Analiza mjere pogreške predviđanja rizika modelom binomne logističke regresije

Ruščić, Toni

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Toni Ruščić

ANALIZA MJERA POGREŠKE PREDVIĐANJA RIZIKA MODELOM BINOMNE LOGISTIČKE REGRESIJE

Diplomski rad

Voditelj rada:
Doc. dr. sc. Vesna Lužar-Stiffler

Zagreb, srpanj 2019.

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

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

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Zahvaljujem se svojoj obitelji, roditeljima te svim prijateljima na podršci za vrijeme cijelog školovanja i pisanja ovog diplomskog rada

Sadržaj

Uvod

Ako je upravljanje financijskim institucijama moguće poistovijetiti s upravljanjem rizicima u financijskim institucijama tada se zbog prirode bankarskog poslovanja može reći kako upravljanje kreditinim rizikom predstavlja najznačajniju aktivnost u upravljanju bankama. Razvojem modernih financijskih tržišta banke sve veću pozornost pridaju kontroli rizičnosti njihovog poslovanja, ali još uvijek središnje mjesto u sustavu upravljanja rizicima pripada kreditnome riziku. Ovaj položaj dodatno naglašavaju domaći i međunarodni regulatorni propisi koji od banaka zahtijevaju konstantno usavršavanje i razvijanje različitih modela za procjenu i upravljanje rizikom. Modeli koje banke koriste polaze od klasične kreditne analize koja se temelji na subjektivnoj procjeni pa sve do kvantitativnih modela.

Kreditno bodovanje je statistička analiza koju obavljaju zajmodavci i financijske institucije za pristup kreditnoj sposobnosti osobe, tj. zajmoprimca. Zajmodavci koriste kreditno bodovanje, između ostalog, da odluče hoće li produžiti ili odbiti kredit. Sustav ocjenjivanja kreditnog rejtinga *Fair Isaac Corporation*, poznat kao FICO score, najraširenij je sustav bodovanja u financijskoj industriji. Kreditini rezultati po FICO rezultatu mogu poprimiti rezultate od 300 do 850. Zajmodavci koriste ocjenu kreditnog bodovanja u određivanju cijena na temelju rizika pri čemu se uvjeti zajma uključujući kamatnu stopu ponuđenu zajmoprimcima temelje na vjerojatnosti otplate. Općenito, što je bolja kreditna ocjena osobe, to je bolja stopa koju financijska institucija nudi pojedincu. Kao tradicionalni pristup analizi kreditnog rizika, ocjenjivanje kreditnog bodovanja je najučinkovitije za male tvrtke i pojedince. Sličan koncept, kreditni rejting, ne treba miješati s bodovanjem kredita. Kreditni rejting primjenjuje se na tvrtke, državna tijela, državne obveznice i vrijednosne papire tih subjekata, kao i na vrijednosne papire osigurane imovinom.

Postoji 5 glavnih čimbenika koji se procjenjuju pri izračunu kreditnog rezultata:

1. povijest plaćanja
2. ukupni iznos dugovanja
3. duljina kreditne povijesti
4. vrste kredita
5. novi kredit

U ovom diplomskom radu koristiti ćemo simulirane podatke uporabom Monte Carlo eksperimenta. Monte Carlo eksperiment predstavlja široku skupinu računalnih algoritama koji se oslanjaju na ponavljanje slučajnog uzorkovanja za dobivanje numeričkih rezultata. Njihova osnovna ideja je da se uz pomoć slučajnih događaja rješavaju problemi koji mogu biti deterministički. Oni se često koriste u fizičkim i matematičkim problemima, te su najkorisniji kada je teško ili nemoguće koristiti druge pristupe. Monte Carlo metode primjenjuju se uglavnom u tri različite klase problem: optimizacija, numeričke integracija i generiranje izvoda iz distribucije vjerojatnosti.

Glavni cilj rada je istražiti utjecaj broja podataka, tipa distribucije i jačine povezanosti prediktorskih varijabli na mjere pogreške predviđanja. Mjere pogreške predviđanja koje smo proučavali su: Somers'Delta, c statistika, stvarna pozitivna stopa (osjetljivost), stvarna negativna stopa (specifičnost) i stopa pogrešne klasifikacije. Binomna logistička regresija je, uz stabla odlučivanja i neuralne mreže, jedan od ključnih metoda za predviđanje rizika i najučestalije korištenih. Kroz ovaj rad objasniti ćemo korištenje binomne logističke regresije pri predviđanju rizika kao i ostale statističke i nestatističke metode koje se mogu koristiti.

Poglavlje 1

Credit scoring: kreditno bodovanje

Kreditno bodovanje je skup modela odluka te njihovih temeljnih tehnika koje pomažu zajmodavcima u odobravanju potrošačkih kredita. Te tehnike se koriste pri odlučivanju tko će dobiti kredit, koliko će kredita dobiti te koje će operativne strategije povećati profitabilnost zajmodavaca. Zajmodavac mora donijeti dvije vrste odluka. Prvo, treba li odobriti kredit novom podnositelju zahtjeva i drugo, kako se nositi s postojećim podnositeljima zahtjeva, uključujući i to trebali povećati svoja kreditna ograničenja. Tehnike koje pomažu u donošenju prve odluke nazivaju se *'credit scoring'*(kreditno bodovanje), dok se tehnike koje pomažu drugoj vrsti odluka nazivaju *'behavioral scoring'*(bodovanje ponašanja).

Kreditno bodovanje je jedna od najuspješnijih aplikacija modeliranja za statistička i operativna istraživanja u financijama i bankarstvu. Cilj kreditnog bodovanja je predvidjeti rizik od nemogućnosti povratka zajma. U ovom poglavlju pričati ćemo o povijesti kreditnog bodovanja te o izgradnji prvih modela kreditnog bodovanja.

1.1 Povijest kreditnog bodovanja

Ocjenjivanje kreditne sposobnosti bilo je ključno za rast potrošačkih kredita u posljednjih 50-60 godina. Bez točnog i automatskog alata za procjenu rizika zajmodavci potrošačkih kredita ne bi mogli proširiti svoje kreditne knjige na način na koji jesu. Kreditno bodovanje bilo je jedno od prvih razvijenih alata za upravljanje financijskim rizicima. Njegova uporaba od strane trgovaca na malo (*'retailers'*) u SAD-u i *'mail-order'* tvrtki u 1950-ima podudara se s ranim primjenama analize portfelja za upravljanje i diversifikaciju rizika svojstvenih investicijskim

portfeljima. Također, može se reći da je kreditno bodavanje preteča rudarenju podataka (*'data mining'*) jer je to bio jedan od prvih alata za korištenje podataka o ponašanju potrošača. U stvari, uobičajene tehnike koje se koriste kod rudarenja podataka - segmentacija, modeliranje sklonosti i klasteriranje, također su tehnike koje su korištene sa značajnim uspjehom i u ocjeni kreditnog bodovanja.

No, krenimo od početka, tijekom 1930-tih neke 'mail-order' tvrtke uvele su numeričke sustave ocjenjivanja kako bi pokušale prevladati nedosljednosti u odlučivanju o zajmovima među kreditinim analitičarima. Početkom drugog svjetskog rata došlo je do nestašice kreditnih analitičara zbog povojačenja te su stoga firme imale napisana pravila o odobravanju zajmova. Zaslugu za pokretanje koncepta kreditnog bodovanja daje se Davidu Durandu. Njegova studija objavljena 1941. godine od strane Nacionalnog biroa za ekonomska istraživanja ispitala je oko 7200 izvješća o dobrim i lošim zajmovima od strane 37 tvrtki. Koristeći hi-kvadrat test, Durand je identificirao varijable koje se značajno razlikuju između dobrih i loših zajmova i razvio 'indeks učinkovitosti' osmišljen kako bi pokazao koliko je varijabla bila učinkovita u razlikovanju dobrih od loših rizika među zajmoprimcima. Zatim je koristio diskriminantnu funkciju kako bi razvio modele kreditnog bodovanja.

Durandov rad u početku nije shvaćen pretjerano ozbiljno te se tradicionalna kreditna procjena oslanjala na 'osjećaj' i procjenu karaktera potencijalnog zajmoprimca, sposobnosti povrata zajma te kolaterala, tj. sigurnosti. To je značilo da potencijalni zajmoprimac nije pristupio banci ili drugoj ustanovi sve dok već nekoliko godina nije štedio ili koristio njene druge usluge. Ovaj proces je bio spor i nekonzistentan, a ponuda kredita je bila mala jer bi potencijalni zajmoprimac imao takav odnos sa samo jednim zajmodavcem. Ekstreman rast obročnih zajmova u poslijeratnom razdoblju daleko je nadmašio sposobnost industrije da zaposli obučeno osoblje koje će pregledati sve podnositelje zahtjeva za kredit. Konsolidacija u industriji učinila je ekonomičnijim razviti sustave ocjenjivanja, a tome je pridonijelo što su računala postala dostupna za mukotrpne izračune potrebne za razvoj sustava kreditnog bodovanja. Tako su se ekonomski pritisci i računalna tehnologija spojili tijekom kasnih 1960-ih i ranih 1970-ih, što je rezultiralo razvijanjem sustava kreditnog bodovanja koji se temelji na empirijskim metodama vrednovanja i koji su postupno prihvaćeni. Prvo savjetovalište koje se time bavilo osnovali su Bill Fair i Earl Isaac početkom 1950-tih.

Tijekom 1980-ih u Velikoj Britaniji dogodile su se mnoge promjene u kreditnom okruženju. Neke od tih promjena bile su sljedeće:

- banke su znatno promijenile svoj tržišni položaj i počele plasirati svoje proizvode. To je značilo da su morali prodavati proizvode kupcima, ne samo onima koje jedva poznaju, već i onima koje su namamili

- pojavio se velik rast u kreditnim karticama. Ovlaštenja za prodaju ovog proizvoda značila su da mora postojati mehanizam za vrlo brzo donošenje odluke o posuđivanju. Također, obujam zahtjeva bio je takav da kreditni analitičari ne bi imali vremena ili mogućnosti da intervjuiraju sve podnositelje zahtjeva

- fokus na potrošačko kreditiranje. Prethodno su se banke gotovo isključivo fokusirale na velike kreditne i korporativne klijente i cilj im je obično bio izbjeći bilo kakve gubitke. Međutim, banke su počele shvaćati da cilj potrošačkog kreditiranja ne bi trebao biti izbjegavanje gubitaka, već da se maksimizira profit preuzimanjem male kontrolirane razine loših dugova i tako proširiti potrošačku kreditnu knjigu. To je i i dalje bio vrijednosno manjinski udio, ali postajao je značajniji.

Sistemi kreditnog bodovanja bazirani su na učincima prethodnih klijenata koji koriste istu uslugu kao i procjenjivani. Ocjenjivanje kreditnog bodovanja ne procjenjuje kreditnu sposobnost pojedinaca, već rizik povezan s grupama ljudi. Dakle, možete upasti u skupinu 'lošeg rizika', iako savjesno plaćate. Takav način procjene je pogađao potrošače jer su promatrani hladno i neosobno.

Događaj koji je osigurao potpuno prihvaćanje kreditnog bodovanja je donošenje *Zakona o jednakim kreditnim mogućnostima* (Equal Credit Opportunity Acts) i njegovih izmjena i dopuna u SAD-u 1975. i 1976. godine. To je zabranilo diskriminaciju u odobravanju kredita, tj. korištenje varijabli kao što su rasa, boja kože, spol, vjera, bračno stanje i slično osim ako je diskriminacija 'empirijski izvedena i statistički valjana'. Odobriti dizajnerima sistema kreditnog bodovanja da razmotre sve zabranjene varijable i uključe ih ako su statistički valjani. Tako bi žena koja je bila odbijena jer je radila sa skraćenim radnim vremenom bila ocijenjena zajedno s drugim ženama koje rade s nepunim radnim vremenom. Neizravna diskriminacija bi bila prigušena, a vjerovnici bi odobravali više kredita ženama u vlastitom ekonomskom interesu.

Dva najpouzdanija alata današnjih kreditnih analitičara logistička regresija i linearno programiranje počinju se koristit 1980-tih godina. U novije vrijeme naglasak je na promjeni ciljeva s minimiziranja mogućnosti da kupac neće moći vratiti određeni proizvod na to da se promatra kako tvrtka može maksimizirati dobit koju može ostvariti od tog klijenta.

1.2 Modeli kreditnog bodovanja

Uzet ćemo jedan konkretan jednostavan primjer [1]. Pretpostavimo da imamo karticu s četiri varijable: stambeni status, dob, svrhu kredita i vrijednost presuda županijskih sudova. (*county court judgements*)

stambeni status		**—dob**	
vlasnik	36	18-25	22
podstanar	10	26-35	25
živi s roditeljima	14	36-43	34
drugo navedeno	20	44-52	39
nema odgovora	16	53+	49

svrha kredita		**—vrijednost CCj-a**	
novi auto	41	$0	32
rabljeni auto	33	$1 - $299	17
preuređivanje doma	36	$300 - $599	9
godišnji odmor	19	$600 - $1199	-2
ostalo	25	$1200+	-17

20-godišnjak koji živi s roditeljima i želi posuditi novac za rabljeni auto te nema nikakvih presuda po ovom sistemu ostvariti će 101 (14+22+33+32) bod. 55-godišnjak koji posjeduje kuću, ima $250 presuda i želi posuditi novac za kćerino vjenčanje ostvariti će 127 (36+49+25+17) bodova. Prilikom uspostavljanja sustava bodovanja potrebno je odrediti koja je prolazna ocjena, tj. prag. Pretpostavimo da je u gornjem primjeru oznaka za prolaz 100 bodova. Dakle, svaki klijent koji ima 100 ili više bodova dobiva preporuku za odobrenje zajma bez obzira na odgovor na četiri pitanja. Prema tome, ocjenjivanje omogućuje kompromis tako da se slabost jednog faktora može nadoknaditi snagom drugih faktora. Neki zajmodavci ponekad umjesto jednog mogu staviti više pragova. Visoki prag za definiranje najboljih kandidata kojima bi se mogao ponuditi nadograđeni proizvod, drugi prag za standardni proizvod po nižoj kamatnoj stopi, treći prag za standardni proizvod po standardnoj cijeni i četvrti niski prag za umanjeni proizvod.

Ranije smo uveli pojam savjetovališta o kreditnom bodovanju. Usluge koje oni mogu pružiti zajmodavcu jest praćenje u vidu *monitoringa* i *trackinga*. Monitoring sustava bodova-

nja je skup aktivnosti koje su uključene u ispitivanje trenutne serije aplikacija i novih računa te procjenu koliko su bliske nekim mjerilima, dok tracking uključuje praćenje grupa računa kako bi se vidjelo kako se izvode i jesu li predviđanja sustava bodovanja ispunjena.

Iz pregleda različitih pristupa kako bi se formulirali ciljevi za odabir kreditnog modela, čini se da postoji nekoliko ključnih elemenata za odluku o odobravanju kredita koja ima važne implikacije za izradu modela kreditnog bodovanja.

Prvo, odluka o odobravanju kredita je zapravo višeperiodni problem koji može ovisno o vrstama uključenih kredita imati dvije dimenzije. U kontekstu Bierman-Hausmanovog modela, odobravanje kredita u jednom periodu je samo dio odnosa s klijentima koji se proteže kroz mnogo perioda. Odluka o odobravanju kredita utječe na vrijednost odnosa s tim klijentom kako tijekom razdoblja kredita, tako i tijekom cijelog trajanja odnosa s klijentom. Sasvim je jasno da je pružanje kredita povezano s pružanjem drugih financijskih i nefinancijskih usluga od strane vjerovnika u razdobljima koja su veća od trajanja samog zajma. Odluka o kreditu je doista višeperiodna odluka tijekom koje zajam generira tijek prihoda sve dok se zajam ne isplati u cijelosti ili prestane izvršavati obvezu. Stoga se vrijednost zajma određuje ne samo prema tome je li zajam plaćen u cijelosti, već i u slučaju neispunjavanja obveza, po dužini trajanja, troškovima naplate i ostvarivoj vrijednosti kolaterala. U takvim okolnostima čak i ako je 100 posto sigurno da zajam ne može biti isplaćen u cijelosti ipak može biti vrijedno odobravanja zajma ako je važeć dovoljno dugo. To je osobito istinito ako su troškovi naplate niski, a kolateral ima dovoljnu vrijednost oporavka.

Drugo, postoji potreba za razmatranjem troškova prikupljanja informacija u formuliranju sheme odobravanja kredita. U onoj mjeri u kojoj su informacije o dosadašnjoj kreditnoj uspješnosti relevantne za procjenu budućih kreditnih rezultata, postoji potreba za vaganjem između vrijednosti prikupljanja dodatnih informacija i veličine troškova prije odlučivanja kada je prikladno vrijeme za donošenje kreditne odluke.

Treće, većina modela kreditne politike izričito ili implicitno razmatra pretpostavku da potencijalni vjerovnik procjenjuje rizičnost zajmoprimca u smislu ili vjerojatnosti neispunjavanja obveza ili oportunitetnih troškova pri odlučivanju o tome hoće li odobriti zajam ili ne.
Konačno, u višeperiodnim modelima s više odluka može postojati potreba za ponovnom procjenom očekivanih vjerojatnosti neispunjavanja obveza tijekom vremena na temelju uspješnosti klijenta.

Općenito, razvijeni modeli kreditnog bodovanja usredotočili su se na dvije kategorije kredita:

- potrošačke kredite, uključujući kreditne kartice za otplatu na rate
- komercijalne zajmove, uključujući zajmove na određeno vrijeme, redovne komercijalne i industrijske zajmove te zajmove manjinama i malim poduzećima.

Tipičan pristup je kategorizirati uzorke kredita u dvije međusobno isključive skupine - 'dobre zajmove' koji su oni koji će biti plaćeni ili su tekući i 'loše zajmove' koji su sporo plaćeni, delinkventni ili u kašnjenju. Obično je diskriminativna funkcija procijenjena iz skupa zajmova koji su već odobreni. Tada se formulira pravilo klasifikacije koje je dizajnirano za razlikovanje između skupina dobrih i loših zajmova umanjujući ukupnu stopu pogreške ili troškove pogrešne klasifikacije.

Većina prvotnih modela pati od statističkih problema koji mogu utjecati na pouzdanost procjena vjerojatnosti neplaćanja. Mogu se kategorizirati u sedam različitih vrsta:

1. kršenje pretpostavke o temeljnim distribucijama varijabli
2. korištenje linearnih diskriminativnih funkcija umjesto kvadratnih funkcija kada su grupne disperzije nejednake
3. neprikladna interpretacija uloge pojedinih varijabli u analizi
4. redukcije dimenzionalnosti
5. problemi u definiranju grupa
6. korištenje neprikladnih a priori vjerojatnosti i / ili troškovi pogrešne klasifikacije
7. problemi u procjeni stope klasifikacijske pogreške za procjenu uspješnosti modela

Poglavlje 2

Metode za izgradnju kreditnih kartica

U ovom poglavlju uvesti ćemo statističke i nestatističke metode za izgradnju kreditnih bodovanja i odluku da li pozajmiti kredit ili ne.

2.1 Statističke metode

Statističke metode su daleko najčešće metode za izgradnju kreditnih ocjena. Njihova prednost je ta što nam dopuštaju koristiti znanja o svojstvima procjenitelja uzoraka i alatima pouzdanih intervala te testiranje hipoteza u kontekstu kreditnog bodovanja. Stoga je moguće komentirati snagu ocjene rezultata i relativnu važnost različitih karakteristika koje su sadržane u ocjenama. Ove statističke tehnike omogućuju nam da identificiramo i uklonimo nevažne karakteristike.

U početku, metode su se temeljile na diskriminativnim metodama koje je predložio Fisher (1936.) za opće probleme klasifikacije. To je vodilo ka linearnim ocjenama baziranim na Fisherovoj linearnoj diskriminativnoj funkciji. Pretpostavke koje su bile potrebne kako bi se osiguralo da je to najbolji način za razlikovanje dobrih i loših potencijalnih kupaca bile su iznimno restriktivne i očigledno se nisu održavale u praksi. Fisherov pristup se mogao promatrati kao oblik linearne regresije, a to je dovelo do istraživanja drugih oblika regresije koje su imale manje restriktivne pretpostavke kako bi se zajamčila njihova optimalnost, ali i zadržala linearnost u pravilima ocjenivanja.
Daleko najuspješnija od njih je logistička regresija koja je i fokus ovog diplomskog rada, a posebno dihotomna, tj. binomna logistička regresija. Osim binomne logističke regresije o kojoj ćemo nešto više kasnije imamo i nominalnu logističku regresiju gdje je zavisna varijabla, tj.

varijabla odgovora kategorijska sa tri ili više moguća ishoda. Također, koristi se i ordinalna logistička regresija u slučaju ako varijabla odgovora sadrži tri ili više kategorija koje imaju prirodan redoslijed, kao što su izrazito neslaganje, neslaganje, neutralnost, suglasnost i potpuna suglasnost.

Uz logističku regresiju, veliku primjenu imaju i stabla odlučivanja, tj. rekurzivni algoritam particioniranja. Skup podataka A prvo se podijeli u dva podskupa tako da su gledajući uzorak prethodnih podnositelja zahtjeva, ta dva nova podskupa daleko više homogena u riziku neispunjavanja obveza podnositelja zahtjeva od izvornog skupa. Svaki od ovih skupova se zatim ponovno dijeli na dva da bi proizveo još homogenije podskupove. Proces se zaustavlja kada podskupovi zadovoljavaju zahtjeve da budu terminalni čvorovi stabla. Svaki terminalni čvor je tada klasificiran kao član dobrih klijenata A_G ili loših A_B i cijela procedura se može grafički prikazati pomoću stabla.
Tri odluke čine proceduru klasifikacijskog stabla:

- pravilo podjele - koje pravilo treba koristiti za podjelu skupa u dva podskupa
- pravilo zaustavljanja - kako odlučiti da je skup terminalni čvor
- kako podijeliti terminalne čvorove u dobre i loše kategorije

Definicija 1. *Prosječan gubitak D je dug nevraćenog zajma nastao pogrešnim klasificiranjem lošeg klijenta kao dobrog.*
Prosječan izgubljena dobit L je izgubljeni profit nastao pogrešnim klasificiranjem dobrog klijenta kao lošeg.

Za odlučiti jeli terminalni čvor pripada dobroj ili lošoj kategoriji potrebno je minimizirati trošak pogrešnog klasificiranja. Terminalni čvor je 'dobar' ako omjer 'dobrih' i 'loših' u uzorku tog čvora premašuje $\frac{D}{L}$.
Najjednostavnija pravila podjele su ona koja gledaju samo jedan korak naprijed od promatranog mjesta podjele. To se čini tako da pronalazimo najbolje podjele za svaku karakteristiku imajući određenu mjeru koliko je neka podjela dobra. Tada odlučujemo koja je karakteristika najbolja s obzirom na podjelu pod određenom mjerom. Najčešća mjera za odabir karakteristike je Kolmogorov-Smirnovljeva statistika, ali često se koriste i:
osnovni indeks nečistoće, Ginijev indeks, indeks entropije i pola sume kvadrata.

U ovu skupinu statističkih metoda ulazi i neparametarski pristup temeljem na najbližim susjedima. Ideja je odabrati metriku u prostoru podataka aplikacije kako bi se izmjerilo koliko

su udaljena dva kandidata. Zatim s uzorkom prošlih podnositelja zahtjeva kao reprezentativnim standardom, novi se podnositelj klasificira kao 'dobar' ili 'loš' ovisno o omjerima 'dobrih' i 'loših' među k najbližim podnositeljima zahtjeva iz reprezentativnog uzorka- najbližim susjedima novog kandidata

Definicija 2. *Neka je $X \neq \emptyset$ neprazan skup i $d : X \times X \to \mathbb{R}$ preslikavanje sa Kartezijevog produkta $d : X \times X$ u skup realnih brojeva $\mathbb{R}$ za koje vrijedi:*

1. $d(a, b) \geq 0, \quad \forall a, b \in X$ *(pozitivnost)*
2. $d(a, b) = 0 \Leftrightarrow a = b$ *(strogost)*
3. $d(a, b) = d(b, a), \quad \forall a, b \in X$ *(simetričnost)*
4. $d(a, c) \leq d(a, b) + d(b, c), \quad \forall a, b, c \in X$ *(nejednakost trokuta)*

Tri parametra potrebna za korištenje ovog pristupa su metrika, veličina skupa najbližih susjeda k te koji udio najbližih susjeda bi trebao biti 'dobar' da podnositelj zahtjeva bude klasificiran kao dobar.
Novi podnositelj zahtjeva se može klasicifirati kao dobar samo ako je najmanje $\frac{D}{D+L}$ najbližih susjeda klasificirano kao dobar.
Henley i Hand (1996) su se koncetrirali na metriku koja je je mješavina Euklidske metrike i udaljenosti u smjeru koji najbolje razdvaja 'dobre' i 'loše'. Smjer se dobije iz Fisherove linearne diskriminativne funkcije, a metrika koji su koristili izgleda ovako:

$$d(x_1, x_2) = \{(x_1 - x_2)^T(\mathbf{I} + D\mathbf{w} \cdot \mathbf{w}^T)(x_1 - x_2)\}^{\frac{1}{2}} \tag{2.1}$$

,gdje je $\mathbf{I}$ identiteta, a $\mathbf{w}$ p-dimenzionalni vektor koji definira smjer.

2.2 Nestatističke metode

Do 1980-ih godina jedini su pristupi bili statistički, ali je tada shvaćeno da se pronalaženje linearne funkcije za karakteristike koje najbolje razlikuju skupine mogu modelirati kao problem linearnog programiranja. Pristup linearnog programiranja mjeri koliko je dobra prilagodba uzimajući zbroj apsolutnih pogrešaka ili maksimalnu pogrešku.
Zajmodavac treba donijeti odluku tako da podijeli skup svih kombinacija vrijednosti A s vrijednostima u skupu $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ u 2 podskupa A_G- odgovori 'dobrih' kandidata i A_B-

odgovori 'loših' kandidata. Pretpostavimo da imamo uzorak od n prijašnjih aplikanata i pretpostavimo da je prvih n_G 'dobrih', a ostalih $n_B \quad i = n_G + 1, \ldots, n_G + n_B$ 'loših'. Pretpostavimo da i-ti aplikant ima karakteristike $(x_{i1}, x_{i2}, \ldots, x_{ip}) \in \mathbf{X}$. Želimo odabrati pondere $(w_1, \ldots, w_p)$ tako da ponderirana suma $w_1X_1 + w_2X_2 + \cdots + w_pX_p$ bude povrh granične vrijednosti c za 'dobre' aplikante i ispod granične vrijednosti za 'loše' aplikante. Tada linerani program koji minimizira zbroj apsolutnih grešaka izgleda ovako:

$$\begin{aligned} min \quad & a_1 + a_2 + \cdots + a_{n_G+n_B} \\ t.d \quad & w_1x_{i1} + w_2x_{i2} + \cdots + w_px_{ip} \geq c - a_i, && 1 \leq i \leq n_G, \\ & w_1x_{i1} + w_2x_{i2} + \cdots + w_px_{ip} \leq c + a_i, && n_G + 1 \leq i \leq n_G + n_B, \\ & a_i \geq 0 && 1 \leq i \leq n_G + n_B. \end{aligned} \tag{2.2}$$

Slično izgleda i linearni program koji minimizira maksimalnu pogrešku.

Modeli linearnog programiranja minimiziraju zbroj odstupanja u kreditnoj ocjeni onih koji su pogrešno klasificirani. Međutim, praktičniji kriterij bi bio minimiziranje broja pogrešnih klasifikacija ili ukupnog troška pogrešnog klasificiranja D i L. Izgradnja modela je slična kao i kod lineranog programiranja samo što će neke varijable biti cijeli brojevi. Ova tehnika se zove cjelobrojno programiranje.

Osamdesetih godina iznenada je do izražaja došla druga varijanta pristupa klasifikaciji problema temeljenih na umjetnoj inteligenciji. Neuronske mreže su načini modeliranja procesa odlučivanja slično načinu na koji stanice u mozgu koriste neurone da se aktiviraju i tako uspostavljaju mehanizme učenja. Sustav procesnih jednica su povezane zajedno, od kojih svaka odašilje izlazni signal kada primi ulazne signale. Sustav pokušava iz ovih podataka naučiti kako reproducirati odnos između ulaznih i izlaznih signala podešavanjem načina na koji svaka procesna jedinica povezuje svoj izlazni signal s odgovarajućim ulaznim signalom. Ako se ulazni signali smatraju karakteristikama kupca, a izlazni signal je da li je kreditna izvedba dobra ili loša, tada možemo ovaj pristup upotrijebiti u kreditnom bodovanju. Gore opisani postupak opisuje jednostruku neuralnu mrežu, dok još postoje i višestruke neuralne mreže. Višeslojni perceptron sastoji se od ulaznog sloja signala, izlaznog sloja izlaznih signala i mnogobrojnih slojeva neurona između nazvanih skriveni slojevi. Izlazni signali iz svakog neurona u skrivenom sloju imaju pondere glavnih karakteristika i postaju ulazi za neurone u sljedećem skrivenom sloju i analogno tako do izlaznog sloja. Najčešće korištena metoda za izračun vektora težina je algoritam povratnog širenja (*back-propagation algorithm*).

Poglavlje 3

Generalizirani linearni modeli

U ovom poglavlju pobliže ćemo opisati generalizirane linearne modele (Vidi [4]), procese u izradi modela te na kraju samu logističku regresiju koju ćmo koristiti u eksperimentu. Generalizirani linerani modeli uključuju posebne slučajeve kao što su linearna regresija, ANOVA (analiza varijance), logit modeli (binomna logistička regresija), probit modeli, log-linear modeli i slični.
Uzmimo jedan vrlo jednostavan model linearne regresije

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

gdje su varijable y i x povezane ravnom linijom parametrima α i β. Varijabla y naziva se i zavisna varijabla, a varijabla x nezavisna varijabla. Cilj nam je zamijeniti podatke y sa prilagođenim vrijednostima $\hat{\mu}$ izvedenih iz modela. Ove prilagođene vrijednosti su izabrane da bi minimizirali kriterije mjere nepodudarnosti ko što je suma kvadrata $\sum_i (y_i - \hat{\mu}_i)^2$.
Koristeći onoliko parametara kao što ih je promatranom modelu možemo napraviti savršeni fit, ali pri tome nismo smanjili složenost modela. Stoga je jednostavnost, tj. škrtost parametara poželjna značajka svakog modela. Ne uključujemo parametre koji su nam neznačajni. Ne samo što 'škrti' model omogućuje istraživaču ili analitičaru podataka da razmisli o svojiim podacima, već onaj koji je u biti ispravan daje bolja predviđanja od onoga koji uključuje nepotrebne dodatne parametre.

Važno svojstvo modela je i njegov opseg (*scope*), tj. raspon uvjeta nad kojima model daje dobra predviđanja. Teško je formalizirati opseg, ali ga je lako prepoznati, a intuitivno je jasno da su opseg i štedljivost donekle povezani. I opseg i štedljivost su povezani s invarijantnosti

parametara, tj. s vrijednostima parametara koji se ili ne mijenjaju s promjenom nekih vanjskih uvjeta ili koji se mijenjaju na predvidiv način.

3.1 Procesi u prilagodbi modela

Razlikujemo 3 procesa u prilagodbi modela.

Prvi proces je odabir modela (*model selection*). Pri odabiru modela potrebno je zadovoljiti dvije pretpostavke.
Prva je pretpostavka da su podaci međusobno neovisni. Kao posljedica toga, izričito su isključeni podaci koji pokazuju autokorelacije vremenskih serija i prostornih procesa. Ova pretpostavka neovisnosti karakteristična je za linearne modele klasične regresijske analize i prenosi se bez modifikacija na širu klasu generaliziranih linearnih modela.
Druga pretpostavka je da u modelu imamo samo jednu varijablu za pogrešku.

Izbor skale za analizu važan je aspekt odabira modela. Uobičajeni izbor je između analize Y, tj. izvorne skale i $\log Y$. U klasičnoj linearnoj regresijskoj analizi dobra skala bi trebala kombinirati konstantost varijance, približnu normalnost pogrešaka i aditivnost sustavnih učinaka. Nema razloga unaprijed vjerovati da takva skala postoji, i nije teško zamisliti slučajeve u kojima to ne postoji. Primjerice, u analizi diskretnih podataka gdje su greške dobro aproksimirane Poissonovom distribucijom, sustavni učinci su često multiplikativni. Ovdje $Y^{\frac{1}{2}}$ daje približnu konstantnost varijance, $Y^{\frac{2}{3}}$ bolje aproksimira približnu simetriju ili normalnost, a $\log Y$ daje aditivnost sustavnih učinaka. Očito, niti jedna skala neće istovremeno proizvoditi sva željena svojstva.
Uvođenjem generaliziranih linearnih modela problemi skaliranja uvelike se smanjuju. Normalnost i postojanost varijance više nisu potrebni iako je način na koji varijanca ovisi o srednjoj vrijednosti mora biti poznat. Aditivnost učinaka, iako je još uvijek važna komponenta svih generaliziranih linearnih modela, može se specificirati transformirana ljestvica ako je potrebno.
Najveći problem u odabiru modela je izbor x-varijabli koje će biti uključene u model. Ako su nam dani kandidati $x_1, \ldots, x_p$, potrebno je odabrati podskup koji je najbolji za izračun prilagođenih vrijednosti.

$$\hat{\mu} = \sum x_j \hat{\beta}_j$$

Potrebno je pronaći ravnotežu između poboljšavanja fita promatranih podataka dodavanjem termina modelu i obično nepoželjnog povećanja složenosti zbog dodavanja dodatnog termina.

Drugi proces je procjena parametara (*parameter estimation*). Procjena se odvija definiranjem mjere ispravnosti prilagodbe (*goodness of fit*), tj. mjere nepodudarnosti između promatranih podataka i prilagođenih vrijednosti koje model generira. Procjene parametara su vrijednosti koje minimiziraju kriterij ispravnosti prilagodbe. Ako je $f(y;\theta)$ funkcija gustoće ili funkcija distribucije vjerojatnosti za observaciju y s obzirom na parametar θ, tada je *log* vjerojatnost iskazana kao funkcija parametra srednje vrijednosti $\mu = E(Y)$:

$$l(\mu; y) = log\, f(y;\theta) \tag{3.1}$$

log vjerojatnost skupa nezavisnih observacija $y_1, \ldots, y_n$ je suma pojedinačnih doprinosa

$$l(\boldsymbol{\mu}; \mathbf{y}) = \sum_i log f_i(y_i; \theta_i)$$

gdje je $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$.
Postoje prednosti u korištenju funkcije gustoće kao kriterij ispravnosti prilagodbe, ali ne u obliku *log* vjerojatnosti $l(\boldsymbol{\mu}; \mathbf{y})$, već u obliku posebne linearne funkcije

$$D^*(\mathbf{y}; \boldsymbol{\mu}) = 2l(\mathbf{y}; \mathbf{y}) - 2l(\boldsymbol{\mu}; \mathbf{y}) \tag{3.2}$$

koju zovemo umanjena devijacija (*scaled deviance*). $l(\mathbf{y}; \mathbf{y})$ je najveća vjerojatnost koja se može postići za točno uklapanje gdje su prilagođene vrijednosti jednake promatranim podacima. Jer $l(\mathbf{y}; \mathbf{y})$ nije ovisna o parametrima tada maksimiziranje $l(\boldsymbol{\mu}; \mathbf{y})$, tj. jednadžbe 3.1 je ekvivalentno minimiziranju umanjene devijacije $D^*(\mathbf{y}; \boldsymbol{\mu})$, tj. jednadžbe 3.2 u odnosu na $\boldsymbol{\mu}$, ovisno o ograničenjima koja nameće model.

Za normalne linearne regresijske modele s poznatom varijancom σ^2 imamo:

$$f(y;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(y-\mu)^2}{2\sigma^2}\Big), \tag{3.3}$$

pa je *log* vjerojatnost:

$$l(\mu; y) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y-\mu)^2}{2\sigma^2},$$

i konačno funkcija umanjene devijacije izgleda ovako:

$$D^*(y;\mu) = \frac{(y-\mu)^2}{\sigma^2}.$$

Osim toga, iz poznatog faktora σ^2, devijacija u ovom slučaju je identična rezidualnoj sumi kvadrata, a minimalna devijacija najmanjem kvadratu.

Treći proces je predviđanje budućih vrijednosti (*prediction of future values*). Predviđanje se bavi tvrdnjama o vjerojatnim vrijednostima nespoznatih događaja, a ne nužno onima u budućnosti. Primjerice, nakon analize učestalosti srčanih oboljenja na nacionalnoj razini, podataka koji se kategoriziraju po regijama i dobnim skupinama, tipično pitanje bi bilo što ako bi određeni grad imao istu dobnu strukturu kao zemlja u cjelini? Bili bile iste predviđene vrijednosti?
Ovdje se riječ kalibracija često koristi za razlikovanje inverznih problema predviđanja od uobičajenijeg tipa u kojima je odgovor fiksan i od nas se traži da pronađemo koje su vrijednosti za x najvjerojatnije.

3.2 Komponente generaliziranih linearnih modela

Generalizirani linearni modeli su nastali kao nastavak klasičnih linearnih modela. Sastoji se od slučajne i sustavne komponente. Slučajna varijabla $\mathbf{Y}$ je vektor zapažanja veličine n čije su vrijednosti neovisno distribuirane sa srednjom vrijednosti $\boldsymbol{\mu}$. Sustavni dio modela je specifikacija vektora $\boldsymbol{\mu}$ u smislu nepoznatih parametara $\beta_1, \ldots, \beta_p$. U slučaju običnih linearnih modela, ova specifikacija ima oblik:

$$\boldsymbol{\mu} = \sum_{j=1}^{p} x_j \beta_j, \tag{3.4}$$

gdje su $\beta_1, \ldots, \beta_p$ parametri čije su vrijednosti obično nepoznate i moraju se procijeniti iz danih podataka. Ako označimo s i indeks opažanja tada se sustavni dio modela može zapisati:

$$E(Y_i) = \mu_i = \sum_{j=1}^{p} x_{ij} \beta_j, \qquad i = 1, \ldots, n \tag{3.5}$$

Matrično to možemo zapisati kao $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ gdje je $\boldsymbol{\mu}$ dimenzije n $\times$ 1, $\mathbf{X}$ dimenzije n $\times$ p i $\boldsymbol{\beta}$ dimenzije p $\times$ 1.

Prelaskom s klasičnih na generalizirane modele, model proširujemo na tri komponenete:

1. slučajna komponenta: $\mathbf{Y}$ se neovisno normalno distribuira s $E(\mathbf{Y}) = \boldsymbol{\mu}$ i konstantnom varijancom σ^2.

2. sustavna komponenta: kovarijable $x_1, \ldots, x_p$ čine linearni prediktor $\boldsymbol{\eta}$:

$$\boldsymbol{\eta} = \sum_{j=1}^{p} x_j \beta_j$$

3. veza između slučajne i sustavne komponente:

$$\boldsymbol{\mu} = \boldsymbol{\eta}$$

Ako to zapišemo kao $\eta_i = g(\mu_i)$ onda $g(\cdot)$ zovemo funkcija povezivanja.

U ovoj formulaciji, klasični linearni modeli imaju normalnu (ili Gaussovu) raspodjelu u komponenti 1 i funkciju identiteta za vezu u komponenti 3. Generalizirani linearni modeli dopuštaju dva proširenja; prvo, raspodjela u komponenti 1 može doći iz eksponencijalne obitelji koja nije normalna i drugo funkcija veze u komponenti 3 može biti bilo koja monotona diferencijabilna funkcija.

Pretpostavimo da svaka komponenta od $\mathbf{Y}$ ima distribuciju iz familije eksponencijalnih funkcija:

$$f_Y(y;\theta;\phi) = \exp\Big(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\Big) \tag{3.6}$$

za neke funkcije $a(\cdot), b(\cdot)$ i $c(\cdot)$.
Tada za normalnu distribuciju iz 3.3 i 3.6 imamo da je

$$\theta = \mu,\ \phi = \sigma^2$$

te

$$a(\phi) = \phi, \quad b(\theta) = \frac{\theta^2}{2}, \quad c(y,\phi) = -\frac{1}{2}\Big(\frac{y^2}{\phi} + \log(2\pi\phi)\Big)$$

Tada iz log vjerojatnosne funkcije $l(\theta,\phi;y) = log f_Y(y;\theta,\phi)$ te iz relacija

$$E\Big(\frac{\partial l}{\partial \theta}\Big) = 0 \tag{3.7}$$

$$E\Big(\frac{\partial^2 l}{\partial \theta^2}\Big) + E\Big(\frac{\partial l}{\partial \theta}\Big)^2 = 0 \tag{3.8}$$

dobivamo da je

$$E(Y) = \mu = b'(\theta) \quad i \quad var(Y) = b''(\theta)\, a(\phi)$$

.
Parametar θ zovemo kanonski parametar, dok ϕ zovemo parametar disperzije.

Funkcija povezivanja povezuje linearni prediktor η s očekivanom vrijednosti μ varijable y. Za binomnu distribuciju imamo da je $0 < \mu < 1$ i veza bi trebala zadovoljiti uvjet da mapira interval (0,1) na skup $\mathbb{R}$. Tri najpoznatije funkcije povezivanja su:

1. *logit* - $\eta = \log \frac{\mu}{1-\mu}$
2. *probit* - $\eta = \Phi^{-1}(\mu)$, gdje je $\Phi(\cdot)$ normalna kumulativna funkcija distribucije
3. *komplemenatarna log-log* - $\eta = \log\{-\log(1-\mu)\}$

od kojih je nama najzanimljivija prva jer ćemo se njome baviti u ovom diplomskom radu.

3.3 Modeli s binarnim odgovorima

Sada ćemo obratiti pozornost na regresijske modele za dihotomne podatke, uključujući logističku regresiju i probit analizu. Ovi modeli su prikladni kada odgovor poprima samo jednu od dvije moguće vrijednosti koje predstavljaju uspjeh i neuspjeh.

Definicija 3. *Pretpostavimo da za svaku pojedinačnu ili eksperimentalnu jedinicu odgovor Y_i može poprimiti samo jednu od dvije moguće vrijednosti 0 i 1. Označimo s*

$$P(Y_i = 0) = 1 - \pi_i, \quad P(Y_i = 1) = \pi_i$$

vjerojatnosti za neuspjeh i uspjeh.

Definicija 4. *Neka je y_i realizacija slučajne varijable Y_i gdje je*

$$y_i = \begin{cases} 1 & , uspjeh \\ 0 & , inače \end{cases}$$

Distribuciju Y_i nazivamo Bernoullijeva distribucija s parametrom π_i i može se zapisati kao

$$P\{Y_i = y_i\} = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

Također, lako se pokaže da su srednja vrijednost i varijanca od Y_i jednaki:

$$E(Y_i) = \mu_i = \pi_i, \qquad Var(Y_i) = \sigma_i^2 = \pi_i(1-\pi_i) \tag{3.9}$$

Imajmo na umu da srednja vrijednost i varijanca ovise o temeljnoj vjerojatnosti π_i. Svaki čimbenik koji utječe na vjerojatnost neće mijenjati samo srednju vrijednost, već i varijancu

opažanja. To sugerira da linearni model koji dopušta prediktorima da utječu na srednju vrijednost, ali pretpostavlja da je varijanca konstantna neće biti prikladan za analizu binarnih podataka.

S praktične točke gledišta važno je napomenuti da ako su prediktori diskretni faktori te rezultati neovisni, tada možemo koristiti Bernoullijevu razdiobu za pojedinačne podatke nula-jedan ili binomnu razdiobu za grupirane podatke koji se sastoje od broja uspjeha u svakoj skupini. Ta dva pristupa su ekvivalentna u smislu da vode do iste funkcije vjerojatnosti i stoga iste procjene parametara i standardne pogreške. Rad s grupiranim podacima kada je to moguće ima dodatnu prednost da je ovisno o veličini grupa moguće testirati ispravnost prilagodbe modela.

Sljedeći korak u definiranju modela za naše podatke odnosi se na strukturu. Željeli bismo da vjerojatnosti π_i ovise o vektoru promatranih kovarijabli x_i. Najjednostavnija ideja bi bila da je

$$\pi_i = \sum_{j=1}^{p} x_{ij}\beta_j \tag{3.10}$$

gdje 3.10 proizlazi iz 3.5 i 3.9. Model 3.10 se naziva *model linearne vjerojatnosti*. Ovaj model se često procjenjuje iz pojedinačnih podataka pomoću metode najmanjih kvadrata.

Jedan od problema s ovim modelom je da vjerojatnost π_i na lijevoj strani mora biti između nule i jedan, ali linearni prediktor $\mathbf{x}\boldsymbol{\beta}$ na desnoj strani može poprimiti bilo koju vrijednost na $\mathbb{R}$ tako da nema jamstava da će predviđene vrijednosti biti u ispravnom rasponu osim ako se ne nametnu neka ograničenja na koeficijente. Jednostavno rješenje ovog problema je transformirati vjerojatnost da bi uklonili ograničenja domene i modelirati transformaciju kao linearnu funkciju kovarijata. To radimo u dva koraka.

Prvo prelazimo s vjerojatnosti π_i na omjer uspjeha i neuspjeha

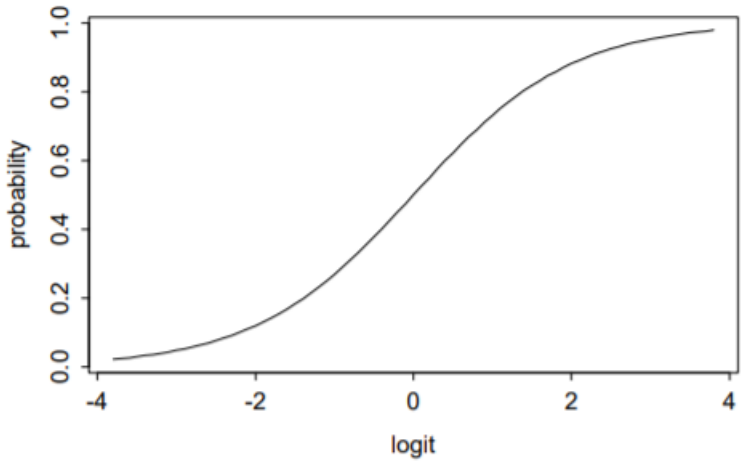
$$omjer_i = \frac{\pi_i}{1-\pi_i} \tag{3.11}$$

Ako je vjerojatnost vrlo mala,kaže se da su izgledi dugački. U nekim kontekstima jezik omjera vjerojatnosti je prirodniji od jezika vjerojatnosti.

Drugo, uzimamo logiratam omjera koji se još naziva i *logit*:

$$\eta_i = \text{logit}(\pi_i) = \log \frac{\pi_i}{1-\pi_i} \tag{3.12}$$

što ima učinak uklanjanja ograničenja, tj. kada vjerojatnost ide prema nuli, omjer vjerojatnosti također ide prema nuli, a logit prema $-\infty$. S druge strane, kada vjerojatnost i omjer idu prema



Slika 3.1: Logit transformacija [5]

1, logit ide prema $+\infty$. Dakle, logit preslikava interval (0,1) na $\mathbb{R}$. Imajte na umu da ako je vjerojatnost $\frac{1}{2}$, omjer je jednak i logit je nula. Kao što možemo vidjeti na Slici 3.1 negativni logovi predstavljaju vjerojatnosti ispod jedne polovine, dok pozitivni logovi odgovaraju vjerojatnostima iznad jedne polovine.
Iz jednadžbe 3.12 lako dobijemo

$$\pi_i = \text{logit}^{-1}(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}} \tag{3.13}$$

Sada smo u mogućnosti definirati model logističke regresije uz pretpostavku da logit vjerojatnosti π_i slijedi linearni model, a ne sama vjerojatnost.

Pretpostavimo da imamo k neovisnih opažanja $y_1, \ldots, y_k$ i i-to opažanje se može tretirati kao realizacija slučajne varijable Y_i. Pretpostavimo da Y_i ima Bernoullijevu distribuciju, tj. da se radi o modelu dihotomne logističke regresije. To znači da sistemska struktura modela izgleda

$$\text{logit}(\pi_i) = \mathbf{x}_i^T \boldsymbol{\beta} \tag{3.14}$$

gdje je $\mathbf{x}_i$ vektor kovarijata, tj. neovisnih varijabli, a $\boldsymbol{\beta}$ vektor regresijskih koeficijenata. Regresijski koeficijenti $\boldsymbol{\beta}$ mogu se tumačiti duž iste linije imajući na umu da je lijeva strana *logit*, a ne prosjek. Dakle, koeficijenti β_j predstavljaju promjenu u logit vjerojatnosti povezanu s promjenom jedinice u j-tom prediktoru držeći sve ostale prediktore konstantnim.

Eksponenciranjem jednadžbe 3.14 dobijemo da je omjer za i-tu jedinicu dan s

$$\frac{\pi_i}{1-\pi_i} = \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \tag{3.15}$$

Ovaj izraz definira multiplikativni model za omjere. Na primjer, ako bismo promijenili j-ti prediktor za jednu jedinicu, dok bi sve ostale varijable bile konstantne, pomnožili bismo omjer vjerojatnosti s $\exp\beta_j$. Dakle, eksponirani koeficijent $\exp\beta_j$ predstavlja omjer vjerojatnosti. Rješavanje vjerojatnosti π_i u logit modelu jednadžbe 3.14 daje složeniji model

$$\pi_i = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^T \boldsymbol{\beta})} \tag{3.16}$$

Na desnoj strani jednakosti je nelinearna funkcija prediktora i ne postoji jednostavan način da se izrazi učinak povećanja prediktora za jednu jedinicu na vjerojatnost dok su ostale varijable konstantne. Približan odgovor možemo dobiti uzimajući derivaciju s obzirom na x_j što naravno ima smisla samo za neprekidne prediktore. Koristeći kvocijentno pravilo dobivamo

$$\frac{d\pi_i}{dx_{ij}} = \beta_j \pi_i (1-\pi_i)$$

Dakle, učinak j-tog prediktora na vjerojatnost π_i ovisi o koeficijentu β_j i vrijednosti vjerojatnosti. Analitičari ponekad procjenjuju π_i na srednju vrijednost uzorka te tada rezultat aproksimira učinak kovarijate u blizini srednje vrijednosti ovisne varijable.

Pretpostavimo sada da se ispitivane jedinice mogu klasificirati prema faktorima interesa u k skupina na takav način da svi pojedinci u istoj skupini imaju identične vrijednosti svim prediktorskim varijablama. Neka n_i označava broj opažanja u skupini i, a y_i označava broj jedinica koje imaju atribut interesa u skupini i. y_i je realizacija slučajne varijable Y_i koja poprima vrijednosti $0, 1, \ldots, n_i$. Ako su promatranja n_i u svakoj skupini neovisna, a svi imaju istu vjerojatnost π_i, tada Y_i ima binomnu distribuciju $Y_i \sim B(n_i, \pi_i)$. Tada je

$$P(Y_i = y_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i} \tag{3.17}$$

Funkcija vjerojatnosti za n neovisnih binomnih opažanja je umnožak funkcija gustoće danih jednadžbom 3.17. Logaritmirajući dobivamo da *log* vjerojatnost je

$$\log L(\beta) = \sum \{y_i \log(\pi_i) + (n_i - y_i)\log(1-\pi_i)\} \tag{3.18}$$

gdje π_i ovisi o kovarijatama x_i i vektoru $\boldsymbol{\beta}$ kroz logit transformaciju jednadžbe 3.14.
S obzriom na trenutnu procjenu parametra $\hat{\boldsymbol{\beta}}$, računamo linearni prediktor $\hat{\boldsymbol{\eta}} = \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$ i prilagođene vrijednosti $\hat{\boldsymbol{\mu}} = \text{logit}^{-1}(\hat{\boldsymbol{\eta}})$. S ovim vrijednostima izračunavamo radno ovisnu varijablu $\mathbf{z}$, koja ima elemente

$$z_i = \hat{\eta}_i + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i(n_i - \hat{\mu}_i)} \tag{3.19}$$

Zatim regresiramo $\mathbf{z}$ na kovarijatama koje izračunavaju novu težinsku procjenu najmanjih kvadrata

$$\hat{\boldsymbol{\beta}} = (X^T W X)^{-1} X^T W \mathbf{z} \tag{3.20}$$

gdje je W dijagonalna matrica težina s vrijednostima, tj. težinama

$$w_{ii} = \frac{\hat{\mu}_i(n_i - \hat{\mu}_i)}{n_i}$$

Procjena parametra β koristi se za dobivanje poboljšanih prilagođenih vrijednosti i postupak se ponavlja do konvergencije. Prikladne početne vrijednosti mogu se dobiti primjenom veze na podatke. Da bismo izbjegli probleme s brojevima 0 ili n_i, empirijske logove računamo dodajući $\frac{1}{2}$ i brojniku i nazivniku.

$$z_i = \log \frac{y_i + 1/2}{n_i - y_i + 1/2}$$

i zatim regresirati ovu količinu na x_i da bi se dobila početna procjena β. Varijanca krajnjeg β parametra na velikom uzorku je

$$var(\hat{\beta}) = (X^T W X)^{-1}$$

gdje je W matrica težina u zadnjoj iteraciji.

Poglavlje 4

Monte-Carlo simulacija

Monte-Carlo metode su stohastičke (determinističke) simulacijske metode i algoritmi koji pomoću slučajnih ili kvazislučajnih brojeva i velikog broja izračuna i ponavljanja predviđaju ponašanje složenih matematičkih sistema koji su previše komplicirani za analitičko rješavanje. Monte Carlo metode uglavnom se koriste u tri klase problema: optimizacija, numerička integracija i generiranje izvoda iz distribucije vjerojatnosti.
Monte Carlo metode variraju, ali imaju tendenciju slijediti određene obrasce:

1. Definiranje domene mogućih ulaza
2. Slučajno generiranje ulaza iz vjerojatnosne distribucije po domeni
3. Izvedba deterministčkog izračuna na ulazima
4. Dobivanje rezultata

Upotreba Monte Carlo metoda zahtijeva velike količine slučajnih brojeva i upravo je njihova uporaba potaknula razvoj generatora pseudoslučajnih brojeva, koji su bili daleko brži od tablica slučajnih brojeva koji su se ranije koristili za statističko uzorkovanje. [8]

4.1 Opis problema

Cilj simulacije je analiza raznih statističkih mjera koji predstavljaju pogreške predikcije rizika s obzirom na razne ulazne parametre.
U ovom odjeljku podrobnije ćemo opisati statističke mjere pogreške koje ćemo kasnije koristiti za izvesti zaključak iz njih. Uspoređivati ćemo pet mjera pogreške predviđanja s obzirom

na ulazne podatke, a to su Somersov D, c statistika, osjetljivost (*'Sensitivity'*), specifičnost (*'Specificity'*) te stopa pogrešne klasifikacije (*'Misclassification rate'*).
Da bi jasnije mogli obrazložiti ove statistike prvo moramo uvesti podjelu na parove.
S obzirom da koristimo binomnu logističku regresiju svaki uzorak ćemo podijeliti na podskupove gdje je $y = 0$ i $y = 1$. Nakon toga ćemo uzimati sve moguće kombinacije parova iz ta dva različita podskupa i uspoređivati prediktorske varijable.

Definicija 5. *Kažemo da je par saglasan ('concordant') ako vrijedi $x_i < x_j$ i $y_i < y_j$, tj. $x_i > x_j$ i $y_i > y_j$*
Kažemo da je par proturječan ('discordant') ako vrijedi $x_i < x_j$ i $y_i > y_j$, tj. $x_i > x_j$ i $y_i < y_j$

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	89.5	Somers' D	0.791
Percent Discordant	10.4	Gamma	0.792
Percent Tied	0.1	Tau-a	0.202
Pairs	11475	c	0.896

Response Profile		
Ordered Value	Bonus	Total Frequency
1	0	255
2	1	45

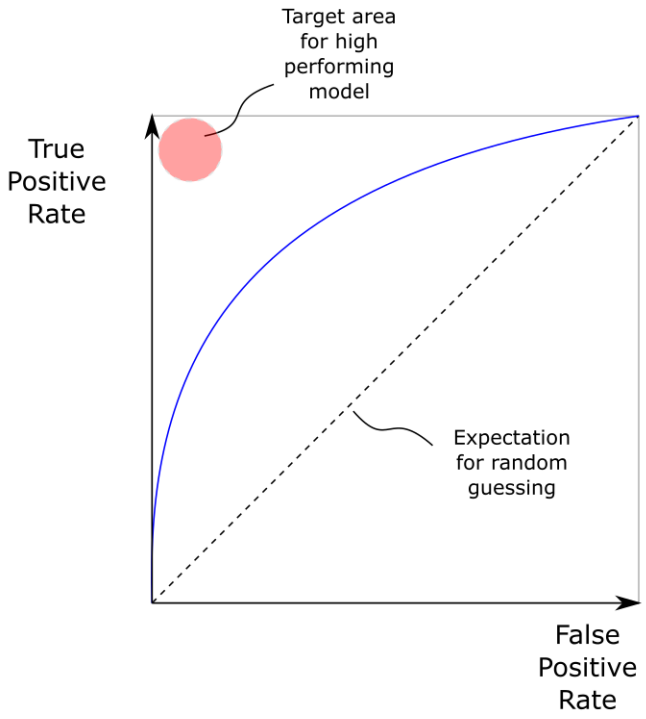
Slika 4.1: Primjer dobivanja saglasnih i proturječnih parova

Definicija 6. *ROC krivulja ('Receiver operating characteristic') je jedan od načina usporedbe dijagnostičkih testova. Na grafu se prikazuje stvarna pozitivna stopa (TPR) naspram lažnoj pozitivnoj stopi (FPR).*
Graf s ROC krivuljom prikazuje:

- *odnos između osjetljivosti (TPR) i specifičnosti (1-FPR) - obrnuto proporcionalni*
- *točnost ispitivanja - što je krivulja bliža gornjem lijevom kutu, test je točniji. Isto tako, što je krivulja bliža dijagonali, test je manje točan*

- *omjer vjerojatnosti - daje derivat na bilo kojoj određenoj točki*

Točnost ispitivanja također je prikazana kao površina ispod krivulje (AUC - 'area under the curve'). Što je veće područje ispod krivulje, točniji je test.



Slika 4.2: ROC krivulja

Na ovih 5 mjera pogreške predviđanja ćemo bazirati naše rezultate i zaključke:

Somersov D - skraćeno od Somersov Delta, prima vrijednosti od -1 do 1, a dobiven je oduzimanjem broja saglasnih s brojem proturječnih parova podijeljen s ukupnim projem parova. Na slici 4.1 to je broj dobiven oduzimanjem postotka saglasnih s postotkom proturječnih parova.

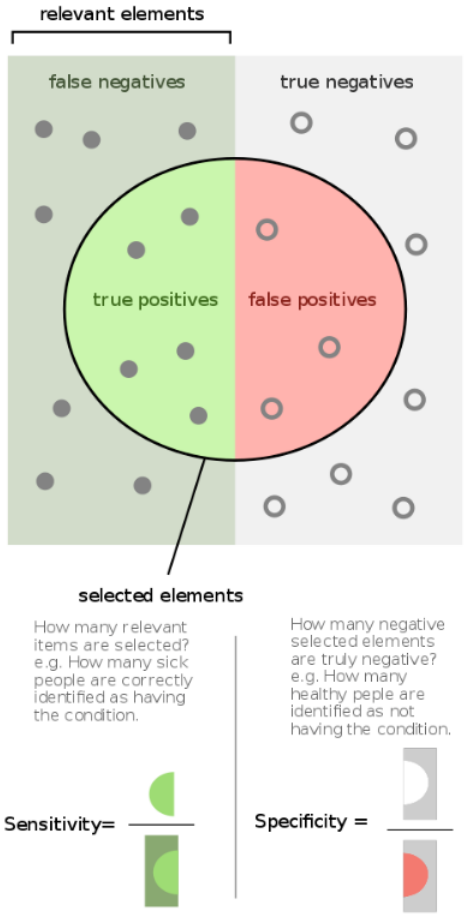
c statistika - slično kao i Somersov D, a na slici 4.1 možemo vidjeti da se računa zbrajanjem postotka saglasnih sa jednom polovinom postotka jednakih od ukupnog broja parova.

C statistika poprima vrijednosti od 0.5 (najlošiji model) do 1 (najbolji model), a može se procijeniti i kao površina ispod ROC krivulje, tj. AUC vrijednost (*'area under the curve*)

Osjetljivost - naziva se i stvarna pozitivna stopa (*'true positive rate'*), mjeri udio stvarnih pozitivnih koji su ispravno identificirani kao takvi.

Specifičnost - naziva se i stvarna negativna stopa (*'true negative rate'*), mjeri udio stvarnih negativnih koji su ispravno identificirani kao takvi.

Stopa pogrešne klasifikacije - kao što sam naziv govori mjeri udio pogrešno klasificiranih pozitivnih i negativnih u cijelom uzorku.



Slika 4.3: Računanje osjetljivosti i specifičnosti

Za stvarnu negativnu i pozitivnu stopu poprimaju se vrijednosti od 0 do 100 te što su vrijednosti bliže 100 tada model ima veću prediktivnu sposobnost. Za svaki test obično postoji kompromis između mjera, npr. u sigurnosti zračne luke skeneri mogu biti postavljeni tako da pokreću alarme na niskorizičnim predmetima kao što su kopče pojasa i ključevi (niska specifičnost) kako bi se povećala vjerojatnost identificiranja opasnih objekata i smanjila opasnost od nestalih objekata koji predstavljaju prijetnju (visoka osjetljivost). Za stopu pogrešne klasifikacije se također postižu vrijednosti od 0 do 100, a što su niže vrijednosti veća je prediktivna sposobnost.

4.2 Ulazni podaci

Rezultate ćemo uspoređivati s obzirom na varijacije u broju prediktorskih varijabli, distribuciji prediktorskih varijabli, korelaciji među njima i veličini uzorka. Svaku simulaciju ulaznih podataka ćemo ponavljati 500 puta kako bi mogli donijeti valjane zaključke. Ulazni podaci koje smo koristili su:

1. VELIČINA UZORKA - 50, 100, 500
2. BROJ PREDIKTORSKIH VARIJABLI - jedna, dvije
3. DISTRIBUCIJA - Normalna (0,1), Gamma (1,1) i Gamma (3,1)
4. KORELACIJA IZMEĐU PREDIKTORSKIH VARIJABLI - 0, 0.25, 0.50, 0.75, 0.95

Kod modela s dvije prediktorske varijable, obje varijable imaju istu distribuciju, dok smo korelaciju naravno koristili samo za modele s dvije prediktorske varijable.
Označimo sa ρ koeficijent korelacije, tada x_1 i x_2 dobivamo ovako:

```
x1=rand("Normal",0,1);
x2=rand("Normal",0,1);
x1n=x1;
x2n=&rho*x1+ sqrt(1-&rho**2)*x2;
```

Slika 4.4: Korištenje koeficijenta korelacije

tj.

$$x_{1n} = x_1 \qquad x_{2n} = \rho x_1 + \sqrt{1-\rho^2}x_2$$

Definicija 7. *Slučajna varijabla X ima normalnu distribuciju ako je njena funkcija gustoće vjerojatnosti:* $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ *i označavamo s* $X \sim N(\mu, \sigma^2)$

μ predstavlja srednju vrijednost, tj. očekivanu vrijednost E(X), a σ standardnu devijaciju. Mi ćemo u ovom MC eksperimentu koristiti standardnu normalnu raspodjelu N(0,1) sa očekivanjem $\mu = 0$. [3]

Definicija 8. *Slučajna varijabla X ima Gamma distribuciju ako je njena funkcija gustoće vjerojatnosti:* $f(x) = \frac{\beta^k x^{k-1} e^{-\beta x}}{\Gamma(k)}$ *i označavamo s* $X \sim \Gamma(k, \beta)$ *gdje k predstavlja parametar oblika (**'shape parameter'**), a* β ***'rate parameter'*** *ili* ***inverzni 'scale parameter'*** *jer vrijedi* $\beta = \frac{1}{\theta}$. *[7]*

Definirajmo i $\Gamma(k)$ funkciju s

$$\Gamma(k) = \int_0^\infty x^{k-1} e^{-x} dx$$

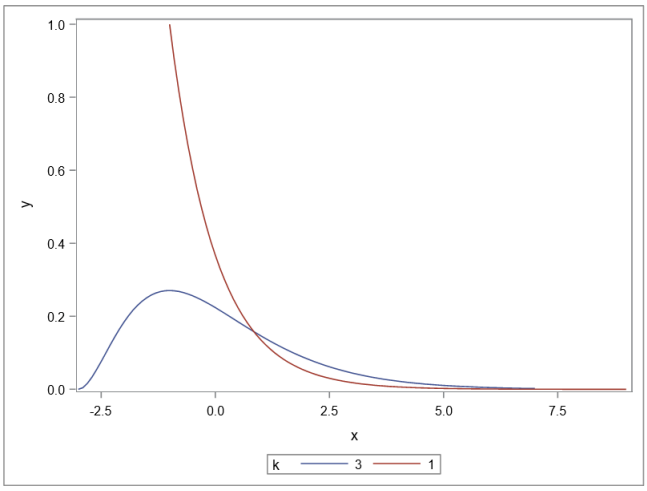
koji ima svojstva:

1. $\forall \alpha > 1 \qquad \Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1) \qquad$ (rekurzija)
2. $\forall n \in \mathbb{N} \qquad \Gamma(n) = (n - 1)! \qquad$ (poopćenje faktorijela)
3. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

U ovom radu za ulazne podatke smo koristili troparametarsku Gamma distribuciju s *rate parametrom* $\beta = 1$, a za parametar oblika k smo uzeli $k = 1$ i $k = 3$. Treći parametar se naziva parametar pomaka (*'threshold' ili 'shift' parameter*), a za njega smo uzeli vrijednost *-k* jer je za Gamma distribuciju očekivanje $E(X) = \frac{k}{\beta} = k$ za $\beta = 1$. To smo uradili kako bi nam očekivanja bila ista za sve korištene distribucije pošto je za standardnu normalnu distribuciju $E(X) = 0$.

```
x1=rand("gamma",&k);
x1t=x1-&k;
x2=rand("gamma",&k);
x2t=x2-&k;
```

Slika 4.5: Simulacija slučajnih varijabli u slučaju gamma distribucije



Slika 4.6: Graf troparametarske gamma distribucije s parametrom pomaka -k

4.3 Rezultati

U ovom odjeljku prikazati ćemo rezultate i interpretirati ih s grafovima. Podijeliti ćemo rezultate na jednoparametarske i dvoparametarske odnosno na modele s jednom prediktorskom varijablom i dvije prediktorske varijable.

Jednoparametarski modeli

Za jednoparametarske modele izabrali smo proizvoljnu formulu

$$\eta = -1 + 2x$$

pa iz 3.13 dobivamo μ kojeg koristimo za simulaciju y-a preko Bernoullijeve distribucije.

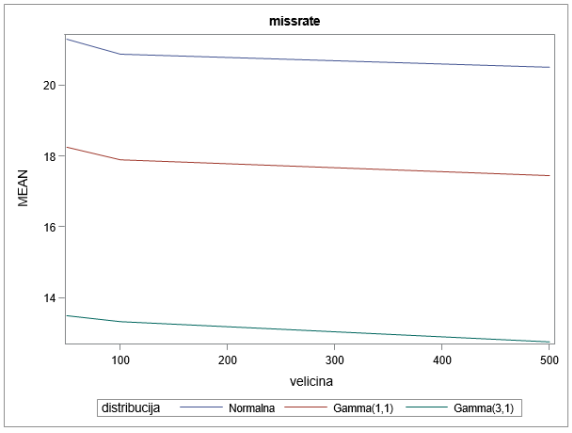
```
eta= -1+2*x1;
mu=exp(eta)/(1+exp(eta));
y=rand("Bernoulli",mu);
```

Slika 4.7: Simuliranje y-a [6]

Na grafovima sa slika 4.10 i 4.11 vidimo da najbolju prediktivnost ima model koji koristi gamma distribuciju sa parametrom oblika *k=3*. Vidimo također da prediktivnost ne ovisi previše o veličini uzorka iako su vrijednosti neznatno veće kako se povećava uzorak.

		distribucija		
		Gamma(1,1)	Gamma(3,1)	Normalna
statistika	veličina			
Sensitivity	50	56.6787	76.0067	61.8989
	100	57.2148	76.8046	63.2098
	500	57.8924	77.8275	64.4709
Somers' D	50	0.7081	0.8743	0.7258
	100	0.7073	0.8735	0.7240
	500	0.7106	0.8770	0.7262
Specificity	50	92.0795	91.3787	86.5922
	100	92.6015	91.6298	87.2329
	500	93.1039	92.2451	87.5620
c	50	0.8541	0.9372	0.8629
	100	0.8536	0.9368	0.8620
	500	0.8553	0.9385	0.8631
missrate	50	18.2440	13.4960	21.2880
	100	17.8900	13.3280	20.8660
	500	17.4440	12.7564	20.4988

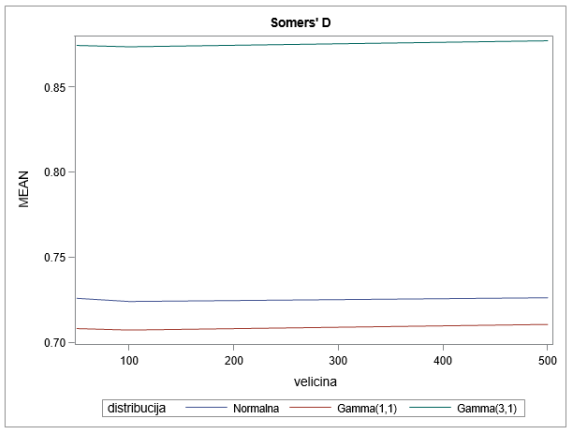
Slika 4.8: Aritmetičke sredine za mjere pogreške jednoparametarskih modela



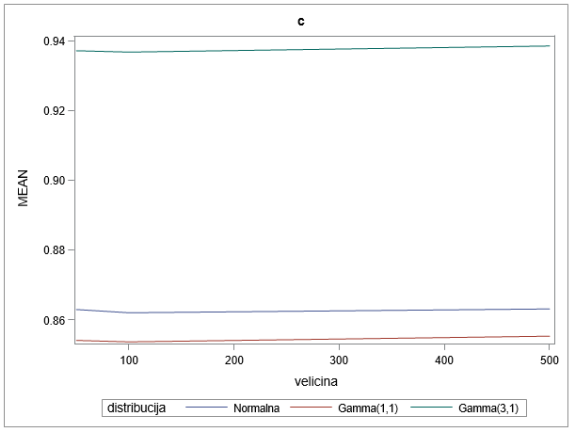
Slika 4.9: Stopa pogrešne klasifikacije

Najveći postotak pogrešne klasifikacije rizika u jednoparametarskom modelu imamo s normalnom distribucijom i to oko 20 posto, dok gamma distribucija s *k=3* ima najmanje i to oko 14 posto. Također povećavanjem veličine uzorka se smanjuje pogrešna klasifikacija.

Na grafovima sa slika 4.12 i 4.13 vidimo da je za jednoparametarski model osjetljivost puno niža od specifičnosti. Osjetljivost predstavlja sposobnost testa da ispravno klasificira pojedinca kao rizičnog. Također tu model s distribucijom *Gamma (3,1)* ima najbolju pre-

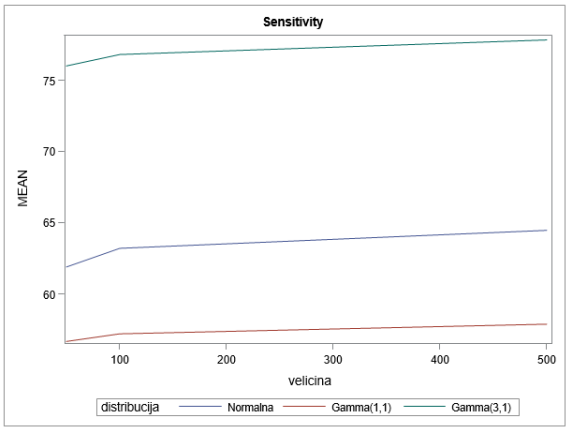


Slika 4.10: Somersov D

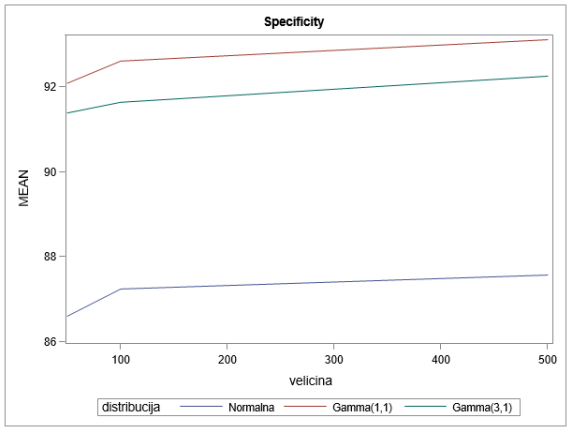


Slika 4.11: c statistika

diktivnu sposobnost. S druge strane specifičnost je visoka za obje gamma distribucije, a nešto malo veće za *Gamma (1,1)* i ona predstavlja sposobnost testa da ispravno klasificira pojedinca kao nerizičnog.



Slika 4.12: Osjetljivost



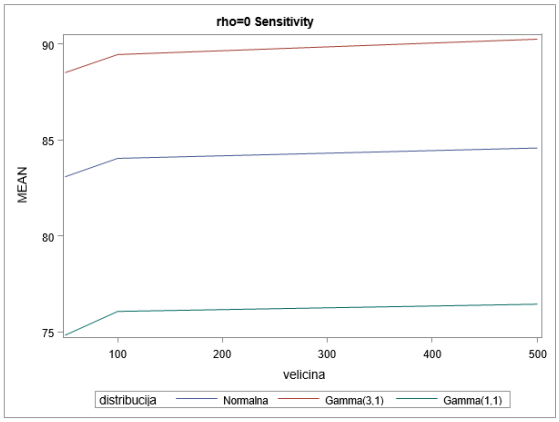
Slika 4.13: Specifičnost

Dvoparametarski modeli

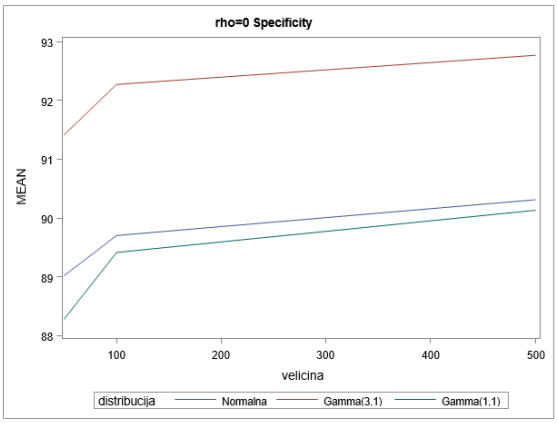
Za dvoparametarske modele osim što imamo dvije prediktorske varijable imamo i dodatan parametar u korelaciji između prediktorskih varijabli ρ. Ovdje smo izabrali proizvoljnu formulu

$$\eta = -1 + 3x_1 - 3x_2$$

pa je očekivana vrijednost $\eta = -1$ jednaka i za jednoparametarske i dvoparametarske modele. Podatke ćemo prikazivati posebno za svaku distribuciju, ali prije toga ćemo usporediti rezultate iz dvoparametarskih modela u kojem je $\rho = 0$ s rezultatima iz jednoparametarskih modela.



Slika 4.14: Osjetljivost za $\rho = 0$



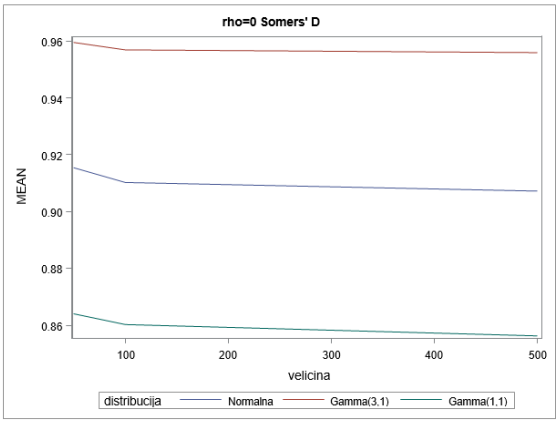
Slika 4.15: Specifičnost za $\rho = 0$

Usporedimo li sliku 4.14 sa slikom 4.12 možemo vidjeti da su distribucije slično rangirane kao i za jednoparametarske modele, ali su mnogo veće vrijednosti za dvoparametarske modele pa nam se samim time nameće zaključak da je prediktivna sposobnost bolja za modele s više prediktorskih varijabli.

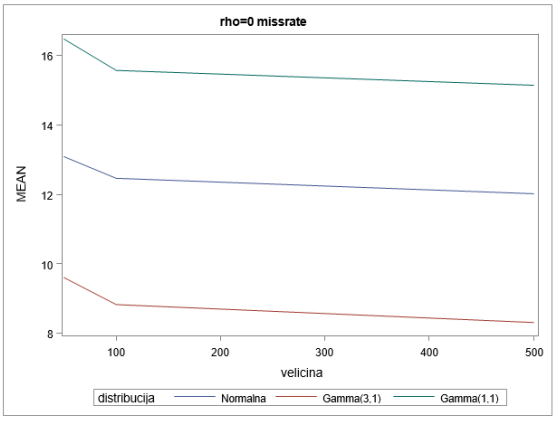
Slično kao i za osjetljivost, Somersov D i c statistika također imaju znatno veće vrijednosti za dvoparametarski model sa $\rho = 0$ nego jednoparametarski model, a također gamma distribucija s parametrom oblika *k=3* ima najbolju prediktivnu sposobnost, dok gamma s *k=1* najlošiju. Usporedbu rezultata se može napraviti i iz tablica na slikama 4.8 i 4.16. Za stopu pogrešne klasifikacije rizika su se također poboljšale vrijednosti za dvoparametarski model, ali usporedbom slika 4.18 i 4.9 možemo uočiti da varijable s normalnom distribucijom imaju bolju prediktivnu sposobnost od *Gamma(1,1)* u dvoparametarskom modelu, dok je u jednoparametarskom modelu bilo obrnuto.

rho=0		distribucija		
		Gamma(1,1)	Gamma(3,1)	Normalna
statistika	veličina			
Sensitivity	50	74.8557	88.5203	83.0984
	100	76.0911	89.4577	84.0519
	500	76.4681	90.2539	84.5908
Somers' D	50	0.8641	0.9595	0.9155
	100	0.8603	0.9569	0.9103
	500	0.8563	0.9560	0.9073
Specificity	50	88.2899	91.4221	89.0288
	100	89.4199	92.2755	89.7087
	500	90.1368	92.7707	90.3162
c	50	0.9320	0.9798	0.9577
	100	0.9301	0.9785	0.9551
	500	0.9281	0.9780	0.9536
missrate	50	16.4680	9.6080	13.0840
	100	15.5680	8.8320	12.4620
	500	15.1392	8.3144	12.0180

Slika 4.16: Rezultati dvoparametarskih modela za $\rho = 0$



Slika 4.17: Somersov D za $\rho = 0$



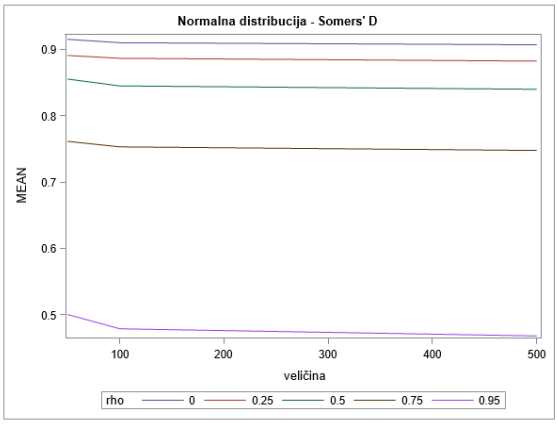
Slika 4.18: pogrešne klasifikacije za $\rho = 0$

Dvoparametarski modeli s normalnom distribucijom

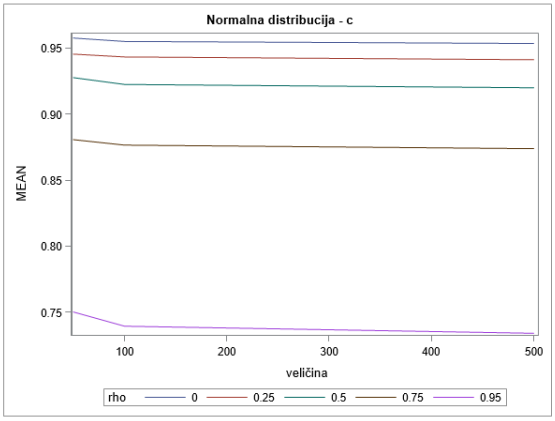
U ovom pododjeljku prikazali smo statističke rezultate za normalno distribuirane varijable ovisno o veličini uzorka na X osi i korelacije između varijabli na Y osi.

Normalna distribucija s dvije varijable		statistika				
		Sensitivity	Somers' D	Specificity	c	missrate
veličina	korelacija					
50	0	83.0984	0.9155	89.0288	0.9577	13.0840
	0.25	80.3540	0.8911	88.2620	0.9456	14.5160
	0.5	75.4849	0.8554	87.4949	0.9277	16.6720
	0.75	64.4355	0.7616	85.7016	0.8808	21.1960
	0.95	29.4008	0.5006	88.2549	0.7503	28.7400
100	0	84.0519	0.9103	89.7087	0.9551	12.4620
	0.25	81.2016	0.8868	89.1899	0.9434	13.8140
	0.5	76.5618	0.8453	87.9845	0.9226	16.1840
	0.75	65.9157	0.7533	86.8623	0.8767	20.2960
	0.95	29.0119	0.4790	90.1298	0.7395	27.8180
500	0	84.5908	0.9073	90.3162	0.9536	12.0180
	0.25	81.9250	0.8825	89.4273	0.9413	13.5552
	0.5	77.2805	0.8402	88.3939	0.9201	15.8544
	0.75	66.8397	0.7479	87.4687	0.8739	19.8484
	0.95	29.7719	0.4682	91.5564	0.7341	26.9104

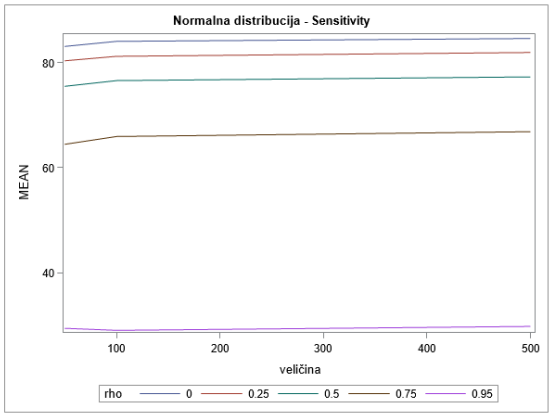
Slika 4.19: Statistički rezultati za normalno distribuirane varijable



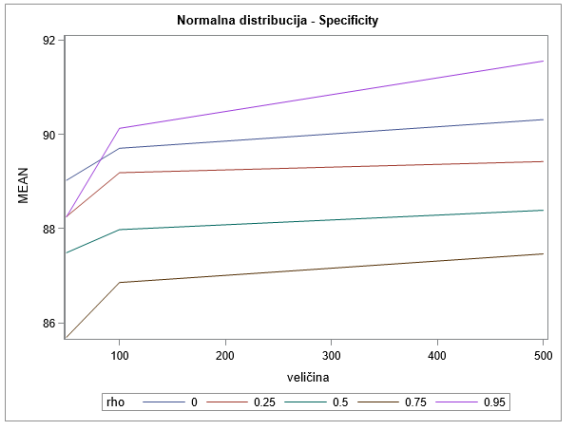
Slika 4.20: Somersov D



Slika 4.21: c statistika



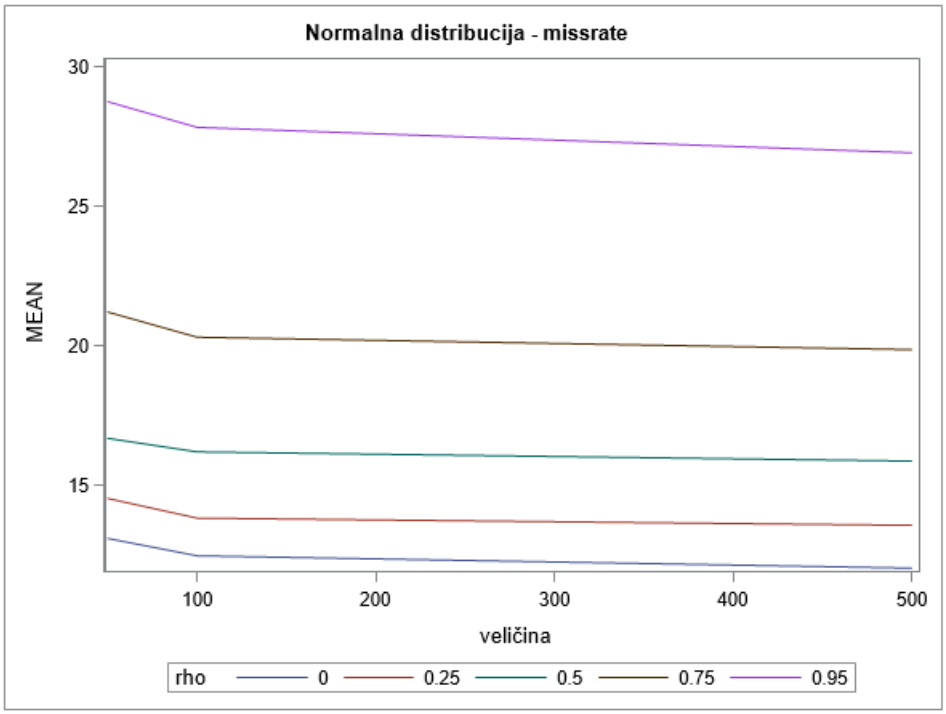
Slika 4.22: Osjetljivost



Slika 4.23: Specifičnost

Iz grafova na slikama 4.20, 4.21 i 4.22 jasno vidimo da se prediktivna sposobnost modela smanjuje kako se povećava korelacija između prediktorskih varijabli, a pogotovo jako niske vrijednosti su za $\rho = 0.95$. Kod specifičnosti se događa mala iznimka za $\rho = 0.95$ te zbog jako velike korelacije dolazi do rasta *'true negative ratea'*.

Stopa pogrešne klasifikacije (slika 4.24) je zadovoljavajuća za $\rho = 0$ i $\rho = 0.25$, dok su za



Slika 4.24: Stopa pogrešne klasifikacije

ostale korelacijske vrijednosti, a posebno za $\rho = 0.75$ i $\rho = 0.95$ prilično velike, lošije čak i od jednoparametarskih modela. Gledajući ove grafove nameće se zaključak da korelacija između prediktorskih varijabli uvelike utječe na rezultate i prediktivnu sposobnost modela.

Dvoparametarski model s gamma distribucijom (k=1)

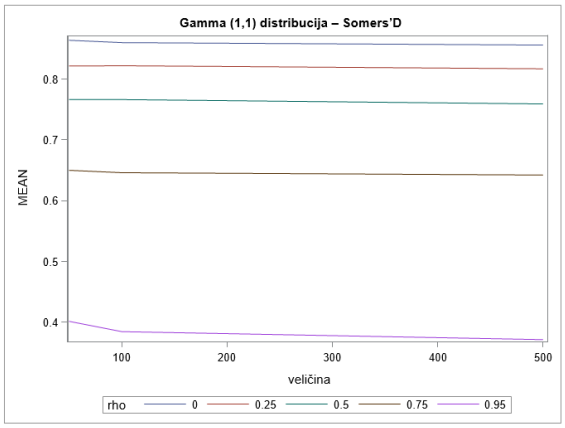
U ovom pododjeljku prikazali smo statističke rezultate za gamma distribuirane varijable s parametrom oblika *k=1* u ovisnosti o veličini uzorka na X osi i korelaciji između varijabli na Y osi. Slično kao i za normalno distribuirane varijable, što je veća koreliranost između

Gamma(1,1) distribucija s dvije varijable		statistika				
		Sensitivity	Somers' D	Specificity	c	missrate
veličina	korelacija					
50	0	74.8557	0.8641	88.2899	0.9320	16.4680
	0.25	72.3943	0.8217	85.1451	0.9109	19.3440
	0.5	68.6491	0.7667	81.4561	0.8834	22.9120
	0.75	57.0106	0.6500	79.1959	0.8250	27.9480
	0.95	17.9339	0.4012	87.4957	0.7006	32.7040
100	0	76.0911	0.8603	89.4199	0.9301	15.5680
	0.25	73.9360	0.8222	86.3320	0.9111	18.3480
	0.5	71.2626	0.7665	82.5449	0.8832	21.6100
	0.75	59.9899	0.6461	80.1900	0.8231	26.8020
	0.95	14.6538	0.3843	91.3973	0.6922	31.5700
500	0	76.4681	0.8563	90.1368	0.9281	15.1392
	0.25	74.4415	0.8172	87.1239	0.9086	17.8956
	0.5	71.9954	0.7592	83.4461	0.8796	21.0484
	0.75	62.1853	0.6423	80.9217	0.8212	25.9560
	0.95	8.3728	0.3712	96.2006	0.6856	30.2412

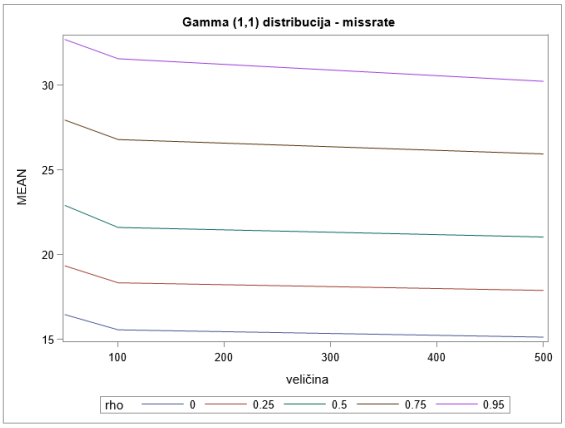
Slika 4.25: Statistički rezultati za gamma distribuirane varijable (k=1)

prediktorskih varijabli rezultati su lošiji. Povećavanjem uzorka vrijednosti Somersov D i c su boljom aproksimacijom u neznatnom padu, dok se stopa pogrešne klasifikacije poboljšava.

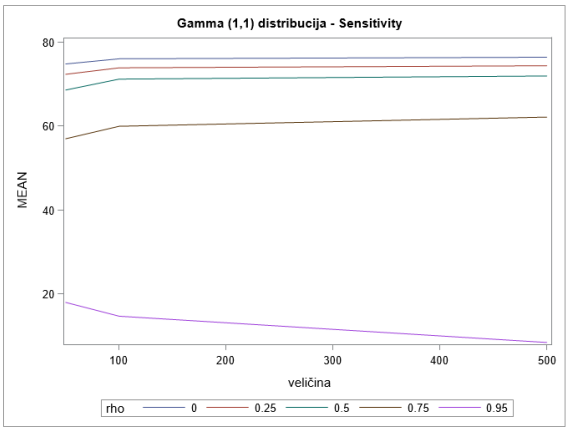
Same vrijednosti su lošije nego kod normalno distribuiranih varijabli te su modeli s ovom distribucijom najlošiji.



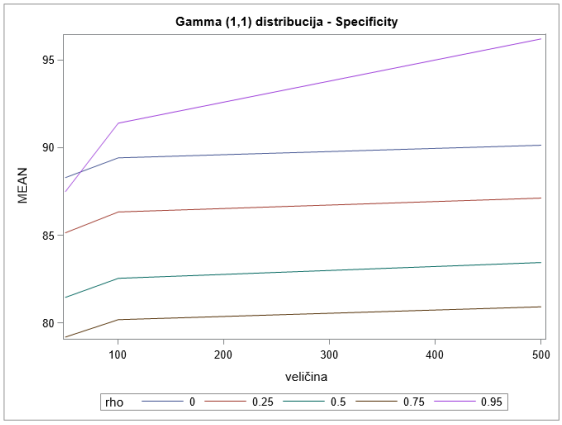
Slika 4.26: Somersov D



Slika 4.27: Stopa pogrešne klasifikacije



Slika 4.28: Osjetljivost



Slika 4.29: Specifičnost

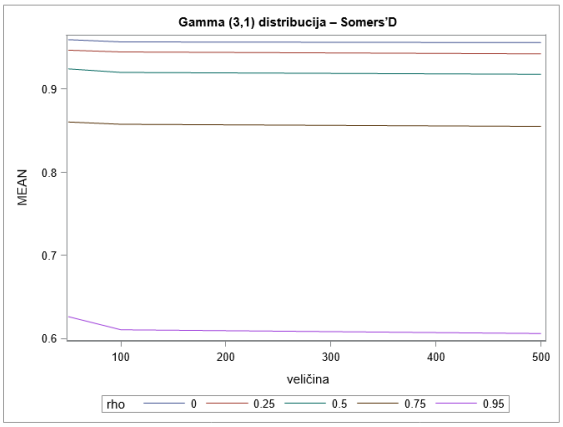
Dvoparametarski model s gamma distribucijom (k=3)

U ovom pododjeljku prikazali smo statističke rezultate za gamma distribuirane varijable s parametrom oblika *k=3* u ovisnosti o veličini uzorka na X osi i korelaciji između varijabli na Y osi. Kao i za prethodne distribucije vidimo da su rezultati najbolji za neovisne varijable, a

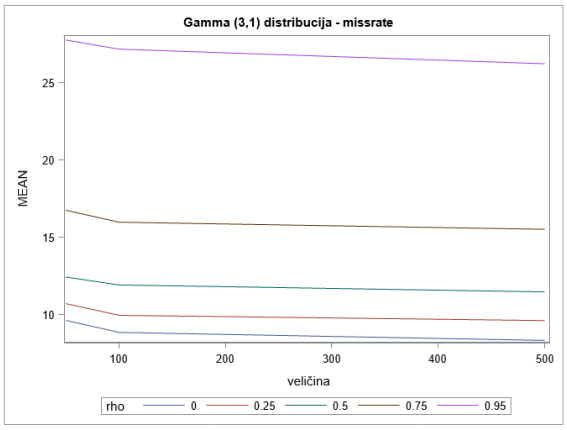
Gamma(3,1) distribucija s dvije varijable		statistika				
		Sensitivity	Somers' D	Specificity	c	missrate
veličina	korelacija					
50	0	88.5203	0.9595	91.4221	0.9798	9.6080
	0.25	87.6123	0.9469	90.1612	0.9734	10.6920
	0.5	85.9462	0.9244	88.2917	0.9622	12.4120
	0.75	80.5884	0.8605	84.4210	0.9303	16.7360
	0.95	51.3890	0.6264	81.7475	0.8132	27.7600
100	0	89.4577	0.9569	92.2755	0.9785	8.8320
	0.25	88.7381	0.9447	90.8548	0.9724	9.9420
	0.5	86.9000	0.9201	88.7660	0.9601	11.9000
	0.75	81.9967	0.8578	85.1419	0.9289	15.9680
	0.95	53.4860	0.6106	82.4580	0.8053	27.1700
500	0	90.2539	0.9560	92.7707	0.9780	8.3144
	0.25	89.2708	0.9425	91.2802	0.9712	9.6024
	0.5	87.6528	0.9179	89.2341	0.9589	11.4516
	0.75	82.9091	0.8553	85.6313	0.9277	15.5092
	0.95	55.8558	0.6061	83.2042	0.8031	26.2300

Slika 4.30: Statistički rezultati za gamma distribuirane varijable (k=3)

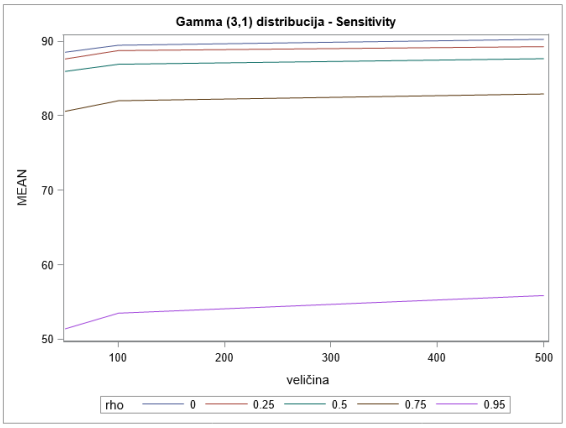
pritom vidimo da varijable s ovom distribucijom imaju najbolju prediktivnu sposobnost. Na slikama 4.31 - 4.34 možemo vidjeti grafove za statističke rezultate.



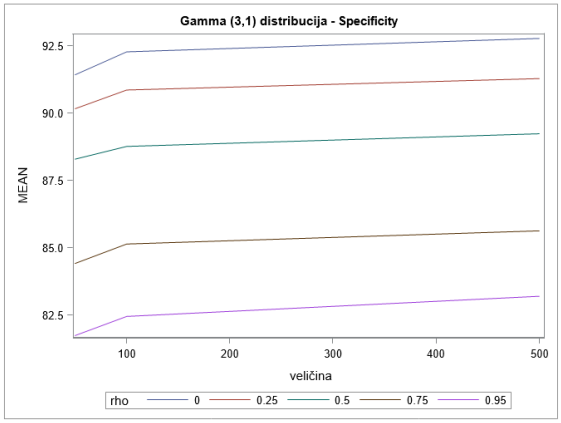
Slika 4.31: Somersov D



Slika 4.32: Stopa pogrešne klasifikacije



Slika 4.33: Osjetljivost



Slika 4.34: Specifičnost

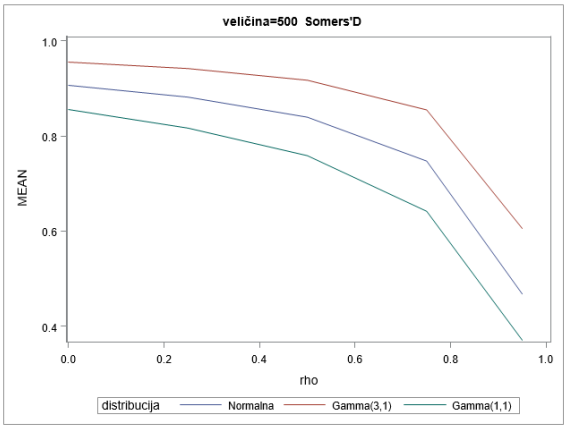
Analiza rezultata za n=500

Primjetili smo da se većina rezultata znatno ne mijenja povećavanjem veličine uzorka. Zato smo u ovom pododjeljku fiksirali veličinu uzorka na $n = 500$ te na taj način grafički prikazali kretanje statističkih rezultata promjenom korelacija i distribucija prediktorskih varijabli. Na idućim grafovima se jasno vidi regres mjera pogreške predviđanja povećavanjem korelacije između varijabli.

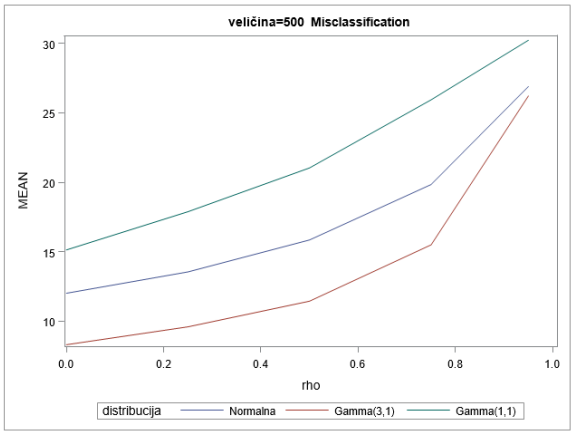
veličina=500		korelacija				
		0	0.25	0.5	0.75	0.95
statistika	distribucija					
Sensitivity	Gamma(1,1)	76.4681	74.4415	71.9954	62.1853	8.3728
	Gamma(3,1)	90.2539	89.2708	87.6528	82.9091	55.8558
	Normalna	84.5908	81.9250	77.2805	66.8397	29.7719
Somers' D	Gamma(1,1)	0.8563	0.8172	0.7592	0.6423	0.3712
	Gamma(3,1)	0.9560	0.9425	0.9179	0.8553	0.6061
	Normalna	0.9073	0.8825	0.8402	0.7479	0.4682
Specificity	Gamma(1,1)	90.1368	87.1239	83.4461	80.9217	96.2006
	Gamma(3,1)	92.7707	91.2802	89.2341	85.6313	83.2042
	Normalna	90.3162	89.4273	88.3939	87.4687	91.5564
c	Gamma(1,1)	0.9281	0.9086	0.8796	0.8212	0.6856
	Gamma(3,1)	0.9780	0.9712	0.9589	0.9277	0.8031
	Normalna	0.9536	0.9413	0.9201	0.8739	0.7341
missrate	Gamma(1,1)	15.1392	17.8956	21.0484	25.9560	30.2412
	Gamma(3,1)	8.3144	9.6024	11.4516	15.5092	26.2300
	Normalna	12.0180	13.5552	15.8544	19.8484	26.9104

Slika 4.35: Statistički rezultati za n=500

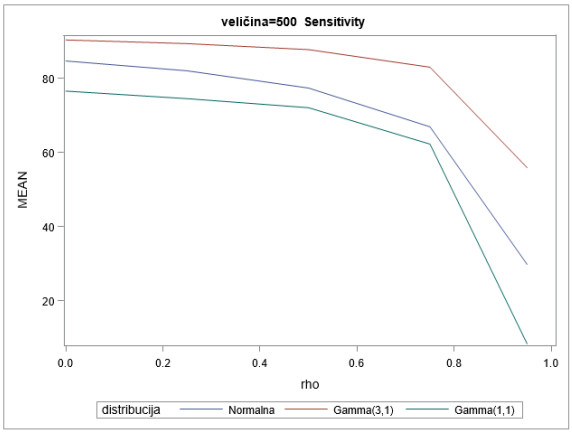
Iz tablice na slici 4.35 možemo vidjeti da je najbolji prediktivni model gdje su varijable neovisno distribuirane *Gamma (3,1)* distribucijom.



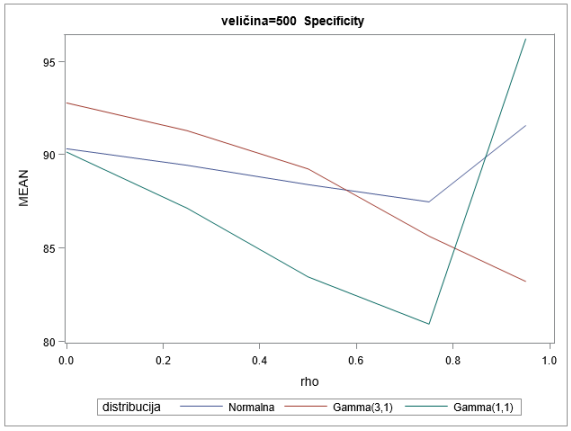
Slika 4.36: Somersov D



Slika 4.37: Stopa pogrešne klasifikacije

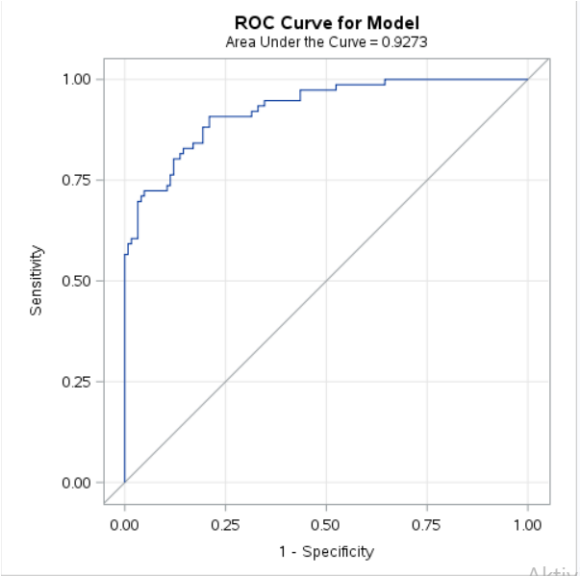


Slika 4.38: Osjetljivost

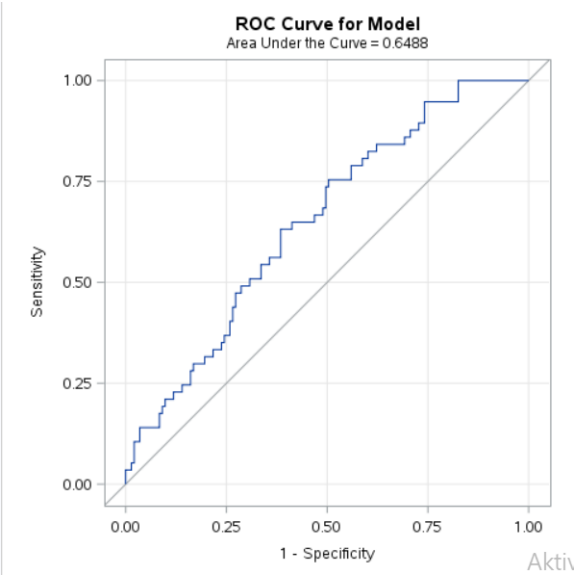


Slika 4.39: Specifičnost

Kako bi što zornije prikazali odnos između rezultata kod nekoreliranih i visokokoreliranih varijabli. na slikama 4.40 i 4.41 usporediti ćemo njihove ROC krivulje (vidi definicija 6). Na ovim slikama vidimo da je ROC krivulja prirodnijeg i oblijeg oblika za $\rho = 0$ te je površina ispod krivulje (c statistika) veća i iznosi 0.9273, dok je za $\rho = 0.95$ prilično mršava te je površina ispod krivulje manja i iznosi 0.6489. To nam pokazuje da je model sa $\rho = 0$ bolji.



Slika 4.40: ROC krivulja za $\rho = 0$



Slika 4.41: ROC krivulja za $\rho = 0.95$

4.4 Zaključak i ograničenja simulacije

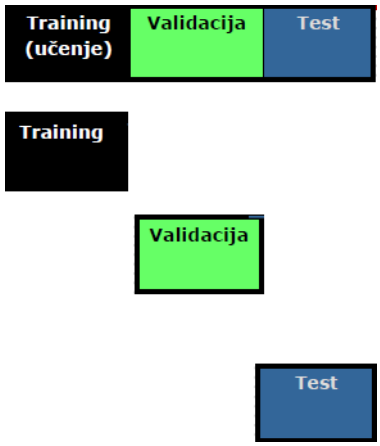
Uzevši u obzir statističke rezultate za mjere pogreške predviđanja u ovisnosti o varijacijama ulaznih podataka možemo zaključiti:

1. VELIČINA UZORKA - u ovoj simulaciji veličina uzorka se nije pokazala presudnom iako su stastički pokazatelji bili bolji i vjerodostojniji kako se veličina uzorka povećavala, no ne u tolikoj mjeri koliko se moglo očekivati prije početka simulacije. To nas pak dovodi do nove diskusije kolika je veličina uzorka dovoljna da bi se rezultati mogli smatrati reprezentativnima za donijeti neke zaključke, a opet da veličina uzorka ne bude nepotrebno prevelika usporavajući tako brzinu izvođenja programa.

2. BROJ PREDIKTORSKIH VARIJABLI - u ovom radu proučavali smo samo modele s jednom i dvije prediktorske varijable te su rezultati i točnost modela puno bolji za dvoparametarske modele. Za neke površne zaključke i odnose između distribucija prediktorski varijabli zbog manjka vremena i brzine izvođenja jednoparametarski modeli bi također poslužili svrsi, ali za izvođenje nekih većih zaključaka potrebni su modeli s više prediktorskih varijabli. Ovdje bi također mogli diskutirati o modelima s 3 i više prediktorskih varijabli, ali bi takav algoritam bio presložen i brzina izvođenja bi bila preduga.

3. - DISTRIBUCIJA - u većini modela koje smo proučavali, modeli s prediktorskim varijablama koje su bile distribuirane gamma distribucijom s parametrom oblika $k=3$ su imali najbolju prediktivnu sposobnost, dok kod modela gdje su varijable distribuirane gamma distribucijom s $k=1$ su imali najlošiju. Iz toga zaključujemo da su gamma distribucije s velikim k bolje za predviđanje rizika. Također smatramo da bi za dublju analizu trebalo testirati i neke druge distribucije kako bi se donijeli valjaniji zaključci.

4. KORELACIJA IZMEĐU PREDIKTORSKIH VARIJABLI - varijacije u korelacijama između prediktorskih varijabli najviše su uzrokovala točnost modela. Modeli u kojima su prediktorske varijable bile neovisne, tj. modeli u kojima je $\rho = 0$ su imali najbolju prediktivnu sposobnost i mjere pogreške predviđanja rizika su imale najbolje statističke rezultate. Povećavanjem korelacije između prediktorskih varijabli statistički rezultati su znatno lošiji i modeli imaju neprihatljivu prediktivnu sposobnost. Najbolja ilustracija za pokazati tu tvrdnju je na slikama 4.40 i 4.41 gdje su prikazane ROC krivulje za $\rho = 0$ i $\rho = 0.95$. Upravo površina ispod ROC krivulje koja predstavlja c statistiku je jedan od bitnijih statističkih pokazatelja koje smo proučavali te je na prvoj slici za $\rho = 0$ ta površina znatno veća.



Slika 4.42: Podjela na nezavisne podatke za razvoj,validaciju i ocjenjivanje modela [2]

Najveće ograničenje ove simulacije je što su sve statistike procjenjivane na istim podacima na kojima su procjenjivani parametri modela. Te su procjene pristrane (previše 'optimistične') jer bi procjene pogrešaka na nezavisnim podacima bile odveć 'pesimistične'. Za procjenu pogrešaka na nezavisnim podacima potrebno je bilo particioniranje podataka, tj. podjela podataka na 3 podskupa. Prvi dio su trening podaci, tj. podaci za *razvoj modela* i aproksimiranje funkcije. Drugi dio podataka bi se trebao odnositi na *validaciju modela*, tj. usporedbu ponašanja različitih aproksimacija sa ciljem odabira najboljeg modela. Treći dio podataka bi se trebao odnositi na *ocjenjivanje modela*, tj. nakon odabira završnog modela, procjenjuje se pogreška predikcije na novim nezavisnim podacima. Uobičajeno je da su podaci podijeljeni u omjeru $50/25/25$, ali u pravilu omjeri mogu biti i drugačiji.

U stvarnim primjenama ovih modela u procjenama rizika, osim na nezavisnim testnim podacima pogreške predikcije se mogu procjenjivati i primjenom krosvalidacije. Proces krosvalidacije se odvija tako da se podaci podijele na N podjednakih dijelove i za svaki $k = 1, \ldots, N$ se izbaci k-ti dio podataka te se računaju pogreške predikcije na prilagođenom modelu bez k-tog dijela. Poslije toga se svih N procjena pogreške predikcije združi i iz toga se donose zaključci o valjanosti modela.

Zbog fokusiranja na specifične i drugačije definirane ciljeve rada te zbog vremenskih i ostalih ograničenja vezanih uz složenost algoritama i brzinu izvođenja simulacije takav način podjele podataka nije pokriven u ovom radu.

Bibliografija

[1] D.B. Edelman L.C. Thomas i J.N. Crook, *Credit scoring and its applications*, Society for industrial and Applied Mathematics, 2002.

[2] Vesna Lužar-Stiffler, *Kvantitativne metode: pregled metoda multivarijantne analize*.

[3] M. Makek, *Statistika i osnovna mjerenja*, http://www.phy.pmf.unizg.hr/~makek/som/predavanja.

[4] P. McCullagh i J.A. Nelder, *Generalized Linear Models Second edition*, Chapman and Hall, 1991.

[5] Princeton University, *Logit models for binary data*, https://data.princeton.edu/wws509/notes/c3.pdf?fbclid=IwAR2qQhOQTXiIyDqHHHsIFlkh_-dZAjXRebJ8u9Ed_W-KafX5miVwW-uJWoY.

[6] R. Wicklin, *Simulating data with SAS*, SAS Institute, 2013.

[7] Wikipedia, *Gamma distribution*, https://en.wikipedia.org/wiki/Gamma_distribution.

[8] S.A. Sivo S.C. Keenan X. Fan, A. Felsovalyi, *SAS for Monte Carlo Studies: A Guide for Quantitative Researches*, SAS Institute Inc., 2002.

Sažetak

U ovom diplomskom radu cilj je bio analizirati mjere pogreške predikcije rizika u bankama i kreditnim institucijama. Za dobivanje rezultata koristili smo metodu binomne logističke regresije. U prvom dijelu diplomskog rada opisana je povijest kreditnog bodovanja, statističke i nestatističke metode koje se mogu koristiti za predikciju rizika te općenito o generaliziranim linearnim modelima i logističkoj regresiji. U drugom dijelu diplomskog rada opisan je Monte Carlo eksperiment pomoću kojeg smo ispitali mjere pogreške predviđanja rizika pri raznim uvjetima kao što su distribucija prediktorskih varijabli, broj prediktorskih varijabli, veličina uzorka i korelacija između varijabli.
Rezultati su pokazali da su najbolji modeli Gamma distribucije s većim parametrom oblika k te da korelacija jako utječe na statističke modele. Modeli u kojem su varijable neovisne te modeli s više prediktorski varijabli su pokazali bolju prediktivnu sposobnost.

Summary

In this thesis, the aim was to analyze risk measures in banks and credit institutions. We used the binomial logistic regression method to obtain results. The first part of the thesis deals with the history of credit scoring, statistical and nonstatistical methods that can be used for risk prediction and about generalized linear models and logistic regression in general. In the second part of the thesis, we have described the Monte Carlo experiment by which we tested risk estimation measures at various perspectives such as probability distribution of predictor variables, number of predictor variables, sample size, and correlation between variables.
The results showed that the best models are when predictor variables have Gamma distribution with the greater shape parameter k and that the correlation strongly influences the statistical models. Models in which variables are independent and models with more predictor variables have shown better predictive capability.

Životopis

Rođen sam 31.03.1994. u Splitu gdje i živim. Tamo sam pohađao Osnovnu školu Lučac koju završavam 2008. godine te potom upisujem III. Gimnaziju u Splitu popularniju kao MIOC. 2012. godine upisujem Prirodoslovno-matematički fakultet u Splitu, smjer matematika. 2016. godine stekao sam titulu sveučilišnog prvostupnika matematike i te sam godine upisao Diplomski sveučilišni studij financijske i poslovne matematike na Prirodoslovno-matematičkome fakultetu u Zagrebu.