EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA HOSPODODÁRSKEJ INFORMATIKY

Evidenčné číslo: 103003/I/2024/36145173626850052

APLIKÁCIA MODELOV S UMELOU ZÁVISLOU PREMENNOU

Diplomová práca

Bratislava, 2024 **Bc. Juraj Špánik**

EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA HOSPODODÁRSKEJ INFORMATIKY

APLIKÁCIA MODELOV S UMELOU ZÁVISLOU PREMENNOU

Diplomová práca

Študijný program:	Data science v ekonómii
Študijný odbor:	Ekonómia a manažment
Školiace pracovisko:	Katedra operačného výskumu a ekonometrie
Školiteľ:	Ing. Adriana Lukáčiková, PhD.

Bratislava, 2024 **Bc. Juraj Špánik**

Ekonomická univerzita v Bratislave
Fakulta hospodárskej informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Juraj Špánik
Študijný program: data science v ekonómii (Jednoodborové štúdium, inžiniersky II. st., denná forma)
Študijný odbor: ekonómia a manažment
Typ záverečnej práce: Inžinierska záverečná práca
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Aplikácia modelov s umelou závislou premennou

Anotácia: V práci budú prezentované rôzne funkčné tvary regresných modelov s umelou závislou premennou. Tieto modely slúžia na analýzu rozhodnutí jednotlivcov alebo firiem a pomáhajú vysvetliť, prečo bola zvolená daná alternatíva reakcie. Modely s umelou závislou premennou sa so vzrastajúcou dostupnosťou údajov získaných z rôznych prieskumov, ankiet a pravidelných sčítaní stali základom výskumu vo všetkých sociálnych vedách, ekonómiu nevynímajúc. V rámci ekonomických vied sú najviac spojené s marketingovým výskumom.

Vedúci: Ing. Adriana Lukáčiková, PhD.
Katedra: KOVE FHI - Katedra operačného výskumu a ekonometrie
Vedúci katedry: prof. Mgr. Juraj Pekár, PhD.

Dátum zadania: 25.10.2021

Dátum schválenia: 13.04.2023

prof. Mgr. Juraj Pekár, PhD.
osoba zodpovedná za realizáciu študijného programu

ČESTNÉ VYHLÁSENIE

Čestne vyhlasujem, že celú diplomovú prácu na tému „Aplikácia modelov s umelou závislou premennou," vrátane všetkých jej príloh a obrázkov, som vypracoval samostatne, a to s použitím literatúry uvedenej v priloženom zozname.

V Bratislave, dňa ..... Bc. Juraj Špánik

Poďakovanie

Touto cestou by som sa chcel srdečne poďakovať mojej vedúcej diplomovej práce Ing. Adriane Lukáčikovej, PhD., ktorá ma svojím odborným prístupom a vecnými radami dokázala vždy správne usmerniť pri tvorbe diplomovej práce, taktiež za jej ochotu a trpezlivosť pri konzultovaní problematických častí práce. V neposlednom rade patrí veľká vďaka aj mojej rodine a blízkemu priateľovi Ing. Liborovi Knapcovi, ktorí ma vždy podporovali a stáli pri mne ako počas celého štúdia, tak i v osobnom živote.

ABSTRAKT

ŠPÁNIK, Juraj: *Aplikácia modelov s umelou závislou premennou.* – Ekonomická univerzita v Bratislave. Fakulta hospodárskej informatiky; Katedra operačného výskumu a ekonometrie. – Vedúca záverečnej práce: Ing., Adriana Lukáčiková, PhD. – Bratislava: FHI EU, 2024, 79.s

Diplomová práca sa zaoberá problematikou aplikácie modelov s umelou závislou premennou – dichotomickou. Hlavným cieľom je demonštrovať aplikačné možnosti vybraných nelineárnych pravdepodobnostných modelov na praktickom príklade z oblasti marketingového výskumu. V praktickej časti práce budú za týmto účelom použité údaje, pochádzajúce z marketingovej kampane nemenovanej portugalskej bankovej inštitúcie. Tie obsahujú sociodemografické a finančné charakteristiky klientov, rovnako tak aj informácie o spôsobe vedenia súčasnej a predchádzajúcich kampaní u oslovených klientov. Modely budú aplikované na týchto údajoch, s cieľom klasifikovať klientov do úspešnej alebo neúspešnej marketingovej skupiny (na základe výsledku kampane). Analýzou aplikovaných modelov budú následne identifikované kľúčové determinanty, ktoré sú spojené s možnou úspešnosťou kampane u osloveného klienta, čím sa poskytne hlbší pohľad na efektívne využitie týchto modelov v praxi. Práca je rozčlenená do štyroch kapitol, obsahuje osem obrázkov, sedemnásť tabuliek a päť príloh. Prvá časť práce je venovaná stručnej charakteristike súčasného stavu riešenej problematiky. V rámci nej budú zahrnuté základné teoretické východiská a všeobecné možnosti ekonomickej aplikácie modelov logit a probit. Uvedené budú aj ďalšie alternatívne modely z oblasti binárnej klasifikácie a prehľad literatúry domácich a zahraničných autorov. Druhá kapitola práce sumarizuje hlavný a sekundárny cieľ práce. Obsah tretej kapitoly je venovaný teoretickej definícií aplikovaných modelov, spôsobu odhadu ich parametrov, validácie štatistickej významnosti – parametrov a modelov ako celku, interpretačným možnostiam odhadnutých parametrov. Okrem toho budú teoreticky definované aj metódy, prostredníctvom ktorých sa bude realizovať validácia dosiahnutých predikčných výsledkov. Jadro práce je obsiahnuté v rámci poslednej, t.j štvrtej kapitoly, v ktorom budú získané teoretické poznatky prevedené v praxi, za účelom naplnenia hlavného cieľa práce. Pri spracovaní tejto časti práce bol využitý programovací jazyk R, ktorý poskytuje bohatú škálu programových knižníc, určených na riešenie rôznych druhov štatistických problémov.

Kľúčové slová
Nelineárne pravdepodobnostné modely. Logit model. Probit model. Umelá závislá premenná. Regresné úlohy. Klasifikačné úlohy.

ABSTRACT

ŠPÁNIK, Juraj: *Application of models with dummy dependent variable*. - University of Economics in Bratislava. Faculty of Economic Informatics; Department of Operational research and Econometrics. – Supervisor of the Diploma thesis: Ing., Adriana Lukáčiková, PhD. – Bratislava: FHI EU, 2024, number of pages 79.

The diploma thesis is devoted to the problem of application of models with dummy dependent variable - dichotomous. The main objective is to demonstrate the application possibilities of selected nonlinear probabilistic models on a practical example from the field of marketing research.In the practical part of the thesis, data originating from a marketing campaign of an unnamed Portuguese banking institution will be used for this purpose. These contain socio-demographic and financial characteristics of the clients, as well as information on the way the current and previous campaigns were conducted among the clients approached. Models will be applied to this data, in order to classify clients into successful or unsuccessful marketing groups (based on the outcome of the campaign). The analysis of the applied models will then identify the key determinants that are associated with the potential success of the campaign with the client approached, providing a deeper insight into the effective use of these models in practice. The thesis is divided into four chapters, contains eight figures, seventeen tables and five appendices. The first part of the thesis is devoted to a brief characterization of the current state of the problem. It will include basic theoretical background, general possibilities of economic application of logit and probit models. Other alternative models in the field of binary classification and a literature review of domestic and foreign authors will also be presented.The second chapter of the thesis summarizes the main and secondary objective of the thesis. The content of the third chapter is devoted to the theoretical definition of the applied models, the method of estimation of their parameters, the validation of statistical significance - parameters and models as a whole, and the interpretation possibilities of the estimated parameters. In addition, the theoretical definition of the methods through which the validation of the achieved prediction results of these models will be carried out. The core of the thesis is contained within the last, i.e. the fourth chapter, in which the theoretical knowledge obtained will be transferred in practice, in order to fulfill the main objective of the thesis. In the development of this part of the thesis, the R programming language has been used, which provides a rich variety of programming libraries designed to solve different kinds of statistical problems.

Keywords
Nonlinear probabilistic models. Logit model. Probit model. Dummy dependent variable. Regression problems. Classification problems.

OBSAH

ZOZNAM ILUSTRÁCIÍ A TABULIEK

Zoznam ilustrácií

Zoznam tabuliek

ZOZNAM SKRATIEK A SYMBOLOV

skratka	anglický názov	slovenský názov
AIC	*Akaike information criterion*	Akaikeho informačné kritérium
AUC	*Area under the ROC Curve*	Plocha pod ROC krivkou
BIC	*Bayesian information criterion*	Bayesovo informačné kritérium
FN	*False negative*	Nesprávne klasifikované prípady negatívnej triedy
FP	*False positive*	Nesprávne klasifikované prípady pozitívnej triedy
FPR	*False positive rate*	Synonimu pre mieru výpadku klasifikácie
GLM	*Generalized linear model*	Zovšeobecnený lineárny model
HQC	*Hannan-Quinn information criterion*	Hannan-Quinnovo informačné kritérium
KNN	*K – Nearest Neighbours*	Model K – najbližších susedov
LFS	*Labour Force Survey*	Výberové zisťovanie pracovných síl
LPM	*Linear probability model*	Lineárny pravdepodobnostný model
LR test	*Likelihood ratio test*	Test pomeru vierohodnosti
MLE	*Maximum likelihood estimation*	Metóda maximálnej vierohodnosti
OR	*Odds ratio*	Pomer šancí
SVM	*Support vector machines*	Metóda podporných vektorov
ROC	*Receiver operating characteristic curve*	Krivka operačnej charakteristiky príjmača – ROC krivka
TN	*True negative*	Správne klasifikované prípady negatívnej triedy
TNR	*True negative rate*	Synonimu pre mieru špecifickosti klasifikácie
TP	*True positive*	Správne klasifikované prípady pozitívnej triedy
TPR	*True positive rate*	Synonimu pre mieru senzitívnosti klasifikácie

SLOVNÍK TERMÍNOV

Premenná dichotomická (binárna) - druh kvalitatívnej premennej, ktorej hodnota môže nadobúdať len jednu z dvoch alternatív, napr. pohlavie – muž, žena. (Klein, 2020)

Premenná diskrétna – druh kvantitatívnej premennej, ktorá disponuje konečným, ale spočítateľným počtom obmien. Všetky obmeny je možné číslovať prirodzeným číslom. Napr. počet detí v rodine, alebo počet dosiahnutých bodov v teste. (Štatistický úrad Slovenskej republiky [ŠÚ SR], 2020)

Premenná kvalitatívna (resp. kategoriálna) – štatistický znak, ktorého hodnota je slovne vyjadrená vlastnosťou štatistickej jednotky. Hodnoty kvalitatívnych znakov bývajú v praxi prekódované číselnými hodnotami. Ďalej sa kvalitatívne členia na nominálne a ordinálne alebo dichotomické (binárne) a multinomiálne. (ŠÚ SR, 2020)

Premenná kvantitatívna (kardinálna, číselná) – štatistický znak, ktorého hodnota je vyjadrená reálnym číslom. O ich hodnotách je možné povedať, či sú rôzne alebo rovnaké, je možné ich usporiadať do poradia a taktiež určiť, o koľko je jedna hodnota väčšia ako druhá. Tieto štatistické znaky majú aj nulovú hodnotu a meraciu jednotku. Ďalej je ich možné členiť na diskrétne a spojité alebo intervalové a racionálne. (ŠÚ SR, 2020)

Premenná multinomiálna – druh kvalitatívnej premennej, ktorej hodnota môže nadobúdať iba jednu z viacerých alternatív (viac ako dvoch), napr. farba očí – hnedé, modré, zelené, atď. (Klein, 2020)

Premenná nominálna – druh kvalitatívnej premennej, ktorej hodnoty možno pomenovať, ale nemožno ich zoradiť do poradia. Možno o nich povedať, či sú rôzne alebo sa rovnajú, napr. pohlavie, farba očí, štátna príslušnosť, atď. (ŠÚ SR, 2020)

Premenná ordinálna - druh kvalitatívnej premennej, ktorej hodnoty možno prirodzene usporiadať do poradia, avšak nie je možné určiť, o koľko je jedna hodnota väčšia ako druhá. Napr. medaila - zlatá, strieborná, bronzová, hodnosť v armáde, kvalitatívne hodnotenie študenta -výborný, veľmi dobrý, dobrý, nevyhovel môžeme vyjadriť číselným hodnotením. (ŠÚ SR, 2020)

Premenná spojitá - druh kvantitatívnej premennej, ktorá môže nadobúdať rôznu číselnú hodnotu z istého intervalu, napr. telesná výška, príjem a mnohé iné. (ŠÚ SR, 2020)

ÚVOD

Diplomová práce je venovaná problematike aplikácie modelov s umelou závislou premennou – dichotomickou. Tieto modely predstavujú spoľahlivý štatistický nástroj na modelovanie situácií, ktorých experimentálnym cieľom je výber jednej z dvoch prípustných alternatív. Ich pomocou je možné odhaliť kľúčové determinanty, ktoré dokážu zdôvodniť uskutočnenie tohto výberu. Vhodne vyvinutý model dokáže prostredníctvom odhalenia vzťahov v historických údajoch presne klasifikovať nové pozorovania do jednej zo skúmaných kategórií. Uplatnenie týchto modelov možno nájsť naprieč všetkými ekonomickými sektormi. V súčasnosti existuje viacero prístupov, ako sa postaviť k riešeniu tejto problematiky. Vzhľadom na veľkú dostupnosť údajov, ktoré so sebou priniesol celosvetový trend digitalizácie procesov, ruka v ruke prekvitá aj sféra strojového učenia, ktorá je v posledných desaťročiach na vzostupe. Tá zároveň ponúka širokú paletu rôznych klasifikačných nástrojov, využiteľných na riešenie tohto typu úloh. Práca sa primárne zameriava na praktické využitie regresných nelineárnych pravdepodobnostných modelov, vhodných na modelovanie pravdepodobnosti príslušnosti pozorovania do jednej z dvoch alternatív. Jadro práce je venované prezentácií aplikačnej možnosti logitovej a probitovej regresie na praktickom príklade z oblasti marketingového výskumu. Cieľom bude klasifikovať klientov zapojených v marketingovej kampani do dvoch kategórií – úspešnej a neúspešnej skupiny marketingovej kampane. To na základe ich sociodemografických charakteristík, indikátorov solventnosti a spôsobu vedenia kampane. V neposlednom rade bude cieľom identifikovať kľúčové determinanty klientových charakteristík, ktoré sú spojené so zvýšením pravdepodobnosti úspechu kampane u osloveného klienta. Diplomová práca je rozdelená do štyroch kapitol.

Prvá kapitola práce pojednáva o súčasnom stave riešenej problematiky. Za týmto účelom bola rozčlenená do troch samostatných podkapitol. V prvej podkapitole práce možno nájsť teoretické východiská riešenej problematiky, s cieľom priblížiť charakter témy práce, uviesť rozdiel medzi úlohami regresie a klasifikácie, a prezentovať všeobecné možnosti ekonomického využitia modelov logitovej a probitovej regresie. V rámci druhej podkapitoly sa bude klásť dôraz na prezentáciu alternatívnych možností, t.j. modelov ktoré je možné využiť za týmto účelom, využívaných vo sfére strojového učenia – naivný bayesovský klasifikátor, model k–najbližších susedov, metóda podporných vektorov, model rozhodovacieho stromu a náhodného lesa. Posledná časť kapitoly je venovaná prehľadu

súčasnej literatúry slovenských i zahraničných autorov, ktorí využili nelineárne pravdepodobnostné modely pri modelovaní rôznych ekonomických problémov, napr. vo sfére predikcie rizika bankrotu ekonomických subjektov, poisťovníctva a bankovníctva. Dôraz bol kladený na získanie prehľadu metód, ktoré sa v súčasnosti využívajú v tejto oblasti predikčného modelovania a môžu byť aplikovateľné pre dosiahnutie hlavného cieľa diplomovej práce.

Obsah druhej kapitoly je venovaný definovaniu primárneho a sekundárneho cieľa diplomovej práce, a zároveň načŕta ich segmentáciu do menších parciálnych cieľov, ktoré boli stanovené na dosiahnutie čo možno najlepších výsledkov práce. Kapitola slúži aj ako stručná sumarizácia kľúčových bodov, ktoré boli následne podrobnejšie rozpracované v jednotlivých častiach práce.

Metódam, ktoré boli využité pri spracovaní praktickej časti práce sa venuje tretia kapitola. Tá pozostáva z troch hlavných podkapitol. V prvej podkapitole budú detailne definované modely logistickej a probitovej regresie, a zároveň bude načrtnutá možnosť modifikácie binárneho logitového modelu na riešenie situácií, v ktorých závislá premenná disponuje viac ako dvomi alternatívami odozvy. Druhá podkapitola je venovaná definícií metódy odhadu parametrov, testovania ich štatistickej významnosti (buď samostatne alebo modelu ako celku) a odvodeniu vzťahov marginálnych efektov, prostredníctvom ktorých je možné priamo interpretovať odhadnuté parametre modelu. Posledná podkapitola je venovaná definícií metód, určených na výber najvhodnejšieho modelu, ktorý bude použitý na predikciu pravdepodobnosti a následnú klasifikáciu klientov. V rámci podkapitoly budú predstavené aj použité metódy, určené na validáciu predikčnej schopnosti testovaných modelov.

Posledná, v poradí štvrtá kapitola predstavuje jadro spracovanej diplomovej práce. Prakticky budú aplikované poznatky danej problematiky, opísané v predošlých podkapitolách. Cieľom bude prostredníctvom využitia logistickej a probitovej regresie klasifikovať klientov do úspešnej alebo neúspešnej marketingovej skupiny a následne segmentovať klientelu zapojenú v kampani do skupín na základe ich predikovanej pravdepodobnosti možného úspechu. Takáto segmentácia by mohla zefektívniť budúce výsledky kampane v prípade, kedy by sa kampaň opakovala na rovnakej vzorke klientov. Ďalším cieľom je objasniť kľúčové determinanty úspechu kampane u klienta, ktoré môžu zefektívniť cielenie marketingovej kampane na špecificky vybranú novú vzorku klientov. Pri spracovaní práce bol využitý programovací jazyk R, ktorý na rozdiel napr. od Pythonu

disponuje rozsiahlejším počtom programových knižníc, priamo určených na riešenie širokého spektra rôznych štatistických problémov. Kapitola je segmentovaná do troch samostatných podkapitol. Údaje a spôsob ich spracovania je obsiahnutý v prvej podkapitole. Explorácia vstupných údajov a preverenie vzťahov medzi nezávislými premennými tvorí obsah druhej podkapitoly. Aplikačná časť je zahrnutá v rámci poslednej podkapitoly. V nej bude detailne zachytený spôsob rozdelenia spracovaného súboru údajov, spôsob výberu vhodných nezávislých premenných a následný výber modelov, vykonanie predikcií a validácia predikčnej schopnosti modelov. Dosiahnuté výsledky budú sumarizované v rámci poslednej časti tejto podkapitoly.

1 SÚČASNÝ STAV RIEŠENEJ PROBLEMATIKY

Obsah prvej kapitoly diplomovej práce je venovaný súčasnému stavu riešenej problematiky, t.j. aplikácií regresných modelov s umelou závislou premennou – binárnou. Pre systematické nahliadnutie do problematiky, je táto kapitola rozčlenená do troch samostatných častí. V rámci podkapitoly 1.1 budú stručne zhrnuté teoretické východiská a kľúčové pojmy riešenej problematiky. Vysvetlené budú základné náležitosti, týkajúce sa témy diplomovej práce. Podkapitola 1.2 sa venuje prehľadu ďalších vhodných modelov zo sféry strojového učenia, ktoré je možné využiť pri riešení tohto typu úloh. Podkapitola 1.3 je venovaná prehľadu súčasnej literatúry autorov, ktorí sa zaoberali problematikou využitia regresných pravdepodobnostných modelov pri modelovaní s umelou závislou premennou.

1.1 Teoretické východiská riešenej problematiky

Matematický model možno vo všeobecnosti charakterizovať ako abstraktný model, využívajúci matematické formulácie na zachytenie správania sa určitého systému. Pod pojmom systém si možno predstaviť napr. reálny objekt, projekt, proces alebo komplex procesov, resp. problémov a mnohé iné. (Hřebíček & Škrdla, 2006)

Pojem prediktívne modelovanie predstavuje proces vytvárania matematického modelu alebo nástroja, ktorý presne predikuje, resp. predpovedá skúmanú náhodnú premennú (označenie - Y), na základe množiny vysvetľujúcich premenných (označenie atribútov - X_1, X_2, ..., X_p). Vysvetľovaná premenná býva v literatúre uvádzaná aj pod pojmom závislá premenná. Na druhú stranu vysvetľujúcu premennú možno analogicky označiť ako nezávislú. Kontext tohto spôsobu označovania premenných vyplýva z toho, že nezávislá premenná predstavuje faktor podrobujúci sa zmenám, s cieľom zistiť, ako tieto zmeny ovplyvňujú závislú premennú. (García Portugués, 2023; Ali & Younas, 2021; Bačíková & Janovská, 2018)

Modely s umelou závislou premennou predstavujú taký typ matematických modelov, ktorých závislá premenná predstavuje kategórie, resp. alternatívy, ktoré sa zvyknú kódovať hodnotami 0 a 1, na základe príslušnosti premennej ku konkrétnej alternatíve. Vhodným príkladom ich použitia je modelovanie situácií s cieľom predikovať, napr. či spotrebiteľ zakúpi daný produkt alebo nie, či jednotlivec participuje na trhu práce alebo nie, situácie s odpoveďami typu áno – nie, úspešný – neúspešný a mnohé iné. (Güneri & Durmuş, 2020)

1.1.1 Rozdiely v modeloch regresných a klasifikačných úloh

Aplikáciou matematických modelov je možné častokrát získať nie len zrozumiteľný opis relevantných faktorov, ovplyvňujúcich skúmaný systém – objekt skúmania, ale taktiež dokážu odhaliť a vysvetliť významné vzťahy medzi jednotlivými atribútmi, ktoré ovplyvňujú takýto systém. Využitím metód **regresnej analýzy** sa ponúka možnosť analyticky vyjadriť vzťah medzi skúmanou premennou (vysvetľovanou premennou, resp. premennou odozvy) a množinou vysvetľujúcich premenných (regresorov), prostredníctvom využitia regresnej funkcie. Inými slovami, funkcia opisujúca závislosť vysvetľovanej premennej (Y) od regresorov ($X_1, X_2, ..., X_p$) sa nazýva regresná funkcia. (Hřebíček & Škrdla, 2006; Neubauer et al., 2021)

Príkladom najjednoduchšieho a zároveň najpoužívanejšieho modelu regresnej analýzy je model lineárnej regresie, ktorý sa využíva na opis lineárneho vzťahu medzi závislou premennou (Y – cieľ predikcie) a jednou, prípadne viacerými nezávislými premennými (regresormi). Inými slovami, predpokladá medzi nimi lineárny vzťah. Možným prípadom použitia modelu lineárnej regresie v ekonomickej sfére môže byť predikcia spotreby zdrojov domácností, podnikov, poprípade celých štátov. Rovnako tak je uplatniteľná pri výpočte množstva exportu a importu komodít pre skúmané krajiny alebo monitorovanie vývoja dopytu a ponuky pracovného trhu. Z uvedeného vyplýva, že model lineárnej regresie sa využíva pri takom modelovaní, kde závislá premenná predstavuje kvantitatívnu spojitú premennú (viď slovník termínov). Z tohto dôvodu, lineárnu regresiu nie je vhodné použiť pri modelovaní s umelou závislou premennou. Tá skúma príslušnosť závislej premennej ku konkrétnej alternatíve, preto je z pravidla definovaná kategoriálnou premennou odozvy (viď slovník termínov). (Gupta et al., 2017; Hope, 2020, Šikyňa, 2019; Machová, 2016)

V prípade modelovania s kategoriálnou závislou premennou, ktorej hodnoty môžu nadobúdať iba dve prípustné alternatívy, bývajú v literatúre často uvádzané pravdepodobnostné regresné modely, ako napr. lineárny pravdepodobnostný model (skrátene LPM) a nelineárne pravdepodobnostné modely - logitový a probitový regresný model. Vo všeobecnosti sa však neodporúča využívať LPM. Problém pri tomto modeli nastáva v tom, že odhadovaná pravdepodobnosť prostredníctvom tohoto modelu sa nachádza mimo rozsahu hodnôt 0 – 1. Na druhú stranu modely, ktoré majú široké uplatnenie v úlohách modelovania s umelou závislou premennou – dichotomickou (viď slovník termínov), predstavujú už spomínaný logitový a probitový model. Tieto modely patria do triedy tzv.

zovšeobecnených lineárnych modelov (skrátene GLM – z angl. „Generalized linear models“), medzi ktoré možno zaradiť už spomínaný model lineárnej regresie, modely analýzy rozptylu (pre spojitú závislú premennú), regresné modely s kategoriálnou závislou premennou, atď. (García Portugués, 2023; Coss, 2015)

Klasifikačné úlohy predstavujú taký typ úloh, pri ktorých sa vynakladá snaha nájsť funkciu, mapujúcu sadu pozorovaní do množiny samostatných tried, alternatív alebo kategórií. Táto funkcia býva označovaná ako klasifikátor a neskôr slúži na klasifikáciu nových prípadov. Z povahy tohto druhu úloh možno dedukovať, že závislá premenná v tomto type úloh je kategoriálna. Pokiaľ sa hovorí o úlohách binárnej klasifikácie, klasifikátor mapuje pozorovania do dvoch preddefinovaných kategórií, resp. alternatív. Proces modelovania klasifikátora sa označuje aj ako trénovanie a algoritmus na to použitý sa nazýva učiaci sa algoritmus. Trénovanie klasifikátora prebieha na množine údajov, ktorá býva označovaná aj ako trénujúci súbor, resp. trénujúca množina údajov. Inak povedané, na tejto množine údajov sa hľadajú jednotlivé vzťahy alebo vzorce medzi jednotlivými atribútmi. Neskôr bude na základe týchto vzťahov klasifikátor triediť nové prípady do jednotlivých tried. Množina nových prípadov sa označuje aj ako testovacia množina údajov, resp. testovací súbor údajov. (Mozos, 2010; Wadi, 2021)

Sféra strojového učenia zaoberajúca sa úlohami klasifikácie a regresie sa súhrnne nazýva učenie s učiteľom (z angl. „Supervised learning“). Tento druh strojového učenia je špecifický tým, že v procese trénovania modelu využíva označené údaje, t.j. hodnota závislej premennej je dopredu známa. Takto natrénovaný model následne na novej údajovej množine (neoznačenej) sa snaží predikovať hodnotu závislej premennej. (Mark & Kaye, 2015)

1.1.2 Všeobecné možnosti aplikácie nelineárnych pravdepodobnostných modelov

V minulej podkapitole už bolo spomenuté, že regresná analýza využíva regresnú funkciu na opísanie závislosti skúmaných javov. To na základe monitorovania týchto javov vo špecifickom časovom rozmedzí – skúmanom období. V budúcnosti je možné vyskúmanú závislosť použiť na predikciu neznámej hodnoty závislej premennej, ktorá reprezentuje požadovaný skúmaný jav. Je však vhodné uviesť, že vo sfére strojového učenia možno nájsť uplatnenie regresnej analýzy rovnako v zmysle predikcie, tak i klasifikácie skúmaného javu do jednotlivých kategórií, resp. tried. Jedným z najčastejších modelov využitých za týmto

účelom, je práve model logistickej regresie – v literatúre predstavuje synonymické označenie pre logitový model. (Machová, 2016)

Všeobecné ekonomické možnosti využitia aplikácie logit modelu, ktorý spolu s modelom probit patria do skupiny nelineárnych pravdepodobnostných modelov, možno zhrnúť do troch základných skupín:

1. Aplikácia, ktorej cieľ predstavuje **určenie závislosti** medzi skúmaným javom (Y) a jeho determinantmi ($X_1, X_2, ..., X_p$) – slúži na overovanie akademických hypotéz. Napríklad overenie predpokladu, či klient s nepriaznivým rizikovým profilom disponuje náklonnosťou k otázke dodatočného zdravotného poistenia. (Lukáčik, 2008)
2. Aplikácia, ktorej cieľom je využiť odhad pravdepodobnosti pre **identifikáciu a následnú segmentáciu do rôznych kategórií.** Dôležitou úlohou je stanovenie kritických hodnôt, ktoré pomôžu segmentovať jednotlivé kategórie. V prípade, že výber kategórie ovplyvňuje skúmanú entitu, možno hovoriť aj o podpore rozhodovania. Príkladom z oblasti marketingového výskumu môže byť segmentácia potenciálnych zákazníkov, na základe ich záujmu o konkrétny produkt. Ďalším príkladom je identifikácia jednotlivcov, ktorí sa uchádzajú o úver a môžu mať ťažkosti so splácaním záväzkov v budúcom období. V neposlednom rade možno model použiť na odhalenie spoločností, ktoré sa nachádzajú na hranici bankrotu. (Lukáčik, 2008)
3. **Aplikácia, ktorej cieľom je predikcia.** V tomto zmysle predikcia konkrétnej pravdepodobnostnej hodnoty nie je veľmi relevantná, pretože dôležitá je možnosť nastatia, resp. nenastatia skúmaného javu. Ten je reprezentovaný pravdepodobnosťou 0 – nenastane a 1 - nastane. Najčastejším a najjednoduchším spôsobom transformácie je porovnanie vypočítanej pravdepodobnosti s hodnotou 0,5. Ak je pravdepodobnosť menšia ako táto hodnota, predpokladá sa, že jav nenastane. Analogicky sa predpokladá, že jav nastane, ak je vypočítaná pravdepodobnosť väčšia ako táto hodnota. Predikčnú schopnosť modelov možno použiť pri analýze:
   - Rozhodnutia jednotlivcov – napr. pri politickom rozhodnutí (pôjde voliť/nepôjde voliť), voľbe poistenia (poistí sa/nepoistí sa), atď.
   - Rozhodnutia skupín jednotlivcov (domácností) – napr. investičné rozhodnutia (zakúpenie/nezakúpenie nehnuteľnosti), schopnosti mobility (ochota sťahovať/nesťahovať sa), vlastnenie tovarov dlhodobej spotreby (používanie/nepoužívanie auta) a iné.

- Rozhodovanie v rámci firemných procesov – napr. rozloženie zdrojov poprípade procesov (zvoliť/nezvoliť outsourcing), atď.
- Rozhodovanie politických subjektov o konkrétnom legislatívnom opatrení na základe prieskumov verejnej mienky (schváliť/neschváliť opatrenie).
- A mnohé iné. (Lukáčik, 2008)

1.2 Prehľad alternatívnych modelov v úlohách binárnej klasifikácie

Sféra strojového učenia s učiteľom ponúka viacero možností, prostredníctvom ktorých je možné pristupovať k riešeniu úloh, založených na predikcii kategoriálnej závislej premennej. V rámci tejto podkapitoly bude predstavených päť alternatívnych modelov, prostredníctvom ktorých možno v súčasnosti riešiť tento typ úloh.

Naivný bayesovský klasifikátor je jednoduchá, ale efektívna technika pre klasifikáciu a zhlukovanie, založená na Bayesovej vete. Model predpokladá, že všetky nezávislé premenné sú medzi sebou nezávislé, čo často nezodpovedá realite, odkiaľ plynie jeho označenie "naivný". Táto metóda sa často využíva v situáciách s vysokou dimenzionalitou údajov, ako je klasifikácia textu, vrátane filtrácie spamu. Klasifikátor vypočítava podmienenú pravdepodobnosť príslušnosti objektu k určitej triede na základe pravdepodobnosti jeho atribútov a triedy, ktorú predstavuje. Objekt je potom klasifikovaný do triedy, pre ktorú dosahuje táto podmienená pravdepodobnosť maximálnu hodnotu. V procese učenia sa určujú podmienené pravdepodobnosti všetkých možných hodnôt závislej premennej, čím sa umožňuje efektívna klasifikácia nových objektov. (Mahesh, 2020; Mohamed et al., 2022; Hendl, 2021)

Model K-najbližších susedov (skrátene KNN, z angl. „K - Nearest Neighbours“**)** je jednoduchá neparametrická metóda strojového učenia vhodná pre klasifikáciu a regresiu, ktorá klasifikuje údajové body na základe tried najbližších susedov. Algoritmus určí triedu údajového bodu podľa prevládajúcej triedy medzi k najbližšími susedmi. Počet susedov – k sa volí v závislosti od typu riešenej úlohy, pričom v binárnej klasifikácii je preferované ich nepárne číslo, aby sa predišlo nerozhodným situáciám. Na meranie vzdialenosti medzi údajovými bodmi a ich susedmi sa najčastejšie používa euklidovská vzdialenosť. V praxi sa možno stretnúť aj s využitím metrík, ako napr. minkowského vzdialenosť, Manhattan, korelácia, $\chi 2$ a pod., čo umožňuje flexibilitu pri aplikácii metódy na rôzne typy údajov. (Alzubi et. al, 2018; Mohamed et al., 2022; Lopez-Bernal et al., 2021)

Metóda podporných vektorov (z angl. „Support Vector Machines“, skrátene SVM) je technika strojového učenia používaná pre klasifikáciu a regresiu, ktorá sa vyvinula v 90. rokoch 20. storočia. Táto metóda sa zameriava na nájdenie optimálnej nadroviny, ktorá rozdeľuje údaje do dvoch tried s cieľom maximalizovať minimálnu vzdialenosť („margin“) medzi údajovými bodmi a nadrovinou. Prostredníctvom toho sa zaistí, že každá trieda leží vo svojej definovanej polrovine. Hlavnými prvkami pri klasifikácii sú podporné vektory, ktoré predstavujú údajové body ležiace najbližšie k nadrovine. SVM je obľúbená vďaka svojej efektívnosti pri práci s malými trénujúcimi súbormi údajov a schopnosti zvládať vysoký počet údajových dimenzií, pričom je možné ju adaptovať na riešenie multinomických problémov. (Mohamed et al., 2022; Hendl, 2021; Laura & Santi, 2017)

Rozhodovací strom je model strojového učenia používaný v klasifikácii aj regresii, ktorý zachytáva rozhodovacie procesy prostredníctvom grafu stromovej štruktúry. Tento model zahŕňa tri hlavné komponenty: rozhodovacie uzly (testujúce hodnoty atribútov), hrany (reprezentujúce výsledky testov) vedúce k ďalším uzlom a listové uzly (určujúce predikciu cieľovej premennej). Strom začína koreňovým uzlom obsahujúcim celý súbor údajov, ktorý sa postupne delí na homogénnejšie podskupiny podľa špecifických atribútov, pričom rozhodovacie kritériá sú zoradené podľa ich dôležitosti. Cieľom je vytvoriť strom, ktorý efektívne klasifikuje vstupné údaje do príslušných kategórií. Na výber atribútov pre vetvenie sa používajú kritériá ako entropia, informačný zisk, Gini index, alebo $\chi 2$. Na ich základe sa následne vyberie najinformatívnejší atribút pre daný uzol. (Mahesh, 2020; Mohamed, 2022; Hendl, 2021)

Pri modelovaní rozhodovacích stromov často dochádza k chybe preučenia, keď model obsahuje príliš veľa uzlov a pracuje s rozsiahlym množstvom atribútov. Toto vedie k dokonalej klasifikácii trénujúcej množiny údajov, ale zhoršenej prediktívnej schopnosti na nových údajoch. Ako riešenie sa ponúka **model náhodného lesa**, ktorý patrí medzi tzv. "ensemble" metódy. Tieto metódy kombinujú viaceré klasifikátory (v tomto prípade rozhodovacie stromy) prostredníctvom agregačných metód, ako napr. hromadné hlasovanie, čím redukujú problém preučenia. Náhodný les vytvára viaceré rozhodovacie stromy na základe náhodne vybraných podvzoriek údajov a nezávislých premenných. Výsledná predikcia je založená na hlasovaní alebo priemerovaní výstupov týchto stromov, čo zlepšuje predikčnú schopnosť modelu na neznámych údajoch. (Hendl, 2021)

1.3 Prehľad literatúry

V rámci tejto podkapitoly bude uvedený prehľad literatúry domácich a zahraničných autorov, ktorí sa venovali aplikácií nelineárnych pravdepodobnostných modelov v kontexte rôznych ekonomických problémov. Cieľom je načrtnutie aplikačných možnosti modelov logistickej a probitovej regresie na konkrétnych príkladoch z praxe, a tým predstaviť poznanie v súčasnosti využívanej metodológie v tomto type úloh.

Prvým príkladom využitia modelu logit je práca kolektívu autorov Ďurica, Valášková, Janošková s názvom *„Logit business failure prediction in V4 countries,"* ktorá vyšla v jedenástom čísle „Engineering Management in Production and Services" z decembra 2019. Štúdia bola zameraná na vytvorenie logitového modelu, ktorého cieľom bola predikcia obchodného neúspechu spoločností v krajinách V4. Autori si vybrali tento model kvôli jeho spoľahlivejším výsledkom v porovnaní s diskriminačnou analýzou. Preferujú ho aj pred niektorými modelmi strojového učenia, z dôvodu lepšej interpretovateľnosti dosiahnutých výsledkov, čo je pre účely štúdie vhodnejšia alternatíva. Model bol úspešne aplikovaný na rozsiahlom súbore údajov, obsahujúcom finančné ukazovatele pre viac ako 173 000 spoločností v sledovanom období rokov 2016 – 2017. Za účelom výberu nezávislých premenných modelu bola zvolená procedúra „stepwise forward selcetion," ktorá iteratívne pridáva do modelu premenné, na základe ich štatistickej významnosti. Výsledný model zahŕňal 17 štatisticky významných atribútov, ako napr. sídlo firmy (krajina), štrukturálna veľkosť spoločnosti, pracovný kapitál a ďalšie. Za účelom testovania štatistickej významnosti parametrov modelu bol zvolený Waldov test a štatistická významnosť modelu ako celku bola založená na teste pomeru vierohodnosti („likelihood ratio test"). Zhodnotenie výsledkov predikcií bolo založené na matici zámen, ukazovateľovi správnosti klasifikácie („accuracy") a na ROC-AUC analýze. Model disponoval 88,1 % správnosťou všetkých predikcií, pričom hodnota AUC (0,939) naznačovala jeho takmer dokonalú klasifikačnú schopnosť.

Vytváranie predikčných modelov bankrotu slovenských spoločností opísal vo svojej práci z decembra 2017 - *„Logit and Probit application for the prediction of bankruptcy in Slovak companies"* kolektív autorov Klieštik a Kováčová. Článok bol uverejnený v dvanástom čísle štvrťročného žurnálu ekonomiky a hospodárskej politiky „Equilibrium". Na rozdiel od Ďuricu, Valáškovej a Janoškovej sa autori venovali nie len vytváraniu modelu logistickej regresie, ale aj probitovej regresie. Autori naznačujú, že napriek určitej podobnosti oboch modelov, existujú značné rozdiely v procese ich tvorby ako aj

v dosiahnutých výsledkoch. Údaje, využité autormi v práci za účelom konštrukcie modelov, boli získané z výročných finančných správ slovenských spoločností (Register finančných výkazov, Ministerstvo financií Slovenskej republiky) z roku 2015. Použitá údajová vzorka pozostávala z 500 spoločností v „defaulte“ a 500 spoločností bez „defaultu,“ pričom boli náhodne vybrané zo základného súboru údajov, ktorý obsahoval údaje o 9 209 spoločnostiach. To bez ohľadu na odvetvie, veľkosť alebo právnu formu spoločností. Na rozdiel od Ďuricu et al., bola použitá procedúra iteratívnej spätnej eliminácie premenných modelu „backward stepwise conditional method“. Prispôsobenie modelu údajom bolo v modeli probit overené prostredníctvom McFaddenovho nepravého koeficientu determinácie ($R^2_{McFadden}$). To dosahovalo hodnotu 87,59 %, čo naznačuje pomerne vhodné prispôsobenie modelu údajom. Pri zhodnocovaní výsledkov klasifikácie bola opäť využitá metrika správnosti, ROC-AUC analýza a v tomto prípade boli zohľadnené aj metriky senzitívnosti a špecifickosti klasifikácie. Správnosť klasifikácie sa ukázala takmer totožná v prípade oboch modelov, pričom na trénovacej množine dosahovala približne 97 % a 86 % na množine testovacej. Model logit zahŕňal osem atribútov, pričom ako najvýznamnejšie sa ukázali parametre premenných: pomer vlastného imania k aktívam, celkový dlh ku celkovým aktívam, pomer obrátenej hodnoty likvidity, čistý výnos z celkových príjmov. Model probit bol omnoho komplexnejší, obsahoval štrnásť atribútov. Štatistická významnosť parametrov sa ukázala v drvivej väčšine ako nevýznamná, avšak autori sa rozhodli pre ich ponechanie v modeli, z dôvodu zachovania jeho komplexnosti. Ako významné parametre probitového modelu sa ukázali: čistý výnos z celkových príjmov, pomer bežnej likvidity, pomer obrátenej hodnoty likvidity, celkový dlh ku celkovým aktívam, pomer nerozdeleného zisku k celkovým aktívam a pomer súčasných aktív k celkovému príjmu.

Doteraz boli rozobraté práce od slovenských autorov, ktorí sa venovali využitiu týchto modelov, pri modelovaní rizika bankrotu rôznych ekonomických subjektov. Znalosti z údajov, ktoré tieto modely pomáhajú odhaliť, sú užitočné aj v sektoroch, akými sú poisťovníctvo, bankovníctvo, marketing a mnohé iné.

Vo sfére poisťovníctva možno spomenúť prácu kolektívu autorov Lobos, Viviani, Schnettler, Muñoz a Reyes, s názvom *„Predicting probability to purchase insurance contracts in the Chilean wine industry: a logit and probit comparative analysis“* z roku 2012. Cieľom štúdie bola identifikácia významných premenných, ktoré ovplyvňujú pravdepodobnosť využitia poistných zmlúv v odvetví vinohradníctva v Čile. Metodologicky sa práca opierala o aplikáciu logistickej a probitovej regresie, na populácií 84 čilských

vinohradníckych spoločností. Údaje o spoločnostiach boli získané prostredníctvom dotazníkového prieskumu, ktorý zahŕňal otázky týkajúce sa rizík v čilskom vinárstve, vrátane otázok o používaní a názoroch na poistenie a ďalších determinantov rizika. Vhodnosť prispôsobenia modelov údajom a ich následné porovnanie bolo realizované prostredníctvom McFaddenovho nepravého koeficientu determinácie ($R^2_{McFadden}$), Akaikeho informačného kritéria (skrátene AIC), Bayesovho informačného kritéria (skrátene BIC) a Hannan-Quinnovho informačného kritéria (skrátene HQC). Na základe dosiahnutých výsledkov, autori vyhodnotili probitový model z pohľadu jeho výkonnostných charakteristík v tomto prípade ako efektívnejší, v porovnaní s logitovým modelom. To na základe už spomínaných štatistik, pri ktorých dosiahol probitový model najvyššiu hodnotu v štatistike $R^2_{McFadden}$ oproti logitovému modelu. Zároveň tak hodnota informačných kritérií AIC, BIC a HQC dosahovala nižšie hodnoty práve v modeli probit. Zhodnotenie výsledkov klasifikácie pre logitové a probitové modely ukazujú globálnu správnosť predikcie 83,33 %, pričom senzitívnosť modelov je 85,71 % a špecifickosť 81,10 % pri prahovej hodnote pravdepodobnosti 0,50. To naznačuje mierne vyššiu úspešnosť v správnosti klasifikácie firiem, ktoré nemajú o poistenie záujem. Taktiež ROC-AUC analýza, pri ktorej hodnota AUC dosahovala 83,10 %, naznačuje vysokú pravdepodobnosť správnej klasifikácie pozitívnych a negatívnych prípadov. Štúdia identifikovala niekoľko kľúčových faktorov, ktoré pozitívne ovplyvňujú pravdepodobnosť uzatvorenia poistných zmlúv v čilskom vinárskom priemysle. Prvým je lepší prístup k informáciám o trhu, ktorý zvyšuje pravdepodobnosť uzavretie poistných zmlúv. Ďalším faktorom je veľkosť vinice – vinohrady s väčšou rozlohou majú vyššiu pravdepodobnosť na uzavretie poistenia.

Vo sfére komerčného bankovníctva možno nájsť široké uplatnenie regresných modelov binárnej odozvy. Možným príkladom ich využitia je identifikácia rizikových faktorov, ovplyvňujúcich klientovu insolvenciu, t.j. neschopnosť splácať jeho finančné záväzky voči bankovej inštitúcií a prostredníctvom vyvinutých modelov identifikovať budúcich rizikových klientov, napr. žiadateľov rôznych druhov úverov.

Efektívnosť využitia modelu logistickej regresie, za účelom hodnotenia kreditnej bonity klientov bankovej inštitúcie, vyzdvihujú vo svojej štúdií s názvom *„The Consumer Loan's Payment Default Predictive Model: an Application of the Logistic Regression and the Discriminant Analysis in a Tunisian Commercial Bank"* autori Abid, Masmoudi a Zouari-Ghorbel. Článok uverejnil Springer v roku 2018. Štúdia sa venovala identifikácií problémov, spojených s prideľovaním úverov insolventným klientom. Údajová vzorka bola

tvorená 603 jednotlivcami – klientami komerčnej banky Tuniska. 341 klientov vzorky boli identifikovaní ako insolventní, zatiaľ čo 262 klientov predstavovalo solventnú časť populácie. Za naplnením cieľa práce sa porovnávala výkonnosť modelov logistickej regresie a diskriminačnej analýzy. Ukázalo sa, že model logistickej regresie dosiahol 99 % správnosť pri klasifikácií jednotlivých klientov, naopak model diskriminačnej analýzy disponoval správnosťou iba 68,49 %. Vyššiu efektívnosť predikcií modelu logistickej regresie v porovnaní s diskriminačnou analýzou spomínali aj Ďurica et al. (2019). Prostredníctvom modelu logistickej regresie autori Abid et al. identifikovali tri významné faktory, ovplyvňujúce pravdepodobnosť nesplácania úveru – výška úveru, nesplatený úver z minulého obdobia a sociálno-profesijná kategória žiadateľa.

Obsah prehľadu literatúry bol doposiaľ venovaný prácam autorov, využívajúcich na modelovanie ekonomických problémov regresné modely, ktorých závislá premenná bola dichotomická (viď slovník termínov). V ekonomickej praxi však môže často nastať prípad, v ktorom modeli s kategoriálnou premennou modelujúcou vzťah medzi dvomi možnými alternatívami závislej premennej na základe nezávislých atribútov nie je postačujúci. V prípade, keď umelá závislá premenná definuje viac ako dve možné alternatívy, t.j. multinomická závislá premenná (viď slovník termínov), je potrebné klasický logitový model modifikovať. Tento modifikovaný model býva v literatúre označovaný pod označeniami multinomický, multinomiálny alebo multilogitový model. Pre lepšie porozumenie tejto modifikácie modelu, je model bližšie vysvetlený v podkapitole 3.1.2.

V publikácií Národnej Banky Slovenska s názvom *„Determinants of labour market flows in Slovakia“* z roku 2021, použil kolektív autorov Klacso a Štulrajterová multinomiálny logitový regresný model, s cieľom analyzovať tok pracovnej sily v rámci Slovenskej republiky. Analýza mapovala tok participantov trhu práce medzi tromi možnými alternatívami – participanti na trhu práce boli zamestnaní, nezamestnaní alebo nečinní (t.j. mimo pracovného trhu) na základe sociálno-demografických charakteristík skúmaných participantov – pohlavie, rodinný stav, vzdelanie, druh ekonomickej aktivity (závislá premenná), zdravotného hendikepu, počtu rokov v aktuálnom zamestnaní a počtu rokov v nezamestnanom stave (oba atribúty v logaritmickom vyjadrení), regiónu a počtu zamestnancov firmy. Výskum bol založený na údajoch z výberového zisťovania pracovných síl (skrátene LFS – z angl. „Labour Force Survey“) v období od 1. štvrťroka 2005 do 1. štvrťrok 2020, a tiež samostatne v období hospodárskej krízy v rokoch 2009 a 2010. S využitím multinomiálneho logit modelu v rámci tejto analýzy autori identifikovali

vzdelanie, rodinný stav a dobu zamestnania, ako kľúčové determinanty výsledkov na pracovnom trhu na Slovensku. Vplyv úrovne dosiahnutého vzdelania sa ukázal významný, najmä v období krízových rokov, ovplyvňujúci pravdepodobnosť udržania si pôvodného zamestnania alebo nájdenia si novej práce. Okrem toho výsledky štúdie zdôrazňujú zistené rozdiely medzi pohlaviami, naznačujúc, že ženy všeobecne disponujú nižšou pravdepodobnosťou nájdenia zamestnania, keď sú nezamestnané alebo nečinné.

2 CIEĽ PRÁCE

Hlavný cieľ diplomovej práce predstavuje prezentáciu aplikačných možností nelineárnych pravdepodobnostných modelov s binárnou závislou premennou (logit a probit) na konkrétnom príklade zo sféry marketingového výskumu. Pri spracovaní praktickej časti práce sa bude vychádzať z historických údajov o klientoch nemenovanej portugalskej bankovej inštitúcie, ktorí participovali v telefonickej marketingovej kampani. Cieľom kampane bolo presvedčiť klienta, aby vložil svoje voľné aktíva na termínovaný vklad, ponúkaný touto bankovou inštitúciou.

V tomto kontexte možno ako primárny cieľ praktickej časti rozumieť zostavenie a analýzu modelov, ktoré umožnia predikovať pravdepodobnosť toho, že oslovený klient akceptuje ponuku banky. Na základe predikovanej pravdepodobnosti bude možné klientov následne klasifikovať do dvoch skupín, z pohľadu výsledku marketingovej kampane – úspešná a neúspešná skupina (parciálny cieľ).

Poznatky vychádzajúce z analýzy bude môcť banka zužitkovať na zefektívnenie svojich marketingových stratégií v dvoch scenároch:

1. **Opakovanie marketingovej kampane (rovnaká klientela)** – na základe predikovanej pravdepodobnosti úspechu kampane u jednotlivých klientov bude možné ich následné rozdelenie do pravdepodobnostných segmentov, pričom sa bude opakovane cieliť na tie, v ktorých sa na základe historických údajov preukázala najvyššia koncentrácia skupiny úspešnej (parciálny cieľ).
2. **Cielenie marketingovej kampane na špecifickú klientelu –** identifikácia kľúčových determinantov maximalizujúcich možnú pravdepodobnosť úspechu u daného klienta, založených na klientových charakteristikách, akými sú sociodemografické alebo finančné, ale aj spôsob vedenia predchádzajúcich kampaní (parciálny cieľ).

Sekundárne ciele práce možno definovať v teoretickom podchytení riešenej problematiky v rámci prvej a tretej kapitoly.

Cieľom prvej kapitoly je adekvátne opísať súčasný stav riešenej problematiky. To v zmysle predstavenia základnej teoretickej roviny skúmaného problému a kľúčových pojmov, predstavenie alternatívnych modelov s možným využitím v kontexte riešeného problému a prehľad literatúry domácich i zahraničných autorov, so zameraním sa na použitú metodológiu.

Cieľ tretej kapitoly predstavuje podrobnú teoretickú definíciu použitých nelineárnych pravdepodobnostných modelov, načrtnutie možnosti modifikácie binárneho logit modelu v prípade multinomickej závislej premennej, opis metódy na odhad parametrov modelu, definovanie testov štatistickej významnosti (parametrov a modelov) a ďalších vybraných validačných techník v zmysle vhodnosti a výkonnosti modelov.

3 METODIKA PRÁCE A METÓDY SKÚMANIA

Táto kapitola je venovaná teoretickej definícii vybraných metód, ktoré budú použité v rámci praktickej časti práce. Za týmto účelom je segmentovaná do štyroch samostatných podkapitol. Prvá podkapitola sa venuje teoretickému vymedzeniu náležitostí, týkajúcich sa modelov aplikovaných v praktickej časti, t.j. dichotomickému logitovému a probitovému modelu. V rámci podkapitoly bude stručne opísaná aj mulitinomická modifikácia logitového modelu, ktorá je uplatniteľná v prípade multinomickej závislej premennej. Druhá podkapitola teoreticky definuje metódu odhadu parametrov vyššie spomínaných modelov, testovanie ich štatistickej významnosti (parametrov) a v neposlednom rade ich interpretácie. V rámci tretej podkapitoly budú rozobraté vybrané validačné techniky, prostredníctvom ktorých sa bude v praktickej časti posudzovať vhodnosť prispôsobenia modelov údajom a ich celková predikčná schopnosť.

3.1 Regresné modely s umelou závislou premennou

Hlavný problém pri modelovaní podmienenej pravdepodobnosti prostredníctvom lineárneho pravdepodobnostného modelu je, že odhadnuté hodnoty modelovanej pravdepodobnosti P_i sa môžu nachádzať mimo požadovaného intervalu pravdepodobnosti (0; 1). (Fox, 2016)

Dichotomické modely logit a probit patria medzi nelineárne pravdepodobnostné modely, ktoré predstavujú vhodné alternatívy na modelovanie vzťahu dichotomickej, resp. binárnej závislej premennej na základe množiny regresorov. Pre odstránenie uvedeného problému, modely využívajú pri modelovaní podmienenej pravdepodobnosti pozitívne monotónnu, t.j. neklesajúcu funkciu, transformujúcu lineárny prediktor $z_i = \beta_0 + \beta_1 x_i$, resp. linéarnu kombináciu prediktorov $z_i = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ji}$ do požadovaného pravdepodobnostného intervalu. Transformácia tohto typu umožňuje zachovať v podstate lineárnu štruktúru modelu, pričom hodnoty P_i nepresahujú interval (0; 1). Súhrne sa tieto transformačné funkcie označujú skratkou KDF, t.j. kumulatívna funkcia pravdepodobnostného rozdelenia. (Fox, 2016)

3.1.1 Logit model

V modeli logitovej, resp. logistickej regresie nie je nutné zohľadňovať predpoklady typické pre model lineárnej regresie (lineárny vzťah medzi závislou premennou

a regresormi, normálne rozdelenie regresorov a náhodnej zložky, nezávislosť náhodnej zložky). To poskytuje značnú flexibilitu využitia tohto modelu.

Vzťah 3.1 zachytáva model lineárnej regresie, v ktorom y_i predstavuje závislú premennú modelu, x_{ji} označuje nezávislé premenné modelu a ϵ_i predstavuje náhodnú zložku. Parameter β_0 možno rozumieť ako úrovňovú konštantu regresného modelu, zatiaľ čo β_j predstavujú parametre nezávislých premenných. (Güneri & Durmuş, 2020)

$$y_i = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ji} + \epsilon_i \tag{3.1}$$

Závislá premenná vo vzťahu 3.1 modeluje kvantitatívnu spojitú premennú, čiže lineárny regresný model možno odhadnúť metódou najmenších štvorcov – skrátene MNŠ. V prípade nelineárnych pravdepodobnostných modelov však MNŠ nepredstavuje vhodnú metóda odhadu parametrov. (Güneri & Durmuş, 2020)

Logitový model je aplikovateľný pri modelovaní pravdepodobnosti konkrétnej triedy závislej binárnej premennej. Možno predpokladať, že latentná, resp. nepozorovaná premenná $y_i^* = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ji} - \epsilon_i$, t.j. generovaná na základe pozorovanej premennej y_i, sa nachádza v rozmedzí hodnôt $-\infty$ a $+\infty$, symbolicky $y_i^* \epsilon (-\infty; +\infty)$. Ďalej možno predpokladať, že generovaná premenná y_i^* je v lineárnom vzťahu s nezávislou premennou x_{ji} v celom modeli, ceteris paribus. Generovaná premenná y_i^* je spojená s binárnou, resp. dichotomickou pozorovanou premennou y_i na základe vzťahu 3.2, v ktorom τ predstavuje prahovú (kritickú) hodnotu, z angl. „Threshold." (Güneri & Durmuş, 2020; Fox, 2016)

$$y_i = \begin{cases} 1, & y_i^* > \tau \\ 0, & y_i^* \leq \tau \end{cases} \tag{3.2}$$

Zo vzťahu 3.2 vyplýva, že ak generovaná premenná $y_i^* > \tau$, závislá premenná y_i bude nadobúdať hodnotu 1. V opačne prípade, t.j. $y_i^* \leq \tau$ bude hodnote závislej premennej y_i priradená 0. Z toho dôvodu možno hovoriť o binárnom logitovom modeli. (Güneri & Durmuş, 2020)

Všeobecne možno zapísať podmienenú pravdepodobnosť latentnej premennej, pri ktorej udalosť nastane vzťahom 3.3.

$$P_i(Y=1) = P_i(y^* > 0) = P_i\left(\beta_0 + \sum_{j=1}^{J} \beta_j x_{ji} - \epsilon_i > 0\right) = \\ = P_i(\epsilon_i < \beta_0 + \sum_{j=1}^{J} \beta_j x_{ji}) \tag{3.3}$$

V prípade modelu logit, P_i možno vypočítať prostredníctvom využitia KDF logistického pravdepodobnostného rozdelenia. Logistická funkcia $[\Lambda(z_i)]$, ktorej výstupom je podmienená pravdepodobnosť lineárneho prediktora (z_i) je znázornená vo vzťahu 3.4, resp. vo vzťahu 3.5 v prípade, ak model zohľadňuje viacero nezávislých premenných vo výpočte podmienenej pravdepodobnosti. (Güneri & Durmuş, 2020; Coss, 2015; Fox, 2016)

Logistická funkcia zohľadňujúca jeden regresor:

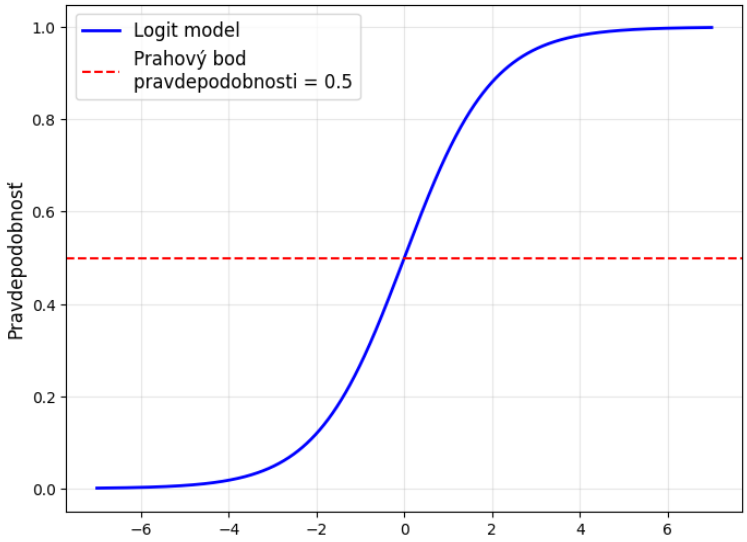
$$\Lambda(z_i) = P_i = E(y = 1|\mathrm{X} = \mathrm{x_i}) = \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}} = = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_i)}} = \frac{1}{1 + e^{-z_i}} \tag{3.4}$$

Logistická funkcia zohľadňujúca viacero regresorov :

$$\Lambda(z_i) = P_i = E\big(y = 1\big|X = x_{ji}\big) = \frac{e^{\beta_0 + \sum_{j=1}^{J} \beta_j x_{ji}}}{1 + e^{\beta_0 + \sum_{j=1}^{J} \beta_j x_{ji}}} = \frac{1}{1 + e^{-\left(\beta_0 + \sum_{j=1}^{J} \beta_j x_{ji}\right)}} \\ = \frac{1}{1 + e^{-z_i}} \tag{3.5}$$

Pre zjednodušenie vzťahov 3.4 a 3.5 bola zavedená premenná z_i, ktorá substituuje vo vzťahu 3.4 úrovňovú konštantu a regresor s príslušným parametrom: $\beta_0 + \beta_1 x_i$. Analogicky pre vzťah 3.5 premenná z_i substituuje úrovňovú konštantu a lineárnu kombináciu regresorov s príslušnými parametrami: $z_i = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ji}$. (Güneri & Durmuş, 2020; James et al., 2013; Hanck, et al., 2021)

Modelovanie podmienenej pravdepodobnosti s využitím logistickej funkcie je možné znázorniť prostredníctvom charakteristickej sigmoidnej krivky v tvare písmena "S", ako je zachytené na obrázku 3.1.



Obrázok 3.1 Priebeh logistickej funkcie (Zdroj: Vlastné spracovanie)

Výhodou sigmoidnej krivky vyobrazenej na obr. 3.1 je, že na rozdiel od klasického lineárneho modelu, modeluje pravdepodobnosť v požadovanom intervale $P_i \, \epsilon \, (0; 1)$ a nie v intervale $(-\infty; +\infty)$. Hodnota τ, pri ktorej latentná premenná nadobúda podmienenú pravdepodobnosť nastatia udalosti rovnú 50 %, t.j. $P_i = 0{,}5$, možno označiť ako prahový bod. Hodnoty, pre ktoré platí $\Lambda(z_i) > 0{,}5$, budú klasifikované y_i = 1. Analogicky pre hodnoty $\Lambda(z_i) \leq 0{,}5$, závislá premenná bude klasifikovaná y_i = 0.

Vzťahy 3.4 a 3.5 definujú P_i, v ktorej skúmaná udalosť nastane, čiže $y = 1$. Keď skúmaná udalosť nenastane, teda $y = 0$, možno pravdepodobnosť tohto prípadu získať odpočítaním pravdepodobnosti, že udalosť nastane od hodnoty 1. Tento princíp je ilustrovaný vzťahom 3.6, kde je pravdepodobnosť nenastania udalosti vyjadrená ako $1 - P_i$.

$$1 - P_i = 1 - \frac{1}{1 + \mathrm{e}^{-z_i}} \tag{3.6}$$

Ak sú rovnice 3.4, resp. 3.5 a 3.7 po dosadení konkrétnych hodnôt proporčné, je možné zaviesť rovnicu 3.7, pre výpočet pomeru pravdepodobnosti – šance, s ktorou nastane skúmaná udalosť. Teda koľkokrát je pravdepodobnejšie, že udalosť nastane oproti tomu, že nenastane.

$$\frac{P_i}{1 - P_i} = \frac{\frac{1}{1 + \mathrm{e}^{-z_i}}}{1 - \frac{1}{1 + \mathrm{e}^{-z_i}}} = \frac{1 + e^{-z_i}}{(1 + e^{-z_i}) * [(1 + e^{-z_i}) - 1]} = e^{z_i} \tag{3.7}$$

Logitový model $[\Lambda^{-1}(P_i)]$ možno následne získať, ako prirodzený logaritmus šance naznačenej vo vzťahu 3.7. Vzťah 3.8 predstavuje jeho symbolický zápis.

$$\Lambda^{-1}(P_i) = Logit_i = ln\left(\frac{P_i}{1-P_i}\right) = z_i = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ji} \tag{3.8}$$

Vzťah 3.8 zachytáva semilogaritmickú funkciu Logit $[\Lambda^{-1}(P_i)]$, ktorá je lineárna s ohľadom na x_{ji} a príslušné parametre β_j. Vstupné hodnoty pravdepodobnosti sú z pravidla asymptotické, teda $P_i \epsilon (0;1)$. Dôvod, na základe ktorého sú hodnoty P_i definované v otvorenom intervale $(0;1)$, možno demonštrovať ich dosadením do vzťahu pre výpočet Logitu $[\Lambda^{-1}(P_i)]$. (Güneri & Durmuş, 2020)

$P_i = 1$:

$$\Lambda^{-1}(P_i) = ln\left(\frac{P_i}{1-P_i}\right) = ln\left(\frac{1}{1-1}\right) = ln\left(\frac{1}{0}\right) = +\infty \tag{3.9}$$

$P_i = 0$:

$$\Lambda^{-1}(P_i) = ln\left(\frac{P_i}{1-P_i}\right) = ln\left(\frac{0}{1-0}\right) = ln\left(\frac{0}{1}\right) = -\infty \tag{3.10}$$

Výpočty uvedené vo vzťahoch 3.9 a 3.10 demonštrujú, že substitúcia krajných hodnôt $P_i = 1$ a $P_i = 0$ by v logitovej funkcii generovala hodnoty výsledkov $+\infty$ a $-\infty$. To by znemožnilo efektívne matematické operácie s takýmito hodnotami, predovšetkým v prípade odhadu parametrov. Preto je pri modelovaní v kontexte logistického regresného modelu nutné zachovať hodnoty v otvorenom intervale (0; 1), aby boli výsledky funkcie jednoznačne definované a prakticky spracovateľné. (Güneri & Durmuş, 2020)

V prípade determinácie logitovej funkcie, odhad úrovňovej konštanty β_0 a parametrov regresorov β_j nemožno použiť MNŠ, ako v prípade modelu klasickej lineárnej regresie. To z dôvodu využitia nelineárnej KDF na transformáciu lineárnej kombinácie prediktorov do požadovaného pravdepodobnostného intervalu. V nelineárnych pravdepodobnostných modeloch sa na odhad parametrov využíva metóda maximálnej vierohodnosti (skrátene – MLE, z angl. „Maximum likelihood estimation“), ktorá je opísaná v podkapitole 3.2.

Parametre regresorov v logit modeli teda nie je možné priamo interpretovať, ako účinok jednotkovej zmeny nezávislej premennej na očakávanú hodnotu závislej premennej.

Za týmto účelom sa pri aplikácií týchto modelov vypočítavajú hodnoty pomeru šancí (skrátene OR, z angl. „Odds ratio“), prípadne hodnoty marginálnych efektov (skrátene ME z angl, „Marginal effects,“ viď podkapitola 3.1.6). (Güneri & Durmuş, 2020)

Logit model (vzťah 3.8) predstavuje prirodzený logaritmus šance. Za účelom konvertovať Logit na šancu je potrebné využiť inverznú funkciu prirodzeného logaritmu – exponenciálnu funkciu: e^{z_i}. V prípade konvertovania logitu na pravdepodobnosť, sa použije inverzná funkcia k funkcií logit – logistická funkcia: $\frac{e^{z_i}}{1+e^{z_i}}$. (Nahhas, 2023)

Interpretácia konštanty $\boldsymbol{\beta}_0$ – za predpokladu, že v rovnici 3.8 bude dosadená hodnota 0 pre všetky regresory x_{ji}, hodnota prirodzeného logaritmu šance bude rovná úrovňovej konštante β_0. Prostredníctvom exponenciálnej funkcie e^{β_0} sa získa hodnota šance výsledku, zároveň pravdepodobnosť výsledku bude získaná inverznou logitovou funkciou $\frac{e^{\beta_0}}{1+e^{\beta_0}}$. (Nahhas, 2023)

Interpretácia parametrov regresorov $\boldsymbol{\beta_j}$ – v modeli lineárnej regresie predstavuje parameter β_j zmenu vo výsledku, t.j. hodnoty závislej premennej, pri zmene príslušného x_{ji} o jednu jednotku, ceteris paribus. Obdobnou interpretáciou disponujú aj parametre regresorov v úlohách logistickej regresie, resp. logit modelu. V tomto prípade β_j predstavuje zmenu vo výslednej hodnote prirodzeného logaritmu šance, pri zmene príslušného regresora x_{ji} o jednu jednotku, ceteris paribus. Matematicky možno odvodiť, že hodnota e^{β_j} predstavuje OR porovnávajúce jednotlivcov, líšiacich sa o jednu jednotku v danom x_{ji}, ceteris paribus. Toto tvrdenie možno overiť odlogaritmovaním oboch strán rovnice vzťahu 3.8, prostredníctvom čoho sa získa šanca: $\frac{P_i}{1-P_i} = e^{\beta_0 + \sum_{j=1}^{J} \beta_j x_{ji}}$. (Nahhas, 2023)

a) V prípade, v ktorom regresor x_{1i} predstavuje **kvantitatívnu spojitú** premennú, pomer šancí s hodnotou $x_{1i} + 1$ bude vyjadrený ako $\frac{e^{(\beta_0 + \beta_1(x_{1i}+1) + \beta_2 x_{2i} + \ldots + \beta_j x_{ji})}}{e^{(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_j x_{ji})}} = e^{\beta_1}$. Teda možno prehlásiť, že v prípade kvantitatívneho spojitého regresoru, jeho príslušný parameter vyjadruje prirodzený logaritmus OR, porovnávajúci jednotlivcov, líšiacich sa o jednu jednotku daného regresoru, pričom ostatné hodnoty regresorov ostávajú konštantné. (Nahhas, 2023)

b) V prípade **kategoriálneho** regresora x_{1i}, OR porovnáva nereferenčnú obmenu s príslušnou referenčnou obmenou, teda $x_{1i} = 1$ a $x_{1i} = 0$. Po dosadení opäť platí: $\frac{e^{(\beta_0 + \beta_1(x_{1i}=1) + \beta_2 x_{2i} + \ldots + \beta_j x_{ji})}}{e^{(\beta_0 + \beta_1(x_{1i}=0) + \beta_2 x_{2i} + \ldots + \beta_j x_{ji})}} = e^{\beta_1}$. Možno prehlásiť, že v prípade kategoriálneho regresoru, jeho príslušný parameter vyjadruje prirodzený logaritmus OR, porovnávajúci jednotlivcov s určitou nereferenčnou obmenou k danej referenčnej obmene regresora, pričom ostatné hodnoty regresorov ostávajú konštanté. (Nahhas, 2023)

V logistickom regresnom modeli regresory, ktorých vypočítané hodnoty OR sú blízke hodnote jedna, neindikujú významný vplyv na zmenu šancí závislej premennej. Na druhej strane, regresory s hodnotami OR väčšími ako jedna, môžu byť považované za významné faktory ovplyvňujúce zvýšenie šancí, za predpokladu, že príslušný koeficient β je štatisticky významný. Hodnoty OR, ktoré sú blízke nule naznačujú, že regresor má zásadný vplyv na zníženie šancí závislej premennej. Aj v tomto prípade je štatistická významnosť koeficientu β kritická na určenie, či regresor skutočne disponuje záporným vplyvom na výslednú hodnotu šancí. (Güneri & Durmuş, 2020)

Pri využívaní modelu logistickej regresie vo výskumnej činnosti je potrebné zohľadniť nasledujúce body:

- Všetky štatisticky významné nezávislé premenné by mali byť zahrnuté do modelu. Ich nedostatočné zahrnutie môže viesť k nárastu chybovosti modelu a jeho nevhodnosť pri ďalšom použití.
- Vyhnutie sa zahŕňaniu štatisticky nevýznamných regresorov do modelu, pretože to môže rovnako spôsobiť nežiadúce komplikácie.
- Každý zahrnutý jednotlivec by mal byť pozorovaný práve raz a nemali by sa opakovať merania.
- Chybovosť v nezávislých premenných spôsobená pri zbere údajov by mala dosahovať minimálne hodnoty, t.j. súbor údajov by mal obsahovať minimálne hodnoty chýbajúcich údajov. Vzniknuté chyby pri zbere údajov môžu skresliť odhadnuté parametre a zapríčiniť nevhodnosť modelu.
- Medzi nezávislými premennými nesmie byť prítomná multikolinearita, t.j. regresory by nemali byť navzájom závislé.

– Súbor údajov by nemal obsahovať extrémne hodnoty, t.j. odľahlé pozorovania. Tie môžu rovnako ako v lineárnej i logistickej regresií negatívne ovplyvniť vhodnosť modelu. (Güneri & Durmuş, 2020)

3.1.2 Multinomiálny logit model

Logitový model definovaný v podkapitole 3.1.1 nachádza svoje praktické využitie v klasifikačných úlohách, pri ktorých závislá premenná je dichotomická. To znamená, že nadobúda práve dve kategórie obmien, resp. alternatív. Symbolicky možno počet kategórií obmien zapísať K = 2. V bežnej praxi je možné stretnúť sa s klasifikačnými úlohami, ktorých závislá premenná nadobúda viac ako dve obmeny. Závislá premenná, ktorej K > 2, býva označovaná ako multinomická, resp. multinomiálna. Za týmto účelom je možné modifikovať klasický model logistickej regresie, ktorý býva pri modelovaní K > 2 alternatív symbolicky označovaný ako multinomiálny logitový model. (James et al., 2013)

Pri modifikácií je potrebné najskôr zvoliť jednu triedu, ktorá bude v modeli považovaná za základnú, resp. referenčnú. V prípade, v ktorom bude vybraná K – tá trieda multinomiálnej závislej premennej ako referenčná, vzťah 3.5 pre výpočet podmienenej pravdepodobnosti je potrebné adaptovať do tvaru vzťahov 3.11 pre porovnávajúce kategórie a vzťahu 3.12 pre referenčnú kategóriu.

$$\mathrm{P}(Y = k|X = x) = \frac{e^{\beta_{k0}+\sum_{j=1}^{J}\beta_{kj}x_j}}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0}+\sum_{j=1}^{J}\beta_{lj}x_j}} \tag{3.11}$$

Pre hodnoty k = 1, ... , K – 1; pre referenčnú kategóriu platí:

$$\mathrm{P}(Y = K|X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0}+\sum_{j=1}^{J}\beta_{lj}x_j}} \tag{3.12}$$

Vypočítanú podmienenú pravdepodobnosť na základe vzorcov 3.11 a 3.12, možno následne vyjadriť ako prirodzený logaritmus šance vo vzťahu 3.13.

$$\ln\left(\frac{\mathrm{P}(Y = k|X = x)}{\mathrm{P}(Y = K|X = x)}\right) = \beta_{k0} + \sum_{j=1}^{J}\beta_{kj}x_j \tag{3.13}$$

Ako je možné si všimnúť, vzťah 3.13 sa nápadne podobá rovnici na výpočet logitu 3.8 v úlohách dichotomickej logistickej regresie. Opäť sa predpokladá, že logaritmus šance

je vo vzťahu lineárny vzhľadom na nezávislé premenné, s príslušnými parametrami. (James et al., 2013)

Je vhodné podotknúť, že výber K – tej triedy, ktorá sa bude považovať za základnú vo vzorcoch 3.11 a 3.12, nie je príliš dôležitý. Pre lepšiu interpretovateľnosť možno uvažovať fiktívny príklad zo zdravotníckeho prostredia. Napríklad pri klasifikácií návštev pohotovosti je potrebné klasifikovať pacientov do troch tried (K = 3) – mozgová príhoda (ozn. MP), predávkovanie drogami (ozn. PD) a epileptické záchvaty (ozn. EZ). Možno predpokladať zvolenie dvoch modelov multinomiálnej logistickej regresie - jeden, kde sa MP považuje za referenčnú a druhý, v ktorom PD vystupuje ako referenčná kategória. Odhady koeficientov sa budú líšiť medzi týmito dvoma modelmi, z dôvodu odlišného výberu referenčnej kategórie, avšak predikcie, logaritmické šance medzi akýmikoľvek dvojicami tried a ďalšie kľúčové výstupy modelu zostávajú rovnaké. (James et al., 2013)

Dôležité je interpretovať koeficienty v modeli multinomiálnej logistickej regresie s ohľadom k viazanému základu.

Napríklad ak bude za základ zvolená trieda EZ, parameter β_{MP0} možno interpretovať ako prirodzený logaritmus šance MP voči EZ pri $x_1 = \cdots = x_J = 0$. Okrem toho jednotkový nárast v x_j je spojený s nárastom o β_{MPj} v prirodzenom logaritme šance MP oproti EZ. Inými slovami, ak sa x_j zvýši o jednotku, potom sa pomer $\frac{\mathrm{P}(\mathrm{Y} = \mathrm{MP}|\mathrm{X} = \mathrm{x})}{\mathrm{P}(\mathrm{Y} = \mathrm{EZ}|\mathrm{X} = \mathrm{x})}$ zvýši presne o $e^{\beta_{MPj}}$. (James et al., 2013)

3.1.3 Probit model

V modeli logistickej regresie, resp. v modeli logit, bolo KDF definované logistickou funkciou, znázornenou vo vzťahoch 3.4, prípadne 3.5. Táto funkcia predstavuje v súčasnosti aj najviac využívanú KDF v úlohách nelineárnych pravdepodobnostných modelov.

V prípade modelu probit sa na modelovanie podmienenej pravdepodobnosti používa KDF normovaného normálneho rozdelenia – jednotkového normálneho rozdelenie. Pre toto pravdepodobnostné rozdelenie je charakteristická nulová stredná hodnota a smerodajná odchýlka rovnajúca sa jednej, symbolicky $N(0; 1)$. (Fox, 2016)

Výpočet podmienenej pravdepodobnosti prostredníctvom distribučnej funkcie normovaného normálneho rozdelenia (Φ) možno formálne zapísať vzťahom 3.14 [lineárny

prediktor obsahujúci jeden regresor (z_i)] a vzťahom 3.15 [lineárna kombinácia prediktorov (z_i)].

Lineárny prediktor (obsahujúci jeden regresor):

$$P_i = E(y = 1|x_i) = \Phi(\beta_0 + \beta_1 x_i) = \Phi(z_i) \tag{3.14}$$

Lineárna kombinácia prediktorov:

$$P_i = E(y = 1|X) = \Phi\left(\beta_0 + \sum_{j=1}^{J} \beta_j x_{ji}\right) = \Phi(z_i) \tag{3.15}$$

Vo vzťahoch 3.14 a 3.15 predstavuje Φ integrál hustoty pravdepodobnosti. Zohľadnením tejto skutočnosti možno uvedené vzťahy rozvinúť, ako je prezentované vo vzťahoch 3.16 a 3.17.

V prípade lineárneho prediktora (jeden regresor):

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z = \beta_0 + \beta_1 x_i} e^{-\frac{1}{2}z^2}\, dz \tag{3.16}$$

V prípade lineárnej kombinácie viacerých prediktorov:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ji}} e^{-\frac{1}{2}z^2}\, dz \tag{3.17}$$

Definovaná distribučná funkcia vyjadruje pravdepodobnosť pre každé reálne z, že náhodná veličina Z nadobudne nižšiu alebo nanajvýš rovnajúcu sa hodnotu z. Symbolicky uvedené možno zapísať, ako $\Phi(z) = P(Z \leq z)$. Premenná Z v tomto kontexte charakterizuje tzv. z-skóre, resp. probitové skóre alebo probitový index. Vyčísliť z-skóre je možné prostredníctvom využitia inverznej funkcie k distribučnej funkcií normovaného normálneho rozdelenia, podľa vzťahu 3.18. (Rublíková et al., 2009)

$$\Phi^{-1}(P_i) = Z = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ji} \tag{3.18}$$

Vo vzťah 3.18 reprezentuje Φ^{-1} inverznú funkciu pre KDF normovaného normálneho pravdepodobnostného rozdelenia. (Güneri & Durmuş, 2020) Vhodné je spomenúť, že inverzná linearizujúca transformácia pre logitový model, t.j. $\Lambda^{-1}(P_i)$ je priamo

interpretovateľná, ako prirodzený logaritmus šance, zatiaľ čo inverzná transformácia pre probitový model, t.j. kvantilová funkcia normovaného normálneho rozdelenia, $\Phi^{-1}(P_i)$, nemá priamu interpretáciu. (Fox, 2016)

Rovnako, ako v prípade logitového modelu možno demonštrovať binárnosť probitového modelu prostredníctvom latentnej premennej. V tomto prípade možno latentnú premennú definovať vzťahom $y^* = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ji} + \epsilon_i$, v ktorom náhodná zložka (ϵ_i) disponuje normálnym rozdelením $N(0; \sigma^2)$. Hodnoty binárnej závislej premennej y_i sú definované prostredníctvom latentnej premennej y_i^* na základe vzťahu 3.19. (Rublíková et al., 2009)

$$y_i = \begin{cases} 1, & y_i^* > 0 \\ 0, & y_i^* \leq 0 \end{cases} \tag{3.19}$$

Pravdepodobnosť, pri ktorej y_i^* nadobudne hodnotu 1 možno vyjadriť nasledovne:

$$\begin{aligned} P_i(y_i = 1) = P_i(y_i^* > 0) = P_i\left(\beta_0 + \sum_{j=1}^{J} \beta_j x_{ji} + \epsilon_i > 0\right) = \\ = P_i\left(\epsilon_i > -\beta_0 - \sum_{j=1}^{J} \beta_j x_{ji}\right) = 1 - F(-\beta_0 - \sum_{j=1}^{J} \beta_j x_{ji}) \end{aligned} \tag{3.20}$$

Vo vzťahu 3.20 demonštruje hodnotu distribučnej funkcie normálneho rozdelenia. Je však potrebné zabezpečiť, aby náhodná zložka bola transformovaná do hodnôt normovaného normálneho rozdelenia. S využitím symetrickosti tohto rozdelenia je možné vzťah 3.20 preformulovať do podoby vzťahu 3.21. (Rublíková et al., 2009)

$$P(y_i = 1) = 1 - \Phi\left(-\frac{\beta_0}{\sigma} - \sum_{j=1}^{J} x_{ji} \frac{\beta_j}{\sigma}\right) = \Phi\left(\frac{\beta_0}{\sigma} \sum_{j=1}^{J} x_{ji} \frac{\beta_j}{\sigma}\right) \tag{3.21}$$

V prípade využitia vzťahu medzi KDF a funkciou hustoty pravdepodobnosti, možno podmienenú pravdepodobnosť nastatia udalosti vyjadriť vzťahom 3.22.

$$P(y_i = 1) = \Phi\left(\frac{\beta_0}{\sigma} \sum_{j=1}^{J} x_{ji} \frac{\beta_j}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_i} e^{-z^2}\, dz \tag{3.22}$$

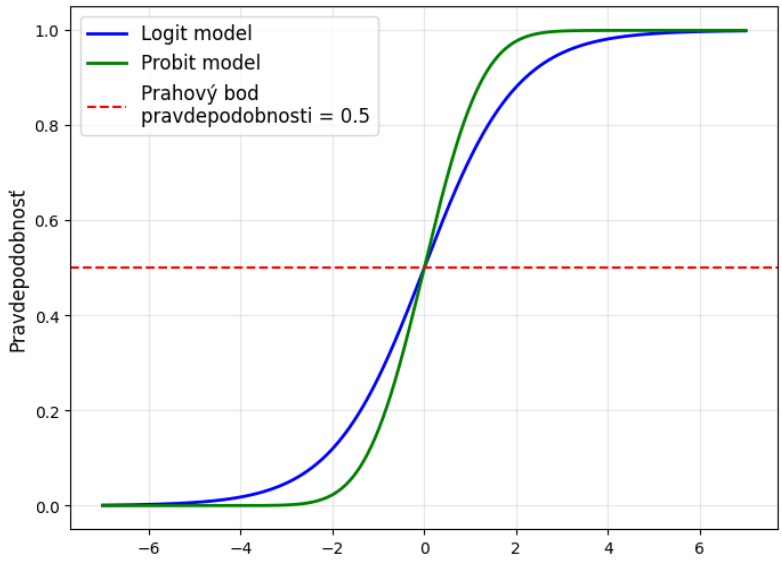
kde $z_i = \frac{\beta_0}{\sigma} \sum_{j=1}^{J} x_{ji} \frac{\beta_j}{\sigma}$

Analogicky možno vyjadriť pravdepodobnosť nenastatia udalosti, teda že náhodná premenná nadobudne nulovú hodnotu vzťahom 3.23.

$$P(y_i = 0) = 1 - \Phi\left(\frac{\beta_0}{\sigma}\sum_{j=1}^{J} x_{ji}\frac{\beta_j}{\sigma}\right) \tag{3.23}$$

Vo všeobecnosti obe distribučné funkcie pravdepodobnostného rozdelenia, resp. KDF funkcie modelov logit a probit disponujú krivkami, ktoré sú už na prvý pohľad veľmi podobné. Krivky oboch funkcií vykazujú podobný sigmoidný charakter – obe sú v tvare písmena „S," monotónne rastúce a k obom existujú inverzné funkcie. (Coss, 2015)

Na obrázku 3.2 je zobrazený priebeh distribučných funkcií oboch modelov. Krivka pre KDF logistického rozdelenia pravdepodobnosti, ktorá sa používa v logit modeli, je znázornená modrou sigmoidnou krivkou. Zelená sigmoidná krivka zodpovedá KDF štandardizovaného normálneho rozdelenia, ktorá sa využíva v probit modeli.



Obrázok 3.2 Porovnanie distribučných funkcií modelov logit a probit (Zdroj: Vlastné spracovanie)

Ďalšou podobnosťou kriviek znázornených na obrázku 3.2 je, že obor hodnôt oboch funkcií je z intervalu $(0;\ 1)$, pričom definičný obor funkcií je z intervalu $(-\infty;\ +\infty)$. Obe krivky sú symetrické v hodnote podmienenej pravdepodobnosti $P_i = 0{,}5$. Badateľným rozdielom kriviek je, že KDF štandardizovaného normálneho rozdelenia vykazuje väčšiu strmosť, teda aj rýchlejšie klesá. (Fox, 2016; Coss, 2015)

KDF logistického aj štandardizovaného normálneho pravdepodobnostného rozdelenia sú veľmi podobné, odhadnuté β parametre modelov prostredníctvom MLE sa

odlišujú výraznejšie, pričom ich nie je možné priamo porovnať. Pri ich porovnávaní musia byť odhady parametrov logit modelu prenásobené $\frac{\sqrt{3}}{\pi}$, kvôli rozdielnemu rozptylu pravdepodobnostných rozdelení (štandardizované normálne rozdelenie má rozptyl rovný hodnote 1, logistické rozdelenie disponuje rozptylom rovným $\frac{\pi^2}{3}$). (Fox, 2016; Coss, 2015)

Pri aplikácií modelu probit musia byť zohľadnené nasledujúce platné predpoklady:

- Závislá premenná je dichotomická (binárna) pre všetky pozorovania, teda symbolicky $y_i \epsilon \{0; 1\}, i = 1, \dots, n$;
- podmienená pravdepodobnosť, že skúmaná udalosť nastane je daná prostredníctvom KDF jednotkového, t.j. normovaného normálneho pravdepodobnostného rozdelenia, teda $P_i = E(Y = 1|x) = \Phi(\beta x_i)$;
- hodnoty všetkých závislých premenných, t.j. $y_1, \dots, y_n$ sú štatisticky nezávislé;
- medzi všetkými nezávislými premennými x_{ji} nie je prítomná multikolinearita. (Güneri & Durmuş, 2020)

3.2 Odhad parametrov modelov

V regresných modeloch β parametre sú neznáme a je potrebné ich odhadnúť na základe údajov z trénujúcej množiny. Už bolo spomenuté, že v prípade klasického modelu lineárnej regresie sú odhady parametrov realizované prostredníctvom MNŠ – metódy najmenších štvorcov. Táto metóda však nie je aplikovateľná v prípade nelineárnych pravdepodobnostných modelov. Pre tento typ úloh sa uplatňuje metóda maximálnej vierohodnosti - MLE. (James et al., 2013)

Programovací jazyk R sprostredkuje MLE odhady parametrov automaticky. Všetky detaily tejto metódy sú príliš komplexné, pričom priamo nesúvisia s témou diplomovej práce. Metódu je však vhodné názorne priblížiť, pretože miera vierohodnosti bude využívaná aj v rámci viacerých validačných techník.

Vzťah 3.24 zachytáva symbolické označenie pravdepodobnosti pre prípad, v ktorom skúmaná situácia nastane $y_i = 1$, resp. nenastane $y_i = 0$.

$$y_i = \begin{cases} 1, & P_i \\ 0, & 1 - P_i \end{cases} \tag{3.24}$$

Funkciou vierohodnosti možno rozumieť funkciu združeného rozdelenia pravdepodobnosti výberu, zároveň je chápaná ako funkcia hodnôt parametrov pri fixovaných

hodnotách výberu. (Lukáčik, 2008) Pre súbor údajov s počtom pozorovaní N, možno sformulovať funkciu vierohodnosti L vo vzťahu 3.25.

$$L = \prod_{i=1}^{N} P_i^{y_i}(1 - P_i)^{1-y_i} \tag{3.25}$$

Model logit alebo probit je založený na výpočte podmienenej pravdepodobnosti P_i. Podmienená pravdepodobnosť pre $X_i'\beta$ je určená prostredníctvom funkcie KDF. Je zrejmé, že v prípade modelu logit ide o KDF logistického pravdepodobnostného rozdelenia. V prípade probit modelu ide o KDF normovaného normálneho rozdelenia. Nech $F(X_i'\beta)$ symbolizuje vybranú KDF, potom možno funkciu vierohodnosti zapísať všeobecne pre oba modely (vzťah 3.26). (Fomby, 2010)

$$L = \prod_{i=1}^{N} F(X_i'\beta)^{y_i}\left(1 - F(X_i'\beta)\right)^{1-y_i} \tag{3.26}$$

Na základe vzťahu 3.26 možno sformulovať rovnicu pre výpočet prirodzeného logaritmu vierohodnosti (l) – „log-likelihood“ – vzťah 3.27.

$$\ln L = l = \sum_{i=1}^{N} [y_i \ln F(X_i'\beta) + (1 - y_i)\ln\left(1 - F(X_i'\beta)\right)] \tag{3.27}$$

Maximalizáciou prirodzeného logaritmu funkcie vierohodnosti (vzťah 3.27) vzhľadom na parametre regresorov, možno získať vzťahy určené pre ich odhady. (Lukáčik, 2008) Podmienky prvého rádu vo vzťahu 3.27 sú nelineárne a nie je možné ich analyticky vyriešiť. Z toho dôvodu sa odhady parametrov získavajú prostredníctvom metód matematickej optimalizácie maximálnej vierohodnosti, napr. prostredníctvom metódy Newton-Raphson. (Fomby, 2010)

3.2.1 Testovanie štatistickej významnosti odhadnutých parametrov

Testovanie štatistickej významnosti odhadnutých β parametrov v logitových a probitových modeloch je možné uskutočniť prostredníctvom Waldovho testu, tiež známeho ako Z-test. Tento test využíva odhady parametrov získané metódou maximálnej vierohodnosti (MLE) a predstavuje analógiu k testovaniu štatistickej významnosti jednotlivých parametrov, prostredníctvom t-testu vo všeobecnom lineárnom modeli. Waldov test testuje nulovú hypotézu, ktorá predpokladá rovnosť odhadnutého parametra s nulovou hodnotou. (Fox, 2016)

Symbolicky možno zapísať nulovú a alternatívnu hypotézu Waldovho testu nasledujúco:

- $H_0: \beta_j = 0$;
- $H_A: \beta_j \neq 0$.

Uvedené hypotézy sa ďalej testujú prostredníctvom vypočítanej hodnoty Waldovej Z – štatistiky podľa vzťahu 3.28.

$$Z = \frac{\widehat{\beta_j}}{SE(\widehat{\beta_j})}, \tag{3.28}$$

Testovacia Z-štatistika zachytená vo vzťahu 3.28 charakterizuje podiel hodnoty odhadnutého parametra $\widehat{\beta_j}$ a jeho príslušnej smerodajnej odchýlky $SE(\widehat{\beta_j})$. Vypočítaná hodnota Z – štatistiky nasleduje asymptotické štandardné normálne rozdelenie. Interval spoľahlivosti pre β_j možno získať použitím štandardného vzorca. Koncové body intervalu spoľahlivosti 100(1 - α) % pre koeficienty β_0 a β_1 možno sformulovať, ako $\widehat{\beta_0} \pm z_{1-\frac{\alpha}{2}} se\widehat{\beta_0}$ a $\widehat{\beta_1} \pm z_{1-\frac{\alpha}{2}} se\widehat{\beta_1}$. Pre hodnotu $z_{1-\frac{\alpha}{2}}$ platí, že Z ~ N(0,1) s pravdepodobnosťou $\frac{\alpha}{2}$ z pravej strane rozdelenia. (Fox, 2016)

Hosmer et al. (2013) odporúčajú pri testovaní štatistickej významnosti odhadnutých parametrov využitie testu pomeru vierohodnosti – LR test, z angl. „Likelihood ratio test," namiesto využitia Waldovho testu.

Test pomeru vierohodnosti (LR test) testuje nulovú hypotézu vychádzajúcu z predpokladu, že všetky odhadnuté β parametre modelu sa súčasne rovnajú nulovej hodnote. V takom prípade je model ako celok považovaný za štatisticky nevýznamný. Analogicky, prijatím alternatívnej hypotézy možno označiť model ako celok za štatisticky významný, t.j. aspoň jeden parameter modelu je rôzny od nulovej hodnoty.

Symbolicky možno zapísať nulovú a alternatívnu hypotézu LR testu nasledujúco:

- $H_0: \beta_1 = \ldots = \beta_j = 0$;
- $H_A: \beta_1 \neq \ldots \neq \beta_j \neq 0$.

Nulová hypotéza je testovaná prostredníctvom testovacej G – štatistiky (vzťah 3.29) Vzťah pre jej výpočet možno odvodiť na základe predpokladu existencie 2 modelov:

- Model 1 – obsahujúci regresory s príslušnými parametrami a s hodnotou maximálnej vierohodnosti L_1;
- Model 0 - obsahujúci iba konštantu β_0 s hodnotou maximálnej vierohodnosti L_0.

Vzhľadom k tomu, že pre hodnotu maximálnej vierohodnosti modelov platí, že $L_1 > L_0$, vzťah pre testovanie nulovej hypotézy bude zapísaný nasledovne:

$$G = 2(lnL_1 - lnL_0) \tag{3.29}$$

Pri nulovej hypotéze má testovacia štatistika G asymptotické rozdelenie χ^2 s počtom q – stupňami voľnosti. (Hosmer et al., 2013; Fox, 2015)

3.2.2 Marginálne efekty

Za účelom interpretácie parametrov v nelineárnych pravdepodobnostných modeloch, v zmysle efektu jednotkovej zmeny regresora, na zmenu výslednej pravdepodobnosti, ceteris paribus, je nutné vypočítať ich marginálne efekty. (Greene, 2012)

V rámci tejto podkapitoly budú názorne odvodené vzťahy pre výpočet marginálnych efektov, prostredníctvom parciálnych derivácií príslušnej KDF funkcie $F(.)$, modelujúcej podmienenú pravdepodobnosť lineárnej kombinácie prediktorov $x'\beta$.

Možno uvažovať všeobecný tvar regresného pravdepodobnostného modelu (vzťah 3.30), ktorého očakávaná stredná hodnota závislej premennej y pri danom regresore x, je rovná príslušnej distribučnej funkcií - *F(x'β),* v ktorej x' predstavuje transponovaný vektor regresorov a β je vektor regresných koeficientov. (Greene, 2012)

$$E[y|x] = F(x'\beta) \tag{3.30}$$

Pre výpočet marginálnych efektov, je potrebné skonštruovať pre vzťah 3.30 parciálne derivácie, podľa príslušných regresorov. Všeobecne platí vzťah 3.31.

$$\frac{\partial E[y|x]}{\partial x} = \left[\frac{dF(x'\beta)}{d(x'\beta)}\right] \times \beta = f(x'\beta) \times \beta \tag{3.31}$$

Funkcia $f(.)$ vo vzťahu 3.31 predstavuje funkciu hustoty pravdepodobnostného rozdelenia pre zodpovedajúcu KDF – $F(.)$. Vzťahom 3.32 možno symbolicky vyjadriť funkciu hustoty logistického pravdepodobnostné rozdelenie (L) a normovaného normálneho rozdelenia (N). (Greene, 2012)

$$f(z) = \begin{cases} L: \Lambda(z)[1 - \Lambda(z)] \\ N: \varphi = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \end{cases} \tag{3.32}$$

Následne možno odvodiť vzťah (3.33) pre výpočet marginálnych efektov probitového modelu:

$$\frac{\partial E[y|x]}{\partial x} = \varphi(x'\beta) \times \beta \tag{3.33}$$

V ktorom φ označuje hustotu normovaného normálneho rozdelenia. Parciálne derivácie logistického rozdelenia sú zachytené vo vzťahu 3.34.

$$\frac{d\Lambda(x'\beta)}{d(x'\beta)} = \frac{e^{(x'\beta)}}{\left[1 + e^{(x'\beta)}\right]^2} = \Lambda(x'\beta)[1 - \Lambda(x'\beta)] \tag{3.34}$$

Teda pre výpočet hodnoty marginálnych efektov logitového modelu platí vzťah 3.35.

$$\frac{\partial E[y|x]}{\partial x} = \Lambda(x'\beta)[1 - \Lambda(x'\beta)]\beta \tag{3.35}$$

Je zrejmé, že hodnoty marginálnych efektov nelineárnych pravdepodobnostných modelov sa budú meniť, na základe hodnoty regresora x. V praxi sa pre zjednodušenie interpretácií považuje za užitočné, vypočítať hodnotu marginálnych efektov na priemerných hodnotách regresorov alebo na referenčnom pozorovaní, s dopredu preddefinovanými hodnotami nezávislých premenných. Ďalšou možnosťou je vypočítať marginálne efekty pre všetky pozorovania, následne ich hodnoty spriemerovať, čoho výsledkom budú priemerné parciálne efekty. V prípade rozsiahlych údajových vzoriek, priemerné parciálne efekty a marginálne efekty priemerných hodnôt regresorov produkujú takmer totožné hodnoty. Výraznejšie odchýlky hodnôt medzi týmito alternatívami možno detegovať na malých a stredných údajových vzorkách. (Greene, 2012)

Komplikácie môžu nastať v prípade výpočtu marginálnych efektov kategoriálnych premenných. Vo všeobecnosti platí, že pre potreby modelovania je nutné ich prekódovanie, resp. nahradenie umelými premennými. Vzhľadom na túto skutočnosť je zrejmé, že derivovanie na základe hodnôt umelých premenných nie je vhodné. Výsledná hodnota marginálneho efektu by vyjadrovala len malú zmenu a neodzrkadľovala tým reálny vplyv zmeny kategórie, na výslednú hodnotu modelovanej pravdepodobnosti. (Greene, 2012)

Možno uvažovať binárnu nezávislú premennú d ϵ $\{0; 1\}$. Vhodný spôsob výpočtu marginálneho efektu pre takýto regresor predstavuje rozdiel v predikovaných pravdepodobnostiach oboch kategórií (vzťah 3.36).

$$ME = P_i\left[Y = 1\middle|\bar{x}_{(d)}, d = 1\right] - P_i[Y = 1|\bar{x}_{(d)}, d = 0] \quad (3.36)$$

Vo vzťahu 3.36 symbolizuje $\bar{x}_{(d)}$ priemerné hodnoty ostatných regresorov v celom modeli, zohľadňujúc pri výpočte pravdepodobnosti konkrétnu kategóriu umelej premennej d. (Greene, 2012)

3.3 Validačné metódy určené na identifikáciu vhodnosti modelu

Všeobecne platí, že po odhadnutí modelov je potrebné overiť jeho vhodnosť („goodness of fit“), t.j. do akej miery model úspešne opisuje modelovaný vzťah závislej premennej od jednej, poprípade viacerých nezávislých premenných príslušnej údajovej množiny. V podkapitole 3.2.1 bol predstavený Waldov test, prostredníctvom ktorého sa overuje štatistická významnosť jednotlivých odhadnutých parametrov regresorov modelu, poprípade test pomeru vierohodnosti, ktorý je preferovanou alternatívou, pri overovaní štatistickej významnosti modelu ako celku. Oba testy sú založené na štatistickom testovaní hypotéz.

V rámci tejto podkapitoly budú opísané validačné techniky, ktoré síce nie sú založené na testovaní štatistických hypotéz, ale budú užitočné v rámci praktickej časti na porovnávanie modelov v štádiu ich vývoja. Definované budú techniky – McFaddenov nepravý koeficient determinácie ($R^2_{McFadden}$), Akaikeho informačné kritérium a Bayesovo informačné kritérium. Prostredníctvom uvedených techník sa bude porovnávať vhodnosť jednotlivých modelov, na základe rôzne zvolených kombinácií regresorov.

Po výbere najvhodnejších modelov prostredníctvom vyššie uvedených technik, bude ich predikčná schopnosť testovaná prostredníctvom validačných metód, bežne používaných v úlohách binárnej klasifikácie. Schopnosť modelov, klasifikovať pozorovania do jednej alebo druhej triedy bude sprostredkovaná prostredníctvom matice zámen, rôznych metrík odvodených z tejto matice, krivkou ROC a plochou pod touto krivkou (hodnota AUC). Uvedené metódy umožnia porovnať predikčnú schopnosť najvhodnejšieho logitového a probitového modelu navzájom.

3.3.1 Nepravý koeficient determinácie - Pseudo R^2

Koeficient determinácie (R^2) predstavuje techniku, ktorá je v úlohách lineárnej regresie využívaná za účelom overenia vhodnosti prispôsobenia modelu údajom, t.j. „goodness of fit“ zvoleného modelu. Koeficient determinácie možno rozumieť, ako podiel

variability $S_T(y) = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ opísanej regresným modelom, k celkovej variabilite vysvetľovanej premennej $S_c(y) = \sum_{i=1}^{n}(y_i - \bar{y})^2$. Vypočítané hodnoty $R^2 \in \langle 0; 1\rangle$, pričom hodnota blížiaca sa k 1 indikuje silnú štatistickú väzbu, naopak hodnota blížiacej sa k 0 napovedá, že štatistická väzba je slabšia. (Neubauer et al., 2021)

Pri regresných modeloch, ktorých závislá premenná je kategoriálna, nie je možné vypočítať tradičný koeficient determinácie. Parametre týchto modelov sa odhadujú pomocou metódy maximálnej vierohodnosti (MLE), pričom cieľom je maximalizovať vierohodnosť modelu, nie minimalizovať sumu druhých mocnín rezíduí, ako v prípade metódy najmenších štvorcov (MNŠ) v lineárnej regresii. Preto sa v týchto modeloch používajú tzv. pseudo koeficienty determinácie – pseudo R^2. (IBM, 2023; Institute for Digital Research and Education, 2021)

Existuje viacero druhov štatistík pseudo R^2, ktorých vzorce sa značne odlišujú. Pri spracovaní diplomovej práce bude použitý nepravý koeficient determinácie, ktorý v roku 1974 opísal McFadden. Vzťah pre jeho výpočet je zahrnutý v rovnici 3.37. (IBM, 2023)

$$R^2_{McFadden} = 1 - \frac{ln\, L_1}{ln\, L_0} \tag{3.37}$$

- L_1 – *vierohodnosť odhadnutého modelu s prediktormi;*
- L_0 – *vierohodnosť odhadnutého modelu bez prediktorov.*

Vypočítané hodnoty pre nepravý koeficient determinácie $R^2_{McFadden}$ sa nachádzajú v intervale $\langle 0; 1\rangle$, pričom hodnoty $0{,}2 \leq R^2_{McFadden} < 0{,}4$ sa považujú za uspokojivé. (Rublíková et al., 2009)

Nepravý koeficient determinácie $R^2_{McFadden}$ nedisponuje priamou interpretáciou. V prípade použitia tejto metriky pri výbere najvhodnejšieho modelu sa uprednostňuje ten, ktorého vypočítaná hodnota $R^2_{McFadden}$ dosahuje najvyššiu hodnotu. (IBM, 2023)

3.3.2 Akaikeho a Bayesovo informačné kritérium

Akaikeho informačné kritérium (skrátene – AIC, z angl. „Akaike information criterion“) a Bayesovo informačné kritérium (skrátene – BIC, z angl. „Bayesian information criterion“) predstavujú metódy, určené na porovnávanie vhodnosti štatistických modelov. Obe kritéria poskytujú referenčný rámec pri výbere najvhodnejšieho modelu z rôznych dostupných alternatív, ktoré sa môžu líšiť v počte a typoch regresorov (atribútov), t.j. líšiacej

sa komplexnosti porovnávaných modelov. Rovnako, ako v prípade $R^2_{McFadden}$, kritéria nedisponujú priamou interpretáciou a nie sú podkladom pre testovanie štatistických hypotéz. Je dôležité uviesť, že použitie obidvoch kritérií na testovanie vhodnosti prispôsobenia jedného konkrétneho modelu k údajom, obvykle nie je veľmi užitočné. Ich uplatnenie spočíva najmä pri spomenutom porovnávaní modelov (odhadnutých z rovnakej údajovej množinu) na základe ich rozdielnej komplexnosti. Ako preferovaný model bude označovaný ten, ktorý disponuje najnižšími hodnotami informačných kritérií v porovnaní s možnými alternatívami (platí pre obe kritéria). (Lobos et al., 2012)

V prípade ich využitia v kontexte nelineárnych pravdepodobnostných modelov, vzťahy pre výpočet oboch kritérií vychádzajú z prirodzeného logaritmu vierohodnosti a penalizačného člena. Pre lepšie porozumenie, vierohodnosť modelu možno zvyšovať postupným pridávaním regresorov do modelu, čo môže v konečnom dôsledku negatívne ovplyvniť model – dôjde k jeho preučeniu. Tento problém sa rieši zavádzaním spomínaného penalizačného člena, prostredníctvom ktorého je model penalizovaný na základe jeho počtu parametrov. (Lobos et al., 2012)

Z uvedeného možno sformulovať vzťah 3.38 pre výpočet AIC a vzťah 3.39 pre výpočet BIC.

$$AIC = -2ln(L) + 2k \tag{3.38}$$

- *ln(L) – prirodzený logaritmus vierohodnosti odhadnutého modelu;*
- *k – počet odhadovaných parametrov v modeli.*

Vo vzťahu 3.38 je penalizačný člen definovaný, ako dvojnásobok počtu parametrov $(2k)$ zahrnutých v modeli.

Penalizácia modelov založená na BIC je ešte prísnejšia, pretože zahŕňa pre výpočet penalizačného člena multiplikátor veľkosti údajovej vzorky.

$$BIC = -2ln(L) + k * ln(n) \tag{3.39}$$

- *ln(L) – prirodzený logaritmus vierohodnosti odhadnutého modelu;*
- *k – počet odhadovaných parametrov v modeli;*
- *n – predstavuje veľkosť vzorky.*

Penalizačný člen vo vzorci 3.39 je definovaný súčinom počtu parametrov modelu a prirodzeného logaritmu veľkosti vzorky, čo poskytuje väčšiu penalizáciu (v porovnaní s AIC) za každý parameter v modeli. To najmä v prípade, ak je veľkosť údajovej vzorky rozsiahla. Tento prístup je efektívny v zamedzení pridania irelevantných parametrov do modelu, čo by mohlo viesť k jeho preučeniu. (Lobos et al., 2012)

3.3.3 Matica zámen

Technika známa ako matica zámen (z angl. „Confusion matrix") patrí medzi často využívané analytické nástroje, určené na hodnotenie výsledkov klasifikačných úloh. Okrem toho, že prostredníctvom tejto metódy možno veľmi podrobne zhodnotiť úspešnosť klasifikácie, tvorí základ pre odvodenie viacerých metrík, ktoré odhaľujú viac o výkonnostných charakteristikách testovaných klasifikačných modelov. V prípade binárnej klasifikácie ju možno chápať ako štvorcovú maticu (2 x 2), pričom jednotlivé bunky matice zachytávajú početnosť klasifikovaných prípadov z testovacej sady údajov nasledovne:

- **Správne klasifikované pozitívne prípady** (skrátene **TP**, z angl. „True positive“) – charakterizuje početnosť prípadov, v ktorých bola predikovaná trieda závislej premennej modelu **Y = 1** správne vyhodnotená, resp. klasifikovaná ako jej skutočná hodnota **Y = 1**.
- **Správne klasifikované negatívne prípady** (skrátene **TN**, z angl. „True negative“) – zachytáva početnosť prípadov, pri ktorých bola predikovaná trieda závislej premennej **Y = 0** správne vyhodnotená, ako jej skutočná hodnota **Y = 0**.
- **Nesprávne klasifikované, ako pozitívne prípady** (skrátene **FP**, z angl. „False positive“) – možno rozumieť početnosť prípadov, v ktorých bola hodnota skutočnej závislej premennej modelu **Y = 0**, nesprávne klasifikovaná ako **Y = 1**.
- **Nesprávne klasifikované, ako negatívne prípady** (skrátene **FN**, z angl. „False negative“) – zachytáva početnosť prípadov, v ktorých skutočná kategória závislej premennej **Y = 1**, bola nesprávne vyhodnotená ako **Y = 0**. (Kelleher et al., 2015)

Pre lepšiu vizualizáciu je matica zámen vyobrazená, ako kontingenčná tabuľka 3.1. Záhlavie riadkov tabuľky odkazuje na predikované binárne hodnoty závislej premennej a záhlavie stĺpcov odkazuje na skutočné binárne hodnoty závislej premennej.

Tabuľka 3.1 Matica zámen

Predikovaná premenná	Cieľová premenná:	Negatívne (0)	Pozitívne (1)
	Negatívne (0)	TN	FN
	Pozitívne (1)	FP	TP

(Zdroj: Vlastné spracovanie)

Tabuľa 3.1. vyobrazuje maticu zámen. Diagonála zvýraznená slabozelenou farbou (hodnoty TN a TP), bližšie konkretizuje správne klasifikované hodnoty cieľovej premennej. Súčet tejto diagonály charakterizuje počet všetkých správne klasifikovaných prípadov. Červená diagonála indikuje prípady, v ktorých bola závislá premenná klasifikovaná nesprávne (hodnoty FN a FP). Analogicky opäť počet všetkých nesprávne klasifikovaných prípadov, je možné vyjadriť súčtom tejto diagonály. (Kelleher, 2015)

Okrem toho, že matica zámen predstavuje užitočnú techniku na mapovanie výkonnosti klasifikačných modelov, tvorí základ pre širokú škálu z nej odvodených výkonnostných metrík. Medzi tie najznámejšie je možné zaradiť, napr.:

– **Správnosť klasifikácie - „Accuracy"** – miera, kvantifikujúca podiel správne klasifikovaných pozorovaní ($TP + TN$), oproti všetkým pozorovaniam predikovanej množiny údajov ($TP + FP + TN + FN$). Správnosť klasifikácie je vyjadrená symbolicky vo vzťahu (3.40). Pokiaľ sa preferuje percentuálne vyjadrenie týchto klasifikačných metrík, je potrebné výslednú hodnotu podielu prenásobiť hodnotou 100. Analogicky mieru chybovosti klasifikácie – „Error rate" možno prepočítať, ako podiel nesprávne klasifikovaných pozorovaní ($FP + FN$), oproti všetkým pozorovaniam ($TP + FP + TN + FN$) predikovanej údajovej množiny. (Hendl, 2021; Obi, 2023)

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{3.40}$$

– **Presnosť klasifikácie – „Precision" –** kvantifikácia pomeru správne klasifikovaných pozitívnych prípadov (TP), v porovnaní so všetkými pozorovaniami predikovanej údajovej množiny, ktoré model klasifikoval ako pozitívne ($TP + FP$). Matematicky vzťah pre výpočet presnosti klasifikácie (3.41). (Hendl, 2021; Obi, 2023)

$$Precision = \frac{TP}{(TP + FP)} \tag{3.41}$$

- **Senzitívnosť klasifikácie –** túto mieru možno v literatúre nájsť aj pod synonymickým označením **„Recall“** alebo **TPR** (z angl. „True Positive Rate“). Senzitívnosť klasifikácie možno rozumieť, ako relatívnu početnosť vyjadrenú pomerom správne klasifikovaných pozitívnych prípadov (TP), oproti všetkým prípadom pozitívnej triedy ($TP + FN$). Vzťah pre výpočet senzitívnosti klasifikácie (3.42). (Hendl, 2021; Singh et al., 2021, Hong, 2021)

$$Senzitívnosť = Recall = TPR = \frac{TP}{(TP + FN)} \tag{3.42}$$

- **Špecifickosť klasifikácie –** niekedy synonymicky označovaná aj ako **TNR** (z angl. „True Negative Rate“). Jedná sa o kvantifikáciu podielu prípadov, ktoré boli správne klasifikované do negatívnej triedy (TN), oproti všetkým prípadom negatívnej triedy ($TN + FP$). Špecifickosť klasifikácie možno charakterizovať matematicky na základe vzťahu (3.43) nižšie. (Singh et al., 2021, Hong, 2021)

$$Špecifickosť = TNR = \frac{TN}{(TN + FP)} \tag{3.43}$$

- **Miera výpadku - „Fall-out“ –** synonymické označovaná **FPR** (z angl. False positive rate) predstavuje mieru, kvantifikujúcu podiel negatívnych prípadov, ktoré model chybne vyhodnotil ako pozitívne (FP), v porovnaní so všetkými pozorovaniami negatívnej triedy ($TN + FP$) predikovanej množiny údajov. Matematická formulácia tejto metriky je znázornená (3.44). (Obi, 2023; Schröder, 2011)

$$FALL - OUT = FPR = \frac{FP}{(TN + FP)} = 1 - Špecifickosť \tag{3.44}$$

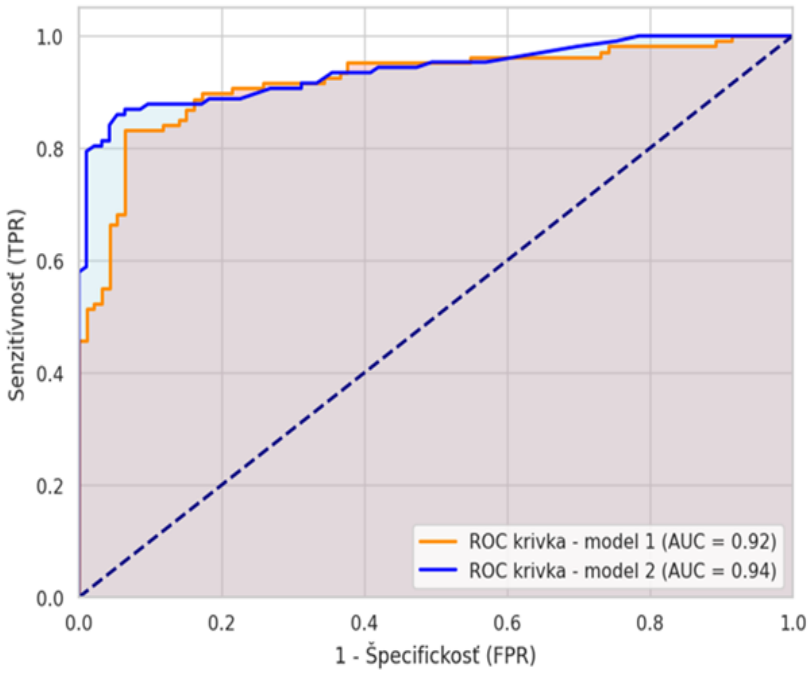
- **F1 Skóre –** miera, charakterizovaná harmonickým priemerom presnosti klasifikácie - „precision“ a senzitívnosti klasifikácie – „recall.“ Matematicky zápis pre výpočet metriky F1 (3.45). (Hendl, 2021)

$$F1 = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \tag{3.45}$$

3.3.4 ROC krivka a hodnota AUC

V úlohách binárnej klasifikácie predstavuje ROC krivka (skrátene z angl. „receiver operator characteristic“) spoľahlivý grafický nástroj, určený na vizualizáciu výkonnostných charakteristík modelu. Výhodou tejto techniky je odhad výkonnosti modelu bez uvedenia špecifickej hodnoty prahového bodu – „threshold“, pričom poskytuje kritéria na výber optimálneho prahového bodu. V prípade multinomickej závislej premennej, počet potenciálnych prahových bodov je o jeden menší, ako počet daných alternatív. Analogicky pre binárnu závislú premennú existuje iba jedna hodnota optimálneho prahového bodu. (Muschelli III, 2020)

Prostredníctvom ROC krivky býva typicky vizualizované, ako sa hodnota miery senzitívnosti (TPR, resp. „recall“) mení, na základe meniacej sa miery špecifickosti modelu [resp. 1 – TNR (špecifickosť), resp. FPR (miera výpadku)], pre rozličné hodnoty prahového bodu prediktora. Na sumarizáciu prediktívnej schopnosti procedúry býva bežne využívaná hodnota pod ROC krivkou – AUC (skrátene z angl. „Area under the curve“). ROC krivka je vizualizovaná na obrázku 3.3. Na osi y je zachytená miera senzitívnosti (TPR, „Recall“) a na osi x tzv. miera „Fall – out,“ (1 - TNR, resp. FPR, 1 - Špecifickosť). (Hendl, 2021; Muschelli III, 2020)



Obrázok 3.3 ROC krivka (Zdroj: Vlastné spracovanie)

Pri pohľade na obr. 3.3 je možné si všimnúť diagonálu, znázornenú prerušovanou čiarou. Diagonála y = x býva označovaná, ako tzv. „referenčná priamka“ a reprezentuje stav,

kedy klasifikačný model generuje rovnakou mierou FP výsledky ako TP výsledky. Dokonalý model dosahuje mieru senzitívnosti a špecifickosti rovnú jednej (reprezentuje ľavý horný roh grafu). Čím je ROC krivka bližšie k ľavému hornému rohu grafu nad referenčnou priamkou, tým je klasifikačná procedúra lepšia. Pre sumarizáciu výsledkov sa vypočíta hodnota AUC. V prípade ak hodnota AUC dosiahne hodnotu 0,5 – hodnota referenčnej krivky, tak reprezentuje test bez schopnosti rozlišovať (t.j. nie lepší ako náhoda), zatiaľ čo AUC hodnota 1,0 – ľavý horný roh grafu, reprezentuje test s dokonalým rozlišovaním. Na obr. 3.3 možno vidieť dve ROC krivky – model 1 (oranžová krivka) a model 2 (modrá krivka). V prípade porovnania modelov prostredníctvom ROC kriviek bolo vypočítané AUC pre oba modely. Možno konštatovať, že modely majú výbornú spoľahlivosť, avšak model 2 dosahuje čiastočne spoľahlivejšie klasifikačné výsledky ($AUC_{model\_1} < AUC_{model\_2}$). (Hendl, 2021; Hoo et al., 2017)

Formálny zápis rovnice na výpočet hodnoty AUC je demonštrovaný vo vzorci (3.46) nižšie.

$$AUC = \int_0^1 f(x)dx \tag{3.46}$$

Vo vzťahu (3.46) predstavuje *f(x)* funkciu ROC krivky. Vzhľadom k tomu, že *f(x)* nemá tendenciu mať integrovateľný tvar ako parabola, odporúčajú sa rôzne metódy aproximácie na výpočet hodnoty AUC. (Bowers & Zhou, 2019)

Vypočítaná hodnotu AUC bude interpretovaná, na základe interpretačnej tabuľky 3.2.

Tabuľka 3.2 Interpretačné intervaly AUC hodnoty

Hodnota plochy pod ROC krivkou (AUC)	Interpretácia výslednej hodnoty
0,9 ≤ AUC	**Výborná**
0,8 ≤ AUC < 0,9	**Dobrá**
0,7 ≤ AUC < 0,8	**Priemerná**
0,6 ≤ AUC < 0,7	**Zlá**
0,5 ≤ AUC < 0,6	**Neúspech**

(Zdroj: Vlastné spracovanie na základe Wahono & Suyana (2014))

Tabuľa 3.2 klasifikuje vypočítanú hodnotu AUC do 5 základných intervalov, ktoré odkazujú na predikčnú spoľahlivosť modelu – výborná $\langle 0{,}9 \leq$ AUC), dobrá $\langle 0{,}8 \leq$ AUC $< 0{,}9)$, priemerná $\langle 0{,}7 \leq$ AUC $< 0{,}8)$, zlá $\langle 0{,}6 \leq$ AUC $< 0{,}7)$ a neúspešná klasifikácia $\langle 0{,}5 \leq$ AUC $< 0{,}6)$.

4 VÝSLEDKY PRÁCE A ZHODNOTENIE

Praktická časť práce prezentovaná v obsahu tejto kapitoly bola venovaná aplikácií nelineárnych pravdepodobnostných modelov, na historických údajoch o klientoch nemenovanej portugalskej bankovej inštitúcie. Tí participovali v rámci marketingovej kampane, ktorej cieľom bolo presvedčiť týchto klientov, aby vložili svoje voľné finančné prostriedky na termínovaný vklad. Modely budú aplikované na týchto údajoch, s cieľom predikovať príslušnosť klienta do úspešnej alebo neúspešnej marketingovej skupiny. Ich pomocou budú následne odhalené aj kľúčové determinanty možnej úspešnosti kampane u konkrétneho klienta.

Kapitola je segmentovaná do troch podkapitol. V rámci prvej podkapitoly sú zahrnuté špecifikácie, týkajúce sa použitých údajov a spôsobu ich spracovania. Druhá podkapitola je venovaná explorácií štruktúry spracovaných údajov, vrátane preverenia vzájomných vzťahov medzi nezávislými premennými. Záver kapitoly je venovaný aplikačnej časti, v ktorej je detailne zachytený spôsob rozdelenia základného súboru údajov, výber nezávislých premenných a následný výber najvhodnejšieho modelu logit a probit, za účelom predikcie pravdepodobnosti úspechu kampane. Ďalej bude opísaný spôsob vykonania predikcií pravdepodobnosti a následná klasifikácia klientov do príslušných skupín. Po validácií klasifikačných charakteristík testovaných modelov nasleduje interpretácia dosiahnutých výsledkov praktickej časti práce.

4.1 Údaje

Údaje o klientoch zapojených v marketingovej kampani, ktoré boli použité pri spracovaní praktickej časti práce, sú verejne dostupné online na viacerých webových sídlach. Pri spracovaní práce boli použité údaje, uverejnené na webovom sídle spoločnosti Kaggle,[1] ktorá predstavuje dcérsku spoločnosť spoločnosti Google LLC. Tie boli na webovom sídle dostupné v dvoch súboroch vo formáte „CSV," ktoré boli dopredu rozdelené na trénovaciu a testovaciu množinu údajov. Konkrétne ide o súbory s názvami *„train.csv" a „test.csv."* Súbor údajov s názvom *„train.csv"* obsahoval celkovo 45 211 pozorovaní, určených na trénovanie modelov. Súbor údajov s názvom *„test.csv"* mal v sebe obsiahnutých 4 521 pozorovaní, určených na testovanie už natrénovaných modelov. Tento súbor vznikol

[1] Údaje využité pri spracovaní diplomovej práce: https://www.kaggle.com/datasets/prakharrathi25/banking-dataset-marketing-targets

náhodným výberom 10 % pozorovaní zo súboru údajov, určených na trénovanie modelov (*„train.csv“*). Rozdelenie základného súboru údajov na trénovaciu a testovaciu množinu bolo realizované prostredníctvom programovacieho jazyka R. V dôsledku toho sa v praktickej časti práce vychádzalo len z údajov obsiahnutých v súbore *„train.csv.“*

Zber údajov prebiehal v období od mája 2008 – do novembra 2010. Základný súbor údajov (nespracovaný) tvorilo 17 atribútov, ktoré možno zoskupiť do viacerých podkategórií, ako napr.:

1. **Premenné sociodemografických indikátorov**, medzi ktoré možno zaradiť vek klienta, jeho zamestnanie, rodinný stav a najvyššie dosiahnuté vzdelanie.
2. **Premenné indikujúce solventnosť klienta**, resp. či je klient schopný splácať svoje záväzky voči banke. Tieto atribúty bližšie špecifikujú zadlženosť klienta v tom zmysle, či je klient v súčasnom stave platiteľom pôžičky na bývanie, osobnej pôžičky alebo či sa v minulosti klient dostal do tzv. „defaultného“ stavu, kedy nebol schopný splácať svoje záväzky voči banke. Do tejto podkategórie možno zaradiť aj výšku finančných prostriedkov na bankovom účte osloveného klienta.
3. **Premenné špecifikujúce spôsob oslovenia klienta banky** v rámci marketingovej kampane. Atribúty, ktoré sú potrebné na priblíženie toho, akým spôsobom bol klient banky kontaktovaný, v akom dátume bol klient naposledy kontaktovaný (deň a mesiac kontaktu) a dĺžka trvania komunikácie s klientom.
4. **Premenné špecifikujúce kampaň** charakterizujú, koľkokrát bol klient kontaktovaný v rámci aktuálnej a predchádzajúcich marketingových kampaní, rovnako tak aj koľko dní uplynulo od posledného kontaktovania tohto klienta alebo výsledná úspešnosť predošlej kampane u daného klienta.
5. Cieľová premenná štúdie, resp. **závislá premenná modelu**, pojednáva o výsledku aktuálnej marketingovej kampane. Úspechom kampane možno rozumieť situáciu, v ktorej klient vložil svoje voľné aktíva na termínovaný vklad banky. Analogicky neúspech predstavuje alternatívu, v ktorej sa klient rozhodol odmietnuť ponuku banky.

Spôsob, akým bola premenná označená v rámci súboru údajov, vrátane jej bližšej špecifikácie, typu konkrétnej premennej, jej unikátne obmeny v rozmedzí nespracovaného súboru údajov a ich charakteristika, je podrobnejšie priblížená v prílohe A. Ako je možné si všimnúť, súbor údajov pred spracovaním obsahuje 7 kvantitatívnych a 10 kategoriálnych premenných (z toho 4 binárne, resp. dichotomické). Atribúty v poradí 1. – 16. reprezentujú

nezávislé (vysvetľujúce) premenné. V poradí 17. atribút predstavuje závislú premennú modelov.

4.1.1 Spracovanie údajov

Táto podkapitola práce mapuje proces spracovania údajov, ktorý je kľúčový pred samotným vývojom a aplikáciou modelov. Vo všeobecnosti proces spracovania údajov zahŕňa niekoľko krokov, ktorých cieľom je zabezpečiť, aby súbor údajov neobsahoval chybné alebo irelevantné údaje. Tie môžu negatívne ovplyvniť presnosť a spoľahlivosť výsledkov analýzy.

V prvom kroku spracovania údajov bolo potrebné zistiť, či súbor údajov neobsahuje neznáme hodnoty, resp. chýbajúce hodnoty. Tie bývajú vo všeobecnosti symbolicky označované ako „NA." Ich prítomnosť v súbore údajov môže byť zavinená chybou záznamu, ktorá nastala pri zbere údajov. Ďalší faktor, ktorý by mohol negatívne ovplyvniť presnosť modelu predstavuje prítomnosť duplicitných pozorovaní.

Možno zhodnotiť, že v rámci základného súboru údajov neboli detegované žiadne duplicitné pozorovania a na prvý pohľad neboli odhalené ani chýbajúce, resp. neznáme hodnoty „NA." Je však potrebné zdôrazniť, že prostredníctvom preskúmania unikátnych hodnôt kategoriálnych atribútov boli detegované neznáme hodnoty (označenie – „unknown") a hodnoty, symbolizujúce iné obmeny, ako štandardizované hodnoty prieskumu (označenie – „others"). Prítomnosť spomínaných hodnôt možno zaznamenať v atribúte „Job," symbolizujúcom zamestnanie osloveného klienta (288 hodnôt „unknown"); v atribúte „Education," - špecifikujúcom najvyššie dosiahnuté vzdelanie klienta (1 857 hodnôt „unknown"); v atribúte „Contact" – spôsob oslovenia klienta v rámci kampane (13 020 hodnôt „unknown") a v atribúte „Poutcome," ktorý špecifikuje výsledok predošlej kampane (1 840 hodnôt „others" a 36 959 hodnôt „unknown").

Agregované percentuálne zastúpenie detegovaných irelevantných hodnôt v rámci nezávislej premennej v základnom súbore údajov možno pozorovať v tabuľke 4.1.

Tabuľka 4.1 Irelevantné hodnoty

Atribút:	Job	Education	Contact	Poutcome
Percentuálne zastúpenie	0,64 %	4,11 %	28,80 %	85,82 %

Zdroj: Vlastné spracovanie

Vzhľadom na nízke percentuálne zastúpenie irelevantných hodnôt (viď tabuľka 4.1) v rámci atribútov „Job" a „Education," budú v týchto odstránené iba pozorovania,

obsahujúce pre analýzu irelevantné hodnoty. Odlišná situácia nastáva v atribúte „Contact," v ktorom tvoria irelevantné hodnoty takmer tretinu všetkých pozorovaní a v atribúte „Poutcome," kde sú irelevantné údaje zaznamenané u takmer 86 % všetkých pozorovaní. Tieto premenné boli zo základného súboru údajov vylúčené.

Zo základného súboru boli vylúčené aj atribúty špecifikujúce deň oslovenia klienta v rámci kampane – „Day" a mesiac, v ktorom bol klient oslovený – „Month." Rozhodnutie nezaradiť dátumové atribúty do analýzy možno zdôvodniť tým, že v súbore údajov nie je prítomná informácia o roku, v ktorom bol záznam uskutočnený. Vzhľadom k tomu, že kampaň trvala dlhšie ako dva roky, interpretovateľnosť dátumových atribútov bez uvedenia roku možno považovať za nevhodné.

V tabuľke 4.2 sú zachytené špecifikácie týkajúce sa základného súboru údajov pred spracovaním a po spracovaní.

Tabuľka 4.2 Porovnanie dimenzionality a počtu pozorovaní nespracovaného a spracovaného súboru údajov

Základný súbor	Nespracovaný súbor	Spracovaný súbor
Počet pozorovaní:	45 211	43 193
Počet atribútov:	17	13

Zdroj: Vlastné spracovanie

Ako vyplýva z tabuľky 4.2, pri procese spracovania údajov bolo zo základného súboru odstránených 2 018 pozorovaní, pričom počet atribútov sa zredukoval z pôvodných sedemnástich na trinásť.

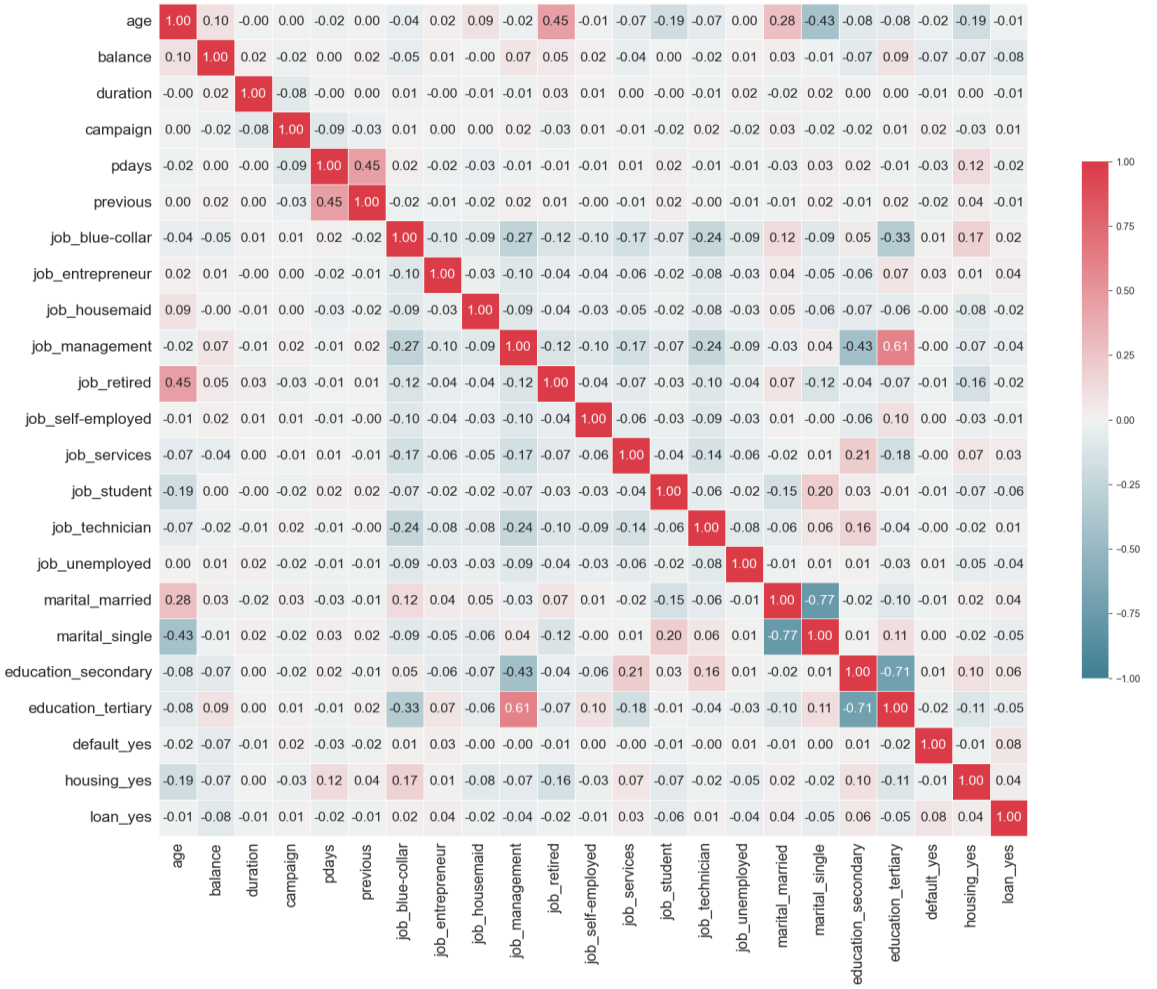
4.2 Exploratívna analýza údajov

V rámci tejto podkapitoly bude uskutočnená exploratívna analýza údajov, ktorej cieľom je priniesť cenné poznatky o štruktúre a povahe údajov. Podkapitola stručne načrtne sociodemografický profil oslovených klientov, profil solventnosti oslovených klientov a profil priebehu kampane. Štruktúra kvantitatívnych premenných bola skúmaná prostredníctvom popisných štatistík (početnosť, priemer, smerodajná odchýlka, minimálna a maximálna hodnota, dolný kvartil, medián a horný kvartil) a prostredníctvom vizualizačných nástrojov – histogram a krabicový graf. V prípade kvantitatívnych údajov bola ich štruktúra skúmaná prostredníctvom kontingenčných tabuliek. Na preskúmanie vzájomných vzťahov medzi nezávislými premennými bola použitá korelačná matica, mapujúca ich vzájomnú mieru korelácie.

Vzhľadom k výraznej obsahovej robustnosti je explorácia štruktúry údajov zahrnutá v rámci príloh diplomovej práce. Náčrt sociodemografického profilu osloveného klienta je obsiahnutý v prílohe B, profil klientovej solventnosti – príloha C a profil priebehu marketingovej kampane je súčasť prílohy D.

4.2.1 Preverenie vzťahov medzi premennými

Za účelom preverenia vzájomných vzťahov, t.j. korelácií medzi nezávislými premennými bola využitá korelačná matica. Tá je vizualizovaná na obrázku 4.1. Premenné disponujúce veľmi silnou koreláciou sú zvýraznené tmavočervenou farbou (v kladnom smere) alebo tmavomodrou (v negatívnom smere). Analogicky podľa stupnice, bledá farba prislúcha premenným, medzi ktorými je len veľmi slabá alebo žiadna závislosť



Obrázok 4.1 Korelačná matica regresorov (Zdroj: Vlastné spracovanie)

Na základe korelačnej matice možno skonštatovať, že vo väčšine nezávislých premenných neexistuje priama závislosť, resp. sa vyznačuje len vo veľmi malej miere. Nájdu sa však aj prípady, u ktorých korelácia naznačuje pomerne silnú závislosť. Tú možno

detegovať medzi jednotlivými obmenami kategoriálnych premenných v negatívnom smere. Jedná sa o premennú charakterizujúcu vzdelanie klienta („Education“), pričom negatívna závislosť existuje medzi klientami s najvyšším dosiahnutým vzdelaním stredoškolským („secondary“) a vysokoškolským („tertiary“). Rovnako existuje silná negatívna závislosť medzi obmenami premennej definujúcej rodinný stav klienta („Marital“), konkrétne medzi slobodnými klientami („single“) a tými, ktorí žijú v manželstve („married“). Silná kladná závislosť bola naopak odhalená medzi klientami pracujúcimi na manažérskej pozícií (premenná - „Job“, obmena „management“) a klientami s ukončeným vysokoškolským vzdelaním. Aby sa predišlo multikolinearite, nemá veľký zmysel agregovať protichodné obmeny kategorických premenných do jednej spoločnej (nelogickosť interpretácií). Nepriaznivé by mohlo byť aj ich prosté odstránenie, keďže by sa stratili tisícky zápisov pozorovaní.

4.3 Aplikácia modelov

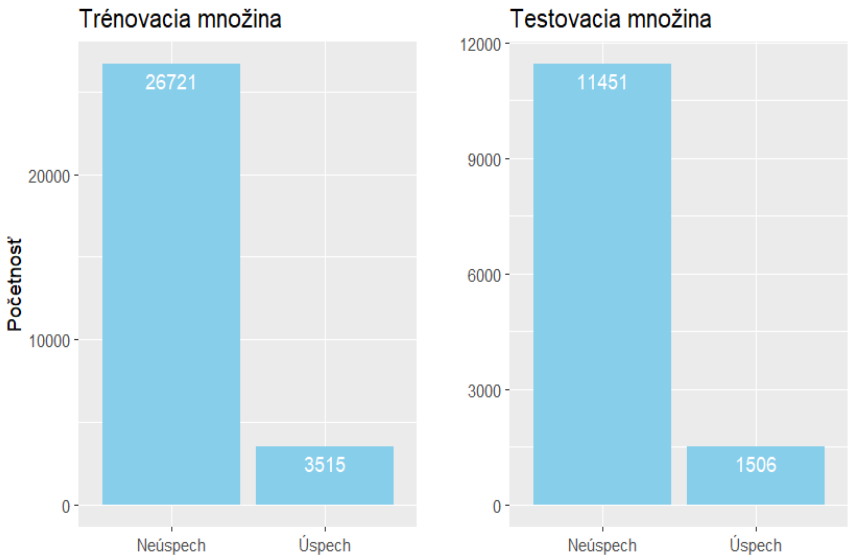
V tejto podkapitole bude detailne zmapovaný proces aplikácie logistickej a probitovej regresie, s cieľom predikovať pravdepodobnosť úspechu kampane u jednotlivých klientov. Za týmto účelom bola kapitola segmentovaná do piatich samostatných podkapitol. Prvá podkapitola opisuje spôsob rozdelenia základného súboru údajov do dvoch samostatných podmnožín – trénovacia množina (odhad modelov), testovacia množina (testovanie modelov, zároveň symbolizuje nové údaje, na ktorých budú vykonané predikcie pravdepodobnosti). Druhá podkapitola zdôvodňuje výber vhodných regresorov pre oba pravdepodobnostné modely. Taktiež zahŕňa odhady viacerých modelov, obsahujúcich rôzne kombinácie zvolených regresorov, z ktorých sa bude vyberať najvhodnejší model logit a najvhodnejší model probit. To na základe porovnania hodnôt AIC, BIC, $R^2_{McFadden}$. Testovanie štatistickej významnosti parametrov bude založené na Waldovom teste. Taktiež sa otestuje významnosť modelov ako celku prostredníctvom testu pomeru vierohodnosti. Tretia podkapitola opisuje spôsob uskutočnenia predikcií pravdepodobnosti a následnú klasifikáciu klientov do úspešnej alebo neúspešnej skupiny. Na zmapovanie výsledkov predikcií poslúži matica zámen. Štvrtá podkapitola sa venuje validácií klasifikačnej výkonnosti oboch modelov, prostredníctvom metrík: správnosť, presnosť, senzitívnosť, špecifickosť, výpadok, F1 Skóre. Zároveň budú pre oba modely skonštruované ROC krivky, sumarizované prostredníctvom AUC indexu. Posledná piata podkapitola je venovaná interpretácií dosiahnutých výsledkov analýzy.

4.3.1 Rozdelenie spracovaného súboru údajov

Rozdelenie základného súboru údajov bolo potrebné realizovať ešte pred samotným výberom vhodných premenných a následnou aplikáciou modelov. To z dôvodu, že pozorovania obsiahnuté v trénovacej množine predstavujú údaje, prostredníctvom ktorých boli odhadované parametre modelov. Pozorovania obsiahnuté v testovacej množine boli použité na vykonanie predikcií, ktorých úspešnosť sa vyhodnocovala prostredníctvom na to určených validačných techník.

V spracovanom súbore údajov bolo obsiahnutých celkovo 43 193 pozorovaní. Ten bol rozdelený náhodným výberom v pomere 70:30, pričom trénovacia množina obsahuje 70 % všetkých pozorovaní základného súboru, zatiaľ čo zvyšných 30 % pozorovaní prislúcha testovacej množine.

Pre lepšiu predstavu sú obe množiny vizualizované na obrázku 4.2.



Obrázok 4.2 Absolútna početnosť kategórií závislej premennej - trénovacia a testovacia množina (Zdroj: Vlastné spracovanie)

Už na prvý pohľad je zrejmé, že obe množiny údajov disponujú značne neproporčným rozdelením pozorovaní v rámci oboch kategórií závislej premennej. Táto skutočnosť je logická od samej podstaty riešeného problému. Na druhú stranu takáto nevyváženosť alternatív medzi klientami môže ovplyvniť výsledky predikcií v zmysle, že model bude mať tendenciu minoritnú triedu vyhodnocovať ako majoritnú. Za účelom prvotného zmapovania výsledkov predikcií oboch modelov bola použitá matica zámen, na základe ktorej sa následne prepočítajú výkonnostné metriky klasifikácie.

4.3.2 Výber nezávislých premenných a následný odhad modelov

V tejto podkapitole bude opísaná realizácia výberu vhodných nezávislých premenných, na základe ktorých boli zostavené modely logistickej a probitovej regresie. Výhoda modelov obsahujúcich iba malý počet nezávislých premenných je spojená s jednoduchosťou interpretácií dosiahnutých výsledkov. Nevýhody spojené s modelmi založených na menšom počte nezávislých premenných môže predstavovať skutočnosť, že môžu byť prehliadané faktory, ktoré majú výrazný vplyv pri modelovaní závislosti, čím môže dôjsť k značnej chybovosti predikcií.

Negatívny vplyv na predikčnú schopnosť modelu môže byť zapríčinený aj jeho vysokou mierou komplexnosti. Modely založené na príliš rozsiahlej množine atribútov môžu viesť k tzv. preučeniu modelu. Ide o situáciu, v ktorej model dokáže predikovať s pomerne vysokou úspešnosťou údaje, na základe ktorých bol odhadnutý, resp. natrénovaný, ale výrazne zlyháva jeho schopnosť správne vyhodnotiť údaje obsiahnuté v testovacej množine. (Kuhn & Johnson, 2013) Zohľadňujúc obe vyššie uvedené skutočnosti, pre výber premenných modelov bol zvolený algoritmus „Stepwise“ po vzore autorov Ďurica et al. (2019) a Klieštik & Kováčová (2017). V tomto prípade však algoritmus neselektoval premenné na základe ich štatistickej významnosti, teda podľa p – hodnoty, ale na základe informačných kritérií (AIC, BIC). Za účelom aplikácie tohto algoritmu bol využitý balíček programovacieho jazyka R s názvom *MASS*. Konkrétne išlo o funkciu *stepAIC()*. Dôvodom výberu tohto algoritmu bola snaha o nájdenie čo možno najvhodnejšej lineárnej kombinácie nezávislých premenných, ktoré dokážu presne objasniť závislosť medzi možnou zaznamenanou úspešnosťou kampane u klienta a tým zabezpečiť správnosť jednotlivých predikcií. Výber premenných bol založený na základe:

- **sociodemografického profilu** (**A:** „Age,“ **J_j :** „Job,“ **M_j :** „Marital,“ **E_j:** „Education“);
- **profilu klientovej solventnosti** (**Def_j:** „Default“, **B:** „Balance,“ **H_j:** „Housing,“ **L_j:** „Loan“);
- **profilu marketingovej kampane** (**Dur:** „Duration,“ **C:** „Campaign,“ **Pd:** „Pdays,“ **Prev:** „Previous“).

Algoritmus „Stepwise“ pri výbere nezávislých premenných bol použitý v dvoch variantoch:

- **„Forward selection“ –** ide o iteratívny prístup, v ktorom sa do modelu bez prediktorových premenných, t.j. obsahujúceho iba úrovňovú konštantu β_0 postupne

pridávajú premenné v krokoch. Odhadnutý model sa následne porovnáva na základe informačného kritéria – AIC a BIC. Kritérium s najnižšou hodnotou indikuje najlepší model.

– **„Backward elimination“ –** v tomto variante iterácie začínajú na modeli so všetkými prediktorovými premennými. Postupným vyradzovaním premenných sa jednotlivé odhadnuté modely opäť vyhodnocujú na základe informačných kritérií – AIC a BIC. Vybraný model disponuje najnižšou hodnotou týchto kritérií.

Výber nezávislých premenných prostredníctvom oboch variantov pre modely logit a probit je zaznamenaný v tabuľke 4.3

Tabuľka 4.3 Tabuľka 4.3 Výber nezávislých premenných modelov - procedúra Stepwise

Výber premenných	Kritérium výberu	Logit model	Probit model
„Backward elimination“	**AIC**	Vylúčené: -A	Vylúčené: -A
	BIC	Vylúčené: -A, - Def_j, -B	Vylúčené: -A, - Def_j
„Forward selection“	**AIC**	Zahrnuté: Všetky premenné	Zahrnuté: Všetky premenné
	BIC	Zahrnuté: Všetky premenné	Zahrnuté: Všetky premenné

Zdroj: Vlastné spracovanie

Pre porozumenie výstupom odhadnutých modelov je nutné spomenúť, že pre prácu s kategoriálnymi premennými v programovacom jazyku R, bolo potrebné tieto hodnoty definovať ako faktor. Následne sa vytvorí pre danú kategoriálnu premennú k – 1 umelých premenných, kde „k“ predstavuje počet obmien kategoriálneho regresora. Za hodnoty umelých premenných sa následne dosadí hodnota 1 v prípade, ak pozorovanie je definované v danej kategórií. Hodnota 0 sa dosadí, ak pozorovanie nepatrí do danej kategórie. Premenné nezahrnuté vo výstupe predstavujú referenčnú kategóriu, s ktorou sa porovnáva ostatných k – 1 umelých premenných. V prípade dichotomických premenných je pomerne jednoduché sa zorientovať v tom, ktorá premenná vystupuje ako referenčná. Avšak, v prípade premennej multinomickej to nemusí byť zrejmé na prvý pohľad. Za týmto účelom bola skonštruovaná tabuľka 4.4, v ktorej je obsiahnuté pracovné označenie referenčnej kategórie, jej definované hodnoty a príslušná množina porovnávaných kategórií.

Tabuľka 4.4 Definovanie referenčnej kategórie k množine porovnávaných kategórií

Označenie	Hodnota referenčnej kategórie	Množina porovnávaných kategórií
J_0:	**Job:** *admin*	**{J_1, J_2, ... , J_{10}}**
M_0:	**Marital:** *divorced*	**{M_1, M_2}**
E_0:	**Education:** *primary*	**{E_1, E_2}**
Def_0:	**Default:** *no*	**{Def_1}**
H_0:	**Housing:** *no*	**{H_1}**
L_0:	**Loan:** *no*	**{L_1}**

Zdroj: Vlastné spracovanie

Ako vyplýva z tabuľky 4.4, referenčná kategória pre J_j je definovaná v obmene zamestnania – administratívny pracovník, pre M_j je definovaná v obmene rodinného stavu – rozvedený, pre E_j je definovaná v obmene dosiahnutého vzdelania – základné. Ostatné premenné sú dichotomické, pričom ich referenčná kategória je nastavená na zápornú alternatívu skúmaného stavu.

Tabuľky 4.5 a 4.6 zachytávajú odhadnuté hodnoty parametrov oboch modelov prostredníctvom metódy maximálnej vierohodnosti, pre vybrané nezávislé premenné z tabuľky 4.3. Hodnoty v tabuľke prezentované v zátvorkách predstavujú smerodajnú odchýlku odhadnutého parametra. Test štatistickej významnosti parametrov je založený na testovaní Waldovej z – štatistiky, pričom symboly „ *, **, *** “ uvádzajú príslušnú hladinu štatistickej významnosti testovaných parametrov, na základe výslednej p - hodnoty. V päte tabuľky sú uvedené hodnoty prirodzeného logaritmu funkcie vierohodnosti (l), AIC (vzťah 3.38), BIC (vzťah 3.39) a $R^2_{McFadden}$ (vzťah 3.37). Tie budú použité na výber najvhodnejšieho modelu logit a probit z troch dostupných kandidátov.

Tabuľka 4.5 Logit modely – odhadnuté

Logit modely									
Y = Úspešnosť		(1)	(2)	(3)			(1)	(2)	(3)
	β_0	-2,799***	-2,755***	-2,739***	M_1:	**marital:**	-0,097	-0,099	-0,092
		(0,166)	(0,114)	(0,113)		*married*	(0,067)	(0,067)	(0,067)
A:	**age**	0,001	-	-	M_2:	**marital:**	0,156*	0,147*	0,148*
		(0,003)				*single*	(0,076)	(0,072)	(0,072)
J_1:	**job:**	-0,515***	-0,517***	-0,520***	E_1:	**education:**	0,232**	0,229**	0,228**
	blue-collar	(0,083)	(0,083)	(0,083)		*secondary*	(0,073)	(0,073)	(0,073)
J_2:	**job:**	-0,802***	-0,800***	-0,808***	E_2:	**education:**	0,585***	0,581***	0,591***
	entrepreneur	(0,153)	(0,153)	(0,153)		*tertiary*	(0,085)	(0,084)	(0,084)
J_3:	**job:**	-0,759***	-0,755***	-0,757***	Def_1:	**default:**	-0,441*	-0,442*	
	housemaid	(0,161)	(0,160)	(0,160)		*yes*	(0,200)	(0,200)	-
J_4:	**job:**	-0,286***	-0,284***	-0,283***	B:	**balance**	0,00002**	0,00002**	
	management	(0,083)	(0,083)	(0,083)			(0,00001)	(0,00001)	-
J_5:	**job:**	0,340**	0,357***	0,373***	H_1:	**housing:**	-1,114***	-1,116***	-1,121***
	retired	(0,108)	(0,097)	(0,097)		*yes*	(0,046)	(0,046)	(0,046)
J_6:	**job:**	-0,374**	-0,373**	-0,369**	L_1:	**loan:**	-0,616***	-0,617***	-0,641***
	self-employed	(0,126)	(0,126)	(0,125)		*yes*	(0,069)	(0,069)	(0,068)
J_7:	**job:**	-0,398***	-0,399***	-0,402***	Dur:	**duration**	0,004***	0,004***	0,004***
	services	(0,095)	(0,095)	(0,095)			(0,0001)	(0,0001)	(0,0001)
J_8:	**job:**	0,658***	0,650***	0,654***	C:	**campaign**	-0,126***	-0,126***	-0,127***
	student	(0,125)	(0,123)	(0,123)			(0,012)	(0,012)	(0,012)
J_9:	**job:**	-0,301***	-0,301***	-0,305***	Pd:	**pdays**	0,003***	0,003***	0,003***
	technician	(0,078)	(0,078)	(0,078)			(0,0002)	(0,0002)	(0,0002)
J_{10}:	**job:**	-0,287*	-0,286*	-0,286*	Prev:	**previous**	0,091***	0,091***	0,092***
	unemployed	(0,128)	(0,128)	(0,128)			(0,010)	(0,010)	(0,009)
Pozorovania:							30 236	30 236	30 236
l:							-8126,267	-8126,333	-8133,789
AIC							16300,533	**16298,666**	16309,578
BIC							16500,136	16489,952	**16484,23**
R^2_{McFadden}							**0,252**	**0,252**	0,251
Poznámka:									*p<0,05; **p<0,01; ***p<0,001

Zdroj: Vlastné spracovanie

Tabuľka 4.6 Probit modely – odhadnuté

Probit modely									
Y = Úspešnosť		(1)	(2)	(3)			(1)	(2)	(3)
	β_0	-1,593***	-1,580***	-1,583***	M_1:	**marital:**	-0,052	-0,053	-0,051
		(0,087)	(0,060)	(0,059)		*married*	(0,035)	(0,035)	(0,035)
A:	**age**	0,0003			M_2:	**marital:**	0,083*	0,080*	0,080*
		(0,001)	-	-		*single*	(0,040)	(0,038)	(0,038)
J_1:	**job:**	-0,276***	-0,276***	-0,278***	E_1:	**education:**	0,128***	0,127***	0,127***
	blue-collar	(0,043)	(0,043)	(0,043)		*secondary*	(0,038)	(0,038)	(0,038)
J_2:	**job:**	-0,391***	-0,390***	-0,396***	E_2:	**education:**	0,305***	0,304***	0,306***
	entrepreneur	(0,078)	(0,078)	(0,077)		*tertiary*	(0,045)	(0,044)	(0,044)
J_3:	**job:**	-0,368***	-0,366***	-0,368***	Def_1:	**default:**	-0,241*	-0,242*	-
	housemaid	(0,081)	(0,081)	(0,081)		*yes*	(0,101)	(0,101)	
J_4:	**job:**	-0,148***	-0,148***	-0,150***	B:	**balance**	0,00001***	0,00001***	0,00001***
	management	(0,044)	(0,044)	(0,044)			(0,00000)	(0,00000)	(0,00000)
J_5:	**job:**	0,199***	0,204***	0,204***	H_1:	**housing:**	-0,568***	-0,568***	-0,568***
	retired	(0,059)	(0,053)	(0,053)		*yes*	(0,024)	(0,024)	(0,024)
J_6:	**job:**	-0,200**	-0,200**	-0,203**	L_1:	**loan:**	-0,317***	-0,317***	-0,322***
	self-employed	(0,067)	(0,067)	(0,067)		*yes*	(0,035)	(0,035)	(0,035)
J_7:	**job:**	-0,223***	-0,223***	-0,223***	Dur:	**duration**	0,002***	0,002***	0,002***
	services	(0,050)	(0,050)	(0,050)			(0,00004)	(0,00004)	(0,00004)
J_8:	**job:**	0,379***	0,376***	0,377***	C:	**campaign**	-0,061***	-0,061***	-0,061***
	student	(0,071)	(0,070)	(0,070)			(0,006)	(0,006)	(0,006)
J_9:	**job:**	-0,158***	-0,158***	-0,159***	Pd:	**pdays**	0,002***	0,002***	0,002***
	technician	(0,041)	(0,041)	(0,041)			(0,0001)	(0,0001)	(0,0001)
J_{10}:	**job:**	-0,146*	-0,146*	-0,147*	Prev:	**previous**	0,052***	0,052***	0,052***
	unemployed	(0,068)	(0,068)	(0,068)			(0,005)	(0,005)	(0,005)
Pozorovania:							30 236	30 236	30 236
l:							-8109,341	-8109,364	-8112,404
AIC							16266,682	**16264,728**	16268,808
BIC							16466,285	16456,014	**16451,778**
R^2_{McFadden}							**0,254**	**0,254**	0,253
Poznámka:									*p<0,05; **p<0,01; ***p<0,001

Zdroj: Vlastné spracovanie

4.3.3 Výber vhodného modelu logit a probit

Pri výbere najvhodnejšieho logit a probit modelu, ktoré boli použité na predikciu pravdepodobnosti úspechu kampane u klientov, sa vychádzalo z vypočítaných hodnôt informačných kritérií – AIC, BIC a nepravého koeficientu determinácie $R^2_{McFadden}$. Pre modely logit možno nájsť ich výsledné hodnoty v päte tabuľky 4.5, v prípade modelu probit sú uvedené v päte tabuľky 4.6.

V prípade modelu logit bol na predikciu pravdepodobnosti vybraný druhý model. Zohľadňujúc skutočnosť, že v prípade tohto modelu vyšla väčšina porovnávajúcich metrík v požadovaných hodnotách. Hodnota AIC (16298,666) v prípade tohto modelu dosahovala najnižšie skóre, zo všetkých modelov. Rovnako tak $R^2_{McFadden}$ dosahovalo najvyššiu hodnotu (0,252), práve v tomto modeli a v modeli (1). Ten však obsahoval štatisticky nevýznamný parameter premennej **A**. V modeloch (2 - 3) bol detegovaný nárast hladiny štatistickej významnosti parametru $\mathbf{J_5}$ z 1 % na 0,1 % hladinu významnosti. Dôležité je spomenúť, že vo všetkých troch modeloch figuruje štatisticky nevýznamný parameter kategórie $\mathbf{M_1}$ čo znamená, že sa jeho hodnota výrazne neodlišuje od hodnoty parametra $\mathbf{M_0}$. Kategória $\mathbf{M_2}$ je vo všetkých modeloch štatisticky významná na hladine významnosti 5 %. Na hladine významnosti 5 % sa ukázali ako štatisticky významné aj parametre kategórie $\mathbf{J_{10}}$ (všetky modely) a kategórie $\mathbf{Def_1}$ [iba model (1 – 2), v modeli (3) nefiguruje]. Na hranici významnosti 1 % sa ukázali ako štatisticky významné parametre kategórie $\mathbf{J_6}$ a premennej **B** (všetky modely). Ostatné parametre sú štatisticky významné na hladine významnosti 0,1 % vo všetkých modeloch.

Pri výbere modelu probit bola situácia rovnaká. Na predikciu pravdepodobnosti bol zvolený druhý model, vzhľadom na najnižšiu hodnotu AIC (16264,728) a najvyššiu hodnotu $R^2_{McFadden}$ (0,254), ktorá bola rovnaká aj v prípade modelu (1). Ten však zohľadňoval štatisticky nevýznamný atribút **A.** Na rozdiel od modelu logit (1), parameter kategórie $\mathbf{J_5}$ bol v modeli probit (1) štatisticky významný na hladine významnosti 0,1 %. Na hladine významnosti 1 % sa v tomto prípade ukázal štatisticky významný iba parameter kategórie $\mathbf{J_6}$ (všetky modely). V prípade všetkých probitových modelov sa parameter premennej **B** ukázal ako štatisticky významný na hladine významnosti 0,1 % (v modeloch logit bola významnosť na hladine 1 %). Vo všetkých troch probitových modeloch figuruje štatisticky nevýznamný parameter kategórie $\mathbf{M_1}$ a teda neexistuje významný rozdiel medzi jeho

hodnotou a hodnotou referenčnej kategórie . Parametra kategórie **M_2** sa opäť preukázal ako štatisticky významný vo všetkých modeloch (hladina významnosti 5 %).

Na tejto hladine významnosti sa ukázali ako významné aj parametre kategórie **J_{10},** a **Def_1** [probit (1 - 2), v modeli probit (3) nefiguruje]**.** Ostatné parametre zahrnuté v modeloch probit sa ukázali ako štatisticky významné na hladine významnosti 0,1 %.

V tabuľkách 4.5 a 4.6 boli jednotlivé odhady parametrov testované prostredníctvom Waldovho testu, ktorý testoval $H_0: \beta_j = 0$; resp. či jednotlivé parametre modelu sú rovné nulovej hodnote. V praxi sa však odporúča testovanie modelu ako celku. Za týmto účelom boli modely testované prostredníctvom testu pomeru vierohodnosti (LR_{test}). Ten testuje $H_0: \beta_1 = \ldots = \beta_j = 0$; teda či všetky parametre modelu ako celku sa súčasne rovnajú nule. Výsledky LR_{test} všetkých odhadnutých modelov sú zachytené v tabuľke 4.7. V záhlaví tabuľky sú zachytené stupne voľnosti [DF(q)], prirodzený logaritmus funkcie vierohodnosti (l), zmena stupňov voľnosti [Δ Df(q)], testovaná G – štatistika [G (χ2)], výsledná p – hodnota [Pr(> χ2)] a štatistická významnosť jednotlivých modelov (ŠV). V prípade označenia modelov, model (0) zachytáva informácie o modeli obsahujúcom iba úrovňovú konštantu β_0, ktorý figuruje pri výpočte testovacej G – štatistiky podľa vzťahu 3.29. Označenie ostatných modelov je totožné s ich označením v tabuľkách 4.5 a 4.6.

Tabuľka 4.7 Výsledky testu pomeru vierohodnosti (LR test)

Model:	DF(q)	l	Δ Df(q)	$G\ (\chi^2)$	Pr(> χ^2)	ŠV
Logit						
(0)	1	-10866,5	-	-	-	-
(1)	24	-8126,27	23	5480,532	<0,001	***
(2)	23	-8126,33	22	5480,399	<0,001	***
(3)	21	-8133,79	20	5465,487	<0,001	***
Probit						
(0)	1	-10866,5	-	-	-	-
(1)	24	-8109,34	23	5514,383	<0,001	***
(2)	23	-8109,36	22	5514,337	<0,001	***
(3)	22	-8112,4	21	5508,257	<0,001	***

Poznámka: *p<0,05; **p<0,01; ***p<0,001

Zdroj: Vlastné spracovanie

Vzhľadom na výsledky testu uvedené v tabuľke 4.7 je zrejmé, že výsledná p – hodnota nadobúdala len veľmi nízke hodnoty. Pre všetky testované modely možno zamietnuť H_0 a prijať alternatívnu hypotézu $H_A: \beta_1 \neq \ldots \neq \beta_j \neq 0$ na 0,1 % hladine

významnosti. Teda všetky testované modely sú ako celok štatisticky významné, t.j. aspoň jeden parameter testovaných modelov je rôzny od nuly.

4.3.4 Predikcie pravdepodobnosti a klasifikácia klientov

Obsah tejto podkapitoly je venovaný predikcií podmienenej pravdepodobnosti (P_i) prostredníctvom vybraného logitového a probitového modelu z predošlej podkapitoly. Hodnota predikovanej P_i bola využitá za účelom klasifikácie jednotlivých klientov do skupín na základe dosiahnutej úspešnosti v rámci marketingovej kampane. Klasifikácia klientov bola realizovaná prostredníctvom porovnania hodnoty P_i klientov s definovaným prahovým bodom, v ktorom hodnota P_i dosahuje úroveň 50 %. To znamená, že klient s predikovanou $P_i \leq 0{,}5$ bude vyhodnotený ako príslušník skupiny, u ktorej marketingová kampaň nezaznamenala úspešnosť. Naopak klient, ktorého predikovaná $P_i > 0{,}5$ bude považovaný za člena úspešnej skupiny.

Predikcie boli realizované na oboch údajových množinách – trénovacej a testovacej. Na základe porovnania klasifikačných výsledkov modelov na oboch údajových množínách možno posúdiť, či odhadnuté modely neprejavujú známky preučenia.

Formálny zápis lineárnej kombinácie (z_i) nezávislých premenných s príslušnými parametrami pre vybraný logitový a probitový model je obsiahnutý vo vzťahoch 4.1 a 4.2.

$$z_i^{Logit} = \left[-2{,}755 + \left(\begin{array}{c} (-0{,}517) * J_{1_i} + (-0{,}8) * J_{2_i} + (-0{,}755) * J_{3_i} + (-0{,}284) * J_{4_i} + \\ + 0{,}357 * J_{5_i} + (-0{,}373) * J_{6_i} + (-0{,}399) * J_{7_i} + 0{,}65 * J_{8_i} + \\ +(-0{,}301) * J_{9_i} + (-0{,}286) * J_{10_i} + \\ +(-0{,}099) * M_{1_i} + 0{,}147 * M_{2_i} + 0{,}229 * E_{1_i} + 0{,}581 * E_{2_i} + \\ +(-0{,}442) * Def_{1_i} + 0{,}00002 * B_i + \\ +(-1{,}116) * H_{1_i} + (-0{,}617) * L_{1_i} + 0{,}004 * Dur_i + \\ +(-0{,}126) * C_i + 0{,}003 * Pd_i + 0{,}091 * Prev_i \end{array} \right) \right] \quad (4.1)$$

$$z_i^{Probit} = \left[-1{,}58 + \left(\begin{array}{c} (-0{,}276) * J_{1_i} + (-0{,}390) * J_{2_i} + (-0{,}366) * J_{3_i} + (-0{,}148) * J_{4_i} + \\ + 0{,}204 * J_{5_i} + (-0{,}2) * J_{6_i} + (-0{,}223) * J_{7_i} + 0{,}376 * J_{8_i} + \\ +(-0{,}158) * J_{9_i} + (-0{,}146) * J_{10_i} + \\ +(-0{,}053) * M_{1_i} + 0{,}08 * M_{2_i} + 0{,}127 * E_{1_i} + 0{,}304 * E_{2_i} + \\ +(-0{,}242) * Def_{1_i} + 0{,}00001 * B_i + \\ +(-0{,}568) * H_{1_i} + (-0{,}317) * L_{1_i} + 0{,}002 * Dur_i + \\ +(-0{,}061) * C_i + 0{,}002 * Pd_i + 0{,}052 * Prev_i \end{array} \right) \right] \quad (4.2)$$

Lineárna kombinácia z_i^{Logit} definovaná vo vzťahu 4.1 bude substituovaná do vzťahu 4.3 a z_i^{Probit} definovaná vo vzťahu 4.2 bude substituované do vzťahu 4.4. Následne možno uskutočniť predikcie P_i pre všetkých klientov (obe údajové množiny).

$$P_i = E(y = 1|X) = \frac{1}{1 + e^{-z_i^{Logit}}} \quad (4.3) \qquad P_i = E(y = 1|X) = \Phi(z_i^{Probit}) \quad (4.4)$$

Za účelom prvotného zmapovania výsledkov klasifikácie klientov na základe ich predikovanej P_i bola použitá matica zámen, spracovaná na základe predlohy v podkapitole 3.3.3. Záhlavie riadkov matice odkazuje na predikovanú príslušnosť klientov, buď do triedy úspešnej alebo neúspešnej. Záhlavie stĺpcov symbolizuje reálnu úspešnosť klientov v marketingovej kampani. Obsah tabuľky 4.8 reprezentuje maticu zámen logitového modelu, obsah tabuľky 4.9 reprezentuje maticu zámen pre probitový model – zahŕňajú výsledky klasifikácie na oboch údajových množinách (trénovacej a testovacej).

Tabuľka 4.8 Matica zámen - Logit model

Predikovaná premenná	Cieľová premenná		
	Trénovacia množina		
	Neúspech (Y = 0)	**Úspech (Y = 1)**	**Spolu:**
Neúspech (Y = 0)	**TN = 26 209**	**FN = 2 747**	**28 956**
Úspech (1)	**FP = 512**	**TP = 768**	**1 280**
Spolu:	**26 721**	**3 515**	**30 236**
	Testovacia množina		
	Neúspech (Y = 0)	**Úspech (Y = 1)**	**Spolu:**
Neúspech (Y = 0)	**TN = 11 237**	**FN = 1 186**	**12 423**
Úspech (Y = 1)	**FP = 214**	**TP = 320**	**534**
Spolu:	**11 451**	**1 506**	**12 957**

Zdroj: Vlastné spracovanie

Tabuľka 4.9 Matica zámen - Probit model

Predikovaná premenná	Cieľová premenná		
	Trénovacia množina		
	Neúspech (Y = 0)	**Úspech (Y = 1)**	**Spolu:**
Neúspech (Y = 0)	**TN = 26 310**	**FN = 2 871**	**29 181**
Úspech (Y = 1)	**FP = 411**	**TP = 644**	**1 055**
Spolu:	**26 721**	**3 515**	**30 236**
	Testovacia množina		
	Neúspech (Y = 0)	**Úspech (Y = 1)**	**Spolu:**
Neúspech (Y = 0)	**TN = 11 273**	**FN = 1 235**	**12 508**
Úspech (Y = 1)	**FP = 178**	**TP = 271**	**449**
Spolu:	**11 451**	**1 506**	**12 957**

Zdroj: Vlastné spracovanie

Výsledky klasifikácie zachytené v tabuľkách 4.8 a 4.9 naznačujú, že oba modely priniesli veľmi podobné predikčné výsledky na oboch údajových množinách.

Počet celkovo správne predikovaných pozorovaní na trénujúcej množine logitového modelu dosahuje hodnotu 26 997, oproti 26 954 správne vyhodnoteným pozorovaniam probitového modelu. Z celkového počtu 30 236 pozorovaní tejto množiny, logit model

vyhodnotil celkovo 3 259 klientov do nesprávnej kategórie úspešnosti, v porovnaní s 3 282 nesprávne vyhodnotenými klientami prostredníctvom modelu probit.

Z pohľadu klasifikačnej úspešnosti modelov na testovacej množine údajov, možno opäť pozorovať značnú podobnosť v dosiahnutých výsledkoch oboch modelov. Na tejto množine model logit celkovo vyhodnotil úspešne 11 557 pozorovaní, oproti 11 544 úspešne vyhodnotených klientov probitovým modelom. Celkový počet neúspešne klasifikovaných klientov prostredníctvom logit modelu predstavuje 1 400 pozorovaní, v porovnaní s celkovým počtom neúspešne klasifikovaných 1 413 pozorovaní probitovým modelom, z celkového počtu 12 957 klientov obsiahnutých v tejto údajovej množine.

Na základe vyššie uvedených skutočností možno konštatovať, že napriek relatívne nízkym rozdielom z pohľadu úspešnosti klasifikácie klientov do jednotlivých skupín, model logit dosiahol nepatrne lepšie výsledky na oboch údajových množinách.

V rámci ďalšej podkapitoly budú prvky matíc zámen ilustrovaných v tabuľkách 4.8 a 4.9 použité, za účelom výpočtu rôznych výkonnostných metrík, určených na validáciu predikčnej schopnosti modelov.

4.3.5 Validácia predikčnej schopnosti modelov

Prvky matíc zámen obsiahnutých v tabuľkách 4.8 – 4.9 demonštrujú absolútne početnosti klientov, ktorí mohli byť na základe výsledkov predikcií modelov vyhodnotení v nasledujúcich štyroch kombináciách:

- klienti neúspešnej skupiny, správne vyhodnotení ako členovia tejto skupiny – **TN**;
- klienti patriaci do neúspešnej skupiny, nesprávne klasifikovaní ako členovia úspešnej skupiny – **FP**;
- klienti úspešnej skupiny, ktorí boli správne klasifikovaní do tejto skupiny – **TP**;
- klienti zo skupiny úspešnej, chybne vyhodnotení ako členovia neúspešnej skupiny – **FN**.

Absolútne početnosti týchto štyroch kombinácií možných výsledkov klasifikácie boli dosadené do vzťahov 3.40 – 3.45 za účelom výpočtu výkonnostných metrík, určených na validáciu predikčnej schopnosti oboch modelov. Ide o metriky mapujúce relatívnu klasifikačnú správnosť, presnosť, senzitívnosť, špecifickosť, mieru výpadku a F1 skóre.

Výsledné hodnoty vypočítaných výkonnostných metrík oboch modelov sú zachytené v tabuľke 4.10. Výkonnostné metriky boli v oboch prípadoch vypočítané na základe

klasifikačných výsledkov oboch údajových množín. Vďaka výsledkom predikčnej schopnosti modelov na oboch údajových množinách bolo možné ich následné porovnanie v zmysle, do akej miery dokázali modely správne predikovať údaje, na základe ktorých boli odhadnuté oproti údajom, ktoré neboli použité pri ich odhade, t.j. na novej údajovej množine.

Tabuľka 4.10 Výsledné hodnoty výkonnostných charakteristík

Model:	Množina:	Správnosť	Presnosť	Senzitívnosť	Špecifickosť	Výpadok	F1 Skóre
Logit	Trénovacia	0,8922	0,6000	0,2185	0,9808	0,0192	0,3203
	Testovacia	0,8920	0,5993	0,2125	0,9813	0,0187	0,3138
Probit	Trénovacia	0,8915	0,6104	0,1832	0,9846	0,0154	0,2818
	Testovacia	0,8909	0,6036	0,1799	0,9845	0,0155	0,2772

Zdroj: Vlastné spracovanie

Z dosiahnutých výsledkov zachytených v tabuľke 4.10 je na prvý pohľad zrejmé, že predikčná schopnosť oboch modelov sa v zásade výrazne nelíšila a dosiahnuté výsledky klasifikácie klientov s využitím oboch modelov priniesli porovnateľné výsledky. Táto skutočnosť sa týkala aj dosiahnutých výsledkov predikcií na oboch údajových množinách, na základe čoho možno skonštatovať, že odhadnuté modely neprejavovali známky preučenia.

Na základe uvedených výsledkov je možné ďalej konštatovať, že oba modely správne klasifikovali cca. 89 % klientov na oboch údajových množinách. Vzhľadom na vysokú nevyváženosť počtu alternatív závislej premennej v údajových množinách (viď obrázok 4.2), môže byť táto metrika značne zavádzajúca. Presnosť klasifikácie klientov úspešnej skupiny bola napr. o poznanie nižšia v prípade oboch modelov. Zo všetkých klientov, ktorí boli vyhodnotení ako členovia úspešnej skupiny, bolo presne vyhodnotených iba okolo 60 % klientov na oboch údajových množinách.

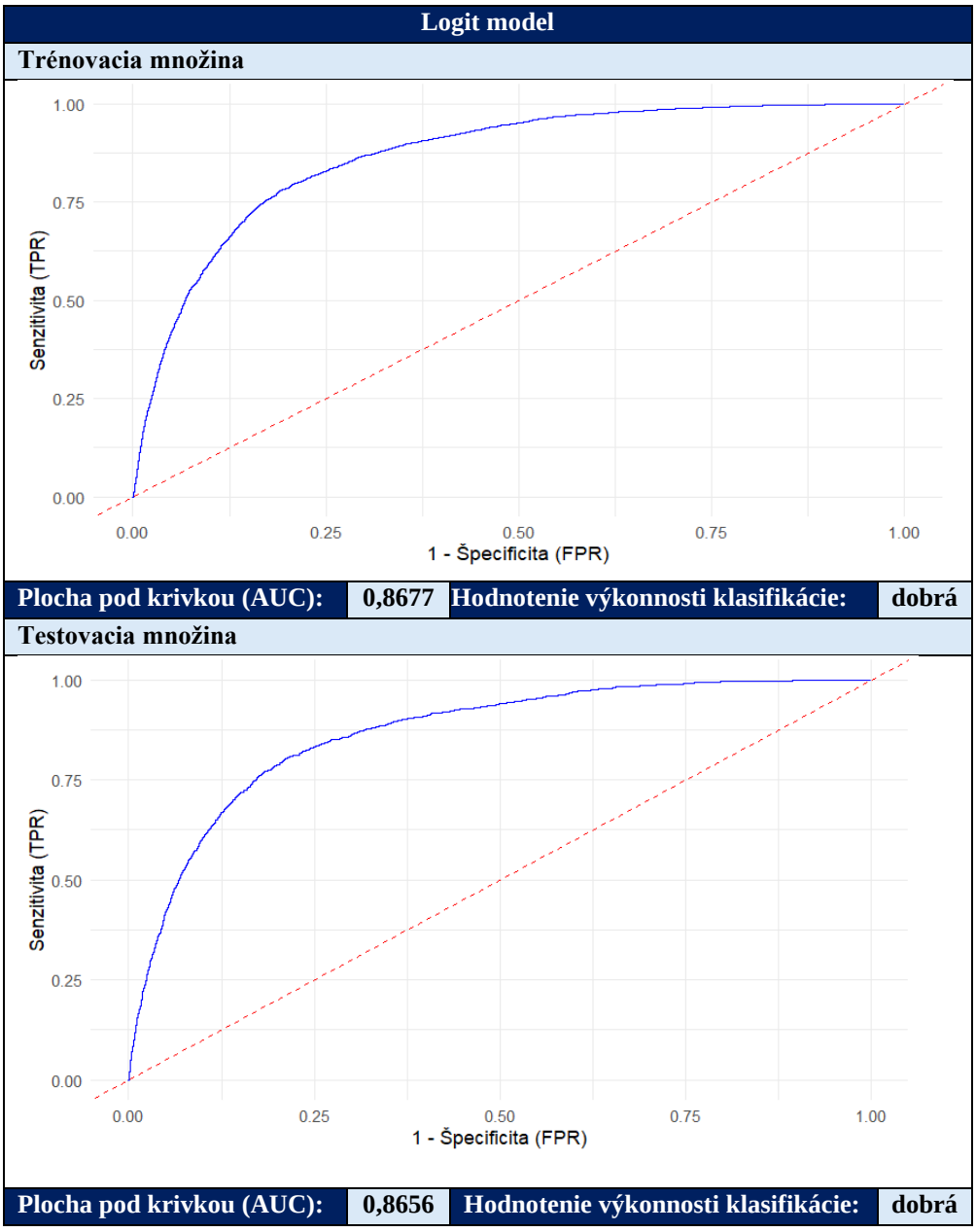
Negatívny vplyv na presnosť klasifikácie klientov úspešnej skupiny mohol byť zapríčinený práve tým, že obe údajové množiny sú značne v nevyváženom pomere z pohľadu klientov patriacich do úspešnej skupiny, oproti klientom neúspešnej skupiny. Toto tvrdenie sa odráža aj vo výsledných hodnotách metrík senzitívnosť a špecifickosť.

Miera senzitívnosti v kontexte riešeného problému kvantifikuje pomer správne klasifikovaných klientov úspešnej marketingovej skupiny, oproti celej populácií úspešnej skupiny klientov obsiahnutých v údajovej množine. Vzhľadom na to, že výsledná miera senzitivity klasifikácie oboch modelov bola menšia ako štvrtina zo všetkých úspešných

klientov je zrejmé, že modely do značnej miery zamieňajú klientov minoritnej (úspešnej) triedy a disponujú tendenciou nesprávne ich vyhodnocovať ako klientov triedy majoritnej (neúspešnej). V tomto prípade logitový model dosahoval mierne uspokojivejšie výsledky. To s ohľadom na skutočnosť, že miera senzitívnosti dosahovala na trénovacej množine o 3,53 % a na testovacej množine o 3,26 % lepšie výsledky ako model probit. Miera špecifickosti, ktorá relatívne kvantifikuje pomer správne vyhodnotených členov neúspešnej marketingovej skupiny oproti všetkým členom populácie neúspešnej skupiny v údajovej množine naopak naznačuje, že oba modely dokázali do veľkej miery správne vyhodnotiť príslušníkov neúspešnej skupiny. Úroveň špecifickosti klasifikácie pre oba modely (na oboch údajových množinách) dosahovala približne 98 %. Táto skutočnosť odôvodňuje aj pomerne nízku hodnotu miery výpadku, ktorá kvantifikuje relatívnu početnosť klientov neúspešnej triedy chybne zamenených za príslušníkov triedy úspešnej. V tomto kontexte možno dedukovať, že model probit dosahoval vzhľadom na jej nižšie hodnoty zanedbateľne lepší výsledok (1,54 % - trénovacia množina, 1,55 % - testovacia množina) oproti modelu logit (1,92 % - trénovacia množina, 1,87 % - testovacia množina).

Identifikácia vhodnejšieho modelu prostredníctvom využitia harmonického priemeru mier presnosti a senzitívnosti klasifikácie (F1 skóre) naznačoval, že model logit by mohol byť vhodnejšia alternatíva pri klasifikácií oboch typov klientov. Avšak tento rozdiel nie je významne odlišný, nakoľko model logit disponuje vyšším F1 skóre na trénovacej množine iba o 3,85 % a na testovacej množine iba o 3,66 %. Je potrebné zdôrazniť, že vzhľadom na pomerne nízke skóre tejto metriky u oboch modelov (cca. 30 %) je zrejmé, že model mal tendenciu do značnej miery zamieňať prípady minoritnej triedy za prípady majoritnej.

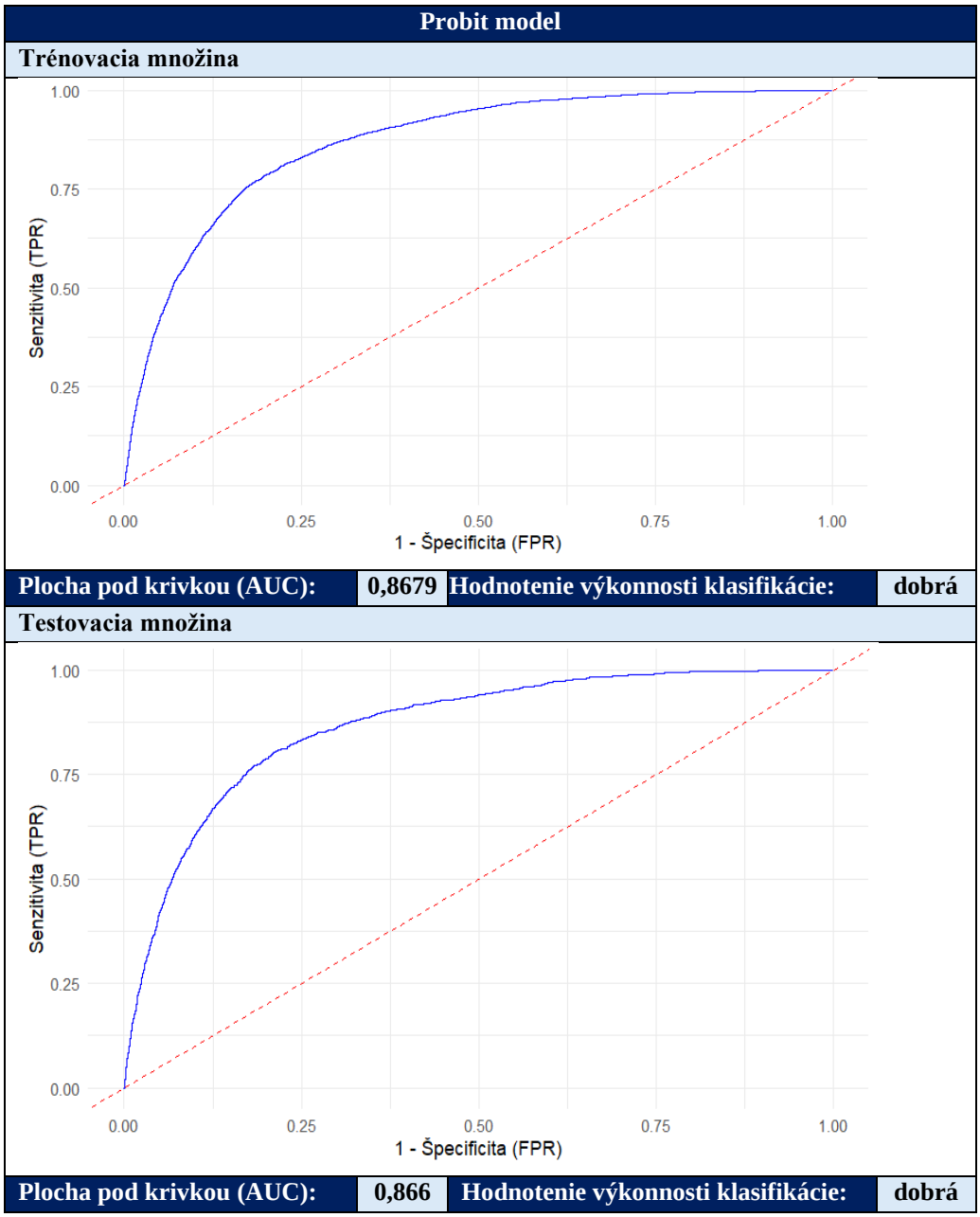
Vzhľadom na limitáciu interpretovateľnosti výsledkov, spôsobenú nevyváženosťou oboch tried závislej premennej, bola za účelom identifikácie vhodnejšieho modelu skonštruovaná ROC krivka. Tá zachytáva vzťah medzi výslednou mierou senzitívnosti a miery výpadku (resp. 1 – špecifickosť), na základe meniacich sa hodnôt prahového bodu pravdepodobnosti. Na sumarizáciu výkonnosti bude použitá plocha pod krivkou (AUC), ktorej výsledná hodnota bola interpretovaná na základe tabuľky 3.2. ROC krivky pre modely logit a probit sú znázornené v obrázkoch 4.3 – 4.4 Referenčná priamka y = x je na grafoch zvýraznená červenou prerušovanou čiarou.



Obrázok 4.3 ROC-AUC analýza - Logit model (Zdroj: Vlastné spracovanie)

ROC krivky logitového modelu (obrázok 4.3, obe údajové množiny) naznačujú, že predikčná schopnosť tohto modelu rozlišovať triedu úspešných a neúspešných klientov, je významne odlišná od náhodnej klasifikácie (ROC krivka nekopíruje referenčnú priamku). Model má však aj svoje mierne predikčné nedostatky, vzhľadom na vychýlenie ROC kriviek od ľavého horného rohu (dokonalá klasifikácia).

Výsledná hodnota AUC (obe údajové množiny) je z intervalu $\langle 0{,}8 \leq \text{AUC} < 0{,}9)$. To naznačuje pomerne dobrú výkonnosť klasifikácie modelu na oboch údajových množinách (možno vylúčiť preučenie modelu).



Obrázok 4.4 ROC-AUC analýza - Probit model (Zdroj: Vlastné spracovanie)

Výsledné ROC krivky modelu probit (obrázok 4.4, obe údajové množiny) sú takmer identické s výslednými ROC krivkami modelu logit (obrázok 4.3). To naznačuje aj sumarizácia ich plochy pod krivkou prostredníctvom AUC indexu, ktorý priniesol takmer totožné výsledky (obe údajové množiny) v porovnaní s výsledkami logit modelu. Zdá sa, že oba modely prinášajú porovnateľné, až takmer identické klasifikačné výsledky. Na základe toho je možné konštatovať, že oba modely sú rovnako vhodné na riešenie tohto problému.

4.3.6 Interpretácia dosiahnutých výsledkov

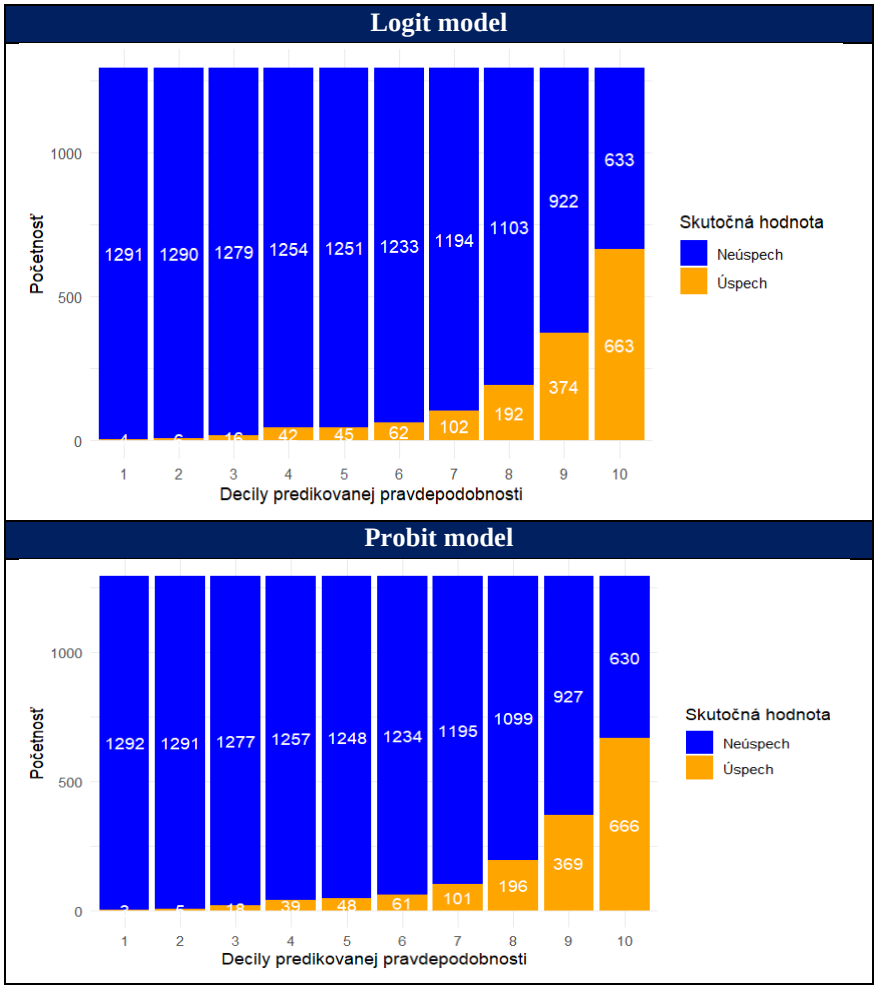
Vo všeobecnosti kampaň zaznamenala u oslovených klientov iba veľmi nízku úspešnosť, pričom drvivá väčšina klientov na ponuku banky nepristúpila. V rámci základného súboru údajov predstavovala relatívna úspešnosť hodnotu 11,6 %, analogicky zvyšných 88,4 % tvorila neúspešná marketingová skupina klientov.

V prípade opakovania marketingovej kampane z tohto pohľadu nie je príliš racionálne oslovovať všetkých klientov zapojených v kampani. Vzhľadom na drvivú väčšinu klientov neúspešnej skupiny, ich opätovné oslovenie by sa mohlo javiť ako neefektívne, vzhľadom k vynaloženým nákladom na prevádzkovanie kampane. Z tohoto dôvodu je z pohľadu banky racionálnejšie cieliť kampaň na skupinu klientov, u ktorých sa prostredníctvom historických údajov preukázala najvyššia pravdepodobnosť úspechu. Tým by sa do značnej miery mohlo podariť bankovej inštitúcií zredukovať náklady spojené s prevádzkovaním kampane a zároveň zvýšiť jej relatívnu úspešnosť.

Vybraný logitový a probitový model, prostredníctvom ktorých bol modelovaný vzťah podmienenej pravdepodobnosti úspechu kampane u klienta, dosahoval v oboch prípadoch veľmi podobné výsledky predikcií, rovnako na historických údajoch (trénovacej množine), tak i nových údajoch (testovacej množine), ktoré neparticipovali pri odhadovaní modelov.

Analýza decilového rozloženia predikovanej pravdepodobnosti

Za účelom odhalenia pravdepodobnostného rozloženia úspešnej skupiny klientov, bola predikovaná pravdepodobnosť (oboma modelmi) na novej vzorke údajov zoradené vzostupne, na základe ich pravdepodobnostných hodnôt. Zoradené pravdepodobnostné hodnoty jednotlivých klientov boli rozdelené do decilov, v ktorých sa následne premietla absolútna početnosť skutočnej príslušnosti klientov do úspešnej a neúspešnej skupiny. Vizualizácia výsledného decilového rozloženia pravdepodobnosti možno pozorovať na obr. 4.5. V prípade oboch grafov os y reflektuje absolútnu početnosť populácie klientov v rámci príslušného decilu, os x odkazuje na jednotlivé decily usporiadaných pravdepodobnostných hodnôt.



Obrázok 4.5 Rozdelenie početnosti úspešnej a neúspešnej skupiny v rámci pravdepodobnostných decilov (Zdroj: Vlastné spracovanie)

Z obrázku 4.5 už na prvý pohľad vyplýva, že rozloženie zoradenej predikovanej pravdepodobnosti v rámci jednotlivých decilov bolo takmer totožné v prípade oboch pravdepodobnostných modelov. Na základe skutočnej príslušnosti klientov do skupiny úspešnej alebo neúspešnej je zrejmé, že najväčšia koncentrácia klientov úspešnej skupiny bola zaznamenaná práve v posledných deciloch pravdepodobnostné rozloženia, kde pravdepodobnosť úspechu dosahuje najvyššie hodnoty. Z toho dôvodu v prvých pravdepodobnostných deciloch, kde predikovaná pravdepodobnosť dosahuje najnižšie hodnoty prevládala majoritná skupina klientov, u ktorej nebol zaznamenaný úspech v rámci kampane.

V prípade opakovania kampane sa z marketingového pohľadu môže zdať efektívne cieliť na klientelu, obsiahnutú v posledných pravdepodobnostných deciloch, v ktorých dosahuje pravdepodobnosť úspechu kampane najvyššie hodnoty. Správnym výberom

cieľovej skupiny klientov sa môžu radikálne znížiť náklady, spojené s prevádzkou kampane. Takéto zníženie nákladov sa môže v konečnom dôsledku odzrkadliť aj na výške čistého zisku banky.

Vzhľadom na uvedené je zrejmé, že cieľová skupina klientov bola obsiahnutá práve v posledných štyroch pravdepodobnostných deciloch. Na základe relatívneho vyčíslenia koncentrácie úspešných klientov vyplýva, že práve v týchto deciloch sa nachádzalo približne 88,4 % populácie úspešných klientov. Koncentrácia neúspešných klientov v tomto prípade tvorila približne 33,6 % z celkovej populácie neúspešných klientov. Z pohľadu druhého, v populácií prvých 60 % zoradených klientov bolo z celkovej populácie úspešných klientov detegovaných iba 11,6 % oproti 66,4 % klientov populácie neúspešnej skupiny. Z tohto pohľadu je zrejmé, že cieliť na skupinu klientov, obsiahnutú v prvých šiestich pravdepodobnostných decilov je do veľkej miery neefektívne. Preto by sa mala kampaň bližšie zamerať práve na posledných 40 % populácie (na základe ich zoradenej predikovanej pravdepodobnosti). Čo môže spoločnosti ušetriť náklady, spojené s kontaktovaním a vedením kampane u 60 % klientov, u ktorých bola zaznamenaná len minimálna pravdepodobnosť úspechu.

Dosiahnuté výsledky boli zaokrúhlené pre oba pravdepodobnostné modely, s ohľadom na ich výraznú podobnosť (líšili sa v desatinách percentuálneho bodu v prospech modelu logit).

Decilová analýza pravdepodobnostných hodnôt dokáže v prípade opakovania kampane (na rovnakej skupine klientov) poskytnúť efektívny nástroj na segmentáciu cieľovej skupiny.

Marginálne efekty

Odhalenie najvýznamnejších faktorov ovplyvňujúcich modelovanú pravdepodobnosť dokáže objasniť, na aký typ klientov sa v budúcnosti oplatí kampaň zacieliť. Za týmto účelom boli využité hodnoty parametrov pravdepodobnostných modelov. Tie však nie je možné priamo interpretovať v zmysle ich efektu na zmenu pravdepodobnosti. Z toho dôvodu boli vypočítané hodnoty ich marginálnych efektov, ktoré predstavujú parciálne derivácie príslušnej KDF podľa jednotlivých regresorov modelu – v prípade kvantitatívnych spojitých premenných. Vzťah pre výpočet marginálnych efektov pre logitový model je obsiahnutý vo vzťahu 3.35, pre probitový model platí vzťah 3.33. Marginálny efekt pri kategoriálnych

premenných odzrkadľuje zmenu v predikovanej pravdepodobnosti výsledku pri zmenení kategórie z referenčnej na porovnávanú, po vzore vzťahu 3.36.

Pre zjednodušenie interpretácie marginálnych efektov, boli pri ich výpočte za hodnoty kvantitatívnych spojitých regresorov dosadené ich priemerné hodnoty (z trénujúcej množiny údajov). Pre lepšiu orientáciu sú priemerné hodnoty zaznamenané v tabuľke 4.11.

Tabuľka 4.11 Priemerné hodnoty kvantitatívnych regresorov - trénovacia množina

Regresor:	**B**	**Dur**	**C**	**Pd**	**Prev**
Priemer	1 362,66	258,2	2,77	40,47	0,5796

Zdroj: Vlastné spracovanie

Výsledné hodnoty marginálnych efektov (dF/dx) sú obsiahnuté v tabuľke 4.12, spolu s hodnotami ich smerodajných odchýlok (Std. Err.), Waldovou testovacou z-štatistikou (z), výslednou p-hodnotou k príslušnej testovacej z-štatistike (P>|z|) a úroveň štatistickej významnosti (ŠV).

Tabuľka 4.12 Výsledné hodnoty marginálnych efektov - Logit a Probit model

Marginálne efekty										
***Y* =**	**Logit model**					**Probit model**				
Úspešnosť	**dF/dx**	**Std. Err.**	**z**	**P>|z|**	**ŠV**	**dF/dx**	**Std. Err.**	**z**	**P>|z|**	**ŠV**
J_1	-0,0284	0,004	-7,0404	<0,001	***	-0.0324	0.0045	-7.2214	<0,001	***
J_2	-0,0363	0,0049	-7,447	<0,001	***	-0.039	0.0056	-6.957	<0,001	***
J_3	-0,0347	0,0053	-6,5922	<0,001	***	-0.0371	0.0061	-6.1296	<0,001	***
J_4	-0,0165	0,0045	-3,67	0,0002	***	-0.0183	0.0051	-3.5632	0.0004	***
J_5	0,0255	0,0079	3,2142	0,0013	**	0.0308	0.0091	3.391	0.0007	***
J_6	-0,02	0,0057	-3,4778	0,0005	***	-0.0229	0.0066	-3.49	0.0005	***
J_7	-0,0216	0,0045	-4,8417	<0,001	***	-0.0256	0.0049	-5.1993	<0,001	***
J_8	0,0531	0,0128	4,144	<0,001	***	0.0642	0.0148	4.3432	<0,001	***
J_9	-0,0172	0,0041	-4,2239	<0,001	***	-0.0192	0.0046	-4.168	<0,001	***
J_{10}	-0,0158	0,0063	-2,5209	0,0117	*	-0.0173	0.0073	-2.3812	0.0173	*
M_1	-0,0062	0,0042	-1,4804	0,1388		-0.007	0.0047	-1.4897	0.1363	
M_2	0,0094	0,0047	1,9948	0,0461	*	0.0108	0.0052	2.0664	0.0388	*
E_1	0,0141	0,0045	3,1737	0,0015	**	0.0166	0.0049	3.3887	0.0007	***
E_2	0,04	0,0064	6,2533	<0,001	***	0.0437	0.0069	6.3257	<0,001	***
Def_1	-0,0229	0,0085	-2,6855	0,0072	**	-0.0267	0.0092	-2.9144	0.0036	**
B	0.000001	0.0000003	3,0456	0,0023	**	0.0000014	0.0000004	3.4769	0.0005	***
H_1	-0,0757	0,0033	-22,8306	<0,001	***	-0.0799	0.0035	-22.9354	<0,001	***
L_1	-0,0322	0,003	-10,7686	<0,001	***	-0.0357	0.0033	-10.8269	<0,001	***
Dur	0,0002	0	43,946	<0,001	***	0.0003	0	45.0222	<0,001	***
C	-0,0078	0,0007	-11,0417	<0,001	***	-0.008	0.0008	-10.5623	<0,001	***
Pd	0,0002	0	14,2042	<0,001	***	0.0002	0	13.6185	<0,001	***
Prev	0,0057	0,0006	9,5404	<0,001	***	0.0069	0.0007	9.6876	<0,001	***

Poznámka: *p<0,05; **p<0,01; ***p<0,001

Zdroj: Vlastné spracovanie

Hodnoty marginálnych efektov zaznamenaných v tabuľke 4.12 reflektujú zmenu predikovanej pravdepodobnosti o túto hodnotu, v dôsledku zmeny konkrétneho regresora o jednu jednotku, za predpokladu ceteris paribus, t.j. za zachovania konštantných hodnôt ostatných regresorov modelu. Znamienko uvedené pred hodnotou marginálneho efektu udáva smer tejto zmeny vo výslednej pravdepodobnosti. V prípade, v ktorom je hodnota marginálneho efektu záporná, možno hovoriť o poklese predikovanej pravdepodobnosti o túto hodnotu, v dôsledku zvýšenia daného regresora o jednu jednotku (ceteris paribus). V opačnom prípade, kladná hodnota marginálneho efektu signalizuje nárast v predikovanej pravdepodobnosti o túto hodnotu, v prípade zvýšenia daného regresora o jednu jednotku (ceteris paribus). Hodnoty marginálnych efektov veľmi blízke nule naznačujú minimálny alebo žiadny vplyv na predikovanú pravdepodobnosť, čo znamená, že zmena v nezávislej premennej nemá významný efekt na jej výslednú hodnotu.

Zohľadňujúc vyššie uvedené skutočnosti, možno na základe výsledných hodnôt marginálnych efektov zahrnutých v tabuľke 4.12 identifikovať a následne interpretovať premenné (ich zmenu o jednotku, resp. v zmene kategórie oproti referenčnej, ceteris paribus), ktorých efekt mal najvýznamnejší vplyv na výslednú hodnotu modelovanej podmienenej pravdepodobnosti. Dôležité je spomenúť, že rozdiel medzi vplyvom marginálnych efektov pre logitový a probitový model na výslednú hodnotu pravdepodobnosti sa ukázal ako minimálny, resp. zanedbateľný.

Sociodemografické charakteristiky

Z pohľadu klientovho zamestnania – $\mathbf{J_j}$ sa ukázalo, že vo väčšine prípadov pravdepodobnosť úspechu klesá v porovnaní s klientami, pracujúcimi ako administratívny pracovníci ($\mathbf{J_0}$). Negatívny vplyv na pokles pravdepodobnosti sa najviac odzrkadlil u klientov podnikateľov – $\mathbf{J_2}$ [- 3,63 % (logit model); - 3,9 % (probit model)] a u klientov v domácnosti – $\mathbf{J_3}$ [- 3,47 % (logit model); - 3,7% (probit model)]. Pravdepodobnosť úspechu kampane u klientov naopak vzrastá v prípade, že ide o klienta dôchodcu - $\mathbf{J_5}$ [+ 2,55 % (logit model); + 3,08 % (probit model)] a klienta študenta – $\mathbf{J_8}$ [+ 5,31 % (logit model); + 6,42 % (probit model)]. S každým dosiahnutým stupňom vzdelania klienta, rastie aj pravdepodobnosť úspechu kampane u daného klienta v porovnaní s tými, ktorí úspešne ukončili iba základné vzdelanie ($\mathbf{E_0}$). Najvýznamnejší pozitívny vplyv na výslednú hodnotu pravdepodobnosti bol zaznamenaný u klientov s dosiahnutým vysokoškolským vzdelaním - $\mathbf{E_2}$ [+ 4 % (logit model); + 4,37 % (probit model)], oproti skupine klientov $\mathbf{E_0}$. Rozdielne statusy v rodinnom stave – $\mathbf{M_j}$ klienta nepriniesli signifikantné rozdiely vo

výslednej hodnote pravdepodobnosti úspechu. Zdá sa však, že slobodní klienti – **M_2** disponujú zanedbateľne vyššou mierou predikovanej pravdepodobnosť úspechu, ako referenčná skupina klientov rozvedených – **M_0**.

Charakteristiky klientovej solventnosti

Najväčší vplyv na pokles vo výslednej hodnote predikovanej pravdepodobnosti úspechu kampane bol zaznamenaný u klientov, ktorí v súčasnosti splácajú pôžičku na bývanie - **H_1** [- 7,57 % (logit model); - 7,99 % (probit model)]. Významný pokles hodnoty pravdepodobnosti úspechu je však spojený aj s klientelou, ktorá v súčasnosti spláca osobnú pôžičku - **L_1** alebo sa v minulosti dostala do situácie, v ktorej nedokázala splácať svoje záväzky, t.j. skupina v minulosti „defaultných“ klientov - **Def_1**. Pozitívny vplyv na pravdepodobnosť úspechu kampane u klienta sa odzrkadľuje aj vo výške zostatku finančných prostriedkov na jeho bankovom účte – **B**. Vzhľadom na skutočnosť, že hodnota marginálneho efektu tohto atribútu dosahuje len veľmi nízke hodnoty [+ 0,0001 % (logit model); + 0,00014 % (probit model)], odzrkadľuje to pozitívnu zmenu vo výslednej pravdepodobnosti úspechu, pri zmene priemernej sumy na bankovom účte o jednu peňažnú jednotku. Čo je samozrejme logické, avšak treba zdôrazniť, že premenná vystupuje ako štatisticky významná. Z toho dôvodu možno očakávať preukázateľné rozdiely vo výslednej pravdepodobnosti úspechu v prípadoch, kedy sa rozdiel v sume na bankovom účte zmení napr. v desiatkach tisícov peňažných jednotiek oproti ich priemernej hodnote.

Charakteristiky priebehu kampane

V kontexte kvantitatívnych premenných súvisiacich s marketingovou kampaňou, kde marginálne efekty boli vypočítané na základe ich priemerných hodnôt (viď tabuľka 4.11), a ich hodnoty sú výrazne blízke nule, možno konštatovať nasledovné: Premenná **C**, ktorá zachytáva počet kontaktov uskutočnených s klientom počas kampane, má mierne negatívny efekt na pravdepodobnosť úspechu v kampani. Tento negatívny vplyv naznačuje, že s každým ďalším kontaktom sa pravdepodobnosť, že klient bude na kampaň reagovať pozitívne, mierne zníži. Ostatné premenné disponujú miernym pozitívnym efektom na modelovanú pravdepodobnosť. Najvýraznejším pozitívnym efektom z nich disponuje premenná **Prev,** ktorá reflektuje počet kontaktov s klientom, uskutočnených pred touto kampaňou.

Pomer šancí – v prípade modelu logit (na rozdiel od probitového modelu) existuje ďalší spôsob, prostredníctvom ktorého je možné interpretovať odhadnuté parametre modelu. Tie

v modeli logit udávajú zmenu v prirodzenom logaritme šance (pri jednotkovej zmene príslušného regresora, ceteris paribus). Šancu (viď vzťah 3.7) možno v tomto kontexte rozumieť ako pomer pravdepodobnosti, že klient bude vyhodnotený do úspešnej skupiny, oproti pravdepodobnosti jeho príslušnosti do neúspešnej skupiny. Teda koľkokrát je pravdepodobnejšie, že kampaň bude u klienta úspešná, oproti možnej pravdepodobnosti neúspechu. Odlogaritmovaním príslušných parametrov logitového modelu je možná ich následná interpretácia v zmysle, koľkokrát sa zmení šanca v prípade jednotkovej zmeny daného regresora, ceteris paribus. V prípade kategoriálnych premenných samozrejme ide o zmenu šance v danej kategórií oproti referenčnej kategórii, ceteris paribus. Týmto spôsobom vypočítané pomery šancí (OR) sú zachytené v tabuľke 4.13. Tabuľka tiež zachytáva hodnoty ich smerodajných odchýlok (Std. Err.), hodnoty testovacej Waldovej z-štatistiky (z), výslednú p-hodnotu k danej z-štatistike (P>|z|) a úroveň štatistickej významnosti (ŠV).

Tabuľka 4.13 Výsledne hodnoty pomeru šancí - Logit model

Pomer šancí					
Y = Úspešnosť'	**OR**	**Std. Err.**	**z**	**P>|z|**	**ŠV**
J_1	0,5965	0,0495	-6,226	<0,001	***
J_2	0,4493	0,0686	-5,2415	<0,001	***
J_3	0,47	0,0753	-4,7146	<0,001	***
J_4	0,7526	0,0624	-3,4249	0,0006	***
J_5	1,4296	0,139	3,6764	0,0002	***
J_6	0,6885	0,0865	-2,9718	0,003	**
J_7	0,6711	0,0639	-4,1908	<0,001	***
J_8	1,9157	0,2354	5,2905	<0,001	***
J_9	0,7401	0,0575	-3,8763	0,0001	***
J_{10}	0,7515	0,0962	-2,2326	0,0256	*
M_1	0,9054	0,0603	-1,4933	0,1353	
M_2	1,1582	0,0829	2,0515	0,0402	*
E_1	1,2575	0,0912	3,1584	0,0016	**
E_2	1,7876	0,15	6,9238	<0,001	***
Def_1	0,6429	0,1287	-2,2068	0,0273	*
B	1	0	3,049	0,0023	**
H_1	0,3275	0,015	-24,4113	<0,001	***
L_1	0,5398	0,0371	-8,9794	<0,001	***
Dur	1,004	0,0001	53,2629	<0,001	***
C	0,8817	0,0105	-10,6109	<0,001	***
Pd	1,0029	0,0002	14,2435	<0,001	***
Prev	1,0956	0,0104	9,6071	<0,001	***

Poznámka: *p<0,05; **p<0,01; ***p<0,001

Zdroj: Vlastné spracovanie

Pred samotnou interpretáciou výsledných hodnôt OR uvedených v tabuľke 4.13 je potrebné zdôrazniť, že hodnoty OR = 1 (resp. blízke 1) nepredstavujú významný faktor vzhľadom na zmenu výslednej šance. Naopak hodnoty OR > 1 predstavujú významný faktor, v zmysle nárastu šance pri danej premennej (pri jednotkovej zmene regresora, prípadne zmeny kategórie, ceteris paribus). Pri hodnotách OR < 1 sa jedná taktiež o významný faktor, avšak je spojený s poklesom šance v tejto premennej. Tieto hodnoty sa pre jednoduchosť interpretácie môžu uvádzať v ich prevrátených (inverzných) hodnotách.

Vzhľadom na komplexnosť logitového modelu budú interpretované iba najvýznamnejšie premenné (ceteris paribus), ktorých hodnoty OR sú čo možno najvýraznejšie odlišné od 1, aby sa zdôraznil ich potenciálny vplyv na závislú premennú – úspešnosť kampane.

Interpretácie pomerov šancí

Na základe klientovho zamestnania sa ukázalo, že šanca na úspech kampane u klienta narastá v prípade, ak oslovený klient predstavuje študenta (**J_8**) alebo dôchodcu (**J_5**), oproti klientom pracujúcim ako administratívny pracovník (**J_0**). V prípade klientov študentov bol nárast šance na úspech v rámci možných zamestnaní najvýraznejší (oproti **J_0**). Šanca úspechu kampane sa u klientov študentov zvýši približne 1,9157 – krát oproti klientom, pracujúcich v administratíve. U klientov dôchodcov sa táto šanca zvyšuje približne 1,4296 – krát.

Z druhého pohľadu, šanca na úspech kampane u klienta administratívneho pracovníka (**J_0**) výrazne narastá, v porovnaní s povolaniami klientov zo skupiny podnikateľov (**J_2**) alebo pracujúcich v domácnosti (**J_3**). Konkrétne, šanca úspechu kampane u administratívneho pracovníka narastá približne 2,2257 – krát, oproti klientovi podnikateľovi a cca. 2,1277 – krát oproti klientovi, ktorý pracuje v domácnosti.

Stupeň klientovho najvyššieho dosiahnutého vzdelania sa taktiež preukázal ako významný faktor, pričom s každým stupňom klientovho dosiahnutého vzdelania šanca na úspech rastie, oproti skupine klientov, ktorý ukončili iba základné vzdelanie (**E_0**). Šanca na úspech kampane sa ukázala najvýraznejšia u vysokoškolsky vzdelaných klientov (**E_2**). Pri tejto skupine klientov šanca úspechu kampane rastie približne 1,7876 – krát oproti klientom, s ukončeným iba základným vzdelaním.

Veľký vplyv na nárast šance úspechu kampane odzrkadľovala aj skupina klientov, ktorí v súčasnosti nie sú platiteľmi pôžičky na bývane (**H_0**). Šanca úspechu sa v tejto skupine

zvyšuje približne 3,0534 – násobne oproti klientom, ktorí v súčasnosti splácajú tento druh pôžičky (**H_1**).

Výpočet pomeru šancí jednotlivých klientov možno demonštrovať aj na konkrétnom príklade. Vzhľadom na komplexnosť modelu existuje enormné množstvo kombinácií charakteristík jednotlivých klientov, ktorých zmeny môžu buď pozitívne alebo negatívne ovplyvniť šancu na úspech kampane u daného klienta.

Vzhľadom k tomu, že je prakticky nemožné venovať sa každej kombinácií klientov individuálne, bol za týmto účelom vytvorený referenčný profil klienta, ktorého charakteristiky boli podrobené zmenám.

Referenčný profil skúmaného klienta obsahuje nasledujúce charakteristiky:

- Povolanie (**J_j**) = administratívny pracovník (**J_0**);
- Rodinný stav (**M_j**) = slobodný (**M_2**);
- Vzdelanie(**E_j**) = vysokoškolské (**E_2**);
- Default (**Def_j**) = nie (**Def_0**);
- Zostatok na bankovom účte (**B**) = 1360 peňažných jednotiek [p.j.];
- Pôžička na bývanie (**H_j**) = nie (**H_0**);
- Osobná pôžička (**L_j**) = nie (**L_0**);
- Dĺžka telefonického rozhovoru (**Dur**) = 600 sekúnd, resp. 10 minút;
- Počet kontaktov v rámci kampane (**C**) = 2;
- Počet uplynutých dní od posledného kontaktu klienta z minulej kampane (**Pd**) = 90 dní;
- Počet predchádzajúcich kontaktov pred aktuálnou kampaňou (**Prev**) = 1.

Prvý príklad preskúma vplyv výšky finančného zostatku na bankovom účte klienta. Je zrejmé, že jednotková zmena tejto charakteristiky nemá výrazný vplyv na zmenu pravdepodobnosti úspechu kampane, ktorá sa priamo premietne do výslednej šance na úspech u referenčného klienta.

Za týmto účelom bola zvolená referenčná hladina zostatku, veľmi blízka priemernej hodnote. Skúmané bolo rapídne zvýšenie tejto charakteristiky, na úroveň 15 000 p.j. a vplyv klientovho najvyššieho vzdelania, ceteris paribus.

Vzhľadom k tomu, že referenčný klient predstavoval administratívneho pracovníka, nemalo zmysel skúmať kategóriu klienta so základným vzdelaním,. Výsledky skúmania sú zachytené v tabuľke 4.14.

Tabuľka 4.14 Vplyv dosiahnutej najvyššej úrovne vzdelania a zostatku finančných prostriedkov na bankovom účte klienta na výslednú šancu úspechu

Premenná = B, E_j		Pravdepodobnosť úspechu	Pravdepodobnosť neúspechu	Šanca	Pomer šancí
Referenčné	**B** = 1 360 p.j., $\mathbf{E_2}$	0,6146	0,3854	1,5950	-
Porovnanie 1	**B** = 1 360 p.j., $\mathbf{E_1}$	0,5287	0,4713	1,1220	0,7034
Porovnanie 2	**B** = 15 000 p.j., $\mathbf{E_2}$	0,6670	0,3330	2,0032	1,2560
Porovnanie 3	**B** = 15 000 p.j., $\mathbf{E_1}$	0,5849	0,4151	1,4092	0,8835

Zdroj: Vlastné spracovanie

Na základe výsledkov uvedených v tabuľke je možné zhodnotiť, že šanca na úspech kampane u referenčného klienta dosahuje hodnotu 1,595. To znamená, že je približne 1,595 – krát pravdepodobnejšie, že kampaň bude u daného klienta úspešná, oproti možnej pravdepodobnosti neúspechu. V prípade, že klient dosiahol najvyššie stredoškolské vzdelanie a disponuje rovnakým finančným zostatkom na bankovom účte možno konštatovať, že u klienta s dosiahnutým vysokoškolským vzdelaním je šanca na úspech 1,422 – krát vyššia oproti tomuto klientovi, ceteris paribus. V prípade, v ktorom referenčný klient disponuje výrazne nadpriemerným finančným zostatkom, v tomto prípade 15 000 p.j., šanca na úspech kampane u daného klienta vzrastie približne 1,256 – krát oproti klientovi, ktorého finančný zostatok je blízky priemernej hodnote klientely, ceteris paribus.

Obzvlášť zaujímavé je zistenie, že klient s nadpriemerným zostatkom 15 000 p.j. so stredoškolským vzdelaním, disponuje nižšou šancou na úspech kampane, ako klient so zostatkom blízkym priemeru s dosiahnutým vysokoškolským vzdelaním. U referenčného klienta je stále šanca na úspech 1,132 – krát vyššia, ako u výrazne majetnejšieho klienta s nižším stupňom vzdelania, ceteris paribus.

Možno konštatovať, že aj napriek nepatrnému zvýšeniu pravdepodobnosti úspechu u výrazne majetnejších klientov, výška zostatku na bankovom účte sama o sebe nezaručuje významne vyššiu pravdepodobnosť úspechu kampane. Naopak, vzdelanie klienta môže mať významný vplyv na úspech kampane, ako to dokazujú porovnania medzi klientami s rovnakým finančným zostatkom, ale rozdielnym stupňom vzdelania.

Finančný zostatok klienta nepredstavuje jediný ukazovateľ jeho solventnosti, ktorý môže ovplyvniť šancu na úspech marketingovej kampane. Zistenia naznačujú, že ďalšie zaťaženia solventnosti klienta, ako sú splátky osobnej pôžičky, úvery na bývanie a minulé ocitnutie sa v stave neschopnosti splácať dlhy („default“), majú preukázateľný vplyv na zníženie pravdepodobnosti úspechu. V tomto príklade budú analyzované rôzne úrovne zaťaženia klientovej solventnosti a ich potenciálne negatívny vplyv na šancu úspechu

kampane. To v porovnaní s referenčným klientom, ktorý nedisponuje žiadnou z uvedených finančných záťaží. Výsledky analýzy sú obsiahnuté v tabuľke 4.15.

Tabuľka 4.15 Vplyv rôznych úrovní záťaže klientovej solventnosti na výslednú hodnotu šancu úspechu

Premenná = Def_j, H_j, L_j		Pravdepodobnosť úspechu	Pravdepodobnosť neúspechu	Šanca	Pomer šancí
Referenčné	**Def_0, H_0, L_0**	0,6146	0,3854	1,5947	-
Porovnanie 1	**Def_1**	0,5063	0,4937	1,0254	0,6430
Porovnanie 2	**H_1**	0,3431	0,6569	0,5223	0,3275
Porovnanie 3	**L_1**	0,4626	0,5374	0,8609	0.5399
Porovnanie 4	**H_1, L_1**	0,2199	0,7801	0,2819	0,1768
Porovnanie 5	**Def_1, H_1**	0,2514	0,7486	0,3358	0,2106
Porovnanie 6	**Def_1, L_1**	0,3563	0,6437	0,5535	0,3471
Porovnanie 7	**Def_1, H_1, L_1**	0,1534	0,8466	0,1813	0,1137

Zdroj: Vlastné spracovanie

Z výsledkov uvedených v tabuľke 4.15 je zrejmé, že s každou úrovňou zaťaženia klientovej solventnosti rapídne klesá aj pravdepodobnosť úspechu kampane. Tá sa logicky priamo premietne aj vo výslednej šanci na úspech kampane u daného klienta. Referenčný profil skúmaného klienta bol konštruovaný s nulovou finančnou záťažou, napriek čomu bola jeho predikovaná pravdepodobnosť úspechu kampane na úrovni 61,46 %. Z pohľadu šance je iba 1,5947 – krát pravdepodobnejšie, že kampaň bude u daného klienta úspešná, oproti možnej pravdepodobnosti neúspechu.

Vo všeobecnosti sa bankovej inštitúcií z pohľadu prevádzkovania kampane neoplatí nadväzovať kontakt so žiadnym klientom, trpiacim akoukoľvek formou finančnej záťaže. Meniace sa úrovne finančnej záťaže, resp. ich vzájomné kombinácie disponujú negatívnym vplyvom, ktorý sa priamo odzrkadľuje v poklese výslednej hodnoty šance úspechu kampane.

Z pohľadu referenčného klienta bez akejkoľvek záťaže jeho solventnosti, najmenej dramatický vplyv na výslednú zmenu šancu úspechu disponujú klienti, ktorý sa v minulosti dostali do „defaultného“ stavu alebo tí, ktorí v súčasnosti splácajú osobnú pôžičku. Konkrétne referenčný klient má 1,555 – krát vyššiu šancu na úspech ako „defaultný“ klient a 1,852 – krát vyššiu šancu oproti rovnakému klientovi, ktorý v súčasnosti spláca osobnú pôžičku.

Oproti referenčnému klientovi bol zaznamenaný najviac devastujúci účinok na výslednú hodnotu šance v prípade klientov, so všetkými tromi stupňami finančnej záťaže. U klienta bez finančnej záťaže je 8,798 – krát väčšia šanca na úspech, oproti klientovi z najhoršej skupiny solventnosti. Referenčný klient ďalej disponuje 5,656 – krát vyššou šancou na úspech ako klient splácajúci pôžičku na bývanie a zároveň osobnú pôžičku;

približne 4,749 – krát vyššou šancou ako klient splácajúci pôžičku na bývane s „defaultnou“ minulosťou; 3,05 – krát vyššou šancou oproti klientovi splácajúcemu pôžičku na bývanie a 2,88 – krát vyššou šancou oproti klientovi s „defaultnou“ minulosťou a súčasne splácajúceho osobnú pôžičku.

Socidemografické charakteristiky a črty klientovej solventnosti nemusia byť jediné faktory, ovplyvňujúce výslednú šancu na úspech v kampani. Podpísať sa na výsledku môže aj spôsob, akým bola daná kampaň vedená.

Tento príklad v jednoduchosti načrtne, do akej miery môže šancu na úspech kampane ovplyvniť cielenie banky na klientov, ktorí:

– doposiaľ neboli oslovení v rámci kampane;
– majú bohatšiu históriu interakcie s bankovou inštitúciou.

Zároveň bolo preskúmané, do akej miery môže šancu na úspech ovplyvniť zintenzívnenie na dĺžke rozhovorov s jednotlivými klientami. Výsledne hodnoty skúmania v porovnaní s referenčným klientom sú obsiahnuté v tabuľke 4.16

Tabuľka 4.16 Vplyv stratégie kampane na výslednú hodnotu šance úspechu

Premenná = Prev, C, Dur		Pravdepodobnosť úspechu	Pravdepodobnosť neúspechu	Šanca	Pomer šancí
Referenčné	**C** = 2 , **Prev** = 1, **Dur** = 600	0,6146	0,3854	1,5950	-
Porovnanie 1	**C** = 0, **Prev** = 1, **Dur** = 600	0,6723	0,3277	2,0515	1,2863
Porovnanie 2	**C** = 0, **Prev** = 1, **Dur** = 1200	0,9565	0,0435	21,9777	13,7794
Porovnanie 3	**C** = 0, **Prev** = 5, **Dur** = 600	0,7472	0,2528	2,9559	1,8533
Porovnanie 4	**C** = 0, **Prev** = 5, **Dur** = 1200,	0,9694	0,0306	31,6662	19,8539

Zdroj: Vlastné spracovanie

Na základe výsledkov skúmania zachytených v tabuľke 4.16 je zrejmé, že šanca na úspech kampane jemne rastie oproti referenčnému klientovi v prípade, ak rovnaký klient doposiaľ nebol oslovený v rámci aktuálnej kampane a v minulosti bol bankovou inštitúciou viac krát kontaktovaný. Konkrétne klient, ktorý nebol doposiaľ oslovený v rámci kampane, disponuje 1,2863 – krát vyššou šancou ako referenčný klient, ceteris paribus. Zároveň klient, ktorý doposiaľ nebol oslovený v rámci kampane a banková inštitúcia nadviazala s týmto klientom kontakt päťkrát v minulosti, disponuje 1,853 – krát vyššou šancou na úspech kampane, oproti referenčnému klientovi. Táto skutočnosť naznačuje, že kampaň môže byť

efektívnejšia u klientov, ktorí neboli doposiaľ oslovení v rámci kampane. Zároveň vyšší počet kontaktov klienta v minulosti sa pozitívne odráža v šanci na úspech aktuálnej kampane. Tento trend naznačuje, že klienti s bohatšou históriou interakcií s bankovou inštitúciou môžu byť kľúčovým faktorom pre úspešnosť marketingových stratégií. Nie však až do takej miery ako v prípade, keby banková inštitúcia zintenzívnila dĺžku rozhovorov s klientami o 100 %. Na základe zvýšenia dĺžky rozhovorov z 10 na 20 minút možno pozorovať najsignifikantnejšie výsledky v zmene šance na úspech kampane u daného klienta. Klient, s ktorým bol vedený rozhovor v tejto dĺžke a v minulosti bol bankou päťkrát kontaktovaný, disponuje 19,854 – krát väčšou šancou na úspech, oproti referenčnému klientovi. Pre porovnanie, klient s ktorým bol vedený 20 minútový rozhovor, pričom bol v minulosti kontaktovaný iba jedenkrát, šanca na úspech kampane je u neho vyššia iba 13,779 – krát oproti referenčnému klientovi. To iba opäť podčiarkuje tvrdenie o dôležitosti bohatej interakčnej histórie bankovej inštitúcie s klientom. Zároveň to naznačuje, že ak je klient ochotný v dialógu zotrvať dlhšie, rastie šanca úspechu kampane u takéhoto klienta.

ZÁVER

Hlavným cieľom diplomovej práce bolo prezentovanie aplikačných možností nelineárnych pravdepodobnostných modelov (s binárnou závislou premennou) v kontexte praktického príkladu z oblasti marketingového výskumu.

Sekundárny cieľ práce predstavoval vhodné definovanie riešenej problematiky v rámci teoretickej časti práce. V obsahu prvej kapitoly boli stručne predstavené základné teoretické východiská a kľúčové pojmy. Vysvetlený bol rozdiel medzi úlohami regresnej analýzy a klasifikačnými úlohami, ktorými sa v súčasnosti zaoberá sféra strojového učenia s učiteľom. Nelineárne pravdepodobnostné modely tvoria prienik práve týchto disciplín, teda je možné ich využiť rovnako v zmysle regresie, tak i klasifikácie. V obsahu tejto kapitoly boli definované aj ich všeobecné možnosti ekonomického uplatnenia. Prezentované boli aj alternatívne modely zo sféry strojového učenia s učiteľom, ktoré je možné aplikovať v oblasti riešenia úloh binárnej klasifikácie. Záver prvej kapitoly bol venovaný prehľadu literatúry domácich a zahraničných autorov, ktorí sa zaoberali praktickou aplikáciou nelineárnych pravdepodobnostných modelov v oblastiach predikcie bankrotu ekonomických subjektov, bankovníctva a poisťovníctva. Tieto práce tvorili aj metodologický prehľad, ktorým bola inšpirovaná aplikačná časť práce. Metódy, ktoré boli využité v rámci praktickej časti práce boli úspešne definované v rámci tretej kapitoly. Tá ponúka teoretické pozadie pre aplikovanie dichotomického modelu logit a probit, metódu odhadu parametrov, testovanie štatistickej významnosti parametrov a modelov ako celku, interpretáciu odhadnutých parametrov a možné spôsoby interpretácie ich výsledkov. Záver kapitoly bol venovaný definícií metód, ktoré boli využité na validáciu predikčnej schopnosti modelov a ich vhodnosti prispôsobenia sa skúmaným údajom.

Spracovanie teoretických náležitostí v rámci prvej a tretej kapitoly boli následne aplikované na praktickom príklade v rámci štvrtej kapitoly práce.

Modely logistickej a probitovej regresie boli aplikované na historických údajoch o klientoch nemenovanej portugalskej bankovej inštitúcie. Tie boli zozbierané v rámci telefonickej marketingovej kampane, ktorej cieľom bolo presvedčiť klienta, aby vložil svoje voľné aktíva na termínovaný vklad. Vzhľadom k tomu, že účelom aplikácie modelov bolo klasifikovať klientov na základe predikovanej pravdepodobnosti úspechu kampane buď do skupiny úspešnej alebo neúspešnej, závislá premenná vyjadrovala výsledok tejto kampane u pozorovaného klienta. Výber vhodnej kombinácie nezávislých premenných,

charakterizujúcich klientove sociodemografické charakteristiky, indikátory solventnosti a indikátory spôsobu vedenia kampane, bol založený na algoritme „stepwise“ v dvoch variantoch – „backward elimination“ a „forward selection.“ Algoritmus v oboch variantoch porovnával modely na základe ich líšiacej sa miery komplexnosti, prostredníctvom ich výsledných hodnôt AIC a BIC. Aplikovaný algoritmus v rôznych variáciách celkovo priniesol šesť kandidátnych modelov (3 pre logit a 3 pre probit), z ktorých bol na predikciu pravdepodobnosti zvolený práve jeden logitový a probitový model. Výber najvhodnejšieho modelu zohľadňoval porovnávanie výsledných hodnôt $R^2_{McFadden}$, AIC a BIC. Testovanie štatistickej významnosti parametrov bolo realizované prostredníctvom Waldovho testu, pričom väčšina testovaných parametrov sa preukázala ako štatisticky významná. Štatistická významnosť modelu ako celku bola realizovaná prostredníctvom testu pomeru vierohodnosti, ktorý preukázal štatistickú významnosť všetkých testovaných modelov.

Pri predikcií pravdepodobnosti a následnej klasifikácií klientov do úspešnej alebo neúspešnej skupiny bolo preukázané, že oba zvolené modely priniesli takmer totožné výsledky. Úspešnosť realizovanej klasifikácie bola prvotne zmapovaná prostredníctvom matice zámen, z ktorej boli následne vypočítané relatívne charakteristiky predikčnej schopnosti modelov. Oba nelineárne pravdepodobnostné modely priniesli takmer totožné predikčné výsledky na trénovacej a testovacej množine údajov, čím bolo možné vylúčiť, že modely disponovali známkami preučenia. Spoločnými charakteristikami pre oba modely bola správnosť klasifikácie všetkých pozorovaní, na úrovni približne 89 %. Kategórie závislej premennej modelov boli v značne nevyváženom pomere, pričom prevládala majoritná trieda neúspešných klientov oproti minoritnej triede úspešných klientov. To sa odzrkadlilo na tendencií modelov chybne zamieňať prípady minoritnej triedy, za prípady triedy majoritnej. Svedčí o tom aj skutočnosť, že oba modely dokázali presne vyhodnotiť klientov úspešnej skupiny iba v približne 60 % prípadov čo naznačuje, že 40 % klientov, ktorí boli vyhodnotení ako členovia úspešnej skupiny, v skutočnosti patrili do triedy neúspešnej. Na druhú stranu, oba modely pomerne správne vyhodnocovali klientov triedy neúspešnej. Svedčí o tom pomerne vysoká miera dosiahnutej miery špecifickosti modelov na úrovni cca. 98 % a pomerne nízka miera výpadku na úrovni približne 2 %. Dosiahnutá nízka miera senzitívnosti modelov naopak opäť podčiarkla tvrdenie, že modely výrazne zlyhávali v prípade správneho vyhodnotenia členov úspešnej marketingovej skupiny. V tomto prípade priniesol logit model mierne uspokojivejšie výsledky (cca. 21 %) oproti modelu probit (cca. 18 %). Výber vhodnejšieho modelu v zmysle dosiahnutých

klasifikačných výsledkov prostredníctvom harmonického priemeru mier presnosť a senzitívnosť naznačoval, že model logit sa prejavoval ako vhodnejší kandidát použitia (dosiahnuté hodnoty cca. 31 – 32 %), oproti modelu probit (cca. 28 %). Vzhľadom na limitácie, spôsobené nevyváženosťou kategórií závislej premennej, bola táto vhodnosť vyhodnocovaná aj prostredníctvom ROC-AUC analýzy. Výsledky z nej plynúce ukázali, že oba modely boli rovnako vhodné, v prípade ich využitia pri klasifikácií klientov, pričom sa prejavovali pomerne dobrou mierou klasifikačnej schopnosti – oba modely dosahovali hodnotu AUC na úrovni cca. 86 %.

Za účelom odhalenia cieľovej skupiny klientov v prípade opakovania kampane bolo analyzované decilové rozdelenie vzostupne zoradenej predikovanej pravdepodobnosti úspechu, do ktorých sa premietla absolútna početnosť skutočnej príslušnosti klientov v oboch skupinách. Z výsledkov analýzy vyplynulo, že pokiaľ chce banková inštitúcia zefektívniť výsledky pri opakovaní marketingovej kampane na rovnakej vzorke klientov, mala by sa zamerať na tých, ktorých predikovaná pravdepodobnosť úspechu sa nachádzala v posledných štyroch pravdepodobnostných deciloch. V nich sa preukázala najvyššia koncentrácia úspešnej skupiny klientov na relatívnej úrovni 88,4 %, zároveň tak aj najnižšia koncentrácia neúspešnej skupiny klientov na úrovní 33,6 %. Cielenie na posledných 40 % populácie by mohlo dramaticky znížiť náklady spojené s prevádzkovaním kampane, čo by sa mohlo odzrkadliť na vyššej návratnosti kampane. Ušetrené náklady by mohla banková inštitúcia pretaviť do rôznych inovatívnych možností vedenia budúcich kampaní, poprípade na optimalizovanie stratégií oslovovania klientov, ktorých charakteristiky sú najviac spojené s možným dosiahnutým úspechom kampane.

Odhalenie kľúčových determinantov na výslednú pravdepodobnosť úspechu kampane u klientov bolo v prípade modelov logit a probit realizované prostredníctvom analyzovania hodnôt ich marginálnych efektov. Tie boli vypočítané na priemerných hodnotách regresorov. V prípade modelu logit boli analyzované aj pomery šancí, v zmysle jednotkovej zmeny určitej klientovej charakteristiky, tak aj na menení rôznych úrovní hodnôt determinantov referenčného klienta. To dopomohlo odhaliť kľúčové faktory, ktoré by mala banková inštitúcia zohľadniť pri oslovovaní nových klientov, v prípade realizácie budúcich marketingových kampaní. Banka by sa mala konkrétne zamerať na pracovníkov administratívy, dôchodcov a študentov, s ideálne najvyšším dosiahnutým vysokoškolským vzdelaním. Z pohľadu finančného profilu klienta by mali byť oslovení tí, ktorí netrpia žiadnou formou alebo kombináciou úrovní finančnej záťaže, v podobe určitej formy úveru,

prípadne „defaultnej“ minulosti. Výška finančného zostatku na bankovom účte klienta sa síce preukázala ako významný faktor, avšak vo výraznej miere nedeterminovala výrazné zvýšenie možnosti úspechu. V prípade zefektívnenia stratégií by sa mala banka zamerať na klientov, ktorí neboli doposiaľ oslovení v rámci aktuálnej kampane a zároveň disponujú bohatou interakčnou minulosťou s bankou v minulých kampaniach. Preukázalo sa, že zintenzívnenie dĺžky telefonátu s klientom ma signifikantný dopad na úspech kampane. Vzhľadom na túto skutočnosť, by mohla banka prísť so spôsobom efektívnej prezentácie a dôkladného, prípadne cieleného vysvetlenia výhod plynúcich z ponúkaného produktu.

Možno konštatovať, že stanovené ciele diplomovej práce sa podarilo úspešne naplniť. Zistené závery práce by mohli bankovej inštitúcií potencionálne dopomôcť pri optimalizovaní marketingových procesov. Zároveň práca môže slúžiť ako podrobne spracovaný podklad pre príbuzné štúdie zaoberajúce sa touto problematikou.

ZOZNAM BIBLIOGRAFICKÝCH ODKAZOV

Knižné zdroje:

Bačíková, M., & Janovská, A. (2018). *Základy metodológie pedagogicko-psychologického výskumu. Sprievodca pre študentov učiteľstva*. ŠafárikPress. ISBN 978-80-8152-695-4.

Fox, J. (2016). *Applied regression analysis and generalized linear models*. Sage Publications. ISBN 978-1-4522-0566-3.

Greene, W. H. (2012). *Econometric analysis (7th ed.)*. Pearson Education. ISBN 978-0-273-75356-8.

Hendl, J. (2021). *Big data: věda o datech-základy a aplikace*. Grada Publishing. ISBN 978-80-271-3031-3.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013*). Applied logistic regression*. John Wiley & Sons. ISBN 978-0-470-58247-3.

Hope, Thomas MH. (2020). *"Linear regression." In Machine Learning*, pp. 67-81. Academic Press. ISBN 978-0128157398.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: Springer. ISBN 978-1-0716-1417-4.

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: algorithms*. *Worked examples, and case studies*, The MIT Press. ISBN 9780262029445.

Klein, D. (2020). *Pokročilé štatistické metódy*. Univerzita Pavla Jozefa Šafárika v Košiciach. ISBN 978-80-8152-915-3.

Laura, I., & Santi, S. (2017). *Introduction to data science: a python approach to concepts, techniques and applications*. Springer. ISBN 978-3-319-50016-4.

Machová, K. (2016). *Nové trendy v strojovom učení: Štatistický prístup. Edícia vedeckých spisov Fakulty elektrotechniky a informatiky*, Technická univerzita v Košiciach. ISBN: 978-80-553-2602-3.

Neubauer, J., Sedlačík, M., & Kříž, O. (2021). *Základy statistiky: aplikace v technických a ekonomických oborech*. Grada Publishing. ISBN 978-80-271-3421-2.

Rubliková E., Labudová V., Sandtnerová S. (2009). *Analýza kategoriálnych údajov*. Ekonóm. ISBN 978-80-225-2710-1

Elektronické zdroje:

Abid, L., Masmoudi, A., & Zouari-Ghorbel, S. (2018). *The consumer loan's payment default predictive model: An application of the logistic regression and the discriminant analysis in a Tunisian commercial bank*. Journal of the Knowledge Economy, 9, 948-962. Dostupné na: https://doi.org/10.1007/s13132-016-0382-8

Ali, P., & Younas, A. (2021). *Understanding and interpreting regression analysis*. Evidence-Based Nursing, 24(4), 116-118. Dostupné na: https://doi.org/10.1136/ebnurs-2021-103425

Alzubi, J., Nayyar, A., & Kumar, A. (2018). *Machine learning from theory to algorithms: An overview*. In Journal of Physics: Conference Series (Vol. 1142, p. 012012). IOP Publishing. Dostupné na: https://iopscience.iop.org/article/10.1088/1742-6596/1142/1/012012/pdf

Bowers, A. J., & Zhou, X. (2019). *Receiver operating characteristic (ROC) area under the curve (AUC): A diagnostic measure for evaluating the accuracy of predictors of education outcomes*. Journal of Education for Students Placed at Risk (JESPAR), 24(1), 20-46. Dostupné na: https://doi.org/10.1080/10824669.2018.1523734

Coss, S. (2015). *Modely s binárnou závislou premennou*. Economics And Informatics, 2015, 13.2. Dostupné na: https://ei.fhi.sk/index.php/EAI/article/view/18

Fomby, T. B. (2010). *Maximum Likelihood Estimation of Logit and Probit Models*. Department of Economics, Southern Methodist University. Dostupné na: https://s2.smu.edu/tfomby/eco6352/Notes/Logit%20and%20Probit%20Notes.pdf

García Portugués, E. (2023). *Notes for predictive modeling*. Version 5.9.12. Dostupné na: https://bookdown.org/egarpor/PM-UC3M/

Güneri, Ö. İ., & Durmuş, B. (2020). *Dependent dummy variable models: An application of logit, probit, and tobit models on survey data*. International Journal of Computational and Experimental Science and Engineering, 6(1), 63-74. Dostupné na: https://dergipark.org.tr/en/download/article-file/1009887

Gupta, A., Sharma, A., & Goel, A. (2017). *Review of regression analysis models*. Int. J. Eng. Res. Technol, 6(08), 58-61. Dostupné na: https://www.ijert.org/research/review-of-regression-analysis-models-IJERTV6IS080060.pdf

Hřebíček, J., & Škrdla, M. (2006). *Úvod do matematického modelování*. Masarykova univerzita, Brno. Dostupné na: https://is.muni.cz/el/1431/podzim2007/Bi3101/um/skripta.pdf

Hong, C. S., & Oh, T. G. (2021). *TPR-TNR plot for confusion matrix.* Communications for Statistical Applications and Methods, 28(2), 161-169. Dostupné na: https://doi.org/10.29220/CSAM.2021.28.2.161

Hoo, Z. H., Candlish, J., & Teare, D. (2017). *What is an ROC curve?*. Emergency Medicine Journal, 34(6), 357-359. Dostupné na: https://doi.org/10.1136/emermed-2017-206735

IBM. (2023). *Pseudo R Square. In SPSS Statistics Subscription documentation*. Dostupné na: https://www.ibm.com/docs/en/spss-statistics/saas?topic=model-pseudo-r-square

Institute for Digital Research and Education. (2021). *FAQ: What are pseudo R-squareds?* Dostupné na: https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/

Klacso, J., & Stulrajterova, E. (2021). *Determinants of labour market flows in Slovakia (No. WP 5/2021).* Research Department, National Bank of Slovakia. ISSN 2585-9269. Dostupné na: https://nbs.sk/_img/documents/publik/wp_5_2021_klacso_stulrajterova_determinants_of_labour_market_flows_in_slovakia.pdf

Kovacova, M., & Kliestik, T. (2017). *Logit and Probit application for the prediction of bankruptcy in Slovak companies*. Equilibrium. Quarterly Journal of Economics and Economic Policy, 12(4), 775–791. doi: 10.24136/eq.v12i4.40

Kuhn, M., Johnson, K. (2013). *Over-Fitting and Model Tuning*. Applied Predictive Modeling. Springer, New York, NY. Dostupné na: https://doi.org/10.1007/978-1-4614-6849-3_4

Lobos, G., Viviani, J.-L., Schnettler, B., Tigero, T., & Reyes, Á. (2012). *Predicting probability to purchase insurance contracts in the Chilean wine industry: a logit and probit comparative analysis.* Ciência e Técnica Vitivinícola, 26, 55-62. Dostupné na:

https://www.researchgate.net/publication/235726265_Predicting_probability_to_purchase_insurance_contracts_in_the_Chilean_wine_industry_a_logit_and_probit_comparative_analysis

Lopez-Bernal, D., Balderas, D., Ponce, P., & Molina, A. (2021). *Education 4.0: teaching the basics of KNN, LDA, and simple perceptron algorithms for binary classification problems.* Future Internet, 13(8), 193. Dostupné na: https://doi.org/10.3390/fi13080193

Lukáčik, M. (2008). *EKONOMICKÉ APLIKÁCIE MODELU LOGIT ECONOMIC APPLICATIONS OF LOGIT MODEL*. MEDZINÁRODNÝ SEMINÁR MLADÝCH. Dostupné na: http://fhi.sk/files/netrinecop/Praha2008.pdf#page=115

Mahesh, B. (2020). *Machine learning algorithms-a review*. International Journal of Science and Research (IJSR), 9(1), 381-386. Dostupné na: https://www.ijsr.net/archive/v9i1/ART20203995.pdf

Mohamed, S., Ashraf, R., Ghanem, A., Sakr, M., & Mohamed, R. (2022). *Supervised Machine Learning Techniques: A Comparison*. Dostupné na: https://www.researchgate.net/publication/363870735_Supervised_Machine_Learning_Techniques_A_Comparison

Mozos, Ó. M. (2010). S*upervised Learning. Semantic Labeling of Places with Mobile Robots*. Springer Tracts in Advanced Robotics, vol 61. Springer, Berlin, Heidelberg. Dostupné na: https://doi.org/10.1007/978-3-642-11210-2_2

Muschelli III, J. (2020). *ROC and AUC with a binary predictor: A potentially misleading metric*. Journal of Classification, 37(3), 696-708. Dostupné na: https://doi.org/10.1007/s00357-019-09345-1

Nahhas, W. (2023). *Introduction to Regression Methods for Public Health Using R.* Dayton: Creative Common. Dostupné na: https://www.bookdown.org/rwnahhas/RMPH/blr-interp.html

Obi, J. C. (2023). *A comparative study of several classification metrics and their performances on data*. World Journal of Advanced Engineering Technology and Sciences, 8(1), 308-314. Dostupné na: https://doi.org/10.30574/wjaets.2023.8.1.0054

Schröder, G., Thiele, M., & Lehner, W. (2011, October). *Setting goals and choosing metrics for recommender system evaluations*. UCERSTI2 workshop at the 5th ACM conference on recommender systems, Chicago, USA (Vol. 23, p. 53). Dostupné na: https://www.researchgate.net/publication/268381252_Setting_Goals_and_Choosing_Metrics_for_Recommender_System_Evaluations

Singh, P., Singh, N., Singh, K. K., & Singh, A. (2021). *Diagnosing of disease using machine learning*. In Machine Learning and the Internet of Medical Things in Healthcare (pp. 89-111). Academic Press. Dostupné na: https://doi.org/10.1016/B978-0-12-821229-5.00003-3

Šikyňa, M. (2019). *Lineárna a polynomiálna regresia*. Course material: UMĚLÁ INTELIGENCE I. Dostupné na: https://nlp.fi.muni.cz/uui/referaty2019/sikyna_matus/referat.pdf

Štatistický úrad Slovenskej republiky. (2020). *Štatistické pojmy*. Dostupné na: https://slovak.statistics.sk/ExportPdf2/PdfExportSrvlt?Document=2347faf6-1064-4abd-a71d-1db75a23065c

Wadi, M. (2021). *Fault detection in power grids based on improved supervised machine learning binary classification*. Journal of Electrical Engineering, 72(5) 315-322. Dostupné na: https://doi.org/10.2478/jee-2021-0044

Wahono, R. S., Herman, N. S., & Ahmad, S. (2014). *A comparison framework of classification models for software defect prediction*. Advanced Science Letters, 20(10-11), 1945-1950. Dostupné na: https://doi.org/10.1166/asl.2014.5640

ZOZNAM PRÍLOH

PRÍLOHA A OPIS ATRIBÚTOV ZÁKLADNÉHO SÚBORU ÚDAJOV

Poradie	Atribút	Popis	Typ premennej	Unikátne obmeny premennej	Popis
1.	Age	Vek klienta	Kvantitatívna	-	-
2.	Job	Zamestnanie klienta	Kategoriálna	1. "admin"	Administratívny pracovník
				2. "unknown"	Neznáme povolanie
				3. "unemployed"	Klient je nezamestnaný
				4. "management"	Práca vo sfére managementu
				5. "housemaid"	Starostlivosť o domácnosť
				6. "entrepreneur"	Klient je podnikateľ
				7. "student"	Klient je študentom
				8. "blue-collar"	Klient pracujúci ako konštrukčný pracovník
				9. "self-employed"	Samostatná zárobková činnosť
				10. "retired"	Klient banky je dôchodca
				11. "technician"	Klient je povolaním technik
				12. "services"	Klient pracujúci v službách
3.	Martial	Rodinný stav klienta	Kategoriálna	1. "married"	Klient je v manželstve
				2. "divorced"	Klient je rozvedený (v údajoch sú zahrnutí aj klienti, ktorí v minulosti ovdoveli)
				3. "single"	Klient je nezadaný
4.	Education	Úroveň najvyššieho dosiahnutého vzdelania klienta	Kategoriálna	1. "unknown"	Vzdelanie klienta je neznáme
				2. "primary"	Klient dosiahol základné vzdelanie
				3. "secondary"	Klient dosiahol sekundárne - stredoškolské vzdelanie
				4. "tertiary"	Klient dosiahol terciárne - vysokoškolské vzdelanie
5.	Default	Defaultný stav klienta v minulosti	Kategoriálna (Binárna)	1. "yes"	Klient sa dostal do defaultného stavu v minulosti
				2. "no"	Klient sa dostal do defaultného stavu v minulosti
6.	Balance	Zostatok finančných prostreiedkov na účte klienta	Kvantitatívna (spojitá)	-	-

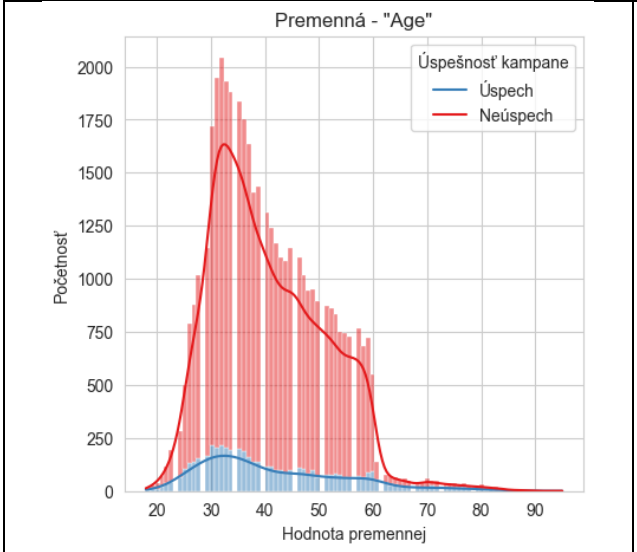
7.	**Housing**	Klient je platiteľ úveru na bývanie	**Kategoriálna (Binárna)**	**1. "yes"**	Klient spláca poskytnutý úver na bývanie
				2. "no	Klient nespláca poskytnutý úver na bývanie
8.	**Loan**	Klient spláca osobnú pôžičku	**Kategoriálna (Binárna)**	**1. "yes"**	Klient pôžičku spláca
				2. "no"	Klient pôžičku nespláca
9.	**Contact**	Spôsob oslovenia klienta v marketingovej kampani banky	**Kategoriálna**	**1. "unknown"**	Hodnota nie je známa
				2. "telephone"	Obmena indikujúca kontaktovanie klienta prostredníctvom telefónu pevnej linky
				3. "cellular"	Obmena indikujúca kontaktovanie klienta prostredníctvom mobilného telefónu
10.	**Day**	Deň v mesiaci, v ktorom bol klient naposledy oslovený v rámci kampane	**Kvantitatívna**	**-**	**-**
11.	**Month**	Mesiac, v ktorom bol klient banky naposledy oslovený v rámci marketingovej kampane	**Kategoriálna**	**1. "jan"**	**-**
				2. "feb"	
				3. "mar"	
				...	
				11. "nov"	
				12. "dec"	
12.	**Duration**	Dĺžka poslednej komunikácie s klientom banky v rámci marketingovej kampane (vyjadrená v sekundách)	**Kvantitatívna**	**-**	**-**
13.	**Campaign**	Počet kontaktov uskutočnených počas kampane u konkrétneho klienta	**Kvantitatívna**	**-**	**-**
14.	**Pdays**	Počet dní, ktoré uplynuli od posledného kontaktovania klienta z predchádzajúcej kampane	**Kvantitatívna**	**-**	**-**
15.	**Previous**	Počet kontaktov pre konkrétneho klienta, uskutočnených pred touto kampaňou	**Kvantitatívna**	**-**	**-**
16.	**Poutcome**	Výsledok predchádzajúcej marketingovej kampane	**Kategoriálna**	**1. "unknown"**	Neznáma hodnota
				2. "other"	Iná hodnota
				3. "failure"	Neúspech minulej kampane
				4. "success"	Úspech minulej kampane
17.	**Y**	Vklad finančných prostriedkov klienta na termínovaný vklad, po jeho oslovení v kampani	**Kategoriálna (Binárna)**	**1. "yes"**	Klient vložil finančné prostriedky na termínovaný vklad
				2. "no"	Klient nevložil finančné prostriedky na termínovaný vklad

Zdroj: Vlastné spracovanie

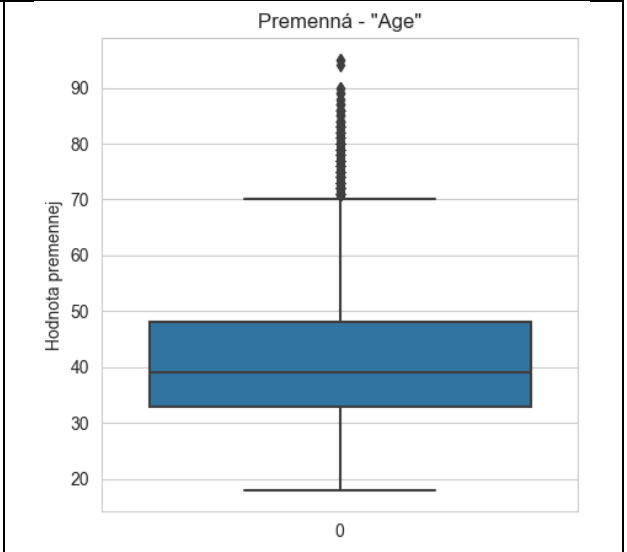
PRÍLOHA B SOCIODEMOGRAFICKÝ PROFIL OSLOVENÝCH KLIENTOV

Vekové rozdelenie klientov banky:

a) Histogram vekového rozdelenia klientov



b) Krabicový graf vekového rozdelenia klientov



c) Popisné štatistiky veku oslovených klientov

Početnosť	Priemer	Smerodajná odchýlka	Minimum	Dolný Kvartil	Medián	Horný kvartil	Maximum
43 193	40, 76	10, 51	18	33	39	48	95

Zdroj: Vlastné spracovanie

Najvyššie dosiahnuté vzdelanie klientov:

Atribút Education	Úspešnosť kampane		Spolu:
	Neúspech (no)	Úspech (yes)	
primary	6 212	588	6 800
secondary	20 690	2 441	23 131
tertiary	11 270	1 992	13 262
Spolu:	38 172	5 021	43 193

Zdroj: Vlastné spracovanie

Zamestnanie klientov banky:

Atribút Job	Úspešnosť kampane		Spolu:
	Neúspech (no)	Úspech (yes)	
admin	4 387	613	5 000
blue-collar	8 603	675	9 278
entrepreneur	1 295	116	1 411
housemaid	1 090	105	1 195
management	7 963	1 253	9 216
retired	1 659	486	2 145
self-employed	1 358	182	1 540
services	3 654	350	4 004
student	549	226	775
technician	6 538	817	7 355
unemployed	1 076	198	1 274
Spolu:	38 172	5 021	43 193

Zdroj: Vlastné spracovanie

Rodinný stav klientov banky:

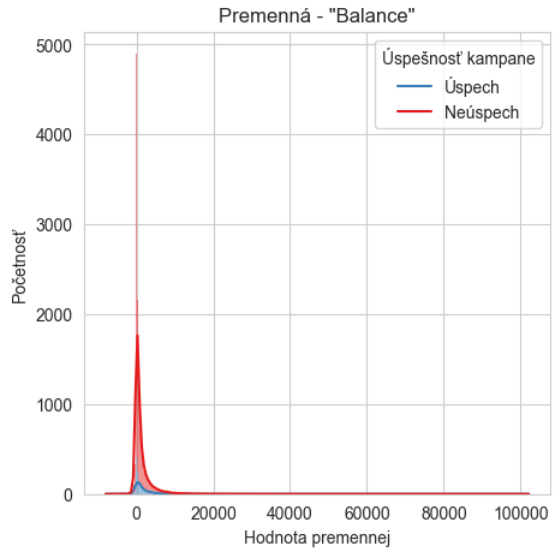
Atribút Marital	Úspešnosť kampane		Spolu:
	Neúspech (no)	Úspech (yes)	
divorced	4 430	598	5 028
married	23 343	2 603	25 946
single	10 399	1 820	12 219
Spolu:	38 172	5 021	43 193

Zdroj: Vlastné spracovanie

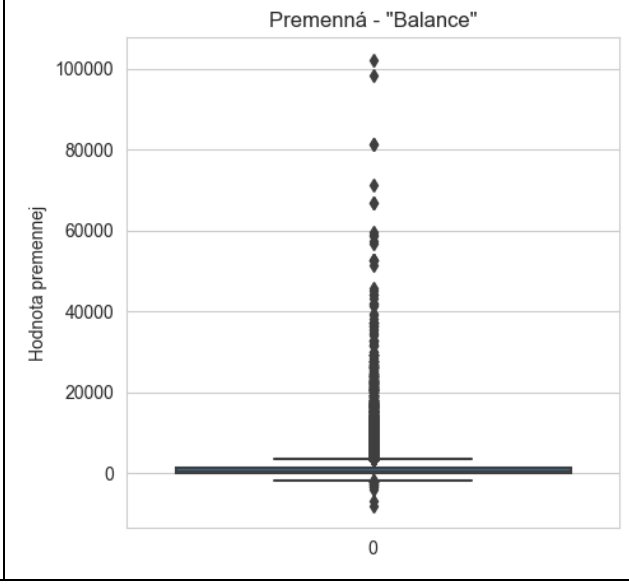
PRÍLOHA C PROFIL SOLVENTNOSTI OSLOVENÝCH KLIENTOV

Zostatok na bankovom účte klienta:

a) Histogram rozdelenia zostatku na bankovom účte klientov



b) Krabicový graf rozdelenia zostatku na bankovom účte klientov



c) Popisné štatistiky zostatku finančných prostriedkov na bankovom účte klienta

Početnosť	Priemer	Smerodajná odchýlka	Minimum	Dolný Kvartil	Medián	Horný kvartil	Maximum
43 193	1 354,03	3 042, 10	- 8 019	71	442	1 412	102 127

Zdroj: Vlastné spracovanie

Kreditná história klienta:

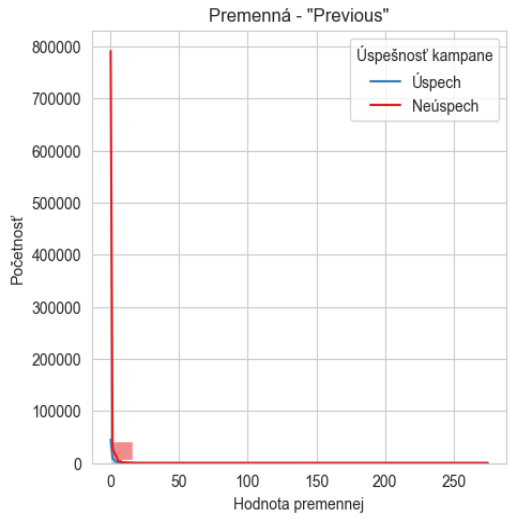
Atribút:	Default			Housing			Loan		
Úspešnosť kampane	Nie (no)	Áno (yes)	Spolu:	Nie (no)	Áno (yes)	Spolu:	Nie (no)	Áno (yes)	Spolu:
Neúspech (no)	37 438	734	38 172	15 754	22 418	38 172	31 538	6 634	38 172
Úspech (yes)	4 973	48	5 021	3 147	1 874	5 021	4 548	473	5 021
Spolu:	42 411	782	43 193	18 901	24 292	43 193	36 086	7 107	43 193

Zdroj: Vlastné spracovanie

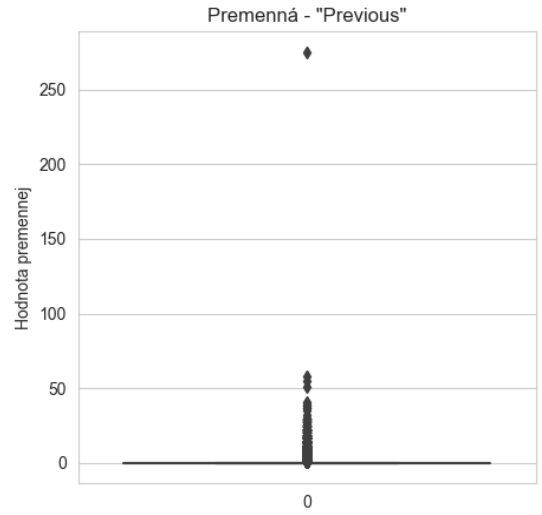
PRÍLOHA D PROFIL PRIEBEHU MARKETINGOVEJ KAMPANE

Počet uskutočnených kontaktov s klientami v minulosti (mimo kampane):

a) Histogram rozdelenia počtu minulých kontaktov



b) Krabicový graf rozdelenia počtu minulých kontaktov



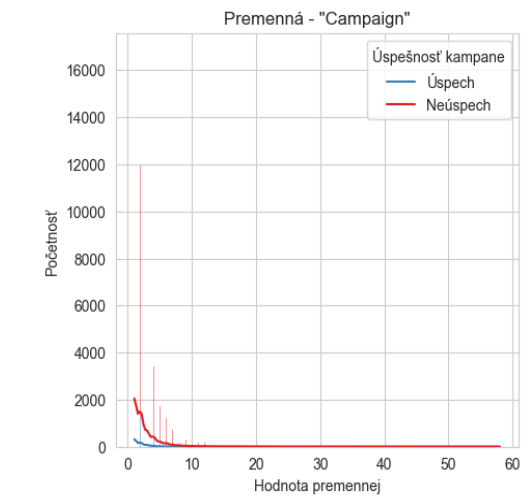
c) Popisné štatistiky počtu uskutočnených minulých kontaktov s klientom

Početnosť	Priemer	Smerodajná odchýlka	Minimum	Dolný Kvartil	Medián	Horný kvartil	Maximum
43 193	0, 58	2, 33	0	0	0	0	275

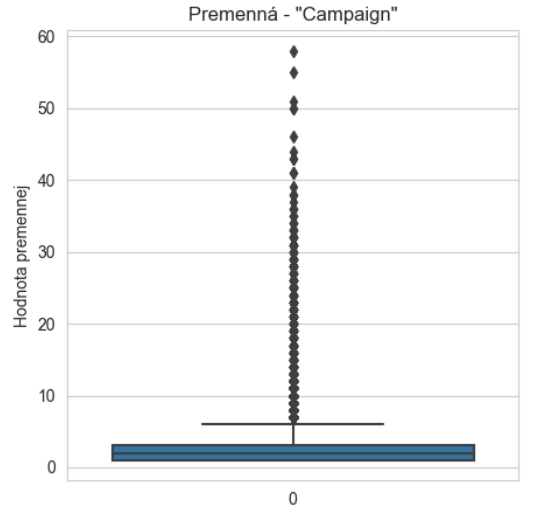
Zdroj: Vlastné spracovanie

Počet uskutočnených kontaktov s klientom v rámci kampane:

a) Histogram rozdelenia počtu uskutočnených kontaktov s klientom v rámci kampane



b) Krabicový graf rozdelenia počtu uskutočnených kontaktov s klientom v rámci kampane

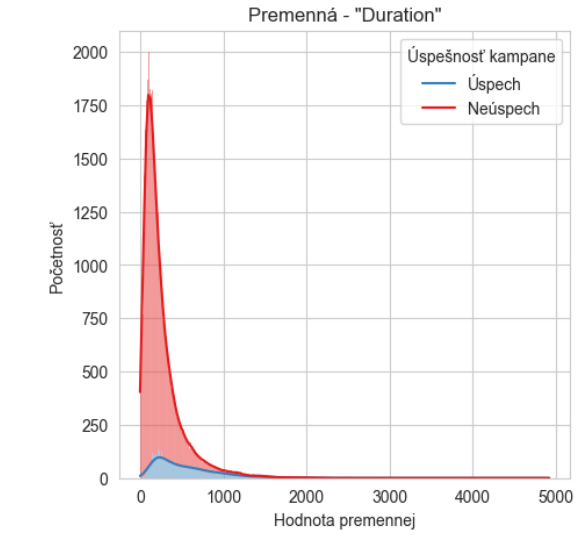
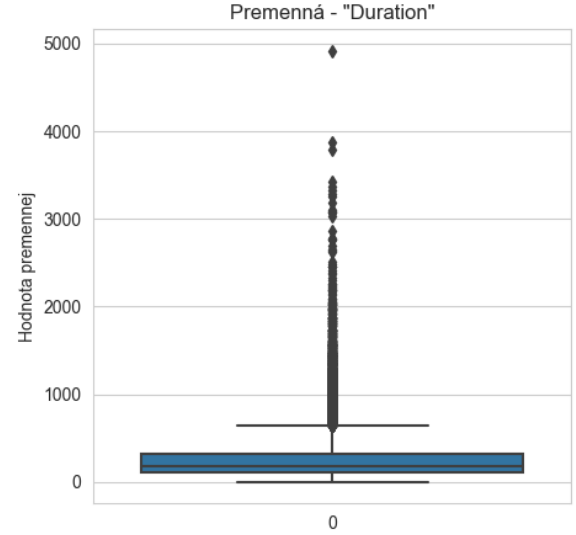


c) Popisné štatistiku pre počet uskutočnených kontaktov s klientov v rámci kampane

Početnosť	Priemer	Smerodajná odchýlka	Minimum	Dolný Kvartil	Medián	Horný kvartil	Maximum
43193	2,76	3,06	1	1	2	3	58

Zdroj: Vlastné spracovanie

Dĺžka poslednej komunikácie s klientom:

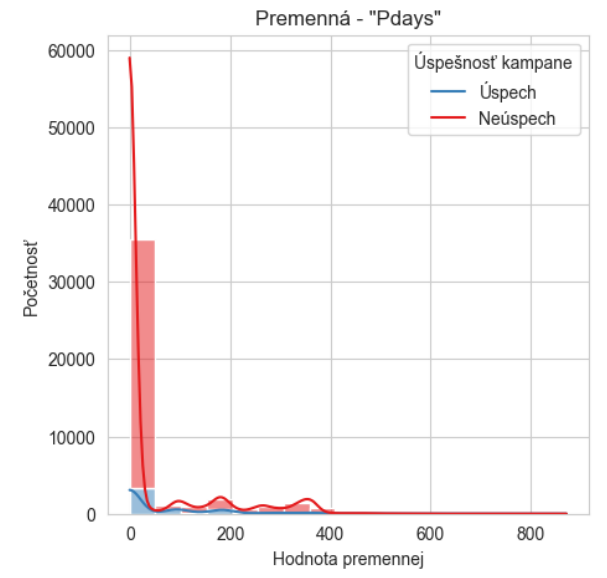
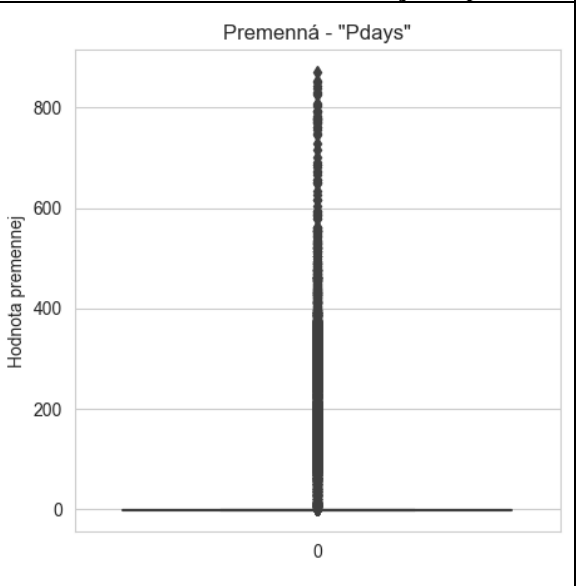
a) Histogram rozdelenia dĺžky komunikácie	b) Krabicový graf rozdelenia dĺžky komunikácie
	

c) Popisné štatistiky pre dĺžku poslednej komunikácie s osloveným klientom

Početnosť	Priemer	Smerodajná odchýlka	Minimum	Dolný Kvartil	Medián	Horný kvartil	Maximum
43 193	258, 32	258, 16	0	103	180	318	4 918

Zdroj: Vlastné spracovanie

Počet dní, ktoré uplynuli od kontaktovania klientov od predchádzajúcej kampane:

a) Histogram rozdelenia uplynutých dní od kontaktovania klienta v minulej kampani	b) Krabicový graf rozdelenia uplynutých dní od kontaktovania klienta v minulej kampani
	

c) Popisné štatistiky počtu uplynutých dní od kontaktovania klienta po minulej kampani

Početnosť	Priemer	Smerodajná odchýlka	Minimum	Dolný Kvartil	Medián	Horný kvartil	Maximum
43 193	40, 40	100, 42	-1	-1	-1	-1	871

Zdroj: Vlastné spracovanie

PRÍLOHA E ZDROJOVÝ KÓD

```
#install.packages(c("caret","ROSE","MASS",
#"dplyr","DescTools","pROC","stargazer", "lmtest", "gridExtra", "mfx"))
library(caret)
library(ROSE)
library(ggplot2)
library(MASS)
library(dplyr)
library(knitr)
library(DescTools)
library(pROC)
library(stargazer)
library(lmtest)
library(gridExtra)
library(mfx)
# Načítanie údajov ##################################################################
base_path <- "C:/Users/juraj/OneDrive/Desktop/Aplikacia modelov s umelou závislou premennou/Data/"
files <- list("prepared_data.csv","DroppedOutliers.csv")
Data <- list()
for (file in files) {
 file_path <- paste0(base_path, file)
 data <- read.csv(file_path)
 data <- data %>% mutate(across(where(is.character), as.factor))
 Data[[file]] <- data
}
#############################################################################################
 # Definovanie pracovných funkcií:
 ## Rozdelenie údajov na trénovaciu a testovaciu množinu ############################
 split_data <- function(data,
                        seed = 1,
                        split_ratio  = 0.1) {
  set.seed(seed)
  train_properties <- createDataPartition(data$y, p = split_ratio, list = FALSE)
  train_data <- data[train_properties, ]
  test_data <- data[-train_properties, ]
  return(list(train_data = train_data, test_data = test_data))
 }
 ##########################################################################################
```

```
## Stepwise Regresia ###################################################################
StepWiseRegression <- function(full_model,
                               train_data,
                               link_function = c("logit", "probit"),
                               criterion = c("AIC", "BIC"),
                               direction = c("forward", "backward"),
                               trace = FALSE) {
  if (!criterion %in% c("AIC", "BIC")) {
    stop("Nesprávne zvolené kritérium, zadaj 'AIC'/'BIC'.")
  }
  if (!direction %in% c("forward", "backward")) {
    stop("Nesprávne zvolený smer, zadaj 'forward'/'backward'.")
  }
  k <- if (criterion == "BIC") log(nrow(train_data)) else 2
  null_model <- if (link_function == "logit") {
    glm(y ~ 1, data = train_data, family = binomial(link = "logit"))
  } else if (link_function == "probit") {
    glm(y ~ 1, data = train_data, family = binomial(link = "probit"))
  } else {
    stop("Nespravne zvolená funkcia hustoty, zadaj logit/probit.")
  }
  result <- stepAIC(full_model, null_model,
                    direction = direction, trace = trace, k = k)
  return(result)
}


## AIC,BIC,McFadden ###################################################################
# Argument - glm(y ~ ., data = train_data, ...)
GoFtable <- function(model) {
aic_value <- round(AIC(model, k = 2), digits = 3)
bic_value <- round(BIC(model), digit = 3)
pseudo_r2_value <- round(PseudoR2(model, which = "McFadden"), digit = 3)
table_data <- data.frame(
  Metric = c("AIC", "BIC", "McFadden - R2"),
  Value = c(aic_value, bic_value, pseudo_r2_value))
return(table_data)
}
```

```
## ROC curve GRAF ###############################################################
# Argument - roc_curve <- roc(test_data$y, predictions)
ROCcurve <- function(roc_curve){
roc_df <- data.frame(
 FPR = 1 - roc_curve$specificities,
 TPR = roc_curve$sensitivities
)
ggplot(roc_df, aes(x = FPR, y = TPR)) +
 geom_line(color = "blue") +
 geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red") +
 labs(title = "ROC Krivka", x = "1 - Špecificita (FPR)", y = "Senzitivita (TPR)") +
 theme_minimal()
}
###Výpočet prvkov Matice zámen ##################################################
# Argument - confusion_matrix_list:
#CM_logit <- matrix(c(11237, 214, 1186, 320), nrow = 2, byrow = TRUE)
#CM_probit <- matrix(c(11273, 178, 1235, 271), nrow = 2, byrow = TRUE)
#CM <- list(CM_logit, CM_probit)
#confusion_matrix_list = #CM
################################################################################
Get_CM_properties <- function(confusion_matrix_list){
 TN <- list()
 TP <- list()
 FN <- list()
 FP <- list()
 CM_properties <- list()
 for (matrix in confusion_matrix_list) {
  TN[[length(TN) + 1]] <- matrix[[1, 1]]
  TP[[length(TP) + 1]] <- matrix[[2, 2]]
  FN[[length(FN) + 1]] <- matrix[[1, 2]]
  FP[[length(FP) + 1]] <- matrix[[2, 1]]
 }
 return(list(TN = TN, TP = TP, FN = FN, FP = FP))
}
```

```
## Metriky odvodené z matice zámen ##############################################

# Argumenty - zoznam prvkov matice zámen z funkcie Get_CM_properties()

CM_performance_metrics <- function(CM_properties) {

 TN <- CM_properties$TN

 TP <- CM_properties$TP

 FN <- CM_properties$FN

 FP <- CM_properties$FP

 Accuracy_model <- list()

 Precision_model <- list()

 Sensitivity_model <- list()

 Specificity_model <- list()

 FallOut_model <- list()

 F1_score <- list()

 for (i in 1:length(TP)) {

  # Správnosť = (TP + TN)/(TP + FP + TN + FN)

  Accuracy_model[[i]] <- round((TP[[i]] + TN[[i]]) / (TP[[i]] + TN[[i]] + FP[[i]] + FN[[i]]), digits = 4)

  # Presnosť = (TP)/(TP + FP)

  Precision_model[[i]] <- round(TP[[i]] / (TP[[i]] + FP[[i]]), digits = 4)

  # Senzitívnosť = TP/(TP + FN)

  Sensitivity_model[[i]] <- round(TP[[i]] / (TP[[i]] + FN[[i]]), digits = 4)

  # Špecifickosť = TN/(TN + FP)

  Specificity_model[[i]] <- round(TN[[i]] / (TN[[i]] + FP[[i]]), digits = 4)

  # Miera výpadku = FP / (FP + TN)

  FallOut_model[[i]] <- round(FP[[i]] / (FP[[i]] + TN[[i]]), digits = 4)

  # F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

  F1_score[[i]] <- round(2 * (Precision_model[[i]] * Sensitivity_model[[i]]) / (Precision_model[[i]] + Sensitivity_model[[i]]), digits = 4)

 }

 return(list(Accuracy_model = Accuracy_model,

        Precision_model = Precision_model,

        Sensitivity_model = Sensitivity_model,

        Specificity_model = Specificity_model,

        FallOut_model = FallOut_model,

        F1_score = F1_score))

}
```

```
 ## Funkcia vracajúca tabuľku metrík z matice zámen ################################
 # Argument - zoznam metrík vrátený z funkcie CM_performance_metrics()
 CM_performence_metrics_table <- function(CM_performance_metrics) {
  Accuracy_model <- CM_performance_metrics$Accuracy_model
  Precision_model <- CM_performance_metrics$Precision_model
  Sensitivity_model <- CM_performance_metrics$Sensitivity_model
  Specificity_model <- CM_performance_metrics$Specificity_model
  FallOut_model <- CM_performance_metrics$FallOut_model
  F1_score <- CM_performance_metrics$F1_score
  metrics_table <- data.frame(
   Accuracy = unlist(Accuracy_model),
   Precision = unlist(Precision_model),
   Sensitivity = unlist(Sensitivity_model),
   Specificity = unlist(Specificity_model),
   FallOut = unlist(FallOut_model),
   F1_Score = unlist(F1_score)
  )
  metrics_table <- round(metrics_table, 4)
  return(metrics_table)
 }
 ###############################################################################
data <- Data$prepared_data.csv
splitted_data <- split_data(data = data, seed = 123, split_ratio = 0.7)
Logit_Full <- glm(y ~ ., data = splitted_data$train_data, family = binomial(link = "logit"))
Logit_SW_B_AIC <- StepWiseRegression(full_model = Logit_Full, train_data = splitted_data$train_data,
                      link_function = "logit", criterion = "AIC",
                      direction = "backward", trace = FALSE)
Logit_SW_B_BIC <- StepWiseRegression(full_model = Logit_Full, train_data = splitted_data$train_data,
                      link_function = "logit", criterion = "BIC",
                      direction = "backward", trace = FALSE)
Logit_SW_F_AIC <- StepWiseRegression(full_model = Logit_Full, train_data = splitted_data$train_data,
                      link_function = "logit", criterion = "AIC",
                      direction = "forward", trace = FALSE)
Logit_SW_F_BIC <- StepWiseRegression(full_model = Logit_Full, train_data = splitted_data$train_data,
                      link_function = "logit", criterion = "BIC",
                      direction = "forward", trace = FALSE)
LogitModels <- list(
 Logit_Full,  Logit_SW_B_AIC, Logit_SW_B_BIC, Logit_SW_F_AIC, Logit_SW_F_BIC
)
GF_Tables_Logit <- list()
for (model in LogitModels) {

 table <- GoFtable(model)
 GF_Tables_Logit[[length(GF_Tables_Logit) + 1]] <- table
}
table_logit <- kable(GF_Tables_Logit)
```

```
Probit_Full <- glm(y ~ ., data = splitted_data$train_data, family = binomial(link = "probit"))
Probit_SW_B_AIC <- StepWiseRegression(full_model = Probit_Full, train_data = splitted_data$train_data,
                    link_function = "probit", criterion = "AIC",
                    direction = "backward", trace = FALSE)
Probit_SW_B_BIC <- StepWiseRegression(full_model = Probit_Full, train_data = splitted_data$train_data,
                    link_function = "probit", criterion = "BIC",
                    direction = "backward", trace = FALSE)
Probit_SW_F_AIC <- StepWiseRegression(full_model = Probit_Full, train_data = splitted_data$train_data,
                    link_function = "probit", criterion = "AIC",
                    direction = "forward", trace = FALSE)
Probit_SW_F_BIC <- StepWiseRegression(full_model = Probit_Full, train_data = splitted_data$train_data,
                    link_function = "probit", criterion = "BIC",
                    direction = "forward", trace = FALSE)
ProbitModels <- list(
 Probit_Full, Probit_SW_B_AIC, Probit_SW_B_BIC, Probit_SW_F_AIC, Probit_SW_F_BIC
)
GF_Tables_Probit <- list()
for (model in ProbitModels) {
 table <- GoFtable(model)
 GF_Tables_Probit[[length(GF_Tables_Probit) + 1]] <- table
}
table_probit <- kable(GF_Tables_Probit)
################################################################################
## Definovanie prázdnych zoznam pre ukladanie údajov o modeloch
################################################################################
ChoosenModels <- list(LogitModels[2], ProbitModels[2])
PredictedClasses <- list()
Probabilities <- list()
ConfusionMatrix <- list()
RocCurves <- list()
AucScore <- list()
RocCurvesGraphically <- list()
################################################################################################################################
#############
# Pre lepšiu orientáciu v kóde:

# Index 1 odkazuje na najlepší vybraný Logit model
# Index 2 odkazuje na najlepší vybraný Probit model
# výpočet pravdepodobnosti pre najlepší logit a probit model
for (model in ChoosenModels) {
 Probabilities[[length(Probabilities) + 1]] <- predict(model, newdata = splitted_data$test_data, type = "response")
}
# Priradiť predikované pravdepodobnosti do príslušných tried a spracovanie Roc krivky pre oba modely
for (probability in Probabilities) {
 predicted_classes <- ifelse(unlist(probability) > 0.5, "yes", "no")
 factorize <- as.factor(predicted_classes)
 PredictedClasses[[length(PredictedClasses) + 1]] <- factorize
 RocCurves[[length(RocCurves) + 1]] <- roc(splitted_data$test_data$y, unlist(probability))
}
```

```
# Vytvorenie Matíc zámen pre oba modely
for (classes in PredictedClasses) {
  ConfusionMatrix[[length(ConfusionMatrix) + 1]] <- confusionMatrix(data = classes, reference = splitted_data$test_data$y,  positive = "yes")
}
# Vizualizácia Roc krivky a výpočet AUC hodnôt
for (roc in RocCurves) {
  RocCurvesGraphically[[length(RocCurvesGraphically) + 1]] <- ROCcurve(roc)
  AucScore[[length(AucScore) + 1]] <- auc(roc)
}
CM_logit <- matrix(c(11237, 1186, 214, 320), nrow = 2, byrow = TRUE)
CM_probit <- matrix(c(11273, 1235, 178, 271), nrow = 2, byrow = TRUE)
CM <- list(CM_logit, CM_probit)
CM_Properties <- Get_CM_properties(CM)
CM_performence <- CM_performance_metrics(CM_Properties)
CM_metrics_table <- CM_performence_metrics_table(CM_performence)
################################################################################
## Definovanie prázdnych zoznam pre ukladanie údajov o modeloch
################################################################################
ChoosenModels_train <- list(LogitModels[2], ProbitModels[2])
PredictedClasses_train <- list()
Probabilities_train <- list()
ConfusionMatrix_train <- list()
RocCurves_train <- list()
AucScore_train <- list()
RocCurvesGraphically_train <- list()
################################################################################################################################
#############
# Pre lepšiu orientáciu v kóde:
# Index 1 odkazuje na najlepší vybraný Logit model
# Index 2 odkazuje na najlepší vybraný Probit model
# výpočet pravdepodobnosti pre najlepší logit a probit model
for (model in ChoosenModels_train) {
  Probabilities_train[[length(Probabilities_train) + 1]] <- predict(model, newdata = splitted_data$train_data, type = "response")
}
# Priradiť predikované pravdepodobnosti do príslušných tried a spracovanie Roc krivky pre oba modely
for (probability in Probabilities_train) {
  predicted_classes <- ifelse(unlist(probability) > 0.5, "yes", "no")
  factorize <- as.factor(predicted_classes)
  PredictedClasses_train[[length(PredictedClasses_train) + 1]] <- factorize
  RocCurves_train[[length(RocCurves_train) + 1]] <- roc(splitted_data$train_data$y, unlist(probability))
}
# Vytvorenie Matíc zámen pre oba modely
for (classes in PredictedClasses_train) {
  ConfusionMatrix_train[[length(ConfusionMatrix_train) + 1]] <- confusionMatrix(data = classes, reference = splitted_data$train_data$y,  positive
= "yes")
}
# Vizualizácia Roc krivky a výpočet AUC hodnôt
for (roc in RocCurves_train) {
  RocCurvesGraphically_train[[length(RocCurvesGraphically_train) + 1]] <- ROCcurve(roc)
  AucScore_train[[length(AucScore_train) + 1]] <- auc(roc)
}
```

```
CM_logit_train <- matrix(c(26209, 2747, 512, 768), nrow = 2, byrow = TRUE)

CM_probit_train <- matrix(c(26310, 2871, 411,  644), nrow = 2, byrow = TRUE)

CM_train <- list(CM_logit_train, CM_probit_train)

CM_Properties_train <- Get_CM_properties(CM_train)

CM_performence_train <- CM_performance_metrics(CM_Properties_train)

CM_metrics_table_train <- CM_performence_metrics_table(CM_performence_train)

#base_path

#tab_logit <- stargazer(LogitModels[1],LogitModels[2], LogitModels[3], align=TRUE, type = "html", out=paste0(base_path,
"LogitSummarytable.htm"))

#tab_probit <- stargazer(ProbitModels[1], ProbitModels[2], ProbitModels[3],align=TRUE, type = "html", out=paste0(base_path,
"ProbitSummarytable.htm"))

LRT_logit <- list()

LRT_probit<- list()

LNull <- glm(y ~ 1, data = splitted_data$train_data,

        family = binomial( link = "logit"))

PNull <- glm(y ~ 1, data = splitted_data$train_data,

        family = binomial( link = "probit"))

for (model in LogitModels) {

  LRT_logit[[length(LRT_logit) + 1]] <- lrtest(LNull, model)

}

for (model in ProbitModels) {

  LRT_probit[[length(LRT_probit) + 1]] <- lrtest(PNull, model)

}

# Výber modelu pre výpočet Marginálnych efektov

logit_model <- ChoosenModels[[1]][[1]]

probit_model <- ChoosenModels[[2]][[1]]

formula_used_l <- formula(logit_model)

formula_used_p <- formula(probit_model)

# MARGINÁLNE EFEKTY

ME_logit <- logitmfx(formula = formula_used_l, data = splitted_data$train_data, atmean = TRUE, robust = FALSE, clustervar1 = NULL, clustervar2
= NULL, start = NULL, control = list())

ME_probit <- probitmfx(formula = formula_used_p, data = splitted_data$train_data, atmean = TRUE, robust = FALSE, clustervar1 = NULL,
clustervar2 = NULL, start = NULL, control = list())

OR <- logitor(formula = formula_used_l, data = splitted_data$train_data, robust = FALSE, clustervar1 = NULL, clustervar2 = NULL, start = NULL,
control = list())


ME_tab_l <- round(as.data.frame(ME_logit$mfxest),8)

ME_tab_p <- round(as.data.frame(ME_probit$mfxest),8)

OR_tab_l <- round(as.data.frame(OR$oddsratio),5)

write.csv(ME_tab_l, file = paste0(base_path, "MeLogitME8.csv"), row.names = TRUE)

write.csv(ME_tab_p, file = paste0(base_path, "MeProbitME8.csv"), row.names = TRUE)

write.csv(OR_tab_l, file = paste0(base_path, "LogitOR5.csv"), row.names = TRUE)

# Demonštrácia výpočtu pomeru šancí pre špecifické pozorovania

used_data <- splitted_data$train_data

observation_1 <- data.frame(job = "admin", marital = "single", education = "tertiary", default= "no",balance = 1360, housing = "no", loan = "no",
duration= 600, campaign = 2, pdays = 90, previous=1)

observation_2  <- data.frame(job = "admin", marital = "single", education = "secondary", default= "no",balance = 1360, housing = "no", loan =
"no", duration= 1200, campaign = 0, pdays = 90, previous=5)
```

```
# Predikcia pravdepodobnosti
prob_obs_1 <- predict(model, newdata = observation_1, type = "response")
prob_obs_2 <- predict(model, newdata = observation_2, type = "response")
prob_value_1 <- prob_obs_1[[1]][1]
prob_value_2 <- prob_obs_2[[1]][1]
# Výpočet šance
odds_1 = prob_value_1 / (1 - prob_value_1)
odds_2 = prob_value_2 / (1 - prob_value_2)
# Pomer šancí
odds_ratio = odds_2 / odds_1
```