Klasifikacija kreditne sposobnosti modelom logističke regresije: problem nebalansiranih kategorija

Akmačić, Dora

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Dora Akmačić

KLASIFIKACIJA KREDITNE SPOSOBNOSTI MODELOM LOGISTIčKE REGRESIJE: PROBLEM NEBALANSIRANIH KATEGORIJA

Diplomski rad

Voditelj rada:
doc. dr. sc. Petra Posedel Šimović
doc. dr. sc. Snježana Lubura Strunjak

Zagreb, Srpanj, 2024.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Hvala mentoricama doc. dr. sc. Petri Posedel Šimović i doc. dr. sc. Snježani Luburi Strunjak na svim savjetima i pomoći pri izradi ovog rada

Sadržaj

Uvod

Jedan od ključnih problema financijskih institucija je procjena kreditne sposobnosti klijenata. Dobra procjena može dovesti do smanjenja kreditnog rizika, troškova i prekomjernog zaduživanja što u konačnici doprinosi stabilnijem i održivijem financijskom sustavu. Iz tog razloga, pravilna klasifikacija klijenata od iznimne je važnosti za financijske institucije. Cilj ovog rada je procijeniti koji su klijenti kreditno sposobni, a koji nisu na temelju informacija koje su o njima dostupne. Za klasifikaciju klijenata koristit će se model logističke regresije, što je jedan je od najpoznatijih statističkih modela koji se koristi za probleme ovog tipa.

U prvom poglavlju definira se pojam logističke regresije i daju se teorijski rezultati vezani uz model. U drugom poglavlju, primjenom znanja iz prvog poglavlja, odabire se model koji najbolje predviđa kreditnu sposobnost na temelju dostupnih podataka o klijentima. Testira se prilagodba odabranog modela podacima, odnosno kroz razne metrike dana je procjena o tome koliko je model adekvatan. Za kraj, u trećem poglavlju, poseban fokus je na rješavanju problema nebalansiranih kategorija koji se javlja u podacima obzirom da je broj klijenata koji su vratili kredit veći od broja klijenata koji nisu. U tu svrhu analizira se značajnost modela uz primjenu SMOTE algoritma te se uspoređuju rezultati s početnim modelom kako bi se utvrdilo poboljšava li algoritam točnost predikcija.

Poglavlje 1

Logistička regresija

1.1 Uvod i osnovni pojmovi

Regresijska analiza je skup metoda kojima se nastoji ispitati i analizirati ovisnost varijabli. Cilj svake regresijske analize je naći model koji će najbolje opisati i kvantificirati odnos između zavisnih i nezavisnih varijabli.
Nezavisne varijable su one varijable pomoću kojih se želi opisati ili predvidjeti zavisna varijabla. Nezavisne varijable još se nazivaju i kovarijatama i označavaju se s $x_1, x_2..., x_k$, dok se zavisne varijable još nazivaju varijablama odziva i označavaju se s $Y_1, Y_2..., Y_n$.

Model može sadržavati proizvoljan broj zavisnih i nezavisnih varijabli, pa se tako model s jednom nezavisnom varijablom zove univarijabilni model, a multivarijabilni ukoliko su barem dvije nezavisne varijable u modelu. Dodatno, ako model ima jednu zavisnu varijablu onda se zove univarijantni model, a ako su barem dvije zavisne varijable riječ je o multivarijatnom modelu.
U ovom radu fokus je isključivo na modelu koji ima samo jednu zavisnu varijablu i više nezavisnih varijabli koje su zadane i neslučajne.

1.2 Linearna regresija

Jedan od najpoznatijih modela regresijske analize je linearna regresija. U njoj je veza između nezavisnih i zavisnih varijabli linearna. Ukoliko je broj međusobno nezavisnih opažanja jednak n i broj nezavisnih varijabli jednak k, model linearne regresije može se zapisati na sljedeći način:

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i, \tag{1.1}$$

pri čemu je x_{ij}, i=1,2,...,n, i-to opažanje j-te nezavisne varijable. Koeficijenti β_j, j=0,1,...,k, su nepoznati parametri, a ϵ_i, i=1,2,...,n su slučajne greške.

Pretpostavka je da su nezavisne varijable zadane, nemaju grešku te da linearno ovise o zavisnoj varijabli. Osim toga, pretpostavlja se da su slučajne greške normalno distribuirane s očekivanjem nula i nekom konstantnom varijancom σ^2.
Prema tome, očekivanje slučajne varijable Y obzirom na i-to opažanje glasi $\mathbb{E}[Y_i] = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$, uz $Var(Y_i) = \sigma^2$, odnosno Y_i je normalno distribuirana slučajna varijabla s očekivanjem $\mathbb{E}[Y_i] = \mu_i$ i varijancom σ^2, u oznaci $Y_i \sim N(\mu_i, \sigma^2)$.

Linearna regresija samo je jedan od modela unutar šire klase modela koja se zove Generalizirani linearni modeli.

1.3 Generalizirani linearni modeli

Iz prošlog potpoglavlja, uz zapis $\mathbb{E}[Y_i] = \mu_i$ gdje je $\mu_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$, i=1,2,...,n, mogu se uočiti tri stvari:
1. Komponente slučajne varijable Y su normalno distribuirane s očekivanjem $\mathbb{E}[Y_i] = \mu_i$ i konstantnom varijancom σ^2, pri čemu je vektor očekivanja jednak $\mu = (\mu_1, \mu_2, ..., \mu_n)$.
2. Postoji neslučajna komponenta $\eta = (\eta_1, \eta_2, ..., \eta_n)$ pri čemu je:

$$\eta_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}. \tag{1.2}$$

3. Veza slučajne i neslučajne komponente je:

$$\mu = \eta. \tag{1.3}$$

Vrijednost η definirana u (1.2) naziva se linearni prediktor.
Ukoliko se (1.3) napiše u obliku $\eta_i = g(\mu_i)$, tada se funkcija g zove funkcija veze ili funkcija povezivanja.
Kod linearne regresije, funkcija g je identiteta. Zadaća funkcije povezivanja je uspostaviti vezu između komponenata linearnog prediktora η i komponenata očekivanja μ.

Generalizirani linearni modeli generaliziraju linearni model na način da komponente slučajne varijable Y ne moraju biti normalno distribuirane, nego je dovoljno da distribucija dolazi iz neke eksponencijalne familije te funkcija veze g može biti proizvoljna funkcija koja je monotona i diferencijabilna.

Definicija 1.3.1. *Model $\mathcal{P} = \{f(\cdot\,;\theta) : \theta \in \Theta\}$ je k-parametarska eksponencijalna familija ako se funkcije gustoće mogu prikazati u obliku:*

$$f(x;\theta) = C(\theta)h(x)e^{\sum_{j=1}^{k} Q_j(\theta)t_j(x)}$$

pri čemu su $t_1, \ldots, t_k$ nekonstantne, linearno nezavisne funkcije $t_i : \mathbb{R}^n \to \mathbb{R}$ te $C : \Theta \to [0, +\infty)$, $h : \mathbb{R}^n \to [0, +\infty)$, $Q_j : \Theta \to \mathbb{R}$, $j = 1, \ldots, k$ izmjerive funkcije.

U kontekstu linearne regresije zavisna varijabla Y definirana je na način da može postići bilo koju realnu vrijednost, no cilj ovog rada je proučiti slučaj kada Y postiže samo dvije vrijednosti, tj. $Y \in \{0, 1\}$. Obzirom da u tom slučaju nema smisla promatrati linearni model jer komponente Y_i nisu normalno distribuirane i obzirom da očekivanje više nema smisla promatrati na cijelom skupu $\mathbb{R}$, potrebno je naći prikladniji model u skupu generaliziranih linearnih modela.
Kako Y_i postiže ili vrijednost nula ili vrijednost jedan, za distribuciju od Y_i potrebno je uzeti Bernoullijevu distribuciju s parametrom p_i. Tada vrijedi da Y_i postiže vrijednost jedan s vjerojatnošću p_i, odnosno nula s vjerojatnošću 1-p_i. To se može zapisati na sljedeći način:

$$\mathbb{P}(Y_i = x) = p_i^x \cdot (1 - p_i)^{1-x}, \quad x \in \{0, 1\}. \tag{1.4}$$

Napomena 1.3.2. *Bernoullijeva funkcija gustoće je 1-parametarska eksponencijalna familija. Za $\theta = p_i$ vrijedi:*

$$f(x;\theta) = \theta^x(1-\theta)^{1-x}\mathbb{1}_{\{0,1\}} = (1-\theta)\left(\frac{\theta}{1-\theta}\right)^x \mathbb{1}_{\{0,1\}} = (1-\theta)\mathbb{1}_{\{0,1\}}e^{x\log\frac{\theta}{1-\theta}}.$$

Skup $\{x\}$ je linearno nezavisan te su $t_1(x) = x$, $C(\theta) = 1 - \theta$, $h(x) = \mathbb{1}_{\{0,1\}}(x)$, i $Q_1(\theta) = \log\frac{\theta}{1-\theta}$ izmjerive funkcije.

Osim funkcije distribucije, još je potrebno definirati novu funkciju veze. Funkcija veze g treba biti definirana tako da preslika interval $(0, 1)$ na skup $\mathbb{R}$. Najpoznatije funkcije veze koje se koriste u tu svrhu su sljedeće funkcije:

1. logit: $g(x) = \log\left(\frac{x}{1-x}\right)$

2. probit: $g(x) = \Phi^{-1}(x)$, pri čemu je Φ^{-1} inverz funkcije distribucije standardne normalne razdiobe

3. komplementarna log-log: $g(x) = \log(-\log(1 - x))$

Ukoliko je funkcija veze u modelu logit funkcija, tada se radi o modelu logističke regresije.

1.4 Model logističke regresije

Logistička regresija je generalizirani linearni model čiji je cilj predvidjeti Y_i koji poprima ili vrijednost 0 ili vrijednost 1 pomoću nezavisnih varijabli $x_1, x_2, ..., x_k$.

Obzirom da komponente zavisne varijable Y imaju Bernoullijevu distribuciju neka, kao i do sada, p_i označava vjerojatnost da je Y_i jednak vrijednosti 1, tj. $p_i = \mathbb{P}(Y_i = 1)$. Očekivanje Bernoullijeve slučajne varijable Y_i jednako je $\mu_i = \mathbb{E}[Y_i] = 0 \cdot (1 - p_i) + 1 \cdot p_i = p_i$.
Kao u izrazu (1.2), linearni prediktor η je oblika $\eta = \sum_{j=1}^{k} \beta_j x_j$ za nepoznate parametre β_j, a funkcija povezivanja je logit funkcija, $logit : (0, 1) \to \mathbb{R}$, $g(x) = logit(x) = \log\left(\frac{x}{1-x}\right)$.
Iz toga slijedi da je:

$$\mathbb{P}(Y_i = 1|X_i = x_i) = p_i = \mathbb{E}(Y_i = 1|X_i = x_i) = \mu_i = logit^{-1}(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}. \tag{1.5}$$

Stoga, model logističke regresije je model oblika:

$$p(x) = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}}{1 + e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}}, \tag{1.6}$$

ili ekvivalentno:

$$logit(p(x)) = \log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k. \tag{1.7}$$

Iz izraza (1.6) i (1.7) se lako vidi da je logistička funkcija $f(x) = \frac{e^x}{1+e^x}$, $f : \mathbb{R} \to (0, 1)$ inverzna funkciji logit. Logistička funkcija pretvara logaritme izgleda (*eng. log odds*) u vjerojatnosti.
Prijelaz iz izraza (1.6) u (1.7) zove se logit transformacija.

Izgled

Razlog zbog kojega je ponekad korisnije za funkciju veze uzeti logit funkciju umjesto probit ili komplementarne log-log je zbog pojma *izgleda* koji se nalazi u logaritmu funkcije logit.

Definicija 1.4.1. *Neka je Y Bernoullijeva slučajna varijabla s parametrom p, $Y \sim B(p)$. Izgled ili šansa (eng. odds) definira se na sljedeći način:*

$$odds(Y) = \frac{p}{1 - p} = \frac{P(Y = 1)}{P(Y = 0)}. \tag{1.8}$$

Pojam izgleda koristi se u sličnim kontekstima kao i pojam vjerojatnosti. Prednost je što nije ograničen na interval $(0, 1)$, a lako ga je interpretirati kao omjer vjerojatnosti da se događaj dogodio i vjerojatnosti da se nije dogodio.
Izgled poprima vrijednosti u intervalu $[0, +\infty)$, pri čemu se vrijednosti manje od jedan postižu ukoliko je vjerojatnost $p < \frac{1}{2}$.
Ako se izgled logaritmira, tada za vjerojatnost $p < \frac{1}{2}$, log(odds) poprima vrijednosti manje od nula, a za $p > \frac{1}{2}$, log(odds) poprima vrijednosti veće od nula. Na taj se način, logit transformacijom, interval $(0, 1)$ proširuje na cijeli skup $\mathbb{R}$.

Definicija 1.4.2. *Neka su Y_1,Y_2 Bernoullijeve slučajne varijable s parametrom p_1 i p_2, $Y_1 \sim B(p_1)$, $Y_2 \sim B(p_2)$. Omjer izgleda ili omjer šansi (eng. odds ratio) definira se na sljedeći način:*

$$OR = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}. \tag{1.9}$$

Na taj se način može izraziti koliko je puta izgled da se dogodio jedan događaj veći ili manji od izgleda da se dogodi neki drugi događaj.

Pojam izgleda omogućava interpretaciju parametara β_j jer p ovisi o linearnom prediktoru η na nelinearan način.
Neka su $x_1, x_2, ..., x_k$ nezavisne varijable. Vrijedi:

$$\text{odds}(Y \mid X = x) := \frac{p(x)}{1 - p(x)} = e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}. \tag{1.10}$$

Neka je $x = x_j$ kvantitativna varijabla i neka je $\hat{x}$ varijabla dobivena na način da se x_j uveća za jedan, pri čemu su ostale nezavisne varijable fiksne. Tada slijedi:

$$\frac{\text{odds}(Y \mid X = \hat{x})}{\text{odds}(Y \mid X = x)} = \frac{\frac{p(\hat{x})}{1-p(\hat{x})}}{\frac{p(x)}{1-p(x)}} = e^{\beta_j}. \tag{1.11}$$

Prema tome, ako se j-ta nezavisna varijabla poveća za jedan, omjer izgleda mijenja se e^{β_j} puta. Interpretacija je smislena sve dok ne postoji varijabla x_i koja je zavisna s x_j. Kvantitativna varijabla je nerijetko takva da je korisnije promatrati promjenu za neku konstantu c, umjesto za jedan. Ukoliko se x_j uveća za c, analogno kao u (1.11) dobije se da je:

$$\frac{\text{odds}(Y \mid X = \hat{x})}{\text{odds}(Y \mid X = x)} = e^{c\beta_j}. \tag{1.12}$$

1.5 Procjena parametara

Nakon odabira prikladnog modela, potrebno je procijeniti koliko koja nezavisna varijabla ima utjecaj na zavisnu varijablu, odnosno ukoliko je model oblika (1.6) potrebno je na neki način dati procjenu za nepoznate parametre β_j.
Metoda za procjenu parametara u logističkom modelu naziva se metoda maksimalne vjerodostojnosti.

1.5.1 Metoda maksimalne vjerodostojnosti

Definicija 1.5.1. *Neka je* $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ *slučajni uzorak duljine n s gustoćom* $f(\cdot|\theta)$, $\theta \in \Theta \subseteq \mathbb{R}^m$ *i neka je* $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ *realizacija tog uzorka. Vjerodostojnost (eng. likelihood) je funkcija* $L : \Theta \to \mathbb{R}$ *definirana na sljedeći način:*

$$L(\theta) \equiv L(\theta|x) := f_X(x;\theta) = \prod_{i=1}^{n} f(x_i;\theta), \quad \theta \in \Theta. \tag{1.13}$$

Definicija 1.5.2. *Vrijednost* $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n) \in \Theta$ *za koju vrijedi*

$$L(\hat{\theta}) = \max_{\theta\in\Theta} L(\theta) \tag{1.14}$$

zove se procjena metodom maksimalne vjerodostojnosti. Statistika $\hat{\theta}(X_1, X_2, \ldots, X_n)$ *je procjenitelj metodom maksimalne vjerodostojnosti ili kraće MLE(eng. Maximum Likelihood Estimation).*

Dakle, kako bi se mogli procijeniti parametri logističkog modela, potrebno je definirati vjerodostojnost za taj model i onda pronaći parametre koji maksimiziraju tu vjerodostojnost.

Neka je $(\mathbf{x}_i, y_i)$ uzorak n nezavisnih opažanja i neka je model logističke regresije definiran kao u (1.6). Obzirom da je zavisna varijabla y_i Bernoullijeva slučajna varijabla slijedi da je vjerodostojnost jednaka:

$$L(\beta) = \prod_{i=1}^{n} \left(p(x_i)^{y_i} \cdot (1 - p(x_i))^{1-y_i} \right). \tag{1.15}$$

Kako je $x \mapsto \ln(x)$ strogo rastuća funkcija maksimum se može naći i ukoliko se prvo vjerodostojnost logaritmira. Takva funkcija vjerodostojnosti naziva se log-vjerodostojnost i označava se s $l(\beta)$.
Dakle, potrebno je maksimizirati sljedeću funkciju:

$$l(\beta) = \log(L(\beta)) = \sum_{i=1}^{n} \left(y_i \log(p(x_i)) + (1 - y_i) \log(1 - p(x_i)) \right). \tag{1.16}$$

Točka maksimuma log-vjerodostojnosti je stacionarna točka funkcije $l(\beta)$, pa je potrebno izračunati parcijalne derivacije po β_j, j=0,1,...k i izjednačiti ih s nulom.
Za β_0 vrijedi:

$$\begin{aligned}\frac{\partial \ell(\beta)}{\partial \beta_0} &= \sum_{i=1}^{n}\left(y_i \frac{\partial \log(p(x_i))}{\partial p(x_i)}\frac{\partial p(x_i)}{\partial \beta_0} + (1-y_i)\frac{\partial \log(1-p(x_i))}{\partial p(x_i)}\frac{\partial p(x_i)}{\partial \beta_0}\right)\\ &= \sum_{i=1}^{n}\left(y_i \frac{1}{p(x_i)}p(x_i)(1-p(x_i)) - \frac{1-y_i}{1-p(x_i)}p(x_i)(1-p(x_i)\right)\\ &= \sum_{i=1}^{n}(y_i(1-p(x_i)) - (1-y_i)p(x_i))\\ &= \sum_{i=1}^{n}(y_i - p(x_i)).\end{aligned} \tag{1.17}$$

Za β_j, j=1,...k vrijedi:

$$\begin{aligned}\frac{\partial \ell(\beta)}{\partial \beta_j} &= \sum_{i=1}^{n}\left(y_i \frac{\partial \log(p(x_i))}{\partial p(x_i)}\frac{\partial p(x_i)}{\partial \beta_j} + (1-y_i)\frac{\partial \log(1-p(x_i))}{\partial p(x_i)}\frac{\partial p(x_i)}{\partial \beta_j}\right)\\ &= \sum_{i=1}^{n}\left(y_i \frac{1}{p(x_i)}x_j p(x_i)(1-p(x_i)) - \frac{1-y_i}{1-p(x_i)}x_j p(x_i)(1-p(x_i)\right)\\ &= \sum_{i=1}^{n}(y_i x_j(1-p(x_i)) - (1-y_i)x_j p(x_i))\\ &= \sum_{i=1}^{n} x_j(y_i - p(x_i)).\end{aligned} \tag{1.18}$$

Prema tome, za n opažanja uzorka oblika $(x_{i1}, x_{i2}, ..., x_{ik}, y_i)$, i=1,...n, optimalni parametri β_j, j=0,1,...,k dobivaju se rješavanjem k+1 jednadžbe vjerodostojnosti dobivene gore raspisanim računom koje glase:

$$\begin{aligned}\sum_{i=1}^{n}(y_i - p(x_i)) &= 0\\ \sum_{i=1}^{n} x_{ij}(y_i - p(x_i)) &= 0, \quad j = 1, 2, ..., k.\end{aligned} \tag{1.19}$$

Parcijalne derivacije drugog reda funkcije log-vjerodostojnosti dane su s:
$\frac{\partial^2 l(\beta)}{\partial \beta_l \partial \beta_d} = -\sum_{i=1}^{n} x_{id}p(x_i)(1-p(x_i))x_{il}$, koristeći da je derivacija logističke funkcije $p(x)$ jednaka izrazu $p(x)(1-p(x))$.

Matrica čiji su elementi parcijalne derivacije drugog reda, tzv. Hessian, je oblika $H = -X^T WX$, pri čemu je $W \in \mathbb{R}^{n\times n}$ definirana kao matrica težina$W = \text{diag}\{p(x_i)(1 - p(x_i))\}$. Obzirom da $p(x_i)$ poprima vrijednosti između nula i jedan, slijedi da je H negativno definitna matrica pa su stacionarne točke zaista maksimumi.
Kako jednadžbe (1.19) nisu linearne u parametrima β_j, j=0,1,...,k, ne postoji rješenje u zatvorenoj formi. Do rješenja se dolazi pomoću iterativnih metoda.

1.5.2 Svojstva procjenitelja maksimalne vjerodostojnosti

Popularnost metode maksimalne vjerodostojnosti proizlazi i iz činjenice da za određene uvjete na parametarski model vrijedi da su dobiveni procjenitelji konzistentni i asimptotski normalni.
U nastavku ovog potpoglavlja, oznaka za procjenitelje metodom maksimalne vjerodostojnosti je $\hat{\theta}$, dok je oznaka za pravu vrijednost koja se želi procijeniti θ_0. Neka je n kao i ranije veličina uzorka.

Definicija 1.5.3. *Niz procjenitelja $\hat{\theta}_n$ za vektor parametara $\theta \in \Theta$ je konzistentan ako vrijedi:* $(\forall \varepsilon > 0)\ \lim_{n\to\infty} P_{\theta_0}(|\hat{\theta}_n - \theta_0| \geq \varepsilon) = 0$*, u oznaci* $\hat{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0$.

Definicija 1.5.4. *Niz procjenitelja $\hat{\theta}_n$ je asimptotski normalan ukoliko $\sqrt{n}(\hat{\theta}_n - \theta_0)$ konvergira po distribuciji prema normalnoj razdiobi(ili multivarijatnoj normalnoj razdiobi ukoliko je dimenzija od θ barem dva) s očekivanjem nula i asimptotskom varijancom procjenitelja $\hat{\theta}_n$.*

Da bi se moglo doći do zaključka da su procjenitelji konzistentni i asimptotski normalni, potrebno je definirati regularan statistički model.

Definicija 1.5.5. *Statistički model $\mathcal{P} = \{f(\cdot;\theta) : \theta \in \Theta\}$ za jednodimenzionalne razdiobe je regularan ako su zadovoljeni sljedeći uvjeti:*

1. *Nosač supp $f(\cdot;\theta) = \{x \in \mathbb{R} : f(x;\theta) > 0\}$ ne ovisi o θ;*
2. *Parametarski prostor Θ je otvoren skup u $\mathbb{R}$;*
3. *Preslikavanje $\theta \mapsto f(x;\theta)$ je neprekidno diferencijabilno na Θ, za svaki x;*
4. *$\sum_{x\in\mathbb{R}} \left(\frac{\partial}{\partial\theta} \log f(x;\theta)\right)^2 f(x;\theta)\, dx \in \langle 0, \infty\rangle$ za sve $\theta \in \Theta$;*
5. *$\frac{\partial}{\partial\theta} \sum_{x\in\mathbb{R}} f(x;\theta) = \sum_{x\in\mathbb{R}} \frac{\partial}{\partial\theta} f(x;\theta) = 0$.*

Treći i peti uvjet mogu se dodatno proširiti tako da je preslikavanje $\theta \mapsto f(x;\theta)$ klase C^3 na Θ te da je $\frac{\partial}{\partial\theta^k} \sum_{x\in\mathbb{R}} f(x;\theta) = \sum_{x\in\mathbb{R}} \frac{\partial^k}{\partial\theta^k} f(x;\theta) = 0$ za $k \in \{1, 2\}$.

Napomena 1.5.6. *Bernoullijev model zadovoljava uvjete regularnosti:*
Nosač je skup {0,1} koji ne ovisi o θ. Parametarski prostor je interval $\langle 0, 1\rangle$ *što je otvoren skup u* $\mathbb{R}$*. Preslikavanje* $\theta \mapsto f(x;\theta)$ *je neprekidno diferencijabilno jer je* $f(x;\theta)$ *jednak* θ *za* $x = 1$, $1-\theta$ *za* $x = 1$ *i nula inače. Nadalje, izraz* $\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)$ *je jednak* $x/\theta-(1-x)/(1-\theta)^2$ *pa je onda izraz dan u četvrtom uvjetu konačan i veći od nula. Zamjena sume i derivacije slijedi iz činjenice da je model diskretan.*

Neka je $X = (X_1, X_2, ...X_n)$ slučajni uzorak sa zajedničkom gustoćom $f_X(x;\theta_0)$, pri čemu su komponente slučajnog uzorka nezavisne i jednako distribuirane.

Kako bi se pokazalo da je procjenitelj $\hat{\theta}$ konzistentan, potrebno je pokazati da konvergira po vjerojatnosti k θ_0.
Vrijednost $\hat{\theta}$ je dobivena metodom maksimalne vjerodostojnosti što znači da maksimizira funkciju $\frac{1}{n}l(\theta) = \frac{1}{n}\sum_{i=1}^{n}\log f(x_i;\theta)$, pri čemu faktor $1/n$ ne utječe na maksimizaciju, a uveden je radi jednostavnosti u kasnijem računu.
Vrijednost θ_0 je prava vrijednost koja se želi procijeniti.
Slabi zakon velikih brojeva kaže da za nezavisan jednako distribuiran uzorak $X_1, X_2, ..., X_n$ takav da je $|\mathbb{E}X_1| < \infty$ vrijedi da $\bar{X}_n = (X_1 + ... + X_n)/n$ konvergira po vjerojatnosti prema $\mathbb{E}X_1$. [7]
Iz toga slijedi da za svaki θ vrijedi:

$$\frac{1}{n}\sum_{i=1}^{n}\log f(X_i;\theta) \to E_{\theta_0}[\log f(X;\theta)]. \tag{1.20}$$

To implicira da vrijednost koja maksimizira lijevu stranu (1.20), $\hat{\theta}$, konvergira po vjerojatnosti k vrijednosti θ koja maksimizira desnu stranu. Tvrdnja je da je vrijednost θ koja maksimizira desnu stranu upravo θ_0.
Zaista, za svaki θ vrijedi:

$$\mathbb{E}_{\theta_0}[\log f(X;\theta)] - \mathbb{E}_{\theta_0}[\log f(X;\theta_0)] = \mathbb{E}_{\theta_0}\left[\frac{\log f(X;\theta)}{f(X;\theta_0)}\right] \le \log\mathbb{E}_{\theta_0}\left[\frac{f(X;\theta)}{f(X;\theta_0)}\right]$$
$$= \log\sum\frac{f(x;\theta)}{f(x;\theta_0)}f(x;\theta_0) = \log\sum f(x;\theta) = 0.$$

Nejednakost slijedi iz činjenice da je $x \mapsto \log x$ konveksna funkcija pa je $\mathbb{E}[\log X] \le \log\mathbb{E}[X]$ primjenom Jensenove nejednakosti.
Dakle funkcija $\theta \mapsto \mathbb{E}_{\theta_0}[\log f(X;\theta)]$ postiže maksimum za $\theta = \theta_0$ iz čega slijedi konzistentnost procjenitelja $\hat{\theta}$.

Za pokazivanje svojstva asimptotske normalnosti potrebno je uvesti nekoliko važnih pojmova.

Definicija 1.5.7. *Fisherova funkcija pogotka je funkcija čije su komponente prve parcijalne derivacije log-vjerodostojnosti:*

$$u(x,\theta) = \frac{\partial}{\partial\theta}\log f(x;\theta) = \frac{\partial l(\theta;x)}{\partial\theta}. \tag{1.21}$$

Definicija 1.5.8. *Fisherova informacijska matrica definirana je na sljedeći način:*

$$I(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right]. \tag{1.22}$$

Za informacijsku matricu vrijedi sljedeća jednakost: $I(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right]$ jer je

$$\begin{aligned}
&\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right] = \sum\left(\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)\right)f(x;\theta)\\
&= \sum\left(\frac{\frac{\partial^2}{\partial\theta^2}f(x;\theta)}{f(x;\theta)} - \left(\frac{\frac{\partial}{\partial\theta}f(x;\theta)}{f(x;\theta)}\right)^2\right)f(x;\theta) = \sum\frac{\partial^2}{\partial\theta^2}f(x;\theta) - \sum\left(\frac{\frac{\partial}{\partial\theta}f(x;\theta)}{f(x;\theta)}\right)^2 f(x;\theta)\\
&= 0 - \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right] = 0 - I(\theta) = -I(\theta).
\end{aligned}$$

Treća jednakost slijedi jer je $(logf(x))'' = \frac{f''(x)}{f(x)} - \left(\frac{f'(x)}{f(x)}\right)^2$, a peta jer je $\sum f(x) = 1$ za sve x pa su prva i druga derivacija tog izraza jednake nuli.
Iz činjenice da $\hat\theta$ maksimizira $l(\theta)$ vrijedi da je $l'(\hat\theta) = 0$. Uz Taylorov razvoj funkcije $l'(\hat\theta)$ oko $\hat\theta = \theta_0$ s dva člana dobije se izraz: $0 \approx l'(\theta_0) + (\hat\theta - \theta_0)l''(\theta_0)$, odnosno:

$$\sqrt{n}(\hat\theta - \theta_0) \approx -\sqrt{n}\frac{l'(\theta_0)}{l''(\theta_0)} = -\frac{\frac{1}{\sqrt{n}}l'(\theta_0)}{\frac{1}{n}l''(\theta_0)}. \tag{1.23}$$

Za nazivnik razlomka iz (1.23), po slabom zakonu velikih brojeva i zbog $I(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right] = -\mathbb{E}_\theta[u'(X,\theta)]$ vrijedi:

$$\frac{1}{n}l''(\theta_0) = \frac{1}{n}\sum_{i=1}^n\frac{\partial^2}{\partial\theta^2}\left[\log f(x_i;\theta)\right]_{\theta=\theta_0} = \frac{1}{n}\sum_{i=1}^n u'(x_i,\theta_0) \to \mathbb{E}_{\theta_0}[u'(X,\theta_0)] = -I(\theta_0). \tag{1.24}$$

Za brojnik vrijedi sljedeće:

$$\frac{1}{\sqrt{n}}l'(\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n\frac{\partial}{\partial\theta}\left[\log f(x_i;\theta)\right]_{\theta=\theta_0} = \frac{1}{\sqrt{n}}\sum_{i=1}^n u(x_i,\theta_0) \to N(0, I(\theta_0)). \tag{1.25}$$

Tvrdnja iz (1.25) slijedi zbog centralnog graničnog teorema i jer su očekivanje i varijanca za $u(X,\theta_0)$ jednaki nula i $I(\theta_0)$.

Centralni granični teorem kaže da za nezavisne jednako distribuirane $X_1, ...X_n$ takve da je $|\mathbb{E}X_1| < \infty$ i $\sigma^2 = Var(X_1) < \infty$ vrijedi $\sqrt{n}(\bar{X}_n - \mathbb{E}X_1) \xrightarrow{d} N(0, \sigma^2)$. [7]

Za funkciju pogotka vrijedi da je $\mathbb{E}_\theta[u(X,\theta)] = 0, \text{Var}_\theta[u(X,\theta)] = -\mathbb{E}[u'(X,\theta)]$.
Prva tvrdnja može se pokazati na sljedeći način:

$$u(x,\theta)f(x;\theta) = \left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)f(x;\theta) = \frac{\frac{\partial}{\partial\theta}f(x;\theta)}{f(x;\theta)}f(x;\theta) = \frac{\partial}{\partial\theta}f(x;\theta)$$

iz čega slijedi:

$$\mathbb{E}_\theta[u(X,\theta)] = \sum u(x,\theta)f(x;\theta) = \sum \frac{\partial}{\partial\theta}f(x;\theta) = \frac{\partial}{\partial\theta}\sum f(x;\theta) = 0.$$

Koristeći dobivenu tvrdnju slijedi da je

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\mathbb{E}_\theta[u(X,\theta)] = \frac{\partial}{\partial\theta}\sum u(x,\theta)f(x;\theta) = \sum (u'(x,\theta)f(x;\theta) + u(x,\theta)\frac{\partial}{\partial\theta}f(x;\theta)) \\ &= \sum (u'(x,\theta)f(x;\theta) + u(x,\theta)^2 f(x;\theta)) = \mathbb{E}_\theta[u'(X,\theta)] + \mathbb{E}_\theta[u(X,\theta)^2] \\ &= \mathbb{E}_\theta[u'(X,\theta)] + \text{Var}_\theta[u(X,\theta)]. \end{aligned}$$

Dakle,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to\sim N\left(0, I(\theta_0)^{-1}\right). \tag{1.26}$$

Općenito, Fisherova informacijska matrica nije poznata, kao ni procjenitelj za θ pa je procijenjena uvrštavanjem procjenitelja $\hat{\theta}$ ili za dovoljno velik uzorak, adekvatna procjena je negativna matrica drugih parcijalnih derivacija log-vjerodostojnosti, odnosno $\hat{I} = -H$.

Procjenitelji dobiveni metodom maksimalne vjerodostojnosti u (1.19) su konzistentni i asimptotski normalni tako da vrijedi:

$$\hat{\beta} \sim AN(\beta, (X^T W X)^{-1}). \tag{1.27}$$

Može se pokazati da isti rezultati vrijede i za slučaj kada $X_1, X_2, ..., X_n$ nemaju istu distribuciju.

1.6 Prilagodba modela podacima

Jednom kada su parametri β_j procijenjeni, sljedeći je korak promotriti koliko je dobro model prilagođen podacima i treba li smatrati da je taj model prihvatljiv. Smatra se da je model dobro prilagođen podacima ukoliko je razlika između danih i dobivenih vrijednosti zavisne varijable što manja. Na temelju koliko se te vrijednosti razlikuju, potrebno je vidjeti koje nezavisne varijable su značajne i dati neke mjere koje pokazuju koliko je model dobar.
Postoje mnoge statistike koje opisuju odstupanje od dane vrijednosti zavisne varijable, a jedna od najvažnijih u logističkoj regresiji je devijanca.

1.6.1 Devijanca

Devijanca (*eng. deviance*), kao jedna od statistika odstupanja, temelji se na funkciji vjerodostojnosti definiranoj kao u (1.16). Vrijednost log-vjerodostojnosti, koja je dobivena uvrštavanjem procijenjenih parametara β_j koji ju maksimiziraju, daje uvid u to koliko je model prilagođen podacima. Obzirom da je ta vrijednost zavisna s brojem opažanja, sama po sebi nije dobar pokazatelj prilagodbe. Za dobru mjeru odstupanja, potrebno je usporediti model s alternativnim modelom kojemu se opažene i dane vrijednosti ne razlikuju. Takav model, koji savršeno opisuje podatke, naziva se saturirani model i sadrži onoliko parametara koliko je opažanja.
Devijanca, u oznaci D, dana je sljedećim izrazom:

$$D = -2\log\left(\frac{\text{vjerodostojnost promatranog modela}}{\text{vjerodostojnost saturiranog modela}}\right). \tag{1.28}$$

Devijanca poprima veće vrijednosti ukoliko je brojnik manji u odnosu na nazivnik što ukazuje na to da model nije dobro prilagođen podacima. Dakle, manja devijanca znači da je prilagodba modela bolja, no ne nužno i da je model bolji obzirom da je moguća prekomjerna prilagodba podacima.

Obzirom da za saturirani model vrijedi da je $p(x_i) = y_i$ i da je vjerodostojnost modela definirana na sljedeći način: $L(\beta) = \prod_{i=1}^{n}\left(p(x_i)^{y_i}\cdot(1-p(x_i))^{1-y_i}\right)$ slijedi da je vjerodostojnost saturiranog modela jednaka 1 jer je $\prod_{i=1}^{n} y_i^{y_i}(1-y_i)^{(1-y_i)} = 1$.

Iz toga slijedi da je devijanca za zavisnu slučajnu varijablu koja poprima vrijednosti nula

ili jedan, tj. devijanca za binarni odziv dana na sljedeći način:

$$\begin{aligned} D &= -2\log(\text{vjerodostojnost promatranog modela}) \\ &= -2\sum_{i=1}^{n}(y_i\log(p(x_i)) + (1-y_i)\log(1-p(x_i)). \end{aligned} \tag{1.29}$$

Devijanca je važna u testu omjera vjerodostojnosti o kojem će više riječi biti kasnije. Ukratko, statistika D koristi se prilikom usporedbe dvaju ugniježđenih modela. Ukoliko se uspoređuju promatrani i saturirani model hipoteze glase:

$$H_0 : \text{Promatrani model je dovoljno dobar}$$
$$H_1 : \text{Promatrani model nije dovoljno dobar,}$$

pri čemu alternativna hipoteza reprezentira saturirani model.
Testna statistika je upravo jednaka devijanci i često se tvrdi da D ima asimptotsku χ^2 distribuciju s $n-k$ stupnjeva slobode, iako to nije točno. Kako bi se mogle uvesti pretpostavke o asimptotskoj χ^2 distribuciji, potrebno je uvesti pojam kovarijantnih razreda.

Kovarijantni razred je grupa podataka čije nezavisne varijable imaju jednake vrijednosti. Ukoliko model ima dvije nezavisne varijable i svaka može poprimiti dvije vrijednosti, tada postoje četiri kovarijantna razreda. Ukoliko je barem jedna nezavisna varijabla neprekidna, tada je broj kovarijantnih razreda jednak broju opažanja.
Neka su dane nezavisne varijable $x_1, x_2, ..., x_k$ i neka je J broj različitih vrijednosti koje se postižu za iste vrijednosti nezavisnih varijabli. Broj subjekata u svakom razredu označen je s n_i te je s y_i označen broj koliko je puta zavisna varijabla jednaka jedan među n_i subjekata, pri čemu je $x = x_i$, $i = 1, .., J$.

Zavisna varijabla više nema Bernoullijevu distribuciju nego binomnu distribuciju.
Općenito, slučajna varijabla Y ima binomnu distribuciju s parametrima n i p, $Y \sim B(n,p)$, ako je razdioba sljedećeg oblika:

$$\mathbb{P}(Y=y) = \binom{n}{y} p^y(1-p)^{n-y}, \quad y \in \{0,1,\ldots,n\}. \tag{1.30}$$

Napomena 1.6.1. *a) Binomni model je 1-parametarska eksponencijalna familija.*
b) Logistički model je definiran na način da je $logit(p_i(x_i)) = \beta_0 + x_1\beta_1 + \ldots + x_k\beta_k$, *pri čemu su zavisne varijable* $Y_i \sim B(n_i, p_i)$ *međusobno nezavisne slučajne varijable,* $i = 1, ..., J$.
c) Vjerodostojnost je dana sljedećim izrazom: $L(\beta) = \prod_{i=1}^{J}\binom{n_i}{y_i}p_i(x_i)^{y_i}(1-p_i(x_i))^{n_i-y_i}$, *odnosno log-vjerodostojnost:* $l(\beta) = \sum_{i=1}^{J}\left(\log\binom{n_i}{y_i} + y_i\log p_i(x_i) + (n_i-y_i)\log(1-p_i(x_i))\right) = \sum_{i=1}^{J}\left(\log\binom{n_i}{y_i} + y_i\log\left(\frac{p_i(x_i)}{1-p_i(x_i)}\right) + n_i\log(1-p_i(x_i))\right)$.

Za procijenjene vrijednosti $\hat{y}_i = n_i\hat{p}_i$ promatranog modela i $y_i = n_i\tilde{p}_i$ saturiranog modela, zanemarivanjem konstantnog člana funkcije log-vjerodostojnosti i po *Napomeni 1.6.1 b)* slijedi da je devijanca modela s binomnim odzivom dana sljedećim izrazom:

$$\begin{aligned} D &= -2\log(\text{vjerodostojnost promatranog modela}) + 2\log(\text{vjerodostojnost saturiranog modela}) \\ &= -2\sum_{i=1}^{J}\left(y_i\log\hat{p}_i + (n_i - y_i)\log(1-\hat{p}_i)\right) + 2\sum_{i=1}^{J}\left(y_i\log\tilde{p}_i + (n_i - y_i)\log(1-\tilde{p}_i)\right) \\ &= 2\sum_{i=1}^{J}\left(y_i\log\frac{\tilde{p}_i}{\hat{p}_i} + (n_i - y_i)\log\frac{1-\tilde{p}_i}{1-\hat{p}_i}\right) \\ &= 2\sum_{i=1}^{J}\left(y_i\log\frac{y_i}{\hat{y}_i} + (n_i - y_i)\log\frac{n_i - y_i}{n_i - \hat{y}_i}\right). \end{aligned} \tag{1.31}$$

Da bi devijanca bila asimptotski χ^2 distribuirana s $J-(k+1)$ stupnjeva slobode, pri čemu je J broj kovarijantnih razreda, a k+1 broj nepoznatih parametara u modelu, trebaju vrijediti sljedeće pretpostavke:
1. Opažanja iz binomne distribucije trebaju biti nezavisna.
2. Za fiksnu vrijednost opažanja n i proizvoljan broj kovarijantnih razreda J vrijedi da $n_i \to \infty$ i $n_i p_i(1-p_i) \to \infty$, za svaki $i = 1, \ldots, J$.

Prva pretpostavka je uvijek zadovoljena jer je to i pretpostavka modela.
Razlog zbog kojega devijanca definirana u (1.29) nije asimptotski χ^2 distribuirana je zbog druge pretpostavke. Naime, u tom slučaju vrijedi da je broj kovarijantnih razreda jednak n pa je $n_i = 1$, za svaki $i = 1, .., J$.

1.6.2 Pearsonova χ^2 i Hosmer-Lemeshowova statistika

Pearsonova χ^2 statistika jedna je od poznatijih alternativa devijanci i definirana je na sljedeći način:

$$X_P^2 = \sum_{i=1}^{J}\frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)}, \tag{1.32}$$

pri čemu su $y_i = n_i p_i$ dane vrijednost, a $\hat{y}_i = n_i\hat{p}_i$ procijenjene vrijednosti za $i = 1, .., J$.
Pod pretpostavkom da je početni model dobar, Pearsonova χ^2 statistika, kao i devijanca, ima asimptotsku χ^2 distribuciju s $J - (k + 1)$ stupnjeva slobode. Vrijednosti tih statistika općenito se razlikuju, ali razlika najčešće nije od praktične važnosti. Velika razlika vrijednosti tih statistika ukazuje da aproksimacija χ^2 distribucijom nije točna.
Ukoliko se procjena parametara radi metodom maksimalne vjerodostojnosti, devijanca

ipak ima prednost ispred Pearsonove χ^2 statistike obzirom da je devijanca s procjeniteljima koji maksimiziraju log-vjerodostojnost minimizirana. Devijanca ima prednost i prilikom usporedbe dvaju ugniježđenih modela, pri čemu razlika devijanci ukazuje na značajnost dodanih varijabli, dok se Pearsonova χ^2 statistika, dodavanjem varijabli, može povećati.

Hosmer-Lemeshowova statistika je, za razliku od devijance i Pearsonove χ^2 statistike, korištena za modeliranje binarnih odziva ili ukoliko je broj n_i, definiran kao broj subjekata u kovarijantnom razredu, manji od pet.
U slučaju da zavisna varijabla nije binarna i podijeljena je u kovarijantne razrede, potrebno je razdvojiti subjekte kako bi broj kovarijantnih razreda J bio jednak broju opažanja n.
Vrijednost statistike dobije se tako da se u prvom koraku poredaju binarna opažanja y_i po vrijednostima procijenjenih vjerojatnosti $p(x_i)$. Tako poredane vrijednosti dijele se u g grupa približno sličnih veličina. Grupe se mogu odabrati korištenjem percentila procijenjenih vjerojatnosti ili unaprijed određenim intervalima. Sumirane vrijednosti opaženih uspjeha (y_i=1) i procijenjenih uspjeha unutar grupe uspoređuju se χ^2 statistikom.
Dakle, za broj opažanja n_i, opaženi broj uspjeha o_i, procijenjeni očekivani broj uspjeha e_i te prosječnu vjerojatnost uspjeha $\hat{p}_i$ u i-toj grupi, pri čemu je $\hat{p}_i = e_i/n_i$, slijedi da je:

$$X^2_{HL} = \sum_{i=1}^{g} \frac{(o_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)}. \tag{1.33}$$

Statistika dana s (1.33) zove se Hosmer-Lemeshowova statistika.
Simulacijske studije su pokazale da pod pretpostavkom da je model dobar, statistika X^2_{HL} ima približno χ^2 distribuciju s $(g-2)$ stupnjeva slobode, pri čemu je g broj grupa.[4]
Problem statistike X^2_{HL} je u tome što ovisi o broju grupa i o broju opažanja unutar grupe.

U praksi, kako je vrijednost X^2_{HL} previše podložna odabiru grupa, a cilj je predvidjeti zavisnu varijablu koja je binarna, predloženo je puno alternativnih statistika. Jedna od njih je Standardizirana Pearsonova χ^2 statistika koja je dana izrazom: $X^2 = \sum_i \frac{(y_i-\hat{p}_i)^2}{\hat{p}_i(1-\hat{p}_i)}$, pri tome X^2 ima asimptotsku normalnu distribuciju.

Navedene statistike koriste se kako bi se pod pretpostavkom da je nulta hipoteza točna, tj. da je navedeni model dobar, mogao donijeti zaključak o modelu, najčešće na temelju p-vrijednosti.

1.6.3 Testiranje hipoteza

Nakon procjene parametara u modelu, potrebno je vidjeti postoje li nezavisne varijable koje nisu značajne za model. Za ispitivanje značajnosti parametara za pojedine nezavisne varijable koriste se statistički testovi. Testiranjem značajnosti više parametara može se doći

do zaključka je li dovoljan promatrani model ili je ipak prikladniji model s više nezavisnih varijabli.
Kako bi se testiralo je li nezavisna varijabla x_j značajna, hipoteze glase:

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_1 &: \beta_j \neq 0. \end{aligned} \tag{1.34}$$

Za usporedbu dva modela, oni moraju biti ugniježđeni. Ukoliko promatrani, manji model ima nezavisne varijable $x_1, x_2, ..., x_k$, to znači da veći model za koji se želi vidjeti je li prikladniji, uz nezavisne varijable $x_{k+1}, ..., x_r$ ima iste varijable $x_1, x_2, ..., x_k$ kao i promatrani model. Hipoteze u tom slučaju glase:

$$\begin{aligned} H_0 &: \text{ manji model je dovoljan} : \beta_{k+1}, .., \beta_r = 0 \\ H_1 &: \text{potreban je veći model.} \end{aligned} \tag{1.35}$$

Najpoznatiji takvi testovi, za logističku regresiju su test omjera vjerodostojnosti i Waldov test.

Test omjera vjerodostojnosti

Test omjera vjerodostojnosti kratko je spomenut prilikom usporedbe promatranog i saturiranog modela, gdje se testiralo je li promatrani model prikladan. Općenito, testom omjera vjerodostojnosti uspoređuju se dva modela, pri čemu oni moraju biti ugniježđeni.
Neka su nezavisne varijable promatranog modela ω dane s $x_1, x_2, ..., x_k$ te neka su nezavisne varijable modela ω', $x_1, ..x_k, x_{k+1}, x_r$. Cilj je vidjeti je li promatrani model ω dovoljan ili je ipak bolji model ω' koji u sebi sadrži i nezavisne varijable $x_{k+1}, ..., x_r$. Hipoteze su dane na isti način kao i u (1.35).
Testna statistika, ukoliko je ω' saturirani model, jednaka je iznosu devijance.
Testna statistika prilikom usporedbe proizvoljna dva ugniježđena modela, dana je sljedećim izrazom:

$$T := 2\left(l(\hat{\beta}_{\omega'}) - l(\hat{\beta}_{\omega})\right) = D_{\omega} - D_{\omega'}, \tag{1.36}$$

pri čemu je s $l(\hat{\beta})$ označena funkcija log-vjerodostojnosti, a D je oznaka za devijancu.
Druga jednakost slijedi iz činjenice da su se log-vjerodostojnosti saturiranog modela pokratile.
Ako je nulta hipoteza istinita, za veliki n, statistika T ima asimptotsku χ^2 razdiobu s $(r-k)$ stupnjeva slobode, neovisno o broju kovarijantnih razreda i broju subjekata unutar kovarijantnih razreda.

Waldov test

Procjenitelj $\hat{\beta}$, kao što je pokazano u (1.27) ima asimptotsku multivarijatnu normalnu razdiobu s očekivanjem koje je jednako stvarnoj vrijednosti β i kovarijacijskom matricom $(X^T\hat{W}X)^{-1}$. Neka je oznaka za kovarijacijsku matricu $\hat{\Sigma}(\hat{\beta})$.
Za testiranje hipoteza danih kao u (1.34) koristi se testna statistika oblika

$$Z = \frac{\beta_j}{\sqrt{(\hat{\Sigma}(\hat{\beta}))_{jj}}}. \tag{1.37}$$

Pod pretpostavkom da je H_0 istinita, statistika Z ima asimptotsku standardnu normalnu razdiobu.
Ukoliko hipoteze glase na sljedeći način:

$$\begin{aligned} H_0 &: \beta = \beta' \\ H_1 &: \beta \neq \beta', \end{aligned} \tag{1.38}$$

za vektor parametara $\beta' = (\beta'_0, \beta'_1, .., \beta'_k)$ tada je testna statistika dana s

$$W = (\hat{\beta} - \beta')^T(\hat{\Sigma}(\hat{\beta}))^{-1}(\hat{\beta} - \beta') \tag{1.39}$$

i ako je H_0 istinita ima χ^2 distribuciju s $(k + 1)$ stupnjeva slobode.
Statistika W zove se Waldova statistika, a statistika Z iz (1.37) je zapravo korijen Waldove statistike.
Za usporedbu dva ugniježđena modela, hipoteze su dane kao i u (1.35). Ukoliko je $\tilde{\beta}$ oznaka za vektor parametara $(\beta_{k+1}, .., \beta_r)$, testna statistika u tom je slučaju dana s

$$W = \tilde{\beta}^T(\hat{\Sigma}(\tilde{\beta}))^{-1}\tilde{\beta} \tag{1.40}$$

i ima χ^2 distribuciju s $(r - k)$ stupnjeva slobode, ako je H_0 istinita.

Waldov test i test omjera vjerodostojnosti daju slične rezultate ukoliko se testiraju na velikom uzorku. Kod manjih uzoraka nije poznato koji je bolji, iako neke simulacijske studije pokazuju da je test omjera vjerodostojnosti precizniji.

1.6.4 Pouzdani intervali

Asimptotska normalnost procjenitelja omogućava i konstrukciju pouzdanih intervala parametara $\beta_j,\ j = 0, 1, ..., k$.
Iz činjenice da je $\hat{\beta} \sim AN(\beta, (X^TWX)^{-1})$ slijedi da je

$$\mathbb{P}(-z_{1-\alpha/2} \leq \frac{\hat{\beta}_j - \beta_j}{\hat{\text{SE}}(\hat{\beta}_j)} \leq z_{1-\alpha/2}) \approx 1 - \alpha, \tag{1.41}$$

pri čemu je $\hat{\text{SE}}(\hat{\beta}_j)$ procjena standardne greške koja je jednaka $\sqrt{(\hat{\Sigma}^{-1}(\hat{\beta}))_{jj}}$ te je $z_{1-\alpha/2}$ $(1-\frac{\alpha}{2})$-kvantil standardne normalne distribucije.

Dakle, za danu razinu značajnosti α, Waldov $(1-\alpha)\cdot 100\%$ pouzdani interval je oblika:

$$\left(\hat{\beta}_j - z_{1-\alpha/2}\sqrt{(\hat{\Sigma}^{-1}(\hat{\beta}))_{jj}}, \hat{\beta}_j + z_{1-\alpha/2}\sqrt{(\hat{\Sigma}^{-1}(\hat{\beta}))_{jj}}\right). \tag{1.42}$$

Ovakva procjena pouzdanih intervala pogodna je za velike uzorke obzirom da se konstruiraju na temelju pretpostavke o asimptotskoj distribuciji.

1.6.5 Generalizirani $\mathbf{R}^2$

Statistički testovi odgovaraju na pitanje može li model biti bolji dodavanjem određenih nezavisnih varijabli, ali ne daje nikakvu numeričku vrijednost kojom se može opisati točnost modela. Takve metrike vrlo su korisne jer daju uvid u to koliko promatrani model dobro predviđa.
R^2 je jedna od najpoznatijih statističkih mjera koja predstavlja udio varijabilnosti zavisne varijable na temelju prilagođenog modela. U slučaju linearne regresije $R^2 = 1 - \frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2}$, pri čemu je y_i dana vrijednost, $\hat{y}_i$ procijenjena vrijednost, a $\bar{y}$ prosječna vrijednost zavisne varijable.Vrijednost R^2 poprima vrijednosti između 0 i 1, pri čemu veće vrijednosti ukazuju da je model dobro prilagođen podacima.
U slučaju binarne zavisne varijable nema smisla koristiti formulu kao i u slučaju linearne regresije. Obzirom da je R^2 vrlo korisna metrika predloženo je nekoliko generalizacija s tim da nema dogovora oko toga koja je generalizacija najbolja.

Jedna generalizacija, takozvani *pseudo* R^2, dan je sljedećim izrazom:

$$R^2 = 1 - \frac{\log L(\hat{\beta})}{\log \hat{L}_0}, \tag{1.43}$$

pri čemu je $\log L(\hat{\beta})$ vjerodostojnost promatranog modela koja je maksimizirana uz procijenjene parametre $\hat{\beta}$, a $\log \hat{L}_0$ je vjerodostojnost za model koji ne sadrži niti jednu nezavisnu varijablu, samo procijenjeni parametar $\hat{\beta}_0$.
Tako definiran R^2 poprima vrijednosti između 0 i 1, pri čemu vrijednost jedan odgovara savršeno prilagođenom modelu.

Druga generalizacija definira se sljedećim izrazom:

$$R^2 = 1 - \left(\frac{\hat{L}_0}{L(\hat{\beta})}\right)^{2/n}, \tag{1.44}$$

pri čemu je n veličina uzorka. Jedan od problema je što je izraz (1.44) uvijek manji od 1, točnije postiže maksimum za $1 - \frac{(L_0)^2}{n}$ pa se može promatrati dodatna generalizacija:

$$R^2 = \frac{1 - \left(\frac{\hat{L}_0}{L(\hat{\beta})}\right)^{2/n}}{1 - \frac{(L_0)^2}{n}}. \tag{1.45}$$

1.6.6 Reziduali

Ključna stvar kod svakog modela je udaljenost dane i procijenjene vrijednosti zavisne varijable.
U modelu linearne regresije, rezidual se definira kao razlika dane i procijenjene vrijednosti $y - \hat{y}$, pri čemu je ta greška normalno distribuirana. Reziduale za generalizirane linearne modele potrebno je poopćiti tako da budu primjenjivi i za ostale distribucije.
U nastavku su navedena dva poopćenja reziduala.

Pearsonov rezidual definira se na sljedeći način:

$$r_p = \frac{y_i - \hat{y}_i}{\sqrt{Var(Y_i)}} = \frac{y_i - \hat{y}_i}{\sqrt{\hat{p}_i(1 - \hat{p}_i)}}, \tag{1.46}$$

pri čemu je Y_i Bernoullijeva slučajna varijabla s parametrom p_i.
Kvadrirani i sumirani izraz Pearsonovog reziduala daje upravo vrijednost Pearsonove χ^2 statistike (1.32).

Reziduali devijance dani su izrazom:

$$r_d = \operatorname{sgn}(y_i - \hat{y}_i) \cdot \left(2\left(y_i \log \frac{y_i}{\hat{y}_i} + (1 - y_i) \log \frac{1 - y_i}{1 - \hat{y}_i}\right)\right)^{1/2}. \tag{1.47}$$

Izraz $\operatorname{sgn}(y_i - \hat{y}_i)$ označava da je predznak reziduala jednak kao i predznak od $y_i - \hat{y}_i$. Također, kvadriranjem i sumiranjem danog izraza dobiva se devijanca definirana u (1.31).

Poglavlje 2

Klasifikacija kreditne sposobnosti

2.1 Opis problema

Cilj ovog poglavlja je za zadane podatke o klijentima, naći model koji će klasificirati klijente u dvije klase: oni koji jesu kreditno sposobni i oni koji nisu.

Klasifikacija je proces razvrstavanja podataka u neku od unaprijed definiranih kategorija, odnosno klasa. Za razliku od regresije gdje su vrijednosti zavisne varijable neprekidne, u klasifikaciji su vrijednosti zavisne varijable diskretne. Ukoliko je broj klasa jednak dva, tada je riječ o binarnoj klasifikaciji.
Jedan od najpoznatijih algoritama za binarnu klasifikaciju je logistička regresija. Logistička regresija sama po sebi nije klasifikator jer procjenjuje vjerojatnost pripadanja određenoj klasi. Ipak, koristi se kao klasifikator na način da se uz neki odabrani prag podaci razvrstaju u klase ovisno je li njihova predviđena vjerojatnost manja ili veća od praga.

Klasifikacija klijenata između onih koji će uspjeti vratiti kredit i onih koji neće jedan je od glavnih problema financijskih institucija. Dobra predikcija može dovesti do sprječavanja velikih gubitaka i do većih prihoda.
Kreditna klasifikacija je metoda procjene kreditne sposobnosti pojedinca ili organizacije na temelju podataka o tim klijentima i služi kao procjena rizika prilikom davanja kredita, polica osiguranja, kreditnih kartica i sl. Proces kreditne klasifikacije vezanog uz kredit može se podijeliti u dva tipa. Jedan tip je kreditna klasifikacija prije samog odobrenja kredita, gdje se procjenjuje sposobnost vraćanja kredita i treba li klijentu kredit biti odobren ili ne. Drugi tip nastupa nakon što je kredit odobren te služi kao procjena da će klijent u nekom razdoblju prestati s vraćanjem kredita.

U ovom radu, promatrat će se problem drugog tipa klasifikacije: na temelju podataka o klijentima kojima je kredit odobren, potrebno je uočiti one za koje se pretpostavlja da će prestati s vraćanjem kredita i njih klasificirati kao potencijalno loše klijente. Smatra se da je došlo do neispunjavanja obaveza, tj. do *defaulta* ako klijent nije podmirio obaveze u razdoblju od tri mjeseca.

Podaci na kojoj će se provesti analiza kreditne sposobnosti preuzeti su s platforme Kaggle (http://www.kaggle.com/c/GiveMeSomeCredit).

2.2 Deskriptivna statistika

Preuzeti podaci sadrže informacije o 150 000 klijenata.
Za svakog klijenta dane su sljedeće varijable:

- **SeriousDlqin2yrs**: Osoba je imala zakašnjenje u plaćanju od 90 dana ili više.
- **Unname**: Brojevi od 1 do 150 000.
- **RevolvingUtilizationOfUnsecuredLines**: Ukupan iznos duga na kreditnim karticama i kreditnim linijama (osim nekretnina i bez obročnih dugova) podijeljen sa zbrojem kreditnih limita.
- **age**: Dob klijenta u godinama.
- **NumberOfTime30-59DaysPastDueNotWorse**: Broj puta koliko je klijent kasnio 30-59 dana s vraćanjem kredita, ali ne više, u zadnje 2 godine.
- **NumberOfTime60-89DaysPastDueNotWorse**: Broj puta koliko je klijent kasnio 60-89 dana s vraćanjem kredita, ali ne više, u zadnje 2 godine.
- **NumberOfTimes90DaysLate**: Broj puta koliko je klijent kasnio s vraćanjem kredita 90 dana ili više.
- **DebtRatio**: Omjer mjesečnih troškova (otplata duga, alimentacija, troškovi života) i mjesečnog bruto prihoda.
- **MonthlyIncome**: Mjesečni prihod.
- **NumberOfOpenCreditLinesAndLoans**: Broj dugova (obročnih poput auto kredita ili hipoteke) i kreditnih linija (npr. kreditne kartice).
- **NumberRealEstateLoansOrLines**: Broj hipotekarnih i stambenih kredita uključujući kreditne linije stambenog kapitala.

- **NumberOfDependents**: Broj uzdržavanih članova obitelji, ne uključujući samog klijenta (supružnik, djeca itd.).

Varijabla *Unname* sadrži samo brojeve, koji služe kao indeks u podacima pa se može izbaciti jer nije korisna.
Sljedeće slike prikazuju statistički sažetak varijabli gledajući minimum i maksimum svake varijable, kvartile, srednju vrijednost i standardnu devijaciju:

	SeriousDlqin2yrs	RevolvingUtilizationOfUnsecuredLines	age	NumberOfTime30-59DaysPastDueNotWorse	DebtRatio	MonthlyIncome
count	150000.000	150000.000	150000.000	150000.000	150000.000	120269.000
mean	0.067	6.048	52.295	0.421	353.005	6670.221
std	0.250	249.755	14.772	4.193	2037.819	14384.674
min	0.000	0.000	0.000	0.000	0.000	0.000
25%	0.000	0.030	41.000	0.000	0.175	3400.000
50%	0.000	0.154	52.000	0.000	0.367	5400.000
75%	0.000	0.559	63.000	0.000	0.868	8249.000
max	1.000	50708.000	109.000	98.000	329664.000	3008750.000

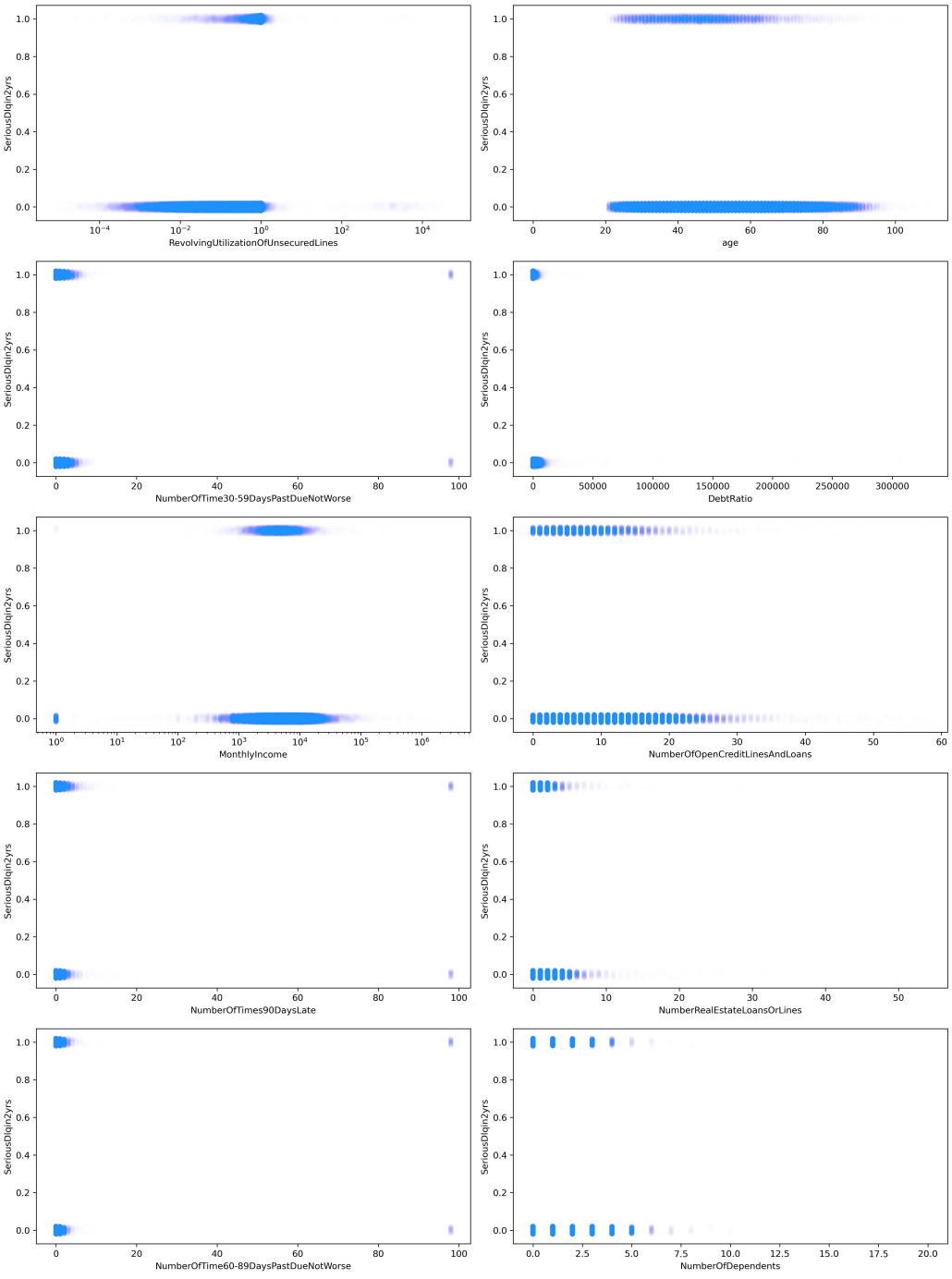
Slika 2.1: Statistički sažetak prvih šest varijabli.

	NumberOfOpenCreditLinesAndLoans	NumberOfTimes90DaysLate	NumberRealEstateLoansOrLines	NumberOfTime60-89DaysPastDueNotWorse	NumberOfDependents
count	150000.000	150000.000	150000.000	150000.000	146076.000
mean	8.453	0.266	1.018	0.240	0.757
std	5.146	4.169	1.130	4.155	1.115
min	0.000	0.000	0.000	0.000	0.000
25%	5.000	0.000	0.000	0.000	0.000
50%	8.000	0.000	1.000	0.000	0.000
75%	11.000	0.000	2.000	0.000	1.000
max	58.000	98.000	54.000	98.000	20.000

Slika 2.2: Statistički sažetak preostalih pet varijabli.

Sve dane varijable su kvantitativne, osim varijable *SeriousDlqin2yrs* koja poprima vrijednosti 0 (u slučaju da klijent vraća kredit) ili 1 (ukoliko je došlo do neispunjavanja obaveza) te će to biti zavisna varijabla u modelu. Broj klijenata koji su prestali vraćati kredit je 10 026, što je otprilike 6.7% ukupnog broja klijenata iz podataka.

Slika 2.3 za svaku varijablu prikazuje graf u ovisnosti s varijablom *SeriousDlqin2yrs*, pri čemu je na y-osi varijabla *SeriousDlqin2yrs*.



Slika 2.3: Ovisnost varijabli obzirom na varijablu SeriousDlqin2yrs

Može se vidjeti da su varijable *NumberOfTime30-59DaysPastDueNotWorse*, *NumberOfTime60-89DaysPastDueNotWorse*, *NumberOfTimes90DaysLate* u prosjeku ispod jedan i da većina ljudi nije kasnila s plaćanjem kredita. Međutim, sve tri varijable poprimaju nekoliko velikih vrijednosti, točnije 96 i 98, a kako se promatra samo period unazad dvije godine, te vrijednosti ne bi trebale biti veće od 24. Iz tog razloga, ti podaci će biti izbačeni.

Za varijablu *RevolvingUtilizationOfUnsecuredLines* koja označava omjer ukupnog duga i dopuštenog limita očekuje se da poprima vrijednosti manje ili jednake jedan. Vrijednost maksimuma te varijable ukazuje na to da postoje ekstremne vrijednosti u podacima. Kako veći broj klijenata (oko 2%) ipak ima vrijednost te varijable iznad jedan, a manji broj vrijednost iznad dva (oko 0.2%), izbačeni su podaci o klijentima čija je vrijednost varijable *RevolvingUtilizationOfUnsecuredLines* veća od dva.

Za neke klijente nedostaju podaci o *MonthlyIncome* i *NumberOfDependents*. Točnije, za 29 731 klijenata nedostaje podatak o mjesečnom prihodu te za 3924 klijenata nedostaje podatak o broju uzdržavanih članova obitelji.
Velike vrijednosti varijable *DebtRatio* odgovaraju upravo klijentima kojima nedostaje podatak o mjesečnom prihodu. Obzirom da je takvih klijenata puno, umjesto izbacivanja, vrijednosti za *MonthlyIncome* zamijenjene su medijanom. Analogno, tim klijentima, i vrijednosti varijable *DebtRatio* postavljene su na vrijednost medijana.

Klijenti kojima nedostaje podatak o varijabli *NumberOfDependents* su izbačeni.
Ostale varijable nisu promijenjene.
Sveukupno, preostali podaci sadrže informacije o 145 481 klijenata.

2.3 Odabir modela i prilagodba modela podacima

Sljedeći je korak pronaći varijable koje su značajne za model te vidjeti koliko se dobro model prilagodio podacima.

Obzirom da varijable, ne uzimajući u obzir *SeriousDlqin2yrs*, međusobno nemaju veliku korelaciju, početni model uključivat će sve dane varijable.
Slika 2.4 prikazuje rezultate provedene logističke regresije.

Dep. Variable:	SeriousDlqin2yrs	**No. Observations:**	97472
Model:	Logit	**Df Residuals:**	97461
Method:	MLE	**Df Model:**	10
Date:	-	**Pseudo R-squ.:**	0.2328
Time:	-	**Log-Likelihood:**	-18065.
converged:	True	**LL-Null:**	-23548.
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.3306	0.069	-48.103	0.000	-3.466	-3.195
RevolvingUtilizationOfUnsecuredLines	1.9700	0.042	47.402	0.000	1.889	2.051
age	-0.0192	0.001	-16.147	0.000	-0.021	-0.017
NumberOfTime30_59DaysPastDueNotWorse	0.4328	0.014	31.094	0.000	0.405	0.460
DebtRatio	-5.574e-05	5.09e-05	-1.096	0.273	-0.000	4.4e-05
MonthlyIncome	-1.983e-05	3.76e-06	-5.280	0.000	-2.72e-05	-1.25e-05
NumberOfOpenCreditLinesAndLoans	0.0274	0.003	8.469	0.000	0.021	0.034
NumberOfTimes90DaysLate	0.6392	0.021	31.079	0.000	0.599	0.680
NumberRealEstateLoansOrLines	0.0945	0.013	7.132	0.000	0.069	0.120
NumberOfTime60_89DaysPastDueNotWorse	0.5779	0.028	20.335	0.000	0.522	0.634
NumberOfDependents	0.0377	0.012	3.071	0.002	0.014	0.062

Slika 2.4: Sažetak rezultata logističkog modela

Značajnost nezavisnih varijabli testira se Waldovim testom, pri čemu testna statistika dana u 1.37 odgovara stupcu naziva z. Nulta hipoteza Waldovog testa je da parametar β_j uz varijablu x_j iznosi nula.
Kako su sve p-vrijednosti, osim za varijablu *DebtRatio*, manje od svih standardnih razina značajnosti nulta hipoteza se može odbaciti. Prema tome, sve varijable su statistički značajne, osim varijable *DebtRatio* čija p-vrijednost iznosi 0.273.
Rezultati modela logističke regresije s izbačenom varijablom *DebtRatio* prikazani su slikom 2.5.

Dep. Variable:	SeriousDlqin2yrs	**No. Observations:**	97472
Model:	Logit	**Df Residuals:**	97462
Method:	MLE	**Df Model:**	9
Date:	-	**Pseudo R-squ.:**	0.2328
Time:	-	**Log-Likelihood:**	-18065.
converged:	True	**LL-Null:**	-23548.
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.3320	0.069	-48.134	0.000	-3.468	-3.196
RevolvingUtilizationOfUnsecuredLines	1.9698	0.042	47.400	0.000	1.888	2.051
age	-0.0192	0.001	-16.167	0.000	-0.022	-0.017
NumberOfTime30_59DaysPastDueNotWorse	0.4328	0.014	31.100	0.000	0.406	0.460
MonthlyIncome	-1.93e-05	3.71e-06	-5.204	0.000	-2.66e-05	-1.2e-05
NumberOfOpenCreditLinesAndLoans	0.0274	0.003	8.449	0.000	0.021	0.034
NumberOfTimes90DaysLate	0.6395	0.021	31.087	0.000	0.599	0.680
NumberRealEstateLoansOrLines	0.0937	0.013	7.086	0.000	0.068	0.120
NumberOfTime60_89DaysPastDueNotWorse	0.5778	0.028	20.330	0.000	0.522	0.634
NumberOfDependents	0.0373	0.012	3.036	0.002	0.013	0.061

Slika 2.5: Sažetak rezultata logističkog modela

Testom omjera vjerodostojnosti može se testirati treba li uzeti model koji ne uključuje varijablu *DebtRatio* ili ipak treba ostaviti početni model.
Nulta hipoteza testa omjera vjerodostojnosti kaže da je dovoljan manji model, dok alternativna kaže da je potreban veći model.
Koristeći dane vrijednosti log-vjerodostojnosti i činjenicu da testna statistika 1.36 ima χ^2 distribuciju s jednim stupnjem slobode, dobije se da je p-vrijednost testa 0.20066. Prema tome, nulta hipoteza se ne može odbaciti pa je zaključak da je manji model dovoljan.

Model logističke regresije za dani skup podataka je oblika:

$$p(Y = 1|X_i) = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_9 X_9}}{1 + e^{\beta_0+\beta_1 X_1+\ldots+\beta_9 X_9}}. \tag{2.1}$$

X_1 je varijabla *RevolvingUtilizationOfUnsecuredLines*, X_2 je varijabla *age*, X_3 je varijabla *NumberOfTime30_59PastDueNotWorse* i tako redom kao na slici 2.5 do X_9 što je *NumberOfDependents*. Vrijednosti procijenjenih parametara β_j, $j = 0, 1, 2, ..., 9$ dane su u stupcu *coef* na slici 2.5.

Izraz 2.1 može se zapisati i obliku:

$$\log\left(\frac{p(Y=1|X_i)}{1-p(Y=1|X_i)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_9 X_9 \tag{2.2}$$

$$\iff \frac{p(Y=1|X_i)}{1-p(Y=1|X_i)} = e^{\beta_0+\beta_1 X_1+\ldots+\beta_9 X_9}, \tag{2.3}$$

što omogućuje interpretaciju parametara. Lijeva strana izraza 2.3 je pojam izgleda koji je definiran u prvom poglavlju. Ozbirom da su sve nezavisne varijable u modelu kvantitativne, pomoću izraza 1.11 slijedi da ukoliko se vrijednost j-te nezavisne varijable poveća za jedan, omjer izgleda mijenja se e^{β_j} puta.

Vrijednosti e^{β_j}, $j = 1, 3, 5, 6, 7, 8, 9$ su veće od jedan što znači da povećanje vrijednosti varijabli *RevolvingUtilizationOfUnsecuredLines, NumberOfTime30-59DaysPastDueNotWorse, NumberOfOpenCreditLinesAndLoans, NumberOfTimes90DaysLate, NumberRealEstateLoansOrLines, NumberOfTime60-89DaysPastDueNotWorse, NumberOfDependents* povećava izgled da će osoba prestati vraćati kredit.
Na primjer, za varijablu *NumberOfTime60-89DaysPastDueNotWorse*, parametar β_8 jednak je 0.5778, odnosno $e^{\beta_8} = 1.782118$. Ako su vrijednosti svih ostalih nezavisnih varijabli ostale iste, a broj puta kada je klijent kasnio s plaćanjem između 60 i 89 dana se poveća za jedan, izgledi defaulta se povećavaju 78%.

Parametri za nezavisne varijable *age* i *MontlyIncome* su negativni, odnosno vrijednosti e^{β_2} i e^{β_4} su manje od jedan. Iz toga slijedi da smanjenje vrijednosti tih varijabli povećava izgled da će osoba prestati vraćati kredit.
Na primjer, izgledi defaulta se povećavaju za faktor $e^{\beta_2} = 0.980989$ ako se broj godina poveća za jedan, a sve ostale nezavisne varijable ostanu iste.

Potrebno je još vidjeti koliko se dobro dobiveni model prilagodio podacima što je ključno za osiguravanje točnosti procijenjenih vjerojatnosti.

Važno je napomenuti kako podaci imaju informacije o 150 000 klijenata, međutim za odabir modela korišteno je njih 100 000 što je dvije trećine podataka. Razlog je taj što metrike za procjenu greške modela nisu objektivne na istom skupu podataka na kojem se model

trenira. Postupak razdvajanja podataka na skup za učenje i ispitni skup naziva se unakrsna validacija.

Dosadašnji model, za svakog klijenta, daje vjerojatnost da će klijent otići u default. Obzirom da je cilj razdvojiti klijente u dvije klase (oni koji neće vratiti kredit i oni koji hoće) potrebno je postaviti prag između nula i jedan. Svi kojima je vjerojatnost defaulta iznad praga bit će klasificirani rizičnima i pretpostavljat će se da ti klijenti neće vratiti kredit. Suprotno vrijedi za klijente čija je vjerojatnost defaulta manja od praga. U ovom primjeru, prag je postavljen na standardnih 0.5, ali se kasnije može vidjeti kako promjena praga utječe na vrijednost pojedine metrike.

Prilagodba modela logističke regresije, najčešće se testira pomoću Hosmer-Lemeshow testa, pri čemu je testna statistika dana s 1.33. Nažalost, test je problematičan za velike uzroke.
Dobivena p-vrijednost na testu za dane podatke je praktički nula, što znači da bi se trebala odbaciti nulta hipoteza koja kaže da je model adekvatan. Takav ishod je očekivan jer na tako velikom skupu podataka procijenjene vjerojatnosti bi trebale biti praktički jednake stvarnima kako p-vrijednost ne bi bila niska.
Osim Hosmer-Lemeshow testa koriste se razne druge metrike od kojih su neke obrađene u nastavku, a korisno je napraviti i rezidualnu analizu.

Mjere vrednovanja definirane su pomoću elemenata matrice zabune.
Matrica zabune (*eng. confusion matrix*) je kvadratna matrica oblika: $\begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix}$, pri čemu su elementi definirani na sljedeći način:

- stvarno pozitivni (*eng. true positive, TP*) - prava vrijednost i predviđena vrijednost zavisne varijable je 1
- lažno pozitivni (*eng. false positive, FP*) - prava vrijednost zavisne varijable je 0, a predviđena vrijednost zavisne varijable je 1
- lažno negativni (*eng. false negative, FN*) - prava vrijednost zavisne varijable je 1, a predviđena vrijednost zavisne varijable je 0
- stvarno negativni (*eng. true negative, TN*) - prava vrijednost i predviđena vrijednost zavisne varijable je 0.

Poznata mjera vrednovanja modela je *točnost (eng. accuracy)* koja označava koliko je od svih podataka njih točno klasificiranih. Takvu mjeru nema smisla promatrati na podacima čija je jedna klasa puno veća od druge jer je tada trivijalno ostvariti veliku točnost.

Ostale mjere koje se mogu koristiti za vrednovanje modela su sljedeće:

- preciznost: $\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$
- ispadanje: $\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP+TN}}$
- odziv: $\mathrm{R} = \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$
- specifičnost: $\mathrm{S} = \frac{\mathrm{TN}}{\mathrm{TN+FP}}$.

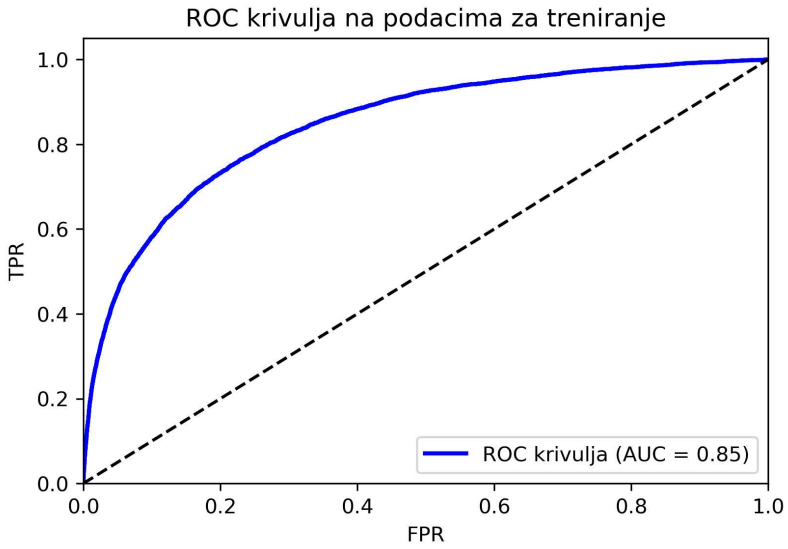
Preciznost je udio zavisnih varijabli koje su točno klasificirane kao 1 u skupu zavisnih varijabli koje su klasificirane kao 1 (i točno i netočno). Odziv je udio zavisnih varijabli koje su točno klasificirane kao 1 u skupu svih zavisnih varijabli čija je prava vrijednost 1.
Na neki način ove dvije metrike su suprotne, pa se može dodatno promotriti njihova harmonijska sredina. Ta metrika naziva se F-mjera i definirana je s: $\mathrm{F} = 2\frac{\mathrm{P \cdot R}}{\mathrm{P+R}}$.

Za odabrani model, dobije se da je matrica zabune za podatke iz skupa za učenje jednaka $\begin{pmatrix} 1018 & 700 \\ 5357 & 90397 \end{pmatrix}$. Matrica zabune za podatke iz ispitnog skupa je $\begin{pmatrix} 504 & 329 \\ 2779 & 44397 \end{pmatrix}$.

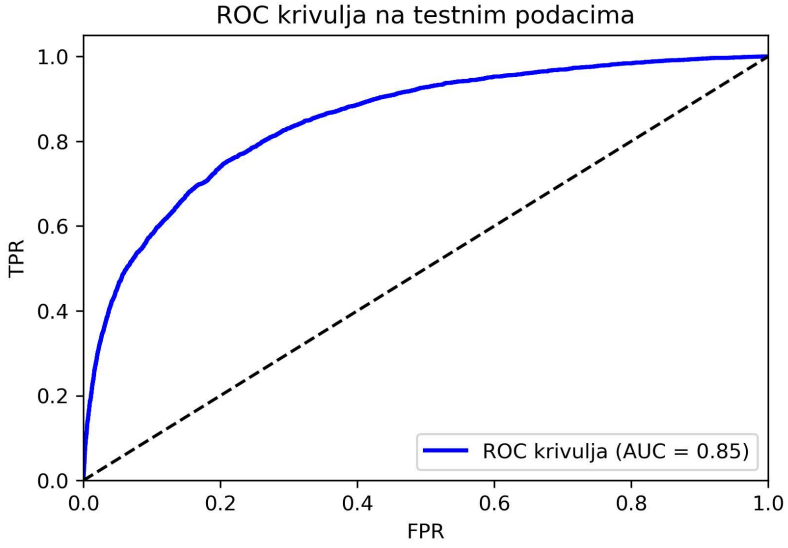
Ranije definirane metrike na ispitnom skupu su jednake: 0.605042 (preciznost), 0.007356 (ispadanje), 0.153518 (odziv), 0.992644 (specifičnost) i 0.244898 (F-mjera).
Naravno, više je slučaja kada je model krivo klasificirao zavisnu varijablu koja poprima vrijednost 1, od one koja poprima 0 jer je u podacima puno više ljudi koji jesu vratili kredit od onih koji nisu.

Jedna od najvažnijih mjera vrednovanja modela klasifikacije je ROC krivulja. Kako rezultati logističke regresije ovise o izboru praga, ova metrika je vrlo korisna jer uzima u obzir varijabilnost praga.
ROC krivulja prikazuje odnos između *ispadanja (FPR)* i *odziva (TPR)* pri različitima pragovima. Što je prag manji to je FPR veći jer je više zavisnih varijabli koje će krivo poprimiti vrijednost 1. Isto tako raste i TPR jer će model prepoznati više zavisnih varijabli čija je vrijednost 1. Poželjno je da je površina ispod ROC krivulje što veća, pa se tako može uvesti mjera čija je vrijednost jednaka toj površini. Ta mjera naziva se AUC.
Pravac y=x, odnosno AUC koji je jednak 0.5 predstavljaju slučaj ukoliko je klasifikator nasumičan, što znači da ukoliko ispadne da je AUC jednak 0.5, model radi jednako dobro kao i nasumičan klasifikator.

Za odabrani model, ROC krivulje za podatke iz skupa za učenje i ispitnog skupa prikazane su slikama 2.6 i 2.7.



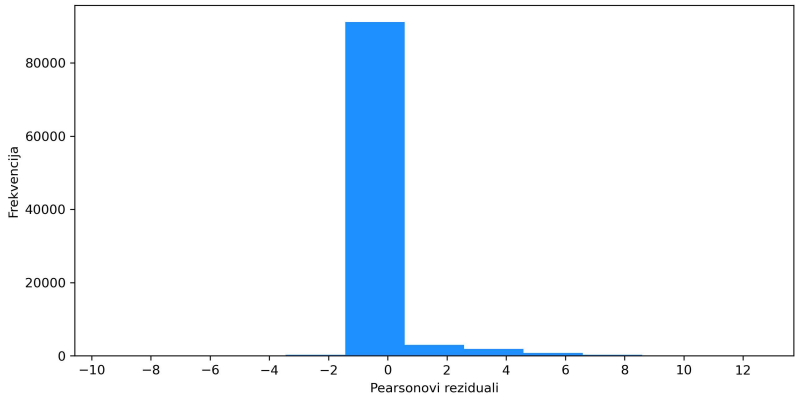
Slika 2.6: ROC i AUC na podacima iz skupa za učenje



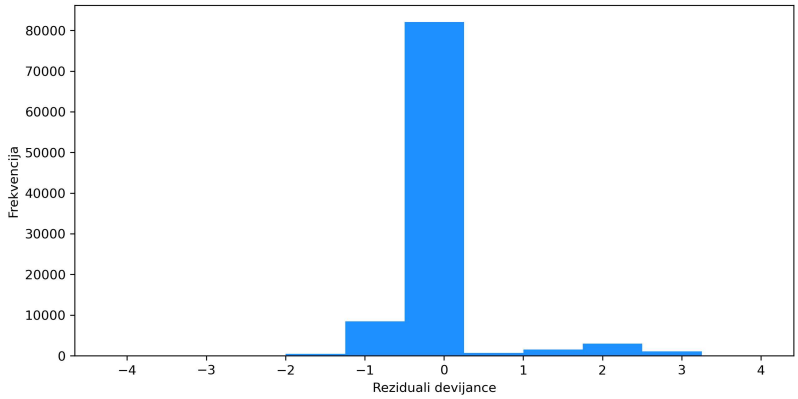
Slika 2.7: ROC i AUC na podacima iz ispitnog skupa

Za oba skupa podataka, rezultati su isti što znači da model jednako dobro radi na podacima na kojima je treniran kao i na neviđenim podacima.

Za provjeru modela, poželjno je napraviti i rezidualnu analizu.
U slučaju logističke regresije, najpoznatije vrste reziduala su Pearsonovi reziduali i reziduali devijance definirani u prvom poglavlju. Ako u modelu ima velik broj reziduala čija je vrijednost velika, to bi moglo sugerirati na činjenicu da model radi velike greške. Slike 2.8 i 2.9 prikazuju histograme reziduala.



Slika 2.8: Histogram Pearsonovih reziduala



Slika 2.9: Histogram reziduala devijance

Srednja vrijednost i standardna devijacija Pearsonovih reziduala jednake su -0.019901 i 1.057633, dok su kod reziduala devijance jednake -0.151320 i 0.589733. Histogram je asimetričan s više vrijednosti ispod nule, što je očekivano obzirom da je velika većina podataka u klasi gdje je zavisna varijabla jednaka 0. Većina reziduala nalazi se u intervalu $(-2, 2)$ a sve izvan toga su outlieri kojih nema puno.

Konačno, uz sve dosad prikazane rezultate i metrike može se zaključiti da se model dobro prilagodio podacima.

Dodatno, može se još promotriti kako se vrijednosti metrika mijenjaju ovisno o vrijednosti praga. Dane vrijednosti dane su slikom 2.10 za pragove između 0.1 i 0.9.

prag	Preciznost	Odziv	Specifičnost	Ispadanje	F	TP	FN	FP	TN
0.1	0.265105	0.630825	0.871641	0.128359	0.373321	2071	1212	5741	38985
0.2	0.445179	0.389583	0.964361	0.035639	0.415530	1279	2004	1594	43132
0.3	0.524161	0.280841	0.981286	0.018714	0.365728	922	2361	837	43889
0.4	0.564677	0.207432	0.988262	0.011738	0.303408	681	2602	525	44201
0.5	0.605042	0.153518	0.992644	0.007356	0.244898	504	2779	329	44397
0.6	0.636667	0.116357	0.995126	0.004874	0.196755	382	2901	218	44508
0.7	0.678832	0.084983	0.997049	0.002951	0.151056	279	3004	132	44594
0.8	0.685393	0.055742	0.998122	0.001878	0.103099	183	3100	84	44642
0.9	0.675000	0.024673	0.999128	0.000872	0.047605	81	3202	39	44687

Slika 2.10: Mjere vrednovanja za različite pragove

Zadnja četiri stupca prikazuju elemente matrice zabune. Može se uočiti da s povećanjem praga preciznost i specifičnost rastu, a odziv i ispadanje se smanjuju. Vrijednost F-mjere je najveća za prag 0.2, no to nužno ne mora biti najbolji odabir.
Prilikom odabira praga, može se uzeti u obzir i omjer broja opažanja između dvije promatrane klase. Kako je u ovom primjeru broj klijenata koji su u defaultu samo 6% od ukupnog broja klijenata, prag bi se mogao postaviti na 0.1. F-mjera ukazuje kako je ipak bolji model uz prag 0.2, što nije veliko odstupanje.
Treba uzeti u obzir i da financijska institucija ima veći gubitak na odobrenim kreditima koji nisu vraćeni, nego na neodobrenim kreditima koji bi bili vraćeni. Minimizacija gubitka se postiže kada je omjer tih gubitaka jednak omjeru lažno negativnih i lažno pozitivnih procjena.

Poglavlje 3

Algoritam SMOTE

U binarnoj klasifikaciji, može se dogoditi da je broj opažanja jedne klase puno veći nego broj opažanja druge klase. Takav problem naziva se problem nebalansiranih kategorija.
Podaci iz prošlog poglavlja odgovaraju tom problemu, što je čest slučaj s podacima vezanim uz kreditnu sposobnost jer je broj klijenata koji su vratili kredit obično puno veći od broja klijenata koji nisu vratili kredit.
Nebalansirane kategorije, odnosno klase uzrokuju da algoritmi za klasifikaciju postaju pristrani prema klasi s više opažanja. U tu svrhu uvedeni su mnogi algoritmi naduzorkovanja i poduzorkovanja kojima je cilj smanjiti razliku u broju opažanja među klasama. Ipak, algoritmi naduzorkovanja pokazali su se boljima jer se ne gubi originalna informacija o danim podacima u većoj klasi. Osim toga, postoje i drugi pristupi kao što je penalizirana logistička regresija koja dodatno rješava problem prekomjerne prilagođenosti.[10]
U ovom radu, za problem nebalansiranih kategorija, koristit će se jedan od poznatijih algoritama naduzorkovanja SMOTE.

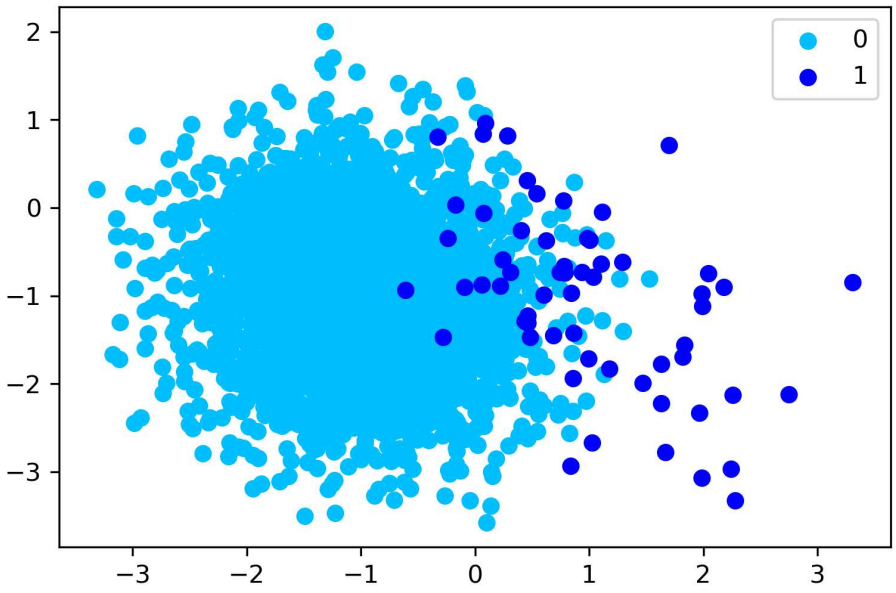
SMOTE (*Synthetic Minority Oversampling Technique*) je algoritam naduzorkovanja koji generira nova opažanja manjinske klase, tj. klase koja ima manje opažanja.
Glavna ideja SMOTE algoritma je da prilikom naduzorkovanja nova opažanja nisu ista dostupnima nego su generirana pomoću tih dostupnih opažanja.
Za dani broj opažanja manjinske klase, količinu SMOTE-a $N\%$ i broj najbližih susjeda k, algoritam generira nove podatke na način da za zadano opažanje manjinske klase i, nasumično odabire jednog od njegovih k najbližih susjeda. Novo opažanje je dobiveno pomoću ta dva opažanja za način da se razlika vrijednosti njihovih kovarijata (nezavisnih varijabli) pomnoži nasumičnim brojem između nula i jedan te se doda zadanom opažanju i.
Postupak se ponavlja za sva opažanja iz manjinske klase do željenog broja novih opažanja.
Broj novih opažanja ovisi o količini SMOTE-a N koji predstavlja koliko posto od broja opažanja manjinske klase je potrebno generirati. Ukoliko se želi broj opažanja manjinske

klase povećati za dva puta, tada je broj N jednak 100%.

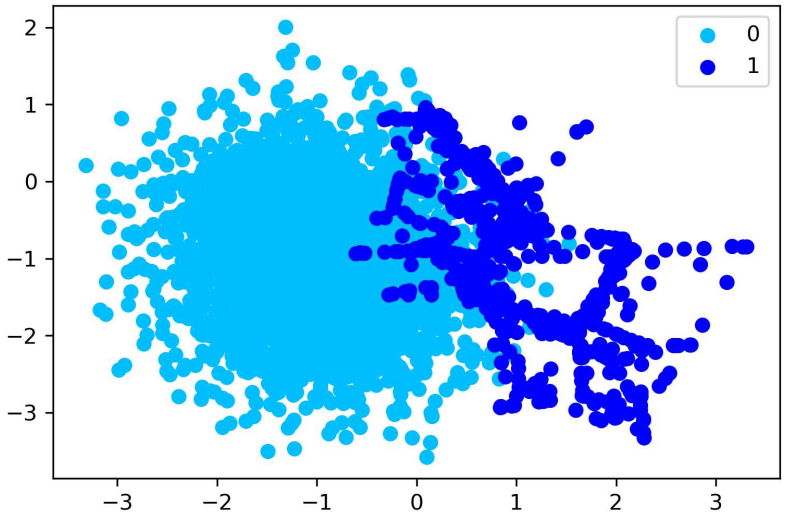
Pseudokod za algoritam SMOTE(T, N, k):
Ulaz: T - broj opažanja manjinske klase; N% količina SMOTE-a; k - broj najbližih susjeda
Izlaz: $(N/100) * T$ sintetičkih uzoraka za manjinsku klasu
1. (Ako je N manji od 100%, nasumično odredi opažanja manjinske klase jer je samo N% njih ulazi u algoritam.)
2. **ako** $N < 100$
3. **onda** nasumično odaberi T opažanja manjinske klase
4. $T = (N/100) * T$
5. $N = 100$
6. **završetak uvjeta**
7. N = (int)$(N/100)$
8. k = broj najbližih susjeda
9. broj_kovarijata = broj kovarijata
10. uzorak[ ][ ]: polje za originalne uzorke manjinske klase
11. novi_indeks: predstavlja broj generiranih sintetičkih uzoraka, inicijaliziran na 0
12. sinteticki[ ][ ]: polje za sintetičke uzorke
13. **za** i od 1 do T
14. izračunaj k najbližih susjeda za opažanje *i* te spremi njihove indekse u *polje_najblizih*
15. generiranje_uzoraka(N, i, polje_najblizih)
16. **kraj petlje**
generiranje_uzoraka(N, i, polje_najblizih) (Funkcija za generiranje sintetičkih uzoraka.)
17. **dok** $N \neq 0$
18. odaberi nasumičan broj između 1 i k, nazovi ga *nn* - ovaj korak odabire jednog od k najbližih susjeda opažanja i
19. **za** kovarijata od 1 do broj_kovarijata
20. izračunaj: razlika = uzorak[polje_najblizih[nn]][kovarijata] − uzorak[i][kovarijata]
21. izračunaj: razmak = nasumičan broj između 0 i 1
22. sinteticki[novi_indeks][kovarijata] = uzorak[i][kovarijata] + razmak ∗ razlika
23. **kraj petlje**
24. novi_indeks++
25. $N = N - 1$
26. **kraj petlje**
27. (kraj funkcije generiranje_uzoraka)
Kraj pseudokoda.

Sljedeće tri slike prikazuju, za umjetno generirane podatke, kako algoritam SMOTE utječe na opažanja manjinske klase.

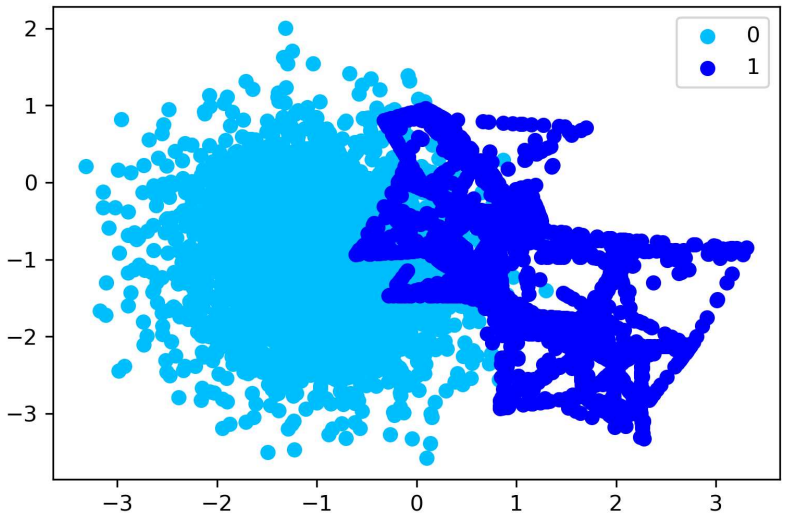
Slika 3.1 prikazuje podatke prije naduzorkovanja. Može se vidjeti kako je broj opažanja klase 1 znatno manji od broja opažanja klase 0. Cilj je primijeniti algoritam SMOTE na te podatke kako bi se povećao broj opažanja manjinske klase, odnosno kako bi se smanjio omjer između opažanja tih dviju klasa.
Slika 3.2 prikazuje podatke na kojima je primijenjen algoritam SMOTE, pri čemu je ukupan broj opažanja manjinske klase jednak trećini broja opažanja klase 0.
Slično, slika 3.3 prikazuje podatke na kojima je primijenjen algoritam SMOTE, ali tako da je broj opažanja klase 0 i klase 1 jednak.



Slika 3.1: Podaci bez primjene SMOTE-a



Slika 3.2: Podaci nakon SMOTE-a u omjeru 1:3



Slika 3.3: Podaci nakon SMOTE-a u omjeru 1:1

Može se vidjeti kako su opažanja manjinske klase koja su generirana SMOTE-om u okolini danih opažanja prije naduzorkovanja, što vizualizira generiranje podataka pomoću k najbližih susjeda.

Model iz prošlog poglavlja odabran je i testiran upravo na podacima čije su klase nebalansirane. Klijenata koji nisu vratiti kredit ima puno manje od klijenata koji su vratili kredit, točnije samo je 6% klijenata u defaultu.
Iako je pokazano kako se model dobro prilagodio podacima, poželjno je vidjeti može li se dodatno poboljšati primjenom algoritma SMOTE.

Slika 3.4 prikazuje mjere vrednovanja modela i elemente matrice zabune za prag 0.5. Prvi stupac, SMOTE %, označava koliki postotak od veće klase iznosi manjinska klasa. Na primjer, vrijednost 50% odgovara slučaju kada je broj opažanja, odnosno klijenata koji nisu vratili kredit duplo manji od broja klijenata koji jesu vratili kredit.

SMOTE %	Preciznost	Odziv	Specifičnost	Ispadanje	F	AUC	TP	FN	FP	TN
Bez SMOTE-a	0.605042	0.153518	0.992644	0.007356	0.244898	0.850086	504	2779	329	44397
10.0%	0.574413	0.201036	0.989067	0.010933	0.297834	0.847146	660	2623	489	44237
30.0%	0.356890	0.461468	0.938962	0.061038	0.402497	0.837091	1515	1768	2730	41996
50.0%	0.255595	0.619251	0.867616	0.132384	0.361840	0.835832	2033	1250	5921	38805
70.0%	0.216940	0.693573	0.816237	0.183763	0.330503	0.835713	2277	1006	8219	36507
100.0%	0.189948	0.750533	0.765058	0.234942	0.303168	0.835395	2464	819	10508	34218

Slika 3.4: Mjere vrednovanja nakon primjene SMOTE-a

Iz priloženog se može zaključiti kako postotak novogeneriranih opažanja može dosta utjecati na model. Vrijednosti AUC su se malo smanjile kako se postotak novogeneriranih opažanja povećavao. F-mjera se također mijenja i za razliku od 0.2449 koliko iznosi za model prije naduzorkovanja, povećava se na vrijednost 0.4025 u slučaju da manjinska klasa iznosi 30% veće klase. Također, u tom se slučaju broj stvarno pozitivnih (TP) povećao s 504 na 1515. To znači da je ovaj model točno prepoznao dodatnih 1011 klijenata koji nisu vratili kredit. Broj stvarno negativnih se u tom slučaju smanjio, kao i broj lažno negativnih obzirom da je sada više opažanja u kojima je klijent u defaultu.
Na slici 2.10 može se vidjeti kako se najveća vrijednost F-mjere postiže za prag 0.2. Slična vrijednost može se postići i za prag 0.5 primjenom SMOTE-a. Dodatno, broj stvarno pozitivnih je veći u slučaju SMOTE-a za čak 239 klijenata, dok je broj stvarno negativnih smanjen za otprilike 2300.

Iako je odabir između ta dva modela temeljen na individualnoj procjeni financijskih institucija, ukoliko je vrijednost stvarno pozitivnih od posebne važnosti, prikladniji model je onaj s pragom 0.5 uz algoritam SMOTE.

Bibliografija

[1] M. Anis i M. Ali, *Investigating the Performance of Smote for Class Imbalanced Learning: A Case Study of Credit Scoring Datasets*, European Scientific Journal **13** (2017), br. 33, 340–353.

[2] D. Collett, *Modelling Binary Data*, Chapman & Hall/CRC, 2014.

[3] J. Friedman, T. Hastie i Tibshirani R., *The Elements of Statistical Learning*, Springer, 2001.

[4] D. Hosmer, S. Lemeshow i R. Sturdivant, *Applied Logistic Regression*, Wiley, 2013.

[5] P. McCullagh i J. A. Nelder, *Generalized, Linear, and Mixed Models*, Chapman and Hall/CRC, 1989.

[6] C. E. McCulloch i S. R. Searle, *Generalized, Linear, and Mixed Models*, Wiley-Interscience, 2001.

[7] N. Sarapa, *Teorija vjerojatnosti*, Školska knjiga, 2012.

[8] S. N. Wood, *Generalized Additive Models: An Introduction with R*, Chapman Hall/CRC, 2017.

[9] J. M. Wooldridge, *Introductory Econometrics: A Modern Approach*, Cengage Learning, 2019.

[10] P. P. Šimović, C. Y. T. Chen i E. W. Sun, *Classifying the Variety of Customers' Online Engagement for Churn Prediction with a Mixed-Penalty Logistic Regression*, Computational Economics **61** (2023), 451–485.

Sažetak

Cilj ovog rada je klasificirati kreditnu sposobnost modelom logističke regresije.
U prvom poglavlju obrađeni su najvažniji pojmovi potrebni za razumijevanje logističke regresije. Prvo je objašnjeno kako se može definirati logistička regresija pomoću generaliziranih linearnih modela i koja je uloga funkcije veze logit. Zatim je dano objašnjenje kako se metodom maksimalne vjerodostojnosti procjenjuju parametri modela. Također, dani su najvažniji rezultati i testovi koji se koriste prilikom prilagodbe modela podacima.
Drugo poglavlje bazirano je na primjeni logističke regresije za klasifikaciju klijenata koji su kreditno sposobni i onih koji nisu. Nakon analize i čišćenja podataka u deskriptivnoj statistici, testirana je značajnost nezavisnih varijabli modela. Nakon odabira modela, dane su metrike kojima se ispitala točnost predikcija modela.
Treće poglavlje uzima u obzir da podaci imaju problem nebalansiranih kategorija. Iz tog razloga uveden je algoritam SMOTE kao pomoćni alat koji smanjuje razliku između broja opažanja dviju klasa. Na kraju se uspoređuju rezultati sa i bez SMOTE-a.

Summary

The aim of this thesis is credit score classification using a logistic regression model. In the first chapter, the key concepts necessary for understanding logistic regression are covered. First, it explains how logistic regression can be defined using generalized linear models and the role of the logit link function. Then, an explanation is given of how the parameters of the model are estimated using the maximum likelihood method. Additionally, the most important results and tests used for the goodness of fit of the model to the data are presented. The second chapter focuses on applying logistic regression for credit score classification. After analyzing and cleaning the data in descriptive statistics, the significance of the independent variables in the model is tested. Following the model selection, metrics are provided to evaluate the performance of the model's predictions. The third chapter addresses the issue of imbalanced data. For this reason, the SMOTE algorithm is introduced as an auxiliary tool to reduce the difference between the number of observations of the two classes. Finally, the results with and without the use of SMOTE are compared.

Životopis

Rođena sam 19. rujna 2000. godine u Zagrebu. Nakon završene Osnovne škole Josip Zorić, 2015. godine upisujem II. gimnaziju u Zagrebu. Obrazovanje nastavljam 2019. godine upisujući preddiplomski sveučilišni studij Matematika na Matematičkom odsjeku Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu te stječem akademski naziv sveučilišne prvostupnice matematike 2022. godine. Nakon toga, na istom odsjeku upisujem diplomski sveučilišni studij Financijske i poslovne matematike.