# Korištenje regresijskih metoda u kreditnom skoringu

**Rajčić, Jana**

**SVEUČILIŠTE U ZAGREBU**

**PRIRODOSLOVNO-MATEMATIČKI FAKULTET**

**MATEMATIČKI ODSJEK**

Jana Rajčič

# KORIŠTENJE REGRESIJSKIH METODA U KREDITNOM SKORINGU

Diplomski rad

Voditelj rada:

Prof. dr. sc. Siniša

Slijepčević

Zagreb, 2018/2019

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

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

Povjerenstvo je rad ocijenilo rad ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

*Mjesto za posve*

# Sadržaj

# Uvod

U prvom djelu rada se proučava linearna regresija, opći oblik, ali i jednostavni linearni model. U drugom djelu se proučava povijest kreditnog skoringa, važnost izgleda samog obrasca koji klijent popunjava pri traženju kredita, te važnost kreditnih registra.

Suština pozajmljivanja novca shvaćena je od samog početka. Potrebno je uspostaviti tko je potencijalni dužnik i kakva je njegova spremnost na plaćanje. Postoje zapisi 4000 godina unazad o korištenju kredita. Postojala je određena mistika o lukavom i izuzetnom bankaru koji bi sa nepogrešivim sudom odredio „kvalitetu“ klijenta. Svaki bankar je trebao imati duboko znanje o svojim klijentima i konkurenciji, što mu je onda omogućavalo da odluči, uz čisti instinkt i uz malo sreće, kome će posuditi novac, koliko će mu posuditi, koliko će to naplatiti i koje kolaterale može zahtijevati. No, tijekom posljednje polovice prošlog stoljeća do današnjeg dana, ova percepcija bankarstva praktički je nestala. Tijekom proteklih dvadeset i pet godina došlo je do radikalne promjene u načinu na koji se rizik (u općem smislu) i kreditni rizik mjere i kako se njima upravlja. Do ovakve drastične promjene u razmišljanju, došlo je zbog povećanja kompleksnosti na tržištu kredita, razvoja tehnologije i financijskog znanja. Zbog toga su bile potrebne metode kojima će se olakšati i preciznije izračunati dužnikova sposobnost otplaćivanja kredita. U trećem poglavlju rada se proučavaju razne statističke metode za izgradnju kreditnih bodovnih kartica. Postoje različite metode i analize kojima se može odrediti da li je osoba „dobar“ ili „loš“ klijent, odnosno da li će moći otplatiti kredit ili ne. U ovom diplomskog radu pokazati ću neke od njih, sa velikim naglaskom na regresijske metode.

# Poglavlje 1

# 1 Linearna regresija

## 1.1 Opći linearni model

Opći linearni model uključuje velik broj popularnih i korisnih modela koji se pojavljuju u primijenjenoj statistici, uključujući modele za višestruku regresiju i analizu varijance. Osnovni model se može ukratko napisati u obliku

$$Y = X\beta + \varepsilon \tag{1.1}$$

gdje je $Y$ varijabla odziva (slučajna varijabla) koja se opaža, t.d. $Y \in \mathbb{R}^n$ , a je $X$ je $n \times p$ matrica poznatih konstanti, $\beta \in \mathbb{R}^p$ je nepoznati parametar, te $\varepsilon$ slučajni vektor u $\mathbb{R}^n$. $\varepsilon$ se interpretira kao slučajna greška ili šum i ona se ne opaža. Inače pretpostavljamo da je $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ slučajni uzorak iz $N(0, \sigma^2)$, gdje je $\sigma^2$ nepoznati parametar, t.d.

$$\varepsilon \sim N(0, \sigma^2 I). \tag{1.2}$$

No neki od rezultata vrijede i za manje restriktivne uvjete, gdje je $E[\varepsilon_i] = 0$ za sve $i = 1,2,\ldots,n$, $Var[\varepsilon_i] = \sigma^2$ za sve $i = 1,2,\ldots,n$, te $cov(\varepsilon_i, \varepsilon_j) = 0$, *za sve* $i \neq j$. Ako to napišemo u matričnom obliku, vrijedi $E[\varepsilon] = 0$ i $cov(\varepsilon) = \sigma^2 I$. Sa normalnom distribucijom za $\varepsilon$ iz (1.2), za $Y$ vrijedi

$$Y \sim N(X\beta, \sigma^2 I) \tag{1.3}$$

Primjer 1. U kvadratnoj regresiji, varijabla odziva je modelirana kvadratnom funkcijom neke eksplanatorne varijable x + slučajna greška. Slijedi

$$Y_i = \beta_1 + \beta_2 x_i + \beta_3 {x_i}^2 + \varepsilon_i, \qquad i = 1,2,\ldots,n$$

Ovdje su eksplanatorne varijable $x_1, x_2, \ldots, x_n$ konstante, $\beta_1, \beta_2$ i $\beta_3$ su nepoznati parametri, a $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ su nezavisno, jednako distribuirane slučajne varijable iz $N(0, \sigma^2)$. Ako definiramo matricu $X$ kao [1]

[1] Robert W. Keener: „Theoretical Statistics“, Debt. Statistics, Universit of Michigan, Ann Arbor, USA, 2010, str. 269

$$X = \begin{bmatrix} 1 & x_1 & x_1^{(2)} \\ 1 & x_2 & x_2^{(2)} \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^{(2)} \end{bmatrix},$$

tada je $Y = X\beta + \varepsilon$, kao i u (1.1).

Pretpostavimo da imamo nezavisni slučajni uzorak iz normalne populacije sa zajedničkom varijancom $\sigma^2$, t.d.

$$Y_i = \begin{cases} N(\beta_{1,}, \sigma^2), & i = 1,2, \dots, n_1 \\ N(\beta_{2,}, \sigma^2), & i = n_1 + 1, \dots, n_1 + n_2 \\ N(\beta_{2,}, \sigma^2), & i = n_1 + n_2 + 1, \dots, n_1 + n_2 + n_3 \stackrel{\text{def}}{=} n \end{cases}$$

Ako definiramo

$$X = \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{pmatrix},$$

tada je $E[Y] = X\beta$, a model ima oblik (1.3).

U zahtjevima se često pojavljuju parametri $\beta_1, \beta_2, \dots, \beta_p$ pri formulaciji modela. Posljedica toga je da ih se jednostavno može interpretirati. No, zbog tehničkih razloga, često je prikladno gledati nepoznato očekivanje od $Y$. Naime,

$$\xi \stackrel{\text{def}}{=} EY = X\beta$$

je nepoznati parametar iz $\mathbb{R}^n$. Ako su $c_1, c_2, \dots, c_p$ stupci od $X$, tada vrijedi

$$\xi = X\beta = \beta_1 c_1 + \beta_2 c_2 + \dots + \beta_p c_p,$$

što pokazuje da $\xi$ mora biti linearna kombinacija stupaca od $X$. Stoga $\xi$ mora biti iz vektorskog prostora

$$\omega \stackrel{\text{def}}{=} span\{c_1, c_2, \dots, c_p\} = \{X\beta : \beta \in \mathbb{R}^p\}.$$

Ako koristimo $\xi$ umjesto $\beta$, vektor nepoznatih parametra je $\theta = (\xi, \sigma)$ koje poprimaju vrijednosti iz $\Omega = \omega \times \langle 0, \infty \rangle$.

Budući da $Y$ ima očekivanje $\xi$, intuitivno je da podaci moraju pružati informacije koje razlikuju bilo koje dvije vrijednosti za $\xi$, budući da je distribucija od $Y$ za dvije različite vrijednosti $\xi$ mora biti različita. Da li to vrijedi i za $\beta$ ovisi o rangu $r$ od $X$. Kako $X$ ima $p$ stupaca, tada je $r$ može najviše biti $p$. Ako je rang od $X$ jednak $p$, tada je svaka vrijednost $\xi \in \omega$ slika jedinstvene vrijednosti $\beta \in \mathbb{R}^p$.No, ako su stupci od $X$ linearno zavisni, tada će netrivijalna linearna kombinacija stupaca od $X$ biti jednaka 0, stoga je $Xv = 0$ za neke $v \neq 0$. No tada slijedi, $X(\beta + v) = X\beta + Xv = X\beta$, a oba parametra, $\beta$ i $\beta^* = \beta + v$, imaju isto očekivanje $\xi$. Ovdje naši podaci $Y$ pružaju informacije za razlikovanje parametara $\beta$ i $\beta^*$.

Pretpostavimo

$$X = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix}.$$

Ovdje su tri stupca matrice $X$ linearno zavisna zato što je prvi stupac suma drugog i trećeg stupca i rang matrice $X$ iznosi 2, $r = 2 < p = 3$. Primijetimo da obje vrijednosti parametra

$$\beta = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} i\ \beta^* = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$

daju

$$\xi = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

## 1.2 Kanonski oblik

Mnogi rezultati o testiranju i procjeni u općem linearnom modelu jednostavno slijede, kada su podaci prikazani u kanonskom obliku. Neka je $v_1, \ldots, v_n$ ortonormirana baza za $\mathbb{R}^n$, odabrana tako da $v_1, \ldots, v_r$ generira $\omega$. Ulazi u kanonskim podacima vektora $Z$ su koeficijenti koji izrazuju $Y$ kao linearnu kombinaciju tih bazičnih vektora, tj.

$$Y = Z_1\, v_1 + \ldots + Z_n v_n \tag{1.4}$$

Algebarski, $Z$ nam može predstaviti $n \times n$ ortogonalnu matricu, $O'O = OO' = I$, a $Y$ i $Z$ su povezani sa

$$Z = O'Y \;\; ili \;\; OZ.$$

Budući da je $Y = \xi + \varepsilon$, $Z = O'(\xi + \varepsilon) = O'\xi + O'\varepsilon$. Ako definiramo $\eta = O'\xi$ i $\varepsilon^* = O'\varepsilon$, tada vrijedi $Z = \eta + \varepsilon^*$. Stoga je je $E\varepsilon^* = EO'\varepsilon = O'E\varepsilon = 0$ i $cov(\varepsilon^*) = cov(O'\varepsilon) = O'cov(\varepsilon)O = O'(\sigma^2 I)O = \sigma^2 O'O = \sigma^2 I$, $\varepsilon^*{\sim}N(0, \sigma^2 I)$ i ${\varepsilon^*}_1, {\varepsilon^*}_2, \ldots, {\varepsilon^*}_n$ su nezavisno, jednako distribuirane iz $N(0, \sigma^2)$. Budući da je $Z = \eta + \varepsilon^*$, vrijedi

$$Z{\sim}N(\eta, \sigma^2 I). \tag{1.5}$$

Nadalje, označimo $c_1, c_2, \ldots, c_p$ kao stupce od matrice $X$. Tada je $\xi = X\beta = \sum_{i=1}^{p} \beta_i\, c_i$ i

$$\eta = O'\xi = \begin{pmatrix} v_1' \\ \vdots \\ v_n' \end{pmatrix} \sum_{i=1}^{p} \beta_i\, c_i = \begin{pmatrix} \sum_{i=1}^{p} \beta_i\, v_1' c_i \\ \vdots \\ \sum_{i=1}^{p} \beta_i v_n'\, c_i \end{pmatrix}.$$

Budući da $c_1, c_2, \ldots, c_p$ leže u vektorskom prostoru $\omega$ i $v_{r+1}, \ldots, v_n$ leže u $\omega^{\perp}$, imamo $v_k'\, c_i = 0$ za $k > r$. Stoga vrijedi

$$\eta_{r+1} = \ldots = \eta_n = 0. \tag{1.6}$$

Kako je $\eta = O'\xi$, slijedi

$$\xi = O\eta = (\, v_1 \quad \ldots \quad v_n) \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_r \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \sum_{i=1}^{r} \eta_i\, v_i.$$

Te jednadžbe utemeljuju jedan-na-jedan odnos između točaka $\xi \in \omega$ i $(\eta_1, \ldots, \eta_r \in \mathbb{R}^r)$.

Budući da je $Z{\sim}N(\eta, \sigma^2 I)$, varijable $Z_1, \ldots, Z_n$ su nezavisne sa distribucijom $Z_i{\sim}N(\eta_i, \sigma^2)$. Gustoća od $Z$, uzimajući u obzir prednost koju nam daje (1.6), jednaka je

$$\frac{1}{\sqrt{2\pi\sigma^2}^n} exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{r}(z_i-\eta_i)^2-\frac{1}{2\sigma^2}\sum_{i=r+1}^{n} {z_i}^2\right]$$

$$= exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n} {z_i}^2+\frac{1}{\sigma^2}\sum_{i=1}^{r}\eta_i z_i-\sum_{i=1}^{r}\frac{{\eta_i}^2}{2\sigma^2}-\frac{n}{2}log(2\pi\sigma^2)\right].$$

Ove gustoće formiraju puni rang $(r+1)$-parametarsku eksponencijalnu familiju sa kompletnom, dovoljnom statistikom

$$\left(Z_1,\dots,Z_{r,}\sum_{i=1}^{n} {Z_i}^2\right) \tag{1.7}$$

## 1.3 Procjena parametra modela

Ako iskoristimo kanonski oblik, mnogi se parametri mogu lako procijeniti. Kako je $EZ_i = \eta_i, i = 1,\dots,r$, $Z_i$ je nepristran procjenitelj od $\eta_i, i = 1,\dots,r$. Također, budući da je $\xi = \sum_{i=1}^{r} \eta_i\, v_i$, slijedi

$$\hat{\xi} = \sum\nolimits_{i=1}^{r} Z_i v_i \tag{1.8}$$

je prirodni procjenitelj od $\xi$. Uočimo da vrijedi

$$E\hat{\xi} = \sum\nolimits_{i=1}^{r} EZ_i v_i = \sum\nolimits_{i=1}^{r} \eta_i\, v_i = \xi,$$

tj. $\hat{\xi}$ je nepristran. Budući da je on funkcija kompletne, dovoljne statistike, on bi trebao biti optimalan. Jedna mjera optimalnosti očekivana kvadratna udaljenost od prave vrijednosti $\xi$. Ako je $\tilde{\xi}$ nepristrani procjenitelj. Tada vrijedi

$$E\left\|\tilde{\xi}-\xi\right\|^2 = \sum\nolimits_{j=1}^{n} E\left(\tilde{\xi}_j-\xi_j\right)^2 = \sum\nolimits_{j=1}^{n} Var(\tilde{\xi}_j). \tag{1.9}$$

Budući da je $\hat{\xi}_j$ nepristran za $\xi_j$, te je funkcija kompletne, dovoljne statistike, tada je $Var(\hat{\xi}_j) \le Var(\tilde{\xi}_j), j = 1,\dots,n$. Stoga $\hat{\xi}$ minimizira svaki izraz u (1.9), pa slijedi

$$\left\|\hat{\xi} - \xi\right\|^2 \leq E\left\|\tilde{\xi} - \xi\right\|^2.$$

No, $\hat{\xi}$ minimizira i očekivanje od bilo kojeg drugog nenegativnog kvadratnog oblika u pogreški procjene, $E\left(\hat{\xi} - \xi\right)' A(\hat{\xi} - \xi)$, među svim nepristranim procjeniteljima.

Iz (1.4), možemo zapisati $Y$ na sljedeći način

$$Y = \sum_{i=1}^{r} Z_i v_i + \sum_{i=r+1}^{n} Z_i v_i = \hat{\xi} + \sum_{i=r+1}^{n} Z_i v_i.$$

U ovom izrazu, prvi pribrojnik, $\hat{\xi}$, leži u vektorskom prostoru $\omega$, a drugi pribrojnik, $Y - \hat{\xi} = \sum_{i=r+1}^{n} Z_i v_i$, leži u $\omega^{\perp}$. Ova razlika $Y - \hat{\xi}$ se zove *vektor reziduala*, koji ćemo označiti sa $e$:

$$e \stackrel{\text{def}}{=} Y - \hat{\xi} = \sum_{i=r+1}^{n} Z_i v_i \tag{1.10}$$

Budući da je $Y = \hat{\xi} + e$, po Pitagorinom teoremu, ako je $\tilde{\xi}$ bilo koja točka u $\omega$, tada je

$$\left\|Y - \tilde{\xi}\right\|^2 = \left\|\hat{\xi} - \tilde{\xi} + e\right\|^2 = \left\|\hat{\xi} - \tilde{\xi}\right\|^2 + \|e\|^2,$$

jer je razlika $\hat{\xi} - \tilde{\xi} \in \omega$ ortogonalna na $e \in \omega^{\perp}$. Iz ove jednadžbe, očito je da je $\hat{\xi}$ jedinstvena točka u $\omega$ najbliža podacima vektora $Y$. Ta najbliža točka se zove *projekcija od $Y$ na $\omega$*.

Vrijedi da je $\hat{\xi} = PY$, gdje je P $n \times n$ *ortogonalna projekcijska matrica na* $\omega$. Budući da $\hat{\xi}$ leži u vektorskom prostoru $\omega$ i $P\hat{\xi} = \hat{\xi}$, vrijedi

$$P^2 Y = P(PY) = P\hat{\xi} = \hat{\xi} = PY.$$

Kako $Y$ poprima arbitrarne vrijednosti na $\mathbb{R}^n$, slijedi $P^2 = P$.

Prisjetimo se da su $c_i, i = 1, \ldots, p$ stupci matrice $X$, koji leže u vektorskom prostoru $\omega$, a $e = Y - \hat{\xi}$ leži u $\omega^{\perp}$. Tada moramo imati $c_i' e = 0$, što povlači činjenicu da je $X'e = 0$. Budući da je $Y = \hat{\xi} + e$, slijedi

$$X'Y = X'(\hat{\xi} + e) = X'\hat{\xi} + X'e = X'\hat{\xi} = X'X\hat{\beta} \tag{1.11}$$

Ako je $X'X$ inverzna, tada ova jednadžba daje sljedeće:

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{1.12}$$

Matrica $X'X$ je inverzna ako je $X$ punog ranga, tj. $r = p$. U ovom slučaju je $X'X$ čak pozitivno definitna. Kako je $X$ punog ranga, slijedi

$$PY = \hat{\xi} = X\hat{\beta} = X(X'X)^{-1}X'Y$$

Tada projekcijsku matricu $P$ možemo zapisati na sljedeći način:

$$P = X(X'X)^{-1}X' \tag{1.13}$$

## 1.4 Gauss-Markovljev teorem

Ovdje ćemo malo ublažiti pretpostavke iz općeg linearnog modela. Model još uvijek ima oblik $Y = X\beta + \varepsilon$, ali sada $\varepsilon_i, i = 1, \ldots, n$ ne moraju biti slučajni uzorci iz $N(0, \sigma^2)$. Umjesto toga, pretpostavimo da $\varepsilon_i, i = 1, \ldots, n$ imaju varijance 0, $E\varepsilon_i = 0, i = 1, \ldots, n$, zajedničku varijancu, $Var(\varepsilon_i) = \sigma^2, , i = 1, \ldots, n$, te $cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j$. U matričnom obliku, ove pretpostavke se mogu zapisati na sljedeći način $E\varepsilon = 0$ i $cov(\varepsilon) = \sigma^2 I$. Tada vrijedi

$$EY = X\beta = \xi \; i \; Cov(Y) = \sigma^2 I \, .$$

Bilo koji procjenitelj oblika $a'Y = a_1Y_1 + \ldots + a_nY_n$, gdje je $a \in \mathbb{R}^n$ vektor konstanta, koji zovemo *linearni procjenitelj*. Prisjetimo se da ako je $Y = f(X)$, gdje je X slučajna varijabla na vjerojatnosnom prostoru $(\varepsilon, B, P)$, gdje je $B$ Borelov skup i $P$ vjerojatnosna mjera, tada je

$$EY = Ef(X) = \int f dP_X \tag{1.14}$$

Sada koristeći (1.14) možemo dobiti sljedeće:

$$Var(a'Y) = Cov(a'Y) = a'Cov(Y)a = a'(\sigma^2 I)a = \sigma^2 a'a = \sigma^2 \|a\|^2, \tag{1.15}$$

Budući da je $\hat{\xi}$ nepristrani procjenitelj, $a'\hat{\xi}$ je nepristrani procjenitelj za $a'\xi$. Procjenitelj je nepristrani procjenitelj sa najmanjom varijancom jer je $\hat{\xi}$ funkcija kompletne, dovoljne statistike. Iz (1.11) slijedi da je $X'Y = X'\hat{\xi}$, stoga iz (1.12), ako je $X$ matrica punog ranga, slijedi

$$\hat{\beta} = (X'X)^{-1}X'\hat{\xi}.$$

Ova jednadžba pokazuje da je $\hat{\beta}_i$ linearna funkcija od $\hat{\xi}$, stoga je $\hat{\beta}_i$ nepristrani procjenitelj od $\beta_i$ sa najmanjom varijancom.

Kako je $EY = \xi$, procjenitelj $a'\hat{\xi}$ je nepristran za $a'\xi$. Budući da je $\hat{\xi} = PY$, $a'\hat{\xi} = a'PY = (Pa)'Y$. Stoga iz (1.15) slijedi

$$Var(a'\hat{\xi}) = \sigma^2\|Pa\|^2, \tag{1.16}$$

Opet iz (1.14) i činjenice da je $P$ je simetrična, tj. vrijedi $P^2 = P$, slijedi

$$Cov(\hat{\xi}) = Cov(PY) = PCov(Y)P = P(\sigma^2 I)P = \sigma^2 P.$$

Kada $X$ ima puni rang, možemo izračunati procjenitelja temeljenog na metodu najmanjih kvadrata $\hat{\beta}$ koristeći (1.14):

$$\begin{aligned} Cov(\hat{\beta}) = Cov((X'X)^{-1}X'Y) &= (X'X)^{-1}X'Cov(Y)X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned} \tag{1.17}$$

**Teorem 1 (Gauss-Markov)** *Pretpostavimo da vrijedi*

$$EY = X\beta = \xi \; i \; Cov(Y) = \sigma^2 I \; .$$

*Tada je procjenitelj temeljen na metodi najmanjih kvadrata $a'\hat{\xi}$ od $a'\xi$ nepristran te ima uniformno minimalnu varijancu među svim linearnim nepristranim procjeniteljima.*

*Dokaz.* Neka je $\delta = b'Y$ nepristrani procjenitelj različit od $a'$. Iz (1.15) i (1.16), varijance od $\delta$ i $a'\hat{\xi}$ su

$$Var(\delta) = \sigma^2\|b\|^2 \text{ i } Var(a'\hat{\xi}) = \sigma^2\|Pa\|^2.$$

Ako $\varepsilon$ dolazi iz normalne distribucije, budući da su oba procjenitelja nepristrana i $a'\hat{\xi}$ je nepristrani procjenitelj sa najmanjom uniformno varijancom (UMVU), tj. $Var(a'\hat{\xi}) \leq Var(\delta)$ ili

$$\sigma^2\|Pa\|^2 \leq \sigma^2\|b\|^2.$$

No, formule za varijancu procjenitelja ne ovise i tome da li su procjenitelji normalno distribuirani ili ne, stoga $Var(a'\hat{\xi}) \leq Var(\delta)$ vrijedi i općenito. [1, str. 276] □

Iako je $a'\hat{\xi}$ „najbolji“ linearni procjenitelj, u nekim primjerima nelinearni procjenitelji mogu biti precizniji.

## 1.5 Jednostavna linearna regresija

Budući da je $X$ neslučajna varijabla , od sada pa nadalje ćemo je označavati sa $x$. Pogledajmo sada kako izgleda linearni regresijski model. Jednostavni linearni regresijski model opisuje odnos među pojavama za koje je svojstveno da svakome jediničnom porastu vrijednosti jedne varijable odgovara približno jednaka linearna promjena druge varijable. Takav model izražava vezu između zavisne i jedne nezavisne varijable. Pretpostavljamo da su varijable $x$ i $Y$ u srednjem linearno povezane, tj.

$E[Y|x] = \beta_0 + \beta_1 x,$ te $Var[Y|x] = \sigma^2$,

Kada su vrijednosti $\beta_0$, $\beta_1$ i $\sigma^2$ poznate, tada je model kompletno opisan. No, vrijednosti $\beta_0$, $\beta_1$ i $\sigma^2$ su najčešće nepoznate. Preciznije to možemo zapisati na sljedeći način:

$$Y = \beta_0 + \beta_1 x + \varepsilon, \tag{1.18}$$

gdje su $\beta$ i $\beta_1$ nepoznate konstante, gdje $\beta_1$ često zovemo *nagib regresijske jednadžbe*, poznate kao koeficijenti ili parametri modela , te je $\varepsilon$ slučajna varijabla t.d. je $E[\varepsilon] = 0$ i ona se ne opaža. $\varepsilon$ se interpretira kao slučajna greška ili šum, a predstavlja razliku između

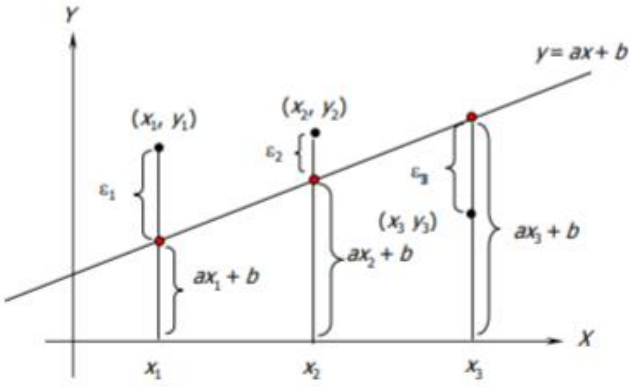
teorijske i eksperimentalne realizacije od $Y$. Pretpostavljamo da je $\varepsilon$ nezavisna i jednoliko distribuirana slučajna varijabla sa očekivanjem 0 i konstantnom varijancom $\sigma^2$.

Da bi znali vrijednosti parametara $\beta_0$, $\beta_1$ i $\sigma^2$ promatramo $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ slučajni uzorak iz linearnog regresijskog modela. Te opservacije zadovoljavaju jednostavni linearni regresijski model, stoga možemo pisati

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1,2, \ldots, n \tag{1.19}$$

Kada bi odnos među varijablama bio funkcionalan (veze se mogu predočiti izrazima na temelju kojih se točno utvrđuje vrijednost jedne za danu vrijednost druge (drugih) vrijednosti $Y$), svaka bi vrijednost varijable $\varepsilon_i$ bila jednaka nuli – geometrijski, sve bi točke s koordinatama $(x_i, Y_i), i = 1,2, \ldots, n$ ležale na istom pravcu. Kako su odnosi među pojavama statistički (jednoj vrijednosti jedne pojave odgovara više vrijednosti druge (drugih) pojava) trebamo odrediti kriterij prema kojem će se izabrati jednadžba pravca $\hat{Y} = \beta_0 + \beta_1 x$ koja će najbolje opisati odnos pojava na temelju njihovih opaženih vrijednosti.

Postoje različite metode za procjenu tih parametra. Jedna od tih metoda je metodom najmanjih kvadrata (bazira se na uvjetu da zbroj kvadrata vertikalnih odstupanja točaka na dijagramu rasipanja od traženog pravca regresije bude minimalan) .




Tablica 1.1 Vertikalna odstupanja od pravca regresije

Želimo minimizirati sljedeću funkciju:

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} \hat{\varepsilon}_i{}^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2, \tag{1.20}$$

Odnosno želimo minimizirati sumu kvadrata odstupanja $Y_i$ i $\widehat{Y}_i$ promatranih opažanih $Y_i$-a od njihovih predviđenih vrijednosti $\widehat{Y}_i = \beta_0 + \beta_1 x_i$.

Rješavanjem sustava

$$\frac{\partial L}{\partial \beta_0}(\beta_0, \beta_1) = 0, \qquad \frac{\partial L}{\partial \beta_1}(\beta_0, \beta_1) = 0$$

dobijemo parametre $\beta_0$ i $\beta_1$. Rješenja te dvije jednadžbe zovu se *direktni regresijski procjenitelji* od $\beta_0$ i $\beta_1$. Parametar $\hat{\beta}_1$ zove se *regresijski koeficijent*. On pokazuje za koliko se u prosjeku promijeni zavisna varijabla ako se nezavisna varijabla promijeni za jedan. Parametar $\hat{\beta}_0$ pokazuje vrijednost zavisne varijable u slučaju kada je nezavisna varijabla jednaka nuli. Pravac prilagođen metodom najmanjih kvadrata podacima
$Y = \beta_0 + \beta_1 x + \varepsilon$ je

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x, \tag{1.21}$$

gdje su

$$\hat{\beta}_1 := \frac{S_{xY}}{S_{xx}}, \quad \hat{\beta}_0 := \bar{Y} - \hat{\beta}_1 \bar{x}, \tag{1.22}$$

te je srednje kvadratno odstupanje varijable $x$ od $\bar{x}$

$$S_{xx} := \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

i uzročna kovarijanca

$$S_{xY} := \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ su aritmetičke sredine varijable $x_i$, odnosno $y_i$.

Uočimo da su $\hat{\beta}_0$ i $\hat{\beta}_1$ linearne kombinacije od $Y_i$, za $i = 1,2, \ldots, n$. Sada možemo izračunati koliko iznosi svako odstupanje teorijske $Y_i$ od eksperimentalne vrijednosti: $\hat{e}_i :=$

$Y_i - \hat{Y}_i$. Razlike između teorijskih vrijednosti $Y_i$ i eksperimentalnih vrijednosti $\hat{Y}_i$ zovu se rezidualima. Slijedi da je mjera kvalitete modela ($SSE$) dana sljedećom formulom

$$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i\right)^2 = \sum_{i=1}^{n} \hat{e}_i{}^2 \tag{1.23}$$

# Poglavlje 2

# 2 Praksa kreditnog skoringa

Sa razvojem srednje klase, 1800-tih, zajmodavci su uvidjeli da postoji brzo rastuće tržište manjih kredita. To dovodi do stvaranja prvih komercijalnih banaka, posrednika u zalagaonicama, pa čak i pošte – sve kako bi služile potrošačkom kreditu. Cijeli proces krenuo je u 1920-tim s mogućnošću velike većine građana da kupi automobil. To je međutim zahtijevalo radikalne promjene u načinu poslovanja. Postojala je potreba za standardizacijom proizvoda i sistematizacijom postupka odobravanja kredita i upravljanja. U današnje vrijeme, to je postalo imperativ. Budući da je tipični dužnik izgubljen u anonimnosti velike mase pojedinaca koji duguju novac bankama ili drugim potrošačkim vjerovnicima, teško je provesti procjenu rizika koji oni predstavljaju, na temelju intuicije ili čistog iskustva. Stoga je neka vrsta automatske klasifikacije "kvalitete" dužnika postala nužnost. [2, str. 9-10]

Kreditni skoring je prvi formalni pristup problem procjene kreditnog rizika jednog dužnika na znanstveni i automatiziran način, kao direktni odgovor na potrebu obrade velikog broja zahtjeva za relativno male kredite[2, str.10]. Kreditni skoring je sistem dodjeljivanja bodova zajmodavcu u cilju dobivanja numeričke vrijednosti koja pokazuje koliko je vjerojatno da zajmotražitelj iskusi neki događaj odnosno izvede neku akciju kao primjerice, kasni u otplati kredita [3, str. 6]. Odnosno, kreditni skoring je proces kojim se određuje koliko je vjerojatno da klijent kasni u otplatama rata kredita. Sustavi kreditnog skoringa pomažu: pojednostavljenju postupka kreditiranja, poboljšanju učinkovitosti kreditnog službenika, povećanju dosljednosti postupka ocjenjivanja, smanjivanju ljudske pristranosti u odluci kreditiranja, omogućavanju bankama da mijenjaju kreditnu politiku u skladu s klasifikacijom rizika, kao što je odobravanje ili praćenje nekih zajmova nižih rizika bez poslovnih inspekcija na licu mjesta, boljem kvantificiranju očekivanih gubitka za različite nivoe rizika zajmoprimaca, te smanjenju vremena provedenog na „zbirkama".

## 2.1 Uvod

Prvo ćemo predstaviti osnovnu operaciju i ideje kreditnog skoringa, iz poslovne perspektive i perspektive kreditiranja. Kao prvo, opisati ćemo na koji način je bila vršena procjena kredita prije rasprostranjenosti koncepta kreditnog skoringa. Nakon toga, na višim razinama smo ispitali što je sustav kreditnog skoringa i na koji način se on uklapa u cjelokupno poslovanje zajmodavca. Također, obratili smo pažnju na činjenicu koji su podaci potrebni i kako njima upravljati. Također razmatramo druge uključene strane - kreditne registre i kreditne konzultantske kuće.

## 2.2 Kreditna procjena prije skoringa

Ne tako davno - sigurno u 1970-ima u Velikoj Britaniji i SAD-u i možda, za neke zajmodavce, čak i u kasnim 1990-ima – kreditni skoring nije bio korišten. Tradicionalna kreditna procjena oslanjala se na "osjećaj intuicije" i procjenu karaktera potencijalnog dužnika, sposobnosti otplate i kolaterala ili osiguranja. To je značilo da potencijalni dužnik nije pristupio upravitelju banke ili upravitelju stambene štedionice sve dok nije koristio usluge štednje i druge usluge istih institucija nekoliko godina prije samog zahtjeva za kredit. Zatim, s određenom dozom strepnje, bio bi zakazan sastanak i, noseći najbolje nedjeljno odijelo, klijent bi tražio da pozajmi nešto novca. Upravitelj bi razmotrio prijedlog i, unatoč dužini odnosa banke sa potencijalnim dužnikom, razmotrio vjerojatnost otplate i procijenio stabilnost, iskrenost pojedinca i njegov karakter. On - upravitelj je uvijek bio muškarac - također bi procijenio predloženo korištenje dobivenih sredstava, a nakon toga imao i mogućnost zatražiti neovisno mišljenje kroz vođe zajednice u kojoj su živjeli ili poslodavca podnositelja zahtjeva. Nakon toga, dodatni sastanak bi bio dogovoren gdje bi odluka bila donesena i kupac bi bio obaviješten. Taj je proces bio spor i nedosljedan. Potisnula je ponudu kredita jer potencijalni zajmoprimac bi imao odnos sa samo jednim zajmodavcem.

Ti nedostaci su postojali dugo vremena. No tijekom 1980-ih u Velikoj Britaniji dogodile su se mnoge promjene u okruženju kreditnog zaduživanja. Neke od tih promjena bile su

sljedeće: banke su znatno promijenile svoj tržišni položaj i počele plasirati svoje proizvode na tržište. Zabilježen je nevjerojatan rast kreditnih kartica. Ovlaštenje za prodaju tog proizvoda uvjetovalo je postojanje mehanizma za donošenje vrlo brze odluke o posudbi. Bankarska praksa promijenila je naglasak. Prije su se banke gotovo isključivo fokusirale na velike zajmove i korporativne klijente. Sada, kreditiranje potrošača je postalo važan i rastući dio banke. I dalje je to bio manji dio gledajući vrijednost, ali je postajao sve značajni. Banke nisu mogle kontrolirati kvalitetu preko mreže stotine ili tisuće poslovnica, i učinjene su pogreške. Uzimajući u obzir prirodu kreditiranje trgovačkih društava, cilj je obično bio izbjeći bilo kakve gubitke. Međutim, banke su počele shvaćati da kod potrošačkog kreditiranja, cilj ne bi trebao biti izbjegavanje bilo kakvih gubitaka nego povećanje profita. Održavanje gubitaka pod kontrolom je veliki dio toga, ali maksimiziranje profita je moguće dobiti preuzimanjem male razine kontroliranih, loših dugova i tako proširiti knjigu potrošačkih kredita.

Bilo je i mnogo drugih razloga zbog kojih je kreditni skoring uveden u UK 1980-ih, iako, kao što smo rekli, neki zajmodavci i dalje ignoriraju skoring. Većina njih su mali zajmodavci koji ne mogu ili ne žele kontrolirano proširiti svoju knjigu zajmova, tako da oni imaju luksuz kao dobro obučeni i iskusni menadžeri koji mogu donositi kritične odluke o preuzimanju rizika. Za veliku većinu tržišta koristi se kreditni skoring u jednom ili drugom obliku. U potrošačkom kreditiranju, to se vjerojatno najviše koristi u kreditnim karticama, zbog volumena zahtjeva i 24-satnog pokrivanja usluge. Također, nakon što je donesena odluka da se odobri kreditna kartica, time proces kreditiranja ne završava. Odluke se mogu zahtijevati ne samo za odobrenja, nego i za kreditna ograničenja klijenta ili čak o tome hoće li se ponovno izdati kartica ili ne i na koliko dugo. Za kredit na rate, odluka o skoringu je malo jednostavnija. Osnovna odluka je odobriti zajam ili ne. U nekim situacijama, odluka može biti složenija jer mogu postojati dodatni parametri za razmatranje koji mogu utjecati na odluku. To može uključivati kamatnu maržu ili kamatnu stopu za naplatu ili osiguranja zahtjeva, ili rok na koji se novac posuđuje. Za prekoračenja limita, situacija je slična onoj za kreditne kartice. Potrebna je odluka što bi trebao biti limit prekoračenja. Ako klijenti upravljaju svojim računima unutar toga ograničenja, nema puno toga za učiniti. Međutim, ako je predočen ček ili je napravljen zahtjev za elektroničko plaćanje i račun bi prešao trenutno prekoračenje limita, donosi se odluka o tome da li će

ček biti plaćen ili će biti vraćen neplaćen. Ako kupac zatraži povećanje limita prekoračenja, zatražena je odluka da li će povećanje biti odobreno ili ne.

## 2.3 Kako se kreditni skoring uklapa u procjenu zajmodavca?

Na visokoj razini, potencijalni zajmoprimac predstavlja zahtjev vjerovniku. Vjerovnik razmatra zahtjev i procjenjuje vezani rizik. Prije su bankari imali određeno znanje koje im je omogućilo da ocijene je li rizik prihvatljivo niske razine. Kroz kreditni skoring, zajmodavac primjenjuje formulu sa ključnim elementima zaprimljenog zahtjeva, a kao izlazni proizvod te formule je obično numerička kvantifikacija rizika. Kako se kreditni skoring uklapa u procjenu zajmodavca može se malo razlikovati od proizvoda do proizvoda. Razmotrimo, na primjer, zahtjev za osobni zajam. U današnje vrijeme, podnositelj zahtjeva popunjava obrazac zahtjeva. To može biti popunjeno ručno ili elektronski. Na kraju, podaci iz zahtjeva će se bodovati. Ne koriste se svi podaci iz zahtjeva u izračunu kreditne ocjene. Međutim, kao što ćemo vidjeti u nastavku, ostale informacije su potrebne u razne svrhe, uključujući identifikaciju, sigurnost i buduću ocjenu bodovne kartice.

Izračun kreditnog skoringa može uključivati i neke informacije iz kreditnog registra. U mnogim slučajevima i u mnogim sredinama rezultat procesa kreditnog skoringa donosi preporuku ili odluku zaprimljenog zahtjeva. Uloga subjektivne ljudske procjene je smanjena na mali postotak slučajeva u kojima postoji istinska mogućnost dodane vrijednosti kroz ljudski faktor. Da bi stvari bile konkretnije, razmotrimo jednostavnu operaciju kreditnog skoringa. Pretpostavimo da imamo bodovnu karticu s četiri varijable (ili karakteristike): stambeni status, dob, svrha zajma i vrijednost presuda županijskih sudova (PžS); vidi Tablica 2.1.

| **STAMBENI STATUS** | | **DOB** | |
|---|---|---|---|
| Vlasnik | 36 | 18-25 | 22 |
| Stanar | 10 | 26-35 | 25 |
| Živi s roditeljima | 14 | 36-43 | 34 |
| Ostalo navedeno | 20 | 44-52 | 39 |
| Nema odgovora | 16 | 53+ | 49 |
| **Svrha kredita** | | **Vrijednost PžS** | |
| Novi auto | 41 | 0 | 32 |
| Rabljeni automobil | 33 | £l-£299 | 17 |
| Uređenje doma | 36 | £300-£599 | 9 |
| Odmor | 19 | £600-£1199 | -2 |
| drugo | 25 | £1200+ | -17 |

Tablica 2.1 *Jednostavna bodovna kartica*

Dvadesetogodišnjak, koji živi sa svojim roditeljima, koji želi posuditi novac za pomoć pri kupnji rabljenog automobila i nikada nije imao PžS, će postići 101 (14 + 22 + 33 + 32). S druge strane, 55-godišnji vlasnik kuće, koji je imao £250 PžS i želi posuditi novac za vjenčanje svoje kćeri, postigao bi 127 (36 + 49 + 25 + 17). Napominjemo da ne tvrdimo da netko stariji od 53 godine ima 27 bodova više od nekoga u dobi od 18 do 25 godina. Za karakterističnu dob, ova razlika od 27 točaka je istinita. Međutim, kao što se može jasno vidjeti, postoje korelacije. Na primjer, za nekog tko je stariji od 53 godine je vjerojatnije da će biti vlasnik kuće nego neka osoba od 18 do 25 godina, dok je manje vjerojatnije naći nekoga u dobi od 53+ koji živi s roditeljima. Dakle, ono što bismo mogli naći je da netko u starijoj starosnoj kategoriji može postići prosječno 40 ili 50 ili 60 bodova više kada su druge karakteristike uzete u obzir. Pri uspostavljanju sustava skoringa , donijet će se odluka o tome što predstavlja prolaznu ocjenu. To je jednostavna stvar koju treba provesti, ali nije nužno jednostavna stvar za utvrditi. Pretpostavimo da je

u gornjem primjeru, oznaka za prolazak 100. Dakle, svaki zahtjev koji bi imao minimum 100 ili više bodova, nosio bi preporuku za odobrenje. To bi bio slučaj bez obzira na odgovor na četiri pitanja. Prema tome, to što omogućuje skoring, je kompromis da se slabost jednog faktora može nadoknaditi snagom drugih čimbenika.. U procjeni zahtjeva za kredit, zajmodavac prikuplja podatke o podnositelju zahtjeva. To može biti iz različitih izvora, uključujući samog podnositelja zahtjeva, kreditnog registra i dosjea zajmodavca o ostalim računima podnositelja zahtjeva. Izvještaji kreditnog registra obično su dostupni elektronski, ali papirna izvješća su još uvijek dostupna (i još uvijek su prilično uobičajena u poslovanju komercijalnog kreditiranja). Vjerovnik će provjeriti i ispitati dostupne informacije i izračunati rezultat. Postoji mnogo načina za korištenje tog rezultata.

Neki zajmodavci rabe vrlo strogu politiku određivanja granice odobrenja/odbijanja zahtjeva. Ako je rezultat veći ili jednak granici, zahtjev je odobren. Ako je niži od granice, zahtjev je odbijen. Neki zajmodavci upotrjebljuju jednostavniju verziju toga. U tom slučaju kao posljedica stvara se referentni pojas ili tzv. sivo područje. To može biti 5 ili 10 točaka na jednoj ili obje strane granice odobrenja/odbijanja zahtjeva. Zahtjevi koji padaju u takvo sivo područje zahtijevaju detaljniju analizu . Takva vrsta analize može uključivati i neke subjektivne vrste osiguravanja ili traženje dodatnih informacija koje zahtijevaju objektivnu procjenu.

Neki zajmodavci primjenjuju pravila koja prisiljavaju potencijalno prihvatljive slučajeve u referentni pojas ili tzv. sivo područje. Na primjer, to može biti slučaj gdje zahtjev postiže graničnu vrijednost, ali u podacima kreditnog registra postoji zabilježba štetnog događaja, npr. stečaj. Drugim riječima, ne bi bilo dopušteno da snaga ili težina ostatka zahtjeva automatski kompenzira slabosti istog.

Neki zajmodavci primjenjuju nešto sto se naziva „super-pass“ i „super-fail“ kategorije. Dolje u nastavku ćemo diskutirati o ulozi kreditnog registra, ali je poznato da će dobivanje izvješća kreditnog registra uzrokovati trošak. Dakle, svakako postoje slučajevi kod kojih je rezultat toliko loš da čak i izvješća najboljeg kreditnog registra neće podići rezultat na iznad granice odobrenja. To bi bilo klasificirano kao slučaj super-fail-a. U drugoj krajnosti, možda imamo slučajeve koji su toliko dobri da čak i najgore izvješće

kreditnog registra, neće smanjiti rezultat koji je ispod granice odobrenja. To je primjer superpass-a.

U načelu, takvi zajmodavci rade sa dva ili tri limita odobrenja/odbijanja zahtjeva: jedan za definiranje superpass-a, jedan za definiranje superfail-a, i između njih, limit koji će se koristiti nakon što se upotrijebe informacije dobivene od kreditnog registra. Alternativno korištenje toga može biti u slučaju gdje je cijena izvješća kreditnog registra niska u usporedbi sa dobivanjem informacija o podnositelja zahtjeva. To se može dogoditi kod zajmodavaca koje se baziraju na tele prodaji svojih proizvoda: izvještaj kreditnog registra mogao bi biti dobiven dovoljno rano, i ako je rezultat skoringa određenog slučaja slab, zahtjev bi mogao biti odbijen veoma brzo. Neki zajmodavci djeluju na određivanju cijena po principu procjene rizika ili diferencijalnih cijena. Ovdje, možda vise nemamo jednostavnu fiksnu cijenu. Umjesto toga, cijena se prilagođava prema riziku (ili potencijalu profita) koji zahtjev nosi. Umjesto jednog limita za odobrenje/odbijanje zahtjeva, vjerovnik ih može imati nekoliko. Može postojati visoki limit za definiranje najboljih zahtjeva kojima bi mogli biti ponuđeni nadograđeni proizvodi, limit za standardni proizvod po nižoj kamatnoj stopi, treći limit za standardni proizvod po standardnoj cijeni, i četvrti limit za degradirani ili proizvod manje vrijednost. U komercijalnom kreditiranju, u određenoj mjeri, u procjenu cijene ulazi i procjena rizika, iako će i druga pitanja, kao što su konkurencija i odnosi s kupcima, imati utjecaja.

## 2.4 Uloga savjetovanja o bodovanju kreditnog rizika

Na svakom tržištu bilo je vrlo malo razvojnih inženjera kreditnog skoringa, ali njihova usluga je bila korisna. Oni su obično imali računala velikih memorija koja su imala dovoljno prostora za obavljanje ogromnih kalkulacija i izračuna.

| **Svrha** | **Primjeri** |
|---|---|
| Identificirati klijenta | Ime, adresa, datum rođenja |
| Mogućnost sklapanja ugovora s kupcem | Ime, adresa, datum rođenja, iznos kredita, raspored otplate |
| Obrada / ocjena zahtjeva | Značajke bodovne kartice |
| Dobivanje izvješća o kreditnom uredu | Ime, adresa, datum rođenja, prethodna adresa |
| Procjena učinkovitosti marketinga | Šifra kampanje, datum primitka zahtjeva, način primanja - pošta, telefon, internet |
| Utjecaj na međubankovni transfer novca | broj bankovnog računa, pojedinosti poslovnice banke |
| Razvitak bodovne kartice | Ostale informacije koje se legalno mogu upotrijebiti sa bodovne kartice. (Zakon može varirati od zemlje do zemlje.) |

Tablica 2.2 *Razlozi za sakupljanje podataka*

Stvari su se promijenile, ali ne u potpunosti. Još uvijek postoje tvrtke za kreditni skoring koje razvijaju nove modele skoringa. One također obavljaju savjetovanje o strategijama i implementacijama i pružaju obuku ukoliko je potrebno. Prednosti prisutnosti vanjskog razvojnog suradnika ostaju - to su veze kreditnih registra i činjenica da vanjski razvojni model za mnoge zajmodavce ima priliku identificirati trendove u industriji ili vrsti portfelja.

S druge strane, značajniji zajmodavci na tržištu su sami izgradili vlastite interne analitičke timove koji mogu razviti bodovnu karticu za puno manji trošak. Rast u snazi kompanije i smanjenje cijena računala olakšali su taj unutarnji razvoj. Kao i korist koja dolazi sa nižim troškovima, interni timovi mogu bolje razumjeti podatke i trebali bi biti bolje pozicionirani unutar strukture da bi mogli predvidjeti izazove u provedbi. Savjetovanja glede skoringa su se pojavila u svrhu da bi se premostio raskorak između

unutarnjeg i vanjskog razvoja. One same ne učestvuju u samom razvoju, nego savjetuju interne razvojne timove i pokušavaju umanjiti pogreške, kako sa praktične tako i sa analitičke strane.

## 2.5 Zahtjevni obrazac

Jasno je da što više podataka trebamo dobiti od podnositelja zahtjeva, toliko će duže vremena trebati za izradu/popunu obrasca za prijavu. Što je obrazac za prijavu duži, to je manja vjerojatnost za podnositelja zahtjeva ili da podnesete zahtjev ili da popuni sve detalje. Stoga je često prisutan pritisak kako bi postupak bio što jednostavniji za podnositelja zahtjeva. Jedan od načina da se to učini je da se obrazac zahtjeva stavi u njegov najmanji / najjednostavniji mogući oblik. Nažalost, ovo ponekad otežava budući razvoj bodovnih kartica.

Drugi način, je pronaći alternativne izvore za iste ili približno ekvivalentne informacije. Kreditni biro bi mogao biti u mogućnosti dostaviti informacije, radije nego ih tražiti od podnositelja zahtjeva. To je osobito slučaj s zajmodavcima koji dijele informacije. U slučaju potvrde duljine boravka na adresi, može se koristit informacija koliko je netko registriran kao glasač na istoj adresi kako bi se ustanovio boravak.

## 2.6 Uloga kreditnog registra

Kreditni registri (ili kreditne agencije) u SAD-u i Velikoj Britaniji su vrlo dobro afirmirani. U drugim zemljama zapadne Europe i ostalim razvijenim zemljama, one su u različitim fazama razvoja. Istočna Europa, na primjer, tek je nedavno počela se suočavati sa pitanjem kako ih razviti. Tamo gdje su dobro uspostavljeni, oni su u državnom vlasništvu ili je mali broj vrlo velikih igrača na tržištu. U SAD-u trenutno postoje tri glavna registra, dok U.K. ima dva. Najpoznatiji kreditni registar u Hrvatskoj poznat je pod nazivom HROK – Hrvatski registar obveza po kreditima.

HROK d.o.o. je tvrtka koju je osnovalo 20 banaka u Republici Hrvatskoj sa svrhom obavljanja djelatnosti potpunog kreditnog registra. Te banke predstavljaju preko 97% tržišta kredita u Republici Hrvatskoj, a očekuje se pristupanje i ostalih banaka. Potpuni kreditni registar je sustav za prikupljanje, obradu i razmjenu informacija o svim kreditnim obvezama klijenata i urednosti njihovog podmirivanja. U registru se obrađuju podaci o svim kreditnim zaduženjima klijenata bankama i stambenih štedionica, bilo da se radi o urednom otplaćivanju kredita ili neurednom.

Da bismo razumjeli ulogu kreditnih registra, treba pogledati kako su nastali. Prije njihove široko rasprostranjene upotrebe, kada se razmatrao zahtjev za zajam, moguće je bilo pisati poslodavcu ili banci za referencu. Svakako, u Velikoj Britaniji, te su reference postale još čuvanije i manje korisne. Nastavno, ako je Banka X, na primjer, postala svjesna činjenice da gospodin Brown aplicira kod Banke Y za kreditnu karticu ili hipoteku, prije odgovora na referencu, Banka X bila je u mogućnosti ponuditi jednu od svojih kreditnih kartica ili hipotekarne pakete. Naravno, referenca od banke bi otkrila samo, u najboljem slučaju, detalje prometa bankovnog računa, generalno mišljenje o karakteru potencijalnog dužnika i bilo kakve druge štetne informacije o strane banke ili podružnice o gospodinu Brown-u. Kao što je raspoloživost kredita bila proširenija, tako su se počeli pojavljivati loši dugovi. Ono što je bilo jako negativno bilo je to što su se ti loši dugovi mogli lako izbjeći jer su indikacije da bi se mogli pojaviti problemi, bili dostupni u vrijeme podnošenja zahtjeva.

Prije samog opisa načina na koji kreditni registri djeluju u UK i USA, trebamo prepoznati vrlo različit položaj USA i UK kreditnih registra. Osnovni dio zakona koji regulira kreditne registre u Velikoj Britaniji jest Zakon o zaštiti podataka. U SAD-u, jedan od ključnih zakona je Zakon o slobodi informiranja. Dakle, dva režima počinju sa suprotnih krajeva spektra. Grubo rečeno, u SAD-u informacije su dostupne, osim ako postoji dobar razlog za njihovo ograničavanje. Isto tako, grubo rečeno, u Velikoj Britaniji, informacije su ograničene ili zaštićene, osim ako postoji dobar razlog za njihovu dostupnost. Tako američki kreditni registri imaju veće bogatstvo informacija o potrošačima. U oba okruženja, agencije za kreditne reference prve su počele javno akumulirati javno dostupne informacije i stavile ga u središnji fokus. Čak i u 1980-ima to se može postići velikom sobom punom ormarića s karticama. Po primitku upita, agent bi

stavio istražitelja na čekanje i trčao gore-dolje po sobi i pristupao relevantnim karticama po potrebi. Očito, tokom vremena to je postalo kompjuterizirano. To znači da se upit može izvršiti i elektronički, fizički putem ljudskog operatera koji poziva službu za upite ili, češće, gdje dva računala razgovaraju jedni s drugima. Ove javne informacije mogu sad sadržavati i podatke o biračim spiskovima i informacije o javnim dugovima nastalim kroz sudske odluke . Koristeći snagu kompjuterizacije, registri također mogu povezati različite adrese tako da, kada potrošač promjeni adresu, dug i potrošač ostaju povezani.

Registri također djeluju kao agenti za zajmodavce. Zajmodavci doprinose registrima u detaljima  trenutno stanje računa svojih zajmoprimaca. Ovi statusi mogu se pregledavati i koristiti drugim zajmodavcima prilikom razmatranja zahtjeva za kredit. Mogu se pregledavati i koristiti drugim zajmodavcima u marketingu, iako s nekim povećanim ograničenjima u korištenju podataka.

Tri su glavne uloge registra. Prva je skupljanje podataka i tu je jasno da što su podaci potpuniji, to će biti veća njihova vrijednost, jer će se dobiti cjelokupna slika o zaduženosti svakog potrošača. Druga uloga je održavanje tih podataka, što obuhvaća ažuriranje banke podataka kako pristižu novi zapisi, kao i provjeru podataka, te mogućnost ispravljanja eventualnih pogrešaka. U tom postupku, možemo reći da se "podaci" pretvaraju u "informacije" jer se uključuju u postojeće zapise o potrošačima i na taj način se dobiva informacija o stanju zaduženosti. Treća uloga je distribucija informacija u obliku kakvog korisnici trebaju za učinkovito donošenje odluka.

Daljnja usluga koju koriste mnogi zajmodavci je generički rezultat. Ovaj se rezultat izračunava na temelju ocjene koju je izradio ured na temelju iskustva s milijunima aplikacija i milijuna zapisa kreditne povijesti. To je osobito korisno u slučajevima kada zajmodavac nije dovoljno velik da bi sam razvio bodovne kartice za ocjenjivanje vlastitih portfelja ili u prve dvije godine novog proizvoda. Također se koristi za dobivanje ažurnog pogleda na kreditnu poziciju dužnika tako što će uključiti nedavnu kreditnu povijest dužnika sa svim davateljima kredita i bilo kakve upite koji su rezultirali kroz nove zahtjeve za kreditiranje. Doista, neki zajmodavci, osobito u portfeljima kreditnih kartica, kupe ocjenu za svakog od svojih vlasnika kartica svaki mjesec te ih upotrebljavaju za procjenu načina rješavanja slučajeva kod kasnog plaćanja ili prelaza ograničenja, i kada i za koliko povećati limit kreditne kartice.

# Poglavlje 3

# 3 Statističke metode za izgradnju kreditnih bodovnih kartica

## 3.1 Uvod

1950-ih i 1960-ih, kada je kreditni skoring prvi put razvijen, jedine korištene metode bile su statistička diskriminacija i metode klasifikacije. Čak i danas statističke metode su daleko najkorištenije metode za izgradnju kreditnih bodovnih kartica. Njihova je prednost da one dopuštaju korištenje znanja o svojstvima procjenitelja uzoraka i alate o pouzdanim intervalima i testiranje hipoteza u kontekstu kreditnog skoringa. Tako je moguće komentirati vjerojatnu diskriminirajuću moć gradnje kreditne bodovne kartice i relativnu važnost različitih karakteristika (varijabli) koje čine bodovnu karticu u njegovoj diskriminaciji. Te statističke tehnike zatim omogućuju utvrđivanje i uklanjanje nevažnih karakteristika i osiguravaju da sve važne karakteristike ostanu u bodovnoj kartici. Ta se informacija može koristiti i kada se gleda koje bi promjene mogle biti potrebne u pitanjima vezanih za potencijalne zajmoprimce.

Iako su statističke metode bile prve koje su se upotrebljavale za izgradnju sustava skoringa i još uvijek ostale najvažnije metode, došlo je do promjena u metodama koje su se upotrebljavale tijekom 40 godina. U početku su se metode temeljile na metodama diskriminacije koje je predložio Fisher (1936) za opće klasifikacijske probleme. To je dovelo do linearne bodovne kartice na temelju Fisherove linearne diskriminantne funkcije. Pretpostavke koje su bile potrebne kako bi se osiguralo da je to bio najbolji način za diskriminaciju dobrih i loših potencijalnih kupaca bile su iznimno restriktivne i jasno, nisu bile održive u praksi, iako su produkti bodovne kartice bili vrlo snažni. Fisherov pristup mogao bi se promatrati kao oblik linearne regresije i to je dovelo do istrage drugih oblika regresije koji su imali manje restriktivne pretpostavke, kako bi se zajamčila njihova optimalnost i koje bi još uvijek dovele do linearnih pravila za skoring.

Daleko najuspješniji od njih je logistička regresija, koja je preuzeta iz linearne regresije — diskriminantan pristup analizi kao najčešća statistička metoda. Drugi pristup koji je pronašao naklonost tijekom posljednjih 20 godina je klasifikacijsko stablo ili rekurzivni particionirani pristup. Takav pristup dijeli skup podnositelja zahtjeva u niz različitih podskupina, ovisno o njihovim atributima, a zatim se svaka podskupina klasificira kao zadovoljavajuća ili nezadovoljavajuća. Iako to ne daje težinu svakom od atributa kao što daje linearna bodovna kartica, rezultat je isti – metoda za odlučivanje hoće li novi podnositelj zahtjeva biti klasificiran kao zadovoljavajući ili nezadovoljavajući. Sve ove metode su korištene u praksi kako bi se osmislile bodovne kartice za komercijalne organizacije, ali još uvijek postoji dosta eksperimentiranja u korištenju statističkih metoda. Jedna od tih eksperimentalnih metoda je neparametarski pristup temeljen na najbližim susjedima.

U ovom poglavlju preispitujemo svaku od tih metoda i statističku pozadinu koja podupire te metode. Počinjemo s diskriminantnom analizom. U sljedeća tri odlomka razmatra se na koji način je linearna diskriminantna funkcija klasificirana po tri različita pristupa problemu. Prvi pristup je donošenje odluka gdje se traži pravilo koje minimizira očekivani trošak prilikom odlučivanja hoće li se prihvatiti novi kupac. Drugi pristup je onaj koji je motivirao Fisherov izvorni rad, pokušavajući pronaći funkciju koja se najbolje razlikuje između dviju skupina zadovoljavajućih (dobrih) i nezadovoljavajućih (loših) kupaca. Treći pristup je razmotriti regresijsku jednadžbu, koja pokušava pronaći najbolju procjenu vjerojatnosti da klijent bude dobar. Svaki od tih pristupa dolazi do iste linearne metode rezultata za određivanje dobrih od loših klijenata. Konačno, postoji rasprava o tome kako bi se neke od metoda mogle proširiti, kako bi sustavi za kreditni skoring imali mogućnost klasificirati klijente u više od dviju kategorija dobrih i loših.

## 3.2 Diskriminantna analiza: pristup teoriji odluke

Postupak odobravanja kredita dovodi do izbora između dviju radnji, dati novom podnositelju zahtjeva kredit ili odbiti taj zahtjev. Kreditni skoring nastoji pomoći ovoj

odluci pronalaženjem onoga što bi bilo najbolje pravilo za primjenu na uzorku ranijih podnositelja zahtjeva. Prednost toga je to što znamo kako su ti podnositelji zahtjeva naknadno provedeni. Ako postoje samo dva moguća djelovanja, prihvati ili odbaci, onda nema prednosti u klasificiranju te izvedbe u više od dvije klase, dobre i loše. Dobro je bilo koja izvedba koja je prihvatljiva organizaciji pozajmljivanja (organizaciji koja pozajmljuje), dok je loše, izvedba u kojoj zajmodavac odbacuje podnositelja zahtjeva. U nekim organizacijama, za lošu izvedbu se uzima izvedba kod koje nedostaje određeni broj uzastopnih plaćanja, dok je u drugima ukupan broj propuštenih plaćanja koji su važni.

U tom pristupu postoji inherentna pristranost u tome da je uzorak uzet iz prethodnih podnositelja zahtjeva kojima je odobren kredit, te ne postoje informacije o izvedbi o podnositeljima zahtjeva koji su u prošlosti bili odbijeni. Tako je uzorak reprezentativan za one koji su prihvaćeni u prošlosti i nije reprezentativan za one koji su se primjenjivali u prošlosti. Na primjer, može se odlučiti zatražiti dodatne informacije o podnositelju zahtjeva ili da kreditni analitičar razmotri zahtjev podnositelja. Međutim, te varijacije imaju više veze s načinima na koje zajmodavac organizira donošenje odluka, a konačna odluka u svakom slučaju bit će, prihvatiti ili odbiti. Nije vrijedno pokušavati klasificirati koji će podnositelji zahtjeva biti u određenoj skupini, kada je odluka o tome jesu li u toj skupini u potpunosti na zajmodavcu. Stoga je čak i u tim višestrukim akcijskim procesima razumno klasificirati podnositelje zahtjeva u samo dvije skupine, dobre i loše, budući da će konačna odluka rezultirati jednim od tih dviju mjera.

Neka je $X = (X_1, X_2, \ldots, X_p)$ skup p nasumičnih varijabli koje opisuju informacije dostupne o podnositelju zahtjeva za kredit iz obrasca zahtjeva i putem provjere kreditnog referentnog registra. Naravno, danas možda ne postoji fizički oblik, već se detalji mogu sakupljati na zaslonu, putem interneta ili bi ih mogao zapisati član zajmodavca u telefonskom pozivu. Koristimo riječi varijabla i karakteristika naizmjenično, da bi opisali tipičan $X_i$ kada želimo naglasiti nasumičnu prirodu ove informacije između podnositelja zahtjeva i potonjeg. Stvarna vrijednost varijabli za određenog podnositelja zahtjeva označena je sa $x = (x_1, x_2, \ldots, x_p)$. U terminologiji kreditnog skoringa različite moguće vrijednosti ili odgovori $x_i$, na varijablu $X_i$, nazivaju se atributi te karakteristike. Dakle,

ako je tipična karakteristika „boravišni status podnositelja“ , onda njegovi atributi mogu biti „vlasnik“, „najam (nenamješten)“, „najam (namješten)“, „živi s roditeljima“, ili drugi. Različiti zajmodavci mogu imati različite grupe atributa za iste karakteristike. Tako drugi zajmodavac može odlučiti klasificirati stambeni status u „vlasnik bez hipoteke“, „vlasnik s hipotekom“, „iznajmljivanje nenamještene imovine“, „iznajmljivanje namještene imovine“, „zakup nekretnine“, „mobilna kuća“, „pruženi smještaj“, „živi s roditeljima“, „živi s drugim (osim roditelja)“ ili drugi statusi. Nije neuobičajeno da se atribut i karakteristika miješaju. Jednostavan način za zapamtiti što je atribut, a što je karakteristika, je taj da je atribut odgovor na pitanje obrasca zahtjeva i karakteristika je pitanje koje je pitano.

Vratimo se na odluku koja mora biti donesena od strane organizacije koja daje zajmove. Pretpostavimo da je A skup svih mogućih vrijednosti koje varijable zahtjeva $X = (X_1, X_2, \ldots, X_p)$ mogu poprimiti, tj. svi različiti načini na koji se obrazac zahtjeva može ispuniti. Cilj je pronaći pravilo koje dijeli skup $A$ u dvije podskupine, $A_G$ $i$ $A_B$, kako bi se minimizirao očekivani trošak zajmodavca. Klasificiranje podnositelja zahtjeva, čiji su odgovori u $A_G$ , kao "dobro" , znači prihvaćanje, dok klasificiranje onih, čiji su odgovori u $A_B$, kao "loše", znači odbijanje. Dvije vrste troškova, odgovaraju dvjema vrstama pogrešaka koje se mogu dogoditi u ovoj odluci. Postoji mogućnost odbacivanja osobe zbog klasificiranja nekoga tko je „dobar“, kao „loš“. U tom slučaju izgubi se potencijalna zarada od tog podnositelja zahtjeva. Pretpostavimo za sada da je očekivani profit isti za svakog podnositelja zahtjeva i označimo ga sa $L$. Druga pogreška je klasificirati nekoga tko je „loš“ kao „dobar“ i tako prihvatiti podnositelja zahtjeva. U tom slučaju nastati će dug kada kupac ne ispuni svoje obveze prema kreditu.

Pretpostavljamo da je očekivani dug koji je nastao, isti za sve kupce i označimo ga sa $D$. Pretpostavimo da je $p_G$ udio podnositelja zahtjeva koji su „dobri“. Isto tako, neka $p_B$ bude udio podnositelja zahtjeva koji su „loši“.

Pretpostavimo da karakteristike zahtjeva imaju konačan broj diskretnih atributa, tako da je $A$ konačan i postoji samo konačan broj različitih atributa $x$. To je kao da se kaže da postoji samo konačan broj načina ispunjavanja obrasca zahtjeva. Neka je $p(x|G)$

vjerojatnost da će „dobar“ podnositelj imati atribute *x*. To je uvjetna vjerojatnost i predstavlja omjer:

$$p(x|G) = \frac{Vjerojatnost(podnositelj\ zahtjeva\ je\ "dobar"\ i\ ima\ atribute\ x)}{Vjerojatnost(podnositelj\ zahtjeva\ je\ "dobar")} \quad (3.1)$$

Slično, definirajmo $p(x|B)$ kao vjerojatnost da će „loš“ podnositelj imati atribute *x*. Ako je $q(G|x)$ definiran kao vjerojatnost da je netko s atributima *x* „dobar“, onda vrijedi sljedeće:

$$q(G|x) = \frac{Vjerojatnost(podnositelj\ zahtjeva\ ima\ atribute\ x\ i\ "dobar"\ je)}{Vjerojatnost(podnositelj\ zahtjeva\ ima\ atribute\ x)} \quad (3.2)$$

I ako je $p(x) = Vjerojatnost(podnositelj\ zahtjeva\ ima\ atribute\ x)$ , onda iz

(3.1) i (3.2) slijedi:

$$Vjerojatnost(podnositelj\ zahtjeva\ ima\ atribute\ x\ i\ "dobar"\ je) = q(G|x)p(x) = p(x|G)p_G \quad (3.3)$$

Stoga, dolazimo do Bayes-ovog teorema, koji kaže

$$q(G|x) = \frac{p(x|G)p_G}{p(x)} \quad (3.4)$$

Sličan rezultat dobijemo i za $q(B|x)$, vjerojatnost da je netko sa atributima x „loš“. Naime,

$$q(B|x) = \frac{p(x|B)p_B}{p(x)} \quad (3.5)$$

Iz (3.4) i (3.5) slijedi

$$\frac{q(G|x)}{q(B|x)} = \frac{p(x|G)p_G}{p(x|B)p_B} \tag{3.6}$$

Očekivani trošak po podnositelju zahtjeva, ako prihvaćamo podnositelje zahtjeva s atributima u $A_G$ i odbacujemo one s atributima u $A_B$, je

$$\begin{aligned} L\sum_{x\in A_B} p(x|G)p_G + D\sum_{x\in A_B} p(x|B)p_B \\ = L\sum_{x\in A_B} q(G|x)p(x) + D\sum_{x\in A_B} q(B|x)p(x) \end{aligned} \tag{3.7}$$

Pravilo koje minimizira očekivani trošak je jednostavno. Razmotrimo koja su to dva troška ako kategoriziramo određeni $x = (x_1, x_2, \ldots, x_p)$ u $A_G$ ili $A_B$. Ako se *x* klasificira u $A_G$, onda postoji trošak samo ako je *x* „loš“ i u tom slučaju očekivani trošak je $Dp(x|B)p_B$. Ako je *x* klasificiran u $A_B$, postoji trošak samo ako je *x* „dobar“ i tada je očekivani trošak $Lp(x|G)p_G$. Slijedi da se *x* klasificira u $A_G$ ako vrijedi $Dp(x|B)p_B \leq Lp(x|G)p_G$. Stoga je pravilo odlučivanja koje minimizira očekivane troškove dan sljedećom jednadžbom:

$$\begin{aligned} A_G = \{x|D\cdot p(x|B)p_B \leq L\cdot p(x|G)p_G\} = \left\{x|\frac{D}{L} \leq \frac{p(x|G)p_G}{p(x|B)p_B}\right\} \\ = \left\{x|\frac{D}{L} \leq \frac{q(G|x)}{q(B|x)}\right\} \end{aligned} \tag{3.8}$$

gdje zadnja jednakost slijedi iz (3.6).

Jedna kritika gore navedenog kriterija je da ovisi o očekivanim troškovima *D* i *L*, koji možda nisu poznati. Dakle, umjesto minimiziranja očekivanog troška može se nastojati minimizirati vjerojatnost da će se počiniti jedna vrsta pogreške, dok se vjerojatnost da će se počiniti druga vrsta pogreške drži na dogovorenom nivou. U kontekstu odobravanja kredita očita stvar za učiniti, jest minimizirati razinu neispunjavanja obveza, dok se postotak prihvaćenih podnositelja zahtjeva drži na nekom dogovorenom nivou.

Potonji zahtjev jednak je održavanju vjerojatnosti odbijenih „dobrih“ podnositelja zahtjeva na nekoj fiksnoj razini. Pretpostavimo da je stopa prihvaćanja $a$. Onda $A_G$ mora zadovoljiti

$$\sum_{x \in A_G} p(x|G)p_G + \sum_{x \in A_G} p(x|B)p_B = a \tag{3.9}$$

gdje se istovremeno minimizira zadana stopa

$$\sum_{x \in A_G} p(x|B)p_B \, .$$

Ako definiramo $b(x) = \; p(x|B)p_B$, za svaki $x \in A$, tada želimo naći skup $A_G$ tako da se

$$\sum_{x \in A_G} b(x) = \sum_{x \in A_G} \left(\frac{b(x)}{p(x)}\right) \cdot p(x)$$

minimizira do

$$\sum_{x \in A_G} p(x) = a \, . \tag{3.10}$$

Koristeći Lagrange-ov množitelj, može se vidjeti da to mora biti skup atributa *x*, gdje je

$$\frac{b(x)}{p(x)} \leq c,$$

gdje je $c$ izabran tako da je zbroj $p(x)$, koji zadovoljava to ograničenje, jednako $a$. Stoga vrijedi

$$A_G = \left\{x| \frac{b(x)}{p(x)} \leq c\right\} = \{x|q(B|x) \leq c\} = \left\{x| \frac{1-c}{c} \leq \frac{p(x|G)p_G}{p(x|B)p_B}\right\} \tag{3.11}$$

gdje druga nejednadžba slijedi iz definicije $p(x)$ i $b(x)$.

Stoga je oblik pravila donošenja odluka, prema ovim kriterijima, jednak kao i kod (3.8) prema kriteriju očekivanog troška za neke odgovarajuće izbore troškova *D* i *L*.

Cijela analiza mogla bi se ponoviti pod pretpostavkom da su karakteristike zahtjeva neprekidne i nediskretne slučajne varijable. Jedina razlika bi bila da se uvjetne distribucije funkcija $p(x|G)$, $p(x|B)$ zamjenjuju se funkcijama uvjetnih gustoća $f(x|G)$, $f(x|B)$ i sume se zamjenjuju integralima. Dakle, očekivani trošak, ako se set $A$ podijeli u setove $A_G$ i $A_B$ te prihvaća samo one u $A_G$, postaje

$$L\int_{X\in A_B} f(x|G)p_G\,dx + D\int_{X\in A_G} f(x|B)p_B\,dx\,, \tag{3.12}$$

a pravilo odluke koje minimizira ovu jednadžbu jednako je (3.8). Naime,

$$A_G = \{x|Df(x|B)p_B \le Lf(x|G)p_G\} = \left\{x|\frac{Dp_B}{Lp_G} \le \frac{f(x|G)}{f(x|B)}\right\} \tag{3.13}$$

### 3.2.1 Univarijatni normalni slučaj

Razmotrimo najjednostavniji mogući slučaj u kojem postoji samo jedna neprekidna varijabla *X* i njena distribucija među „dobrim“ $f(x|G)$ je normalna sa očekivanjem $\mu_G$ i varijancom $\sigma^2$, dok je distribucija među „lošima“ normalna sa očekivanjem $\mu_B$ i varijancom $\sigma^2$. Tada slijedi

$$f(x|G) = (2\pi\sigma^2)^{-\frac{1}{2}} exp\left(\frac{-(x-\mu_G)^2}{2\sigma^2}\right),$$

pa pravilo (3.13) postaje

$$\begin{gathered}\frac{f(x|G)}{f(x|B)} = \frac{exp\left(\frac{-(x-\mu_G)^2}{2\sigma^2}\right)}{exp\left(\frac{-(x-\mu_B)^2}{2\sigma^2}\right)} = exp\left(\frac{-(x-\mu_G)^2 + -(x-\mu_B)^2}{2\sigma^2}\right) \ge \frac{Dp_B}{Lp_G} \\ \Rightarrow x(\mu_G - \mu_B) \ge \frac{{\mu_G}^2 - {\mu_B}^2}{2} + \sigma^2 log\left(\frac{Dp_B}{Lp_G}\right).\end{gathered} \tag{3.14}$$

Stoga, pravilo postaje „prihvati ako je vrijednost *x* dovoljno velika“.

### 3.2.2 Multivarijatni normalni slučaj s zajedničkom kovarijancom

Jedan realističniji primjer je kada imamo p karakteristika u zahtjevu sa informacijama te ishodi među „dobrima“ i „lošima“ tvore multivarijatnu normalnu distribuciju. Pretpostavimo da je $\mu_G$ očekivanje među „dobrima“, te $\mu_B$ očekivanje među „lošima“, sa zajedničkom kovarijacijskom matricom $\Sigma$. To znači da je $E(X_i|G) = \mu_{G,i}$ , $E(X_i|B) = \mu_{B,i}$ i $E(X_iX_j|G) = E(X_iX_j|B) = \Sigma_{ij}$.

Odgovarajuća funkcija gustoće u ovom slučaju je

$$f(x|G) = (2\pi)^{-\frac{p}{2}}(det(\Sigma))^{-\frac{1}{2}}exp\left(\frac{-(x-\mu_G)\Sigma^{-1}(x-\mu_G)^T}{2}\right), \qquad (3.15)$$

gdje je $(x-\mu_G)$ vektor sa jednim retkom i $p$ stupaca, a $(x-\mu_G)^T$ je transponirani $(x-\mu_G)$ sa istim brojevima reprezentiranim u vektoru sa $p$ redaka i jednim stupcem. Prateći izračun iz (3.14) dobijemo

$$\begin{gathered}\frac{\mathrm{f(x|G)}}{\mathrm{f(x|B)}} \geq \frac{\mathrm{Dp_B}}{\mathrm{Lp_G}} \\ \Rightarrow \mathrm{x}\Sigma^{-1}(\mu_\mathrm{G}-\mu_\mathrm{B})^\mathrm{T} \geq \frac{\mu_\mathrm{G}\Sigma^{-1}{\mu_\mathrm{B}}^\mathrm{T} - \mu_\mathrm{B}\Sigma^{-1}{\mu_\mathrm{G}}^\mathrm{T}}{2} + \log\left(\frac{\mathrm{Dp_B}}{\mathrm{Lp_G}}\right).\end{gathered} \qquad (3.16)$$

Lijeva strana implikacije (3.16) je ponderirana suma vrijednosti varijabli izgleda : $x_1w_1 + x_2w_2 + \ldots + x_pw_p$. Dok je desna strana konstanta. Stoga, iz (3.16) slijedi pravilo linearnog skoringa, poznato kao linearna diskriminantna funkcija.

Gornji primjer je pretpostavio da su očekivanja i kovarijance iz distribucije koju znamo. To je rijetko slučaj, te je normalnije da ih zamijenimo sa procjenama, gdje su procijenjena očekivanja $m_G$ i $m_B$, a procijenjena kovarijacijska matrica je S. Pravilo donošenje odluke (3.16) sada postaje

$$xS^{-1}(m_G - m_B)^T \geq \frac{m_G S^{-1} {m_B}^T - m_B S^{-1} {m_G}^T}{2} + log\left(\frac{Dp_B}{Lp_G}\right). \tag{3.17}$$

### 3.2.3 Multivarijatni normalni slučaj s različitom kovarijacijskom matricom

Još jedna očita restrikcija prethodnog slučaja jest da su kovarijacijske matrice jednake za populaciju „dobrih" i „loših". Pretpostavimo da je kovarijacijska matrica za populaciju „dobrih" $\Sigma_G$, a za populaciju „loših" je $\Sigma_B$. U ovom slučaju (3.16) postaje

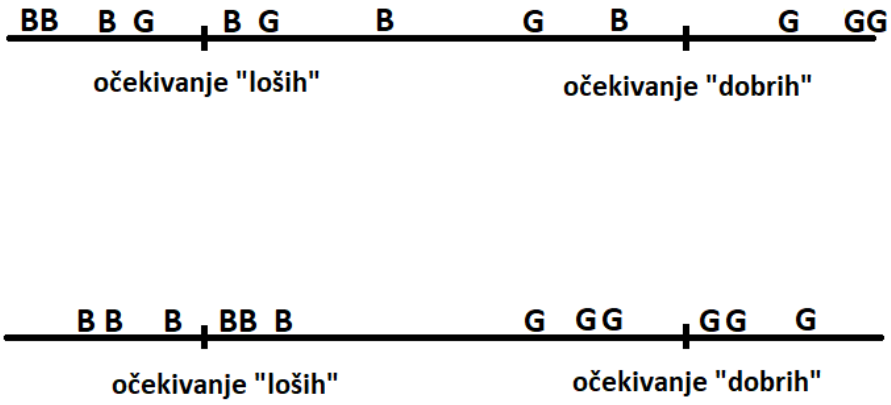
$$\begin{aligned} &\frac{f(x|G)}{f(x|B)} \geq \frac{D \cdot p_B}{L \cdot p_G} \\ \Rightarrow\ & exp\left\{-\frac{1}{2}((x-\mu_G){\Sigma_G}^{-1}(x-\mu_G)^T - (x-\mu_B){\Sigma_B}^{-1}(x-\mu_B)^T)\right\} \\ &\geq \frac{D \cdot p_B}{L \cdot p_G} \\ \Rightarrow\ & (x({\Sigma_G}^{-1} - {\Sigma_B}^{-1})x^T + 2x({\Sigma_G}^{-1}{\mu_G}^T - {\Sigma_B}^{-1}{\mu_B}^T) \\ &\geq (\mu_G {\Sigma_G}^{-1}{\mu_G}^T + \mu_B {\Sigma_B}^{-1}{\mu_B}^T) + 2log\left(\frac{D \cdot p_B}{L \cdot p_G}\right). \end{aligned} \tag{3.18}$$

Lijeva strana je kvadratna u vrijednostima $x_1, x_2, \ldots, x_p$. Čini se da je to općenitije pravilo donošenja odluka, pa se očekuje da je bolje nego linearno pravilo donošenja odluka. Međutim, u praksi je potrebno procijeniti dvostruki broj parametra $\Sigma_B$ i $\Sigma_G$.

Dodatna nesigurnost povezana sa tim procjenama čini ovo kvadratno pravilo donošenja odluka manje snažnim od linearnog pravila, te u većini slučajeva se ne isplati dobivanje malo bolje točnosti koja može proizaći iz kvadratnog pravila.

## 3.3 Diskriminantna analiza: Razdvajanje dviju skupina

U Fisher-ovom izvornom radu (1936),u kojem je obznanio linearnu diskriminantnu funkciju, cilj je bio pronaći kombinaciju varijabli koja najbolje odvaja dvije skupine čija su obilježja bila dostupna. U kontekstu ocjenjivanja kredita, dvije skupine su klasificirane od strane zajmodavca kao „dobri“ i „loši“, a karakteristike su detalji zahtjeva za prijavu i informacije iz kreditnog registra. Neka je $Y = w_1X_1 + w_2X_2 + \ldots + w_pX_p$ bilo koja linearna kombinacija karakteristika $X = (X_1, X_2, \ldots, X_p)$. Jedna očita mjera razdvajanja je kako su različita očekivanja od $Y$ za dvije različite grupe „dobrih“ i „loših“ u uzorku. Dakle, gledamo razliku između $E(Y|G)$ i $E(Y|B)$, te biramo težinu $w_i$ t.d. $\sum_i w_i = 1$, koja maksimizira tu razliku. Međutim, ovo je malo naivno, jer to govori da su grupe na slici (Slika 3.3.1) jednako udaljene. Ono što taj primjer pokazuje jest da bi trebalo omogućiti i koliko se svaka od dviju skupina blisko okuplja kada se raspravlja o njihovom odvojenosti.




Slika 3.3.1 *Dva primjera kada su očekivanja "dobrih" i "loših" jednaka*

Fisher je predložio da ako pretpostavimo da dvije grupe imaju zajedničku uzoračku varijancu, onda je razumna mjera razdvajanja

$$M = \frac{Udaljenost\ između\ procijenjenih\ očekivanja\ uzoraka\ dviju\ grupa}{(uzoračka\ varijanca\ svake\ grupe)^{\frac{1}{2}}}.$$

Kako bi mjerna skala bila neovisna, može se cijeli izraz podijeliti sa korijenom uzoračke varijance. Ako se promijeni varijabla $Y$ u $cY$, tada se mjera $M$ ne mijenja.

Pretpostavimo da su procijenjena očekivanja dana sa $m_G$ i $m_B$ za „dobre“ i „loše“, te je $S$ zajednička uzoračka varijanca. Ako je $Y = w_1X_1 + w_2X_2 + \ldots + w_pX_p$, tada je pridružena razdvajajuća udaljenost $M$ dana sa

$$M = w^T \frac{m_G - m_B}{(w^T \cdot S \cdot w)^{\frac{1}{2}}}. \tag{3.19}$$

To slijedi iz $E(Y|G) = wm_G{}^T$, $E(Y|B) = wm_B{}^T$ i $Var(Y) = wSw^T$. Deriviranjem po $w$ i izjednačavanjem te derivacije s 0, vidimo da je vrijednost $M$ maksimalna kada

$$\frac{\mathrm{m_G} - \mathrm{m_B}}{(\mathrm{w^T \cdot S \cdot w})^{\frac{1}{2}}} - \frac{(\mathrm{w \cdot (m_G - m_B)^T (Sw^T)})}{(\mathrm{w^T \cdot S \cdot w})^{\frac{1}{2}}} = 0, \tag{3.20}$$

$$\Rightarrow (\mathrm{m_G - m_B})(\mathrm{w \cdot S \cdot w^T}) = (\mathrm{Sw^T})(\mathrm{w \cdot (m_G - m_B)^T}).$$

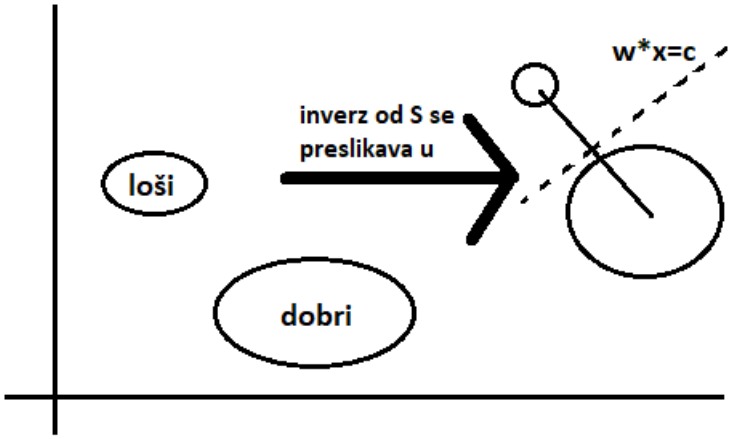
U stvari, sve to pokazuje da je to prekretnica, ali činjenica da drugi derivat od $M$ po $w$ formira pozitivnu određenu matricu jamči da je to minimum. Budući da je $\frac{w \cdot S \cdot w^T}{w \cdot (m_G - m_B)^T}$ skalar $\lambda$, to daje

$$w^T \alpha (S^{-1}(m_G - m_B)^T). \tag{3.21}$$

Stoga su težine iste kao i one dobivene u (3.17), iako ovaj put nije bilo pretpostavke o normalnosti. To je samo najbolji separator „dobrih“ i „loših“ pod ovim kriterijem, bez obzira na njihovu distribuciju. Ovaj rezultat je dobar za sve distribucije, jer mjera udaljenosti M uključuje samo očekivanje i varijancu distribucija i tako daje iste rezultate za sve distribucije s istim očekivanjem i varijancom.

Slika 3.3.2 pokazuje grafički ono što kartica bodovanja (3.21) nastoji učiniti. $S^{-1}$ pojam standardizira dvije grupe tako da imaju istu disperziju u svim smjerovima. w je tada

smjer pridružen očekivanjima „dobrih“ i „loših“ nakon što su bili standardizirani. Tada crta okomita na ovu crtu spaja dva očekivanja. Rezultat rezanja je sredina udaljenosti između očekivanja standardiziranih grupa.




Slika 3.3.2 *Linija koja se podudara sa kreditnom bodovnom karticom*

## 3.4 Diskriminantna analiza: Oblik linearne regresije

Još jedan pristup kreditnom skoringu, koji također proizlazi iz linearne diskriminatne funkcije, jest linearna regresija. Ovdje se pokušava naći najbolja linearna kombinacija karakteristika $w_0 + w_1X_1 + w_2X_2 + \ldots + w_pX_p = w^* \cdot X^{*T}$, gdje je

$$w^* = (w_0, w_1, w_2, \ldots, w_p), \qquad X^* = (1, X_1, X_2, \ldots, X_p)$$

što objašnjava vjerojatnost neispunjavanja obveza podnositelja. Ako je $p_i$ vjerojatnost da podnositelj zahtjeva i uzorka nije ispunio obveze, želimo pronaći $w^*$ koji najbolje aproksimira

$$p_i = w_0 + x_{i1}w_1 + x_{i2}w_2 + \ldots + x_{ip}w_p, \qquad \text{za svaki } i \ . \tag{3.22}$$

Pretpostavimo da je $n_G$ dio uzorka koji je „dobar“. Zbog lakše notacije pretpostavljamo da su to prvi $n_G$ u uzorku. Stoga slijedi $p_i = 1$ za $i = 1, \ldots, n_G$ . Ostalih

$n_B$ uzorka $i = n_G + 1, \ldots, n_G + n_B$ su „loši“, stoga za njih vrijedi $p_i = 0$ gdje je $n_G + n_B = n$.

U linearnoj regresiji odabiremo koeficijent koji minimizira srednju kvadratnu pogrešku između lijeve i desne strane jednadžbe (3.22). To odgovara minimizaciji

$$\sum_{i=1}^{n_G} \left(1 - \sum_{j=0}^{p} w_j x_{ij}\right)^2 + \sum_{n_G+1}^{n_G+n_B} \left(\sum_{j=0}^{p} w_j x_{ij}\right)^2. \tag{3.23}$$

Ako to zapišemo kao vektore, (3.22) se može zapisati na sljedeći način:

$$\begin{pmatrix} 1 & X_G \\ 1 & X_B \end{pmatrix} \begin{pmatrix} w_0 \\ w \end{pmatrix} = \begin{pmatrix} \mathbf{1}_G \\ 0 \end{pmatrix} \quad \text{ili } Yw^T = b^T, \tag{3.24}$$

gdje je

$$Y = \begin{pmatrix} \mathbf{1}_G & X_G \\ \mathbf{1}_B & X_B \end{pmatrix}$$

matrica sa *n* redaka i $(p + 1)$ stupaca,

$$X_G = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{n_G 1} & \cdots & x_{n_G p} \end{pmatrix}$$

je $n_G \times p$ matrica,

$$X_B = \begin{pmatrix} x_{n_G+11} & \cdots & x_{n_G+1p} \\ \vdots & \vdots & \vdots \\ x_{n_G+n_B 1} & \cdots & x_{n_G+n_B p} \end{pmatrix}$$

je $n_B \times p$ matrica, a

$$b^T = \begin{pmatrix} \mathbf{1}_G \\ 0 \end{pmatrix},$$

gdje je $\mathbf{1}_\mathrm{G}(\mathbf{1}_B)$ $1 \times n_G(n_B)$ vektor u kojem su sve stavke jedinice.

Pronalaženje koeficijenata linearne regresije odgovara, kao i u jednadžbi (3.23), minimizaciji

$$(Yw^T - b^T)^T(Yw^T - b^T) \quad . \tag{3.25}$$

Deriviranjem po *w*, to je minimizirano kada je derivacija jednaka 0, t.d.

$$Y^T(Yw^T - b^T) = 0 \quad ili \quad Y^TYw^T = Y^Tb^T,$$

$$Y^Tb^T = \begin{pmatrix} 1 & 1 \\ X_G & X_B \end{pmatrix}\begin{pmatrix} \mathbf{1}_G \\ 0 \end{pmatrix} = \begin{pmatrix} n_G \\ n_G m_G \end{pmatrix},$$

$$Y^TY = \begin{pmatrix} 1 & 1 \\ X_G & X_B \end{pmatrix}\begin{pmatrix} 1 & X_G \\ 1 & X_B \end{pmatrix} = \begin{pmatrix} n & n_G m_G + n_B m_B \\ n_G {m_G}^T + n_B {m_B}^T & {X_G}^T X_G + {X_B}^T X_B \end{pmatrix}. \tag{3.26}$$

Ako u svrhu obrazloženja označimo procijenjeno očekivanje kao stvarno očekivanje, tada dobijemo

$${X_G}^T X_G + {X_B}^T X_B = nE[X_i X_j] = nCov(X_i, X_j) + n_G m_G {m_G}^T + n_B m_B {m_B}^T.$$

Ako je S uzoračka kovarijacijska matrica, dobijemo

$${X_G}^T X_G + {X_B}^T X_B = nS + n_G m_G {m_G}^T + n_B m_B {m_B}^T. \tag{3.27}$$

Koristeći (3.27), možemo proširiti (3.26) do

$$nw_0 + (n_G m_G + n_B m_B)w^T = n_G,$$

$$\Rightarrow (n_G {m_G}^T + n_B {m_B}^T)w_0 + (nS + n_G m_G {m_G}^T + n_B m_B {m_B}^T)w^T = n_G {m_G}^T. \tag{3.28}$$

Ako supstituiramo prvu jednakost iz (3.28) u drugu jednakost, dobijemo

$$\begin{aligned} &\left((n_G {m_G}^T + n_B {m_B}^T)(n_G - (n_G m_G + n_B m_B)w^T)/n\right) \\ &\qquad + (n_G m_G {m_G}^T + n_B m_B {m_B}^T)w^T + nSw^T = n_G {m_G}^T. \end{aligned} \tag{3.29}$$

Stoga vrijedi sljedeće:

$$\left(\frac{n_G n_B}{n}\right)(m_G - m_B)w^T + nSw^T = \left(\frac{n_G n_B}{n}\right)(m_G - m_B)^T. \quad (3.30)$$

Stoga je

$$Sw^T = c(m_G - m_B)^T \ . \quad (3.31)$$

(3.29) nam daje najbolji odabir $w = (w_1, w_2, \ldots, w_p)$ za koeficijente linearne regresije. To je isti w kao i u (3.21) linearne diskriminante funkcije. Ovaj pristup pokazuje da možemo dobiti koeficijente linearne kreditne bodovne kartice manje-kvadratnim pristupom od linearne regresije.

Imamo očite lijeve strane u regresijskoj jednadžbi (3.22), gdje „dobri“ poprimaju vrijednost 1, a „loši“ vrijednost 0. To nam daje skup konstanti koje ćemo označiti sa $w(1,0)^*$. Ako uzmemo bilo koje druge vrijednosti, tako da „dobri“ imaju lijevu stranu $g$, a „loši“ imaju lijevu stranu $b$, tada se koeficijenti $w(g,b)^*$ u regresiji razlikuju samo u konstanti $w_0$, jer je

$$w(a,b)^* = b + (g - b)w(1,0)^*. \quad (3.32)$$

## 3.5 Logistička regresija

Regresijski pristup linearnoj diskriminaciji ima jednu očiglednu manu. U (3.22) desna strana može poprimiti bilo koju vrijednost od $-\infty$ do $+\infty$, ali lijeva strana je vjerojatnost. Stoga poprima samo vrijednosti između 0 i 1. Bilo bi bolje da je lijeva strana funkcija $p_i$, koja bi mogla poprimiti širi raspon vrijednosti. Tada ne bi bio problem da sve točke podataka imaju vrlo slične vrijednosti zavisnih varijabli ili da regresijska jednadžba predviđa vjerojatnosti koje su manje od 0 ili veće od 1. Jedna takva funkcija je logaritam vjerojatnosti . To dovodi do pristupa logističkoj regresiji, na kojem je Wiginton (1980) bio jedan od prvih koji je objavio rezultate kreditnog skoringa. U logističkoj regresiji logaritam vjerojatnosti se izjednačava sa linearnom kombinacijom karakteristika, t.d.

$$log\left(\frac{p_i}{1 - p_i}\right) = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_p x_p = wx^T. \quad (3.33)$$

Budući da $\frac{p_i}{1- p_i}$ poprima vrijednosti između 0 i ∞, tada $log\left(\frac{p_i}{1- p_i}\right)$ poprima vrijednosti između $-\infty$ i $+\infty$. Ako eksponiramo obje strane u izrazu (3.31), dobijemo jednadžbu

$$p_i = \frac{e^{wx}}{1 + e^{wx}}.$$

Ovo je pretpostavka logističke regresije. Zanimljivo je napomenuti da ako pretpostavljamo da je distribucija vrijednosti karakteristika „dobrih“ i „loših“ multivarijatna normalna, kao što je predloženo u dijelu 3.2, onda ovaj primjer zadovoljava pretpostavku logističke regresije. Opet pretpostavimo da su očekivanja $\mu_G$ među „dobrim“ i $\mu_B$ među „lošim“ sa zajedničkom kovarijskom matricom $\Sigma$. To znači da je $E(X_i|G) = \mu_{G,i}$, $E(X_i|B) = \mu_{B,i}$, te $E\left(X_iX_j|G\right) = E\left(X_iX_j|B\right) = \Sigma_{i,j}$.

Odgovarajuća funkcija gustoće u ovom slučaju (kao i u (4.15)) je

$$f(x|G) = (2\pi)^{-\frac{p}{2}}(det(\Sigma))^{-\frac{1}{2}}exp\left(\frac{-(x-\mu_G)\Sigma^{-1}(x-\mu_G)^T}{2}\right), \qquad (3.34)$$

gdje je $(x-\mu_G)$ vektor sa jednim retkom i p stupaca, a $(x-\mu_G)^T$ je transponirani $(x-\mu_G)$ sa istim brojevima reprezentiranim u vektoru sa $p$ redaka i jednim stupcem. Pretpostavimo da je $p_G$ udio podnositelja zahtjeva koji su „dobri“. Isto tako, neka $p_B$ bude udio podnositelja zahtjeva koji su „loši“. Tada je logaritam vjerojatnosti za klijenta i koji ima karakteristike *x*

$$\begin{aligned} log\left(\frac{p_i}{1-\ p_i}\right) &= log\left(\frac{p_G f(x|G)}{p_B f(x|B)}\right) \\ &= x\Sigma^{-1}2(\mu_B - \mu_G)^T + (\mu_G\Sigma^{-1}{\mu_G}^T + \mu_B\Sigma^{-1}{\mu_B}^T) \\ &+ log\left(\frac{p_G}{p_B}\right). \end{aligned} \qquad (3.35)$$

Budući da je to linearna kombinacija od $x_i$, jednadžba zadovoljava pretpostavku logističke regresije. Međutim, druge klase distribucija također zadovoljavaju pretpostavku logističke regresije, uključujući i one koje ne vode do linearne determinantne funkcije ako se primjeni Bayes-ov pristup iz 3.2. Uzmimo npr. slučaj kada su sve karakteristike binarne i međusobno nezavisne. To znači da

$$P(X_i = 1|G) = p_G(i); \qquad P(X_i = 0|G) = 1 - p_G(i);$$

$$P(X_i = 1|B) = p_B(i); \qquad P(X_i = 0|B) = 1 - p_B(i).$$

Stoga vrijedi, ako su $p_G$, $p_B$ prethodne vjerojatnosti „dobrih“ i „loših“ populacija

$$P(G|x) = \frac{P(x|G)p_G}{P(x)} = \frac{\prod_i p_G(i)^{x_i}\left(1 - p_G(i)\right)^{1-x_i} p_G}{P(x)}, \tag{3.36}$$

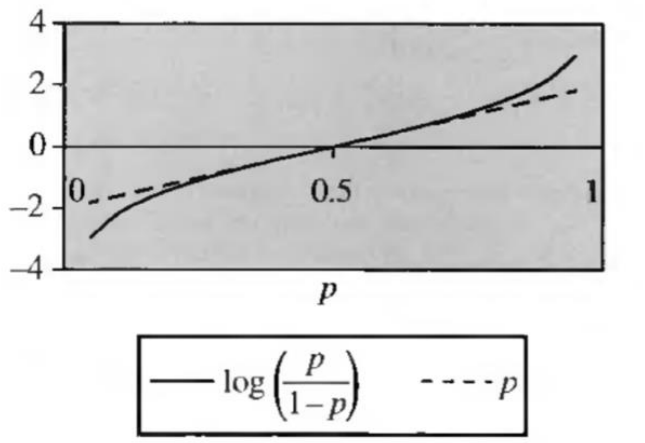
tada je

$$\begin{aligned} log\left(\frac{P(G|x)}{P(B|x)}\right) &= \Sigma_i x_i \left(log\left(p_G(i)\right) - log\left(p_B(i)\right)\right) \\ &\quad + \Sigma_i (1 - x_i)\left(log\left(1 - p_G(i)\right) - log\left(1 - p_B(i)\right)\right) \\ &\quad + log\left(\frac{p_G}{p_B}\right) \\ &= \Sigma_i x_i \left( log\left(\frac{p_G(i)\left(1 - p_B(i)\right)}{p_B(i)\left(1 - p_G(i)\right)}\right)\right) + \Sigma_i log\left(\frac{1 - p_G(i)}{1 - p_B(i)}\right) + log\left(\frac{p_G}{p_B}\right). \end{aligned} \tag{3.37}$$

To je opet oblika (3.31), stoga zadovoljava pretpostavku logističke regresije. Jedina poteškoća s logističkom regresijom u odnosu na običnu regresiju je da nije moguće koristiti obični pristup metodom najmanjih kvadrata za izračun koeficijenata *w*. Umjesto toga, mora se koristiti pristup maksimalne vjerojatnosti da bismo dobili procjene za te koeficijente. To dovodi do iterativne Newton-Raphson metode za rješavanje jednadžbi koje se pojavljuju. Uz moć modernih računala to nije problem, čak i za velike uzorke koji su često dostupni prilikom izgradnje kreditne bodovne kartice.

Jedan od iznenađujućih rezultata je da, iako je teoretski logistička regresija optimalna, za mnogo širu klasu distribucija od linearne regresije u svrhu klasifikacije, kada se rade usporedbe na razvijenim bodovnim karticama, koristeći dvije različite metode na istom skupu podataka, postoji vrlo mala razlika u njihovim klasifikacijama. Razlika je u tome što linearna regresija pokušava uklopiti vjerojatnost p (nemogućnosti ispunjavanja obaveza) u linearnu kombinaciju atributa, dok logistička regresija pokušava

uklopiti $log\left(\frac{p_i}{1-p_i}\right)$ u linearnu kombinaciju atributa. Slika 3.5.1 pokazuje, ako mapiramo linearni pomak od $p$ i od $log\left(\frac{p_i}{1-p_i}\right)$, tada su oni jako slični, sve dok $p$ ne dođe blizu 0 ili blizu 1. U kontekstu kreditnog skoringa, znači da su rezultati dobiveni tim dvjema metodama jako slični, osim za one vjerojatnosti gdje je vjerojatnost da će se ispuniti obveza, vrlo niska ili vrlo visoka. To su podnositelji zahtjeva kod kojih bi trebalo biti jednostavno predvidjeti njihovu nemogućnost ispunjavanja obveza. Na još kompliciranijem mjestu vjerojatnosti neispunjavanja obveza, oko $p = 0.5$, krivulje su jako slične. To može objasniti zašto postoji manje varijacija kod metoda nego što se može očekivati.



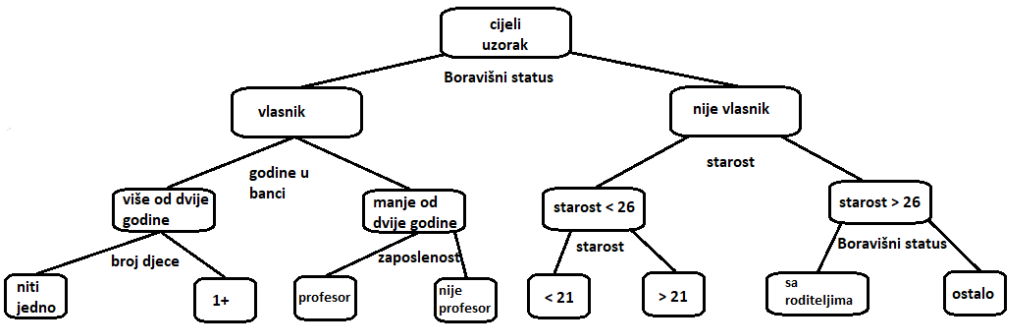

Slika 3.5.1 *Graf od log(p/(1-p)) i ap+b*

## 3.6 Klasifikacijsko stablo (rekurzivni particionirani pristup)

Potpuno drugačiji statistički pristup klasifikaciji i diskriminaciji je ideja razvrstavanja stabala, koji se ponekad naziva rekurzivni particionirani algoritam (RPA). Ideja je podijeliti skup odgovora zahtjeva u različite podskupove, a zatim identificirati svaki od tih podskupova kao „dobar“ ili „loš“, ovisno o tome što je većina u tom skupu. Ideja je

razvijena za opće klasifikacijske probleme neovisno od strane Breiman-a i Friedman-a u 1973, koji su opisali niz statističkih zahtjeva (ali ne u kreditnom skoringu) u svojoj knjizi (Breiman et al. 1984). Njegova uporaba u kreditnom skoringu je brzo uslijedila (Makowski 1985, Coffman 1986). Ideja je također sakupljena u literaturu umjetne inteligencije. Slične ideje korištene su u problemima klasificiranja, a razvijen je koristan računalni softver za njegovu provedbu. Iako softver ima različita imena, CHAID i C5, osnovni koraci jednaki su u obje literature (Safavian i Landgrebe 1991).

Skup podataka zahtjeva $A$ prvo se dijeli na dvije podskupine. Stoga gledajući uzorak ranijih podnositelja zahtjeva, ta dva nova podskupa atributa zahtjeva su daleko homogenija, u riziku neispunjavanja obveza od strane podnositelja zahtjeva, od izvornog skupa. Svaki od tih skupova ponovno se dijeli na dva kako bi se proizvele još homogenije podskupine, te se proces ponavlja. Zato se pristup zove rekurzivno particioniranje. Proces prestaje kada podskupine ispunjavaju zahtjeve terminalnih čvorova stabla. Svaki terminalni čvor se zatim klasificira kao član $A_G$ ili $A_B$ i cijeli postupak se može prikazati grafički kao stablo, kao na slici (Slika 3.6.1).




Slika 3.6.1 *Klasifikacijsko stablo*

Tri odluke čine proceduru klasifikacijskog stabla:

- koje pravilo koristiti za podjelu skupova na dva – pravilo podjele;
- kako odlučiti da je skup terminalni čvor — pravilo zaustavljanja

- kako dodijeliti terminalne čvorove u „dobre“ i „loše“ kategorije.

Dobra-loša odluka o dodjeli terminalnih čvorova je najlakše napraviti. Normalno, čvor je određen kao „dobar“ ako su većina uzoraka u tom čvoru „dobri“. Alternativa je da se minimizira trošak pogrešne klasifikacije. Ako je *D* dug nastao pogrešnom klasifikacijom „loših“ kao „dobre“ i *L* je izgubljena zarada uzrokovana pogrešnom klasifikacijom „dobrih“ kao „loš“, onda minimiziramo trošak ako razvrstavamo čvor kao „dobar“, kada omjer „dobrih“ naspram „loših“ u tom čvoru u uzorku premašuje $\frac{D}{L}$.

Najjednostavnija pravila razdvajanja su ona koja gledaju samo jedan korak unaprijed na rezultat predložene podjele. To se radi pronalaženjem najbolje podjele za svaku karakteristiku pojedinačno, imajući na umu neke mjere o tome koliko je podjela dobra. Tada se odlučuje koja je podjela karakteristika najbolja, uzimajući u obzir tu mjeru. Za sve neprekidne karakteristične $X_i$, promatramo podjele $\{x_i < s\}, \{x_i \geq s\}$, za sve vrijednosti $s$, te pronalazimo vrijednost s na mjestu gdje je mjera najbolja. Ako $X_i$ je kategorična varijabla, tada promatramo sve moguće podjele kategorija na dva podskupa i provjeravamo mjeru u okviru tih različitih razdjelnih skupina.

Najčešća mjera za podjelu karakteristika je Kolmogorov-Smirnovljeva statistika. No postoje još barem četiri druge – jednostavni indeks nečistoće, Ginijev indeks, indeks entropije i polusuma kvadrata. Razmatrat ćemo svaki od njih pojedinačno, te ćemo ih koristiti kako bi dobili izračun najbolje podjele u jednostavnom primjeru 3.1.

**Primjer 3.1** Stambeni status ima 3 atributa sa brojevima „dobrih“ i „loših“ u svakom od atributa u uzorku prethodnog klijenta, kao što je prikazano u Tablica 3.1 . Ako podijelimo stablo na njegove karakteristike, kako bi trebala izgledati ta podjela?

| **Stambeni status** | **Vlasnik** | **Zakupac** | **Nadređeni** |
|---|---|---|---|
| **Broj “dobrih”** | 1000 | 400 | 80 |
| **Broj “loših”** | 200 | 200 | 120 |
| **Omjer dobro:loše** | 5:1 | 2:1 | 0.67:1 |

Tablica 3.1

### 3.6.1 Kolmogorov-Smirnovljeva statistika

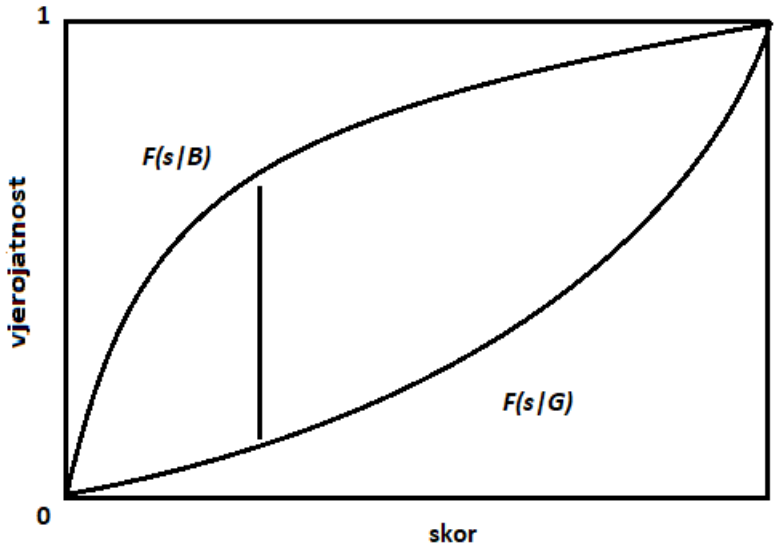
Za neprekidnu karakteristiku $X_i$ , neka je $F(s|G)$ je kumulativna funkcija distribucije od $X_i$ među „dobrima“ i $F(s|B)$ kumulativna funkcija distribucije od $X_i$ među „lošima“. Uzeći u obzir da „loši“ imaju veću sklonost poprimanja niže vrijednosti karakteristika $X_i$, nego „dobri“, te da su definicije troškova $D$ i $L$ definirane kao u prošlom paragrafu, kratkovidno pravilo bi bilo da se podjela vrši na vrijednosti $s$, koja minimizira

$$LF(s|G)p_G + D\big(1 - F(s|B)\big)p_B. \tag{3.38}$$

Ako je

$$Lp_G = Dp_B,$$

to je isto kao da odaberemo Kolmogorov-Smirnovljevu udaljenost između dvije distribucije, kao što je prikazano na slici (Slika 3.6.2).




Slika 3.6.2 *Kolmogorov-Smirnovljeva udaljenost*

Želimo minimizirati $F(s|G) - F(s|B)$ ili još očitije maksimizirati $F(s|B) - F(s|G)$. Ako npr. zamislimo dvije podgrupe, lijeva ($l$) i desna ($d$), to je isto kao da

maksimiziramo razliku između $p(l|B)$ (ili $F(s|B)$ u neprekidnom slučaju) , vjerojatnost da se „loši“ pojave u lijevoj grupi, te $p(l|G)$ (ili $F(s|G)$ u neprekidnom slučaju), vjerojatnost da se „dobri“ pojave u lijevoj grupi. Koristeći Bayesov teorem, možemo $p(l|B)$ zapisati kao $\frac{p(B|l)p(l)}{p(B)}$ , te nam je sada to lakše izračunati. Stoga za kategorične ili neprekidne varijable Kolmogorov-Smirnovljev (KS) kriterij postaje: pronađi lijevo-desni prijelom koji maksimizira

$$KS = |p(l|B) - p(r|G)| = \left|\frac{p(B|l)}{p(B)} - \frac{p(G|l)}{p(G)}\right| p(l) \ . \tag{3.39}$$

Ako pogledamo **Primjer 3.1** očito je iz omjera dobro:loše da najbolji prijelom mora biti ili sa nadređenim u jednoj grupi i vlasnik + zakupac u drugoj grupi ili zakupac + nadređeni u jednoj grupi i vlasnik u drugoj:

l = nadređeni, r = vlasnik + zakupac:

$$p(l|B) = \frac{120}{520} = 0.231, \qquad p(l|G) = \frac{80}{1480} = 0.054, \quad KS = 0.177,$$

l = nadređeni + zakupac, r = vlasnik:

$$p(l|B) = \frac{320}{520} = 0.615, \qquad p(l|G) = \frac{480}{1480} = 0.324, \quad KS = 0.291.$$

Stoga je najbolja podjela nadređeni + zakupac u jednoj grupi i vlasnik u drugoj.

### 3.6.2 Jednostavni indeks nečistoće *i(v)*

Postoji cijela klasa mjera indeksa nečistoće kojima je cilj procijeniti koliko je nečist svaki čvor stabla, gdje se čistoća podudara sa time da su čvorovi u istoj klasi. Ako sada podijelimo čvor na lijevi čvor ($l$) i desni čvor ($d$), gdje je proporcija koja ide u $l$ $p(l)$ i proporcija koja ide u $d$ $p(d)$, tada možemo izmjeriti promjenu u nečistoći nastaloj podjelom:

$$I = i(v) - p(l)i(l) - p(r)i(r) \ . \tag{3.40}$$

Što je veća razlika, to je veća promjena u nečistoći, što povlači da su čvorovi puno čišći. To je ono što mi želimo, stoga odabiremo podjelu koja maksimizira taj izraz. To je ekvivalentno minimizaciji $p(l)i(l) + p(r)i(r)$. Jasno je da ako ne postoji podjela sa pozitivnom razlikom, tada se uopće ne radi podjela čvora.

Najjednostavniji indeks nečistoće jest uzeti $i(v)$ kao proporciju manje grupe u tom čvoru tako da vrijedi:

$$\begin{aligned} i(v) &= p(G|v) \quad ako\ p(G|v) \leq 0.5, \\ i(v) &= p(B|v) \quad ako\ p(B|v) < 0.5. \end{aligned} \tag{3.41}$$

Na **Primjer 3.1**, to nam daje sljedeći izračun za indekse kako bismo vidjeli koji od njih je najbolji prijelom, gdje je *v* skup cijelih brojeva prije bilo kojeg prijeloma:

l = nadređeni, r = vlasnik + zakupac:

$$i(v) = \frac{520}{2000} = 0.26, \ \ p(l) = \frac{200}{2000} = 0.1, \ \ i(l) = \frac{80}{200} = 0.4,$$

$$p(r) = \frac{1800}{2000} = 0.9, \ \ i(r) = \frac{400}{1800} = 0.22,$$

$$I = 0.26 - 0.1(0.4) - 0.9(0.22) = 0.02;$$

l = nadređeni + zakupac, r = vlasnik:

$$i(v) = \frac{520}{2000} = 0.26, \ \ p(l) = \frac{800}{2000} = 0.4, \ \ i(l) = \frac{320}{800} = 0.4,$$

$$p(r) = \frac{1200}{2000} = 0.6, \ \ i(r) = \frac{200}{1200} = 0.167,$$

$$I = 0.26 - 0.4(0.4) - 0.6(0.167) = 0.$$

To nam sugerira da je najbolja podjela, nadređeni u jednoj grupi, te vlasnik + zakupac u drugoj.

Iako to izgleda beskorisno, izgled može varati. *I* je u drugoj podjeli jednak 0, zato što su „loši“ u manjini u sva tri čvora *v, l* i *r*. To će se uvijek dogoditi ako su iste grupe u manjini u sva tri čvora što je slučaj za mnoge kreditne situacije. Budući da sve podjele daju istu razliku 0, taj kriterij je beskoristan za odlučivanje onoga koji je najbolji. Također, djelo Brieman el.al (1984) je dalo primjer u kojemu je bilo 400 „dobrih“ i 400 „loših“. Jedna particija dijeli to na jednu klasu od 300 „dobrih“ i 100 „loših“ i drugu klasu na 100 „dobrih“ i 300 „loših“, dok druga particija dijeli na jednu klasu od 200 „dobrih“ i drugu klasu od 200“dobrih“ i 400 „loših“. Obje podjele imaju razliku indeksa *i(v)*, ali većina ljudi bi mislila da je druga particija, koja je u mogućnosti identificirati cijelu grupu „dobrih“, bolja podjela. Ono što gledamo je indeks koji nagrađuje čistije čvorove nego ova.

### 3.6.3 Ginijev koeficijent (indeks)

Umjesto linearne proporcije vjerojatnosti nečistoća, Ginijev indeks je kvadratni, stoga daje veću težinu na čišće čvorove. On je definiran na sljedeći način:

$$i(v) = p(G|v)p(B|v),$$

pa slijedi

$$I = \ i(v) - p(l)i(l) - p(r)i(r) \ . \tag{3.42}$$

U **Primjer 3.1** koji nas jako interesira, to nam daje

l = nadređeni, r = vlasnik + zakupac:

$$i(v) = \left(\frac{1480}{2000}\right)\left(\frac{520}{2000}\right) = 0.1924,$$

$$p(l) = \frac{200}{2000} = 0.1, \ \ i(l) = \left(\frac{80}{200}\right)\left(\frac{120}{200}\right) = 0.24,$$

$$p(r) = \frac{1800}{2000} = 0.9, \ \ i(r) = \left(\frac{400}{1800}\right)\left(\frac{1400}{1800}\right) = 0.1728,$$

$$I = 0.1924 - 0.1(0.24) - 0.9(0.1728) = 0.01288;$$

l = nadređeni + zakupac, r = vlasnik:

$$i(v) = \left(\frac{520}{2000}\right)\left(\frac{1480}{2000}\right) = 0.1924,$$

$$p(l) = \frac{800}{2000} = 0.4, \;\; i(l) = \left(\frac{320}{800}\right)\left(\frac{480}{800}\right) = 0.24,$$

$$p(r) = \frac{1200}{2000} = 0.6, \;\; i(r) = \left(\frac{200}{1200}\right)\left(\frac{1000}{1200}\right) = 0.1389,$$

$$I = 0.1924 - 0.4(0.24) - 0.6(0.1389) = 0.01306;$$

Indeks sugerira da je najbolja podjela nadređeni + zakupac u jednom čvoru i vlasnik u drugom čvoru.

### 3.6.4 Indeks entropije

Još jedan nelinearni indeks je indeks entropije, gdje je

$$i(v) = -p(G|v)ln(p(G|v)) - p(B|v)ln(p(B|v)). \qquad (3.43)$$

Kao što samo ime indeksa sugerira, to je povezano sa entropijom ili količinom informacija u podjeli između „dobrih“ i „loših“ u čvoru. On je mjera koja kaže na koliko različitih načina se može doći do podjela „dobrih“ i „loših“ u čvoru. Koristeći tu mjeru, podjele u **Primjer 3.1** su izmjerene na sljedeći način:

l = nadređeni, r = vlasnik + zakupac:

$$i(v) = -\left(\frac{520}{2000}\right) ln\left(\frac{520}{2000}\right) - \left(\frac{1480}{2000}\right) ln\left(\frac{1480}{2000}\right) = 0.573,$$

$$p(l) = \frac{200}{2000} = 0.1, \;\; i(l) = -\left(\frac{80}{200}\right) ln\left(\frac{80}{200}\right) - \left(\frac{120}{200}\right) ln\left(\frac{120}{200}\right) = 0.673,$$

$$p(r) = \frac{1800}{2000} = 0.9,\ \ i(r) = -\left(\frac{400}{1800}\right) ln\left(\frac{400}{1800}\right) - \left(\frac{1400}{1800}\right) ln\left(\frac{1400}{1800}\right) = 0.530,$$

$$I = 0.573 - 0.1(0.673) - 0.9(0.530) = 0.0287;$$

l = nadređeni + zakupac, r = vlasnik:

$$i(v) = -\left(\frac{520}{2000}\right) ln\left(\frac{520}{2000}\right) - \left(\frac{1480}{2000}\right) ln\left(\frac{1480}{2000}\right) = 0.573,$$

$$p(l) = \frac{800}{2000} = 0.4,\ \ i(l) = -\left(\frac{320}{800}\right) ln\left(\frac{320}{800}\right) - \left(\frac{480}{800}\right) ln\left(\frac{480}{800}\right) = 0.673,$$

$$p(r) = \frac{1200}{2000} = 0.6,\ \ i(r) = -\left(\frac{200}{1200}\right) ln\left(\frac{200}{1200}\right) - \left(\frac{1000}{1200}\right) ln\left(\frac{1000}{1200}\right) = 0.451,$$

$$I = 0.573 - 0.4(0.673) - 0.6(0.451) = 0.0332;$$

Indeks također sugerira da je najbolja podjela nadređeni + zakupac u jednom čvoru i vlasnik u drugom čvoru.

### 3.6.5 Maksimizirana polusuma kvadrata

Zadnja mjera koju promatramo nije indeks, ali dolazi iz $\chi^2$testa, koji provjerava jesu li proporcije „dobrih“ iste u dva podčvora na koja dijelimo. Ako je $\chi^2$ statistika velika, tada kažemo da hipoteza nije istinita tj. dvije proporcije nisu iste. Što je veća vrijednost, to je manje vjerojatno da su proporcije iste, no potonji se može interpretirati ina sljedeći način: što je vrijednost veća, tada postoji veća razlika između njih. To je ono što tražimo od podjele, stoga dolazimo do sljedećeg testa.

Ako su *n(l)* i *n(d)* suma brojeva na lijevom i desnom čvoru, tada maksimiziramo

$$Chi = n(l)n(d) - \frac{(p(G|l) - p(G|d))^2}{n(l) + n(d)}.$$

Ako primijenimo taj test na podatke iz **Primjer 3.1** dobijemo:

l = nadređeni, r = vlasnik + zakupac:

$$n(l) = 200,\ \ p(G|l) = \frac{80}{200} = 0.4,$$

$$n(r) = 1800,\ \ p(G|r) = \frac{1400}{1800} = 0.777,$$

$$stoga\ je\ Chi = 25.69;$$

l = nadređeni + zakupac, r = vlasnik:

$$n(l) = 800,\ \ p(G|l) = \frac{480}{800} = 0.6,$$

$$n(r) = 1200,\ \ p(G|r) = \frac{1000}{1200} = 0.833,$$

$$stoga\ je\ Chi = 26.13;$$

Stoga je mjera maksimizirana kada su „sa nadređenim“ klasa i „zakupac“ klasa stavljeni zajedno.

Postoje i druge metode podjele. Breiman et al. (1984) sugerira da bi bolji kriterij, od gledanja sljedeće podjele, bio uzeti u obzir koja je situacija poslije $r$ novih generacija podjele. Takav pristup uzima u obzir ne samo srednje poboljšanje prouzrokovano trenutačnom podjelom, nego i dugoročnu strategijsku važnost podjele. Radi se najčešće o slučaju kada se podjela vrši koristeći različite karakteristike na različitim nivoima stabla, te također različite karakteristike koje dolaze ispred različitih podskupova na istim nivoima stabla, kao što je ilustrirano na Slika 3.6.1. Na taj način stablo pronalazi nelinearne veze između karakteristika zahtjeva.

Iako govorimo o tome kada prekidamo stablo i prepoznajemo čvor kao terminalni čvor, možda je bolje govoriti o pravilu zaustavljanja i odrezivanja. To ističe da očekujemo da je početno stablo preveliko i da će biti potrebno srezati ga na stablo koje je snažno. Ako imamo stablo kod kojeg svi terminalni čvorovi imaju samo jedan slučaj iz uzorka za treniranje modela, tada je stablo savršen diskriminator na uzorku za treniranje modela, ali je jako loš klasifikator na bilo kojem drugom skupu. Stoga, čvor postaje

terminalni čvor iz dva razloga. Prvi razlog je toliko mala količina slučajeva na čvoru da nema smisla dalje ga dijeliti. To je najčešće kada čvor ima manje od 10 slučajeva**.** Drugi je razlog da vrijednost mjere koja dijeli čvor na dva manja čvora, nema gotovo nikakve razlike od vrijednosti mjere ako držimo čvor onakvim kakav je. No, moramo jasno definirati što znači „nema gotovo nikakve razlike“ u kontekstu, a može biti da je razlika u mjeri niža od nekog propisanog nivoa $\beta$.

Sada kada smo dobili tako veliko stablo, možemo ga smanjiti tako da maknemo neke od podjela. Najbolji način da to napravimo je da koristimo ogledni uzorak, koji nije korišten u izgradnji stabla. Taj uzorak je korišten pri empirijskoj procjeni očekivanih gubitaka za različite mogućnosti rezanja stabla. Koristeći ogledni uzorak i klasifikaciju stabla *T,* definirat ćemo $T_G$ $(T_B)$ kao skup čvorova koji su klasificirani kao „dobri“ („loši“). Neka je $r(t, B)$ proporcija oglednog uzorka koji je u čvoru *t,* koji su klasificirani kao „loši“, te $r(t, G)$ neka je proporcija oglednog uzorka koji je u čvoru *t,* koji su klasificirani kao „dobri“. Tada je procjena očekivanog gubitka

$$r(T) = \sum_{t \in T_G} Dr(t, B) + \sum_{t \in T_B} Lr(t, G) \tag{3.44}$$

Ako je $n(T)$ broj čvorova u stablu *T,* definirajmo $c(T) = r(T) + dn(T)$ te podrežemo stablo $T^*$ promatrajući sva podstabla od $T^*$ i odaberemo stablo $T$ koje minimizira $c(T)$. Ako je $d = 0$, dolazimo do početnog neodrezanog stabla. Što se $d$ povećava, to se stablo sastoji samo od jednog čvora. Stoga odabir $d$ nam daje pogled na to koliko želimo da nam bude veliko stablo.

## 3.7 Metoda najbližeg susjeda

Metoda najbližeg susjeda je standardni neparametarski pristup klasifikaciji problema koji je prvi predložen od strane Fix-a i Hodges-a (1952). Prvi put je primijenjen u kontekstu kreditnog skoringa od strane Chatterjee-a i Barcun -a (1970), a kasnije i od strane Henley-

a i Hand-a (1996). Ideja je odabrati metriku na prostoru podataka zahtjeva kako bi se izmjerilo koliko su međusobno udaljena dva podnositelja zahtjeva. Zatim, s uzorkom prijašnjih podnositelja zahtjeva kao reprezentativnog standarda, novi podnositelj zahtjeva klasificira se kao „dobar“ ili „loš“ ovisno o proporcijama „dobrih“ i „loših“ među $k$ najbližih podnositelja zahtjeva iz reprezentativnog uzorka – najbližim susjedima podnositelja zahtjeva.

Tri parametra potrebna za pokretanje ovog pristupa su metrika, koliko podnositelja zahtjeva $k$ čine skup najbližih susjeda, a koje su dobre proporcije ta dva parametra tako da podnositelj zahtjeva bude klasificiran kao „dobar“. Normalno, odgovor na ovo posljednje pitanje je, da ako su većina susjeda „dobri“, podnositelj zahtjeva je klasificiran kao „dobar“, u suprotnom, podnositelj zahtjeva klasificiran je kao „loš“. Međutim, ako je kao u dijelu 3.2, prosječni zadani gubitak $D$ i prosječni izgubljeni profit nastao odbijanjem „dobrog“ jednak $L$, tada se novi podnositelj zahtjeva može klasificirati kao „dobar“, samo ako su barem $\frac{D}{D+L}$ najbližih susjeda „dobri“. Ovaj kriterij minimizira očekivani gubitak, ako je vjerojatnost da je novi podnositelj zahtjev „dobar“ proporcija susjeda koji su „dobri“.

Izbor metrike je ključan. Fukanaga i Flick (1984) uveli su opću metriku obrasca

$$d(x_1, x_2) = (x_1 - x_2)A(x_1)((x_1 - x_2)^T)^{\frac{1}{2}}, \tag{3.45}$$

gdje je $A(x)$ $p \times p$ simetrična pozitivno definitna matrica. $A(x)$ se zove lokalna metrika ako ovisi o $x$, a globalna metrika ako ne ovisi o x. Problem lokalne metrike je to što uzima obilježja skupa za treniranje modela koji nisu prikladni općenito, stoga se većina autora koncentrira na globalnu metriku. Najdetaljnije ispitivanje pristupa najbližeg susjeda u kontekstu kreditnog skoringa je bilo od strane Henley-a i Hand-a (1996), koji su se koncentrirali na metrike koje su mješavina euklidske udaljenosti i udaljenosti u smjeru koja najbolje separira „dobre“ i „loše“. Jednu takvu udaljenost možemo dobiti iz Fisherove linearne diskriminantne funkcije iz odjeljka 3.3. Stoga, ako je $w$ $p$-dimenzionalni vektor koji definira taj smjer, dan u (3.21), Henley i Hand sugeriraju metriku oblika

$$d(x_1, x_2) = \{(x_1 - x_2)^T (I + Dww^T)(x_1 - x_2)\}^{\frac{1}{2}}, \quad (3.46)$$

gdje je *I* jedinična matrica. Oni rade veliki broj ispitivanja kako bi identificirali najbolji odabir *D*. Odabiru *k*, idealan broj najbližih susjeda, eksperimentirajući s velikim brojem *k*-ova. Iako ne postoji veliki broj varijacija u rezultatu, najbolji izbor *D* je u intervalu od 1.4 do 1.8. Izbor *k* ovisi očito o veličini uzorka za treniranje modela, stoga u nekim slučajevima mijenjajući *k* za samo jedan, čini veliku razliku. Međutim, kao što Slika 3.7.1 sugerira, ako pogledamo širu sliku, ne postoji velika razlika u procjeni krivo klasificiranih „loših" za fiksiranu dopuštenu procjenu, dok *k* varira u velikom intervali od 100 do 1000 (sa uzorkom za treniranje modela od 3000). Kako bi izbjegli odabir lokalno loše vrijednosti k, možemo „izravnati" *k* odabirom distribucije od *k*. Stoga za svaku točku može postojati različiti broj različitih najbližih susjeda.

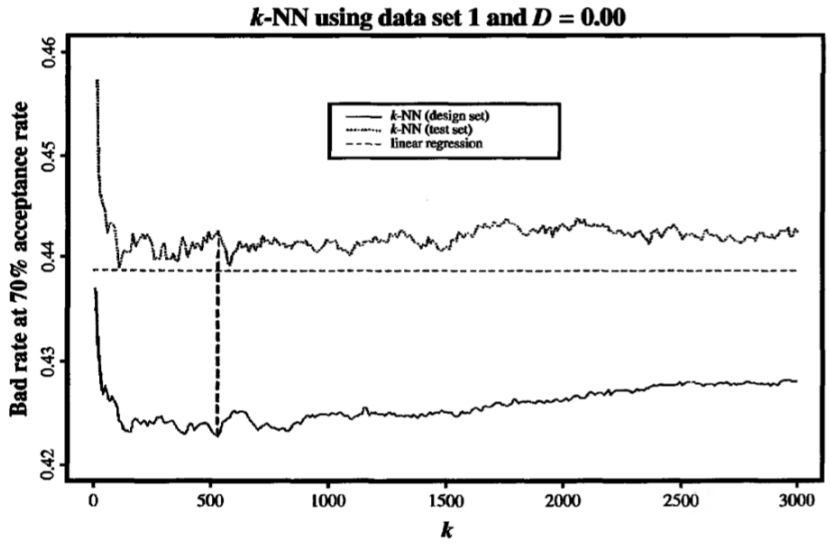



**Figure 4.6.** *Default (bad) rate in nearest-neighbor systems as k—size of neighbourhood—varies.*

Slika 3.7.1 *Mjera kašnjenja plaćanja u sustavu najbližeg susjeda kada k-veličina susjedstva varira*

Pristupi najbližeg susjeda, iako nisu široko korišteni u kreditnom skoringu, kao što su linearna i logistička regresija, imaju potencijalno atraktivna obilježja za stvarnu

uporabu. Bilo bi jednostavno dinamički modernizirati uzorak za treniranje modela, tako da se dodaju novi slučajevi uzorku za treniranje modela ako znamo da su ti slučajevi „dobri“ ili „loši“, te da se izbace oni slučajevi koji se nalaze najduže u uzorku. To bi donekle nadvladalo potrebu za redovnom modernizacijom skoring sustava, zbog promjena u populaciji. Metrika $d$ (kao u (3.46)) se također treba modernizirati , uzimajući u obzir promjene u populaciji, a to se ne može napraviti dinamički. Zabrinutost za to što bi veliki broj izračuna trebao biti dobar u oba slučaja („dobar“ ili „loš“), pronalazeći koji su $k$ najbliži susjedi u uzorku za treniranje modela, je nepotrebna , jer moderna računala mogu raditi takve izračune u nekoliko sekundi. Pronalazak dobre metrike je skoro ekvivalentno regresijskom pristupu za izgradnju bodovne kartice. Kao i u klasifikacijskom pristupu, činjenica da pristup najbližeg susjeda nije u mogućnosti dati rezultat za svaku karakteristiku podnositelja zahtjeva, uskraćuje korisnike utočištem koje im pomaže da misle da razumiju kako sustav ustvari funkcionira.

# Poglavlje 4

# 4 Praktična pitanja razvoja rezultata bodovne kartice

## 4.1 Odabir uzorka

Sve metodologije za kredit i bihevioralni skoring (Bihevioralni skoring gleda ponašanje klijenata kako bi se poboljšao kreditni portfolijski menadžment. On pomaže boljem shvaćanju klijenata, jer što se bolje poznaje klijent to se bolje može odgovoriti na njegove zahtjeve.) zahtijevaju uzorak prethodnih klijenata i njihove povijesti da bi se razvio sustav skoringa. U odabiru takvog uzorka postoje dva proturječna cilja. Prvo, uzorak bi trebao biti predstavnik onih ljudi kod kojih postoji veliki izgled da će se prijaviti za kredit u budućnosti. Drugo, trebalo bi obuhvaćati dovoljnu količinu različitih vrsta ponašanja otplate (tj. „dobrih“ i „loših“) kako bi se omogućilo utvrđivanje koje karakteristike odražavaju to ponašanje u općoj populaciji. Grupa koja mora biti najbliža ovoj populaciji su prethodni podnositelji zahtjeva za taj proizvod pozajmljivanja. Sukob nastaje zbog toga što da bi se približili što je moguće bliže budućoj populaciji, želimo da uzorak grupa bude što je moguće noviji. Međutim, za razlikovanje „dobro“ i „loše“ ponašanje otplate, trebamo razumnu povijest otplate, stoga i razumno vrijeme u kojemu se skupina uzoraka primjenjivala. To je posebno slučaj s bihevioralnim skoringom, gdje je također potrebno i razumno razdoblje provođenja kako bi se identificirale transakcijske karakteristike, ali i razdoblje ishoda. Dogovoreno kompromisno razdoblje ishoda je obično period od 12 mjeseci za sustav za aplikacijski skoring. U bihevioralnom skoringu, obično je potrebno od 18 do 24 mjeseca povijesti i kada se to podijeli na dva razdoblja dobije se 9 do 12 mjeseci povijesti provođenja i 9 do 12 mjeseci razdoblja ishoda. Ta se razdoblja razlikuju ovisno o proizvodu jer je za hipoteke za razdoblje ishoda možda potrebno nekoliko godina.

Sljedeća pitanja su koliko bi veliki uzorak trebao biti i koja bi podjela trebala biti između broja „dobrih“ i broja „loših“ u uzorku. Treba li biti jednak broj „dobrih“ i „loših“ u uzorku, ili bi uzorak trebao odražavati dobar : loš koeficijent u populaciji u cjelini? U pravilu, potonji je tako snažno usmjeren na „dobre“ (recimo, 20:1), da bi uzimajući iste koeficijente u uzorku značilo da možda neće biti dovoljno „loše“ podpopulacije za utvrđivanje njihovih karakteristika. Iz tog razloga, uzorak ima tendenciju da bude ili 50:50 ili negdje između 50:50 i pravog udjela populacije „dobrih“ naspram „loših“. Ako distribucija „dobrih“ i „loših“ u uzorku nema jednaku distribuciju kao ona u populaciji u cjelini, tada je potrebno prilagoditi rezultate dobivene uzorkom kako bi se to omogućilo. U regresijskom pristupu, to se radi automatskim budući da se vjerojatnost „dobrih“ i „loših“ u pravoj populaciji, $p_G \; i \; p_B$, koristi u izračunima. U drugim pristupima, to mora biti učinjeno naknadno, tako da ako klasifikacijsko stablo izgrađeno na uzorku, u kojem su „dobri“ 50% populacije (ali od prave populacije su 90%), ima čvor u kojem je dobar : loš omjer 3:1 ili 75% : 25%, tada su pravi koeficijenti

$$\frac{(koeficijenti\ u\ čvoru)(koeficijenti\ u\ pravoj\ populaciji)}{(koeficijenti\ u\ uzorku)} = \frac{\frac{3}{1} \cdot \frac{9}{1}}{\frac{1}{1}} \qquad (4.1)$$

$$= 27{:}\,1.$$

Što se tiče broja u uzorku, Lewis (1992) je predložio da 1,500 „dobrih“ i 1,500 „loših“ može biti dovoljno. U praksi se koriste mnogo veći uzorci, iako je Makuch (1999) iznio dobru činjenicu, da ako imamo 100,000 „dobrih“, nema potrebe za mnogo više informacija o „dobrima“. Tako bi tipična situacija bila da uzmemo sve „loše“ koje možemo uzeti u uzorak, te uzmemo 100,000+ „dobrih“. Ovaj uzorak se onda nasumično podijeli na dva. Jedan dio se koristi za razvoj skoring sustava, a drugi se koristi kao ogledni uzorak za testiranje.

Ako je uzorak odabran nasumično od postojeće populacije podnositelja zahtjeva, moramo biti sigurni da je izbor stvarno slučajan. To nije preteško ako postoji središnji popis zahtjeva koji se čuva s podnositeljima zahtjeva koji su naloženi s obzirom na vrijeme

primjene. Odabir svakog desetog „dobrog“ na popisu trebalo bi dati razumno slučajni uzorak od 10% „dobrih“. Ako, međutim, moramo ići na razinu grananja kako bi stvorili popis, prvo moramo nasumično odabrati grane kako bi se osigurala dobra mješavina urbanih i ruralnih grana i prikladno širenje socioekonomskih uvjeta i geografije. Tek onda se mora nasumično odabrati na razini grananja. No, treba biti oprezan. Odluka, da se uzmu svi kupci koji se prijavljuju u određenom mjesecu, može se na površini činiti razumna, ali ako je to mjesec kada počinje semestar na sveučilištu, onda će puno više studenata biti u populaciji podnositelja zahtjeva nego u pravom slučajnom uzorku. Ponekad je možda potrebno staviti takvu pristranost u uzorak jer na primjer, novi proizvod je više za mlade ljude, nego za one postojeće i stoga želimo veći udio mladih ljudi u uzorku nego u izvornoj populaciji. Cilj je uvijek dobiti uzorak koji će najbolje reflektirati najvjerojatniju populaciju koja je trenutačno zainteresirana za novi proizvod. U stvari, to nije sasvim točno; ono što želimo je uzorak populacije kroz vrata koja je trenutačno zainteresirani za neki proizvod, koje će zajmodavac uzeti u obzir za proces bodovanja. Stoga oni koji ne bi dobili kredite iz političkih razloga trebali bi biti uklonjeni iz uzorka, kao i oni koji bi ih automatski dobili. Bivši uzorak može uključivati maloljetne podnositelje zahtjeva, bankrote i one bez dosjea kreditnog registra. Potonje može uključivati kupce s posebnim štednim proizvodima ili zaposlenicima zajmodavca.

Sav taj trud oko odabira uzorka pretpostavlja da znamo definirati „dobre“ i „loše“. Stoga ćemo u sljedećem poglavlju, promatrati koje bi to mogle biti definicije i što raditi s njima u uzorku kada oni ne upadaju u taj uzorak ili u tu kategoriju.

## 4.2 Definicije „dobrih“ i „loših“

Kao dio razvoja kreditnih bodovnih kartica, potrebno je odlučiti kako definirati „dobro“ i „loše“. Definiranje „lošeg“ ne znači nužno da su svi ostali slučajevi „dobri“. Često u razvoju kreditnih bodovnih kartica, mogu se identificirati još najmanje dvije vrste slučajeva. Prvi bi mogao biti nazvan „neodređeni“. To su slučajevi koji su između tj. nisu ni „dobri“ ni „loši“. Drugi može biti nazvan „nedovoljno iskusan“.

U razvoju kreditnih bodovnih kartica za portfelj kreditnih kartica, uobičajena definicija „lošeg“ je slučaj kada je klijent u nekom trenutku tri plaćanja (rate) u zaostatku. To se često naziva "ever 3 + Down" ili "worst 3 + down." Neodređeni slučajevi mogu biti oni kod kojih postoji dva plaćanja u zaostatku. Takav slučaj može izazvati neke probleme i neke dodatne aktivnosti– možda ako su bili na dva plaćanja u zaostatku u više navrata – ali nikada nisu postali tri plaćanja u zaostatku. Onda bismo te slučajeve mogli identificirati kao „nedovoljno iskusni“. Pretpostavimo da imamo neki prozor za razvoj uzorka od 12 mjeseci prijava, te promatračku točku godinu dana kasnije, tako da su slučajevi izloženi 12 do 24 mjeseca. Tada bismo mogli označiti „nedovoljno iskusne“ one koji imaju tri ili manje mjeseci kupnje ili aktivnosti podizanja gotovine ,odnosno novčanog avansa. Drugim riječima, račun nije „loš“, ali je bio korišten prilično rijetko i stoga bi bilo prerano definirati ga kao „dobrog“. Ostatak bi bio kategoriziran kao „dobar“ račun.

Ova klasifikacija je samo primjer. Moguće su mnoge varijacije. Možemo definirati klijenta kao „lošeg“ kada je tri plaćanja u zaostatku, tj. "ever 3 + Down" ili „twice 2 down“ . Tada bismo mogli označiti „nedovoljno iskusne“ one koji su imali manje od šest mjeseci debitnih aktivnosti (dugovanja). Možemo uključiti slučajeve koji su preskočili jedno plaćanje u “neodređeno“.

Na rate kredita, situacija bi mogla biti malo jasnija. Ovdje, možemo definirati „loše“ one klijente koji su u zaostatku od tri plaćanja ili one klijente koji su u zaostatku od dva plaćanja ili nešto između, kao što je prethodno predloženo. „Neodređeni“ bi mogli bile oni slučajevi kod kojih je došlo do preskakanja jednog plaćanja ili kod kojih je došlo do preskakanja jednog plaćanja nekoliko puta ili gdje je došlo do preskakanja dva plaćanja. Ako smo odabrali ogledni prozor pažljivo, onda ne može biti slučajeva s „nedovoljno iskustva“. Međutim, ako netko isplati zajam, tj. otplati kredit u paušalnom iznosu, nakon samo nekoliko mjeseci, možemo klasificirati kredit u tu kategoriju, pogotovo ako odsječen kredit znači da nismo stvorili profit ili smo napravili vrlo malu zaradu.

Kada se preselimo u osigurano kreditiranje, kao što je hipoteka, naše se definicije mogu značajno promijeniti. Ovdje imamo neku sigurnost, stoga naša definicija „dobrog“ i „lošeg“ može biti značajno pogođena time da li je slučaj generirao gubitak. Dakle, ako nismo ispunili svoje obveze nad hipotekom i imovina je ponovno preuzeta, te mi vratimo sve naše kredite, naše naplate i troškove parnice, može se reći da to nije „loš“ račun. Neki

bi zajmodavci klasificirati to kao „loše“ u razvoju kreditnih bodovnih kartica, dok drugi mogu smatrati „neodređeni“ prikladnom oznakom. Slično kao i kredit za rate, ako je hipoteka otplaćena prilično rano u svom životu, možda ćemo željeti to klasificirati kao „neodređen“ ili „nedovoljno iskusan“.

U slučaju tekućih računa s prekoračenjem, prema potrebi, ponovno se moraju izmijeniti definicije. Ne postoji fiksna očekivana mjesečna otplata, tako da nositelj računa ne mogu biti u zakašnjenju. Stoga je potrebno uvesti potpuno drugačiji skup definicija. Mogli bismo kategorizirati „loš“ račun onaj u kojem postoji neovlašteno pozajmljivanje, tj. posuđivanje iznad dogovorenog ograničenja prekoračenja, ako ograničenje uopće postoji.

Bez obzira koju definiciju odaberemo , ne postoji učinak na metodologiju kreditne bodovne kartice. (To pretpostavlja da definicije stvaraju particiju; tj. svi mogući slučajevi spadaju u točno jednu klasifikaciju.) Obično bismo odbacili „neodređene“ i one s „nedovoljno iskustva“, te bi izgradili kreditnu bodovnu karticu s „dobrim“ i „lošim“. Naravno, kako definiramo „dobre“ i „loše“ će jasno imati učinak na razvoj kreditne bodovne kartice. Različite definicije mogu stvarati različite kreditne bodovne kartice. Međutim, to ne znači da će rezultati biti vrlo različiti. Doista, različite definicije „dobrog“ i „loše“ mogu generirati vrlo različite kreditne bodovne kartice, ali još uvijek rezultirati velikom sličnošću u slučajevima koji su prihvaćeni i odbijeni.

Općenito, iako je važno razviti dobre i razumne definicije „dobrog“ i „lošeg“, to može učiniti samo marginalnu razliku u učinkovitosti kreditne bodovne kartice koja se zapravo razvija na temelju definicija.

## 4.3 Karakteristike kreditnog registra

Kreditne referentne agencije ili kreditni registri postoje u mnogim zemljama. Njihove uloge nisu identične od zemlje do zemlje, a nisu ni zakonodavni okviri u kojima posluju.

Stoga ne iznenađuje da se pohranjeni i dostupni podaci razlikuju od zemlje do zemlje, pa čak i unutar zemalja.

U Velikoj Britaniji postoje dva glavna registra za informacije o potrošačima – Experian i Equifax. Informacijama se pristupa putem imena i adrese, iako postoje različite razine tih podudaranja. Informacije koje imaju na raspolaganju o potrošačima dolaze u nekoliko vrsta, a mi ćemo se baviti svakim od njih pojedinačno -
javno dostupne informacije, prethodna pretraživanja, zajednički doprinijele informacije (kroz zatvorene grupe korisnika, mnogi zajmodavci dijele informacije o uspješnosti svojih kupaca), zbirne informacije (na temelju prikupljenih podataka, kao što su podaci na razini poštanskog broja) i dodana vrijednost registra.

### 4.3.1 Dostupne informacije

U Velikoj Britaniji, javno dostupne informacije o dvije vrste. Prvi je izborni popis ili glasački popis. Ovo je popis svih stanovnika koji su registrirani za glasovanje. U Velikoj Britaniji, to nije kompletan popis svih odraslih osoba jer ne postoji obaveza glasovanje, kao što je u nekim drugim zemljama. Te informacije uključuju i godinu u kojoj je netko bio registriran za glasovanje na adresi. To je korisna informacija jer se može koristiti za provjeru vremena provedenog na adresi na kojoj je netko živio. Na primjer, ako oni tvrde da su živjeli na adresi tri godine, ali su tamo registrirani za glasovanje na dvanaest godina, očito je došlo do nepodudarnosti informacija.

U Velikoj Britaniji postoji rasprava između Office of the Data Protection Register (ODPR) i regulatornih i industrijskih tijela. Ta je rasprava nastala jer ODPR želi da glasači mogu odlučiti hoće li njihova registracija birača biti dostupna za upotrebu u kreditne ili marketinške svrhe. Vjerojatno je da će se provesti neka ograničenja. Međutim, iako se to radi kako bi se potrošačima pružila veća zaštita, moguće je da će učinak oslabiti sposobnost zajmodavca da razlikuje prihvatljive i neprihvatljive kreditne rizike. Posljedica toga je da potrošač može zapravo patiti ili zbog toga što je odbijen ili je prisiljen platiti više u kamatnim stopama kako bi pokrio dodatne rizike.

Na lokalnim vijećima ne postoji zakonska obveza pružanja uredu izborni popis. Izborni popis dostupan je, na papiru, političkim agentima u pogledu izbora i često je dostupan za inspekciju u lokalnim bibliotekama ili registrima vijeća. Međutim, u gotovo svim slučajevima, izborni popis isporučuje se registrima elektronski u zamjenu za znatnu naknadu.

Druga vrsta javnih informacija je javna sudska informacija. To su detalji sudske presude okružnog suda (CCJs). Jedna od opcija, kao dio postupka za naplatu duga, je otići na okružni sud i podići tužbu  na presudu okružnog suda. Time se uspostavlja dug i može se prisiliti dužnika na djelovanje . Ako dužnik zatim otplati dug i te se informacije prenose u registar, presuda okružnog suda ne nestaje, ali će biti pokazana kao zadovoljena. Slično tome, ako postoji spor koji se odnosi na presudu okružnog suda, koja je dokazana u korist tužitelja, presuda okružnog suda može biti prikazana kao ispravljena. To se može dogoditi na primjer, ako je sudska tužba podignuta za zajednički dug kada su obveze jednog od dužnika riješene. U takvom slučaju, često s mužem i ženom, jedan od dužnika može podići ispravak kako bi pokazao da nepodmireni dug nije njihov, nego od njihovog partnera.

Presude okružnog suda su informacije u javnoj domeni. Neki županijski sudovi mogu dostaviti informacije elektroničkim putem, no u mnogim slučajevima, informacije će se unijeti iz papirnatih obavijesti i zapisa.

Važna stvar za shvatiti o ovoj kategoriji informacija – i izbornih popis i presuda okružnog suda- da je sve (ili gotovo sve) dostupno javnosti. Vrijednost koju registri dodaju jest mogućnost učitavanja tih informacija elektronički, jer time uvelike ubrzaju proces. Stoga, umjesto da se obratite na nekoliko lokalnih mjesta, što može potrajati nekoliko dana, ako ne i tjednima, sada možete pristupiti informacijama u nekoliko sekundi.

### 4.3.2 Prethodna pretraživanja

Kada zajmodavac istraži kreditnu referentnu agenciju, ta se istraga bilježi u datoteku potrošača. Kada drugi zajmodavac napravi naknadnu istragu, zapis o prethodnim

pretragama bit će vidljiv. Prethodna pretraživanja nose datum i pojedinosti o vrsti organizacije koja ga je provela – banka, osiguravajuće društvo, tvrtka kreditnih kartica, komunalne usluge, poštanski nalog itd.

Ono što datoteka potrošača ne otkriva je ishod istrage. Stoga potrošač može imati osam pretraga zabilježenih u periodu od dva tjedna od strane brojnih trgovačkih društava. Ne možemo reći koji od njih su povezani s ponudama od zajmodavaca i u kojim slučajevima je zajmodavac odbio zahtjev. Za one slučajeve u kojima je dana ponuda, ne možemo reći je li podnositelj zahtjeva prihvatio ponudu. Na primjer, potrošač bi mogao jednostavno kupovati kod nekoliko zajmodavaca kako bi našao najbolju ponudu, ali će uzeti samo jedan kredit. Ili postoji mogućnost da se potrošač seli, a istrage služe kao podrška financiranja kupnje namještaja, televizije, kuhinjskih jedinica itd. Daljnja mogućnost je da potrošaču očajnički nedostaje novca, stoga podnosi zahtjev za razne kredite ili kreditne kartice i namjerava uzeti sve što može dobiti.

Stoga, iako broj i uzorak prethodnih pretraživanja mogu biti od interesa, bilo zbog subjektivne procjene ili u kreditnoj bodovnoj kartici, to zahtijeva pažljivo tumačenje. U razvoju kreditne bodovne kartice, karakteristike koje bi mogle biti uključene su broj pretraga u posljednja 3 mjeseca, 6 mjeseci, 12 mjeseci i 24 mjeseca, kao i vrijeme od zadnjeg upita.

### 4.3.3 Zajednički doprinošene informacije

Prije mnogo godina, zajmodavci i registri su shvatili da postoji vrijednost u razmjeni informacija o ponašanju potrošača na računima. Stoga, na najjednostavniji način, nekoliko zajmodavaca može doprinositi detaljima trenutačnog provođenja njihovih osobnih kredita. Ako potrošač podnese zahtjev za osobnu pozajmicu, a oni trenutno imaju već jednu s jednim od suradnika, istraga u registru će osigurati pojedinosti o tome je li njihov postojeći kredit ažuran ili je u zaostatku, te nekoliko kratkih detalja o povijesti plaćanja.

To je razvijeno na mnogo načina, ali temeljne smjernice se nalaze u dokumentu na koji se zajmodavci i registri „pretplate“ – načela reciprociteta. U osnovi, zajmodavci i registri mogu vidjeti samo vrstu informacija koju i on sam doprinosi. Neki zajmodavci

doprinose podatke samo o nekim njihovim proizvodima, stoga vide samo detalje drugih zajmodavaca koji doprinose informacije za iste proizvode. Neki zajmodavci ne daju pojedinosti o svim svojim računima. Ti zajmodavci prilikom provođenja upita koji se odnose na ovaj proizvod , moći će vidjeti samo zadane informacije dostavljene od strane drugih zajmodavaca. To je slučaj čak i ako su i drugi zajmodavci također doprinijeli pojedinostima svih njihovih računa koji nisu u ozbiljnim nepovratima.
Kada zajmodavac provodi istragu, neće biti u mogućnosti otkriti s kojom tvrtkom su vezani postojeći objekti i proizvodi. Međutim, može vidjeti pojedinosti o vrsti proizvoda — revolving kredit, mail nalog, kreditna kartica, itd.

Načela reciprociteta ne samo da diktira da smo u mogućnosti vidjeti istu vrstu informacija kao što doprinosimo. Također diktira ograničenja stavljena na pristup informacijama ovisno o svrsi na koju će se staviti. Kao gruba smjernica, ograničenja su najmanja kada se informacije koriste za upravljanje postojećim računom s postojećim klijentom. Daljnja ograničenja uvode se kada se podaci upotrebljavaju za ciljanje postojećeg kupca na novi proizvod, tj., proizvod koji nemaju. Čak se i daljnja ograničenja mogu staviti na korištenje zajedničkih podataka za marketinške proizvode nekupcima.

### 4.3.4 Agregirane (zbirne) informacije

Nakon što su registri prikupili podatke iz izbornih popisa, te time pridonijeli zapisima od mnogih zajmodavaca, registri su u povoljnom položaju za stvaranje novih mjera koje bi mogle biti od koristi u kreditnoj procjeni. S dubinom informacija koje isporučuju zajmodavci, zajedno s izbornim popisima i zapisima poštanskog registra, stvaraju varijable na razini poštanskog broja. (Svaki pojedinačni poštanski broj ima između 15 i 25 dodijeljenih kuća. U velikim blokovima stanova, može biti jedan poštanski broj dodijeljen za svaki blok.) Stoga, ako se objedine te informacija, registri mogu stvoriti i izračunati mjere kao što su: postotak kuća na poštanskom broju s CCJ-om, postotak računa na poštanskom broju koji su ažurirani, postotak računa na poštanskom broju koji

su tri ili više plaćanja u zakašnjenju, postotak računa na poštanskom broju koji su zapisani u posljednjih 12 mjeseci.

Očito, zajmodavac pojedinačno ne može vidjeti ove informacije u cijelosti jer je pretraživanje indeksirana adresom koja je unesena. Međutim, registri su u stanju stvoriti takve mjere koje se, u nekim promjenama bodovanja, dokazuju kao vrijednima.

### 4.3.5 Dodana vrijednost registra

Kao što se može vidjeti iz gore navedene rasprave, kreditni registri pohranjuju ogromne količine informacija. Koristeći to, oni su u mogućnosti stvoriti i izračunati mjere. Međutim, oni također mogu razviti generičke kreditne bodovne kartice. Postoji ogromna količina podataka, stoga se uzima definicija „lošeg“. Ove kreditne bodovne kartice ne odnose se na specifično iskustvo zajmodavca. Oni se također ne odnose na specifične proizvode. Nadalje, loša definicija možda nije prikladna za specifičnu situaciju. Međutim, oni mogu biti iznimno korisni u najmanje tri okruženja.

Prvo okruženje je mjesto gdje je zajmodavac premalen da bi mogao izgraditi vlastitu kreditnu bodovnu karticu. Kako bi se omogućila neka od prednosti skoringa, zajmodavac kalibrira vlastito iskustvo protiv generičke kreditne bodovne kartice. Drugo okruženje je kada postoji novi proizvod. U takvim slučajevima, generička kreditna bodovna kartica može biti od koristi čak i za veće zajmodavce jer ne bi bilo dovoljno informacija za izgradnju prilagođene kreditne bodovne kartice za taj proizvod. Treće okruženje je u općem radu. Iako zajmodavac može imati kreditnu bodovnu karticu koja snažno diskriminira „dobre“ i „loše“ slučajeve, generička kreditna bodovna kartica omogućuje zajmodavcu da učini najmanje dvije stvari. Prva stvar je da je generička kreditna bodovna kartica može dodati na snagu zajmodavca vlastite kreditne bodovne kartice kako bi generička kreditna bodovna kartica bila u mogućnosti otkriti ako podnositelj zahtjeva ima poteškoća s drugim zajmodavcima. Druga stvar je da se generička kreditna bodovna kartica može koristiti za kalibriranje kvalitete zahtjeva koje zajmodavac prima. Na primjer, ako prosječni generički skor jedan mjesec je 5% manji nego u prethodnom mjesecu, zajmodavac ima kvazi-neovisnu mjeru kreditne kvalitete nedavne

serije zahtjeva. Zbog različitih ograničenja postavljenih na korištenje zajedničkih podataka, svaki registar gradi različite generičke kreditne bodovne kartice koristeći različite razine podataka.

Još jedan primjer razvoja i upotrebe generičke kreditne bodovne kartice pojavljuje se u Leonardu (2000). U ovom primjeru, kreditni registar je izgradio generičku kreditnu bodovnu karticu, ali je uključio samo slučajeve koji su tri plaćanja u zakašnjenju. Računi su klasificirani kao „dobri“ ako dođe do isplate u idućih 90 dana. Ova generička krditna bodovna kartica namijenjena je samo za diskriminaciju među računima kod kojih već postoji već tri plaćanja u zakašnjenju. Ne smije se koristiti za račune koji nisu u obračunu niti za nove zahtjeve.

Jedan problem o kojem moramo raspraviti je podudaranje nivoa. Kada se provodi kreditni referentni upit, zajmodavac podnosi ime i adresu. Međutim, u mnogim slučajevima, to se ne podudara točno s zapisom iz baze podataka zajmodavca. Na primjer, podnositelj zahtjeva možda je na obrascu za prijavu napisao "S. Jones", ali u zapisu je spremljeno pod nazivom Steven Jones. Adresa može biti napisana kao "The Old Mill, 63 High Street, London," ali zapis u bazi podataka je 63 High Street, London. Stoga, registri moraju imati neki način za podudaranje slučajeva i vraćanje slučajeve koji imaju podudaranje, ili skoro podudaranje, ili moguće podudaranje. Dakle, ili zajmodavac određuje razinu podudaranja ili registri mogu vratiti elemente upita s zastavom koja detaljno opisuje razinu podudaranja koja je postignuta.

U Velikoj Britaniji, kreditni registri djeluju u okviru pravnog okvira. Na primjer, svi potrošači imaju pravo, na plaćanje male naknade, primiti ispis onoga što se drži od njihovih datoteka u kreditnom registru. Podnositelj zahtjeva mora dati odobrenje za zahtjev prije upita. Također, ako zajmodavac želi imati ovlast za ponavljanje upita tijekom držanja računa, ovo odobrenje mora biti unaprijed dano. Ponovljeni upiti često se rade u upravljanju ograničenjima kreditnih kartica ili u procjeni odgovarajućih mjera koje treba poduzeti kada kupac padne u zakašnjenje.

Registri izgrađuju modele za procjenu kreditne sposobnosti poslovanja. To se često naziva kao kreditni limit. Drugim riječima, ako poduzeće ima kreditni limit od £10.000, to je najveći iznos potreban kreditorima za proširenje posla. Registar također

može procijeniti snagu poslovanja ne samo u smislu svoje financijske uspješnosti, ali i u pogledu trendova relevantne industrije i gospodarstva.

Za mala poduzeća dostupne informacije slične onima za potrošače. Što se više bližimo većim tvrtkama, javni podaci postaju standardiziraniji u njihovom formatu i sadržaju. Za nacionalna i multinacionalna poduzeća, vrijednost koju registar može dodati se umanjuje, stoga društva imaju tendenciju da uzmu ubrzani put do javnih informacija.

# Bibliografija


1. Robert W. Keener: „Theoretical Statistics“, Debt. Statistics, University of Michigan, Ann Arbor, USA, 2010
2. Javier Márquez: „An Introduction to Credit Scoring For Small and Medium Size Enterprises“, 2008, dostupno na https://siteresources.worldbank.org/EXTLACOFFICEOFCE/Resources/870892-1206537144004/MarquezIntroductionCreditScoring.pdf
3. Prof.dr.sc. Šarlija Nataša : Kreditna analiza, Osijek, 2008 http://www.efos.unios.hr/kreditna-analiza/wp-content/uploads/sites/252/2013/04/1_kreditna-politika-poduzeca.doc.pdf
4. Lyn C. Thomas, David B. Edelman, Jonathan N. Crook, “Credit Scoring and its Application”, 2002
5. Lewis,E.M.: “An Introduction to Credit Scoring”, Fair Isaac and Co., Inc, San Rafael, 1992.
6. Mays, E., editor: “Handbook of Credit Scoring”, Glenlake Publishing Company, Ltd., Chicago, 2001.
7. https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/linear_regression.pdf
8. Skalabh: „Simple Linear Regression Analysis“, Chapter 2 , ITT Kanpur, dostupno na http://home.iitk.ac.in/~shalab/regression/Chapter2-Regression-SimpleLinearRegressionAnalysis.pdf (04.kolovoz 2019.)
9. https://www.pmf.unizg.hr/_download/repository/PREDAVANJE11.pdf
10. Huzak M.,: kolegij Statistički praktikum, Linearna regresija, skripte sa vježbi, PMF, Zagreb, 2019, dostupno na https://web.math.pmf.unizg.hr/nastava/statpr/files/linearna.pdf
11. Štambuk A., Biljan-August M.: Regresijska i korelacijska analiza, Rijeka, 2013, dostupno na

https://www.veleri.hr/files/datotekep/nastavni_materijali/k_poduzetnistvo_s1/Kvantitativne_za_poduzetnike_Pr2_Izv.pdf

12. Cajner H.: Korelacija i regresija, skripte s predavanja, 2012
13. Huzak M., prof.dr.sc. Slijepčević S.: kolegij statistika, Regresijska analiza, skripte s predavanja, PMF, Zagreb, 2019, dostupno na https://web.math.pmf.unizg.hr/nastava/stat/files/StatRegresija.pdf
14. Prof.dr.sc. Šarlija Nataša : kolegij Upravljanje kreditnim rizicima, Osijek, 2008, dostupno na https://www.mathos.unios.hr/upravljanjekr/materijali/Kredit%20scoring%20modeli%20za%20stanovnistvo%20(tekst).pdf
15. http://www.poslovniforum.hr/poljoprivreda/hrok-krediti.asp

# Sažetak

Cilj ovog rada je bio upoznati se sa linearnom regresijom, te kako korištenjem regresijskih modela možemo odrediti klijentovu vjerojatnost otplaćivanja obveze na temelju prijašnjih podnositelja zahtjeva. Glavna pitanja su bila koliko bi trebao biti veliki uzorak s kojim radimo i kako uopće treba izgledati podjela na „dobre“ i „loše“ klijente, odnosno kako glasi definicija „dobrih“ i „loših“ klijenata, te je bilo potrebno odrediti razumnu povijest otplate. U radu smo spomenule neke od statističkih metoda za izgradnju kreditne bodovne kartice. Zaključak rada je da bez obzira koju definiciju „dobrih“, odnosno „loših“ odaberemo, ne postoji učinak na metodologiju kreditne bodovne kartice jer se ona zapravo razvija na temelju definicija.

# Summary

The aim of this work was to introduce the notion of linear regression, and explain how one can, by using regression models, determine the probability of a customer to pay off an obligation, based on behavior of previous applicants. The main questions were on how large the sample for analysis should be, and how we should divide data into "good" and "bad" clients, i.e. what does the definition of "good" and "bad" clients looks like, and it was necessary to determine a reasonable history of repayment. In the thesis we mentioned some of the statistical methods for building credit score cards. The conclusion of the work is that no matter what definition of "good" or "bad" we choose, there is no effect on the methodology of credit score cards because it actually develops on the basis of definitions.

# ŽIVOTOPIS

Rođena sam 09. prosinca 1994. godine u Sisku. Godine 2013. završavam Opću VII. gimnaziju u Zagrebu i upisujem Prirodoslovno-matematički fakultet u Zagrebu, smjer matematika nastavnički. Godine 2017. završavam diplomski studij i stječem titulu univ. bacc. educ. math. Iste godine upisujem Diplomski studij Financijska i poslovna matematika na Prirodoslovno-matematičkom fakultetu u Zagrebu.

Od 10.2015. do 07.2019. godine sam radila u Zagrebačkom gradskom kazalištu Komedija.