Multinomna logistička regresija u kreditnom skoringu

Andabaka, Zana

Sveučilište J.J. Strossmayera u Osijeku
Odjel za matematiku

Zana Andabaka

Multinomna logistička regresija u kreditnom skoringu

Diplomski rad

Osijek, 2017.

Sveučilište J.J. Strossmayera u Osijeku
Odjel za matematiku

Zana Andabaka

Multinomna logistička regresija u kreditnom skoringu

Diplomski rad

Mentor:
prof. dr. sc. Nataša Šarlija

Komentor:
prof. dr. sc. Mirta Benšić

Osijek, 2017.

Sadržaj

1 Uvod

Primarni cilj ovoga rada je kreirati model multinomne logističke regresije koji će najbolje predviđati kreditnu rizičnost klijenata jedne banke. Koristimo bazu podataka klijenata u kojoj su klijenti klasificirani kao *dobri*, *srednji* ili *loši*. *Dobre* klijente smo definirali kao klijente koji niti jednom nisu kasnili pri otplati kredita, *srednji* nisu nikad kasnili više od 31 dan, a *loši* su barem jednom zakasnili više od 31 dan.

U kreditnoj praksi za kreiranje modela kreditnog rizika najčešće se koristi binomna logistička regresija pri čemu se za kreiranje modela uzimaju podaci samo o *dobrim* i *lošim* klijentima, dok se *srednji* izostavljaju. Razlog tome je što izostavljajući *srednje* klijente iz analize, divergencija između definicija *dobrih* i *loših* klijenata je veća, a na temelju rezultata modela za svakog klijenta dobiva se jednostavan odgovor - vjerojatnost da klijent bude *loš*.

No, uvodeći *srednje* klijente u model kreditnog rizika dobiva se kompletna slika o klijentima pojedine banke. Razlog zbog kojeg ih je korisno uključiti u model je i taj što se u bazi podataka nalaze i klijenti čiji su krediti još aktivni, te ukoliko je skup podataka mali, svaki raspoloživi podatak može biti značajan.

U drugom poglavlju definirani su eksponencijalna familija distribucija, generalizirani linearni modeli te binomna logistička regresija.

Treće poglavlje je proširenje binomne logističke regresije na multinomnu gdje ovisna varijabla može poprimiti jednu od J kategorija, $J > 2$. Napravljeno je proširenje generaliziranih linernih modela, definirana multinomna distribucija, te objašnjena razlika između nominalne i ordinalne logističke regresije. U konačnici definiran je model multinomne logističke regresije (nominalne i ordinalne), opisan proces procjene parametara koja se vrši metodom maksimalne vjerodostojnosti, te neki načini ocjene kvalitete modela.

Četvrto poglavlje je primjena trećeg poglavlja na problem kreditnog skoringa. Definirana je baza podataka na kojoj je provedena analiza. Za svaku neovisnu varijablu je napravljena kratka analiza kako bi detektirali koje neovisne varijable najbolje razlikuju kategorije ovisne varijable. Na temelju dobivenih rezultata kreiran je model multinomne logističke regresije, te je na validacijskom uzorku, koji se sastoji od podataka koji nisu korišteni za kreiranje modela, izvršena provjera točnosti modela. Također, kreiran je i binomni model samo s *dobrim* i *lošim* klijentima, te su uspoređeni rezultati oba modela.

Analiza podataka i kreiranje modela provedeni su u programskom jeziku R.

2 Logistička regresija

Ovisna varijabla kod logističke regresije je diskretna (kategorička) i dihotomna, tj. može poprimiti jednu od dvije vrijednosti od kojih su obe nekakve kategorije. Neovisne varijable mogu biti i kategoričke i metričke.

Model logističke regresije pripada familiji generaliziranih linearnih modela. Stoga ćemo se prije definicije modela upoznati s pojmom *eksponencijane familije* koja je važna za razumjevanje *generaliziranog linearnog modela* kojeg ćemo u nastavku također definirati.

2.1 Eksponencijalna familija distribucija

Definicija 2.1.1 *Neka je Y slučajna varijabla čija vjerojatnosna distribucija ovisi o parametru θ. Distribucija slučajne varijable Y pripada* eksponencijalnoj familiji *ako njena funkcija gustoće može biti zapisana u sljedećem obliku:*

$$f(y;\theta) = s(y)t(\theta)e^{a(y)b(\theta)}, \tag{2.1}$$

gdje su a, b, s i t poznate funkcije. (Barnett, Dobson, 2008.)

Definiramo li $s(y)$ kao $s(y) = e^{d(y)}$, a funkciju $t(\theta)$ kao $t(\theta) = e^{c(\theta)}$, prethodna funkcija poprima sljedeći oblik,

$$f(y;\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]. \tag{2.2}$$

Ukoliko vrijedi $a(y) = y$, kažemo da distribucija ima *kanonsku (standardnu) formu*, dok član $b(\theta)$ nazivamo i *prirodni parametar* distribucije. Ako postoje i drugi parametri, dodatni uz θ koji je od interesa, oni se smatraju poznatima i ne procjenjujemo ih.
Neke od distribucija koje pripadaju eksponencijalnoj familiji su Bernoullijeva i binomna.

Bernoulijeva distribucija ima funkciju gustoće oblika

$$f(y) = \pi^y(1-\pi)^{1-y}, \quad y \in \{0,1\}, \tag{2.3}$$

gdje je $\pi = P(y = 1)$ vjerojatnost uspjeha. Neovisnim ponavljanjem istog Bernoullijevog pokusa $n \in \mathbf{N}$ puta, i pri tome bilježeći realizacije uspjeha, kreira se binomna slučajna varijabla Y sa slikom $\mathcal{R}(Y) = \{0, 1, \ldots, n\}$ koja ima sljedeću funkciju gustoće:

$$h(y) = \binom{n}{y}\pi^y(1-\pi)^{n-y}, \quad y \leq n. \tag{2.4}$$

Pokažimo da je funkcija $h(y)$ iz eksponencijalne familije:

$$\begin{aligned}\binom{n}{y}\pi^y(1-\pi)^{n-y} &= \exp\left[\ln\binom{n}{y}+y\ln\pi+(n-y)\ln(1-\pi)\right]\\ &= \exp\left[\ln\frac{n!}{y!(n-y)!}+y(\ln\pi-\ln(1-\pi))+n\ln(1-\pi)\right]\\ &= \exp\left[y\ln\frac{\pi}{1-\pi}+n\ln(1-\pi)+\ln\frac{n!}{y!(n-y)!}\right]. \end{aligned} \tag{2.5}$$

Usporedimo li dobiveni izraz s jednadžbom (2.2), članove izraza (2.5) možemo pridružiti odgovarajućim funkcijama iz (2.2) na sljedeći način:

$$\begin{aligned} a(y) = y \qquad & b(\pi) = ln\frac{\pi}{1-\pi}\\ c(\pi) = n\ln(1-\pi) \qquad & d(y) = \ln\frac{n!}{y!(n-y)!}. \end{aligned} \tag{2.6}$$

Dakle, binomna distribucija pripada eksponencijalnoj familiji distribucija. Primjetimo ukoliko u jednadžbi (2.5) vrijedi $n = 1$ i $y \in \{0,1\}$, jednadžba se svodi na Bernoullijevu distribuciju. U tom slučaju vrijedi $d(y) = 0$ i $c(\pi) = \ln(1-\pi)$, dok funkcije a i b ostaju iste kao u (2.6). Slijedi da je i Bernoullijeva distribucija iz eksponencijalne familije distribucija.

2.2 Generalizirani linearni model

Generalizirani linearni model definiran je u terminima skupa međusobno neovisnih slučajnih varijabli $Y_1, Y_2, \ldots, Y_N$ čije distribucije ne moraju biti jednake, ali distribucija svake od varijabli $Y_1, \ldots, Y_N$ pripada eksponencijalnoj familiji i vrijede sljedeća svojstva (Barnett, Dobson, 2008.):

1) Distribucija od Y_i ima kanonsku formu i ovisi samo o jednom parametru θ_i (θ_i-evi ne moraju svi nužno biti jednaki), tada za sve $i = 1, \ldots, N$ vrijedi:

$$f_{y_i}(y;\theta_i) = \exp[yb_i(\theta_i) + c_i(\theta_i) + d_i(y)]$$

2) Distribucije svih Y_i-eva su istog tipa, stoga funkcije b, c i d ne ovise o indeksu i.
Tada zajednička funkcija gustoće varijabli $Y_1, \ldots, Y_N$ ima sljedeći oblik:

$$\begin{aligned} f(y_1,\ldots,y_N;\theta_1,\ldots,\theta_N) &= \prod_{i=1}^{N}\exp[y_ib(\theta_i)+c(\theta_i)+d(y_i)]\\ &= \exp\Big[\sum_{i=1}^{N}y_ib(\theta_i)+\sum_{i=1}^{N}c(\theta_i)+\sum_{i=1}^{N}d(y_i)\Big]. \end{aligned} \tag{2.7}$$

Obzirom da skup parametara $\{\theta_1, \ldots, \theta_N\}$ inače nije od direktnog interesa za model zamjenit ćemo ga drugim skupom parametara $\{\beta_1, \ldots, \beta_K\}$, koji je od interesa za model, gdje je K broj

neovisnih varijabli.
Definirajmo matricu dizajna $\mathbf{X}$, tj. matricu u kojoj su prikazane mjerene vrijednosti neovisnih varijabli $X_1, X_2, \ldots, X_K$.

Neka je vektor $\mathbf{X_i}$ K-dimenzionalni vektor mjerenih vrijednosti K neovisnih varijabli pri i-tom promatranju, $\mathbf{X_i} = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{iK} \end{bmatrix}$.

Vektor $\mathbf{X_i^T}$ je i-ti redak matrice dizajna $\mathbf{X}$,

$$\mathbf{X} = \begin{bmatrix} \mathbf{X_1}^T \\ \vdots \\ \mathbf{X_N}^T \end{bmatrix} = \begin{bmatrix} x_{11} & \ldots & x_{1K} \\ \vdots & & \vdots \\ x_{N1} & \ldots & x_{NK} \end{bmatrix}.$$

$\boldsymbol{\beta}$ je $K-$dimenzionalni vektor parametara, $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}$.

Sada možemo definirati tzv. *linearni prediktor* $\eta_i = \mathbf{X}_i^T \boldsymbol{\beta}$.

Uvjetno očekivanje $E[Y_i|X_i] = \mu_i$ povezano je s linearnim prediktorom η_i na sljedeći način (Fahrmeir, Tutz, 2001.):

$$\mu_i = h(\eta_i), \text{ odnosno } \eta_i = g(\mu_i).$$

Pretpostavka je da je h injekcija i dovoljno glatka funkcija. Funkcija g definirana je kao inverz funkcije h i nazivamo je *link funkcija*. Ona je monotona i derivabilna i opisuje kako očekivanje ovisi o linearnom prediktoru. Link funkciju obično zapisujemo u sljedećem obliku:

$$g(\mu_i) = \mathbf{X}_i^T \boldsymbol{\beta}.$$

Dakle, očekivanje $E[Y_i|X_i] = \mu_i$ je povezano s vektorom parametara $\boldsymbol{\beta}$ i modelirano na temelju skupa neovisnih varijabli X_i, $i = 1, \ldots, K$.

U konačnici, možemo reći da su generalizirani linearni modeli definirani kroz tri komponente (Barnett, Dobson, 2008.):

1. ovisne varijable $Y_1, \ldots, Y_N$ koje dijele isti tip vjerojatnosne distribucije iz ekponencijalne familije.

2. Skup paramatera $\boldsymbol{\beta}$ i neovisne varijable čije mjerene vrijednosti kreiraju matricu dizajna $\mathbf{X}$.

3. Monotona link funkcija g.

2.3 Model logističke regresije

Model binomne logističke regresije koristi se za procjenu vjerojatnosti realizacije jedne od kategorija binomne ovisne varijable na temelju poznatih vrijednosti jedne ili više neovisnih varijabli. Link funkcija kojom je određen model logističke regresije je sljedeća:

$$b(\pi_i) = \ln \frac{\pi_i}{1-\pi_i} = \mathbf{X}_i^T \boldsymbol{\beta}, \quad i = 1, \ldots, N \tag{2.8}$$

gdje je π_i vjerojatnost realizacije uspjeha pri $i-$tom promatranju, $\mathbf{X}_i$ je vektor vrijednosti neovisnih varijabli, a $\boldsymbol{\beta}$ vektor parametara kao u prethodnom poglavlju. Navedena funkcija naziva se i *logit funkcija*. Primjetimo da je logit funkcija prirodni parametar binomne distribucije (vidi 2.6). Izraz pod logaritmom nazivamo i *sklonost* (enlg. *odds*). U ovom slučaju sklonost možemo definirati kao omjer vjerojatnosti realizacije uspjeha i vjerojatnosti realizacije neuspjeha.
Izrazimo li iz jednadžbe (2.8) vjerojatnost π_i, dobivamo jednadžbu kojom je definiran model logističke regresije:

$$\pi_i = \frac{e^{\mathbf{X}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{X}_i^T \boldsymbol{\beta}}}, \quad i = 1, \ldots, N. \tag{2.9}$$

Logit funkcija je monotona po y_i i poprima vrijednosti iz čitavog slupa $\mathbf{R}$, a vjerojatnost iz jednakosti (2.9) poprima vrijednosti iz intervala $\langle 0, 1 \rangle$.
Procjena parametara provodi se metodom maksimalne vjerodostojnosti (vidi [7]).

3 Multinomna logistička regresija

U prethodnom poglavlju naveli smo jedan način rješavanja problema svrstavanja jedinice promatranja u jednu od dvije kategorije. Ukoliko postoji više od dvije kategorije ovisne varijable možemo koristiti *multinomnu logističku regresiju*.
U ovisnosti o kakvim se kategorijama radi, odnosno kakva je slika ovisne varijable, razlikujemo dvije vrste multinomne logističke regresije: logistička regresija s *nominalnom* ovisnom varijablom i logistička regresija s *ordinalnom* ovisnom varijablom.
Nominalna varijabla ima sliku u kojoj ne postoji skala ili redosljed po kojoj je neka kategorija više ili manje vrijedna od druge, kao što je npr. plava, zelena, crvena, žuta boja, ili, odgovor na pitanje u nekom upitniku: da, ne, ne znam, nema odgovora.
Ordinalna varijabla je ona koja poprima vrijednosti iz skupa u kojem je prisutna skala kvalitete ili prirodni slijed, kao npr. *dobar*, *srednji* ili *loš* klijent.
Bazna distribucija za multinomnu logističku regresiju je multinomna distribucija.

3.1 Multinomna distribucija

Neka je Y slučajna varijabla s $J \geq 2$ kategorija. Neka su $\pi_j = P(Y = j)$, $j = 1, \ldots, J$, odgovarajuće vjerojatnosti tako da vrijedi $\sum_{j=1}^{J} \pi_j = 1$. Promotrimo N neovisnih realizacija slučajne varijable Y tako što ćemo za rezulat bilježiti broj realizacija svake od kategorija. Broj realizacija kategorije 1 označit ćemo s y_1, broj realizacija kategorije 2 s y_2, i tako dalje. Na taj način dolazimo do vektora $\mathbf{y} = [y_1 \;\; y_2 \;\; \ldots \;\; y_J]^T$, gdje je $y_j \in \{0, 1, \ldots, N\}$, tako da vrijedi $\sum_{j=1}^{J} y_j = N$. (Agresti, 2002.)

Sukladno s navedenim oznakama, definirajmo multinomnu distribuciju.

Definicija 3.1.1 *Vektor* $\mathbf{Y}$ *ima* multinomnu distribuciju *s parametrima* $N \in \mathbf{N}$ *i* $\pi \in \mathbf{R}^J$ *ako ima sljedeću funkciju gustoće:*

$$f(\mathbf{y}|N) = \frac{N!}{y_1! y_2! \ldots y_J!} \pi_1^{y_1} \pi_2^{y_2} \ldots \pi_J^{y_J}, \tag{3.1}$$

i označavamo s $\mathbf{M}(N, \pi_1, \ldots, \pi_J)$.

Primjetimo da ukoliko $J = 2$ distribucija se svodi na binomnu distribuciju (vidi jednadžbu 2.4).

3.2 Multivarijatno proširenje generaliziranih linearnih modela

Multinomni logistički modeli su specijalni slučaj multivarijatnih generaliziranih linearnih modela. Slično kao kod univarijatnih, multivarijatni generalizirani modeli bazirani su na distribucijskoj i strukturnoj pretpostavci.

Neka su X_1, X_2, $\dots X_K$ neovisne varijable. Ovisna varijabla Y_i je $J-$dimenzionalni vektor s očekivanjem $\mu_i = E(Y_i|\mathbf{X}_i)$, gdje je ${\mathbf{X}_i}^T$ $i-$ti redak matrice dizajna kao u poglavlju (2.2).
Distribucijska pretpostavka je da su $\mathbf{X}_i$ i Y_i (uvjetno) neovisni i Y_i ima distribuciju koja pripada eksponencijalnoj familiji, tj. ima sljedeću formu (Barnett, Dobson, 2008.):

$$f(y_i;\theta_i) = \exp[y_i b(\theta_i) + c(\theta_i) + d(y_i)].$$

Odnosno, zajednička funkcija gustoće je oblika:

$$f(y_1,\dots,y_N;\theta_1,\dots,\theta_N) = \exp\Big[\sum_{i=1}^{N} y_i b(\theta_i) + \sum_{i=1}^{N} c(\theta_i) + \sum_{i=1}^{N} d(y_i)\Big]. \tag{3.2}$$

Strukturna pretpostavka je da je očekivanje μ_i određeno linearnim prediktorom $\eta_i = \mathbf{X}_i^T\boldsymbol{\beta}_i$ u sljedećem obliku:

$$\mu_i = h(\eta_i),$$

odnosno definirana je *link funkcija* kao inverz funkcije h, $g(\mu_i) = \eta_i$, gdje je

- $\mathbf{X}$ matrica dizajna koja se sastoji od vektora $\mathbf{X}_1, \dots, \mathbf{X}_N$
- $\boldsymbol{\beta}$ matrica nepoznatih parametara i sastoji se od vektora $\boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_K$

Vjerojatnosti možemo zapisati kao $\pi_{ij} = P(Y_i = j)$.

3.3 Model multinomne logističke regresije

3.3.1 Nominalna logistička regresija

Promotrimo varijablu Y koja ima multinomnu vjerojatnosnu distribuciju s $J \geq 2$ kategorija. Označimo s N broj promatranja slučajne varijable Y. Ako je svako od N promatranja slučajne varijable neovisno, tada svaki Y_i, $i = 1,\dots,N$, ima multinomnu distribuciju.
Obzirom da se pri svakom promatranju realizira jedna od J mogućih vrijednosti ovisne varijable Y, neka je $\mathbf{y}$ matrica realizacija s N redaka i $J-1$ stupaca.

$$\mathbf{y} = \begin{bmatrix} y_{11} & \cdots & y_{1(J-1)} \\ \vdots & & \vdots \\ y_{N1} & \cdots & y_{N(J-1)} \end{bmatrix}$$

Svaki y_{ij} predstavlja realizaciju kategorije j pri i-tom promatranju. Ukoliko se kategorija j realizirala, y_{ij} će poprimiti vrijednost 1, a 0 inače, i vrijedi $\sum_{j=1}^{J} y_{ij} = 1$.
Neka je π matrica dimenzija $N \times (J-1)$, gdje je svaki π_{ij} vjerojatnost realizacije $j-$te vrijednosti u $i-$tom promatranju ovisne varijable, tj. $\pi_{ij} = P(Y_i = j)$ i vrijedi $\sum_{j=1}^{J} \pi_{ij} = 1, \forall i \in \{1,\dots,N\}$.

$$\pi = \begin{bmatrix} \pi_{11} & \dots & \pi_{1(J-1)} \\ \vdots & & \vdots \\ \pi_{N1} & \dots & \pi_{N(J-1)} \end{bmatrix}$$

Matrica dizajna neovisnih varijabli, $\mathbf{X}$, je dimenzije $N \times (K+1)$, gdje je K broj neovisnih varijabli. Prvi stupac matrice sadrži samo jedinice, $x_{i0} = 1$, $\forall i \in \{1, \dots, N\}$, jer se veže uz slobodni član (*engl. intercept*).

$$\mathbf{X} = \begin{bmatrix} x_{10} & \dots & x_{1K} \\ \vdots & & \vdots \\ x_{N0} & \dots & x_{NK} \end{bmatrix}$$

Matrica parametara $\boldsymbol{\beta}$ je dimenzija $(K+1) \times (J-1)$. Parametar β_{kj} veže se uz $k-$tu neovisnu varijablu i $j-$tu vrijednost ovisne varijable, $\forall k \in \{0, \dots, K\}$ i $\forall j \in \{1, \dots, J-1\}$.

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{01} & \dots & \beta_{0(J-1)} \\ \vdots & & \vdots \\ \beta_{K1} & \dots & \beta_{K(J-1)} \end{bmatrix}$$

Slično kao kod binomne logističke regresije, kod multinomne linearnu komponentu izjednačavamo s logaritmom od sklonosti. Kod nominalne regresije za baznu kategoriju možemo uzeti bilo koju od J kategorija.
Uzmimo kategoriju J kao baznu kategoriju. Logaritam od sklonosti prvih $J-1$ kategorija bi imao sljedeći oblik:

$$ln\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = ln\left(\frac{\pi_{ij}}{1 - \sum_{j=1}^{J-1} \pi_{ij}}\right) = \sum_{k=0}^{K} x_{ik}\beta_{kj}. \quad i = 1, \dots, N. \tag{3.3}$$

Rješavanjem po π_{ij} dobivamo

$$\pi_{ij} = \frac{e^{\sum_{k=0}^{K} x_{ik}\beta_{kj}}}{1 + \sum_{j=1}^{J-1} e^{\sum_{k=0}^{K} x_{ik}\beta_{kj}}}, \qquad j < J \tag{3.4}$$

odnosno, za baznu kategoriju vrijedi

$$\pi_{iJ} = 1 - \sum_{j=1}^{J-1} \pi_{ij} = \frac{1}{1 + \sum_{j=1}^{J-1} e^{\sum_{k=0}^{K} x_{ik}\beta_{kj}}}. \tag{3.5}$$

Za ordinalnu logističku regresiju vrijede neke modifikacije.

3.3.2 Ordinalna logistička regresija

Ukoliko postoji više od dvije kategorije ovisne varijable koje međusobno čine nekakav uređeni slijed, možemo koristiti ordinalnu logističku regresiju.
Postoji više modela ordinalne logističke regresije, a onaj koji se u praksi najviše koristi je *kumulativni* model logističke regresije, odnosno *proportional odds* model, te ćemo njega koristiti. Ostali modeli su modifikacije navedenog modela (vidi Barnett, Dobson, 2008, str. 157).

Neka je Y_i, $i = 1,\dots,N$, ovisna ordinalna varijabla koja pri jednom promatranju poprima jednu od J kategorija. Odgovarajuće vjerojatnosti realizacija svake od kategorija pri $i-tom$ promatranju su $\pi_{i1} = P(Y_i = 1), \dots, \pi_{iJ} = P(Y_i = J)$. Distribucija slučajne varijable Y_i je multinomna s parametrom $\pi_i = (\pi_{i1},\dots,\pi_{iJ})$.
Za ordinalnu logističku regresiju ključne su *kumulativne vjerojatnosti* koje definiramo na sljedeći način:

$$P(Y_i \le j) = \pi_{i1} + \pi_{i2} + \cdots + \pi_{ij}, \tag{3.6}$$

gdje je $j = 1,\dots,J$ realizirana kategorija ovisne varijable. Kod ovog modela zanima nas sklonost za svaku pojedinu kategoriju ovisne varijable, ali na način da gledamo omjer sume vjerojatnosti realizacije manjih kategorija i sume vjerojatnosti realizacije kategorija većih od njih.

Prema (3.6) svaku sklonost možemo zapisati na sljedeći način:

$$\frac{P(Y_i \le j)}{P(Y_i > j)} = \frac{P(Y_i \le j)}{1 - P(Y_i \le j)} = \frac{\pi_{i1} + \pi_{i2} + \cdots + \pi_{ij}}{\pi_{i(j+1)} + \cdots + \pi_{iJ}}, \tag{3.7}$$

iz čega možemo izraziti logaritam sklonosti dvije kumulativne vjerojatnosti:

$$ln\left(\frac{P(Y_i \le j)}{1 - P(Y_i \le j)}\right) = ln\left(\frac{\pi_{i1} + \pi_{i2} + \cdots + \pi_{ij}}{\pi_{i(j+1)} + \cdots + \pi_{iJ}}\right). \tag{3.8}$$

Time mjerimo kolika je šansa realizacije kategorije koja je $\le j$ u odnosu na realizaciju kategorije koja je $> j$.

Uključimo sada i neovisne varijable u model čime dobivamo konačni *kumulativni* model logističke regresije:

$$ln\left(\frac{P(Y_i \le j)}{1 - P(Y_i \le j)}\right) = \beta_{0j} + \beta_{1j}x_{i1} + \beta_{2j}x_{i2} + \cdots + \beta_{Kj}x_{iK}, \quad j \in \{1,\dots,J-1\}, \tag{3.9}$$

gdje je K broj neovisnih varijabli. Na taj način kreiramo $J-1$ funkciju.

Važna pretpostavka ovog modela je da koeficijenti β_{1j}, β_{2j}, $\dots$, β_{Kj} ne ovise o kategoriji j, nego su jednaki za svaki $j = 1,\dots,J$. Dok koeficijenti β_{0j} ovise o kategoriji j i variraju za svaku od funkcija. Koeficijenti β_{0j} su kao slobodni član u linearnom regresijskom modelu. Iz tog razloga prethodnu jednakost možemo zapisati na sljedeći način gdje ćemo slobodni član β_{0j} zamjeniti oznakom α_j,

$$ln\left(\frac{P(Y_i \leq j)}{1 - P(Y_i \leq j)}\right) = \alpha_j + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} \tag{3.10}$$

Dakle, vektor α je oblika $(\alpha_1, \alpha_2, \ldots, \alpha_{J-1})$.

Iz jednadžbe (3.10) možemo izraziti kumulativne vjerojatnosti u ovisnosti o poznatim vrijednostima neovisnih varijabli:

$$P(Y_i \leq j) = \frac{e^{\alpha_j + \beta_1 x_{i1} + \cdots + \beta_K x_{iK}}}{1 + e^{\alpha_j + \beta_1 x_{i1} + \cdots + \beta_K x_{iK}}}. \tag{3.11}$$

Svaka se pojedina vjerojatnost može izraziti pomoću kumulativnih iz (3.11):

$$P(Y_i = j) = P(Y_i \leq j) - P(Y_i \leq j - 1). \tag{3.12}$$

Pokažimo da za slobodne članove vrijedi $\alpha_j < \alpha_{j+1}$, $\forall j \in \{1, \ldots, J-1\}$.

Iz nejednakosti $P(Y_i \leq j) < P(Y_i \leq j+1)$ i jednadžbe (3.11) slijedi:

$$\frac{e^{\alpha_j + \beta_1 x_{i1} + \cdots + \beta_K x_{iK}}}{1 + e^{\alpha_j + \beta_1 x_{i1} + \cdots + \beta_K x_{iK}}} < \frac{e^{\alpha_{j+1} + \beta_1 x_{i1} + \cdots + \beta_K x_{iK}}}{1 + e^{\alpha_{j+1} + \beta_1 x_{i1} + \cdots + \beta_K x_{iK}}}.$$

Obzirom da član $\beta_1 x_{i1} + \cdots + \beta_K x_{iK}$ ne utječe na nejednakost možemo ga tretirati kao konstantu i izuzeti iz nejednadžbe:

$$\frac{e^{\alpha_j}}{1 + e^{\alpha_j}} < \frac{e^{\alpha_{j+1}}}{1 + e^{\alpha_{j+1}}},$$

množenjem dolazimo do sljedeće nejednakosti:

$$e^{\alpha_j}(1 + e^{\alpha_{j+1}}) < (1 + e^{\alpha_j})e^{\alpha_{j+1}},$$

$$e^{\alpha_j} < e^{\alpha_{j+1}},$$

iz čega slijedi da je $\alpha_j < \alpha_{j+1}$.

3.4 Procjena parametara

Procjenu parametara multinomnog logističkog modela možemo provesti *metodom maksimalne vjerodostojnosti* (engl. *Maximum Likelihood Estimation*) (S.Czepiel [7]). Želimo procijeniti skup parametara β_{kj} iz jednadžbe (3.4).

Neka su $Y_1, \ldots, Y_N$ međusobno nezavisne varijable koje zadovoljavaju svojstva generaliziranih linearnih modela. Ako je $f(y_1, \ldots, y_N)$ njihova zajednička funkcija gustoće, *funkciju maksimalne vjerodostojnosti* definiramo kao funkciju parametara $\boldsymbol{\beta}$ na sljedeći način:

$$\mathcal{L}(\boldsymbol{\beta}) = f(y_1, \ldots, y_N|\boldsymbol{\beta}). \tag{3.13}$$

Obzirom da su varijable međusobno nezavisne i jednako distribuirane, izraz (3.13) možemo pojednostaviti:

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{N} f(y_i|\boldsymbol{\beta}). \tag{3.14}$$

Sukladno definiciji multinomne distribucije (3.1), zajednička funkcija gustoće, odnosno funkcija maksimalne vjerodostojnosti multinomne distribucije ima sljedeći oblik:

$$f(\mathbf{y}|\boldsymbol{\beta}) = \prod_{i=1}^{N}\left[\frac{n_i!}{\prod_{j=1}^{J} y_{ij}!}\prod_{j=1}^{J}\pi_{ij}^{y_{ij}}\right], \tag{3.15}$$

gdje je N broj realizacija, a J broj kategorija ovisne varijable. Obzirom da želimo maksimizirati gornju funkciju u terminima parametra $\boldsymbol{\beta}$ koji ne ovisi o n_i niti y_{ij}, faktorijalne izraze možemo tretirati kao konstantu. Iz tog razloga dovoljno je maksimizirati "jezgru" funkcije maksimalne vjerodostojnosti,

$$f(\mathbf{y}|\boldsymbol{\beta}) \simeq \prod_{i=1}^{N}\prod_{j=1}^{J}\pi_{ij}^{y_{ij}}, \tag{3.16}$$

što možemo raspisati na sljedeći način:

$$\begin{aligned}
&\prod_{i=1}^{N}\prod_{j=1}^{J-1}\pi_{ij}^{y_{ij}}\pi_{iJ}^{n_i-\sum_{j=1}^{J-1}y_{ij}} \\
= &\prod_{i=1}^{N}\prod_{j=1}^{J-1}\pi_{ij}^{y_{ij}}\frac{\pi_{iJ}^{n_i}}{\pi_{iJ}^{\sum_{j=1}^{J-1}y_{ij}}} \\
= &\prod_{i=1}^{N}\prod_{j=1}^{J-1}\pi_{ij}^{y_{ij}}\frac{\pi_{iJ}^{n_i}}{\prod_{j=1}^{J-1}\pi_{iJ}^{y_{ij}}} \\
= &\prod_{i=1}^{N}\prod_{j=1}^{J-1}\left(\frac{\pi_{ij}}{\pi_{iJ}}\right)^{y_{ij}}\pi_{iJ}^{n_i}.
\end{aligned} \tag{3.17}$$

Uvrštavajući u prethodnu jednakost vrijednosti za π_{ij} i π_{iJ} iz (3.4) i (3.5) dobivamo:

$$\prod_{i=1}^{N}\prod_{j=1}^{J-1}(e^{\sum_{k=0}^{K}x_{ik}\beta_{kj}})^{y_{ij}}\left(\frac{1}{1+\sum_{j=1}^{J-1}e^{\sum_{k=0}^{K}x_{ik}\beta_{kj}}}\right)^{n_i}$$
$$= \prod_{i=1}^{N}\prod_{j=1}^{J-1}e^{y_{ij}\sum_{k=0}^{K}x_{ik}\beta_{kj}}\Big(1+\sum_{j=1}^{J-1}e^{\sum_{k=0}^{K}x_{ik}\beta_{kj}}\Big)^{-n_i}. \tag{3.18}$$

Definicija 3.4.1 Procjenitelj maksimalne vjerodostojnosti *parametra $\boldsymbol{\beta}$, je vrijednost spomenutog parametra koji maksimizira funkciju vjerodostojnosti,*

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbf{R}^{(K+1)\times(J-1)}}{\operatorname{argmax}} \mathcal{L}(\boldsymbol{\beta}). \tag{3.19}$$

Umjesto maksimiziranja produkta (3.14) često se koristi činjenica da je logaritam rastuća funkcija, te će logaritam funkcije vjerodostojnosti, i sama funkcija vjerodostojnosti, postići maksimum za istu vrijednost parametra. Stoga je ekvivalentno maksimizirati logaritam funkcije vjerodostojnosti,

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N} ln(f(y_i|\boldsymbol{\beta})). \tag{3.20}$$

Logaritmirajući jednakost (3.18) dobivamo logaritam funkcije vjerodostojnosti modela multinomne logističke regresije:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N}\sum_{j=1}^{J-1}\left[\Big(y_{ij}\sum_{k=0}^{K}x_{ik}\beta_{kj}\Big) - n_i\ln\Big(1+\sum_{j=1}^{J-1}e^{\sum_{k=0}^{K}x_{ik}\beta_{kj}}\Big)\right]. \tag{3.21}$$

Želimo pronaći vrijednosti za parametre β_{kj} koji će maksimizirati funkciju (3.21).
Za procijenu parametara koristit ćemo Newton-Raphsonovu metodu (Agresti, 2002, str. 143).
Definirajmo funkciju u koja odgovara derivaciji logaritma funkcije vjerodostojnosti:

$$u(\boldsymbol{\beta}) = \frac{\partial l(\beta)}{\partial\beta_{kj}} = \sum_{i=1}^{N}\left[y_{ij}x_{ik} - n_i\frac{1}{1+\sum_{j=1}^{J-1}e^{\sum_{k=0}^{K}x_{ik}\beta_{kj}}}e^{\sum_{k=0}^{K}x_{ik}\beta_{kj}}x_{ik}\right]$$
$$= \sum_{i=1}^{N}[y_{ij}x_{ik} - n_i\pi_{ij}x_{ik}], \quad j\in\{1,\ldots,J-1\}, \quad k\in\{0,\ldots,K\} \tag{3.22}$$

Dobili smo $(J-1)\times(K+1)$ jednadžbi oblika (3.22), koje želimo izjednačiti s nulom i riješiti po β_{kj}. Svako od rješenja jednadžbi (3.22), ako postoji, će biti maksimum ukoliko je matrica drugih derivacija pozitivno definitna, odnosno ako je svaki element na dijagonali matrice manji od nule. Elementi martice drugih derivacija imaju sljedeći oblik:

$$
\begin{aligned}
u''(\boldsymbol{\beta}) = \frac{\partial^2 l(\boldsymbol{\beta})}{\partial\beta_{kj}\partial\beta_{k'j'}} &= \frac{\partial}{\partial\beta_{k'j'}}\sum_{i=1}^{N}\left[y_{ij}x_{ik} - n_i\pi_{ij}x_{ik}\right] \\
&= \frac{\partial}{\partial\beta_{k'j'}}\sum_{i=1}^{N} -n_i\pi_{ij}x_{ik} \\
&= -\sum_{i=1}^{N} n_i x_{ik}\frac{\partial}{\partial\beta_{k'j'}}\left(\frac{e^{\sum_{k=0}^{K}x_{ik}\beta_{kj}}}{1+\sum_{j=1}^{J-1}e^{\sum_{k=0}^{K}x_{ik}\beta_{kj}}}\right). \qquad (3.23)
\end{aligned}
$$

Možemo primjetiti da gornja derivacija ovisi o tome je li $j = j'$ ili $j \neq j'$, stoga računamo svaki slučaj posebno.

Ako je $j = j'$, parcijalna derivacija (3.23) je jednaka

$$
\begin{aligned}
\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\beta_{kj}\partial\beta_{k'j'}} &= -\sum_{i=1}^{N} n_i x_{ik}\frac{e^{\sum_{k=0}^{K}x_{ik}\beta_{kj}}x_{ik'}(1+\sum_{j=1}^{J-1}e^{\sum_{k=0}^{K}x_{ik}\beta_{kj}}) - e^{\sum_{k=0}^{K}x_{ik}\beta_{kj}}e^{\sum_{k=0}^{K}x_{ik}\beta_{kj}}x_{ik'}}{(1+\sum_{j=1}^{J-1}e^{\sum_{k=0}^{K}x_{ik}\beta_{kj}})^2} \\
&= -\sum_{i=1}^{N} n_i x_{ik}\frac{e^{\sum_{k=0}^{K}x_{ik}\beta_{kj}}x_{ik'}(1+\sum_{j=1}^{J-1}e^{\sum_{k=0}^{K}x_{ik}\beta_{kj}} - e^{\sum_{k=0}^{K}x_{ik}\beta_{kj}})}{(1+\sum_{j=1}^{J-1}e^{\sum_{k=0}^{K}x_{ik}\beta_{kj}})^2} \\
&= -\sum_{i=1}^{N} n_i x_{ik}\pi_{ij}x_{ik'}(1-\pi_{ij}). \qquad (3.24)
\end{aligned}
$$

Ako $j \neq j'$, tada je parcijalna derivacija (3.23) jednaka

$$
\begin{aligned}
\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\beta_{kj}\partial\beta_{k'j'}} &= -\sum_{i=1}^{N} n_i x_{ik}\frac{0 - e^{\sum_{k=0}^{K}x_{ik}\beta_{kj}}e^{\sum_{k=0}^{K}x_{ik}\beta_{kj'}}x_{ik'}}{(1+\sum_{j=1}^{J-1}e^{\sum_{k=0}^{K}x_{ik}\beta_{kj}})^2} \\
&= \sum_{i=1}^{N} n_i x_{ik}\pi_{ij}x_{ik'}\pi_{ij'}. \qquad (3.25)
\end{aligned}
$$

Rješavanje sustava nelinearnih jednadžbi nije uvijek jednostavno, stoga jednadžbe $u(\boldsymbol{\beta}) = 0$ rješavamo iterativnim postupkom.

Prvi korak Newtonove metode je izabrati početnu aproksimaciju rješenja $\boldsymbol{\beta}^{(0)}$, te koristeći prvi stupanj Taylorovog polinoma[1] za aproksimaciju funkcije u, razvijamo funkciju u u okolini točke $\boldsymbol{\beta}^{(0)}$. Prva aproksimacija za $u(\boldsymbol{\beta})$ ima sljedeći oblik:

[1]Taylorov polinom stupnja n za funkciju f u točki x_0 je definiran kao prvih n članova Taylorovog reda za fukciju f i jednak je $\sum_{i=0}^{n}\frac{f^{(i)}(x_0)}{i!}(x-x_0)^i$, uz pretpostavku da prvih n derivacija funkcije f u točki x_0 postoji. (S.Czepiel [7])

$$u_1(\boldsymbol{\beta}) = u(\boldsymbol{\beta}^{(0)}) + u'(\boldsymbol{\beta}^{(0)})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}). \tag{3.26}$$

Primjetimo da je fukcija $u_1(\boldsymbol{\beta})$ jednadžba tangente funkcije $u(\boldsymbol{\beta})$ u točki $(\boldsymbol{\beta}^{(0)}, u(\boldsymbol{\beta}^{(0)}))$. Točka $(\boldsymbol{\beta}^{(1)}, 0)$, u kojoj tangenta siječe os x, odnosno rješenje jednadžbe $u_1(\boldsymbol{\beta}) = 0$, će biti korištena kao sljedeća aproksimacija nultočke funkcije $u(\boldsymbol{\beta})$:

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} - \frac{u(\boldsymbol{\beta}^{(0)})}{u'(\boldsymbol{\beta}^{(0)})}. \tag{3.27}$$

Ponavljajući navedeni proces dobivamo $\boldsymbol{\beta}^{(2)}, \boldsymbol{\beta}^{(3)}, \boldsymbol{\beta}^{(4)} \ldots$ sve dok aproksimacija ne konvergira (ako konvergira) ka jednoj vrijednosti koje je rješenje jednadžbe $u(\boldsymbol{\beta}) = 0$. Niz dobivenih aproksimacija možemo zapisati rekurzivnom formulom:

$$\boldsymbol{\beta}^{(m)} = \boldsymbol{\beta}^{(m-1)} - \frac{u(\boldsymbol{\beta}^{(m-1)})}{u'(\boldsymbol{\beta}^{(m-1)})}, \quad m \in \mathbf{N}. \tag{3.28}$$

Definirajmo matricu očekivanja $\boldsymbol{\mu}$ s N redaka i $J-1$ stupaca s elementima $n_i\pi_{ij}$, istih dimenzija kao i matrice $\mathbf{y}$ i $\boldsymbol{\pi}$. Može se pokazati da je

$$u(\boldsymbol{\beta}) = \mathbf{X}^{\mathbf{T}}(\mathbf{y} - \boldsymbol{\mu}), \tag{3.29}$$

matrica dimenzija $(K+1) \times (J-1)$ s elementima $\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_{kj}}$, gdje je $\mathbf{X}$ matrica dizajna, a $\mathbf{y}$ matrica realizacija ovisne slučajne varijable. Primjetimo da je $u(\boldsymbol{\beta})$ matrica dimenzija jednakih kao matrica $\boldsymbol{\beta}$.

Definirajmo sada kvadratnu matricu $\mathbf{W}$ reda N. Obzirom da matrica drugih derivacija ovisi o j, za $j = j'$ neka je matrica $\mathbf{W}$ dijagonalna s elementima $n_i\pi_{ij}(1-\pi_{ij})$ na dijagonali, a za $j \neq j'$ neka je $\mathbf{W}$ dijagonalna matrica čiji su elementi na dijagonali jednaki $n_i\pi_{ij}\pi_{ij'}$.

Množenjem matrica može se pokazati da je

$$u'(\boldsymbol{\beta}) = -\mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{X} \tag{3.30}$$

matrica dimenzija $(K+1) \times (K+1)$ s elementima $\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_{kj} \beta_{k'j'}}$. Koristeći navedenu dualnu formulaciju matrice $\mathbf{W}$, jednadžbu (3.27) možemo zapisati na sljedeći način:

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} + [\mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{X}]^{-1}\mathbf{X}^{\mathbf{T}}(\mathbf{y} - \boldsymbol{\mu}). \tag{3.31}$$

Ponavljajući primjenu prethodne jednadžbe sve dok vrijednosti za $\boldsymbol{\beta}$ od iteracije do iteracije ne odstupaju, odnosno konvergiraju u $\hat{\boldsymbol{\beta}}$. U tom slučaju kažemo da procjenitelj maksimalne vjerodostojnosti konvergira.

3.5 Kvaliteta modela

Nakon procjene parametara modela potrebno je povesti ocjenu njegove kvalitete. Postoje razni statistički testovi za ocjenu kvalitete modela, no u ovom potpoglavlju ćemo navesti neke metode koje su naveli Barnett i Dobson (2008.), a neke od njih ćemo i koristiti u nastavku rada.

ANOVA (analysis of variance) je termin korišten za statističku metodu kojom uspoređujemo očekivanja dvije ili više grupa. Nulhipoteza ovog testa su jednaka očekivanja promatranih grupa. (više o ANOVA vidi Barnett, Dobson, 2008, str. 102).

Kvalitetu modela možemo procjeniti i *rezidualima* koji se definiraju na sljedeći način:

$$r_i = \frac{o_i - e_i}{\sqrt{e_i}}, \tag{3.32}$$

gdje je o_i promatrana vrijednost, a e_i očekivana odnosno procjenjena vrijednost za i-to promatranje. Što su reziduali manji, model je bolji.

Devijancu D definiramo u terminima maksimalne vrijednosti funkcije vjerodostojnosti $l(\boldsymbol{\beta})$, i modela s maksimalnim brojem parametara:

$$D = 2[l(\boldsymbol{\beta}_{\max}) - l(\boldsymbol{\beta})], \tag{3.33}$$

gdje je $l(\boldsymbol{\beta}_{\max})$ maksimalna vrijednost logaritma funkcije vjerodostojnosti za model s najvećim brojem parametara.

Još neki kriteriji za uspoređivanje modela su AIC (Akaike inormacijski kriterij)

$$AIC = 2p - 2l(\boldsymbol{\beta}), \tag{3.34}$$

gdje je p broj procjenjenih parametara, te BIC (Bayesov informacijski kriterij) koji je dobar za velike N,

$$BIC = 2p \times \ln(N) - 2l(\boldsymbol{\beta}), \tag{3.35}$$

gdje je N broj promatranja. Kod oba posljednja kriterija poželjna je što manja vrijednost.

4 Primjena multinomne logističke regresije u kreditnom skoringu

U kreditnoj praksi cilj analitičara je predvidjeti hoće li potencijalni klijent otplatiti dug koji bi nastao ugovornom obvezom. Dakle, analitičar procjenjuje *kreditni rizik*. Kreditni rizik se može definirati kao mogućnost da zajmoprimac, odnosno druga ugovorna strana banke neće ispuniti svoje obveze u skladu s ugovorenim uvjetima (Bazelski odbor za superviziju banaka, *Načela za upravljanje kreditnim rizikom*, HNB 2000.).
Opis kreditnog rizika može se provesti kvalitativno i kvantitativno (Šarlija 2008.). Kvalitativan opis provodi se *klasičnom kreditnom analizom* kojom se daje opisna procjena rizičnosti. Vrši je skupina kvalitetno obučenih eksperata koji na temelju određenih karakteristika potencijalnog klijenta i subjektivne procjene odlučuju o davanju kredita. Takav način procjene pokazao se nepraktičan iz puno razloga među kojima su i sljedeći:

- kvaliteta analize ovisi o stručnosti eksperta,
- banka mora uvijek imati dovoljan broj eksperata za nastali volumen posla,
- sustav za provođenje ovakve analize je skup za održavanje,
- primjenjujući dugo ista pravila banka se može uljuljkati u lažnu sigurnost.

Iz navedenih razloga i povećanja potražnje za kreditima javlja se potreba za kvantificiranjem kreditnog rizika. Uvode se *kredit skoring sustavi*, odnosno modeli kreditnih rizika koji za rezultat daju vjerojatnost da klijent neće ispuniti svoju obvezu. Na temelju dobivene vrijednosti klijenta se svrstava u jednu od dvije kategorije - *dobar* ili *loš* klijent. *Dobar* je klijent koji niti jednom nije kasnio pri plaćanju rate kredita, dok je *loš* klijent onaj koji je barem jednom kasnio više od 31 dan. Upravo iz razloga što klijent može biti svrstan u jednu od dvije kategorije, u većini istraživanja je ovisna varijabla modela binomna. Kao što je spomenuto u prethodnim poglavljima, za probleme takvog tipa najčešće se koristi (binomna) logistička regresija.

Ali, osim *dobrih* i *loših* klijenata u bazi podataka svake banke nalaze se i *srednji* klijenti. Oni ne odgovaraju definiciji niti *dobrih* niti *loših* i uglavnom ih se izostavlja pri procesu kreditnog skoringa. Takvi klijenti u pravilu otplate svoj dug, ali uz kašnjenje u otplati koje za banku nije u potpunosti neprihvatljivo, kao što je to slučaj s kategorijom *loših* klijenata čije kašnjenje u otplati banci stvara veće gubitke. Možemo reći da klijenti koji nisu u kategoriji *dobrih* nisu nužno *loši*, i obrnuto. Uvodeći *srednje* klijente dobiva se uvid u čitav uzorak klijenata koji su poslovali s bankom, što je jedan od razloga zbog kojih ih je zanimljivo uključiti u analizu. Ukoliko je uzorak mali i sadrži klijente koji čiji su krediti još aktivni, svakako je korisno i *srednje* klijente uključiti u analizu jer nam je u tom slučaju svaki raspoloživi podatak koristan. (Benšić, Bohaček, Šarlija, 2004.)

4.1 Prethodna istraživanja

Benšić, Bohaček i Šarlija (2004) proveli su analizu korisnosti uvođenja kategorije *srednjih* u model kreditnog skoringa uspoređujući rezultate multinomnog i tri binomna modela. Korištena je baza klijenata kojima su odobreni potrošački krediti. Kreirana su tri binomna modela. Pri kreiranju prvog modela korišteni su samo *dobri* i *loši*, bez *srednjih*, u drugom modelu *srednji* klijenti su pridruženi *dobrima*, a u trećem su *srednji* pridruženi *lošima*. Prvi model bez *srednjih* pokazuje se kao najuspješniji pri predikciji, a najlošiji je onaj gdje su *srednji* pridruženi *lošima*. U nastavku kreiran je multinomni model koji je imao najnižu prediktivnost, a *srednje* klijente gotovo i ne prepoznaje, dok *loše* klijente bolje detektira od binomnog modela u kojem su *srednji* pripojeni *lošima*. Na kraju kreiran je model koji opravdava korištenje kategorije *srednjih* u kreditnom skoringu. Binomnim modelom koji je kreiran bez *srednjih*, izvršena je procjena *srednjih* od kojih je njih 39.57% procjenjeno kao *loši*, a 60.43% kao *dobri*. Zatim se na temelju skupa podataka *dobrih*, *loših* i *srednjih* koji su procijenjeni kao *dobri* ili *loši*, ponovno kreirao binomni model koji je dao najbolje rezultate.

Tsai (2012) u svom radu vrši usporedbu binomnog i multinomnog modela u predviđanju korporativnog bankrota tvrtki. Uvodi kategoriju *srednjih*, kao podkategoriju loših kako bi pokušao detektirati različite rizične faktore pri objašnjavanju različitih tipova korporativnog bankrota. *Srednji* su definirani kao nestabilne tvrtke, odnosno one koje su pod utjecajem raznih poteškoća u poslovanju, ali još nisu bankrotne. Za kreiranje modela korišteni su podaci od 2003. – 2006., a za validaciju podaci od 2007. – 2008. Predikcija za 2007. i 2008. je bolja binomnim modelom, gdje je točnost 91.18% i 70.46% redom, dok je kod multinomnog modela točnost predikcije 61.87% za 2007. godinu, a 56.41% za 2008. Za podatke iz 2008. su velike greške u predikciji za oba modela što govori da je predikcija bolja ako se predviđa za bližu budućnost. *Srednje* i *loše* firme pokazuju se kao slične po karakteristikama, a rezultat predikcije nije ništa bolji ako za loše tvrtke uvedemo više kategorija (ovdje su to bile nestabilne i bankrotne tvrtke).

4.2 Varijable i podaci

Za provođenje istraživanja u ovom radu na raspolaganju smo imali bazu podataka od 3165 klijenata neke banke. Od toga je 2293 klijenata ocijenjeno kao *dobri*, 673 kao *srednji*, te 199 kao *loši*.

Za kreiranje modela korišteni su podaci od slučajno odabranih 2165 klijenata od čega je 1526 *dobrih*, 470 *srednjih* i 169 *loših*. Pri odabiru uzorka vodili smo računa da struktura uzorka bude slična strukturi baze, odnosno da udjeli *dobrih*, *srednjih* i *loših* budu približno jednaki udjelima navedenih kategorija u bazi. Za validaciju korišteni su podaci preostalih 1000 klijenata od kojih je 767 *dobrih*, 203 *srednja* i 30 *loših*.

Klijenti koji niti pri jednom plaćanju rate kredita nisu kasnili u otplati definirani su kao *dobri*, klijenti koji nikada nisu kasnili više od 31 dan su definirani kao *srednji*, dok su *loši* klijenti oni

koji su barem jednom zakasnili više od 31 dan pri plaćanju obveza prema banci. Dakle, ovisna varijabla LOSI ima tri kategorije - *dobri*, *srednji* i *loši*, numerički označene s 0, 1 i 2 redom.
Baza podataka sadrži 13 (input) neovisnih varijabli kojima je okarakterizitran svaki od 3165 klijenata. U nastavku je opis svake od varijabli, te kategorizacija nekih od njih. Za kategorijalne varijable su u zagradama navedene numeričke oznake za svaku kategoriju.

- DOB - kontinuirana varijabla koja opisuje starosnu dob klijenta.
- STAŽ - radni staž klijenta u mjesecima. Varijabla je kategorizirana jer skupina klijenata nije dala podatak o radnom stažu. Kategorije: nema podatka (-1); ≤ 24 (1); $25-48$ (2); $49-72$ (3); $73-96$ (4); $97-120$ (5); > 120 (6);
- BRAČNO - bračno stanje klijenta. Kategorije: neudana/neoženjen ili živi u zajednici (1); udana/oženjen (2); udovac/udovica (3); razveden/a (4)
- CIJENA - kontinuirana varijabla koja opisuje iznos cijene robe za koju klijent traži kredit.
- GOTOVINA - kontinuirana varijabla, iznos gotovine koju klijent daje pri kupovini robe.
- STAN - stambeno stanje klijenta. Kategorije: unajmljen (1); stanarsko pravo (2); društveni stan (3); vlastita kuća/stan (4); s roditeljima (5); ostalo (6).
- RATA - kontinuirana varijabla, opisuje iznos rate kredita.
- DOHODAK - iznos mjesečnog dohotka klijenta.
- ZANIMNJE - zanimanje klijenta: poduzetnik (1); umirovljenik (2); zaposlenik (3)
- NAČIN - način plaćanja: uplatnice (1); trajni nalog (2)
- ROBA - roba koju klijent kupuje: bijela tehnika (1); crna tehnika, foto, kino, glazbala (2); kućne potrepštine, kućni tekstil (3); mobiteli, telefonska tehnika (4); namještaj, kuhinje, sanitarije (5); računala, uredska tehnika (6); ostalo (7)
- ISKUSTVO - dihotomna varijabla, prima vrijednost (1) ukoliko je novi klijent, a vrijednost (2) ukoliko banka već ima iskustava s klijentom koji podnosi zahtjev za kredit
- KVALITETA - kvaliteta dućana u kojem klijent kupuje robu. Razvrstani su u 6 kategorija: (A) najboljih 10% (1); (B) iznad prosjeka (2); (C) ispod prosjeka (3); (D) najlošijih 10% (4); (E) novi dućani s manje od 6 mjeseci poslovanja (5); (F) mali dućani koji sklapaju do 6 ugovora mjesečno (6).

Uz navedene varijable, uvodimo još 3 neovisne varijable, izvedene pomoću nekih prethodnih. To su sljedeći omjeri:

- gotovina/cijena - predstavlja omjer gotovine i cijene. Izražen je u postotku.
- rata/cijena - omjer rate kredita i cijene robe koju klijent kupuje, izraženo u postotku.
- cijena/dohodak - omjer cijene robe i mjesečnog dohotka klijenta. Obzirom da postoji skupina klijenata koji imaju dohodak jednak nuli, ovu varijablu ćemo kategorizirati: klijenti bez dohotka (0); $0-0.5$ (1); $0.5-1$ (2); $1-1.5$ (3); >1.5 (4).

4.3 Analiza karakteristika

U većini slučajeva svaka neovisna varijabla ne daje informaciju kojom bi se moglo naslutiti kojoj kategoriji ovisne varijable klijent pripada. Stoga, prije kreiranja skoring modela za svaku od neovisnih varijabli koje smo naveli u prethodnom poglavlju, potrebno je ispitati prediktivnost. Ukoliko se varijabla pokaže kao neprediktivna, nema ju smisla uključivati u model, no ako se pokaže kao prediktivna, i dalje ne znamo kako će se utjecati na točnost predikcije skoring modela, ali je ima smisla pokušati uključiti u model.

Za svaku neovisnu varijablu napravit ćemo tablicu frekvencija da dobijemo kratak uvid o ponašanju klijenata po kategorijama neovisne varijable. Izračunat ćemo postotak svake od kategorija ovisne varijable po kategorijama neovisnih, te ćemo izračunati *informacijski omjer (engl. information odds) dobrih* u odnosu na *loše* i informacijski omjer (dalje u tekstu *inf. omjer*) *srednjih* u odnosu na *loše*. Za naš slučaj inf. omjer možemo definirati kao omjer postotka jedne od 'viših' kategorija ovisne varijable i postotka bazne kategorije, u našem slučaju bazna kategorija su *loši* klijenti. Inf. omjer možemo interpretirati na sljedeći način, koliko klijenata jedne kategorije ovisne varijable dolazi na jednog klijenta iz bazne kategorije ovisne varijable.
Analizu prediktivnosti ćemo provesti ANOVA testom, na temelju kojeg ćemo izvesti zaključak o korisnosti uključivanja pojedine neovisne varijable u skoring model.

ANOVA analizu provest ćemo usporedbom dva modela - modela bez ijedne neovisne varijable (nulmodel) i modela koji sadrži samo jednu neovisnu varijablu za koju vršimo analizu. Cilj ove analize je ispitati nul-hipotezu o jednakim očekivanjima navedena dva modela. Ukoliko je dobivena $p-$vrijednost manja od 0.05, smatrat ćemo da na razini značajnosti 5% možemo odbaciti nul-hipotezu o jednakim očekivanjima i pretpostaviti da model s neovisnom vrijablom bolje razlikuje kategorije ovisne varijable. Dodatni pokazatelj bolje prediktivnosti modela s neovisnom varijablom je manja AIC vrijednost od AIC vrijednosti nulmodela, te što veća razlika reziduala između dva modela, odnosno LR statistika.

Rezultati ANOVA testova prikazani su u tablici (1).

Navedena analiza provedena je u programskom jeziku *R*. Modeli su kreirani naredbom *polr* iz paketa $MASS$ koja za rezultat (*output*) daje prametar β procjenjen metodom maksimalne vjerodostojnosti, standardnu grešku i vrijednost Waldove statistike. ANOVA analiza provedena je

naredbom *anova* iz paketa *stats*.

Tablica 1: Test značajnosti neovisnih varijabli

R.Br.	Model	Resid df	Resid. Dev	Test	Df	LR stat	Pr(Chi)	AIC
1	Nulmodel	2163	3365,29					3369,29
2	dob	2162	3339,82	1 vs 2	1	25,46	4,50E-07	3345,82
3	zanimanje	2161	3351,40	1 vs 3	2	13,89	0,0010	3359,40
4	bračno	2160	3351,20	1 vs 4	3	14,10	0,0028	3361,20
5	cijena/dohodak	2159	3344,82	1 vs 5	4	20,47	0,0004	3356,82
6	cijena	1356	2340.86	1 vs 6	807	1024.43	2,78E-07	3958,86
7	dohodak	720	1488,82	1 vs 7	1443	1876,47	7,32E-14	4378,82
8	gotovina/cijena	1682	2737,93	1 vs 8	481	627,36	7,40E-06	3703,93
9	gotovina	2162	3364,45	1 vs 9	1	0,84	0,3606	3370,45
10	iskustvo	2162	3317,88	1 vs 10	1	47,41	5,77E-12	3323,88
11	kvaliteta	2158	3331,85	1 vs 11	5	33,44	3,07E-06	3345,85
12	način	2162	3365,29	1 vs 12	1	0,003	0,9556	3371,29
13	rata/cijena	1570	2574,65	1 vs 13	593	790,64	8,89E-08	3764,65
14	rata	1257	2161,28	1 vs 14	3906	1204,01	9,52E-11	3977,28
15	roba	2157	3348,16	1 vs 16	6	17,13	0,0088	3364,16
16	stan	2158	3339,63	1 vs 17	5	25,66	0,0001	3353,63
17	staž	2162	3349,69	1 vs 18	1	15,60	7,81E-05	3355,69

I. Dob

Varijabla DOB opisuje starosnu dob svakog klijenta. Najmlađi klijent ima 18, a najstariji 82 godine, dok je srednja vrijednost svih klijenata 38.83 godina. U tablici (2) podjelili smo klijente u 7 starosnih kategorija kako bi detektirali koje dobi su klijenti koji su bolji, a koji lošiji pri otplati kredita.

Najveći udio *dobrih* klijenata u odnosu na *loše* je u kategorji klijenata koji su stariji od 50 godina, inf. omjer iznosi 2.06 što možemo interpretirati da na jednog *lošeg* dolaze 2 *dobra*. Udio *srednjih* najveći je u kategoriji klijenata između 31 i 35 godina, gdje na jednog *lošeg* dolazi 1.85 *srednjih*. Inf. omjeri najmanji su za klijente mlađe od 26 godina, tj. jedan *dobar* klijent dolazi na 2 *loša* (inf. omjer *dobrih* je 0.49), te jedan *srednji* na 2 *loša* (inf. omjer *srednjih* je 0.58).

Možemo zaključiti da su klijenti mlađi od 26 najrizičniji pri otplati kredita, dok klijenti stariji od 50 imaju najveću šansu biti *dobri*.

Usporedbom nulmodela i modela koji sadrži samo varijablu DOB, model s varijablom DOB se pokazuje bolji. Razlika u rezidualnoj devijaciji je 25.46, a dobivena p-vrijednost je jednaka 4.5×10^{-7} na temelju čega odbacujemo nul-hipotezu o jednakim očekivanjima i pretpostavljamo da je ova varijabla prediktivna, odnosno da dobro razlikuje kategorije ovisne varijable i ima ju

Tablica 2: Varijabla DOB

Kategorija	Dobri	Srednji	Loši	%D	%S	%L	i.omjer D	i.omjer S
<26	235	86	53	15.40%	18.30%	31.36%	0.49	0.58
26-30	191	65	31	12.52%	13.83%	18.34%	0.68	0.75
31-35	185	72	14	12.12%	15.32%	8.28%	1.46	1.85
36-40	210	57	15	13.76%	12.13%	8.88%	1.55	1.37
41-45	211	62	20	13.83%	13.19%	11.83%	1.17	1.12
46-50	196	61	20	12.84%	12.98%	11.83%	1.09	1.10
>50	298	67	16	19.53%	14.26%	9.47%	2.06	1.51

smisla pokušati uključiti u model. Također AIC vrijednost modela s varijablom DOB je manja, što potvrđuje prethodnu pretpotavku.

II. Staž

Varijabla STAŽ daje informaciju o radnom stažu klijenta. Među kategorijama prema kojima smo prema radnom stažu podjelili klijente, ova varijabla ima kategoriju koja ne daje informaciju o stažu, nego klijent iz nekog razloga nije ponudio taj podatak. Stoga se se u toj kategoriji mogu naći klijenti koji bi mogli biti svrstani u bilo koju od preostalih, a često može značiti i da klijent nema radnog iskustva.

Tablica 3: Varijabla STAŽ

Kateg.	Dobri	Srednji	Loši	%D	%S	%L	i.omjer D	i.omjer S
nema podatka	323	129	33	21.17%	27.45%	19.53%	1.08	1.41
≤ 24	394	123	76	25.82%	26.17%	44.97%	0.57	0.58
25-48	251	70	29	16.45%	14.90%	17.16%	0.96	0.87
49-72	173	42	9	11.34%	8.94%	5.33%	2.13	1.68
73-96	122	40	5	7.99%	8.51%	2.96%	2.70	2.88
97-120	65	17	9	4.26%	3.62%	5.33%	0.80	0.68
>120	198	49	8	12.98%	10.43%	4.73%	2.74	2.20

Za sve kategorije ovisne varijable najveći udio klijenata je u kategoriji s radnim stažem manjim od 24 mjeseca. Udjeli *dobrih* i *srednjih* klijenata po kategorijama ove neovisne varijable se slično ponašaju. Gotovo polovica *loših* klijenata pripada kategoriji s manje od 24 mjeseca staža. Za navedenu kategoriju su inf. omjeri najmanji i iznose 0.57 za *dobre* i *0.58* za *srednje*. Dakle, za klijente mlađe do 24 na jednog *dobrog* klijenta dolaze 2 *loša*, kao i na jednog *srednjeg*. Najbolji klijenti se pokazuju oni s više od 10 godina radnog staža, te oni između 6 i 8 godina. Za te dvije kategorije inf. omjeri su veći od 2, dakle šansa da klijent bude *dobar*, ili da bude *srednji*, je najveća za navedene dvije kategorije.
ANOVA testom dobivena je p-vrijednost 7.81×10^{-5} što je na razini značajnosti 5% dovoljno mala vrijednost da odbacimo nul-hipotezu o jednakim očekivanjima nulmodela i modela koji sadrži

samo varijablu STAŽ. AIC vrijednost modela s varijablom STAŽ je manja od AIC vrijednosti nulmodela. Sukladno dobivenim rezultatima možemo pretpostaviti da je varijabla STAŽ prediktivna i možemo je pokušati uključiti u model.

III. Bračno

Ova varijabla opisuje bračno stanje klijenta. Najveći udjeli svih kategorija su u kategoriji klijenata koji su u braku. Veliki udio neudanih/neoženjenih (36.69%) je u kategoriji *loših*. *Dobri* i *srednji* klijenti se slično ponašaju u ovisnosti o kategoriji bračnog statusa.

Tablica 4: Varijabla BRAČNO

Kateg.	Dobri	Srednji	Loši	%D	%S	%L	i.omjer D	i.omjer S
neud./neož.	302	101	62	19.79%	21.49%	36.69%	0.54	0.59
u braku	928	296	78	60.81%	62.98%	46.15%	1.32	1.36
udovac(ica)	212	55	22	13.89%	11.70%	13.02%	1.07	0.90
razveden(a)	84	18	7	5.50%	3.83%	4.14%	1.33	0.92

Najveći inf. omjeri su za kategorije razvedenih i onih koji su u braku, a najmanje za neudane/neoženjene. Dakle, klijenti koji su u braku ili su razvedeni imaju veću šansu biti *dobri*, ili biti *srednji*. Dok je za neudane/neoženjene najveća šansa da budu *loši*.
ANOVA testom dobivena je $p-$vrijednost 0.0028, što je na razini značajnosti 5% dovoljno mala vrijednost da odbacimo pretpostvaku o jednakim očekivanjima. AIC vrijednost modela s varijablom BRAČNO manja je od AIC vrijednosti nulmodela. Ovu varijablu možemo pokušati uključiti u model.

IV. Cijena

Sljedeća varijabla je cijena robe koju klijent kupuje. Iz tablice frekvencija ove varijable je teško uočiti koja bi kategorija najbolje opisivala *dobre, srednje* ili *loše* klijente. Trend udjela po kategorijama ove varijable i *dobrih* i *srednjih* i *loših* je sličan. Najveći broj klijenata svih kategorija ovisne varijable kupuje robu cijene od $3000-6000$.
Prema vrijednostima inf. omjera možemo vidjeti da se klijenti koji kupuju robu cijene manje ili jednake 1500 pokazuju kao najbolji, odnosno u odnosu na *loše* klijente imaju najveću šansu biti *dobri*, dok klijetni koji kupuju robu skuplju od 9000 imaju najveću šansu biti *srednji*. Najmanji inf. omjer je za kategoriju cijene robe od $6001-9000$, iz čega možemo zaključiti da je ta kategorija najrizičnija.
Provedbom ANOVA testa dobivena je $p-$vrijednost 2.78×10^{-7}. AIC vrijednost modela je veća do AIC vrijednosti nulmodela. Razidualna devijanca je za 1024 manja za model s varijablom cijena.

Tablica 5: Varijabla CIJENA

Kateg.	Dobri	Srednji	Loši	%D	%S	%L	i.omjer D	i.omjer S
≤ 1500	92	16	6	6.03%	3.40%	3.55%	1.70	0.96
1501-3000	431	124	42	28.24%	26.38%	24.85%	1.14	1.06
3001-6000	676	202	77	44.30%	42.98%	45.56%	0.97	0.94
6001-9000	222	76	30	14.55%	16.17%	17.75%	0.82	0.91
>9000	105	52	14	6.88%	11.06%	8.28%	0.83	1.34

Dobiveni rezultati upućuju da ova neovisna varijabla razlikuje kategorije ovisne varijable. Možemo ju pokušati uključiti u model.

V. Gotovina

Ova varijabla daje nam informaciju o gotovini koju klijent daje pri kupnji robe. Promotrimo li tablicu frekvecija ove varijable, može se uočiti sličan trend svih kategorija ovisne varijable po kategorijama varijable GOTOVINA. Najveći inf. omjeri su za klijente koji pri kupnji daju 800 i više, te za njih možemo reći da je najveća šansa da budu *dobri* ili da budu *srednji*. No, nema značajnog odstupanja među inf. omjerima među kategorijama ove varijable.

Tablica 6: Varijabla GOTOVINA

Kateg.	Dobri	Srednji	Loši	%D	%S	%L	i.omjer D	i.omjer S
< 200	298	116	30	19.53%	24.68%	17.75%	1.10	1.39
200 – 400	509	112	58	33.36%	23.83%	34.32%	0.97	0.69
400 – 600	216	65	22	14.15%	13.83%	13.02%	1.09	1.06
600 – 800	129	37	23	8.45%	7.87%	13.61%	0.62	0.58
800 – 1000	67	24	6	4.39%	5.11%	3.55%	1.24	1.44
≥ 1000	307	116	30	20.12%	24.68%	17.75%	1.13	1.39

Dobivena $p-$vrijednost je 0.3606. Na razini značajnosti 5% odbacujemo nul-hipotezu o jednakim očekivanjima nulmodela i modela koji sadrži samo varijablu GOTOVINA. Razlika u AIC vrijednostima, te razlika reziduala su zanemarivi, na temelju čega možemo pretpostaviti da oba modela slično razlikuju kategorije ovisne varijable. Ovu varijablu nema smisla uključivati u model.

VI. Omjer gotovine i cijene

Obzirom da se prethodna varijabla nije pokazala suviše prediktivna, pomoću nje možemo izvesti novu varijablu i provjeriti njenu prediktivnost. Pokušat ćemo provjeriti značajnost udjela gotovine koju klijent daje pri kupnji robe u odnosu na ukupnu cijenu robe.

Prema tablici frekvencija vidimo da najveći udio *loših* klijenata pripada kategoriji klijenata kojima je udio gotovine 10% – 20% u odnosu na cijenu robe. U istoj kategoriji je i najveći udio *dobrih* klijenata (39.25%). *Srednjih* klijenata je najviše u kategoriji 0% – 10% (29.15%).
Najbolji klijenti pokazuju se oni kojima je ovaj omjer veći od 30%, šansa da klijent bude *dobar*, a ne *loš* je 2.67 : 1, a da bude *srednji* a ne *loš* gotovo 3 : 1.

Tablica 7: Gotovina/Cijena

Kateg.	Dobri	Srednji	Loši	%D	%S	%L	i.omjer D	i.omjer S
0%	160	95	12	10.48%	20.21%	7.10%	1.48	2.85
0-10%	422	137	56	27.65%	29.15%	33.14%	0.83	0.88
10-20%	599	113	71	39.25%	24.04%	42.01%	0.93	0.57
20-30%	176	68	23	11.53%	14.47%	13.61%	0.85	1.06
>30%	169	57	7	11.07%	12.13%	4.14%	2.67	2.93

ANOVA testom dobivena $p-$vrijednost je 7.4×10^{-6}. AIC vrijednost modela s ovom varijablom je veća od AIC vrijednosti nulmodela, a razlika reziduala je 627.36. Zbog dobivene $p-$vrijednosti možemo pretpostaviti da je ova varijabla prediktivna i pokušat ćemo je uključiti u model.

VII. Stan

U ovoj varijabli sadržan je podatak o stambenom statusu klijenta. Najveći broj klijenata živi u društvenom stanu, 40.85% *dobrih*, 42.74% *srednjih* i 25.44% *loših*. Društveni stan je ujedno kategorija u kojoj je najviše *dobrih* i *srednjih* klijenata. *Loših* klijenata najviše je među onima koji žive u vlastitom stanu.
Prema inf. omjerima, najbolji se pokazuju klijenti koji žive u društvenom stanu i oni koji žive s roditeljima. Najmanji inf. omjer je za klijente koji imaju stanarsko pravo, te oni koji žive u vlastitom stanu.

Tablica 8: Varijabla STAN

Kateg.	Dobri	Srednji	Loši	%D	%S	%L	i.omjer D	i.omjer S
Unajmljen	154	40	21	10.09%	8.51%	12.43%	0.81	0.68
Stan. Pravo	58	19	11	3.80%	4.04%	6.51%	0.58	0.62
Društ. Stan	661	192	43	40.85%	42.74%	25.44%	1.70	1.61
Vlastiti	220	78	46	14.42%	16.60%	27.22%	0.53	0.61
S roditeljima	323	95	26	21.17%	20.21%	15.38%	1.38	1.31
Ostalo	110	46	22	7.21%	9.79%	13.02%	0.55	0.75

ANOVA testom dobivena je p-vrijednost 0.0001, na temelju čega odbacujemo nul-hipotezu o jednakim očekivanjima numlmodela i modela s varijablom STAN. AIC vrijednost modela je manja od AIC vrijednosti nulmodela, a razlika reziduala je 25.66. Možemo zaključiti da bi ovu varijablu

imalo smisla pokušati uključiti u model.

VIII. Dohodak

Varijabla DOHODAK nosi podatak o mjesečnom dohotku klijenta. Klijenti koji imaju najveću šansu biti *dobri* imaju dohodak od 4500 – 5500, gdje 1.87 *dobrih* dolazi na jednog *lošeg*. Klijenti s najvećim inf. omjerima *srednjih* su klijenti s dohotkom 4500 – 5500, te klijenti bez dohotka. Najmanji inf. omjeri su za kategoriju klijenata s dohotkom manjim od 2500, te se oni pokazuju kao najrizičniji.

Tablica 9: Varijabla DOHODAK

Kateg.	Dobri	Srednji	Loši	%D	%S	%L	i.omjer D	i.omjer S
0	141	81	18	9.24%	17.23%	10.65%	0.87	1.62
< 2500	156	49	23	10.22%	10.43%	13.61%	0.75	0.77
2500-3500	393	104	44	25.75%	22.13%	26.04%	0.99	0.85
3500-4500	387	96	44	25.36%	20.43%	26.04%	0.97	078
4500-5500	236	57	14	15.47%	12.13%	8.28%	1.87	1.46
≥ 5500	213	83	26	13.96%	17.66%	15.38%	0.91	1.15

Provedbom statističkog testa dobivena p vrijednost je 7.32×10^{14} na temelju čega odbacujemo nul-hipotezu o jednakim očekivanjima nulmodela i modela s varijablom DOHODAK. Rezidualna devijanca modela s varijablom DOHODAK znatno je manja od devijance nulmodela.
Sukladno dobivenim rezultatima, možemo zaključiti da ovu varijablu možemo pokušati uključiti u model.

IX. Omjer cijene i dohotka

Omjer cijene i dohotka je još jedna izvedena varijabla čiju ćemo prediktivnost ispitati. Iz tablice frekvencija vidimo da je najviše klijenata u kategoriji gdje cijena čini 50% – 100% dohotka. Za navedenu kategoriju je i najveća šansa da klijent bude *dobar*, inf. omjer je 1.25.

Tablica 10: Omjer cijene i dohotka

Kateg.	Dobri	Srednji	Loši	%D	%S	%L	i.omjer D	i.omjer S
0	141	81	18	9.24%	17.23%	10.65%	0.87	1.62
0-0.5	203	42	20	13.30%	8.94%	11.83%	1.12	0.76
0.5-1	543	150	48	35.58%	31.91%	28.40%	1.25	1.12
1-1.5	316	94	37	20.71%	20.00%	21.89%	0.95	0.91
>1.5	323	103	46	21.17%	21.91%	27.22%	0.78	0.81

Klijenti bez dohotka imaju najveći inf. omjer *srednjih*, dok se kategorija sa cijenom većom od dohotka za više od 150% pokazuje kao najrizičnija.

ANOVA analizom dobivena je $p-$vrijednost 0.0004 na temelju čega možemo zaključiti da ova varijabla dobro razlikuje kategorije ovisne varijable i ima je smisla uključiti u model. AIC vrijednost modela s ovom varijablom je manja od AIC vrijednosti nulmodela, a razlika reziduala je 20.47.

X. Rata

Udjeli klijenata po kategorijama mjesečne rate kredita se ponašaju slično po svim kategorijama ovisne varijable. Najviše klijenata je u kategoriji sa ratom kredita $200-400$, a najmanje s ratom većom od 800.
Najbolji se pokazuju klijenti sa ratom $400-600$, a najlošiji oni sa ratom većom od 800.

Tablica 11: Varijabla RATA

Kateg.	Dobri	Srednji	Loši	%D	%S	%L	i.omjer D	i.omjer S
< 200	258	75	29	16.91%	15.96%	16.57%	1.02	0.96
200-400	683	208	78	44.76%	44.26%	46.15%	0.97	0.96
400-600	371	121	36	24.31%	25.74%	21.30%	1.14	1.21
600-800	128	45	16	8.39%	9.57%	9.47%	0.89	1.01
≥ 800	86	21	11	5.64%	4.47%	6.51%	0.87	0.69

ANOVA analizom dobivena je $p-$vrijednost 9.52×10^{-11} na temelju čega odbacujemo nul-hipotezu o jednakim očekivanjima modela s varijablom RATA i nulmodela. Razlika reziduala je 1204, a AIC vrijednost nulmodela je manja od AIC vrijednosti modela s varijablom RATA. Iako AIC vrijednost modela ne potvrđuje kvalitetu modela s varijablom RATA, na temelju dobivene $p-$vrijednosti ovu varijablu možemo pokušati uključiti u model.

XI. Omjer rate i cijene

Posljednja izvedena varijabla daje nam omjer rate i cijene robe koju klijent kupuje. Omjer je izražen u postotku. Rata najvećeg dijela klijenata čini $5\% - 10\%$ ukupne cijene robe koju klijent kupuje. U toj kategoriji je više od 50% *dobrih*, *srednjih* i *loših* klijenata.
Prema inf. omjerima u odnosu na *loše* klijente, najveću šansu biti *dobri* imaju klijenti kojima rata čini $5\% - 10\%$ cijene robe koju, a najveću šansu biti *srednji* imaju klijenti s ratom manjom od 5% cijene. Najlošiji se pokazuju klijenti s ratom većom od 15% cijene robe koju kupuju.

ANOVA testom dobivena je $p-$vrijednost 8.89×10^{-8} na temelju čega odbacujemo nul-hipotezu o jednakim očekivanjima i pretpostavljamo da je ova varijabla prediktivna i ima je smisla pokušati uključiti u model. AIC vrijednost je veća od AIC vrijednosti nulmodela, a razlika u rezidualima je 790.64.

Tablica 12: Omjer rate i cijene

Kateg.	Dobri	Srednji	Loši	%D	%S	%L	i.omjer D	i.omjer S
<5%	107	59	13	7.01%	12.55%	7.69%	0.91	1.63
5-10%	825	252	94	54.06%	53.62%	55.62%	0.97	0.96
10-15%	485	119	47	31.78%	25.32%	27.81%	1.14	0.91
>15%	109	40	15	7.14%	8.51%	8.88%	0.80	0.96

XII. Zanimanje

Prema tablici frekvencija najviše klijenata je u kategoriji zaposlenih. *Dobrih* je najmanje među poduzetnicima, a *loših* i *srednjih* klijenata je najmanje među umirovljenicima.
Kod ove varijable umirovljenici su skupina koja ima najveću šansu biti *dobri*. Poduzetnici imaju najveću šansu biti *srednji*, dok je za zaposlenike najveća rizičnost za ulazak u stanje neispunjavanja obveza, za koje 0.98 *dobrih* i 0.91 *srednjih* dolazi na jednog *lošeg*.

Tablica 13: Varijabla ZANIMANJE

Kateg.	Dobri	Srednji	Loši	%D	%S	%L	i.omjer D	i.omjer S
poduzetnik	142	81	18	9.31%	17.23%	10.65%	0.87	1.62
umirovljenik	187	49	16	12.25%	10.43%	9.47%	1.29	1.10
zaposlenik	1197	340	135	78.44%	72.34%	79.88%	0.98	0.91

Provodeći ANOVA test dobivena je $p-$vrijednost 0.001, te odbacujemo nul-hipotezu o jednakim očekivanjima nulmodela i modela s varijablom ZANIMANJE. AIC vrijednost modela s ovom varijablom je manja od AIC vrijednosti nulmodela što je dodatna potvrda da model s varijablom ZANIMANJE bolje razlikuje kategorije ovisne varijable. Ovu varijablu možemo pokušati uključiti u model.

XIII. Način

Način na koji klijent plaća kredit je također informacija s kojom rapolažemo i možemo ispitati njenu prediktivnost. Prema tablici frekvencija ove varijable vidimo da su veći udjeli klijenata koji plaćaju uplatnicama po svim kategorijama ovisne varijable. Taj je udio nešto manji kod *loših* klijenata od kojih 42.0% plaća trajnim nalogom. Prema inf. omjerima klijenti koji plaćaju uplatnicama pokazuju se kao manje rizični nego oni koji plaćaju trajnim nalogom.

ANOVA testom dobivena je $p-$vrijednost 0.9556, razlika reziduala modela s varijablom NAČIN i nulmodela je 0. AIC vrijednost modela s ovom varijablom je veća od AIC vrijednosti nulmodela. Prema navedenim rezultatima zaključujemo da ova varijabla ne razlikuje kategorije ovisne varijable i nema je smisla uključivati u model.

Tablica 14: Varijabla NAČIN

Kateg.	Dobri	Srednji	Loši	%D	%S	%L	i.omjer D	i.omjer S
uplatnice	987	323	97	64.68%	68.72%	57.40%	1.13	1.20
trajni nalog	539	147	72	35.32%	31.28%	42.60%	0.83	0.73

XIV. Roba

Obzirom na vrstu robe koju klijenti kupuju, najviše *dobrih* (32.18%) i *srednjih* klijenata (30.43%) kupuje crnu tehniku. Najviše *loših* klijenata (33.73%) kupuje mobitele. Roba koja se najmanje kupuje po svim kategorijama ovisne varijable su kućne potrepštine, namještaj, kuhinje, računala i uredska tehnika.
Najmanje rizična kategorija se pokazuje kategorija klijenata koji kupuju bijelu tehniku gdje na jednog *lošeg* klijenta dolazi više od 2 *dobra* i više od 2 *srednja*. Najmanji inf. omjeri su za klijente koji kupuju mobitele, računala i uredsku tehniku, te se oni pokazuju kao najrizičniji.

Tablica 15: Varijabla ROBA

Kateg.	Dobri	Srednji	Loši	%D	%S	%L	i.omjer D	i.omjer S
Bijela teh.	278	79	14	18.22%	16.81%	8.28%	2.20	2.03
Crna teh.	491	143	51	32.18%	30.43%	30.18%	1.07	1.01
Kućne potr.	55	28	5	3.60%	5.96%	2.96%	1.22	2.01
Mobiteli	375	85	57	24.57%	18.09%	33.73%	0.73	0.54
Kuhinje	83	45	10	5.44%	9.57%	5.92%	0.92	1.62
Rač. i ured	47	21	10	3.08%	4.47%	5.92%	0.52	0.76
Ostalo	197	69	22	12.91%	14.68%	13.02%	0.99	1.13

Dobivena $p-$vrijednost je 0.0088 na temelju čega odbacujemo nul-hipotezu o jednakim očekivanjima nulmodela i modela s varijablom ROBA. Razlika u rezidualima je 17.13, a AIC vrijednost modela s ovom varijablom je manja od AIC vrijednosti nulmodela. Sukladno dobivenim rezultatima možemo pretpostaviti da je ova varijabla prediktivna i pokušat ćemo je uključiti u model.

XV. Iskustvo

Varijabla ISKUSTVO nosi podatak o tome je li banka već imala iskustvo s klijentom. Klijent koji je otplatio jedan kredit ili ga još otplaćuje u bazi će biti označen kao novi klijent, dok klijenti koji imaju zaključen bar jedan kredit i trenutno otplaćuju drugi ili više, pri analizi podataka bit će klasificirani kao stari klijenti.
Prema tablici frekvencija je znatno veći broj novih klijenata. Inf. omjeri za nove klijente su približno jednaki 1, što znači da na jednog *lošeg* dolazi jedan *dobar* i jedan *srednji*. Za stare

klijente inf. omjer *dobrih* je 0.49, dakle za kategoriju starih klijenata je veća šansa da budu *dobri* nego *loši*. Dok je inf. omjer *srednjih* 1.31, te je veća šansa da stari klijenti budu *srednji*, a ne *loši*.

Tablica 16: Varijabla ISKUSTVO

Kateg.	Dobri	Srednji	Loši	%D	%S	%L	i.omjer D	i.omjer S
Novi	1410	375	143	92.40%	79.79%	84.62%	1.09	0.94
Stari	116	95	26	7.60%	20.21%	15.38%	0.49	1.31

Provedbom ANOVA testa dobivena $p-$vrijednost je 5.77×10^{-12} te odbacujemo nul-hipotezu i pretpostavljamo da je ova varijabla prediktivna. Razlika rezidula je 47.41. AIC vrijednost modela s varijablom ISKUSTVO je znatno manja od AIC vrijednosti nulmodela.
Prema navedenim rezultatima naslućujemo da ova varijabla dobro razlikuje kategorije ovisne varijable i pokušat ćemo je uključiti u skoring model.

XVII. Kvaliteta

Prema kvaliteti dućana u kojima klijenti kupuju robu za koju traže zajam, klijenti su podjeljeni u 6 kategorija. Najveći broj klijenata kupuje u dućanima koji su nešto ispod prosjeka, 52.56% *dobrih*, 38.72% *srednjih* i 55.62% *loših*. Najmanje je klijenata koji kupuju robu u najboljih 10% i najlošijih 10% dućana.
Prema inf. omjerima, najmanje rizična se pokazuje kategorija klijenata koji kupuju u 10% najboljih dućana, a najrizičnija je skupina klijenata koji kupuju u 10% najlošijih dućana.

Tablica 17: Varijabla KVALITETA

Kateg.	Dobri	Srednji	Loši	%D	%S	%L	i.omjer D	i.omjer S
A	24	12	1	1.57%	2.55%	0.59%	2.66	4.31
B	524	176	38	34.34%	37.45%	22.49%	1.53	1.67
C	802	182	94	52.56%	38.72%	55.62%	0.94	0.70
D	28	14	8	1.83%	2.98%	4.73%	0.39	0.63
E	78	49	15	5.11%	10.43%	8.88%	0.58	1.17
F	70	37	13	4.59%	7.87%	7.69%	0.60	1.02

ANOVA testom dobivena je $p-$vrijednost 3.07×10^{-6} na temelju čega pretpostavljamo da ova varijabla dobro razlikuje kategorije ovisne varijable. Razlika reziduala je 33.44, a AIC vrijednost modela s varijablom KVALITETA je manja od AIC vrijednosti nulmodela. Dakle, više je pokazatelja da ova varijabla bolje razlikuje *dobre*, *srednje* i *loše* klijente od nulmodela i možemo je pokušati uključiti u model.

4.4 Skoring model

Sukladno analizi iz prethodnog poglavlja odabiremo 14 prediktivnih varijabli koje ćemo pokušati uključiti u model: DOB, STAŽ, BRAČNO, CIJENA, GOTOVINA/CIJENA, STAN, DOHODAK, CIJENA/DOHODAK, RATA, RATA/CIJENA, ZANIMANJE, ROBA, ISKUSTVO i KVALITETA. Koristeći navedene varijable kreirana su 3 modela. Prvi model sadrži svih 14 varijabli, drugi njih 7: DOB, STAŽ, CIJENA, STAN, RATA, ISKUSTVO i KVALITETA, a treći model odabranih 6 varijabli: DOB, STAŽ, CIJENA, STAN, ISKUSTVO i KVALITETA.
Dakle, počeli smo sa modelom koji sadrži sve varijable koje su se pokazale značajne u prethodnom poglavlju. Zatim smo izbacujući po jednu varijablu kreirali 14 modela s 13 varijabli i ANOVA analizom proveli usporedbu modela sa 14 varijabli sa svakim od kreiranih modela koji nije sadržavao po jednu varijablu iz modela s 14 varijabli. Ukoliko je za neki model ANOVA analizom dobivena p-vrijednost manja od 0.05, znači da je model s tom varijablom bolji od onoga koji ju ne sadrži. Na taj način 7 varijabli se pokazalo značajnim i s njima smo kreirali model 2.
Isti postupak smo napravili s modelom 2, i samo model bez varijable RATA se na razini značajnosti 5% pokazao bolji od modela sa svih 7 varijabli. Model 3 kreiran je sa 6 varijabli iz modela 2 bez varijable RATA.
U tablici (18) prikazana je ANOVA analiza navedena 3 modela. Model 3 je model s najmanje neovisnih varijabli, model 2 je model sa 7 varijabli, a model 1 sa svih 14 neovisnih varijabli.

Tablica 18: Usporedba modela

	br. parametara	AIC	logLik	LR stat	Test	df	Pr(Chi)
Model 3	21	3258	-1608				
Model 2	22	3256	-1606	3.82	Model 3 vs Model 2	1	0.0508
Model 1	40	3274	-1597	21.85	Model 3 vs Model 1	19	0.2919

Usporedbom modela 3 s modelom 2 dobivena je $p-$vrijednost 0.0508 na temelju čega na razini značajnosti 5% ne odbacujemo nul-hipotezu o jednakim očekivanjima dva modela, te pretpostavljamo da je model s manje varijabli bolji. Kod usporedbe modela 3 s modelom 1, dobivena je $p-$vrijednost 0.2919 na temelju čega ne odbacujemo nul-hipotezu o jednakim očekivanjima modela 1 i 3. Sukladno dobivenim rezultatima možemo zaključiti da je model 3 najbolji izbor, te ćemo ga u nastavku detaljnije analizirati.

4.4.1 Analiza modela

Model je kreiran funkcijom *clm* (cumulative logistic model) iz paketa *ordinal* u programskom jeziku R. Funkcija *clm* vrši procjenu parametara metodom maksimalne vjerodostojnosti, računa standardnu grešku, vrijednosti Waldove statistike za pojedni procjenjeni parametar, te odgovarajuću p vrijednost. Dobiveni rezultati za model 3 su u tablici (19).

Tablica 19: Procjena parametara

	parametar	st. greška	z vrijednost	p-vrijednost
dob	0.0170	0.0050	3.407	0.000657
staž1	0.0063	0.1481	0.043	0.965989
staž2	0.3001	0.1651	1.818	0.069050
staž3	0.5939	1.9310	3.076	0.002101
staž4	0.4033	0.2054	1.963	0.049618
staž5	0.2307	0.2598	0.888	0.374517
staž6	0.5148	0.1825	2.821	0.004787
cijena	-4.84e-05	1.51e-05	-3.217	0.001296
stan2	-0.3187	0.2742	-1.162	0.245089
stan3	0.1278	0.1741	0.734	0.462842
stan4	-0.1203	0.1978	-0.608	0.543285
stan5	0.0990	0.1897	0.522	0.601840
stan6	-0.4420	0.2196	-2.022	0.043133
iskus2	-0.9733	0.1374	-7.082	1.42e-12
kval2	0.2656	0.3516	0.756	0.449938
kval3	0.4301	0.3495	1.231	0.218372
kval4	-0.2723	0.4441	-0.613	0.539756
kval5	-0.2798	0.3801	-0.736	0.461610
kval6	-0.5623	0.3888	-0.145	0.885019
0|1	0.1098	0.4474	0.245	
1|2	1.7917	0.4512	3.971	

Za kategoričke varijable, za svaku kategoriju neovisnih varijabli je procjenjen parametar, osim za prvu koja je referentna i procjenitelj za nju je sadržan u slobodnom članu. Za kontinurane varijable procjenjen je po jedan parametar.

Uvrštavanjem odgovarajućih parametara β u jednadžbu (3.11) dobivamo vjerojatnost da pojedini klijent bude *dobar*:

$$P(Y_i \leq 0) = \frac{e^{\alpha_1 + \beta_{[dob]} x_{i1} + \cdots + \beta_{[kval]} x_{i6}}}{1 + e^{\alpha_1 + \beta_{[dob]} x_{i1} + \cdots + \beta_{[kval]} x_{i6}}}. \tag{4.1}$$

odnosno, vjerojatnost da bude *dobar* ili *srednji*:

$$P(Y_i \leq 1) = \frac{e^{\alpha_2 + \beta_{[dob]} x_{i1} + \cdots + \beta_{[kval]} x_{i6}}}{1 + e^{\alpha_2 + \beta_{[dob]} x_{i1} + \cdots + \beta_{[kval]} x_{i6}}}. \tag{4.2}$$

Analizirajući svaki parametar modela 3 možemo zaključiti kako će se kretati vjerojatnosti u ovisnosti o vrijednostima neovisnih varijabli.

- Procjenjeni parametar za varijablu DOB je 0.017. Vrijednost parametra je veća od nule, te možemo zaključiti da će povećanje dobi klijenta utjecati na porast vjerojatnosti (4.1) i (4.2). Dakle, stariji klijenti će imati veću šansu upasti u kategoriju *dobrih* ili u kategoriju *srednjih*. Dok će se smanjenjem dobi klijenta povećavati šansa da klijent bude *loš*.

- Svi procjenjeni parametri za varijablu STAŽ su veći od nule. Obzirom na referentnu kategoriju, a to su klijenti koji nisu dali podatak o stažu, svi ostali imaju veću šansu upasti

u kategoriju *dobrih* ili u kategoriju *srednjih* nego klijenti koji nisu dali podatak o stažu. Najveću šansu biti *dobri* imaju klijenti sa stažem od 49 – 75 mjeseci, za koje je parametar jednak 0.5939, te klijenti sa stažem većim od 120 mjeseci za koje je procjenjeni parametar jednak 0.5148. Nakon klijenata koji nisu dali podatak o stažu, najmanju šansu biti *dobri* ili biti *srednji* imaju klijenti sa stažem manjim od 24 mjeseca, parametar za njih je 0.0063, te klijenti sa stažem od 57 – 120 mjeseci s parametrom 0.2307.

- Za kontinuranu varijablu CIJENA procjenjeni parametar je manji od nule. Povećanje cijene robe koju klijent kupuje utjecat će na smanjenje vjerojatnosti (4.1) i (4.2). Možemo zaključiti da se povećanjem cijene robe povećava rizičnost za otplatu kredita.

- Za varijablu STAN referentna kategorija je unajmljen stan. Klijenti koji imaju stanarsko pravo imaju najmanju šansu upasti u neku od viših kategorija ovisne varijable, procjenjen parametar za njih iznosi −0.3187. Klijenti koji žive u vlastitom stanu ne pokazuju se bolji od klijenata koji žive u unajmljenom stanu, parametar za njih je jednak −0.1203. Najveću šansu da budu procjenjeni kao *dobri* ili *srednji* imaju klijenti koji žive u društvenom stanu, parametar za njih je 0.1278, zatim slijede klijenti koji žive s roditeljima, procjenjeni parametar je 0.099. Klijenti iz kategorije *ostalo* pokazuju se najlošiji.

- Varijabla ISKUSTVO ima dvije kategorije. Referentna kategorija su novi klijenti, a procjenjen parametar za stare je −0.9733. Prema tom rezultatu stari klijetni se pokazuju rizičniji pri otplati kredita, odnosno imaju manju šansu od novih klijenata upasti u neku od viših kategorija ovisne varijable.

- Referentna kategorija varijable KVALITETA su klijenti koji kupuju u dućnima A kvalitete (10% najboljih dućana). Klijenti koji se pokazuju bolji od klijenata iz referentne kategorije su klijenti koji kupuju u dućanima kvalitete B i C, procjenjeni parametri za njih su 0.2656 i 0.4301 redom. Klijenti koji kupuju u preostalim kategorijama dućana imaju manju šansu biti *dobri* ili *srednji* od klijenata koji kupuju u dućanima *A* kvalitete. Najrizičniji se pokazuju klijenti koji kupuju u dućanima kvalitete F (mali dućni koji sklapaju do 6 ugovora mjesečno), procjenjen parametar za njih je −0.5623.

Uvrštavajući procjenjene parametre u jednadžbe (4.1) i (4.2) za svakog klijenta dobivamo vjerojatnosti $P(Y \leq 0)$ i $P(Y \leq 1)$. Uvedimo oznake $p_0 = P(Y = 0)$, $p_1 = P(Y = 1)$ i $p_2 = P(Y = 2)$. Obzirom da je Y diskretna varijabla, vrijedi $P(Y \leq 0)=p_0$, što odgovara vjerojatnosti da je klijent *dobar*. Pomoću $P(Y \leq 1)$ izražavamo vjerojatnost da je klijent *srednji*, $p_1 = P(Y \leq 1) - p_0$, te vjerojatnost da je klijent *loš*, $p_2 = 1 - P(Y \leq 1)$.

U tablici (20) prikazane su procjenjene vjerojatnosti za nekoliko klijenata.

Iz dobivenih vjerojatnosti želimo zaključiti kojoj kategoriji ovisne varijable klijent pripada. Velika većina klijenata ima najveću vjerojatnost da bude *dobra*, a najmanju da bude *loša* što je

Tablica 20: Procjenjene vjerojatnosti

	p_0	p_1	p_2
1	0.7059	0.2222	0.0719
2	0.4359	0.3701	0.1940
3	0.7304	0.2054	0.0643
4	0.5660	0.3091	0.1249
5	0.6661	0.2486	0.0853
⋮	⋮	⋮	⋮
2164	0.7564	0.1871	0.0565
2165	0.4707	0.3563	0.1730

posljedica uzorka na kojem smo kreirali model koji sadrži 70% *dobrih* klijenata. Bilo bi prirodno gledati maksimalnu od tri dobivene vjerojatnosti i na temelju toga zaključiti u koju kategoriju klijenta svrstati. Uzimajući taj kriterij, 2 klijenta bi bila procjenjena kao *loša*, 33 kao *srednji*, a preostalih 2130 bi bili *dobri*. Od toga niti jedan *loš* klijent nije točno procjenjen. Točno je procjenjeno 3.40% *srednjih* i 99.34% *dobrih*.
Cilj skoring modela je prije svega detektirati *loše* klijente koji bi banci mogli prouzročiti probleme, što nam ovaj kriterij ne omogućava. Ukupna stopa pogodaka (*engl. hit rate*) je 70.76%, no to su uglavnom točno procjenjeni *dobri* klijenti, dok *srednji* i *loši* gotovo uopće nisu prepoznati.

Definirajmo greške tipa *I* i *II* koje smo u ovom radu podjelili na *male* i *velike*. Velika greška tipa *I* je ukoliko je klijent koji je stvarno *loš* procjenjen kao *dobar*, a mala greška je ukoliko je klijent procjenjen za jednu kategoriju više. Slično za grešku tipa *II*, ukoliko je stvarno *dobar* klijent procjenjen kao *loš* smatrat ćemo to velikom greškom, a ukoliko je procjenjen za kategoriju niže nego što stvarno jeste reći ćemo da se radi o maloj greški tipa *II*.

Zbog velikih grešaka tipa *I*, pokušat ćemo pronaći adekvatne *pragove (engl. cut-off* ili *cut points)* kojima ćemo minimizirati greške, i istovremeno povećati stope pogodaka *srednjih* i *loših* klijenata.

Prag je proizvoljna vjerojatnost unutar jedne od kategorija. Ukoliko je vjerojatnost da klijent pripada toj kategoriji veća od odabranog praga, smatrat ćemo da klijent pripada toj kategoriji. Prvi prag uzet je za *loše* klijente. Ukoliko je p_2 veća od odabranog praga, klijenta ćemo svrstati u kategoriju *loših*. U suprotnom biramo prag za kategoriju *srednjih*. Ukoliko je p_1 veća od odabranog drugog praga, klijenta svrstavamo u *srednju* kategoriju, a ukoliko niti taj uvijet nije zadovoljen, klijent će biti procjenjen kao *dobar*.

U tablici (21) prikazani su uređeni parovi pragova, greške i stope pogodaka za svaki uređeni par. Krenuli smo s pragom 0.09 za *loš*e klijente jer nema razloga gledati manje od njega obzirom da 29.42% klijenata ima p_2 veći od 0.09, a znamo da je stvarno *loših* klijenata iz uzorka tek 7.81%. Još većim smanjenjem tog praga povećala bi se stopa pogodaka *loših*, ali bi se povećale greške tipa *II* i smanjile stope pogodaka za preostale dvije kategorije.

Tablica 21: Točnost procjene multinomnog modela za različite parove pragova

Pragovi		Greške Tipa I		Greške Tipa II		Stope pogodaka			
loši	srednji	mala	velika	mala	velika	dobri	srednji	loši	ukupno
0.09	0.20	33.96%	22.49%	28.46%	21.69%	53.47%	26.17%	42.60%	46.70%
0.10	0.20	36.15%	22.49%	31.41%	15.66%	53.47%	33.19%	34.32%	47.58%
0.11	0.20	37.56%	22.49%	33.72%	11.34%	53.47%	37.45%	28.99%	48.08%
0.12	0.20	38.34%	22.49%	34.02%	9.63%	53.47%	41.70%	26.04%	48.78%
0.13	0.20	39.28%	22.49%	34.67%	7.54%	53.47%	45.74%	22.49%	49.38%
0.14	0.20	40.06%	22.49%	34.92%	5.96%	53.47%	49.79%	19.53%	50.02%
0.15	0.20	40.85%	22.49%	35.12%	4.91%	53.47%	52.34%	16.57%	50.35%
0.09	0.21	35.37%	29.59%	24.60%	21.69%	58.52%	21.70%	42.60%	49.28%
0.10	0.21	37.56%	29.59%	27.56%	15.66%	58.52%	28.72%	34.32%	50.16%
0.11	0.21	38.97%	29.59%	29.86%	11.34%	58.52%	32.98%	28.99%	50.67%
0.12	0.21	39.75%	29.59%	30.16%	9.63%	58.52%	37.23%	26.04%	51.36%
0.13	0.21	40.69%	29.59%	30.81%	7.54%	58.52%	41.28%	22.49%	51.96%
0.14	0.21	41.47%	29.59%	31.06%	5.96%	58.52%	45.32%	19.53%	52.61%
0.15	0.21	42.25%	29.59%	31.26%	4.91%	58.52%	47.87%	16.57%	52.93%
0.09	0.22	36.15%	35.50%	21.29%	21.69%	62.84%	18.51%	42.60%	51.64%
0.10	0.22	38.34%	35.50%	24.25%	15.66%	62.84%	25.53%	34.32%	52.52%
0.11	0.22	39.75%	35.50%	26.55%	11.34%	62.84%	29.79%	28.99%	53.03%
0.12	0.22	40.53%	35.50%	26.85%	9.63%	62.84%	34.04%	26.04%	53.72%
0.13	0.22	41.47%	35.50%	27.51%	7.54%	62.84%	38.09%	22.49%	54.32%
0.14	0.22	42.25%	35.50%	27.76%	5.96%	62.84%	42.13%	19.53%	54.97%
0.15	0.22	43.04%	35.50%	27.96%	4.91%	62.84%	44.68%	16.57%	55.29%
0.09	0.23	39.12%	40.24%	17.59%	21.69%	67.69%	12.77%	42.60%	53.81%
0.10	0.23	41.31%	40.24%	20.54%	15.66%	67.69%	19.79%	34.32%	54.69%
0.11	0.23	42.72%	40.24%	22.85%	11.34%	67.69%	24.04%	28.99%	55.20%
0.12	0.23	43.51%	40.24%	23.15%	9.63%	67.69%	28.30%	26.04%	55.89%
0.13	0.23	44.44%	40.24%	23.80%	7.54%	67.69%	32.34%	22.49%	56.49%
0.14	0.23	45.23%	40.24%	24.05%	5.96%	67.69%	36.38%	19.53%	57.14%
0.15	0.23	46.01%	40.24%	24.25%	4.91%	67.69%	38.94%	16.57%	57.46%
0.09	0.24	41.78%	43.20%	14.43%	21.69%	71.82%	8.09%	42.60%	55.70%
0.10	0.24	43.97%	43.20%	17.38%	15.66%	71.82%	15.11%	34.32%	56.58%
0.11	0.24	45.38%	43.20%	19.69%	11.34%	71.82%	19.36%	28.99%	57.09%
0.12	0.24	46.17%	43.20%	19.99%	9.63%	71.82%	23.62%	26.04%	57.78%
0.13	0.24	47.10%	43.20%	20.64%	7.54%	71.82%	27.66%	22.49%	58.38%
0.14	0.24	47.89%	43.20%	20.89%	5.96%	71.82%	31.70%	19.53%	59.03%
0.15	0.24	48.67%	43.20%	21.09%	4.91%	71.82%	34.26%	16.57%	59.35%
0.09	0.25	43.35%	49.11%	11.77%	21.69%	75.29%	3.83%	42.60%	57.23%
0.10	0.25	45.54%	49.11%	14.73%	15.66%	75.29%	10.85%	34.32%	58.11%
0.11	0.25	46.95%	49.11%	17.03%	11.34%	75.29%	15.11%	28.99%	58.61%
0.12	0.25	47.73%	49.11%	17.33%	9.63%	75.29%	19.36%	26.04%	59.31%
0.13	0.25	48.67%	49.11%	17.99%	7.54%	75.29%	23.40%	22.49%	59.91%
0.14	0.25	49.45%	49.11%	18.24%	5.96%	75.29%	27.45%	19.53%	60.55%
0.15	0.25	50.23%	49.11%	18.44%	4.91%	75.29%	30.00%	16.57%	60.88%

Povećanjem spomenutog praga stopa pogodaka *loših* i velika greška tipa *II* opadaju, stopa pogodaka *srednjih* raste, a na stopu pogodaka *dobrih* nema utjecaj, kao niti na veliku grešku tipa *I*. Povećanjem prvog praga do 0.15 znatno se smanjila stopa pogodaka *loših* klijenata. Samo

16.57% *loših* je točno procjenjeno, stoga nismo promatrali rezultate koji bi se realizirali još većim porastom ovog praga.

Za prag *srednje* kategorije prvo je uzeta vrijednost 0.20 iz razloga što svi klijenti koji imaju p_1 manju od 0.20, imaju p_0 vjerojatnosti veće od 0.74, što je dovoljno veliko da takve klijente svrstamo u kategoriju *dobrih*.
Povećanje praga za *srednje* klijente nema utjecaj na stopu pogodaka *loših* niti na veliku grešku tipa II, dok velika greška tipa I raste, stopa pogodaka *srednjih* opada, a *dobrih* raste. Povećanjem drugog praga do 0.25 znatno se povećala greška tipa I, posebno velika greška do gotovo 50%, što znači da bi gotovo 50% *loših* klijenata bilo procjenjeno kao *dobri*. Iz tog razloga nismo promatrali daljnji rast drugog praga.

Maksimalna stopa pogodaka *loših* postignuta je za sve uređene parove kojima je prag *loših* 0.09 i iznosi 42.60%. *Srednji* klijenti najbolje su prepoznati za uređeni par pragova $(0.15, 0.20)$ za koje stopa pogodaka *srednjih* iznosi 52.34%.

Koristeći rezultate iz tablice (21) potrebno je odabrati uređene parove pragova pomoću kojih bi došli do najtočnije procjene ponašanja klijenta pri otplati kredita.
Važni kriteriji za odabir su što veće stope pogodaka *dobrih*, *srednjih* i *loših*, zatim što manje greške tipa I i II. U našem slučaju stopa pogodaka *dobrih* je u većini slučajeva velika i povlači za sobom ukupnu stopu pogodaka. Iz tog razloga nam kriterij za točno procjenjene *dobre* klijente neće biti od presudne važnosti.

Korišteni kriteriji za odabir pragova su sljedeći:

- Stopa pogodaka *loših* veća od 28%
- Stopa pogodaka *srednjih* veća od 20%
- Velika greška tipa I manja od 30%

Navedene kriterije smo birali u skladu s dobivenim razultatima u tablici (21). Stopu pogodaka *loših* smatramo najvažnijim kriterijem i poželjno je da je što veća, no povećavajući je smanjuje se stopa pogodaka *srednjih*. Velika greška tipa II je za sve uređene parove pragova relativno mala te ju nismo uključili u kriterije za odabir. Velika greška tipa I sljedeći je važan kriterij, te smo u obzir uzeli dvije najmanje vrijednosti. Male greške nećemo promatrati.

Prema navedenom kriteriju dobili smo 4 para pragova: $(0.09, 0.20)$, $(0.10, 0.20)$, $(0.09, 0.21)$, $(0.10, 0.21)$.
Za navedene pragove ćemo provesti validaciju na uzorku klijenata čiji podaci nisu korišteni za kreiranje modela.

4.4.2 Validacija modela

U ovom potpoglavlju ćemo provesti validaciju prethodno analiziranog modela na 1000 klijentana koji nisu korišteni za izradu modela. Validacijski uzorak sastoji se od 767 *dobrih*, 203 *srednja* i 30 *loših* klijenata.

Uvrstimo li u jednadžbe (4.1) i (4.2) odgovarajuće procjenjene parametre iz tablice (19) za svakog od 1000 klijenata, možemo izraziti vjerojatnost za svakog klijenta da bude *dobar*, *srednji* i *loš*. Nakon što smo dobili procjenjene vjerojatnosti, svrstavamo ih u odgovarajuće kategorije ovisne varijable prema odabranim pragovima.
U prethodnom potpoglavlju odabrali smo 4 uređena para pragova koje smatramo najoptimalnijima. U tablici (22) prikazani su rezultati procjene za svaki od odabranih parova pragova.

Tablica 22: Točnost procjene multinomnog modela na validacijskom uzorku

Pragovi		Greške Tipa I		Greške Tipa II		Stope pogodaka			
loši	srednji	mala	velika	mala	velika	dobri	srednji	loši	Ukupno
0.09	0.20	38.63%	26.67%	28.56%	27.25%	44.98%	30.05%	33.33%	41.60%
0.10	0.20	41.20%	26.67%	30.41%	23.08%	44.98%	36.95%	13.33%	42.40%
0.09	0.21	45.49%	33.33%	24.33%	27.25%	50.33%	21.18%	33.33%	43.90%
0.10	0.21	48.07%	33.33%	26.19%	23.08%	50.33%	28.08%	13.33%	44.70%

Možemo primjetiti da je ukupna stopa pogodaka manja za sve parove pragova nego za iste pragove za uzorak na kojem je kreiran model. Nadalje, greške su veće, stope pogodaka *dobrih* i *loših* su manje za sve pragove, dok za *srednje* klijente stopa pogodaka ne odstupa više od 4% niti za jedan par.

Niti za jedan prag ne možemo reći da je ovaj model dobar. Ukupne stope pogodaka ne dostižu niti 50%, a *srednje* i *loše* klijente model gotovo ne prepoznaje.

Promotrimo još *matricu konfuzije* za jedan par, npr. za $(0.09, 0.21)$.

Tablica 23: Matrica konfuzije za uređeni par pragova $(0.09, 0.21)$

Procjena	Stvarno 0	Stvarno 1	Stvarno 2
0	386	96	10
1	172	43	10
2	209	64	10

Prema tablici (23) veliki broj *dobrih* klijenata je procjenjen kao *loš*, što banci ne bi prouzročilo gubitak, ali bi propustila priliku za zaradu. Veći problem čine *loši* klijenti koji bi bili procjenjeni kao *dobri* te bi banci prouzročili gubitke. Za naš model to je trećina *loših* klijenata. Najveći dio *srednjih* je procjenjen kao *dobar*, a najmanje ih je točno procjenjeno.

Možemo zaključiti da ovim modelom nismo dobili željene rezultate kojima bi mogli predvidjeti ponašanje klijenata pri otplati kredita. Iz tog razloga u idućem poglavlju ćemo pokušati kreirati binomni model bez *srednjih* klijenata kojim bi dobili bolje rezultate procjene.

4.5 Binomni model

Na isti način kao i za multinomni model proveli smo analizu varijabli. Kao prediktivne, na razini značajnosti 5%, se pokazalo sljedećih 7 varijabli: DOB, STAŽ, BRAČNO, STAN, ROBA, ISKUSTVO i KVALITETA. Vidi tablicu (24).

Tablica 24: Značajnosti neovisnih varijabli za binomni model

R. Broj	Model	Resid df	Resid. Dev	Test	Df	Deviance	$Pr(>\|Chi\|)$
1	Nulmodel	1694	1099.80				
2	dob	1693	1072.00	1 vs 2	1	27.80	1.34e-07
3	zanimanje	1692	1098.50	1 vs 3	2	1.38	0.5024
4	bračno	1691	1076.20	1 vs 4	3	23.63	2.98e-05
5	cijena/dohodak	1690	10954.40	1 vs 5	4	5.45	0.2444
6	cijena	1693	1096.80	1 vs 6	1	3.01	0.0825
7	dohodak	1693	1099.80	1 vs 7	1	0.02	0.8790
8	gotovina/cijena	1693	1096.40	1 vs 8	1	3.41	0.0650
9	gotovina	1693	1099.80	1 vs 9	1	0.00	0.9664
10	iskustvo	1693	1089.80	1 vs 10	1	10.08	0.0015
11	kvaliteta	1689	1080.50	1 vs 11	5	19.32	0.0017
12	način	1693	1096.40	1 vs 12	1	3.43	0.0640
13	rata/cijena	1693	1099.60	1 vs 13	1	0.26	0.6071
14	rata	1693	1099.70	1 vs 14	1	0.12	0.7257
15	roba	1688	1081.20	1 vs 15	6	18.66	0.0048
16	stan	1689	1061.70	1 vs 16	5	38.14	3.54e-07
17	staž	1688	1057.80	1 vs 17	6	41.98	1.86e-07

Tablica 25: Usporedba dva binomna modela

	Resid. Df	Resid. Dev	Df	Devinace	$P(>\|Chi\|)$
Model 5 var	1676	999.54			
Model 7 var	1667	986.08	9	13.466	0.1426

Kreirajući binomne modele s navedenim varijablama, te provedbom ANOVA analize, model sa sljedećim varijablama se pokazao kao najbolji: DOB, STAŽ, STAN, ISKUSTVO i KVALITETA. ANOVA usporedba navedenog modela i modela sa svih 7 varijabli je u tablici (25). Procjenjeni parametri modela kreiranim s navedenih 5 varijabli su u tablici (26).

Uvrstimo li u jednadžbu (2.9) procjenjene parametre i odgovarajuće vrijednosti neovisnih varijabli, za svakog klijenta dobivamo vjerojatnost da bude *dobar*. Ukoliko je pojedini procjenjeni

parametar veći od nule, to će povećati šansu da klijent bude *dobar*.

Tablica 26: Procjena parametara binomnog modela

	parametar	st. greška	z vrijednost	p-vrijednost
dob	0.0211	0.0091	2.320	0.020318
staž_k1	-0.3292	0.2595	-1.269	0.204519
staž_k2	0.0861	0.2928	0.294	0.768730
staž_k3	0.7806	0.4020	1.942	0.052189
staž_k4	0.9949	0.5041	1.974	0.048423
staž_k5	-0.1936	0.4192	-0.462	0.644176
staž_k6	0.8973	0.4124	2.176	0.029581
stan2	-0.3882	0.4185	-0.927	0.353704
stan3	0.5175	0.2899	1.785	0.074212
stan4	-0.0677	0.3051	-0.222	0.824383
stan5	0.3929	0.3182	1.235	0.216951
stan6	-0.4208	0.3425	-1.229	0.219251
iskus2	-0.8347	0.2498	-3.342	0.000833
kval2	- 0.4703	1.0511	-0.447	0.654548
kval3	-0.8067	1.0429	-0.774	0.439196
kval4	-1.6807	1.1199	-1.501	0.133415
kval5	-1.3842	1.0785	-1.283	0.199336
kval6	-1.3339	1.0836	-1.231	0.218341
(Intercept)	2.0602	1.1405	1.806	0.070846

- Procjenjeni parametar za varijablu DOB je 0.0211, te možemo zaključiti da povećanjem dobi raste šansa da klijent bude *dobar*.

- Za varijablu STAŽ, referentna kategorija su klijenti koji nisu dali podatak o stažu. Klijenti sa stažem manjim od 24 mjeseca, te klijenti sa stažem od 97 – 120 mjeseci imaju manju šansu biti *dobri* nego oni koji nisu dali podatak o stažu. Najveću šansu biti *dobri* imaju klijenti sa stažem od 73 – 96 mjeseci, te klijenti sa stažem većim od 120 mjeseci.

- Referentna kategorija varijable STAN su klijenti s unajmljenim stanom. Obzirom na njih, najveću šansu biti *dobri* imaju klijenti koji žive u društvenom stanu, te klijenti koji žive s roditeljima. Najlošiji se pokazuju klijenti iz kategorije *ostalo*, te oni koji imaju stanarsko pravo.

- Procjenjeni parametar za varijablu ISKUSTVO za kategoriju starih klijenata je manji od nule. Dakle, novi klijenti se pokazuju bolji od klijenata s kojima je banka već poslovala.

- Referentna kategorija varijable KVALITETA su klijenti koji kupuju u najboljih 10% dućana. Procjenjeni parametri su za sve ostale kategorije manji od nule, što znači da se klijenti iz referentne kategorije pokazuju najbolji pri otplati kredita. Najlošiji klijenti, s najmanjim procjenjenim parametrima, su iz najlošije tri kategorije dućana.

Uzorak sadrži puno veći broj *dobrih*, njih 1526, a samo 169 *loših*. Zbog strukture uzorka, kao i kod multinomnog modela, vjerojatnosti da je klijent *dobar* su gotovo za sve klijente veće od vjerojatnosti da je klijent *loš*. Iz tog razloga, promotrit ćemo rezultate za nekoliko pragova s ciljem pronalaska najoptimalnijeg kojim bi povećali stopu pogodaka *loših*. U tablici (27) prikazani su rezultati za neke pragove. Obzirom na spomenutu strukturu uzorka u kojoj je samo 10% klijenata u kategoriji *loših*, promotrili smo pragove od 0.08 nadalje. Još većim smanjenjem pada ukupna stopa pogodaka, ali raste stopa pogodaka *loših*. Povećanjem praga od 0.08 do 0.25, ukupna stopa pogodaka, stopa pogodaka *dobrih* i greška tipa I rastu, dok stopa pogodaka *loših* i greška tipa II padaju.

Tablica 27: Točnost procjene binomnog modela za različite pragove

Prag	Stopa pogodaka dobrih	Greška tipa I	Greška tipa II	Stopa pogodaka loših	Ukupna stopa pogodaka
0.08	56.09 %	21.89%	43.91%	78.11 %	58.29%
0.09	61.21 %	28.40%	38.79%	71.60 %	62.24 %
0.10	65.73 %	33.14 %	34.27 %	66.86 %	65.84 %
0.11	69.53 %	34.91 %	30.47%	65.09 %	69.09 %
0.12	73.13 %	39.05 %	26.87%	60.95 %	71.92 %
0.13	75.69%	43.79%	24.31%	56.21 %	73.75%
0.14	78.44%	50.30 %	21.56%	49.70 %	75.58 %
0.15	81.06%	53.85%	18.94%	46.15%	77.58%
0.16	83.88%	57.99%	16.12%	42.01%	79.71%
0.17	86.11%	62.72%	13.89%	37.28%	81.24%
0.18	88.01%	64.50%	11.99%	35.50%	82.77%
0.19	89.65%	68.05%	10.35%	31.95%	83.89%
0.20	90.83%	71.01%	9.17%	28.99%	84.66%
0.21	92.66%	76.92%	7.34%	23.08%	85.72%
0.22	94.36%	80.47%	5.64%	19.53%	86.90%
0.23	95.22%	82.84%	4.78%	17.16%	87.43%
0.24	95.74%	83.43%	4.26%	16.57%	87.85%
0.25	95.94%	84.62%	4.06%	15.38%	87.91%

Promotrimo li rezultate multinomnog modela iz tablice (21) i rezulate binomnog modela iz tablice (27) možemo vidjeti da je postignuta veća točnost procjene za binomni model. No, niti za binomni model ne možemo reći da je dobar zbog greške tipa *I* koja je za sve pragove binomnog modela velika, dok je za visoku stopu pogodaka *dobrih*, stopa pogodaka *loših* preniska i obrnuto.

4.5.1 Validacija binomnog modela

Za validaciju modela kreiranog u prethodnom potpoglavlju, korišteni su podaci 797 klijenata koji nisu korišteni za izradu modela, od njih je 767 *dobrih* i 30 *loših*. Za te klijente ćemo provesti validaciju za pragove koji su se pokazali najbolji za uzorak na kojem je kreiran model. Uzmimo pragove od $0.10 - 0.20$ jer se za one manje od 0.10 ukupna stopa pogodaka *dobrih* pokazala preniska, a za one iznad 0.20 je stopa pogodaka *loših* preniska.

Navedeni rezultati su u tablici (28).

Tablica 28: Točnost procjene za binomni model na validacijskom uzorku

Prag	Stopa pogodaka dobrih	Greška tipa I	Greška tipa II	Stopa pogodaka loših	Ukupna stopa pogodaka
0.10	68.71%	36.67%	31.29%	63.33%	68.51%
0.11	71.97%	40.00%	28.03%	60.00%	71.52%
0.12	74.71%	46.67%	25.29%	53.33%	73.90%
0.13	78.36%	46.67%	21.64%	53.33%	77.42%
0.14	80.44%	53.33%	19.56%	46.67%	79.17%
0.15	82.40%	66.67%	17.60%	33.33%	80.55%
0.16	83.96%	73.33%	16.04%	26.67%	81.81%
0.17	87.35%	76.67%	12.65%	23.33%	84.94%
0.18	88.40%	86.67%	11.60%	13.33%	85.57%
0.19	89.57%	86.67%	10.43%	13.33%	86.70%
0.20	89.83%	86.67%	10.17%	13.33%	86.95%

Obzirom da je velik udio *dobrih* klijenata u uzorku na kojem su kreirani modeli, točnost procjene *dobrih* uvelike utječe na ukupnu stopu pogodaka, te je za visoke ukupne stope pogodaka nedovoljna točnost procjene *loših* klijenata. Npr. za ukupnu stopu pogodaka 80%, koju bi mogli uzeti kao donju granicu kvalitete modela, samo je trećina *loših* klijenata točno procjenjena.

Usporedimo li rezultate na validacijskom uzorku binomnog modela (tablica 28) i rezultate multinomnog modela (tablica 22), možemo zaključiti da je točnost procjene kvalitete klijenata bolja za binomni model. Najbolja 'raspodjela' stope pogodaka *dobrih* i *loših* bi bila za pragove 0.10 i 0.11 binomnog modela. No, ne možemo reći ni za kreirani binomni model da dobro razlikuje *dobre* i *loše* klijente.

5 Zaključak

Uvodeći *srednje* klijente u model kreditnog skoringa htjeli samo dobiti širu sliku klijenata neke banke. Odnosno, dobiti precizniju procjenu kvalitete pojedinog klijenta u smislu da ih ne tretiramo samo kao *dobre* ili *loše*. S ciljem svrstavanja klijenta u jednu od 3 kategorije, *dobre*, *srednje* ili *loše*, kreirana su 3 modela ordinalne multinomne logističke regresije sa različitim brojem neovisnih varijabli. Za modele smo proveli analizu kvalitete na temelju koje se model s najmanje neovisnih varijabli pokazao najboljim, te je on i iznesen u radu.

Obzirom da uzorak na kojem je kreiran model sadrži veliki broj *dobrih* klijenata, njih više od 70%, procjenjene vjerojatnosti su za gotovo sve klijente najveće za kategoriju *dobrih*. Iz tog razloga bilo je potrebno pronaći adekvatne pragove kojima bi razdvojili kategorije ovisne varijable, te je rezultat konačne procjene ovisio je o izboru pragova.

Dobiveni rezultati modela su i dalje bili nezadovoljavajući. *Srednji* i *loši* klijenti gotovo uopće nisu prepoznati, stopa pogodaka *dobrih* je visoka, no uz velike greške tipa *I*.

Pokušavajući pronaći bolji model, kreiran je binomni model logističke regresije bez *srednjih* klijenata. No, niti taj model nije dao zadovoljavajuće rezultate. *Loši* klijenti nisu prepoznati, dok je stopa pogodaka *dobrih* visoka, ali i dalje uz visoku grešku tipa *I*.

Očito je da struktura podataka ili uzorak kojim raspolažemo nisu adekvatni za kreiranje modela kojim bi dobili rezultate na temelju kojih bi klijenta mogli s nekom sigurnošću svrstati u jednu od kategorija ovisne varijable.

Usporedbom rezultata multinomnog i binomnog modela možemo zaključiti da binomni model daje rezultate koji su nam prihvatljiviji. Samim time što je granica između definicija *dobrih* i *loših* veća, binomni model ih lakše razlikuje, što nije slučaj kod multinomnog modela gdje se definicije ovisnih varijabli razlikuju u svega jednom danu.

Također postoji mogući problem u interpretaciji neovisnih varijabli zbog moguće multikolinearnosti nekih od njih, odnosno moguća je visoka koreliranost dvije ovisne varijable (recimo > 0.8). Tada procjenjeni koeficijenti za te varijable mogu biti neprecizni (s velikim standardnim greškama). U tom slučaju potrebno je promotriti izbacivanje jedne od njih iz analize ili ih zamjeniti trećom varijablom koja bi bila kombinacija te dvije. Primjer varijabli za koje postoji rizik multikolinearnosti su npr. RATA/CIJENA i RATA, CIJENA/DOHODAK i DOHODAK... No, nije isključivo da bilo koje dvije neovisne varijable mogu biti visoko korelirane.

Prijedlog za daljnje istraživanje je pokušati kreirati ovakve modele na uzorku koji sadrži sličan broj klijenata svake od kategorija ovisne varijable. Obzirom da se binomni model pokazuje bolji, *srednje* klijente bi mogli uključiti u model na drugačiji način. Npr. za *srednje* klijente provesti procjenu kojom bi ih svrstali u *dobre* ili *loše* binomnim modelom koji je kreiran samo s *dobrim* i *lošim* klijentima, te s dobivenim podacima, odnosno sada proširenim skupovima *dobrih* i *loših*, kreirati konačan binomni model (kao što su to napravili Bohaček, Benšić, Šarlija, 2004.).

Multinomni model se nije pokazao kao najbolje rješenje za detekciju *srednjih* klijenata niti kod prethodnih istraživanja (Bohaček, Benšić, Šarlija, 2004. i Tsai, 2012.). Korisno ih je uključiti u model da bi dobili reprezentativniji uzorak klijenata banke, no sam multinomni model logističke regresije nije dobro rješenje kojim bi ih mogli detektirati. Bilo bi interesantno pokušati s nekim drugim metodama, npr. neuralne mreže, diskriminacijska analiza, Markovljevi lanci, matematičko programiranje, genetički algoritmi ili druge metode.

Literatura

[1] A. Agresti, *Categorical Data Analysis*, University of Florida, Gainesville, Florida, 2002.

[2] A. G. Barnett, A. J. Dobson, *An introduction to generalized linear models*, CRC Press, Boca Raton, 2008.

[3] M. Benšić, Z. Bohaček, N. Šarlija, *Multinomial model in consumer credit scoring*, 10th International conference on OR - KOI 2004.

[4] M. Benšić, N. Šuvak, *Uvod u vjerojatnost i statistiku*, Sveučilište J. J. Strossmayera, Odjel za matematiku, Osijek, 2014.

[5] C. Bolton, *Logistic regression and its application in credit scoring*, University of Pretoria, 2009.

[6] R. H. B Christensen, *Analysis of ordinal data with cumulative link models - estimation with the R-package* ordinal, June 28, 2015.,
https://cran.r-project.org/web/packages/ordinal/vignettes/clm_tutorial.pdf, 25/9/2015

[7] S. A. Czepiel, *Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation*, http://www.czep.net/stat/mlelr.pdf, 12/4/2015

[8] L. Fahrmeir, G. Tutz, *Multivariate statistical modelling based on generalized linear models*, Springer, New York, 2001.

[9] G. De Rossi Ayres, W. Wei, *Credit Scoring Model Applications: Testing Multinomial Targets*, Orebro University, May 2014.

[10] N. Šarlija, *Predavanja za kolegij 'Upravljanje kreditnim rizicima'*, Odjel za matematiku, Osijek 2008.

[11] B. Tsai, *Comparison of Binary Logit Model and Multinomial Logit Model in Predicting Corporate Failure*, Review of Economics & Finance, April 2012.

Sažetak

Jedna od metoda kvantitativne kreditne analize je logistička regresija kojom se klijent svrstava u jednu od dvije kategorije, u *dobre* ili *loše*. Podaci o *srednjim* klijentima, iako postoje u bazi klijenata svake banke, uglavnom su isključeni u procesu kreditnog skoringa. Mnoge studije do sada nisu pokazale značajnost upotrebe *srednjih* klijenata pri kreiranju kredit skoring modela, uglavnom zbog toga što je slabija granica među definicijama *dobrih*, *srednjih* i *loših*, te su konačne procjene manje točne od procjena koje bi dao logistički model u kojem su *srednji* izostavljeni.

U ovom radu ispitat ćemo značajnost upotrebe *srednjih* na bazi klijenata jedne banke, na način da ćemo kreirati multinomni logistički model u kojem ovisna varijabla ima tri kategorije, *dobre*, *srednje* i *loše*, te binomni model samo s *dobrim* i *lošim* klijentima. Te ćemo usporediti rezultate.

Ključne riječi:

generalizirani linearni model, multinomna distribucija, ordinalna varijabla, metoda maksimalne vjerodostojnosti, kreditni skoring, pragovi

Abstract

One of quantitative credit analysis methods is logistic regression where the client is classified into the one of two categories, *good* or *bad*.

Although there are data of *poor* clients in the database of any bank. They are generally excluded in the process of credit scoring. Many studies so far haven't shown the importance of using *poor* clients in credit scoring model, mainly because the lower boundary between definitions of *good*, *poor* and *bad*, so the final estimates are less accurate than estimates obtained by logistic model in which *poor* are omitted.

In this article we will examine the significance of use *poor* clients of a bank, in a way that will create a multinomial logistic model in which the dependent variable has three categories, *good*, *poor* and *bad*, and the binomial model only with *good* and *bad* customers. Than the results will be compared.

Key words:

generalized linear model, multinomial distribution, ordinal covariate, maximum likelihood estimation, credit scoring, cut points

Životopis

Zana Andabaka rođena je 20. prosinca 1989. godine u Vinkovcima. Osnovnu školu pohađa u Otoku, a prirodoslovno-matematičku gimnaziju u Vinkovcima. 2008. godine upisuje preddiplomski studij matematike na Odjelu za matematiku u Osijeku. 2011. stječe status prvostupnika matematike i iste godine upisuje diplomski studij Financijske i poslovne matematike također na Odjelu za matematiku u Osijeku. Od 2015. godine zaposlena je kao analitičar u Odjelu logistike u Konzumu u Zagrebu.