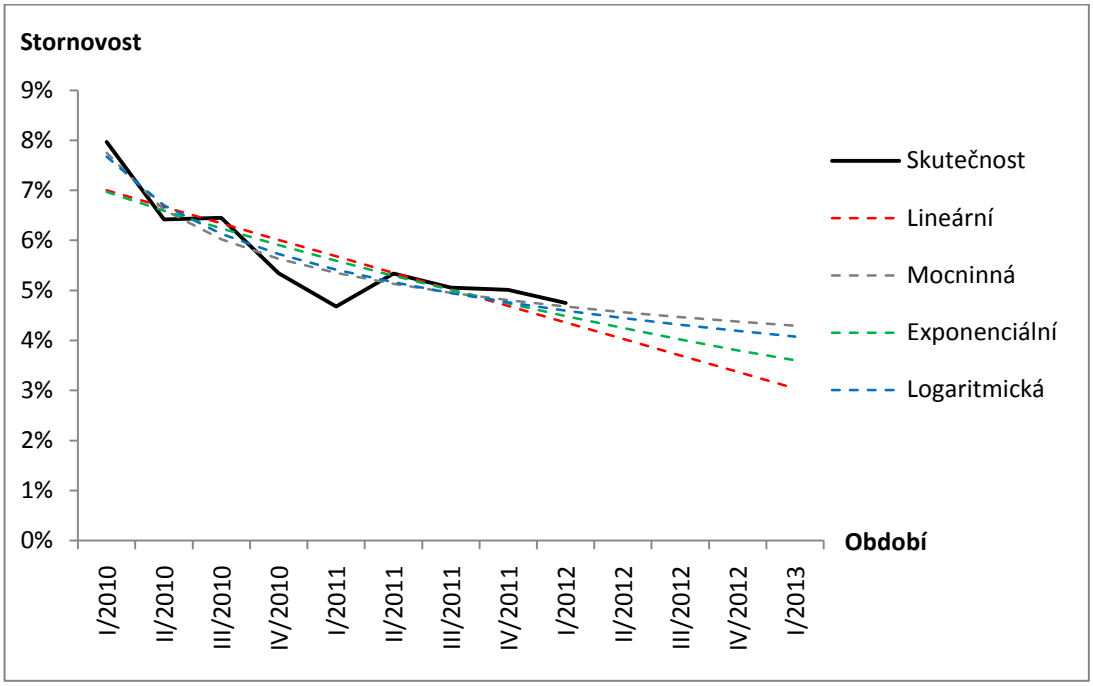
$$y = \alpha \exp(\beta x)$$

a logaritmickou trendovou funkci danou předpisem

$$y = \alpha \ln(x) + \beta .$$

Ve všech případech použiji k odhadu parametrů metodu nejmenších čtverců. V grafu 11 je znázorněna skutečná stornovost i všechny 4 trendové křivky. V tabulce 46 pak odhadnuté stornovosti pro příští rok vzniklé jako aritmetické průměry odhadů pro následující 4 čtvrtletí. Je patrné, že stornovost v posledních 9 čtvrtletích postupně klesala. I s přihlédnutím k faktu, že v posledních čtvrtletích již je pokles mírnější, zvolím

nejkonzervativnější odhad, tedy ten, který vzešel z extrapolace pomocí mocninné trendové funkce.



Graf 11: Vývoj stornovosti

Trendová funkce:	lineární	mocninná	exponenciální	logaritmická
Stornovost:	3,54 %	4,43 %	3,92 %	4,26 %

Tabulka 46: Odhady průměrné stornovosti pro příští rok pomocí jednotlivých trendových funkcí

Než samotný vývoj stornovosti mě ovšem z hlediska dodatečné úpravy hodnot vystupujících z modelu více zajímá vývoj rozdílu mezi skutečnou a odhadovanou stornovostí. Neboli vývoj stornovosti, který model nezachycuje. Na grafu 12 je opět skutečná stornovost extrapolovaná mocninnou trendovou funkcí o 4 období dopředu a současně s ní také stejnou metodou extrapolované hodnoty odhadované stornovosti. Z tohoto grafu lze vyčíst, že pro nejstarší období model pravděpodobnosti storna mírně podhodnocuje, s postupem času se tento trend pozvolna obrací a u extrapolovaných hodnot pro nejbližší čtyři čtvrtletí již je patrné, že model pravděpodobnosti storna nadhodnocuje.

Abych model lépe připravil pro budoucí použití, pokusím se tomuto nadhodnocování předejít transformací hodnot, které z modelu vystupují. Nejprve spočítám průměrný rozdíl mezi extrapolovanými skutečnými a odhadnutými stornovostmi v příštích čtyřech obdobích.

$$Posun.stornovosti = \frac{1}{4}\sum_{i=1}^{4} ext.odhad_i - ext.skutečnost_i = 0{,}0014\,,$$

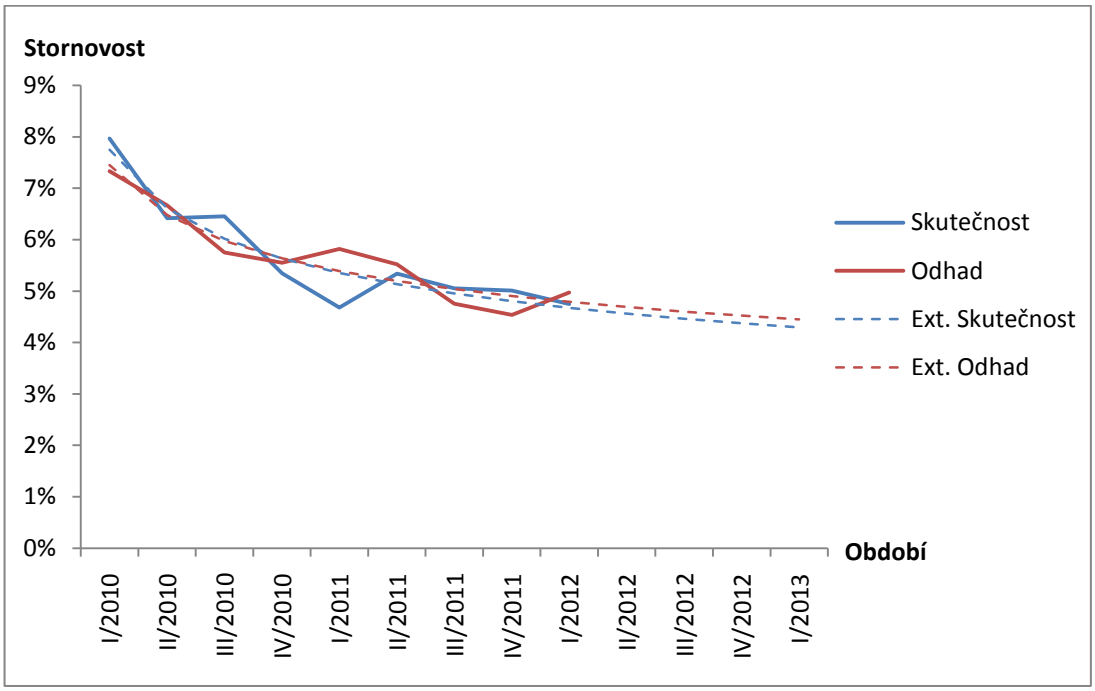
kde i označuje pořadí extrapolovaných období. Pro výpočet transformační funkce poté využiji průměrný odhad stornovosti na validačním vzorku, $\bar{\pi}$, a jako cílovou hodnotu tento průměrný odhad snížený o *Posun.stornovosti*, který označím $\bar{\pi}_T$. Jako kalibrační funkci použiji transformaci odvozenou v sekci 2.12.

Transformační funkce (2.10) bude mít tvar:

$$\pi_{j_C} = \frac{\pi_j(1-\bar{\pi})\bar{\pi}_T}{(1-\pi_j)\bar{\pi}(1-\bar{\pi}_T)+\pi_j(1-\bar{\pi})\bar{\pi}_T} =$$

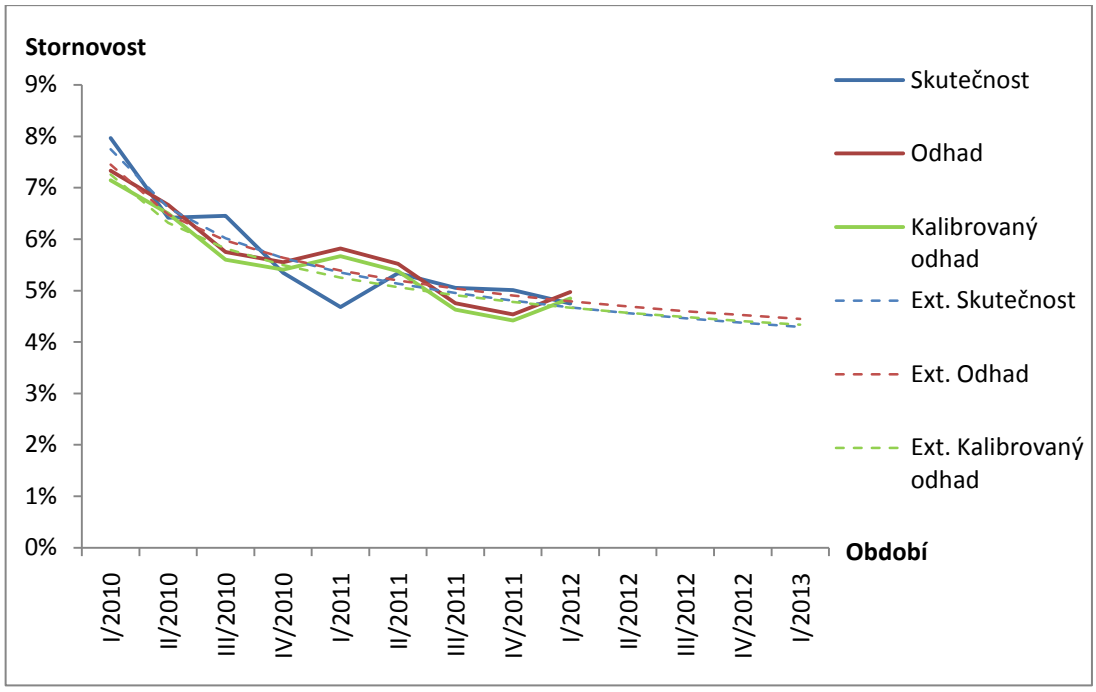
$$= \frac{\pi_j(1-0{,}049747)0{,}048346}{(1-\pi_j)0{,}049747(1-0{,}048346)+\pi_j(1-0{,}049747)0{,}048346}$$

$$= \frac{0{,}045941\pi_j}{0{,}47342-0{,}0014\pi_j} = \frac{-1108{,}075875}{\pi_j-33{,}791528} - 32{,}791528$$



Graf 12: Porovnání vývoje skutečné a odhadované stornovosti

Finálním výstupem tedy je hodnota odhadnutá model transformovaná výše uvedeným vztahem. Na diverzifikační sílu modelu tato úprava nemá žádný vliv. *Hosmer-Lemeshow test* na validačním vzorku pro kalibrované hodnoty dosahuje p-hodnoty 17 %, což je mírné zlepšení. Pokud přiřadím finální kalibrované hodnoty ke všem pozorováním vývojového, testovacího i validačního vzorku, mohu provést srovnání se skutečnými stornovostmi a s původními odhady. Toto porovnání obsahuje graf 13. Můžeme pozorovat, že kalibrované odhady jsou lehce pod skutečnými hodnotami a že extrapolované hodnoty kalibrovaných odhadů téměř kopírují extrapolované hodnoty skutečné stornovosti, což bylo cílem kalibrace.



Graf 13: Porovnání vývoje skutečné, odhadované stornovosti a kalibrované odhadované stornovosti

Závěr

Cílem této práce bylo analyzovat storno pojištění odpovědnosti z provozu motorového vozidla během obnovy smlouvy a pomocí logistické regrese vyvinout model, který by pojišťovně umožnil s předstihem identifikovat ohrožené smlouvy. Pro další postup bylo nutné nejprve popsat, z jakých důvodů může ze zákona ke stornu smlouvy dojít, a které z nich v praxi nejčastěji nastávají. Z jednotlivých typů storen jsem pak vybral ty, které je možné modelovat a u nichž to má pro pojišťovnu smysl. Postupně jsem se tedy omezil pouze na storno pro nezaplacení pojistného.

V druhé kapitole jsem připravil teoretický základ pro praktickou část práce. Jednalo se o pomocné ukazatele pro analýzu proměnných (*Weight of evidence, Giniho statistika, Informační hodnota, Population stability index*), logistickou regresi, testy významnosti jednotlivých regresních koeficientů i celých modelů, míry těsnosti modelu a kalibraci.

Poté jsem na vzorku skutečných smluv s obnovami v letech 2010 a 2011 provedl podrobnou analýzu proměnných, na jejímž základě jsem sestavil širší a užší množinu kandidátů na vstup do modelu. Pomocí metod *Stepwise selection* a *Best subsets* s různě nastavenými parametry a obou množin kandidátů jsem sestavil škálu modelů, které jsem porovnal z hlediska jejich složitosti, diverzifikační síly a míry těsnosti. Na základě takto získaných poznatků jsem manuálně vystavěl model, u kterého jsem vliv jednotlivých proměnných a jejich příspěvek k diverzifikační síle modelu zkoumal detailněji. Takto vzniklý finální model se nejvíce podobá modelu sestavenému metodou *Stepwise selection* s nejpřísnější použitou variantou nastavení parametrů pro vstup a výstup z modelu s výběrem z užší množiny kandidátů. Finální model sestává z proměnných: výše pojistného na nové smlouvě, výše bonusu (případně malusu) na nové smlouvě, změna výše pojistného oproti původní smlouvě, stáří vozidla, věk hlavního řidiče vozidla, existence dalších produktů mimo povinné ručení a pořadí obnovy (po kolikáté již se klientovi smlouva obnovuje).

Následně jsem provedl validaci modelu na obnovách smluv z období leden až březen 2012, tedy na jiném vzorku smluv z jiného období, než na kterém byl model vyvinut. Model se celkově ukázal jako stabilní s diverzifikační silou vyjádřenou pomocí *Giniho statistiky* necelých 59 %. Mírné změny v síle jednotlivých proměnných se mi podařilo vysvětlit pomocí rozdílů ve složení vzorků. Pokud by ovšem tyto změny pokračovaly

ve stejném trendu, hrozilo by, že se proměnná bonus na nové smlouvě stane nevýznamnou. Stejně jako na testovacím, ani na tomto vzorku jsem pomocí *Hosmer-Lemeshow testu* nezamítnul hypotézu o shodě odhadů se skutečnými hodnotami.

Na závěr jsem provedl kalibraci modelu, jejímž cílem byl posun průměrné odhadované pravděpodobnosti storna směrem k očekávané průměrné stornovosti v období duben 2012 až březen 2013. Tak, aby výstupy z modelu i v budoucnosti sloužily nejen k porovnání klientů mezi sebou, ale aby vypovídaly o skutečné pravděpodobnosti, zda klient svou smlouvu během obnovy stornuje.

Ve své práci jsem popsal a prakticky předvedl metody pro analýzu dat, sestavení a výběr modelu logistické regrese a jeho validaci a kalibraci. Na skutečných datech jsem pak vyvinul model, který dokáže dobře identifikovat smlouvy se zvýšenou pravděpodobností storna pro nezaplacení pojistného po obnově. A který má předpoklady pro to, aby tyto pravděpodobnosti správně odhadoval i v budoucnu. Stanovené cíle se mi tedy podařilo naplnit.

Literatura

[1] Anděl J., *Statistické metody*, MATFYZPRESS, 2003. ISBN 80-85863-27-8

[2] Cook R. D., Weisberg S.: *Applied Regression Including Computing and Graphics*, John Wiley & Sons, Inc., 1999. ISBN 978-0-471-31711-1

[3] Česká kancelář pojistitelů [online], http://www.ckp.cz [cit. 28. 7. 2013]

[4] Česká asociace pojišťoven [online], http://www.cap.cz [cit 28. 7. 2013]

[5] Furnival G. M., Wilson R. W.: Regression by Leaps and Bounds, *Technometrics*, 16, 1974

[6] Hosmer D. W., Lemeshow S.: *Applied Logistic Regression* Second Edition, John Wiley & Sons, Inc., 2000. ISBN 0-471-35632-8

[7] Homser D. W., Jovanovic B., Lemeshow S.: Best Subsets Logistic Regression, *Biometrics*, 45, 1989

[8] Jedlička P., *ČKP a škody na zdraví v povinném ručení*, SAV 2. 12. 2011 http://www.actuaria.cz/upload/CKP%20prezentace%20%C5%A1z%20SAV%20web.ppt.

[9] Mallows, C. L.: Some comments on Cp, *Technometrics*, 15, 1973

[10] MDČR, Statistiky vyplývající z centrálního registru vozidel [online], www.mdcr.cz/cs/Silnicni_doprava/Dovoz_registrace_a_schvalovani_vozidel/ [cit 28. 7. 2013]

[11] Pruit R.: *The Applied Use of Population Stability Index (PSI) in SAS Enterprise Miner*, Premier Bankcard, LLC, Sioux Falls, SD, 2010, http://support.sas.com/resources/papers/proceedings10/288-2010.pdf.

[12] Řezáč M., Řezáč F.: *Measuring the Quality of Credit Scoring Models*, 2009, http://www.crc.man.ed.ac.uk/conference/archive/2009/presentations/Paper-65-Presentation.pdf.

[13] Sas Institute Inc., *SAS/STAT® 9.22 User´s Guide*, Cary, NC: Sas Institue Inc., 2000, http://support.sas.com/documentation/cdl/en/statug/63347/PDF/default/statug.pdf.

[14] WolframMathWorld, *Newton's Method* [online], http://mathworld.wolfram.com/NewtonsMethod.html [cit 28.7.2013]

[15] Zákon č. 168/1999 Sb. o pojištění odpovědnosti z provozu vozidla, http://portal.gov.cz/app/zakony/download?idBiblio=47910&nr=168~2F1999~20Sb.&ft=pdf.

[16] Zákon č. 37/2004 Sb. o pojistné smlouvě, http://portal.gov.cz/app/zakony/download?idBiblio=57259&nr=37~2F2004~20Sb.&ft=pdf.

[17] Zvára K., *Regrese*, MATFYZPRESS, 2008. ISBN 987-80-7378-041-8