VYSOKÁ ŠKOLA EKONOMICKÁ V PRAZE
FAKULTA INFORMATIKY A STATISTIKY

BAKALÁŘSKÁ PRÁCE

2018 Mariya Oleynik

VYSOKÁ ŠKOLA EKONOMICKÁ V PRAZE
FAKULTA INFORMATIKY A STATISTIKY

Název bakalářské práce:

Attrition analýza pomocí metod strojového učení

Autor: **Mariya Oleynik**
Katedra: **Katedra ekonometrie**
Obor: **Matematické metody v ekonomii**
Vedoucí práce: **Ing. Lukáš Frýd**

Prohlášení:
Prohlašuji, že jsem bakalářskou práci na téma „Attrition analýza pomocí metod strojového učení“ zpracovala samostatně. Veškerou použitou literaturu a další podkladové materiály uvádím v seznamu použité literatury.

V Praze dne 6. května 2018 .....

Mariya Oleynik

Poděkování:
Ráda bych na tomto místě poděkovala Ing. Lukáši Frýdovi za vedení mé bakalářské práce, za pomoc při zpracování a za podnětné návrhy, které ji obohatily. Také bych chtěla poděkovat Mgr. Petru Weissovi a oddělení Risk Monitoring za vstřícný přístup a za poskytnutí dat. Dále bych poděkovala své rodině za podporu v průběhu celého studia.

Abstrakt

Název práce: Attrition analýza pomocí metod strojového učení
Autor: Mariya Oleynik
Katedra: Katedra ekonometrie
Vedoucí práce: Ing. Lukáš Frýd

Customer attrition se zabývá analýzou chování odcházejících klientů. Cílem bakalářské práce je analýza a predikce spotřebitelského chování klientů u jejich běžných bankovních účtů pomocí standartních ekonometrických nástrojů a dále pomocí metod strojového učení. Jako standartní ekonometrický model byl využit model logistické regrese. Ze skupiny metod strojového učení byl zvolen zobecněný aditivní model (GAM). GAM byl schopen zachytit nelinearitu v proměnných, která byla odvozená pomocí testu ANOVA pro neparametrické efekty a pomocí grafického zobrazení závislosti odchodu klienta na jednotlivých proměnných. Ve výsledku bylo prokázáno, že GAM poskytuje lepší predikční schopnost odchodu klienta než model logistické regrese.

Klíčová slova: logistický model, zobecněný aditivní model, transakce, ztráta klienta, predikce

Abstract

Title: Attrition analysis using machine learning methods
Author: Mariya Oleynik
Department: Department of Econometrics
Supervisor: Ing. Lukáš Frýd

Customer attrition deals with analysis of the loss of clients. The aim of the bachelor thesis is to analyse and predict consumer behaviour of clients in their current bank accounts using standard econometric tools and also using machine learning methods. The model of logistic regression was used as a standard econometric model. A generalized additive model (GAM) was chosen from the group of machine learning methods. GAM was able to reveal nonlinearity in variables using the ANOVA test for nonparametric effects and graphical depiction of customer attrition dependence on individual variables. As a result, GAM provides better prediction performance of customer attrition than the logistic regression model.

Keywords: logistic model, generalized additive model, transaction, loss of client, prediction

Obsah

Úvod

Customer attrition je přirozeným pohybem klientů na trhu služeb. S rostoucí konkurencí a širší nabídkou produktů se firmy snaží nejenom přilákat nového zákazníka, ale také udržet stávajícího klienta. Ztráta zákazníka může mít negativní dopad na tržní podíl a zisk banky. Různé výzkumy poukazují na to, že pokud zákazník změní poskytovatele služeb, ztrácí se potenciál dodatečných zisků a banka čelí dodatečným nákladům na získání náhradního zákazníka. Právě proto je jedním z hlavních bodů strategie banky individuální předpověď pravděpodobnosti odchodu klienta v budoucnosti.

Táto bakalářská práce popisuje způsoby predikce odchodu stávajícího klienta bankovního domu. V této práci bude aplikován logistický model a zobecněný aditivní logistický model, které následně budou mezi sebou porovnané. Cílem práce je zjistit, který z uvedených modelů je schopen lepším způsobem analyzovat a predikovat odchod stávajícího klienta.

V první kapitole bude uvedená příslušná literatura a její hlavní myšlenky spojené se způsoby predikce ztráty klientů v různých oblastech podnikaní.

Druhá kapitola se zabývá teoretickým základem logistického modelu. V této kapitole bude uvedena metoda pro odhady parametrů, způsoby jejich interpretace a také způsoby hodnocení modelu.

Následující kapitola obsahuje úvod do zobecněných aditivních modelů. Dále v ní budou popsané postupy řešení zobecněného aditivního modelu a logistického aditivního modelu.

Čtvrtá kapitola se bude věnovat popisu sběru dat a jejich zpracovaní. V této kapitole bude datově definován také odchod klienta.

Následující pátá kapitola se zaměřuje na attrition analýzu pomocí logistického modelu. Součásti kapitoly je vysvětlení vlivů jednotlivých faktorů na pravděpodobnost odchodu klienta.

Poslední šestá kapitola se zabývá aplikací zobecněného aditivního logistického modelu. Také zde budou popsané vztahy závislostí pravděpodobnosti odchodu klienta na jednotlivých vysvětlujících proměnných.

1 Přehled literatury

K. Coussement, D.F. Benoit a D. Van Den Poel se ve svém článku „Improved marketing decision making in a customer churn prediction context using generalized additive models“ (2010) zabývají analýzou spotřebitelského chování klientů. Ve své práci se autoři zaměřují na predikci prodloužení předplatného časopisů největší Belgické vydavatelské společnosti. Daná analýza byla provedená pomocí logistické regrese a zobecněného aditivního modelu. Pro hodnocení modelů byly použité AUC a Top-decile lift. Na základě dosažených výsledků autoři potvrzují, že GAM zvyšuje predikční schopnost modelů kvůli zachycení nelineárních vztahů mezi proměnnými a zároveň poskytuje lepší představu o účincích vysvětlujících proměnných.

Analýza odchodu klienta byla také provedená v práci W. Buckinx a D. Van Den Poel „Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting“ (2005). Práce se zabývá predikcí odcházejícího klienta od obchodníka rychloobrátkového spotřebního zboží. Pro analýzu autoři použili logistickou regresi, metodu ARD, která patří do metod neuronových sítí, a náhodný les. Hodnocení modelů se provádí pomocí kritéria správného klasifikovaného procenta (PCC) a AUC. Z výsledku provedené práce lze odvodit, že všechny klasifikační techniky mají téměř stejnou predikční schopnost. V závěru práce autoři uvádí několik výhod náhodného lesu, kvůli čemuž má náhodný les přednost před ostatními modely.

T. Vafeiadis, K.I. Diamantaras, G. Sarigiannidis, K.Ch. Chatzisavvas se ve svém článku „A comparison of machine learning techniques for customer churn prediction“ zabývají analýzou loajality klientů v odvětví telekomunikací. Ve své práci autoři porovnávají umělé neuronové sítě, rozhodovací stromy, Support Vector Machines, klasifikátory Naïve Bayes a logistickou regresi. Zároveň na každou uvedenou metodu strojového učení byl aplikován boosting algoritmus, který zlepšuje predikční schopnost prostřednictvím kombinace rozhodnutí z mnoha klasifikačních modelů. Ve výsledku práce se pro každou použitou metodu uvádějí F - hodnota a přesnost predikce, na základě čehož autoři dávají přednost boosting Support Vector Machines.

2 Logistická regrese

Jedním z nejpoužívanějších modelů předpovědí, že klient zůstane nebo odejde ke konkurenci, je logistická regrese, kde závislou proměnnou je binární proměnná.

2.1 Logistický model

Vektorem **x** se označuje vektor vysvětlujících proměnných, a pokud je stav $y = 1$ chápan jako žádoucí varianta, potom lze podmíněnou pravděpodobnost úspěchu p zapsat následovně: [10]

$$p = Pr\{y = 1|\mathbf{x}\}. \tag{1}$$

Při modelování pravděpodobností úspěchu pomocí lineárního regresního modelu nastává problém, že pravděpodobnost nabývá hodnot mimo interval <0,1>. [8] Pro řešení tohoto problému se zavádí tzv. poměr šancí (odds ratio), který vyjadřuje poměr mezi pravděpodobností úspěchu a pravděpodobností neúspěchu a zároveň nabývá hodnot v intervalu od 0 do ∞:

$$Poměr\ šancí = \frac{p}{1-p}. \tag{2}$$

Logit je logaritmus poměru šancí, který nabývá hodnot v intervalu od $-\infty$ do ∞:

$$\text{logit}(p) = \log(poměr\ šancí) = \log\left(\frac{p}{1-p}\right). \tag{3}$$

Někdy se model logistické regrese zapisuje jako: [10]

$$\text{logit}(p) = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}, \tag{4}$$

kde $\boldsymbol{\beta}^{\mathrm{T}}$ je transponovaný vektor regresních parametrů,
$\mathbf{X}$ je matice vysvětlujících proměnných.

Zároveň rovnice (4) je inverzní funkcí ke klasické logistické funkci, která vypadá následovně:

$$\sigma(z) = \frac{e^{z}}{e^{z}+1} = \frac{1}{1+e^{-z}}, \tag{5}$$

a splňuje základní požadavky kladené na distribuční funkce: [10]

1. je neklesající,
2. $\sigma(z) \to 0$ při x $\to -\infty$, a $\sigma(z) \to 1$ při x $\to \infty$,
3. integrál přes $\mathbb{R}$ je roven 1,

kde $z = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ je tzv. identita.

Nyní lze pravděpodobnost úspěchu ze vztahu (1) přepsat jako:

$$p = Pr\{y = 1|\mathbf{x}\} = F(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}), \tag{6}$$

kde $F(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ je distribuční funkce logistického rozdělení, která vypadá následovně:

$$F(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = \frac{1}{1+e^{-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}}}. \tag{7}$$

2.2 Metoda maximální věrohodnosti

Nyní je pro řešení logistické regrese nezbytné najít odhady neznámých parametrů β_i. Pro tento účel se používá metoda, která se nazývá metoda maximální věrohodnosti. Pravděpodobnosti pro jednotlivé hodnoty binární proměnné Y_i je možné zapsat jako:

$$\Pr(Y_i = y|\boldsymbol{x}_i) = \begin{cases} p_i & \text{když } y = 1 \\ 1 - p_i & \text{když } y = 0 \end{cases} \tag{8}$$

Pokud je $p(x_i)$ pravděpodobnost nabytí závislé proměnné y_i hodnoty 1, potom věrohodnost tohoto pozorovaní je $F(\boldsymbol{b}^{\mathrm{T}}\mathbf{x}_i)$, ale pokud proměnná y_i nabývá hodnoty 0, potom se věrohodnost rovná $(1 - F(\boldsymbol{b}^{\mathrm{T}}\mathbf{x}_i))$. Za předpokladu, že pozorovaní jsou nezávislá, celkovou věrohodnost pozorovaných dat vzhledem k vektoru koeficientů $\boldsymbol{b}$ lze vyjádřit jako: [8]

$$L(b) = \prod_i F(\boldsymbol{b}^{\mathrm{T}}\mathbf{x}_i)^{y_i} \cdot (1 - F(\boldsymbol{b}^{\mathrm{T}}\mathbf{x}_i))^{(1-y_i)}, \tag{9}$$

kde $L(b)$ je věrohodnostní funkce,

$F(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}_i)$ je věrohodnost pozorovaní $y_i = 1$.

Častěji se ale používá logaritmus věrohodnostní funkce, který převádí součin na součet:

$$lnL(b) = \sum_i \left[y_i * \ln F(\boldsymbol{b}^{\mathrm{T}}\mathbf{x}_i) + (1 - y_i) \cdot \ln\left(1 - F(\boldsymbol{b}^{\mathrm{T}}\mathbf{x}_i)\right)\right]. \tag{10}$$

Nyní je potřeba najít extrémy funkcí pomocí parciální derivace log-věrohodnostní funkce podle všech parametrů $\boldsymbol{b}$:

$$\frac{\partial l}{\partial b_j} = \sum_i y_i - F(\boldsymbol{b}^{\mathrm{T}}\mathbf{x}_i)) \cdot \mathbf{x}_{i,j} = 0, \tag{11}$$

kde j=0,...,k. Vzhledem k náročností výpočtů se pro řešení podobné rovnice většinou používají statistické softwary. Nalezené body jsou odhady parametrů β_i.

2.3 Hodnocení modelu

Následujícím krokem je testovaní regresních parametrů. Je potřeba prozkoumat, zda vysvětlující proměnné mají na vysvětlovanou proměnnou statisticky významný vliv, nebo nejsou-li jednotlivé parametry nulové.

Významnost jednotlivých vysvětlujících proměnných se dá jednoduše zjistit pomocí Waldovy statistiky, která má asymptoticky normální rozdělení $N(0,1)$. To znamená, že bude nulová hypotéza zamítnutá na hladině významnosti α, pokud $|W| \geq \Phi^{-1}(1 - \alpha/2)$.

Testové kritérium Waldovy statistiky vypadá následovně: [8]

$$Wald = \frac{\widehat{\beta_J}}{\widehat{SE_J}}, \tag{12}$$

kde j=0,…,k,

$\widehat{SE_j}$ je standardní chyba odhadu parametrů β_j.

Nulová hypotéza Waldova testu předpokládá, že parametr β_j bude roven nule, a alternativní hypotéza říká, že β_i bude od nuly odlišný:

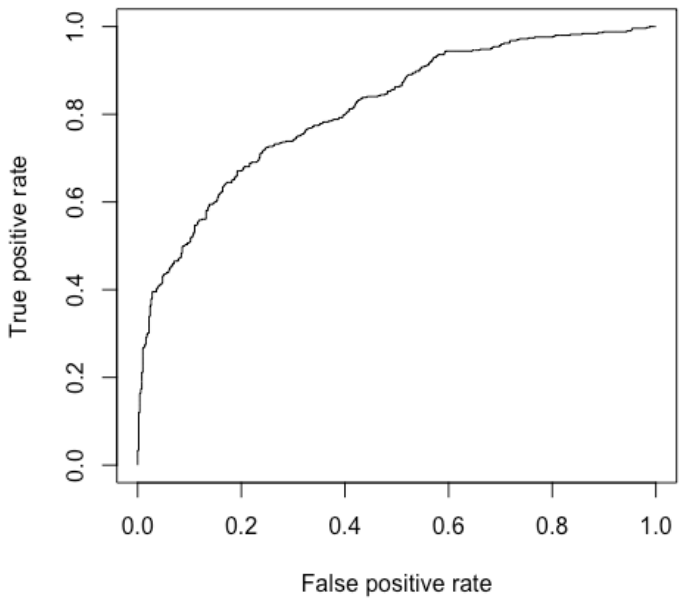
$$H_0: \beta_j = 0$$
$$H_1: \beta_j \neq 0$$

Při zamítnutí nulové hypotézy se předpokládá, že β_j je významným parametrem v modelu, a naopak. [8]

Waldův test lze také použít pro hodnocení významnosti celkového modelu. V tomto případě nulová hypotéza předpokládá rovnost všech parametrů β_j s nulou, a alternativní hypotéza předpokládá alespoň jeden nenulový parametr.

Vzhledem k tomu, že rozdělení je asymptotické, Waldův test se dá použít pouze v případě velkého počtů pozorování, jinak výsledky testu mohou být zavádějící.

Další možnost hodnocení modelu je jeho schopnost správného přiřazení respondenta k binární vysvětlované proměnné *y*. Pro tento účel slouží ROC (Receiver operating characteristic), která je zobrazená v grafu 1.1. Na horizontální osu ROC je vynášená falešně pozitivní míra a na vertikální ose je skutečně pozitivní míra. Citlivost, nebo skutečně pozitivní míra, je chápana jako pravděpodobnost detekce. Falešně pozitivní míra se vypočítává jako (1− specificita). Nejlepší možná předpověď by se nacházela v levém horním rohu a měla by souřadnici (0,1), což by znamenalo 100% citlivost a 100% specificitu modelu. Pokud se ROC blíží k diagonální přímce, jedná se o příklad náhodného hádání. [8], [13]



Graf 2.1: ROC, vlastní výpočet a zpracovaní v RStudio

Pro numerickou interpretaci ROC slouží AUC (Area Under the Curve). Pro výpočet AUC se zavádí p_1, kterou se označuje pravděpodobnost úspěšného přiřazení respondenta k binární vysvětlované proměnné *y*; p_2 je pravděpodobnost chybného přiřazení, a p_3 je pravděpodobnost neschopnosti přiřazení respondenta k žádné vysvětlované proměnné y, přičemž $p_1+p_2+p_3=1$. [8] Potom je AUC definovaná jako:

$$AUC = p_1 + \frac{p_3}{2}. \tag{13}$$

Hodnoty AUC leží v intervalu <0,1>. Vzhledem k tomu, že diagonální ROC spojuje body (0,0) a (1,1), a tento případ je příkladem náhodného hádaní, proto i AUC při realistických klasifikátorech nabývá hodnot v intervalu <0.5, 1>. [2]

2.4 Interpretace parametrů

Na rozdíl od lineárního regresního modelu nelze koeficient β interpretovat jako výši průměrné změny proměnné *y* při jednotkové změně vysvětlující proměnné *x*. [11] U logistické regrese se pro tento účel používá mezní efekt proměnné x_j, který vyjadřuje změnu pravděpodobnosti úspěchu při jednotkové změně x_j. V případě spojité proměnné lze mezní efekt vypočítat jako

$$\frac{\partial p}{\partial x_j} = f(\mathbf{x\beta})\beta_j, \tag{14}$$

kde $f(\mathbf{x\beta}) = F'(\mathbf{x\beta})$.

Hodnota mezního efektu logistického modelu není konstantní, jelikož je závislá na hodnotách vysvětlujících proměnných, proto se zavádí mezní efekt pro průměrné pozorování:

$$MEM_j = f(\bar{\mathbf{x}}\widehat{\boldsymbol{\beta}})\widehat{\beta_J}, \tag{15}$$

kde $\bar{\mathbf{x}} = \frac{1}{N}\sum_1^n \mathbf{x}_i$.

Pro zjištění průměrného mezního efektu v celé populaci se používá následující vzorec:

$$AME_j = \frac{1}{N}\sum\nolimits_1^n f(\mathbf{x_i}\widehat{\boldsymbol{\beta}})\widehat{\beta_J}. \tag{16}$$

Výsledné hodnoty se interpretují jako změny pravděpodobnosti úspěchu.

3 Zobecněný aditivní model

Doposud byly předpokládané pouze lineární vztahy mezi vysvětlovanou proměnnou a vysvětlujícími proměnnými. Většinou v praxi nemají reálná data lineární vztahy, a proto je potřeba zachytit také případnou nelinearitu v proměnných. Pro tento účel slouží tzv. zobecněné aditivní modely, které byly vyvinuté Trevorem Hastie a Robertem Tibshirani.

V případě regrese lze zobecněný aditivní model zapsat ve tvaru: [3]

$$E(Y|x_1, \dots, x_p) = \alpha + f_1(x_1) + \cdots + f_p(x_p), \tag{17}$$

kde $x_1, \dots, x_p$ jsou vysvětlující proměnné,
Y je vysvětlovaná proměnná,
α je intercept,
f_p jsou hladké neparametrické funkce.

3.1 Kubický splajn smoother a vyrovnávací parametr

Nejprve je zapotřebí definovat hladkou funkci neboli smoother jako nástroj, který popisuje trend závislosti vysvětlované proměnné jako funkci vysvětlujících proměnných.[6] Smoother poskytuje odhad trendu, který je méně proměnlivý než *y*. Trend smoother lze jednoduše odvodit z grafu. Také je potřeba dodat, že smoother odhaduje závislost střední hodnoty vysvětlované proměnné na vysvětlujících proměnných.

Hladká funkce je neparametrická, což je velkým rozdílem při porovnaní se zobecněnými lineárními modely. V podstatě smoother aproximuje celý model pomocí součtu různých funkcí, které jsou schopné lepším způsobem popsat trend závislosti *y* na každé jednotlivé nezávislé proměnné.

Kubický splajn smoother je nástrojem, který najde z druhých spojitých derivací všech funkcí $f(x_i)$ jedinou funkci minimalizující penalizované nejmenší čtverce (18).

$$\sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda \int f''(x)^2 dx, \tag{18}$$

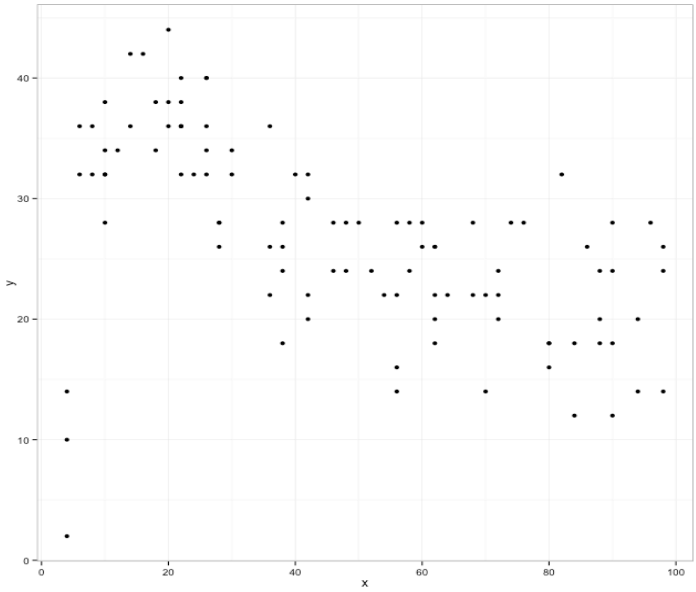
kde λ je vyrovnávací parametr.

První složka vzorce (18) představuje klasickou metodu nejmenších čtverců. Bez použití druhé části vzorce by výsledkem byla interpolační křivka, procházející v blízkosti změřených dat. Druhá část vystihuje zlom této křivky. Pokud $f(x)$ je linearní funkce, potom $\int f''(x)^2 = 0$.

λ je nezáporným vyrovnávacím parametrem, který dává do souladu data s vyhlazováním. Při $\lambda \to \infty$ křivka bude více připomínat přímku, proloženou metodou nejmenších čtverců, protože parametr λ přibližuje $f''(x)$ k nule. Pokud $\lambda \to 0$, penalizace se stává nevýznamnou, a velkou roli budou hrát funkce druhých derivací. Tím pádem čím větší je parametr λ, tím hladší bude křivka, a naopak menší parametr λ poskytuje ohnutější křivku.

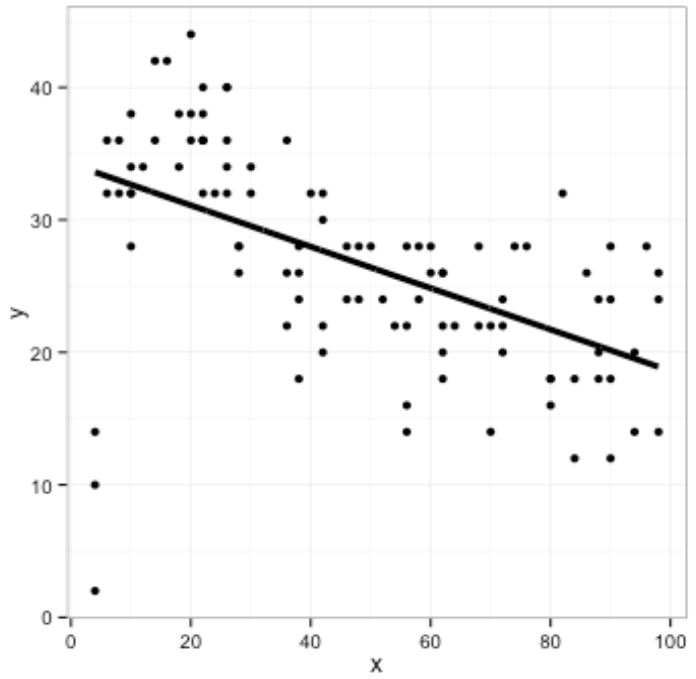
Je zřejmé, že pokud jsou všechny vyrovnávací parametry velmi vysoké, pak bude tento model poměrně nepružný a bude mít velmi málo stupňů volnosti. Jedním ze způsobů měření flexibility modelu je definovat počet stupňů volnosti jako $df_\lambda = tr(S_\lambda)$.[6]

Pro lepší představu je rozdíl mezi kubickým splajnem smoother a metodou nejmenších čtverců zobrazen na grafech 3.1-3.3.

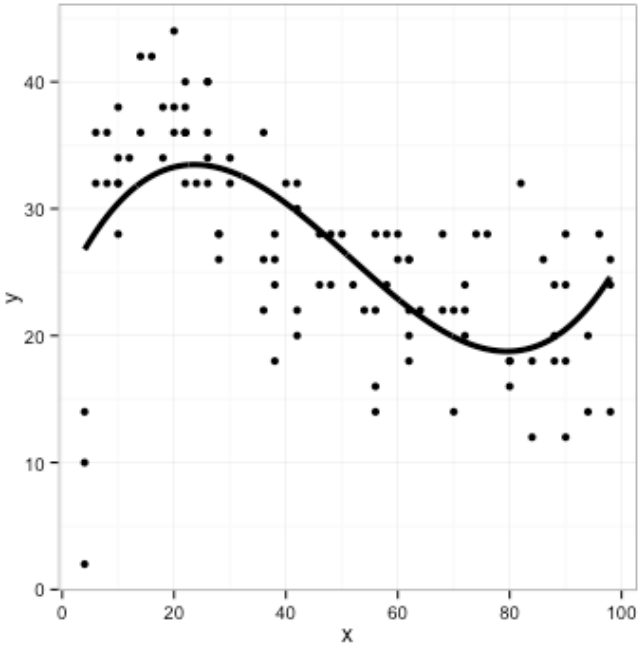


Graf 3.1: Bodový graf pozorování (*x, y)*, zdroj: [12]

Graf 3.1 znázorňuje výsledné hodnoty proměnné *y*, která je závislá na vysvětlující proměnné *x*. Následující graf 3.2 je grafickým znázorněním metody nejmenších čtverců.



Graf 3.2: Metoda nejmenších čtverců, zdroj: [12]



Graf 3.3: Kubický splajn smoother, zdroj: [12]

Graf 3.3 je ukázkou kubického splajnu smoother, který popisuje trend závislostí *y* na proměnné *x*.

Nyní je zapotřebí určit parametr λ. Pokud λ bude příliš vysoká, data budou nadměrně vyhlazená, a pokud bude λ příliš nizká, data nebudou dostatečně vyhlazená. V obou případech to bude znamenat, že se odhady hladkých funkcí nebudou blížit skutečným funkcím. Ideálním případem by bylo vybrat λ takovým způsobem, aby se odhady hladkých funkcí $\hat{f}$ co nejvíce přibližovaly skutečným funkcím. Vhodným kritériem může být volba λ, která minimalizuje střední průměrnou čtvercovou chybu AMSE: [9]

$$AMSE(\lambda) = \frac{1}{n}\sum_{i=1}^{n} E\left\{\widehat{f_\lambda}(x_i) - f(x_i)\right\}^2, \tag{19}$$

kde $y_i = f(x_i) + \varepsilon_i$,

$\widehat{f_\lambda}(x_i)$ je odhad $f(x)$.

Jelikož $f(x)$ je neznámá, lze odvodit střední prediktivní čtvercovou chybu (PSE), která se od AMSE liší pouze konstantní funkcí σ^2.

$$PSE(\lambda) = \frac{1}{n}\sum_{i=1}^{n} E\left\{Y_i^* - \widehat{f_\lambda}(x_i)\right\}^2, \tag{20}$$

kde Y_i^* je nové pozorovaní x_i,

$Y_i^* = f(x_i) + \varepsilon_i{}^*$,

kde $E(\varepsilon_i{}^*) = 0$.

Nyní lze dokázat, že $PSE = AMSE + \sigma^2$:

$$\begin{aligned} & E\left\{Y_i^* - \widehat{f_\lambda}(x_i)\right\}^2 = \\ & = E\left\{Y_i^* - \widehat{f_\lambda}(x_i) + f(x_i) - f(x_i)\right\}^2 = \\ & = E\left\{Y_i^* - f(x_i)\right\}^2 + E\left\{\widehat{f_\lambda}(x_i) - f(x_i)\right\}^2 + 2E\left\{Y_i^* - f(x_i)\right\}\left\{f(x_i) - \widehat{f_\lambda}(x_i)\right\} = \\ & = E\left\{Y_i^* - f(x_i)\right\}^2 + E\left\{\widehat{f_\lambda}(x_i) - f(x_i)\right\}^2 + 2E(\varepsilon_i{}^*)E(\varepsilon_i) = \end{aligned} \tag{21}$$

$$= \delta_i^2 + E\{\widehat{f_\lambda}(x_i) - f(x_i)\}^2.$$

Potom $PSE(\lambda) =$

$$\begin{aligned} &= \frac{1}{n}\sum_{i=1}^{n} E\left\{Y_i^* - \widehat{f_\lambda}(x_i)\right\}^2 = \\ &= \frac{1}{n}\sum_{i=1}^{n} (\delta_i^2 + E\left\{\widehat{f_\lambda}(x_i) - f(x_i)\right\}^2) = \\ &= \delta_i^2 + \frac{1}{n}\sum_{i=1}^{n} E\left\{\widehat{f_\lambda}(x_i) - f(x_i)\right\}^2 = \\ &= AMSE + \sigma^2. \end{aligned} \tag{22}$$

Na základě důkazu lze odvodit, že je jednodušší odhadnout PSE, která se od AMSE liší pouze konstantní funkcí σ^2, což je očekávaná kvadratická chyba při predikci nové proměnné. Předpověď nové proměnné se vypočítává pomocí křížové validace, která se spočívá v postupném vynechání každého vzorku dat, následném odhadu modelu na zbylých datech a vypočtu čtvercové chyby mezi chybějícím údajem a jeho predikovanou hodnotou. [9]

3.1.1 Křížová validace

Dalším krokem určení vyrovnávacího parametru λ je křížová validace, která představuje statistickou metodu rozdělení dat do dvou podmnožin, trénovací a testovací. [6]

V praxi se používá metoda leave-one-out (LOOCV) křížové validace vzhledem k zjednodušení a použití většího objemu dat v trénovací části. Hlavní myšlenkou LOOCV je vynechání bodu (x_i, y_i) jeden po druhém jako testovací podmnožinu a určení odhadu hladké funkce v bodě x_i na základě zbývajících $n - 1$ pozorování.

Součet čtverců křížové validace se konstruuje následujícím způsobem:

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \widehat{f_\lambda^{-\iota}}(x_i)\right\}^2, \tag{23}$$

kde $\widehat{f_\lambda^{-\iota}}(x_i)$ identifikuje vhodnost odhadu x_i spočítaného při vynechaní i-tého pozorování. S použitím stejného principu, který byl popsán výše, se postupuje dále:

$$\begin{aligned} &E\left\{y_i - \widehat{f_\lambda^{-\iota}}(x_i)\right\}^2 = \\ &= E\left\{y_i + f(x_i) - f(x_i) - \widehat{f_\lambda^{-\iota}}(x_i)\right\}^2 = \\ &= E\left\{y_i - f(x_i)\right\}^2 + E\left\{f(x_i) - \widehat{f_\lambda^{-\iota}}(x_i)\right\}^2 + E\{y_i - f(x_i)\}E\left\{f(x_i) - \widehat{f_\lambda^{-\iota}}(x_i)\right\} = \\ &= \delta_i^2 + E\{f(x_i) - \widehat{f_\lambda^{-\iota}}(x_i)\}^2, \end{aligned} \tag{24}$$

kde $E\{y_i - f(x_i)\}E\left\{f(x_i) - \widehat{f_\lambda^{-\iota}}(x_i)\right\} = 0$ z toho důvodu, že $\widehat{f_\lambda^{-\iota}}(x_i)$ nezahrnuje y_i, a také se předpokládá, že $\widehat{f_\lambda^{-\iota}}(x_i) \approx \widehat{f_\lambda}(x_i)$, potom:

$$\begin{aligned} &E\{CV(\lambda)\} = \\ &= E\left\{\frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \widehat{f_\lambda^{-\iota}}(x_i)\right\}^2\right\} = \end{aligned} \tag{25}$$

$$= \delta_i^2 + E\left\{f(x_i) - \widehat{f_\lambda^{-i}}(x_i)\right\}^2 =$$
$$= \delta_i^2 + E\left\{\widehat{f_\lambda}(x_i) - f(x_i)\right\}^2 =$$
$$= PSE$$

Minimalizace $CV(\lambda)$ je ekvivalentní minimalizaci $PSE(\lambda)$. Parametr λ, který minimalizuje $CV(\lambda)$, lze označit jako vyhlazující parametr. [6]

3.2 Odhad aditivního modelu

Klasický zápis aditivního modelu vypadá následovně: [3]

$$Y = \alpha + \sum_{j=1}^{p} f_j(\mathrm{x}_j) + \varepsilon, \tag{26}$$

kde $\varepsilon \sim i.i.d.(0, \sigma^2)$.

V případě aditivního modelu se při daných pozorovaní x_i, y_i zavádí penalizované nejmenší čtverce,

$$PRSS(\alpha, f_1, f_2, \dots, f_p) = \sum_{i=1}^{N} (y_i - \alpha - \sum_{j=1}^{p} f(x_{ij}))^2 + \sum_{j=1}^{p} \lambda_j \int f_j{}''(x_j)^2 dx_j, \tag{27}$$

kde $\lambda_j \geq 0$ je vyrovnávací parametr, který kontroluje soulad mezi vhodností a hladkostí modelu, a funkce f_j je kubickým splajnem smoother.

3.2.1 Algoritmus zpětného vybavení pro aditivní modely

Je nezbytné přidat další omezení modelu, aby řešení bylo jedinečné. Obvykle se předpokládá, že $\sum_{i=1}^{N} f_j(x_{ij}) = 0\ \forall j$. Z toho vyplývá, že $\hat{\alpha} = ave(y_i)$. [3]

Pro odhad aditivních modelů se používá algoritmus zpětného vybavení pro aditivní modely. Prvním krokem algoritmu je inicializace $\hat{\alpha}$:

1. $\hat{\alpha} = \frac{1}{N}\sum_1^N y_i,\ \widehat{f_j} \equiv 0, \forall i, j$ (28)

Druhým krokem je cyklus: kubický vyhlazující splajn se aplikuje pro získaní nového odhadu $\widehat{f_j}$ takovým způsobem, že se pro každý prediktor v pořadí pomocí aktuálního odhadu ostatních funkcí $\widehat{f_k}$ počítá $y_i - \hat{\alpha} - \sum_{k \neq j} \widehat{f_k}(x_{ik})$. Proces se opakuje, dokud se odhady $\widehat{f_k}$ nestabilizují.

2. $\widehat{f_j} \leftarrow S_j\left[\left\{y_i - \hat{\alpha} - \sum_{k \neq j} \widehat{f_k}(x_{ij})\right\}_1^N\right],$ (29)
$$\widehat{f_j} \leftarrow \widehat{f_j} - \frac{1}{N}\sum_{i=1}^{N} \widehat{f_j}(x_{ij}).$$

Odhadnuté funkce $\widehat{f_j}$ odhalí nelineární vztah vysvětlujících proměnných x_j. Ne všechny funkce f_j musejí být nelineární, to znamená, že je možné kombinovat lineární a jiné parametrické formy s nelineární funkcí. [3]

3.3 Zobecněný aditivní logistický model

Zobecněné aditivní modely lze také použít v případě binární vysvětlované proměnné *y*. Za předpokladu, že podmíněná střední hodnota vysvětlované proměnné je lineární funkcí vysvětlujících proměnných, $\mu(X) = Pr\{y = 1|X\}$, potom zmíněnou rovnici logistického modelu (4) lze přepsat do tvaru: [3]

$$ln\left(\frac{\mu(X)}{1-\mu(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p. \tag{30}$$

Aditivní logistický model nahradí každou lineární složku obecnější funkční formou:

$$ln\left(\frac{\mu(X)}{1-\mu(X)}\right) = f_0 + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p). \tag{31}$$

Obecně platí, že podmíněná střední hodnota vysvětlované proměnné *y* souvisí s aditivní funkcí prediktorů pomocí linkovací funkce *g*:

$$g[\mu(X)] = f_0 + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p). \tag{32}$$

Příklady klasických linkovacích funkcí jsou v tabulce 3.1: [3]

Linkovací funkce	$g(\mu_i)$
Identity	μ_i
Log	$log(\mu_i)$
Logit	$log\frac{\mu_i}{1-\mu_i}$
Probit	$\Phi^{-1}(\mu_i)$

Tabulka 3.1: Linkovací funkce, vlastní zpracování

- $g(\mu) = \mu$, představuje lineární pravděpodobnostní model.
- $g(\mu) = F_{logit}(\mu)$, kde $F_{logit}(\mu)$ je distribuční funkcí logistického rozdělení.
- $g(\mu) = F_{probit}(\mu)$, kde $F_{probit}(\mu)$ představuje distribuční funkcí normálního rozdělení.
- $g(\mu) = log(\mu)$ pro log-lineární nebo log-aditivní modely, kde se *y* řídí Poissonovým rozdělením.

Všechny tři funkce patří do exponenciální rodiny rozdělení.
Obecně platí $E(Y|X) = \mu,\ (\mu_i = 1 * p_i + 0 * (1 - p_i) = p_i)$. [6]

$$\eta(x) = g(\mu) = \ln\left(\frac{p(x)}{1-p(x)}\right), \tag{33}$$

kde η je funkce p proměnných.

3.4 Lokální scoringový algoritmus

Postup řešení aditivního logistického modelu má stejný princip jako klasický aditivní model. Lze ale uvést některé rozdíly, způsobené nula-jedničkovou vysvětlovanou proměnnou.

Prvním a důležitým rozdílem je, že penalizované nejmenší čtverce obsahují váhy pro jednotlivá pozorování a vypadají následovně: [3]

$$\min RSS(f,\lambda) = \sum_{i=1}^{N} w_i\{y_i - f(x_i)\}^2 + \lambda \int \{f''(t_j)\}^2 dt_j, \tag{34}$$

kde $w_i \geq 0$ jsou váhy pozorování.

Druhá odlišnost aditivní logistické regrese spočívá v algoritmu minimalizace nejmenších čtverců. Pro aditivní logistický model existuje tzv. lokální scoringový algoritmus. Prvním krokem se inicializuje $\hat{\alpha}$ tímto způsobem: [3]

$$\hat{\alpha} = log[\bar{y}/(1-\bar{y})], \tag{35}$$

kde $\bar{y} = ave(\bar{y})$,

$\hat{f}_j \equiv 0, \forall j$.

Ve druhém kroku je zapotřebí nadefinovat

$$\hat{\eta}_\iota = \hat{\alpha} + \sum_j \hat{f}_j(x_{ij}) \tag{36}$$

a

$$\hat{p}_\iota = 1/[1 + \exp(-\hat{\eta}_\iota)]. \tag{37}$$

V další iteraci je nutné konstruovat cílovou proměnnou:

1. $z_i = \hat{\eta}_\iota + \frac{(y_1 - \hat{p}_\iota)}{\hat{p}_\iota(1-\hat{p}_\iota)}$. (38)

Potom se konstruují váhy následujícím způsobem:

2. $w_i = \hat{p}_\iota(1-\hat{p}_\iota)$. (39)

Ve třetí iteraci se aditivní model přizpůsobí k z_i s vahami w_i pomocí algoritmu zpětného vybavení s vahami. Ve výsledku budou zjištěné odhady pro $\hat{\alpha}, \hat{f}_j, \forall j$. Nyní lze vypočítat konvergenční kritérium:

$$\Delta(\eta^1, \eta^0) = \frac{\sum_{j=1}^{p} \left\| f_j^1 - f_j^0 \right\|}{\sum_{j=1}^{p} \left\| f_j^0 \right\|} \tag{40}$$

Druhý krok se opakuje, dokud $\Delta(\eta^1, \eta^0)$ neklesne pod předem zvolenou hranici, například 0,001, nebo defaultně se horní hranice nastavuje na 10^{-8}. [6]

4 Sběr dat

V praktické části mé bakalářské práce se budu zabývat customer attrition analýzou pro klienty bankovního domu. Customer attrition je známý také jako obrat klientů nebo odliv zákazníků.

Podle mého názoru lze změnu spotřebitelského chování retailových klientů banky prozkoumat přednostně pomocí analýzy jejich příchozích a odchozích transakcí.

Většinou příchozí transakce představují příjmy retailových klientů za předpokladu, že mzda a další sociální dávky se pravidelně posílají zaměstnavatelem a sociálními úřady na určitý učet, který byl předem nahlášen. Změna daného účtu pro převod mzdy a sociálních dávek zabere nějaký čas, čímž je vyloučeno rychlé střídání bank zákazníkem, a rozhodně se projeví v celkových měsíčních částkách a počtu příchozích transakcí klienta.

Co se týká odchozích transakcí, klient má také pravidelné měsíční platby, které jsou schopné popsat jeho stabilnost vůči bance. Na příklad měsíční platby za elektřinu, telefon, internet, nájemné za byty, platby v obchodech a další budou pravidelně odesílané z bankovního účtu klienta. Pokud se klient rozhodne změnit banku, tato změna chování se také odrazí na celkových částkách a také počtů odchozích transakcí.

Ztrátou klienta nebo odchodem ze strany klienta rozumím rozhodnutí, že klient zastaví nebo do určité míry omezí využívání bankovních služeb.

4.1 Sběr dat

V této části se budu věnovat popisu sběru a zpracování dat. Zde popíšu způsob sběru dat pro jednotlivé proměnné, které by ve výsledku mohly ovlivnit pravděpodobnost odchodu klienta.

Všechna data jsem sbírala měsíčně v časovém intervalu od ledna do září roku 2017. Moje práce je zaměřená na analýzu spotřebitelského chovaní retailových klientů, což znamená, že klienti jsou fyzické osoby nepodnikající. Z databáze klientů jsem pro konstrukci modelů použila výběr v rozsahu 5000 klientů.

4.1.1 Vysvětlující proměnné

Z původních dat jsem byla schopná vybrat měsíční počty a sumy odchozích a příchozích transakcí pro každou protistranu jednotlivého klienta zvlášť. Jak je zobrazeno v tabulce 4.1, náhodně zvolený klient přijal během jednoho měsíce od Starbucks jednu platbu ve výši 15000 Kč, jelikož se Počet příchozích transakcí rovná 1. Dále je vidět, že klient zaplatil 5 krát v Kauflandu v celkové výši 2500 Kč, a jednou platbou zaplatil 650 Kč Vysoké škole ekonomické.

Číslo klienta	Suma příchozích transakcí	Počet příchozích transakcí	Suma odchozích transakcí	Počet odchozích transakcí	Protistrana
1	15000	1	0	0	Starbucks
1	0	0	-2500	5	Kaufland
1	0	0	-650	1	Vysoká škola ekonomická

Tabulka 4.1: Příklad dat, vlastní výpočty a zpracování v Excel

Dále z výše uvedených dat jsem vypočítala průměr počtů odchozích transakcí za 6 měsíců od února do července roku 2017. Následně byl spočítán průměr počtů odchozích transakcí za poslední dva měsíce, které jsem měla k dispozici, a to je za srpen a září. Pro posouzení, zda-li došlo ke snížení průměrného počtu transakcí během posledních dvou měsíců ve vztahu k předcházejícímu půl roku, jsem spočítala poměr obou průměrů:

$$\frac{Průměrný\ počet\ odchozích\ transakcí\ za\ období\ 2.2017-7.2017}{Průměrný\ počet\ odchozích\ transakcí\ za\ období\ 8.2017-9.2017}.$$

Tento průměr jsem označila jako vysvětlující proměnnou *podil* a následně používám pro konstrukci modelů, které budou popsané níže.

Podil	Absolutní četnost	Relativní četnost (%)
do 1	1942	39%
do 1,5	1820	36%
do 2	507	10%
do 3	456	9%
nad 3	275	6%

Tabulka 4.2: Četnosti vysvětlované proměnné *podíl*, vlastní výpočty a zpracování v Excel

V tabulce 4.2 jsou uvedené absolutní a relativní četnosti proměnné *podil*, ze kterých je vidět, že v 39 % pozorování se průměrný počet odchozích transakce buď nezměnil nebo se naopak zvýšil, což znamená zvětšení aktivity u 1942 klientů. Další číslo říká, že 36 % pozorovaných klientů snížilo počet odchozích plateb přibližně o čtvrtinu. Desetina klientů, což v přepočtu činí 507 klientů, omezila odchozí transakce na 50 %. Dalších 15% snížilo své průměrné platby během posledních dvou měsíců přibližně 3 a vícekrát.

Stejný princip jsem použila pro zjištění změn sumy a počtu příchozích transakcí:

$$\frac{Průměrná\ suma\ příchozích\ transakcí\ za\ období\ 2.2017-7.2017}{Průměrná\ suma\ příchozích\ transakcí\ za\ období\ 8.2017-9.2017},$$

$$\frac{Průměrný\ počet\ příchozích\ transakcí\ za\ období\ 2.2017-7.2017}{Průměrný\ počet\ příchozích\ transakcí\ za\ období\ 8.2017-9.2017}.$$

Poměr sum příchozích transakcí je další vysvětlující proměnnou, která je v databázi pojmenovaná jako *suma_podil_pr*.

Suma_podil_pr	Absolutní četnost	Relativní četnost (%)
do 1	1709	34%
do 1,5	1453	29%
do 2	505	10%
do 3	529	11%
nad 3	804	16%

Tabulka 4.3: Četnosti vysvětlující proměnné *suma_podil_pr*, vlastní výpočty a zpracování v Excel

V tabulce 4.3, kde jsou zobrazené absolutní a relativní četnosti poměru změn průměrných sum příchozích transakcí, je vidět, že 1709 respondentů dostávalo během posledních dvou měsíců na svůj účet stejné nebo větší částky než v předcházejícím půl roce. U 29 % klientů příjem klesl o cca 25 %. U desetiny pozorovaných klientů je zaznamenán pokles příchozích plateb skoro o polovinu. Dalším 37 % respondentům poklesla suma celkových příchozích plateb 3 a vícekrát.

Poměr počtů příchozích transakcí je vysvětlující proměnnou s názvem *podil_pr*.

Podil_pr	Absolutní četnost	Relativní četnost (%)
do 1	2102	42%
do 1,5	1866	37%
do 2	622	12%
do 3	302	6%
nad 3	108	2%

Tabulka 4.4: Četnosti vysvětlující proměnné *podil_pr*, vlastní výpočty a zpracování v Excel

Tabulka 4.4 zobrazuje přehled absolutních a relativních četností poměru průměrných počtů příchozích transakcí. V posledních dvou měsících zůstalo 2102 respondentů v počtu průměrných příchozích transakcí na stejné nebo vyšší úrovni než v předcházejícím půl roce. Dalším 37 % klientům klesly transakce přibližně o čtvrtinu. U 12 % klientů byl zaznamenán pokles v počtu transakcí o 2 krát. 500 respondentů snížilo své příchozí transakce 3 a vícekrát.

Další vysvětlující proměnnou je věk klienta, který je pojmenován jako *client_age*. V tabulce 4.5 jsou uvedené četnosti pro jednotlivé věkové skupiny. Okolo 44 % pozorovaných klientů je ve věku 36-50 let, druhá největší skupina klientů je ve věku 51-65 let a tvoří 30 % všech pozorovaných klientů.

Client_age	Absolutní četnost	Relativní četnost (%)
do 20	19	0,4%
21-35	700	14,0%
36-50	2187	43,7%
51-65	1498	30,0%
nad 65	493	9,9%

Tabulka 4.5: Četnosti vysvětlující proměnné *client_age*, vlastní výpočty a zpracování v Excel

Také jsem zahrnula proměnnou *hist_uctu*, kterou se označuje délka vedení účtu u každého klienta. Jelikož v databázi bylo datum založení účtu, přepočítala jsem to ke dni sběru dat následujícím způsobem:

$$\frac{Datum\ sběru\ dat - datum\ založení\ účtu}{365},$$

a všechny výsledky jsou zaokrouhlené setiny.

Následující důležitou proměnnou je existence kreditní karty, která může v dostatečné míře zvětšit aktivitu klienta. Proměnná se v databázi jmenuje jako *credit_card* a v případě, že klient disponuje kreditní kartou, je mu kódována 1, v opačném případě, je kódována 0. Přehled absolutních a relativních četností proměnné *credit_card* je zobrazen v tabulce 4.6.

Credit_card	Absolutní četnost	Relativní četnost (%)
Ano (1)	1461	29%
Ne (0)	3539	71%

Tabulka 4.6: Četnosti vysvětlované proměnné *credit_card*, vlastní výpočty a zpracování v Excel

V tabulce 4.6 je vidět, že téměř třetina klientů disponuje kreditní kartou.

Další vysvětlující proměnnou je riziková kategorie, která se v bance člení na několik podskupin, např. kategorie klientů, kteří se nacházejí ve fázi Early Collection nebo Late Collection. V rámci bakalářské práce jsem se rozhodla, že tato proměnná bude binární. V případě, že klient spadá do nějaké rizikové kategorie, bude kódována 1. Pokud je klient považován jako standardní, bude kódována 0. Tato proměnná je označená jako *risk_category*.

Risk_category	Absolutní četnost	Relativní četnost (%)
Ano (1)	538	11%
Ne (0)	4462	89%

Tabulka 4.7: Četnosti vysvětlované proměnné *risk_category*, vlastní výpočty a zpracování v Excel

Z tabulky 4.7 lze odvodit, že více než 10 % pozorovaných klientů má nějakou rizikovou kategorii.

4.1.2 Vysvětlovaná proměnná

Nyní je zapotřebí definovat předběžný pokles aktivity klienta ve využívání bankovních služeb. Pro definici ztráty zájmu v bankou poskytovaných služeb jsem se rozhodla použít celkové sumy odchozích transakcí, jelikož počet plateb může klesnout, ale celková odeslaná částka zůstane beze změny. Pro výpočet poměrů průměrných sum odchozích transakcí používám stejný princip, který byl použit pro vysvětlující proměnné.

$$\frac{Průměrná\ suma\ odchozích\ transakcí\ za\ období\ 2.2017-7.2017}{Průměrná\ suma\ odchozích\ transakcí\ za\ období\ 8.2017-9.2017}.$$

V tabulce 4.8 uvádím absolutní a relativní četnosti poměrů průměrných sum odchozích transakcí.

Suma_podil	Absolutní četnost	Relativní četnost (%)
do 1	1195	24%
do 1,5	970	19%
do 2	354	7%
do 3	1166	23%
nad 3	1315	26%

Tabulka 4.8: Četnosti vysvětlované proměnné *suma_podil*, vlastní výpočty a zpracování v Excel

Nyní potřebuji určit, zda klient z banky odchází nebo zůstává, a proto musím nadefinovat vysvětlovanou proměnnou jako binární. V rámci své bakalářské práce jsem se rozhodla, že odcházející klient bude snižovat celkovou měsíční částku odchozích plateb o 2 a vícekrát, což je dostatečně významný pokles aktivity ve využívání služeb. Pokud bude zaznamenán podobný pokles, dá se předpokládat, že chybějící platby klient pravděpodobně provede v jiné bance. Podobným klientům jsem v rámci binární proměnné přiřadila 1 a pro ostatní klienty, kteří mají výše zmíněný poměr menší než 2, což předpokládá stabilitu ve využívání bankovních služeb, jsem přiřadila 0. Tuto vysvětlovanou proměnnou jsem pojmenovala jako *Y*. Jelikož jsem pro definici vysvětlované proměnné použila průměrné sumy odchozích transakcí, nebudu tyto poměry používat pro konstrukci modelů.

5 Aplikace logistické regrese

V této kapitole mé bakalářské práce se pokusím aplikovat teoretické znalosti logistického modelu, který byl popsán v první kapitole. Veškeré výpočty budu provádět pomocí statistického software RStudio, kde jsou potřebné balíčky pro analýzu dat k volnému stažení.

5.1 Odhad logistického modelu

Soubor s daty jsem rozdělila na trénovací a testovací podmnožiny v poměru 80/20. Trénovací podmnožinu budu používat pro konstrukci modelu, což znamená, že všechny základní výpočty a manipulace s modelem budu provádět na trénovací podmnožině souboru. Testovací podmnožinu budu používat pro hodnocení modelu pomocí ROC a AUC.

V prvním kroku vytvořím model se všemi vysvětlujícími proměnnými. V tabulce 5.1 jsou zobrazené P-hodnoty získané na základě Waldovy statistiky. P-hodnota posuzuje významnost jednotlivých parametrů, které byly do modelu zahrnuté.

Proměnná	P-hodnota
podil	<0,001
podil_pr	<0,001
suma_podil_pr	0,005
client_age	<0,001
hist_uctu	<0,001
credit_card	<0,001
risk_category	0,044

Tabulka 5.1: Výsledky pro logistický model se všemi vysvětlujícími proměnnými, vlastní výpočty a zpracování

Z výsledků v tabulce 5.1 je vidět, že koeficienty všech proměnných jsou statisticky významné na 5% hladině významnosti, jelikož je jejich P-hodnota menší než 0,05.

5.2 Mezní efekty

V teoretické části logistického modelu byly zmíněné mezní efekty, pomocí kterých lze interpretovat vlivy jednotkových změn vysvětlujících proměnných na změnu pravděpodobnosti úspěchu. [11]

Proměnná	MEM efekt
podil	0,385
podil_pr	0,058
suma_podil_pr	0,005
client_age	-0,013
hist_uctu	0,019
credit_card	-0,213
risk_category	-0,066

Tabulka 5.2: MEM pro logistický model se všemi vysvětlujícími proměnnými, vlastní výpočty a zpracování

V tabulce 5.2 jsou vidět mezní efekty pro průměrné pozorování pro každou proměnnou. Vzhledem k tomu, že MEM není tak vhodný pro interpretaci, popíšu interpretaci pro průměrný mezní efekt v celé populaci.

Proměnná	AME efekt
podil	0,287
podil_pr	0,043
suma_podil_pr	0,004
client_age	-0,009
hist_uctu	0,014
credit_card	-0,159
risk_category	-0,048

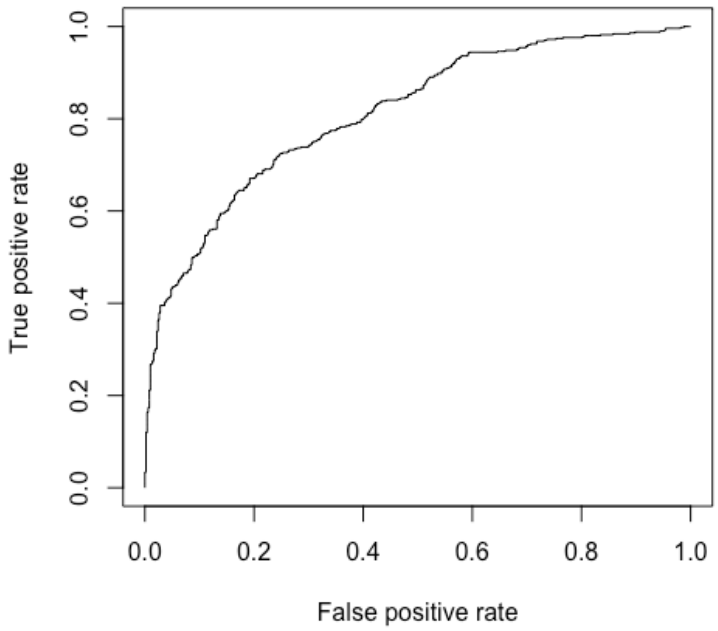
Tabulka 5.3: AME pro logistický model se všemi vysvětlujícími proměnnými, vlastní výpočty a zpracování

V tabulce 5.3 jsou zobrazené průměrné mezní efekty v celé populaci. Při zvýšení proměnné *podil* o jednotku se pravděpodobnost odchodu klienta zvýší o cca 28,7 %. Pokud chci interpretovat výsledek bez použití názvů proměnných, tak se pravděpodobnost odchodu klienta z bankovního domu zvýší o 28,7 % při jednotkovém růstu poměru průměrných počtů transakcí. Při jednotkovém růstu poměru průměrných počtů příchozích transakcí se pravděpodobnost ztráty klienta zvětšuje o 4,3 %. Zvýšení o jednotku poměru průměrných sum příchozích plateb zvětšuje pravděpodobnost odchodu klienta o 0,4 %. Se zvýšením věku klienta o jednotku klesá pravděpodobnost odchodu klienta z banky o 0,9 %. Při zvýšení o jednotku délky vedení účtu se zvětšuje pravděpodobnost odchodu klienta o 1,4 %. Pokud klient vlastní kreditní kartu,

pravděpodobnost odchodu klienta se snižuje o 15,9 %. Jestliže je klientovi přiřazena riziková kategorie, klesá pravděpodobnost odchodu o 4,8 %.[11]

5.3 Hodnocení modelu

Pro hodnocení sestaveného modelu prozkoumám jeho schopnost správně přiřadit respondenta k binární vysvětlované proměnné. Pro tento účel používám ROC. Hodnocení provádím na druhé podmnožině datového souboru, který byl předem určen pro testovací část.



Graf 5.1: ROC pro logistický model, vlastní výpočet a zpracování v RStudio

Na grafu 5.1 je vidět průběh ROC, kterou model odhadnul. Pro numerickou interpretaci použiji AUC.

Ve výsledku dostávám hodnotu AUC, která se rovná 0,8118962. To znamená, že model byl schopen v 81,18 procentech správně přiřadit respondenta k příslušné vysvětlované proměnné, což také znamená, že výsledný logistický model může v 81,18 % případů předpovědět odchod klienta.

6 Aplikace zobecněného aditivního modelu

V této kapitole se pokusím aplikovat teoretické poznatky z kapitoly „Zobecněný aditivní model“ pro predikci odchodu klienta. Již dřív bylo zmíněno, zobecněný aditivní model by měl zachytit případnou nelinearitu mezi vysvětlující a vysvětlovanou proměnnou, což by v daném případě znamenalo lepší predikční schopnost v porovnání s logistickým modelem.

Stejně jako při aplikaci logistického modelu je zapotřebí rozdělit soubor s daty na dvě podmnožiny, trénovací a testovací.

6.1 GAM se všemi vysvětlujícími proměnnými

Nejprve vytvořím zobecněný model se všemi vysvětlujícími proměnnými. Pro konstrukci modelu aplikuji trénovací podmnožinu.

Prvním krokem jsem pro všechny kvantitativní proměnné (*podil, podil_pr, suma_podil_pr, client_age, hist_uctu*) použila vyhlazování kubickým splajnem. Vzhledem k tomu, že zbývající dvě proměnné jsou binární, nejlepší volbou pro vyhlazovací funkci je lineární závislost. V tabulce 6.1 jsou zobrazené P-hodnoty testu ANOVA pro parametrické efekty. V daném případě se pomocí P-hodnoty posuzují předpokládané lineární vztahy.

Proměnná	P-hodnota
s(podil)	<0,001
s(podil_pr)	<0,001
s(suma_podil_pr)	0,165
s(client_age)	<0,001
s(hist_uctu)	<0,001
credit_card	<0,001
risk_category	0,074

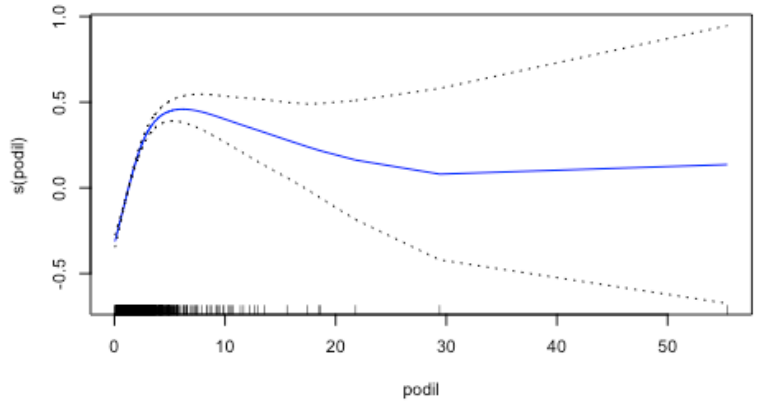
Tabulka 6.1: ANOVA pro parametrické efekty aditivního logistického modelu se všemi vysvětlujícími proměnnými, vlastní výpočty a zpracování

Proměnná	P-hodnota
s(podil)	<0,001
s(podil_pr)	<0,04
s(suma_podil_pr)	<0,001
s(client_age)	<0,001
s(hist_uctu)	<0,001
credit_card	
risk_category	

Tabulka 6.2: ANOVA pro neparametrické efekty aditivního logistického modelu se všemi vysvětlujícími proměnnými, vlastní výpočty a zpracování

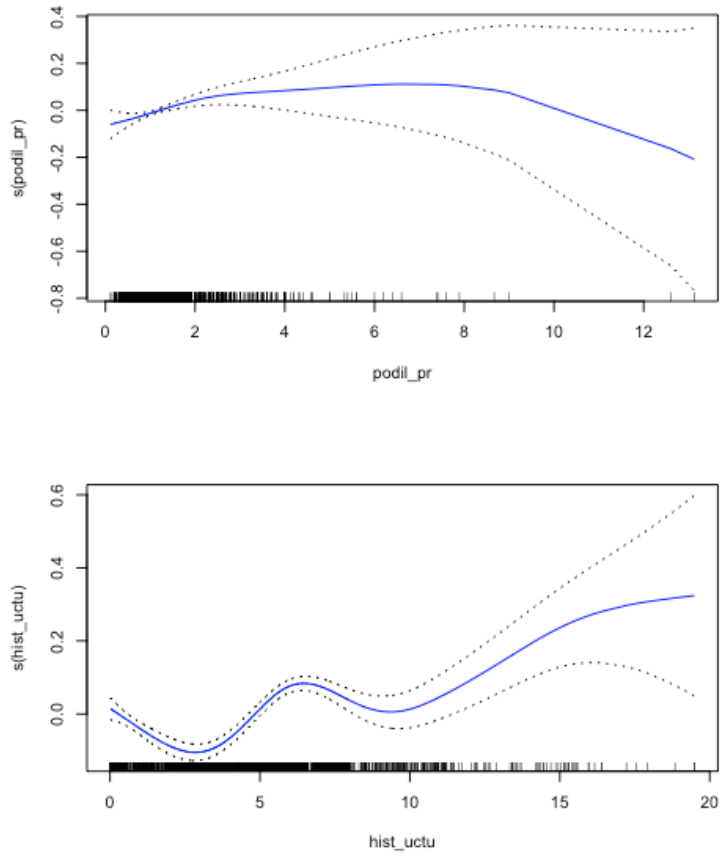
V tabulce 6.2 je vidět, že pomocí ANOVA pro neparametrické efekty posuzuje předpokládaný nelineární vztah mezi vysvětlovanou a vysvětlující proměnnou. Jelikož jsem pro proměnné *credit_card* a *risk_category* nepoužila vyhlazování kubickým splajnem, tyto proměnné nemají v daném testu uvedenou P-hodnotu.

Pro lepší pochopení používám také grafické znázornění pro každou proměnnou zvlášť.



Graf 6.1: Závislost na vysvětlující proměnné *podil*, vlastní výpočet a zpracování v RStudio

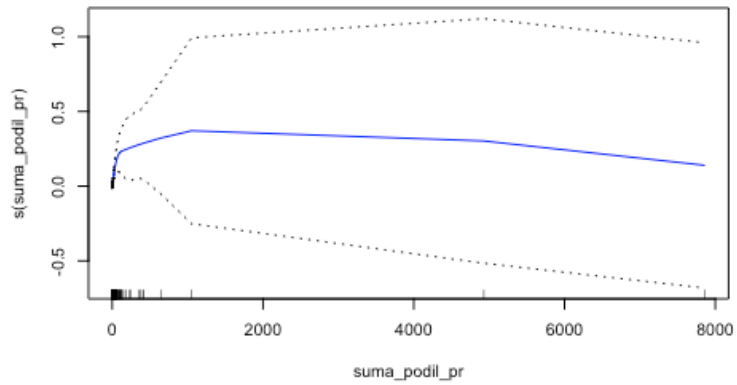
Podle malé p-hodnoty testu ANOVA pro neparametrické efekty z tabulky 6.2 a grafu 6.1 lze odvodit, že závislost mezi vysvětlovanou a vysvětlující proměnnou má složitější vztah než lineární. Na grafu 6.1 je vidět, že na začátku zvětšení proměnné *podil* vede k růstu pravděpodobností odhodu klienta. Potom se postupně pravděpodobnost odchodu klienta s rostoucí hodnotou proměnné *podil* klesá a při hodnotách větší než 30 se pravděpodobnost odchodu klienta stabilizuje. Podobné neparamatrické trendy spadají do kategorie invertovaného U trendu. [2]



Graf 6.2: Závislost na vysvětlující proměnné *podil_pr* a *hist_uctu*, vlastní výpočet a zpracování v RStudio

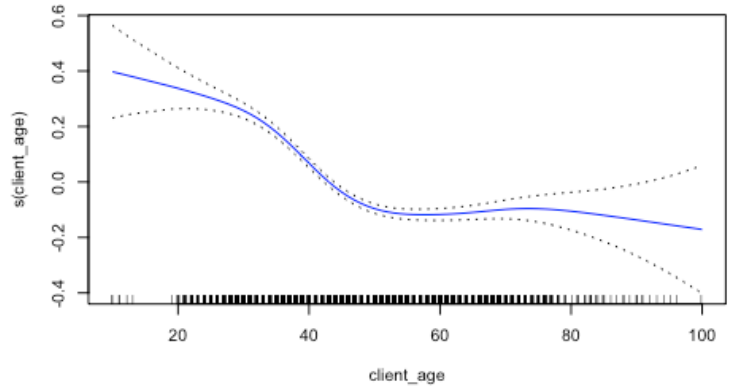
Podobně lze odvodit vztahy pro další dvě vysvětlující proměnné *podil_pr* a *hist_uctu*. Jejich p-hodnoty testu ANOVA pro neparametrické efekty jsou menší než 0,05 a graf 6.2 zobrazuje nelineární vztahy pro každou proměnnou. Na grafu závislosti na vysvětlující proměnné *podil_pr* je vidět, že vysoké a nízké hodnoty vedou k menší pravděpodobnosti odchodu klienta, zatímco centrální hodnoty zvyšují pravděpodobnost odchodu klienta. Tato proměnná je dalším příkladem invertovaného U trendu.

Graf závislosti na proměnné *hist_uctu* představuje kategorii komplexního trendu. Proměnné v této kategorii zvyšují prediktivní sílu modelu. Na začátku se pravděpodobnost odchodu klienta mění. Následně s hodnotou vysvětlující proměnné větší než 10 roste pravděpodobnost odchodu klienta.



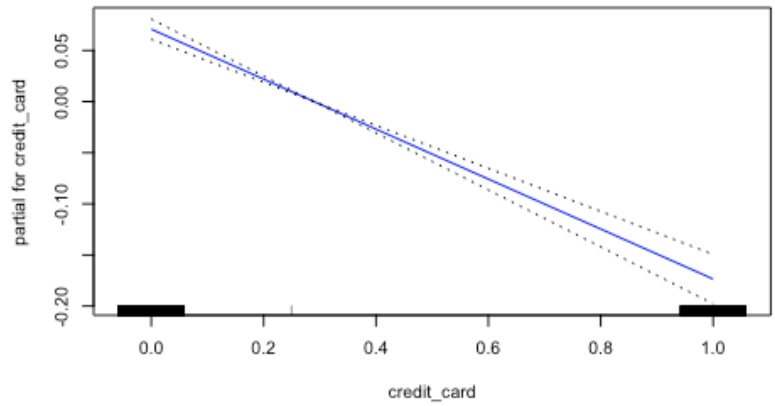
Graf 6.3: Závislost na vysvětlující proměnné *suma_podil_pr*, vlastní výpočet a zpracování v RStudio

Odlišným případem je závislost vysvětlované proměnné na vysvětlující proměnné *suma_podil_pr*. I když je její p-hodnota v testu ANOVA pro neparametrické efekty menší než 0,05, v testu ANOVA pro parametrické efekty se p-hodnota rovná 0,165. Zároveň graf 6.3 připomíná lineární vztah závislosti. Vzhledem k výše uvedenému jsem se rozhodla, že při konstrukci dalšího zlepšeného modelu odstraním vyhlazovaní kubickým splajnem pro tuto proměnnou.

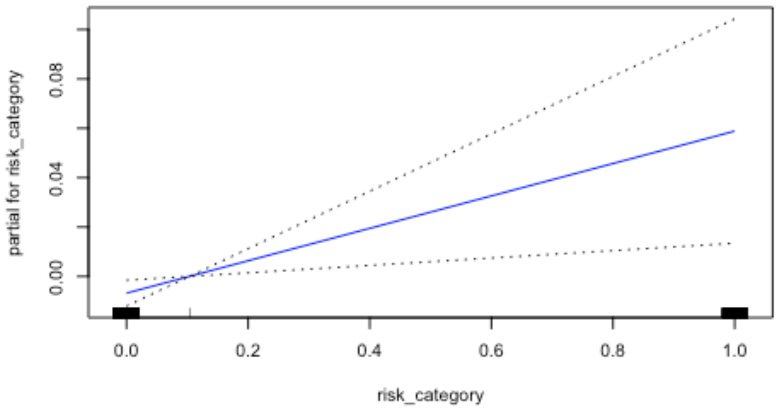


Graf 6.4: Závislost na vysvětlující proměnné *client_age*, vlastní výpočet a zpracování v RStudio

Další proměnná *client_age* také patří do skupiny proměnných, u kterých má závislost nelineární charakter. Graf 6.4 to krásně znázorňuje a p-hodnota v tabulce 6.2 to potvrzuje. Je vidět, že malé hodnoty proměnné *client_age* vedou k velké pravděpodobnosti odchodu klienta, zatímco zvyšující se hodnoty vysvětlující proměnné vedou ke snížení pravděpodobnosti odchodu klienta. Podobný trend má název kvazi-exponenciální trend. [2]



Graf 6.5: Závislost na vysvětlující proměnné *credit_card*, vlastní výpočet a zpracování v RStudio



Graf 6.6: Závislost na vysvětlující proměnné *risk_category*, vlastní výpočet a zpracování v RStudio

Grafy 6.5 a 6.6 popisují lineární vztahy binárních vysvětlujících proměnných, jelikož při konstrukci modelu jsem pro tyto dvě proměnné nepoužívala vyhlazování pomocí kubického splajnu, což je také důvodem, proč tyto proměnné nemají p-hodnoty v testu pro neparametrické efekty. Pravděpodobnost odchodu klienta má nepřímou lineární závislost na proměnné *credit_card*. S rostoucí hodnotou vysvětlující proměnné klesá pravděpodobnost odchodu klienta. Zároveň proměnná *risk_category* má přímý vliv na pravděpodobnost odchodu klienta. Zvyšující se hodnoty vysvětlující proměnné vedou ke zvýšení pravděpodobnosti odchodu klienta.

Credit_card a *risk_category* mají velký rozdíl v p-hodnotách tesu ANOVA pro parametrické efekty. Proměnná *credit_card* je významně nižší než 0,05. P-hodnota proměnné *risk_category* se rovná 0,074. V dalším kroku zlepšení modelu jsem se rozhodla tuto proměnnou z modelu odstranit.

6.2 Úpravy modelu

Nyní zlepším původní zobecněný aditivní model tím způsobem, že odstraním vyhlazování proměnné *suma_podil_pr* kubickým splajnem a zároveň nebudu zahrnovat proměnnou *risk_category*, jelikož v původním modelu nebyla její významnost prokázaná.

Proměnná	P-hodnota
s(podil)	<0,001
s(podil_pr)	<0,001
suma_podil_pr	0,192
s(client_age)	<0,001
s(hist_uctu)	<0,001
credit_card	<0,001

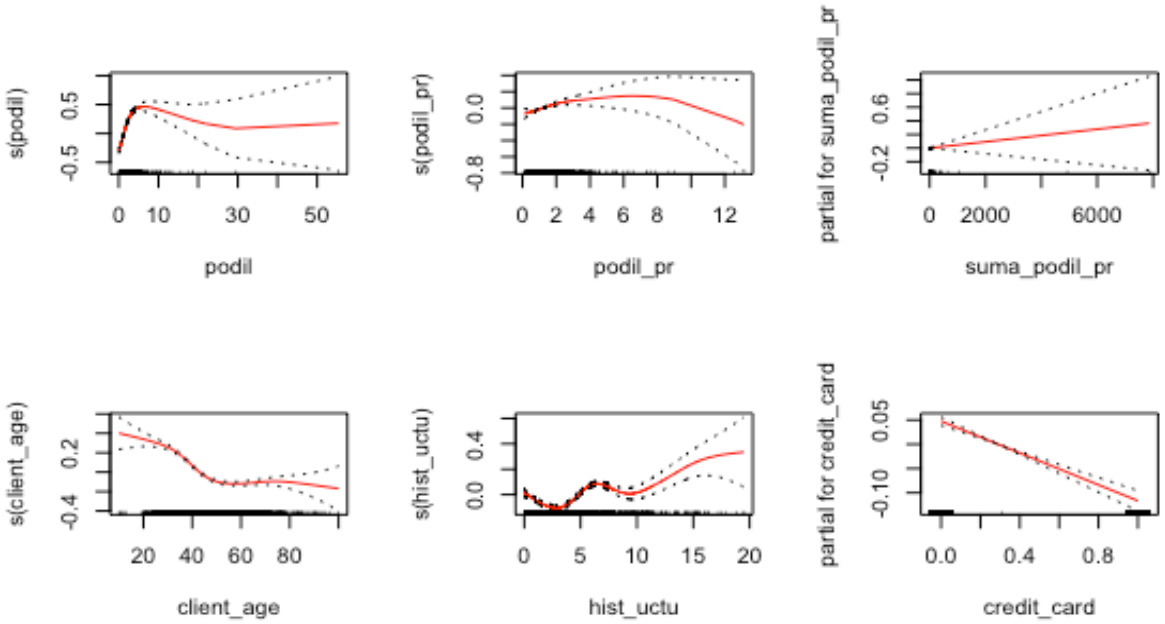
Tabulka 6.3: ANOVA pro parametrické efekty aditivního logistického modelu po úpravě, vlastní výpočty a zpracování

Z výstupu v tabulce 6.3 je vidět, že *suma_podil_pr* s lineárním vztahem závislosti má také vysokou p-hodnotu v testu ANOVA pro parametrické efekty.

Proměnná	P-hodnota
s(podil)	<0,001
s(podil_pr)	0,016
suma_podil_pr	
s(client_age)	<0,001
s(hist_uctu)	<0,001
credit_card	

Tabulka 6.4: ANOVA pro neparametrické efekty aditivního logistického modelu po úpravě, vlastní výpočty a zpracování

V tabulce 6.4 jsou uvedené p-hodnoty testu ANOVA pro neparametrické efekty a je vidět, že jsou všechny proměnné s vyhlazováním kubickým splajnem mají prokázanou nelineární závislost.



Graf 6.7: Závislost na vysvětlujících proměnných po první úpravě modelu, vlastní výpočet a zpracovaní v RStudio

Na grafu 6.7 jsou zobrazené jednotlivé grafy závislosti vysvětlované proměnné na každé z vysvětlujících proměnných z upraveného modelu.

Při dalším zlepšení modelu odstraním proměnnou *suma_podil_pr* a všechny tři modely porovnám mezi sebou pomocí testu ANOVA, který se používá pro hodnocení zobecněných aditivních modelů.

Proměnná	P-hodnota
s(podil)	<0,001
s(podil_pr)	<0,001
s(client_age)	<0,001
s(hist_uctu)	<0,001
credit_card	<0,001

Tabulka 6.5: ANOVA pro parametrické efekty aditivního logistického modelu po finální úpravě, vlastní výpočty a zpracování

Proměnná	P-hodnota
s(podil)	<0,001
s(podil_pr)	0,016
s(client_age)	<0,001
s(hist_uctu)	<0,001
credit_card	<0,001

Tabulka 6.6: ANOVA pro neparametrické efekty aditivního logistického modelu po finální úpravě, vlastní výpočty a zpracování

Nyní je z tabulek 6.5 a 6.6 vidět, že všechny proměnné dávají smysl a jsou významné v rámci daného aditivního logistického modelu.

6.3 Výběr modelu pomocí ANOVA

Dalším krokem je zapotřebí vybrat jeden nejvhodnější model. Pro tento účel se používá test ANOVA.

Je zapotřebí dodat, že se modely do testu ANOVA zapisují v pořadí s rostoucím počtem proměnných. [8] V daném případě bude třetí finální model zapsán jako první, jelikož model obsahuje pouze 5 proměnných. Následně zapíšu zbývající modely.

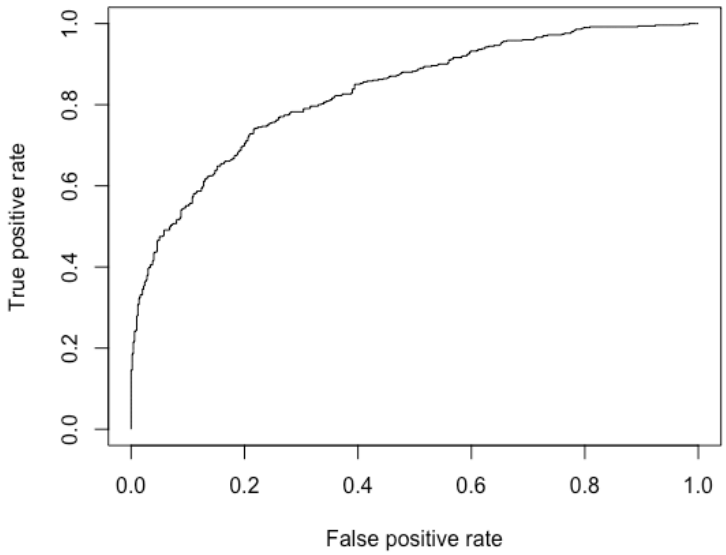
Ve výstupu v tabulce 6.7 jsou uvedené P-hodnoty, podle kterých můžu posoudit, jaký model lepším způsobem popisuje závislost mezi vysvětlovanou a vysvětlujícími proměnnými. Druhý model, který nezahrnuje všechny vysvětlující proměnné, je lepší než finální třetí model podle P-hodnoty, která se rovná 0,347. P-hodnota pro původní model je nižší než P-hodnota druhého modelu, což znamená, že odstranění vyhlazování proměnné *suma_podil_pr* bylo zbytečné. [4]

Model	P-hodnota
model_3	
model_2	0,347
model_1	<0,001

Tabulka 6.7: ANOVA pro hodnocení modelů, vlastní výpočty a zpracování

6.4 Hodnocení modelu

Nyní lze zhodnotit predikční schopnost zobecněného aditivního modelu pomocí aplikace ROC a AUC. Pro hodnocení modelu a jeho schopnosti správně přiřadit respondenta k binární vysvětlované proměnné y vybírám nejlepší model podle testu ANOVA.



Graf 6.8: ROC pro zobecněný aditivní model, vlastní výpočet a zpracovaní v RStudio

Na grafu 6.8 je zobrazena ROC, kterou byl vybraný model schopen odhadnout. Nyní používám funkci AUC pro numerickou interpretaci ROC.

Ve výsledku z výstupu dostávám hodnotu AUC, která se rovná 0,8348942. To znamená, že model byl schopen v 83,48 procentech správně přiřadit respondenta k příslušné vysvětlované proměnné, což také znamená, že výsledný zobecněný aditivní model může v 83,48 % případů předpovědět odchod klienta.

Závěr

Cílem mé bakalářské práce byla analýza chovaní odcházejícího klienta. Odchodem klienta z banky jsem rozuměla snížení do určité míry nebo ukončení využívání služeb ze strany zákazníka. Vhodným nástrojem pro měření aktivity klienta bankovního domu je jeho využívání vlastního bankovního účtu, což jsou příchozí a odchozí transakce, které klient během určitého časového intervalu pravidelně provádí.

Pro analýzu chování odcházejícího a stávajícího klienta jsem zavedla binární vysvětlovanou proměnnou, což se stalo rozhodujícím faktorem pro výběr logistického a aditivního logistického modelu.

V praktické časti mé bakalářské práce jsem sestrojila oba modely. Pro model logistické regrese jsem interpretovala vlivy změn každé vysvětlující proměnné na vysvětlovanou proměnnou pomocí mezních efektů. Také jsem zjistila predikční schopnost sestaveného logistického modelu pomocí AUC, která se rovná 0,8118962. To znamená, že model byl schopen v 81,18 % případů správně přiřadit respondenta k příslušné vysvětlované proměnné.

V případě aditivního logistického modelu se nejvíc používá grafická interpretace závislosti proměnných, jelikož se provádí vyhlazování kubickým splajnem funkce některých proměnných. Z výše uvedeného důvodu jsem do poslední páté kapitoly zahrnula grafy pro ukázku funkce závislosti. Také bylo provedeno hodnocení výsledných modelů pomocí testu ANOVA, na základě kterého byl mnou vybrán jeden model, který byl následně použit pro analýzu predikční schopnosti. V případě aditivního logistického modelu se AUC rovná 0,8348942.

Výsledkem mnou provedené práce je, že na základě AUC je aditivní logistický model schopen lepším způsobem identifikovat odcházejícího klienta, jelikož zobecněné aditivní modely zachycují nelinearitu v závislosti vysvětlované a vysvětlujících proměnných.

Literatura

[1] BUCKINX, Wouter a Dirk VAN DEN POEL. Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research* [online]. 2005, 27 February 2004, **164**(1), 252-268 [cit. 2018-03-29]. DOI: 10.1016/j.ejor.2003.12.010. ISSN 03772217. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S0377221703009184

[2] COUSSEMENT, Kristof, Dries F. BENOIT a Dirk VAN DEN POEL. Improved marketing decision making in a customer churn prediction context using generalized additive models. *Expert Systems with Applications* [online]. 2010, **37**(3), 2132-2143 [cit. 2018-03-29]. DOI: 10.1016/j.eswa.2009.07.029. ISSN 09574174. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S0957417409007325

[3] HASTIE, Trevor, Robert TIBSHIRANI a J. H. FRIEDMAN. *The elements of statistical learning: data mining, inference, and prediction*. Corrected ed. New York: Springer, 2003. ISBN 03-879-5284-5.

[4] HASTIE, Trevor a Robert TIBSHIRANI. Generalized Additive Models. *Statistical Science*. 1986, **1986**(3), 297-318.

[5] JAMES, Gareth, Daniela WITTEN, Trevor HASTIE a Robert TIBSHIRANI. *An introduction to statistical learning: with applications in R* [online]. New York: Springer, 2013 [cit. 2018-03-29]. Springer texts in statistics, 103. ISBN 978-1-4614-7138-7.

[6] LIU, Huimin. Generalized Additive Models. In: *StAtliNK* [online]. Department of Mathematics and Statistics University of Minnesota Duluth, Duluth, MN 55812, 2008 [cit. 2018-05-02]. Dostupné z: http://www.d.umn.edu/math/Technical%20Reports/Technical%20Reports%202007-/TR%202007-2008/TR_2008_8.pdf

[7] VAFEIADIS, T., K.I. DIAMANTARAS, G. SARIGIANNIDIS a K.Ch. CHATZISAVVAS. A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory* [online]. 2015, **2015**(55), 1-9 [cit. 2018-05-20]. Dostupné z: https://doi.org/10.1016/j.simpat.2015.03.003

[8] WITZANY, Jiří. *Credit risk management and modeling*. Ed. 1st. Praha: Oeconomica, 2010. Odborná kniha s vědeckou redakcí. ISBN 978-802-4516-820.)

[9] WOOD, Simon N. *Generalized additive models: an introduction with R*. Boca Raton, Fla., 2006. Texts in statistical science. ISBN 15-848-8474-6.

Internetové zdroje:

[10] 4EK216 Ekonometrie: prezentace k přednáškám. *Ekonometrie - VŠE* [online]. [cit. 2018-05-01]. Dostupné z: https://sites.google.com/site/ekonometrievse/4ek216-ekonometrie

[11] *Modely binární volby v R* [online]. [cit. 2018-05-01]. Dostupné z: http://nb.vse.cz/~zouharj/ekon/ekon_tyden_10.html

[12] Polynomials and Splines: Fractional polynomials example data set. *RPubs* [online]. [cit. 2018-05-02]. Dostupné z: http://rstudio-pubs-static.s3.amazonaws.com/11310_bf69b70c6ce64c1baa19dd89b4539eba.html

[13] ROC curve analysis. *MEDCALC® easy-to-use statistical software* [online]. [cit. 2018-05-02]. Dostupné z: https://www.medcalc.org/manual/roc-curves.php

Seznam grafů, obrázků a tabulek