UNIVERZITET U NOVOM SADU
PRIRODNO-MATEMATIČKI
FAKULTET
DEPARTMAN ZA MATEMATIKU I
INFORMATIKU

Bejzova logistička regresija kao alat za procenu kreditnog rizika

- MASTER RAD -

Mentor:

Prof. dr Zagorka Lozanov-Crvenković

Student:

Lazar Beić

183m/17

Novi Sad, 2023.

Sadržaj

Predgovor

Cilj ovog rada je da istraži primenu Bejzove logističke regresije u kontekstu ocenjivanja kreditnog skoringa. Kreditni skoring igra ključnu ulogu u finansijskoj industriji, a najpre u bankarstvu. On omogućava kreditnim institucijama procenu kreditnog rizika potencijalnih klijenata.

U uvodnom delu rada biće pružen pregled osnovnih pojmova vezanih za kreditni rizik kao i pregled i istorijat modela kreditnog skoringa. Nakon toga biće predstavljeni osnovni principi logističke regresije kao najčešće metode modeliranja kreditnog skoringa.

Drugi deo rada uvodi principe Bejzove statistike i zaključivanja kao i teorijske osnove za njihovu primenu u logističkoj regresiji. Uvode se teorijske osnove Monte Karlo lanaca Markova, probablističkih metoda neophodnih za Bejzovu logisitčku regresiju.

Ono što sledi su predstavljanje i opis podataka klijenata koji su korišćeni za modeliranja. Nakon toga biće prikazani modeli dobijeni pomoću standardne logističke regresije kao i model dobijen Bejzovom logističkom regresijom.

Na kraju ćemo izložiti zaključke iz našeg istraživanja i uporediti performanse modela dobijenih na dva pomenuta načina. Takođe ćemo diskutovati o prednostima i ograničenjima primene Bejzove logističke regresije u ocenjivanju kreditnog skoringa. Potencijalne smernice za dalja istraživanja će naposletku biti navedene.

Izuzetnu zahvalnost dugujem svom mentor Prof. dr Zagorki Lozanov – Crvenković za pruženu priliku za mentorstvo, kao i za sve savete, pomoć tokom izrade rada, i za prenetu ljubav prema statistici i statističkom modeliranju tokom svih godina studija.

Iskoristio bih priliku da se zahvalim svojoj majci ocu i bratu za nesebičnu podršku tokom svog školovanja. Posebno bih ovaj rad posvetio svojoj porodici: supruzi Bojani i tek rođenom sinu Bogdanu koji svakodnevno predstavljaju neiscrpnu inspiraciju i daju podstrek svemu što radim.

1. Uvod

Cilj ovog rada je da istraži potencijalnu primenu Bejzove logističke regresije u ocenjivanju kreditnog skoringa klijenata banke.

U uvodnom delu rada pružen je pregled osnovnih pojmova vezanih za kreditni rizik, pregled i istorijat modela kreditnog skoringa, kao i diskusija o metodama modeliranja skoringa.

Nakon toga u drugom delu predstavljeni su osnovni principi standardne logističke regresije kao najčešće metode modeliranja kreditnog skoringa.

Treći deo rada uvodi principe Bejzove statistike i zaključivanja kao i teorijske osnove za njihovu primenu u logističkoj regresiji. Bejzova logistička regresija predstavlja kao prirodno proširenje standardne metodologije, koja inkorporira prethodne informacije radi poboljšanja tačnosti procene parametara. Ističe se ključna karakteristika u tretiranju koeficijenata regresije kao slučajnih promenljivih, uz uzimanje u obzir informacija iz uzorka (verovatnoća pojavljivanja podataka) i prethodnih saznanja. Objašnjena je metodologija korišćenja MCMC (Markov Chain Monte Carlo) metoda radi uzorkovanja iz posteriornih raspodela parametara modela procene posteriorne distribucije.

Ono što sledi u narednom delu je predstavljanje samih podataka koji su korišćeni za modeliranje, opis, vizuelizacija i statistički pokazatelji korišćenih promenljivih. Predstavljena je i podela inicijalnog skupa na skup na kome će se primenjivati standardna logistička regresija i na skup na kome će se primenjivati Bejzova logistička regresija, koja koristi priorna znanja o koeficijentima modela iz standardnog modela.

Četvrti deo rada predstavlja model dobijen standardnom logističkom regresijom, zajedno sa statističkim ocenama parametara i metrikama koje pokazuju kvalitet modela i prilagođenost modela podacima kao što su Gini i KS, tačnost, osetljivost, preciznost modela i drugi pokazatelji. Nakon toga je predstavljen i model dobijen Bejzovom logističkom regresijom i dato je poređenje modela sa modelom standardne logističke regresije na testnom skupu, nezavisnom od skupova na kome su prethodni modeli razvijani. Svo modeliranje i manipulacija podacima rađena je kroz porgramski jezik Python.

Na kraju su izloženi zaključci iz istraživanja i diskutovano je o prednostima i ograničenjima primene Bejzove logističke regresije u ocenjivanju kreditnog skoringa. Potencijalne smernice za dalja istraživanja su naposletku navedene.

2. Kreditni rizik

"Kredit je ugovorni sporazum po kojem klijent - zajmotražioc (dužnik) prima novac od zajmodavca (poverioca) uz dogovor da u utvrđenom roku taj novac vrati." [1]

Kada finansijska institucija odobrava kredit zajmotražiocu, ona obavlja svoju fundamentalnu funkciju: prihvatanje kreditnog rizika. Uspeh banke leži u njenoj sposobnosti da predvidi i kvantifikuje ukupan rizik. Bankovni menadžment se suočava sa šest temeljnih rizika: kreditni rizik, rizik likvidnosti, tržišni rizik, operativni rizik, regulatorni rizik i rizik ljudskog faktora.[10]

Kreditni rizik je rizik spremnosti i mogućnosti dužnika da izmiri obaveze prema banci delimično ili u potpunosti prema uslovima iz ugovora kojim se regulišu obligacioni odnosi u vezi određene finansijske transakcije. Dakle, ovaj rizik potencijalno ima dva izvorišta:

- subjektivnu spremnost i volju dužnika da odgovori svojim obavezama iz obligacionog odnosa sa bankom i
- objektivnu mogućnost da izmiruje obaveze prema banci iz ostvarenog prihoda ili na drugi, za banku prihvatljiv način.

Reč je o dve podjednako važne komponente kreditne sposobnosti dužnika, koje daju specifičnu dimenziju ovom riziku.[12]

Kreditni rizik banke predstavlja verovatnoću da banka neće biti u stanju da naplati svoja ukupna potraživanja po osnovu glavnice duga i po osnovu ugovorene kamate što za posledicu ima oslabljen kapital banke i negativan uticaj na finansijski rezultat banke.

Kreditni rizik je suštinski rizik koji prati svako bankarsko poslovanje, a upravljanje njime predstavlja osnovu za uspešno delovanje banke. Ideja o kreditnom riziku je postojala od samih početaka bankarstva, jer pozajmljivanje novca drugoj strani uvek nosi mogućnost da sredstva neće biti vraćena. Iz tog razloga, prepoznata je potreba za efikasnim upravljanjem ovim rizikom kako bi se izbegle negativne posledice po bankarsko poslovanje, te su se preduzimale različite aktivnosti u skladu sa tim. U prošlosti, mehanizmi za upravljanje kreditnim rizikom bili su primitivniji u odnosu na današnje, ali su već tada bankarske institucije prepoznavale značaj ovog rizika.

Danas, banke su obavezne da identifikuju, mere i procenjuju kreditni rizik na osnovu kreditne sposobnosti dužnika, kao i na osnovu kvaliteta sredstava obezbeđenja koja se koriste (kolaterala). Upravljanje kreditnim rizikom postalo je mnogo sofisticiranije, uz primenu naprednih analitičkih metoda i tehnologija. Cilj banaka je da pravilno procene rizik svakog

[1] ıhttp://www.nbs.rs/internet/cirilica/glossary.html

zajmoprimca kako bi smanjile šanse za potencijalne gubitke i osigurale stabilnost svog poslovanja.

Kreditna sposobnost dužnika se odnosi na njegovu sposobnost da uredno servisira obaveze prema banci i vrati ih u ugovorenom roku. Osim što služe kao pokriće poverioca (banke) u slučaju neizvršenja obaveza, sredstva obezbeđenja takođe predstavljaju podsticaj dužniku da kredit otplaćuje prema ugovoru. Postojanje sredstava obezbeđenja umanjuje kreditni rizik, ukoliko se sredstvo obezbeđenja može lako realizovati, odnosno ukoliko je lako utrživo. Sredstva obezbeđenja su široko prisutna u praksi i koriste se kao instrumenti zaštite od kreditnog rizika.

Najčešći oblici sredstava obezbeđenja uključuju sledeće:

- **Založno pravo na imovinu**:

Banka može tražiti zalog na imovini dužnika, kao što su nekretnine, vozila, oprema ili drugi vredni predmeti. U slučaju neplaćanja, banka ima pravo da proda založenu imovinu kako bi nadoknadila nenaplaćeni dug.

- **Hipoteka**:

Hipoteka je poseban oblik založnog prava koji se odnosi na nekretnine. Banka postaje hipotekarni poverilac i ima pravo na prodaju hipotekarno založene nekretnine ako dužnik ne izmiri obaveze prema banci.

- **Bankarske garancije i akreditivi**:

Banka može izdati garancije ili akreditive u korist dužnika, što pruža dodatno obezbeđenje drugoj strani u poslovnoj transakciji. Ovo osigurava isplatu u slučaju da dužnik ne ispuni svoje obaveze.

- **Zalog novca ili vrednosnih papira**:

Banka može zahtevati zalog u novcu ili vrednosnim papirima kao obezbeđenje za odobreni kredit. Ukoliko dužnik ne ispuni obaveze, banka može naplatiti sredstva sa računa ili prodati založene vrednosne papire kako bi povratila sredstva.

- **Lične garancije**:

Ponekad, banka može tražiti lične garancije od drugih lica kao dodatnu sigurnost za odobreni kredit. To znači da će treća osoba biti odgovorna za vraćanje duga ukoliko dužnik ne bude u mogućnosti.

Korišćenje sredstava obezbeđenja može smanjiti rizik za banku jer pruža dodatnu sigurnost u slučaju neplaćanja. Međutim, važno je da banke pažljivo procenjuju vrednost i likvidnost obezbeđenja kako bi bili sigurni da će biti u mogućnosti da povrate sredstva u slučaju potrebe.

U nekim situacijama, vrednost obezbeđenja može varirati i podložna je promenama na tržištu, što bankama zahteva pažljivo praćenje kako bi održale adekvatnu zaštitu od kreditnog rizika.

Metodologija utvrđivanja kreditnog rizika nalazi se u središtu kreditne analize, a osnovna ideja kreditne politike banke je minimizacija kreditnog rizika i maksimizacija profita. Prioritet banke je sposobnost pokrivanja nastalih gubitaka na osnovu plasiranih kredita. Cilj je adekvatno kontrolisati izloženost riziku i pratiti eventualna pogoršanja, preduzimajući preventivne mere kako bi se nepovoljne situacije sprečile.

Efikasan sistem upravljanja kreditnim rizikom ima ključni uticaj na eliminaciju većine problema prisutnih u bankama. Upravljanje kreditnim rizikom se sprovodi primenom različitih metoda, uključujući postavljanje limita, pažljivu selekciju kreditnih zahteva, diversifikaciju plasmana i korišćenje odgovarajućih sredstava obezbeđenja za plasmane.

Zaštita od kreditnog rizika postaje još značajnija u periodima ekonomske nestabilnosti ili finansijskih kriza. U takvim vremenima, banke posebno obraćaju pažnju na adekvatno upravljanje rizikom i primenu odgovarajućih mera kako bi sačuvale svoju stabilnost i sigurnost poslovanja.

Upravljanje kreditnim rizikom je složen i dinamičan proces, zahtevajući stalno praćenje i prilagođavanje u skladu sa promenama u ekonomskom okruženju i tržišnim uslovima. Njegova uspešna primena obezbeđuje bankama veću stabilnost i poverenje kako klijenata, tako i investitora.

Kreditni rizik se može posmatrati na nivou pojedinačnih plasmana, na nivou klijenta/dužnika i na nivou celokupnog portfolija. Na nivou pojedinačnih plasmana, banka može preduzeti sledeće korake kako bi umanjila kreditni rizik:

- **Definisanje vremenskog perioda kreditiranja**:

Postavljanjem jasnog vremenskog okvira za vraćanje kredita, banka smanjuje mogućnost zastoja u plaćanju i osigurava povraćaj sredstava u predviđenom roku.

- **Procena kreditne sposobnosti dužnika**:

Banka treba pažljivo proceniti kreditnu sposobnost potencijalnih dužnika, uzimajući u obzir njihove finansijske podatke, istoriju kreditnih transakcija i druge relevantne faktore. Ovo omogućava da se odobre krediti samo onima koji su sposobni da ih redovno servisiraju.

- **Postavljanje kreditnog limita dužnika**:

Definisanjem maksimalnog iznosa koji dužnik može pozajmiti, banka smanjuje potencijalne gubitke u slučaju neplaćanja ili nesolventnosti dužnika.

- **Kontrola korišćenja kredita**:

Banka može postaviti određene uslove i ograničenja u vezi sa namenom kredita kako bi osigurala da se sredstva koriste u skladu sa dogovorenim uslovima.

- **Obezbeđenje povraćaja kredita**:

Traženje adekvatnog sredstva obezbeđenja, kao što su zalozi, hipoteke ili bankarske garancije, povećava šanse za povraćaj kredita čak i u slučaju neplaćanja dužnika.

Na nivou portfolija banke, umanjenje rizika se postiže primenom različitih strategija:

- **Limitiranje veličine kredita prema vrsti korisnika kredita**:

Banka može postaviti maksimalne granice za iznos kredita koje će odobravati određenoj grupi klijenata, što pomaže u diversifikaciji rizika.

- **Restrikcija odobravanja kredita za pojedina regionalna područja**:

Banka može postaviti ograničenja na odobravanje kredita za određena geografska područja kako bi smanjila koncentraciju rizika na određenim tržištima.

- **Polaganje depozita**:

Banka može zahtevati određeni nivo depozita od klijenata pre odobravanja kredita, čime se smanjuje broj kreditnih zahteva i povećava sigurnost portfolija.

Ove strategije pomažu bankama da održe zdrav i stabilan portfolio, minimizirajući potencijalne gubitke i osiguravajući dugoročnu održivost poslovanja.

Da bi se adekvatno procenio kreditni rizik, koriste se različiti parametri koji pomažu banci da predvidi očekivane gubitke u slučaju neizvršenja obaveza klijenta.

Jedna od jednostavnijih formula koja sumira najvažnije paramtre kreditnog rizika je sledeća:

$$EL = EAD * PD * LGD$$

Ovi parametri su ključni za razumevanje i upravljanje kreditnim rizikom i predstavljaju sledeće:

1. **EAD (Exposure at Default):**

Ova vrednost predstavlja izloženost banke u trenutku neizvršenja obaveza klijenta. U suštini, to je iznos glavnice kredita, zajedno sa pripadajućom kamatom, koje banka još uvek nije naplatila kroz proces amortizacije kredita u trenutku kada klijent prestane izvršavati svoje obaveze.

2. **PD (Probability of Default):**

Verovatnoća neizvršenja obaveza je numerička vrednost koja označava šansu da klijent neće otplatiti tri dospela mesečna anuiteta. Ovaj parametar predstavlja procenu verovatnoće da će klijent postati nesposoban da ispuni svoje obaveze prema banci.

3. **LGD (Loss Given Default):**

LGD predstavlja procenat gubitka koji će banka pretrpeti ukoliko klijent ne ispuni svoje obaveze i dođe do neizvršenja obaveza. Ovaj parametar odražava koliki deo iznosa glavnice kredita će banka izgubiti u slučaju da dođe do neplaćanja.

Pravilna procena i upravljanje ovim parametrima omogućava banci da bolje razume rizik i pravilno prilagodi strategije i uslove kreditiranja. Kroz analizu ovih parametara, banka može donositi informisane odluke o odobravanju kredita, postaviti odgovarajuće uslove i obezbediti adekvatno obezbeđenje kako bi smanjila potencijalne gubitke i osigurala stabilnost svog portfolija.

Neizvršenje kreditnih obaveza dužnika obično podrazumeva kašnjenje od 90 dana sa materijalno značajnim iznosom duga. U praksi, "materijalno značajan iznos duga" može se tumačiti kroz dva glavna pristupa: apsolutni i relativni prag materijalne značajnosti.

Apsolutni prag se odnosi na apsolutnu vrednost duga, dok relativni prag usmerava pažnju na odnos duga u odnosu na ukupne finansijske resurse dužnika. Važno je napomenuti da se ove definicije mogu razlikovati od države do države i zavise od regulativa i praksi u bankarskom sektoru u svakoj od njih. Takođe, smatra se da je dužnik u neizvršenju obaveza ukoliko prekrši neku od zaštitnih klauzula u kreditnom ugovoru, što automatski pokreće pregovore između banke i dužnika. U suprotnom, banka ima pravo da zahteva da dužnik odmah vrati celokupan dug. Važno je napomenuti da ova definicija neizvršenja obaveza od 90 dana sa značajnim iznosom duga predstavlja najčešće korišćenu definiciju u bankarstvu.

Kreditni proces započinje primanjem kreditnih zahteva od kompanija ili pojedinaca (fizičkih lica). Ti zahtevi prolaze kroz proceduru koja ima za cilj adekvatnu analizu kreditnog rizika. Banka se oslanja na različite izvore informacija relevantnih za ocenu kreditne sposobnosti, uključujući podatke koje podnosi tražilac kredita, interne baze podataka banke, kao i spoljne izvore informacija koje banka prikuplja. Interne baze podataka su posebno važne jer pružaju istorijske podatke o sličnim klijentima ili sličnim kreditnim proizvodima, kao i informacije o prethodnoj saradnji s tražiocem kredita. Na osnovu svih prikupljenih informacija, banka donosi kreditnu odluku u vidu odobrenja ili odbijenice, sa ciljem da proceni stepen kreditnog rizika i time se zaštiti.

Kreditna analiza je ključni proces koji banka obavlja prilikom odobravanja kredita tražiocu, s ciljem utvrđivanja njegove kreditne sposobnosti i procene stepena kreditnog rizika. Svaki odobreni kredit nosi određeni nivo kreditnog rizika, pa banka mora jasno definisati koliko rizika može prihvatiti u svojoj kreditnoj politici. U današnjem savremenom bankarstvu, proces kreditiranja postao je kompleksniji nego ranije. Značaj korišćenja više izvora informacija za tražioce kredita, posebno za kompanije, leži u potrebi za verifikacijom i proverom tačnosti podataka putem različitih izvora informacija. Povezivanjem i ukrštanjem ovih informacija, banka obezbeđuje bolju procenu kreditne sposobnosti i smanjenje rizika pri odobravanju kredita.

Savremeno bankarstvo koristi dva ključna pristupa za ispitivanje kreditne sposobnosti klijenata:

- **Klasična kreditna analiza** (kvalitativni pristup):

Ovaj pristup podrazumeva detaljnu analizu finansijskih izveštaja i poslovnih performansi tražioca kredita. Banka pažljivo procenjuje istoriju plaćanja, bilanse, izveštaje o dobiti i gubicima, kao i druge relevantne finansijske pokazatelje. Proučavanjem ovih informacija, banka stiče opisnu ocenu rizika klijenta, pružajući detaljan uvid u njegovu sposobnost da redovno izmiruje obaveze. Ovaj kvalitativni pristup omogućava bankama da dublje razumeju sve aspekte kreditne sposobnosti klijenta, kao i eventualne izazove s kojima se suočava.

- **Kreditni skoring/rejting model** (kvantitativni pristup):

Ovde se koriste sofisticirani kreditni skoring ili rejting modeli koji se oslanjaju na kvantitativne podatke i statističku analizu. Modeli koriste veliki broj različitih parametara, uključujući istoriju kreditnog ponašanja, ukupne prihode, zaduženost, godine poslovanja i mnoge druge faktore kako bi izračunali numeričku ocenu rizika za svakog klijenta. Ova kvantitativna procena omogućava objektivno i precizno merenje verovatnoće neizvršenja obaveza klijenta. Banka koristi ove modele da bi na osnovu statističkih podataka brže i efikasnije donela kreditne odluke.

Klasična kreditna analiza, koja i dalje ima primenu prevashodno prilikom odobravanja kredita većim kompanijama, sastoji se od temeljne analize finansijskih izveštaja kako bi se dobio opis rizičnosti tražioca kredita. Ova analiza obuhvata detaljnu procenu poslovnih aktivnosti, istorije poslovanja, broja zaposlenih, odnos sa povezanim licima, ponuđenih sredstava obezbeđenja, analizu platnog prometa, saradnju sa drugim bankama, strukturu prihoda, ključne finansijske indikatore, vlasničku strukturu tražioca kredita i rukovodstva, kao i imovinu u njihovom vlasništvu. Takođe, analizira se promena ovih faktora u najmanje dva poslednja finansijska izveštaja.

Klasična kreditna analiza zasniva se na subjektivnoj proceni kreditnog analitičara, koji koristi svoje znanje i iskustvo kako bi ocenio kreditnu sposobnost tražioca kredita. Međutim, ovaj pristup je poznat po tome što zahteva značajan vremenski i finansijski angažman, a oslanjanje na subjektivne stavove analitičara može dovesti do različitih ocena.

U cilju poboljšanja i ubrzanja procesa donošenja odluka, kao i smanjenja zavisnosti od subjektivnih faktora, banke su se okrenule razvoju kreditnih skoring modela. Ovi modeli koriste kvantitativne parametre kako bi automatski izračunali ocenu kreditnog rizika za svakog klijenta. Ovaj napredniji pristup omogućava bržu i precizniju ocenu kreditne sposobnosti i doprinosi efikasnosti i objektivnosti u procesu donošenja kreditnih odluka.

Kreditni rejting je termin koji obuhvata različite kriterijume kojima se procenjuje kreditna sposobnost dužnika da redovno servisira kredit. Ova ocena uključuje formalne kriterijume zasnovane na kreditnoj istoriji pojedinca ili kompanije, ali i apstraktne faktore poput reputacije ili životnih navika fizičkih lica ili ugleda kompanija. Čak i politička nestabilnost država može biti deo kreditnog rejtinga. Primenom rejting modela, važno je razlikovati eksterne i interne kreditne rejtinge.

Eksterni kreditni rejting obezbeđuju Agencije za kreditni rejting, od kojih su najpoznatije: "Moody's Investors Service", "Standard & Poor's", "Fitch Group" i "Dominion Bond Rating Service". Ove agencije nezavisno procenjuju kreditni rizik dužnika i dodeljuju ocene na osnovu informacija koje prikupljaju i analiziraju. Ove ocene služe kao referenca za investitore i kreditore da procene pouzdanost dužnika.

Interni rejting modeli, s druge strane, su razvijeni od strane samih banaka i koriste se za procenu kreditnog rizika njihovih klijenata (ili potencijalnih klijenata). Ovi modeli su rezultat dugogodišnjeg iskustva i podataka koje banke imaju o svojim klijentima. Banka koristi interne kriterijume kako bi procenila rizik svakog klijenta pojedinačno i donela odluku o odobravanju kredita.

Eksterni i interni kreditni rejting zajedno pružaju sveobuhvatnu sliku o kreditnoj sposobnosti dužnika. Eksterni rejtingi pružaju nezavisne i standardizovane ocene, dok interni rejting modeli omogućavaju bankama da prilagode procenu rizika u skladu sa njihovim specifičnim potrebama i internim podacima. Oba pristupa igraju ključnu ulogu u upravljanju kreditnim rizikom banaka i obezbeđivanju stabilnosti u finansijskom sektoru.

Kreditni skoring i metode ocenjivanja kreditnog skoringa

Kreditni skoring i samo njegovo modeliranje, bilo da su razvijeni interno ili eksterno, se široko primenjuju u kreditnoj industriji. Skoring modeli sažimaju relevantne informacije o zajmotražiocu ili klijentu banke i predstavljaju ih kroz niz poređanih kategorija (skorova) koji predviđaju ishod. Skor potrošača je numerički prikaz njihovog procenjenog rizika u datom trenutku. Ovi modeli omogućavaju brzo, efikasno i objektivno donošenje kreditnih odluka zasnovanih na iskustvu banke i/ili celokupne finansijske industrije.

Kreditni skoring modeli koriste se u različite svrhe, uključujući upravljanje rizikom, određivanje adekvatnih cena kreditiranja, smanjenje gubitaka, evaluaciju novih programa kredita, ubrzavanje procesa odobravanja kredita, osiguravanje dosledne primene kreditnih kriterijuma i povećanje profitabilnosti.

Ovi modeli se razvijaju statističkom analizom i odabiranjem karakteristika korisnika kredita koje se smatraju povezanim sa kreditnom sposobnošću. Različite metode tretiranja podataka koriste se za kreiranje skorova, od vrlo jednostavnih modela sa malo ulaznih podataka koji predviđaju jedan ishod, do vrlo složenih modela sa više ulaznih podataka koji predviđaju više ishoda.

Korišćenje skoring modela omogućava efikasnost, ali ne sme da zameni neophodnu dokumentaciju za odobravanje kredita ili osnovne kreditne kriterijume. Neispravno, nepravilno ili loše razvijeni skoring modeli mogu izazvati probleme u selekciji klijenata i upravljanju potraživanjima, što može dovesti do povećanog kreditnog rizika, smanjenja profitabilnosti i otežanog poslovanja banke.

Dobro vođeni skoring modeli moraju biti pažljivo razvijeni, implementirani, testirani i održavani, jer mogu značajno uticati na različite faze životnog ciklusa kredita. Banka treba

redovno da validira modele i prilagodi ih u skladu sa promenama u kreditnom okruženju. Takođe, očekuje se da se skoring modeli primenjuju u skladu sa relevantnim zakonima i propisima.

Napredak u tehnologiji i analitičkim alatima omogućava bankama da sve efikasnije procenjuju kreditni rizik i donose informisane odluke. Analitički modeli, uključujući metode kreditnog skoringa i statističke analize, pružaju sveobuhvatniji uvid u kreditnu sposobnost dužnika. Pored toga, bankarske institucije takođe koriste i ekonomske indikatore, makroekonomske projekcije i ostale relevantne podatke kako bi bolje razumeli rizike sa kojima se suočavaju.

2.1 Istorijat kreditnog skoringa

Kreditni skoring suštinski predstavlja problem klasifikacije, gde se aplikanti svrstavaju u različite grupe. Prema Thomasu [21], tehnike statističke klasifikacije su započele kada je Fisher (1936) razvio jedan od prvih uspešnih modela za klasifikaciju tri različite vrste cveta iris. On je koristio merenja različitih karateristika cveta kako bi razlikovao njegove tri vrste. Zatim je Durand [6] prvi prepoznao da ove statističke tehnike klasifikacije mogu biti primenjene za klasifikaciju dobrih i loših kredita. Pre toga, kako navodi Thomas [21], finansijske institucije su donosile odluke o odobravanju kredita subjektivno.

Uvođenjem kreditnih kartica tokom 1960-ih, počela je da se shvata korisnost kreditnog skoringa. Zbog velikog broja ljudi koji su aplicirali za kreditne kartice, automatizacija postupka obrade kreditnih zahteva se činila kao jedino rešenje. Kada su finansijske institucije uvele kreditne skoring modele, primetile su da su ovi modeli daleko efikasniji od prethodnih (subjektivnih) metoda odlučivanja. Rezultat je bio da su, kako Thomas [21] navodi, stope nesolventnosti pale za 50% ili više.

U 1980-ima, uspeh kreditnog skoringa kod kreditnih kartica doveo je do toga da su finansijske institucije počele da koriste skoring metode i za druge proizvode, kao što su gotovinski krediti, stambeni krediti i poslovni krediti.

Sa razvojem računara i njihovom sveprisutnom primenom u svakodnevnom životu, tokom 80-ih godina prošlog veka, omogućena je implementacija linearnog programiranja i logističke regresije u skoring modele. U poslednjim godinama, kombinacija standardnih metoda sa tehnikama veštačke inteligencije i neuronskih mreža postala je sve zastupljenija.

Godinama unazad, primarni cilj razvoja i korišćenja kreditnih skoring modela bio je razlikovanje dobrih od loših plasmana kako bi se minimizirao rizik ulaganja u loše klijente. Međutim, ovaj cilj se danas modifikovao u postizanje maksimalnog profita za finansijske institucije na svakom odobrenom kreditu.[6]

2.1.1 Tradicionalni skoring modeli

U delu koji sledi, dat je pregled tradicionalnih skoring modela, koji se smatraju pionirskim koracima u modeliranju kreditnog skora.

- **"5C" model**

Tradicionalni model "5C" postavlja pet osnovnih kriterijuma za kreditnu analizu ("5 Cs of Credit"[16])

- Karakteristike tražioca kredita (***Character***)

Procena karaktera dužnika obuhvata analizu ličnih osobina, poslovnog ugleda i rukovodstva tražioca kredita. Ova procena predstavlja subjektivnu ocenu finansijske institucije, koja se temeljno analizira kako bi se utvrdila poslovna reputacija, vrsta delatnosti i pravni status tražioca kredita. Cilj je identifikacija odgovornosti, integriteta i tačnosti u izmirivanju obaveza i doslednosti u vođenju poslovnih knjiga. Na osnovu ovog procesa zaključujemo spremnost i želju tražioca kredita da adekvatno servisira svoje obaveze koje proizilaze iz odobrenog kredita

- Kapacitet ili sposobnost otplate (***Capacity***)

Pre odobravanja kredita, finansijska institucija mora identifikovati najmanje dva jasno razdvojena i nezavisna izvora otplate kredita. Ovi izvori mogu uključivati dobit, prihod od prodaje imovine, prihod od prodaje akcija ili sredstva dobijena od drugih finansijskih institucija.Prilikom analize kapaciteta tražioca kredita, potrebno je pažljivo analizirati njihovu dobit, očekivanu buduću dobit (uzimajući u obzir obim poslovanja i njegovu prirodu), postojeći dug i strukturu troškova. Na osnovu ovih informacija, zaključujemo o sposobnosti tražioca kredita da otplati svoje kreditne obaveze na osnovu aktuelnih prihoda koji će biti generisani tokom perioda ugovorene otplate kredita.

- Kapital ili imovina dužnika (***Capital***)

Kapital predstavlja jedan od izvora kojim dužnik može otplatiti kredit, stoga je kreditni analitičar u procesu finansijske analize dužan da adekvatno proceni njegovu stvarnu vrednost. Kapital tražioca kredita odražava neto imovinu, što je jedan od pokazatelja finansijskog stanja u prethodnom periodu. Neto imovina se dobija kao razlika između ukupnih sredstava i ukupnih obaveza. Na osnovu kapitala tražioca kredita, banka ograničava iznos kredita koji može da odobri i uslove pod kojima će kredit biti odobren. Veći iznos stalnog kapitala predstavlja manji kreditni rizik za banku. Kako bi se zaštitila od rizika koji proizilazi iz netačnih finansijskih izveštaja, neophodno je da revizori koji su izvršili pregled pruže pozitivno mišljenje. Ovo omogućava finansijskim institucijama da donose informisane odluke o odobravanju kredita i utvrde potencijalni kreditni rizik koji nosi tražilac kredita.

- Kolaterali ili ostala sredstva obezbeđenja (***Collateral***)

Zaloga ili kolateral predstavlja stvarno sredstvo koje banka zahteva kao obezbeđenje od kreditnog rizika, a takođe služi kao sekundarni izvor otplate kredita. Kolateral obično predstavlja uslov za odobravanje kredita, pri čemu je neophodno pružiti jasne dokaze o vlasništvu nad tim sredstvima, jedinstvenu identifikaciju i dokazanu tržišnu vrednost.

Kako bi se banka zaštitila od mogućeg gubitka, praksa je da se uzima određeni procenat tržišne vrednosti kolaterala, što je jasno definisano kreditnom politikom banke. Na taj način, u slučaju da dužnik ne može izmiriti kreditne obaveze, banka ima pravo da proda kolateral kako bi se obezbedila naplata duga.

Kolateral pruža dodatnu sigurnost banci i smanjuje kreditni rizik, omogućavajući bankama da odobravaju kredite uz povoljnije uslove i kamatne stope, jer imaju siguran način da se zaštite u slučaju da dužnik ne ispuni svoje obaveze.

- Uslovi u okruženju i poslovanju (***Conditions***)

Ekonomski uslovi okruženja mogu imati značajan uticaj kako na tražioca kredita, tako i na banku. Nepovoljni ili promenljivi makroekonomski uslovi mogu dovesti do većih gubitaka i ugroziti sposobnost tražioca kredita da izmiri svoje kreditne obaveze. Stoga je veoma važno na početku proceniti trenutne tržišne uslove i situaciju na tržištu kojoj je izložen tražilac kredita, identifikovati potencijalne konkurente ili prilike za zapošljavanje u preduzeću.

Budući uslovi poslovanja tražioca kredita takođe treba da se uzmu u obzir, posebno u odnosu na rokove vraćanja kredita. Za kratkoročne kredite, lakše je proceniti trendove budućih promena i njihovih efekata na poslovanje preduzeća. Međutim, što je period kreditiranja duži, to je teže realno sagledati buduće tržišne kretanja.

U svakom slučaju, banka mora pažljivo analizirati sve ekonomske uslove i faktore koji mogu uticati na sposobnost tražioca kredita da otplati kreditne obaveze. Ovo je ključno za donošenje informisane odluke o odobravanju kredita i za smanjenje rizika za banku.

Od svih karakteristika kreditne sposobnosti, najveći značaj se pridaje prvom karakteru - volji tražioca kredita da redovno servisira svoje dospelu obaveze prema banci. Banka može odobriti kredit i komitentu čija je kreditna sposobnost na granici prihvatljivog, ali u takvim slučajevima preuzima veći rizik i može tražiti plaćanje veće kamate nego što je uobičajeno. U praksi, kada banka odobrava kredit rizičnijem tražiocu kredita, ona uzima jaka sredstva obezbeđenja kako bi umanjila kreditni rizik koliko god je to moguće. Obezbeđenje može biti u vidu imovine ili nekih drugih vrednih sredstava koja se koriste kao garancija da će banka biti nadoknađena u slučaju da tražilac kredita ne izmiri svoje obaveze. S obzirom na veći rizik koji nosi ovakva vrsta kreditiranja, banka ima pravo da postavi strože uslove i traži dodatne garancije kako bi se zaštitila od potencijalnih gubitaka. To uključuje i mogućnost naplate veće kamatne stope kako bi se kompenzovao povećani rizik.

- **Beaver-ov model**

Prvi statistički modeli su bili univarijantni, obično bazirani na računovodstvenim podacima. Beaver [2] je prezentovao svoj model koji koristi kombinacije finansijskih pokazatelja I upoređuje takve pokazatelje tražioca kredita sa standardima u industriji u kojoj tražilac kredita pripada.

Beaver je svoj model za procenu finansijskog neuspeha bazirao na sledeća tri pokazatelja:

- $\frac{tok\ novca}{ukupna\ imovina}$
- $\frac{čist\ prihod}{ukupni\ dugovi}$
- $\frac{tok\ novca}{ukupni\ dugovi}$

Beaver je za svaki pojedinačni pokazatelj izračunao graničnu vrednost, a na osnovu tog rezultata, klasifikovao tražioce kredita u grupu potencijalno uspešnih ili potencijalno neuspešnih. Ako je finansijski pokazatelj bio iznad propisane granične vrednosti, tražilac kredita je svrstan u grupu potencijalno uspešnih, dok je ako je bio ispod granične vrednosti svrstan u grupu potencijalno neuspešnih. Korišćenjem univarijantnih modela koji uključuju samo jedan pojedinačni pokazatelj, zaključivanje je bilo izuzetno teško i ograničeno.

- **Z-skor model**

Z-skor model je kreirao Edward Altman, koristeći standardnu linearnu regresiju. Kao nezavisne promenljive, Altman je koristio 30 finansijskih pokazatelja, ali usavršavanjem modela došao je do finalnih pet indikatora:

1. $X_1 = \frac{tekuća\ aktiva}{ukupna\ aktiva}$
2. $X_2 = \frac{zadržani\ dobitak}{ukupna\ aktiva}$
3. $X_3 = \frac{operativni\ dobitak}{ukupna\ aktiva}$
4. $X_4 = \frac{tržišna\ vrednost\ glavnice}{knjigovodstvena\ vrednost\ ukupnog\ duga}$
5. $X_5 = \frac{prihod\ od\ prodaje}{ukupna\ aktiva}$

Zavisna promenljiva, koju je Altman nazvao Z-skor predstavlja skor kreditnog rizika koji meri uspeh ili neuspeh tražioca kredita u otplati odobrenog kredita. Ocenjene koeficijente za svaku promenljivu dobio je na osnovu analize uzorka koji je uključivao 33 uspešna i 33 neuspešna preduzeća. Opšta formula Z-skor modela je sledeća:

$$Z = 0.012\,X_1 + 0.014\,X_2 + 0.033\,X_3 + 0.006\,X_4 + 0.010\,X_5$$

Granice za promenljivu Z su sledeće:

- *Zona bankrotstva*: $Z < 1.81$
- *Siva zona* : $1.81 \leq Z < 2.99$
- *Bezbedna zona* : $Z < 2.99$

U zavisnosti od delatnosti tražioca kredita postoje i posebne Altmanove formule koje nećemo navoditi.

Nakon Z-skor modela, Altman, Haldeman i Harayanan su razvili ZETA model. Glavni cilj ovog modela bio je analizirati i testirati klasifikaciju preduzeća na one koja će bankrotirati i one koja neće. Na početku su analizirali 27 promenljivih, ali konačni ZETA model je uključivao samo 7 promenljivih. U poređenju sa Z-skor modelom, ZETA model se pokazao preciznijim u predviđanju neuspešnih preduzeća 2 do 5 godina pre nego što zapravo bankrotiraju, dok su za prvu godinu tačnost oba modela gotovo jednaka.

2.2 Tipovi kreditnog skoringa

Danas, sve veći broj banaka koristi više vrsta kreditnih skorova kako bi procenile bonitet pojedinaca ili pravnih lica. U ovom delu predstavljamo različite skorove koji se često koriste u finansijskoj, a pre svega u bankarskoj industriji. Tradicionalno, svaki skor i model su se smatrali odvojenim alatima za procenu kreditne sposobnosti. Međutim, primećuje se trend integrisanja ovih modela i skorova tokom celog životnog ciklusa računa odnosno finansijskog proizvoda.[21]

Ako posmatramo podatke koji se koriste u njihovom razvoju postoje 2 vrste kreditnih skoring modela:

1. Generički kreditni skoring modeli

Ovi kreditni modeli se baziraju na podacima kreditnih biroa, koji raspolažu veoma velikom bazom podataka o kreditnoj istoriji klijenata. Korišćenjem takve baze podataka primenom različitih metoda kreiraju se kreditni skoring modeli koji obuhvataju one karakteristike potencijalnih klijenata koje najbolje predviđaju buduće ponašanje u otplati kredita.

Najpoznatiji i najčešće korišćeni skorovi kreditnih biroa nazivaju se FICO skorovi. Ovi skorovi potiču iz modeliranja koje je započela kompanija Fair, Isaac and Company (sada poznata kao Fair Isaac Corporation) (Fair Isaac), odakle i naziv "FICO" skor.

FICO skorovi uzimaju u obzir podatke u pet oblasti kako bi odredili kreditnu sposobnost zajmoprimca: istoriju plaćanja, trenutni nivo zaduženosti, vrste korišćenih kredita, dužinu kreditne istorije i nove kreditne račune[11]. Da bi se odredili kreditni skorovi, FICO različito vrednuje svaku od pomenutih kategorija, za svakog pojedinca. Međutim, generalno, istorija plaćanja čini 35% skora, iznos dugovanja čini 30%, dužina kreditne istorije 15%, novi krediti 10%, a miks kredita 10%.[11] Opseg FICO skora je između 300 i 850 gde veći broj poena predstavlja veći kreditni rizik.

2. Kreditni skoring modeli prilagođeni korisniku

Bazirani su na podacima o klijentima (ili potencijalnim klijentima) konkretne finansijske institucije. Dakle, razvijaju se zasebno za svaku finanansijsku instituciju (banku). Procedure zasnovane na statističkim metodama se primenjuju na podatke koje banka poseduje, gde se izdvajaju one karakteristike klijenta koje su značajne za otplatu kredita ili njegovo ponašanje uopšteno.[23]

Zavisno od namene neki od tipova kreditnog skoringa su:

- **Aplikativni skoring model**

Aplikativni skoring uključuje dodeljivanje vrednosti prediktivnim varijablama u toku aplikacije za kredit, pre donošenja odluka o odobravanju kredita. Tipični aplikativni podaci uključuju stavke poput dužine zaposlenje, dužine vremena boravka na trenutnoj adresi, načina stanovanja (podstanar ili sopstveno prebivalište) i nivo prihoda. Najčešće se bodovi dodeljeni vrednostima varijabli sumiraju gde se dobija aplikativni skor klijenta (moguće je da za različite kreditne proizvode klijent ima različit skor). Rezultati aplikativnog skora, pored odluke o odobravanju kredita mogu odrediti i uslove kredita (maksimalni odobreni limit itd.).

- **Bihevioralni skoring model (modeli ponašanja)**

Ovi modeli se koriste za interno rangiranje klijenata koje se dobija na osnovu istorijskih podataka dostupnih za pojedinačnog klijenta iz prethodne saradnje sa bankom. Ovakvi modeli pomažu pri oceni kreditnog rizika jer je banka već upoznata sa ranijim ponašanjem klijenta. Podaci dobijeni ovakvim modelom se mogu koristiti i pri odobravanju plasmana, prilikom računanja rezervisanja kao i za proces naplate.

- **Skoring model za predviđanje bankrota**

Ova vrsta modela služi za predikciju koji klijenti će najverovatnije bankrotirati. Svrha ovakvih modela je sprovođenje radnji koje mogu minimizovati potencijalni gubitak.

- **Skoring modeli za detekciju prevare (fraud modeli)**

Skoring modeli za detekciju prevare nastoje da identifikuju račune ili aplikacije za kredit sa potencijalnom prevarnom aktivnošću. Prevara i dalje ostaje rasprostranjena najviše u industriji kreditnih kartica, pa zatim keš kredita, sa posebnim akcentom na kredite plasirane putem digitalnih kanala. Otkrivanje potencijalnih prevarnih aktivnosti može pomoći u identifikaciji i kontroli gubitaka, kao i pomoći rukovodstvu u razvoju kontrola za sprečavanje prevare.

Otkrivanje prevare je od suštinskog značaja jer može zaštititi banku od neovlašćenih transakcija i finansijskih gubitaka. Kroz analizu transakcionih uzoraka i neuobičajenih obrazaca ponašanja, skorovi za detekciju prevare pomažu u identifikaciji računa koji zahtevaju dodatnu proveru. Osim toga, ovakvi skorovi omogućavaju finansijskim institucijama da razvijaju sofisticirane strategije za sprečavanje prevare, uključujući dinamičke promene u autentifikaciji transakcija i automatsko obaveštavanje korisnika o sumnjivim aktivnostima na njihovim računima.

Važno je naglasiti da je prevencija prevare konstantan izazov zbog evolucije tehnika prevaranata, stoga je upotreba naprednih analitičkih alata ključna komponenta efikasnog upravljanja rizikom prevare u bankarskoj industriji.

- **Skoring modeli za projekciju plaćanja**

Ovi modeli koriste interne podatke u banci u cilju rangiranja klijenata obično po relativnom procentu duga koji će se otplatiti. Neki modeli predviđaju procenat otplate duga, dok drugi klasikuju verovatnoću otplate duga. Rezultati ovog modela se obično koriste u početnim i srednjim stupnjevima delikvencije.

- **Skoring modeli odziva (response modeli)**

Modeli za ocenjivanje odziva koriste se za upravljanje troškovima akvizicije novih klijenata. Identifikacijom potencijalnih klijenata koji će se najverovatnije odazvati, banka može prilagoditi svoje marketinške kampanje kako bi ciljala upravo tu grupaciju. Istovremeno, banka izbegava trošenje marketinških sredstava na potencijalne klijente za koje je manje verovatno da će uspostaviti saradnju sa bankom.

- **Skoring modeli za optimizaciju prihoda**

Ova vrsta modela se primenjuje kako bi se maksimizovali prihodi uzimajući u obzir nivo rizika koji nosi svaki klijent. To znači da se za klijente sa niskim rizikom da neće ispunjavati svoje obaveze nude niže kamatne stope, dok se za klijente sa visokim rizikom povećavaju kamatne stope. Ovakav pristup omogućava banci da optimalno upravlja svojim portfoliom, nagrađujući pouzdane klijente povoljnijim uslovima, dok istovremeno kompenzuje veći rizik sa višim kamatama za one klijente kod kojih postoji veća verovatnoća neizvršenja obaveza.

- **Attrition scoring**

Skoring modeli za predviđanje prekida saradnje sa bankom pokušavaju da identifikuju potrošače koji su najverovatnije skloni zatvaranju svojih računa, prevođenju računa u neaktivan

status ili značajno smanjenju njihovih preostalih salda. Identifikacija takvih klijenata omogućava upravi banke da preduzme proaktivne mere kako bi na efikasan način zadržala te klijente i nastavila saradnju. Ovi modeli prevashodo imaju za cilj zadržavanje dobrih klijenata banke.

- **Skoring modeli za naplatu**

Ovi modeli se kako bi se utvrdile strategije naplate, dodeljivanje redosleda za naplatu, dodeljivanje redosleda za automatske pozive, kao i postupci saradnje sa agencijama za naplatu i slično. Ocena naplate se obično koristi u srednjem i kasnom stadijumu kašnjenja plaćanja.

2.3 Metode kreditnog skoringa

U svetu finansija, razvoj kreditnih skoring modela ima dugu istoriju. Početak je bio obeležen tradicionalnim metodama kao što su analiza diskriminante i logistička regresija. Ove metode su se dokazale kao dragocen alat za procenu kreditne sposobnosti klijenata. Analiza diskriminante je fokusirana na identifikaciju ključnih faktora koji razdvajaju dobre i loše klijente, dok logistička regresija modelira verovatnoću ishoda na osnovu više varijabli. Ovi pristupi su dugo bili i može se reći da i dalje jesu standardni izbor u bankarskom sektoru zbog svoje interpretabilnosti i mogućnosti da se analiziraju uticaji različitih faktora na donošenje odluka.

U poslednjim decenijama, razvoj tehnologije i dostupnost velikih količina podataka (*big data*) doneli su novu eru u oblasti kreditnog skoringa. Modeli kao što su stabla odlučivanja i neuronske mreže postali su sveprisutni.[23] Stabla odlučivanja pružaju fleksibilnost i sposobnost rada sa kompleksnim podacima. Ona konstruišu hijerarhijske odluke na osnovu serije pitanja i kriterijuma. Sa druge strane, neuronske mreže koriste duboko učenje da identifikuju obrasce i veze u podacima. Njihova sposobnost da prepoznaju kompleksne obrasce daje im prednost u analizi podataka visokih dimenzija.

Razlikovanje između parametarskih i neparametarskih modela je takođe važno. Parametarski modeli, kao što su logistička regresija i analiza diskriminante, pretpostavljaju određenu funkcionalnu formu i estimiraju parametre te funkcije. Ovi modeli su lakši za interpretaciju, ali su podložni pretpostavkama o raspodeli podataka. Nasuprot tome, neparametarski modeli, kao što su stabla odlučivanja i neuronske mreže, ne pretpostavljaju određenu funkcionalnu formu i imaju veću fleksibilnost u radu sa različitim tipovima podataka. Međutim, njihova složenost čini ih težim za interpretaciju.

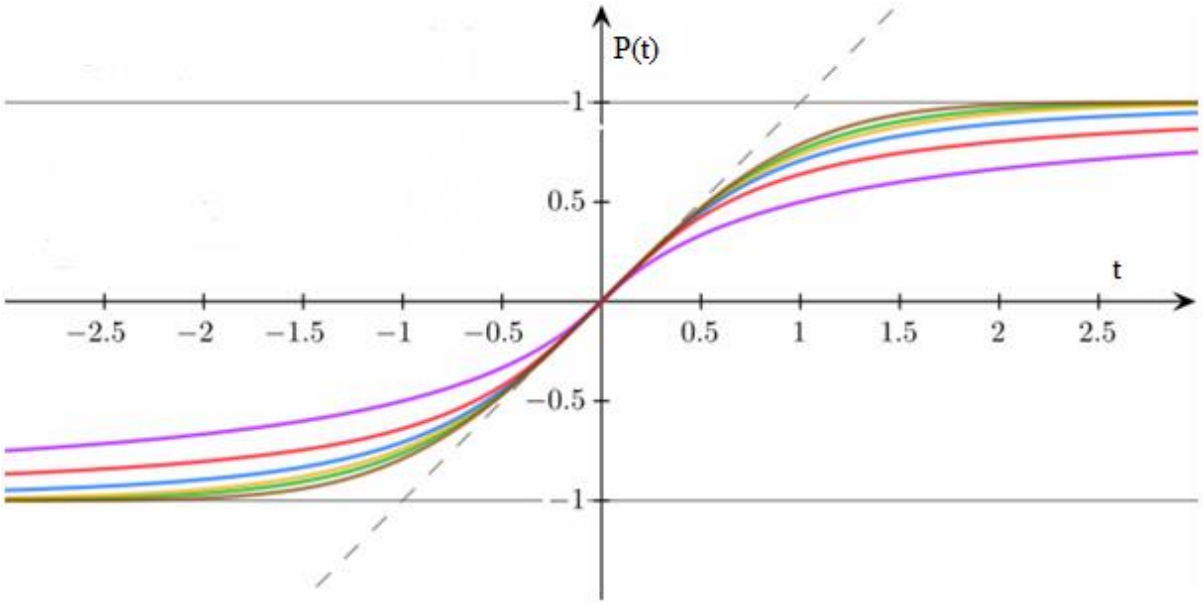
U suštini, dok su tradicionalni modeli kao analiza diskriminante i logistička regresija i dalje značajni i koriste se, moderni pristupi poput stabala odlučivanja i neuronskih mreža donose novu dubinu i širinu analize. Bankarski sektor sve više prepoznaje važnost ovih naprednih metoda u pravovremenom i tačnom donošenju odluka o kreditnoj sposobnosti klijenata.

U nastavku je pregled najčešćih metoda za razvoj kreditnih skoring modela:

- **Logistička regresija**

Logistička regresija je klasična metoda koja se široko koristi u modeliranju kreditnih skoringa. Ova tehnika omogućava predviđanje verovatnoće da će se određeni događaj (kao što je neizvršenje obaveza) dogoditi na osnovu karakteristika klijenata.

Osnova logističke regresije je logistička funkcija, koja transformiše linearnu kombinaciju ulaznih karakteristika u vrednost između 0 i 1, koja se interpretira kao verovatnoća (Slika 1). Ulazne karakteristike se množe sa odgovarajućim težinskim koeficijentima, koji se prilagođavaju tokom procesa treniranja modela kako bi se postigla što bolja predikcija. Ovi težinski koeficijenti predstavljaju stepen uticaja svake karakteristike na krajnji rezultat.



Slika 1 – Logistička funkcija

Izvor: Gorica Gvozdić, Primenjena logistička regresija

Prednost logističke regresije leži u njenoj interpretabilnosti. Kako se svaka karakteristika vrednuje svojim težinskim faktorom, moguće je analizirati kako svaka karakteristika doprinosi konačnoj oceni klijenta. Ovo je posebno važno u finansijskom sektoru gde transparentnost odluka ima ključnu ulogu.

Uprkos izazovima, logistička regresija ostaje važan alat u kreditnom skoringu zbog svoje interpretabilnosti i sposobnosti da pruži razumne rezultate čak i sa manje podataka. Kombinovanje logističke regresije sa drugim tehnikama može dodatno poboljšati tačnost i predikciju modela.

Napomenimo da u ovom poglavlju nismo predstavljali matematičke osnove logističke regresije, jer je to tema jednog od narednih poglavlja.

- **Analiza diskriminante**

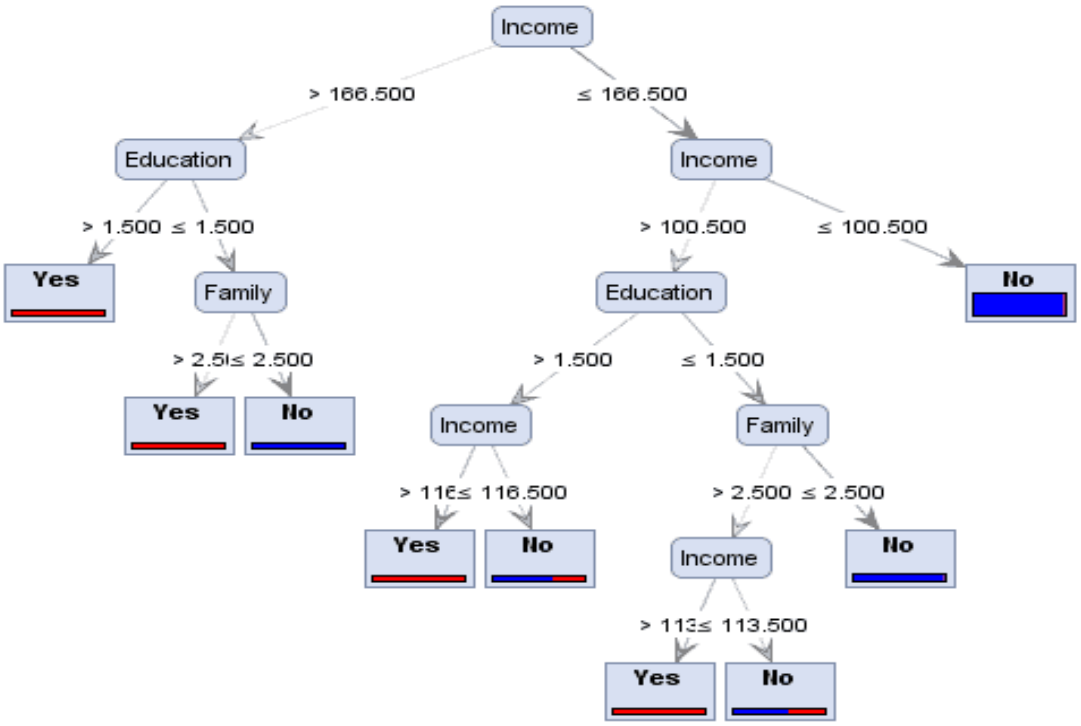
Analiza diskriminante je klasična tehnika klasifikacije koja se često koristi za predviđanje bankrota kod korporativnih klijenata. Osnova ove metode je stvaranje linearnih diskriminatnih funkcija koje pomažu razdvajanju različitih grupa klijenata, kao što su dobri i loši klijenti, na osnovu njihovih karakteristika. Ove funkcije se zasnivaju na dostupnim podacima o klijentima i oblikuju se kao linearne kombinacije tih podataka. Rezultat analize diskriminante, poznat kao diskriminacioni skor, ukazuje na očekivanu verovatnoću da će klijent ispoštovati svoje finansijske obaveze, pri čemu viši skor implicira veći rizik. Cilj ove metode je maksimizirati razlike između različitih grupa klijenata, dok se istovremeno minimiziraju varijacije unutar svake grupe. Važno je napomenuti da rezultati ove analize ne mogu direktno da se tumače kao verovatnoće, već se koriste za relativno rangiranje klijenata i upoređivanje njihovog rizika.

- **Stabla odlučivanja (Decision trees)**

Još jedan izuzetno koristan metod u razvoju modela kreditnog skoringa su stabla odlučivanja (Slika 2). Ova tehnika, takođe poznata kao drvo raspodele ili stabla klasifikacije, pruža široku primenu u analizi podataka. Stabla odlučivanja su modeli koji se sastoje od niza if- then-else (ako-tada-u suprotnom) uslova, koristeći se za klasifikaciju slučajeva u različite grupe.

Osnovna ideja ovog metoda je da se početni skup podataka podeli na manje grupe u skladu sa nezavisnim promenljivama. Na primer, za binarnu klasifikaciju, svaki čvor stabla odgovara određenom uslovu i deli podatke na dve podgrupe. Kroz ovaj proces, posmatranja se kreću naniže kroz drvo, u skladu s uslovima donošenja odluka, sve dok se ne dostigne krajnji čvor koji predstavlja klasifikaciju za to posmatranje.

Stabla odlučivanja su posebno korisna zbog svoje sposobnosti razumevanja i interpretacije veza između promenljivih. Ona takođe mogu raditi sa neprekidnim promenljivima i već postojećim kategorijalnim promenljivima. Važno je napomenuti da se sve nezavisne promenljive tretiraju kao kategorijalne, što predstavlja ključnu razliku u odnosu na parametarske modele.



Slika 2 – Primer stabla odlučivanja

Izvor: https://www.simafore.com/how-to-use-decision-trees-for-credit-scoring-part-3-of-4/

Iako stabla odlučivanja pružaju razumljive rezultate i mogućnost interpretacije, ona takođe mogu biti sklona preprilagođavanju (*overfitting*) podacima za obuku. Ovo znači da se model može preterano prilagoditi trening podacima i imati smanjenu sposobnost generalizacije na nove podatke. Kako bi se to rešilo, često se koriste metode kao što su ograničenja dubine stabla i upotreba više stabala (poput Random Forest-a) za stvaranje boljih i pouzdanijih modela za predviđanje kreditnog rizika.

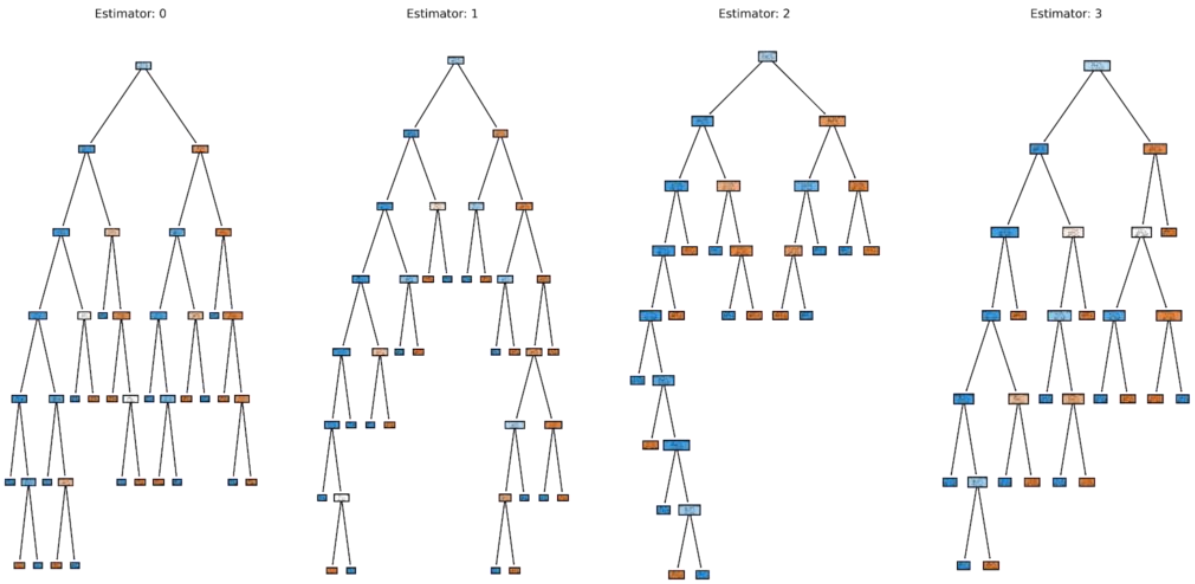
- **Random forest**

Random Forest (Slika 3) je dalji korak u evoluciji metoda kreditnog skoringa i predstavlja svojevrsno poboljšanje u odnosu na stabla odlučivanja. Ova tehnika se zasniva na konceptu konstruisanja više stabala odlučivanja i kombinovanja njihovih rezultata kako bi se postigla bolja predikcija.

Zamislite da se svako stablo odlučivanja konstruiše na drugačiji način, koristeći različite podskupove trening podataka i neke nasumične uslove za svaki čvor. Kada se svi ovi modeli kombinuju, efikasno se smanjuje tendencija preprilagođavanja (*overfitting*) i poboljšava se generalizacija modela. Time se postiže bolja sposobnost modela da se nosi sa novim, nepoznatim podacima, kao i da bolje identifikuje ključne veze između promenljivih. U odnosu na čistu upotrebu stabala odlučivanja, Random Forest donosi više pouzdanosti i stabilnosti u rezultate. To je zato što se, uzimajući u obzir različite načine konstrukcije svakog stabla, model lakše prilagođava različitim tipovima podataka i složenim relacijama među promenljivama.

Dodatno, Random Forest takođe može rukovati sa većim brojem nezavisnih promenljivih, bez potrebe za pretpostavkama o njihovoj raspodeli. Ovaj algoritam se takođe koristi za rešavanje problema neprekidnih i kategoričkih promenljivih.

U suštini, Random Forest je kao tim stručnjaka koji donose nezavisne odluke, a zatim kombinuju svoje odgovore kako bi se postigla najbolja moguća odluka. Ova tehnika pruža stabilnost, preciznost i sposobnost tumačenja rezultata, čineći je snažnim alatom u razvoju modela kreditnog skoringa. Međutim, potencijalni izazov u primeni Random Forest-a za kreditni skoring može biti teže tumačenje rezultata, budući da kombinacija više stabala često dovodi do kompleksnih modela čija interpretacija može biti zahtevnija.



Slika 3 – Primer za Random Forest

Izvor: https://brunaw.com/slides/rladies-dublin/RF/intro-to-rf.html#1

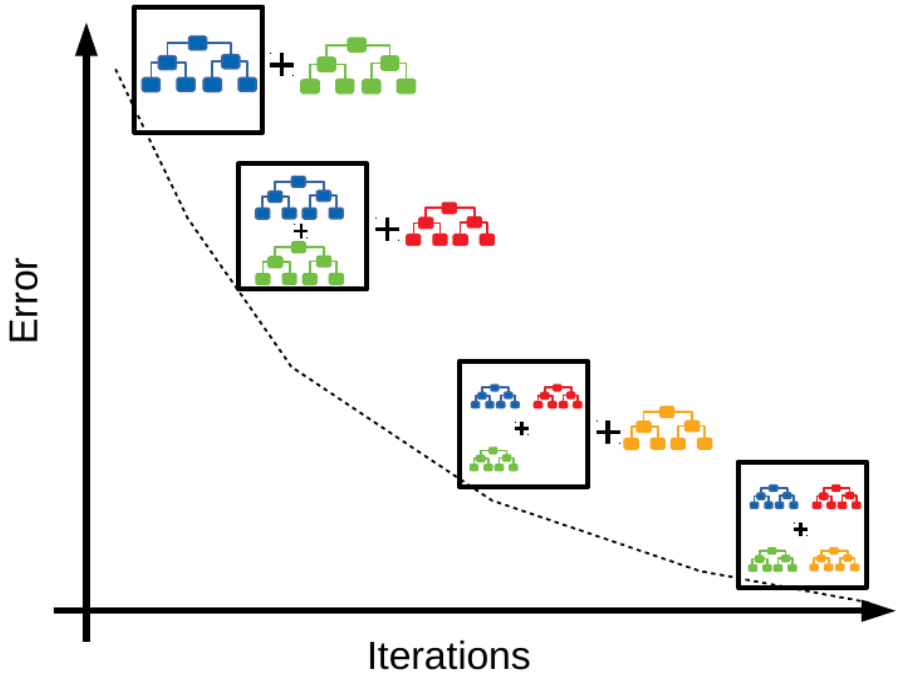
- **Gradient Boosting**

Gradient boosting je moćan algoritam mašinskog učenja koji se često primenjuje u modeliranju kreditnog skoringa. Ovaj algoritam funkcioniše tako što kombinuje više slabijih modela (ensamble tip modela kao i pomenuti random forest) kako bi se formirao jedan snažan model. Glavna ideja iza gradient boostinga je da svaki novi model pokuša popraviti greške koje prethodni modeli nisu uspeli ispraviti.

Postupak gradient boostinga počinje od izgradnje osnovnog modela, često se koristi jednostavna tehnika poput stabla odlučivanja. Nakon toga, analizira se kako su se ovaj model ponašao i gde su greške nastale. Novi model se zatim gradi sa ciljem da ispravi ove greške. Važno je napomenuti da svaki novi model ne zamenjuje prethodni, već ga dodaje i doprinosi ukupnoj preciznosti.

U kontekstu kreditnog skoringa, gradient boosting se može primeniti na različite načine. Na primer, može se koristiti za poboljšanje preciznosti postojećih skoring modela, posebno ako se primećuje da postoje određene greške ili nedostaci u modelima. Takođe, gradient boosting može efikasno raditi sa velikim i kompleksnim skupovima podataka, što ga čini pogodnim za analizu velikih portfolija klijenata.

Prednost gradient boostinga je u tome što može obuhvatiti nelinearne veze i interakcije među različitim karakteristikama, što može biti ključno za preciznije predviđanje kreditne sposobnosti klijenata. Međutim, kao i svaka metoda, i gradient boosting ima svoje izazove. Potrebno je vreme i trud za podešavanje parametara kako bi se postigao optimalan rezultat (Slika 4). Takođe, zbog kompleksnosti algoritma, postoji opasnost od preprilagođavanja (*overfitting*) podacima.



Slika 4 – Gradient Boosting algoritam

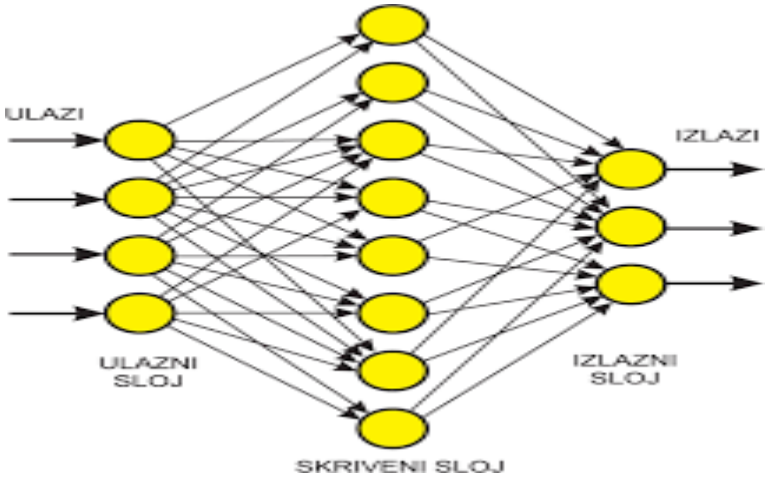
Izvor: https://medium.com/swlh/gradient-boosting-trees-for-classification-a-beginners-guide-596b594a14ea

- **Neuronske mreže**

Neuronske mreže su moćan alat inspirisan biološkim nervnim sistemom. One su alternativna metoda parametarskim tehnikama sa mnogo prednosti. One pružaju fleksibilniji dizajn i sposobnost modeliranja kompleksnih odnosa između faktora. Osim toga, neuronske mreže omogućavaju objektivan pristup bez inherentnih pretpostavki koje karakterišu parametarske modele. One su takođe netransparentne, što znači da njihove odluke nisu lako interpretirane, ali

upravo to omogućava ‘hvatanje’ nijansiranih i složenih veza među nezavisnim i zavisnim promenljivim.

Sistem neuronskih mreža sastoji se iz tri osnovna sloja parametara: ulaznih, skrivenih i izlaznih (Slika 5). Ulazni parametri obrađuju početne karakteristike kroz slojeve skrivenih parametara. Ovi skriveni parametri koriste funkcije aktivacije, kao što su hiperbolička tangenta ili logistička funkcija, kako bi pravilno utežili informacije pre nego što se proslede izlaznim parametrima. Kroz ovaj proces, kreiraju se mnogi čvorovi koji omogućavaju otkrivanje složenih nelinearnih veza među podacima.



Slika 5 – Primer neuronske mreže

Izvor: http://pc.pcpress.rs/tekst.php?id=9688

Iako neuronske mreže imaju mnoge prednosti, važno je uzeti u obzir i njihove nedostatke. Za uspešno postavljanje i treniranje neuronskih mreža, potrebna je značajna količina podataka i tehničko znanje. Takođe, zbog njihove netransparentnosti, interpretacija rezultata može biti složena, što može predstavljati problem u okruženjima gde je razumljivost odluka od suštinskog značaja.

- **Bejzove metode**

Bejzove metode pružaju moćan okvir za modeliranje i analizu rizika u kreditnom skoringu. Bejzova statistika temelji se na Bejzovoj teoremi, koja omogućava ažuriranje verovatnoća na osnovu novih informacija. U kontekstu kreditnog skoringa, Bejsove metode mogu biti korisne jer omogućavaju inkorporiranje a priori znanja i iterativno ažuriranje modela kako novi podaci postanu dostupni.[17]

Bejzova logistička regresija predstavlja jedan od pristupa koji kombinuje Bejzovu statistiku sa klasičnom logističkom regresijom. U ovom pristupu, počinjemo sa a priori raspodelom koja opisuje naše početno znanje o modelu. Nakon toga, kroz iterativni proces, ažuriramo ovu

raspodelu na osnovu novih podataka pomoću Bajesove teoreme. Ovo omogućava modelu da evoluira kako se sticanje novih informacija povećava.

Prednost Bejzove logističke regresije je ta što može efikasno rukovati sa malim uzorcima podataka i širokim spektrom promenljivih. Takođe, ovaj pristup može obuhvatiti nesigurnosti u parametrima modela, što može dovesti do realističnijih procena rizika.

Primena Bejzoovskih metoda u kreditnom skoringu omogućava bolju integraciju a priori informacija o klijentima, kao i stalno ažuriranje modela kako se nova saznanja stiču. Ovo je posebno korisno u situacijama gde su dostupni podaci ograničeni i gde želimo da modeliramo rizik na temelju svih dostupnih informacija.

Važno je napomenuti da primena Bejzovskih metoda zahteva duboko razumevanje Bejzove statistike i odgovarajuće tehničke veštine. Osim toga, kao i svaka metodologija, i Bejzove metode imaju svoje ograničenja, uključujući potrebu za izborom odgovarajućih prior raspodela i složenosti računanja.

U suštini, Bejzove metode pružaju sofisticiranu i prilagodljivu platformu za modeliranje rizika u kreditnom skoringu, omogućavajući bankama da donose bolje informisane odluke na osnovu sveobuhvatnih informacija i ažuriranih analiza.

Napomenimo da u ovom poglavlju nismo predstavljali matematičke osnove Bejzovog modeliranja, jer je to tema većeg dela ostatka rada.

Razna istraživanja su uporedila metode u kreditnom skoringu. Altman i saradnici [1] su analizirali neuronske mreže u poređenju sa linearnom diskriminantnom analizom (LDA) i utvrdili da je LDA dala bolje rezultate. Desai i saradnici [4] su, koristeći set podataka kreditne unije, pokazali da se neuronske mreže pokazuju bolje od LDA, ali ne znatno bolje od logističke regresije. U studiji Komorád-a [13], logistička regresija je upoređena sa neuronskim mrežama kao što su višeslojni perceptron i neuronske mreže sa radijalnom baznom funkcijom. Ovi modeli su trenirani na podacima iz francuske banke i pokazalo se da su neuronska mreža višeslojnog perceptrona i neuronska mreža sa radijalnom baznom funkcijom dale veoma slične rezultate, dok je logistička regresija bila najefikasnija.

Logistička regresija je često korišćena metoda za izradu skoring modela. Ona spada u širu klasu generalizovanih linearnih modela (GLM), što je čini fleksibilnom. Međutim, neuronske mreže, kao što su višeslojni perceptron i neuronske mreže sa radijalnom baznom funkcijom, predstavljaju složeniju alternativu logističkoj regresiji. Ove mreže mogu bolje modelirati nelinearne odnose u podacima, što je često prisutno u kreditnom skoringu.

Još jedan napredan pristup u kreditnom skoringu je Bejzov pristup. Ovaj pristup omogućava ažuriranje modela kako stižu nove informacije. Ovo je korisno jer se modeli u realnom svetu suočavaju sa promenama tokom vremena. Bejzov pristup omogućava modelima da se prilagode novim uslovima i poboljšaju performanse tokom vremena.

Bejzov pristup u kreiranju kreditnih skoring modela je jedna od centralnih tačaka ovog rada i u nastavku će biti više reči o njemu.

2.4 Ograničenja i izazovi kreditnog skoringa

Kreditni skoring modeli omogućavaju bankama koje ih primenjuju da pruže povoljnije uslove za kreditne proizvode dobrim klijentima (onima koji imaju značajnu imovinu i dobru kreditnu istoriju). Ovo se postiže uz niže troškove za banku u poređenju sa tradicionalnim metodama odobravanja kredita.[23] Postoje argumenti kako za, tako i protiv upotrebe kreditnih skoring modela. Sa jedne strane, skoring proces čini kreditnu analizu efikasnijom i ubrzava celokupan postupak. Takođe, eliminisanje subjektivnosti koja može postojati kod ručne procene kreditnih referenata doprinosi objektivnosti i doslednosti u odlučivanju. Međutim, s druge strane, skoring modeli mogu propustiti uzimanje u obzir određenih jedinstvenih karakteristika koje bi ručno ocenjivanje možda uočilo. Ovo se posebno odnosi na specifične situacije koje se teško uklapaju u uobičajene modele i obrasce. Uprkos tome, skoring modeli ostaju važan alat za bankarski sektor, omogućavajući brže donošenje odluka sa smanjenjem rizika i troškova.

Neke od prednosti korišćenja kreditnih skoring modela su sledeće:

- Smanjuje se operativni rizik – process je automatizovan i rizik greške kreditnog referenta je minimizovan.

- Kreditni skoring modeli su objektivni, konzistentni i efikasni, u smislu da je pristrasnost kreditnog analitičara otklonjena, mogućnost njegove greške pri manuelnom radu (u smisli ovog dela posla) otklonjena i sam referent ima više vremena za ostale aktivnosti.

- Kreditni skoring modeli su relativno jeftini. Potreban je manji broj kreditnih referenata, što utiče na manju cenu koštanja kredita za banku.

- Zadovoljstvo klijenata je veće usled efikasnijeg i bržeg procesa obrade kreditnog zahteva

- Kako kreditni skoring modeli određuju verovatnoću da li će klijent kasniti po otplati kredita ili ne, cenu kredita je moguće prilagoditi riziku klijenta. Na ovaj način banka povećava profit, a dobrim klijentima šalje pozitivan signal.

- Banka je u mogućnosti da odredi količinu kredita koju će plasirati u skladu sa kreditnom politikom banke. Definisanjem granične rizičnosti klijenta (izražena verovatnoćom neservisiranja obaveza po kreditu) banka je u mogućnosti da kontorliše tržišnu aktivnost.

- Kreditni skoring modeli su relativno jednostavni za shvatanje i interpretaciju. Metodologija koja se koristi za izradu modela je takođe lako shvatljiva. Jednostavnost skoring modela koji se najčešće koriste je delom nametnuta potrebom za objašnivošću modela menadžmentu banke.

Kreditini skoring se može računati pomoću generičkih modela, polu-prilagođenih modela ili prilagođenih modela. Kada se pravilo dizajniraju, modeli su obično pouzdaniji od subjektivnih metoda (ekspertsko mišljenje kreditnog analitičara). Međutim, razvoj i implementacija skoring modela, kao i njihova revizija, nose sa sobom inherentne izazove. Ovi modeli nikada neće biti savršeno tačni i korisni su samo ako ih korisnici u potpunosti razumeju. Dodatno, greške u konstrukciji modela mogu dovesti do netačnog ocenjivanja, što može rezultirati odobravanjem rizičnijih računa nego što je planirano i/ili neuspehom u pravilnom prepoznavanju i rešavanju povećanog kreditnog rizika unutar portfolija kredita. Greške u konstrukciji mogu varirati od osnovnih formula, preko pristrasnosti uzorka, do korišćenja neprikladnih prediktivnih varijabli.

Neki od nedostataka primene skoring modela, koji postavljaju svojevrsne izazove su:

- Tokom vremena, modeli najčešće degradiraju. Ukoliko se grupacija klijenata za koju se primenjuje skoring model promeni u odnosu na uzorak na kome je model razvijan, model će imati slabiju prediktivnu moć.

- Da bi se razvio model potrebno je imati veliki broj uzoraka. Takođe potrebno je odvojiti i uzorak za validaciju modela koji je nezavistan od razvojnog skupa.
  Razvojni skup treba da sadrži dovoljan broj kako ‘dobrih’ tako i ‘loših’ klijenata.
  Ove definicije zavise od same zavisne varijable modela, npr. za aplikativne modele to su najčešće klijenti koji su kasnili 90 dana u kontinuitetu i materijalno značajnom iznosu u period od godinu dana nakon odobrenje kredita. Najčešće je ovakvih klijenata daleko manje nego ‘dobrih’, što u uslovima ekstremno malog broja takvih (što se još zove i *low default portfolio*) predstavlja izazov za modeliranje, jer modeli zahtevaju dovoljnu količinu ovakvih klijenata da bi mogli da ‘uče’ iz podataka.

- Verovatnoće neizvršenja obaveza, koje su izračunate putem internih kreditnih skoring modela, ne mogu se koristiti za procenu rizika na tržištu. Ako želimo proceniti rizičnost klijenta na tržištu, postoje javno dostupne informacije o kreditnom rejtingu putem rejting agencija generički kreditni skoring modeli (o kojima je bilo reči. Verovatnoća neizvršenja obaveza je interni podatak koji ukazuje na verovatnoću da će klijent kasniti sa plaćanjem za određeni kredit za koji je podneo zahtev.

- Jedan od značajnih problema u kreditnom skoringu je problem inferencije odbijanja (*reject inferencing*). Mok [15] objašnjava da potpuni podaci postoje samo za prihvaćene aplikante Važno je napomenuti da rizik od nesolventnosti odbijenih aplikacija nije poznat, budući da se performansa kredita aplikacije ne može posmatrati kada je odbijena. Budući da su prihvaćeni aplikanti već prošli kroz postojeći skoring model, imamo pristrasne podatke. Bilo bi bolje izgraditi model gde su svi prihvaćeni ili gde se

prihvatanje vrši potpuno slučajno i gde se posmatra njihovo ponašanje [15]. Međutim, to nije izvodljivo za banke. Zato se koriste metode **inferencije odbijanja** kako bi se rešio ovaj problem pristrasnosti. Ovaj postupak zasniva se na određenim pretpostavkama o tome kako bi se odbijeni aplikanti ponašali da su bili prihvaćeni i nastoji umanjiti pristrasni uzorak samo prihvaćenih aplikacija. Ovaj proces se koristi kako bi se izbeglo trošenje značajnih resursa i potencijalno štetne posledice odobravanja zajmova potrošačima koji bi inače bili odbijeni samo u svrhu poboljšanja modela.

- Razvijeni kreditni skoring modeli primenjuju se samo na kategorije klijenata koje su bile uključene u procesu njihovog razvoja. Na primer, ako banka planira da pruža kreditne kartice penzionerima, ali koristi kreditni skoring model koji je stvoren na osnovu podataka koji ne uključuju penzionere, taj model neće precizno razlikovati dobre i loše klijenate. Ovo se dešava zbog nedostatka relevantnih podataka za određenu kategoriju klijenata u samom razvojnom uzorku modela. Kako bi model bio efikasan za određeni tip klijenata, potrebno je uključiti raznovrsne informacije i karakteristike tih klijenata u proces razvoja modela. Inače, model neće pružiti tačne i pouzdane rezultate, čime se smanjuje njegova praktična primenljivost i sposobnost za donošenje ispravnih odluka.

- U sličnom tonu, tokom perioda snažnog ekonomskog rasta ili ekonomskih fluktuacija, modeli se mogu pokazati nedovoljno sposobnim da predviđaju performanse zaduženih lica u posebno ako istorijski period posmatranja korišćen za modeliranje nije obuhvatao takve ekonomske uslove. Postoje različita ponašanja koja mogu uticati na efikasnost modela tokom recesijskih vremena. Jedno od njih je da potrošači mogu prioritetno izdvajati sredstva za otplatu obezbeđenih dugova (npr. stambenih kredita) pre nego za dugove po gotovinskim kreditima ili kreditnim karticama. U teškim vremenima, to može ostaviti banku koja drži dugove potrošača na kreditnim karticama npr. kao jednog od poslednjih kojima će biti plaćeno, ako uopšte bude plaćeno. Jedan od načina da se ovo predupredi je stres testitranje (stress testing).

  Stres testiranje u modelima kreditnog skoringa predstavlja proces ispitivanja stabilnosti modela pod ekstremnim ekonomskim scenarijima. Ova tehnika omogućava procenu ponašanja modela tokom kriznih uslova i identifikaciju slabosti. Stres testovi otkrivaju kako bi se model ponašao u recesiji, visokoj nezaposlenosti ili likvidnosti, itd. Ova praksa doprinosi boljem upravljanju kreditnim rizikom i odlučivanju u nestabilnim uslovima.

3. Logistička regresija

Logistička regresija (ranije pomenuta) je statistička metoda koja se koristi za modeliranje i predviđanje verovatnoća događaja koji imaju binarni ili kategorički izlaz, kao što je "da" ili "ne", "uspeh" ili "neuspeh". Ova metoda se koristi u mnogim disciplinama, uključujući finansije, medicinu, ekonomiju, marketing i društvene nauke.[8]

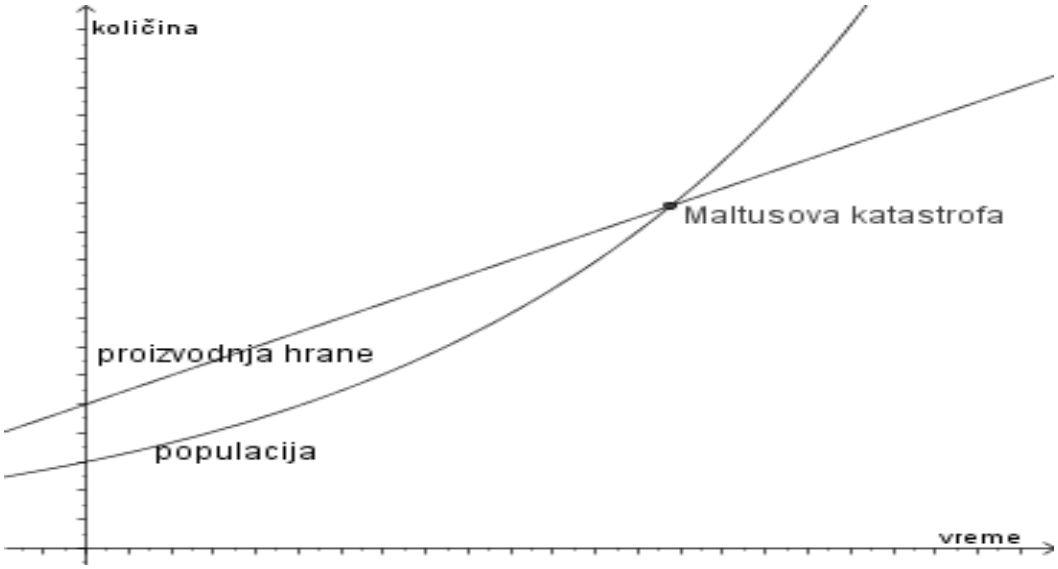
Osnova logističke regresije je *logistička funkcija*, poznata i kao sigmoidna kriva, koja se koristi za transformaciju linearne kombinacije nezavisnih promenljivih u raspon verovatnoća između 0 i 1. Parametri logističke regresije se procenjuju tako da se model prilagodi podacima, čime se omogućava predviđanje verovatnoća pripadnosti određenoj klasi ili kategoriji.

Važnost logističke regresije leži u njenoj sposobnosti da se nosi sa složenim međusobnim odnosima između promenljivih i da pruža razumljive rezultate. Model može biti interpretiran kroz koeficijente, koji ukazuju kako na usmerenost (zavisnih promenljivih u odnosu na nezavisne) tako i na i snagu veze između zavisnih i nezavisnih promenljivih. Ova interpretacija omogućava donosiocima odluka da razumeju faktore koji utiču na verovatnoću ishoda.

U suštini, logistička regresija je moćan alat za modeliranje i predviđanje verovatnoća događaja u okruženjima gde imamo binarne ili kategoričke izlaze. Njena primena pruža dublje razumevanje odnosa među promenljivima i pomaže u donošenju informisanih odluka.

3.1 Poreklo logističke funkcije

Poreklo logističke funkcije seže u 18.vek, kada je engleski ekonomista Tomas Robert Maltus u svom radu *''An essay on the principle of population as it affects the future improvement of society''* izneo krajnje zabrinjavajuće zapažanje. Naime, primetio je da se broj stanovnika na Zemlji konstantno povećava kao i potreba za osnovnim resursima kao što su hrana i voda. Maltus tvrdi da se i količina resursa povećava, ali aritmetičkom progresijom, dok povećanje broja stanovnika Zemlje prati geometrijsku progresiju. Kako geometrijska progresija za bilo koju konstantu veću od 1 može davati veće vrednosti od bilo kog aritmetičkog niza, posle dovoljnog broja članova, neminovno će doći do situacije kada će zavladati oskudica. Popularno se ovaj hipotetički događaj naziva i *demografska (Maltusova) katastrofa* (Slika 6). Ovim rezonovanjem Maltus dolazi se do zaključka da treba smanjiti priraštaj stanovništva i ne samo da treba prevenirati glad kugu i slične pošasti već ih treba i podsticati.



Slika 6 – Maltusova katastrofa

Izvor: Gorica Gvozdić, Primenjena logistička regresija

Osnovni Maltusov model

Pretpostavimo da u nekom trenutku t_0 na Zemlji živi $p(0)$ stanovnika. U sledećem vremenskom trenutku koji posmatramo veličina populacije je $p(1) = rp(0)$, gde je r koeficijent geometrijske progresije.

Dalje, neka je stopa priraštaja $\lambda = \gamma - \delta$, gde je γ konstanta koja predstavlja brzinu rađanja u jedinci vremena, a δ brzinu umiranja u jedinici vremena.

Ako sa $p(t)$ označimo broj stanovnika u trenutku t, onda je on posle nekog vremenskog intervala Δt jednak:

$$p(t + \Delta t) = p(t) + \lambda p(t)\Delta t$$

Znajući definiciju izvoda dobijamo sledeću diferencijalnu jednačinu:

$$p'(t) = \lambda p(t)$$

$$p(0) = p_0$$

Rešavajući ovu ODJ dobijamo:

$$\frac{dp(t)}{p(t)} = \lambda dt$$

$$ln\,|p(t)| = \lambda t + c$$

$$p(t) = e^{\lambda t}e^{c}$$

$$p(t) = Ce^{\lambda t}$$

Kada primenimo početne uslove dobijamo da je rešenje diferencijalne jednačine:

$$p(t) = p_0 e^{\lambda t}$$

Gde je p_0 broj stanovnika u početnom trenutku posmatranja. Iz rešenja se vidi da populacija raste eksponencijalno sa vremenom.

Modifikacija Maltusovog modela

...all models are approximations. Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind ... George E.P.Box

Imajuću u vidu navedeni citat, i činjenicu da su modeli samo naše nesavršeno shvatanje realnosti, postajemo svesni njihovih potencijalnih manjkavosti. S tim u vezi, i Maltusov populacioni model je imao svoje kritičare.

Pjer Fransoa Verhlust (1804-1849) je unapredio model, naglašavajući da nijedna sredina ne može podržavati neograničen broj jedinki. Prema njemu, rast populacije bi trebalo ograničiti do neke maksimalne fiksne vrednosti koja je karakteristična za posmatrani sistem. To je poznato kao maksimalni nosivi kapacitet sredine, označen sa K. Ograničeni resursi usporavaju rast populacije, te populacija teži ka tački zasićenja. Osim toga, stopa rađanja i umiranja nisu konstantne već se menjaju vremenom, i to je opisano preko sledećeg:

$$\gamma(t) = \gamma_0 - \gamma_1 p(t)$$

$$\delta(t) = \delta_0 + \delta_1 p(t)$$

$$\gamma_0 > \delta_0 > 0, \qquad \gamma_1, \delta_1 > 0$$

Opisane jednačine podržavaju smanjenje brzine rađanja odnosno povećavaju brzinu umiranja sa porastom populacije.

Inicijalni (ujedno i najveći) priraštaj označićemo sa a, gde je

$$a = \gamma_0 - \delta_0$$

Sada važi da je prirodni priraštaj

$$\lambda(t) = (\gamma_0 - \delta_0) - (\gamma_1 + \delta_1)p(t) = a - bp(t)$$

gde smo sa b označili $b = \gamma_1 + \delta_1$

Maltusova jednačina sada ima oblik

$$p'(t) = \lambda(t)p(t)$$

$$p'(t) = ap(t) - bp^2(t)$$

$$p'(t) = a\left(1 - \frac{b}{a}p(t)\right)p(t)$$

$$p'(t) = a\left(1 - \frac{1}{K}p(t)\right)p(t)$$

$$a > b > 0, \qquad p(0) = p_0$$

Populacija P u početku raste eksponencijalno sa stopom rasta *a*, ali se taj rast smanjuje kako se populacija približava maksimalnom (nosivom) kapacitetu sistema $\frac{a}{b} = K$. Matematički takvo ponašanje možemo modelirati *logističkom jednačinom*:

$$\frac{dp(t)}{dt} = ap(t)(1 - \frac{p(t)}{K})$$

$$p(0) = p_0$$

Odnosno važi da kada ja populacija P mala u odnosu na kapacitet K, populacija se ponaša prema Maltusovom populacionom modelu. Kada se populacija približi maksimalnom kapacitetu, tada izraz u zagradi teži nula što usporava rast populacije. Kada rešimo jednačinu:

$$\frac{dp(t)}{dt} = ap(t)(1 - \frac{p(t)}{K})$$

$$\int \frac{1}{p(1 - \frac{p}{K})}\,dp = at + c$$

$$\ln\left|\frac{K - p}{p}\right| = -at - c$$

$$\left|\frac{K - p}{p}\right| = e^{-at}e^{-c}$$

$$\frac{K}{p} - 1 = Ce^{-at}$$

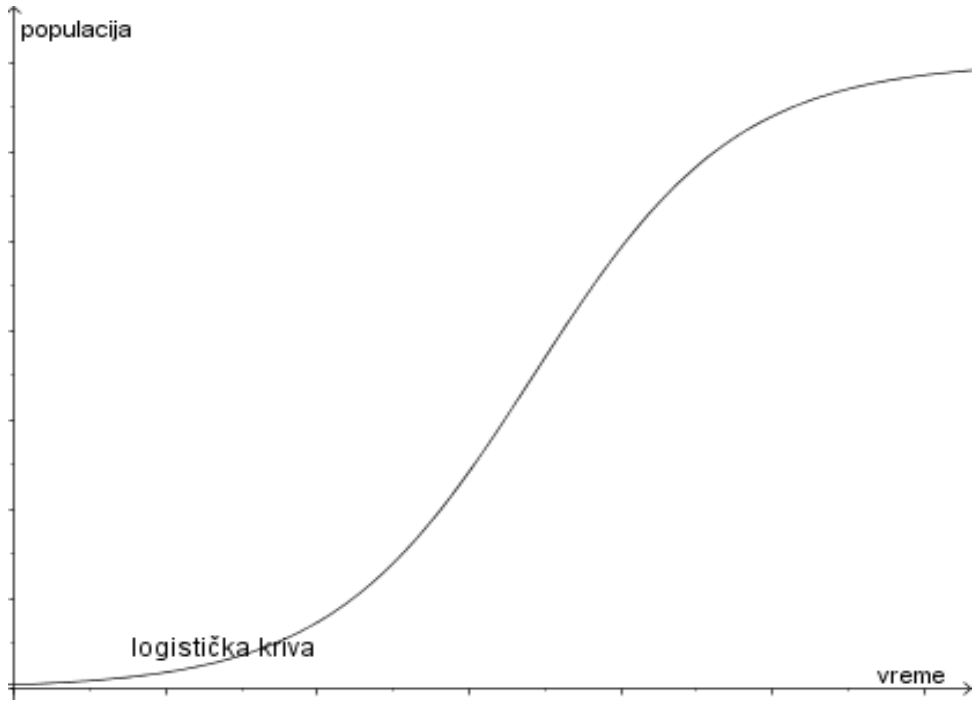
$$p(t) = \frac{K}{1 + Ce^{-at}}$$

Vidimo da kada $t \to \infty$ funkcija $p(t) \to K$. Opšte rešenje ove jednačine je *logistička funkcija*.

Konstantu C dobijamo iz početnog uslova:

$$p(0) = p_0 = \frac{K}{1+C} \rightarrow C = \frac{K - p_0}{p_0}$$

Kriva $p(t)$ ima oblik slova S i naziva se *logistička kriva* ili *sigmoida* (Slika 7). Ovaj model je kompleksniji i realističniji Maltusov model, ali kao i svaki model ima nedostatke.



Slika 7 – Sigmoida

Izvor: Gorica Gvozdić, Primenjena logistička regresija

3.2 Sigmoidna funkcija

Sigmoidna funkicija ili sigmoid je funkcija čiji grafik je kriva karakterističnog oblika koja podseća na latinično slovo 'S'. Najčešće se misli na standardnu sigmoid funkciju ili standardnu logističku funkciju čija je jednačina:

$$P(t) = \frac{1}{1 + e^{-t}}$$

Ova funkcija je rešenje sledeće diferencijalne jednačine:

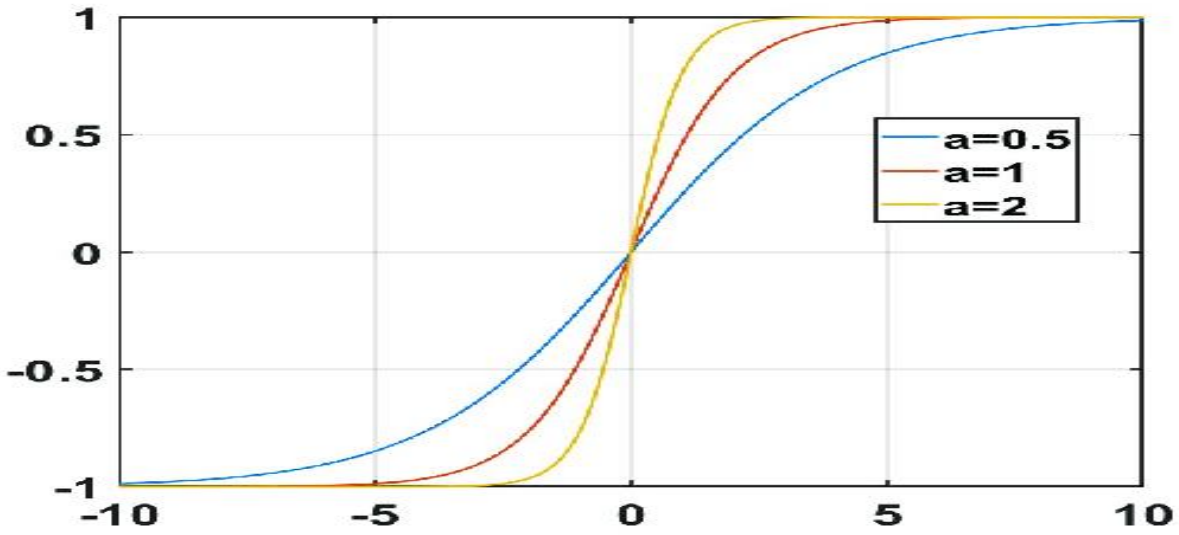
$$\frac{dP}{dt} = P(1 - P)$$

$$P(0) = \frac{1}{2}$$

Opštije, ova funkcija je najčešći prestavnik familije funkcija sledećeg oblika:

$$P(t) = \frac{1}{1 + e^{-at}}$$

gde je a parameter (Slika 8).



Slika 8 – Familija sigmoidnih funkcija

Izvor: https://www.slideshare.net/Devansh16/sigmoid-function-machine-learning-made-simple

Sigmoidne funkcije imaju domen na celom skupu realnih brojeva, a skup vrednosti je otvoreni interval (0,1). Zbog ove osobine skupa slika sigmoidna funkcija je pogodna za predstavljanje verovatnoće.

Ako posmatramo standardnu logističku funkciju, kao što rekosmo ona može predstavljati verovatnoću događaja u zavisnosti od promenljive t koja se definiše kao linearna kombinacija više faktora:

t = $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ gde su $\beta_i, i = 1 \ldots k$ regresioni koeficijenti. Oni određuju u kojoj meri i u kom smeru će promena odgovarajućeg im faktora uticati na verovatnoću pojave koju opisuju.

3.3 Uopšteni linearni modeli (GLM)

Uopšteni linearni modeli (*generalized linear models*) su široko korišćen koncept u statistici i mašinskom učenju koji generalizuje klasične linearno-regresione modele kako bi se obuhvatile različite distribucije zavisne promenljive i različiti tipovi veza između nezavisnih i zavisnih promenljivih. Ova fleksibilnost GLM-ova omogućava modeliranje širokog spektra problema koji prevazilaze ograničenja klasične linearno-regresione analize.

- **Osvrt na linearnu regresiju**

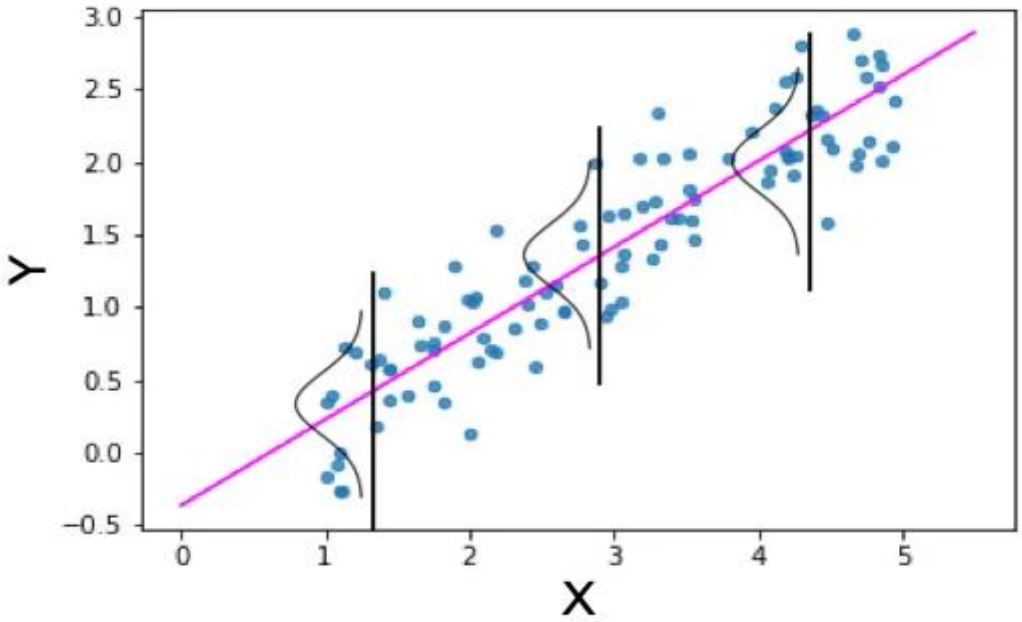
Linearna regresija se koristi za predikciju vrednosti neprekidne zavisne promenljive y kao linearne kombinacije nezavisnih promenljivih $x = (x_1, \ldots x_n)$.

U univarijantnom slučaju linearna regresija se može predstaviti na sledeći način:

$$\mu_i = b_0 + b_1 x_i$$

$$y_i \sim N(\mu_i, \varepsilon),$$

gde i predstavlja indeks instance. Model pretpostavlja normalnu raspodelu greške. To se može uočiti sa slike ispod:



Slika 9 – Linearna regresija

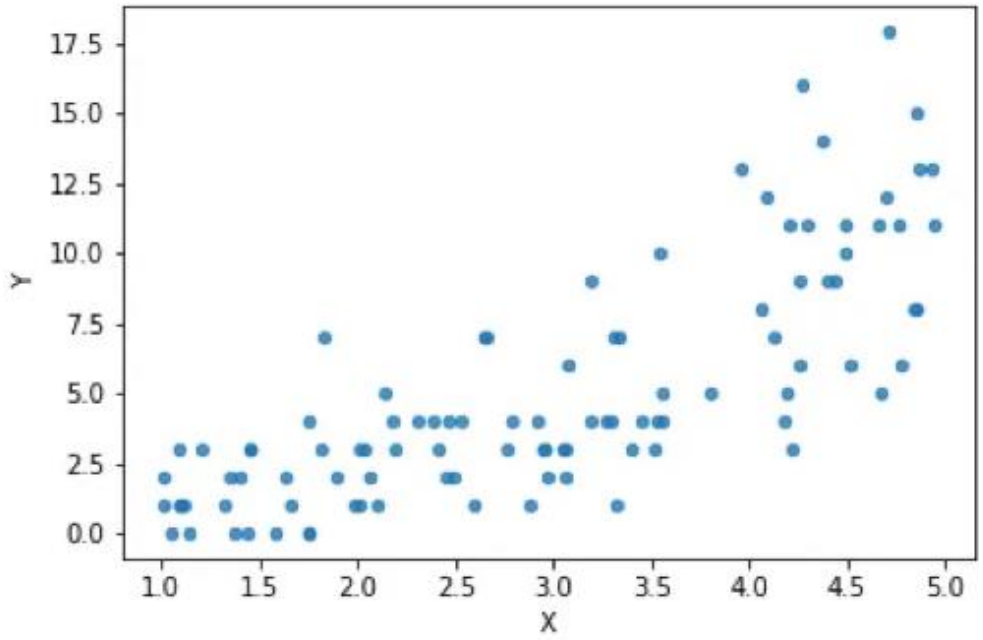
Izvor: https://towardsdatascience.com/generalized-linear-models-9cbf848bb8ab

Kao što primećujemo zavisna promenljiva ima fiksnu varijansu na svakom nivou nezavisne promenljive.

Kao što se moglo i pretpostaviti, linearna regresija premda ima veliku primenu, nije pogodna za regresije svih vrsta podataka.

Posmatrajmo sledeći primer.

Posmatra se broj neispravnih proizvoda (Y) u zavisnosti od broja senzora (X) kao nezavisnom varijablom. Grafikon rasipanja izgleda kao na slici ispod:



Slika 10 – Dijagram rasejanja (scatter plot)

Izvor: https://towardsdatascience.com/generalized-linear-models-9cbf848bb8ab

Postoji nekoliko problema ukoliko bismo želeli da zavisnost modeliramo linearnom regresijom:

- Veza između X i Y ne izgleda linearno, može se reći da liči na eksponencijalnu zavisnost
- Varijansa od Y ne deluje konstantno, već deluje da se povećava kako se X povećava

Kao što smo rekli, X predstavlja broj proizvoda, dakle u pitanju je pozitivan broj i X je diskretna promenljiva. U linearnoj regresiji zahteva se da zavisna promenljiva ima normalnu raspodelu.

Model koji bi bio više pogodan ovom primeru je model *Poasonove logističke regresije*.

Upravo je pomenuta regresija primer uopštenih linearnih modela (GLM).

GLM se sastoje iz 3 komponente:

1. **Komponente slučajnosti**

Ova komponenta vezuje se za raspodelu verovatnoće zavisne promenljive. Po pretpostavci GLM realizacije zavisne promenljive $Y_1, Y_2, \ldots Y_n$ su međusobno nezavisne i potiču iz pomenute raspodele. Realizacije od Y mogu biti dihotomne (binarne, sa samo dva moguća ishoda) tada

zavisna promenljiva Y ima binomnu raspodelu ili se realizacije mogu dobiti prebrojavanjem tada zavisna promenljiva ima Poasonovu raspodelu.

Komponenta sistematičnosti

Predstavlja lineranu kombinaciju nezavisnih promenljivih koje opisuju zavisnu promenljivu i parametara:

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Ovakva linerana kombinacija se naziva linearno predviđanje (linearni predictor), i promenljive x_i ne moraju biti linerano nezavisne.

2. Funkcija veze (Link funkcija)

Funkcija veze kao što samo ime kaže, povezuje komponentu slučajnosti i komponentu sistematičnosti. Ona predstavlja neku funkciju $g(\cdot)$ tako da važi:

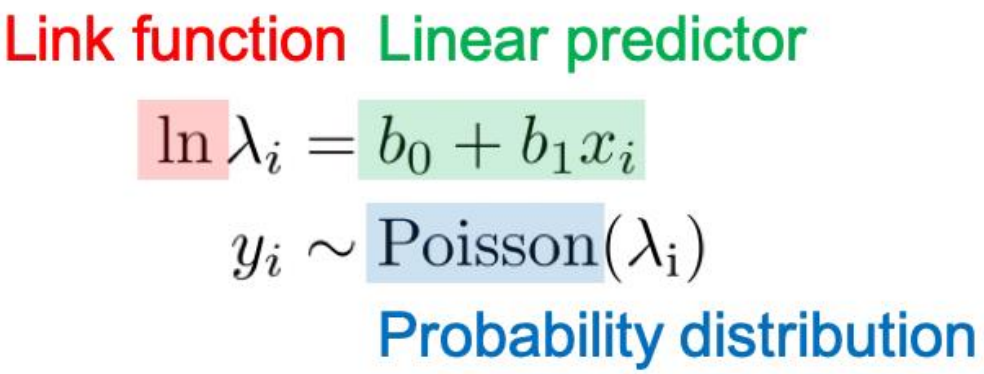
$$g(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

gde je $\mu = E(Y)$, za zavisnu promenljivu Y.

Funkcija $g(\cdot)$ je monotona i ne mora da bude linearno preslikavanje.

Vratimo se na prethodni primer. Adekvatan način za opisivanje zavisnosti zavisne i nezavisne promenljive može biti *Poasonova regresija.*

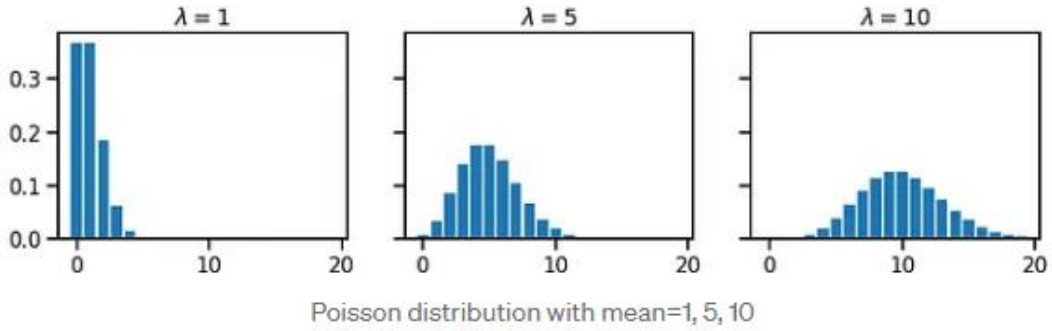
Poasonova regresija je specijalni slučaj GLM gde su pomenute 3 komponente prikazane na slici ispod.



Slika 11 – Poasonova regresija - komponente

Izvor: https://towardsdatascience.com/generalized-linear-models-9cbf848bb8ab

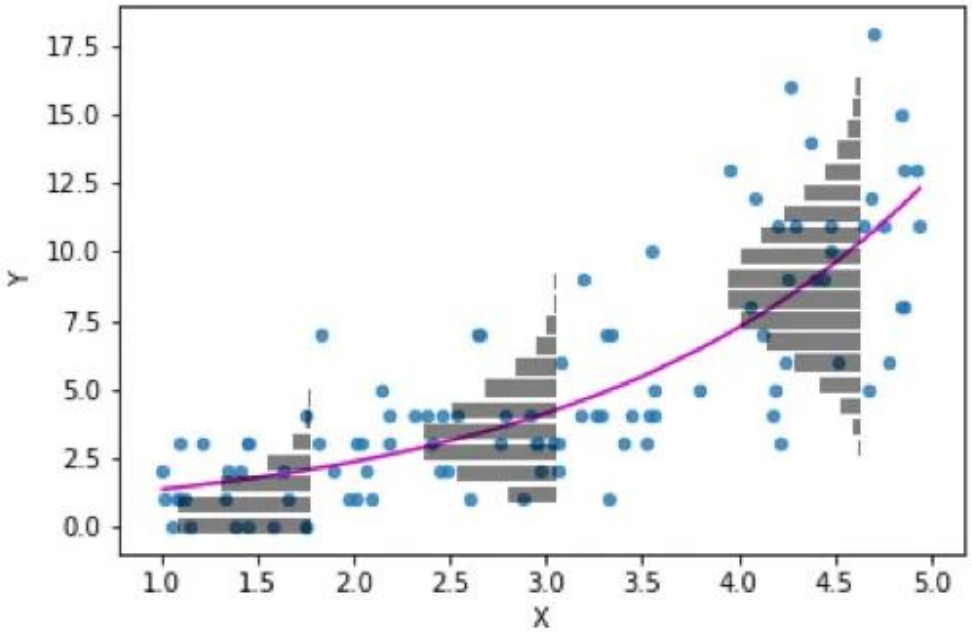
Poasonova raspodela se koristi za modeliranje podataka gde se radi o prebrojavanju. Ima samo jedan parametar koji predstavlja i srednju vrednost i standardnu devijaciju raspodele. To znači da što je veća srednja vrednost, to je veća i standardna devijacija. Ovo možemo videti i na slici ispod.



Slika 12 – Poasonova raspodela za različito λ

Izvor: https://towardsdatascience.com/generalized-linear-models-9cbf848bb8ab

Nakon primene Poasonove regresije rezultati izgledaju kao na sledećoj slici.



Slika 13 – Poasonova regresija - vizualizacija

Izvor: https://towardsdatascience.com/generalized-linear-models-9cbf848bb8ab

Kriva predstavljena ljubičastom bojom predstavlja predviđanje Poasonove regresije. Dodat je i stubičasti dijagram funkcije gustine verovatnoće Poasonove raspodele kako bi se jasno prikazala razlika u odnosu na linearnu regresiju.

Kriva predviđanja je eksponencijalna funkcija kao inverzna funkcija logaritamske link funkcije (Slika 14). Iz ovoga je takođe jasno da je parametar za Poasonovu regresiju izračunat pomoću linearnog prediktora garantovano pozitivan.

$$\ln \lambda_i = b_0 + b_1 x_i$$
$$\Leftrightarrow \lambda_i = \exp(b_0 + b_1 x_i)$$

Inverse of log link function

Slika 14 – Link funkija za Poasonovu regresiju

Izvor: https://towardsdatascience.com/generalized-linear-models-9cbf848bb8ab

Predstavimo još 2 primera uopštenih linearnih modela.

- **Obična linearna regresija**

Obična linearna regresija modelira kako srednja vrednost neprekidne zavisne promenljive zavisi od skupa objašnjavajućih promenljivih, gde indeksira svaku posmatranu vrednost:

$$\mu_i = \beta_0 + \beta_1 x_i$$

- *Komponenta slučajnosti*

Zavisna promenljiva ima normalnu raspodelu sa očekivanjem μ i konstantnom varijansom σ^2.

- *Kompomenta sistematičnosti*

Koristi se linearni prediktor oblika $\beta_0 + \beta_1 x_i$ gde x može biti kako neprekidna tako i diskretna. Može se proširiti na višestruku linearnu regresiju, gde se koristi više nezavisnih promenljivih. Takođe, same nezavisne varijable mogu biti transformisane ($ln\ (x)$ na primer) pod uslovom da se kombinuju sa koeficijentima parametara linearno.

-*Link funkcija* – Identično preslikavanje kao najjednostavniji oblik link funkcije:

$$g(E(Y)) = E(Y)$$

- **Binarna logistička regresija**

Binarna logistička regresija modeluje kako odnos šansi "uspeha" i "neuspeha" (odds ratio) za odgovarajuću binarnu promenljivu zavisi od skupa objašnjavajućih promenljivih:

$$\frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

- *Komponenta slučajnosti*

Zavisna promenljiva ima Bernulijevu raspodelu sa verovatnoćom uspeha $E(Y) = \pi$

- *Kompomenta sistematičnosti*

Isto kao i za linearnu regresiju.

- *Link funkcija*

$$g(E(Y)) = g(\pi) = ln\ \frac{\pi}{1-\pi}$$

4. Logistički regresioni model

Regresione metode su sastavni deo većine analize podataka čiji je cilj da se odredi veza između zavisnih i nezavisnih promenljivih. Cilj je naći model koji je najbolje prilagođen podacima, a ujedno je interpretabilan (objašnjiv) i ekonomičan, kako u smislu prikupljanja podataka tako i u smislu računske složenosti korišćenih algoritama.[8]

Logistička regresija se pored predviđanja zavisne promenljive u odnosu na vrednosti nezavisnih promenljivih koristi i za procenu rangiranja nezavisnih promenljivih po važnosti uticaja na zavisnu promenljivu, kao i za procenu efekata interakcije između zavisnih promenljivih.

Zavisna promenljiva u logističkom regresionom modelu je diskretna, najčešće binarna, dok se u ređem broju javlja korišćenje zavisne promenljive sa više od dve kategorije. Ona u praksi može predstavljati da li je određeni tretman delovao na pacijenta, da li je pristigao mejl spam, da li je određeni proizvod faličan ili ne, da li je klijent banke izmirio dugovanja prema istoj i slično. Nezavisne promenljive mogu biti kategorijalne, neprekidne kao i kombinacija pomenuta dva. Važno je pomenuti da u logističkoj regresiji ne postoji pretpostavka o raspodeli nezavisnih promenljivih. Zavisnu promenljivu označavaćemo sa Y, dok nezavisne označavamo sa x. Praksa je da se vrednosti zavisne promenljive kodiraju sa 0 i 1. Na primer posmatrani mejl je spam – 1, posmarani mejl nije spam – 0.

U regresionom modelu ključno je odrediti očekivanu vrednost zavisne promenljive za datu vrednost nezavisne promenljive, u oznaci $E(Y/x)$. Kako je zavisna promenljiva dihotomna (može imati dve vrednosti), za uslovno očekivanje važi: $0 \leq E(Y/x) \leq 1$. Promena u $E(Y/x)$ po jedinici promene za x postaje progresivno manja kako uslovna sredina postaje bliža 0 ili 1.

Kako je zavisna promenljiva dihotomna i uzima vrednosti 0 i 1, recimo da ista uzima vrednost 1 sa verovatnoćom π, a vrednost 0 sa verovatnoćom $1-\pi$, tj.

$$Y = \begin{pmatrix} 0 & 1 \\ 1-\pi & \pi \end{pmatrix}$$

Slučajna promenljiva Y/x će takođe uzimati vrednosti 0 i 1, sa verovatnoćama

$$1-\pi(x),\ \pi(x) \text{ redom, tj. } Y|x = \begin{pmatrix} 0 & 1 \\ 1-\pi & \pi \end{pmatrix}.$$

Kako nas interesuje očekivana vrednost od Y za dato x, izračunajmo je:

$$E(Y|x) = 0 \cdot \big(1-\pi(x)\big) + 1 \cdot \pi(x) = \pi(x)$$

Zbog ovoga, ubuduće ćemo koristiti oznaku $\pi(x)$ za prikazivanje uslovne sredine od Y za dato x kada se koristi logistička raspodela.

Poseban oblik regresionog modela koji koristimo je:

$$\pi(x) = \frac{e^{\beta_0 + \sum_k \beta_k x_k}}{1 + e^{\beta_0 + \sum_k \beta_k x_k}}$$

Kod logističke regresije, vrednost rezultujuće promenljive za dato x možemo izraziti kao $Y|x = \pi(x) + \varepsilon$, gde je ε greška koja ima binomnu raspodelu. Objasnimo i zašto.

Promenljiva ε može uzeti vrednost $-\pi(x)$ i $1 - \pi(x)$ i to vrednost $-\pi(x)$ uzima kada promenljiva $Y|x$ uzme vrednost 0, a vrednost $1 - \pi(x)$ uzima kada $Y|x$ uzme vrednost 1.

Kako slučajna promenljiva $Y|x$ uzima vrednost 0 sa verovatnoćom $1 - \pi(x)$, a vrednost 1 sa verovatnoćom $\pi(x)$, sledi da će i ε uzeti odgovarajuće vrednosti sa tim verovatnoćama, tj.

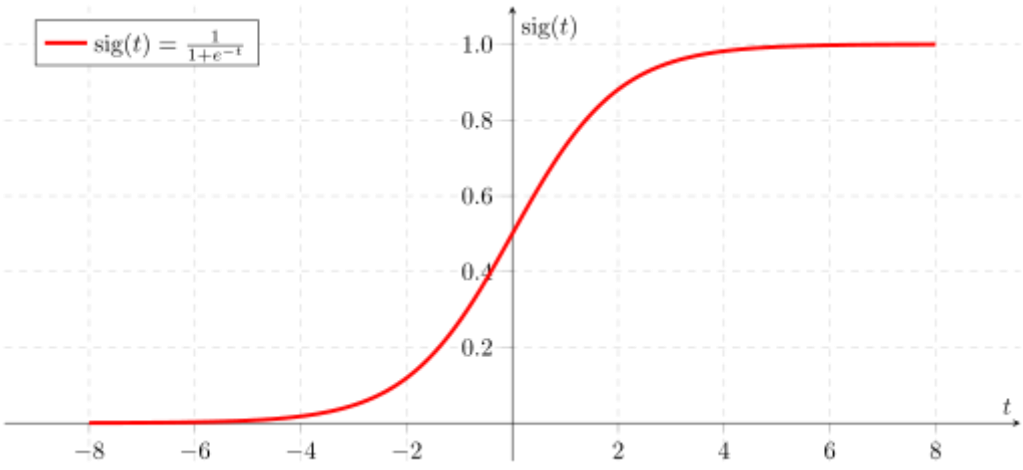
$$\varepsilon = \begin{pmatrix} -\pi(x) & 1 - \pi(x) \\ 1 - \pi(x) & \pi(x) \end{pmatrix}$$

Dakle, ε zaista ima binomnu raspodelu sa sredinom $E(\varepsilon) = 0$ i varijansom:

$$Var(\varepsilon) = \pi(x)\big(1 - \pi(x)\big)$$

4.1 Logit model

Iz do sada navedenog, jasno je da se u slučaju jedne nezavisne promenljive odnos između verovatnoće π i nezavisne promenljive x može predstaviti preko logističkog regresionog modela, koji se predstavlja preko S-krive, pomenute sigmoide (Slika 15).



Slika 15 – Sigmoidna kriva

Izvor: Gorica Gvozdić, Primenjena logistička regresija

Uočljivo je da verovatnoća postepeno raste sa porastom vrednosti nezavisne promenljive po stopi rasta i da je ograničena sa 0 odozdo i sa 1 odozgo. Stopa rasta verovatnoće raste do sredine sigmoidne krive gde nakon toga postepeno opada kako vrednost verovatnoće ide ka 1.

Verovatnoća π se može predstaviti formulom:

$$\pi = P(Y = 1|X = x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

Model može biti uopšten za slučaj kada postoji više nezavisnih promenljivih i onda izgleda ovako:

$$\pi = P\left(Y = 1|X_1 = x_1, X_2 = x_2, \dots, X_p = x_p\right) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}$$

Ova jednakost se naziva logistička regresiona funkcija. Ona nije linearna po parametrima $\beta_i, i = 0 \dots p$, ali se može linearizovati odgovarajućom logit transformacijom. Tada važi:

$$1 - \pi = P\left(Y = 0|X_1 = x_1, X_2 = x_2, \dots, X_p = x_p\right) = \frac{1}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}$$

Dalje imamo da je:

$$\frac{\pi}{1 - \pi} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}$$

Ako na prethodni izraz primenimo prirodni logaritam na obe strane dobijamo:

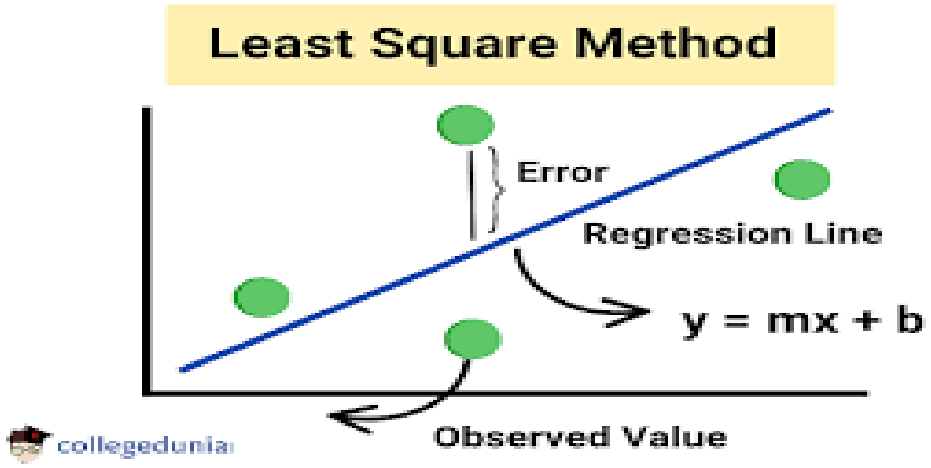
$$ln\ \frac{\pi}{1 - \pi} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Ova jednakost se naziva ***logit*** i ona je linearna po komponentama $\beta_i, i = 1 \dots p$. Primetimo još da vrednost od π pripada intervalu $[0,1]$, dok se vrednost logita kreće od $(-\infty, +\infty)$, pa je logit funkcija najprikladniji izbor za link funkciju.

4.2 Slaganje logističkog modela sa podacima

U linearnoj regresiji polazni metod za ocenjivanje regresionih parametara je metod najmanjih kvadrata. U tom metodu, biramo one vrednosti regresionih koeficijenata, koje minimiziraju sumu kvadrata odstupanja registrovane vrednosti za Y od predviđene vrednosti dobijene na osnovu modela (Slika 16). Pod uobičajenim pretpostavkama za linearnu regresiju, metod najmanjih kvadrata daje ocene sa mnoštvom poželjnih statističkih svojstava. Međutim, kada se

metod najmanjih kvadrata primeni na model sa dihotomnim ishodom, ocene više nemaju te iste osobine.



Slika 16 – Metod najmanjih kvadrata

Izvor: https://collegedunia.com/exams/least-square-method-mathematics-articleid-7402

Kada je u pitanju logistička regresija za ocenjivanje regresionih koeficijenata najčešće se koristi metod maksimalne verodostojnosti -**ML** (**maximum likelyhood**). Ovaj metod daje vrednosti za β_i, i=0… p, koje maksimiziraju verovatnoću dobijanja registrovanog skupa podataka. Utvrđuje se verodostojnost (verovatnoće) registrovanih podataka za različite kombinacije vrednosti regresionih koeficijenata, za razliku od metode najmanjih kvadrata. Ovaj metod zahteva numerički iterativni postupak izračunavanja.

Da bismo opisali ML metod, potrebno je da se upoznamo sa funkcijom verodostojnosti (**likelihood**). Ovo je funkcija regresionih koeficijenata u oznaci $L(\beta)$, gde je $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ i pomenute parameter smatra nepoznatima dok vrednosti nezavisnih promenljivih uzima kao date.

Ako zavisna promenljiva ima sledeću raspodelu $Y: \begin{pmatrix} 0 & 1 \\ 1-\pi & \pi \end{pmatrix}$, tada izraz:

$$\pi(x_i) = \frac{e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}{1+e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}$$

za proizvoljnu vrednost $\beta = (\beta_0, \beta_1, \dots, \beta_p)$, daje uslovnu verovatnoću $P\{Y = 1|x_i\} = \pi(x_i)$ i $P\{Y = 0|x_i\} = 1 - \pi(x_i)$, gde je $x_i = (1, x_{1i}, x_{2i} \dots x_{pi})$, $i = 1 \dots n$

Za one parove (x_i, y_i) gde je $y_i = 1$ doprinos funkciji verodostojnosti je $\pi(x_i)$, a za one parove (x_i, y_i) gde je $y_i = 1$ doprinos funkciji verodostojnosti je $1 - \pi(x_i)$.

Dakle, za par (x_i, y_i) doprinos funkciji verodostojnosti je dat sledećim izrazom:

$$\pi(x_i)^{y_i}\big(1 - \pi(x_i)\big)^{1-y_i}$$

S obzirom da je pretpostavka da su registrovane vrednosti nezavisne, funkcija verodostojnosti je dobijena kao proizvod gornjih izraza.

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{p} \pi(x_i)^{y_i}\big(1 - \pi(x_i)\big)^{1-y_i} \qquad (1)$$

Verodostojnost se može zapisati u ekvivalentnom obliku i kao:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{p} \left(\frac{\pi(x_i)}{1 - \pi(x_i)}\right)^{y_i} \big(1 - \pi(x_i)\big)$$

gde se izraz $\frac{\pi(x_i)}{1-\pi(x_i)}$ naziva šansa(*odds*) za $P\{Y = 1|x_i\}$ i jednak je

$$\frac{\pi(x_i)}{1 - \pi(x_i)} = e^{e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}}} = e^{x_i' \beta}$$

odnosno verodostojnost predstavlja funkciju registrovanih vrednosti zavisne i nezavisnih promenljivih i nepoznatih parametara.

Radi jednostavnosti obično se koristi logaritam ove funkcije, tj. logaritam verodostojnosti:

$$L(\boldsymbol{\beta}) = \ln l(\boldsymbol{\beta}) = \sum_{i=1}^{p} [y_i \ln \pi(x_i) + (1 - y_i) \ln \pi(x_i)]$$

odnosno:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{p} \left[y_i \boldsymbol{x_i'\beta} - \ln(1 + e^{\boldsymbol{x_i'\beta}})\right]$$

Ocene parametara tražimo tako da maksimiziraju funkciju verodostojnosti. Da bismo našli $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ koji maksimizira funkciju $L(\beta)$ diferenciraćemo $L(\beta)$ u odnosu na $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ i dobijene jednačine ćemo izjednačiti sa nulom, odnosno važi:

$$0 = \frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{p} \left(y_i - \frac{e^{\boldsymbol{x_i'\beta}}}{1 + e^{\boldsymbol{x_i'\beta}}} \right) \boldsymbol{x_i'} \qquad (2)$$

Ove jednačine su nelinearne po $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$, pa se rešavaju nekim od iterativnih numeričkih postupaka.

Jedan od najčešće korišćenih iterativnih postupaka za rešavanje jednačine (2) je Njutn-Rapsonov (**Newton–Raphson**) metod.[8] Radi lakšeg rada sistem (2) ćemo napisati u ekvivalentnom matričnom zapisu, odnosno važi:

$$\frac{\partial L(\beta)}{\partial \beta} = X'(y - p)$$

gde je $p = P\{Y = 1|x_i\} = \pi(x_i) = \pi_i = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}$ i $X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & x_{np} \end{bmatrix}$

Neka je $W = diag(p_i(1 - p_i))$, odnosno:

$$W = \begin{bmatrix} \pi_1(1-\pi_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \pi_n(1-\pi_n) \end{bmatrix}$$

odakle sledi da je:

$$\frac{\partial^2 L(\beta)}{\partial \beta^2} = -X'WX$$

Neka je $\beta^{(0)}$ vektor početnih aproksimacija za svako $\beta^{(k)}$, tada je prva iteracija Njutn-Rapsonovog postupka:

$$\beta^{(1)} = \beta^{(0)} + \left(-\frac{\partial^2 L\left(\beta^{(0)}\right)}{\partial {\beta^{(0)}}^2} \right)^{-1} \frac{\partial L\left(\beta^{(0)}\right)}{\partial \beta^{(0)}}$$

Odnosno:

$$\beta^{(1)} = \beta^{(0)} + \left(X'W^{(0)}X\right)^{-1} X'(y - p^{(0)})$$

Svaku $l + 1$ iteraciju dobijamo:

$$\beta^{(l+1)} = \beta^{(l)} + \left(X'W^{(l)}X\right)^{-1} X'(y - p^{(l)})$$

Vrednost koja se dobija kao rešenje ovih iteracija naziva se ocena maksimalne verodostojnosti i označava se sa $\hat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, \dots, \widehat{\beta}_p)$.

4.2.1 Testiranje značajnosti koeficijenata regresionog modela

Nakon ocene koeficijenata logističkog regresionog modela, koraci koji slede se uopšteno odnose na ocenjivanje značajnosti promenljivih u modelu. Ovo obično uključuje formulisanje i testiranje statističkih hipoteza za određivanje da li su nezavisne promenljive u modelu značajne odnosno u značajnoj meri povezane sa rezultujućom promenljivom. Pitanje koje ovde postavljamo je sledeće: Da li nam model koji sadrži promenljivu, govori više o rezultujućoj promenljivoj nego model koji ne sadrži tu promenljivu?[8]

Odgovor na ovo pitanje je dobijen upoređivanjem registrovane vrednosti rezultujuće promenljive sa predviđenom vrednosti pomoću svakog od dva modela; prvi sa, i drugi bez te promenljive. Ako su predviđene vrednosti na osnovu modela koji sadrži određenu promenljivu "bolje", ili tačnije u nekom smislu, nego vrednosti koje su predviđene na osnovu modela koji ne sadrža tu promenljivu, tada kažemo da je promenljiva u modelu značajna.

4.2.1.1 Test količnika verodostojnosti

Poređenje registrovane i predviđene vrednosti dobijene iz modela koji sadrži nezavisnu promenljivu i modela koji je ne sadrži, je bazirano na logaritmu funkcije verodostojnosti. Pri tome se smatra da je registrovana vrednost zavisne promenljive ona predviđena vrednost koja se dobija iz zasićenog modela. Zasićen model je onaj model koji sadrži toliko mnogo parametara koliko ima podataka. [8]

Za poređenje registrovanih sa predviđenim vrednostima na osnovu modela koristimo funkcije verodostojnosti:

$$D = -2\ln\frac{l_f}{l_z} \tag{3}$$

Ovde je l_f-verodostojnost fitovanog modela, dok je l_z-verodostojnost zasićenog (**saturated**) modela, dok se izraz $\frac{l_f}{l_z}$ naziva **količnik verodostojnosti**.

Koristili smo -2ln da bismo dobili veličinu čija nam je raspodela poznata, tako da ovu statistiku možemo koristiti za testiranje hipoteza.

Korišćenjem izraza (1) izraz (3) postaje:

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\widehat{\pi_i}}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\widehat{\pi_i}}{1-y_i}\right)\right] \qquad (4)$$

gde je $\widehat{\pi_i} = \widehat{\pi_i}(x_i)$

Statistika D, u jednakosti (4), se naziva odstupanje (deviance).

U cilju procenjivanja značajnosti nezavisne promenljive, upoređujemo vrednost D za model koji sadrži nezavisnu promenljivu i model koji je ne sadrži. Promena u D koja nastaje zbog uključivanja nezavisne promenljive u model je data sa:

$$G = D\,(model\ bez\ nezavisne\ promenljive) - D\,(model\ sa\ nezavisnom\ promenljivom)$$

Kako obe vrednosti D imaju isti imenilac (verodostojnost zasićenog modela), G se može se izraziti kao:

$$G = -2\ln\left(\frac{verodostojnost\ modela\ bez\ nezavisne\ promenljive}{verodostojnost\ modela\ sa\ nezavisnom\ promenljive}\right)$$

Kada je u pitanju univarijabilni slučaj lako se pokazuje da kada promenljiva nije u modelu maksimalna verodostojnost od β_0 je $\ln\frac{n_1}{n_0}$, gde je $n_1 = \sum y_i$, a $n_0 = n - n_1$, dok je predviđena vrednost konstantna i iznosi $\frac{n_1}{n}$. U tom slučaju vrednost G je:

$$G = -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\widehat{\pi(x_i)}^{y_i}\left(1-\widehat{\pi(x_i)}\right)^{1-y_i}}\right]$$

odnosno

$$G = 2\left\{\sum_{i=1}^{n}\left[y_i \ln\widehat{\pi(x_i)} + (1-y_i)\ln\widehat{\pi(x_i)}\right] - [n_1\ln(n_1) + n_0\ln(n_0) - n\ln(n)]\right\}$$

Pod hipotezom da je β_1 jednako nuli, statistika G ima hi-kvadrat raspodelu sa jednim stepenom slobode.

Kada je u pitanju multivarijabilna logistička regresija, test količnika verodostojnosti za ukupnu značajnost k koeficijenata za nezavisne promenljive u modelu je izveden na isti način

kao i u univarijabilnom slučaju. Jedina razlika je da su fitovane vrednosti za model, $\hat{\pi}$, bazirane na vektoru $\widehat{\boldsymbol{\beta}}$ koji sadrži $k+1$ parametar. Tada G ima hi-kvadrat raspodelu sa k stepeni slobode pod nultom hipotezom da je svih k koeficijenata nagiba za nezavisne promenljive u modelu jednako 0.

4.2.1.2 Wald-ov test

Još jedan od pristupa ispitivanju značajnosti koeficijenata jeste da koristimo test koji povezuje koeficijente sa njihovim standardnim greškama. Wald test predstavlja količnik ocene maksimalne verodostojnosti koeficijenta $\hat{\beta}$ sa njegovom standardnom greškom $S_{\widehat{\beta}}$ i statistički ima približno standardnu normalnu raspodelu $N(0,1)$ pod hipotezom da je $\beta = 0$. Kvadrat ove Z statistike za univarijabilni slučaj ima približno χ^2 raspodelu sa jednim stepenom slobode. Odnosno Wald statistika za univarijantni slučaj je:

$$Z = \frac{\hat{\beta}}{S_{\widehat{\beta}}} : N(0,1)$$

$$Z^2 : \chi_1^2$$

Test statistika količnika verodostojnosti i Wald statistika daju približno iste vrednosti kad su u pitanju veliki uzorci, pa ako je neka studija dovoljno obimna nije bitno koju statistiku koristimo, međutim ako su uzorci malog obima statisitke mogu značajno da se razlikuju i pokazano je da je test statistika količnika verodostojnosti u ovakvim situacijama tačnija.

Waldova test statistika za multivarijabilni slučaj je:

$$W = \widehat{\boldsymbol{\beta}}' \left[\widehat{Var}(\widehat{\boldsymbol{\beta}})\right]^{-1} \widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}' (\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}) \widehat{\boldsymbol{\beta}}$$

i ima χ^2 raspodelu sa $k+1$ stepenom slobode pod početnom hipotezom da je svaki od $k+1$ koeficijenata jednak nuli. Statistiku za samo k koeficijenata nagiba dobijamo kad eliminišemo $\widehat{\beta_0}$ iz vektora $\widehat{\boldsymbol{\beta}}$ kao i odgovarajuće redove, odnosno kolone iz $\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}$.

Wald-ov test često ima nedostatak da se ne odbacuje nulta hipoteza iako su koeficijenti značajni tako da je preporučljivije koristiti test količnika verodostojnosti.

4.2.2 Intervali poverenja za parametre logističkog regresionog modela

Nakon testiranja značajnosi koeficijenata interpretiraćemo testiranje intervala poverenja za parametre koji nas interesuju. Prvo razmatramo intervale poverenja za univarijabilni slučaj.

Baza za konstrukciju intervala ocene je ista statistička teorija koju smo koristili za formulisanje testa za značajnost modela. Intervali poverenja ocene za nagib i odsečak su bazirani na njihovim odgovarajućim Wald testovima. Krajnje tačke za 100(1 – α)% interval poverenja za ocenjeni koeficijent nagiba su:

$$\widehat{\beta_1} \mp z_{1-\alpha/2} S_{\widehat{\beta_1}} \tag{5}$$

a za odsečak (**intercept**):

$$\widehat{\beta_0} \mp z_{1-\alpha/2} S_{\widehat{\beta_0}} \tag{6}$$

gde je sa $S_{\widehat{\beta_0}}$ označena ocena standardne greške (zasnovane na modelu) za odgovarajuću ocenu parametra, a $z_{1-\alpha/2}$ je gornja tačka standardne normalne raspodele.

Logit je linearan deo logističkog regresionog modela, i kao takav podseća na fitovanu pravu u linearnoj regresiji. Ocena za logit je

$$\widehat{g(x)} = \widehat{\beta_0} + \widehat{\beta_1} x$$

Ocena varijanse za ocenjeni logit zahteva korišćenje varijansu od sume varijansi. U tom slučaju je :

$$\widehat{Var}\left(\widehat{g(x)}\right) = \widehat{Var}\left(\widehat{\beta_0}\right) + x^2 \widehat{Var}\left(\widehat{\beta_1}\right) + 2x\widehat{Cov}\left(\widehat{\beta_0}, \widehat{\beta_1}\right) \tag{7}$$

Krajnje tačke za 100(1 – α)% interval poverenja za logit (na osnovu Wald statistike) su:

$$\widehat{g(x)} \mp z_{1-\alpha/2} S_{\widehat{g(x)}}$$

gde je $S_{\widehat{g(x)}}$ pozitivan kvadratni koren varijanse ocene u (7).

Ocena za logit i njegov interval poverenja su osnova za ocenu fitovanih vrednosti, u ovom slučaju logističke verovatnoće i njenog intervala poverenja. Konkretno, korišćenjem dobija se ocena za logit:

$$\hat{\pi}(x) = \frac{e^{\widehat{g}(x)}}{1 + e^{\widehat{g}(x)}}$$

Krajnje tačke 95% intervala poverenja za verovatnoću su dobijene iz odgovarajućeh krajnjih tačaka intervala poverenja za logit. Tako da su krajnje tačke za $100(1-\alpha)\%$ interval poverenja (baziran na Wald-ovom testu) za fitovanu vrednost:

$$\frac{e^{\widehat{g(x)} \mp z_{1-\alpha/2} \hat{S}_{\widehat{g(x)}}}}{1 + e^{\widehat{g(x)} \mp z_{1-\alpha/2} \hat{S}_{\widehat{g(x)}}}}$$

Fitovana vrednost izračunata u (8) je analogna odgovarajućoj tački na pravoj dobijenoj linearnom regresijom, U linearnoj regresiji svaka tačka na fitovanoj pravoj predstavlja ocenu proporcije zavisne promenljive u populaciji sa kovarijatom x.

Kad je u pitanju multivarijabilni slučaj, odnosno višestruka regresija, računanje intervala poverenja za koeficijente se izvodi na isti način kao i za univarijabilni slučaj iznad. Kod računanja intervala poverenja za logit koristimo istu osnovnu ideju kao i kod univarijabilnog slučaja sa tom razlikom da sada imamo više izraza uključenih u sumiranje. Opšti izraz za ocenu logita koji sadrži *p* kovarijati je:

$$\widehat{g(x)} = \widehat{\beta_0} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \widehat{\beta_p} x_p$$

odnosno:

$$\hat{g}(\boldsymbol{x}) = \boldsymbol{x}' \widehat{\boldsymbol{\beta}}$$

gde je $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$, a $\widehat{\boldsymbol{x}} = (x_0, x_1, \ldots, x_p)$ je konstantni vektor gde je $x_0 = 1$ (vrednost pridružena intercept-u)

Tada važi da je:

$$\widehat{Var}\left(\widehat{g(x)}\right) = \sum_{j=0}^{p} x_j^2 \widehat{Var}\left(\widehat{\beta_j}\right) + \sum_{j=0}^{p} \sum_{k=j+1}^{p} 2x_j x_k \widehat{Cov}\left(\widehat{\beta_0}, \widehat{\beta_1}\right)$$

U matričnom obliku dobijamo:

$$\widehat{Var}\left(\widehat{g(x)}\right) = \boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{x}$$

Dalje izračunavanje se vrši analogno univarijabilnom slučaju.

4.2.3 Interpretacija ocenjenog logističkog modela

Osnovna pretpostavka je da je logistički regresioni model prilagođen podacima, odnosno da je fitovan i da su promenljive u modelu značajne, tj. da su odgovarajući regresioni koeficijenti različiti od nule.

Interpretacija fitovanog modela predstavlja izvođenje zaključaka na osnovu ocenjenih koeficijenata u modelu. Postavlja se pitanje šta nam ocenjeni koeficijenti govore o pitanjima zbog kojih je i započeto istraživanje? Prilikom interpretacije modela posmatra se više aspekata problema a to su: određivanje funkcionalne veze između zavisne i nezavisne promenljive kao i stepen jačine te veze, kao i definisanje odgovarajuće jedinice promene za nezavisnu promenljivu.

Kao što smo videli, vezu između zavisne i nezavisne promenljive u logističkom regresionom modelu daje logit funkcija, tj.

$$g(x) = ln\,\frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

U logističkom regresionom modelu sa jednom nezavisnom promenljivom koeficijent nagiba β_1 predstavlja promenu u logitu po jedinici promene nezavisne promenljive, tj. :

$$\beta_1 = g(x+1) - g(x)$$

Slučaj kada je nezavisna promenljiva u logističkom regresionom modelu dihotomna predstavlja osnovu za druge slučajeve i podrazumeva da nezavisan pronmenljiva može uzeti dve vrednosti. Ovaj slučaj je u domenu našeg interesovanja. U našem slučaju neka je nezavisna promenljiva kodirana sa 0 i 1.

Kako koeficijent β_1 predstavlja stopu promene zavisne promenljive po jedinici promene nezavisne promenljive, važi da je:

$$g(1) - g(0) = (\beta_0 + \beta_1) - \beta_0 = \beta_1$$

Da bismo mogli interpretirati dobijeni rezultat uvešćemo pojam *odnos šansi* (**odds ratio**), koji daje meru povezanosti nezavisne promenljive sa ishodom.

Šansa je odnos verovatnoća da se događaj desi prema verovatnoći da se događaj ne desi.

Šansa da je zavisna promenljiva uzela vrednost 1, kada nezavisna promenljiva uzme vrednost 1 je:

$$\frac{P(Y = 1|X = 1)}{P(Y = 0|X = 1)} = \frac{\pi(1)}{1 - \pi(1)}$$

Kada nezavisna promenljiva uzme vrednost 0, šansa da je zavisna promenljiva uzela vrednost 1 je:

$$\frac{P(Y = 1|X = 0)}{P(Y = 0|X = 0)} = \frac{\pi(0)}{1 - \pi(0)}$$

Odnos šansi (unakrsni odnos šansi), u oznaci OR, je definisan kao odnos ove dve šanse, tj.

$$OR = \frac{\frac{\pi(1)}{1 - \pi(1)}}{\frac{\pi(0)}{1 - \pi(0)}} \tag{8}$$

Moguće vrednosti verovatnoća se mogu predstaviti tablicom $2x2$ na sledeći način:

Rezultujuća promenljiva (*Y*)	**Nezavisna promenljiva (*X*)**	
	$x = 1$	$x = 0$
$y = 1$	$\pi(1) = \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}}$	$\pi(0) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$
$y = 0$	$1 - \pi(1) = \frac{1}{1 + e^{\beta_0 + \beta_1}}$	$1 - \pi(0) = \frac{1}{1 + e^{\beta_0}}$
Ukupno	1	1

Tabela 1 – Vrednosti nezavisne i zavisne promenljive

Ova tabela opravdava to što se odnos šansi OR još naziva i unakrsni odnos šansi, jer vidimo da se OR dobija kao odnos unakrsnog proizvoda elemenata na glavnoj dijagonali date tabele i elemenata na sporednoj dijagonali.

Zamenom izraza iz tabele u (8) dobijamo:

$$OR = \frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}}{\frac{1}{1+e^{\beta_0+\beta_1}}} \Big/ \frac{\frac{e^{\beta_0}}{1+e^{\beta_0}}}{\frac{1}{1+e^{\beta_0}}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

Ovim dobijamo uvid u jednostavnu veza između koeficijenta i odnosa šansi.

4.3 Metodologija građenja modela logističke regresije

Prethodna poglavlja bavila su se nastankom i motivacijom za nastanak logističke regresije, kao i statističkim osnovama za njenu teoriju. Iako se u teorijske svrhe navode primeri sa manjim brojem nezavisnih promenljiih, u realnim primenama se koristi i više destina pa čak i stotina nezavisnih pormenljivih koje se inicijalno razmatraju.

Kriterijumi za uključivanje promenljivih u model mogu varirati od jednog problema do drugog i od jedne naučne discipline (ili sfere primene) do druge. Građenje statističkog modela uključuje težnju ka modelu sa što manjim brojem promenljivih koji ipak objašnjavaju podatke. Više je razloga za težnju za manjim brojem promenljivih. Možemo reći da je jedan razlog to što će model sa manjim brojem promenljivih biti numerički stabilniji, promenljive će najverovatnije biti manje korelisane i slično. Ukoliko je više promenljivih uključeno u model, ocene standardne greške postaju veće, i model postaje više zavisan od registrovanih podataka. Drugi aspekt težnje je sa biznis strane. Manji broj promenljivih zahteva manje operativnog rada za ekstrakciju podataka, kreiranje samih promenljivih i monitoring samog modela.

Postoji nekoliko koraka koje možemo pratiti kao pomoć pri izboru promenljivih za logistički regresioni model. Postupak za izbor modela je prilično sličan onom koji se koristi u linearnoj regresiji.

Koraci koji se sprovode pri modeliranju se mogu podvesti pod tri skupa: **sređivanje podataka, univarijantnu** i **multivarijantnu** analizu.

-Sređivanje podataka-

Pre svega smatramo da imamo podatke, najčešće u tabelarnom obliku koji sadrže sve promenljive koje dolaze u obzir za kreiranje modela logističke regresije. Pored samih promenljivih koje se koriste za modeliranje tu su i pomoćne promenljive koje koje mogu služiti da identifikuju određenu instancu (red u tabeli) korišćenog skupa (na primer JMBG osobe, ime prezime), vremenske promenljive koje mogu predstavljati datum ili vreme registrovanja posmatranog događaja (datum aplikacije za kreditni proizvod, godina izrade nekog proizvoda i slično).

Podatke koje inicijalno posedujemo nisu savršeni i skoro uvek sadrže određene manjkavosti, nelogične vrednosti (jedan od primera su outlieri), nedostajuće vrednosti i slično.

Koraci koji se u ovoj fazi su sledeći:

- *Vizuelizacija podataka*

Savremeni programski jezici i statistički alati pružaju pregršt načina da se podaci predstave vizuelno (histogrami, pie chart-ovi, bar plotov-i, box plotov-i i slično). Premda se odluke o tretmanu podataka zasnivaju na statističkim vrednostima i pokazateljima, vizuelizaicija je koristan metod za uočavanje pravilnosti u podacima trendu kretanja promenjive kroz vreme i slično.

- *Tretman outlier-a*

Outlier je podatak koji se primetno razlikuje od ostalih. Oni najčešće predstavljaju greške u merenju ili rezultat loše prikupljanje podataka. Postoji više tehnika za detekciju outlier-a. Najčešće korišćena posmatra 25-procentni (q25) i 75-procentni (q75) kvartil i definiše inter-kvartilni razmak (IQR- interquartile range) kao razliku pomenuta dva. Po ovom načinu detekcije, one vrednsosti promenljive koje su manje od q25- 1.5 * IQR i one koje su veće od q75 + 1.5 * IQR se smatraju outlier-ima. Takve vrednosti se najčešće zanemariju ili se smatraju nedostajućim vrednostima. Napomenimo da treba imati u vidu prirodu podataka i da na osnovu toga treba oceniti šta su stvarni outlier-i, bez obzira šta je sistem detekcije prikazao.

- *Tretman nedostajućih vrednosti*

Nedostajuće vrednosti u podacima predstavljaju izazov koji retko zaobilazi ijedan realan skup podataka. Mnogi su razlozi za postojanje istih: nedostupnost podataka, faktor ljudske greške pri unosu podatka i slično. Veliki broj statističkih postupaka ne trpi postojanje nedostajućih vrednosti u podacima, ili u najmanju ruku zahteva jasno grupisanje takvih vrednosti pod jedinstvenom oznakom u okviru jedne promenljive.

Kada je u pitanju tretman nedostajućih vrednsosti najčešće se koristi nešto od sledećeg:

- Uklanjanje redova sa nedostajućim vrednostima: U ovom slučaju, redovi u kojima nedostaju vrednosti se jednostavno izbacuju iz analize. Pri korišćenju ovog postupka morate biti pažljivi da ne izgubite previše podataka i da ne uvedete pristrasnost u analizu.

- Zamena nedostajućih vrednosti: Ovde se nedostajuće vrednosti zamenjuju odgovarajućim vrednostima, kao što su srednje vrednosti, mediane ili modalne vrednosti za numeričke varijable, ili najčešće korišćene kategorije za kategoričke varijable.

-Univarijantna analiza-

Ovo je analiza koja se fokusira samo na jednu nezavisnu promenljivu u odnosu na zavisnu promenljivu, bez uzimanja u obzir drugih nezavisnih promenljivih. U univarijantnoj analizi logističke regresije, istraživač posmatra kako se promena jedne nezavisne promenljive (npr. starost) odnosi na verovatnoću ili log-odds da će se dogoditi određeni događaj (npr. kupovina proizvoda). Ovo je osnovna analiza koja pomaže razumevanju pojedinačnih uticaj svake nezavisne na zavisnu promenljvu.

-Multivarijantna analiza-

Ovde istraživač razmatra više od jedne nezavisne promenljive istovremeno u modelu logističke regresije. Ovo je važno jer omogućava analizu uticaja više faktora zajedno na zavisnu promenljivu, uzimanje u obzir međusobnih interakcija između tih faktora. U multivarijantnoj analizi logističke regresije, istraživač pokušava identifikovati kako različite nezavisne promenljive zajedno utiču na verovatnoću ili log-odds događaja od interesa. Ova analiza omogućava bolje razumevanje kompleksnih veza između faktora i ciljne promenljive.

4.4 Metode za logističku regresiju

Nakon primene univarijantne analize dobija se skup promenljivih koje su kandidati za fitovanje modelom logističke regresije

Bilo koji postupak za izbor ili eliminisanje promenljivih iz modela je bazirana na statističkom algoritmu koji proverava značajnost promenljivih, te ih uključuje ili isključuje iz modela na osnovu utvrđenog pravila odlučivanja. Značajnost promenljive je definisana pomoću statističke značajnosti njenog koeficijenta. Statistika koja je korišćena zavisi od pretpostavki modela. U linearnoj regresiji korak po korak je korišćen F test, jer je pretpostavka da greške imaju normalnu raspodelu. U logističkoj regresiji pretpostavka je da greške imaju binomnu raspodelu, a značajnost je ocenjena putem hi- kvadrat testa količnika verodostojnosti. Dakle, u bilo kom koraku procedure, najvažnija promenljiva, u statističkim terminima je ona koja prouzrokuje najveću promenu u logaritmu verodostojnosti za model koji sadrži promenljivu u odnosu na

onaj koji ne sadrži promenljivu (to jest onaj koji bi trebalo rezultirati najvećom statistikom količnika verodostojnosti, *G*).

Navešćemo nekoliko algoritama za izbor finalnih promenljivih u logističkom postupku.

- **Stepwise** – 'Korak po korak' metod izbora promenljivih gradi logistički model u koracima, počevši od početnog modela bez ulaznih promenljivih u jednačini osim odsečka (intercept). U svakom koraku se evaluiraju ulazne promenljive koje još nisu dodate modelu, a ako najbolja od tih ulaznih varijabli značajno doprinosi prediktivnoj moći modela, ona se dodaje. Pored toga, ulazne varijable koje se trenutno nalaze u modelu se ponovo procenjuju kako bi se utvrdilo da li se bilo koja od njih može ukloniti bez značajnog odstupanja od modela. Ako je tako, uklanjaju se i proces se ponavlja sve dok se više ne mogu dodati/ukloniti varijable da bi se poboljšao model ili smanjio broj promenljivih u modelu. Na kraju se generiše konačni model.
- **Forwards** – ili metod unared metod počinje bez ulaznih promenljivih u početnom modelu, (samo odsečak), a promenljive se mogu samo dodati u model. U svakom koraku, ulazne promenljive koje još nisu u modelu se testiraju u smisli koliko bi one doprinele modelu i najbolja od tih promenljivih se dodaje modelu. Kada se više ne može dodati promenljiva ili najbolja kandidat promenljiva ne proizvodi dovoljno veliko poboljšanje u modelu, generiše se konačni model.
- **Backwards** – ili metod unazad počinje sa svim ulaznim promenljivama u početnom modelu, a promenljive se mogu samo ukloniti iz modela. Ulazne promenljivekoje malo doprinose modelu uklanjaju se jedna po jedna sve dok nijedna promenljiva više ne može biti uklonjena bez značajnog smanjenja kvaliteta modela, čime se dobija konačni model.
- **Backwards Stepwise** – Postepeni odabir unazad uključuje početak pristupa unazad, a zatim potencijalno dodavanje promenljivih ako se kasnije ispostavi da su značajne. Proces se sastoji od naizmeničnih izbora najmanje značajne promenljive za izbacivanje zatim ponovnog razmatranja svih ispuštenih varijabli (osim poslednje izbačene) za ponovno uvođenje u model. To znači da se moraju izabrati dva odvojena nivoa značajnosti za brisanje iz modela i za dodavanje u model. Drugi nivo značajnosti mora biti stroži od prvog.

Postoje i drugi algoritmi ali se mahom zasnivaju na nabrojanim. Predstavićemo algoritam koji će biti korišćen u praktičnom delu rada, a predstavlja izvesnu modifikaciju **backwards** algoritma.

Pretpostavimo da na raspolaganju imamo ukupno m nezavisnih promenljivih.

Korak (1):

Počinjemo sa fitovanjem modela koji ima svih m nezavisnih pormenljivih i odsečak.Svaka promenljiva koja ima p vrednost veću od 0.05 (najčešće korišćeni prag) se izbacuje iz modela.

Pokreće se model logističke regresije bez izbačenih pormenljivih i evidentiraju nove p vrednosti za promenljive preostale u modelu. Ukoliko ne postoje promenljive čija je p vrednost veća od 0.05 prelazi se na korak(2). U suprotnom ponavlja se korak (1)

Korak (2):

Izračunava se AIC (Akaike informacioni kriterijum modela) trenutnog modela. AIC nagrađuje gof (goodness of fit) modlela (koji procenjuje funkcijom verodostojnosti), ali takođe uključuje penal koja je rastuća funkcija broja procenjenih parametara. Kazna obeshrabruje overfiting, što je poželjno jer povećanje broja parametara u modelu skoro uvek poboljšava gof.

Od preostalih promenljivih bira se ona sa najvećom p vrednosti i ona je kandidat za izbacivanje. Ukoliko je AIC modela bez posmatrane promenljive veći od AIC a dosadašnjeg modela, promenljiva se vraća u model i dobija se finalni model. U suprotnom promenljiva se izbacuje. U slučaju izbacivanja promenljive ponavlja se korak 1.

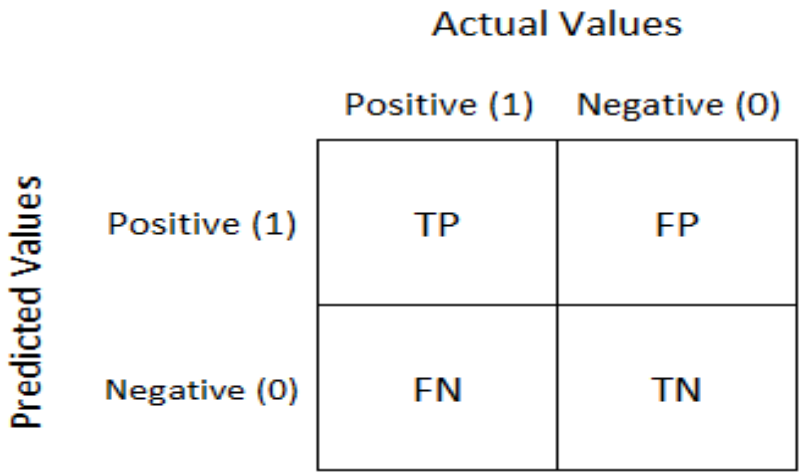
4.5 Ocene kvaliteta modela logističke regresije

U ovom delu će biti predstavaljene osnovne metrike koje govore o kvalitetu dobijenog modela primenom logističke regresije, ali se upošteno mogu primeniti na bilo koji model klasifikacije.

Rezultat logističke regresije je verovatnoća da ocenjena instance uzima 1 kao vrednost zavisne promenljive. Pored verovatnoće najčešće je potrebno dati i ocenu zavisne promenljive $\hat{y}$ koja kao i realizovana vrednost zavisne promenljive uzima vrednost 0 ili 1. Potrebno je postaviti prag c (**treshold**) ili u domenu kreditnog skoringa ova prag se zove **PD cut-off.**

Uzećemo da važi da je $\hat{y} = 1$ ukoliko je $\hat{\pi}_i > c$ tj. $\hat{y} = 0$ ukoliko je $\hat{\pi}_i \leq c$. Postoji više načina da se odredi optimalni PD cut-off, a mi ćemo ga odrediti na način da je stvarna default stopa na uzorku jednaka predviđenoj default stopi. Ukoliko je stopa defaulta na uzorku d %, c će biti (100-d) postotni percentil reallizovanih verovatnoća.

Kada smo odredili šta su vrednosti zavisne promenljive predviđene modelom uvedimo pojam **matrice konfuzije** (**confusion matrix**). Ona predstavlja tabelu sa 4 različite kombinacije predviđenih i stvarnih vrednosti. (Slika 17)



Slika 17 – Matrica konfuzije

Izvor: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Interpretacija je sledeća

True Positive - stvarno pozitivni: Model je predvideo pozitivan ishod i to je tačno.

True Negative - stvarno negativni: Model je predvideo negativan ishod i to je tačno.

False Positive – lažno pozitivni (Greška tipa 1): Model je predvideo pozitivan ishod i to nije tačno.

False Negative – lažno negativni: (Greška tipa 2): Model je predvideo negativan ishod i to nije tačno.

Postoje 4 značajne metrike koje se mogu dobiti iz matrice klasifikacije: Odziv - **Recall**, Preciznost - **Precision**, Tačnost - **Accuracy** i **F skor** ili F metrika.

Odziv se može tumačiti na sledeći način: koliko je pozitivnih slučajeva detektovano modelom.

$$Recall = \frac{TP}{TP + FN}$$

Preciznost se može tumačiti kao odnos broja stvarnih pozitivnih slučajeva i broja predviđenih pozitivnih slučajeva.

$$Precision = \frac{TP}{TP + FP}$$

Tačnost meri udeo tačno predviđenih slučajeva, sveukupno.

$$Accuracy = \frac{TP + FN}{TP + FP + TN + FN}$$

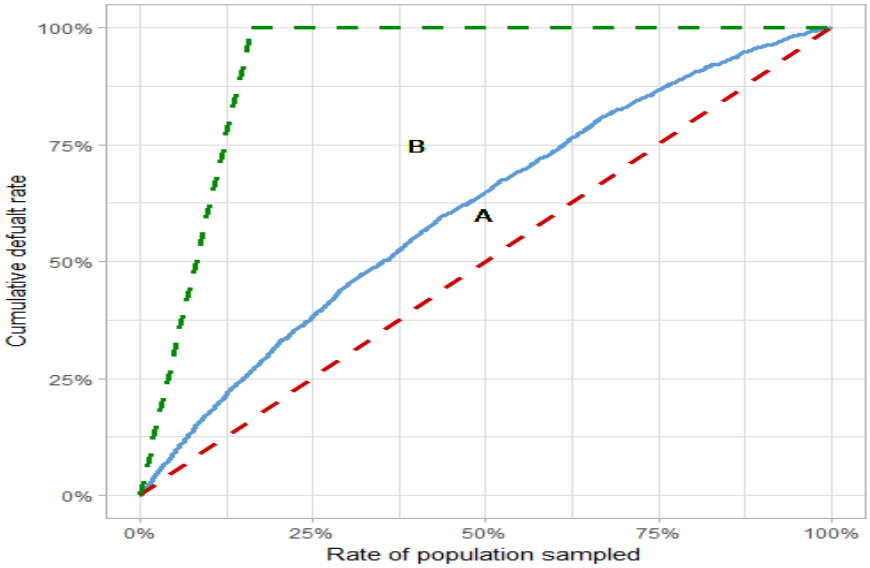
Teško je uporediti dva modela sa malom preciznošću i visokim odzivom ili obrnuto. Da bi bili uporedivi, koristi se F-skor ili F-metrika. Ona pomaže u merenju odziva i preciznosti u isto vreme. Koristi harmonijsku sredinu umesto aritmetičke sredine tako što više kažnjava ekstremne vrednosti (sledi iz nejednakosti između aritmetičke i harmonijske sredine).

$$F - measure = \frac{2 * Recall * Precision}{Recall + Precision}$$

Navedimo 2 važne metrike široko rasprostranjene u oceni kvaliteta modela logisitčke regresije a posebno modela u domenu kreditnog skoringa. U pitanju su **Gini** koeficijent i **KS** (Kologorov-Smirnov) statistika.

Gini koeficijent prikazuje koliko dobro razvijen model razlikuje dobre od loših slučajeva kao i efikasnost razvijenog modela.

Ginijev koeficijent meri stepen razdvajanja izlaznog rezultata modela. Gini koeficijent daje broj između 0 i 1, pri čemu 0 označava savršenu nasumičnost, a 1 ukazuje na savršeno razdvajanje (a time i savršenu diskriminatornu moć modela). Da bi se ilustrovao mehanizam Gini koeficijenta, obično se koristi kriva kumulativnog profila tačnosti (**CAP**), koja je kumulativni prikaz procenta ukupnih slučajeva naspram procenta detektovanih pozitivnih vrednosti zavisne promenljive. Dijagonalna linija na slici predstavlja liniju savršene slučajnosti. Ovo je ekvivalentno nasumičnom dodeljivanju rezultata (predviđanja) zapažanjima i zatim merenju koliko dobro predviđaju ciljnu promenljivu. Slika (Slika 18) ilustruje da tačke dalje od dijagonale odgovaraju visokoj diskriminatornoj moći. Gini koeficijent poredi površinu između krive i prave (A) sa celokupnom površinom trougla iznad prave (A+B).



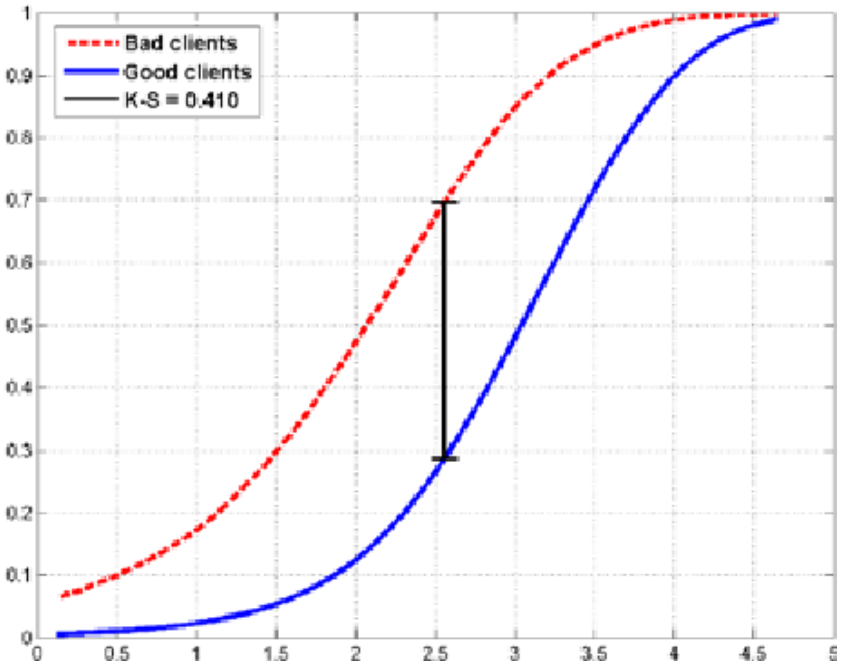
Slika 18 – Gini koeficijent - ilustracija

Izvor: Originalni grafik autora

KS statistika je definisana kao maksimalno rastojanje između kumulativne raspodele dobrih i loših slučajeva (Slika 19), sa predviđenim rezultatom (verovatnoćom) na x-osi. Stoga se može tumačiti kao mera razdvajanja između dobrih i loših slučajeva. Formalno, statistika KS je data formulom:

$$KS = |F(G) - F(s|B)|$$

Gde $F(G)$ i $F(B)$ predstavljaju kumulativne raspodele dobrih i loših slučajeva.



Slika 19 – KS koeficijent - ilustracija

Izvor: https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/

U praksi su prihvatljive vrednosti za Gini > 50% dok je za KS> 40% prihvatljivo.

5 Bejzov pristup logističkoj regresiji

5.1 Bejzova statistika

U osnovi Bejzove (*Bayes*), statistike je Bejzova teorema, koja se lako izvodi, ako se poznaje pojam **uslovne verovatnoće**.[23]

Primer:

Svake godine oko 18250 ljudi u Srbiji doživi srčani udar. Prema popisu iz 2016. godine, u Srbiji ima 7.057.000 ljudi, dakle, verovatnoća da slučajno izabran stanovnik Srbije doživi srčani udar u narednoj godini je približno 0.3%. Međutim, niko od nas nije slučajno izabran stanovnik Srbije. Lekari izdvajaju brojne faktore rizika za srčani udar: pol, starost, nivo holesterola, krvni protisak i slično. Ako posmatramo muškarca od 45 godina, pušača, sa normalnim krvnim pritiskom, kolika je njegova verovatnoća srčanog udara? Da li dati uslovi menjaju verovatnoću od 0.3% ?

Dakle, traži se uslovna verovatnoća događaja A, pod uslovom da se desio događaj B.

$$P(A \mid B) = \frac{P(AB)}{P(A)}$$

Znači, ako je osoba pušač, kolika je verovatnoća da doživi srčani udar? Međutim, može se postaviti i obrnuto pitanje: ako je neko doživeo srčani udar, kolika je verovatnoća da je pušač? Ovakva pitanja se često pojavljuju prilikom dijagnostifikovanja bolesti nekom medicinskom metodom tj. testom. Ako postoji bolest, kolika je verovatnoća pozitivnog testa i obrnuto, ako je test pozitivan, kolika je verovatnoća postojanja bolesti?

Englez **Tomas Bejz** (Thomas Bayes 1701-1761) bio je sveštenik Prezbiterijanske crkve, religijske denominacije koja je progonjena zato što je odbila da podrži Englesku crkvu. Malo je poznato o njegovom životu. Publikovao je dva rada tokom svog života. Jedan je bio iz teologije, a drugi ("An Introduction to the Doctrine of Fluxions, and a Defense of the Mathematicians Against the Objections of the Author of the Analyst"), je bio odbrana Njutnovo infinitezimalnog računa od kritike biskupa Berklija (George Berkeley, 1685-1753). Ovaj drugi rad je bio toliko impresivan de je bio izabran u Kraljevsko Društvo (Royal Society).

Međutim, njegovo najznačajnije otkriće je objavljeno posle njegove smrti. "An Essay towards Solving a Problem in the Doctrine of Chances" u "Philosophical Transaction of the Royal Society of London", štampano je zahvaljujući Ričardu Prajsu (Richard Price) 1763.godine.

Naime, tokom četrdesetih godina XVIII veka, Tomas Bejz došao je do genijalnog otkrića koje danas nosi njegovo ime. Ovo je kasnije ponovo nezavisno otkriveno od strane poznatog

francuskog matematičara, Pjera Simona Laplasa (Pierre-Simon Laplace, 1749-1827), koji je istom otkriću dao moderni matematički oblik i naučnu primenu.

U vreme kada je Bejz živeo pojam verovatnoće jedva da je i postojao. Jedina primena tog pojma bila je u kockanju, a i tu se primenjivala jedino na osnovna pitanja, kao što je verovatnoća izvlačenja četiri asa u jednoj ruci u pokeru ili verovatnoća vezana za bacanje kockica za igru. Bejz je izveo formulu za računanje uslovne verovatnoće događaja "posle" pod uslovom "ranije, pre". Verovatnoću da padne šestica ako je kocla ispravna. Verovatnoća posledice ako se desio uzrok.

Ali, formula je posedovala unutrašnju simetriju. Bilo je moguće izračunati verovatnoću "pre" pod uslovom "posle". Ako je pala šestica, koja je verovatnoća da je kocka ispravna? Verovatnoću uzroka ako se desila posledica. Ovo se u to vreme činilo potpuno besmisleno, pa Bejz nije nastavio istraživanje.

Na konceptualnom nivou Bejzov sistem je bio dosta prost: Ljudi imaju svoje subjektivno mišljenje, verovanje o nekom događaju i to predstavlja prvu, **priornu** verovatnoću tog događaja. Zatim, ljudi modifikuju svoja mišljenja, verovanja u skladu sa objektivnim informacijama, odnosno:

Početna verovanja + skorašnji objektivni podaci = nova, poboljšana verovanja.

On je uveo i specifične termine kojima je obrazložio svoju argumentaciju:

(1) **prior** - verovatnoća početnog verovanja

(2) **izvesnost** - verovatnoća drugih hipoteza ukoliko se u obzir uzmu novi objektivni podaci,

(3) **posterior** - verovatnoća novog revizovanog verovanja.

Ono što je danas poznato kao Bejzovska statistika ima svoje uspone i padove od 1763. Iako su Bejzovski metod preuzeli Laplas i drugi vodeći probabilisti njegovog vremena, taj metod je u devetnaestom veku bio veoma kritikovan i osporavan, jer nisu znali kako da se izbore sa priornim verovatnoćama.

Osim što je Bejzovo učenje bilo osporavano, bilo je i računski veoma izazovno sve do kasnih osamdesetih i devedesetih godina, kada su postali dostupni moćni računari i razvijeni novi računarski metodi, kao što su Markovljevi lanci Monte Karlo (MCMC). Interesovanje za Bejzovu statistiku je eksplodiralo, što je rezultiralo ne samo intenzivnim istraživanjem Bejzovske metodologije, već i njenom primenom u različitim oblastima kao što su astrofizika, meteorologija, medicina, arheologija, krivično pravo, psihologija, sport, politika (izbori) i borba protiv terorizma.

5.1.1 Poređenje sa frekvencionističkim pristupom

Tokom vremena, razvila su se dva osnovna pristupa verovatnoći i statističkom zaključivanju: frekvencionistički i Bejzovski. Frekvencionistički pristup je bio dominantan u 20. veku, a temelji se na proceni verovatnoće događaja putem ponavljanja eksperimenata pod istim uslovima. Ovde se verovatnoća smatra objektivnom i nezavisnom od subjektivnog verovanja.[23]

S druge strane, Bejzovski pristup posmatra verovatnoću kao stepen ličnog verovanja pojedinca u tačnost nekog iskaza. Ovde se verovatnoća tumači kao kvantifikacija nesigurnosti. Primeri uključuju procenu krivice optužene osobe, uspeh akcija na berzi, tačnost odgovora studenata na testu, ili efikasnost leka na pacijenta.[23]

Bejzovski pristup može se smatrati generalizacijom frekvencionističkog. Oba pristupa imaju svoje prednosti i mane. Frekvencionistički pristup može dovesti do različitih zaključaka u zavisnosti od obima uzorka, dok Bejzovski pristup omogućava dodeljivanje verovatnoća samim hipotezama i tretira neodređenost na drugačiji način. Ova dva pristupa se razlikuju u tome što frekvencionistički govori o varijabilnosti, dok Bejzovski govori o neodređenosti.

5.1.2 Bejzovska verovatnoća

Posmatrajmo iskaz "u Srbiji živi više od 10 miliona ljudi”. Za nekoga, ko živi u Srbiji, verovatnoća da je ovaj iskaz tačan je 0. Ali, neko, ko ne zna ništa o Srbiji, bi ovom iskazu dodelio verovatnoću 0.5.

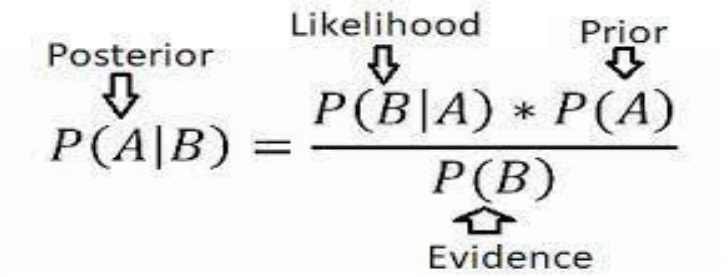
Dakle, različiti ljudi mogu imati, davati različite verovatnoće istim iskazima. To je prirodno, jer mogu imati različite informacije ili različite pretpostavke o iskazu.

Suština Bejzovske statistike je da, kada su dostupne nove informacije, (neki podaci), verovatnoća se menja, tj. "apdejtuje” (Tabela 2 Slika 20). Početna, "priorna" verovatnoća iskaza, hipoteze je P(H). Ako se dobije nova informacija A, nova, ažurirana, apdejtovana, verovatnoća je P(H|A). Ako se dobije i informacija B, nova verovatnoća je P(H|AB). Ove verovatnoće se računaju po **Bejzovoj formuli**:

$$P(H \mid D) = \frac{P(H) * P(D \mid H)}{P(D)}$$

$P(H)$	Početna verovatnoća, "pior"
$P(D \mid H)$	Verovatnoća da se dobiju podaci ako je hipoteza tačna, verodostojnost, likelihood
$P(D)$	Verovatnoća da se dobiju podaci, bez obzira na hipotezu, "magrinalna verodostojnost
$P(H \mid D)$	Ažurirana verovatnoća, "posterior"

Tabela 2 – Elementi Bejzove teoreme



Slika 20 – Elementi Bejzove teoreme

Izvor: https://benjaminwhiteside.com/2020/10/25/bayes-theorem/

5.1.3 Bejzovske ocene parametara

Naučne hipoteze se obično izražavaju pomoću raspodela verovatnoća za podatke koji se posmatraju. Ove raspodele verovatnoća zavise od nepoznatih veličina koje se nazivaju parametri i označavaju sa θ. Na primer, očekivana vrednost, proporcija, medijana, standardna devijacija mogu biti nepoznazi parametri.

U frekvencionstičkom pristupu verovatnoći i statističkom zaključivanju postavljaju se hipoteze o parametrima, a parametri se smatraju za fiksirane konstante. Zaključivanje se vrši na osnovu p -vrednosti, uvode se intervali poverenja.

U suštini Bejzovog statističkog zaključivanja je proces ažuriranja (update) nečijeg verovanja (prior) o pojavi koja se posmatra u svetlu ocenjivanja izvesnosti, (**evidence**), pomoću Bejzove teoreme i na taj način formiranja novog verovanja (**posterior**). Ovde se parametri smatraju za slučajne promenljive, sa unapred datom raspodelom (**prior**). Svo znanje o parametru koje se trenutno poseduje izražava se preko raspodele verovatnoća, "prior distribution"- priorne raspodele:

$$p(\theta)$$

Kad su dostupni podaci y, informaciju koju oni sadrže o parametrima populacije se izražava kao verodostojnost "likelihood", koja je proporcionalana raspodeli opserviranih podataka ako su dati parametri modela, što se zapisuje kao:

$$p(\boldsymbol{y} \mid \theta)$$

Ova informacija se sada kombinuje sa priornom raspodelom verovatnoća, da bi se dobila ažurirana „posteriorna raspodela“ na kojoj se zasniva Bejzovsko zaključivanje. Posteriorna raspodela se dobija primenom Bejzove teoreme.

$$p(\theta \mid \boldsymbol{y}) = \frac{p(\theta) * p(\boldsymbol{y} \mid \theta)}{\int_{\theta}\ p(\theta) * p(\boldsymbol{y} \mid \theta) d\theta}$$

Takođe, imamo da je posteriorna raspodela proporcionalna prouzvodu priorne raspodele i verodostojnosti:

$$p(\theta \mid \boldsymbol{y}) \propto p(\theta) * p(\boldsymbol{y} \mid \theta).$$

Verodostojnost (likelihood) obično se određuje tako što se prvo izračuna "raspodela uzorkovanja" (sampling distribution) - $p(\boldsymbol{y} \mid \theta) = P(D \mid \theta)$, koja predstavlja verovatnoću podataka ako je određeni parametar poznat. Ova raspodela uzorkovanja se bazira na podacima koji su već prikupljeni. Zatim, kako bismo dobili verodostojnost, koristimo tu raspodelu uzorkovanja i zamenjujemo konkretne podatke (D) u nju. Drugim rečima, verodostojnost je verovatnoća da su konkretni podaci koji su nam dostupni generisani iz te raspodele, s obzirom na vrednost parametra θ.

U suštini, verodostojnost je način da kvantifikujemo koliko su naši konkretni podaci usklađeni sa određenim parametrima, koristeći verovatnoću raspodele podataka pod tim parametrima.

U slučaju diskretnih priornih raspodela hipoteze o parametru mogu biti $\theta = 2, \theta = 3 \ldots, \theta = 5$…Sledeći Bejzov princip, svaki od ovih iskaza ima neku priornu verovatnoću, $P(\theta = 2)$, odnosno $P(\theta = 3) \ldots P(\theta = 5)$. Može se kazati da parametar ima priornu raspodelu verovatnoća.

Ako su registrovani podaci $x = 2$, potrebno je odrediti posterioru raspodelu verovatnoća:

$$P(\theta = 1 \mid x = 2) = \frac{P(\theta = 1) * P(x = 2 \mid \theta = 1)}{P(x = 2)}$$

$$P(\theta = 3 \mid x = 2) = \frac{P(\theta = 3) * P(x = 2 \mid \theta = 3)}{P(x = 2)}$$

$$\ldots$$

$$P(\theta = 5 \mid x = 2) = \frac{P(\theta = 5) * P(x = 2 \mid \theta = 5)}{P(x = 2)}$$

U slučaju neprekidnih priornih raspodela i raspodela verodostojnosti, stvari su mnogo komplikovanije. Polazi se od priorne raspodele

$$p(\theta)$$

Kad su dostupni podaci $\boldsymbol{y}$, informaciju koju oni sadrže o parametrima populacije se izražava ko verodostojnost "likelihood", koja je proporcionalana raspodeli opserviranih podataka ako su dati parametri modela, što se zapisuje kao

$$p(\boldsymbol{y} \mid \theta)$$

Ova informacija se sada kombinuje sa priornom raspodelom verovatnoća, da bi se dobila ažurirana „posteriorna raspodela“ na kojoj se zasniva Bejzovsko zaključivanje. Posteriorna raspodela se dobija primenom Bejzove teoreme.

$$p(\theta \mid \boldsymbol{y}) = \frac{p(\theta) * p(\boldsymbol{y} \mid \theta)}{p(\boldsymbol{y})}$$

Ovde se $p(\boldsymbol{y})$ dobija tako da je $p(\theta) * p(\boldsymbol{y} \mid \theta)$ raspodela verovatnoća, to jest kao normalizacija:

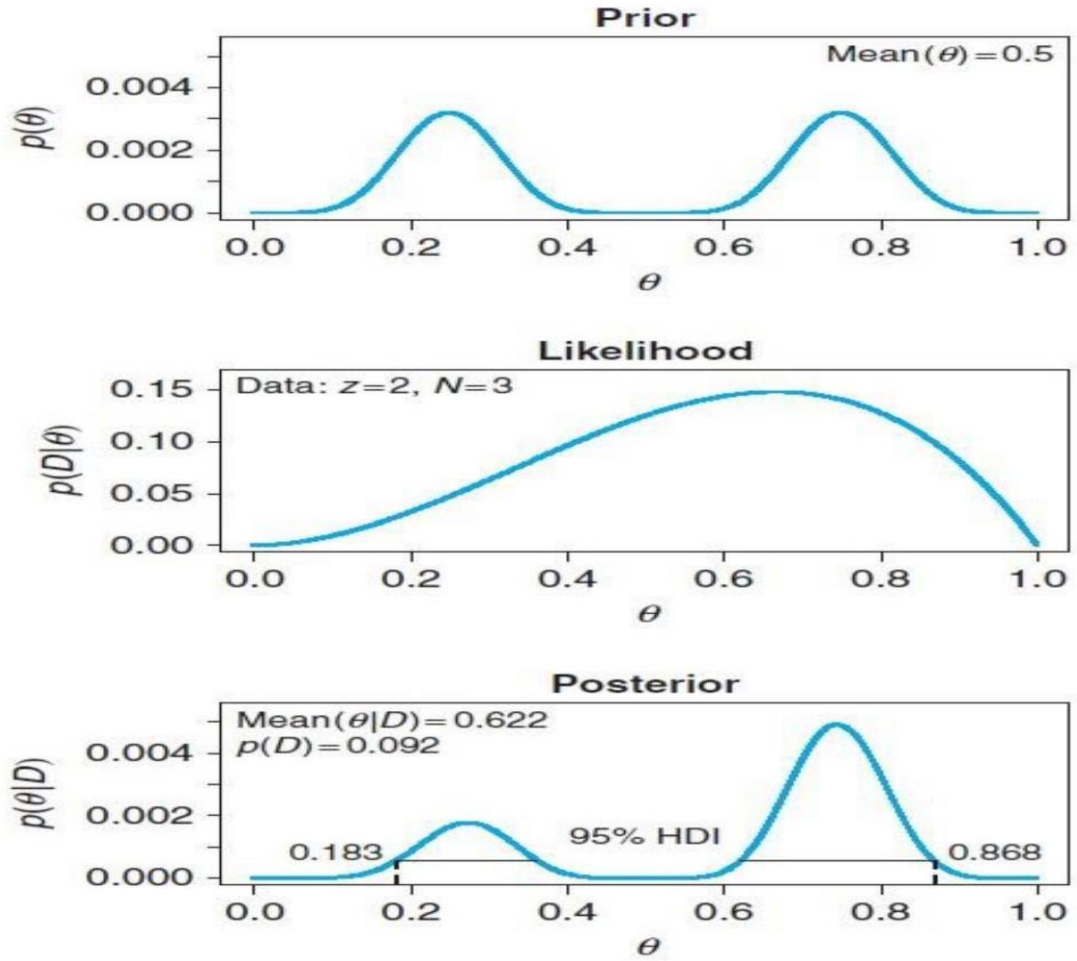
$$\int_{\theta} p(\theta) * p(\boldsymbol{y} \mid \theta) d\theta.$$

Konačno,

$$p(\theta \mid \boldsymbol{y}) = \frac{p(\theta) * p(\boldsymbol{y} \mid \theta)}{\int_{\theta} p(\theta) * p(\boldsymbol{y} \mid \theta) d\theta}$$

$p(\theta \mid \boldsymbol{y}) \propto p(\theta) * p(\boldsymbol{y} \mid \theta)$ to jest posterior $\propto$ prior $\times$ likelihood.

Na slici ispod (Slika 21) se vidi kako se priorna znanja o raspodeli promenljive mogu pormeniti znajući nove informacije.



Slika 21 – Primer priorne i posteriorne raspodele

Izvor: Z.L. Crvenković: Bejzova statistika

Teorijski, posteriorna raspodela (raspodela parametara nakon što su uzeti u obzir podaci) uvek postoji, ali za realne i kompleksne modele, analitički proračuni često postaju izuzetno teški ili čak nemogući. Krajem osamdesetih i devedesetih godina prošlog veka, razvijeni su metodi koji omogućavaju generisanje uzoraka iz ove posteriorne raspodele. Danas se često koriste Monte Carlo metodi Markovskih lanaca za ovo generisanje, poznati kao Markovljevi lanac Monte Karlo (MCMC).

Postoji mnogo razloga za usvajanje Bejzovog pristupa i metoda, a primene se pojavljuju u različitim oblastima. Mnogi se opredeljuju za Bejzovski pristup zbog njegove filozofske doslednosti.[23] Neke osnovne teoreme sugerišu da je jedini način da se donesu dosledne i ispravne odluke korišćenje Bejzovskog pristupa. Drugi ukazuju na logičke probleme unutar frekvencionističkog pristupa koji se ne pojavljuju u Bejzovskom. S druge strane, priorne raspodele u Bejzovskom pristupu su subjektivne - svako ima svoje prethodno verovanje i informaciju. Mnogi statističari to smatraju glavnim nedostatkom Bejzovskih metode.

5.2 Bejzova logistička regresija

Bejzovo zaključivanje za model logističke regresije zahteva priorne raspodele za parameter modela. Wilhelmsen [22] i Ziemba [24] koriste normalnu raspodelu parametara predstavljenu na sledeći način:

$$\pi(\beta_i) = N(\beta_{0i}, \sigma_i^2) \tag{9}$$

Posteriorna raspodela je srazmerna proizvodu priorne raspodele i verodostojnosti, $\pi(\beta \mid y) \propto L(y \mid \beta)\pi(\beta)$. Dakle iz izraza, (1) I (9), imamo:

$$\pi(\beta \mid y) \propto \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi\sigma_j^2}} exp\left(-\frac{(\beta_j-\beta_{0j})^2}{2\sigma_j^2}\right) \tag{10}$$

ili

$$\pi(\beta \mid y) = \frac{\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{(\beta_j-\beta_{0j})^2}{2\sigma_j^2}\right)}{\int_{-\infty}^{\infty} \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{(\beta_j-\beta_{0j})^2}{2\sigma_j^2}\right) d\beta} \tag{11}$$

kada u obzir uzmemo faktor normalizacije.

Ovakav oblik posteriorne raspodele, jednačina (10), sugeriše da prior nije deo konjugovane porodice raspodela (**conjugate family**). U stvari, ne postoji konjugovani prior za Bejzovski model logističke regresije. Normalizaciona konstanta, integral u imeniocu, ne može se izračunati eksplicitno. U ovakvoj situaciji, potrebno je koristiti simulacione metode kako bismo dobili posteriorne raspodele parametara. **Metode Markovljevih lanaca Monte Karlo** (**MCMC**) se koriste, gde se generiše Markovljev lanac sa stacionarnom raspodelom koja odgovara posteriornoj raspodeli vektora **β.**

5.3 Monte Karlo metode

Simulacije su znatno proširile obim Bejzovog zaključivanja. Metode Markovljevih lanaca

Monte Karlo (MCMC) omogućavaju uzorkovanje iz nestandardne distribucije. Prema tome, Bejzovo zaključivanje se može izvesti u širokom spektru oblika posteriorne distribucije, na primer jednačina (11). Ideja je da se generiše Markovljev lanac čija je granična (stacionarna) raspodela jednaka posteriornoj raspodeli. Ovaj odeljak će opisati tehnike simulacije, pružiti uvod u Markovljeve lance, a zatim objasniti ulogu i svrhu Markovljevih lanaca Monte Karlo.

U Bejzovom zaključivanju, potrebna je simulacija za procenu integrala. Da bi se to uradilo, neophodno je da se mogu generisati nasumični (**random**) podaci. Generisanje slučajnih promenljivih i sve druge Monte Karlo metode se oslanjaju na generisanje uniformnih slučajnih promenljivih na intervalu (0,1).

Polazna tačka za definisanje slučajnih pormenljivih je generisanje slučajnih brojeva na intervalu (0,1).

Prema Kroese [14] dva odlična generatora koji imaju veoma dobre performanse su:

• **Combined multiple-recursive** generatori.

• **Twisted general feedback** shift register generatori.

Srećom, savremeni programski jezici koriste veoma dobre generatore pseudo brojeva. Programski jezik **Python** korišćen za praktičan deo ovog rada koristi **Mersenne Twister** generator, koji je jedan od nakorišćenijih i ujedno najtestiranijih generatora.

5.3.1 Generator slučajnih promenljivih

Dva uobičajena generator slučajnih promenljivih su metod inverzne transformacije (**inverse transform method**) i **accept-reject** metod.

5.3.1.1 Metod inverzne transformacije

Ovaj metod se uvodi na sledeći način:

Neka je X slučajna promenljia sa funkcijom raspodele (**cumulative distribution function - cdf**) $F(x) = P(X \leq x)$. Kako je F is neopadajuća funkcija, njena inverzna funkcija može biti definisana sa:

$$F^{-1}(y) = \begin{cases} \min\{x: F(x) \geq y\} & \text{if } y > 0 \\ -\infty & \text{if } y = 0 \end{cases}.$$

Dalje, ako imamo slučajni promenljivu U iz uniormne raspodele na $(0,1)$, tj. $U \sim \text{unif}(0,1)$, tada postoji inverzna funkcija $F^{-1}(U)$ data sa:

$$P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x).$$

Dakle da bismo definisali slučajnu promenljivu X sa funkcijom raspodele $F(x)$, generišemo U sa uniformnom raspodelom na $(0,1)$ I izvršavamo inverznu transformaciju $x = F^{-1}(u)$. Prema tome algoritam je sledeći:

1. Generisanje U iz unif (0,1);
2. Vraćanje $X = F^{-1}(U)$.

Ova metoda se koristi za uzorkovanje slučajne promenjlive sa kontinualnom raspodelom. Očigledno, ovaj metod funkcioniše samo kada možemo da odredimo inverznu funkciju funkcije raspodele F.

5.3.1.2 Accept-reject metoda

Metoda inverzne transformacije nije od koristi kada se ne može dobiti inverzna funkcija funkcije raspodele. Opštija metoda je metoda 'prihvata-odbaci' koja se može koristiti za uzorkovanje iz opštijih distribucija.

Po Greenberg-u [9], the accept-reject metoda se može koristiti za simuliranje slučajne promenljive sa funkcijom gustine $f(x)$ kada je moguće simulrati vrednosti iz neke druge funkcije gustine $g(x)$, i ako postoji broj $M \geq 1$ tako da he $f(x) \leq Mg(x)$ za svako x. Funkcija gustine $g(x)$ is se naziva instrumental ili kandidat gustina. Da bi simulirali slučajnu promenljivu X sa gustinom $f(x)$ nezavisno generišemo dve slučajne promenljive i $Y \sim g(x)$ i $U \sim unif(0,1)$. Tada ako

$$U \leq \frac{1}{M}\frac{f(Y)}{g(Y)}$$

setujemo $X = Y$. Ako ne, odbacujemo Y. Ovo predstavlja **accept-reject** algoritam

1. Generišemo Y iz $g(x)$;
2. Generišemo U iz unif $(0,1)$, nezavisno od Y;
3. Prihvatamo $X = Y$ ako je $U \leq \frac{1}{M}\frac{f(Y)}{g(Y)}$, inače odbacujemo Y;
4. Povratak na korak 1.

Po Robertu i Casella [18], funkcija raspodele $P\left(Y \leq x \mid U \leq \frac{f(Y)}{Mg(Y)}\right)$ je upravo funkcija raspodele od X. To vidimo iz sledećeg

$$P\left(Y \le x \middle| U \le \frac{f(Y)}{Mg(Y)}\right) = \frac{P\left(Y \le x, U \le \frac{f(Y)}{Mg(Y)}\right)}{P\left(U \le \frac{f(Y)}{Mg(Y)}\right)}$$

$$= \frac{\int_{-\infty}^{x}\left[\int_{0}^{\frac{f(y)}{Mg(y)}} du\right] g(y)dy}{\int_{-\infty}^{\infty}\left[\int_{0}^{\frac{f(y)}{Mg(y)}} du\right] g(y)dy}$$

$$= \frac{\int_{-\infty}^{x} \frac{f(y)}{Mg(y)} g(y)dy}{\int_{-\infty}^{\infty} \frac{f(y)}{Mg(y)} g(y)dy}$$

$$P\left(Y \le x \middle| U \le \frac{f(Y)}{Mg(Y)}\right) = \frac{\int_{-\infty}^{x} f(y)dy}{\int_{-\infty}^{\infty} f(y)dy} = \int_{-\infty}^{x} f(y)dy = P(Y \le y).$$

Čime je tvrdnja potvrđena.

Posmatrajmo efikasnost ove metode. Primetimo da verovatnoća da generisana tačka bude prihvaćena data sa

$$P(\text{ accept }) = P\left(U \le \frac{f(y)}{Mg(y)}\right) = \int_{-\infty}^{\infty}\left(\int_{0}^{\frac{f(y)}{Mg(y)}} 1du\right) g(y)dy$$

$$= \int_{-\infty}^{\infty} \frac{f(y)}{Mg(y)} g(y)dy$$

$$= \int_{-\infty}^{\infty} \frac{1}{M} f(y)dy = \frac{1}{M}$$

Ovo implicira da bi trebalo da izaberemo što manji M kako bismo maksimizirali verovatnoću prihvatanja. Algoritam je efikasan kada je g što je moguće bliže f. Maksimiziranje verovatnoće prihvatanja je važno jer odbačene vrednosti koriste računarsko vreme bez dodavanja uzorku, dakle, smanjuje efikasnost.

5.3.2 Monte Karlo integracija

Monte Karlo integracija je statistička tehnika za aproksimaciju integrala. Koristi simulaciju za

dobijanje procene integrala koji ima srednju vrednost i varijansu. Jedan od metoda Monte Karlo integracije je pristup srednje vrednosti uzorka **(sample mean approach**). Ova metoda je opisana u nastavku za procenu integrala, $I = \int_a^b f(x)dx$.. Sledeći pristup je razmatran u Suess i Trumbo [19] .

Ako ima uniformnu raspodelu, $X \sim$ unif (a, b) tada $E(f(X)) = \int_a^b \left(\frac{1}{b-a}\right) f(x)dx = \frac{1}{b-a}\int_a^b f(x)dx$. Prema tome,

$$\int_a^b f(x)dx = (b-a)E(f(x))$$

Dakle, integral $\int_a^b f(x)dx$ može biti aproksimiran sa

$$\theta = \frac{b-a}{N}\sum_{k=1}^{N} f(u_k) \tag{12}$$

Gde su $u_1, u_2, \dots, u_N$ slučajni brojevi iz unif (a, b). Srednja vrednost i varijansa ovog estimatora se dobijaju na sledeći način:

$$E(\theta) = E\left(\frac{b-a}{N}\sum_{k=1}^{N} f(u_k)\right) = \frac{b-a}{N}NE(f(U)) = \frac{b-a}{N}N\frac{1}{b-a}\int_a^b f(u)du = I$$

Prema tome, predloženi estimator iz, (12) je nepristrasan estimator za integral , $I = \int_a^b f(x)dx$. Dalje, varijansa

$$\begin{aligned}
\text{var}\,(\theta) &= \text{var}\left(\frac{b-a}{N}\sum_{k=1}^{N} f(u_k)\right) \\
&= \frac{(b-a)^2}{N^2}\text{Nvar}\,(f(U)) \\
&= \frac{(b-a)^2}{N}[(E(f(U)^2) - (E(f(U)))^2] \\
&= \frac{(b-a)^2}{N}\left[\frac{1}{b-a}\int_a^b (f(u))^2du - \left(\frac{1}{b-a}\int_a^b f(u)du\right)^2\right] \\
&= \frac{1}{N}(b-a)\left[\int_a^b (f(u))^2dx - \left(\int_a^b f(u)dx\right)^2\right].
\end{aligned} \tag{13}$$

Dakle važi da je var $(\theta) \propto \frac{1}{N}$.

5.3.2.1 Importance sampling

Importance sampling se koristi da redukuje varijansu Monte Karlo procene integrala. Iz (13) vidimo da je standardna devijacija estimatora integrala , std $(\theta) \propto \frac{1}{\sqrt{N}}$. Prema tome standardna devijacija estimatora opada kako N raste, ali sa opadajućom stopom. Ovo znači da ako

povećamo broj slučajnih tački sa $N = 10^2$ na $N = 10^4$ standardna devijacija se poboljša od reda veličine $\frac{1}{10}$ do $\frac{1}{100}$. Dakle veliki broj slučajnih tački je potreban za bitno poboljšanje u tačnosti estimatora. Importance Sampling ima za cilj da unapredi standardnu devijaciju Monte Karlo estimatora. Ideja koja sledi je predstavljena u i Robert i Casella [18] .

Posmatrajmo funkciju gustine $p(x)$ na $[a, b]$ gde je $p(x) > 0$ kad god je $f(x) \neq 0$. Tada

$$\int_a^b f(x)dx = \int_a^b \frac{f(x)}{p(x)} p(x)dx = E_p\left(\frac{f(X)}{p(X)}\right) \text{ ako } X \sim p(x).$$

Prema tome, da bismo dobili estimator za $\int_a^b f(x)dx$ koristeći importance sampling, uzorkujemo $x_1, x_2, \dots, x_N$ iz $p(x)$ i dajemo procenu:

$$\int_a^b f(x)dx \approx \frac{1}{N}\sum_{k=1}^N \frac{f(x_k)}{p(x_k)}$$

Novi estimator je dat sa $\vartheta = \frac{1}{N}\sum_{k=1}^N \frac{f(X_k)}{p(X_k)}$ I varijansa od ϑ je data sa

$$\begin{aligned}
\text{var}\,(\vartheta) &= \text{var}\left(\frac{1}{N}\sum_{k=1}^{N}\frac{f(X_k)}{p(X_k)}\right) \\
&= \frac{1}{N^2}\,\text{Nvar}\left(\frac{f(X)}{p(X)}\right) \\
&= \frac{1}{N}\left[E\left(\frac{f(X)}{p(X)}\right)^2 - \left(E\left(\frac{f(X)}{p(X)}\right)\right)^2\right] \\
&= \frac{1}{N}\left[\int_a^b \frac{(f(x))^2}{(p(x))^2}p(x)dx - \left(\int_a^b f(x)dx\right)\right)^2\Bigg].
\end{aligned}$$

Da bismo minimizirali varijansu potrebno je da minimiziramo član $E\left(\frac{f(X)}{p(X)}\right)^2$. Koristeći Jensen-ovu nejednakost

$$E\left(\frac{f(X)}{p(X)}\right)^2 \geq \left[E\left(\frac{|f(X)|}{p(X)}\right)\right]^2 = \left(\int_a^b \frac{|f(x)|}{p(x)}p(x)dx\right)^2 = \left(\int_a^b |f(x)|dx\right)^2 \qquad (14)$$

Dobijamo donje ograničenje za posmatrani član, koje ne zavisi od izbora $p(x)$.

Teorema 1

Ako je $\vartheta = \frac{1}{N}\sum_{k=1}^N \frac{f(X_k)}{p(X_k)}$ gde su $X_1, X_2, \dots, X_N$ nezavisne slučajne promenljive sa funkcijom gustine $p(x)$ tako da je $p(x) > 0$ kad god je $f(x) > 0$ i $\int_a^b p(x)dx = 1$ tada je

1. $E(\vartheta) = \int_a^b f(x)dx.$

2. var (ϑ) je minimizirana za $p(x) = \frac{|f(x)|}{\int_a^b |f(x)|dx}$.

Dokaz:

1. $E(\vartheta) = E\left(\frac{1}{N}\sum_{k=1}^{N} \frac{f(X_k)}{p(X_k)}\right) = E\left(\frac{f(X)}{p(X)}\right) = \int_a^b \frac{f(x)}{p(x)} p(x)dx = \int_a^b f(x)dx.$

2. Iz jednačine (14) $\left(\int_a^b |f(x)|dx\right)^2$ je donja granica za $E\left(\frac{f(X)}{p(X)}\right)^2$ što smo i želeli da minimizujemo. Ako je $p(x) = \frac{|f(x)|}{\int_a^b f(x)dx}$, tada

$$\begin{aligned} E\left(\frac{f(X)}{p(X)}\right)^2 &= \int_a^b (f(x))^2 \frac{\int_a^b |f(x)|dx}{|f(x)|} dx \\ &= \int_a^b |f(x)|dx \int_a^b |f(x)|dx \\ &= \left(\int_a^b |f(x)|dx\right)^2. \end{aligned}$$

Dakle, za ovaj izbor $p(x)$ dosegnuto je donje ograničenje i varijansa je minimizirana. Praktična korisnost ove teoreme je veoma mala. Ovo je zbog toga što je potrebno da znamo integral $\int_a^b |f(x)|dx$ koji je za $f(x) \geq 0$ isti kao i $\int_a^b f(x)dx$ što je integral koji smo prvenstveno želeli da procenimo. Kako god ova teorema nam pomaže u dobrom izboru $p(x)$. Željeni cilj je postići da $\frac{f(x)}{p(x)}$ bude približno konstantno. Prema tome potrebno je uzorkovati više tačaka u regionima gde $f(x)$ uzima veće vrednosti. Time će "važni"(important) delovi integral biti estimirani bolje. Ovo je razlog zašto metod nosi naziv importance sampling.

5.2 Lanci Markova

U ovom odeljku dat je pregled pojma Markovljevih lanaca. Prethodno opisane metode simulacije ne mogu se lako primeniti u svim slučajevima. Monte Karlo integracija i importance sampling mogu se primeniti kada imamo posla sa standardnim distribucijama. Međutim, kada se suočimo sa nestandardnom distribucijom (kao što je slučaj sa Bajesovom logističkom regresijom), prethodne tehnike simulacije se ne mogu lako koristiti za dobijanje uzoraka iz bilo koje posteriorne distribucije. Ako se koriste, podležu velikim praktičnim poteškoćama. Metode Markovljevih lanca Monte Karlo (MCMC) pružaju izlaz.

Metode Monte Karla Markovljevog lanca su u velikoj meri poboljšale obim za Bejzovo zaključivanje. Pošto se MCMC oslanja na Markovljeve lance, oni su predstavljeni i sa diskretnim i sa kontinualnim prostorima stanja.

5.2.1 Diskretan prostor stanja

Sledi definicija lanca Markova:

Definicija 1

Neka je $(X_0, X_1, X_2, \dots)$ stohastički proces indeksiran sa t (najčešće vreme) koji uzima vrednosti iz konačnog skupa $S = \{1,2, \dots, s\}$ (konačni prostor stanja) ili $S = \{1,2, \dots\}$ (prebrojivo beskonačan prostor stanja). Ako važi Markovljev uslov:

$$P(X_{t+1} = j \mid X_t = k, X_{t-1} = k_{t-1}, \dots, X_1 = k_1, X_0 = k_0) = P(X_{t+1} = j \mid X_t = k) = p_{kj}$$

za sva stanja $j, k, k_{t-1}, \dots, k_1, k_0 \in S$ i za sve korake u vremenu $t = 0,1,2, \dots$, tada se $(X_0, X_1, X_2, \dots)$ zove **Markovljev lanac**.

Dakle trenutno stanje lanca Markova uzrokovano je samo pređašnjim. Verovatnoće p_{kj} se nazivaju **verovatnoće tranzicije** (transition probabilities). Verovatnoće tranzicije ne zavise od vremena t. Kakos u p_{kj} verovatnoće važi da je, $p_{kj} \geq 0$ kako vrednosti procesa pripadaju S

$$\sum_{j=1}^{s} p_{kj} = 1$$

Verovatnoće tranzicije su veoma važne i čuvaju se u $s \times s$ **tranzicionoj matrici**:

$$\boldsymbol{P} = \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1s} \\ p_{21} & p_{22} & \cdots & p_{2s} \\ \vdots & & \ddots & \vdots \\ p_{s1} & p_{s2} & \cdots & p_{ss} \end{pmatrix}$$

i –ti red u P, specificira raspodelu procesa u trenutku $t + 1$, uzimajući da je process u stanju i u trenutku t. Na primer, p_{22} predstavlja verovatnoću ostanka u stanju 2.

Posmatrajmo višekoračne tranzicione verovatnoće $p_{kj}^{(n)}$, koje se definišu sa

$$p_{kj}^{(n)} = P(X_n = j \mid X_0 = k) = P(X_{m+n} = j \mid X_m = k).$$

Računanje višekoračnih verovatnoća lako se dobija korišćenjem leme Chapman-Kolmogorova:

Lema 1

Neka je $(X_0, X_1, X_2, \dots)$ lanac Markova sa prostorom stanja $S = \{1,2,3, \dots\}$. Tada za višekoračne tranzicione verovatnoće važi:

$$p_{kj}^{(m+n)} = \sum_{i \in S} p_{ki}^{(m)} p_{ij}^{(n)}$$

Dokaz:

$$\begin{aligned} p_{kj}^{(m+n)} &= P(X_{m+n} = j \mid X_0 = k) \\ &= \sum_{i \in S} P(X_{m+n} = j, X_m = i \mid X_0 = k) \end{aligned}$$

$$= \sum_{i \in S} \frac{P(X_{m+n} = j, X_m = i, X_0 = k)}{P(X_0 = k)} \frac{P(X_m = i, X_0 = k)}{P(X_m = i, X_0 = k)}$$

$$= \sum_{i \in S} P(X_{m+n} = j \mid X_m = i, X_0 = k) P(X_m = i \mid X_0 = k).$$

Sada, koristeći osobinu Markova

$$p_{kj}^{(m+n)} = \sum_{i \in S} P(X_{m+n} = j \mid X_m = i) P(X_m = i \mid X_0 = k).$$

Što implicira,

$$p_{kj}^{(m+n)} = \sum_{i \in S} p_{ki}^{(m)} p_{ij}^{(n)}$$

The Chapman-Kolmogorov lema se može zapisati i u matričnoj formi

$$\boldsymbol{P}^{m+n} = \boldsymbol{P}^m \boldsymbol{P}^n.$$

Sada dajemo dajemo klasifikaciju mogućih stanja lanca.

Neka od stanja će biti posećena iznova i iznova beskonačno puta dok će neka biti posećena konačan broj puta i neće biti posećena ponovo. Definišimo za stanje $i \in S$

$$T_i = \min\{n \geq 1 : X_n = i\}$$

Dakle, T_i je vreme prve posete stanju i. Takođe definišimo

$$f_i = P(T_i < \infty \mid X_0 = i)$$

Što predstavlja verovatnoću da će se lanac Markova vratiti u stanje that the Markov chain will i odakle je lanac započet. Dva su moguća slučaja za f_i :

1. $f_i = 1$. Ovo znači da će se lanac izvesno neprekidno vraćati u stanje i. Ovakvo stanje se naziva **rekurentno** i biće posećeno beskonačno mnogo puta.

2. $f_i < 1$. Ovo znači da postoji pozitivna verovatnoća da se lanac nikada neće vratit u stanje i. Ovakvo stanje se naziva **prolazno** i biće posečeno konačan broj puta.

Kažemo da je stanje j **pristupačno** (accessible) iz stanja i ako postoji $n \geq 0$ tako da je $p_{ij}^{(n)} > 0$ i to označavamo sa $i \to j$. Stanje j is je pristupačno iz stanja i ukoliko u konačnom broju koraka konačnom broju koraka možemo stići iz i u j. Takođe, ako $i \to j$ I $j \to i$ kažemo da stanja i I j **komuniciraju** i označavamo sa $i \leftrightarrow j$. Može se pokazati da je **komuniciranje** relacija ekvivalencije između stanja na skupu stanja S. To znači da za sva stanja $i, j, k \in S$ važi

- $i \leftrightarrow i$ (**refleksivnost**);
- Ako $i \leftrightarrow j$ tada je i $j \leftrightarrow i$ (simetričnost);
- Ako $i \leftrightarrow j$ i $j \leftrightarrow k$, tada je i $i \leftrightarrow k$ (tranzitivnost).

Markovljev lanac je poznat kao **ireducibilan** ako postoji samo jedna komunikaciona klasa. To znači da sva stanja komuniciraju (proces može dostići bilo koje drugo stanje sa pozitivnom verovatnoćom). Ovo bi impliciralo da su za ireducibilan Markovljev lanac sa konačnim prostorom stanja sva stanja rekurentna.

Raspodela verovatnoća $\pi = (\pi_1, \pi_2, \dots)$ se naziva **stacionarnom (stationary, limiting)** raspodelom ako **je** $\boldsymbol{\pi = \pi P}$. Stacionarna raspodela, $\pi = \lim_{t\to\infty} \pi^t$ postoji ako je lanac Markova ireducibilan i sva stanja su **aperiodična** (najveći zajednički delilac skupova $A_i = \left\{t \geq 1 : p_{ii}^{(t)} > 0\right\}$ je 1)

U nastavku predstavljamo lance Markova za kontinualni prostor stanja.

5.2.2 Kontinualni prostor stanja

Pretpostavimo da imamo stohastički proces $(X_0, X_1, X_2, \dots)$sa disktretnim vremenima ali sa kontinualnim prostorom stanja $S \subseteq \mathbb{R}^d$ i da sve raspodele imaju gustine.

Za lanac Markova sa kontinualnim prostorom stanja tranzicione verovatnoće $P(X_{t+1} = x_{t+1} \mid X_t = x_t)$ su uvek jednake nuli. Prema tome za određenu tačku $x \in S$, definisanje trazicionih verovatnoća nema smisla. Zbog toga posmatraju se podskupovi $A \subseteq S$. To dovodi do sledeće definicije:

Definicija 2

Neka je $(X_0, X_1, X_2, \dots)$ stohastički proces sa kontinualnim prostorom stanja $S \subseteq \mathbb{R}^d$. Ako za sve $A \subseteq S$ i za sva stanja $x_0, x_1, \dots, x_t \in S$ važi

$$P(X_{t+1} \in A \mid X_t = x_t, X_{t-1} = x_{t-1}, \dots, X_1 = x_1, X_0 = x_0) = P(X_{t+1} \in A \mid X_t = x_t),$$

Stohastički process nazivamo lancem Markova sa kontinualnim prostorom stanja.

Pretpostavka je da možemo odrediti tranzicione verovatnoće

$P(X_{t+1} \in A \mid X_t = x_t)$ koristeći tranzicioni kernel $K: S \times S \to \mathbb{R}_{\geq 0}$ zadat sa

$$P(X_{t+1} \in A \mid X_t = x_t) = \int_{x_{t+1} \in \mathrm{A}} K(x_t, x_{t+1}) dx_{t+1}.$$

Tranzicioni kernel ima sledeća svojstva

- $K(x_t, x_{t+1}) \geq 0$ for all $x_t, x_{t+1} \in S$
- $\int_{x_{t+1} \in \mathrm{S}} K(x_t, x_{t+1}) dx_{t+1} = 1.$

Može se pokazati da je dvo-koračna tranziciona verovatnoća data sa

$$\begin{aligned} P(X_2 \in A \mid X_0 = x_0) \quad &= P(X_2 \in A, X_1 \in S \mid X_0 = x_0) \\ &= \int_{x_2 \in A} \int_{x_1 \in S} K(x_0, x_1) K(x_1, x_2) dx_1 dx_2. \end{aligned}$$

Dakle dvo-koračni tranzicioni kernel je

$$K^{(2)}(x_0, x_2) = \int_{x_1 \in S} K(x_0, x_1) K(x_1, x_2) dx_1.$$

Ovo možemo generalizovati na T-koračne tranzicije (višekoračne tranzicije).

$$P(X_T \in A \mid X_0 = x_0) =$$

$$\int_{x_T \in A} \int_{x_{T-1} \in S} \dots \int_{x_1 \in S} K(x_0, x_1) K(x_1, x_2) \dots K(x_{T-1}, x_T) dx_1 \dots dx_{T-1} dx_T.$$

Kako imao T-koračni tranzicioni kernel

$$K^{(T)}(x_0, x_T) = \int_{x_{T-1} \in S} \dots \int_{x_1 \in S} K(x_0, x_1) K(x_1, x_2) \dots K(x_{T-1}, x_T) dx_1 \dots dx_{T-1}$$

Imamo i verovatnoću $P(X_T \in A \mid X_0 = x_0) = \int_{x_T \in A} K^{(T)}(x_0, x_T) dx_T.$

Chapman-Kolmogorov lema takođe važi i za kontinualni prostor stanja

$$K^{(T+S)}(x_0, x_{T+S}) = \int_{x_T \in S} K^{(T)}(x_0, x_T) K^{(S)}(x_T, x_{T+S}) dx_T$$

Koncept ireducibilnog Markovljevog lanca (kao što se razmatra pod diskretnim prostorima stanja) je isti za neprekidni prostor stanja. Dakle, definicije rekurentnih i prolaznih

Markovljevih lanaca, komunikacionih i aperiodičnih takođe se primenjuju na kontinuirani slučaj.

Sada se razmatra koncept stacionarne raspodele za Markovljev lanac sa neprekidnim prostorom stanja.

Pretpostavimo da za lanac Markova $(X_0, X_1, X_2, \dots)$ sa tranzicionim kernelom K, raspodela od X_t ima gustinu p_t. Ako je $p_{t+1}(x_{t+1})$ gustina od X_{t+1} tada:

$$p_{t+1}(x_{t+1}) = p(x_{t+1} \mid x_t)p_t(x_t) = p_t(x_t)K(x_t, x_{t+1})$$

Prema tome, za raspodelu od X_{t+1} za svaki $A \subseteq S$ važi:

$$P(X_{t+1} \in A) = \int_{x_{t+1}\in A} p_{t+1}(x_{t+1})dx_{t+1} = \int_{x_{t+1}\in A}\int_{x_t\in S} p_t(x_t)K(x_t, x_{t+1})dx_t dx_{t+1}.$$

Gustina za X_{t+1} je dakle data sa:

$$p_{t+1}(x_{t+1}) = \int_{x_t\in S} p_t(x_t)K(x_t, x_{t+1})dx_t$$

Uvedimo sledeću definiciju:

Definicija 3

Raspodela verovatnoća π sa gustinom p naziva se **stacionarnom raspodelom** za Markovljev lanac $(X_0, X_1, X_2, \dots)$ sa tranzicionim kernelom K ako

$$p(y) = \int_{x\in S} p(x)K(x, y)dx$$

za svako $y \in S$ osim za skupove $A \subseteq S$ gde je $\pi(A) = 0$- skupovi mere nula. Ova raspodela se naziva invarijantna raspodela. Sada prelazimo na pojam **zvani uslov detaljnog balansa (detailed balance condition)**

Lema 2

Neka je $(X_0, X_1, X_2, \dots)$ Markovljev lanac sa tranzicionim kernelom K. Ako za funkciju gustine, p, važi uslov detaljnog balansa:

$$p(y)K(y, x) = p(x)K(x, y) \text{ za sve } x, y \in S$$

Tada je p funkcija gustine stacionarne raspodele Markovljevog lanca.

Dokaz:

Imamo da važi:

$$\int_{x\in S} p(x)K(x, y)dx = \int_{x\in S} p(y)K(y, x)dx = p(y)\int_{x\in S} K(y, x)dx = p(y).$$

Definicija

Neka je $(X_0, X_1, X_2, \dots)$ Markovljev lanac sa kontinualnim prostorom stanja S. Neka je v funkcija raspodele na S. Markovljev lanac se naziva i **v-ireducibilan** ako za sve $x_0 \in S$ i za sve $A \subseteq S$ gde je $v(A) > 0$ postoji $T \in \mathbb{N}$ tako da je

$$P(X_T \in A \mid X_0 = x_0) = \int_{y \in A} K^{(T)}(x_0, y)dy > 0.$$

Ako je $T = 1$ Markovljev lanac se naziva **jako v-ireducibilan** (strongly v-irreducible). Osobina lanca Markova implicira da svaki skup sa pozitivnom verovatnoćom $v(A) > 0$ može biti posećen iz bilo koje tačke $x_0 \in S$ u konačnom vremenu. Dakle ako ova osobina važi sva stanja komuniciraju.

Neka $\eta_A = \sum_{t=1}^{\infty} 1_A(X_t)$ označava broj poseta Markovljevog lanca skupu A.

Definicija 4

Neka je $(X_0, X_1, X_2, \dots)$ Markovljev lanac I neka je $A \subseteq S$. Tada kažemo da je

- Skup A rekurentan ako za sve $x_0 \in A$ važi $E(\eta_A \mid X_0 = x_0) = \infty$.
- Markov chain **rekurentan** ako je v-ireducibilan for za neki raspodelu verovatnoće v i kad god je $v(A) > 0$, tada je A rekurentan.

Data je i jača definicija rekurentnosti

Definicija 5

Neka je $(X_0, X_1, X_2, \dots)$ Markovljev lanac I neka je $A \subseteq S$. Tada kažemo da je

- Skup A Harris-rekurentan ako za sve $x_0 \in A$ važi $P(\eta_A = \infty \mid X_0 = x_0) = 1$.
- Markov lanac Harris-rekurentan ako je v-ireducibilan za neku raspodelu verovtnoće v i kad god je $v(A) > 0$, tada je A Harris-rekuretan.

Lema 3

Neka je $(X_0, X_1, X_2, \dots)$ Markovljev lanac sa stacionarnom raspodelom π (sa gustinom p). Ako $X \sim p$ i ako je lanac Markova π-ireducibilan i rekurentan tada za svaku integrabilnu funkciju $h: S \to \mathbb{R}$ imamo (sa verovatnoćom 1)

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} h(X_t) = \int_S h(x)p(x)dx = E_p(h(X))$$

Za skoro svaku početnu vrednost $X_0 = x_0$. Ako je lanac Markova Harris-rekurentan, tada jednakost važi za sve $x_0 \in S$.

5.5 Markovljevi lanci Monte Karlo (Markov chain Monte Carlo)

Markovljev lanci Monte Karlo metoda konstruiše Markovljev lanac koji za ciljnu raspodelu ima stacionarnu raspodelu. To radi konstruisanjem ireducibilnog Markovljevog lanca, koji osigurava da je većina Markovljevih lanaca koji su rezultat MCMC algoritma rekurentna ili čak Haris-rekurentna. Kao što je objašnjeno, Harisova rekurentnost obezbeđuje da Markovljev lanac konvergira ka stacionarnoj raspodeli za svaku početnu vrednost umesto za skoro svaku početnu vrednost. Stoga nam je potrebna Harrisova rekurentnost da bismo osigurali da MCMC algoritam konvergira. MCMC algoritmi konstruišu tranzicioni kernel koje rezultira Markovljevim lancem koji je rekurentan i konvergira ka ciljnoj raspodeli. Opšti princip za ovo je **Metropolis-Hastings (MH)** algoritam. Gibsov sampler (Gibbs sampler) je poseban slučaj MH algoritma.

5.5.1 Metropolis-Hastings algoritam

Cilj nam je konstruisanje lanca Markova koji ima stacionarnu raspodelu jednaku ciljanoj raspodeli.

Rezultati iz prethodnih poglavlja su korišćeni za konstrukciju tranzicionog kernel, $K(x, y)$, koji ima invarijantnu gustinu jednaku ciljanoj gustini. Razmatramo kontinualni slučaj. Metropolis-Hastings algoritam je uopšteni algoritam za uzorkovanje iz bilo koje posteriorne raspodele.

Metropolis-Hastings (MH) algoritam koristi sledeće leme: lemu 2 i lemu 3. Lema 3 suštinski znači da možemo da uzorkujemo zavisne uzorke iz Markovljevog lanca i da pri tom korsitmo $\frac{1}{T}\sum_{t=1}^{T} h(X_t)$ kao procenu za $E_p(h(X))$.

Koristimo Lemu 2 (**detailed balance condition**). Tranzicioni kernel koji se javlja u ovoj lemi je poznat kao reverzibilni kernel i rezultira stacionarnom raspodelom. Ova lema može poslužiti za nalaženje kernela sa željenom ciljnom raspodelom. Po Chib i Greenberg [9], od ireverzibilnog kreiramo reverzibilni kernel. Ako kernel nije reverzibilan za neki par (x, y) imamo da važi (bez uticaja na opštost):

$$p(x)K(x, y) > p(y)K(y, x).$$

Cilje je od nejednakosti napraviti jednakost. U ovom slučaju postoji više poteza iz x u y nego iz y u x. Da bismo postigli jednakost, pre nego što napravimo potez iz x u y uvodimo verovatnoću $\alpha(x, y) < 1$ po kojoj će takav potez biti prihvaćen. Za verovatnoću $\alpha(x, y)$ mora da važi $\alpha(x, y)p(x)K(x, y) = p(y)K(y, x)$. To znači da je

$$\alpha(x,y) = \min\left(\frac{p(y)K(y,x)}{p(x)K(x,y)}, 1\right)$$

Navedeno osigurava da uslov detaljnog balansa (detailed balance condition) važi.

Ovo dovodi do opisa Metropolis-Hastings algoritma:

- 1. Bira se tranzicioni kernel q gde važi $q(x,y) > 0$ za sva stanja $x, y \in S$;
- 2. Početak je za $t = 0$ iz proizvoljnog stanja $X_t = x_t \in S$;
- 3. Ako je $X_t = x_t$, generisati slučajni promenljivu $Y \sim q(x_t, *)$ i $U \sim$ unif (0,1);
- 4. Ako je $Y = y$ (i $X_t = x_t$) definišemo

$$X_{t+1} = \begin{cases} y & \text{ako je } U \leq \alpha(x_t, y) - prihvatanje\ (\boldsymbol{accept})\ novog\ stanja \\ x_t & \text{u suprotnom} \quad - odbijanje\ (\boldsymbol{reject})\ novog\ stanje \end{cases}$$

- 5. **Inkrementacija**: $t = t + 1$, zatim povratak na tačku 3.

Ovde važi da je $\alpha(x,y) = \min\left(\frac{p(y)q(y,x)}{p(x)q(x,y)}, 1\right)$.

Metropolis-Hastings algoritam je fundamentalni algoritam koji se koristi za Bejzovu logističku regresiju.

Tranzicioni kernel, q se naziva i kernel **predloga** (**proposal kernel**). Postoji velika sloboda u izboru kernela predloga. Međutim, i dalje treba voditi računa da biste izabrali one koji su zapravo korisni. Na primer, ukoliko kernel predloga ne "pretražuje" ceo prostor stanja od $p(x)$ tada se odeđene vrednosti ne mogu uzorkovati. Dva su uobičajena izbora za kernel predloga: **independence sampler** i **random walk sampler**.

Izbor kernela predloga utiče na stope prihvatanja algoritma (*accept/reject*). Prema Ntzoufrasu (2009), [17] varijansa kernel predloga kontroliše brzinu konvergencije algoritma. Male varijacije kernela predloga će rezultirati visokim stopama prihvatanja, ali niskom konvergencijom pošto će algoritmu trebati veliki broj iteracija da bi istražio ceo prostor parametara. Suprotno tome, velika varijansa će rezultirati niskim stopama prihvatanja i visoko koreliranim uzorkom. Optimalna stopa prihvatanja je između 20% i 40% [17]. Za modele sa velikim brojem parametara stopa prihvatanja treba da bude prema donjoj granici, za univarijantni model stopa prihvatanja treba da bude prema gornjoj granici. Način da se dobije stopa prihvatanja u ovom opsegu je podešavanjem varijanse kernel predloga. Metropolis-Hastings algoritmi uključuju parametar podešavanja. Ovaj parametar je "podešen" (**tuned**) tako da je stopa prihvatanja između 20-40%.

Independence sampler

Ako predlog kernel $q(x,y)$ ne zavisi od y, to jest $q(x,y) = g(x)$ za svako x tada je verovatnoća prihvatanja:

$$\alpha(x,y) = \min\left(\frac{p(y)g(x)}{p(x)g(y)}, 1\right)$$

Independence sampler je veoma sličan **accept-reject** metodi u poglavlju 5.3.1.2. Kao i kod accept-reject methode, važno je da je kernel predloga g, blizu ciljanoj raspodeli f da bi došlo do efikasne simulacije. Kakogod, independence sampler proizvodi zavisne uzorke. Takođe, ako postoji konstanta M takva da $p(x) \leq Mg(x)$, tada je očekivana stopa prihvatanja (acceptance rate) najmanje $1/M$ kada je lanac Markov stacionaran. Sledi dokaz:

$$\begin{aligned}
E\big(\alpha(X_t, Y)\big) = \quad & E\left(\min\left(\frac{p(Y)g(X_t)}{p(X_t)g(Y)}, 1\right)\right) = \iint \min\left(\frac{p(y)g(x)}{p(x)g(y)}, 1\right) p(x)g(y)dxdy \\
= \quad & \iint_{(x,y):\frac{p(y)g(x)}{p(x)g(y)}>1} p(x)g(y)dxdy \\
& + \iint_{(x,y):\frac{p(y)g(x)}{p(x)g(y)}\leq 1} \frac{p(y)g(x)}{p(x)g(y)} p(x)g(y)dxdy \\
= \quad & \iint_{(x,y):\frac{p(y)g(x)}{p(x)g(y)}>1} p(x)g(y)dxdy + \iint_{(x,y):\frac{p(y)g(x)}{p(x)g(y)}\leq 1} p(y)g(x)dxdy \\
= \quad & 2\iint_{(x,y):\frac{p(y)g(x)}{p(x)g(y)}\geq 1} p(x)g(y)dxdy \\
\geq \quad & \frac{2}{M}\int_{(x,y):\frac{p(y)g(x)}{p(x)g(y)}\geq 1} p(x)p(y)dxdy \\
= \quad & \frac{2}{M}\int_{(x,y):\frac{p(y)}{g(y)}\geq\frac{p(x)}{g(x)}} p(x)p(y)dxdy.
\end{aligned}$$

Ako je $h(x) = \frac{p(x)}{g(x)}$, poslednji integral je $P(h(U) \geq h(V))$ za U, V nezavisne sa raspodelom $p(x)$. Prema tome, jednako je verovatno da $h(U) \geq h(V)$kao i da $h(V) \geq h(U)$ ako s U, V nezavisne i sa identičnom raspodelom. Prema tome,

$$E(\alpha(x,y)) \geq \frac{2}{M}\frac{1}{2} = \frac{1}{M}$$

Random walk sampler

Drugi uobičajeni izbor je korišćenje trenutne simulirane vrednosti za generisanje sledeće vrednosti. Na taj način se istražuje susedstvo trenutne vrednosti Markovljevog lanca. Kernel predloga koje ovo dozvoljava je simetrični kernel $q(x,y) = q(y,x)$. Ovo dovodi do verovatnoće prihvatanja

$$\alpha(x,y) = \min\left(\frac{p(y)}{p(x)}, 1\right)$$

Predložena vrednost $Y = y$ je tada prihvaćena sa verovatnoćom 1 ako je $p(y) \geq p(x)$. Prema tome, tačke koje su više 'verovatne' nego prethodna tačka x_t će uvek biti prihvaćene. Međutim, prihvataju se i tačke koje su manje 'verovatne'sa određenom verovatnoćom.

MCMC dijagnostika

Markovljev lanac, iz kojeg uzimamo uzorke, treba da konvergira ciljnoj distribuciji. Ako se Markovljev lanac nije konvergirao, biće uzeti uzorci koji nisu iz željene ciljne distribucije. Da bi se osiguralo da se uzorci uzimaju samo iz stacionarne distribucije, koristi se **period sagorevanja** (**burn – in period**) [17]. Period sagorevanja je broj uzoraka koji se eliminišu da bismo obezbedili da uzorkujemo samo iz stacionarne distribucije. Dobson i Barnett [5] navode da je način za procenu konvergencije Markovljevog lanca posmatranjem dijagrama vremenskih serija (tragova - **trace**). Ovi grafikoni prikazuju istoriju Markovljevog lanca. Lanac koji je konvergirao treba da bude stabilan i da pokazuje razuman stepen slučajnosti između iteracija.

Drugi način za procenu konvergencije je posmatranje autokorelacione funkcije (**ACF**) lanca. U idealnom slučaju bismo želeli da uzorci budu nezavisni, ali sa MCMC algoritmima to se ne može dogoditi. Stoga prihvatamo neku autokorelaciju. Ako su ACF vrednosti niske, to ukazuje da je Markovljev lanac uspešno konvergirao.

Geveke (1992) je predložio dijagnostički test za procenu konvergencije srednje vrednosti svakog parametra. On posmatra simulirani Markovljev lanac (dobijen iz MCMC izlaza) kao vremensku seriju i primenjuje z-test da bi proverio da li su srednje vrednosti iz dva različita poduzorka jednake. Ovi poduzorci dolaze sa početka i kraja generisanog lanca. Obično se prvih 10% lanca koristi kao početni uzorak, a poslednjih 50% se koristi kao krajnji uzorak. Koristeći ovaj z-test, parametri sa $|z|>2$ ukazuju na dokaze o značajnim razlikama između srednjih vrednosti prvog i poslednjeg skupa iteracija i nekonvergencije lanca u smislu srednjih vrednosti.

R-hat metrika, takođe poznata kao "Gelman-Rubin" statistika ili "potvrda konvergencije" (convergence diagnostic), takođe sekoristi se za procenu konvergencije Markov Chain Monte Carlo (MCMC) lanaca u Bejzovskim statističkim modelima. Ova metrika je posebno korisna kada koristite više nezavisnih MCMC lanaca kako bi se procenila posteriorna raspodela parametara u modelima. Osnovna ideja R-hat metrike je da uporedi varijaciju između lanaca (unutar-varijabilnost) sa varijacijom unutar svakog pojedinačnog lanca. Ako su lanci konvergirali ka istoj posterior raspodeli, očekuje se da će njihova varijacija biti slična varijaciji unutar svakog pojedinačnog lanca. Ako se lanci ne konvergiraju, može doći do značajnih razlika u varijaciji između lanaca. Konkretno, R-hat metrika se izračunava tako što se poredi odnos između ukupne varijacije između lanaca i prosečne varijacije unutar lanaca. Ako su lanci konvergirali, R-hat će težiti ka vrednosti 1. Veće vrednosti R-hat-a ukazuju na nedostatak konvergencije.

Kada se koristi R-hat metrika, preporučuje se da se vrednosti R-hat-a blizu 1 smatraju dobrim indikatorom konvergencije. Međutim, nema jasnog konsenzusa o tačnoj granici ispod koje se

smatra da su lanci konvergirali, pa je važno uzeti u obzir kontekst i proučiti rezultate MCMC analize kako bi se bolje razumela konvergencija.

R-hat metrika je posebno korisna u Bejzovim analizama jer omogućava da procenu da li se MCMC lančani uzorci uzorkuju iz stabilne posterior raspodele, što je ključno za pouzdano zaključivanje o parametrima i njihovim intervalima poverenja.

6 Izrada modela

Metodologija izrade modela

Praktični deo ovog rada ima za cilj da ispita mogućnost korišćena principa Bejzovog zaključivanja na kreiranje bankarskih skoring modela. Model koji razvijamo u ovom radu je aplikativni skoring model za gotovinske (keš) kredite.

Cilj je kreirati model primenom standardne logističke regresije (što predstavlja pristup koji većina banaka i dalje koristi pre svega zbog interpreatbilnosti modela) i uporediti takav model sa modelom Bejzove logističke regresije. U narednom delu će biti predstavljeni i sami realni bankarski podaci dobijeni od jedne domaće komercijalne banke.

Postavka modela je sledeća. Ceo skup podataka ima vremensku komponentu tj. datum kada je aplikacija za gotovinski kredit nastala. Pretopstavka je da je banka u jednom trenutku poslujući na istom tržištu akvizirala jednu ili više drugih komercijalnih banki. Jasno je da ovo dovodi do izmene samog portfolija banke. Pretpostavka je da je ‘stara banka’ imala aplikativni skoring model za gotovinske kredite koji je i dalje u upotrebi i nakon akvizicije nove banke. Kada je prošlo dovoljno dugo vremena, tako da se mogu posmatrati performanse klijenata kojima su odobreni plasmani banka ima sledeću dilemu.

Menadžment “nove banke” razmatra kreiranje novog modela standardne logističke regresije na podacima nakon integracije kao jednu opciju i alternativno kreiranje modela pomoću Bejzove logističke regresije koji će koristiti priorna znanja o koeficijentima modela iz modela korišćenog u staroj banci da bi se na osnovu samih podataka dobije posteriorne raspodele parametara. Osnovni cilj našeg istraživanja je uporedti ova dva pristupa. Na testnom skupu koji je deo potfolija, porediće se performanse razvijenih modela. Bitno je napomenuti da su iako različiti, portfoliji stare i nove banke dovoljno slični, pre svega zbog zajedničkih klijenata, a onda i zbog istog geografskog reona na kome banke posluju. Ova pretpostavka je važna za pomenutu ideju priornih znanja o koeficijentima logističkog modela.

6.1 Predstavljanje podataka

Celokupni skup podataka, kao što smo pomenuli predstavlja podatke o aplikacijama za gotovinske kredite fizičkim licima. U pitanju su odobreni gotovinski krediti čiji je vremenski opseg datuma odobrenja od 03.05.2019 do 28.02.2022. Radi se o **229,865** kredita u pomenutom vremenskom periodu.

Pomenimo i da skup sadrži **4,928** instancu minorne klase zavisne pormenljive, tj. toliko aplikacija za gotovinski kredit je rezultovalo kašnjenjem od 90 ili više dana u materijalno značajnom iznosu u periodu od 12 meseci nakon odobrenja plasmana.

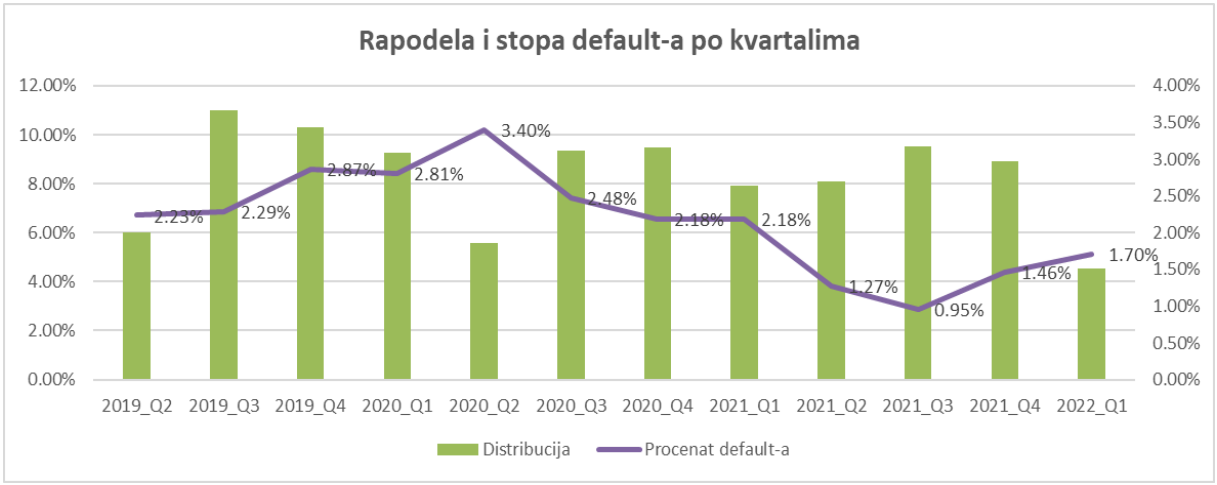
U nastavku je tabela u kojoj je dat opis korišćenim promenljivih. Pored promenljivih korišćenih za modeliranje i zavisne promenljive, prisutna je i pormenljiva APPR_DT koja predstavlja datum odobrenja kredita.

Promenljiva	Tip promenljive	Opis promenljive
APPR_DT	Datum	Datum odobrenja kredita
NEW_CLIENT	Binarna	Oznaka da li je klijent u trenutku aplikacije bio novi ili postojeći klijent banke
INCM	Neprekidna	Prosečna primanja u poslednja 3 meseca u trenutku aplikacije izražena u dinarima
INCM_RATE	Neprekidna	Odnos prosečnog prihoda u poslednja 3 meseca u trenutku apliciranja i iznosa kredita za koji klijent aplicira
AGE	Neprekidna	Broj godina klijenta u trenutku aplikacije
GENDER	Binarna	Oznaka pola klijenta
EDUCATION_LEVEL	Ordinalna	Nivo obrazovanja klijenta
WORK_EXP	Neprekidna	Radno iskustvo klijenta izraženo u godinama. (U slučaju pezionera iznosi 40)
LN_NUM	Neprekidna	Ukupan broj kreditnih proizvoda (kredit, kreditna kartica, dozvoljeni minus) klijenta u istoriji banke
CB_RQST_30D_CNT	Neprekidna	Ukupan broj zahteva za izveštaj kreditnog biroa u poslednjih 30 dana
CB_LN_REPAID_CNT	Neprekidna	Ukupan broj otplaćenih kredita u KB izveštaju
CB_LN_PD_MAX_AMT	Neprekidna	Maksimalni iznos u kašnjenju po kreditima u istoriji u izveštaju KB
CB_LN_PD_DIFF_CNT	Neprekidna	Ukupan broj kredita u kašnjenju u istoriji u izveštaju KB
CB_LNCC_PD_3Y_M_FLG	Binarna	Oznaka da li je klijent u poslednje 3 godine od datuma izveštaja kreditnog biroa bio u materijalno značajnom kašnjenju (iznos

		preko 1000 RSD) po kreditima i kreditnim karticama
CB_PD_MAX_M_CNT	Neprekidna	Maksimalan broj dana kašnjenja u materijalno značajnom kašnjenju (iznos preko 1000 RSD) sa dospelim iznosom većimu istoriji u izveštaju KB
CB_PD_MAX_M_AMT	Neprekidna	Maksimalni materijalno značajan iznos kašnjenja (iznos preko 1000 RSD) u istoriji u izveštaju KB
DEF_FLG	Binarna	Oznaka da li je klijent dostigao 90 dana kašnjenja u materijalno značajnom iznosu (iznos preko 1000 RSD) u prvih 12 meseci od odobrenja kredita

Tabela 3- Opis korišćenih promenljivih

Prosečna stopa default-a iznosi **2.14**% na celokupnom skupu. Sledi predstavljanje kretanja stope default-a po kvartalima. Primetno je da od drugog kvartala 2020-te godine stopa defaulta umereno opada (Slika 22).



Slika 22 – Prikaz stope default-a po kvartalima

Izvor: Originalni grafik autora

Napomenimo da je za potrebe modeliranja i primene logističke regresije neophodno da sve promenljve imaju numeričku reprezentaciju.

Sledi predstavljanje mapiranja promenljvih: NEW_CLIENT, GENDER, EDUCATION_LEVEL, CB_LNCC_PD_3Y_M_FLG, DEF_FLG.

- **NEW_CLIENT**
  0 – U trenutku apliciranja klijent je bio postojeći klijent banke

1- U trenutku apliciranja klijent nije bio klijent banke

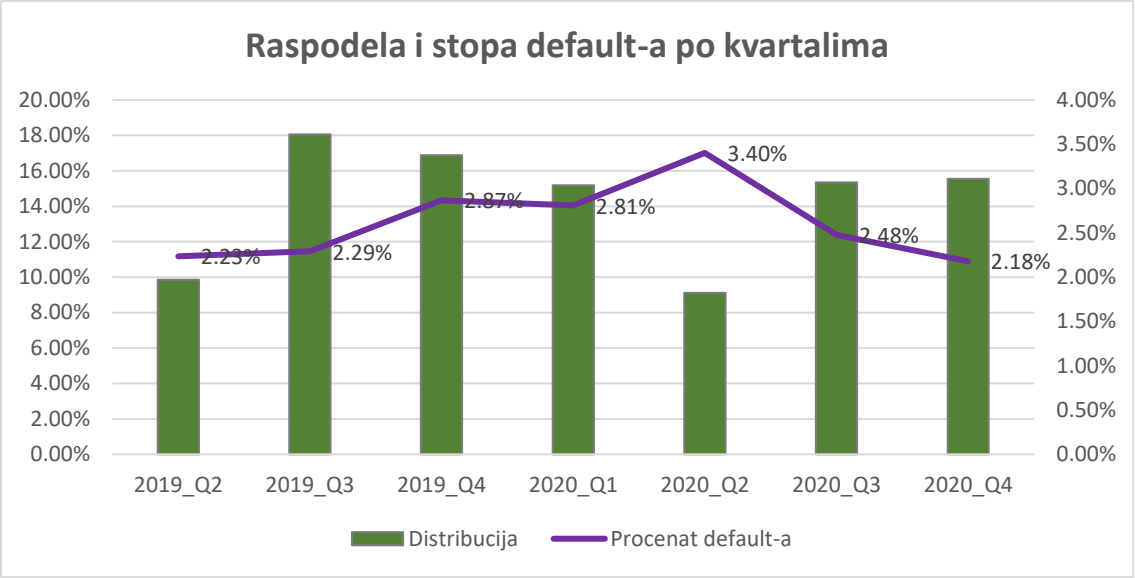
- **GENDER**
  0 – Klijent je ženskog pola
  1 – Klijent je muškog pola

- **EDUCATION_LEVEL**
  1 – Završena osnovna škola
  2 – Završena srednja škola
  3 – Završen fakultet
  4 – Završene master ili doktorske studije

- **CB_LNCC_PD_3Y_M_FLG**
  0 – Klijent u poslednje 3 godine od datuma izveštaja kreditnog biroa nije bio u materijalno značajnom kašnjenju (iznos preko 1000 RSD) po kreditima i kreditnim karticama
  1 – Klijent je u poslednje 3 godine od datuma izveštaja kreditnog biroa bio u materijalno značajnom kašnjenju (iznos preko 1000 RSD) po kreditima i kreditnim karticama
- **DEF_FLG**
  0 – Klijent nije dostigao 90 dana kašnjenja u materijalno značajnom iznosu (iznos preko 1000 RSD) u prvih 12 meseci od odobrenja kredita
  1 – Klijent je dostigao 90 dana kašnjenja u materijalno značajnom iznosu (iznos preko 1000 RSD) u prvih 12 meseci od odobrenja kredita

6.2 Razvoj modela logističke regresije stare banke (Model 1)

Prvobitini model logističke regresije se razvija na podacima stare banke. Ovo su aplikacije iz inicijalnog skupa zaključno sa datumom 31.12.2020. Ovaj skup sadrži **140,232** instance (aplikacije) od koji je **3,609** rezultiralo default-om. Dakle stopa default-a na ovom skupu je **2.57%**. Sledi predstavljanje kretanja stope default-a po kvartalima.



Slika 23 – Prikaz stope default-a po kvartalima na skupu stare banke

Izvor: Originalni grafik autora

Predstavljene su i neke od osnovnih karakteristika promenljivh na posmatranom skupu.

	NEW_CLIENT	INCM	INCM_RATE	AGE	GENDER	EDUCATION_LEVEL
count	140232.0	133415.0	133415.0	140232.0	140232.0	133393.0
mean	0.2	48026.6	0.2	49.3	0.5	2.4
std	0.4	29863.6	0.2	14.5	0.5	0.8
min	0.0	10000.0	0.0	20.0	0.0	1.0
25%	0.0	30780.6	0.1	37.0	0.0	2.0
50%	0.0	40467.0	0.1	48.0	1.0	2.0
75%	0.0	56851.6	0.3	63.0	1.0	3.0
max	1.0	596742.9	4.4	84.0	1.0	4.0

Tabela 4 – Statistički pokazatelji promenljivh na skupu stare banke – 1. deo

	WORK_EXP	LN_NUM	CB_RQST_30D_CNT	CB_LN_REPAID_CNT
count	140232.0	140232.0	139803.0	140232.0
mean	20.4	3.8	0.1	1.7
std	15.2	2.7	0.4	1.7
min	0.0	0.0	0.0	0.0
25%	6.0	1.0	0.0	0.0

50%	18.0	3.0	0.0	1.0
75%	40.0	6.0	0.0	3.0
max	40.0	10.0	13.0	10.0

Tabela 5 - Statistički pokazatelji promenljivh na skupu stare banke – 2. deo

	CB_LN_PD_MAX_AMT	CB_LN_PD_DIFF_CNT	CB_LNCC_PD_3Y_M_FLG	CB_PD_MAX_M_CNT	CB_PD_MAX_M_AMT	DEF_FLG
count	140232.0	140232.0	140232.0	140232.0	140232.0	140232.0
mean	2332.2	0.1	0.0	94.4	4985.9	0.0
std	19180.0	0.4	0.2	346.6	23309.3	0.2
min	0.0	0.0	0.0	0.0	0.0	0.0
25%	0.0	0.0	0.0	0.0	0.0	0.0
50%	0.0	0.0	0.0	0.0	0.0	0.0
75%	0.0	0.0	0.0	0.0	0.0	0.0
max	908038.5	8.0	1.0	5611.0	975205.0	1.0

Tabela 6 - Statistički pokazatelji promenljivh na skupu stare banke – 3. deo

Sledeći korak je provera nedostajućih vrednosti. U pristupu logističke regresije koji koristimo skup na kome se model obučava i skup na kome se model testira ne smeju imati nedostajuće vrednosti. Uočavamo da 4 promenljive sadrže nedostajuće vrednosti (Slika 24).

```
[ ]  #Missing values
     df_old_bank.isna().sum()

     APPR_DT                0
     NEW_CLIENT             0
     INCM                6817
     INCM_RATE           6817
     AGE                    0
     GENDER                 0
     EDUCATION_LEVEL     6839
     WORK_EXP               0
     LN_NUM                 0
     CB_RQST_30D_CNT      429
     CB_LN_REPAID_CNT       0
     CB_LN_PD_MAX_AMT       0
     CB_LN_PD_DIFF_CNT      0
     CB_LNCC_PD_3Y_M_FLG    0
     CB_PD_MAX_M_CNT        0
     CB_PD_MAX_M_AMT        0
     DEF_FLG                0
     dtype: int64
```

Slika 24 – Prikaz nedostajućih vrednosti za skup stare banke

Izvor: Sopstveni proračun autora dobijen korišćenjem programskog koda u Python-u

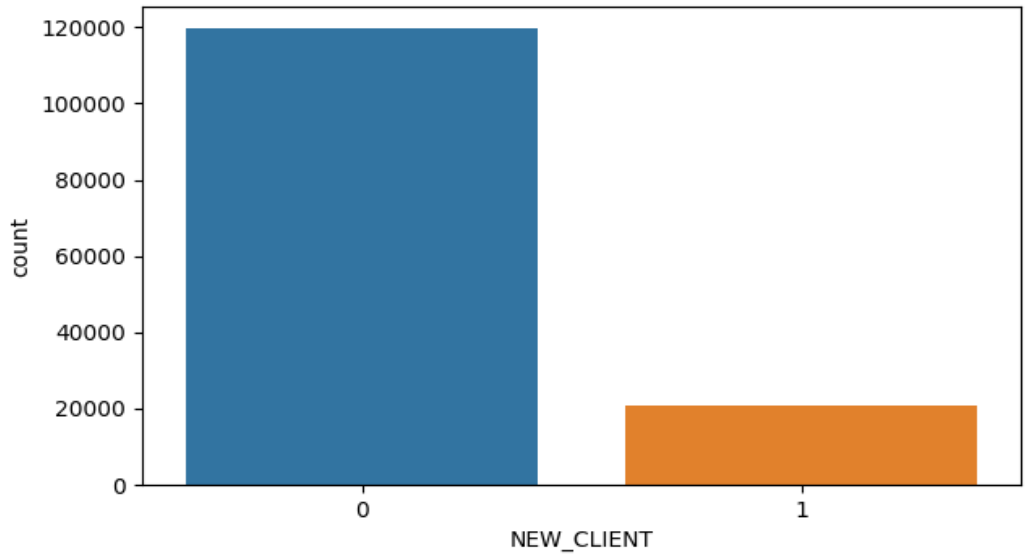
Neretko se dešava da se promenljiva ukoliko sadrži značajan broj nedostajućih vrednosti (više od 25 ili 30%) izbacuje iz daljeg razmatranja. Kako vidimo da to nije slučaj nedostajuće vrednosti ćemo menjati medijanom za neprekidne promenljive odnosno mod-om za kategoričke promenljive.

Za promenljive koje imaju nedostajuće vrednosti one su zamenjene na sledeći način:

- **INCM** nedostajuće vrednosti zamenjene medijanom - 40467.0
- **INCM_RATE** nedostajuće vrednosti zamenjene medijanom - 0.10560243
- **EDUCATION_LEVEL –** nedostajuće vrednosti zamenjene modom – 2
- **CB_RQST_30D_CNT –** nedostajuće vrednosti zamenjene sa medijanom – 0

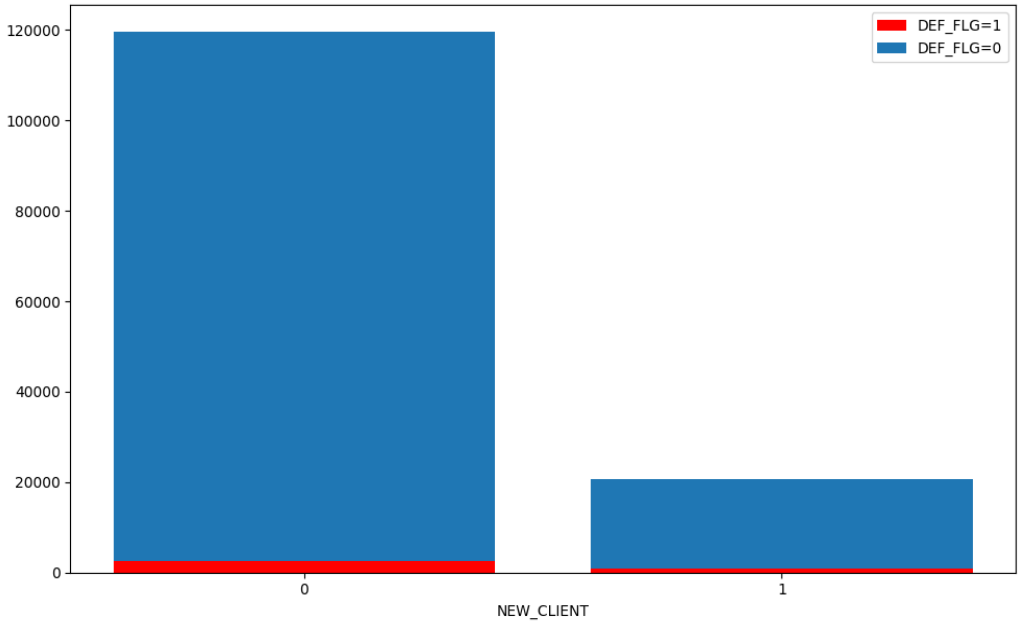
Dalje slede grafička predstavljanja pojedinačnih promenljivih. Za kategoričke promenljive data je raspodela po kategorijama, raspodela po kategorijama sa udelom default- instanci, kao i promena udela kategorija promenljive kroz vreme. Za neprekidne promenljive dat je histogram vrendnosti, sa udelom default instance, kao i box-plot vrednosti promenljve.

- **NEW_CLIENT**



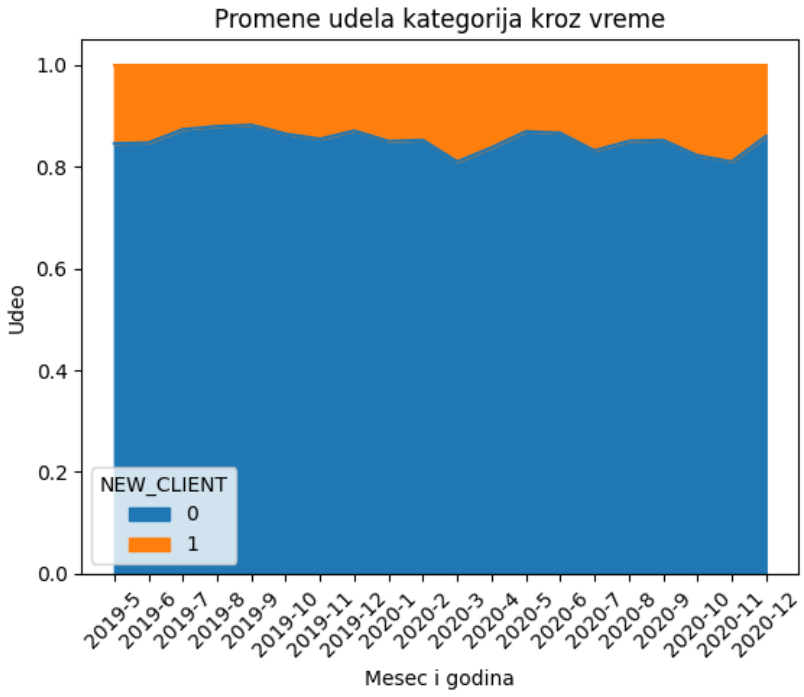
Slika 25 – Bar plot za promenljivu NEW_CLIENT

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u



Slika 26 - Bar plot za promenljivu NEW_CLIENT sa prikazom default klijenata

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

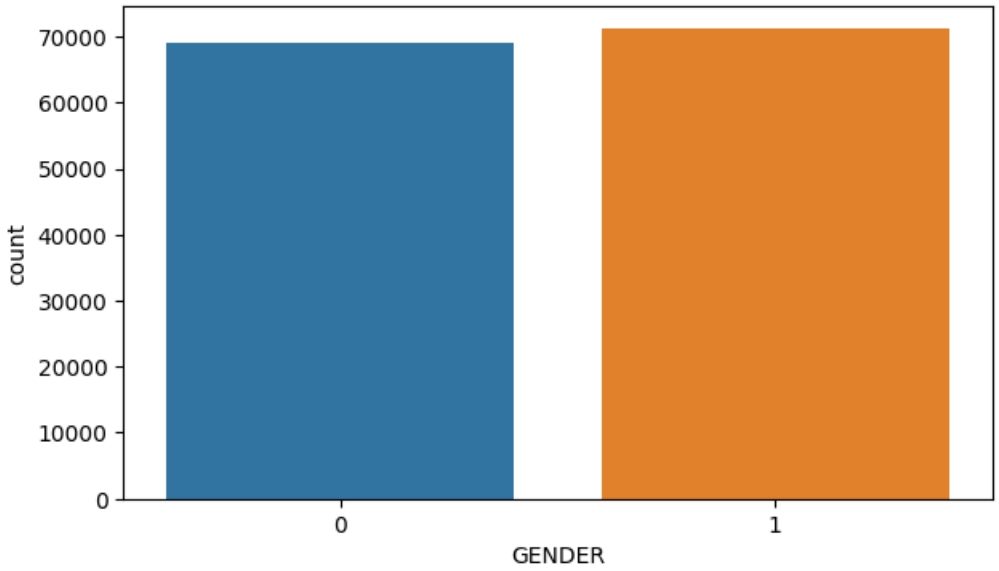


Slika 27 – Prikaz kretanja vrednosti promenljive NEW_CLIENT kroz vreme

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

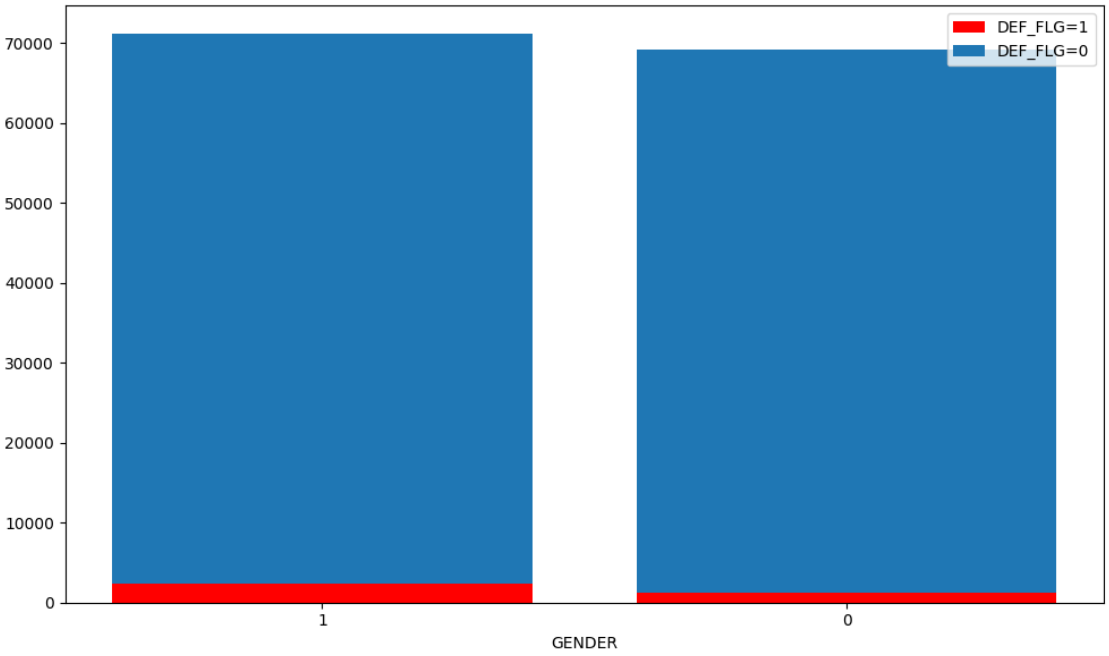
Da se primetiti da je veći udeo klijenata koji su bili postojeći klijenti banke.

- **GENDER**



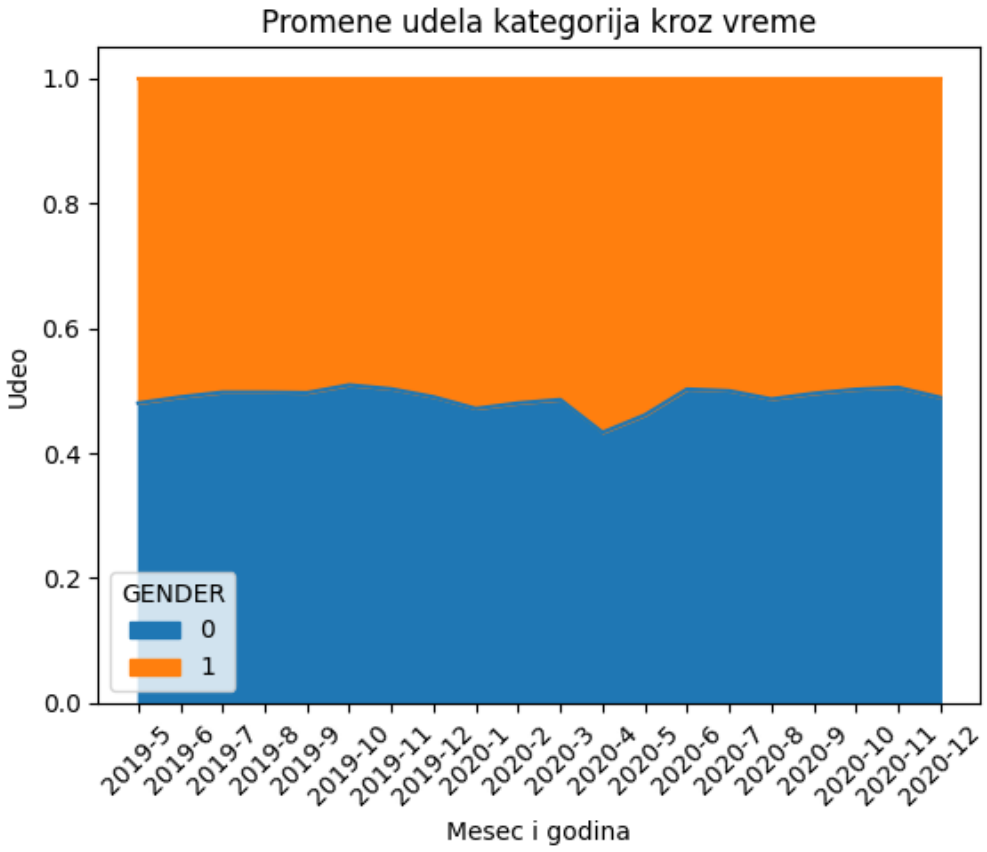
Slika 28 - Bar plot za promenljivu GENDER

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u



Slika 29 - Bar plot za promenljivu GENDER sa prikazom default klijenata

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

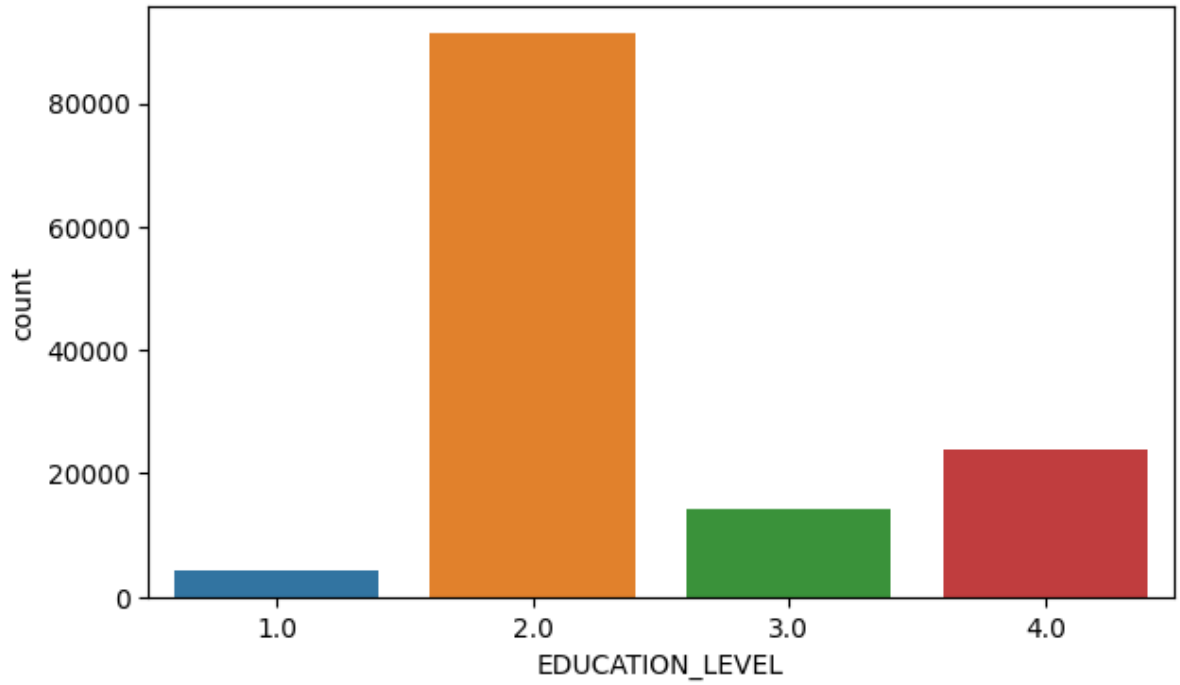


Slika 30 - Prikaz kretanja vrednosti promenljive GENDER kroz vreme

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

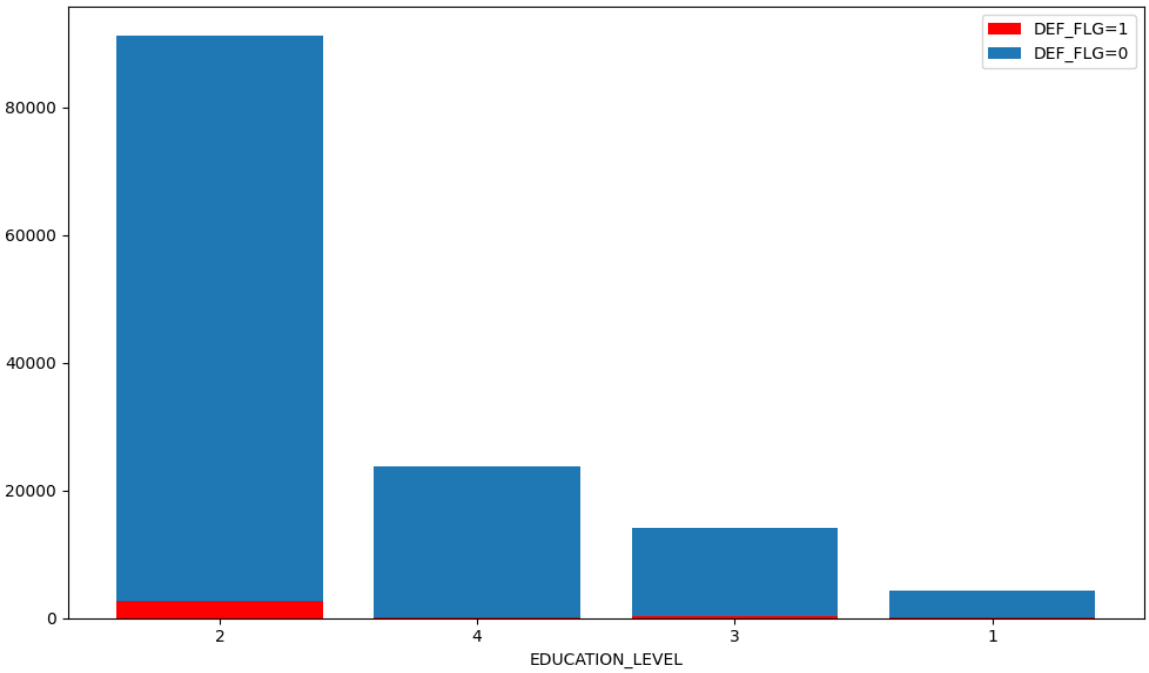
Primetno je da je veću udeo žena, kao i da muškarci predstavljaju rizičniju kategoriju.

- **EDUCATION_LEVEL**



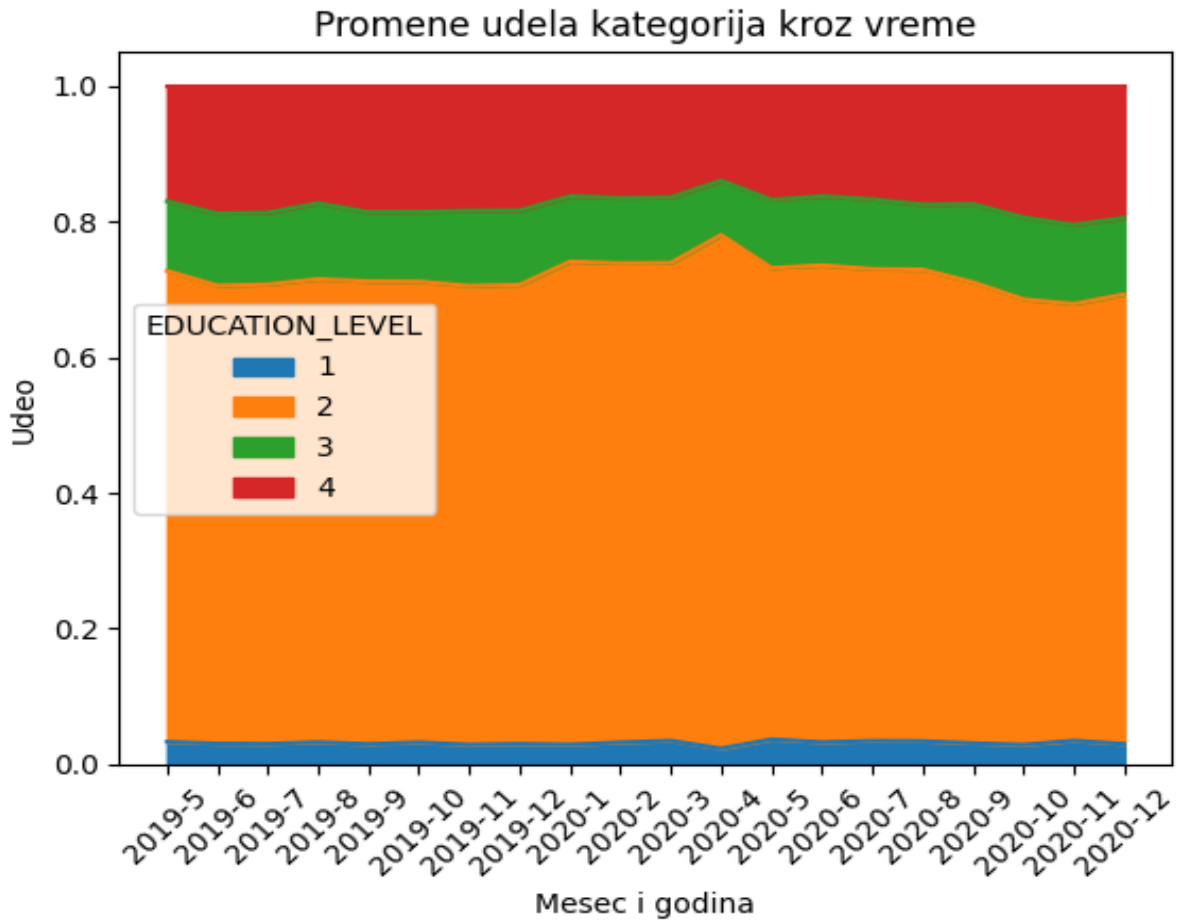
Slika 31 - Bar plot za promenljivu EDUCATION_LEVEL

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u



Slika 32 - Bar plot za promenljivu EDUCATION_LEVEL sa prikazom default klijenata

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

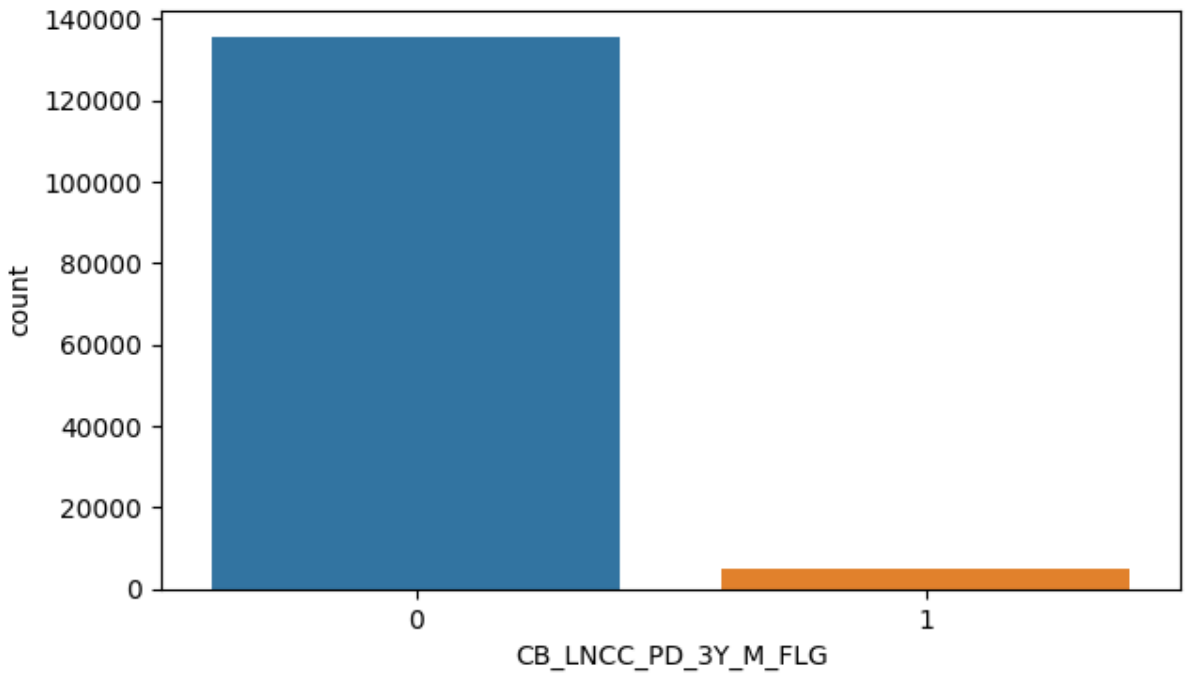


Slika 33 - Prikaz kretanja vrednosti promenljive EDUCATION_LEVEL kroz vreme

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

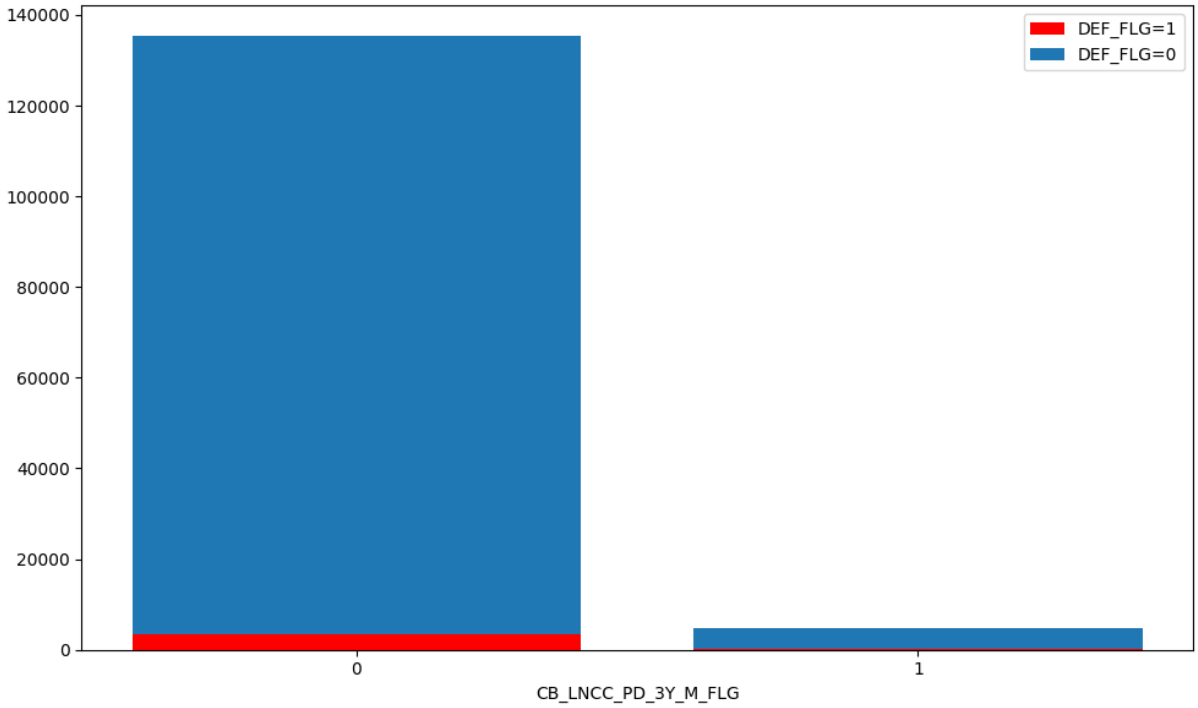
Primetimo da je najveći udeo klijenata sa srednjoškolskim obrazovanjem.

- **CB_LNCC_PD_3Y_M_FLG**



Slika 34 - Bar plot za promenljivu CB_LNCC_PD_3Y_M_FLG

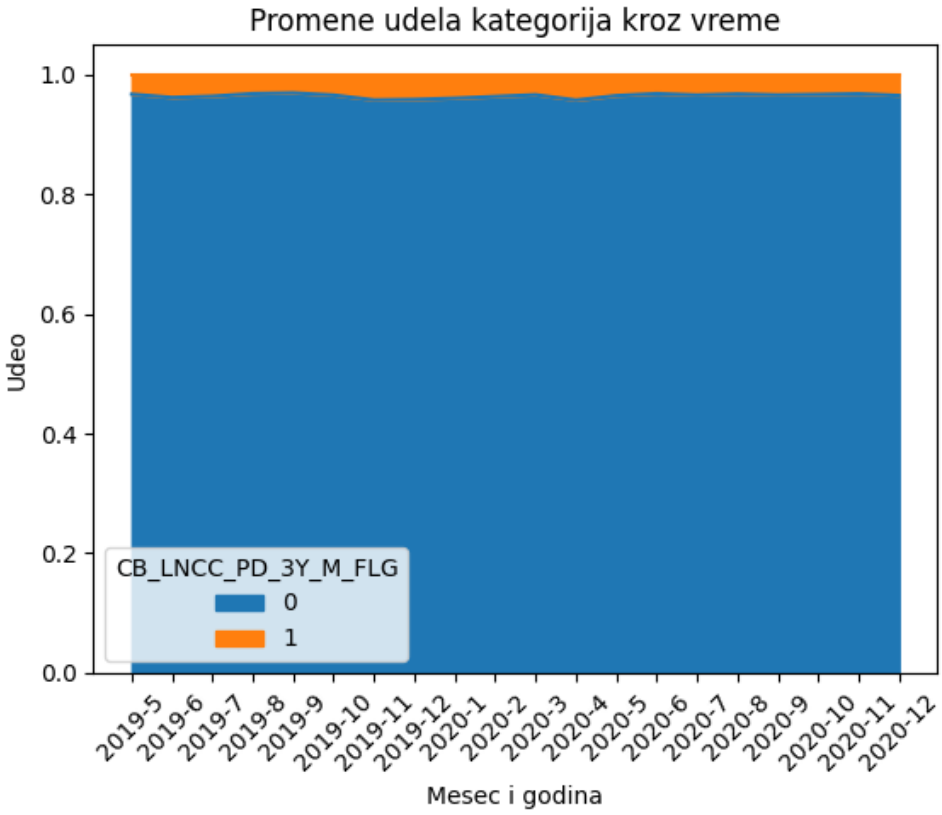
Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u



Slika 35 - Bar plot za promenljivu CB_LNCC_PD_3Y_M_FLG sa prikazom default klijenata

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

Primetno je da većina klijenata nisu imali materijalno značajno kašnjenje po kreditima i kreditnim karticama u poslednje 3 godine.

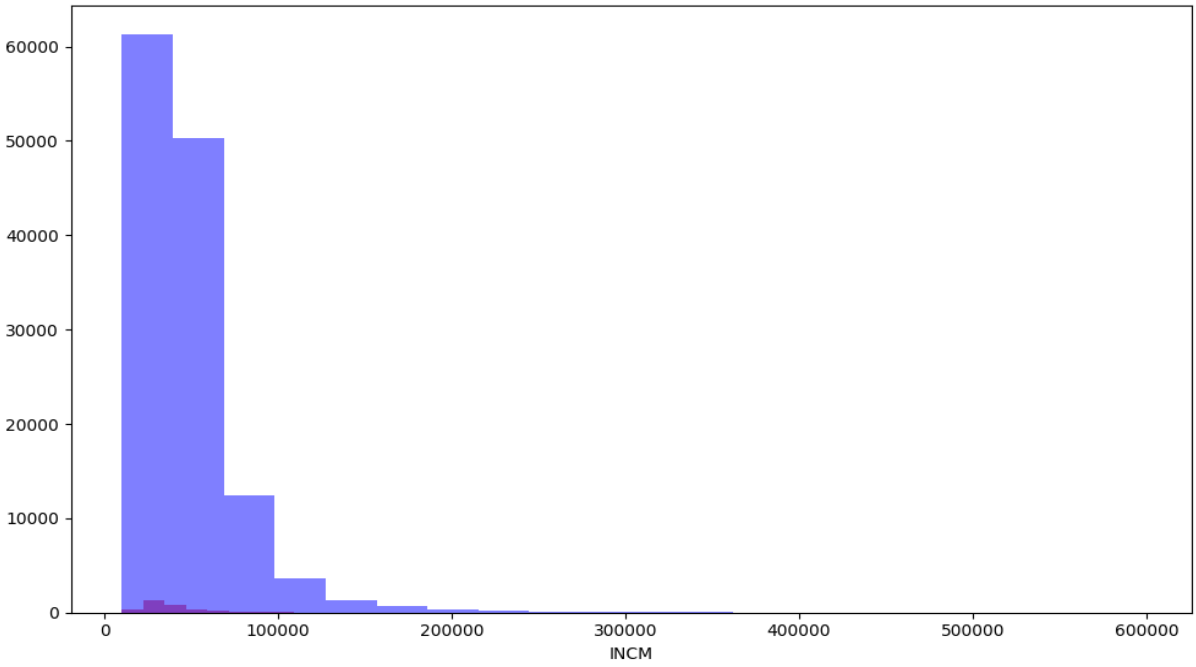


Slika 36 - Prikaz kretanja vrednosti promenljive CB_LNCC_PD_3Y_M_FLG kroz vreme

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

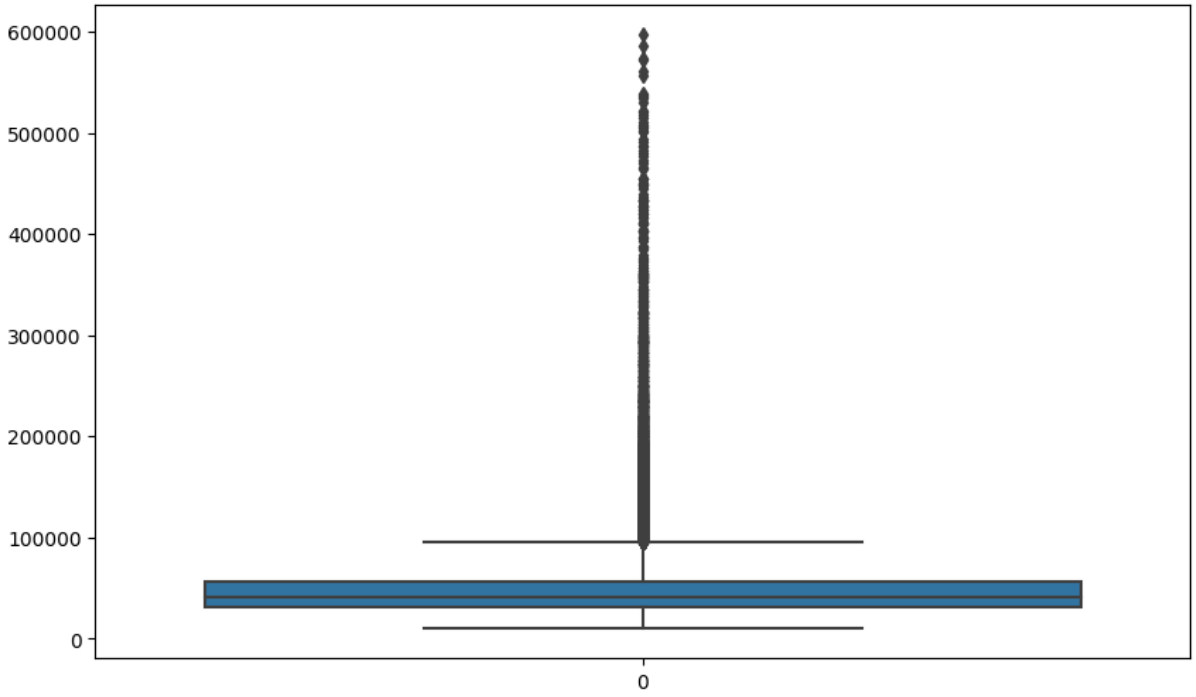
Sledi predstavljanje neprekidnih promenljivih.

- **INCM**



Slika 37- Histogram za promenljivu INCM

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

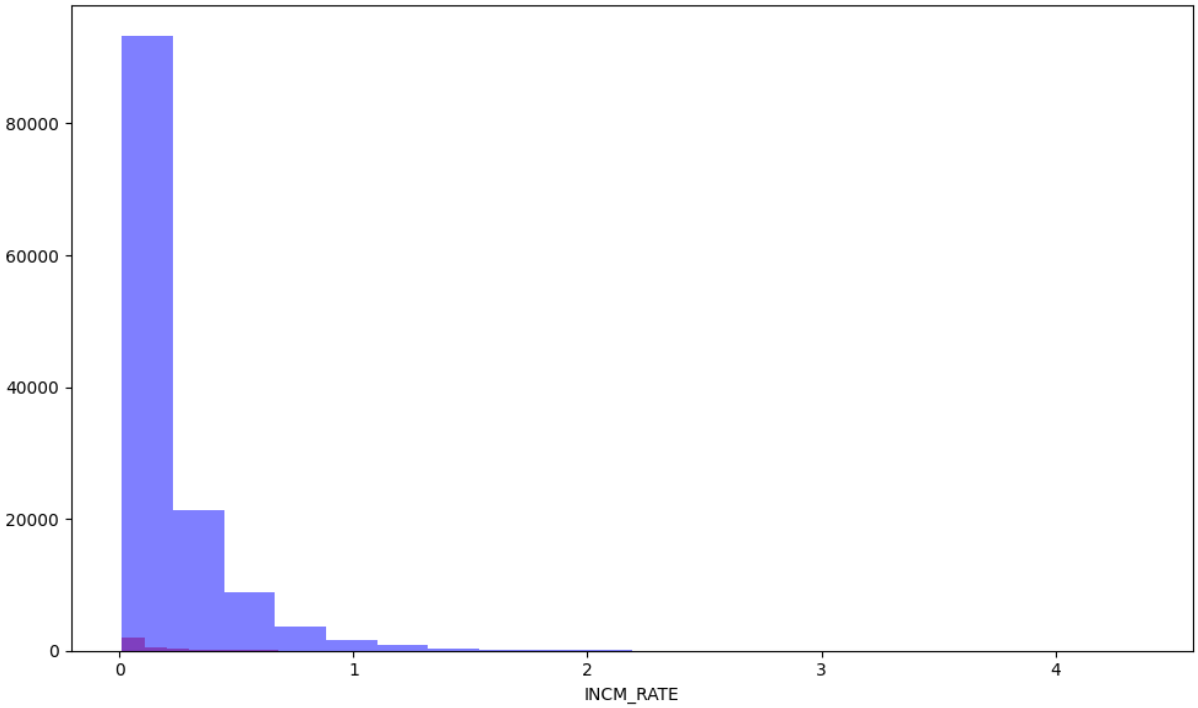


Slika 38 – Box plot za promenljivu INCM

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

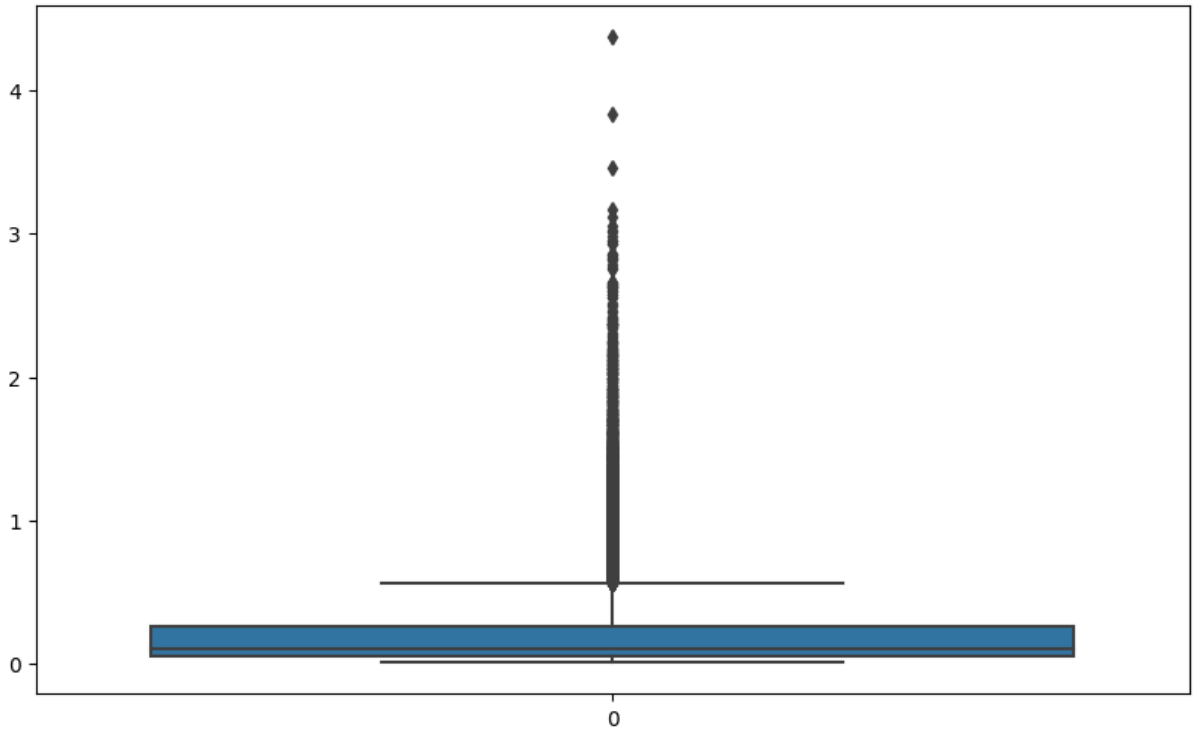
Uočavamo da su klijenti sa nižim primanjima rizičniji što je intuitivno.

- **INCM_RATE**



Slika 39 - Histogram za promenljivu INCM_RATE

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

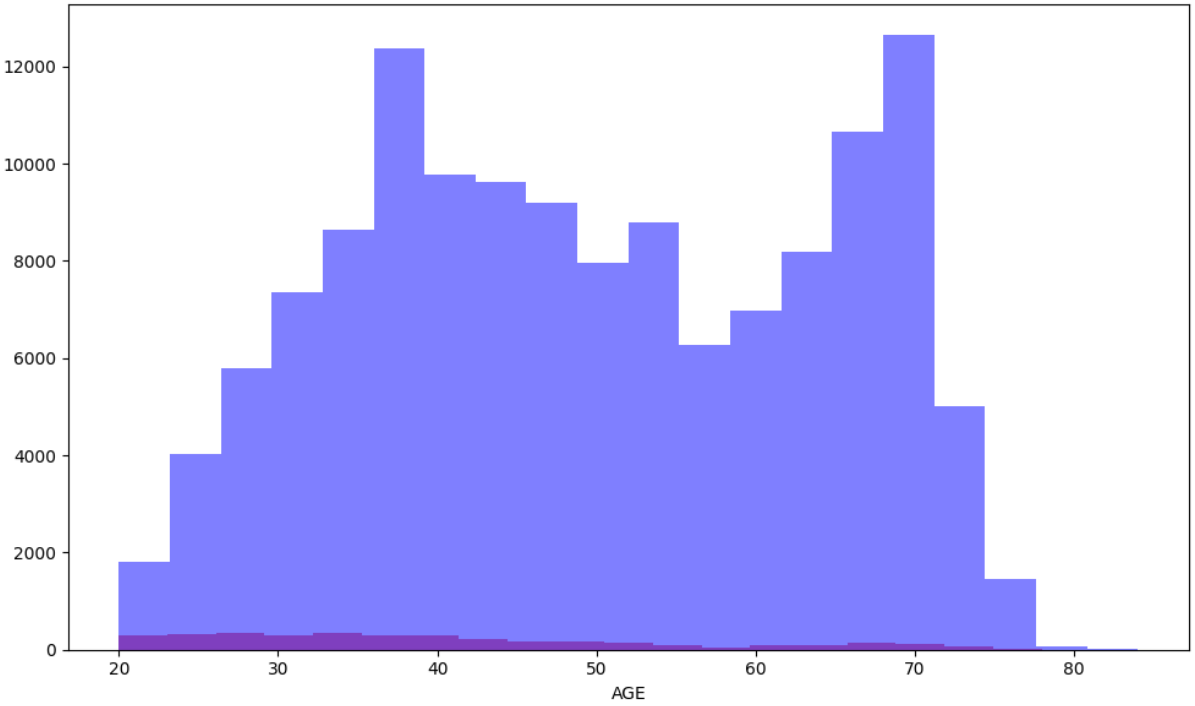


Slika 40 - Box plot za promenljivu INCM_RATE

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

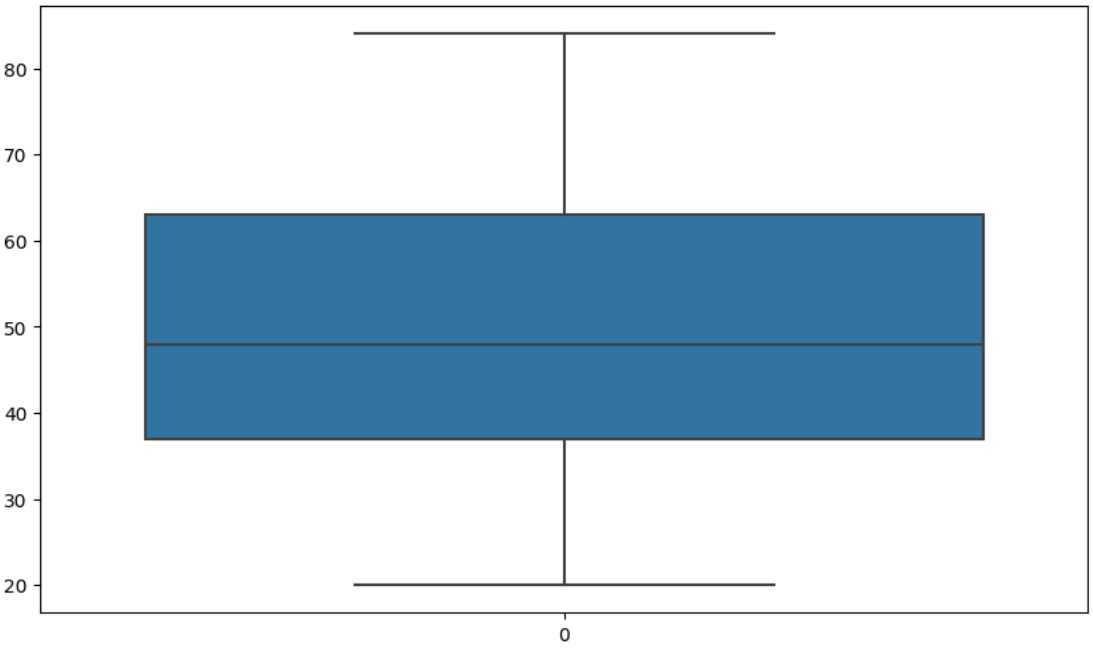
Primetimo da su klijenti čiji je odnos primanja i iznosa manji rizičniji što je za očekivati.

- **AGE**



Slika 41 - Histogram za promenljivu AGE

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u



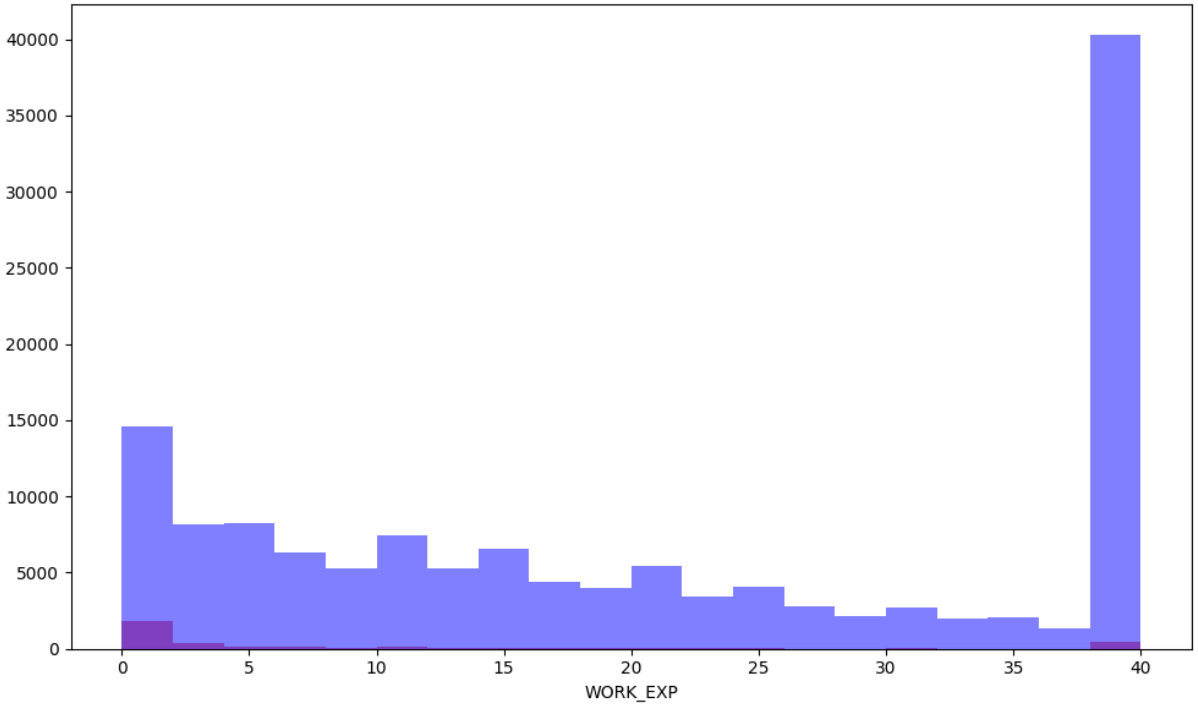
Slika 42 - Box plot za promenljivu AGE

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

Iz grafičkih prikaza ove promenljive uočava se veća rizičnost mlađih klijenata, što je i potvrđeno u bankarskoj praksi.

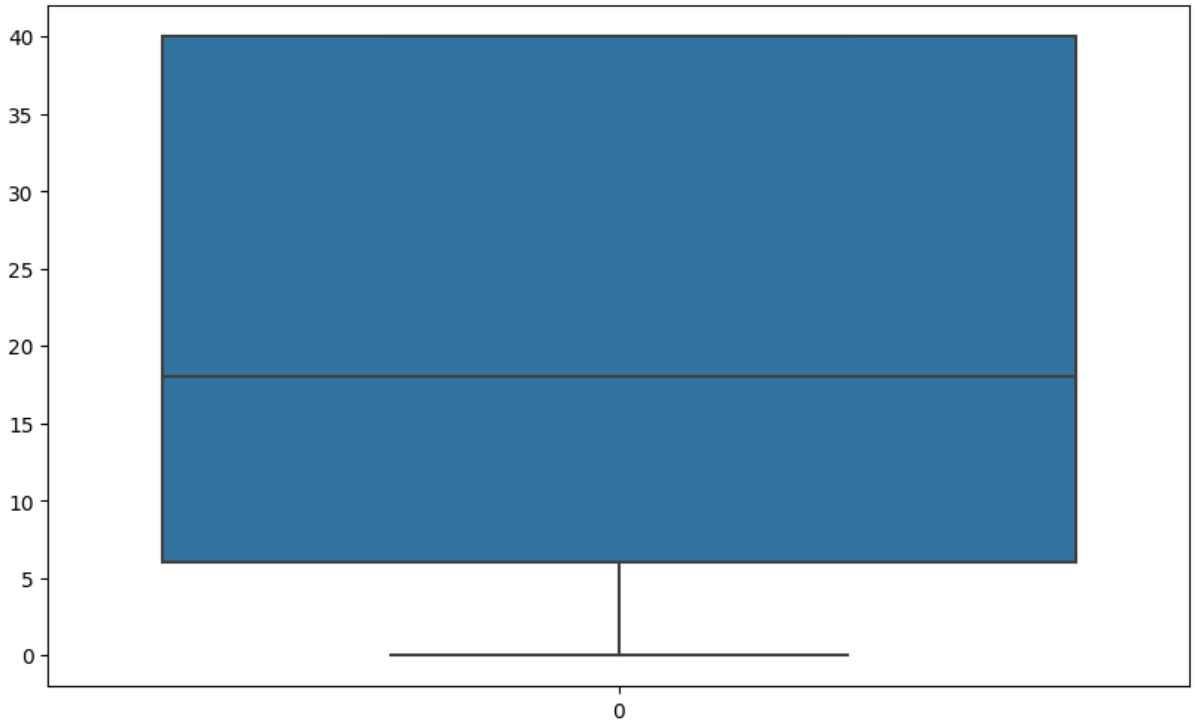
- **WORK_EXP**

Napomenimo da je za klijente koji su penzioneri vrednost ove promenljive postavljena na 40 godina.



Slika 43 - Histogram za promenljivu WORK_EXP

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

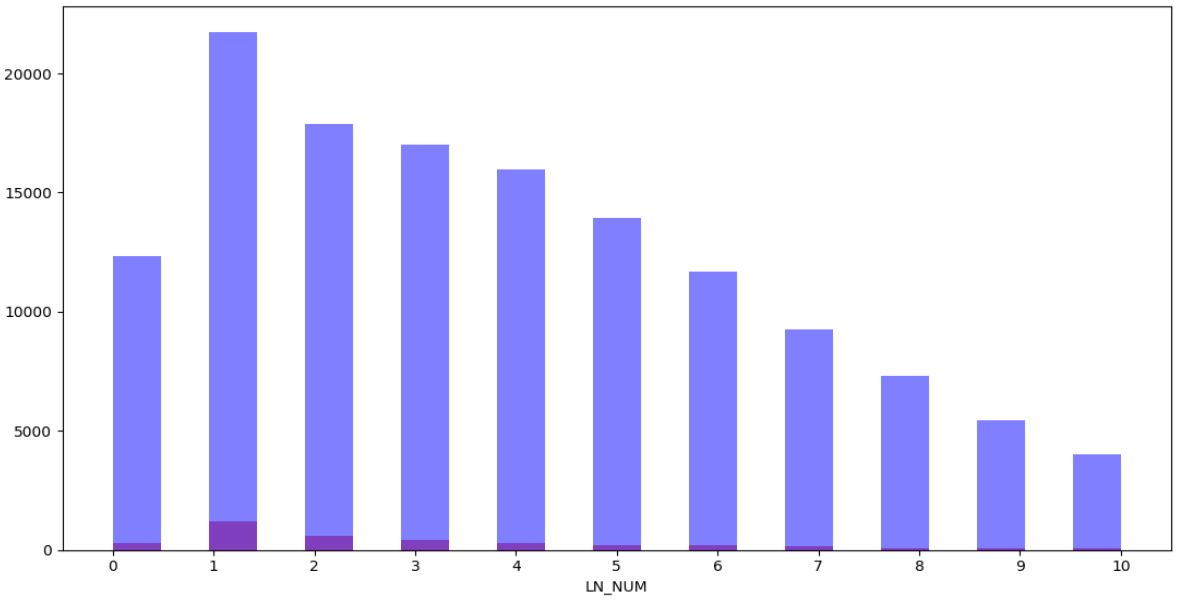


Slika 44 - Box plot za promenljivu WORK_EXP

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

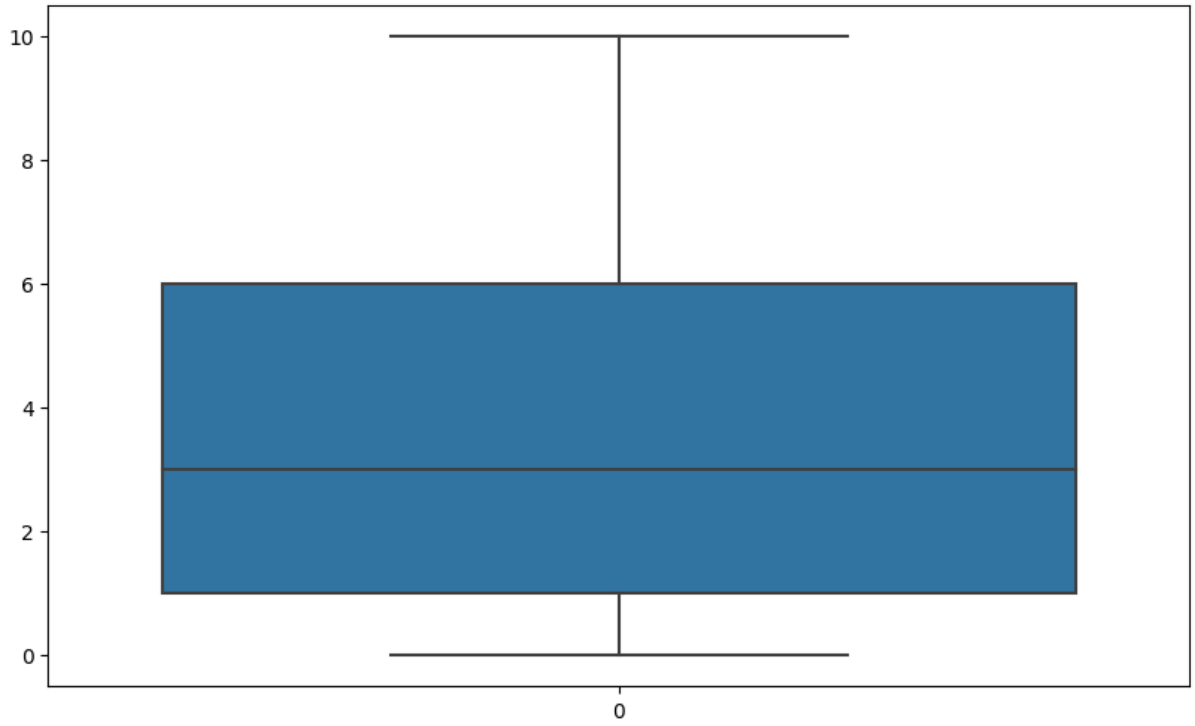
Uočavamo da klijenti sa manje radnog iskustva predstavljaju rizičniju grupu klijenata.

- **LN_NUM**



Slika 45 - Histogram za promenljivu LN_NUM

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

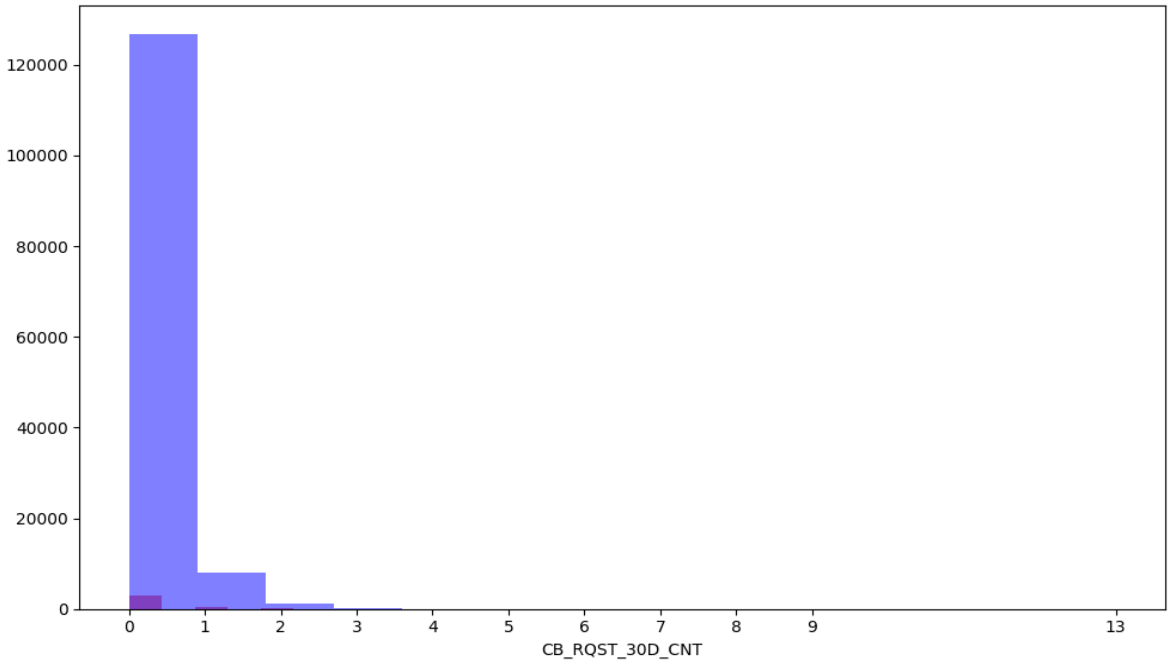


Slika 46 - Box plot za promenljivu LN_NUM

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

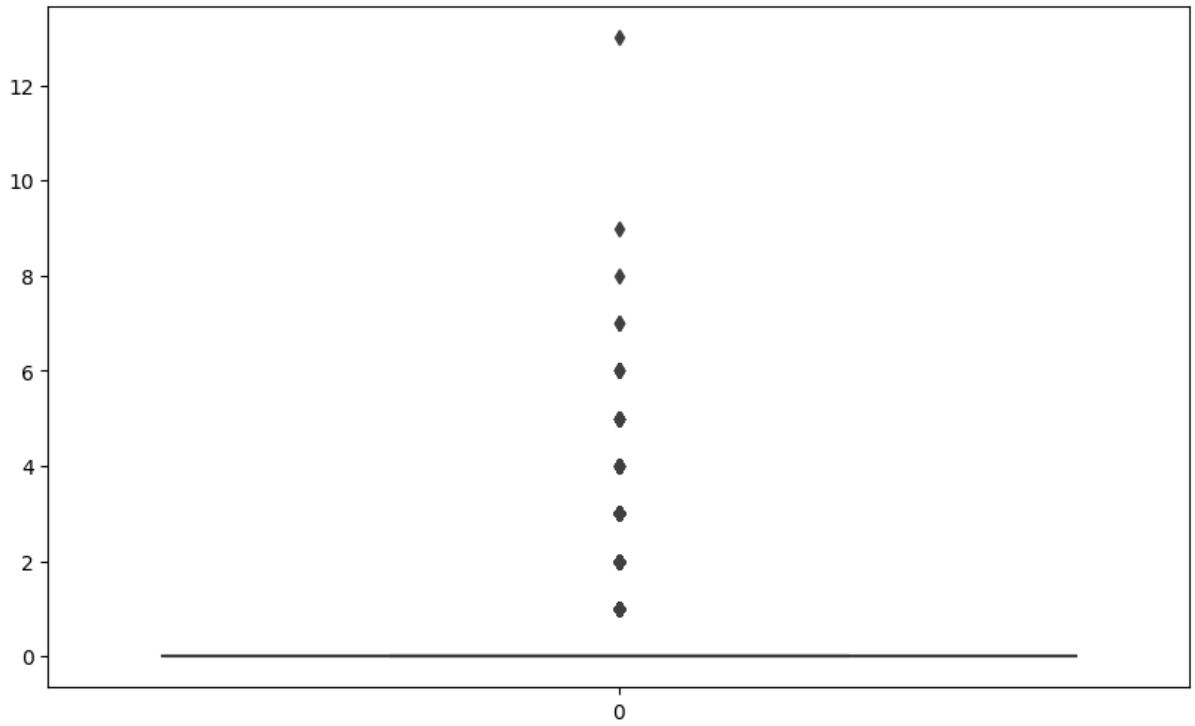
Klijenti sa većim broje kredita imaju kreiranu kreditnu istoriju i logično predstavljaju manje rizične klijente. Primetimo da su klijenti bez kreditnih proizvoda nisu deo trenda.

- **CB_RQST_30D_CNT**



Slika 47 - Histogram za promenljivu CB_RQST_30D_CNT

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

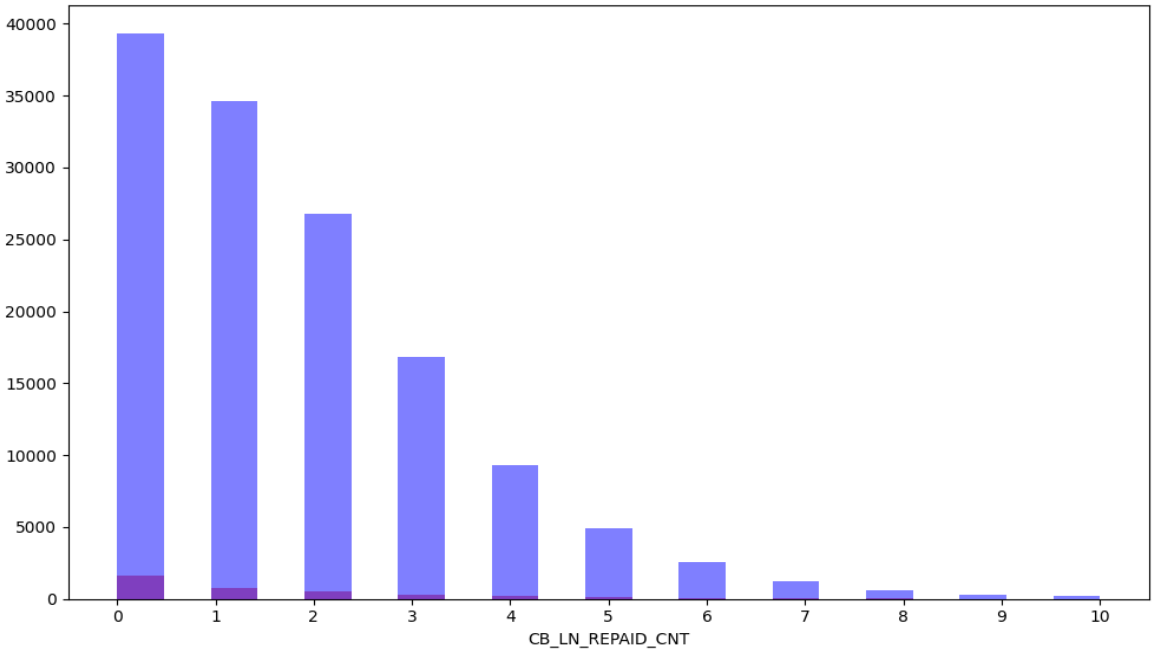


Slika 48 - Box plot za promenljivu CB_RQST_30D_CNT

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

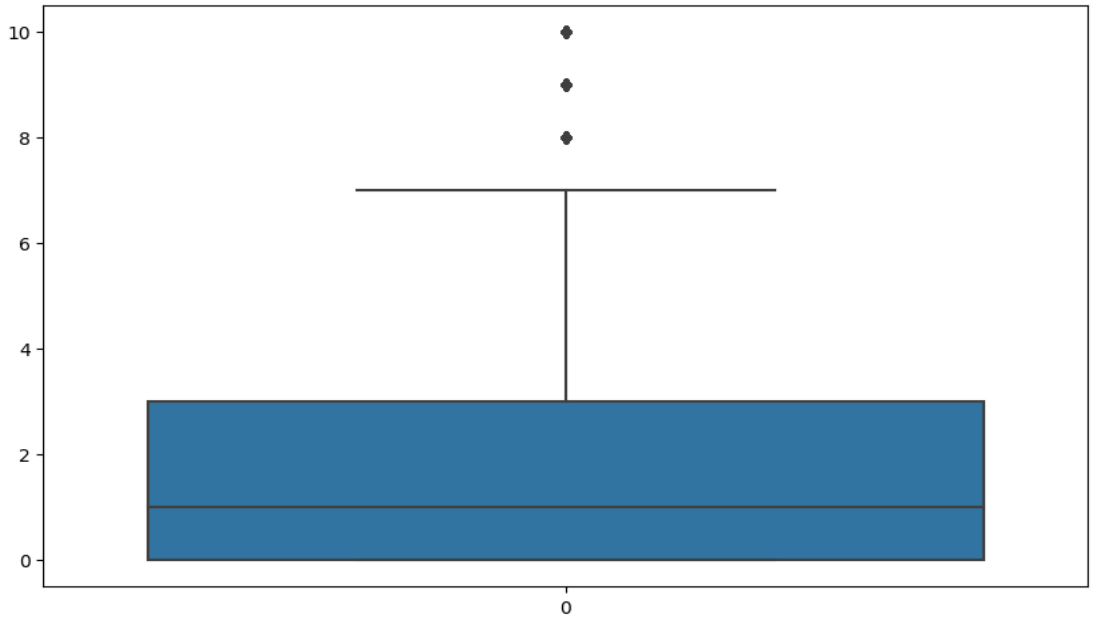
Ukoliko je klijent u kratkom vremenskom roku aplicirao na više mesta i samim time tražio izveštaj kreditnog biroa, to može biti znak da klijent nije pogodan za kreditiranje da je samim tim rizičniji.

- **CB_LN_REPAID_CNT**



Slika 49 - Histogram za promenljivu CB_LN_REPAID_CNT

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

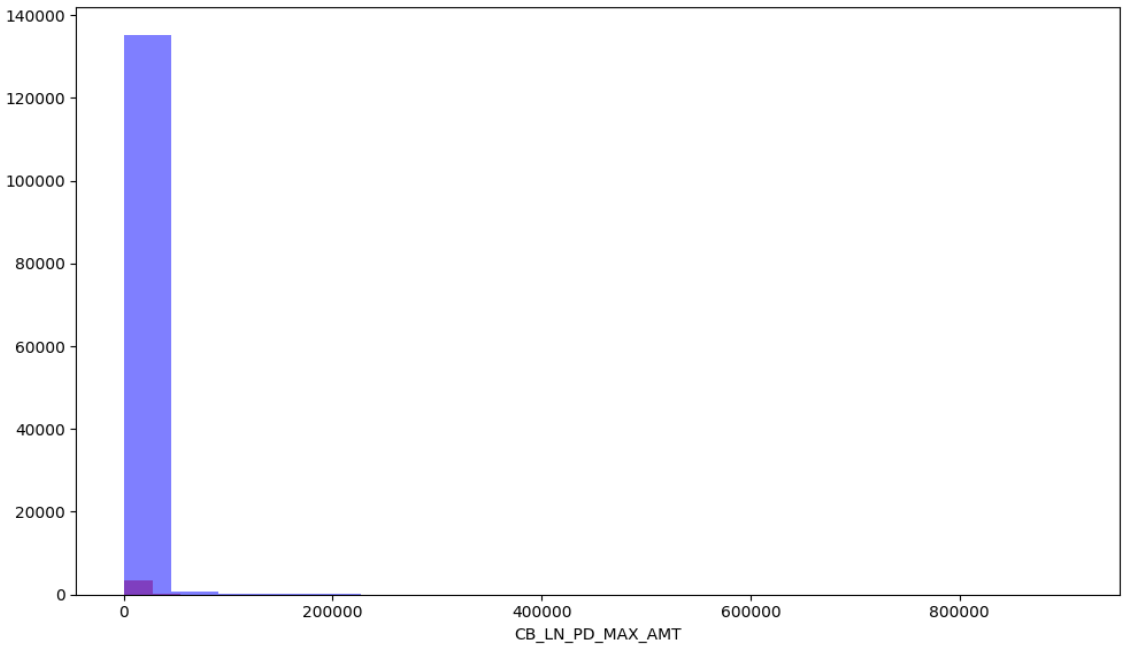


Slika 50 - Box plot za promenljivu CB_LN_REPAID_CNT

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

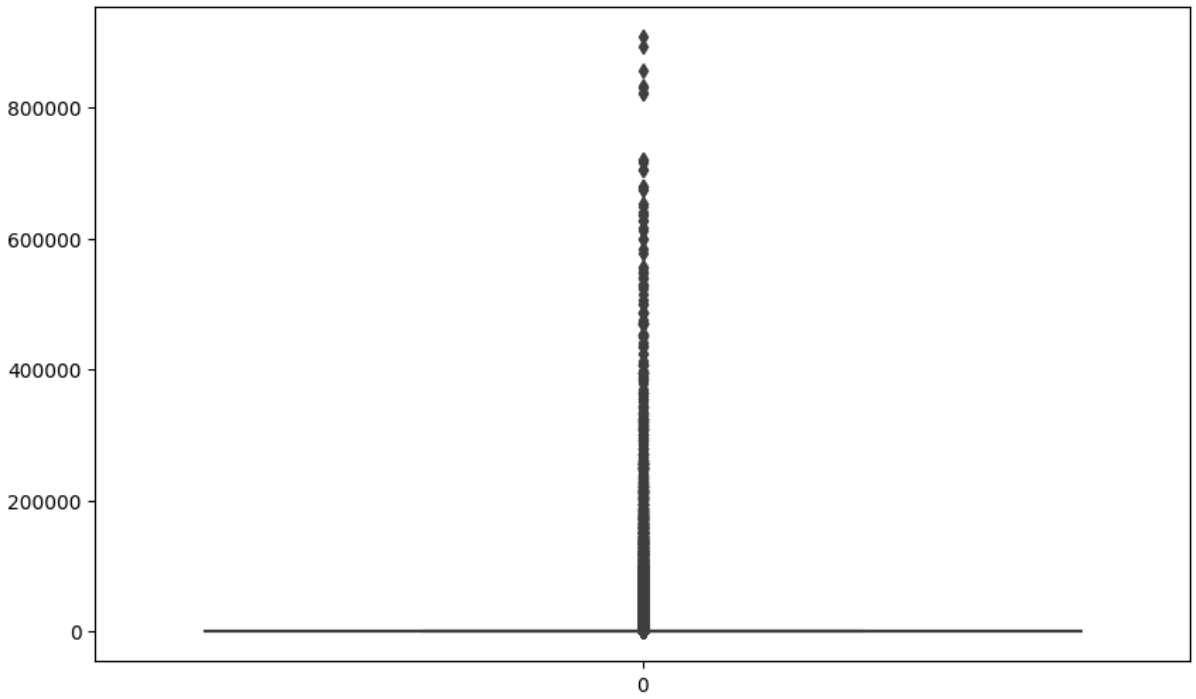
Zaključak za ovu promenljivu je sličan kao i za promenljivu LN_NUM.

- **CB_LN_PD_MAX_AMT**



Slika 51 - Histogram za promenljivu CB_LN_PD_MAX_AMT

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

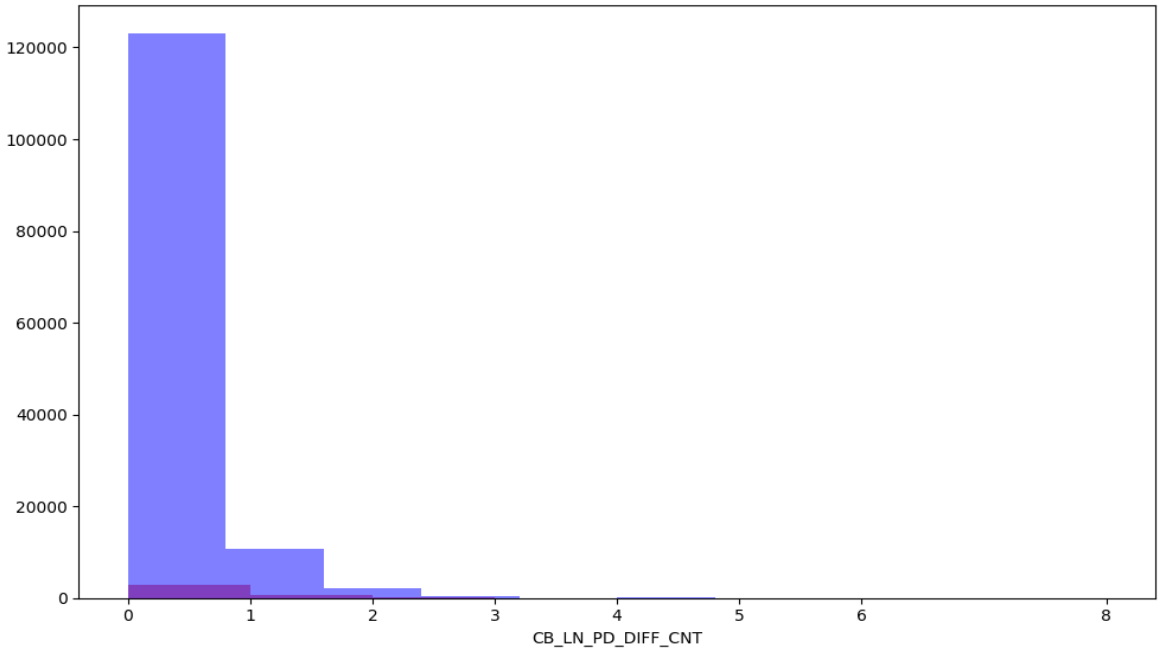


Slika 52 - Box plot za promenljivu CB_LN_PD_MAX_AMT

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

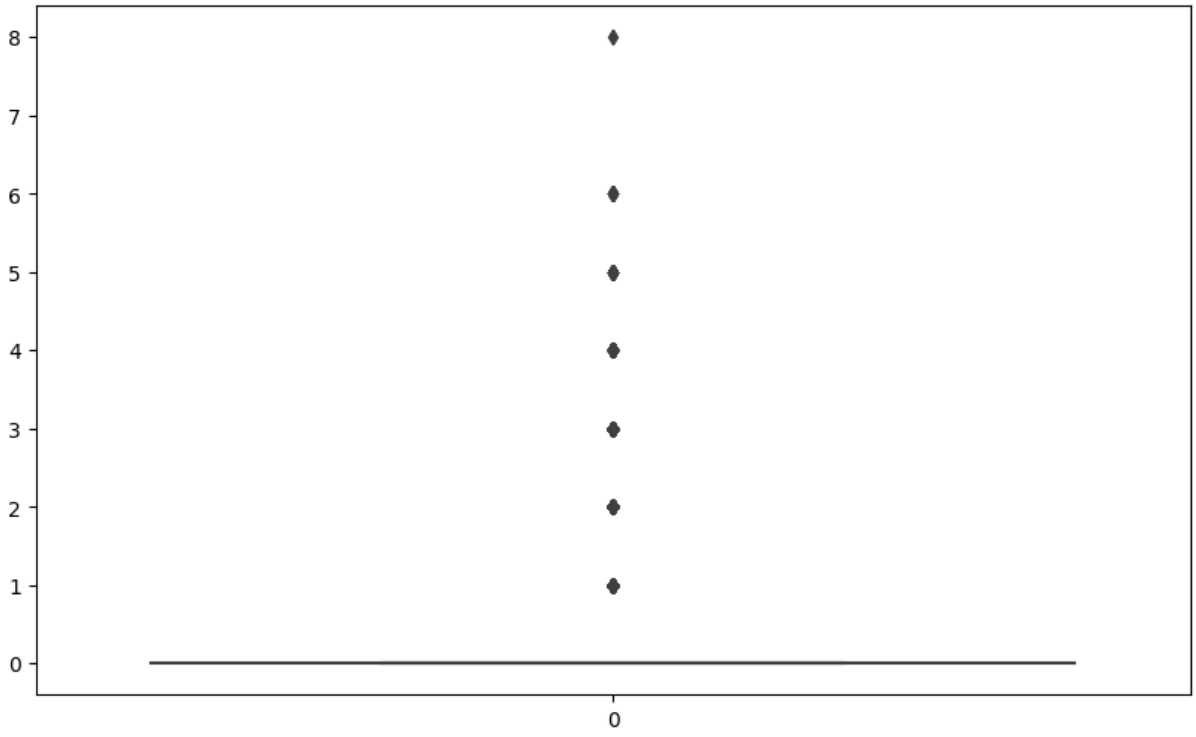
Primetimo da je mali broj klijenata uopšte imao kašnjenja po kreditima u izveštaju kreditnog biroa. Kakogod, takvi klijenti mogu biti značajni za učenje modela.

- **CB_LN_PD_DIFF_CNT**



Slika 53 - Histogram za promenljivu CB_LN_PD_DIFF_CNT

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

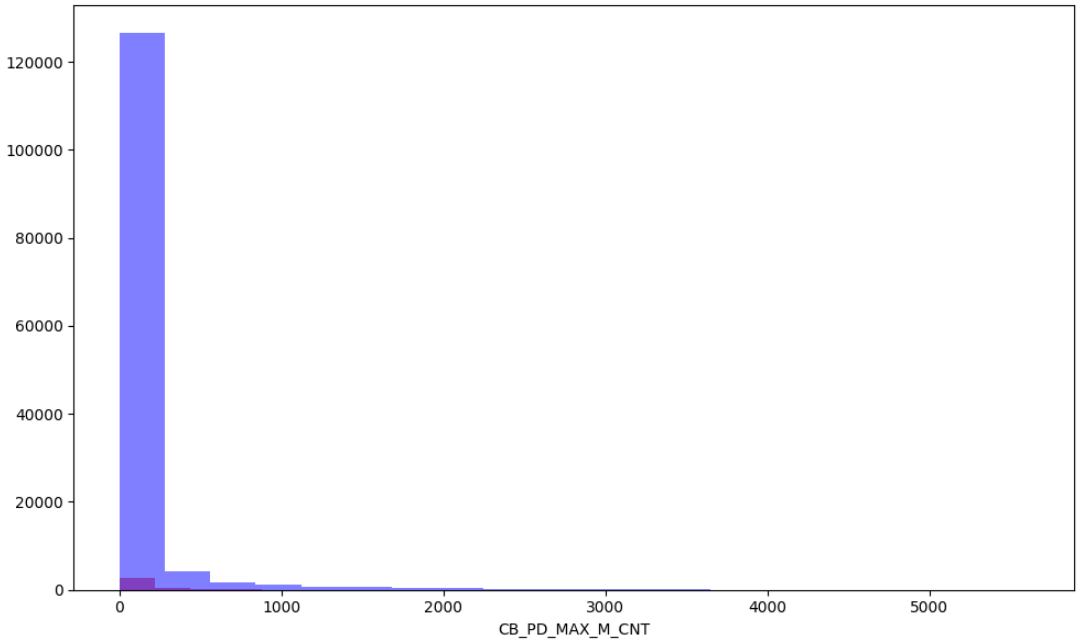


Slika 54 - Box plot za promenljivu CB_LN_PD_DIFF_CNT

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

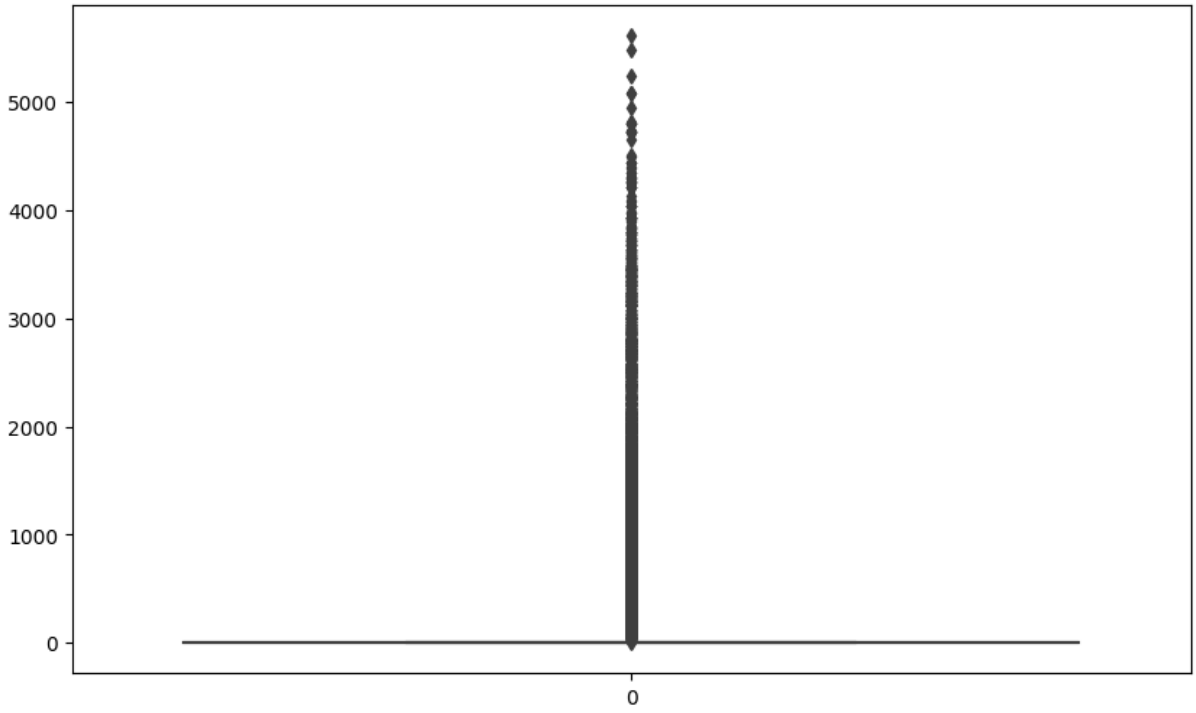
Zaključak je sličan kao i za prethodnu promenljivu stim da se ovde posmatra broj kredita po kojima je klijent kasnio a ne iznos.

- **CB_PD_MAX_M_CNT**



Slika 55 - Histogram za promenljivu CB_PD_MAX_M_CNT

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

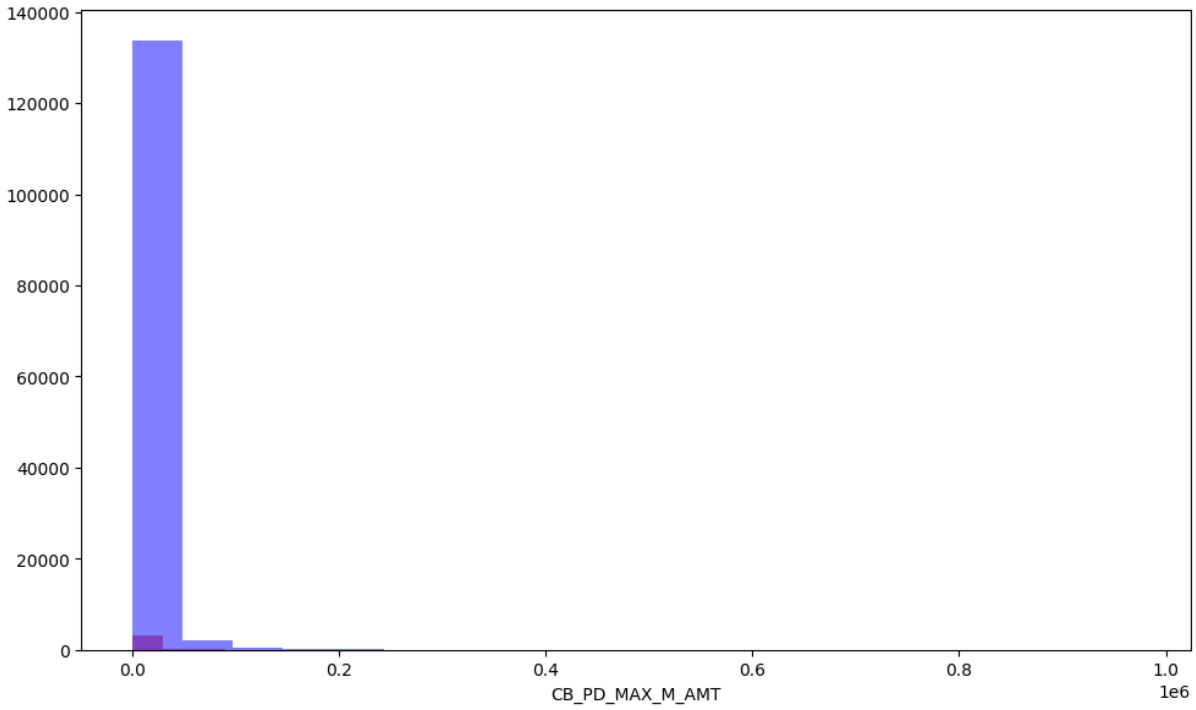


Slika 56 - Box plot za promenljivu CB_PD_MAX_M_CNT

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

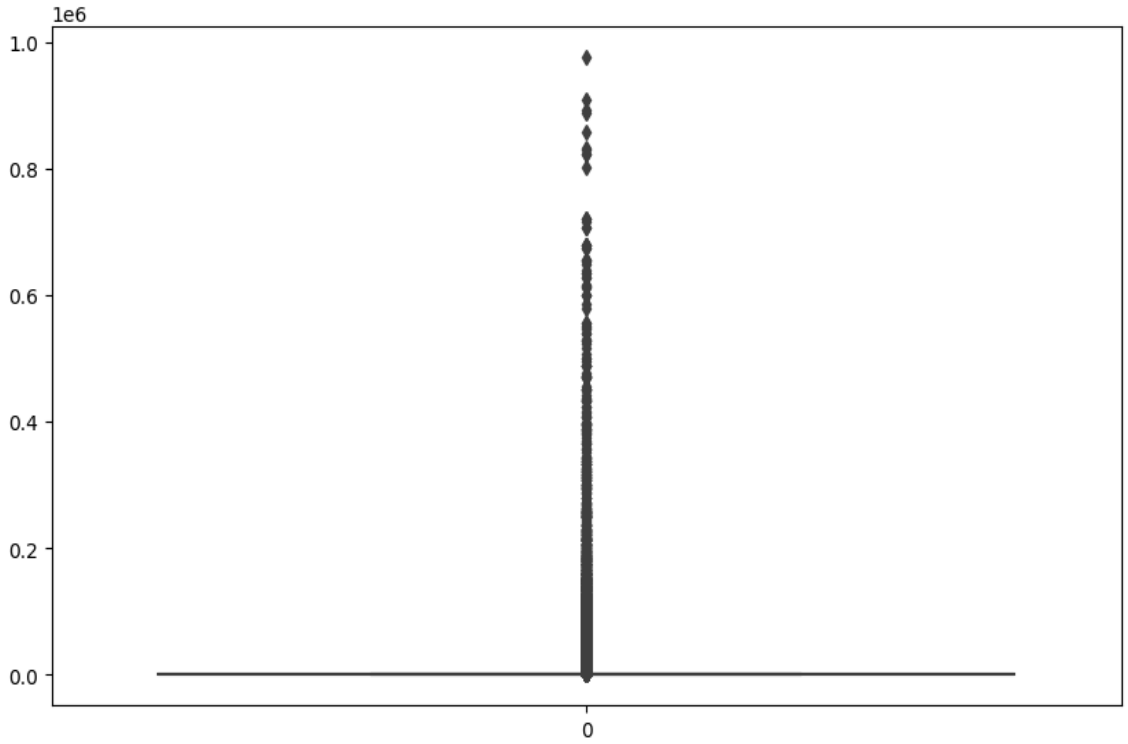
Zaključak je sličan kao i za promenljivu CB_LN_PD_MAX_AMT.

- **CB_PD_MAX_M_AMT**



Slika 57 - Histogram za promenljivu CB_PD_MAX_M_AMT

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u



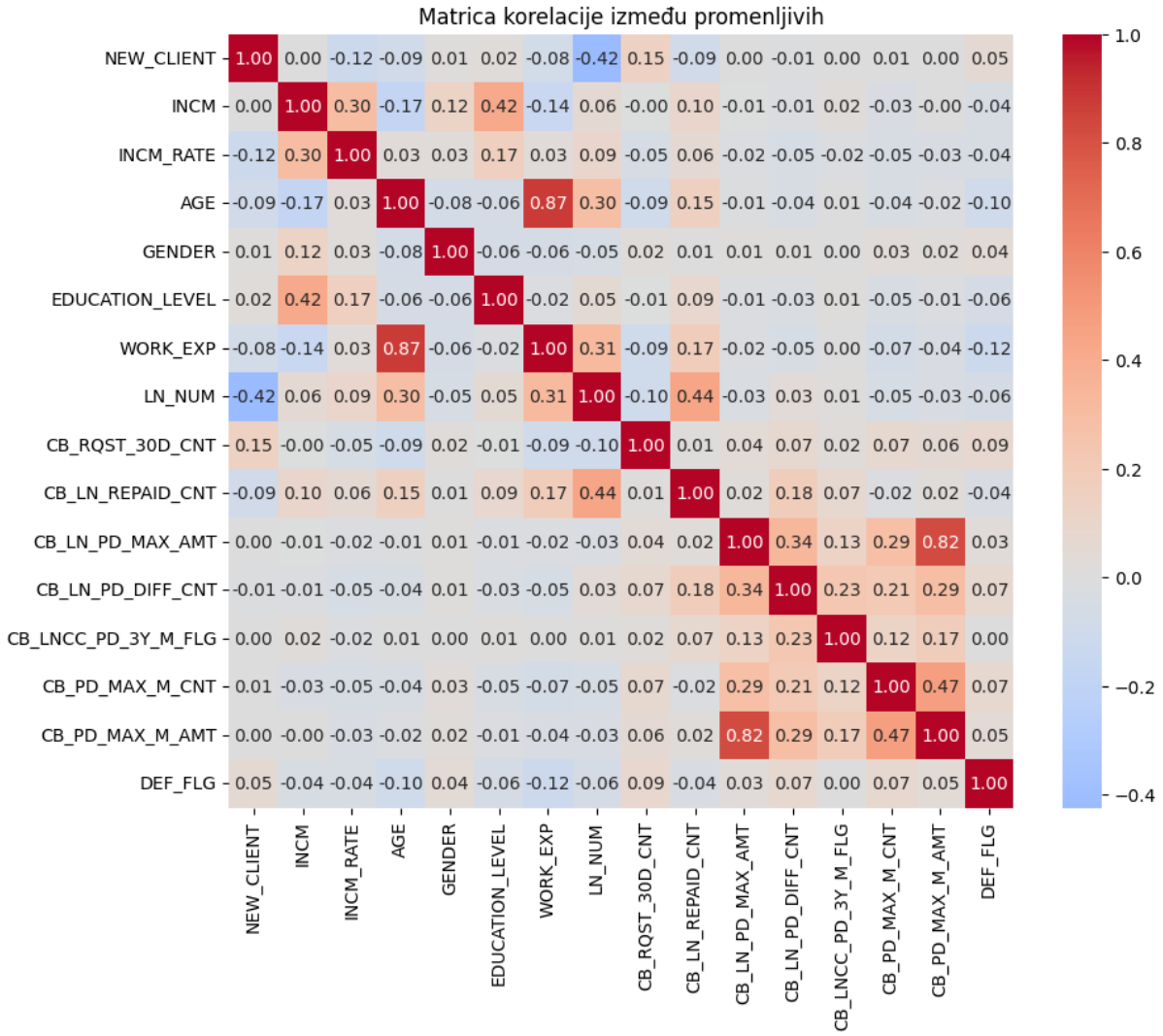
Slika 58 - Box plot za promenljivu CB_PD_MAX_M_AMT

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

Zaključak je sličan kao i za promenljivu CB_LN_PD_MAX_AMT, s tim da ova promenljiva posmatra i druge kreditne proizvode osim kredita (dozvoljeni minus, kreditne kartice itd.)

U kreiranju modela logističke regresije, slično kao i za model linearne regresije, poželjno je izostaviti promenljive koje su visoko korelisane. Takođe potrebno je razmotriti i multikolinearnost.

Prikažimo matricu korelacije za razmatrane promenljive. Posmatra se Pirsonov (Pearson) koeficijent korealcije.



Slika 59 – Matrica korelacije promenljivih u Modelu 1

Izvor: Sopstveni proračun autora dobijen korišćenjem programskog koda u Python-u

Uočimo dva para promenljivih sa većom korelacijom:

- **AGE** i **WORK_EXP** – koeficijent korelacije 0.87
- **CB_LN_PD_MAX_AMT** i **CB_PD_MAX_M_AMT** – koeficijent korelacije 0.82

Imperativ je da finalni model ne sadrži promenljive koje su visokokorelisane (više od 90%).

Skup koji posmatramo za ovaj model podelićemo na dva skupa:

- **Train skup** ili skup za treniranje, na kome obučavamo model
- **Validacioni skup**, na kome ćemo odrediti PD cutoff.

Testni skup (**test set**) koji treba da bude potpuno nezavisan od prethodna dva ne postoji, jer akcenat na samim parametrima ovog modela koji će dati priporne raspodele koeficijenata za model Bejzove logističke regresije.

Podela na pomenuta dva skupa je izvršena u odnosu 80:20, tako da oba skupa zadrže jednaku stopu defaulta. Train skup sadrži **112,185** instanci dok validacioni sadrži **28,047** instanci. Na oba skupa stopa defaulta iznosi **2.57**%.

Nadalje, koristićemo train skup za treniranje modela logističke regresije.

Nakon što smo definisali skup za treniranje modela primenjujemo modifikovani **backwards** algoritam za selekciju promenljivih u modelu logističke regresije, definisan u 4.4.

Ispod su rezultati dobijenog modela logističke regresije:

```
Optimization terminated successfully.
         Current function value: 0.103783
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:                DEF_FLG   No. Observations:               112185
Model:                          Logit   Df Residuals:                   112171
Method:                           MLE   Df Model:                           13
Date:                Wed, 20 Sep 2023   Pseudo R-squ.:                 0.1321
Time:                        20:45:57   Log-Likelihood:                -11643.
converged:                       True   LL-Null:                       -13416.
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================
                           coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const                  -1.6823      0.089    -18.826      0.000      -1.857      -1.507
NEW_CLIENT              0.5319      0.046     11.631      0.000       0.442       0.622
INCM                -1.964e-05   1.42e-06    -13.840      0.000   -2.24e-05   -1.69e-05
INCM_RATE              -0.6613      0.133     -4.984      0.000      -0.921      -0.401
GENDER                  0.5257      0.041     12.939      0.000       0.446       0.605
EDUCATION_LEVEL        -0.3429      0.035     -9.841      0.000      -0.411      -0.275
WORK_EXP               -0.0563      0.002    -32.641      0.000      -0.060      -0.053
CB_RQST_30D_CNT         0.4176      0.031     13.307      0.000       0.356       0.479
CB_LN_REPAID_CNT       -0.0332      0.014     -2.446      0.014      -0.060      -0.007
CB_LN_PD_MAX_AMT    -4.553e-06    1.1e-06     -4.149      0.000    -6.7e-06    -2.4e-06
CB_LN_PD_DIFF_CNT       0.5142      0.032     15.901      0.000       0.451       0.578
CB_LNCC_PD_3Y_M_FLG    -0.2287      0.105     -2.169      0.030      -0.435      -0.022
CB_PD_MAX_M_CNT         0.0003   3.75e-05      8.459      0.000       0.000       0.000
CB_PD_MAX_M_AMT       4.12e-06   8.92e-07      4.617      0.000    2.37e-06    5.87e-06
================================================================================
```

Slika 60 – Rezultati logističke regresije za Model 1

Izvor: Sopstveni proračun autora dobijen korišćenjem programskog koda u Python-u

Primetimo da su kao rezultat pomenutog algoritma sve promenljive preostale u modelu statistički značajne.

Kao što smo pomenuli prisustvo multikolinearnosti u uzorku prouzrokuje veće standardne greške koeficijenata, kao i to da male promene u podacima dovode do velikih promena u koeficijentima, čak i do promene smera uticaja na zavisnu promenljivu. Neki od uzroka pojave multikolinearnosti su mali obim uzorka, uključivanje u model nezavisne promenljive koja je

kombinacija drugih nezavisnih promenljivih iz modela, uključivanje u model dve iste ili skoro iste promenljive. Radi identifikovanja prisustva multikolinearnosti često se koristi VIF test, koji je dat formulom:

$$VIF = \frac{1}{1 - R_J^2}$$

i predstavlja procenat varijanse jedne nezavisne promenljive objašnjen varijansom ostalih nezavisnih promenljivih. U formuli R_j^2 predstavlja koeficijent determinacije za model u kome je j-ta nezavisna promenljiva prikazana korišćenjem linearne regresije sa preostalim nezavisnim promenljivama.

Najbolji slučaj podrazumeva da je $VIF = 1$, s obzirom da to znači da se određena nezavisna promenljiva ne može objasniti pomoću ostalih. U praksi se često pretpostavlja da, ukoliko je $VIF > 10$, postoji snažna multikolinearnost.

	feature	VIF
0	NEW_CLIENT	1.209471
1	INCM	5.016815
2	INCM_RATE	1.901956
3	GENDER	1.899900
4	EDUCATION_LEVEL	6.813398
5	WORK_EXP	2.601396
6	CB_RQST_30D_CNT	1.100278
7	CB_LN_REPAID_CNT	2.183745
8	CB_LN_PD_MAX_AMT	3.270023
9	CB_LN_PD_DIFF_CNT	1.356537
10	CB_LNCC_PD_3Y_M_FLG	1.113888
11	CB_PD_MAX_M_CNT	1.453119
12	CB_PD_MAX_M_AMT	3.834918

Slika 61 – VIF za promenljive odabrane u Modelu 1

Izvor: Sopstveni proračun autora dobijen korišćenjem programskog koda u Python-u

Uočimo da nije prisutna snažna multikolinearnost. Pomenuti validacioni skup ima ulogu određivanja optimalnog PD cut-off-a posmatrajmo neke od metrika kvaliteta modela na ovom skupu. GINI koefiicijent na validacionom skupu iznosi 57.17% dok KS koeficijent iznosi 45.97% što predtsavlja relativno zadovoljavajuće rezultate.

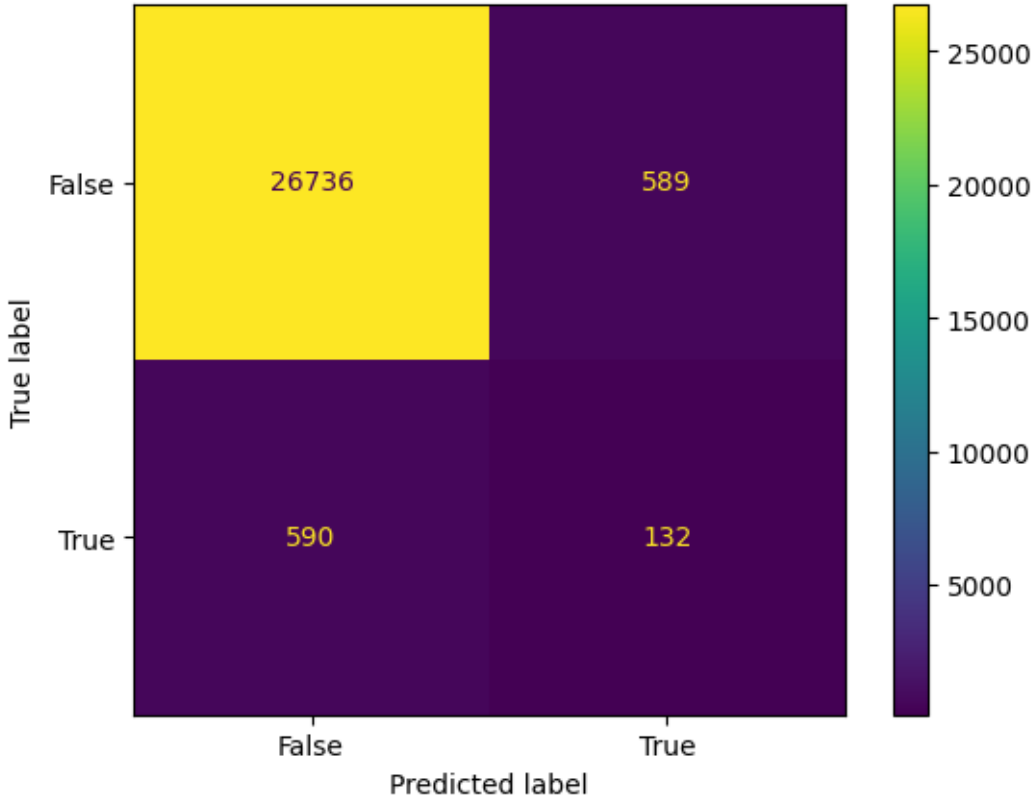
U 4.5 smo opisali način na koji ćemo odrediti optimalni PD cut-off. Kako je stopa defaulta na validacionom skupu 2.57%, vrednost praga će biti 97.34 postotni percentile vrednosti verovatnoća iznosi 0.11454065361364345. Napomenimo da ćemo ovu vrednost praga koristiti i u ostalim modelima za metrike kvaliteta modela koje zahtevaju procenjenu vrednost zavisne promenljive: odziv preciznost, tačnost i f-skor. Napomenimo da ova odluka ima logike iz dva razloga. Jedan je što banke definišu pragove iznad kojih ne kreditiraju klijente ili su iznosi koje

klijent može dobiti manji nego sa boljim skorom. Ove skale sa granicama vrednosti PD-a se nazivaju master skale i nisu podložne čestim promenama. Drugi razlog je da bi poređenje modela po prethodno pomenute 4 metrike bilo konzistentno. Pogledajmo vrednosti ovih metirka za model 1 na validacionom skupu.

Accuracy	0.95796
Recall	0.18283
Precision	0.18308
F1 Score	0.18295

Tabela 7 – Metrike Modela 1

Data je i pomenuta matrica konfizije koja prikazuje odnos predviđenih i stvarnih vrednosti zavisne promenljive iz koje se i dobijaju prethodne metrike.



Slika 62 – Matrica konfuzije za Model 1

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

6.3 Razvoj modela logističke regresije nove banke (Model 2)

Još jedan model logističke regresije se razvija ovaj put na podacima nove banke. Ovo su aplikacije iz inicijalnog skupa zaključno između 01.01.2021 i 28.02.2022. Ovaj skup sadrži **89,633** instance (aplikacije) od koji je **1,319** rezultiralo default-om. Dakle stopa default-a na ovom skupu je **1.47%**.

Predstavljene su i neke od osnovnih karakteristika promenljivh na posmatranom skupu.

	NEW_CLIENT	INCM	INCM_RATE	AGE	GENDER	EDUCATION_LEVEL
count	89633	87851	87851	89633	89633	89053
mean	0.17	56419.46	0.21	48.53	0.52	2.45
std	0.38	34553.99	0.25	14.12	0.5	0.82
min	0	10093	0.01	20	0	1
25%	0	35478	0.06	37	0	2
50%	0	48651.43	0.12	47	1	2
75%	0	67000	0.27	61	1	3
max	1	596584	3.8	79	1	4

Tabela 8 - Statistički pokazatelji promenljivh na skupu nove banke – 1. deo

	WORK_EXP	LN_NUM	CB_RQST_30D_CNT	CB_LN_REPAID_CNT
count	89633	89633	89584	89633
mean	20.2	3.53	0.06	1.56
std	13.96	2.91	0.29	1.65
min	0	0	0	0
25%	8	1	0	0
50%	17	3	0	1
75%	36	6	0	2
max	40	10	8	10

Tabela 9 - Statistički pokazatelji promenljivh na skupu nove banke – 2. deo

	CB_LN_PD_MAX_AMT	CB_LN_PD_DIFF_CNT	CB_LNCC_PD_3Y_M_FLG	CB_PD_MAX_M_CNT	CB_PD_MAX_M_AMT	DEF_FLG
count	89633	89633	89633	89633	89633	89633
mean	1097.97	0.09	0.03	53.23	2844.72	0.01
std	11204.41	0.35	0.17	244.06	14996.81	0.12

min	0	0	0	0	0	0
25%	0	0	0	0	0	0
50%	0	0	0	0	0	0
75%	0	0	0	0	0	0
max	892860.4	6	1	5081	892860.4	1

Tabela 10 - Statistički pokazatelji promenljivh na skupu nove banke – 3. deo

Napomenimo da je ovaj skup nove banke podeljen na trening skup i na testni skup. Na trening skupu ćemo razvijati model logističke regresije kao i model Bejzove logističke regresije gde se posmatraju priorne raspodele koeficijenata regresije iz modela 1. Testni skup će služiti za poređenje performansi sva 3 modela. Podela na train i test skup skupa nove banke je urađena u odnosu 80:20 gde je stopa defaulta u oba skupa jednaka i iznosi 1.47% kao i na celom skupu.

Predstavićemo dobijeni model logističke regresije smatrajući da su svi koraci isti kao u razvoju Modela 1 i oni neće biti navođeni.

Rezultati dobijenog logističkog modela su sledeći:

Logit Regression Results

Dep. Variable:	DEF_FLG	No. Observations:	71706
Model:	Logit	Df Residuals:	71693
Method:	MLE	Df Model:	12
Date:	Wed, 20 Sep 2023	Pseudo R-squ.:	0.09880
Time:	20:46:55	Log-Likelihood:	-4955.1
converged:	True	LL-Null:	-5498.3
Covariance Type:	nonrobust	LLR p-value:	4.822e-225

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.2508	0.210	-10.700	0.000	-2.663	-1.838
NEW_CLIENT	0.7569	0.070	10.857	0.000	0.620	0.894
INCM	-1.617e-05	1.91e-06	-8.483	0.000	-1.99e-05	-1.24e-05
INCM_RATE	-1.4215	0.244	-5.825	0.000	-1.900	-0.943
AGE	-0.0091	0.004	-2.106	0.035	-0.018	-0.001
GENDER	0.5561	0.067	8.258	0.000	0.424	0.688
EDUCATION_LEVEL	-0.2379	0.051	-4.663	0.000	-0.338	-0.138
WORK_EXP	-0.0417	0.005	-9.026	0.000	-0.051	-0.033
CB_RQST_30D_CNT	0.4449	0.058	7.735	0.000	0.332	0.558
CB_LN_REPAID_CNT	0.0337	0.021	1.623	0.105	-0.007	0.074
CB_LN_PD_DIFF_CNT	0.4233	0.063	6.720	0.000	0.300	0.547
CB_PD_MAX_M_CNT	0.0002	8.28e-05	2.985	0.003	8.48e-05	0.000
CB_PD_MAX_M_AMT	4.9e-06	1.22e-06	4.016	0.000	2.51e-06	7.29e-06

Slika 63 - Rezultati logističke regresije za Model 2

Izvor: Sopstveni proračun autora dobijen korišćenjem programskog koda u Python-u

Primetimo da Model 2 ne sadrži iste promenljive kao model 1. Performanse modela na testnom skupu ćemo posmatrati u odeljku *Rezultati i poređenja*.

6.4 Razvoj modela Bejzove logističke regresije (Model 3)

U ovom modelu cilj nam je da razvijamo logističku regrsiju na trening skupu nove banke ali koristeći priorna prethodna znanja. Koristimo ocenjene koeficijente modela 1 i smatramo da koeficijenti regresije imaju priorne normalne raspodele sa očekivanjem i standardnom devijacijom dobijenim u modelu 1. Da bismo dobili posteriorne ocene parametara koristili generisali smo lanace Markova za parameter modela koristeći pomenuti Metropolis- Hastings algoritam uz random walk sampler. Generisani su lanci od 55 000 uzoraka, gde je 5000 predstavljamo burn-in period koji služi da pruži dovoljno vremena lancu da konvergira ka stacionarnoj raspodeli. Korišćena je Python PYMC biblioteka.

Na slici ispod su dati rezultati pomenutih MCMC algoritama

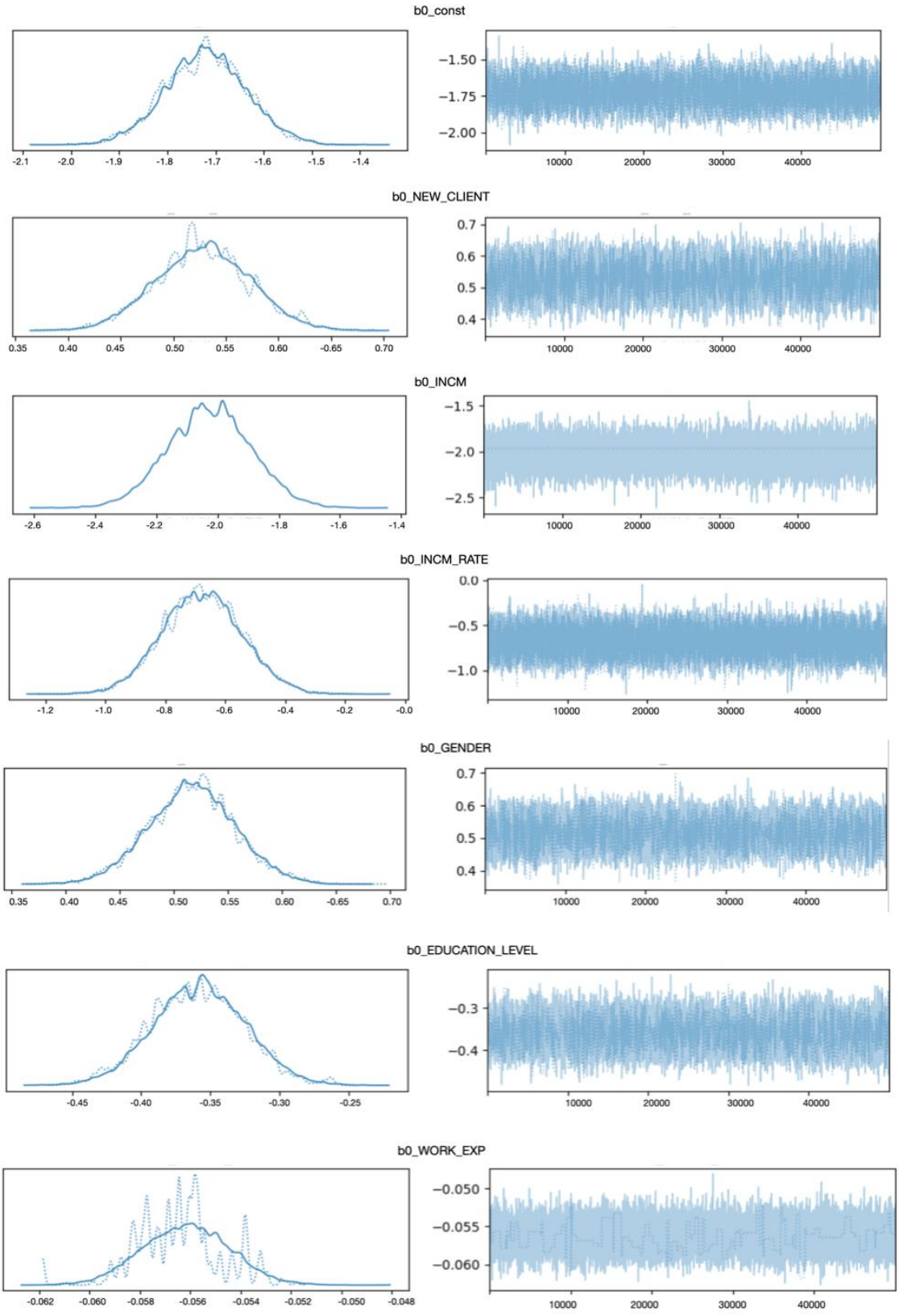
	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd
b0_const	-1.721	0.084	-1.875	-1.562	0.001	0.001
b0_NEW_CLIENT	0.529	0.045	0.443	0.614	0.001	0.000
b0_INCM	-0.000	0.000	-0.000	-0.000	0.000	0.000
b0_INCM_RATE	-0.677	0.133	-0.924	-0.424	0.001	0.001
b0_GENDER	0.516	0.040	0.442	0.595	0.001	0.000
b0_EDUCATION_LEVEL	-0.359	0.034	-0.420	-0.294	0.000	0.000
b0_WORK_EXP	-0.056	0.002	-0.059	-0.053	0.000	0.000
b0_CB_RQST_30D_CNT	0.412	0.031	0.353	0.468	0.000	0.000
b0_CB_LN_REPAID_CNT	-0.034	0.013	-0.060	-0.009	0.000	0.000
b0_CB_LN_PD_MAX_AMT	-0.000	0.000	-0.000	-0.000	0.000	0.000
b0_CB_LN_PD_DIFF_CNT	0.516	0.032	0.456	0.577	0.000	0.000
b0_CB_LNCC_PD_3Y_M_FLG	-0.228	0.104	-0.421	-0.028	0.001	0.001
b0_CB_PD_MAX_M_CNT	0.000	0.000	0.000	0.000	0.000	0.000
b0_CB_PD_MAX_M_AMT	0.000	0.000	0.000	0.000	0.000	0.000
p[0]	0.001	0.000	0.000	0.001	0.000	0.000

	ess_bulk	ess_tail	r_hat
b0_const	8878.0	10251.0	1.00
b0_NEW_CLIENT	6086.0	4998.0	1.00
b0_INCM	36.0	42.0	2.23
b0_INCM_RATE	14309.0	13525.0	1.00
b0_GENDER	5711.0	5011.0	1.00
b0_EDUCATION_LEVEL	4873.0	4182.0	1.00
b0_WORK_EXP	342.0	394.0	1.00
b0_CB_RQST_30D_CNT	5207.0	4946.0	1.00
b0_CB_LN_REPAID_CNT	2222.0	1581.0	1.00
b0_CB_LN_PD_MAX_AMT	16999.0	42.0	2.23
b0_CB_LN_PD_DIFF_CNT	5217.0	4358.0	1.00
b0_CB_LNCC_PD_3Y_M_FLG	11117.0	11891.0	1.00
b0_CB_PD_MAX_M_CNT	2729.0	40.0	2.12
b0_CB_PD_MAX_M_AMT	17265.0	41.0	2.23

Slika 64 – Rezultati MCMC algoritma

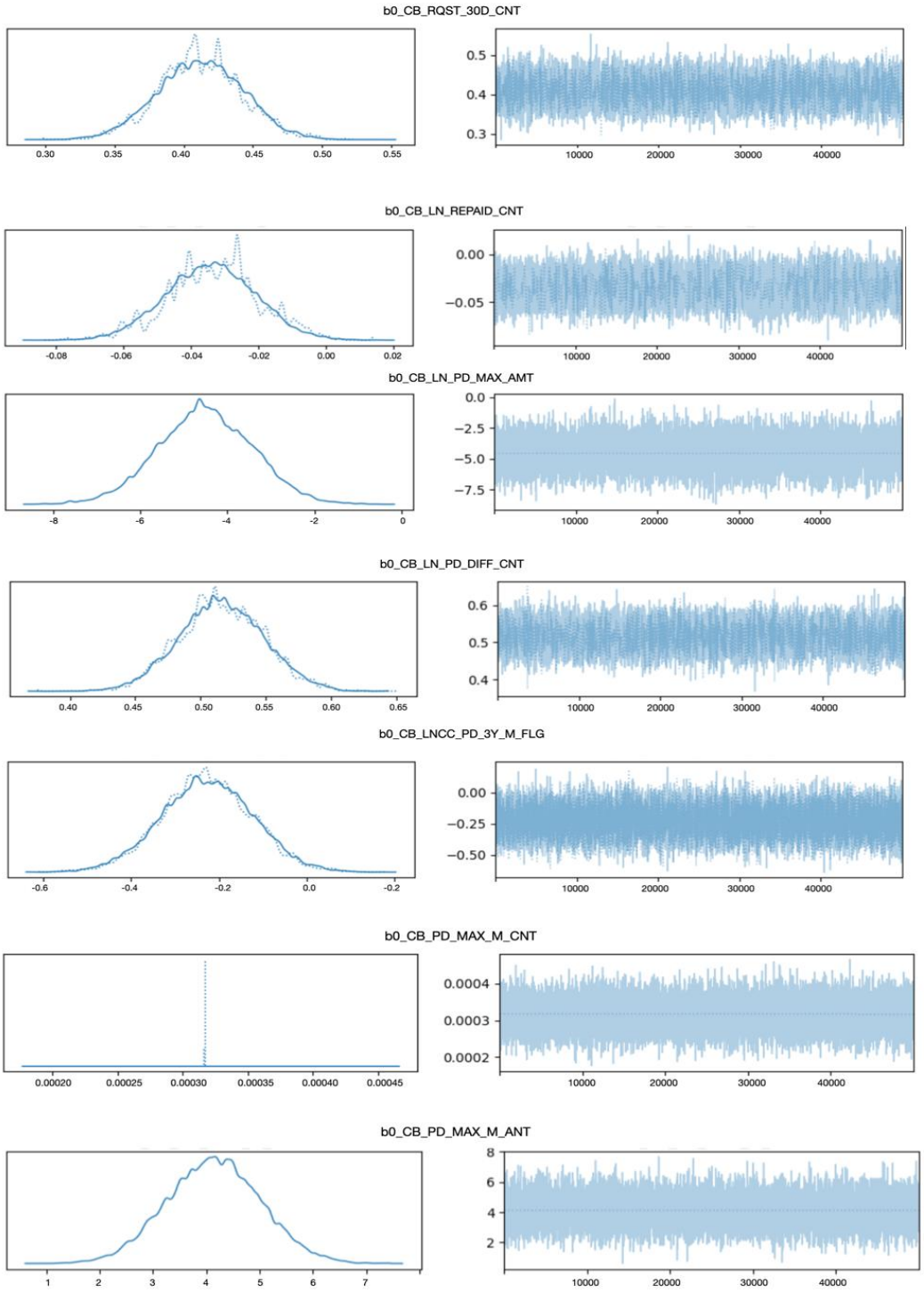
Izvor: Sopstveni proračun autora dobijen korišćenjem programskog koda u Python-u

Posmatrajući pomenutu R-hat metriku možemo uočiti da su lanci Markova za promenljive INCM, CB_LN_PD_MAX_AMT, CB_PD_MAX_M_CNT, CB_PD_MAX_M_AMT. Ovo znači da lanci Markova za ove koeficijente nisu konvergirali. Primetimo da je ocenjena srednja vrednost ovih raspodela 0. Ove promenljive neće biti korišćene u finalnom modelu. Na slikama ispod se može posmatrati vizuelni prikaz priorne i posteriorne rasopdele za svaki koeficijenat kao i trag (trace) lanca Markova koji predtstavlja vremensku seriju uzorkovanih vrednosti.



Slika 65 – Prikaz traga i posteriornih raspodela promenljivih – 1.deo

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u



Slika 66 - Prikaz traga i posteriornih raspodela promenljivih – 2.deo

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

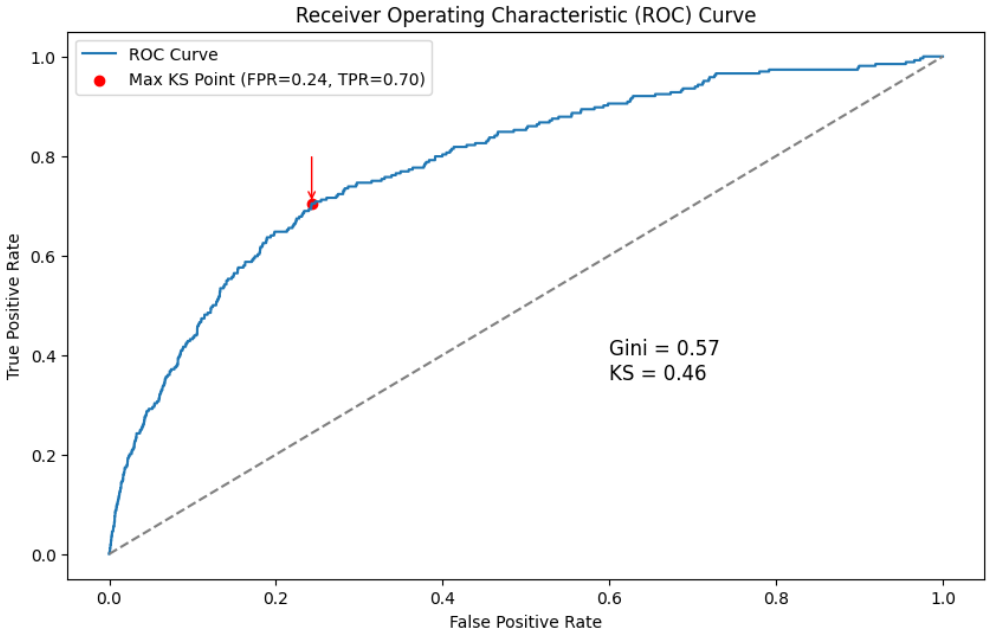
Kako je rezultat Bejzovog pristupa posteriorna raspodela, koeficijenti regresije se ne posmatraju kao tačkaste ocene već kao slučajne promenljive. Za ocenu koeficijenata biramo očekivanu vrednost slučajne promenljive, tj vrednosti iz kolone mean.

7. Rezultati i poređenja

U ovom poglavlju predstavićemo rezultate dobijenih modela, tj. posmatraćemo performanse 3 modela na testnom skupu, koji je potiče iz skupa nove banke. Sa biznis strane se izbor modela može opisati na sledeći način: nova banka pred sobom ima tri moguća skoring modela. Jedan model (Model 1) predtsavlja korišćenje modela stare banke. Model 2 predstavlja razvijen model logističke regresije na novim podacima, dok treći model, Model 3 koristi promenljive iz Modela 1 stim da su koeficijenti regresije prilagođeni novim podacima.

- **Model 1**

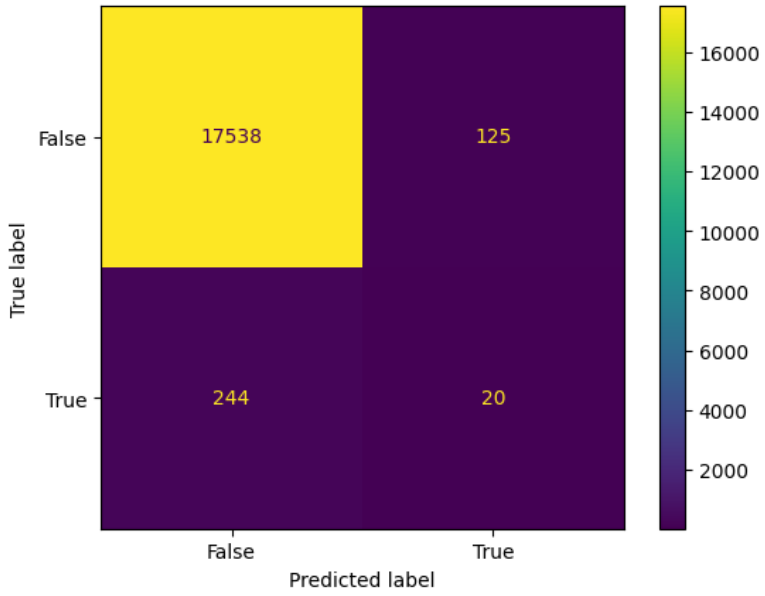
Gini i KS koeficijenti za model 1 na testnom kupu iznose 57.00% odnosno 46.16%



Slika 67 – ROC kriva sa GINI i KS vrednostima za Model 1

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

Na slici ispod je predstavljena matrica konfuzije ukoliko za PD cut-off koristimo vrednost koju smo ranije utvrdili.



Slika 68 - Matrica konfuzije za Model 1 – testni skup

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

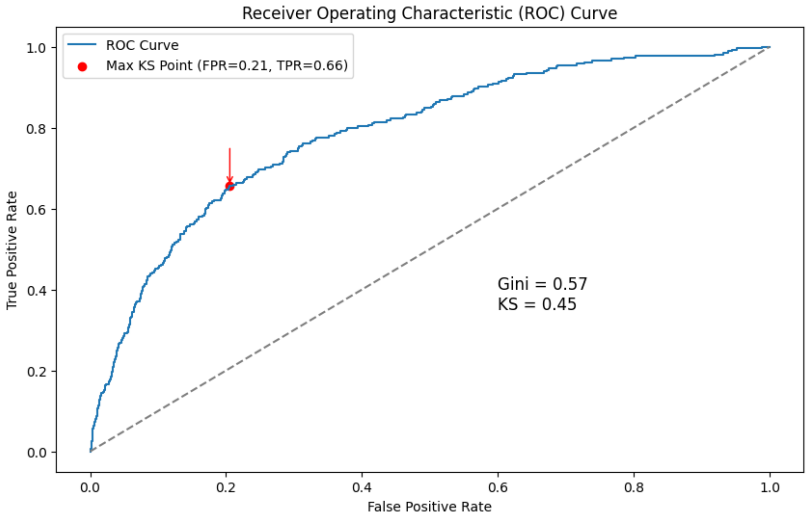
Vrednosti metrika dobijenih iz ove matrice su sleedeće

Accuracy	0.97942
Recall	0.07576
Precision	0.13793
F1 Score	0.0978

Tabela 11 - Metrike Modela 1-testni skup

- **Model 2**

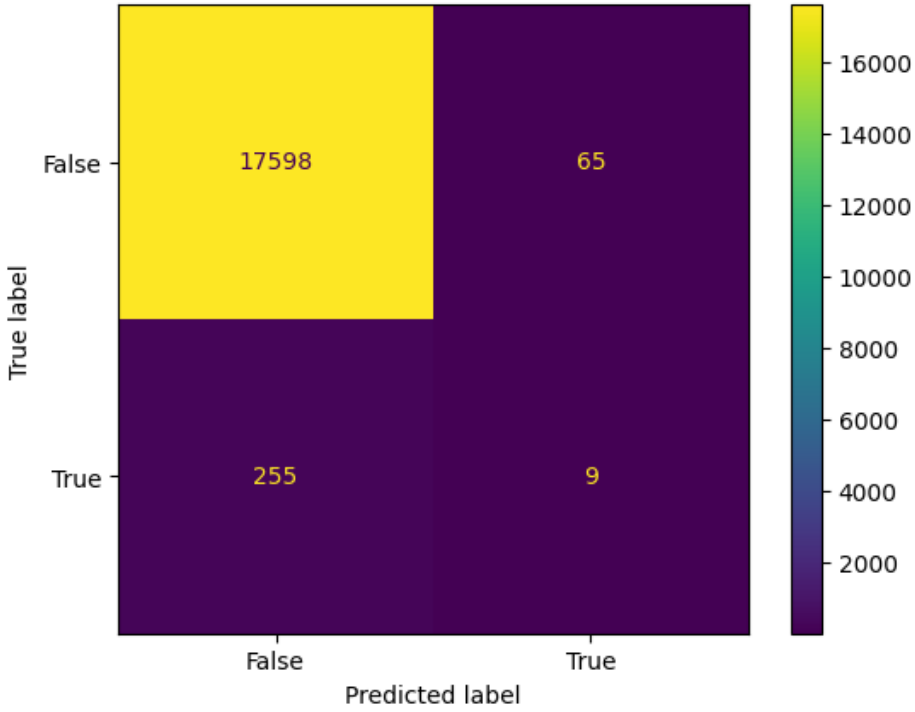
Gini i KS koeficijenti za model 2 na testnom kupu iznose 57.48% odnosno 44.96%



Slika 69 - ROC kriva sa GINI I KS vrednostima za Model 2

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

Na slici ispod je predstavljena matrica konfuzije ukoliko za PD cut-off koristimo vrednost koju smo ranije utvrdili.



Slika 70 - Matrica konfuzije za Model 2

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

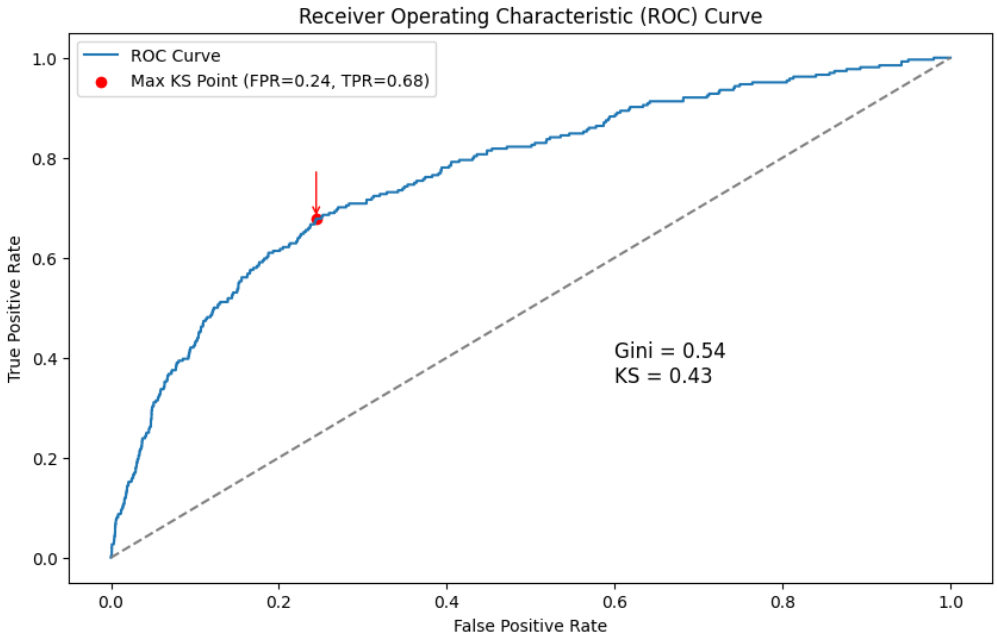
Vrednosti metrika dobijenih iz ove matrice su sleedeće

Accuracy	0.98215
Recall	0.03409
Precision	0.12162
F1 Score	0.05325

Tabela 12 - Metrike Modela 2

- **Model 3**

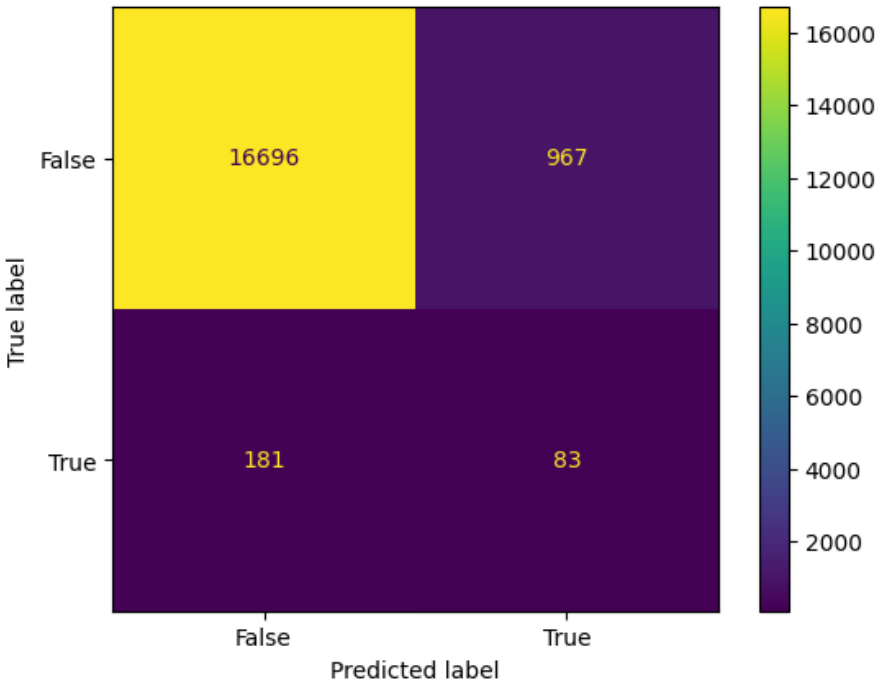
Gini i KS koeficijenti za model 1 na testnom kupu iznose 53.66% odnosno 43.36%



Slika 71 - ROC kriva sa GINI I KS vrednostima za Model 3

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

Na slici ispod je predstavljena matrica konfuzije ukoliko za PD cut-off koristimo vrednost koju smo ranije utvrdili.



Slika 72 - Matrica konfuzije za Model 3

Izvor: Grafički prikaz autora dobijen korišćenjem programskog koda u Python-u

Vrednosti metrika dobijenih iz ove matrice su sleedeće

Accuracy	0.93596
Recall	0.31439
Precision	0.07905
F1 Score	0.12633

Tabela 13 - Metrike Modela 2

Predstavimo sada rezultate sumarno. Sa jedne strane posmatramo Gini I KS koeficijente koji ne zahtevaju određen prag za PD cut-off i ocenjuju diskriminatornu moć modela za različite pragove za klasifikaciju. Sa druge strane posmatramo 4 pomenute metrike koje govore o kvalitetu modela nakon određivanja PD cut-off-a. Napomenimo da bi način na koji smo odredili granicu za PD imao i biznis smisao i da se sa novim modelima koji se javljaju jednom u godinu ili dve najverovatnije granica ne bi menjala već bi bila konstantna duži niz godina.

Posmatrajmo Gini i KS pokazatelje. Modeli 1 i 2 daju bolje rezultate po ove dve metrike. Model 1 ima neznatno niži Gini i veći KS koeficijent nego Model 2. Razlog za ovo može da leži u tome što su portfoliji stare i nove banke i dalje u dobroj meri slični, a Model 1 je treniran na većem skupu podataka. Važno je da primetimo da Model 3, dobijen korišćenjem Bejzovim zaključivanjem ne zaostaje bitno po 2 pomenute performanse u odnosu na 2 altrnativna modela.

	Gini	KS
Model 1	57.00%	46.16%
Model 2	57.48%	44.96%
Model 3	53.66%	43.36%

Tabela 14 - Metrike Modela 3

Posmatrajmo sada 4 metrike dobijene iz matrice konfuzije

	Accuracy	Recall	Precision	F1 Score
Model 1	0.97942	0.07576	0.13793	0.09780
Model 2	0.98215	0.03409	0.12162	0.05325
Model 3	0.93596	0.31439	0.07905	0.12633

Tabela 15 – Poređenje metrika modela

Premda bi se ovi rezultati razlikovali da smo drugačije odredili prag za PD, ovo predstavlja realnu sliku izazova sa kojim se banka može susreti. Sva tri modela imaju veliku tačnost, što u problemima skoringa ne igra veliku ulogu, jer se radi o nebalansiranim problemima klasifijkacije. Većina instance pripada negativnoj klasi i u tom slučaju tačnost ne predstavlja merodavnu metriku. Primetino da Model 3 ima daleko najbolji odziv, uz šta ide i niža preciznost. Ovo sa poslovne strane znači da bi model otkrio oko 31% default slučajeva, po cenu da za samo oko 8% slučajeva koji se procene kao default to zaista i važi. Banka ima slobodu da definiše strategiju kojom će optimizovati određene metrike. Ukoliko posmatramo F1 skor kao harmonijsku sredinu preciznosti i odziva, Model 3 sa ovim postavkama pokazuje najbolje performanse.

8. Zaključak

Posmatrajući rezultate dobijene u ovom radu vidimo zašto je logisitčka regresija i dalje glavni alat za modeliranje kreditnog skoringa. Uočavamo da ona daje jasne logične i dovoljno dobre rezultate. U teorijskom delu rada naveli smo i alternativne metode modeliranja kreditnog rizika. Njihov glavni nedostatak bi pored teže implementacije bio i nedostatak interpretabilnosti. Za razliku od toga logistička i statističke metode koje idu uz nju pruža mogućnost za procenu veličine efekta određene nezavisne promenljive na krajnji ishod kao i ocenu međusobne interakcije nezavisnih promenljivih.

Ovaj rad je imao za cilj da pored standardnog načina primene logističke regresije istu pokuša da posmatra iz drugog ugla, tj. ne na standardni frekvencionistički način već pomoću Bejzovog zaključivanja, metoda koje zadjih decenija postaju sve popularnije i popularnije. Ideja je da

poštujući ekspertska znanja, u ovom slučaju istorijska znanja o raspodeli iz koje potiču parametric model, na osnovu dostupnih podataka pokušamo da formiramo posteriorne ocene parametara logističke regresije. Iza svega stoji Bejzova teorema, a kao nužni alat nameću se Monte Karlo lanci Markova kao probablističko rešenje koje služi za uzorkovanje iz posteriorne raspodele.

Rad se završava poređenjem više modela, gde su dva dobijena standardnom logističkom regresijom a jedan primenom Bejzove logističke regresije. Ono što se nameće kao zaključak je da je Bejzov pristup logističkoj regresiji uporediv sa pristupom standardne logističke regresije, i da u našem slučaju daje skoro pa jednako dobre rezultate. Napomenimo da je model dobijen Bejzovom logističkom regresijom sadržao manje promenljivh uz slične performanse, što je u modeliranju skoringa veoma poželjno.

Smatramo da bi banke trebalo da razmtotre korišćenje Bejzove logističke regresije jer pored rezultata koji bi u zavisnosti od konkretnog korišćenog MCMC algoritma, veličine razvojnih, validacionih i testnih skupova mogli da budu i bolji, sami operativni troškovi mogu biti smanjeni. Ukolio se utvrdi da Bejzova logistička regresija daje rezultate ona može poslužiti za reviziju bankarskih modela, što bitno smanjuje troškve banke u poređenju sa razvojem modela od početka.

Ono što bi moglo biti predmet daljih istraživanja su poređenje različitih MCMC metoda u Bejozovoj logističkoj regresiji. Takođe veličine skupova i stopa default-a samih skupova korišććenih u modeliranju/validaciji/testiranju bi trebalo da bude predmet dublje analize. Na posletku kako se u bankama često koristi WoE (weight of evidence) (pristup logističkoj regresiji, gde se svaka promenljiva deli na kategorije, vrednosti koje su slične po rizičnosti) bilo bi interesantno posmatrati kako se i da li se priorna znanja o samim granicama kategorija ili WoE vrednostima mogu podvrgnuti Bejzovom zaključivanju.

Literatura

[1] Altman, E., Marco, G., and Varetto, F. (1994). Corporate distress diagnostics: Comparison using linear discriminant analysis and neural networks (the Italian Experience). Journal of Banking and Finance 18: 505-529.

[2] Beaver W."Financial ratios as predictors of failure", Empirical Research in Accounting, 1966

[3] D.W. Hosmer, S. Lemeshow: Applied Logistic Regression, 2nd Edition, John Wiley & Sons, Inc.,2000

[4] Desai, V.S., Crook, J.N., and Overstreet, G. (1996). Credit scoring models in credit union environment. European Journal of Operational Research 95: 24-35

[5] Dobson, A.J., Barnett, A.G. (2008). An Introduction to Generalized Linear Models. 3rd Edition. Taylor & Francis Group, Boca Raton, Florida.

[6] Durand, D. Credit-rating formulae. In: Risk Elements in Consumer Instalment Financing. pp. 83-91. National Bureau of Economic Research, Inc. Massachusetts (1941)

[7] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor: An Introduction to Statistical Learning with Applications in Python, Springer, 2023

[8] Gorica Gvozdić, Primenjena logistička regresija, Master rad, Univerzitet u Novom Sadu, Prirodno-matematički fakultet, Novi Sad (2011)

[9] Greenberg, E. (2008). Introduction to Bayesian Econometrics. Cambridge University Press, New York

[10] http://www.efos.unios.hr/kreditna-analiza/wp-content/uploads/sites/252/2013/04/4_rizici-u-bankama.doc.pdf

[11] https://www.ficoscore.com/ficoscore/pdf/Frequently-Asked-Questions-About-FICO-Scores.pdf

[12] https://www.ubs-asb.com/Portals/0/Casopis/2008/1_2/B01-02-2008-Ekoleks.pdf

[13] Komorád, K. (2002). On Credit Scoring Estimation. Master's Thesis. Humboldt University, Berlin

[14] Kroese, D.P., Taimre, T., Botev, Z.I. (2011). Handbook of Monte Carlo Methods. John Wiley & Sons, Inc., Hoboken, New Jersey.

[15] Mok, J-M. (2009). Reject Inference in Credit Scoring.

[16] N. Vunjak, Lj. Kovačević, Bankarski menadžment, Ekonomski fakultet Univerziteta u Beogradu, 2016

[17] Ntzoufras, I. (2009). Bayesian Modeling Using WinBUGS. John Wiley & Sons, Inc., Hoboken, New Jersey.

[18] Robert, C.P., Casella, G. (2004). Monte Carlo Statistical Methods. 2nd Edition. Springer-Verlag, New York.

[19] Sandra Rackov, Primena Koksovog PH modela u analizi kredintog rizika, Master rad, Univerzitetu Novom Sadu, Prirodno-matematički fakultet, Novi Sad (2013)

[20] Suess, E.A. and Trumbo, B.E. (2010). Introduction to Probability Simulation and Gibbs Sampling with R. Springer-Verlag, New York.

[21] Thomas, L.C, Consumer Credit Models: Pricing, Profit and Portfolios. Oxford University Press, Oxford (2009)

[22] Wilhelmsen, M., Dimakos, X.K., Husebø, T., Fiskaaen, M. (2009). Bayesian Modelling of Credit Risk using Integrated Nested Laplace Approximations

[23] Z.L. Crvenković: Bejzova statistika, Univerzitet u Novom Sadu, Prirodno-matematički fakultet, Novi Sad, 2015

[24] Ziemba, A. (2005). Bayesian Updating of Generic Scoring Models.

Kratka biografija

Autor ovog rada, Lazar Beić rođen je 12. Decembra 1994. godine u Valjevu. Osnovnu šlolu “Desanka Maksimović“ pohađao je u periodu od 2001. do 2009. godine, nakon čega upisuje specijalizovano matematičko odeljenje Valjevske gimazije. Po završetku gimnazije 2013. godine upisuje osnovne studije primenjene matematike na Prirodno-matamatičkom fakultetu u Novom Sadu. Završava ih 2017. godine nakon čega upisuje master studije na istom fakultetu.

UNIVERZITET U NOVOM SADU PRIRODNO-MATEMATIČKI FAKULTET KLJUČNA DOKUMENTACIJSKA INFORMACIJA

Redni broj:

RBR

Identifikacioni broj:

IBR

Tip dokumentacije: *Monografska dokumentacija*

TD

Tip zapisa: *Tekstualni štampani materijal*

TZ

Vrsta rada: *Master rad*

VR

Autor: *Lazar Beić*

AU

Mentor: *Prof. dr Zagorka Lozanov Crvenković*

MN

Naslov rada*: Bejzova logistička regresija kao alat za procenu kreditnog rizika*

NR

Jezik publikacije: *Srpski (latinica)*

JP

Jezik izvoda: *s/e*

JI

Zemlja publikovanja: *Republika Srbija*

ZP

Uže geografsko područje: *Vojvodina*

UGP

Godina: *2023.*

GO

Izdavač: *Autorski reprint*

IZ

Mesto i adresa: *Novi Sad, Trg Dositeja Obradovića 4*

MA

Fizički opis rada: *8 poglavlja, 141 stranica, 24 reference, 15 tabela, 72 slike*

FO

Naučna oblast: *Primenjena matematika*

NO

Naučna disciplina: *Statistička analiza*

ND

Ključne reči: *Logistička regresija, Bejzova statistika, Lanci Markova, Monte Karlo metode*

PO

UDK

Čuva se*: Biblioteka Departmana za matematiku i informatiku Prirodno- matematičkog fakulteta u Novom Sadu*

ČU

Važna napomena:

VN

Izvod: *Cilj ovog rada je bio da istraži potencijalnu primenu Bejzove logističke regresije u ocenjivanju kreditnog skoringa klijenata banke.U radu su predstavljeni osnovni principi logističke regresije kao najčešće metode za modeliranje kreditnog skoringa. Dobijeni su modeli logističke regresije na realnim podacima. Nakon toga je predstavljen koncept primene Bejzove teorije u modelu logističke regresije zajedno sa Monte Karlo simulacijama i Markovljevim lancima. Predstavljen je model logističke regresije koji se zasniva na Bejzovom zaključivanju, tačnije na proceni posteriornih ocena parametara regresije koristeći priorna znanja.*

IZ

Datum prihvatanja teme od strane NN veća: *29.08.2023.*

DP

Datum odbrane:

DO

Članovi komisije:

KO

Predsednik: *dr Ivana Štajner Papuga, redovni profesor*

Mentor: *dr Zagorka Lozanov Crvenković, redovni profesor*

Član: *dr Mirjana Štrboja, redovni profesor*

UNIVERSITY OF NOVI SAD FACULTY OF SCIENCES KEY WORD DOCUMENTATION

Accession number:

ANO

Identification number:

INO

Document type: *Monograph type*

DT

Type of record: *Printed text*

TR

Contents code: *Master thesis*

CC

Author: *Lazar Beić*

AU

Mentor: *Zagorka Lozanov Crvenković, PhD*

MN

Title: *Bayesian logistic regression as a tool for credit risk assessment*

XI

Language of text: *Serbian (latin)*

LT

Language of abstract: *s/e*

LA

Country of publication: *Republic of Serbia*

CP

Locality of publication: *Vojvodina*

LP

Publication year: 2023

PY

Publisher: Author’s reprint

PU

Publ. place: Novi Sad, Trg Dositeja Obradovića 4

PP

Physical description: 8 chapters, 141 pages, 24 references, 15 tables, 72 pictures

PD

Scientific field: Applied mathematics

SF

Scientific discipline: Statistical analysis

SD

Key words: Logistic regression, Bayesian statistics, Markov chains, Monte Carlo methods

UC

Holding data: Department of Mathematics and Informatics' Library, Faculty of Sciences, Novi Sad

HD

Note:

N

Abstract: The aim of this, master thesis was to investigate the potential application of Bayesian logistic regression in evaluating the credit scoring of bank clients.

The thesis presents the basic principles of logistic regression as the most common method for modeling credit scoring. Logistic regression models were obtained on real data. After that, the concept of applying the Bayesian theory in the logistic regression model together with Monte Carlo simulations and Markov chains is presented. A logistic regression model that is based on Bayesian inference is presented, more precisely on the evaluation of posterior estimates of regression parameters using prior knowledge.

AB

Accepted by the Scientific Board on: August 29, 2023

ASB

Defended:

DE

Thesis defended board:

DB

President: *Prof. Ivana Štajner Papuga, PhD*

Mentor: *Prof. Zagorka Lozanov Crvenković, PhD*

Member: *Prof. Mirjana Štrboja, PhD*