UNIVERZITET U NOVOM SADU
PRIRODNO – MATEMATICKI FAKULTET
DEPARTMAN
ZA MATEMATIKU I INFORMATIKU

RAZLICITI PRISTUPI KREDITNOM SKORING SISTEMU

- master rad -

Profesor: dr. Zorana Lužanin

Autor: Jelena Burgijašev

Novi Sad, Septembar 2009.

Sadržaj:

1. Uvod

Kako se kreditno tržište sve više razvija, sve je veca potreba za dobro razvijenim sistemom za zaštitu od rizika i gubitaka, i u tu svrhu je uveden kreditni skoring sistem. U ovom radu su prezentovani razliciti pristupi kreditnom skoring sistemu koji su razvijeni primenom matematickog aparata. Ovaj model se najcešce srece u bankama kada odobravaju kredit, pri proceni kreditne sposobnosti klijenta i rizika koji taj klijent nosi. U zavisnosti od kreditne politike banke postoje razni pristupi ovom problemu, kao što su na primer minimizacija rizika, maksimizacija profita ili nešto trece.

Za razvijanje modela najcešce se koriste statisticke i ekonometrijske tehnike, kao što su na primer logit i probit analiza. Metode separacije nam služe za klasifikaciju klijenata u dve grupe: dobre i loše, tj. one koji redovno izmiruju svoje obaveze i one koji imaju problema u poslovanju pa samim tim kasne u otplati kredita. Cilj je što tacnije klasifikovati klijente, pošto se time u isto vreme minimiziraju troškovi koji nastaju pri pogrešnoj klasifikaciji. Prilikom separacije pomocu linearne, odnosno kvadratne površi primenjuje se problem linearnog programiranja za minimizaciju greške klasifikacije. U jednom primeru cemo uporedivanjem genetskog programiranja i probit analize pokazati da se uporedo sa razvojem kreditnog tržišta razvijaju i precizniji modeli.

Kroz date primere se vidi da se u razlicitim zemljama primenjuju razliciti modeli i pristupi kreditnom skoringu, koji u nekim slucajevima daju i nelogicne rezultate sa naše tacke gledišta, kao što cemo videti na primeru Vijetnamskog bankarskog tržišta. Vecina modela se zasniva na istorijskim podacima, i zbog toga je bitno da uzorak bude što veci da bi dobili što bolji i realniji model. Kao problem može da se javi nedovoljna kolicina podataka, jer je kreditno tržište još uvek u fazi razvoja, kao i odredena pristrasnost modela pošto se posmatraju samo klijenti kojima je odobren kredit. Ovo su problemi kojima se treba baviti u buducnosti, da bi se razvio što pouzdaniji i realniji model.

Pojam kreditnog skoringa se uglavnom vezuje za analizu kreditne sposobnosti pojedinacnog klijenta, medutim sve je veca njegova primena i pri ocenjivanju rizika koji nose firme. Ovaj tip kreditnog skoringa je još uvek u fazi razvoja i još uvek nije toliko zastupljen u praksi, iako cemo kroz primer videti da on daje dosta precizne rezultate.

2. Osnove kreditnog skoring sistema

Kada jedan bankar pita drugog „Koji je skor?“, vecina ljudi ce pomisliti da ovaj prvi nije gledao sinocnu utakmicu. Medutim, oni samo rade svoj posao, tacnije, raspravljaju o tome koliki je kreditni skor klijenta koji podnosi zahtev za kredit. ***Kreditni skoring*** je statisticki metod koji se koristi da se oceni koliki je rizik da ce novi klijent koji podnosi zahtev za kredit ili vec postojeci klijent kasniti u otplati kredita ili da uopšte nece biti u stanju da isplati kredit. Drugim recima, kreditni skoring je metod za odredivanje kreditnog rizika koji nosi klijent. Koristeci istorijske podatke i statisticke tehnike kreditni skoring pokušava da izoluje efekte odredenih karakteristika koje klijenta vode u situaciju da ne može da otplacuje kredit. Rezultat ovog metoda je skor koji banka koristi da rangira klijente na osnovu rizika koji oni nose. Na osnovu toga koliki rizik banka želi da prihvati, odreduje se granicni skor, tako da ce klijentima koji imaju veci skor od granicnog biti odobren kredit, a oni koji imaju manji skor ce biti odbijeni. Da bi napravili model, analiticari analiziraju istorijske podatke ranije odobrenih kredita, tj. analiziraju kako su se raniji klijenti ponašali pri otplati kredita, da bi odredili karakteristike koje su korisne pri oceni da li je klijent sposoban da redovno otplacuje kredit. Samim tim su i klijenti podeljeni na dobre i loše. Klijenta definišemo da je ***loš*** ukoliko nije platio tri uzastopne rate, tj. u kašnjenju je dužem od devedeset dana. Problem se javlja kada se pravi model za relativno nove proizvode zbog nedovoljnog obima istorijskih podataka na osnovu kojih bi se razvio model. U tom slucaju se model može praviti na malom uzorku, ili na nekom slicnom proizvodu, ali tada rezultati nece biti dovoljno dosledni. Dobro napravljen model bi trebao da dodeli veci skor klijentima koji ce redovno otplacivati kredit, a niži skor klijentima koji ce kasniti u otplacivanju kredita. Medutim, nijedan model nije savršen, pa se dešava da neki loši klijenti dobiju veci skor od nekih dobrih klijenata.

Razlikujemo dva modela kreditnog skoring sistema:

1. model koji pomaže analiticaru da donese odluku da li da odobri kredit klijentu ili da ga odbije
2. model koji pomaže u analiziranju ponašanja vec postojecih klijenata

Mi cemo se u ovom radu fokusirati na prvi model kreditnog skoring sistema.

Pri oceni kreditne sposobnosti klijenta koriste se dva izvora informacija: prvi je aplikacioni formular koji klijent popunjava kada podnosi zahtev za kredit, a drugi je kreditni biro. ***Kreditni biro*** je institucija koja poseduje veliku bazu podataka o svim klijentima, kako licnih podataka tako i informacije o urednosti otplate kredita. Ova informacija se naziva ***kreditna istorija***. Kada klijent podnosi zahtev za kredit, kreditni analiticar proverava sa kreditnim biroom kakva je klijentova kreditna istorija. Ako je kreditna istorija «siromašna», analiticar može odbiti da odobri kredit. U situaciji kada analiticar i odluci da odobri kredit, sigurno ce odrediti i vecu kamatnu stopu.

Najzastupljeniji statisticki modeli koji se koriste da bi se napravio kreditni skoring sistem su linearni model, logit model, probit model i model diskriminantne analize. Prva tri modela su standardne statisticke tehnike koje na osnovu istorijskih podataka i karakteristika klijenta ocenjuju verovatnocu da ce taj klijent biti loš.

Razlika medu ovim modelima je ta što linearni model pretpostavlja linearnu zavisnost izmedu verovatnoce da ce klijent biti loš i karakteristika koje se koriste u modelu, logit model pretpostavlja da ova verovatnoca ima logaritamsku raspodelu, a probit model da ima normalnu raspodelu. Diskriminantna analiza se od ovih statistickih modela razlikuje po tome što umesto da ocenjuje verovatnocu da ce klijent biti loš, ona deli klijente u dve klase, manje i više rizicne. Na jednom primeru cemo prikazati primenu genetskog programiranja u kreditnom skoring modeliranju kao i primenu analize obavijanja podataka, ali necemo ulaziti u detalje, pošto to izlazi iz opsega ovog rada. Napomenimo da za kreditni skoring model mogu da se koriste i razni neparametarki i modeli veštacke inteligencije.

Koliko god model bio dobar, on nece sa sigurnošcu predvideti kako ce se klijent ponašati pri otplati kredita, ali bi ipak trebalo da dâ dovoljno tacnu ocenu verovatnoce da klijent sa odredenim karakteristikama nece biti u mogucnosti da otplati kredit. Takode treba uzeti u obzir da postoji i odredena pristrasnost u uzorku, pošto se uzimaju u obzir samo klijenti kojima je odobren kredit. Ovo se javlja zbog toga što se ne zna kako bi se odbijeni klijenti ponašali da im je odobren kredit.

Prednosti kreditnog skoringa koje su dovele do njegove sve vece primene su kao prvo ušteda vremena koje je potrebno da se donese odluka u procesu odobravanja kredita, što je vrlo bitna karakteristika u trenutnoj ekspanziji kreditnog tržišta. Takode, ovaj sistem donosi odluku na osnovu istih kriterijuma za sve klijente, nezavisno od pola, rase ili nekog drugog kriterijuma koji može dovesti do odredene diskriminacije. Da bi se neka od ovih karakteristika koristila u modelu, moraju da se imaju dovoljno jaki argumenti zašto je ona toliko bitna za model.

Prvo cemo se upoznati sa matematickim aparatom koji koristimo da bismo razvili model i došli do željenih rezultata, a zatim cemo kroz primere objasniti ove matematicke modele.

3. Matematicki aparat

3.1 Linearno programiranje

Problem linearnog programiranja služi za modeliranje tzv. uslovne optimizacije u kojoj treba naci optimalno rešenje, tj. ono rešenje za koje se postiže najbolja vrednost nekog cilja na skupu svih mogucih alternativnih rešenja problema, pri cemu svako rešenje iz ovog skupa zadovoljava zadate uslove (ogranicenja).

Opšti zadatak linearnog programiranja se može iskazati u formalizovanom razvijenom obliku kao

$$\begin{Bmatrix}\max\\ \min\end{Bmatrix} f(x) = c_1x_1 + c_2x_2 + ... + c_nx_n$$

uz ogranicenja

$$\begin{aligned} &a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n \overset{<}{=}_{>} b_1 \\ &a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n \overset{<}{=}_{>} b_2 \qquad (*) \\ &\quad\vdots \\ &a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n \overset{<}{=}_{>} b_m \\ &x_1 \geq 0, x_2 \geq 0, ..., x_n \geq 0 \qquad (**) \end{aligned}$$

ili krace

$$\begin{Bmatrix}\max\\ \min\end{Bmatrix} f(x) = \sum_{i=1}^{n} c_i x_i$$

$$\sum_{j=1}^{n} a_{ij}x_j \overset{<}{=}_{>} b_i, \quad i = 1,...., m$$

$$x_j \geq 0, \qquad j = 1,...., n$$

gde su a_{ij}, c_i, b_j, $i = 1,2,..,m$, $j = 1,2,...,n$ zadati realni brojevi. Funkcija $f(x)$ se naziva ***funkcija cilja***, dok se m ogranicenja iz skupa (*) nazivaju jednostavno ***ogranicenja***. Ogranicenja (**) se nazivaju ***uslovi nenegativnosti***.

Dopustiva oblast skupa (*) - (**) se naziva ***dopustiva oblast*** problema linearnog programiranja, a svaka tacka dopustive oblasti predstavlja ***dopustivo rešenje*** ovog problema. Ono dopustivo rešenje u kome funkcija dostiže svoj ekstrem se naziva ***optimalno rešenje*** problema. Matrica $A = [a_{ij}]_{mxn}$ se cesto naziva matricom ogranicenja, a b_i, $i = 1,2,..., m$ slobodnim clanom i-tog ogranicenja iz sistema (*).

Ako su sva ogranicenja u (*) istog tipa i $c = (c_1, c_2,..., c_n)$, $b = (b_1, b_2,..., b_m)$ i $x = (x_1, x_2,..., x_n)$ tada se problem linearnog programiranja može prikazati i u matricno-vektorskoj formi. Ako su sva ogranicenja tipa jednakosti, tada imamo ***standardni oblik*** linearnog programiranja koji ima sledecu matricno-vektorsku formu:

$$\begin{Bmatrix}\max\\ \min\end{Bmatrix} c^T x$$

$$Ax = b \qquad (***)$$

$$x \geq 0$$

Pretpostavljamo da ovaj problem zadovoljava sledece dodatne uslove:

1. svi slobodni clanovi b_i, $i = 1,2,..,m$ su nenegativni, tj. $b \geq 0$
2. $m < n$ i $rangA = m$, tj. sve jednacine ogranicenja su linearno nezavisne

Posmatramo sistem linearnih jednacina:

$$Ax = b \qquad (\&)$$

Definicija ***Bazno rešenje*** sistema (&) predstavlja ono rešenje ovog sistema za koje postoji neka *mxm* podmatrica A_B matrice A koja je regularna, tj. detA≠0 i takva da je u ovom rešenju svaka promenljiva cija se kolona koeficijenata ne nalazi u A_B jednaka nuli. Promenljive koje odgovaraju kolonama podmatrice A_B se nazivaju ***bazne promenljive***, a preostale promenljive su nebazne. Skup svih baznih promenljivih cini ***bazu*** baznog rešenja, dok je A_B ***matrica baze***.

Primetimo da rešenje ovakvog sistema uvek postoji pošto je A_B regularna podmatrica. Kako je $rangA = m$, postoji bar jedna regularna *mxm* podmatrica od A, pa time i bar jedno bazno rešenje sistema (&).

Definicija Ako neko bazno rešenje problema (&) zadovoljava i uslove nenegativnosti onda ono predstavlja ***bazno dopustivo rešenje*** problema (***).

Definicija Dva bazna dopustiva rešenja problema (&) su ***susedna*** ako im se baze razlikuju samo u jednoj promenljivoj.

Jedan od razloga tako rasprostranjene primene modela linearnog programiranja u praksi je i postojanje efikasnih procedura za njihovo rešavanje koje sa uspehom rešavaju i probleme velikih dimenzija. Medu njima najpoznatija je ***Simpleks metoda***. Ona predstavlja algebarsku proceduru koja nalazi niz baznih dopustivih rešenja x_k problema (&) sa odgovarajucim bazama B_k, $k = 0,1,2,...$ tako da je vrednost funkcije cilja u x_{k+1} veca od njene vrednosti u x_k, sve dok trenutno generisano bazno dopustivo rešenje ne zadovolji kriterijum optimalnosti.

Kriterijum optimalnosti Ako je $c_j^k \leq 0$, za svako $j = 1,2,...,n$ tada je x_k optimalno rešenje problema linearnog programiranja, a f_k maksimalna vrednost funkcije cilja.

Drugim recima, x_k je optimalno ako se nikakvim povecanjem vrednosti nebaznih promenljivih ovog rešenja ne može povecati vrednost funkcije cilja. Necemo dublje ulaziti u postupak Simpleks metoda.

3.2 Problem separacije i klasifikacije

Problem separacije je problem nalaženja kriterijuma za razdvajanje elemenata u dva disjunktna skupa objekata. Osnovni problem separacije se može definisati na sledeci nacin: neka su data dva skupa objekata *A* i *B*, pri cemu skup *A* sadrži *m* objekata, a skup *B* sadrži *k* objekata i pri tome je svaki objekat jedan *n*-dimenzionalni vektor. Problem separacije se tada svodi na pronalaženje površi u *n*-dimenzionalnom prostoru tako da se sve tacke koje predstavljaju objekte iz skupa *A* nalaze sa jedne strane ove površi, dok se sve tacke koje predstavljaju objekte iz skupa *B* nalaze sa druge strane površi.

3.2.1 Separacija pomocu linearne površi

Definišemo *skup objekata* kao nepraznu matricu realnih brojeva, gde svaka vrsta definiše pojedinacni objekat. Objekat se sastoji od *n* realnih brojeva koji se nazivaju *opservacije*. Ako matricu objekata oznacimo sa *A*, pojedinacne objekte kao vektor vrste a_i, tada je *j*-ta opservacija *i*-tog objekta oznacena sa a_{ij}. Osnovni problem ***linearne separacije*** glasi: za data dva skupa objekata koja su definisana matricama $A \in \boldsymbol{R}^{mxn}$ i $B \in \boldsymbol{R}^{kxn}$, odrediti hiperravan u prostoru $\boldsymbol{R}^n$ tako da ako *m* vrsta matrice *A* i *k* vrsta matrice *B* posmatramo kao tacke u ovom prostoru, tada se one moraju nalaziti sa razlicitih strana ove ravni. Neka su a_i, $i=1,...,m$ objekti koji pripadaju matrici *A*, a b_j, $j=1,..,k$ objekti koji pripadaju matirici *B*. Ako postoje nenula vektori $c \in \boldsymbol{R}^n$ i $c_0 \in \boldsymbol{R}^n$ takvi da je

$$c^T a_i > c_0 \text{ , } i=1,\ldots m$$
$$c^T b_j < c_0 \text{ , } j=1,\ldots,k$$

tada se $H(c,c_0)=\{x \in \boldsymbol{R}^n \mid c^T x=c_0\}$ naziva ***hiperravan.*** Ako $x_0 \in H(c,c_0)$ tada važi da je $c^T x_0 = c_0$ pa je

$$H(c,c_0) = \{ x \in \boldsymbol{R}^n \mid c^T x=c^T x_0=c_0\} = \{ x \in \boldsymbol{R}^n \mid c^T (x-x_0)=c_0\}$$

Izraz $c^T (x-x_0) = 0$ predstavlja ortogonalnost vektora *c* i $x-x_0$ i zbog toga se vektor *c* naziva normalni vektor hiperravni $H(c,c_0)$.

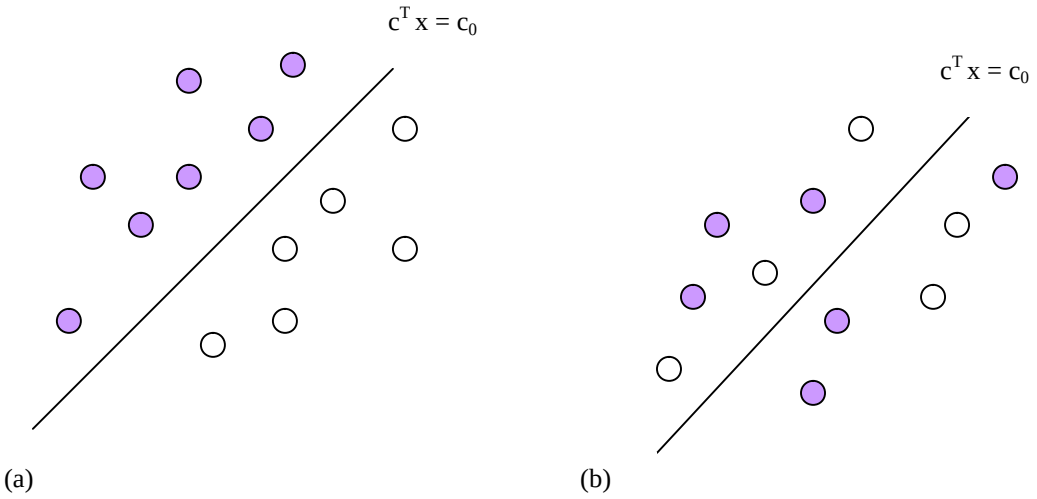
Ako uvedemo granicu debljine δ izmedu klasa, onda važi:

$$c^T a_i \geq c_0 + d \text{ , } i = 1,...,m$$
$$c^T b_j \leq c_0 - d \text{ , } j = 1,...,k$$

U opštem slucaju ove nejednakosti nisu zadovoljene, ali mi težimo da one budu bar približno zadovoljene.

Poluprostore definišemo na sledeci nacin:

$$H_+(c,c_0) = \{x \in \boldsymbol{R}^n \mid c^T x>c_0 \}$$
$$H_-(c,c_0) = \{x \in \boldsymbol{R}^n \mid c^T x<c_0 \}$$



(a) Hiperravan potpuno razdvaja objekte dve razlicite klase;

(b) Hiperravan koja razdvaja objekte razlicitih klasa ne postoji

Klasifikacija može da se definiše kao funkcija klasifikacije koja svakom objektu dodeljuje broj, odnosno svaki objekat «smešta» u jednu od klasa, u ovom slucaju u jedan poluprostor. Za objekat A_i za koji je $a_i \in H_-(c,c_0)$ kažemo da je *pogrešno klasifikovan objekat prvog reda*. Za objekat B_j za koji je $b_j \in H_+(c,c_0)$ kažemo da je *pogrešno klasifikovan objekat drugog reda*. Sa y_i oznacavamo rastojanje $a_i \in H_-(c,c_0)$ od hiperravni $H(c,c_0+d)$, a sa z_j rastojanje $b_j \in H_+(c,c_0)$ od hiperravni $H(c,c_0-d)$. Klasifikaciju vršimo tako što minimiziramo težinsku sumu rastojanja pogrešno klasifikovanih objekata, tj. rešavamo sledeci problem linearnog programiranja:

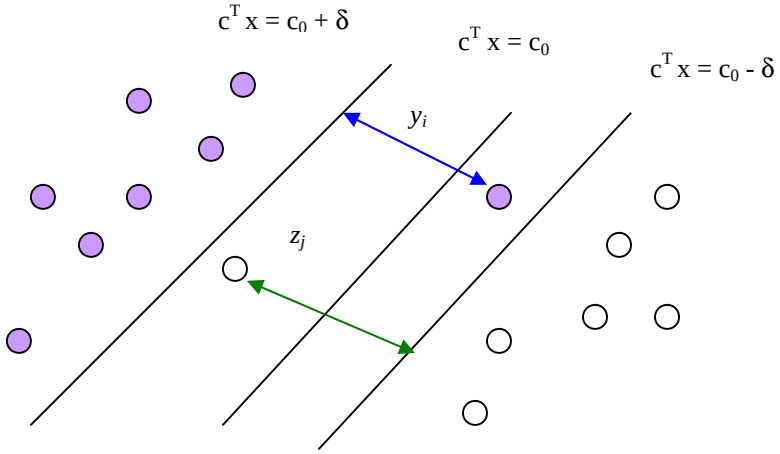
$$\min_{y,z} a\sum_{i=1}^{m} y_i + b\sum_{j=1}^{k} z_j$$

$$c^T a_i + y_i \le c_0 + d\ ,\ i=1,...,m$$

$$c^T b_j - z_j \le c_0 - d\ ,\ j = 1,...,k$$

$$y_i,\ z_j \ge 0$$

gde su α, β težinski koeficijenti. Ovaj problem linearnog programiranja ima optimalno rešenje (c^*, c_0^*, y_1^*,..., y_m^*, z_1^*,...., z_k^*) i pri tome je $H(c^*,c_0^*)$ hiperravan sa najmanjom greškom klasifikacije.



Pogrešno klasifikovani objekti

3.2.2 Separacija pomocu kvadratne površi

Ponekad nije moguce da se dva skupa objekata razdvoje pomocu ravni. Zbog toga se za separaciju koriste površi koje su nelinearne, takva je na primer kvadratna separacija. Neka je $x=(x_1,...,x_n)^T$ vektor koji predstavlja tacku u n-dimenzionalnom prostoru. Problem ***kvadratne separacije*** skupova A i B sastoji se u odredivanju jedinstvene kvadratne površi:

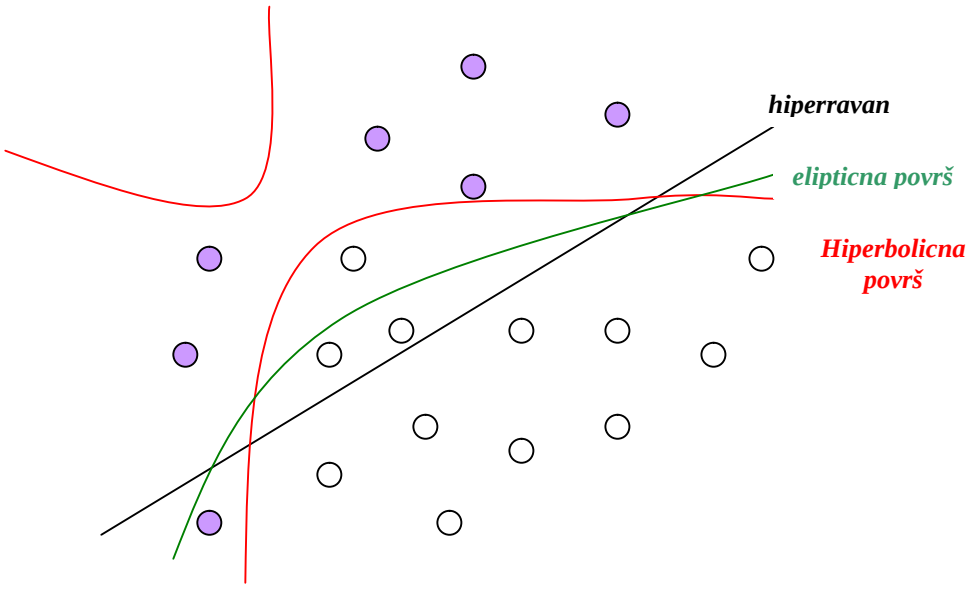
$$x^T Dx + x^T c - c_0 = 0$$

gde su $D \in \boldsymbol{R}^{nxn}$, $c \in \boldsymbol{R}^n$, $c_0 \in \boldsymbol{R}$, tako da je

$$a_i^T Da_i + a_i^T c - c_0 > 0,\ i = 1,...,m$$

$$b_j^T Db_j + b_j^T c - c_0 < 0,\ j = 1,...,k$$

Dva skupa A i B su ***kvadratno separabilna*** ako i samo ako postoje D, c i c_0 takvi da su gornje jednakosti zadovoljene.



Neke kvadratne površi

Sledeci problem linearnog programiranja minimizira grešku klasifikacije:

$$\min_{y,z} a\sum_{i=1}^{m} y_i + b\sum_{j=1}^{k} z_j$$

$$a_i^T D a_i + c^T a_i + y_i \geq c_0 + d\ ,\ i=1,...,m$$

$$b_j^T D b_j + c^T b_j - z_j \leq c_0 - d\ ,\ j=1,...,k$$

$$y_i,\ z_j \geq 0$$

gde su $D \in \boldsymbol{R}^{nxn}$, $c \in \boldsymbol{R}^n$, $c_0 \in \boldsymbol{R}$, y_i, $z_j \in \boldsymbol{R}$ promenljive koje treba odrediti. Promenljive y_i, z_j predstavljaju rastojanje pogrešno klasifikovanih objekata od kvadratne površi.

3.2.3 Model klasifikacije uz pomoc funkcije korisnosti

Prvo se podsetimo pojma relacije preferencije i funkcije korisnosti.

Definicija Binarna relacija $\succ$ na skupu X koja ima osobine:

1) refleksivnost: $\forall\, x \in X\ \ x \succ x$
2) kompletnost: $\forall\, x,y \in X\ \ x \succ y \vee y \succ x$
3) tranzitivnost: $\forall\, x,y,z \in X\ \ x \succ y \wedge y \succ z \Rightarrow x \succ z$

naziva se ***relacija preferencije*** na skupu X.

Definicija Neka je $\succ$ relacija preferencije na skupu X. Funkcija u: $X \rightarrow \boldsymbol{R}$ za koju važi:

$$x \succ y \Leftrightarrow u(x) \geq u(y)$$

naziva se ***funkcija korisnosti*** relacije preferencije $\succ$.

Pošto je funkcija klasifikacije prekidna funkcija, za razdvajanje objekata koji pripadaju razlicitim klasama koristi se neprekidna funkcija korisnosti. Najbolja funkcija korisnosti neke klase se odreduje minimiziranjem greške klasifikacije. Ako tražimo funkciju korisnosti koja je linearna kombinacija nekih drugih funkcija onda problem možemo formulisati kao problem linearnog programiranja.

Sada posmatramo problem klasifikacije u J klasa. Pretpostavimo da postoji skup objekata $I=\{1,\ldots,m\}$ sa poznatim klasifikacijama i svaki objekat je predstavljen tackom iz $\boldsymbol{R}^n$. Ovaj skup tacaka $X=\{x^i | i \in I\}$ se naziva *training skup*. Pretpostavimo da razlaganje skupa I, $\{I_j\}_{j=1}^{J}$ $(I_{j1} \cap I_{j2} = \varnothing, j_1 \neq j_2, I = \bigcup_{j=1}^{J} I_j)$ definiše klasifikaciju datih m objekata. Kažemo da tacka x^i pripada klasi k ako $i \in I_k$.

Ako je data funkcija korisnosti $u(x)$ koja je definisana na skupu $X=\{x^i | i \in I\}$ i skup pragova $U_p=\{u_0, u_1,\ldots, u_{j-1}, u_j\}$ tada je $u_0 = -¥ < u_1 < \ldots < u_{j-1} < u_j = ¥$ i ako važi

$$i_j = \{i \in I \mid u_{j-1} < u(x^i) < u_j\}$$

tada funkcija $u(x)$ i skup U_p definišu razlaganje skupa I, $\{I_j\}_{j=1}^{J}$ i na taj nacin klasifikuju training skup. Medutim, cešci je slucaj da nam je poznata klasifikacija i, $\{I_j\}_{j=1}^{J}$, a da je potrebno pronaci funkciju korisnosti i skup pragova koji definišu ovu klasifikaciju. Neka $u^g(x)$ predstavlja klasu funkcija korisnosti koje su odredene parametrom $\gamma \in \boldsymbol{R^n}$. Pretpostavimo da je training skup $X=\{x^i | i \in I\}$ podeljen klasama $\{I_j\}_{j=1}^{J}$. Sa $F(s^+, s^-)$ oznacicemo ukupnu kaznu za sve pogrešno klasifikovane tacke. Za training skup X, sledeci problem optimizacije pronalazi najbolju funkciju korisnosti $u^g(x)$ u prethodno datoj klasi funkcija:

$$\min_{g,u,s^+,s^-} F(s^+, s^-)$$

$$u_{j-1} - s_i^- + d \le u^g(x^i) \le u_j + s_i^+, i \in I_j, j = 1,..., J$$

$$u_j - u_{j-1} \le s, j = 2,..., J-1$$

$$s_i^+ \ge 0, s_i^- \ge 0$$

$$g \in R^n$$

Funkcija $F(s^+, s^-)$ je neopadajuca u odnosu na greške klasifikacije s_i^+, s_i^-, $i \in I$. Veliko odstupanje od savršene klasifikacije dovodi do velikih kazni. Posmatramo linearnu kaznenu funkciju

$$F(s^+, s^-) = \sum_{i=1}^{m} (a_i^- s_i^- + a_i^+ s_i^+)$$

gde su $a_i^-, a_i^+ \ge 0$ koeficijenti kazne za tacki i. Ako objekat x^i pripada klasi I_j gornji problem matematickog programiranja implicira da je:

$$s_i^+ = \max\{0, u^g(x^i) - u_j\}$$

$$s_i^- = \max\{0, u_{j-1} + d - u^g(x^i)\}$$

Ako funkcija korisnosti $u^g(x)$ u tacki $x^i \in i_j$ premašuje gornji prag u_j, tada je greška s_i^+ jednaka razlici izmedu $u^g(x^i)$ i gornjeg praga u_j. U suprotnom $s_i^+ = 0$. Slicno, ako je vrednost funkcije korisnosti $u^g(x^i)$ ispod $u_{j-1}+d$ tada je kazna s_i^- jednaka razlici izmedu $u_{j-1}+d$ i $u^g(x^i)$, inace je $s_i^- = 0$.

3.2.4 Klasifikacija pomocu linearne funkcije korisnosti

Linearna funkcija korisnosti ima sledeci oblik:

$$u^c(x) = c^T x = \sum_{k=1}^{n} c_k x_k \ , c \in \boldsymbol{R}^n$$

Ako imamo linearnu kaznenu funkciju i linearnu funkciju korisnosti, tada se klasifikacija vrši rešavajuci sledeci problem:

$$\min_{c,u,s^-,s^+} \sum_{i=1}^{n} (a_i^- s_i^- + a_i^+ s_i^+)$$

$$\sum_{k=1}^{n} c_k x_k^i - u_{j-1} + s_i^- \geq s, i \in I_j, j = 1,...,J$$

$$\sum_{k=1}^{n} c_k x_k^i - u_j - s_i^+ \leq 0, i \in I_j, j = 1,...,J$$

$$u_j - u_{j-1} \leq s$$

$$s_i^+ \geq 0, s_i^- \geq 0, i = 1,...,m$$

$$c \in R^n$$

3.2.5 Klasifikacija uz pomoc kvadratne funkcije korisnosti

Posmatramo sledecu kvadratnu funkciju korisnosti:

$$u^{D',c}(x) = x^T D' x + c^T x = \sum_{k=1}^{n} \sum_{l=1}^{n} d'_{kl} x_k x_l + \sum_{k=1}^{n} c_k x_k$$

Ova funkcija ima $n^2 + n$ elemenata, pošto matrica D' ima n^2 elemenata, a vektor c ima n koeficijenata. Ekvivalentna definicija kvadratne funkcije je sledeca:

$$u^{D',c}(x) = \sum_{k=1}^{n} d'_{kk} x_k^2 + \sum_{k=1}^{n-1} \sum_{l=k+1}^{n} (d'_{kl} + d'_{lk}) x_k x_l + \sum_{k=1}^{n} c_k x_k$$

Ako uvedemo nove oznake:

$d_k = d'_{kk}$

$d_{kl} = d'_{kl} + d'_{lk}, k < l$

kvadratna funkcija korisnosti može biti zapisana na sledeci nacin:

$$u^{D',c}(x) = \sum_{k=1}^{n-1} d_k x_k^2 + \sum_{k=1}^{n-1} \sum_{l=k+1}^{n} d_{kl} x_k x_l + \sum_{k=1}^{n} c_k x_k$$

Ako imamo linearnu kaznenu funkciju i kvadratnu funkciju korisnosti rešavamo sledeci problem:

$$\min_{D,c,u,s^+,s^-} F = \sum_{i=1}^{m} (a_i^- s_i^- + a_i^+ s_i^+)$$

$$\sum_{k=1}^{n} d_k {x_k^i}^2 + \sum_{k=1}^{n-1} \sum_{l=k+1}^{n} d_{kl} x_k^i x_l^i + \sum_{k=1}^{n} c_k x_k^i - u_{j-1} + s_i^- \geq d, i \in I_j, j = 1,...,J$$

$$\sum_{k=1}^{n} d_k {x_k^i}^2 + \sum_{k=1}^{n-1} \sum_{l=k+1}^{n} d_{kl} x_k^i x_l^i + \sum_{k=1}^{n} c_k x_k^i - u_j - s_i^+ \leq 0, i \in I_j, j = 1,...,J$$

$$u_j - u_{j-1} \leq s$$

$$s_i^+ \geq 0, s_i^- \geq 0, i = 1,...,m, \quad d \in R^n, c \in R^n$$

3.3 Logit analiza

Regresiona analiza se bavi proucavanjem odnosa izmedu zavisne i jedne ili više nezavisnih promenljivih. Zavisna promenljiva se uglavnom oznacava sa *y*, a nezavisne sa $x_1,..., x_n$. Jedan od mogucih ciljeva regresione analize je:

1. da oceni srednju odnosno ocekivanu vrednost zavisne promenljive, pri cemu su date vrednosti nezavisnih promenljivih
2. da testira hipoteze o prirodi zavisnosti izmedu promenljivih
3. da predvidi ocekivanu vrednost zavisne promenljive, pri cemu su date vrednosti nezavisnih promenljivih
4. kombinovano više prethodnih ciljeva

Linearna regresija je tip regresije koja pretpostavlja da postoji linearna zavisnost izmedu zavisne i nezavisnih promenljivih. Jednodimenzionalna linearna regresija se koristi u slucaju kada imamo jednu zavisnu i jednu nezavisnu promenljivu. Ukoliko imamo više nezavisnih promenljivih koristimo višedimenzionalnu regresiju.

U slucaju ***jednodimenzionalne regresije*** posmatramo sledecu jednacinu:

$$y_i = a + bx_i + u_i \ , \ i = 1,..., n$$

gde su α (*odsecak* ocekivana vrednost za *y* ukoliko je *x* jednako nula) i β (ocekivana vrednost za *y* ukoliko se *x* promeni za jednu jedinicu) *koeficijenti regresije* koje treba oceniti ako je dato *n* opservacija za *y* i *x*, a *u* je *greška regresije*. Da bi mogli da ocenimo parametre moramo da navedemo nekoliko pretpostavki:

1. $E(u_i) = 0, \forall i$
2. $\text{var}(u_i) = s^2, \forall i$
3. u_i i u_j su nezavisne za svako $i \neq j$
4. u_i i x_j su nezavisne za svako *i* i *j*
5. u_i imaju normalnu raspodelu za svako *i* sa nula ocekivanjem i σ^2 varijansom (što sledi iz 1. i 2. pretpostavke)
6. $\sum (x_i - \bar{x})^2 \neq 0$
7. $\sum (x_i - \bar{x})^2 < \infty$

Sada možemo da zapišemo kako izgleda *funkcija jednodimenzionalne regresije*:

$$\hat{y}_i = \hat{a} + \hat{b}x_i$$

gde je sa kapicom oznacena ocekivana vrednost, a znamo da je $E(u_i) = 0, \forall i$. Stohasticka verzija gornje jednacine je

$$y_i = \hat{a} + \hat{b}x_i + \hat{u}_i$$

gde $\hat{u}_i$ predstavlja ocenjenu grešku koja se naziva *rezidual*. Iz ove jednacine se dobijaju ocene za parametre regresije α i β. Tri metode koje se najcešce koriste za ocenu parametara su:

1. metoda najmanjih kvadrata
2. metoda momenta
3. metoda maksimalne verovatnoce

Ovim metodama dobijamo nepristrasne parametre koji imaju najmanju varijansu u klasi linearnih nepristrasnih ocenjenih parametara. Mi u ovom radu necemo objašnjavati ove metode, sem trece koja se koristi pri oceni parametara u slucaju logit regresije.

U slucaju ***višedimenzionalne regresije*** se posmatra odnos zavisne *y* i više nezavisnih $x_1,...,x_n$ promenljivih, pa imamo sledeci model:

$$y_i = a + b_1 x_{1i} + b_2 x_{2i} + ... + b_k x_{ki} + u_i, i = 1,..., n$$

gde je α odsecak (ocekivana promena u *y* ako su svi x_i, $i = 1,...n$ jednaki nula), a β_i parcijalni koeficijenti korelacije (ocekivana vrednost za *y* ukoliko dode do jedinicne promene u x_i pri cemu se pretpostavlja da su svi ostali *x*–evi nepromenjeni). Važe iste pretpostavke kao i u slucaju jednodimenzionalne regresije, samo što je dodata još jedna koja kaže da medu nezavisnim promenljivama nema linearne zavisnosti. Analogno imamo stohasticku jednacinu:

$$y_i = \hat{a} + \hat{b}_1 x_{1i} + \hat{b}_2 x_{2i} + ... + \hat{b}_k x_{ki} + \hat{u}_i, i = 1,..., n$$

Metod najmanjih kvadrata daje ocene za α, $\beta_1,...$, β_k koje su nepristrasne i imaju najmanju varijansu u klasi svih linearnih nepristrasnih ocenjenih parametara.

Medutim, dešava se da je zavisna promenljiva binarna, tj. da uzima samo dve vrednosti, na primer uspeh i neuspeh. Postavlja se pitanje kako se modelira odnos nezavisnih promenljivih i binarne zavisne promenljive? Odgovor na to pitanje nam daje ***logit regresija.***

Pretpostavljemo da je *Y* Bernulijeva slucajna promenljiva koja uzima vrednosti 0 ili 1, u zavisnosti da li je ishod dobar ili loš. Verovatnoca da ce ishod biti loš u zavisnosti od datih nezavisnih promenljivih, tj. da ce *Y* biti jednako 1, se definiše kao $\pi = P(Y = 1| X = x)$, a verovatoca da ce ishod biti dobar sa $1-\pi = P(Y = 0 | X = x)$. Posmatramo odnos (*odds*) ove dve verovatnoce:

$$odds(x) = \frac{P(Y = 1 | X = x)}{P(Y = 0 | X = x)} = \frac{p}{1-p}$$

Logaritam ovog odnosa, tzv. *logit* je linearna funkcija nezavisnih promenljivih x_i, $i = 0,1,2,...,n$:

$$\ln(odds(x)) = b_0 + b_1 x_1 + ... + b_n x_n$$

Dakle,

$$\ln(\frac{p}{1-p}) = b_0 + b_1 x_1 + ... + b_n x_n$$

Jednostavnim eksponencijalnim transformacijama dobijamo da je:

$$p = \frac{\exp(b_0 + b_1 x_1 + ... + b_n x_n)}{1 + \exp(b_0 + b_1 x_1 + ... + b_n x_n)}$$

Ako je promenljiva $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$, $i = 1,2,....N$, tada je verovatnoca pozitivnog ishoda jednaka:

$$p(\mathbf{x}_i) = \frac{\exp(\sum_{k=1}^{p} b_k x_{ik})}{1 + \sum_{k=1}^{p} b_k x_{ik}}$$

Pretpostavke koje važe kod logit regresije su sledece:

1. Y ima Bernulijevu raspodelu sa parametrom $\pi(x)$, tj.

$$Y : \begin{pmatrix} 0 & 1 \\ 1-p(x) & p(x) \end{pmatrix}$$

2. Nijedna promenljiva od znacaja nije izostavljena i nijedna promenljiva koja nema znacaja nije ukljucena
3. Logaritam nezavisnih promenljivih i zavisna promenljiva su linearno nezavisne
4. Nema znacajne korelacije izmedu nezavisnih promenljivih

3.3.1 Selekcioni kriterijumi

U situaciji kada imamo veliki broj nezavisnih promenljivih koje mogu, a i ne moraju biti relevantne za donošenje pretpostavki o zavisnoj promenljivoj, korisno je imati mogucnost redukcije modela, tako da u njemu ostanu samo promenljive koje nam obezbeduju važne informacije o zavisnoj promenljivoj. Nije uvek trivijalno odluciti koju promenljivu treba ostaviti u modelu. ***Maksimalni model*** definišemo kao model koji sadrži sve nezavisne promenljive koje mogu biti prisutne u modelu. Neka k predstavlja maksimalni broj nezavisnih promenljivih, tada maksimalni model ima oblik:

$$Y_i = b_0 + b_1 x_{i1} + ... + b_k x_{ik} + e_i$$

gde su $x_{i1}, ..., x_{ik}$ nezavisne promenljive, a e_i su greške za koje važi da su nezavisne, da imaju normalnu raspodelu sa nula ocekivanjem i zajednickom varijansom.

Kada se definiše maksimalni model, važno je da on sadrži sve nezavisne promenljive koje mogu uticati na zavisnu promenljivu, ali se mora paziti da u model ne ude previše nezavisnih promenljivih koje nemaju znacaja. Ukoliko model sadrži previše nezavisnih promenljivih u poredenju sa brojem opservacija, standardne greške u ocenjenim parametrima regresije mogu da budu izuzetno velike, što dovodi do nepreciznih rezultata. Takode, što je veci broj nezavisnih promenljivih to je veci rizik da dode do medusobne korelacije izmedu promenljivih. U opštem slucaju, treba da se

uzme u obzir velicina uzorka, što je manji uzorak, to treba da bude manji i maksimalni model. Postoje razna pravila o tome kakav ovaj odnos treba da bude. Tako na primer jedno od takvih pravila kaže da bi trebalo da bude najmanje pet opservacija na jednu nezavisnu promenljivu, tj. $n \geq 5k$.

Kada je definisan maksimalan model, sledeci korak je uporediti dva modela i odrediti koji je od njih bolji. U tu svrhu koristimo ***selekcione kriterijume***, ciji je zadatak da porede maksimalni model

$$Y = b_0 + b_1 x_1 + ... + b_m x_m + b_{m+1} x_{m+1} + ... + b_k x_k + e$$

sa redukovanim modelom

$$Y = b_0 + b_1 x_1 + ... + b_m x_m$$

koji se dobija od maksimalnog modela. Cilj je videti da li redukovani model odgovara podacima podjednako dobro kao maksimalni model, i u tom slucaju cemo se odluciti da koristimo redukovani model umesto maksimalnog. Sada cemo navesti par selekcionih kriterijuma.

- R_a^2 ***kriterijum***. Koeficijent determinacije R^2 odreduje koliko dobro model odgovara podacima i on se racuna na sledeci nacin:

$$R^2 = \frac{S_{yy} - RSS}{S_{yy}}$$

gde je $S_{yy} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$, $RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ je suma kvadrata reziduala. Ako sa R_k^2 oznacimo sumu kvadrata reziduala maksimalnog modela, a sa R_m^2 sumu kvadrata reziduala redukovanog modela, tada je

$$R_j^{\ 2} = \frac{S_{yy} - RSS_j}{S_{yy}}$$

gde je

$$RSS_j = \sum_{i=1}^{n} (Y_i - \hat{b}_{j0} - \hat{b}_{j1} x_{i1} - \hat{b}_{j2} x_{i2} - ... - \hat{b}_{jj} x_{ij})^2 \qquad (*)$$

gde $\hat{b}_{ji}$ oznacava ocenu za parametar b_{ji} koja se dobija primenom metode najmanjih kvadrata. Što model bolje odgovara podacima, to je veci R^2. Dakle, jedna od mogucnosti za poredenje dva modela je da se uporede odgovarajuci koeficijenti determinacije, pri cemu je model sa vecim koeficijentom determinacije precizniji. Medutim, kao problem sa javlja cinjenica da model sa vecim brojem nezavisnih promenljivih ima veci ovaj koeficijent, nezavisno od toga kakav uticaj te nezavisne promenljive imaju na zavisnu promenljivu. Da bi se ovaj problem izbegao uvodi se pojam *prilagodenog koeficijenta determinacije* R_a^2 koji se dobija na sledeci nacin:

$$R_a^2 = 1 - \frac{n-1}{n-k-1}(1 - R^2)$$

Kao što možemo da vidimo, sa porastom broja promenljivih, ne mora da znaci da ce i R_a^2 porasti. Prema R_a^2 kriterijumu zakljucujemo da *treba da se izabere model sa najvecim* R_a^2.

- ***F-test kriterijum.*** Ideja ovog kriterijuma je testirati znacajnost *k-m* nezavisnih promenljivih $x_{m+1},...,x_k$ u maksimalnom modelu, sa ciljem da se dobije redukovani model. Dakle, treba da testiramo nultu hipotezu:

$$H_0 : b_{m+1} = ... = b_k = 0$$

F-test statistika je data sa:

$$F_m = \frac{(RSS_m - RSS_k)/(k-m)}{RSS_k/(n-k-1)}$$

gde su RSS_m i RSS_k definisane sa (*). Ako se hipoteza H_0 prihvati, redukovani model odgovara podacima podjednako dobro kao maksimalni, pa samim tim možemo da koristimo redukovani umesto maksimalnog modela. *F-test kriterijum za selekovanje promenljivih nalazi najmanji podskup nezavisnih promenljivih* $x_1,x_2,..,x_m$ *takvih da test statistika* F_m *nije znacajna.*

Kada smo se upoznali sa selekcionim kriterijumima, možemo da navedemo i ***selekcione metode*** koje koriste ove se kriterijume da bi odredile da li se u modelu nalazi optimalan broj promenljivih.

Jedna ovakva metoda je tzv. ***eliminacija unazad.*** Ova metoda je u suštini niz testova znacajnosti nezavisnih promenljivih. Dakle, zakljucujemo da ona koristi *F*-test kriterijum. Eliminacija unazad pocinje sa maksimalnim modelom i u svakom koraku eliminiše promenljivu sa najvecom *p*-vrednosti, pri cemu je unapred odreden nivo znacajnosti. Ova metoda se završava onda kada nemamo mogucnost više nijednu promenljivu da iskljucimo iz modela.

Druga metoda je ***metoda unapred***, koja na kraju treba da da iste rezultate kao i eliminacija unazad. Ova metoda pocinje sa praznim modelom i u svakom koraku dodaje promenljivu koja je najznacajnija i tako sve dok se dode do situacije da se više nema koristi od dodavanja novih promenljivih.

Kombinacija ove dve metode je predstavljena u ***metodi po etapama***. U prvom koraku ove metode se dodaje promenljiva promenljiva sa najvecom vrednošcu koeficijenta determinacije R^2. Koeficijenti determinacije ostalih promenljivih se ispituju da se vidi da li one obezbeduju neke dodatne informacije za model. Kada je dodata nova promenljiva u model, pomocu *F*-testa se proverava znacajnost promenljivih koje se u tom trenutku nalaze u modelu. One promenljive koje ulaskom nove promenljive nisu više znacajne, izlaze iz modela. Može se desiti da promenljiva koja ušla u model u prethodnom koraku, u nekom od sledecih koraka izgubi znacaj, i u tom slucaju ona izlazi iz modela. Ovaj proces se nastavlja sve dok se ne dode do toga da nema više koristi dodavati nove promenljive.

3.3.2 Metod maksimalne verovatnoce

Kada imamo više od jednog opažanja za promenljvu $\boldsymbol{x}_i$, tada nam je bitno da znamo broj opažanja n_i i broj uspeha koje oznacavamo sa y_i. Tada su $\{Y_1,\ldots.Y_N\}$ nezavisne binarne promenljive sa ocekivanjem $E(Y_i) = n_i p(\boldsymbol{x}_i)$, gde je $n_1 +....+ n_N = n$. Tada je njihova zajednicka funkcija verovatnoce jednaka:

$$L(b) = \prod_{i=1}^{N} p(x_i)^{y_i} (1 - p(x_i))^{n_i - y_i}$$

Primenom metode maksimalne verovatnoce dobijamo ocene za β. Kao što sama rec kaže, traži se maksimalna vrednost ove funkcije, koja se dobija kada prvi izvod izjednacimo sa nulom. Dobijamo sistem nelinearnih jednacina po β_k koji treba da rešimo:

$$\sum_{i=1}^{N} y_i x_{ik} - \sum_{i=1}^{N} n_i \hat{p}_i x_{ik} = 0\,,\ k = 1,..., p\ (*)$$

gde je

$$\hat{p}_i = \frac{\exp(\sum_{k=1}^{p} \hat{b}_k x_{ik})}{1 + \exp(\sum_{k=1}^{p} \hat{b}_k x_{ik})}$$

ocena za $\pi(\boldsymbol{x}_i)$. Za rešavanje ovog sistema koristi se neki iterativni postupak, najcešce Newton-Raphson algoritam.

Kada ovaj sistem rešimo po β_k, svako ovakvo rešenje, ako postoji, odreduje jednu kriticnu tacku, ili minimum ili maksimum. Kriticna tacka ce biti maksimum ako je matrica drugih parcijalnih izvoda negativno definitna, tj. ako je svaki element na dijagonali matrice manji od nule. Ova matrica ce imati elemente sledeceg oblika:

$$-\sum_{i=1}^{n} x_{ik} x_{ia} n_i p_i (1 - p_i)$$

Matrica $-\frac{\partial^2 L(b)}{\partial b_a \partial b_b}$ se naziva *Fisher-ova matrica podataka*.

Matrica ocenjenih kovarijansi ima sledeci oblik:

$$\hat{\text{cov}}(\hat{b}) = (\mathbf{X}' diag(n_i \hat{p}_i (1 - \hat{p}_i))\mathbf{X})^{-1}$$

gde $diag(n_i \hat{p}_i (1 - \hat{p}_i))$ predstavlja N x N dijagonalnu matricu sa elementima $(n_i \hat{p}_i (1 - \hat{p}_i))$ na glavnoj dijagonali. Kvadratni koreni elemenata sa glavne dijagonale matrice kovarijansi su *ocenjene standardne greške koeficijenata* $\hat{b}$.

Rezultati logit regresije mogu da se interpretiraju na više nacina. Kao prvo, ako posmatramo jednacinu u obliku

$$\ln(\frac{\hat{p}}{1 - \hat{p}}) = b_0 + b_1 x_1 + ... + b_n x_n$$

tada se sa jedinicnim porastom (smanjenjem) nezavisne promenljive x_i logaritam odnosa verovatnoca povecava (smanjuje) za β_i. Naravno, u slucaju kada su sve nezavisne promenljive jednake nuli, ovaj odnos ce biti jednak konstanti β_0. Medutim, ova jednacina može da se transformiše eksponenciranjem:

$$\frac{\hat{p}}{1-\hat{p}} = \exp(b_0 + b_1 x_1 + ... + b_n x_n) = \exp(b_0) * \exp(b_1 x_1) * ... * \exp(b_n x_n)$$

u ovom slucaju ako x_i poraste za jednu jedinicu, ocenjeni odnos ce porasti exp(β_i) puta. Ukoliko su sve nezavisne promenljive jednake nuli, tada ce ovaj odnos biti jednak exp(β_0). A ako posmatramo sledecu jednacinu:

$$\hat{p} = \frac{\exp(\sum_{k=1}^{p} \hat{b}_k x_{ik})}{1 + \exp(\sum_{k=1}^{p} \hat{b}_k x_{ik})}$$

nije lako protumaciti rezultate logit regresije.

3.3.3 Newton-Raphson algoritam (NRA)

Newton-Raphson algoritam je iterativni postupak za rešavanje nelinearnih jednacina. Sada cemo pokazati kako NRA odreduje vrednost $\hat{b}$ kada se maksimizira funkcija L(b). Neka je $\mathbf{u}' = (\partial L(b)/\partial b_1, ..., \partial L(b)/\partial b_p)$, a sa **H** oznacavamo Hesijan matricu koja ima sledece elemente $h_{ab} = \partial^2 L(b)/\partial b_a \partial b_b$. Neka je sa $\mathbf{u}^{(t)}$ i $\mathbf{H}^{(t)}$ oznaceno **u** i **H** u $b^{(t)}$, što predstavlja *t*-ti pokušaj za $\hat{b}$. Korak *t* iterativnog procesa (*t* = 0,1,2,...) koji aproksimira L(b) u blizini $b^{(t)}$ Tejlorovim polinomom drugog reda je:

$$L(b) \approx L(b^{(t)}) + \mathbf{u}^{(t)'}(b - b^{(t)}) + \frac{1}{2}(b - b^{(t)})'\mathbf{H}^{(t)}(b - b^{(t)})$$

Rešavajuci $\partial L(b)/\partial b \approx \mathbf{u}^{(t)} + \mathbf{H}^{(t)}(b - b^{(t)}) = 0$ po b dolazimo do sledece tacke, što može da se zapiše na sledeci nacin:

$$b^{(t+1)} = b^{(t)} - (\mathbf{H}^{(t)})^{-1}\mathbf{u}^{(t)} \quad (*)$$

i pretpostavljamo da $\mathbf{H}^{(t)}$ nije singularna matrica.

U slucaju logaritamske regresije imamo sledeci slucaj:

$$u_j^{(t)} = \frac{\partial L(b)}{\partial b_j}|_{b(t)} = \sum_i (y_i - n_i p_i^{(t)}) x_{ij}$$

$$h_{ab}^{(t)} = \frac{\partial^2 L(b)}{\partial b_a \partial b_b}|_{b(t)} = -\sum_i x_{ia} x_{ib} n_i p_i^{(t)} (1 - p_i^{(t)})$$

gde je $p^{(t)}$ *t*-ta aproksimacija za $\hat{p}$ dobijena uz pomoc $b^{(t)}$

$$p_i^{(t)} = \frac{\exp(\sum_{k=1}^{p} b_k^{(t)} x_{ik})}{1+\exp(\sum_{k=1}^{p} b_k^{(t)} x_{ik})}$$

Koristimo jednacinu (*) da bi dobili sledecu vrednost $b^{(t+1)}$:

$$b^{(t+1)} = b^{(t)} + (X' diag(n_i p_i^{(t)}(1-p_i^{(t)}))X)^{-1} X'(y - m^{(t)})$$

gde je $m_i^{(t)} = n_i p_i^{(t)}$. Ova vrednost se dalje koristi da se dobije $p^{(t+1)}$ i tako dalje. Vrednosti $p^{(t)}$ i $b^{(t)}$ konvergiraju ka ocenama koje se dobijaju uz pomoc funkcije maksimalne verovatnoce $\hat{p}$ i $\hat{b}$. A Hesijan $\mathbf{H}^{(t)}$ konvergira ka $\hat{H} = -X' diag(n_i \hat{p}_i (1-\hat{p}_i))X$.

3.3.4 Testiranje hipoteza

Za testitiranje hipoteza koristi se ***Waldov test*** koji se najcešce koristi da pokaže da li efekat postoji ili ne, tj. on pokazuje da li nezavisna promenljiva ima statisticki znacajan odnos sa zavisnom promenljivom. Waldov test poredi ocene $\hat{b}$ parametara od znacaja β sa predloženim vrednostima u nultoj hipotezi β_0, pod pretpostavkom da ce se razlika ove dve vrednosti aproksimirati normalnom raspodelom, pa se kvadrat ove razlike aproksimira χ^2 – raspodelom. Dakle, Waldova statistika ima sledeci oblik:

$$\frac{(\hat{b}_j - b_{j0})^2}{\mathrm{var}(\hat{b})} \sim c_1^{\,2}$$

ili $$\frac{\hat{b}_j - b_{j0}}{\hat{se}(\hat{b})} \sim N(0,1)$$

gde se $\hat{se}(\hat{b})$ dobija tako što se uzme inverzni element iz ocenjene Fisher-ove matrice podataka. Ukoliko testiramo hipotezu za više parametara u isto vreme, pod pretpostavkom da je H_0: $\beta = \beta_0$, Waldova statistika ima sledeci oblik:

$$W = (\hat{b} - b_0)' \left[\mathrm{cov}(\hat{b})\right]^{-1} (\hat{b} - b_0)$$

koja prati χ^2 raspodelu, a broj stepena slobode je jednak rangu matrice kovarijansi.

Drugi test koji se koristi za testiranje parametara je ***test odnosa verovatnoca (likelihood-ratio test).*** Obicno se obeležava sa grckim slovom Λ (velikim lambda) i predstavlja odnos dve maksimalne vrednosti: prva je maksimum dobijen od parametara pod pretpostavkom nulte hipoteze koji se oznacava sa l_0 i opšteg maksimuma, tj. maksimuma dobijenog pod pretpostavkom $H_0 \bigcup H_A$ i taj maksimum oznacavamo sa l_1. Odnos $\Lambda = l_0 / l_1$ ne može da bude veci od jedan. Wilks je pokazao da $-2\log\Lambda$ prati χ^2 raspodelu gde je broj stepeni slobode jednak razlici dimenzija parametara pod pretpostavkama $H_0 \bigcup H_A$ i H_0. Dakle,

$$- 2\log\Lambda = - 2\log(l_0 / l_1) = - 2 (L_0 - L_1)$$

Treci nacin za testiranje hipoteza je ***skor test***. To je najmocniji test kada je prava vrednost parametra blizu ocenjene vrednosti. Neka je *L* funkcija verovatnoce koja zavisi od parametra β, tada je skor funkcija:

$$u(b) = \frac{\partial L(b)}{\partial b}$$

Fisher-ova matrica podataka je:

$$I(b) = -\frac{\partial^2 L(b)}{\partial b^2}$$

Tada je statistika koja testira hipotezu H_0: $\beta = \beta_0$ data sa:

$$S(b) = \frac{u(b_0)^2}{I(b_0)} \sim c_1^2$$

3.3.5 Konstruisanje intervala poverenja

Neka z_a predstavlja *z*-vrednost standardizovane normalne raspodele koja ima verovatnocu *a*, tj. to je 100(1-*a*)% rapodele. Neka $\chi^2_{df}(a)$ predstavlja 100(1-*a*)% χ^2 – raspodele sa *df* stepeni slobode. Tada je interval poverenja za Waldov test skup β_0 za koje je:

$$\frac{\left|\hat{b} - b_0\right|}{SE} < z_{a/2}$$

Što daje sledeci interval: $\hat{b} \pm z_{a/2} SE$.

Za test odnosa verovatnoca interval poverenja je skup β_0 za koje je

$$-2\left[L(b_0) - L(\hat{b})\right] < c_1^2(a)$$

Podsetimo se da je $c_1^2(a) = z_{a/2}^2$.

Dok je za skor test interval poverenja skup β_0 za koje je

$$\left|\frac{u(b_0)^2}{I(b_0)}\right| < c_1^2(a)$$

3.3.6 Primena softvera

Uz pomoc statistickog paket *Statgraphics* cemo prikazati jedan jednostavan primer logit regresije. Pretpostavimo da imamo zavisnu promenljivu *Loš* koja uzima vrednost 1 ukoliko je klijent loš i vrednost 0 ako je klijent dobar, i pet nezavisnih promenljivih:

- *Starost*: koja pokazuje koliko klijent ima godina
- *Muško*: binarna promenljiva koja uzima vrednost 1 ukoliko je klijent muško
- *Prihodi*: koja predstavlja klijentove prihode
- *Oženjen*: binarna promenljiva koja uzima vrednost 1 ukoliko je klijent oženjen
- *Vlasnik stana*: binarna promenljiva koja uzima vrednost 1 ukoliko je klijent vlasnik stana u kome živi

	Los	Starost	Musko	Prihodi	Ozenjen	Vlasnik stana
1	1	25	1	500	0	1
2	0	40	1	1000	1	0
3	1	45	1	1500	1	0
4	0	21	0	300	0	0
5	0	25	1	300	0	0
6	1	35	0	550	1	1
7	0	30	0	600	0	0
8	1	40	1	1200	1	1
9	1	42	0	1200	1	0
10	1	37	1	700	0	1
11	0	51	0	1500	1	1
12	1	22	1	600	0	1
13	0	21	1	300	0	0
14	0	29	1	900	0	0
15	1	33	1	1000	1	1

Promenljive koje ulaze u model

Kada smo uneli ove podatke u radni list, možemo da primenimo logit regresiju i da vidimo kakve cemo rezultate dobiti. Prvo dobijamo prozor u kome se nalazi tabela sa ocenjenim parametrima regresije uz pomoc funkcije maksimalne verovatnoce, njihovim standardnim greškama i ocenjenim «*odds*» odnosom koji se dobija na sledeci nacin:

$$odds\_ratio = \exp(\hat{b}_i)$$

i on predstavlja procenat u kome se povecava $odds = \frac{p}{1-p}$ ukoliko dode do jedinicnog poveca promenljive X. Takode dobijemo i tabelu sa rezultatima testa znacajnosti promenljivih. Kao što možemo da primetimo, ovaj paket koristi test odnosa verovatnoca da testira znacajnost promenljivih u modelu. Na kraju imamo kratak opis rezultata, kao i regresionu jednacinu.

Logistic Regression - Los

Dependent variable: Los
Factors:
- Musko
- Ozenjen
- Prihodi
- Starost
- Vlasnik stana

Estimated Regression Model (Maximum Likelihood)

		Standard	*Estimated*
Parameter	*Estimate*	*Error*	*Odds Ratio*
CONSTANT	0.913344	5.13631	
Musko	1.37705	1.86523	3.96318
Ozenjen	3.32295	2.91391	27.742
Prihodi	0.00275722	0.00472383	1.00276
Starost	-0.20772	0.254429	0.812435
Vlasnik stana	3.53999	1.94519	34.4667

Analysis of Deviance

Source	*Deviance*	*Df*	*P-Value*
Model	8.85067	5	0.1152
Residual	11.877	9	0.2203
Total (corr.)	20.7277	14	

Percentage of deviance explained by model = 42.6997
Adjusted percentage = 0.0

Likelihood Ratio Tests

Factor	*Chi-Squared*	*Df*	*P-Value*
Musko	0.614137	1	0.4332
Ozenjen	1.58047	1	0.2087
Prihodi	0.357043	1	0.5502
Starost	0.797817	1	0.3717
Vlasnik stana	5.79497	1	0.0161

The StatAdvisor

The output shows the results of fitting a logistic regression model to describe the relationship between Los and 5 independent variable(s). The equation of the fitted model is

Los = exp(eta)/(1+exp(eta))

where

eta = 0.913344 + 1.37705*Musko + 3.32295*Ozenjen + 0.00275722*Prihodi - 0.20772*Starost + 3.53999*Vlasnik stana

Because the P-value for the model in the Analysis of Deviance table is greater or equal to 0.05, there is not a statistically significant relationship between the variables at the 95.0% or higher confidence level. In addition, the P-value for the residuals is greater than or equal to 0.05, indicating that the model is not significantly worse than the best possible model for this data at the 95.0% or higher confidence level.

The pane also shows that the percentage of deviance in Los explained by the model equals 42.6997%. This statistic is similar to the usual R-Squared statistic. The adjusted percentage, which is more suitable for comparing models with different numbers of independent variables, is 0.0%.

In determining whether the model can be simplified, notice that the highest P-value for the likelihood ratio tests is 0.5502, belonging to Prihodi. Because the P-value is greater or equal to 0.05, that term is not statistically significant at the 95.0% or higher confidence level. Consequently, you should consider removing Prihodi from the model.

Statgrahics nam daje opciju da vidimo i koliki su 95% intervali poverenja za ocenjene parametre.

95.0% confidence intervals for coefficient estimates

		Standard		
Parameter	*Estimate*	*Error*	*Lower Limit*	*Upper Limit*
CONSTANT	0.913344	5.13631	-10.7058	12.5325
Musko	1.37705	1.86523	-2.8424	5.5965
Ozenjen	3.32295	2.91391	-3.2688	9.91469
Prihodi	0.00275722	0.00472383	-0.00792885	0.0134433
Starost	-0.20772	0.254429	-0.783279	0.36784
Vlasnik stana	3.53999	1.94519	-0.860334	7.94032

95.0% confidence intervals for odds ratios

Parameter	*Estimate*	*Lower Limit*	*Upper Limit*
Musko	3.96318	0.0582854	269.481
Ozenjen	27.742	0.0380521	20225.4
Prihodi	1.00276	0.992102	1.01353
Starost	0.812435	0.456905	1.44461
Vlasnik stana	34.4667	0.423021	2808.26

The StatAdvisor

This table shows 95.0% confidence intervals for the coefficients in the model. Confidence intervals show how precisely the coefficients can be estimated given the amount of available data and the noise which is present. Also shown are confidence intervals for the odds ratios. The odds ratio equals the inverse natural logarithm of the coefficient and shows the proportional change in the response variable as the independent variable is increased by 1 unit.

Da bi videli da li postoji odredena medusobna korelacija izmedu promenljivih možemo da koristimo opciju za dobijanje korelacione matrice:

Correlation matrix for coefficient estimates

	CONSTANT	Musko	Ozenjen	Prihodi
CONSTANT	1.0000	-0.3334	0.4237	0.5799
Musko	-0.3334	1.0000	0.3533	-0.1267
Ozenjen	0.4237	0.3533	1.0000	0.0392
Prihodi	0.5799	-0.1267	0.0392	1.0000
Starost	-0.8735	0.0267	-0.5098	-0.8154
Vlasnik stana	0.3512	0.2026	0.4478	0.4347

	Starost	Vlasnik stana
CONSTANT	-0.8735	0.3512
Musko	0.0267	0.2026
Ozenjen	-0.5098	0.4478
Prihodi	-0.8154	0.4347
Starost	1.0000	-0.5831
Vlasnik stana	-0.5831	1.0000

The StatAdvisor
This table shows estimated correlations between the coefficients in the fitted model. These correlations can be used to detect the presence of serious multicollinearity, i.e., correlation amongst the predictor variables. In this case, there are 3 correlations with absolute values greater than 0.5.

Takode, imamo mogucnost primene ranije opisanih selekcionih metoda i to metode unapred i metode eliminacije unazad i za naš primer dobijamo sledece rezultate. Za metod unapred:

Logistic Regression - Los

Estimated Regression Model (Maximum Likelihood)

		Standard	*Estimated*
Parameter	*Estimate*	*Error*	*Odds Ratio*
CONSTANT	-1.09861	0.816497	
Vlasnik stana	2.89037	1.35398	18.0

Analysis of Deviance

Source	*Deviance*	*Df*	*P-Value*
Model	5.98871	1	0.0144
Residual	14.739	13	0.3239
Total (corr.)	20.7277	14	

Percentage of deviance explained by model = 28.8923
Adjusted percentage = 9.59445

Likelihood Ratio Tests

Factor	*Chi-Squared*	*Df*	*P-Value*
Vlasnik stana	5.98871	1	0.0144

Stepwise factor selection
Method: forward selection
P-to-enter: 0.05
P-to-remove: 0.05
Step 0:
0 factors in the model. 14 d.f. for error.
Percentage of deviance explained = 0.00% Adjusted percentage = 0.00%
Step 1:
Adding factor Vlasnik stana with P-to-enter = 0.0143954
1 factors in the model. 13 d.f. for error.
Percentage of deviance explained = 28.89% Adjusted percentage = 9.59%

Final model selected.

The StatAdvisor
The output shows the results of fitting a logistic regression model to describe the relationship between Los and 5 independent variable(s). The equation of the fitted model is

Los = exp(eta)/(1+exp(eta))

where

eta = -1.09861 + 2.89037*Vlasnik stana

Because the P-value for the model in the Analysis of Deviance table is less than 0.05, there is a statistically significant relationship between the variables at the 95.0% confidence level. In addition, the P-value for the residuals is greater than or equal to 0.05, indicating that the model is not significantly worse than the best possible model for this data at the 95.0% or higher confidence level.

The pane also shows that the percentage of deviance in Los explained by the model equals 28.8923%. This statistic is similar to the usual R-Squared statistic. The adjusted percentage, which is more suitable for comparing models with different numbers of independent variables, is 9.59445%.

In determining whether the model can be simplified, notice that the highest P-value for the likelihood ratio tests is 0.0144, belonging to Vlasnik stana. Because the P-value is less than 0.05, that term is statistically significant at the 95.0% confidence level. Consequently, you probably don't want to remove any variables from the model.

A za metod eliminacije unazad:

Logistic Regression - Los

Estimated Regression Model (Maximum Likelihood)

		Standard	*Estimated*
Parameter	*Estimate*	*Error*	*Odds Ratio*
CONSTANT	-1.09861	0.816497	
Vlasnik stana	2.89037	1.35398	18.0

Analysis of Deviance

Source	*Deviance*	*Df*	*P-Value*
Model	5.98871	1	0.0144
Residual	14.739	13	0.3239
Total (corr.)	20.7277	14	

Percentage of deviance explained by model = 28.8923
Adjusted percentage = 9.59445

Likelihood Ratio Tests

Factor	*Chi-Squared*	*Df*	*P-Value*
Vlasnik stana	5.98871	1	0.0144

Stepwise factor selection
Method: backward selection
P-to-enter: 0.05
P-to-remove: 0.05

Step 0:
5 factors in the model. 9 d.f. for error.
Percentage of deviance explained = 42.70% Adjusted percentage = 0.00%

Step 1:
Removing factor Prihodi with P-to-remove = 0.550152
4 factors in the model. 10 d.f. for error.
Percentage of deviance explained = 40.98% Adjusted percentage = 0.00%

Step 2:
Removing factor Starost with P-to-remove = 0.496249
3 factors in the model. 11 d.f. for error.
Percentage of deviance explained = 38.74% Adjusted percentage = 0.15%

Step 3:
Removing factor Musko with P-to-remove = 0.355254
2 factors in the model. 12 d.f. for error.
Percentage of deviance explained = 34.62% Adjusted percentage = 5.67%

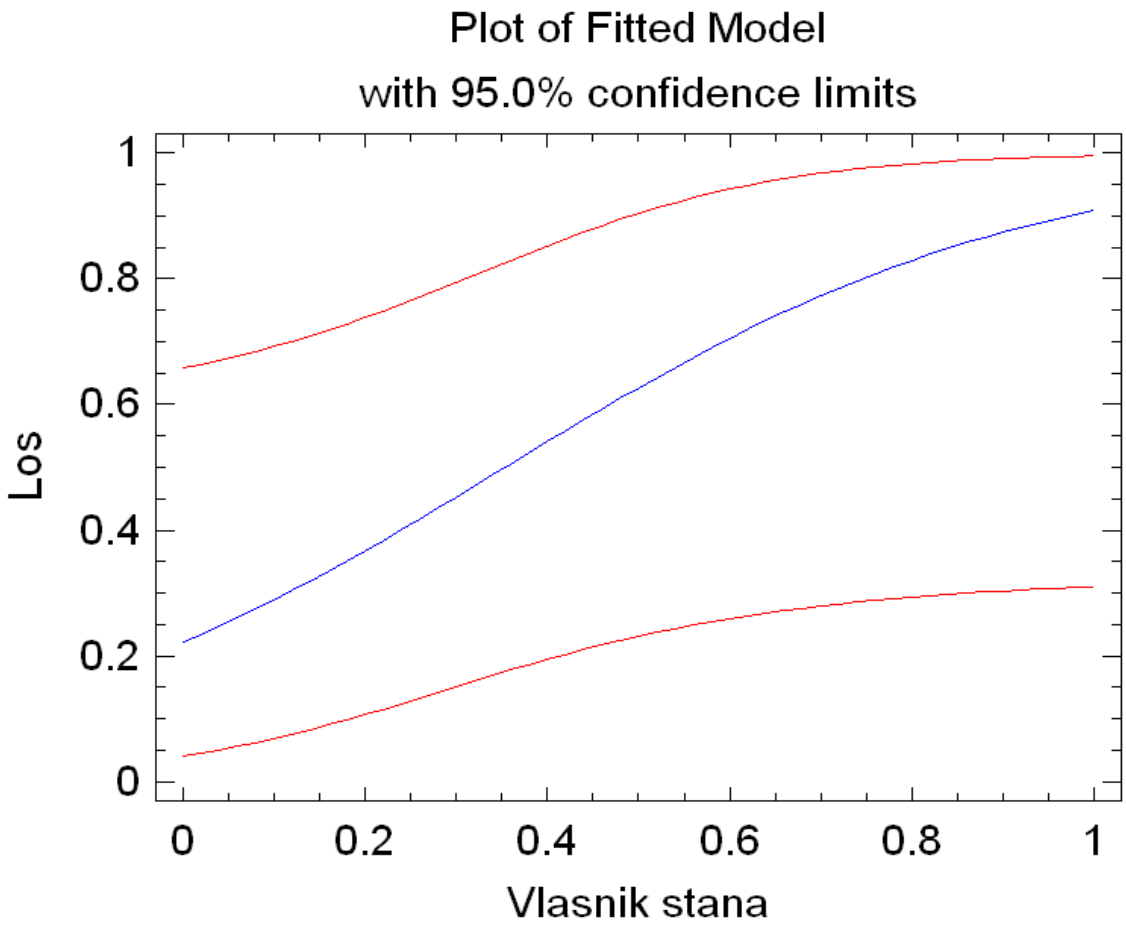
Step 4:
Removing factor Ozenjen with P-to-remove = 0.275858
1 factors in the model. 13 d.f. for error.
Percentage of deviance explained = 28.89% Adjusted percentage = 9.59%

Final model selected.

The StatAdvisor

The output shows the results of fitting a logistic regression model to describe the relationship between Los and 5 independent variable(s). The equation of the fitted model is
Los = exp(eta)/(1+exp(eta))
where
eta = -1.09861 + 2.89037*Vlasnik stana
Because the P-value for the model in the Analysis of Deviance table is less than 0.05, there is a statistically significant relationship between the variables at the 95.0% confidence level. In addition, the P-value for the residuals is greater than or equal to 0.05, indicating that the model is not significantly worse than the best possible model for this data at the 95.0% or higher confidence level.
The pane also shows that the percentage of deviance in Los explained by the model equals 28.8923%. This statistic is similar to the usual R-Squared statistic. The adjusted percentage, which is more suitable for comparing models with different numbers of independent variables, is 9.59445%.
In determining whether the model can be simplified, notice that the highest P-value for the likelihood ratio tests is 0.0144, belonging to Vlasnik stana. Because the P-value is less than 0.05, that term is statistically significant at the 95.0% confidence level. Consequently, you probably don't want to remove any variables from the model.

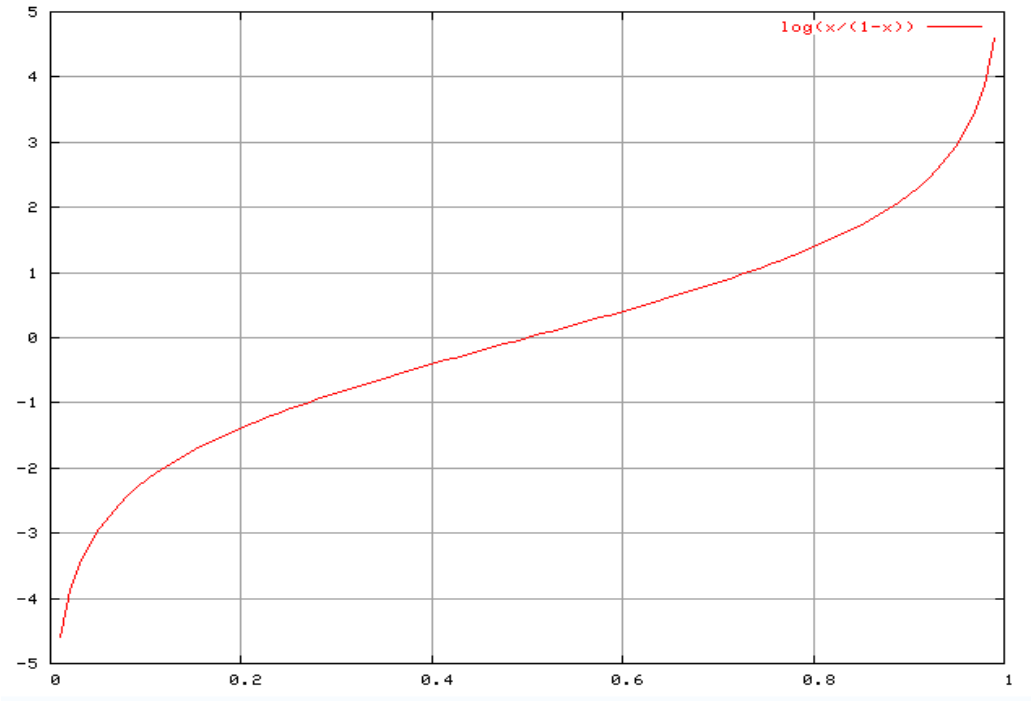
Statgraphics nam omogucava da graficki prikažemo ocenjenu verovatnocu da je klijent loš u odnosu na pojedinacnu nezavisnu promenljivu, dok ostale nezavine promenljive ostaju konstantne. Mi necemo dati grafike za sve nezavisne promenljive, nego samo za jednu, na primer *Vlasnik stana*:



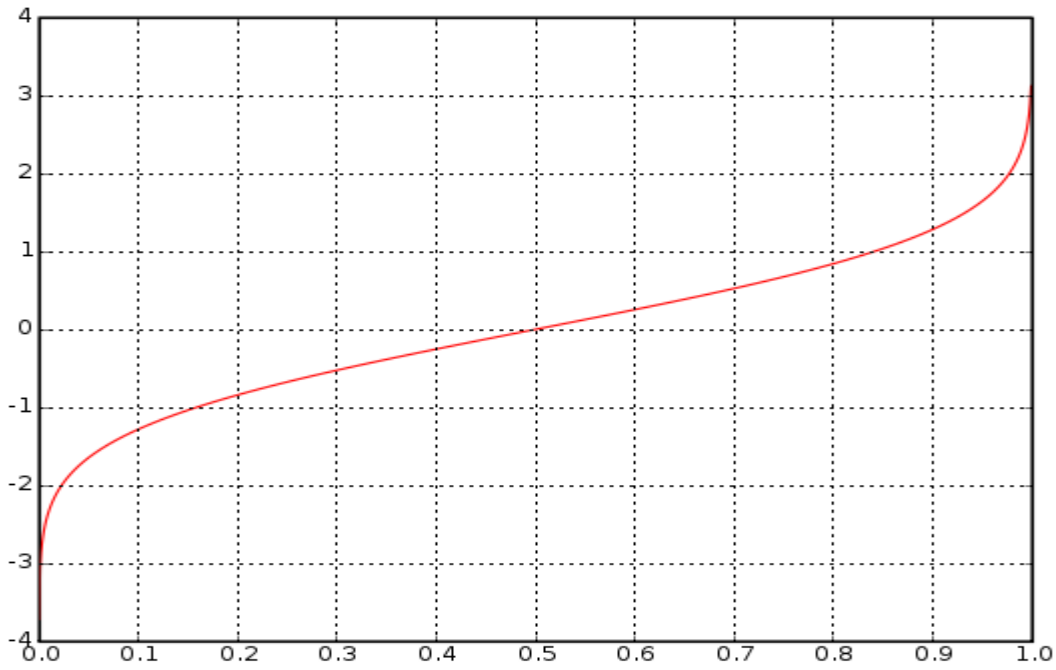
Paket *Statgrahics* je vrlo koristan softver koji nam pored ovih, gore navedenih, omogucava još mnoštvo drugih analiza na podacima koje nam mogu biti od koristi.

3.4 *Probit analiza*

Probit analiza je alternativa logit analizi. Logit i probit analizu su jako slicne, samo što logit analiza koristi logaritam proporcije verovatnoca, dok probit koristi kumulativnu normalnu raspodelu. Grafik kumulativne normalne raspodele ima oblik slova S i krece se u granicama od 0 do 1, što je vrlo slicno grafiku logit funkcije. Dakle, zakljucujemo da ove dve metode daju vrlo slicne rezultate.



Logit kriva



Probit kriva

Probit funkcija je inverzna funkcija raspodele koja je povezana sa standardnom normalnom raspodelom. ***Probit model*** je specijalni oblik generalizovanog linearnog modela, i uglavnom se koristi u regresiji. Neka je Y zavisna promenljiva, a X vektor nezavisnih promenljivih, tada probit model kaže da je:

$$P(Y = 1 \mid X = x) = \Phi(x'b)$$

gde je Φ funkcija raspodele standardne normalne raspodele. Parametri β se ocenjuju pomocu funkcije maksimalne verovatnoce. *Višedimenzionalni probit model* je generalizacija probit modela koji se koristi kada treba zajedno da se oceni nekoliko povezanih binarnih promenljivih. Neka su Y_1 i Y_2 binarne zavisne promenljive tako da je

$$Y_1 = 1(Y_1^* > 0)$$
$$Y_2 = 1(Y_2^* > 0)$$

gde je

$$Y_1^* = Xb_1 + e_1$$
$$Y_2^* = Xb_2 + e_2$$

gde je X vektor nezavisnih promenljivih i pretpostavlja se da one imaju normalnu raspodelu sa nula ocekivanjem, jedinicnom varijansom i koeficijentom korelacije ρ.

Logaritamska funkcija verovatoce za probit model ima sledeci oblik:

$$\ln l = \sum_{i=1}^{N} (1 - y_{1i}) \ln(1 - f(\mathbf{x}_{1i}a_1)) + \sum_{i=1}^{N} y_{1i}(1 - y_{2i}) \ln(f(\mathbf{x}_{1i}a_1) - f_2(\mathbf{x}_{1i}a_1, \mathbf{x}_{2i}a_2; r)) +$$
$$\sum_{i=1}^{N} y_{1i}y_{2i} \ln f_2(\mathbf{x}_{1i}a_1, \mathbf{x}_{2i}a_2; r)$$

gde $\phi(\bullet)$ i $\phi(\bullet, \bullet; \rho)$ predstavljaju jednodimenzionalnu i binarnu standardnu normalnu funkciju raspodele. Ova fukcija nam dalje služi za ocenu parametara β_1, β_2, ρ uz pomoc funkcije maksimalne verovatoce.

3.4.1 Primena softvera

I u slucaju probit analize, izuzetno nam je koristan paket Statgraphics. Koristicemo se primerom koji smo dali u odeljku 3.3.6. Kada na date podatke primenimo probit analizu, dobijamo sledece rezultate:

Probit Analysis - Los
Dependent variable: Los
Factors:
Musko
Ozenjen
Prihodi
Starost
Vlasnik stana

Estimated Regression Model (Maximum Likelihood)

		Standard
Parameter	*Estimate*	*Error*
CONSTANT	0.616715	2.90843
Musko	0.830676	1.06687
Ozenjen	2.10205	1.70154
Prihodi	0.00159893	0.00270175
Starost	-0.127185	0.144372
Vlasnik stana	2.12783	1.05592

Analysis of Deviance

Source	*Deviance*	*Df*	*P-Value*
Model	9.01866	5	0.1083
Residual	11.709	9	0.2302
Total (corr.)	20.7277	14	

Percentage of deviance explained by model = 43.5102
Adjusted percentage = 0.0

Likelihood Ratio Tests

Factor	*Chi-Squared*	*Df*	*P-Value*
Musko	0.706922	1	0.4005
Ozenjen	1.74547	1	0.1864
Prihodi	0.334239	1	0.5632
Starost	0.826978	1	0.3631
Vlasnik stana	5.95822	1	0.0146

The StatAdvisor
The output shows the results of fitting a probit regression model to describe the relationship between Los and 5 independent variable(s). The equation of the fitted model is
Los = normal(eta)
where
eta = 0.616715 + 0.830676*Musko + 2.10205*Ozenjen + 0.00159893*Prihodi - 0.127185*Starost + 2.12783*Vlasnik stana

Because the P-value for the model in the Analysis of Deviance table is greater or equal to 0.05, there is not a statistically significant relationship between the variables at the 95.0% or higher confidence level. In addition, the P-value for the residuals is greater than or equal to 0.05, indicating that the model is not significantly worse than the best possible model for this data at the 95.0% or higher confidence level.
The pane also shows that the percentage of deviance in Los explained by the model equals 43.5102%. This statistic is similar to the usual R-Squared statistic. The adjusted percentage, which is more suitable for comparing models with different numbers of independent variables, is 0.0%.
In determining whether the model can be simplified, notice that the highest P-value for the likelihood ratio tests is 0.5632, belonging to Prihodi. Because the P-value is greater or equal to 0.05, that term is not statistically significant at the 95.0% or higher confidence level. Consequently, you should consider removing Prihodi from the model.

Kada uporedimo ove rezultate sa rezultatima koje smo dobili putem logaritamske regresije, možemo da primetimo da su oni vrlo slicni. I korelaciona matrica daje slicne rezultate:

Correlation matrix for coefficient estimates

	CONSTANT	Musko	Ozenjen	Prihodi
CONSTANT	1.0000	-0.3495	0.4296	0.5856
Musko	-0.3495	1.0000	0.3466	-0.1661
Ozenjen	0.4296	0.3466	1.0000	0.0303
Prihodi	0.5856	-0.1661	0.0303	1.0000
Starost	-0.8801	0.0620	-0.5079	-0.8121
Vlasnik stana	0.3070	0.1652	0.3887	0.3873

	Starost	Vlasnik stana
CONSTANT	-0.8801	0.3070
Musko	0.0620	0.1652
Ozenjen	-0.5079	0.3887
Prihodi	-0.8121	0.3873
Starost	1.0000	-0.5239
Vlasnik stana	-0.5239	1.0000

The StatAdvisor
This table shows estimated correlations between the coefficients in the fitted model. These correlations can be used to detect the presence of serious multicollinearity, i.e., correlation amongst the predictor variables. In this case, there are 3 correlations with absolute values greater than 0.5.

I u slucaju probit analize možemo da primenimo selekcione metode. Pri tome, metoda unapred daje sledece rezultate:

Probit Analysis - Los

Dependent variable: Los
Factors:
Musko
Ozenjen
Prihodi
Starost
Vlasnik stana

Estimated Regression Model (Maximum Likelihood)

		Standard
Parameter	*Estimate*	*Error*
CONSTANT	-0.674486	0.482718
Vlasnik stana	1.74205	0.760325

Analysis of Deviance

Source	*Deviance*	*Df*	*P-Value*
Model	5.98871	1	0.0144
Residual	14.739	13	0.3239
Total (corr.)	20.7277	14	

Percentage of deviance explained by model = 28.8923
Adjusted percentage = 9.59445

Likelihood Ratio Tests

Factor	*Chi-Squared*	*Df*	*P-Value*
Vlasnik stana	5.98871	1	0.0144

Stepwise factor selection
Method: forward selection
P-to-enter: 0.05
P-to-remove: 0.05
Step 0:
0 factors in the model. 14 d.f. for error.
Percentage of deviance explained = 0.00% Adjusted percentage = 0.00%
Step 1:
Adding factor Vlasnik stana with P-to-enter = 0.0143954
1 factors in the model. 13 d.f. for error.
Percentage of deviance explained = 28.89% Adjusted percentage = 9.59%
Final model selected.

The StatAdvisor

The output shows the results of fitting a probit regression model to describe the relationship between Los and 5 independent variable(s). The equation of the fitted model is

Los = normal(eta)

where

eta = -0.674486 + 1.74205*Vlasnik stana

Because the P-value for the model in the Analysis of Deviance table is less than 0.05, there is a statistically significant relationship between the variables at the 95.0% confidence level. In addition, the P-value for the residuals is greater than or equal to 0.05, indicating that the model is not significantly worse than the best possible model for this data at the 95.0% or higher confidence level.
The pane also shows that the percentage of deviance in Los explained by the model equals 28.8923%. This statistic is similar to the usual R-Squared statistic. The adjusted percentage, which is more suitable for comparing models with different numbers of independent variables, is 9.59445%.
In determining whether the model can be simplified, notice that the highest P-value for the likelihood ratio tests is 0.0144, belonging to Vlasnik stana. Because the P-value is less than 0.05, that term is statistically significant at the 95.0% confidence level. Consequently, you probably don't want to remove any variables from the model.

Dok metoda eliminacije unazad u slucaju probit analize izgleda ovako:

Probit Analysis - Los

Estimated Regression Model (Maximum Likelihood)

		Standard
Parameter	*Estimate*	*Error*
CONSTANT	-0.674486	0.482718
Vlasnik stana	1.74205	0.760325

Analysis of Deviance

Source	*Deviance*	*Df*	*P-Value*
Model	5.98871	1	0.0144
Residual	14.739	13	0.3239
Total (corr.)	20.7277	14	

Percentage of deviance explained by model = 28.8923
Adjusted percentage = 9.59445

Likelihood Ratio Tests

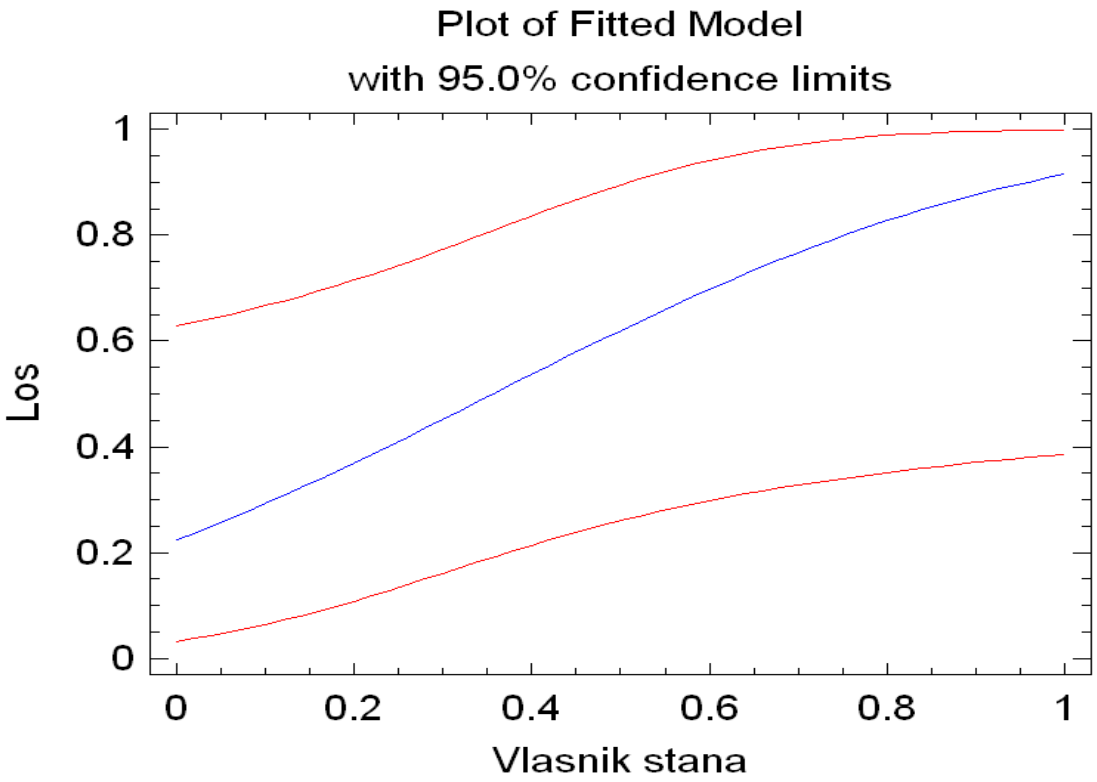
Factor	*Chi-Squared*	*Df*	*P-Value*
Vlasnik stana	5.98871	1	0.0144

Stepwise factor selection
Method: backward selection
P-to-enter: 0.05
P-to-remove: 0.05
Step 0:
5 factors in the model. 9 d.f. for error.
Percentage of deviance explained = 43.51% Adjusted percentage = 0.00%
Step 1:
Removing factor Prihodi with P-to-remove = 0.563173
4 factors in the model. 10 d.f. for error.
Percentage of deviance explained = 41.90% Adjusted percentage = 0.00%
Step 2:
Removing factor Starost with P-to-remove = 0.458766
3 factors in the model. 11 d.f. for error.
Percentage of deviance explained = 39.25% Adjusted percentage = 0.65%
Step 3:
Removing factor Musko with P-to-remove = 0.287249
2 factors in the model. 12 d.f. for error.
Percentage of deviance explained = 33.79% Adjusted percentage = 4.84%
Step 4:
Removing factor Ozenjen with P-to-remove = 0.313862
1 factors in the model. 13 d.f. for error.
Percentage of deviance explained = 28.89% Adjusted percentage = 9.59%
Final model selected.

The StatAdvisor

The output shows the results of fitting a probit regression model to describe the relationship between Los and 5 independent variable(s). The equation of the fitted model is
Los = normal(eta)
where
eta = -0.674486 + 1.74205*Vlasnik stana
Because the P-value for the model in the Analysis of Deviance table is less than 0.05, there is a statistically significant relationship between the variables at the 95.0% confidence level. In addition, the P-value for the residuals is greater than or equal to 0.05, indicating that the model is not significantly worse than the best possible model for this data at the 95.0% or higher confidence level.
The pane also shows that the percentage of deviance in Los explained by the model equals 28.8923%. This statistic is similar to the usual R-Squared statistic. The adjusted percentage, which is more suitable for comparing models with different numbers of independent variables, is 9.59445%.
In determining whether the model can be simplified, notice that the highest P-value for the likelihood ratio tests is 0.0144, belonging to Vlasnik stana. Because the P-value is less than 0.05, that term is statistically significant at the 95.0% confidence level. Consequently, you probably don't want to remove any variables from the model.

Kada graficki predstavimo, vidimo da su i grafici slicni logit regresiji, što smo vec napomenuli u teorijskom delu, a ovo je samo potvrda toj cinjenici:



3.5 Tobit model

Tobit model je tip regresije sa ogranicenjem koji odreduje odnos zavisne y_i i nezavisnih promenljivih x_i. Tobit model ima sledeci oblik:

$$y_i = \begin{cases} y_i^*, y_i^* > 0 \\ 0, y_i^* \le 0 \end{cases}$$

gde je y_i^* neopažena promenljiva, tako da je:

$$y_i^* = bx_i + u_i,\ u_i \sim N(0, \sigma^2)$$

Cesta modifikacija Tobit modela je da se odredi prag y_L koji je razlicit od nule, i tada posmatramo sledeci slucaj:

$$y_i = \begin{cases} y_i^*, y_i^* > y_L \\ y_L, y_i^* \le y_L \end{cases}$$

Logaritamska funkcija verovatoce za tobit model ima sledeci oblik:

$$\ln L = \sum_{i=1}^{N} (d_i(-\ln s + \ln f(\frac{y_i - bx_i}{s})) + (1 - d_i)\ln(1 - f(\frac{bx_i}{s})))$$

gde je ϕ funkcija normalne raspodele, a d_i je indikatorska promenljiva:

$$d_i = \begin{cases} 0, y_i^* \le y_L \\ 1, y_i^* > y_L \end{cases}$$

Uz pomoc funkcije maksimalne verovatnoce ocenjujemo parametre β.

4. Modeli kreditnog skoringa sistema

4.1 Primena problema separacije

Sada cemo na primeru videti kako se problem separacije primenjuje na kreditni skoring sistem. Koristimo se primerom datim u radu «*Credit Cards Scoring with Quadratic Utility Function*» [6]. Neka je zahtev za kreditnu karticu objekat, a na primer stavka «bracno stanje» koje se popunjava u zahtevu za kredit je indikatorska promenljiva. Pretpostavimo da imamo skup od *m* objekata i neka je svaki objekat predstavljen vektorom $x^i = (x_1^i,....,x_n^i)$, i = 1,...,*m* gde su $x_1^i,....,x_n^i$ indikatorske promenljive. Neka je Ω_k skup svih mogucih vrednosti *k*-tog indikatora. U slucaju kada imamo diskretan indikator, pretpostavljamo da klasifikacija pocinje sa nulom i prima uzastopne cele vrednosti. Tako na primer, «bracno stanje» možemo kodirati na sledeci nacin: 0 – razvedena, 1 – neudata, 2 – udata. U slucaju kada imamo neprekidnu promenljivu, nju možemo zameniti skupom diskretnih indikatora. Pretpostavimo da želimo da diskretizujemo *k*-tu indikatorsku promenljivu za objekat $x^i = (x_1^i,....,x_n^i)$. Neka *k*-ti indikator prima vrednosti iz intervala (0, *t*), gde je *t* ceo broj. Na primer, za neprekidni indikator kao što je «starost», možemo uvesti sledecu aproksimaciju:

godine	<18	18-20	21-24	25-29	30-55	56-59	60-75	76-84	>84
kod	0	1	2	3	4	5	6	7	8

Posmatrane funkcije korisnosti, pogotovo kvadratna funkcija korisnosti, mogu biti suviše fleksibilne (tj. mogu imati previše stepeni slobode) za skup podataka sa malim brojem tacaka. Uvodenje ogranicenja, kao što je na primer ogranicenje monotonosti na indikatorske promenljive može da smanji prekomernu fleksibilnost modela. Za linearnu funkciju korisnosti, uslov monotonosti za indikatorske promenljive se definiše na sledeci nacin: linearna funkcija $u^c(x) = c^T x = \sum_{k=1}^{n} c_k x_k$ je rastuca po x_k ako i samo ako je $c_k > 0$. Na primer, za problem skoring sistema možemo nametnuti uslov monotonosti za promenljive koje odgovaraju indikatoru «bracno stanje»:

$$\text{razvedena} \prec \text{neudata} \prec \text{udata}$$

Za kvadratnu funkciju je mnogo teže nametnuti ogranicenje monotonosti. Posmatramo potklasu monotonih kvadratnih funkcija sa nenegativnim elementima matrice *D* i vektora *c*. Ovakva funkcija je monotona po svakoj promenljivoj na $\boldsymbol{R}_+^n = \{x=(x_1,...,x_n) | x_k \geq 0, \ " \ k \in \{1,...,n\}\}$. Da bismo mogli da koristimo takvu klasu funkcija, potrebno je da indikatorske promenljive budu nenegativne i rastuce po preferenciji.

Za neke indikatore je moguce nametnuti ogranicenje monotonosti, dok je za druge bolje koristiti neke druge osobine, kao što su konveksnost i konkavnost. Ako

posmatramo neprekidan indikator «starost», pretpostavimo da su srednje godine bolje od mladih i starijih i neka je preferencija data sa:

$$\{0\text{-}18\} \prec \{18\text{-}20\} \prec \{21\text{-}24\} \prec \{25\text{-}29\} \prec \{85\text{-}\infty\} \prec$$
$$\{76\text{-}84\} \prec \{60\text{-}75\} \prec \{56\text{-}59\} \prec \{30\text{-}55\}$$

Preferencija za godine je odredena pomocu nemonotonog kriterijuma nad celim opsegom godina. Ovo možemo uvrstiti u model kreditnog bodovnom sistema na sledeci nacin:

1) Možemo da pretpostavimo da je funkcija korisnosti konkavna po datoj promenljivoj. Ovo možemo obezbediti racunanjem drugog izvoda po datoj promenljivoj i uvodenjem sledeceg ogranicenja u model linearnog programiranja:

$$\frac{\partial^2 u(x)}{\partial x_i^2} \leq 0$$

2) Kodiranje indikator promenljive «starost» možemo izvesti na sledeci nacin:

godine	<18	18-20	21-24	25-29	30-55	56-59	60-75	76-84	>84
kod	0	1	2	3	4	3	2	1	0

Na kraju želimo da odredimo pragove u_1 i u_2 i funkciju $u(x)$ koja klasifikuje skup svih zahteva za kredit $X_i = \{x^i | i \in I\}$ na sledeci nacin:

- Zahtev i je odobren ako i samo ako je $u_2 \leq u(x^i)$
- Odluka o zahtevu i je odložena ako i samo ako je $u_1 \leq u(x^i) < u_2$
- Zahtev je odbijen ako i samo ako je $u(x^i) < u_1$

Razmatramo funkciju korisnosti koja je linearna po promenljivama koje su korišcene za odlucivanje, i linearna ili kvadratna po indikatorskim promenljivama. Da bismo našli funkciju korisnosti koja klasifikuje skup objekata sa minimalnom greškom rešavamo ranije objašnjene probleme linearnog programiranja, u odeljku 3.2.4. i 3.2.5.

4.2 Dve faze kreditnog ocenjivanja u procesu odobravanja kredita

Pri podnošenju zahteva za kredit, od potencijalnog klijenta se zahteva da dostavi osnovne licne podatke, vlasništvo nad kucom ili nepokretnom imovinom, podatke o zaposlenju, kreditnu istoriju i stanje finansijske aktive i pasive. Pored ovih podataka dostavljenih od klijenta, banka traži izveštaj o klijentu i iz kreditnog biroa. Na osnovu svih ovih podataka, banka donosi odluku o kreditnoj sposobnosti klijenta. Ovaj postupak se najcešce zove *ekspertni sistem kreditne procene.* Ekspertni sistem je kasnije zamenjen *kreditnim skoring sistemom*, koji je matematicki model koji koristi karakteristike podnosioca zahteva da bi izracunao skor klijenta, koji se povezuje sa verovatnocom da klijent nece biti u mogucnosti da vraca kredit (*probability of default - PD*), ili da rangira klijente na osnovu njihovog *default rizika* (rizika da nece vratiti kredit). U ovom delu rada, koji se poziva na rad «*Two stages credit evaluation in bank loan appraisal*» [5], cemo videti da banka u procesu odlucivanja prolazi kroz dve faze, pri cemu cemo navesti i kriterijume koji se uzimaju u obzir da bi se ocenilo da li se isplati izvoditi drugu fazu odlucivanja, pošto se prva uvek izvodi.

Kao što smo vec napomenuli klijente klasifikujemo u dve grupe: dobre i loše (*G* i *B*) u zavisnosti od toga da li ce otplatiti kredit kada on dode na naplatu. Pretpostavlja se da je procenat dobrih *G* klijenata koji su kreditno sposobni i koji ce skoro sigurno otplatiti kredit jednak α, gde je $0<\alpha<1$. Dok je procenat loših *B* klijenata koji ce kasniti u otplati kredita i verovatno ga nece ni otplatiti jednak $1-\alpha$. Medutim, kada banka odobrava kredit ona ne može sa sigurnošcu da tvrdi kojoj grupi ce pripasti klijent. Nakon što banka primi kreditni zahtev sa zahtevanim dokumentima, ona ce doneti preliminarnu odluku na osnovu kreditnog skoring modela zasnovanog na standardizovanim podacima. Ova faza ocene se naziva *prva faza kreditne procene (e_1).* Banka zatim može da traži dodatne informacije o klijentu i tada se prelazi u *drugu fazu kreditne procene (e_2).*

Stopu preciznosti za prvu fazu kreditne procene cemo oznacavati sa q_1, to je verovatnoca da ce klijent iz grupe *G* (odnosno *B*) biti tacno klasifikovan u grupu *G* (odnosno *B*). Ali takode postoji $1-q_1$ šanse da banka klasifikuje klijenta iz grupe *G* u grupu *B* i obrnuto.

Troškovi rezervisanja (zaštite) od gubitaka su jednaki procenatu *k* koji se primenjuje na iznos odobrenog kredita. Banke moraju za svakog klijenta da odvoje odredeni iznos koji ce im barem deliminicno pokriti gubitke ukoliko klijent ne bude u mogucnosti da otplacuje kredit. Pretpostavlja se da su troškovi prve faze zanemarljivi. Pošto banaka u drugoj fazi zahteva dodatne informacije o klijentu, pretpostavlja se da ce stopa preciznosti druge faze q_2 biti veca od q_1. Kao i u prvoj fazi, postoji $1-q_2$ šanse da ce banka pogrešno klasifikovati klijenta.

Odluka o tome da li ce se banka odluciti za drugu fazu je problem koristi i troškova. Troškovi procene *A* druge faze i stope preciznosti q_1 i q_2 igraju odlucujucu ulogu. Ukoliko dolazi do velikog povecanja preciznosti u drugoj fazi u odnosu na prvu fazu, korisno je izvoditi drugu fazu. Medutim, ako je pomak mali, cak i ako su troškovi *A* relativno niski, banka ce odluciti da nije dovoljno ekonomicno da se izvodi druga faza.

Verovatnoca da li ce banka sprovesti drugu fazu procene zavisi od rezultata prve faze. Ako je klijent klasifikovan kao dobar u prvoj fazi, verovatnoca da ce banka sprovesti drugu fazu je jednaka a_G, a ukoliko je klasifikovan kao loš verovatnoca da ce biti sprovedena druga faza je jednaka a_B.

Model se sastoji iz sledecih koraka:

Korak 1: klijent podnosi zahtev za kredit sa zahtevanom dokumentacijom

Korak 2: banka sprovodi prvu fazu kreditne procene i deli klijente na dobre i loše

Korak 3: banka odlucuje da li da odobri kredit na osnovu rezultata iz prve faze ili da nastavi sa drugom fazom. Ako se sprovodi samo prva faze, kredit velicine *D* sa kamatnom stopom *i* ce biti odobren samo klijentima koji su svrstani u grupu *G*, dok ce klijenti koji su svrstani u grupu *B* odbijeni.

Korak 4: banka sprovodi drugu fazu kreditne procene.

Korak 5: kreditne odluke se donose na bazi rezultata iz druge faze procene, tj. klijenti koji su klasifikovani kao dobri, tj. u grupu *G* su dobili kredit, a oni koji su klasifikovani kao loši, tj. u grupu *B* su odbijeni.

Pretpostavlja se da je kamatna stopa *i* veca od procenta troškova rezervisanja *k*. Stopa povracaja banke zavisi od visine kamatne stope (*i*), procenta rezervisanja za eventualne gubitke (*k*) i velicine kredita (*D*). Da bi se model pojednostavio pretpostavlja se da je ocekivana stopa povracaja od davanja kredita dobrom klijentu ($E(r|G)$) pozitivna, a lošem klijentu ($E(r|B)$) negativna, tj.

$$E(r|G) = D(i-k) > 0$$

$$E(r|B) = -D(1+k) < 0$$

Lema 1 Bez upotrebe sistema procene kreditnog rizika, banka daje kredit sa kamatnom stopom $i \geq \frac{1+k-a}{a} \equiv i^*$, gde se i^* može posmatrati kao kamatna stopa u savršeno konkurentnom kreditnom tržištu koja osigurava da ce ocekivana stopa povracaja za banku biti nenegativna.

Dokaz Bez upotrebe sistema za ocenu kreditnog rizika, banka ce odobriti kredit ako i samo ako je ocekivana stopa povracaja $E(r)$ za banku nenegativna.

$$E(r) = aE(r \mid G) + (1-a)E(r \mid B) \geq 0 \Leftrightarrow aD(i-k) - (1-a)D(1+k) \geq 0$$

$$\Leftrightarrow D(ai - 1 - k + a) \geq 0, D \geq 0$$

$$\Leftrightarrow i \geq \frac{1+k-a}{a} \equiv i^*$$

U praksi, banka ce uvek sprovesti prvu fazu da bi procenila kreditnu sposobnost klijenta. Pretpostavlja se da su troškovi sprovodenja prve faze zanemarljivi i da banka uvek izvodi ovu fazu. Lema 2 postavlja donju granicu na stopu preciznosti prve faze q_1.

Lema 2 Pretpostavljamo da su troškovi prve faze procenjivanja zanemarljivi, sledeci uslov za stopu preciznosti prve faze mora da bude zadovoljen da bi se sprovela prva faza procene:

$$q_1 \geq \frac{a(i-k)}{a(i-k)+(1-a)(1+k)} \equiv q_1^*$$

U uslovima savršene konkurencije na kreditnom tržištu, q_1^* je jednako ½.

Dokaz Da bi banka sprovela prvu fazu procene, ocekivani povracaj koji zavisi od procene prve faze $E(r|e_1)$ mora da bude veci ili jednak ocekivanom povracaju $E(r)$ bez primene sistema za ocenjivanje.

$$E(r \mid e_1) \geq E(r) \Leftrightarrow aq_1 D(i-k) - (1-a)(1-q_1)D(1+k) \geq aD(i-k) - (1-a)D(1+k)$$

$$\Leftrightarrow a(i-k)q_1 + (1-a)(1+k)q_1 \geq a(i-k)$$

$$\Leftrightarrow q_1 \geq \frac{a(i-k)}{a(i-k)+(1-a)(1+k)} \equiv q_1^*$$

U uslovima savršene konkurencije, kamatna stopa $i = \frac{1+k-a}{a}$. Kada to uvrstimo u q_1^* dobijamo da je $q_1^* = 1/2$.

Nakon što su ispunjeni uslovi iz Leme 1 i Leme 2, banka ce odluciti da li joj se isplati da sprovodi drugu fazu, ili da donese odluku na osnovu rezultata iz prve faze. Teorema 1 nam daje uslov na osnovu koga se banka odlucuje za drugu fazu u situaciji kada je klijent klasifikovan kao dobar u prvoj fazi. Dok nam Teorema 2 daje uslov na osnovu koga se banka odlucuje za drugu fazu u situaciji kada je klijent klasifikovan kao loš u prvoj fazi.

Teorema 1 Neka je klijent klasifikovan kao dobar, tj. u grupu G, $q_1 \geq q_1^*$ i $i = i^*$, ako je $A < A_G^* \equiv \frac{(1-a)(q_2-q_1)(1+k)D}{aq_1+(1-a)(1-q_1)}$ tada je verovatnoca da ce se sprovesti druga faza procene jednaka $a_G^* = 1$, u suprotnom je $a_G^* = 0$.

Dokaz Ako je klijent u prvoj fazi klasifikovan kao dobar, što oznacavamo sa e_{1G} i ako je verovatnoca izvodenja druge faze procene definisana sa a_G, tada je ocekivani povracaj za banku nakon što se izvede druga faza jednak:

$$E(r \mid e_{1G}, e_2) = a_G\left(\frac{aq_1}{aq_1+(1-a)(1-q_1)}(q_2(i-k)+(1-q_2)\cdot 0)D + \frac{(1-a)(1-q_1)}{aq_1+(1-a)(1-q_1)}(-(1-q_2)(1+k)\right.$$

$$\left. + q_2 \cdot 0)D - A\right) + (1-a_G)\left(\frac{aq_1}{aq_1+(1-a)(1-q_1)}(i-k)D - \frac{(1-a)(1-q_1)}{aq_1+(1-a)(1-q_1)}(1+k)D\right)$$

Banka ce odluciti da izvede drugu fazu procene jedino ako povecanje verovatnoce sprovodenja druge faze procene ima za posledicu vece ocekivane prinose, tj. $\frac{\partial E(r \mid e_{1G}, e_2)}{\partial a_G} > 0$, u suprotnom banka nece sprovesti drugu fazu, tj. $a_G^* = 0$.

$$\frac{\partial E(r \mid e_{1G}, e_2)}{\partial a_G} = (\frac{D}{aq_1 + (1-a)(1-q_1)})(aq_1q_2(i-k) - aq_1(i-k) - (1-a)(1-q_1)(1-q_2)(1+k) +$$
$$+ (1-a)(1-q_1)(1+k)) - A > 0$$
$$\Leftrightarrow A < ((1-a)(1-q_1)q_2(1+k) - aq_1(1-q_2)(i-k))(\frac{D}{aq_1 + (1-a)(1-q_1)})$$
$$\Leftrightarrow A < \frac{(1-a)(q_2-q_1)(1+k)D}{aq_1 + (1-a)(1-q_1)} \equiv A_G^*$$

gde je $i = i^* = \frac{1+k-a}{a}$.

Teorema 2 Neka je klijent klasifikovan kao loš, tj. u grupu *B*, $q_1 \geq q_1^*$ i $i = i^*$, ako je $A < A_B^* \equiv \frac{(1-a)(q_2-q_1)(1+k)D}{a(1-q_1) + (1-a)q_1}$ tada je verovatnoca da ce se sprovesti druga faza procene jednaka $a_B^* = 1$, u suprotnom je $a_B^* = 0$.

Dokaz Ako je klijent u prvoj fazi klasifikovan kao loš, što oznacavamo sa e_{1B} i ako je verovatnoca izvodenja druge faze procene definisana sa a_B, tada je ocekivani povracaj za banku nakon što se izvede druga faza jednak:

$$E(r \mid e_{1B}, e_2) = a_B(\frac{a(1-q_1)}{a(1-q_1) + (1-a)q_1}(q_2(i-k) + (1-q_2)\cdot 0)D + \frac{(1-a)q_1}{a(1-q_1) + (1-a)q_1}(-(1-q_2)(1+k)$$
$$+ q_2 \cdot 0)D - A) + (1-a_B)\cdot 0$$

Banka ce odluciti da izvede drugu fazu procene jedino ako povecanje verovatnoce sprovodenja druge faze procene ima za posledicu vece ocekivane prinose, tj. $\frac{\partial E(r \mid e_{1B}, e_2)}{\partial a_B} > 0$, u suprotnom banka nece sprovesti drugu fazu, tj. $a_B^* = 0$.

$$\frac{\partial E(r \mid e_{1B}, e_2)}{\partial a_B} = (\frac{D}{a(1-q_1) + (1-a)q_1})(a(1-q_1)q_2(i-k) - aq_1(i-k) - (1-a)q_1(1-q_2)(1+k)) - A > 0$$
$$\Leftrightarrow A < (a(1-q_1)q_2(i-k) - (1-a)q_1(1-q_2)(1+k))(\frac{D}{a(1-q_1) + (1-a)q_1})$$
$$\Leftrightarrow A < \frac{(1-a)(q_2-q_1)(1+k)D}{a(1-q_1) + (1-a)q_1} \equiv A_B^*$$

gde je $i = i^* = \frac{1+k-a}{a}$.

Teoreme 1 i 2 daju granicne troškove druge faze ukoliko banka odluci da je sprovede. Ako su stvarni troškovi druge faze procene manji od granicnih troškova, isplati se sprovoditi drugu fazu. Brojilac granicnih troškova $A_G{}^*$ i $A_B{}^*$ je korist od izbegavanja gubitaka koji se stvaraju davanjem kredita lošim klijentima, a imenilac predstavlja procenat klijenata koji su klasifikovani kao dobri odnosno kao loši u prvoj fazi.

U ovom modelu, pretpostavlja se da banka izjednacava kamatnu stopu sa kamatnom stopom u uslovima savršene konkurencije na tržištu kapitala, tj. $i = i^* = \frac{1+k-a}{a}$.

Lema 3 Ako je $\alpha < ½$ i $q_1 \geq ½$ tada je $A_G^* \geq A_B^*$. Ali ako je $\alpha > ½$ i $q_1 \geq ½$ tada je $A_G^* \leq A_B^*$.

Dokaz Ako je $A_G^* \geq A_B^*$, tada je

$$\frac{(1-a)(q_2-q_1)(1+k)}{aq_1+(1-a)(1-q_1)} \geq \frac{(1-a)(q_2-q_1)(1+k)D}{a(1-q_1)+(1-a)q_1}$$

$$\Leftrightarrow a(1-q_1)+(1-a)q_1 \geq aq_1+(1-a)(1-q_1)$$

$$\Leftrightarrow (1-2a)q_1 \geq (1-2a)(1-q_1)$$

$$\Leftrightarrow q_1 \geq 1/2 \text{ ako je } a < 1/2$$

Dakle, ako je $q_1 \geq 1/2$ i $a < 1/2$ tada je $A_G^* \geq A_B^*$.

Ako je $A_B^* \geq A_G^*$, tada je

$$\frac{(1-a)(q_2-q_1)(1+k)}{aq_1+(1-a)(1-q_1)} \leq \frac{(1-a)(q_2-q_1)(1+k)D}{a(1-q_1)+(1-a)q_1}$$

$$\Leftrightarrow a(1-q_1)+(1-a)q_1 \leq aq_1+(1-a)(1-q_1)$$

$$\Leftrightarrow (1-2a)q_1 \leq (1-2a)(1-q_1)$$

$$\Leftrightarrow q_1 \geq 1/2 \text{ ako je } a > 1/2$$

Dakle, ako je $q_1 \geq 1/2$ i $a > 1/2$ tada je $A_B^* \geq A_G^*$.

Lema 3 pokazuje da ako je procenat klijenata koji otplacuju kredit manji (veci) od onih koji kasne u otplati, granicni troškovi za dobre klijente su viši (niži) od tih troškova za loše klijente.

Teorema 3 Neka je $\alpha = ½$, $i = i^*$ i $q_1 \geq q_1^* = ½$,

(1) ako je $A < A_G^* = A_B^*$, tada je $a_G^* = a_B^* = 1$

(2) ako je $A \geq A_G^* = A_B^*$, tada je $a_G^* = a_B^* = 0$, gde je $A_G^* = \frac{(1-a)(q_2-q_1)(1+k)D}{aq_1+(1-a)(1-q_1)}$,

a $A_B^* = \frac{(1-a)(q_2-q_1)(1+k)D}{a(1-q_1)+(1-a)q_1}$.

Dokaz Ako je $a = 1/2$ tada je

$$A_G^* = \frac{(1-a)(q_2-q_1)(1+k)D}{aq_1+(1-a)(1-q_1)} = (q_2-q_1)(1+k)D$$

$$A_B^* = \frac{(1-a)(q_2-q_1)(1+k)D}{a(1-q_1)+(1-a)q_1} = (q_2-q_1)(1+k)D$$

Pošto je $A_G{}^* = A_B{}^*$,

(1) ako je $A < A_G{}^* = A_B{}^*$ onda je $a_G{}^* = a_B{}^* = 1$ na osnovu Teoreme 1 i Teoreme 2

(2) ako je $A \geq A_G{}^* = A_B{}^*$, tada je $a_G{}^* = a_B{}^* = 0$ na osnovu Teoreme 1 i Teoreme 2.

Teorema 3 pokazuje da kada je procenat dobrih i loših klijenata jednak, tada su i granicni troškovi jednaki.

Teorema 4 Neka je $\alpha < ½$, $i = i^*$ i $q_1 \geq q_1^* = ½$,

(1) ako je $A < A_B{}^*$, tada je $a_G{}^* = a_B{}^* = 1$;

(2) ako je $A_B{}^* \leq A < A_G{}^*$, tada je $a_G{}^* = 1$ i $a_B{}^* = 0$;

(3) ako je $A \geq A_G{}^*$, tada je $a_G{}^* = a_B{}^* = 0$, gde je $A_G^* = \frac{(1-a)(q_2-q_1)(1+k)D}{aq_1+(1-a)(1-q_1)}$ i $A_B^* = \frac{(1-a)(q_2-q_1)(1+k)D}{a(1-q_1)+(1-a)q_1}$.

Dokaz Na osnovu Leme 3, ako je $a < 1/2$ i $q_1 \geq 1/2$ tada je $A_G^* \geq A_B^*$. Zatim,

(1) ako je $A < A_B{}^*$, tada je $a_G{}^* = a_B{}^* = 1$ na osnovu Teoreme 1 i Teoreme 2

(2) ako je $A_B{}^* \leq A < A_G{}^*$, tada je $a_G{}^* = 1$ i $a_B{}^* = 0$ na osnovu Teoreme 1 i Teoreme 2

(3) ako je $A \geq A_G{}^*$, tada je $a_G{}^* = a_B{}^* = 0$ na osnovu Teoreme 1 i Teoreme 2

Teorema 5 Neka je $\alpha > ½$, $i = i^*$ i $q_1 \geq q_1^* = ½$,

(1) ako je $A < A_G{}^*$, tada je $a_G{}^* = a_B{}^* = 1$;

(2) ako je $A_G{}^* \leq A < A_B{}^*$, tada je $a_G{}^* = 0$ i $a_B{}^* = 1$;

(3) ako je $A \geq A_B{}^*$, tada je $a_G{}^* = a_B{}^* = 0$, gde je $A_G^* = \frac{(1-a)(q_2-q_1)(1+k)D}{aq_1+(1-a)(1-q_1)}$ i $A_B^* = \frac{(1-a)(q_2-q_1)(1+k)D}{a(1-q_1)+(1-a)q_1}$.

Dokaz Na osnovu Leme 3, ako je $a > 1/2$ i $q_1 \geq 1/2$ tada je $A_G^* \leq A_B^*$. Zatim,

(1) ako je $A < A_G{}^*$, tada je $a_G{}^* = a_B{}^* = 1$ na osnovu Teoreme 1 i Teoreme 2

(2) ako je $A_G{}^* \leq A < A_B{}^*$, tada je $a_G{}^* = 0$ i $a_B{}^* = 1$ na osnovu Teoreme 1 i Teoreme 2

(3) ako je $A \geq A_B{}^*$, tada je $a_G{}^* = a_B{}^* = 0$ na osnovu Teoreme 1 i Teoreme 2

Teoreme 4 i 5 demonstriraju, da u zavisnosti od odnosa dobrih i losih klijenata, banka sprovodi optimalnu strategiju da li da izvede drugu fazu na (1) obe grupe, (2) samo jednoj grupi ili (3) ni na jednoj grupi. Ako je procenat klijenata koji otplacuju kredit manji od procenta klijenata koji kasne u otplati, granicni troškovi za dobre klijente su viši od onih za loše klijente. U ovom slucaju postoji mogucnost (ako je $A_B^* \leq A < A_G^*$) da banka odluci da sprovede drugu fazu samo na dobre klijente. Tada banke teže da sprovedu drugu fazu procene na klijente koji su klasifikovani kao dobri, u suprotnom, namera da se dalje procenjuje je mnogo slabija.

Ukoliko je procenat dobrih klijenata viši od procenta loših klijenata, granicni troškovi ce biti niži od onih za loše klijente. U tom slucaju postoji mogucnost (ako je $A_G^* \leq A < A_B^*$) da banka odluci da sprovede drugu fazu samo na loše klijente. Tada banke teže da sprovedu drugu fazu procene na klijente koji su klasifikovani kao loši, u suprotnom, namera da se dalje procenjuje je mnogo slabija.

Sada cemo na kratkom numerickom primeru videti kakva je osetljivost rezultata u zavisnosti od vrednosti parametara (k, D, q_1, q_2)

Pretpostavke: k = 5%, D = \$100, q_1 = 0.6, q_2 = 0.7		
Rezultati	a = **0.95**	a = **0.40**
i^*	10.53%	162.5%
q_1^*	0.50	0.50
A_G^*	\$ 0.89	\$ 13.13
A_B^*	\$ 1.28	\$ 12.12
Pretpostavke: k = 5%, D = \$100, q_1 = 0.6, q_2 = 0.9		
i^*	10.53%	162.50%
q_1^*	0.50	0.50
A_G^*	\$ 2.67	\$ 39.38
A_B^*	\$ 3.84	\$ 36.35

Kao što vidimo iz tabele, ako se pretpostavi da je $\alpha = 0.95$ i da je mala promena u stopi preciznosti, A_G^* i A_B^* su veoma niski pri cemu je A_G^* niža od A_B^*, što je u saglasnosti sa Teoremom 5. Ako procenat dobrih klijenata padne ispod 0.50 (α = 0.40), zajedno sa A_G^* i A_B^* se drasticno povecava i u tom slucaju je A_G^* vece od A_B^*, što je u saglasnosti sa Teoremom 4. Medutim, u tom slucaju, kamatna stopa mora da bude izuzetno visoka (162.5 %). U drugom primeru vidimo da znacajan napredak u stopi preciznosti rezultira u višim granicnim troškovima A_G^* i A_B^*, a samim tim je veca težnja da se sprovede duga faza procenjivanja.

4.3 Primena probit i tobit modela

U praksi, vecina kreditnih skoring sistema pati od pristrasnosti pri odabiru uzorka zbog toga što se modeli ocenjuju na uzorku koji se sastoji od odobrenih kredita, pri cemu se ne uzima u obzir kriterijum po kojem se klijenti odbijaju. Glavni razlog tome je nedostatak podataka o odbijenim klijentima koji su dostupni javnosti. Boyes je izbegao ovu pristrasnost tako što je formirao binarni probit model sa dva uzastopna dogadaja kao zavisne promenljive: odluka da li odobriti kredit ili ne, i zavisno od toga da li je kredit odobren, klijentova sposobnost da ga otplati. On je, koristeci svoj model, došao da zakljucka da odobravanje kredita nije u saglasnosti sa minimizacijom rizika da ce klijent biti loš.

Za pocetak treba oceniti nepristrasan kreditni skoring sistem, tj. binarni probit model. U ovom primeru korišcena je velika baza podataka koja sadrži veliki skup finansijskih i licnih informacija kako o odobrenim tako i o odbijenim klijentima.

Skup podataka se sastoji od 13,338 zahteva za kredit u jednoj Švedskoj banci u periodu izmedu septembra 1994. i avgusta 1995. godine (pozivamo se na rad «*Bank lending policy, credit scoring and value-at-risk*» [1]). Krediti su davani u obliku kreditnih kartica. To su revolving krediti tako da nemaju odredeno vreme do kada treba da se otplate. U ovom primeru, kredit se klasifikuje kao *loš* ukoliko je u kašnjenju dužem od devedeset dana i samim tim je završio u agenciji za naplatu kredita koji su u problemu. Od pocetnih 57 promenljivih, korišceno je samo 17 za konacnu ocenu modelu. Mnoge promenljive su iskljucene iz modela zbog odsustva odnosa sa promenljivama od znacaja - odluka da li odobriti kredit i ponašanje pri otplati kredita. Drugi razlog je izuzetno jaka korelacija sa nekom drugom promenljivom koja se odnosi na istu stvar (kao što su na primer bruto i neto prihodi), a ova druga ima vecu objašnjavajucu moc. Promenljive koje se nalaze u konacnom modelu i njihove definicije su sledece:

1. *Starost* – koliko godina ima klijent
2. *Muško* – binarna promenljiva, uzima vrednost 1 ako je klijent muško
3. *Razveden* - binarna promenljiva, uzima vrednost 1 ako je klijent razveden
4. *Kuca*- binarna promenljiva, uzima vrednost 1 ako je klijent vlasnik kuce
5. *Veliki grad* - binarna promenljiva, uzima vrednost 1 ako klijent živi u jednom od tri najveca grada
6. *Broj zahteva* – broj zahteva za informacije o klijentu koje je kreditni biro primio u poslednjih 36 meseci
7. *Preduzece* - binarna promenljiva, uzima vrednost 1 ako klijent ima oporezujuci prihod od registrovanog posla
8. *Prihod* – prijavljeni godišnji prihod od plata
9. *Razlika prihoda* – razlika izmedu godišnjih prihoda tekuce i prethodne godine
10. *Kapital* - binarna promenljiva, uzima vrednost 1 ako klijent ima oporezujuci prihod od kapitala

11. *Balans* – odnos prihoda i ukupnih kreditnih kapaciteta, izraženo u procentima
12. *Nula limit* - binarna promenljiva, uzima vrednost 1 ako klijent nema neizmirenih obaveza po kreditima
13. *Limit* – ukupan odobreni iznos kredita
14. *Broj kredita* – broj kredita
15. *Upotrebljen limit* - procenat *Limita* koji je iskorišcen
16. *Iznos kredita* – iznos odobrenog kredita
17. *Jemac* - binarna promenljiva, uzima vrednost 1 ako klijent ima jemca

Od svih 13,338 klijenata, 6899 ili 51.7% je odbijeno za kredit. Dakle, 6439 klijenata je dobilo kredit. U *Tabeli 1* su date deskriptivne statistike za gore navedene promenljive.

Tabela 1: Deskriptivne statistike za promenljive korišcene u modelu

Promenljive	**Odobreni klijenti (*N* = 6899)**				**Odbijeni klijenti (*N* = 6439)**			
	Srednja vrednost	*Standardna devijacija*	*Minimalna vrednost*	*Maksimalna vrednost*	*Srednja vrednost*	*Standardna devijacija*	*Minimalna vrednost*	*Maksimalna vrednost*
Starost	38.65	12.76	18	84	41.02	12.08	20	83
Muško	0.62	0.48	0	1	0.65	0.48	0	1
Razveden	0.13	0.34	0	1	0.14	0.35	0	1
Kuca	0.34	0.47	0	1	0.47	0.50	0	1
Veliki grad	0.41	0.49	0	1	0.37	0.48	0	1
Broj zahteva	4.69	2.60	1	10	4.81	2.68	1	19
Preduzece	0.04	0.21	0	1	0.02	0.16	0	1
Prihod	129.93	70.38	0	737.9	189.47	75.70	0	1093.0
Razlika prihoda	5.37	34.06	-438.5	252.6	9.03	34.63	-6226	5006.0
Kapital	0.12	0.32	0	1	0.07	0.25	0	1
Balans	91.04	894.53	0	41533	31.01	386.15	0	22387
Nula limit	0.15	0.36	0	1	<0.01	0.05	0	1
Limit	79.89	93.69	0	1703.0	50.47	51.07	0	949.2
Broj kredita	2.99	2.42	0	18	3.65	2.04	0	16
Upotrebljen limit	64.34	38.88	0	278.0	53.22	33.94	0	124.0
Jemac	0.16	0.36	0	1	0.14	0.35	0	1

Primetimo da je u ovoj tabeli izostavljena promenljiva «*Iznos kredita*», razlog je taj što ona nema vrednost za odbijene klijente, ali se nalazi u sledecoj tabeli u kojoj su predstavljene vrednosti za odobrene klijente.

U oktobru 1996. je uraden monitoring klijenata kojima je odobren kredit i došlo se do zakljucka da je 6% ili 388 klijenata zapalo u problem nemogucnosti otplacivanja

kredita, tj. bili su prosledeni u agenciju za naplatu duga, dok su ostali klijenti redovno otplacivali kredit. U *Tabeli 2* predstavljene su deskriptivne statistike za ove klijente:

Tabela 2: Deskriptivne statistike za odobrene kredite

Promenljive	**Loši klijenti (*N* = 388)**				**Dobri klijenti (*N* = 6051)**			
	Srednja vrednost	*Standardna devijacija*	*Minimalna vrednost*	*Maksimalna vrednost*	*Srednja vrednost*	*Standardna devijacija*	*Minimalna vrednost*	*Maksimalna vrednost*
Starost	36.11	11.03	21	75	41.33	12.07	20	83
Muško	0.67	0.47	0	1	0.65	0.48	0	1
Razveden	0.20	0.40	0	1	0.14	0.35	0	1
Kuca	0.28	0.45	0	1	0.48	0.50	0	1
Veliki grad	0.41	0.49	0	1	0.36	0.48	0	1
Broj zahteva	6.15	2.85	1	14	4.72	2.64	1	19
Preduzece	0.02	0.13	0	1	0.03	0.16	0	1
Prihod	165.36	82.35	0	1093.0	191.01	75.00	0	1031.7
Razlika prihoda	3.52	39.01	-135.0	439.7	9.38	34.30	-622.6	500.6
Kapital	0.04	0.20	0	1	0.07	0.26	0	1
Balans	39.92	313.51	0	6041	30.44	390.36	0	22387
Nula limit	0.04	0.20	0	1	<0.01	0.02	0	1
Limit	41.44	57.98	0	511.5	51.05	50.54	0	949.2
Broj kredita	2.34	1.64	0	11	3.74	2.04	0	16
Upotrebljen limit	75.69	33.37	0	124.0	51.78	33.47	0	112.0
Iznos kredita	7.08	3.95	3.0	24.5	7.12	3.83	3.0	30.0
Jemac	0.07	0.26	0	1	0.14	0.35	0	1

U ovom primeru koristi se binarni probit model koji se sastoji od dve simultane jednacine. Prva y_{1i} za odluku da li odobriti kredit ili ne, a druga y_{2i} za binarni rezultat, «dobar» ili «loš» klijent. Neka je sa * oznacena neopažena promenljiva i pretpostavljamo da je:

$$y_{1i}^{*} = \mathbf{x}_{1i}\alpha_1 + e_{1i}$$

$$y_{2i}^{*} = \mathbf{x}_{2i}\alpha_2 + e_{2i} \quad \text{za} \quad i = 1, 2, ..., N$$

gde su $\mathbf{x}_{ji}$, j = 1,2, 1 x k_j vektori nezavisnih promenljivih i pretpostavlja se da one imaju normalnu raspodelu sa nula ocekivanjem, jedinicnom varijansom i koeficijentom korelacije ρ.

Binarna promenljiva y_{1i} uzima vrednost 1 ako je kredit odobren, a ukoliko je klijent odbijen uzima vrednost 0:

$$y_{1i} = \begin{cases} 0, odbijen(y_{1i}^{*} < 0) \\ 1, odobren(y_{1i}^{*} \geq 0) \end{cases}$$

Binarna promenljiva y_{2i} uzima vrednost 1 ako je klijent dobar, a 0 ukoliko je klijent loš:

$$y_{2i} = \begin{cases} 0, loš(y_{2i}^* < 0) \\ 1, dobar(y_{2i}^* \geq 0) \end{cases}$$

Ocenjeni parametri i njihove standardne greške su predstavljene u *Tabeli 3*.

Tabela 3: ocenjeni parametri binarnog probit modela

Promenljive	**P (da je kredit odobren)**		**P (da je klijent dobar)**	
	$\hat{a}_1$	*t* - statistika	$\hat{a}_2$	*t* - statistika
Konstatnta	-0.2374	-3.57***	2.2900	15.65***
Starost	-0.004303	-3.69***	0.006892	2.63***
Muško	-0.2003	-7.10***	-0.02456	-0.43
Razveden	-0.02588	-0.70	-0.2380	-3.34***
Kuca	0.06391	2.32**	-0.02019	0.35
Veliki grad	-0.2382	-8.96***	-0.03724	-0.69
Broj zahteva	-0.008123	-1.58	-0.1000	-9.84***
Preduzece	0.5223	8.30***	0.2065	1.28
Prihod	0.008929	49.17***	-0.002392	-4.81***
Razlika prihoda	-0.002336	-6.78***	0.002233	3.04***
Kapital	-0.2776	-5.48***	0.1189	0.97
Balans	0.00006548	3.48***	-0.00009135	-1.54
Nula limit	-2.244	-54.94***	0.005064	-2.23**
Limit	-0.008381	-54.94***	0.005064	10.28***
Broj kredita	0.08420	12.23***	0.2698	14.41***
Upotrebljen limit	-0.007746	-17.72***	-0.01197	-12.91***
Iznos kredita	-	-	-0.006637	-0.98
Jemac	0.1300	3.83***	0.4374	4.50***
r	-	-	-0.9234	-17.34***

gde *,**,*** predstavljaju 10%, 5% i 1% nivo znacajnosti.

Primetimo da «*Iznos kredita*» ne može da se koristi kao nezavisna promenljiva u prvoj jednacini, pošto ne postoje podaci za ovu promenljivu u slucaju odbijenih klijenata. Uticaj vecine promenljivih na verovatnocu da ce klijent biti loš je u saglasnosti sa politikom banke. Tako na primer «*Prihodi*», «*Kuca*», «*Preduzece*», «*Broj kredita*» i «*Jemac*» su vrlo bitni faktori koji pozitivno uticu, dok «*Nula limit*», «*Limit*» i «*Upotrebljen limit*» imaju negativan uticaj kada banka donosi odluku.

Promenljiva «*Iznos kredita*» nema znacajan uticaj na rizik da ce klijent biti loš, dok «*Limit*» ima mnogo veci uticaj. Zato se mora biti obazriv u generalizaciji ovih

rezultata u slucaju kada je odnos «*Iznosa kredita*» i «*Limita*» blizu jedan. Ali ipak, rezultati pokazuju da kada se posmatraju manji krediti, iznos kredita ne utice na rizik koji nosi taj plasman.

Koeficijent korelacije -0.9234 pokazuje skoro savršenu korelaciju izmedu odluke da se odobri kredit i rizika da ce klijent biti loš.

Kada se posmatraju rezultati u *Tabeli 3*, dolazi se do sledecih zakljucaka: prvo, vecina promenljivih koje pozitivno uticu na odobravanje kredita nisu medu onima koje smanjuju rizik da ce klijent biti loš, dakle, banka ne teži ka tome da minimizira rizik; takode, rezultati pokazuju da iznos kredita ne utice na rizik. Dakle, banke teže da odobravaju vece kredite cak i ako su oni rizicniji da bi maksimizirale profit, pošto se za rizicnije klijente obracunava viša kamatna stopa, pa je samim tim i viša ocekivana stopa prinosa.

Pored toga da li ce klijent biti loš, za banke je vrlo bitno da znaju i kada ce se desiti da klijent upadne u problem i prestane da otplacuje kredit, tj. da ocene vreme **«preživljavanja kredita»**. Binarni Tobit model je vrlo efikasan u razdvajanju klijenata koji brzo postanu loši i onih koji imaju duže vreme preživljavanja ili nikada ne postanu loši. Uzeti su rezultati iz rada «*Bank lending policy, credit scoring and the survival of loans*» [2].

Koristimo se vec datim primerom, jedino što je u modelu umesto promenljive «*Starost*» ukljucena binarna promenljiva «*Oženjen*» koja uzima vrednost jedan ako je klijent razveden, dok u ovom slucaju promenljiva «*Razveden*» znaci da je klijent udovac ili da je razdvojen. U ovom slucaju, klijent je loš ako kasni u otplati kredita duže od devedeset dana, a dobar ako i dalje izmiruje svoje obaveze, tj. i dalje je aktivan. Za loše klijente, vreme preživljavanja predstavlja broj dana izmedu odobravanja kredita i dana kada je postao loš, dok je za dobre klijente vreme preživljavanja razlika izmedu dana kada je odobren kredit i dana kada je raden monitoring, u ovom slucaju 9.oktobra 1996.

Model se sastoji od dve simultane jednacine, prva za odluku da li odobriti kredit ili ga odbiti, y_i, a druga je prirodni logaritam od vremena preživljavanja kredita (izraženo u danima), t_i. Neka je sa * oznacena neopažena promenljiva i pretpostavljamo da je:

$$y_i^* = \mathbf{x}_{1i} b_1 + e_{1i}$$

$$t_i^* = \mathbf{x}_{2i} b_2 + e_{2i} \quad \text{za} \quad i = 1, 2, ..., N \qquad (*)$$

i pretpostavimo da su raspodele dvodimezionalne normalne raspodele:

$$\begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} \sim N\begin{pmatrix} 0 & 1 & s_{12} \\ 0, & s_{12} & s_2^2 \end{pmatrix}$$

Binarna promenljiva y_i uzima vrednost 1 ako je kredit odobren, a ukoliko je klijent odbijen uzima vrednost 0:

$$y_i = \begin{cases} 0, odbijen(y_i^* < 0) \\ 1, odobren(y_i^* \geq 0) \end{cases}$$

Za loše kredite, može da se uoci tacno vreme preživaljavanja. Dok za kredite koji se i dalje redovno otplacuju u vreme monitoringa, preživljavanje je ocenjeno pošto ne znamo da li ce i kada ce klijent postati loš. Tako na primer, kreditu koji je odobren 1.septembra 1994. godine prag za ocenjivanje ce biti 768 dana, a za kredit koji je odobren 31.avgusta 1995. godine on ce biti 434 dana. Ova vrednost praga za ocenjivanje ce biti oznacena sa $\bar{t}_i$. Dakle, pravilo za ocenjivanje je sledece:

$$t_i = \begin{cases} t_i^*, ako, t_i^* < \bar{t}_i \\ \bar{t}_i, ako, t_i^* \geq \bar{t}_i \end{cases}$$

Binarna promenljiva d_i razdvaja skup odobrenih kredita na dobre i loše i to na sledeci nacin, ukoliko je vreme preživljavanja kredita manje od praga $\bar{t}_i$ tada je klijent loš, u suprotnom je dobar, tj.

$$d_i = \begin{cases} 0, ako, t_i^* \leq \bar{t}_i \\ 1, ako, t_i^* > \bar{t}_i \end{cases}$$

Za ocenjivanje parametara koristimo sledecu jednacinu:

$$\ln l = \sum_{i=1}^{N} (1-y_i) \ln(1-\Phi(x_{1i} b_1)) + \sum_{i=1}^{N} y_i (1-d_i) \left\{ \ln \Phi\left(\frac{x_{1i} b_1 - \frac{s_{12}}{s_2^2}(t_i - x_{2i} b_2)}{\sqrt{(1-r^2)}}\right) - \frac{1}{2}\ln 2p + \ln\left(\frac{1}{s_2}\right) - \frac{1}{2}\left(\frac{t_i - x_{2i} b_2}{s_2}\right)^2 \right.$$

$$+ \sum_{i=1}^{N} y_i d_i \ln \Phi_2\left(x_{1i} b_1, \frac{x_{2i} b_2 - \bar{t}_i}{s_2}; r\right)$$

gde $\Phi(\bullet)$ i $\Phi(\bullet,\bullet; \rho)$ predstavljaju jednodimenzionalnu i binarnu standardnu normalnu funkciju raspodele, poslednja sa koeficijentom korelacije ρ.

U *Tabeli 4* su prikazani jednodimenzionalne i binarne ocene za parametre β_1 i β_2 koje su dobijene korišcenjem funkcije maksimalne verovatnoce. Jednodimenzionalne ocene za β_1 i β_2 se dobijaju kada se pojedinacno ocenjuju prva i druga jednacina u (*), pri cemu se pretpostavlja da je $\rho = 0$. Dvodimenzionalne ocene su dobijene kada su jednacine u (*) rešavane zajedno, tako da su uzeti u obzir efekti odabira uzorka, takode neophodna je i ocena koeficijenta korelacije ρ. Binarna probit ocena α_2 se dobija kada se u (*) umesto druge jednacine koristi jednacina koja meri default rizik, tj. rizik da ce klijent biti loš.

Tabela 4: Jednodimenzionalne i binarne ocene za parametre; *standardne greške su date u zagradama, a promenljive koje su znacajne na nivou znacajnosti 10% su podebljane*

Jednacine					
Promenljive	**Odobravanje kredita**		**Preživljavanje kredita**		**Nije loš**
	b_1		b_2		a_2
	Jednodimenzionalna	**Binarna**	**Jednodimenzionalna**	**Binarna**	**Binarna**
Konstatnta	**-0.336**	**-0.328**	**8.246**	**9.065**	**2.460**
	(0.051)	(0.051)	(0.156)	(0.193)	(0.113)
Veliki grad	**-0.232**	**-0.222**	**-0.128**	0.233	-0.042
	(0.027)	(0.027)	(0.058)	(0.053)	(0.055)
Muško	**-0.207**	**-0.196**	**-0.106**	0.024	0.024
	(0.028)	(0.028)	(0.061)	(0.058)	(0.059)
Razveden	**-0.186**	**-0.179**	-0.124	-0.009	-0.071
	(0.040)	(0.039)	(0.078)	(0.069)	(0.074)
Kuca	**0.110**	**0.102**	0.061	-0.023	-0.011
	(0.028)	(0.028)	(0.061)	(0.062)	(0.060)
Oženjen	**-0.242**	**-0.233**	**0.187**	**0.344**	**0.253**
	(0.030)	(0.030)	(0.068)	(0.066)	(0.066)
Broj zahteva	-0.007	-0.004	**-0.116**	**-0.097**	**-0.104**
	(0.005)	(0.005)	(0.012)	(0.011)	(0.010)
Preduzece	**0.570**	**0.570**	0.135	0.162	0.147
	(0.064)	(0.064)	(0.181)	(0.160)	(0.157)
Prihod	**0.901**	**0.886**	0.038	**-0.286**	**-0.226**
	(0.018)	(0.018)	(0.042)	(0.050)	(0.050)
Razlika prihoda	**-0.243**	**-0.237**	**0.147**	**0.185**	**0.206**
	(0.035)	(0.035)	(0.075)	(0.079)	(0.073)
Kapital	**-0.284**	**-0.272**	-0.057	**0.195**	0.142
	(0.051)	(0.050)	(0.123)	(0.100)	(0.127)
Nula limit	**-2.253**	**-2.218**	**-2.244**	**-0.328**	**-0.703**
	(0.106)	(0.114)	(0.401)	(0.140)	(0.306)
Limit	**-0.861**	**-0.848**	0.006	**0.561**	**0.486**
	(0.019)	(0.021)	(0.059)	(0.050)	(0.056)
Broj kredita	**0.086**	**0.086**	**0.329**	**0.259**	**0.270**
	(0.007)	(0.007)	(0.026)	(0.023)	(0.020)
Upotrebljen limit	**-0.747**	**-0.759**	**-1.295**	**-1.223**	**-1.210**
	(0.045)	(0.045)	(0.120)	(0.116)	(0.093)
Iznos kredita	-	-	-0.069	-0.070	-0.067
			(0.073)	(0.070)	(0.069)
s^2	-	-	**0.919**	**0.196**	-
			(0.045)	(0.057)	
ρ	-	-	-	**-0.986**	**-0.911**
				(0.021)	(0.056)

U *Tabeli 4* vidimo da osobe sa vecim prihodom imaju veci rizik da ce postati loši od osoba sa manjim prihodom. Ostale promenljive koje imaju koeficijente od znacaja u ovom primeru su «*Broj zahteva*», «*Nula limit*», «*Broj kredita*», «*Limit*» i «*Upotrebljeni limit*». Broj zahteva za informacije o klijentu koje kreditni biro primi je samo znak koliko je klijentu potreban novi kredit, a to samim tim negativno utice na preživljavanje tog kredita, jer je to rani znak da je klijent u problemu. Ukoliko klijent do sada nema kredita, što je predstavljeno promenljivom «*Nula limit*», je znak da je klijent neiskusan u otplati kredita, što može dovesti do toga da brzo ode u default, pa samim tim negativno utice na vreme preživljavanja. Za klijente koji vec imaju kredit se pretpostavlja da vec imaju dovoljno iskustva u servisiranju duga, pa samim tim «*Broj kredita*» i «*Limit*» pozitivno uticu na preživljavanje kredita. Promenljiva «*Upotrebljeni limit*» oznacava u kom opsegu klijent koristi odobreni kredit, a samim tim i utice na smanjenje preživljavanja.

Ako posmatramo poslednju kolonu *Tabele 4* vidimo da probit parametri koji odreduju verovatnocu da klijent biti loš (α_2) imaju isti znak, sem u slucaju promenljive «*Kapital*», kao parametri β_2 u tobit modelu koji ocenjuje vreme preživljavanja. Ipak, promenljive kao što su «*Muško*», «*Razveden*», «*Kuca*», «*Veliki grad*» i «*Preduzece*», koje su znacajne u donošenju odluke nemaju uticaja pri analizi rizika da ce klijent biti loš niti pri analizi vremena preživljavanja, dok promenljiva «*Broj kredita*» ima uticaja na preživljavanje kredita, a ne igra nikakvu ulogu pri odlucivanju da li treba odobriti kredit. Pošto svi sem jednog parametra imaju isti znak u jednacini preživljavanja kao odgovarajuci parametri u binarnom probit modelu, ove promenljive u isto vreme smanjuju, odnosno povecavaju ocekivano vreme preživljavanja i stopu prinosa od kredita. Drugim recima, ako banka ne teži da minimizira rizik da ce klijent biti loš, razlog tome je taj što krediti sa vecim rizikom imaju duže vreme preživljavanja, a samim tim se ocekuju i veci prinosi.

Ovi rezultati su potvrdili da banka ne teži da minimizira rizik. Neke promenljive koje povecavaju (smanjuju) verovatnocu da klijentu bude odobren kredit u isto vreme smanjuju (povecavaju) ocekivano vreme preživljavanja (a samim tim i prinos) i povecavaju (smanjuju) verovatnocu da ce klijent biti loš. Takode, pokazalo se da na odluku banke ne utice iznos kredita. Ovo nas vodi do zakljucka da banka ne teži ni ka tome da maksimizira profit, a samim tim ni vreme preživljavanja kredita. Ali sigurno je da banka u svojoj politici ima definisan cilj kome teži, ka maksimizaciji ili minimizaciji nekih drugih parametara, kao što su na primer broj klijenata, profit od odredene vrste proizvoda, ili slicno.

4.4 Primena logit modela

Kako se bankarsko tržište sve više širi, ne postoji konkurencija samo medu domacim bankama, nego je ona sve jaca i izmedu stranih banka. Sve je veca tražnja za potrošackim kreditima, pa samim tim je neophodno usmeriti pažnju na upravljenje rizikom, što nas vodi do razvijanja dobrog kreditnog skoring modela kojim se ocenjuje kreditna sposobnost klijenta, kolika je verovatnoca da ce isti redovno otplatiti.

Kreditni skoring sistem koristi istorijske podatke o kreditima i podatke o klijentu da odredi koje klijentove karakteristike najbolje razdvajaju dobre od loših klijenata. Kada se napravi jedan ovakav model, on se zatim može primeniti na nove klijente za koje nije poznato sa kojom verovatnocom ce postati loši *PD* (*probability of default*). Na osnovu ocenjenog kreditnog skoring modela mogu da se izracunaju skorovi za svakog novog klijenta, pri cemu viši skor oznacava manji PD. Ovaj skor zatim treba da se uporedi sa granicnom vrednošcu kreditnog skoring sistema koja odreduje koji zahtev za kredit ce biti odobren, a koji odbijen, a koju odreduje sama banka u zavisnosti od toga koliki je rizik spremna da prihvati.

Kao što smo vec videli postoje razni metodi koji mogu da se koriste za razvijanje kreditnog skoring modela, kao što su na primer probit analiza, diskriminantna analiza, linearna regresija i slicno, medutim ovde se koristi logaritamska regresija, tj. *logit model* primenjen na uzorak klijenata koji cine Vijetnamsko finansijsko tržište («*A credit scoring model for Vietnam's retail banking market*» [5]).

Da bi mogao da se primeni ovaj model prvo treba definisati koje ce karakteristike da udu u model. Ne postoji nikakvo pisano pravilo o broju i tipu promenljivih koje ce uci u model. Uglavnom se tu nalaze karakteristike koje pokazuju finansijske mogucnosti klijenta kao što se prihodi i vrednost kuce koja je u njegovom vlasništvu, a takode i promenljive koje to indirektno pokazuju, kao što su obrazovanje, zatim broj nekretnina, broj godina na istom poslu i slicno. Treba obratiti pažnju i na zakon koji u nekim zemljama zabranjuje da se neke karakteristike, kao što su pol, religija i slicno, ukljuce u model zbog mogucnosti diskriminacije.

Nakon što se ocene koeficijenti β_i uz pomoc metoda ocene maksimalne verovatnoce treba testirati model i odrediti sa kolikom preciznošcu je ocenjen model. Ovaj proces se naziva ***validacija***. Uglavnom se uzima uzorak koji nije korišcen za izvodenje modela, *out-of-sample*. Prvo se ocenjuju verovatnoce da ce klijent biti loš za ovaj uzorak. Ovi PD-evi se upoređuju sa granicnom vrednošcu da bi odredili da li ce podnosilac zahteva biti dobar ili loš klijent. Granicnu vrednost odreduje sama banka, na osnovu toga koliki rizik je spremna da prihvati. Ako na primer izaberemo da je granicna vrednost 50%, tada ce klijent ciji je ocenjeni PD veci (manji) od 50% biti klasifikovan kao loš (dobar). U ovoj fazi se koristi ***matrica klasifikacije***, gde G_g predstavlja broj tacno klasifikovanih dobrih kredita, a G_b broj dobrih kredita koji su pogrešno klasifikovani. Slicno, B_b predstavlja broj dobro klasifikovanih loših kredita, a B_g broj pogrešno klasifikovanih loših kredita. Procenat tacno klasifikovanih kredita (PCC) služi kao mera preciznosti. Procenat tacno klasifikovanih dobrih kredita

(PCC_{good}) je definisan kao odnos tacno klasifikovanih dobrih kredita i ukupnog broja posmatranih kredita. Slicno, procenat tacno klasifikovanih loših kredita (PCC_{bad}) je definisan kao odnos tacno klasifikovanih loših kredita i ukupnog broja posmatranih kredita. I na kraju, procenat tacno klasifikovanih kredita (PCC_{total}) je definisan kao odnos tacno klasifikovanih kredita i ukupnog broja kredita.

Matrica klasifikacije

Stvarna opažanja	**Ocekivana opažanja**		**PCC**
	dobri	**loši**	
dobri	G_g	G_b	$PCC_{good}=G_g/(G_g+G_b)$
loši	B_g	B_b	$PCC_{bad}=B_b/(B_b+B_g)$
ukupno			$PCC_{total}=(G_g+B_g)/(G_g+G_b+B_b+B_g)$

Banka možda želi da minimizira obe greške, B_g i G_b. Medutim, smanjivanje B_g dovodi do povecanja G_b i suprotno. Iz tog razloga treba razmotriti kakav je odnos u troškovima (gubicima) koji nastaju pri pogrešnoj klasifikaciji i na osnovu tog kriterijuma odluciti koju grešku treba minimizirati.

Sada cemo se upoznati sa primerom vijetnamskog bankarskog tržišta koje je i dalje dosta nerazvijeno i još uvek u mnogome zaostaje za razvijenim industrijskim zemljama. Bankarski sistem i dalje pati od nedostatka kapitala, adekvatne zaštite od mogucih gubitaka, niske profitabilnosti. Ali on se konstantno razvija.

Da bi razvili kreditni skoring sistem uzeti su podaci o kreditima iz jedne Vijetnamske komercijalne banke u periodu izmedu 1992. i 2005. godine. Dobijamo uzorak od 56,037 kredita. Banka klasifikuje klijenta kao lošeg ukoliko je u kašnjenju dužem od 90 dana. Naš uzorak sadrži 3.3% loših klijenata. Pošto uzorak sadrži samo informacije o klijentima kojima je kredit odobren, on nije dovoljno reprezentativan za buduce klijente. Takode, ne možemo da predvidimo kako bi se odbijeni klijenti ponašali da su odobreni. Jedno od rešenja je out-of-sample kalibracija, pri cemu uzorak od 56,037 kradita delimo na dva poduzorka. Pocetni uzorak sadrži 30,994 kredita od cega su 1026 (3.3%) loši krediti i *hold-out* uzorak koji sadrži 25,043 kredita od cega su 798 (3.2%) loši krediti.

Da bi se ocenio kreditni skoring sistem, koristi se *stepenasta metoda unapred*. Ova metoda pocinje sa modelom koji ne sadrži nijednu nezavisnu promenljivu i postepeno dodaje promenljive. U svakom koraku, dodaje po jedna promenljiva koja ima najvecu moc u poboljšanju tacnosti ocenjivanja, i pri tome se posmatra nivo znacajnosti koji je manji od 5%. Proces se zaustavlja kada se više ne može dodati nijedna promenljiva sa nivoom znacajnosti manjim od 5%. Na ovaj nacin smo od pocetne 22 promenljive, skup sveli na 16 nezavisnih promenljivih. Da bi se osigurali da su ovo stvarno najznacajnije promenljive, koristi se stepenasti metod unazad. Ovaj model u pocetku sadrži 22 promenljive, i u svakom koraku izbacuje po jednu koja ima najslabiji uticaj i ovako dobijamo konacnih 16 promenljivih kao i stepenastom metodom unapred.

Promenljive koje su korišcene u modelu su sledece:

- *Obrazovanje* – možemo da ocekujemo da su obrazovaniji ljudi mnogo stabilniji, imaju vece prihode i samim tim niži PD.
- *Pol* – u nekim zemljama ova promenljiva mora da se iskljuci iz modela, zbog mogucnosti diskriminacije.
- *Region* – predstavlja deo zemlje u kojoj klijent živi. Smatra se da ljudi sa slicnim bogatstvom žive na istoj lokaciji. Dakle, geografski kriterijum može da pokazuje nivo finansijskog bogatstva klijenta
- *Vreme na trenutnoj adresi* – predstavlja broj godina koliko klijent živi na istoj adresi. Ova promenljiva može da oznacava klijentovu zrelost, stabilnost i izbegavanje rizika. Medutim, u Vijetnamu PD se povecava sa brojem godina na trenutnoj adresi, pošto ljudi što više zaraduju i napreduju, teže da žive u što boljim uslovima, pa samim tim i cesto menjaju adresu stanovanja.
- *Stambeno pitanje* – pokazuje da li je klijent vlasnik svog doma, iznajmljuje stan, ili živi sa roditeljima.
- *Bracno stanje* – u našem uzorku verovatnoca da ce klijent biti loš je veca kod ljudi koji su u braku nego za samce.
- *Broj ljudi koje klijent izdržava* – što je veci broj ovih zavisnika, to se i PD povecava, jer je vece optercenje na klijentove prihode.
- *Telefon* – meri da li klijent ima kucni telefon, a time i koliko lako banka može da održava kontakt sa klijentom. Ukoliko klijent nema telefon, to mu je viši PD.
- *Svrha kredita* – opisuje kako ce odobrena sredstva biti iskorišcena
- *Tip kolaterala* – opisuje koji tip obezbedenja pokriva kredit.
- *Vrednost kolaterala*
- *Trajanje kredita* – u Vijetnamu trajanje kredita predlaže klijent, pa samim tim i ova promenljiva pokazuje koliko je klijent spreman da prihvati rizik, kao i samoocenjivanje sposobnosti otplate kredita.
- *Vreme u banci* – meri koliko godina klijent posluje preko te banke.
- *Broj kredita* – racuna broj kredita koliko je klijent primio od kada je u banci
- *Tekuci racun* – je binarna promenljiva koja predstavlja da li klijent održava tekuci racun
- *Štedni racun* - je binarna promenljiva koja predstavlja da li klijent održava štedni racun

Tabela 1: *ocenjeni kreditni skoring model*

Promenljive	***Ocenjeni koeficijenti***	***Standardne greške***	***Nivo znacajnosti***
Vreme u banci	-1.774	0.121	0.0%
Pol	-1.557	0.222	1.0%
Broj kredita	-0.938	0.051	1.4%
Trajanje kredita	-0.845	0.080	3.7%
Štedni racun	-0.750	0.104	3.1%
Region	-0.652	0.030	13.6%
Stambeno pitanje	-0.551	0.278	44.6%
Tekuci racun	-0.492	0.208	10.4%
Vrednost kolaterala	-0.402	0.096	9.8%
Broj zavisnika	-0.356	0.096	9.9%
Vreme na trenutnoj adresi	-0.285	0.054	2.5%
Bracno stanje	-0.233	0.101	68.1%
Tip kolaterala	-0.190	0.057	53.0%
Telefon	-0.181	0.047	3.4%
Obrazovanje	-0.156	0.067	60.3%
Svrha kredita	-0.125	0.054	3.3%
Konstanta	-3.176	0.058	4.6%

Kao što vidimo u ovoj tabeli, od datih 16 promenljivih, vreme u banci je najznacajniji pokazatelj, a zatim slede pol, broj kredita i trajanje kredita. Negativni koeficijent koji stoji uz promenljivu pol znaci da žene imaju manju verovatnocu da ce kasniti sa placanjem od muškaraca. Trajanje kredita kao mera klijentove spremnosti prihvatanja rizika i samoocenjivanja je jedinstveno u Vijetnamu. Ovo pokazuje da banke ne mogu da se pouzdaju samo u sopstvene procene nego se oslanjaju i na klijentovu poštenu ocenu trenutnog stanja.

Da bi pravilno ocenili preciznost modela koristimo out-of-sample. Tada je srednja vrednost predvidjene verovatnoce da ce klijent biti loš jednaka 1.73% za dobre klijente za razliku od 49.05% za loše klijente. Za loše klijente PD se krece u opsegu od 0.01% do 96.81%, a za dobre od 0.00% do 73.54%. U sledecoj tabeli su date vrednosti koje ukazuju na preciznost predvidanja modela.

Tabela 2: *preciznost predvidanja sa granicnom vrednošcu 0.50*

Uoceno	**Predvidene vrednosti**			
	Dobri	**Loši**	**PCC**	
Dobri	24,136	109	99.55%	= PCC*good*
Loši	397	401	50.25%	= PCC*bad*
Ukupno			97.98%	= PCC*total*

4.5 Primer genetskog programiranja

Genetsko programiranje je proširenje tehnike genetskih algoritama. Mi necemo dublje ulaziti u ovaj problem, jer to prevazilazi okvire ovog rada, nego cemo samo navesti primer (uzet iz rada «*Genetic programming for credit scoring: the case of Egyptian public sector banks*» [7]) u kome cemo uporediti rezultate koji se dobijaju genetskim programiranjem i probit analizom. Koristimo dva tipa modela genetskog programiranja: *prost model* i *tim model* koji je kombinacija više prostih modela ciji je cilj da se postignu bolji rezultati nego kad se primenjuje pojedinacni prost model.

Za potrebe pravljenja modela korišcena je baza podataka o potrošackim kreditima dostavljena od strane Egipatskog bankarskog sektora. Baza podataka se sastoji od 1262 kredita, od cega je 851 kredit (67.43%) dobar, a 411 (32.57%) loš.

Prvo se pravi model na celom uzorku, a zatim se za potrebe testiranja tacnosti ocenjivanja skoring sistema koristi *training uzorak* koji sadrži 846 (67%) slucajeva, a testiranje modela se vrši na *uzorku za testiranje* koji sadrži 416 (33%) slucaj.

Za uporedivanje rezultata uzeta su u obzir cetiri kriterijuma: procenat tacne klasifikacije (*PCC*), kriterijum ocenjenih troškova koji nastaju pogrešnom klasifikacijom (*OT*) sa koeficijentom troškova pogrešne klasifikacije 5:1, kriterijum *OT* sa koeficijentom troškova pogrešne klasifikacije 7:1, kriterijum *OT* sa koeficijentom troškova pogrešne klasifikacije 10:1.

Sa pojmom procenta pogrešne klasifikacije i matricom klasifikacije smo se upoznali u delu 4.3.2, a sada cemo dati osnovu ideju kriterijuma ocenjenih troškova koji nastaju pogrešnom klasifikacijom. Ovaj kriterijum daje ocenu efikasnosti skoring sistema. Za izracunavanje *OT*-a koristi se sledeca jednacina:

$$OT=C(I)x(G_b/TG)x(TG/TN) + C(II)x(B_g/TB)x(TB/TN)$$

gde je *C(I)* trošak nastao pogrešnom klasifikacijom koja je povezana sa greškom prvog reda; (G_b/TG) je verovatnoca da ce se napraviti greška prvog reda, predstavljena kao odnos broja dobrih klijenata koji su ocenjeni kao loši (G_b) i ukupnog broja dobrih klijenata; *TG/TN* je verovatnoca da je klijent dobar, tj. odnos broja dobrih *TG* i ukupnog broja klijenata *TN*; *C(II)* je trošak nastao pogrešnom klasifikacijom koja je povezana sa greškom drugog reda; (B_g/TB) je verovatnoca da ce se napraviti greška drugog reda, predstavljena kao odnos broja loših klijenata koji su ocenjeni kao dobri (B_g) i ukupnog broja loših klijenata; *TB/TN* je verovatnoca da je klijent loš, tj. odnos broja loših *TB* i ukupnog broja klijenata *TN*.

Gornja jednacina može da se napiše i na sledeci nacin:

$$OT=C(I)x(G_b/TN) + C(II)x(B_g/TN)$$

4.5.1 Uporedivanje genetskog programiranja i probit analize na celom uzorku

Tabela 1 sumira stope tacne klasifikacije i ocenjene troškove koji nastaju pogrešnom klasifikacijom, i to primenom genetskog programiranja i probit analize. Vidimo da GP_t (tim model genetskog programiranja) ima najvecu stopu tacne klasifikacije. GP_p (prost model genetskog programiranja) je model koji najbolje klasifikuje dobre klijente (91.89%), a GP_t je model koji najbolje klasifikuje loše klijente (74.94%). GP_t je izabrani model pošto ima najnižu ocenu troškova pogrešne klasifikacije, a takode ima i najbolju stopu tacne klasifikacije.

***Tabela 1**: uporedivanja rezultata klasifikacije, grešaka i ocenjenih troškova na celom uzorku*

Model	**Tacno klasifikovani rezultati**			**Rezultati o greškama**		**OT**	**OT**	**OT**
	G%	**B%**	**PCC**	**I tipa**	**II tipa**	**(5:1)**	**(7:1)**	**(10:1)**
Ceo uzorak								
PA	88.95	67.40	81.93	0.1105	0.3260	0.6054	0.8178	1.1363
GP_p	91.89	65.45	83.28	0.0811	0.3455	0.6173	0.8424	1.1800
GP_t	91.07	74.94	**85.82**	0.0893	0.2506	**0.4683**	**0.6316**	**0.8764**

* G – dobri klijenti; B – loši klijenti

4.5.2 Uporedivanje genetskog programiranja i probit analize na poduzorku

Koristi se uzorak za testiranje da bi se testirala moc ocenjivanja razvijenog skoring modela. Vidimo da GP_t ima najvišu stopu tacne klasifikacije. Takode, GP_t model pravi najvecu grešku pri klasifikaciji dobrih klijenata, a GP_p pri klasifikaciji loših klijenata. U ovom slucaju se nije lako odluciti koji model je najbolji , pošto na primer GP_t ima najvišu stopu tacne klasifikacije ali u isto vreme i dosta visoke troškove pogrešne klasifikacije. U obzir bi mogao doci GP_p model pošto ima podnošljivo male troškove pogrešne klasifikacije i dosta visoku stopu tacne klasifikacije.

***Tabela 2** : uporedivanja rezultata klasifikacije, grešaka i ocenjenih troškova na poduzorku*

Model	**Tacno klasifikovani rezultati**			**Rezultati o greškama**		**OT**	**OT**	**OT**
	G%	**B%**	**PCC%**	**I tipa**	**II tipa**	**(5:1)**	**(7:1)**	**(10:1)**
Poduzorak								
PA	90.00	65.87	82.69	0.1000	0.3413	0.6232	0.8456	1.1790
GP_p	88.62	69.84	82.93	0.1138	0.3016	**0.5679**	**0.7644**	**1.0590**
GP_t	94.48	59.52	**83.89**	0.0552	0.4048	0.6964	0.9601	1.3557

* G – dobri klijenti; B – loši klijenti

4.6 Primena analize obavijanja podataka (DEA)

U dosadašnjim primerima smo videli kako se kreditni skoring model primenjuje kada se analizira kreditna sposobnost pojedinacnog klijenta. U ovom poglavlju cemo se malo upoznati i sa primerom kreditnog skoring sistema u situaciji kada firma podnosi zahtev za kredit. Kao što smo videli, u slucaju individue posmatraju se karakteristike kao što su starost klijenta, njegovo finansijsko stanje i obaveze i slicno. U slucaju firmi, bitni su nam podaci kao što o njihovoj aktivi, pasivi, rashodima i prihodima, i na osnovu tih podataka dolazimo do zakljucka o likvidnosti te firme, njihovoj profitabilnosti, produktivnosti, kao i o strukturi troškova. U bankama još uvek nije zastupljen ovaj tip kreditnog skoringa, pošto je on i dalje u fazi razvoja, ali samo je pitanje trenutka kada ce kreditni skoring poceti da se primenjuje i za analizu kreditne sposobnosti firmi. U ovom radu cemo predstaviti kako se pitem analize obavijanja podataka dobijaju skorovi koji nam kasnije služe za rangiranje firmi.

DEA (Data Envelopment Analysis), tj. analiza obavijanja podataka je metoda linearnog programiranja za ocenu relativne efikasnosti organizacionih jedinica koje koriste više razlicitih inputa za stvaranje više razlicitih outputa i ona za razliku od predašnjih modela kojima trebaju istorijski podaci da bi napravili model, zahteva samo *observed* (uoceni) skup ulaznih i izlaznih podataka da bi izracunala kreditni skor. DEA analiza je neparametarski metod operacionih istraživanja koji se koristi u ekonomiji za ocenjivanje graniicne proizvodnje.

Proces se sastoji od šest koraka, od cega se prva tri bave odabirom firmi koje ce uci u model i indikatora koji se koriste za ocenu finansijske situacije u toj firmi, cetvrti korak koristi DEA da bi se izracunali kreditni skorovi firmi, peti korak proverava valjanost ovako dobijenih skorova tako što ih poredi sa onima koji su dobijeni putem regresije i diskriminantne analize, i konacno šesti korak predlaže metod za kreditni rejting uz pomoc raspodele odnosa dobrih i loših klijenata.

Uzimaju se u obzir firme koje podnose zahtev za novi kredit ili one koje hoce da promene vec postojeci kreditni limit. Na pocetku istraživanja posmatrano je 1400 firmi, ali *outlajeri*, tj. firme koje su imale koeficijente koji su znacajno odstupali (više od dve standardne devijacije) od odgovarajucih srednjih vrednosti su iskljucene iz modela i ostala je 1061 firma.

4.6.1 Odabir finansijskih koeficijenata

Za definisanje finansijskih koeficijenata uglavnom se koriste opšte prihvacene finansijske dimenzije kao što su rast prihoda, likvidnost, profitabilnost, produktivnost i struktura troškova. Tako da se razliciti finansijski koeficijenti grupišu po ovim dimenzijama. U ovom primeru («*A practical approach to credit scoring*» [3]) je odabrano šest koeficijenata koji su klasifikovani kao ulazne i izlazne promenljive za DEA.

Ulazni podaci, tj. koeficijenti koji treba da se minimiziraju su *FE*-finansijski troškovi u odnosu na prodaju, *CL*-koeficijent obaveza (trenutne obaveze u odnosu na kapital) i *TB*-ukupno zaduženje u odnosu na ukupnu aktivu. Koeficijent finansijskih troškova u odnosu na prodaju (FE) pokazuje sposobnost firme da plati svoje troškove,

što pokazuje kreditnu sposobnost firme. Koeficijent obaveza (CL) je indikator stabilnosti strukture kapitala. Povecanje ovog koeficijenta pokazuje nestabilnost strukture kapitala i likvidnosti firmi. Ukoliko se koeficijent ukupne zaduženosti (TB) povecava to pokazuje da se profitabilnost i stabilnost firme smanjuju.

Izlazni podaci koji treba da se maksimiziraju su koeficijent adekvatnosti kapitala (*CA*) koji predstavlja odnos kapitala firme i ukupne aktive, zatim pokazatelj trenutne likvidnosti (*CR*) i koeficijent pokrivenosti kamata (*IC*). Što više firme pokriva obaveze iz sopstvenih sredstava, tj. što je viši CA, to je ona ocenjena kao manje rizicna. Takode, što je veci CR to je firma više likvidna i manja je verovatnoca da ce upasti u probleme. Koeficijent pokrivenosti kamata pokazuje sposobnost firme da placa troškove kamata iz operativnog prihoda, a samim tim što je ovaj koeficijent veci to je veca profitabilnost firme.

4.6.2 Racunanje skorova uz pomoc DEA

Kada smo definisali ulazne i izlazne promenljive, možemo da primenimo *CCR model* da bi izracunali DEA skorove:

$$\min q_k - e\sum_{r=1}^{s} s_r^+ - e\sum_{i=1}^{m} s_i^-$$

$$\text{uz uslove } \sum_{j=1}^{N} l_j x_{ij} = q_k x_{ik} - s_i^-; i = 1,...,m$$

$$\sum_{j=1}^{N} l_j y_{ij} = y_{rk} + s_r^+; r = 1,...,s$$

$$l_j, s_r^+, s_i^- \geq 0; \forall j, r, i, q_k$$

gde su

θ_k – skor koji pokazuje kreditne sposobnost za firmu *k*
N – broj firmi u uzorku
λ_j – težišna vrednost firme *j*
y_{rj} – *r*-ti izlazni koeficijent za firmu *j*
y_{rk} – *r*-ti izlazni koeficijent za firmu *k*
x_{ij} – *i*-ti ulazni koeficijent za firmu *j*
x_{ik} – *i*-ti ulazni koeficijent za firmu *k*
s_r^+, s_i^- - dopunske promenljive za *r*-to ogranicenje i *i*-to ogranicenje

DEA skorovi su dati u obliku procenta, pri cemu firme sa DEA skorom od 100 predstavljaju najbolje firme i one pripadaju «DEA efikasnoj granici».

4.6.3 Provera valjanosti DEA skorova

Cilj ovog koraka je da oceni u kom procentu se DEA skorovi poklapaju sa skorovima koji se dobijaju putem regresione ili diskriminantne analize.

Linearna regresija se koristi kao kriterijum za testiranje objašnjavajuce moci indikator promenljivih u DEA. Za ovu svrhu, DEA skorovi se uzimaju kao zavisne, a DEA koeficijenti kao nezavisne promenljive.

Tabela 1: *rezultati regresione analize*

R^2 = 0.741; F = 491.803				
	Koeficijenti	**Standardne greške**	***t*-vrednost**	***p*-vrednost**
Konstanta	72.94035	3.772205	19.33626	0.0000
FE	-109.653	21.6798	-5.05785	0.0000
CL	-50.8614	3.123891	-16.2814	0.0000
TB	-56.3548	3.039185	-18.5427	0.0000
CA	47.45822	4.30213	11.03133	0.0000
CR	17.36333	1.085729	15.99233	0.0000
IC	1.146138	0.184531	6.211091	0.0000

Kao što se vidi u Tabeli 1 znaci koeficijenata koji stoje uz promenljive su ocekivani (tj. minus uz promenljive koje negativno uticu i plus uz one koje pozitivno uticu) i sve su statisticki znacajne (p – vrednost je 0.0000), što pokazuje da je DEA algoritam uspešno izracunao vrednosti za ovih šest koeficijenata. Kada vrednosti koje smo dobili putem regresije uvrstimo u jednacinu, imamo:

$$DEA = 73 – 109.7\ FE – 50.9\ CL – 56.4\ TB + 47.5\ CA + 17.4\ CR + 1.2\ IC$$

Koristeci ovu jednacinu može da se izracuna linearna aproksimacija DEA skora, tako da kada se pojavi novi klijent ne mora da se ponavlja ceo DEA algoritam.

Diskriminanatna analiza se koristi da utvrdi koliko dobro DEA skorovi klasifikuju firme u dve grupe: dobre i loše. Firme su podeljene u dve grupe na osnovu DEA skorova. Granicna vrednost izmedu ovih grupa se odreduje uzimajuci u obzir raspodelu DEA skorova. U ovom primeru za granicnu vrednost je uzeta srednja vrednost, dakle, 518 firme su klasifikovane kao dobre, a ostatak kao loše. Diskriminantna analiza izvodi funkciju diskriminacije koja ukuljucuje pet od šest koeficijenata kao nezavisne promenljive (koeficijent IC je iskljucen).

Table 2: *rezultati diskriminantne analize*

Prognozirana grupa	Odabrana grupa		
	dobri	**loši**	**ukupno**
dobri	400 (84.9%)	57 (11.0%)	497
loši	78 (15.1%)	461 (89.0%)	539
ukupno	518	518	1036

Tabela 2 pokazuje da je ukupno dobro klasifikovano 87.0% populacije ((440+461)/(518+518) = 0.87). Dakle, greška klasifikacije je 13%.

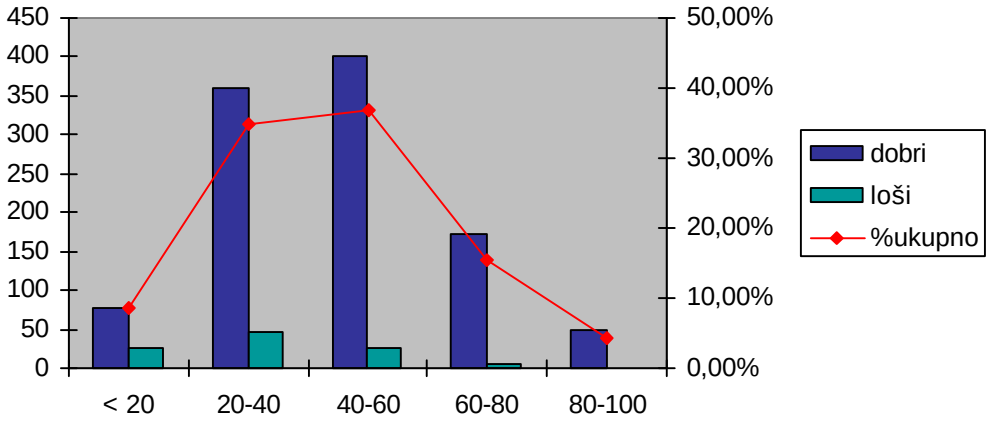
4.6.4 Metod kreditnog rejtinga

DEA skorovi klijenta bi trebalo da budu klasifikovani u klase kao što su A, B, C i tako dalje. Pošto se DEA skorovi krecu u intervalu od 0 do 100, može se napraviti podela da skorovi od 0 do 20 pripadaju jednoj klasi, od 20 do 40 drugoj klasi i tako dalje.

***Tabela 3**: raspodela ucestalosti dobrih i loših firmi*

Klasa	***dobri***	***loši***	***Ukupno***	***% loših***	***% ukupno***
Ispod 20	76	25	101	24.8%	8.7%
20-40	360	46	406	11.3%	34.9%
40-60	402	27	429	6.3%	36.9%
60-80	173	5	178	2.8%	15.3%
80-100	50	0	50	0.0%	4.3%
Ukupno	1061	103	1164	8.8%	

Rezultati se mogu prikazati i graficki, tako što je stubicima prikazan broj dobrih i loših klijenata, a kriva prikazuje raspodelu DEA skorova.

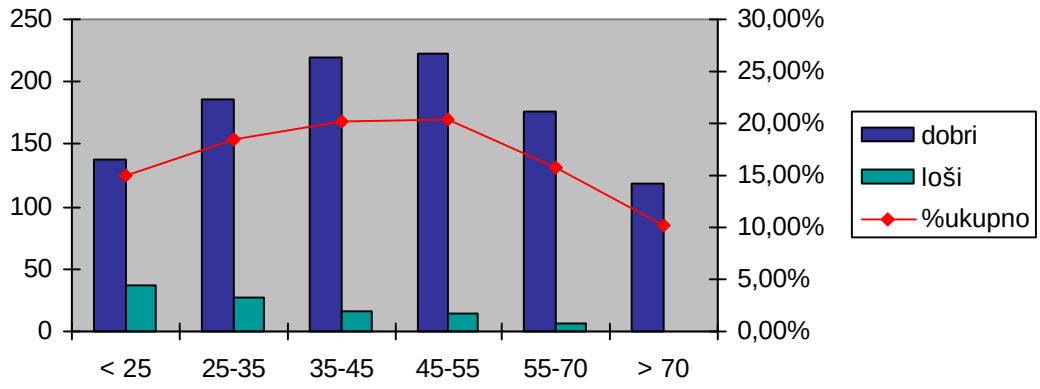


Kao što se vidi u ovoj tabeli, ucestalost loših klijenata (%loših) opada kako DEA skorovi rastu. To znaci da ovi skorovi služe kao veoma korisni pokazatelji za rangiranje klijenata po njihovoj kreditnoj sposobnosti. Tako na primer, klasa 80-100 može da bude oznacena kao «A», klasa 60-80 kao «B» i tako dalje. Tabela 3 (kao i grafik) pokazuje da je rapodela DEA skorova (%ukupno) iskošena na desno, što ukazuje na to da mnogo firmi ima niske skorove. A vecina komercijalnih banaka teže ka tome da rangiraju klijente tako da dobiju normalnu raspodelu. Ovaj cilj se postiže modifikacijom klasa DEA skorova. U Tabeli 4 su prikazane modifikovane klase kojima se postiže normalna raspodela.

Tabela 4: *modifikovana raspodela ucestalosti dobrih i loših klijenata*

Klasa	***dobri***	***loši***	***Ukupno***	***% loših***	***% ukupno***
Ispod 25	138	37	175	21.1%	15.0%
25-35	186	28	214	13.1%	18.4%
35-45	219	16	235	6.8%	20.2%
45-55	222	15	237	6.3%	20.4%
55-70	177	7	184	3.8%	15.8%
Iznad 70	119	0	119	0.0%	10.2%
Ukupno	1061	103	1164	8.8%	

Kada rezultate predstavimo graficki vidimo da je raspodela DEA skorova stvarno normalna.



5. Zakljucak

Kreditni skoring sistem ne daje objašnjenje zašto je neki klijent odbijen, nego samo daje ocenu koliko je neki klijent rizican i upozorava da bi tog klijenta trebalo odbiti. Medutim, ko nereskira taj neprofitira. Zato je na banci da odluci koliki rizik želi da prihvati, a rezultati kreditnog skoringa služe kao odredene smernice.

Kao što smo videli kroz rad, postoje dva osnovna pristupa kredit skoringu. Prvi je pomocu separacije, a drugi je putem regresije. Cilj nam je da što tacnije klasifikujemo klijente na dobre i loše, jer nam pogrešna klasifikacija donosi i vece troškove koji tom prilikom nastaju. Ukoliko nekog klijenta koji je dobar klasifikujemo kao loš, samim tim ce taj klijent biti odbijen i time gubimo profit koji smo mogli da ostvarimo da smo mu odobrili kredit. Medutim, mnogo je veca greška da klijenta koji je loš klasifikujemo kao dobrog i odobrimo mu kredit, jer time nastaju odredeni gubici kada klijent više ne bude u mogucnosti da otplacuje kredit. Zato nam je vrlo bitno da ove greške svedemo na minimum. Medutim, ni to nije tako jednostavno. Smanjenje greške koja nastaje pogrešnom klasifikacijom loših klijenata dovodi do povecanja greške pogrešne klasifikacije dobrih klijenata i obrnuto. Iz tog razloga treba razmotriti kakav je odnos u gubicima koji nastaju pri pogrešnoj klasifikaciji i na osnovu tog kriterijuma odluciti koju grešku treba minimizirati.

U procesu minimizacije troškova može da nam pomogne metod koji se sastoji iz dve faze odlucivanja. Kada banka dobije zahtev za kredit, ona donosi preliminarnu odluku o tome da li ce klijenta svrstati u grupu dobrih ili grupu loših klijenata. Zatim, na osnovu procene koristi i troškova donosi odluku da li ce tražiti dodatne informacije o klijentu i na osnovu njih doneti konacnu odluku. Na osnovu toga koliko se stopa preciznosti povecala u drugoj fazi i koliki su troškovi, banka odlucuje da li joj se isplati da izvede drugu fazu ili da donese odluku na osnovu prve faze ciji su troškovi zanemarljivi i koju banka uvek izvodi da bi ocenila kreditnu sposobnost klijenta.

U današnje vreme, u vecini banaka, sve je veca upotreba modela baziranih na regresiji, i to su najcešce u primeni logit modeli. Iako deluju vrlo jednostavno i razumljivo, da bi se napravio dobar i precizan model mora mu se posvetiti puno vremena i pažnje. Vrlo je bitno dobro izabrati promenljive koje ce uci u model. Polazi se od velikog skupa promenljivih, medutim taj broj se na kraju drasticno smanjuje, najcešce zbog jake korelacije medu promenljivama, zatim zbog toga što neke promenljive predstavljaju istu ili slicnu stvar ali ipak jedna od njih ima veci znacaj. Dati primeri nam daju vrlo realne i logicne rezultate. Medutim, ima nekih stvari koje nas navode na razmišljanje. Tako na primer, posmatrajuci Tabelu 3 u odeljku 4.3 vidimo da vecina promenljivih koje pozitivno uticu na odobravanje kredita nisu medu onima koje smanjuju rizik da ce klijent biti loš, što nas navodi na zakljucak da banka ne teži ka tome da minimizira rizik. Kada uzmemo u razmatranje i vreme preživljavanja kredita, vidimo da neke promenljive koje povecavaju verovatnocu da klijentu bude odobren kredit, u isto vreme smanjuju ocekivano vreme preživljavanja kredita i povecavaju verovatnocu da ce klijent biti loš. Tako na primer, osobe sa vecim prihodom imaju veci rizik da ce postati loše od osoba sa manjim prihodom, dok je verovatnoca da im bude odobren kredit veca. Treba biti vrlo obazriv kod ovakvih

situacija i ukoliko je moguce, izbaciti ih iz modela. Takode, pokazalo se da na odluku banke ne utice iznos kredita, što nas vodi do zakljucka da banka ne teži ka tome da maksimizira profit, a samim tim ni vreme preživljavanja kredita. Ali sigurno je da banka ima neku strategiju kojom se vodi pri odobravanju kredita, bila to minimizacija ili maksimizacija nekih drugih parametara, kao što su na primer broj klijenata, profit od odredene vrste proizvoda ili nešto slicno.

U primeru iz Vijetnama, u koje je kreditno tržište još uvek u fazi razvoja, dolazimo do vrlo zanimljivih rezultata. Na primer, promenljiva «Vreme na trenutnoj adresi» koja oznacava broj godina koje je klijent proveo na istoj adresi, u vecini zemalja oznacava klijentovu stabilnost i zrelost, tako da se teži da ova promenljiva bude što veca. Medutim, u Vijetnamu je malo drugacija situacija, tamo se verovatnoca da ce klijent biti loš povecava sa brojem godina koje je proveo na istoj adresi, pošto ljudi što više zaraduju i napreduju teže da žive u što boljim uslovima, pa samim tim i cesto menjaju adresu stanovanja. U Vijetnamu, klijent je taj koji predlaže trajanje kredita, on sam procenjuje u kom roku je sposoban da otplati kredit i koliki rizik je spreman da prihvati. Ovo pokazuje da banke ne mogu da se pouzdaju u sopstvenu procenu nego se oslanjaju i na klijentovu poštenu ocenu trenutnog stanja.

Uporedivanjem genetskog programiranja i probit analize videli smo da nam model genetskog programiranja daje preciznije rezultate. To je samo znak da se razvojem kreditnog tržišta razvijaju i precizniji i pouzdaniji modeli kreditnog skoringa.

Iako se za sada kreditni skoring najcešce primenjuje u slucaju kada treba da se odredi kreditna sposobnost pojedinacnog klijenta, sve je veca primena skoringa i pri oceni rizika koje nose firme. U tom slucaju možemo da primenimo analizu obavijanja podataka (DEA), koja nam daje dosta tacne rezultate, što smo videli tako što smo rezultate uvrstili u linearnu regresiju i diskriminantnu analizu. Takode smo videli da nam DEA skorovi služe kao veoma dobri pokazatelji pri rangiranju klijenata na osnovu njihove kreditne sposobnosti.

Opšte pravilo ne postoji, modeli se razlikuju od banke do banke, od proizvoda do proizvoda. Pošto se uzimaju istorijski podaci, a model se periodicno ažurira, dešava se da je neka promenljiva u jednom periodu bila znacajna, dok je u sledecem bila iskljucena iz modela.

6. Literatura

1. Tor Jacobson, Kasper Roszbach, *"Bank lending policy, credit scoring and value-at-risk"*; Journal of Banking & Finance 27 (2003) 615-633

2. Kasper Roszbach, *"Bank lending policy, credit scoring and the survival of loans"*; Sveriges Riksbank Working Paper Series No.154 (2003)

3. Jae H. Min, Young-Chan Lee, *"A practical approach to credit scoring"*; Expert System with Applications 35 (2008) 1762-1770

4. Yenpao Chen, Ruey-Ji Guo, Rao-Li Huang, *"Two stages credit evaluation in bank loan appraisal"*; Economic Modelling 26 (2009) 63-70

5. Thi Huyen Thanh Dinh, Stefanie Kleimeier, *"A credit scoring model for Vietnam's retail banking market"*; International Review of Financial Analisys 16 (2007) 471-495

6. Vladimir Bugera, Hiroshi Konno, Stanislav Uryasev, *"Credit Cards Scoring with Quadratic Utility Function"*; Journal of Multi-criteria Decision Ananlysis 11 (2002) 197-211

7. Hussein A.Abdou, *"Genetic programming for credit scoring: the case of Egyptian public sector banks"*; Expert System with Applications (2009)

Biografija

Jelena Burgijašev je rodena 09.10.1984. godine u Novom Sadu od oca Mladena i majke Branislave. Završila je osnovnu školu «Jovan Popovic» u Novom Sadu. Gimnaziju «Jovan Jovanovic Zmaj» , prirodno-matematicki smer je završila 2003. godine sa odlicnim prosekom.

Studije na Prirodno-matematickom fakultetu, Univerzitet u Novom Sadu, smer matematika-finansija je upisala 2003. godine. Diplomirala je 2007. godine sa prosekom 9,17. U školskoj 2005/2006. godini je bila stipendista Ministarstva prosvete i sporta, a u školskoj 2006/2007. godini primala stipendiju iz Fonda za mlade talente. Master studije, smer diplomirani matematicar-master matematike finansija, upisala 2007. godine.

Posle diplomiranja, 2007. godine se zaposlila u Ministarstvu ekonomije i regionalnog razvoja, kao saradnik u Sektoru za strana ulaganja i koncesije. Od avgusta 2008. godine radi u Raiffeisen banci, kao mladi saradnik u Sektoru za Upravljanje rizicima u poslovanju sa malim i srednjim preduzecima.

UNIVERZITET U NOVOM SADU
PRIRODNO-MATEMATICKI FAKULTET
KLJUCNA DOKUMENTACIJSKA INFORMACIJA

Redni broj:
RBR

Identifikacioni broj:
IBR

Tip dokumentacije: Monografska dokumentacija
TZ

Vrsta rada: Master rad
VR

Autor: Jelena Burgijašev
AU

Mentor: Prof. Dr. Zorana Lužanin
MN

Naslov rada: Razliciti pristupi kreditnom skoring sistemu
NR

Jezik publikacije: srpski
JI

Zemlja publikacije: Srbija
ZP

Uže geografsko podrucje: Vojvodina
UGP

Godine: 2009.
GO

Izdavac: Autorski reprint
IZ

Mesto izdavanja: Novi Sad, Departman za matematiku i informatiku, PMF, Trg Dositeja Obradovica 4
MA

Fizicki opis rada: (6, 66, 0, 0, 34, 0)
FO

Naucna oblast: Matematika
NO

Naucna disciplina: Ekonometrija
ND

Predmetna odrednica: Kreditni skoring sistem, Logit regresija, Probit i Tobit model
PO

UKD:
Cuva se: U biblioteci Departmana za matematiku i informatiku
CU

Važna napomena:
VN

Izvod: U ovom radu su prezentovani razliciti pristupi kreditnom skoring sistemu, primenom logit regresije, probit i tobit analize, problema separacije. Ovaj model se najcešce srece u bankama kada odobravaju kredit, pri proceni kreditne sposobnosti klijenta i rizika koji taj klijent nosi
IZ

Datum prihvatanja teme od stane NN veca:
DP

Datum odbrane:
DO

Clanovi komisije:
KO
Dr. Danijela Rajter-Ciric, *redovni profesor*
Dr. Zorana Lužanin, *redovni profesor*
Dr. Andreja Tepavcevic, *redovni profesor*

UNIVERSITY OF NOVI SAD
FACULTY OF NATURAL SCIENCES AND MATHEMATICS
KEY WORD DOCUMENTATION

Accession number:
ANO

Identification number:
IDN

Type of record: Monograph type
DT

Contents code: Master thesis
CC

Author: Jelena Burgijašev
AU

Mentor: Prof. Dr. Zorana Lužanin
MN

Title: Different approaches to credit scoring
TI

Language of text: Serbian
LT

Language of abstract: English
LA

Country of publication: Serbia
CP

Locality of publication: Vojvodina
LP

Publication year: 2009.
PY

Publisher: Author's reprint
PU

Publ. place: Novi Sad, Department of Mathematics and Informatics, Faculty of Sciences and Mathematics, Trg Dositeja Obradovica 4
PP

Phisical description: (6, 66, 0, 0, 34, 0)
PD

Scientific field: Mathematics
SF

Scientific discipline: Econometrics
SD

Subject – Key words: Credit scoring system, Logit regression, Probit i Tobit model
SKW

UC:
Holding data: Library of Department of Mathematics and Informatics
HD

Note:
N

Abstract: Different approaches to credit scoring are presented in this thesis, using logit regression, probit and tobit model and separation. This model is used in banks for ranking clients who appling for credit. Their aim is to determine whether an applicant has the capacity to repay his debt.
AB

Accepted by the Scientifis Board on:
ASB

Defended:
DE

Thesis defened board:
DB
Dr. Danijela Rajter-Ciric, *full professor*
Dr. Zorana Lužanin, *full professor*
Dr. Andreja Tepavcevic, *full professor*