UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN
INFORMACIJSKE TEHNOLOGIJE

Zaključna naloga

# Uporaba logistične regresije za napovedovanje razreda, ko je število enot v preučevanih razredih različno

(Using logistic regression for class-predicition in a setting with unbalanced class distribution)

Ime in priimek: Šejla Žujo
Študijski program: Matematika v ekonomiji in financah
Mentor: doc. dr. Rok Blagus

**Koper, september 2015**

# Ključna dokumentacijska informacija

Ime in PRIIMEK: Šejla ŽUJO

Naslov zaključne naloge: Uporaba logistične regresije za napovedovanje razreda, ko je število enot v preučevanih razredih različno

Kraj: Koper

Leto: 2015

Število listov: 39 Število tabel: 4

Število prilog: 2 Število strani prilog: 7 Število referenc: 4

Mentor: doc. dr. Rok Blagus

Ključne besede: logistična regresija, napovedna točnost, učna množica, slučajno zmanjšanje, slučajno podvajanje, simulacije.

Math. Subj. Class. (2010): 62J12

**Izvleček:**
V nalogi raziskujemo, kako dobre so napovedi logistične regresije, ko obravnavamo neuravnotežene podatke. Problem neravnotežja se pojavi, ko imamo v učni množici veliko enot z določeno lastnostjo in malo enot brez nje. V nalogi predstavljamo osnovne lastnosti logistične regresije kot najuporabnejše metode za napovedovanje vrednosti odvisne spremenljivke, ki lahko zavzame le dve vrednosti. Za oceno vpliva velikosti neravnotežja na ocene mer točnosti smo uporabili simulacije. Pogledali smo celotno napovedno točnost, napovedne točnosti za posamezen razred in *G-povprečje*. Kot popravke za neravnotežje smo obravnavali metodo slučajnega zmanjšanja velikosti večjega razreda, slučajenga podvajanja enot iz manjšega razreda, poleg tega smo uporabili drugačno mejo za uvrščanje: delež enot manjšega razreda v učni množici. S pomočjo simulacij smo tudi preučevali spreminjanje napovednih točnosti za posamezen razred, ko spreminjamo velikost učne množice. Prav tako smo si ogledali rezultate simulacij, ko smo upoštevali različne velikosti razlike med razredoma. Na koncu smo predstavili rezultate analize pravih podatkov. Simulacije in analizo pravih podatkov smo izvedli s pomočjo programskega jezika R.

# Key words documentation

Name and SURNAME: Šejla ŽUJO

Title of final project paper: Using logistic regression for class-predicition in a setting with unbalanced class distribution

Place: Koper

Year: 2015

Number of pages: 39 Number of tables: 4

Number of appendices: 2 Number of appendix pages: 7 Number of references: 4

Mentor: Assist. Prof. Rok Blagus, PhD

Keywords: logistic regression, predictive accuracy, train set, down-sizing, over-sampling, simulations

Math. Subj. Class. (2010): 62J12

**Abstract:**
In the final project paper, we are researching, how good are predictions of logistic regression, when we deal with imbalanced data. The problem of imbalanced data appears when we have a training set with a large number of units which have given property and a few units without same property. In the final project we present basic properties of logistic regression as the most useful method for predicting the value of the outcome variable, which may take only two values. Using simulations, we checked the impact of the size of imbalance on the accuracy measures. We checked the overall predictive accuracy, sensitivity, specificity and G-means. As corrections for the class-imbalanced data we have considered and compared approaches like down-sizing and over-sampling, moreover, we used different boundary for class-prediction: the proportion of units from smaller class in the training set. Using simulations, we also studied change of the predictive accuracies as we change the size of the training set. Also, we looked results of simulations, when we considered different size of difference between classes. At the end, we represented the results of analysis of real data. Simulations and analysis of real data were performed by programming language R.

# Zahvala

Zahvaljujem se vsem, ki so verjeli vame, ko sama nisem.

# Kazalo vsebine

**1 Uvod** **1**

**2 Logistična regresija** **2**
2.1 Ideja in formalna oblika . . . . . 2
2.2 Interpretacija koeficientov . . . . . 5

**3 Uporaba logistične regresije za napovedovanje razreda** **8**
3.1 Meja za uvrščanje . . . . . 8
3.2 Mere točnosti . . . . . 9
3.3 Popravki za neravnotežje . . . . . 11

**4 Simulacije** **12**
4.1 Rezultati simulacij . . . . . 13
4.2 Povzetek najpomembnejših ugotovitev . . . . . 18

**5 Pravi podatki** **20**

**6 Zaključek** **22**

**7 Literatura in viri** **23**

# Kazalo tabel

1 V tabeli je prikazana celotna napovedna točnost ($PA$), napovedna točnost za razred 0 ($PA_0$) in za razred 1 ($PA_1$) ter *G-povprečje* ($G_{pov}$) za model logistične regresije v primeru nespremenjene učne množice (NC), metode slučajnega zmanjšanja (DS) in metode slučajnega podvajanja (OS). Uporabljeni so različni nivoji neravnotežja ($k_1$ = 0,05, 0,1, 0,3 in 0,5) in različne velikosti učne množice ($n$ = 100, 500 in 1000). $\Delta = 0$. . 13

2 V tabeli je prikazana celotna napovedna točnost ($PA$), napovedna točnost za razred 0 ($PA_0$) in za razred 1 ($PA_1$) ter *G-povprečje* ($G_{pov}$) za model logistične regresije v primeru nespremenjene učne množice (NC), metode slučajnega zmanjšanja (DS) in metode slučajnega podvajanja (OS). Uporabljeni so različni nivoji neravnotežja ($k_1$= 0,05, 0,1, 0,3 in 0,5) in različne velikosti učne množice ($n$ = 100, 500 in 1000). $\Delta = 0{,}5$. 15

3 V tabeli je prikazana celotna napovedna točnost ($PA$), napovedna točnost za razred 0 ($PA_0$) in za razred 1 ($PA_1$) ter *G-povprečje* ($G_{pov}$) za model logistične regresije v primeru nespremenjene učne množice (NC), metode slučajnega zmanjšanja (DS) in metode slučajnega podvajanja (OS). Uporabljeni so različni nivoji neravnotežja ($k_1$ = 0,05, 0,1, 0,3 in 0,5) in različne velikosti učne množice ($n = 100, 500$ in 1000). $\Delta = 1$. . 17

4 Celotna napovedna točnost ($PA$), napovedna točnost za razred 0 ($PA_0$) in za razred 1 ($PA_1$) ter *G-povprečje* ($G_{pov}$) za model logistične regresije v primeru nespremenjene učne množice (NC), metode slučajnega zmanjšanja (DS) in metode slučajnega podvajanja (OS) za $n$ = 100 in 300. . . . . 21

# Kazalo prilog

A Celotni rezultati simulacij
B Koda

# Seznam kratic

*ipd.* in podobno
*npr.* na primer
*OLS* Ordinary Least Squares
*oz.* oziroma
*idr.* in drugo
*gl.* glej

# 1 Uvod

Logistična regresija je najpogosteje uporabljena metoda za analizo podatkov, ki imajo binarni izid oz. pri katerih ima odvisna spremenljivka le dve vrednosti. Pogosto se ocenjeni logistični regresijski model uporablja za napovedovanje pripadnosti razreda, za enote pri katerih so vrednosti neodvisnih spremenljivk znane, ni pa znana vrednost odvisne spremenljivke (ang. *class-prediction*). Pomembna naloga pri napovedovanju razreda je ocena napovedne točnosti. V nalogi bomo s pomočjo simulacij in pravih podatkov raziskali, kako na ocene napovedne točnosti vpliva nivo neuravnoteženosti podatkov. Posebna pozornost bo namenjena uporabi ustrezne mere točnosti. Pričakujemo, da bo najprimernješa mera za ta namen *G-povprečje* (geometrijska sredina med napovedanima točnostima za posamezen razred). V simulacijah bomo preučevali tudi vpliv velikosti razlike med razredoma, ki je določena glede na velikost populacijskih regresijskih koeficientov. Obravnavali bomo tudi nekatere pristope za analizo neuravnoteženenih podatkov, in sicer slučajno zmanjšanje večjega razreda in slučajno povečanje manjšega razreda. Dobljene rezultate simulacij bomo prikazali s pomočjo pravih podatkov.
V drugem poglavlju zaključne naloge se bomo spoznali z formalno obliko logistične regresije in načina interpretacije njenih koeficientov. V tretjem poglavlju bomo predstavili uporabo logistične regresije za napovedovanje razreda. Poglavlje štiri obsegajo simulacije in rezultati simulacij. V petem poglavlju bomo obravnavali prave podatke.

# 2 Logistična regresija

V naslednjih dveh podpoglavjih bomo podali osnovno idejo za uporabo logistične regresije, njeno formalno obliko in interpretacijo njenih koeficientov.

## 2.1 Ideja in formalna oblika

Za koliko se spremeni število kadilcev, če se davek na cigarete poveča? Kako število učencev na učitelja vpliva na rezultate preverjanja znanja v srednjih šolah? Kaj se bo zgodilo z rastjo bruto domačega proizvoda, če se stopnja brezposlenosti poveča?

Odgovore na omenjena in podobna vprašanja lahko najdemo s pomočjo ocene modela linearne regresije. Njen cilj je oceniti povezanost neodvisne spremenljivke (ali več njih) z odvisno. Da bi ta vpliv čim bolje ocenili, se pri ocenjevanju parametrov modela linearne regresije uporablja metoda najmanjših kvadratov (ang. *Ordinary Least Squares-OLS*). Pri določenih predpostavkah o povezavi med odvisno in neodvisno spremenljivko in način vzorčenja (pridobivanja njihovih vrednosti) imajo cenilke parametrov, dobljene po OSL-metodi, množico koristnih in zaželenih statističnih lastnosti (za več gl. [2]).

Zdaj si lahko poskusimo odgovoriti še na nekatera druga vprašanja. Glede na višino dohodka osebe se lahko vprašamo, ali bo ta posojilo vrnila ali ne. Glede na spol in potovalni razred osebe na ladji Titanik se lahko vprašamo, ali je oseba preživela nesrečo ali ne. Glede na različne gene se lahko vprašamo, ali ima neka oseba določeno bolezen ali je nima. Opazimo lahko, da sta pri takih vprašanjih možna dva odgovora: da ali ne, kar jih znatno razlikuje od vprašanj na začetku poglavja. Tudi pri takih vprašanjih želimo oceniti povezanost neodvisne spremenljivke (ali več njih) z odvisno spremenljivko, ki pa lahko zavzame le dve vrednosti. V tem primeru ne moremo uporabiti OLS-metode za določanje nepristranskih cenilk parametrov, saj če lahko odvisna spremenljivka zavzame le dve vrednosti, cenilke, izračunane po tej metodi, nimajo več zaželenih statističnih značilnosti. [1] Vse to nakazuje na potrebo po novem modelu, s katerim bomo lahko dobili odgovore tudi na tovrstna vprašanja. Pri modelu logistične regresije ocenjujemo torej povezanost neodvisnih spremenljivk z odvisno spremenljivko, ki je dihotomna (ang. *indicator, binary, dummy variable*) oz. lahko zavzame le dve vrednosti.

*Opomba* 2.1. Cenilka parametra predstavlja pravilo, s pomočjo katerega ocenjujemo vrednosti parametra. Ta je nepristranska, ko je njena pričakovana vrednost enaka dejanski vrednosti parametra, ki ga ocenjujemo.

*Opomba* 2.2. Jasno bi moralo biti, da s pomočjo logistične regresije ni mogoče točno določiti eno od dveh vrednosti odvisne spremenljivke. Lahko pa z njo ocenimo verjetnost, s katero bo odvisna spremenljivka zavzela eno od dveh vrednosti, kar bomo predstavili v nadaljevanju.

Pri linernem in logističnem modelu regresije nas zanima pričakovana vrednost odvisne spremenljivke pri dani vrednosti neodvisne spremenljivke. Z nekaj oznak iz verjetnosti lahko ta sklep zapišemo kot $E(Y|X = x)$, kjer je $Y$ odvisna spremenljivka in je $x$ določena vrednost neodvisne spremenljivke $X$. Količina $E(Y|X = x)$ predstavlja *pogojno pričakovano vrednost $Y$ pri dani vrednosti $x$ neodvisne spremenljivke $X$.*

*Opomba* 2.3. Zaradi enostavnosti bomo namesto $E(Y|X = x)$ pisali $E(Y|x)$.

Zaradi boljšega razumevanja je najprej navedena formalna oblika modela linearne regresije, potem pa je predstavljen še model logistične regresije.

Enačba za univariantno linearno regresijo je definirana kot:

$$Y = \beta_0 + \beta_1 X,$$

pri čemer je $Y$ odvisna spremenljivka, $\beta_0$ in $\beta_1$ so populacijski parametri, ki jih je treba oceniti, in $X$ je neodvisna spremenljivka.

V pomenu pogojne pričakovane vrednosti lahko zapišemo:

$$E(Y|x) = \beta_0 + \beta_1 x.$$

Glede na to da pri linearni regresiji v smislu neodvisnih spremenljivk nismo omejeni z vrsto neodvisne spremenljivke (lahko je zvezna, diskretna idr.), je pričakovati, da je količna $E(Y|x)$ lahko katerakoli vrednost, ko se vrednost $x$ nahaja na intervalu $(-\infty, +\infty)$.

Če želimo govoriti o verjetnosti, bi pri logistični regresiji količina $E(Y|x)$ morala biti omejena z 0 in 1 oz. $0 \leq E(Y|x) \leq 1$.

V ta namen označimo $p(x) = E(Y|x)$ in definirajmo $p(x)$ kot logistično funkcijo:

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}. \tag{2.1}$$

Logistična funkcija je zelo uporabna, saj lahko kot vhodno vrednost dobi katerokoli vrednost z intervala $(-\infty, +\infty)$, ampak kot izhod vedno vrne vrednost med 0 in 1.

Enačbo (2.1) lahko transformiramo tako, da vzamemo inverzno funkcijo logistične funkcije iz enačbe (2.1). Tako dobimo

$$\begin{aligned} g(x) &= \ln\Big[\frac{p(x)}{1-p(x)}\Big] \\ &= \beta_0 + \beta_1 x. \end{aligned} \tag{2.2}$$

Transformacija $g(x)$ se imenuje logit transformacija in predstavlja izhodišče za analizo logistične regresije. Opazimo, da je $g(x)$ linearen v svojih parametrih in ima lahko vrednosti na intervalu $(-\infty, +\infty)$ glede na vrednost $x$ podobno kot model linearne regresije.

V nadaljevanju lahko ocenimo koeficiente logističnega modela. Na začetku uvedimo nekaj oznak in predpostavk. Predpostavimo, da imamo vzorec velikosti $n$ oz. $n$ opazovanih parov $(x_i, y_i)$, kjer $i = 1, 2, \ldots, n$, pri čemer $y_i$ predstavlja eno od dveh vrednosti odvisne spremenljivke $y$ za $i$-to opazovanje, $x_i$ pa vrednost neodvisne spremenljivke za $i$-to opazovanje. Pomembno je predpostaviti, da so pari $(x_i, y_i)$ in $(x_j, y_j)$ za $i, j \in \{1, 2, \ldots, n\}$, $i \neq j$ neodvisni. Za razumevanje te neodvisnosti lahko npr. predpostavljamo, da prisotnost bolezni $i$-tega v populaciji ni odvisna od prisotnosti bolezni $j$-tega v populaciji. Predpostavimo še, da lahko naša odvisna spremenljivka zavzame vrednost 1 ali 0 njeno odsotnost. Recimo, da vrednost 1 predstavlja prisotnost določene lastnosti in 0 odsotnost te. Omenjene vrednosti se bodo uporabljale v nadaljevanju.

*Opomba* 2.4. Namesto vrednosti 0 in 1 bi lahko vzeli katerekoli drugi dve vrednosti.

Pri ocenjevanju modela v enačbi (2.1) je treba na podlagi vzorca oceniti populacijske parametre $\beta_0$ in $\beta_1$, tako da se model čim bolj prilagaja populacijskim podatkom. Za ocenjevanje neznanih parametrov modela logistične regresije bomo uporabili *metodo največjega verjetja* (ang. *Maximum Likelihood Estimation*).

Za uporabo te metode moramo definirati funkcijo, s katero bomo izrazili verjetnost opazovanih vrednosti kot funkcijo neznanih parametrov. Cenilke neznanih parametrov bomo dobili kot vrednosti, ki maksimizirajo funkcijo največjega verjetja.

Poglejmo si bolj natančno enakost v (2.1). Ko odvisna spremenljivka Y zavzame vrednost 1, potem izraz $p(x)$ določa pogojno verjetnost, da je $Y$ enak 1, pri dani vrednosti $x$ oz. $p(x) = P(Y = 1|x)$. Potem je $1 - p(x) = P(Y = 0|x)$, kjer $P(Y = 0|x)$ predstavlja pogojno verjetnost, da je $Y$ enak 0, pri dani vrednosti $x$. Ker smo na začetku predpostavili, da so opazovanja neodvisna, lahko zapišemo:

$$\mathrm{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} p(x_i)^{y_i}[1 - p(x_i)]^{1-y_i}. \tag{2.3}$$

Funkcija (2.3) se imenuje *funkcija največjega verjetja*. Na njeni podlagi bomo izpeljali cenilke za neznane parametre.

*Opomba* 2.5. $\boldsymbol{\beta}$ predstavlja vektor parametrov in v primeru ene neodvisne spremenljivke je $\boldsymbol{\beta} = (\beta_0, \beta_1)$.

Metoda največjega verjetja zahteva, da kot oceno vektorja $\boldsymbol{\beta}$ določimo vrednost, ki maksimizira enačbo (2.3). Zaradi lažje obravnave smo enačbo (2.3) logaritmirali in z uporabo nekaj osnovnih lastnosti logaritmov zapisali:

$$l(\boldsymbol{\beta}) = \ln[\mathrm{L}(\boldsymbol{\beta})] = \sum_{i=n}^{n} \{y_i \ln[p(x_i)] + (1 - y_i)\ln[1 - p(x_i)]\}. \tag{2.4}$$

Vrednost $\boldsymbol{\beta}$, ki maksimizira $L(\boldsymbol{\beta})$, dobimo, ko odvajamo $L(\boldsymbol{\beta})$ glede na neznane parametre $\beta_0$ in $\beta_1$ ter dobljene odvode izenačimo z 0.

Z uporabo lastnosti odvoda dobimo naslednje enačbe za $\beta_0$ in $\beta_1$, zaporedoma:

$$\sum_{i=1}^{n} [y_i - p(x_i)] = 0, \tag{2.5}$$

$$\sum_{i=1}^{n} x_i [y_i - \rho(x_i)] = 0. \tag{2.6}$$

Enačbi (2.5) in (2.6) nista linearni v parametrih $\beta_0$ in $\beta_1$. Zaradi tega je postopek izračuna rešitev teh enačb otežen, toda z večino statističnih programov lahko izračun opravimo.

Ko enkrat dobimo rešitve teh enačb, vektor $\boldsymbol{\beta}$, določen s temi vrednostmi, predstavlja oceno po metodi največjega verjetja in se označi z $\hat{\boldsymbol{\beta}}$.

## 2.2 Interpretacija koeficientov

Podobno kot pri linearnem modelu nas tudi pri logističnem zanima sprememba v odvisni spremenljivki zaradi ustrezne enotske spremembe v neodvisni spremenljivki. Če si ponovno ogledamo enačbo (2.2)

$$\ln\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x,$$

vidimo, da koeficient $\beta_1$ pri vrednosti $x$ predstavlja vpliv te vrednosti na logit. Sledi, če želimo oceniti vpliv te vrednosti na odvisno spremenljivko, nam to dejstvo ne pomaga preveč. Iz tega razloga se spomnimo, da je $p(x) = P(Y = 1|x)$ in $[1 - p(x)] = P(Y = 0|x)$. Če te oznake vstavimo v enačbo (2.1), dobimo:

$$\ln\left[\frac{P(Y = 1|x)}{1 - P(Y = 0|x)}\right] = \beta_0 + \beta_1 x. \tag{2.7}$$

Izraz na levi strani enačbe (2.7) predstavlja logaritmiran obet, da ima posamezen subjekt v populaciji določeno lastnost oz. ima odvisna spremenljivka $Y$ v njegovem primeru vrednost 1. Na podlagi enačbe (2.7) dobimo povezavo med verjetnostjo in vrednostjo neodvisne spremenljivke, kar nam pomaga pri interpretaciji koeficientov.

Zaradi boljšega pregleda bomo v nadeljevanju interpretirali koeficiente modela logistične regresije, ko je neodvisna spremenljivka binarna in ko je zvezna.

**Primer 2.6.** (*Binarna neodvisna spremenljivka*)

Odvisno spremenljivko označimo z $Y$, ki je lahko 1 (prisotnost lastnosti) ali 0 (odsotnost lastnosti). Predpostavimo, da imamo v logističnem modelu eno neodvisno binarno spremenljivko $X$ z vrednostima 1 in 0. V tem primeru so obeti, da je lastnost prisotna med subjekti z neodvisno spremenljivko $X = 1$ enaki $P(Y = 1|X = 1)/[1 - P(Y = 1|X = 1]$. Za subjekte z neodvisno spremenljivko $X = 0$ so obeti za prisotnost lastnosti enaki $P(Y = 1|X = 0)/[1 - P(Y = 1|X = 0)]$. Glavno orodje pri interpretaciji koeficientov je razmerje obetov (ang. *odds ratio-OR*). Dobimo ga kot razmerje med obeti za $X = 1$ in $X = 0$. OR zapišemo z enačbo:

$$\mathrm{OR} = \frac{\frac{P(Y=1|X=1)}{1-P(Y=1|X=1)}}{\frac{P(Y=1|X=0)}{1-P(Y=1|X=0)}}. \tag{2.8}$$

Z zamenjavo verjetnosti v (2.8) z začetno definicijo $p(x)$ v (2.1) in nekaj računanja dobimo, da je:

$$\mathrm{OR} = e^{\beta_1}. \tag{2.9}$$

Z enačbo (2.9) smo torej dobili povezavo med razmerjem obetov in regresijskim koeficientom $\beta_1$. Recimo, da je $Y$= 1, če študent opravi izpit, in $Y$= 0, če ne. Naj bo $X$ neodvisna spremenljivka, ki določa pripravo študenta za izpit, in naj bo enaka 1, če se je študent učil, sicer pa 0. Recimo, da smo ocenili regresijske koeficiente, izračunali OR in dobili, da je OR = 3. To pomeni, da so obeti za neopravljeni izpit trikrat večji za študente, ki se niso učili, v primerjavi s tistimi, ki so se.

**Primer 2.7.** (*Zvezna neodvisna spremenljivka*)

Recimo, da nas zanima vpliv enotske spremembe v zvezni neodvisni spremenljivki $X$ oz. sprememba $X$ iz $x$ v $x + 1$ na odvisno spremenljivko, ki je dihotomna. Kot neodvisno spremenljivko vzemimo število ur priprav na izpit. Odvisna spremenljivka naj bo uspešnost študenta na izpitu. Če je izpit opravljen, bo odvisna spremenljivka imela vrednost 1, sicer pa 0. V tem primeru nas torej zanima vpliv povečanja priprave na izpit (u urah) na njegovo uspešnost. Zdaj predpostavimo, da smo ocenili neznane populacijske parametre $\beta_0$ in $\beta_1$ ter dobili, da je $\hat{\beta}_0$ = -4,5 in $\hat{\beta}_1$ = 0,6. Glede na ocenjene koeficiente dobimo naslednjo enačbo za logit:

$$g(x) = \ln\left[\frac{P(Y=1|x)}{1-P(Y=1|x)}\right] = -4{,}5 + 0{,}6x$$

kjer je $x$ poljubna vrednost spremenljivke število ur priprav na izpit.

Glede na ocenjeno vrednost za konstanto, $\hat{\beta}_0$, lahko rečemo, da so log obeti za opravljeni izpit za študenta, ki se ni pripravljal, $(X = 0)$, enaki $-4{,}5$. Drugače povedano: obeti za opravljeni izpit za študenta, ki se je pripravljal na izpit 0 ur, so enaki $e^{-4{,}5} = 0{,}011$.

Za interpretacijo ocenjenega koeficienta primerjajmo vrednosti $g(x)$ (enačba (2.10)) pri $x = 6$ in $x = 7$. Če odštejemo vrednost $g(6)$ od vrednosti $g(7)$, dobimo:

$$\ln\left(\frac{p(7)}{1-p(7)}\right) - \ln\left(\frac{p(6)}{1-p(6)}\right) = 0{,}6.$$

Z odštevanjem teh dveh enačb lahko sklepamo, da je koeficient pri neodvisni spremenljivki enak razliki med log obetoma. Če priprave na izpit podaljšamo za eno uro, je pričakovana sprememba v log obetu enaka 0,6. Opazimo, da je $\widehat{\mathrm{OR}} = e^{0{,}6}$ oz. $e^{\hat{\beta}_1}$ predstavlja cenilko za razmerje obetov.

Do zdaj smo obravnavali logistično regresijo v primeru le ene neodvisne spremenljivke. Navadno pa je treba preučiti vpliv več neodvisnih spremenljivk. Ko je odvisna spremenljivka dihotomna in je v analizo vključenih več neodvisnih spremenljivk, govorimo o *multivariatni logistični regresiji.*

Predpostavimo, da imamo $p$ neodvisnih spremenljivk in jih označimo z vektorjem $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$. Pogojno verjetnost, da odvisna spremenljivka $Y$ zavzame vrednost 1 glede na poljubno vrednost $\mathbf{x}$ vektorja $\mathbf{x}'$, označimo s $P(Y = 1|\mathbf{x}) = p(\mathbf{x})$. Sledi, da je logit transformacija v primeru multivariatne logistične regresije:

$$g(\mathbf{x}) = \ln\left(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p.$$

Podobno kot v primeru ene neodvisne spremenljivke imamo, da je:

$$p(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}}.$$

Ocene parametrov tudi tu dobimo po metodi največjega verjetja.

# 3 Uporaba logistične regresije za napovedovanje razreda

Do zdaj smo obravnavali obliko in značilnosti logistične regresije, vendar še nismo podali odgovorov na vprašanja, ki smo si jih zastavili na začetku. Dejansko nas zanima, kateremu razredu bo pripadla posamezna enota iz opazovane populacije. V ta namen potrebujemo pravilo, na podlagi katerega bomo ocenili pripadnost razredu. Imeti moramo torej pravilo, s pomočjo katerega bomo uvrščali enote iz preučevane populacije v ustrezni razred. Kot smo prej poudarili, ni mogoče točno določiti pripadnosti razredu, ampak le verjetnost, s katero bo posamezna enota pripadla enemu od razredov. Zaradi tega se v nadaljevanju ukvarjamo z oceno verjetnosti.

## 3.1 Meja za uvrščanje

Za oceno verjetnosti pripadnosti posameznemu razredu imamo vse, kar potrebujemo, in sicer imamo logit transformacijo in cenilke za populacijske parametre po metodi največjega verjetja.

Ko nas zanima razred ($y_i \in \{0, 1\}$) za $i$-to enoto v populaciji z $x_i$ vrednostjo neodvisne spremenljivke, ocenjene vrednosti za neznane parametre vstavimo v enakost za logit transformacjo (enačba (2.2), str. 4) in izračunamo verjetnost, da bo enota pripadla razredu 1 ($p(x_i) = P(Y = 1|X = x_i)$) ali razredu 0 ($1 - p(x_i)$). Dobljeno verjetnost moramo torej na neki način spremeniti v opisno spremenljivko oz. določiti pripadnost razredu. V ta namen moramo določiti mejo ($\rho$), tako da enoto uvrstimo v razred 1, ko je $P(y_i = 1|x_i) > \rho$, sicer v razred 0. Enota se uvrsti slučajno, ko je $P(y_i = 1|x_i) = p$. [4]

Večina statističnih programov uporablja mejo, ki je enaka 0,5, vendar bi v nekaterih primerih morali biti pri določanju meje bolj previdni. V nalogi preučujemo primere, kjer je število enot v razredih različno, oz. se ukvarjamo z uvrščanjem neuravnoteženih podatkov. Za uvrščanje enot v razrede velja, da je celoten nabor podatkov oz. že obstoječih opazovanj običajno razdeljen na učno in testno množico (ang. *learning set* in *test set*). Večina podatkov se uporablja za učno množico. Na podlagi učne množice se oceni koeficiente modela, njegovo delovanje pa se ovrednosti na testni množici. Ne-

ravnotežje v učni množici se pojavi, ko je en razred veliko manjši od drugega oz. porazdelitev števila enot v razredih ni enaka. V literaturi je neuravnoteženost znana kot *class imbalance problem*.

Intuitivno sledi, da bi morala biti meja pri uvrstitvi enot v primeru neuravnotežene učne množice drugačna. Za ponazoritev tega dejstva si oglejmo primer cenilke za regresijsko konstanto $\beta_0$ (formula (2.5)), ko imamo logistično regresijo brez neodvisnih spremenljivk. Naj bo $y_i \in \{0, 1\}$ razred za enoto $i$, $i = 1, 2, \ldots, n$, kjer je n velikost učne množice. Naj $\frac{1}{n}\sum_{i=1}^{n} y_i = k_1$ predstavlja delež enot iz učne množice, ki imajo $y_i = 1$.

Cenilko za $\beta_0$ v takšnem primeru dobimo z računanjem iz enakosti

$$\frac{\partial l(\beta_0)}{\beta_0} = \sum_{i=1}^{n} y_i - n \cdot \frac{e^{\hat{\beta_0}}}{1 + e^{\hat{\beta_0}}} = 0.$$

Dobljeno enačbo delimo z n in dobimo:

$$\frac{1}{n}\sum_{i=1}^{n} y_i - \frac{e^{\hat{\beta_0}}}{1 + e^{\hat{\beta_0}}} = 0.$$

Sledi

$$\hat{\beta_0} = \ln\Big(\frac{k_1}{1 - k_1}\Big).$$

Sklepamo, da je $P(y_i = 1) = k_1$, če v modelu ni nobene neodvisne spremenljivke, kjer je $k_1$ delež enot iz učne množice, ki imajo $y_i = 1$. Če bi se za uvrščanje uporabljala običajna meja 0,5, bi to povzročilo, da bi bile vse enote uvrščene v razred $y = 0$, če velja, da je v učni množici enot z $y = 1$ manj kot enot z $y = 0$.

## 3.2 Mere točnosti

Imeti model in verjetnosti na podlagi katerih bomo napovedali razred, ne pomeni, da imamo končen odgovor na zastavljena vprašanja. Lahko si jih sicer kot take vzamemo za končne, ampak do zdaj ostaja neznanka, kako natančno smo napovedali pripadnost razredu. Oceniti moramo torej, kako dobro model logistične regresije napoveduje razred. Pri logistični regresiji (tudi pri ostalih klasifikatorjih) sta najpogosteje uporabljeni meri kakovosti napovedovanja razreda: napovedana točnost in stopnja napake. Zaradi znanega izida opazovanj v testni množici je enostavno oceniti, ali so napovedi modela točne in v kolikšni meri. Glede na to da se v nalogi ukvarjamo z uvrščanjem v dva razreda, lahko definiramo dejanski pozitivni in negativni razred $\{+, -\}$. Ker na podlagi nabora podatkov iz testne množice vemo, katere enote smo uvrstili pravilno in katere ne, lahko vzpostavimo povezavo med napovedanim in dejanskim razredom.

V tem primeru lahko delovanje logistične regresije za napovedovanje razreda predstavimo s pomočjo naslednjih skupin:
- v skupini $DP$ (dejansko pozitiven) so vse enote, za katere smo pravilno napovedali pozitivni razred (dejanski in napovedani razred sta pozitivna),
- v skupini $DN$ (dejansko negativen) so vse enote, za katere smo pravilno napovedali negativni razred,
- v skupini $LP$ (lažno pozitiven) so vse enote, za katere smo napovedali pozitivni razred, ampak je dejanski razred negativen,
- v skupini $LN$ (lažno negativen) pa so vse enote, za katere smo napovedali negativni razred, ampak je dejanski razred pozitiven.

Glede na te oznake lahko definiramo napovedno točnost (ang. *Predictive Accuracy-PA*) kot:

$$PA = \frac{DP + DN}{DP + LP + DN + LN}$$

in stopnjo napake na naslednji način:

$$\textit{Stopnja napake} = 1 - PA.$$

Čeprav sta omenjeni meri preprosti, obstajajo primeri pri katerih podajata napačno oceno delovanja modela. Take težave se pojavijo pri neuravnoteženi učni množici. Pokažimo, da sta taki meri v tem primeru lahko zavajajoči. Denimo, da želimo napovedati razred in učna množica vsebuje 5 % pozitivnih in 95 % negativnih enot. Z naivnim uvrščanjem vseh enot v negativni razred bi dobili točnost 95 %, kar nakazuje na odlično delovanje modela logistične regresije. Vendar pa nismo pravilno uvrstili nobene enote iz manjšega razreda. [4] Iz tega razloga moramo upoštevati druge mere točnosti napovedi, da bi na čim boljši način prikazali delovanje uporabljenega modela. Pri učenju na neuravnoteženih podatkih sta znana še *G-povprečje* (ang.*G-mean*) in *F-mera* (ang. *F-measure*). Kot ustrezno mero točnosti napovedi logističnega modela v primeru neuravnoteženih podatkov bomo uporabli *G-povprečje*, ki je definirano kot:

$$\textit{G-povprečje} = \sqrt{\frac{DP}{DP + LN} \times \frac{DN}{DN + LP}}. \tag{3.1}$$

Izraz $DP/(DP + LN)$ je občutljivost modela (ang. *sensitivity*) in predstavlja verjetnost, da pravilno uvrstimo pozitivno enoto. Specifičnost modela (ang. *specificity*) je verjetnost, da pravilno uvrstimo negativno enoto, in je enaka $DN/(LP + DN)$. *G-povprečje* torej lahko zapišemo kot:

$$\textit{G-povprečje} = \sqrt{\textit{Občutljivost} \times \textit{Specifičnost}}. \tag{3.2}$$

Občutljivost in specifičnost sta meri, ki dajeta vpogled v samo delovanje klasifikatorja, medtem ko je *G-povprečje* mera, ustrezna za primerjavo med različnimi metodami za uvrščanje. [4]

## 3.3 Popravki za neravnotežje

V zadnjih letih se neuravnoteženosti podatkov posveča veliko pozornosti. Kot rezultat imamo danes več pristopov za reševanje tega problema. Nekateri znani pristopi so: metode zmanjšanja velikosti večjega razreda, metode povečanja velikosti manjšega razreda in metode združevanja klasifikatorjev. V nalogi bomo predstavili in uporabili dve metodi, in sicer metodo slučajnega zmanjšanja (ang. *down-sizing*) in metodo slučajnega podvajanja (ang. *over-sampling*). Cilj obeh metod je popravljanje porazdelitve enot v učni množici oz. ravnotežja. Recimo, da je učna množica razdeljena na manjši in večji razred. Velikost manjšega razreda označimo z $n_1$ in velikost večjega razreda z $n_2$.

**Slučajno zmanjšanje**
Pri metodi slučajnega zmanjšanja se osredinimo na slučajno vzorčenje iz večjega razreda. Iz njega izberemo podmnožico enot, na podlagi katere bomo določili želeno pravilo za uvrščanje. Cilj izbire podmnožice enot je pridobitev uravnotežene učne množice. Tako pridobljena učna množica naj bi imela velikost $2n_1$.

**Slučajno podvajanje**
Uravnoteženo učno množico s pomočjo metode slučajnega podvajanja dobimo tako, da ponavljamo slučajno izbrane enote iz manjšega razreda. V tem primeru naj bi nova učna množica imela velikost $2n_2$.

Pri obeh metodah bomo pravilo za uvrščanje izgradili na novi učni množici, delovanje logistične regresije pa bomo ocenili s pomočjo testne množice.

# 4 Simulacije

V tem poglavlju bomo poročali in primerjali ocenjene napovedne točnosti logističnega modela glede na različne pristope za učenje na neuravnoteženih podatkih. Ogledali smo si eno metodo zmanjšanja velikosti večjega razreda in eno metodo povečanja velikosti manjšega razred, in sicer slučajno zmanjšanje velikosti večjega razreda in slučajno podvanjanje enot iz manjšega razreda. Za večji razred si predstavljamo razred z enotami za katere velja, da je $Y = 0$ (razred 0) in za manjši obratno. Uporabili smo različne nastavitve za simulacije s ciljem čim boljše ocene delovanja logistične regresije za napovedovanje pripadnosti razredu. Pogledali smo različne nivoje neravnotežja v učni množici oz. pogledali smo ocenjene napovedne točnosti za primere, ko je bil v učni množici delež enot z $Y = 1$ ($k_1$) enak 0,05, 0,1, 0,2, 0,3, 0,4 in $k_1 = 0,5$ oz. ko je bila učna množica uravnotežena. Velikost učne množice ($n$) smo tudi spreminjali in pogledali rezultate v primerih, ko je $n$ =100, 200, 500, 1000. Pri vsaki vrednosti $n$ in $k_1$ smo pogledali, kaj se dogaja, ko spreminjamo velikost razlike med razredoma. Pogledali smo si primer, ko je bila razlika med razredoma ($\Delta$) enaka 0 oz. med razredoma ni nobene razlike, in primere, kjer je $\Delta$ = 0,5, 0,7, 1. Število neodvisnih spremenljivk ($p$) smo nastavili na 10 in ga nismo spreminjali. Predpostavljali smo, da so vse spremenljivke med seboj neodvisne. Za razred 0 smo vse spremenljivke za vse enote generirali iz standardne normalne porazdelitve (povprečje 0, varianca 1), za razred 0 pa iz normalne porazdelitve s povprečjem $\Delta$ in varianco 1. Velikost testne množice je bila vedno 1000 enot. Neravnotežje v testni množici je bilo enako kot v učni množici. Vsak korak simulacije smo ponovili 1000-krat. Mejo za uvrščanje smo nastavili na $k_1$ pri nespremenjeni učni množici. Pri metodi slučajnega zmanjšanja in slučajnega podvajana, ko že imamo novo uravnoteženo učno množico, je bila meja enaka 0,5. S pomočjo simulacij smo dobili oceno celotne napovedne točnosti, specifičnosti, občutljivosti in *G-povprečja* za vsako metodo. V nadaljevanju bomo uporabljali pojme napovedna točnost za večji in napovedna točnost za manjši razred namesto specifičnost in občutljivost.

Koda za simulacije (Priloga B) je zapisana v programskem jeziku R. [3]

## 4.1 Rezultati simulacij

Zaradi preglednosti naloge v rezultatih ne bomo predstavili in opisali vseh nastavitev simulacij, ki smo jih obdelali, ker so zelo podobni ostalimi.

*Opomba* 4.1. Rezultati vseh nastavitev za simulacije so v prilogi A.

Vse mere točnosti so podane v odstotkih.

| | $n = 100$ | | | | $n = 500$ | | | | $n = 1000$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $PA$ | $PA_0$ | $PA_1$ | $G_{pov}$ | $PA$ | $PA_0$ | $PA_1$ | $G_{pov}$ | $PA$ | $PA_0$ | $PA_1$ | $G_{pov}$ |
| $k_1 = 0{,}05$ | | | | | | | | | | | | |
| NC | 72 | 75 | 25 | 42 | 60 | 61 | 39 | 48 | 57 | 58 | 42 | 49 |
| DS | 50 | 50 | 50 | 48 | 49 | 50 | 50 | 49 | 50 | 50 | 50 | 50 |
| OS | 74 | 77 | 23 | 41 | 60 | 61 | 39 | 48 | 57 | 58 | 42 | 49 |
| $k_1 = 0{,}1$ | | | | | | | | | | | | |
| NC | 63 | 66 | 34 | 47 | 56 | 57 | 43 | 49 | 54 | 55 | 45 | 50 |
| DS | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |
| OS | 63 | 67 | 33 | 47 | 56 | 58 | 42 | 49 | 54 | 55 | 45 | 50 |
| $k_1 = 0{,}3$ | | | | | | | | | | | | |
| NC | 52 | 55 | 45 | 50 | 51 | 52 | 47 | 50 | 51 | 52 | 48 | 50 |
| DS | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |
| OS | 53 | 57 | 42 | 49 | 51 | 54 | 46 | 50 | 51 | 53 | 47 | 50 |
| $k_1 = 0{,}5$ | | | | | | | | | | | | |
| NC | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |
| DS | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |
| OS | 50 | 54 | 46 | 49 | 50 | 52 | 48 | 50 | 50 | 51 | 49 | 50 |

Tabela 1: V tabeli je prikazana celotna napovedna točnost ($PA$), napovedna točnost za razred 0 ($PA_0$) in za razred 1 ($PA_1$) ter *G-povprečje* ($G_{pov}$) za model logistične regresije v primeru nespremenjene učne množice (NC), metode slučajnega zmanjšanja (DS) in metode slučajnega podvajanja (OS). Uporabljeni so različni nivoji neravnotežja ($k_1 =$ 0,05, 0,1, 0,3 in 0,5) in različne velikosti učne množice ($n =$ 100, 500 in 1000). $\Delta = 0$.

Najprej smo pogledali ocene mer točnosti, ko je razlika med razredoma bila enaka 0 zato, ker tu vemo, da moramo dobiti $PA_0 = 0{,}5 = PA_1$. Vsi odmiki od tega, kažejo na pristransko metodo. Opazimo, da v primeru odsotnosti razlik med razredoma, nespremenjene prvotne učne množice in prisotnosti velikega neravnotežja ($k_1 = 0{,}05$, 0,1) dobimo veliko napovedno točnost za večji razred in majhno za manjši razred. Tako v primeru, ko je velikost učne množice 100 enot pri $k_1 = 0{,}05$, dobimo napovedno

točnost 75 % za večji razred in le 25 % za manjši, čeprav med razredoma ni razlike. Iz tabele 1 vidimo, da lahko v tem primeru z večanjem učne množice (za $n = 500$ in 1000) zmanjšamo razliko med napovednima točnostima, torej zmanjšamo problem neravnotežja. Natančneje: ko je bila velikost učne množice 100, je razlika med napovednima točnostima bila 54 odstotnih točk in v primeru, ko je $n$ bil 1000 je ta bila 16 odstotnih točk. Pri metodi slučajnega podvajanja enot iz manjšega razreda (OS) smo za vsako vrednost neravnotežja in velikost učne množice dobili skoraj enake napovedne točnosti kot v primeru nespremenjene učne množice, kjer smo kot mejo za uvrščanje uporabili delež enot z $Y = 1$. Tudi v tem primeru smo ugotovili, da lahko napovedno točnost za manjši razred izboljšamo, ko povečujemo velikost učne množice. Za veliko neravnotežje v učni množici pri nespremenjeni učni množici in učni množici, uravnoteženi s pomočjo slučajnega podvajanja enot iz manjšega razreda, dobimo velike razlike med napovednima točnostma za večji in manjši razred. Rezultati kažejo, da z uporabo metode slučajnega podvajanja dobimo za večji razred za 54 odstotnih točk večjo napovedno točnost od napovedne točnosti za manjši razred (pri $k_1 = 0{,}05$, $n = 100$) in za 34 odstotnih točk večjo, če je $k_1 = 0{,}1$. Ko ni razlik med razredoma, lahko razlike med napovednima točnostima za posamezen razred zmanjšamo z večjo učno množico (za $k_1 = 0{,}05$ in $n = 1000$ s slučajnim podvajanjem dobimo razliko 16 odstotnih točk).

Ko smo učno množico uravnotežili s pomočjo metode slučajnega zmanjšanja, smo dobili pričakovane napovedne točnosti. Ne glede na velikost neravnotežja in velikost učne množice so bile vse napovedne točnosti 50-odstotne. Pri uravnoteženi učni množici ($k_1 = 0{,}5$) smo v primeru nespremenjene prvotne učne množice in metode slučajnega zmanjšanja dobili vse napovedne točnosti enake 50 % ne glede na velikost učne množice. Z metodo slučajnega podvajanja pa smo v nekaterih primerih ($n = 500$) dobili napovedno točnost za manjši razred manjšo kot 50 %. Kot smo že omenili v poglavlju 3, je lahko celotna napovedna točnost zavajajoča mera in iz tega razloga smo preverili še *G-povprečje* kot mero točnosti. Pri $k_1 = 0{,}05$ in $n = 1000$ z metodo slučajnega podvajanja dobimo npr. celotno napovedno točnost enako 74 % in hkrati *G-povprečje* 41 %. Ne glede na uporabljeno metodo je bilo *G-povprečje* v vsakem primeru nad 41 %, in ne večje od 50 %. Pri metodi slučajnega zmanjšanja je bilo *G-povprečje* v vsakem primeru enako 50 %. Opazili smo, da tudi na *G-povprečje* pozitivno vpliva velikost učne množice, ne glede na prvotno neravnotežje in ne glede na uporabljeno metodo. Za uravnoteženo učno množico smo pri vsaki metodi in za vsak $n$ dobili *G-povprečje* enako 50 %, kar smo tudi pričakovali.

| | $n = 100$ | | | | $n = 500$ | | | | $n = 1000$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $PA$ | $PA_0$ | $PA_1$ | $G_{pov}$ | $PA$ | $PA_0$ | $PA_1$ | $G_{pov}$ | $PA$ | $PA_0$ | $PA_1$ | $G_{pov}$ |
| $k_1 = 0{,}05$ | | | | | | | | | | | | |
| NC | 85 | 87 | 42 | 59 | 80 | 80 | 72 | 76 | 79 | 79 | 75 | 77 |
| DS | 54 | 54 | 54 | 53 | 72 | 72 | 72 | 72 | 75 | 75 | 76 | 75 |
| OS | 86 | 89 | 38 | 57 | 79 | 80 | 70 | 75 | 79 | 79 | 75 | 77 |
| $k_1 = 0{,}1$ | | | | | | | | | | | | |
| NC | 80 | 82 | 60 | 70 | 79 | 79 | 75 | 77 | 78 | 79 | 77 | 78 |
| DS | 64 | 64 | 64 | 63 | 75 | 75 | 75 | 75 | 77 | 77 | 77 | 77 |
| OS | 80 | 83 | 57 | 68 | 78 | 79 | 75 | 77 | 78 | 79 | 77 | 78 |
| $k_1 = 0{,}3$ | | | | | | | | | | | | |
| NC | 76 | 77 | 73 | 75 | 78 | 78 | 77 | 78 | 78 | 78 | 78 | 78 |
| DS | 74 | 74 | 73 | 73 | 77 | 77 | 78 | 77 | 78 | 78 | 78 | 78 |
| OS | 76 | 78 | 70 | 74 | 78 | 78 | 77 | 77 | 78 | 78 | 78 | 78 |
| $k_1 = 0{,}5$ | | | | | | | | | | | | |
| NC | 76 | 76 | 75 | 75 | 78 | 78 | 78 | 78 | 78 | 78 | 78 | 78 |
| DS | 75 | 75 | 75 | 75 | 78 | 78 | 78 | 78 | 78 | 78 | 78 | 78 |
| OS | 75 | 77 | 72 | 74 | 78 | 78 | 77 | 77 | 78 | 78 | 78 | 78 |

Tabela 2: V tabeli je prikazana celotna napovedna točnost ($PA$), napovedna točnost za razred 0 ($PA_0$) in za razred 1 ($PA_1$) ter *G-povprečje* ($G_{pov}$) za model logistične regresije v primeru nespremenjene učne množice (NC), metode slučajnega zmanjšanja (DS) in metode slučajnega podvajanja (OS). Uporabljeni so različni nivoji neravnotežja ($k_1$= 0,05, 0,1, 0,3 in 0,5) in različne velikosti učne množice ($n$ = 100, 500 in 1000). $\Delta$ = 0,5.

Če primerjamo ista neravnotežja in iste velikosti učne množice v primeru, ko med razredoma ni razlike (tablela 1) in ko predpostavljamo razliko 0,5 (tabela 2), lahko opazimo, da v vsakem primeru dobimo večje mere točnosti. Tudi v primeru razlike med razredoma dobimo dosti manjše napovedne točnosti za manjši razred, če je prisotno veliko neravnotežje v prvotni učni množici ($k_1$ = 0,05, 0,1) in ko pravilo za uvrščanje izgradimo na nespremenjeni učni množici ter s pomočjo metode slučajnega podvajanja. Opazimo, da so se se razlike med napovednima točnostima za posamezen razred zmanjšale, ko med razredoma obstaja razlika. V primeru odsotnosti razlik med razredoma (tabela 1) za $n$ = 500 in $k_1$ = 0,05 dobimo napovedno točnost za manjši razred 39 % in za večji 61 %, medtem ko v primeru prisotnosti razlik (razlika = 0,5, tabela 2) pri istem neravnotežju in isti velikosti učne množice dobimo napovedno točnost za manjši razred 70 % in za večji 80 %. Kot v primeru ničelne razlike med razredoma tudi

v primeru njene prisotnosti dobimo rezultate, ki kažejo na večjo napovedno točnost za manjši razred pri večji velikosti učne množice. Vse napovedne točnosti in *G-povprečje* so bili podobni v primeru nespremenjene učne množice in metode slučajnega podvajanja. Iz tabele 2 lahko vidimo, da smo z uporabo metode slučajnega podvajanja pri velikem neravnotežju ($k_1 = 0{,}05$) in majhni velikosti učne množice ($n = 100$) dobili nerealistično celotno napovedno točnost. Če jo primerjamo z *G-povprečjem*, opazimo razliko za skoraj 30 odstotnih točk. Sklepamo lahko, da z metodo slučajnega zmanjšanja za iste vrednosti ($k_1 = 0{,}05$, $n = 100$) dobimo bolj smislene ocene točnosti. Pri metodi slučajnega zmanjšanja je celotna napovedna točnost 54 % in le 1 odstotno točko manjše *G-povprečje*. Z uporabo metode slučajnega zmanjšanja dobimo podobne vrednosti za vse mere točnosti ($PA, PA_0, PA_1$ in $G_{povp}$) in opazimo povečanje vseh mer točnosti (tudi *G-povprečje*) za približno 20 odstotnih točk, če velikost učne množice povečamo s 100 na 1000 pri neratvnotežju 0,05. Pri manjšem neravnotežju ($k_1 = 0{,}3$, $0{,}4$) dobimo vse mere točnosti med 70 % in 78 %. Za $k_1 = 0{,}3$, $0{,}4$ in $0{,}5$ pri $n = 1000$ dobimo za vsako metode vse mere točnosti enake.

| | $n = 100$ | | | | $n = 500$ | | | | $n = 1000$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $PA$ | $PA_0$ | $PA_1$ | $G_{pov}$ | $PA$ | $PA_0$ | $PA_1$ | $G_{pov}$ | $PA$ | $PA_0$ | $PA_1$ | $G_{pov}$ |
| $k_1 = 0{,}05$ | | | | | | | | | | | | |
| NC | 95 | 96 | 67 | 80 | 95 | 95 | 88 | 91 | 94 | 95 | 92 | 93 |
| DS | 62 | 62 | 62 | 61 | 87 | 87 | 87 | 87 | 87 | 91 | 91 | 91 |
| OS | 95 | 96 | 63 | 78 | 95 | 95 | 85 | 90 | 94 | 95 | 91 | 93 |
| $k_1 = 0{,}1$ | | | | | | | | | | | | |
| NC | 93 | 95 | 76 | 85 | 94 | 94 | 92 | 93 | 94 | 94 | 93 | 94 |
| DS | 80 | 80 | 80 | 80 | 91 | 91 | 91 | 91 | 93 | 93 | 93 | 93 |
| OS | 93 | 96 | 74 | 84 | 94 | 95 | 91 | 92 | 94 | 94 | 93 | 93 |
| $k_1 = 0{,}3$ | | | | | | | | | | | | |
| NC | 91 | 93 | 87 | 90 | 94 | 94 | 93 | 94 | 94 | 94 | 94 | 94 |
| DS | 88 | 88 | 88 | 88 | 93 | 93 | 93 | 93 | 94 | 94 | 94 | 94 |
| OS | 91 | 93 | 86 | 89 | 94 | 94 | 93 | 94 | 94 | 94 | 93 | 94 |
| $k_1 = 0{,}5$ | | | | | | | | | | | | |
| NC | 91 | 91 | 91 | 91 | 94 | 94 | 94 | 94 | 94 | 94 | 94 | 94 |
| DS | 91 | 91 | 91 | 91 | 94 | 94 | 94 | 94 | 94 | 94 | 94 | 94 |
| OS | 90 | 91 | 88 | 89 | 94 | 94 | 93 | 94 | 94 | 94 | 94 | 94 |

Tabela 3: V tabeli je prikazana celotna napovedna točnost ($PA$), napovedna točnost za razred 0 ($PA_0$) in za razred 1 ($PA_1$) ter *G-povprečje* ($G_{pov}$) za model logistične regresije v primeru nespremenjene učne množice (NC), metode slučajnega zmanjšanja (DS) in metode slučajnega podvajanja (OS). Uporabljeni so različni nivoji neravnotežja ($k_1 =$ 0,05, 0,1, 0,3 in 0,5) in različne velikosti učne množice ($n = 100, 500$ in 1000). $\Delta = 1$.

Na koncu simulacij smo pogledali rezultate v primeru, ko je razlika med razredoma enaka 1 (tabela 3). Podobno kot prej, ko smo višali razliko med razredoma (z 0 na 0,5 in z 0,5 na 0,7), smo v vsakem primeru (za vsak $k_1$ in vsak $n$) dobili večje napovedne točnosti in večje *G-povprečje*. Primerjali smo tudi razlike med merami točnosti znotraj posamezne metode in pri razliki 1 dobili najmanjša odstopanja. Za metodo slučajnega podvajanja je bila razlika med *G-povprečjem* in celotno napovedno točnostjo pri $k_1 =$ 0,05 in $n = 100$ enaka 17 odstotnih točk, medtem ko je ta razlika za iste vrednosti $k_1$ in $n$ bila 29 odstotnih točk, ko smo predpostavljali, da je razlika med razredoma enaka 0,5 (tabela 2).

Metoda slučajnega podvajanja je spet napovedala manjšo točnost za manjši razred pri velikem neravnotežju. Pri $n = 100$ za $k_1 = 0{,}05$ je bila na primer napovedna točnost za 33 odstotnih točk večja od napovedne točnosti za manjši razred. Pri večji

učni množici ($n = 500, 1000$) in pri istem neravnotežju smo opazili povečanje napovedne točnosti za manjši razred, in sicer, ko je $n$ bil 1000, je bila napovedna točnost za manjši razred enaka 91 %, kar predstavlja povečanje za 28 odstotnih točk. Celotna napovedna točnost metode slučajnega podvajanja je bila za $k_1 = 0{,}05$ in $n = 100$ 95 %, kar je morda naivno pričakovati pri neravnotežju 0,05, ampak se lahko pojasni z veliko razliko med razredoma. Z večanjem velikosti in manjšim neravnotežjem opazimo, da je pri metodi slučajnega zmanjšanja celotna napovedna točnost večja od 90 %. Kot pri drugih vrednosti razlike med razredoma (0, 0,5 in 0,7) tudi pri razliki 1 s povečanjem velikosti učne množice dosežemo večje *G-povprečje* in manjše razlike glede na celotno napovedno točnost.

Tudi pri razliki 1 smo z metodo slučajnega zmanjšanja pri velikem neravnotežju ($k_1 = 0{,}05$) dobili manjše ocene za napovedne točnosti v primerjavi z napovednimi točnostmi z metodo slučajnega podvajanja in pri nespremenjeni učni množici. Tako smo za $n = 100$ dobili vse napovedne točnosti enake 62 %, ki so za najmanj 30 odstotnih točk manjše od napovednih točnosti pri metodi slučajnega podvajanja. Rezultati so ponovno pokazali boljše mere točnosti, ko imamo večjo učno množico in še več: za večje $n$ ($n = 500, 1000$) razlike med merami točnosti sploh ni bilo. Mere točnosti pri metodi slučajnega zmanjšanja so se že pri $k_1 = 0{,}1$ izboljšale za približno 18 odstotnih točk v primerjavi z neravnotežjem 0,05 in pri majhni učni množici ($n = 100$). Pri manjšem neravnotežju ($k_1 = 0{,}3$, 0,4) in večji učni množici ($n = 500, 1000$) so vse mere točnosti bile 93 % ali 94 %. Enake rezultate smo dobili tudi pri metodi slučajnega podvajanja in ko učne množice nismo spreminjali. Pri uravnoteženi učni množici ($k_1 = 0{,}5$) z metodo slučajnega zmanjšanja in pri nespremenjeni prvotni učni množici smo dobili popolnoma enake pričakovane mere točnosti. Ko je bil $n$ 100, so vse mere točnosti bile enake 91 % in 94 %, ko je bil $n$ enak 500 in 1000. Skoraj enake mere točnosti smo dobili tudi v primeru slučajnega podvajanja, vendar je pri manjših $n$-jih napovedna točnost za manjši razred bila malo manjša od napovedne točnosti za večji razred (91 % za večji in 88 % za manjši razred, ko je $n = 100$).

## 4.2 Povzetek najpomembnejših ugotovitev

Na podlagi predstavljenih rezultatov, dobljenih s pomočjo simulacijskih nastavitev, lahko pridemo do različnih ugotovitev glede ocene delovanja logistične regresije za uvrščanje enot v razred (večji − razred 0 ali manjši − razred 1). Pri predpostavki, da med razredoma ni razlik, smo pričakovali, da bodo vse napovedne točnosti enake 50 %, ker se zdi inutitivno, da se v primeru odsotnosti razlik med razredoma enote uvrščajo v enega od razredov slučajno. Ko smo pravilo za uvrščanje izgradili na prvotni učni

množici in kjer je bila nova meja za uvrščanje delež enot z $Y = 1$ ($k_1$), smo dobili slabo napovedno točnost za manjši razred in presenetljivo veliko celotno napovedno točnost v primeru velikega neravnotežja ($k_1 = 0{,}05,\ 0{,}1$). Podobne ocene smo dobili z metodo slučajnega podvajanja. Pristranost teh ocen smo lahko opazili tudi na podlagi *G-povprečja*, ki se je v teh primerih zelo razlikoval od celotne napovedne točnosti. Po drugi stani smo v primeru odsotnosti razlik med razredoma pri metodi slučajnega zmanjšanja dobili vse napovedne točnosti (celotno in za posamezen razred) enake 50 % ne glede na velikost učne množice in kar je še pomembneje: ne glede na prvotno neravnotežje v učni množici. Zanesljivost metode slučajnega zmanjšanja smo lahko preverili s pomočjo rezultatov za oceno *G-povprečja*, ki je v vsakem primeru bil med 48 % in 50 %.

Ugotovimo, da lahko z večjo razliko med razredoma zmanjšamo razlike med merami točnosti znotraj posamezne metode. Napovedne točnosti pri metodi slučajnega zmanjšanja so bile tudi v primeru prisotnosti razlik (0,5, 0,7 in 1) skoraj enake za določeno velikost učne množice. S povečanjem razlike med razredoma smo opazili povečanje vseh mer točnosti. Ko je bila razlika enaka 1 in smo za izgradnjo pravila uvrščanja uporabili metodo slučajnega podvajanja in nespremenjeno učno množico, smo dobili veliko celotno napovedno točnost. Celotna napovedna točnost v obeh primerih je pri neravnotežju 0,05 bila enaka 95 %. Z uporabo metode slučajnega zmanjšanja pri večji učni množici ($n = 500, 1000$) in ne glede na neravnotežje smo dobili podobne napovedne točnosti, kot so te bile pri drugih dveh metodah. V praksi ni pogosto, da so razlike med razredoma velike in zatorej ni pogosto, da so mere točnosti tako velike. Če bi razlike med razredoma bile dejansko velike, potem bi bilo očitno, kateremu razredu bo pripadla enota. Ne glede na uporabljeno metodo smo ugotovili, da je mogoče z večjo učno množico doseči boljšo napovedno točnost za manjši razred.

# 5 Pravi podatki

Ugotovitve, dobljene na podlagi simulacij, smo preverili s pomočjo analize pravih podatkov. Na podlagi podatkov o osebah, ki so dobile posojilo, smo želeli napovedovati tveganje za nevračilo posojila banki. Natančneje: točnosti napovedi tega tveganja. Uporabili smo nemške kreditne podatke, ki so v celoti javno dostopni na naslednji spletni strani: http://archive.ics.uci.edu/ml/datasets.html. Za analizo nismo uporabili vseh podatkov, in sicer sicer zaradi manjkajočih informacij nismo uporabili vseh pojasnjevalnih spremenljivk. Celotna množica podatkov vsebuje podatke o 1000 osebah, pri čemer za vsako osebo poznamo vrednosti 14 neodvisnih spremenljivk, ki so vključene v analizo, in še vrednost odvisne spremenljivke. Slednja lahko zavzame vrednost 0 ali 1. Oseba z odvisno spremenljivko enako 1 je oseba, za katero se na podlagi znanih lastnosti ugotovi, da predstavlja tveganje oz. možnost nevračila posojila. Po drugi strani, je oseba z odvisno spremenljivko 0 oseba, za katero se je ugotovilo, da ne predstavlja tveganja za nevračilo posojila.

Neravnotežje v celotni množici podatkov je bilo 0,3 (oseb z odvisno spremenljivko enako 1 je 300). Iz tega sledi, da je oseb z odvisno spremenljivko enako 1 manj kot oseb z odvisno spremenljivko enako 0 oz. osebe z odvisno spremenljivko enako 1 predstavljajo manjši razred. Spremenljivke, vključene v analizo, se nanašajo na starost, trenutno stanje na tekočem računu, podatki o preteklem zadolževanju, garancije za posojilo ipd. Med spremenljivkami smo imeli 8 dihotomnih in 6 diskretnih spremenljivk.

S pomočjo funkcije `sample` v R programu smo naredili slučajno razdelitev celotne množice podatkov na učno in testno množico, pri čemer smo poskrbeli, da med pojasnjevalnimi spremenljivkami v testni množici ni vrednosti za odvisno spremenljivko. Ogledali smo si primera, ko je bila velikost učne množice 100 in 300.

Neravtnotežje v učni množici je bilo v obeh primerih 0,3.

| | $n = 100$ | | | | $n = 300$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $PA$ | $PA_0$ | $PA_1$ | $G_{pov}$ | $PA$ | $PA_0$ | $PA_1$ | $G_{pov}$ |
| $k_1 = 0{,}3$ | | | | | | | | |
| NC | 65 | 67 | 60 | 63 | 68 | 69 | 65 | 67 |
| DS | 56 | 51 | 66 | 58 | 68 | 68 | 66 | 67 |
| OS | 67 | 74 | 50 | 61 | 70 | 72 | 66 | 69 |

Tabela 4: Celotna napovedna točnost ($PA$), napovedna točnost za razred 0 ($PA_0$) in za razred 1 ($PA_1$) ter *G-povprečje* ($G_{pov}$) za model logistične regresije v primeru nespremenjene učne množice (NC), metode slučajnega zmanjšanja (DS) in metode slučajnega podvajanja (OS) za $n = 100$ in 300.

Tudi na pravih podatkih pri nespremenjeni učni množici in metodi slučajnega podvajanja opazimo večjo napovedno točnost za večji razred. Opazimo, da je bila največja napovedna točnost za večji razred pri metodi slučajnega podvanja za $n = 100$, in sicer 74 %. V istem primeru je bila napovedna točnost za manjši razred manjša za 24 odstotnih točk, čeprav neravnotežje v učni množici ni veliko ($k_1 = 0{,}3$) kar kaže na mogočo pristranost metode slučajnega podvajanja (enako kot pri simulacijah). Opazimo, da se pri večji učni množici ($n = 300$) razlika med napovednimi točnostmi za posamezni razred zmanjša, in sicer za 18 odstotnih točk.

Pri metodi slučajnega podvajanja *G-povprečje* ni preveč odstopalo od celotne napovedne točnosti, ampak je zopet bilo manjše. V simulacijah smo dobili skoraj enake ocene napovednih točnosti pri metodi slučajnega podvajanja in nespremenjeni učni množici, medtem ko so se pri pravih podatkih te ocene bolj razlikovale. Tako je napovedna točnost za manjši razred pri nespremenjeni učni množici za $n = 100$ bila enaka 60 % in pri metodi slučajnega podvajanja 10 odstotnih točk manjša. Podobno je bila napovedna točnost za večji razred pri metodi slučajnega podvajanja za 23 odstotnih točk večja kot pri metodi slučajnega zmanjšanja. Te razlike se zmanjšajo pri večji učni množici (npr. pri $n = 300$).

Z metodo slučajnega zmanjšanja tudi pri majhnih velikosti učne množice ($n = 100$) dosežemo boljšo napovedno točnost za manjši razred. Vidimo, da je ta večja od *G-povprečja* za 8 odstotnih točk. V primeru majhne učne množice ($n = 100$) opazimo razlike med meram točnosti glede na različne metode. Že pri velikosti 300 vidimo, da se te razlike zmanjšajo.

# 6 Zaključek

V prvem delu zaključne naloge smo spoznali osnovne značilnosti logistične regresije. Spoznali smo tudi, kako jo lahko uporabimo, da napovemo pripadnost razredu. Potem smo se osredotočili na uvrščanje enot v enega izmed dveh razredov v primeru, ko imamo neuravnoteženo učno množico, kar je dejansko tema naše naloge. Takšna učna množica lahko povzroči težave, ko želimo oceniti točnost napovedi logističnega modela. V ta namen smo se izognili uporabi naivne meje za uvrščanje enot, ki jo uporablja večina statističnih programov, in smo določili novo; delež enot manjšega razreda v učni množici. Predstavili smo nekatere znane popravke za neravnotežje, in sicer metodo slučajnega zmanjšanja velikosti večjega razreda in metodo slučajnega podvajanja enot iz manjšega razreda. Pokazali smo, da sta v primeru neravnoteženih podatkov tako celotna napovedna točnost kot mera točnosti modela lahko zavajajoči. Zato smo si ogledali še napovedne točnosti za večji in manjši razred posebej ter *G-povprečje*. Na podlagi rezultatov simulacij smo ugotovili, da so ocene napovednih točnosti, dobljene s pomočjo metode slučajnega podvajanja, zelo občutljive na neravnotežje v prvotni učni množici. Podobno smo sklepali tudi za napovedne točnosti, dobljene s pomočjo pravila, izgrajenega na podlagi nespremenjene učne množice (z novo mejo). Videli smo, da se je občutljivost zmanjšala, ko smo predpostavljali večje razlike med razredoma, kar smo tudi pričakovali. Opazili smo, da je metoda slučajnega zmanjšanja najmanj občutljiva na neravnotežje v prvotni učni množici. Za razliko od drugih dveh metod smo pri slučajnem zmanjšanju dobili manjše razlike med merami točnosti ne glede na neravnotežje in velikost učne množice. Naj posebej opozorimo na velike razlike med celotno napovedno točnostjo in *G-povprečjem* pri nespremenjeni učni množici in metodi slučajnega zmanjšanja, kar dokazuje, da je bil naš dvom o celotni napovedni točnosti upravičen. Pri vseh primerih, ki smo jih izvedli, smo ugotovili, da lahko s pomočjo metode slučajnega zmanjšanja izboljšamo napovedno točnost za manjši razred v vsakem primeru, kar se ni pokazalo pri drugih dveh metodah. Ko smo večali velikost učne množice, smo opazili zmanjšanje razlike med merami točnosti znotraj posamezne metode (tudi med različnim metodami) ne glede na velikost neravnotežja. Pri veliki razliki med razredoma in veliki učni množici so vse metode napovedovale podobne točnosti.

Na koncu naloge smo obravnavani problem predstavili še s pravimi podatki.

# 7 Literatura in viri

[1] D. W. Hosmer, R. X. Sturdivant in S. Lemeshow, *Applied Logistic Regression*. John Wiley & Sons, New Jersey, 2013. *(Citirano na strani 2.)*

[2] J. A. Rice, *Mathematical Statistics and Data Analysis*, Thomson Brooks/Cole, Third Edition, 2007. *(Citirano na strani 2.)*

[3] *Programski jezik R*, `https://www.r-project.org/`. (Datum ogleda: 1. 8. 2015.) *(Citirano na strani 12.)*

[4] R. Blagus, *Razvrščanje visoko-razsežnih neuravnoteženih podatkov*, doktorsko delo, Univerza v Ljubljani, 2011. *(Citirano na straneh 8, 10 in 11.)*

# Priloge

# A Celotni rezultati simulacij

| | $n = 100$ | | | | $n = 200$ | | | | $n = 500$ | | | | $n = 1000$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $PA$ | $PA_0$ | $PA_1$ | $G_{pov}$ | $PA$ | $PA_0$ | $PA_1$ | $G_{pov}$ | $PA$ | $PA_0$ | $PA_1$ | $G_{pov}$ | $PA$ | $PA_0$ | $PA_1$ | $G_{pov}$ |
| **$\Delta = 0$** | | | | | | | | | | | | | | | | |
| $k_1 = 0{,}05$ | | | | | | | | | | | | | | | | |
| NC | 72 | 75 | 25 | 42 | 65 | 67 | 33 | 46 | 60 | 61 | 39 | 48 | 57 | 58 | 42 | 49 |
| DS | 50 | 50 | 50 | 48 | 50 | 50 | 50 | 49 | 49 | 50 | 50 | 49 | 50 | 50 | 50 | 50 |
| OS | 74 | 77 | 23 | 41 | 66 | 68 | 32 | 46 | 60 | 61 | 39 | 48 | 57 | 58 | 42 | 49 |
| $k_1 = 0{,}1$ | | | | | | | | | | | | | | | | |
| NC | 63 | 66 | 34 | 47 | 60 | 61 | 39 | 48 | 56 | 57 | 43 | 49 | 54 | 55 | 45 | 50 |
| DS | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |
| OS | 63 | 67 | 33 | 47 | 60 | 62 | 38 | 48 | 56 | 58 | 42 | 49 | 54 | 55 | 45 | 50 |
| $k_1 = 0{,}2$ | | | | | | | | | | | | | | | | |
| NC | 55 | 59 | 41 | 49 | 54 | 56 | 44 | 49 | 52 | 54 | 46 | 50 | 52 | 53 | 47 | 50 |
| DS | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |
| OS | 56 | 61 | 39 | 49 | 54 | 57 | 42 | 49 | 53 | 55 | 45 | 50 | 52 | 53 | 47 | 50 |
| $k_1 = 0{,}3$ | | | | | | | | | | | | | | | | |
| NC | 52 | 55 | 45 | 50 | 50 | 54 | 46 | 50 | 51 | 52 | 47 | 50 | 51 | 52 | 48 | 50 |
| DS | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |
| OS | 53 | 57 | 42 | 49 | 52 | 55 | 44 | 50 | 51 | 54 | 46 | 50 | 51 | 53 | 47 | 50 |
| $k_1 = 0{,}4$ | | | | | | | | | | | | | | | | |
| NC | 51 | 52 | 48 | 50 | 50 | 52 | 48 | 50 | 50 | 51 | 49 | 50 | 50 | 51 | 49 | 50 |
| DS | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |
| OS | 51 | 56 | 44 | 49 | 51 | 54 | 46 | 50 | 50 | 53 | 47 | 49 | 50 | 52 | 48 | 50 |
| $k_1 = 0{,}5$ | | | | | | | | | | | | | | | | |
| NC | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |
| DS | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |
| OS | 50 | 54 | 46 | 49 | 50 | 53 | 47 | 50 | 50 | 52 | 48 | 50 | 50 | 51 | 49 | 50 |
| **$\Delta = 0{,}5$** | | | | | | | | | | | | | | | | |
| $k_1 = 0{,}05$ | | | | | | | | | | | | | | | | |
| NC | 85 | 87 | 42 | 59 | 81 | 82 | 62 | 71 | 80 | 80 | 72 | 76 | 79 | 79 | 75 | 77 |
| DS | 54 | 54 | 54 | 53 | 64 | 64 | 64 | 63 | 72 | 72 | 72 | 72 | 75 | 75 | 76 | 75 |
| OS | 86 | 89 | 38 | 57 | 82 | 83 | 58 | 69 | 79 | 80 | 70 | 75 | 79 | 79 | 75 | 77 |
| $k_1 = 0{,}1$ | | | | | | | | | | | | | | | | |
| NC | 80 | 82 | 60 | 70 | 79 | 80 | 70 | 75 | 79 | 79 | 75 | 77 | 78 | 79 | 77 | 78 |
| DS | 64 | 64 | 64 | 63 | 71 | 71 | 70 | 70 | 75 | 75 | 75 | 75 | 77 | 77 | 77 | 77 |
| OS | 80 | 83 | 57 | 68 | 79 | 80 | 68 | 74 | 78 | 79 | 75 | 77 | 78 | 79 | 77 | 78 |
| $k_1 = 0{,}2$ | | | | | | | | | | | | | | | | |
| NC | 77 | 79 | 70 | 74 | 78 | 78 | 74 | 76 | 78 | 78 | 77 | 78 | 78 | 78 | 77 | 78 |
| DS | 71 | 71 | 71 | 70 | 75 | 74 | 75 | 75 | 77 | 77 | 77 | 77 | 78 | 78 | 78 | 78 |
| OS | 77 | 79 | 68 | 73 | 77 | 78 | 73 | 76 | 78 | 78 | 76 | 77 | 78 | 79 | 77 | 78 |
| $k_1 = 0{,}3$ | | | | | | | | | | | | | | | | |
| NC | 76 | 77 | 73 | 75 | 77 | 78 | 75 | 77 | 78 | 78 | 77 | 78 | 78 | 78 | 78 | 78 |
| DS | 74 | 74 | 73 | 73 | 76 | 76 | 76 | 76 | 77 | 77 | 78 | 77 | 78 | 78 | 78 | 78 |
| OS | 76 | 78 | 70 | 74 | 77 | 78 | 74 | 76 | 78 | 78 | 77 | 77 | 78 | 78 | 78 | 78 |
| $k_1 = 0{,}4$ | | | | | | | | | | | | | | | | |
| NC | 76 | 76 | 75 | 75 | 77 | 77 | 77 | 77 | 78 | 78 | 78 | 78 | 78 | 78 | 78 | 78 |
| DS | 75 | 75 | 75 | 75 | 77 | 77 | 77 | 77 | 78 | 78 | 78 | 78 | 78 | 78 | 78 | 78 |
| OS | 75 | 77 | 72 | 74 | 77 | 78 | 75 | 76 | 78 | 78 | 77 | 77 | 78 | 78 | 78 | 78 |
| $k_1 = 0{,}5$ | | | | | | | | | | | | | | | | |
| NC | 75 | 76 | 75 | 75 | 77 | 77 | 77 | 77 | 78 | 78 | 78 | 78 | 78 | 78 | 78 | 78 |
| DS | 75 | 76 | 75 | 75 | 77 | 77 | 77 | 77 | 78 | 78 | 78 | 78 | 78 | 78 | 78 | 78 |
| OS | 74 | 77 | 72 | 74 | 76 | 77 | 75 | 76 | 78 | 78 | 77 | 78 | 78 | 78 | 78 | 78 |

| | $n = 100$ | | | | $n = 200$ | | | | $n = 500$ | | | | $n = 1000$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $PA$ | $PA_0$ | $PA_1$ | $G_{pov}$ | $PA$ | $PA_0$ | $PA_1$ | $G_{pov}$ | $PA$ | $PA_0$ | $PA_1$ | $G_{pov}$ | $PA$ | $PA_0$ | $PA_1$ | $G_{pov}$ |
| **$\Delta$ = 0,7** | | | | | | | | | | | | | | | | |
| $k_1$ = 0,05 | | | | | | | | | | | | | | | | |
| NC | 90 | 93 | 49 | 66 | 89 | 90 | 69 | 78 | 87 | 88 | 81 | 84 | 87 | 87 | 84 | 86 |
| DS | 58 | 58 | 58 | 56 | 71 | 71 | 70 | 70 | 79 | 79 | 80 | 79 | 84 | 84 | 84 | 84 |
| OS | 90 | 93 | 46 | 65 | 90 | 90 | 64 | 76 | 87 | 88 | 79 | 83 | 87 | 87 | 83 | 85 |
| $k_1$ = 0,1 | | | | | | | | | | | | | | | | |
| NC | 87 | 89 | 66 | 76 | 87 | 88 | 79 | 83 | 87 | 87 | 84 | 85 | 87 | 87 | 85 | 86 |
| DS | 71 | 71 | 70 | 70 | 78 | 78 | 78 | 78 | 84 | 84 | 84 | 84 | 85 | 85 | 85 | 85 |
| OS | 87 | 90 | 63 | 75 | 87 | 88 | 77 | 82 | 86 | 87 | 83 | 85 | 86 | 87 | 85 | 86 |
| $k_1$ = 0,2 | | | | | | | | | | | | | | | | |
| NC | 85 | 87 | 77 | 81 | 86 | 86 | 83 | 85 | 86 | 86 | 85 | 86 | 86 | 86 | 86 | 86 |
| DS | 78 | 78 | 78 | 78 | 83 | 83 | 83 | 83 | 85 | 85 | 85 | 85 | 86 | 86 | 86 | 86 |
| OS | 85 | 87 | 75 | 80 | 86 | 87 | 82 | 84 | 86 | 86 | 85 | 85 | 86 | 86 | 85 | 86 |
| $k_1$ = 0,3 | | | | | | | | | | | | | | | | |
| NC | 84 | 85 | 81 | 83 | 85 | 86 | 84 | 85 | 86 | 86 | 86 | 86 | 86 | 86 | 86 | 86 |
| DS | 81 | 81 | 81 | 81 | 84 | 84 | 84 | 84 | 86 | 86 | 86 | 86 | 86 | 86 | 86 6 86 | |
| OS | 84 | 86 | 79 | 82 | 85 | 86 | 83 | 84 | 86 | 86 | 85 | 86 | 86 | 86 | 86 | 86 |
| $k_1$ = 0,4 | | | | | | | | | | | | | | | | |
| NC | 84 | 84 | 83 | 83 | 85 | 85 | 85 | 85 | 86 | 86 | 86 | 86 | 86 | 86 | 86 | 86 |
| DS | 83 | 82 | 83 | 83 | 85 | 85 | 85 | 85 | 86 | 86 | 86 | 86 | 86 | 86 | 86 | 86 |
| OS | 83 | 85 | 80 | 83 | 85 | 86 | 84 | 85 | 86 | 86 | 85 | 86 | 86 | 86 | 86 | 86 |
| $k_1$ = 0,5 | | | | | | | | | | | | | | | | |
| NC | 84 | 84 | 83 | 84 | 85 | 85 | 85 | 85 | 86 | 86 | 86 | 86 | 86 | 86 | 86 | 86 |
| DS | 84 | 84 | 83 | 84 | 85 | 85 | 85 | 85 | 86 | 86 | 86 | 86 | 86 | 86 | 86 | 86 |
| OS | 82 | 84 | 80 | 82 | 85 | 86 | 84 | 85 | 86 | 86 | 85 | 86 | 86 | 86 | 86 | 86 |
| **$\Delta$ = 1** | | | | | | | | | | | | | | | | |
| $k_1$ = 0,05 | | | | | | | | | | | | | | | | |
| NC | 95 | 96 | 67 | 80 | 95 | 96 | 73 | 84 | 95 | 95 | 88 | 91 | 94 | 95 | 92 | 93 |
| DS | 62 | 62 | 62 | 61 | 81 | 81 | 80 | 80 | 87 | 87 | 87 | 87 | 91 | 91 | 91 | 91 |
| OS | 95 | 96 | 63 | 78 | 96 | 97 | 71 | 83 | 95 | 95 | 85 | 90 | 94 | 95 | 91 | 93 |
| $k_1$ = 0,1 | | | | | | | | | | | | | | | | |
| NC | 93 | 95 | 76 | 85 | 94 | 95 | 83 | 89 | 94 | 94 | 92 | 93 | 94 | 94 | 93 | 94 |
| DS | 80 | 80 | 80 | 80 | 86 | 87 | 86 | 86 | 91 | 91 | 91 | 91 | 93 | 93 | 93 | 93 |
| OS | 93 | 96 | 74 | 84 | 94 | 96 | 81 | 88 | 94 | 95 | 91 | 92 | 94 | 94 | 93 | 93 |
| $k_1$ = 0,2 | | | | | | | | | | | | | | | | |
| NC | 92 | 94 | 83 | 88 | 93 | 94 | 89 | 92 | 94 | 94 | 93 | 94 | 94 | 94 | 94 | 94 |
| DS | 86 | 86 | 86 | 86 | 90 | 90 | 89 | 89 | 93 | 93 | 93 | 93 | 94 | 94 | 94 | 94 |
| OS | 92 | 94 | 82 | 88 | 93 | 94 | 88 | 91 | 94 | 94 | 92 | 93 | 94 | 94 | 93 | 93 |
| $k_1$ = 0,3 | | | | | | | | | | | | | | | | |
| NC | 91 | 93 | 87 | 90 | 93 | 94 | 92 | 93 | 94 | 94 | 93 | 94 | 94 | 94 | 94 | 94 |
| DS | 88 | 88 | 88 | 88 | 91 | 91 | 91 | 91 | 93 | 93 | 93 | 93 | 94 | 94 | 94 | 94 |
| OS | 91 | 93 | 86 | 89 | 93 | 94 | 91 | 92 | 94 | 94 | 93 | 94 | 94 | 94 | 94 | 94 |
| $k_1$ = 0,4 | | | | | | | | | | | | | | | | |
| NC | 91 | 92 | 89 | 90 | 93 | 93 | 93 | 93 | 94 | 94 | 94 | 94 | 94 | 94 | 94 | 94 |
| DS | 90 | 90 | 90 | 90 | 92 | 92 | 92 | 92 | 94 | 94 | 94 | 94 | 94 | 94 | 94 | 94 |
| OS | 90 | 92 | 87 | 89 | 93 | 93 | 91 | 92 | 94 | 94 | 93 | 94 | 94 | 94 | 94 | 94 |
| $k_1$ = 0,5 | | | | | | | | | | | | | | | | |
| NC | 91 | 91 | 91 | 91 | 93 | 93 | 93 | 93 | 94 | 94 | 94 | 94 | 94 | 94 | 94 | 94 |
| DS | 91 | 91 | 91 | 91 | 93 | 93 | 93 | 93 | 94 | 94 | 94 | 94 | 94 | 94 | 94 | 94 |
| OS | 90 | 91 | 88 | 89 | 92 | 93 | 91 | 92 | 93 | 94 | 93 | 94 | 94 | 94 | 94 | 94 |

# B Koda

```
###dodatne funkcije:

my.paa<-function(pred,true){
pa<-sum(pred==true)/length(true)
pa0<-sum(pred[true==0]==true[true==0])/length(true[true==0])
pa1<-sum(pred[true==1]==true[true==1])/length(true[true==1])
gm<-sqrt(pa0*pa1)

c(pa,pa0,pa1,gm)
}

###parametri za simulacije:

n=1000 #velikost učne množice
k1=0.05 #delež enot z y=1 v učni množici
p=10 #število spremenljivk
razlika=0 #razlika med razredoma

n.test=1000 #velikost testne množice
k1.test=k1 #neravnotežje v testni množici (enako kot v učni množici)


B=1000 #število ponovitev simulacije

acc<-acc.ds<-acc.over<-matrix(NA,nrow=B,ncol=4) #mere točnosti za vsak
korak simulacije, imamo 3 matrike, eno za primer ko uporabljamo
neuravnoteženo učno množico (acc),
 eno za downsizing (acc.ds) in eno za oversampling (acc.over)
```

```
##začetek simulacije

for ( ii in 1:B) {
##generiranje podatkov za učno množico. Vse spremenljivke so
simulirane iz normalne porazdelitve in so med seboj neodvisne:

n0<-floor(n*(1-k1)) #število enot z y=0 (učna množica)
n1<-n-n0 #število enot z y=1 (učna množica)

train0<-matrix(rnorm(n0*p),ncol=p) #podatki za enote z y=0 (učna množica)
train1<-matrix(rnorm(n1*p,mean=razlika),ncol=p) #podatki za
enote z y=1 (učna množica)

train<-rbind(train0,train1) #podatki za celtno učno množico

class.train<-c(rep(0,n0),rep(1,n1))  #razred za enote
iz učne množice (y)

##generiranje podatkov za testno množico:

n0test<-floor(n.test*(1-k1.test)) #število enot z y=0 (testna množica)
n1test<-n.test-n0test #število enot z y=1 (testna množica)

test0<-matrix(rnorm(n0test*p),ncol=p)#podatki za enote
z y=0 (testna množica)
test1<-matrix(rnorm(n1test*p,mean=razlika),ncol=p) #podatki za enote
z y=1 (testna množica)

test<-rbind(test0,test1) #podatki za celtno testno množico

class.test<-c(rep(0,n0test),rep(1,n1test))  #razred za enote iz testne
množice (y)

###ocena regresijskega modela za enote iz učne množice:
trdf<-data.frame(y=class.train,train) #funkcija glm rabi, da so podatki
v obliki data.frame, zato rabimo to vrstico

fit<-glm(y~.,data=trdf,family=binomial(link="logit"))
```

```
###ocena verjetnosti p(y=1) za enote iz testne množice:

tedf<-data.frame(y=class.test,test) #funkcija predict želi, da so podatki
v obliki data.frame

pred.prob<-predict(fit,tedf,type="response")

###ocenjene verjetnosti spremeni v razred, kot mejo za uvrščanje
uporablja k1:

pred.class<-ifelse(pred.prob>k1,1,0)
pred.class[pred.prob==k1]<-sample(c(0,1),sum(pred.prob==k1),replace=T) #tiste
enote za katere velja P(y=1)=k1 uvrsti naključno

##izračuna mere točnosti in jih shrani:

acc[ii,]<-my.paa(pred.class,class.test)

#####del za down-sizing:

##izbera enot iz učne množice tako, da bo nova učna množica uravnotežena

id<-c(sample(1:n0,n1),which(class.train==1)) #indeksi izbranih enot

fit.ds<-glm(y~.,data=trdf[id,],family=binomial(link="logit")) #ocena
 regresijskega modela samo za izbrane enote

##oceni verjetnosti za enote iz testne množice

pred.prob.ds<-predict(fit.ds,tedf,type="response")

###ocenjene verjetnosti spremeni v razred, kot mejo za uvrščanje
uporablja 0.5:

pred.class.ds<-ifelse(pred.prob.ds>0.5,1,0)
pred.class.ds[pred.prob.ds==0.5]<-sample(c(0,1),
sum(pred.prob.ds==0.5),replace=T)
#tiste enote za katere velja P(y=1)=0.5 uvrsti naključno
```

```
##izračun in shranjevanje mer točnosti:

acc.ds[ii,]<-my.paa(pred.class.ds,class.test)

######del za over-sampling

id.over<-c(1:n0,sample(which(class.train==1),n0,replace=T)) #v novo učno
množico vključi v enote z y=0 in toliko kopij z y=1 da bo
nova mnočica uravnotežena

fit.over<-glm(y~.,data=trdf[id.over,],family=binomial(link="logit"))
#ocena regresijskega modela samo za izbrane enote

##ocena verjetnosti za enote iz testne množice

pred.prob.over<-predict(fit.over,tedf,type="response")

###ocenjene verjetnosti spremeni v razred, kot mejo za uvrščanje
uporablja 0.5:

pred.class.over<-ifelse(pred.prob.over>0.5,1,0)
pred.class.over[pred.prob.over==0.5]<-sample(
c(0,1),sum(pred.prob.over==0.5),replace=T) #tiste enote
za katere velja P(y=1)=0.5 uvrsti naključno

##izračun in shranjevanje mer točnosti:

acc.over[ii,]<-my.paa(pred.class.over,class.test)

} konec simulacije

##povzetek

apply(acc,2,mean) #to so pričakovane mere točnosti za te simulacijske
nastavitve za neuravnoteženo učno množico.
Prva vrednost je napovedna točnost,
```

```
druga vrednost je točnost za y=0 (PA0),
tretja točnost za y=1 (PA1) in četrta je g-povprečje
apply(acc,2,sd) #simulacijska variabilnost

##downsizing:
apply(acc.ds,2,mean) #to so pričakovane mere točnosti za te simulacijske
nastavitve, če uporabimo down-sizing.
Prva vrednost je napovedna točnost,
druga vrednost je točnost za y=0 (PA0),
tretja točnost za y=1 (PA1) in četrta je g-povprečje
apply(acc.ds,2,sd) #simulacijska variabilnost, ko uporabimo downsizing

##over-sampling:
apply(acc.over,2,mean) #to so pričakovane mere točnosti za te simulacijske
nastavitve, če uporabimo over-sampling.
Prva vrednost je napovedna točnost,
druga vrednost je točnost za y=0 (PA0),
tretja točnost za y=1 (PA1) in četrta je g-povprečje
apply(acc.over,2,sd) #simulacijska variabilnost, ko uporabimo oversampling

#vse skupaj:

mm<-rbind(apply(acc,2,mean),apply(acc.ds,2,mean),apply(acc.over,2,mean))
colnames(mm)<-c("PA","PA0","PA1","g-means")
rownames(mm)<-c("no correction","downsizing","oversampling")mm

par(mfrow=c(1,3))
boxplot(acc,names=c("PA",expression(PA[0]),expression(PA[1]),
"g-povprečje"),col="light gray",ylim=c(0,1))

boxplot(acc.ds,names=c("PA",expression(PA[0]),expression(PA[1]),
"g-povprečje"),col="light gray",ylim=c(0,1), main="down-sizing")
boxplot(acc.over,names=c("PA",expression(PA[0]),expression(PA[1]),
"g-povprečje"),col="light gray",ylim=c(0,1),main="over-sampling")
```