Primjena Bootstrap metode u kreditnom scoringu

Tomičević, Tena

Sveučilište J. J. Strossmayera u Osijeku
Fakultet primijenjene matematike i informatike
Sveučilišni diplomski studij matematike,
smjer: financijska matematika i statistika

Tena Tomičević

Primjena bootstrap metode u kreditnom scoringu

Diplomski rad

Osijek, 2023.

Sveučilište J. J. Strossmayera u Osijeku
Fakultet primijenjene matematike i informatike
Sveučilišni diplomski studij matematike,
smjer: financijska matematika i statistika

Tena Tomičević

Primjena bootstrap metode u kreditnom scoringu

Diplomski rad

Mentor: prof. dr. sc. Nataša Šarlija
Komentor: prof. dr. sc. Mirta Benšić

Osijek, 2023.

Sadržaj

1. Uvod

Istražujući podatke o udjelu malih i velikih poduzeća u gospodarstvu u Europi, podaci pokazuju kako je preko 98 % malih, a tek manje od 1 % velikih poduzeća (vidi [18]). Slična je situacija i u Hrvatskoj. Prema dostupnim podacima iz 2022. godine (vidi [19]), u Hrvatskoj su mala i srednja poduzeća činila udio od 99,7 % u gospodarstvu. Udio je zaposlenih u tim poduzećima također vrlo visok, čak 72,9 %. Ovi brojevi dovoljno govore o velikom značaju takvih poduzeća. Prilikom poslovanja, poduzeća često imaju potrebe za financijskim sredstvima. Neki se zadužuju radi razvitka poslovanja, a neki kako bi se izvukli iz problema. U slučaju kada se radi o malim i srednjim poduzećima, banke često imaju problem procjene rizika za takva poduzeća s obzirom da ona često ne raspolažu dovoljno velikim bazama podataka. Upravo je takva situacija opisana u ovom radu. Analizirana je baza podataka iz jedne banke koja sadrži podatke o financijskim izvještajima malih i srednjih poduzeća na temelju se kojih želi procijeniti kreditna sposobnost tih poduzeća. Postoji više načina kojima banke procjenjuju kreditnu sposobnost svojih klijenata, a jedan je od njih kreditni scoring koji je korišten u ovom slučaju. U programskom je jeziku R izrađen model za procjenu kreditnog rizika. Riječ je o modelu logističke regresije koji ima široku primjenu u praksi. Kako bi se procijenila kvaliteta modela, pouzdanih intervala i varijance, korištena je bootstrap metoda. Radi se o metodi koja se temelji na konceptu ponovnog uzorkovanja, tj. za odabir su uzorka korišteni podaci iz empirijske distribucije podataka. Glavna je ideja metode uzorkovanjem da se zaključci dobiveni na nekom uzorku generaliziraju na čitavu populaciju, pri čemu je naglasak na reprezentativnosti uzorka. Ukoliko to nije zadovoljeno, sama je vjerodostojnost dobivenih zaključaka upitna. Neke od takvih metoda su: jackknife metoda koja se koristi za procjenu varijanci[1], bootstrap metoda te „sub-sampling", odnosno poduzorkovanje[2](vidi [8]).

Bootstrap se metoda najčešće koristi za procjenu varijanci, kreiranje intervala pouzdanosti i testiranje hipoteza. U teorijskom je dijelu rada opisana sama metoda i bitni algoritmi, a primjena se bootstrap metode u kreditnom scoringu nalazi u empirijskom dijelu rada.

[1]Za procjenu se varijanci može koristiti i bootstrap metoda.

[2]Distribucija dobivena procjenom poduzoraka (uzorkovanje bez vraćanja) iz skupa podataka.

2. Teorijski dio: Bootstrap metoda

Bootstrap metoda

Princip je na kojem se temelji bootstrap metoda nastao zahvaljujući radu američkog statističara Bradleya Efrona 1979. godine. Ono je što je kod ove metode bitno za istaknuti da funkcionira i kada pretpostavka o normalnoj distribuciji promatranog uzorka nije zadovoljena (vidi [8]). Niti jedna metoda nije savršena pa tako i ova ima svoje nedostatke. Slučajevi kada bootstrap metoda nije pouzdana su:

- baza je podataka na kojoj se primjenjuje metoda nepotpuna, tj. neki podaci nisu izmjereni ili nisu dostupni,
- među promatranim se podacima nalaze stršeće vrijednosti, tj. neki podaci čije vrijednosti jako odudaraju od ostalih,
- postoji zavisnost među jedinkama uzorka (vidi [8]).

Poznato je nekoliko vrsta i načina za definiranje i opis bootstrap metode. Ovisno o tome je li unaprijed poznata distribucija parametara koji se procjenjuju, razlikuju se dvije vrste boostrapa, a to su parametarski i neparametarski. Obje će vrste biti objašnjene u nastavku ovog poglavlja (vidi [8]).
Kao što je rečeno u uvodu, ova se metoda oslanja na ponovljeno uzorkovanje s vraćanjem iz originalnog skupa podataka. Ta se tehnika može koristiti za procjenu standardne pogreške bilo koje statistike i dobivanje intervala pouzdanosti za parametar koji se procjenjuje. Inače se u literaturi slovom B označava broj bootstrap replikacija i uobičajene su vrijednosti koje se uzimaju 1000, 5000 ili čak 10000. Ne postoji univerzalni odgovor kolika je vrijednost najbolja, no činjenica je da se povećanjem B dobiva preciznija procjena. Promatrajući standardne greške i $p-$ vrijednosti statističkog testa primjenom bootstrap algoritma, dobivaju se informacije o preciznosti procjene (vidi [8]). Primjena je vrlo intuitivna, jer postoji algoritam koji se može primijeniti u ovisnosti o tome što se traži i što je zadano. Sama se metoda pokazala uspješnom u mnogim situacijama te je prihvaćena kao alternativa asimptotskim metodama. U nekim je situacijama uspješnija čak i od standardne normalne aproksimacije. Treba spomenuti i da postoje kontraprimjeri koji pokazuju kako bootstrap metoda može dati pogrešne zaključke u slučaju nekonzistentnih procjenitelja. Za iskaz i objašnjenje algoritma koji se u ovoj metodi primjenjuje, potrebno je definirati određene pojmove i uvesti oznake.

Definicija 1 *Statistički model zovemo model* ***jednostavnog slučajnog uzorka*** *iz funkcije distribucije F ako za slučajni vektor $\mathbf{X} = (X_1, \ldots, X_n)$ čiju realizaciju čine podaci $\mathbf{x} = (x_1, \ldots, x_n)$ vrijedi:*

- *slučajne su varijable $X_1, \ldots, X_n$ međusobno nezavisne,*
- *slučajne varijable $X_1, \ldots, X_n$ imaju jednaku funkciju distribucije F. (vidi [1], str. 201).*

Statistički modeli mogu biti parametarski i neparametarski. Kako je u radu korišten model logističke regresije[3], koji pripada skupini parametarskih statističkih modela, slijedi definicija upravo takvih modela.

[3]Detaljnije u poglavlju 2.7.

Definicija 2 *Statistički model čini familija funkcija distribucije*

$$\mathcal{P} = \{F_\theta : \theta \in \boldsymbol{\theta} \subseteq \mathbb{R}^n\},$$

pri čemu je F_θ funkcija distribucije slučajnog uzorka za dani θ, a $\boldsymbol{\theta}$ skup svih dozvoljenih vrijednosti parametra θ. Kako je model $\mathcal{P}$ indeksiran parametrom θ, ovakav statistički model naziva se ***parametarski*** *(vidi [1], str. 202).*

Prije definiranja neparametarskog bootstrapa, još je preostalo definirati empirijsku funkciju distribucije.

Definicija 3 *Neka je $\mathbf{X} = (X_1, \ldots, X_n)$ jednostavni slučajni uzorak iz nepoznate funkcije distribucije F slučajne varijable X, a $\boldsymbol{x} = (x_1, \ldots, x_n)$ njegova realizacija, tj. podaci (vidi [1], str. 215).*
Empirijska se funkcija distribucije *na temelju podataka $\boldsymbol{x}$ definira kao:*

$$\hat{F}(\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(x_i \leq \boldsymbol{u}),$$

pri čemu je $\boldsymbol{x} = (x_1, \ldots, x_n)$. Prava se funkcija distribucije za X definira kao:

$$F(\boldsymbol{u}) = \mathcal{P}(\boldsymbol{X} \leq \boldsymbol{u}).$$

Za dani $\boldsymbol{u} \in \mathbb{R}^n$, procjenitelj je $\hat{F}(\boldsymbol{u})$ konzistentan i nepristran za $F(\boldsymbol{u})$ (vidi [8]).

Neparametarski bootstrap

Za slučajni je uzorak $\mathbf{X} = (X_1, \ldots, X_n)$ iz nepoznate vjerojatnosne distribucije F pitanje kako procijeniti nepoznati parametar θ, pri čemu je parametar neka funkcija populacijske funkcije distribucije, tj. $\theta = t(F)$, a njegov procjenitelj ima oznaku $\hat{\theta}$. Računanjem vrijednosti procjenitelja $\hat{\theta}$ koji je funkcija slučajnog uzorka $\boldsymbol{X}$, $\hat{\theta} = s(\mathbf{X})$, dobiva se procjena parametra θ. Pitanje je koliko je ta procjena precizna? Cilj je dobiti procjenu čija je vrijednost što bliža vrijednosti pravog parametra θ. Pritom se pojavljuju dva ključna problema:

- F je nepoznata, a
- čak i da je F poznata, postoji mogućnost da je $\hat{\theta}$ toliko kompliciran funkcija od $\mathbf{X}$ da je traženje njegove distribucije izvan opusa ovoga rada. (vidi [17], str.7)

Ideja je neparametarskog bootstrapa simulacija podataka iz empirijske funkcije distribucije $\hat{F}$. Za svaki se $x \in \mathbb{R}$, vrijednost $F(\mathbf{x})$ procjenjuje s $\hat{F}(\mathbf{x})$. Uzorak se dobiven bootstrap metodom definira kao slučajni uzorak veličine n iz empirijske distribucije $\hat{F}(\mathbf{x})$ i bit će označen s $(x_1^*, \ldots, x_n^*) = \mathbf{x}^*$, što se može zapisati kao:

$$\hat{F} \to (x_1^*, \ldots, x_n^*). \tag{1}$$

Dakle, bootstrap se uzorak dobiva uzorkovanjem s vraćanjem iz $\mathbf{x} = (x_1, \ldots, x_n)$, a realizacije su označene s $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$. Pritom nije neobično da se neki elementi u uzorku pojave jednom ili više puta, a neki niti jednom. Moguća je primjerice situacija: $x_1^* = x_7, x_2^* = x_3, x_3^* = x_3, \ldots, x_n^* = x_7$ (vidi [6], str. 45). Analogno gornjim oznakama, bootstrap će replikacije od $\hat{\theta}$ biti označene s $\hat{\theta}^* = s(\mathbf{x}^*)$.

2.1. Bootstrap algoritam

Neka je $(X_1, \ldots, X_n)$ model jednostavnog slučajnog uzorka[4] iz funkcije distribucije F, pri čemu je F nepoznata distribucija. Poznato je kako je aritmetička sredina uzorka (oznaka $\bar{X}_n$) jedan nepristran i konzistentan procjenitelj za očekivanje. (vidi [1], str. 221). Postupak kada se za procjenu parametra populacijske distribucije F koriste svojstva distribucije $\hat{F}$ se u literaturi još naziva i „plug-in" princip i važno ga je shvatiti s obzirom da se na njemu zasniva i bootstrap metoda. Pretpostavka je da je parametar θ populacijsko očekivanje, tj. $\theta = t(F)$, što se može zapisati kao: $\theta = t(F) = E_F(X)$. „Plug-in" se procjenitelj parametra θ definira kao: $\hat{\theta} = t(\hat{F})$. Slijedi da je:

$$\hat{\theta} = E_{\hat{F}}(X) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}. \tag{2}$$

Potrebno je zamijetiti da se na temelju uzorka može dobiti i više od same procjene $\bar{X}$. Štoviše, može se izračunati i $\hat{\sigma}$, čime se dobiva procjena preciznosti za procjenitelja očekivanja, tj.

$$\hat{\sigma} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(X_i - \bar{X})^2}, \tag{3}$$

pri čemu je $\hat{\sigma}$ procjenitelj standardne devijacije od $\bar{X}$, odnosno korijen procijenjene srednje kvadratne greške. Jedina je mana ove formule što se ne može proširiti na druge procjenitelje. Imaju li procjenitelji ista svojstva ako su u pitanju bootstrap uzorci ili ne? Odgovor slijedi u nastavku.
Za dani bootstrap uzorak $\boldsymbol{X}^*$ iz (1) slijedi da se za

$$\bar{X}^* = \frac{1}{n}\sum_{i=1}^{n} X_i^*$$

može izračunati uvjetna varijanca[5] $Var^*\bar{X}^*$ koja se odnosi na varijancu bootstrap uzorka $\boldsymbol{X}^*$:

$$Var^*\bar{X}^* = \frac{1}{n^2}\sum_{i=1}^{n}(X_i^* - \bar{X}^*)^2. \tag{4}$$

Slijedi da se bootstrap procjenitelj standardne devijacije za $\hat{\theta}$ računa kao:

$$\hat{\sigma}^* = \sqrt{Var^*\hat{\theta}(X_1^*, \ldots, X_n^*)}. \tag{5}$$

Ukoliko se usporede (3) i (5) može se vidjeti da je za $\hat{\theta} = \bar{X}$, $\hat{\sigma}^* = \sqrt{\frac{n-1}{n}}\hat{\sigma}$, tj. ako se (5) pomnoži s $\sqrt{\frac{n}{n-1}}$, slijedi da je $\hat{\sigma}^* = \hat{\sigma}$. Još je preostalo prikazati kako bi izgledala bootstrap procjena standardne greške od $\hat{\theta}$. Analogno gornjim rezultatima slijedi kako je bootstrap procjena $se_F(\hat{\theta})$ definirana sa $se_{\hat{F}}(\hat{\theta^*})$. Ukoliko se za $\hat{\theta}$ uzme uzoračka sredina, kao pod (2) dobiva se jako precizna procjena. Problem je što za druge odabire $\hat{\theta}$ ne postoji neka formula za koju bi ta procjena bila jednako precizna.

[4] Definicija 1.
[5] Varijanca uvjetna na empirijsku distribuciju podataka.

Tu u priču ulazi bootstrap algoritam koji u računalnim izračunima daje vrlo dobre aproksimacije numeričkim vrijednostima standardne greške $se_{\hat{F}}(\hat{\theta}^*)$.

Bootstrap algoritam za procjenu standardne greške:

1. Odabire se B nezavisnih bootstrap uzoraka $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_B^*$. Ti su podaci dobiveni uzorkovanjem s vraćanjem iz originalnog uzorka $\mathbf{x}$.
2. Procjenom se bootstrap replikacije svakog bootstrap uzorka dobije: $\hat{\theta}^*(b) = s(\mathbf{x}^*)$, $b = 1, \ldots, B$.
3. Standardna se greška $se_F(\hat{\theta})$ procjenjuje standardnom devijacijom uzorka B replikacija:

$$\hat{se}_B = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}^*(b) - \hat{\theta}^*(\cdot) \right) \right\}^{\frac{1}{2}} \tag{6}$$

pri čemu je

$$\hat{\theta}^*(\cdot) = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^*(b).$$

Za jako je veliki B,$\hat{se}_B$ približno jednaka $se_{\hat{F}}$, odnosno s velikim je brojem ponavljanja empirijska standardna devijacija uzorka približno jednaka standardnoj greški bootstrap uzorka (vidi [3], str. 47).

Parametarski bootstrap

Osnovna je razlika između parametarskog i neparametarskog bootstrapa što je za parametarski funkcija distribucije poznata, a u neparametarskom se ne pretpostavlja tip distribucije, već se simulira iz empirijske distribucije. Dakle, pretpostavlja se da je oblik funkcije distribucije F nepoznat.
Kao i ranije, neka je $\theta = t(F)$, a s $\hat{\theta}$ je označena vrijednost procjenitelja koji je funkcija slučajnog uzorka $\boldsymbol{X}$, $\hat{\theta} = s(\boldsymbol{X})$. Kako bi se izračunao $\hat{\theta}$, koriste se komplicirani i dugotrajni izračuni koji se mogu izbjeći korištenjem simulacija na računalu. Generiranjem velikog broja uzoraka (B), pri čemu je svaki uzorak veličine n iz distribucije F, za svaki se uzorak računa $\hat{\theta}$. U slučaju kada je polazni model parametarski[6], funkcija će distribucije biti označena s F_θ i tada se za procjenu koristi parametarski bootstrap. Ovdje se bootstrap uzorak dobiva simulacijom iz pretpostavljenog oblika funkcije distribucije F_θ s procijenjenim parametrima $\hat{\theta}_1, \ldots, \hat{\theta}_B$. Slijedi algoritamski prikaz parametarskog bootstrapa uz pretpostavku da je poznat oblik fukcije distribucije F_θ, a parametar θ koji dolazi iz skupa svih dozvoljenih vrijednosti parametara $\boldsymbol{\theta}$ je nepoznat.

Parametarski bootstrap algoritam

1. U parametarskom slučaju, umjesto iz F_θ, B nezavisnih uzoraka se generira iz $F_{\hat{\theta}}$, čime se dobiva bootstrap uzorak $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$.

2. Procjenom se bootstrap replikacije svakog bootstrap uzorka metodom maksimalne vjerodostojnosti dobivaju vrijednosti nepoznatih parametara $\hat{\theta}_1, \ldots, \hat{\theta}_B$, tj.
$\hat{\theta}^*(b) = s(\mathbf{x}^*)$, $b = 1, \ldots, B$.

3. Iz pretpostavljene funkcije distribucije pod (1) s procijenjenim parametrima $\hat{\theta}^*(b)$ iz koraka (2) slijedi simulacija podataka (vidi [3], str. 53-54).

[6] Definicija 2.

2.2. Distribucija bootstrap uzorka

Neka je $\mathbf{X}^*$ nasumično odabran iz uzorka $\{\mathbf{X_1}, \ldots, \mathbf{X_n}\}$. Postavlja se pitanje kako bi izgledala distribucija od $\mathbf{X}^*$, uvjetovana promatranim podacima? Distribucija je od $\mathbf{X}^*$ uvjetovana podacima iz $\mathbf{X}$. S obzirom kako je empirijska funkcija distribucije $\hat{F}$ diskretna distribucija s vjerojatnošću $1/n$ za svaki $\mathbf{X_i}$, čak i ako originalni podaci dolaze iz neprekidne distribucije, bootstrap je distribucija podataka svakako diskretna.
Uvjetna se srednja vrijednost i uvjetna varijanca od $\mathbf{X}^*$ izračunavaju iz empirijske funkcije $\hat{F}(x)$ za svaki $x \in \mathbb{R}$ i jednaki su srednjoj vrijednosti uzorka i varijanci podataka, što u zapisu izgleda ovako:

$$\mathbb{E}^*(\mathbf{X}^*) = \sum_{i=1}^{n} \mathbf{X_i} \mathbb{P}^*(\mathbf{X}^* = \mathbf{X_i}) = \sum_{i=1}^{n} \mathbf{X_i} \frac{1}{n} = \overline{\mathbf{X}} \tag{7}$$

$$\begin{aligned} Var^*(\mathbf{X}^*) &= \mathbb{E}^*(\mathbf{X}^*\mathbf{X}^{*'}) - (\mathbb{E}^*(\mathbf{X}^*))(\mathbb{E}^*(\mathbf{X}^*))' \\ &= \sum_{i=1}^{n} \mathbf{X_i}\mathbf{X_i}' \mathbb{P}^*(\mathbf{X}^* = \mathbf{X_i}) - \overline{\mathbf{X}}\overline{\mathbf{X}}' \\ &= \sum_{i=1}^{n} \mathbf{X_i}\mathbf{X_i}' \frac{1}{n} - \overline{\mathbf{X}}\overline{\mathbf{X}}' = \hat{\sum} \end{aligned} \tag{8}$$

Dakle, uvjetna je distribucija od $\mathbf{X}^*$ uz danu funkciju $\hat{F}$ diskretna distribucija na $\{\mathbf{X_1}, \ldots, \mathbf{X_n}\}$ s očekivanjem $\overline{\mathbf{X}}$ i matricom kovarijanci $\hat{\sum}$, za svaki $x \in \mathbb{R}$. (vidi [8]).

Distribucija očekivanja bootstrap uzorka

Kako bi se procijenile srednja vrijednost i varijanca na temelju realizacije bootstrap uzorka, korišteni su određeni procjenitelji (poglavlje 2.1.). Poznato je kako su aritmetička sredina uzorka i korigirana varijanca uzorka nepristrani i konzistentni procjenitelji za očekivanje i varijancu. Aritmetička je sredina realizacije bootstrap uzorka dana izrazom:

$$\bar{\mathbf{X}}^* = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{X_i^*}),$$

te je moguće izračunati njezino (uvjetno) očekivanje i varijancu. Očekivanje se može izračunati koristeći (7) na sljedeći način:

$$\mathbb{E}^*(\bar{\mathbf{X}}^*) = \mathbb{E}^* \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X_i^*} \right) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}^*(\mathbf{X_i^*}) = \frac{1}{n} \sum_{i=1}^{n} \bar{\mathbf{X}} = \bar{\mathbf{X}}. \tag{9}$$

Dakle, očekivanje bootstrap uzorka $\bar{\mathbf{X}}^*$ ima distribuciju centriranu oko aritmetičke sredine originalnog uzorka $\bar{\mathbf{X}}$. To je zato što su bootstrap opservacije $\mathbf{X_i^*}$ izvučene iz bootstrap populacije koja empirijsku funkciju smatra istinitom, a očekivanje je spomenute distribucije $\bar{\mathbf{X}}$. Primjenom (8), (uvjetna) varijanca srednje vrijednosti bootstrap uzorka izgleda ovako:

$$\mathrm{Var}^*(\bar{\mathbf{X}}^*) - \mathrm{Var}^* \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X_i^*} \right) - \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}^*(\mathbf{X_i^*}) - \frac{1}{n^2} \sum_{i=1}^{n} \bar{\Sigma} - \frac{1}{n} \bar{\Sigma}.$$

U skalarnom bi se slučaju dobilo da je $\text{Var}^*(\bar{X}^*) = \frac{\hat{\sigma^2}}{n}$ što pokazuje kako je i varijanca od $\bar{\mathbf{X}}^*$ opisana varijancom uzorka originalnih podataka. Razlog je tome ponovno činjenica da bootstrap uzorak $\mathbf{X}_\mathbf{i}^*$ dolazi iz bootstrap populacije.
U nastavku će biti navedeni još neki asimptotski rezultati bootstrap metode i način njenog korištenja. Stoga je potrebno uvesti oznaku za standardiziranu srednju vrijednost bootstrap uzorka. Neka je to $z_n^* = \sqrt{n}(\bar{\mathbf{x}}^* - \bar{\mathbf{x}})$, pri čemu je $\bar{\mathbf{x}}^*$ srednja vrijednost bootstrap uzorka, a $\bar{\mathbf{x}}$ srednja vrijednost originalnog uzorka (vidi [8]).

2.3. Interval percentila

Bootstrap metoda ima mnoge primjene u statističkim izračunima pa se između ostalog može koristiti i za procjenu varijance uzorka (vidi [2]). Kao što je ranije navedeno u poglavlju 2.1., za skalarni procjenitelj $\hat{\theta}$ bootstrap je standardna pogreška kvadratni korijen bootstrap procjenitelja varijance. Ovu je statistiku prilično lako izračunati i vrlo je česta uporaba bootstrap metode u primijenjenoj ekonometrijskoj praksi. Standardne se pogreške često koriste za konstrukciju pouzdanih intervala. Bootstrap se standardne pogreške također mogu koristiti u tu svrhu. Pouzdani je interval normalne aproksimacije zapisan na sljedeći način:

$$C^{nb} = \left[\hat{\theta} - z_{(1-\alpha)/2}\hat{se}_B, \hat{\theta} + z_{(1-\alpha)/2}\hat{se}_B\right],$$

pri čemu je $z_{(1-\alpha)/2}$ $(1-\alpha)/2$ kvantil standardne normalne $\mathcal{N}(0,1)$ distribucije. Ovaj je pouzdani interval po formulaciji jednak asimptotskom intervalu pouzdanosti, samo je razlika u tome što se umjesto asimptotske standardne greške ovdje javlja bootstrap standardna greška. Postoje i druge metode, kao što je primjerice BC percentilna metoda[7] koja također nije komplicirana za primjenu, ali ima bolji učinak. Nužno je naglasiti kako bi se bootstrap standardne pogreške trebale više koristiti kao procjene preciznosti, nego kao alat za konstruiranje pouzdanih intervala. Također, treba biti oprezan kod uspoređivanja standardnih grešaka dobivenih različitim metodama. Ukoliko su razlike u procjeni standardnih grešaka velike, to znači da bi odabrane metode mogle biti nepouzdane.
Sve su bootstrap statistike procjene te su nasumične i konačne pošto je B konačan. Njihove vrijednosti variraju kroz različiti izbor za B. Stoga ne treba očekivati da će dobivene bootstrap standardne pogreške biti jednake kao i kod nekih drugih istraživača pri replikaciji njihovih rezultata. Trebali bi se dobiti slični, ali ne i potpuno jednaki rezultati (vidi [8]).

Kao što je ranije spomenuto, bootstrap se metoda vrlo često koristi za izračunavanje intervala pouzdanosti. Jedna je od vrlo popularnih metoda za to interval percentila. Radi se o jednostavnoj metodi koja se temelji na kvantilima bootstrap distribucije.
Bootstrap se algoritmom dobiva nezavisni jednako distribuirani uzorak bootstrap procjena $\left\{\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*\right\}$parametra θ. Radi jednostavnosti, neka je θ skalarni parametar. Za bilo koji $0 < \alpha < 1$ može se izračunati empirijski kvantil $\hat{q}_\alpha^*$ dobivenih bootstrap procjena. To je broj za koji vrijedi da su $n\alpha$ bootstrap procjene manje od $\hat{q}_\alpha^*$.
Bootstrap percentil $100(1-\alpha)\%$ pouzdanog intervala je

$$C^{pc} = \left[q_{\alpha/2}^*, q_{(1-\alpha)/2}^*\right].$$

[7]Detaljnije u [8].

Primjerice, ako je $B = 1000$, $\alpha = 0.05$, koristeći empirijski procjenitelj kvantila dobiva se da je $C^{pc} = \left[\hat{\theta}^*_{(25)}, \hat{\theta}^*_{(975)}\right]$. Percentilni interval ima prednost jer ne zahtjeva izračun standardne pogreške. On je naprosto nusproizvod bootstrap algoritma i ne uzrokuje značajne računalne troškove (vidi [8]).

2.4. Bootstrap asimptotika

Prema ranije je dobivenim rezultatima zaljučak kako je bootstrap procjena očekivanja srednje vrijednosti uzorka $\bar{\mathbf{x}}^*$ jednaka prosjeku n nezavisnih i jednako distribuiranih slučajnih varijabli iz originalnog uzorka (poglavlje 2.1.). Slijedi da bi srednja vrijednost bootstrap uzorka[8] trebala konvergirati po vjerojatnosti prema očekivanju originalnog uzorka. Međutim, treba biti na oprezu s obzirom da je ranije pokazano kako srednja vrijednost bootstrap uzorka ima uvjetnu distribuciju uz dane podatke, stoga će i konvergencija po vjerojatnosti biti definirana za uvjetne distribucije.

Definicija 4 (Bootstrap konvergencija po vjerojatnosti) *Kažemo da slučajni vektor z_n^* konvergira po bootstrap vjerojatnosti u z kada $n \longrightarrow \infty$ (pišemo $z_n^* \xrightarrow{p^*} z$), ako za svaki $\varepsilon > 0$*

$$\mathbb{P}^*(\| z_n^* - z \| > \varepsilon) \xrightarrow{p} 0.$$

Za dovoljno veliki uzorak n, vjerojatnost da je z_n daleko od z postaje sve manja, odnosno teži u 0. Glavni su alati koji se koriste u asimptotskoj teoriji **slabi zakon velikih brojeva** (SZVB) i **centralni granični teorem**(CGT) (vidi [8]). Primjenom tih teorijskih rezultata mogu se aproksimirati uzoračke distribucije procjenitelja. Sada slijede teoremi čiji se općeniti iskaz može naći u [8], a ovdje će se odnositi na bootstrap uzorke.

Teorem 1 *Ako $z_n \xrightarrow{p} z$, kada $n \to \infty$, onda $z_n \xrightarrow{p^*} z$.*

Teorem 2 (Bootstrap Slabi zakon velikih brojeva) *Ako su $\mathbf{x_i}$ nezavisni i uniformno integrabilni, tada*

$$\bar{\mathbf{x}}^* - \bar{\mathbf{x}} \xrightarrow{p^*} 0 \ \ i \ \ \bar{\mathbf{x}}^* \xrightarrow{p^*} \mathbb{E}(\mathbf{x_i}),$$

kada $n \to \infty$.

Treba imati na umu kako su uvjeti za bootstrap SZVB jednaki kao i za općeniti SZVB (vidi [1], str. 164). Osim toga, važno je naglasiti kako se razlika srednje vrijednosti bootstrap uzorka $\bar{\mathbf{x}}^*$ i srednje vrijednosti originalnog uzorka $\bar{\mathbf{x}}$ smanjuje kako se veličina uzorka mijenja. S obzirom da srednja vrijednost originalnog uzorka konvergira po vjerojatnosti u očekivanje tog uzorka, ne treba čuditi kako i srednja vrijednost bootstrap uzorka teži prema istom očekivanju. Kako bi se pokazalo da srednja vrijednost bootstrap uzorka ima asimptotski normalnu distribuciju, potrebno je definirati uvjetnu konvergenciju po distribuciji.

Definicija 5 (Konvegrencija po bootstrap distribuciji) *Neka je z_n^* slučajni vektor s uvjetnom distribucijom $\mathrm{G}_\mathrm{n}^*(\mathrm{u}) = \mathbb{P}^*(z_n^* \leqslant u)$. Kažemo da z_n^* konvergira po bootstrap distribuciji prema z, kada $n \to \infty$ [pišemo $z_n^* \xrightarrow{d^*} z$], ako za svaki u za koji je $\mathrm{G}(\mathrm{u}) = \mathbb{P}(z \leqslant u)$ neprekidna funkcija vrijedi:*

$$\mathrm{G}_\mathrm{n}^*(\mathrm{u}) \xrightarrow{p} \mathrm{G}(\mathrm{u}),$$

kada $n \to \infty$.

[8]U nastavku će istom oznakom biti označene slučajna varijabla i njena realizacija. Iz načina će korištenja tih oznaka biti jasno je li riječ o svojstvima slučajne varijable ili realnog broja.

Sada se može iskazati centralni granični teorem za bootstrap uzorke.

Teorem 3 (Bootstrap centralni granični teorem) *Ako su* $\mathbf{x_i}$ *nezavisne,* $\| \mathbf{x_i} \|^2$ *uniformno integrabilna i* $\mathbf{\Sigma} = Var(\mathbf{x}) > 0$*, tada*

$$\sqrt{n}(\bar{\mathbf{x}}^* - \bar{\mathbf{x}}) \xrightarrow{d^*} N(\mathbf{0}, \mathbf{\Sigma}),$$

kada $n \to \infty$.

Time je pokazano da $\bar{\mathbf{x}}^*$ ima asimptotski normalnu distribuciju s očekivanjem 0 i varijancom $\mathbf{\Sigma}$). Slijedi da je distribucija bootstrap uzorka asimptotski jednaka distribuciji originalnog uzorka. Slijedi iskaz teorema o delta metodi u bootstrap smislu (vidi [8]).

Teorem 4 (Bootstrap Delta metoda) *Ako* $\hat{\boldsymbol{\mu}} \xrightarrow{p} \boldsymbol{\mu}$, $\sqrt{n}(\hat{\boldsymbol{\mu}}^* - \hat{\boldsymbol{\mu}}) \xrightarrow{d^*} \mathbf{Z}$ *i* $g(u)$ *neprekidna diferencijabilna funkcija u okolini* $\boldsymbol{\mu}$*, kada* $n \to \infty$ *vrijedi:*

$$\sqrt{n}(g(\hat{\boldsymbol{\mu}^*}) - g(\hat{\boldsymbol{\mu}})) \xrightarrow{d^*} \mathbf{G}'\mathbf{Z},$$

pri čemu je $\mathbf{G}(u) = \frac{\partial}{\partial u} g(u)'$ *i* $\mathbf{G} = \mathbf{G}(\boldsymbol{\mu})$.
Nadalje, ako $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$*, vrijedi:*

$$\sqrt{n}(g(\hat{\boldsymbol{\mu}^*}) - g(\hat{\boldsymbol{\mu}})) \xrightarrow{d^*} \mathcal{N}(\mathbf{0}, \mathbf{G}'\mathbf{V}\mathbf{G}),$$

kada $n \to \infty$.

Teorem 5 *Ako su* $\mathbf{x_i}$ *nezavisne i jednako distribuirane,*
$\boldsymbol{\mu} = \mathbb{E}(\mathbf{h}(\mathbf{x}))$, $\theta = \mathbf{g}(\boldsymbol{\mu})$, $\mathbb{E} \| \mathbf{h}(\mathbf{x}) \|^2 < \infty$*, te je* $G(u) = \frac{\partial}{\partial u} g(u)'$ *neprekidna u okolini* $\boldsymbol{\mu}$ *za* $\hat{\theta} = g(\hat{\boldsymbol{\mu}})$*, pri čemu je* $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{h}(\mathbf{x_i})$ *i* $\hat{\theta}^* = g(\hat{\boldsymbol{\mu}}^*)$*, te* $\hat{\boldsymbol{\mu}}^* = \frac{1}{n}\sum_{i=1}^{n} \mathbf{h}(\mathbf{x_i^*})$*, kada* $n \to \infty$ *vrijedi:*
$\sqrt{n}\left(\hat{\theta}^* - \hat{\theta}\right) \xrightarrow{d^*} \mathcal{N}(\mathbf{0}, \mathbf{V}_\theta)$*, pri čemu je* $\mathbf{V}_\theta = \mathbf{G}'\mathbf{V}\mathbf{G}$,
a $\mathbf{V} = \mathbb{E}\left((\mathbf{h}(\mathbf{x}) - \boldsymbol{\mu})(\mathbf{h}(\mathbf{x}) - \boldsymbol{\mu})'\right)$ *i* $\mathbf{G} = \mathbf{G}(\boldsymbol{\mu})$.

(vidi [8]). Ovo je vrlo važan teorem, jer govori kako je distribucija bootstrap procjenitelja $\hat{\theta}^*$ asimptotski normalna. Dakle, moguće je izvesti približno točne zaključke o distribuciji $\hat{\theta}$ iz bootstrap distribucije. Već je ranije rečeno kako se bootstrap primjenjuje i za procjenu varijance. S obzirom da je procjenitelj za $\mathbf{V}$ zapravo $\hat{\mathbf{V}}_\theta = \hat{\mathbf{G}}'\hat{\mathbf{V}}\hat{\mathbf{G}}$, pri čemu je $\hat{\mathbf{G}} = \mathbf{G}(\hat{\boldsymbol{\mu}})$ i

$$\hat{\mathbf{V}} = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{h}(\mathbf{x_i}) - \hat{\boldsymbol{\mu}})(\mathbf{h}(\mathbf{x_i}) - \hat{\boldsymbol{\mu}})'.$$

Sada možemo zapisati kako to izgleda u bootstrap verziji.

$$\hat{\mathbf{V}}_\theta^* = \hat{\mathbf{G}^*}'\hat{\mathbf{V}}^*\hat{\mathbf{G}}^*$$
$$\hat{\mathbf{G}}^* = \mathbf{G}^*(\hat{\boldsymbol{\mu}^*})$$
$$\hat{\mathbf{V}}^* = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{h}(\mathbf{x_i^*}) - \hat{\boldsymbol{\mu}^*}\right)\left(\mathbf{h}(\mathbf{x_i^*}) - \hat{\boldsymbol{\mu}^*}\right)'.$$

Primjenom bootstrap verzije SZVB i centralnog graničnog teorema slijedi kako je $\hat{\mathbf{V}}^*$ konzistentan procjenitelj varijance $\mathbf{V}_\theta$ o čemu govori i sljedeći teorem.

Teorem 6 *Ako vrijede pretpostavke iz prethodnog teorema 5, tada* $\hat{\mathbf{V}}_\theta^* \xrightarrow{p^*} \mathbf{V}_\theta$*, kada* $n \to \infty$.

(vidi [8]).

2.5. Konzistentnost Bootstrap procjene varijance

Istražimo pod kojim je uvjetima bootstrap procjenitelj varijance konzistentan za asimptotsku varijancu procjenitelja $\hat{\theta}$. Pretpostavka je da je x_i skalar, a $z_n^* = \sqrt{n}(\overline{x^*} - \overline{x})$. Neka za niz a_n vrijedi sljedeće:

$$z_n = a_n(\hat{\theta} - \theta) \xrightarrow{d} \xi \tag{10}$$

i

$$z_n^* = a_n(\hat{\theta}^* - \hat{\theta}) \xrightarrow{d^*} \xi \tag{11}$$

za neku ograničenu distribuciju ξ. Znači da za neke normalizacije, $\hat{\theta}$ i $\hat{\theta}^*$ imaju asimptotski jednaku distribuciju što je prilično općenito, jer su u model uključene glatke funkcije. Standardni je bootstrap procjenitelj varijance od z_n varijanca uzorka izvučenog iz bootstrap uzorka $\{z_n^*(b) : b = 1, \ldots, B\}$. Stoga je bootstrap procjenitelj varijance od z_n:

$$\hat{V}_{\hat{\theta}}^* = \frac{1}{B-1}\sum_{b=1}^{B}\left(z_n^*(b) - \bar{z}_n^*\right)\left(z_n^*(b) - \bar{z}_n^*\right)'$$

$$\bar{z}_n^*(b) = \frac{1}{B}\sum_{b=1}^{B} z_n^*(b). \tag{12}$$

Treba primijetiti kako je procjenitelj indeksiran brojem bootstrap replikacija B. Kako z_n^* konvergira po bootstrap distribuciji prema istoj asimptotskoj distribuciji kao i z_n prirodno je pretpostaviti kako će varijanca od z_n^* konvergirati prema varijanci od ξ. Za konvergenciju varijance potrebno je da niz z_n^* bude uniformno kvadratno integrabilan.

Teorem 7 *Uz pretpostavku da za niz a_n vrijede* (10) *i* (11)*, te da je $\| z_n^* \|^2$ uniformno integrabilna i $B \to \infty$*

$$\hat{V}_\theta^{*,B} \xrightarrow{p^*} \hat{V}_\theta^* = \mathrm{Var}(z_n^*),$$

i kada $n \to \infty$

$$\hat{V}_\theta^* \xrightarrow{p^*} \hat{V}_\theta = \mathrm{Var}(\xi).$$

Slijedi da ako x_i ima konačnu varijancu i vrijede svi uvjeti iz prethodnog teorema, tada je procjenitelj varijance bootstrap uzorka konzistentan. Dakle, osim konzistentnosti bootstrap procjenitelja, pokazali smo i da niz bootstrap procjenitelja ima asimptotski normalnu distribuciju (vidi [8]).

2.6. Bootstrap regresija

Prije samog objašnjavanja bootstrap regresije, prirodno je definirati i opisati neke pojmove vezane uz regresiju. Sama je regresijska analiza vrlo popularna metoda koja proučava ovisnost među varijablama. Ona omogućuje istraživanje zavisnosti između dvije ili više varijabli. U slučaju zavisnosti varijable y o nezavisnoj varijabli x[9] dobiva se linearna regresija, što je ujedno i najjednostavniji slučaj koji se pojavljuje. Bitna je pretpostavka da postoje koeficijenti α i β takvi da se zavisna varijabla y može zapisati na sljedeći način:

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

[9]Nezavisne se varijable u regresiji nazivaju i regresori.

pri čemu e predstavlja grešku modela. Koeficijenti α i β procjenjuju se metodom najmanjih kvadrata[10].
U slučaju zavisnosti jedne varijable Y o više nezavisnih varijabli $x_1, \ldots, x_n$ dobiva se multivarijatna regresija. Neka je $\{(y_i, \boldsymbol{x}_i) : i = 1, \ldots, n\}$, gdje je n veličina uzorka, slučajan uzorak. Greške su slučajne varijable koje su normalno distribuirane s varijancom σ^2 i očekivanjem 0, a bit će označene s e_i. Linearni se projekcijski model i koeficijent linearne projekcije (β) mogu zapisati kao :

$$\begin{aligned} y_i &= \beta \mathbf{x}_i^{'} + e_i, \\ \mathbb{E}(\mathbf{x}_i e_i) &= 0 \\ \beta &= \underset{b \in \mathbb{R}^n}{\arg\min}\, S(b) \\ S(\beta) &= E((y_i - \mathbf{x}_i^{'}\beta)^2) \\ \beta &= [E(\mathbf{x}_i \mathbf{x}_i^{'})]^{-1} E(\mathbf{x}_i y_i), \end{aligned} \tag{13}$$

pri čemu je S minimizator očekivane kvadratne greške modela. Sada se LS metodom može procijeniti minimizator S koji će biti označen sa $\hat{S}$, a kako je prosjek[11] nepristran i konzistentan procjenitelj za očekivanje, slijedi da je

$$\hat{S} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^{'}\beta)^2 = \frac{1}{n} SSE(\beta). \tag{14}$$

Sa SSE[12] označena je suma kvadrata grešaka modela. Procjenitelj koeficijenta linearne projekcije dobivenog LS metodom označen je s $\hat{\beta}$ pa iz toga slijedi:

$$\hat{\beta} = \underset{\beta}{\arg\min}\, S(\beta).$$

Daljnjim se izračunima i primjenom izraza navedenih u (13) i (14) dolazi do formule za $\hat{\beta}$:

$$\hat{\beta} = \Big(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{'} \Big)^{-1} \Big(\sum_{i=1}^{n} \mathbf{x}_i y_i \Big) \tag{15}$$

(detaljnije u [8]). Interesantno bi bilo usporediti što se događa kada je u fokusu promatranja bootstrap uzorak. Kao što je ranije navedeno, neparametarski bootstrap algoritam uzorkuje opservacije nasumično uz zamjenu skupa podataka, čime nastaje bootstrap uzorak. Sada bootstrap uzorak izgleda kao parovi od $\boldsymbol{x}_i$ i y_i pa pišemo: $\{(y_1^*, \boldsymbol{x}_1^*), \ldots, (y_n^*, \boldsymbol{x}_n^*)\}$ što se može zapisati i u matričnom zapisu kao $(\mathbf{y}^*, \boldsymbol{x}^*)$. Važno je primijetiti kako su svi parovi u uzorku od (y_i i $\boldsymbol{x}_i$) uzorkovani. S obzirom da su parovi elementi bootstrap uzorka, često se nazivaju i *bootstrap parovi*. Baš kao i ranije, računaju se procjenitelji regresijskih koeficijenata, samo što treba voditi računa da se sada polazi od bootstrap uzorka. Procjenitelj će regresijskog koeficijenta bootstrap uzorka imati oznaku $\hat{\beta}^*$ i jednak je:

$$\hat{\beta}^* = \Big(\sum_{i=1}^{n} \mathbf{x}_i^{*'} \mathbf{x}_i^* \Big)^{-1} \Big(\sum_{i=1}^{n} \mathbf{x}_i^{*'} y_i^* \Big). \tag{16}$$

[10] Least square method (LS metoda).
[11] Aritmetička sredina uzorka.
[12] Sum squared error.

Postupak se ponavlja B puta. Za podatke je iz bootstrap uzorka $(y_i^*, \boldsymbol{x}_i^*)$ distribucija diskretna s vjerojatnošću $1/n$ za svaki par iz originalnog uzorka $(y_i, \boldsymbol{x}_i)$. Stoga je stvarna vrijednost procijenjenog parametra u ovoj bootstrap populaciji jednaka:

$$\big(\mathbb{E}^*(\mathbf{x}_i^*\mathbf{x}_i^{*'})\big)^{-1}\big(\mathbb{E}^*(\mathbf{x}_i^* y_i^*)\big) = \big(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\big)^{-1}\big(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i y_i\big) = \hat{\beta}, \tag{17}$$

pri čemu je $\hat{\beta}$ procjenitelj dobiven metodom najmanjih kvadrata. Podaci iz bootstrap uzorka zadovoljavaju jednadžbu linearnog projekcijskog modela:

$$y_i^* = \mathbf{x}_i^{*'}\hat{\beta} + e_i^* \tag{18}$$

pri čemu i ovdje vrijedi da je $\mathbb{E}^*(\mathbf{x}_i^* e_i^*) = 0$. Slijedi da je svaki bootstrap par $(y_i^*, \mathbf{x}_i^*)$ korespodentan stvarnoj opservaciji $(y_j, \mathbf{x}_j)$ pa je i bootstrap greška e_i^* jednaka rezidualima[13] iz originalnog skupa podataka. Problem je koji se u simuliranom bootstrap uzorku može pojaviti slučaj kada procjenitelj $\hat{\beta}^*$ ne može biti definiran, a to je mogućnost da $\mathbf{x}^{*'}\mathbf{x}^*$ bude singularna[14]. Standardni pristup za izbjegavanje ovog problema podrazumijeva izračun $\hat{\beta}^*$ samo ako $\mathbf{x}^{*'}\mathbf{x}^*$ nije singularna matrica. Kao bolje se rješenje definira tolerancija τ koja oslobađa $\mathbf{x}^{*'}\mathbf{x}^*$ nesingularnosti. Za vrijednost se tolerancije preporuča uzeti $\tau = \frac{1}{2}$. S λ^* označimo omjer najmanje svojstvene vrijednosti bootstrap matrice dizajna i originalne matrice dizajna:

$$\lambda^* = \frac{\lambda_{min}(\mathbf{x}^{*'}\mathbf{x}^*)}{\lambda_{min}(\mathbf{x}'\mathbf{x})}. \tag{19}$$

Ako u danom uzorku bootstrap replikacije vrijedi da je $\lambda^* < \tau$, onda se procjenitelj smatra nepostojećim ili se može definirati sljedeće pravilo:

$$\hat{\beta}^* = \begin{cases} \hat{\beta}^* \text{, ako je } \lambda^* \geq \tau \\ \hat{\beta} \text{ , ako je } \lambda^* < \tau \end{cases} \tag{20}$$

(vidi [8]).

[13] dobivenih metodom najmanjih kvadrata (LS)

[14] http://www.mathos.unios.hr/la1/p8_2014.pdf, Definicija 3.1.8.

2.7. Osnovni pojmovi u izradi i validaciji modela

Kada poduzeće od banke zatraži kredit, banka želi na što uspješniji način procijeniti koliko je potencijalni klijent pogodan, odnosno želi procijeniti rizik od nemogućnosti podmirenja obveza tog klijenta. Jedan je od mogućih načina kako to procijeniti primjenom scoring modela. Osnovni je cilj minimizirati taj rizik na način da se dođe do modela koji dobro raspoznaje „dobre“ i „loše“ klijente. Krucijalno je da model što bolje prediktira klijente, kako bi banke smanjile štetu koju snose kada lošim klijentima odobre kredite. Lošim se klijentima smatraju oni koji ne mogu podmiriti svoje obaveze u roku od 90 dana, a dobrim se klijentima smatraju oni koji to mogu. Status neispunjenja obveza (engl. default) klijenta nastaje ukoliko je ispunjen jedan od uvjeta:

- *„kreditna institucija smatra vjerojatnim da druga ugovorna strana neće u cijelosti podmiriti svoje obveze nastale na osnovi ugovora na temelju kojih su kreditna institucija ili bilo koja članica grupe kreditnih institucija kojoj ta kreditna institucija pripada izložene kreditnom riziku ne uzimajući u obzir mogućnost naplate iz instrumenata osiguranja (ako je obveza osigurana instrumentima osiguranja);”*

- *„druga ugovorna strana više od 90 dana nije ispunila svoju dospjelu obvezu po bilo kojoj materijalno značajnoj kreditnoj obvezi prema kreditnoj instituciji ili bilo kojoj članici grupe kreditnih institucija kojoj ta kreditna institucija pripada.”* (vidi [9]).

Upravo kada imamo takvu situaciju da je zavisna varijabla u modelu binarna, primjenjuje se metoda logističke regresije. Takva će varijabla u nastavku biti označena s Y.

Generalizirani linearni model i logistička regresija

Za modeliranje se binarne ovisne varijable koristi generalizirani linearni model s Bernoullijevom distribucijom ovisne varijable i prikladno odabranom link funkcijom. U ovom se radu koristi logistička regresija kod koje se kao link funkcija koristi logistička funkcija. Model je opisan u nastavku.
Neka je Y binarna slučajna varijabla iz Bernoulijeve distribucije s parametrom p takva da vrijedi:
$\mathtt{p}(Y = 1) = \mathrm{p}$, $\mathtt{p}(Y = 0) = 1 - \mathrm{p}$, pri čemu je vjerojatnost uspjeha jednaka p, a vjerojatnost neuspjeha $1 - p$. Važno je napomenuti da su „dobri“ klijenti označeni s 0, a „loši“ s 1.
Poznato je da Bernoulijeva dristribucija pripada eksponencijalnoj familiji distribucija (vidi [2], str. 46). Neka su $Y_1, \ldots, Y_n$ nezavisne slučajne varijable, pri čemu svaka od njih ima Bernoulijevu distribuciju. Distribucija od Y_i, ima kanonsku formu i ovisi o jednom parametru p_i, $i \in 1, \ldots, n$. Tada funkciju gustoće od Y_i možemo zapisati kao:

$$f(y_i; p_i) = exp\left[y_i \ln \frac{p_i}{1 - p_i} + \ln(1 - p_i)\right], i \in \{1, \ldots, n\}. \tag{21}$$

Za linearni model za koji je $E[Y_i] = \boldsymbol{x_i}^{\boldsymbol{T}}\boldsymbol{\beta}$ vrijedi da je $\mu_i = E[Y_i]$ očekivanje od Y_i uvjetno na $\boldsymbol{x_i}$,$i \in \{1, \ldots, k\}$. Uočimo da je ovdje $\mu_i = p_i$.
$\mathbf{X}$ će biti matrica koja sadrži vrijednosti nezavisnih varijabli (prediktora) $X_1, \ldots, X_k$ (*matrica dizajna*). U tom će slučaju vektor $\boldsymbol{x}_i$ sadržavati vrijednosti prediktora i-tog mjerenja, $i \in \{1, \ldots, k\}$. Transponiranjem ovog k-dimenzionalnog vektora dobiva se i-ti redak matrice dizajna:

$$\mathbf{X} = \begin{pmatrix} \boldsymbol{x_1^T} \\ \boldsymbol{x_2^T} \\ \vdots \\ \boldsymbol{x_n^T} \end{pmatrix}. \tag{22}$$

Treba imati na umu da je β k-dimenzionalan vektor baš kao i $\mathbf{X_i}$, stoga ga možemo zapisati kao:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}.$$

Kako bi se dobila veza između parametra β i očekivanja, potrebna je funkcija koja će imati ulogu funkcije veze („link funkcija“). Neka je g takva funkcija,
$g : \langle 0, 1\rangle \longrightarrow \mathbb{R}$, pri čemu je ona monotona i diferencijabilna.
Slijedi da je $g(\mu_i) = \mathbf{x_i}^T\boldsymbol{\beta} = \ln\left(\dfrac{p_i}{1-p_i}\right), i \in \{1, \ldots, n\}$. Navedeni se logaritam definira kao logaritam šansi (*eng.odds*) te se računa kao omjer vjerojatnosti realizacije uspjeha i neuspjeha.
Kako familiji generaliziranih linearnih modela pripada i model logističke regresije, da bi procijenili parametre β koristi se algoritam metode maksimalne vjerodostojnosti (eng. maximum-Likelihood estimate - MLE). Osnovu za metodu maksimalne vjerodostojnosti čini skor funkcija koja će biti označena s U. Slijedi postupak kako doći do nje.
Logaritmiranjem funkcije gustoće pod (21) dobiva se funkcija vjerodostojnosti za parametar β:

$$l(y_1, \ldots, y_n; p_1, \ldots, p_n) = \sum_{i=1}^{n}\left[y_i \ln\frac{p_i}{1-p_i} + \ln(1-p_i)\right], \tag{23}$$

pri čemu je $p_i = \dfrac{1}{1+e^{-\mathbf{x}_i^\tau\boldsymbol{\beta}}}$. Slijedi:

$$l(y_1, \ldots, y_n; \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n}\left[y_i\mathbf{x}_i^\tau\boldsymbol{\beta} - \ln\left(1+e^{\mathbf{x}_i^\tau\boldsymbol{\beta}}\right)\right]. \tag{24}$$

Potrebno je još odrediti parcijalnu derivaciju izraza pod (24) i rezultat izjednačiti s 0 kako bi se dobili procjenitelji regresijskih koeficijenata metodom maksimalne vjerodostojnosti.
$U(\beta) = \dfrac{\partial l}{\partial\beta} = 0$. Za $k = 1$ se dobiva:

$$U(\beta) - \frac{\partial l}{\partial\beta} - \sum_{i=1}^{n}\left[y_i x_i - \frac{x_i}{1+e^{-x_i\beta}}\right] = 0. \tag{25}$$

(vidi [2], str. 55-58).
Kako bi se dobilo numeričko rješenje ove nelinearne jednadžbe, potrebno je koristiti Newton-Raphson metodu za m iteracija. Slijedi da je:

$$\beta^m = \beta^{(m-1)} - \frac{U(\beta^{(m-1)})}{U'(\beta^{(m-1)})} \tag{26}$$

pri čemu je $U'(\beta)$ derivacija skor funkcije. Deriviranjem (25) se dobiva:

$$U'(\beta) = \sum_{i=1}^{n} \frac{x_i^2}{(1+e^{x_i\beta})^2} e^{x_i\beta}. \tag{27}$$

Može se primijetiti da gornji izraz ovisi samo o x, pa primjenom SZVB slijedi da je $\mathbb{E}_\beta(U(\beta)) = 0$, a $\mathrm{Var}_\beta(U(\beta)) = \mathfrak{I}$, pri čemu je $\mathfrak{I}$ oznaka za informaciju Fishera. Primjenom definicije varijance slijedi da je $\mathrm{Var}_\beta(U(\beta)) = \mathbb{E}_\beta(U^2) - \mathbb{E}_\beta(U')$, iz čega se jasno vidi da je $\mathfrak{I} = -\mathbb{E}_\beta(U')$. Uvrštavajući dobivene rezultate u (26) slijedi da je:

$$\beta^m = \beta^{(m-1)} + \frac{U(\beta^{(m-1)})}{\mathfrak{I}(\beta^{(m-1)})}.$$

Ovaj se postupak još naziva i skor metoda (vidi [2], str. 67-72). Procjena je regresijskih koeficijenata važna za odabir prediktora koji ulaze u jednadžbu modela, a to će biti oni koji se pokažu statistički značajnima. Interpretacija će se modela uz pomoć parametara $\beta_i, k \in 1, \ldots, n$ i prediktora $X_1, \ldots, X_n$ razlikovati za kategorijalne i numeričke varijable.

Kod numeričkih varijabli:
Ako je $\beta_i > 0$, tada se s porastom X_i povećava logaritam šansi da klijent ode u default.
Ako je $\beta_i < 0$, tada se s porastom X_i smanjuje logaritam šansi da klijent ode u default.

Kod kategorijalnih varijabli:
Ako je $\beta_i > 0$, veći je logaritam šansi da klijent bude loš za promatranu kategoriju u odnosu na baznu kategoriju.
Ako je $\beta_i < 0$, manji je logaritam šansi da klijent bude loš za promatranu kategoriju u odnosu na baznu kategoriju (vidi [15]).

Mjere kvalitete modela

Kod izračunavanja je stope pogodaka i grešaka modela, potrebno definirati klasifikacijsku matricu. Ona daje sve informacije koje su potrebne za izračunavanje grešaka modela, stopu koja govori o točnosti predviđanja modela te stope koje govore o točnosti predikcije dobrih, odnosno loših klijenata. Radi se o matrici koja ima dva retka i dva stupca. Po retcima se upisuju stvarne klasifikacije, a po stupcima predviđene. Općeniti prikaz klasifikacijske matrice izgleda ovako:

	Predviđeni	
Stvarni	**0**	**1**
0	TP	NN
1	NP	TN

Pritom su kratice u tablici oznake za:
TP = Točno pozitivni = „dobri“ klijenti koji su ispravno klasificirani kao „dobri“
NP = Netočno pozitivni = „loši“ klijenti koji su pogrešno klasificirani kao „dobri“
NN = Netočno negativni = „dobri“ klijenti koji su pogrešno klasificirani kao „loši“
TN = Točno negativni = „loši“ klijenti koji su ispravno klasificirani kao „loši“.
Definirajmo i ostale pojmove vezane uz klasifikacijsku matricu, pri čemu će ukupan broj dobrih klijenata imati oznaku „P“, a loših „N“.
Greška je tipa 1 odobravanje kredita lošem klijentu. Vjerojatnost se te greške dobije tako da se broj lažno pozitivnih klijenata podijeli s ukupnim brojem loših klijenata, tj. $\frac{NP}{NN+TN}$.
Greška je tipa 2 neodobravanje kredita dobrom klijentu. Vjerojatnost se te greške dobije tako da se broj lažno negativnih klijenata podijeli s ukupnim brojem dobrih klijenata, tj. $\frac{NN}{TP+NP}$.
Stopa točnog predviđanja modela (engl. *total hit rate*) = omjer ukupnog broja klijenata koje je model ispravno klasificirao i ukupnog broja klijenata, tj. $\frac{TP}{P+N}$.
Stopa točnog predviđanja dobrih klijenata (engl. *good hit rate*) = omjer dobrih klijenata koje je model ispravno klasificirao i stvarno dobrih klijenata, tj. $\frac{TP}{P}$.
Stopa točnog predviđanja loših klijenata (engl. *bad hit rate*) = omjer loših klijenata koje je model ispravno klasificirao i stvarno loših klijenata, tj. $\frac{TN}{N}$ (vidi [4], str. 861-862).
Krivulja se koja prikazuje odnos stope točnog predviđanja dobrih klijenata i stope točnog predviđanja loših klijenata naziva ROC krivulja (engl. *reciever operating characteristic curve*). Može se definirati i kao parametarski zadana krivulja koja prikazuje uređeni par relativnih frekvencija dobrih i loših klijenata u ovisnosti o graničnoj vrijednosti. Tada se na x-osi nalaze relativne frekvencije „dobrih“, a na y-osi relativne frekvencije „loših“ klijenata. U idealnom se slučaju, kada model savršeno klasificira, krivulja kreće od točke $(0,0)$ do točke $(1,1)$. Ukoliko se relativne frekvencije "dobrih" i "loših" ne razlikuju, tada ROC krivulja leži na pravcu (stopa točnog predviđanja dobrih klijenata = stopa predviđanja loših klijenata), što implicira da model ne razlikuje "dobre" od "loših" klijenata, već radi nasumičnu klasifikaciju (vidi [4], str. 863-864.). Površina se ispod ROC krivulje označava s AUC (engl. *area under curve*). Poželjna je vrijednost što bliža 1 i tada ROC krivulja ima idealan oblik, a ukoliko se dobije vrijednost manja od 0.5, to ukazuje na obratnu klasifikaciju klijenata. Može se još računati i Gini koeficijent kao $2 * AUC - 1$ kojim se mjeri sposobnost modela

da rangira rizik. Njegova se vrijednost također kreće između 0 i 1. Što je veća vrijednost dobivena, to je model bolji. Poželjni modeli imaju Gini koeficijent oko 0.6 (vidi [4], str. 868). Na kraju se može izračunati i Kolmogorov-Smirnov statistika (kraće KS statistika) čija vrijednost označava maksimalnu apsolutnu razliku između relativnih frekvencija dobrih i loših klijenata po svim graničnim vrijednostima. Što je ta razlika veća, dobiva se bolja diskriminacija između dobrih i loših klijenata.
Postoje neke granične vrijednosti koje govore o kvaliteti modela s obzirom na vrijednost KS statistike:

- KS > 50 %, model jako dobro klasificira,
- 20 % $<$ KS < 50 %, opadanjem vrijednosti KS statistike model sve lošije klasificira,
- KS < 20 %, model ne klasificira uopće.

(vidi [7], str. 4-5).

3. Empirijski dio: Primjena bootstrap metode u izradi i validaciji scoring modela

3.1. Kreditni scoring

Koristeći povijesne podatke i različite statističke tehnike, scoring je sistem dizajniran da prepozna utjecaj pojedinih karakteristika (varijabli) koje značajno djeluju na podmirenje obveza. Na temelju odabrane granične vrijednosti, primjenom ove metode banke sortiraju poduzeća po grupama u ovisnosti o njihovim razinama rizika. U težnji za izgradnjom scoring modela, analizirani su povijesni podaci (prikupljeni na temelju prijašnjih odobrenih kredita) i odabrane varijable za predikciju otplate duga. Kvalitetan bi model trebao odobriti što više dobrih klijenata, a odbiti što više loših klijenata. Na taj bi se način trebala dobiti bolja procjena rizika poduzeća, nego što bi se dobila bez primjene scoring modela (vidi [12], str. 3–16). *„Kreditni je scoring sistem dodjeljivanja bodova klijentu čiji zbroj predstavlja numeričku vrijednost koja pokazuje koliko je vjerojatno da klijent kasni u otplati kredita. Dodjeljuje jednu kvantitativnu mjeru, nazvanu skor, potencijalnom klijentu predstavljajući buduće ponašanje u otplati dodijeljenog kredita.“* (vidi [15]).
U članku je (vidi [11]) opisan statistički pristup za procjenu kreditnog rizika za srednja i mala poduzeća u Europi. Analizirane su varijable koje predstavljaju financijske omjere. Podaci su prikupljeni iz financijskih i knjigovodstvenih izvještaja europskih malih i srednjih proizvodnih poduzeća (eng. SME's) iz baze ORBIS koja sadrži poslovne i financijske podatke više od 50 milijuna europskih poduzeća. Analizirani podaci sadrže informacije iz 6 europskih zemalja (Ujedinjeno Kraljevstvo, Njemačka, Francuska, Belgija, Italija i Španjolska) te se odnose na vremenski period od 2012. do 2014. godine. Ukupno se u bazi nalaze podaci iz 25875 poduzeća. Ona su podijeljena u dvije kategorije s obzirom na svoj status likvidnosti: rizična (Distressed) i sigurna (Active). Rizična su poduzeća sva ona koja su u tom periodu bankrotirala ili postala nelikvidna, dok su aktivna poduzeća ona čije je poslovanje bilo uspješno u periodu prikupljanja podataka. Tablica 4. prikazuje raspodjelu poduzeća s obzirom na zemlju u kojoj se nalazi, status likvidnosti u danom vremenskom periodu i postotak rizičnih poduzeća.

Godina	**2014**			**2013**			**2012**		
Država/ Kategorija (A ili D)/ %D	A	D	%D	A	D	%D	A	D	%D
Belgija	434	7	1.59	519	16	3	549	10	1.79
Francuska	1140	49	4.12	1075	86	7.4	1062	46	4.16
Njemačka	839	8	0.94	1000	16	1.57	1006	6	0.59
Italija	3091	376	10.85	3245	135	4	3398	130	3.68
Španjolska	1140	70	5.79	1288	87	6.33	1377	88	6.01
Ujedinjeno Kraljevstvo	1210	11	0.9	1239	10	0.8	1102	10	0.9
Ukupno	8375			8716			8784		

Tablica 4: Izvor: https://www.bvdinfo.com/en-us/our-products/company-information/internationalproducts/orbis

Može se primijetiti kako se Italija ističe kao zemlja s najviše poduzeća (sigurnih i rizičnih) u svim godinama, a naročito u 2012. godini (ukupno 10375). Usporedbom se postotka loših poduzeća dolazi do zaključka kako je 2012. godine najveći postotak imala Španjolska, 2013. godine Francuska, a 2014. je godine na vrhu bila Italija. Kako bi pronašao što bolji model, autor je bazu podijelio na dva zasebna uzorka. Za samu je izradu modela koristio jedan

uzorak („training sample“) i tu uključio opservacije iz 2012. i 2013. godine, a za testiranje modela drugi uzorak („testing sample“) u kojem se nalaze podaci iz 2014. godine. Uzorak je koji se koristi za validaciju i testiranje skup podataka na kojem se provjerava pouzdanost modela koji je izrađen na „training“ uzorku. Izvještaji su poduzeća koji su uključeni u validaciju razvrstani s obzirom na zemlju u kojoj se nalazi poduzeće i status likvidnosti u 2014. godini. Ukupno je u „test“ uzorku 8375 podataka od kojih se 93, 77 % odnosi na izvještaje sigurnih poduzeća, a 6, 23 % na izvještaje rizičnih poduzeća. U „training“ se uzorku nalazi 17500 poduzeća od kojih je 96, 3 % sigurnih i 3, 7 % rizičnih. Sljedeća tablica sadrži opis 12 financijskih omjera korištenih za izradu modela.

	Financijski pokazatelj	**Formula**
X_1	Koeficijent tekuće likvidnosti	kratkotrajna imovina/kratkoročne obveze
X_2	Koeficijent ubrzane likvidnosti	(kratkotrajna imovina-zalihe)/kratkoročne obveze
X_3	Koeficijent trenutne likvidnosti	Novac/Kratkoročne obveze
X_4	Neto stopa povrata imovine (ROA[15])	Dobit razdoblja/ imovina
X_5	Koeficijent obrtaja zaliha	Prihodi od prodaje/zalihe
X_6	Dani vezivanja zaliha	365/(prihod od prodaje/zalihe)
X_7	Trajanje naplate potraživanja	365/(prihod od prodaje/ potraživanja)
X_8	Koeficijent likvidnosti (temeljen na imovini)	(Neto dobit + amortizacija)/ukupna imovina
X_9	Dobit prije kamata, poreza i amortizacije (EbitDaMarža)	EBITDA/(prihodi od prodaje)
X_{10}	Dobit prije kamata i poreza(EbitMarža)	EBIT/prihodi od prodaje
X_{11}	Profit po zaposleniku	Neto prihod/Prosječan broj zaposlenih
X_{12}	Koeficijent zaduženosti	Ukupne obveze/imovina

Tablica 5: Izvor: K. R. Subramanyam (2014): „Financial Statement Analysis“, 11th Edition, McGraw-Hill.

Omjeri su iz Tablice 5. iz „training“ uzorka i oni su zapravo prediktori u modelu koji primjenjuje logističku regresiju. Provedbom je Kruskall-Wallis testa na svim financijskim omjerima uz pouzdanost od 95% zaključeno kako su svi omjeri statistički značajni. Odabrano je 8 varijabli koje ulaze u jednadžbu modela. To su ROA (X_4), EBITDA marža (X_9), EBIT marža (X_{10}), Dani vezivanja zaliha (X_6), Koeficijent tekuće likvidnosti (X_1), Koeficijent likvidnosti (X_8), Koeficijent zaduženosti (X_{12}) i Profit po zaposleniku (X_{11}). Izradom je klasifikacijske matrice za oba uzorka, autor došao do sljedećih rezultata. Točnost klasifikacije na „training“ uzorku iznosi 82 %, dok na uzorku za validaciju modela ona iznosi 78, 9 %. U „training“ uzorku model postiže najbolju preciznost u razlikovanju sigurnih i rizičnih poduzeća koja iznosi 82, 8 %. Zatim su izračunate vrijednosti za AUC (Area Under the Curve) i KS-test[16], za oba uzorka. U „training“ uzorku AUC iznosi 75, 3 %,a u uzorku za validaciju 73, 2 %. Vrijednost KS-statistike i iznosi 15, 2 %. Ovi podaci govore kako se metodom logističke regresije došlo do modela koji ima dobru prediktivnu moć, što je i bio cilj.

U sljedećem je istraživanju (vidi [10]) proučavano ispitivanje utjecaja korištenja scoring modela u kooprerativnim bankama u Poljskoj. Pritom se analizira legitimnost korištenja scoring

[16]Stranica 18.

modela u bankarskim aktivnostima zajedno s ocjenom njegove učinkovitosti u reduciranju visokih vrijednosti NPL[17] omjera u poljskim kooperativnim bankama. Glavni je predmet ovog istraživanja ispitivanje utjecaja korištenja scoring modela na količinu kreditnih portfelja (mjerenih NPL omjerom) u poljskim koooperativnim bankama između 2004. i 2020. godine. Testirane su dvije hipoteze:
1.) Korištenje scoring modela u bankama u procjeni kreditnog rizika rezultira statistički značajnim smanjenjem kreditnih zahtjeva klijenata.
2.) Učinkovitost je internih scoring modela izrađenih u kooperativnim bankama u Poljskoj niža, nego učinkovitost modela koje predlažee BIK[18] (poljski ured u kojem se prikupljaju, pohranjuju i po potrebi dijele informacije o kreditnoj povijesti klijenata banaka).
U promatranom su razdoblju opažene promjene u dinamici podizanja kredita u Europi, tj. dolazi do smanjenja zahtjeva za kreditima, a kooperativne su banke u Poljskoj u tom smislu vrlo zanimljive za analiziranje. Obuhvaćene su grupe od 530 poduzeća (što je 10 puta veći broj nego u komercijalnim bankama). Poduzeća su jako neujednačena što se tiče veličine, područja operacija i financijskih uvjeta. Prosječna je razina neprofitnih kredita (krediti kod kojih banke ne ostvaruju povrat) u kooperativnim bankama u Poljskoj u prosjeku porasla 220 % između 2009. i 2020. godine. Ovaj se rastući trend stabilizirao tek u razdoblju između 2018. i 2020. godine. Razlika se u tendenciji može primijetiti u komercijalnim bankama u Poljskoj, ali i u kooperativnim bankama drugih zemalja Europske Unije (Njemačka, Francuska). Jedan je od razloga tog fenomena činjenica da je kriza 2007.-2009. godine zaustavila aktivnost uzimanja zajmova u poljskim kooperativnim bankama, koje su u tom periodu ušle u poslovanje s malim i srednjim poduzećima što ranije nisu prakticirali. Istraživanje je potvrdilo značajan utjecaj korištenja kreditnog scoring modela na vrijednosti NPL omjera u kooperativnim bankama. Banke su, koje su koristile kreditni scoring model predložen od strane poljskog ureda za informacije o kreditima (BIK), imale značajno niže vrijednosti NPL omjera kroz cijeli promatrani vremenski period, nego banke koje su koristile druge scoring modele. Također je potvrđen i značajan negativan utjecaj scoring modela na razinu NPL omjera predloženih od strane BIK-a. Potvrđena je učinkovitost korištenja scoring modela kao alata za ograničavanje kreditnog rizika u kooperativnim bankama u Poljskoj. Pokazano je da manje, lokalne banke imaju više prednosti od korištenja modela koje preporučuju kreditne institucije, od onih koje samostalno stvaraju modele na malim uzorcima (čija učinkovitost nije dokazana). Ukoliko bi se ovo istraživanje nastavilo u budućnosti, za preciznije bi rezultate svakako bilo dobro povećati promatrani uzorak.
Prethodni članci opisuju upotrebu kreditnog scoringa i njegove prednosti, ali i nedostatke. Njihova je upotreba dovela do skraćivanja vremena odobravanja kredita, jednostavnijeg pokretanja zahtjeva za kreditiranjem te odobravanje kredita većem broju zajmotražitelja (vidi [15]). Prilikom se upotrebe scoring modela za mala i srednja poduzeća nailazi na određene probleme. Jedan je od problema koji se javlja odabir varijabli koje ulaze u model. Statistički se značajne varijable u modelu, čiji podaci dolaze iz uzorka nekog većeg poduzeća, ne podudaraju s onima iz podataka manjih poduzeća (vidi [5], str. 18-24). Drugi je problem modeliranje na malim bazama podataka. Osim male količine podataka, problem su i dostupnost te kvaliteta danih podataka. Još je jedan problem koji se javlja starost samog poduzeća što također vodi malom broju podataka ukoliko se radi o mlađem poduzeću (vidi [15]). U tim je situacijama poželjno da se među podacima koji su poslani banci nalaze financijski pokazatelji ili informacije o djelatnosti poduzeća te se oni odabiru za varijable koje ulaze u model. U nastavku će biti opisana primjena kreditnog scoringa na jednu manju bazu

[17]Udio nepodmirenih zajmova u ukupnom portfelju zajmova banke.
[18]Polish Credit Information Bureau.

podataka.

3.2. Analiza varijabli za modeliranje i opis varijabli

Baza podataka koja se analizira u radu potječe iz jedne banke u Hrvatskoj i sadrži 119 podataka iz financijskih izvještaja poduzeća kojima je odobren kredit u toj banci. Među 32 varijable koje predstavljaju financijske omjere i stanje „dobar"/„loš" nalazi se i jedna kategorijalna (NKD[19]) varijabla. Financijski se omjeri prema svom značenju mogu rasporediti u pet grupa: aktivnost, ekonomičnost, profitabilnost, likvidnost i zaduženost. Kao što je već ranije objašnjeno (vidi str. 14) zavisna varijabla poprima vrijednosti 0 ukoliko se radi o klijentu okarakteriziranom kao „dobar" ili 1 ako je klijent „loš" te predstavlja kreditnu sposobnost klijenta. Uzorak sadrži 58 % dobrih i 42 % loših klijenata. Selekcijskom je procedurom „forward" dobiveno 6 varijabli koje ulaze u model i njihov se opis nalazi u Tablici 6.

Financijski pokazatelji	**Formule**
Ekonomičnost ukupnog poslovanja	Ukupni prihodi/Ukupni rashodi
Neto stopa povrata kapitala (NROE[20])	Dobit razdoblja/Kapital *100
Omjer zadržane dobiti i imovine	Zadržana dobit ili gubitak/Imovina
Koeficijent ubrzane likvidnosti	(Kratkotrajna imovina-zalihe)/Kratkoročne obveze
Koeficijent zaduženosti	Ukupne obveze/Imovina
Koeficijent obrta ukupne imovine	Ukupni prihodi/Imovina

Tablica 6: Financijski pokazatelji

Ekonomičnost ukupnog poslovanja mjeri odnos prihoda i rashoda i pokazuje koliko prihoda ostvaruje poduzeće po jedinici rashoda. Poželjne su vrijednosti veće od 1. Pokazatelj je koji govori koliko novčanih jedinica ostvaruje poduzeće na jednu jedinicu vlastitog kapitala stopa povrata kapitala (vidi [16]). U literaturi se često označava kraticom (ROE[21]). Vrijednosti omjera zadržane dobiti i imovine upućuju na način financiranja nekog poduzeća. Niže vrijednosti impliciraju financiranje zaduživanjem, dok više vrijednosti ukazuju na bolju profitabilnost poduzeća i otpornost poslovanja u nepovoljnim okolnostima. Koeficijent ubrzane likvidnosti pokazuje mogućnost podmirenja obveza poduzeća bez prodaje zaliha. Poželjna je vrijednost 1, a minimalno bi trebala biti 0.9. Sljedeći koeficijent u tablici govori koliki je postotak imovine poduzeća stečen zaduživanjem. Poželjna je vrijednost koeficijenta zaduženosti 50 % (0.5) ili manja. Veća vrijednost koeficijenta ukazuje na veću zaduženost što implicira veći rizik nemogućnosti podmirivanja obveza poduzeća (vidi [14]). Koeficijent obrta ukupne imovine ukazuje na to koliko se puta tijekom jedne godine obrne imovina poduzeća (vidi [16]).
Kako bi se ispitala multikolinearnost među varijablama, izračunati su faktori inflacije varijance. Sljedeća tablica prikazuje vrijednosti ovog faktora.

[19]Nacionalna klasifikacija djelatnosti= označava djelatnost prema nacionalnoj klasifikaciji djelatnosti (vidi [13]).
[21]Return on equity.

Varijable	Faktor inflacije varijance
Ekonomičnost ukupnog poslovanja (EUP)	1.133237
Neto stopa povrata kapitala(NROE)	1.114835
Omjer zadržane dobiti i imovine (OZDI)	3.340761
Koeficijent ubrzane likvidnosti (KUL)	1.423516
Koeficijent zaduženosti (KZ)	4.047671
Koeficijent obrta ukupne imovine (KOUI)	1.275404

Tablica 7: Provjera multikolinearnosti

Može se vidjeti kako su svi faktori inflacije manji od 5, što govori da problem multikolinearnosti među varijablama ovdje nije prisutan. U Tablici se 8. nalazi deskriptivna statistika varijabli s obzirom na „dobre“ i „loše“ klijente.

Varijable	Min	Donji kvartil	Med	Očekivanje	SD	Gornji kvartil	Max
EUP							
Dobri = ”0”	0.81	1.01	1.02	1.04	0.49	1.05	1.35
Loši = ”1”	0.00	1.00	1.01	0.99	0.27	1.05	1.73
NROE							
Dobri = ”0”	-107.06	3.12	11.12	17.79	30.53	28.74	154.39
Loši = ”1”	-105.13	0.00	1.42	1.02	20.01	6.47	42.99
OZDI							
Dobri = ”0”	-0.61	0.03	0.07	0.09	0.19	0.20	0.55
Loši = ”1”	-2.90	-0.06	0.05	-0.08	0.73	0.19	0.76
KUL							
Dobri = ”0”	0.04	0.53	0.80	0.99	0.8	1.20	4.89
Loši = ”1”	0.04	0.31	0.67	1.02	1.42	1.25	9.47
KZ							
Dobri = ”0”	0.07	0.53	0.70	0.66	0.23	0.84	1.29
Loši = ”1”	0.04	0.64	0.75	0.93	0.74	0.98	3.75
KOUI							
Dobri = ”0”	0.14	0.74	1.29	1.59	1.14	2.13	5.71
Loši = ”1”	0.05	0.31	0.64	0.96	2.33	1.55	4.25

Tablica 8: Deskriptivna statistika varijabli koje se nalaze u modelu

U Tablici se 8. mogu uočiti rezultati koji odstupaju od vrijednosti koje su definirane kod financijskih pokazatelja. Očekivana je vrijednost koeficijenta ubrzane likvidnosti poduzeća, koja su okarakterizirana kao dobra, niža, nego očekivana vrijednost istog koeficijenta loših poduzeća, što ukazuje da dobra poduzeća ne mogu podmiriti obveze bez prodaje zaliha. Koeficijent je zaduženosti iznad poželjne vrijednosti, što ukazuje da i dobra i loša poduzeća imaju problema s podmirivanjem obveza.

3.3. Izrada modela

Metoda je koja se primjenjuje za izradu modela logistička regresija. U sljedećim su tablicama prikazani procijenjeni regresijski koeficijenti.

Varijable	Parametri (β_i)	Standardna greška	z-vrijednost	p-vrijednost
Intercept[22]	0.9204	1.9258	0.478	0.6327
NROE	-0.0249	0.0126	-1.973	0.0484 *
KZ	4.4716	1.3059	3.424	0.0006 ***
KUL	0.7766	0.3725	2.085	0.0371 *
KOUI	-1.0098	0.3239	-3.117	0.0018 **
EUP	-3.9234	1.7021	-2.305	0.0212 *
OZDI	2.7699	1.2773	2.169	0.0301 *

Tablica 9: Procjene parametara i standardne greške modela

Varijable	Pouzdani interval
Intercept	$(-2.615, 5.263)$
NROE	$(-0.057, -0.005)$
KZ	$(2.123, 7.282)$
KUL	$(0.125, 1.548)$
KOUI	$(-1.697, -0.418)$
EUP	$(-8.126, -1.001)$
OZDI	$(0.339, 5.420)$

Tablica 10: Pouzdani intervali

Poudzanost za koeficijente u gornjoj tablici iznosi 95 %, što znači da će se u 95 % svih realizacija tih intervala prava vrijednost od β_i nalaziti unutar granica tih intervala. Svi će koeficijenti za koje je p-vrijednost manja od 0.05 (označeni s *) i kod kojih se 0 ne nalazi u pouzdanom intervalu, biti statistički značajni. Rezultati u tablicama pokazuju kako su sve varijable (osim slobodnog člana) statistički značajne, iz čega slijedi da će u jednadžbu modela biti uključene sljedeće varijable: neto stopa povrata kapitala, koeficijent zaduženosti, koeficijent ubrzane likvidnosti, koeficijent obrta ukupne imovine, ekonomičnost ukupnog poslovanja te omjer zadržane dobiti i imovine. Stoga će jednadžba odabranog modela izgledati ovako:

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6, \tag{28}$$

pri čemu su:
$\beta_0 = 0.9204$ slobodni član,
$\beta_1 = -0.0249$, X_1 = neto stopa povrata kapitala (NROE),
$\beta_2 = 4.4716$, X_2 = koeficijent zaduženosti (KZ),
$\beta_3 = 0.7766$, X_3 = koeficijent ubrzane likvidnosti (KUL),
$\beta_4 = -1.0098$, X_4 = koeficijent obrta ukupne imovine (KOUI),
$\beta_5 = -3.9234$, X_5 = ekonomičnost ukupnog poslovanja (EUP),
$\beta_6 = 2.7699$, X_6 = omjer zadržane dobiti i imovine (OZDI). Dakle, konačna jednadžba ima sljedeći oblik:

$$\ln(\frac{p}{1-p}) = 0.92 - 0.02 * NROE + 4.47 * KZ + 0.78 * KUL - 1.01 * KOUI - 3.92 * EUP + 2.77 * OZDI \tag{29}$$

Interpretacija:

- Ako se neto stopa povrata kapitala poveća za 1, logaritam se (kvocijenta) šansi da klijent ode u stanje nemogućnosti otplate kredita smanjuje za 0.0249. Dakle, poduzeća koja ostvaruju veći povrat na kapital, imaju nižu vjerojatnost neuspjeha.

- Ako se koeficijent zaduženosti poveća za 1, logaritam se (kvocijenta) šansi da klijent ode u stanje nemogućnosti otplate kredita povećava za 4.4716. Ukoliko poduzeće većim dijelom ostvaruje imovinu financiranjem iz tuđih izvora (zaduživanjem), smatrat će se rizičnijim.

- Ako se koeficijent ubrzane likvidnosti poveća za 1, logaritam se (kvocijenta) šansi da klijent ode u stanje nemogućnosti otplate kredita povećava za 0.7766. S obzirom da je kod „loših“ poduzeća medijan ovog koeficijenta 0.67, a poželjna je vrijednost 1, povećanje bi koeficijenta trebalo ukazivati na bolje poslovanje poduzeća, čime bi ono bilo i manje rizično. Dobiveni rezultat nije u skladu s očekivanjima, no može se objasniti činjenicom da postoje uspješna poduzeća koja jako puno ulažu u poslovanje te se zbog toga zadužuju.

- Ako se koeficijent obrta ukupne imovine poveća za 1, logaritam se (kvocijenta) šansi da klijent ode u stanje nemogućnosti otplate kredita smanjuje za 1.0098. Manje rizična poduzeća ostvaruju veći koeficijent obrta ukupne imovine. Ukoliko ostvare veće prihode, moguće je da se imovina obrne više puta tijekom godine. Ovo također može ovisiti o djelatnosti kojom se poduzeće bavi.

- Ako se koeficijent ekonomičnosti ukupnog poslovanja smanji za 1, logaritam se (kvocijenta) šansi da klijent ode u stanje nemogućnosti otplate kredita povećava za 3.9234. Poduzeća se koja u svom poslovanju ostvaruju veće rashode od prihoda smatraju rizičnima.

- Ukoliko se omjer zadržane dobiti i imovine poveća za 1, logaritam se (kvocijenta) šansi da klijent ode u stanje nemogućnosti otplate kredita povećava za 2.7699. Ovaj rezultat također nije u skladu s očekivanjima, no može se objasniti činjenicom da poduzeće svu svoju zadržanu dobit koristi za ulaganje u poslovanje.

3.4. Validacija modela

Kvaliteta je modela provjerena izračunavanjem informacijskih kriterija AIC[23] i BIC[24]. Njihove su vrijednosti 132.91 i 152.37 redom. Slijedi prikaz klasifikacijske matrice i kvantitativnih pokazatelja kvalitete modela u sljedećim tablicama.

	Predviđeni	
Stvarni	**0**	**1**
0	73.91 %	26.09 %
1	28 %	72 %

Tablica 11: Klasifikacijska matrica

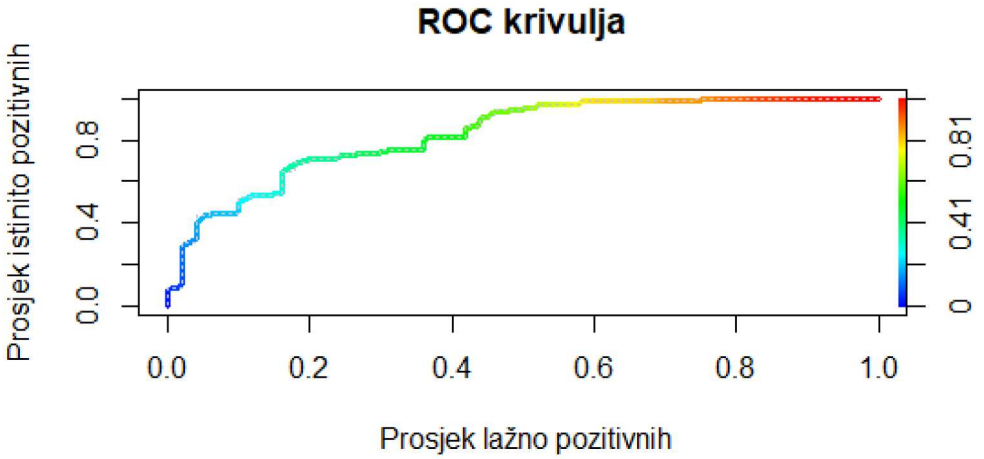
[23] Akaike Information Criterion.
[24] Bayes Information Criterion.

Od ukupnog broja „dobrih“ klijenata, njih je 26, 09 % neispravno prediktirano kao „loši“ što se još naziva i greškom tipa 2. 28 % „loših“ klijenata je neispravno prediktirano kao „dobri“ što se naziva greškom tipa 1[25]. Ispravno su klasificirani „dobri“ klijenti s točnošću od 73, 91 % (Good hit rate), dok točnost ispravno prediktiranih „loših“ klijenata iznosi 72 % (Bad hit rate).

Pokazatelji	**Izračunate vrijednosti**
Total hit rate[26]	73.11 %
AUC[27]	81.97 %
GINI koeficijent	63.94 %
KS-statistika	50.12 %

Tablica 12: Pokazatelji kvalitete modela

Stopa točnog predviđanja modela iznosi 73, 11 %. Površina ispod krivulje iznosi 81, 97 %, a GINI koeficijent 63, 94 % što ukazuje na to da model ima dobru moć diskriminacije „dobrih“ i „loših“ klijenata. Isto je potvrđeno Kolmogorov-Smirnov (KS) statistikom koja iznosi 50, 12 %, a prema vrijednostima navedenim u uvodnom dijelu, ovaj rezultat govori da model radi vrlo dobro. Slijedi prikaz ROC krivulje na Slici 1, te grafički prikaz relativnih frekvencija „dobrih“ i „loših“ klijenata na Slici 2.



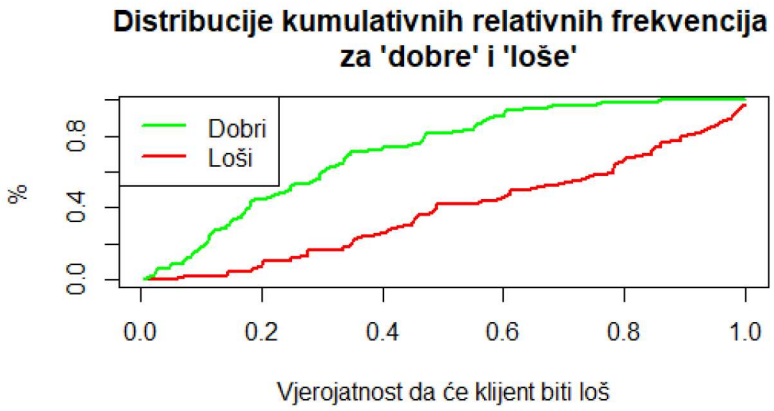
Slika 1: ROC krivulja

Na Slici 2. se na x-osi nalaze vrijednosti scorova, dok su na y-osi prikazane relativne frekvencije „dobrih“ i „loših“ klijenata, pri čemu zelena linija predstavlja frekvencije „dobrih“, a crvena linija frekvencije „loših“ komitenata. Poželjna je što veća udaljenost među grafovima ovih funkcija, jer to ukazuje na bolju diskriminaciju između „dobrih“ i „loših“, što je ovdje slučaj. U sljedećem će poglavlju na ovaj model biti primijenjena bootstrap metoda te će na kraju ti rezultati biti uspoređeni s ovim dosadašnjim.

3.5. Primjena bootstrap metode

Za procjenu je regresijskih koeficijenata iz opisanog modela uzeto 1000 replikacija. Svaki uzorak sadrži 119 podataka te je odabrano 100 uzoraka iz kojih je dobivena bootstrap distribucija procijenjenih parametara danog modela. Generirani se percentilni pouzdani intervali s

[25]Definicije za greške tipa 1 i 2 nalaze se u uvodu.



Slika 2: Graf s relativnim frekvencijama dobrih i loših klijenata

pouzdanošću od 95% mogu vidjeti u Tablici 13. Očekivanje i varijanca bootstrap distribucije procijenjenih parametara se mogu vidjeti u Tablici 14.

Varijable	**Pouzdani interval**
Intercept	$(-3.259, 8.393)$
NROE	$(-0.144, -0.008)$
KZ	$(2.482, 8.527)$
KUL	$(0.269, 1.821)$
KOUI	$(-2.112, -0.131)$
EUP	$(-11.770, -0.917)$
OZDI	$(0.194, 6.303)$

Tablica 13: Bootstrap pouzdani intervali

Treba zamijetiti da su ovi pouzdani intervali širi nego oni dobiveni prijašnjom metodom u Tablici 10. Bootstrapom je procijenjeno kako su svi koeficijenti osim slobodnog člana statistički značajni. Taj se rezultat podudara s onim koji daje logistička regresija, stoga je dobivena potvrda da iste varijable budu uključene u model. Bootstrap metoda omogućuje izračun pouzdanih intervala za mjere kvalitete modela koje su prikazane Tablicom 14. što nam daje bolji uvid u kvalitetu modela.

Pokazatelji	**Očekivanje**	**Varijanca**	**Pouzdani intervali**
Total hit rate	0.7089	0.0008	$(0.6471, 0.7563)$
Good hit rate	0.7079	0.0057	$(0.5366, 0.8261)$
Bad hit rate	0.7132	0.0055	$(0.56, 0.84)$
AUC[28]	0.8001	0.0003	$(0.7597, 0.8206)$
GINI koeficijent	0.5986	0.0013	$(0.5090, 0.6406)$
KS-statistika	0.4833	0.0013	$(0.4102, 0.5536)$

Tablica 14: Karakteristike bootstrap distribucije procijenjenih pokazatelja

4. Zaključak

Procjena rizičnosti poduzeća nije jednostavna, a kod malih je i srednjih poduzeća to još izazovnije. Jedan je od uzroka što takva poduzeća ne raspolažu velikim bazama podataka što može dovesti do pogrešnih zaključaka. Banke imaju svoje mehanizme koje koriste kako bi procijenile kreditnu sposobnost klijenata, a jedan je od njih kreditni scoring. U ovom je radu kreditni scoring primijenjen na bazi podataka iz jedne banke u Hrvatskoj koja sadrži 119 podataka o financijskim izvještajima malih i srednjih poduzeća. Varijable koje se nalaze u bazi podataka predstavljaju financijske omjere koji se mogu razvrstati u 5 grupa: aktivnost, ekonomičnost, profitabilnost, likvidnost i zaduženost. Za procjenu je rizika korišten model logističke regresije koji sadrži 6 varijabli: ekonomičnost ukupnog poslovanja, neto stopa povrata kapitala, omjer zadržane dobiti i imovine, koeficijent ubrzane likvidnosti, koeficijent zaduženosti i koeficijent obrta ukupne imovine. S obzirom na veličinu baze podataka, za izračun je pouzdanih intervala i provjeru kvalitete dobivenog modela, korištena bootstrap metoda. Validacijom se modela došlo do sljedećih rezultata: ukupna stopa predviđanja modela iznosi $70,89$ %, površina ispod krivulje (AUC) 80 % i GINI koeficijent iznosi $59,86$ %. Izračunata je i KS statistika koja iznosi $48,33$ %. Ti rezultati govore kako dobiveni model ima dobru moć predikcije, efikasnost i kako precizno diskriminira „dobre“ i „loše“ klijente. Kako je u radu korištena mala baza podataka, korisno je provjeriti kvalitetu modela koji se dobio primjenom logističke regresije. U tu je svrhu korištena bootstrap metoda, čime je potvrđena kvaliteta modela.

5. Literatura

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

[2] A. J. Dobson, A. G. Barnett: An introduction to generalized linear models, Fourth Edition, Boca Raton, 2018.

[3] B. Efron, R. J. Tibshirani: An Introduction to the Bootstrap, 1993.

[4] T. Fawcet: Introduction to ROC analysis, 2006.

[5] R. Feldman: Small Business Loans, Small Banks and a Big Change in Technology Called Credit Scoring. 11(3), Region 1997.

[6] J. Fox: Bootstrapping regression models (An R and S-PLUS Companion to Applied Regression: A Web Appendix to the Book Sage, Thousand Oaks), 2002.

[7] B. Gadidov, B. McBurnett: Population Stability and Model Performance Metrics Replication for Business Model at SunTrust Bank, Kennesaw State University; Georgia Institute of Technology, 2015.

[8] B. E. Hansen: Econometrics, Princeton University Press, 2022.

[9] Hrvatska narodna banka: Narodne novine, Članak 2., Odluka o adekvatnosti jamstvenog kapitala kreditnih institucija NN 1/2009.

[10] K. Kil, R. Ciukaj, J. Chrzanowska: Scoring models and credit risk: The case of cooperative banks in Poland, 2021.

[11] G. Kyriazopoulos: Credit risk evaluation and rating for SME's using statistical approaches: the case of European SME's manufacturing sector, 2019.

[12] L. Mester: What's the point of credit scoring? Business Review 3, 1997.

[13] Nacionalna klasifikacija djelatnosti 2007. – NKD 2007.

[14] Ross, S. A., Westerfield, R.W., Jordan, B.D. : "Fundamentals of Corporate Finance", IRWIN, Chicago, 1995.

[15] N. Šarlija, Predavanja za kolegij Upravljanje kreditnim rizicima, Odjel za matematiku, Osijek 2018.

[16] K. Žager, L. Žager: Analiza financijskih izvještaja, Masmedia, Zagreb, 1999.

[17] http://www.math.ntu.edu.tw/ hchen/teaching/LargeSample/notes/notebootstrap.pdf

[18] https://www.europarl.europa.eu/erpl-app-public/factsheets/pdf/hr/FTU_2.4.2.pdf

[19] https://www.fina.hr/-/rezultati-poslovanja-poduzetnika-u-2022.-godini-razvrstani-po-velicini

Sažetak

U ovom je radu opisana primjena bootstrap metode na kreditni scoring. Jedan je od načina kako banke procjenjuju kreditnu sposobnost svojih klijenata upotreba kreditnog scoringa. Na bazi je podataka iz jedne banke u Hrvatskoj, koja sadrži 119 podataka, u svrhu procjene rizika korišten model logističke regresije. Kako bi se provjerila kvaliteta dobivenog modela korištena je bootstrap metoda, čija je upotreba česta kada se radi o bazama s malo podataka kao u ovom slučaju. U praksi se najčešće koriste neparametarski i parametarski bootstrap koji su opisani u radu. Za potrebe ovog rada je korišten neparametarski bootstrap koji je primijenjen na parametarskom modelu. Konačni su rezultati pokazali da dobiven model dobro diskriminira „dobre“ i „loše“ klijente i da ima dobru moć predikcije.

Ključne riječi: bootstrap metoda, financijski pokazatelji, kreditni scoring, logistička regresija, validacija modela

Application of the bootstrap method in credit scoring

Summary

This paper discribes application of the bootstrap method of credit scoring. Credit scoring is one of the methods used by banks to determine the creditworthiness of their clients. For risk assessment, a logistic regression model was used, based on the data of the bank in Croatia, which contains 119 data. To check the quality of the obtained model, the bootstrap method was used, which is often used when dealing with databases with little data, as in this case. In practise, nonparametric and parametric bootstraps are often used and are described in this paper. For this paper, the nonparametric bootstrap was used and applied to parametric model. The final results have shown that the obtained model is able to discriminate the non-defaulters from the defaulters and has good predictive power.

Keywords: bootstrap method, financial indicators, credit scoring, logistic regression, model validation

Životopis

Rođena sam 17. siječnja 1993. godine u Osijeku. Završila sam osnovnu školu M.P. Katančića u Valpovu. Nakon završetka osnovne škole upisala sam opću gimnaziju u Srednjoj školi Valpovo. Kroz cijelu osnovnu i srednju školu trenirala sam odbojku i prisustvovala na natjecanjima iz matematike i LiDraNu. Osim toga, u srednjoj školi sam glumila u Amaterskom kazalištu Belišće. Zahvaljujući svojoj razrednici koja je ujedno bila i profesorica matematike, odlučila sam ići na matematički fakultet. 2011. godine upisujem preddiplomski studij matematike na Odjelu za matematiku u Osijeku. Završavam ga 2016. godine s temom završnog rada „Polja i prsteni“ pod mentorstvom izv. prof. dr.sc. Ivana Matića. Naposlijetku iste godine upisujem diplomski studij, također na Odjelu za matematiku, smjer Financijska matematika i statistika.