УНИВЕРЗИТЕТ У БЕОГРАДУ
МАТЕМАТИЧКИ ФАКУЛТЕТ

Ана Томић

Логистички регресиони модели кредитног ризика

— мастер рад —

Београд, 2017.

Матрматички факултет
Универзитет у Београду

Мастер рад

Логистички регресиони модели кредитног ризика

Студент:
Ана Томић 1058/2014

Ментор:
др Бојана Милошевић

септембар 2017.

Предговор

Развој савременог банкарства и привреде повећава изложеност различитим ризицима. Ризици, као могућност апсолутног или релативног губитка у односу на очекивани, су карактеристика банкарског пословања. Њихова идентифкација као и адекватне мере заштите постале су важан фактор успешног пословања.

Да би се ризик избегао или барем довео у прихватљиво стање, потребно је њиме управљати. Успешност финансијских институција заснива се на адекватној процени изложености кредитном ризику.

У првом поглављу овог рада описаћемо кредитни ризик. Упознаћемо се са параматрима које конфигуришу код кредитног ризика, описаћемо вероватноћу неиспуњења обавеза (вероватноћа неплаћања) и вероватноћу испуњења обавеза (вероватноћу опстанка). У наставку описаћемо моделе кредитног ризика, који за циљ имају смањење ризика на најмањи могући ниво.

У другом поглављу упознаћемо се са логистичком регресијом, описаћемо порекло логистичке функције. У овом поглављу представићемо логистичке регресионе моделе, креираћемо модел, одредити интервалне оцене, и представити процене слагања модела са подацима.

У трећем поглављу описаћемо моделе стабла одлучивања. Представити процес креирања стабла одлука, начин гранања, и избор оптималне величине стабла одлука.

У последњем поглављу приказаћемо на примеру претходно описане моделе. Описаћемо податке које ћемо користити, искористићемо логистичко регресионе моделе и моделе стабла одлучивања. И као крајњи резултат представићемо резултате добијене из модела и међусобно их упоредити.

Садржај

Поглавље 1

Кредитни ризик

1.1 Увод

Уопштено говорећи, кредитни ризик представља ризик да дужник *(obligator, reference entity)* не испуни своје обавезе у договореном (унапред одређеном) временском периоду T. Ако се овај догађај догоди, кажемо да се догодио догађај неиспуњења обавеза или догађај неизвршења финансијских обавеза *(default event)*.

Кредитни ризик је присутан у свакодневном животу. На пример, посматрајмо особу који долази у банку и аплицира за кредит како би купила одређену некретнину. Претпоставимо да је кредит одобрен од стране банке, која има договор са клијентом (дужником) да ће новац вратити после испуњења одређених критеријума и унапред одређеном временском периоду. У овој ситуацији кредитна институција (банка) је изложена ризику да та особа неће вратити кредит (део суме или цео износ узетог кредита), или особа неће испунити утврђене критеријуме.

Врста ризика са којом се банка суочава је управо кредитни ризик. Дужник је особа која аплицира за позајмицу (кредит). Догађај неиспуњења обавеза *(default event)* се остварује на дан када дужник није у стању да испуни своје обавезе.

Овај пример показује главне карактеристике кредитног ризика. Можемо видети да су овде укључене две стране: са једне је банка, која

се излаже ризику, док је на другој страни дужник (понекад се назива и поверилац) који мора да испуни низ обавеза. Приликом аплицирања за кредит, постоји низ критеријума који се овом приликом утврђују, као и којим редоследом ће ове обавезе бити испуњене, тј. низ критеријума који идентификују дужника. На крају, ризик је одређен у унапред задатом временском периоду $[0, T]$, где се T назива време доспећа *(maturity, the time horizon)*.

Овај пример такође показује и да постоје различити елементи које банке не познају у тренутку када издају кредит. Прво, банка не зна вероватноћу догађаја краха. Банке овај проблем покушавају да превазиђу прикупљањем разних информација о будућем дужнику, како би одредили вероватноћу да он неће бити у могућности да врати новац. Чак и ако претпоставимо да ће се десити догађај краха, неизвесно је када се он може догодити. Такође, и износ губитка је неодређен.

Не постоји јединствена дефиниција кредитног ризика. Дефиниција зависи од контекста и сврхе за коју неко жели да дефинише овај појам. Самим тим, можемо рећи да се ризик дефинише као вероватносна мера да се догоди догађај неиспуњења обавеза. На тај начин дефинисан ризик изражава опасност да ефекти будућих исхода догађаја одступају од очекиваних исхода.

Опште прихваћена дефиниција кредитног ризика гласи:

Дефиниција 1.1.1. *Кредитни ризик је специфична врста ризика који настаје при инвестирању финансијских средстава. Кредитни ризик је ризик да зајмопримац у финансијском уговору неће извршити обавезу делимично или у целини, што ће изазвати да инвеститор претрпи финансијски губитак.*

У финансијском свету, кредитни ризик се може окарактерисати у терминима: дужник (референтна особа), скуп критеријума који дефинишу догађај краха (неиспуњења обавеза), и временски интервал у којем је заступљен кредитни ризик.

Често, уместо дужника и кредита, говори се о компанијама и обвезницама. У овом случају крах може бити дефинисан на различите начине.

Још неки од примера догађаја краха могу бити: неплаћена обвезница[1] (на пример купон обвезнице), реструктуирање компаније, или спајање са другом корпорацијом [1].

У свету је мали број краха компанија, али имају велики утицај на финансијско тржиште. Кредитне агенције[2], као што су Муди, Стандард и Пур, свакодневно рачунају кредитни ризик за различита предузећа са тржишта [1]. Промена рејтинга компаније утицаће на цене свих сродних финансијских инструмената, као што су приноси корпоративних обвезница.

Вишеструки крахови су изузетно ретки, неки од њих су, на пример, природне катастрофе, системски крахови, политички догађаји, терористички догађаји.

У наставку описаћемо Базелски комитет, који је основан са циљем управљања кредитним ризиком, основне параметре који конфигуришу код кредитног ризика и моделе кредитног ризика.

1.2 Базелски комитет

Крајем 1974. гувернери земаља чланице G10 основали су Базелски комитет за банкарски надзор, који има за циљ унапређење банкарског надзора на нивоу целог света. Данас чланство у комитету има тринаест земаља: Белгија, Холандија, Француска, Канада, Јапан, Луксембург, Немачка, Италија, Шпанија, Велика Британија, САД, Шведска и Швајцарска [2].

Базелски комитет нема наднационални ауторитет контроле, и његови закључци немају правну снагу. Он формулише основне стандарде и препоручује најбоље праксе у својим документима, у циљу да ће их супервизори широм света применити на одговарајући начин за њихове

[1] Обвезница *(bond)* је финансијски инструмент, односно дужничка хартија од вредности, којом издавалац (дужник) признаје да има финансијску обавезу ка повериоцу, а служи повериоцу као доказ за то и као средство прикупљања наплате.

[2] Агенција за кредитне рејтинге (Credit Rating Agencies - CRA) је компанија која додељује кредитне рејтинге клијентима, тако што оцењује способност дужника да исплати дуг правовременим исплатама камата и главнице и вероватноћом неизвршења обавеза.

националне системе. Базелски комитет има лидерску улогу у успостављању смерница процене управљања ризиком банака.

Стандарди из међународно усаглашеног оквира за мерење адекватности капитала (Базел II) представљају обухватнији начин третирања изложености банака кредитном ризику у односу на Споразум о капиталу(Базел I), иницирајући код банака већу осетљивост у односу на изложеност и потребу да у континуитету развијају оквир за управљање овим ризиком.

Посебни квалитети Базелског споразума када је у питању кредитни ризик су:

1. Развојни пут приступа мерења изложености кредитном ризику и калкулације капитала, као и могућност избора између различитих приступа;

2. Велики нагласак на потреби да банке развијају своје интерне моделе за мерење изложености кредитном ризику и калкулацију економског капитала;

3. Нове технике за ублажавање изложености кредитном ризику, могућност избора, од једноставнијих, до сложенијих приступа овим техникама.

Понуђени сет приступа за мерење изложености кредитном ризику разликује се по нивоу софистицираности, прихватајући да нису све банке подједнако спремне да мере своју изложеност овом ризику, али и жељу комитета да и Базелски споразум буде широко примењен у светској банкарској индустрији и да постане светски стандард у домену мерења капиталне адекватности банака. Постоје три нивоа приступа [2]:

- Стандардизовани приступ кредитном ризику, који је најједноставнији,

- Основни виши приступ кредитном ризику, као први ниво приступа кредитном ризику који подразумева примену интерних методологија банака за мерење изложености и процену адекватности капитала само у односу на једну компоненту (вероватноћу неизвршења обавеза од стране дужника). За остале компоненте ризика банке морају да се ослоне на процену супервизора,

- Виши приступ кредитном ризику, као највећи ниво приступа кредитном ризику, који у потпуности омогућује мерење изложености и процену адекватности капитала применом интерних методологија банака.

Након Базеловог споразума *(Basel II Accord)* 2004. године, банке морају издвојити одређен износ капитала за покриће ризика који је својствен њиховим кредитним портфолијима. Предмет овог споразума је управо испуњење минималних условима и захтева за управљање ризиком. Компоненте ризика укључују процене вероватноће краха, стопе опоравка и изложеност краху. Ово, наравно, подстиче банке да инвестирају у моделима кредитног ризика са што већим бројем приступа.

Крајем јуна 2017. године у Србији је усвојен и у примени нов споразум (Базел III), чију израду су банке потврдиле током 2013. године. Циљ новог споразума је боље управљање ризицима, а детаљније се може видети у [3].

1.3 Параметри за мерење изложености кредитном ризику

Главни циљ овог рада је креирање модела за процену изложености ризику разних финансијских инструмената чија је вредност повезана са вероватноћом да ће се одређени дужник (или група дужника) доживети крах у одређеном временском интервалу $[0, T]$. Са овим циљем дефинисане су различите мере које помажу у одређивању изложености кредитном ризику. Вероватноћа неплаћања (Probability at defualt - PD) и додатне две компоненте ризика губици у случају неплаћања (Loss Given Default - LGD) и изложеност (Exposure at default - EAD) су кључни улазни параметри калкулације капитала. Самим тим валидација ових компоненти је кључна у процесу заштите од ризика.

У наставку описаћемо детаљније ове компоненте.

1.3.1 Вероватноћа неплаћања и вероватноћа опстанка

Задатак додељивања вероватноће неплаћања (Probability at default - PD) сваком клијенту, није уопште једноставан задатак. Постоје два основна приступа дефинисања ове вероватноће :

- Одређивање вероватноћа губитака на основу података са тржишта (Calibration of default probabilities from market data). Најпознатији концепт је очекиване фреквенције неизвршења финансијских обавеза[3], коју су представили Килхофер, МекКоун и Вашичек (KMB модел),

- Одређивање вероватноћа губитака на основу рангирања[4]. У овом случају су вероватноће неплаћања повезане са рејтингом.

Можемо дефинисати вероватноћу испуњења и неиспуњења обавеза [1]:

Дефиниција 1.3.1. *Вероватноћа испуњења обавеза - опстанка (survival probability - $P_{Surv}(t)$) је вероватноћа да дужник испуни одговарајуће обавезе у одређеном интервалу $[0,t]$, тј. да се крах не догоди у временском интревалу између 0 и t. Нека случајна променљива X представља временски тренутак у коме се десио крах. Тада је*

$$P_{Surv}(t) = P\{X > t\}.$$

У складу с тим, вероватноћа неиспуњења обавеза - краха (default probability - $P_{Def}(t)$) је вероватноћа да дужник не испуни обавезе у временском интервалу $[0,t]$. Тада је

$$P_{Def(t)} = P\{X \leq t\}.$$

1.3.2 Рангирање клијената

У основи рангирање клијента описује кредитну способност клијената. Користе се квантитативне и квалитативне информације за процену клијента. У пракси, поступак оцењивања се често више заснива на процени и искуству рангирања аналитичара, него на чисто математичким процедурама са строго дефинисаним исходима. У Сједињеним Америчким Државама и Канади, већина апликаната за кредит су оцењени барем од

[3]Очекивана фреквенција губитака (Expected Default Frequencies - EDF) је вероватноћа да компанија у неком тренутку неће извршити обевезе, тако што неће уплатити камату или главницу.

[4]Вероватноћа губитака на основу рангирања (Calibration of default probabilites from ratings) представља метод доделе ранга клијентима, рејтинг може бити додељен на основу агенција за доделу рејтинга (на пример Moody's Investors Services, Standard & Poor's (S&P)), или неким интерним банкарским приступом.

стране две рејтинг агенције[5].

На неразвијеним тржиштима банке морају користити свој систем рангирања. Особе које додељују рејтинг клијентима су кредитни аналитичари банака. Они морају узети у обзир многе различите параметре наведене компаније, као на пример: дуг, краткорочне и дугорочне обавезе, капитал компаније, ликвидност компаније, стање државе у којој се компанија налази, стање на тржишту...

Рејтинг је атрибут кредитне способности који се не може строго математички дефинисати. Представља резултат банкарских процена клијента на основу историјских података о клијенту (као што је историја плаћања, дуговни износ, број рачуна у банци, висина кредита) употребом статистичких алата[6], и након одређеног периода врши се поновна процена (ново рангирање истих као и нових клијената).

1.3.3 Изложеност

Изложеност (Exposure at default - EAD) је параметар који се користи за израчунавање економског капитала или регулаторног капитала под Базелским II комитетом за банкарску институцију. Представља бруто износ који дужник није вратио у тренутку догађаја неиспуњења обавеза, изражено у валути.

Изван Базел II, овај концепт је познат и као кредитна изложеност (Credit exposure - CE). Она представља непосредни губитак који би зајмодавац изгубио ако дужник (уговорна страна) у потпуности не исплати свој дуг.

Генерално гледано, EAD се посматра као процена у којој мери би банка могла бити изложена уговорној страни у случају и у тренутку неизвршења обавеза партнера. EAD је једнак текућем износу у случају неизмирених обавеза, као што су дугорочни кредити [4].

[5] У Сједињеним Америчким Државама постоји велики број агенција које раде процену кредитне спопобности клијената (неке од агенција су Moodi's, S & P i Fitch)

[6] Рејтинг клијената *(credit scoring)* је процена кредитне способности клијената, и за њено израчунавање је неопходна помоћ статистичких софтвера, најчешће коршћени су: SAS Business Intelligence, R, Stata, RapidMiner, итд.

1.3.4 Губитак услед неиспуњења обавеза

Губитак услед неиспуњења обавеза, краха (Loss Given Default - LGD), је удео губитка у тренутну када дужник не може да измири своје дугове, у односу на укупну изложеност, тј. представља проценат изложености губитку.

Ово је један од главних параметара код кредитног ризика, повезан је са параметром очекиваног губитка (EL - Expected Loss), као и регулаторног капитала на основу Базелског комитета.

Губитак услед неиспуњења обавеза (LGD) представља удео изгубљеног капитала у односу на укупан износ, који је изгубљен у тренутку неиспуњења обавеза дужника. Стопа спаса (Recovery Rate - RR) се дефинише $RR = LGD - 1$ и представља удео средстава који је враћен од тренутка када се десио крах.

Пример 1. *Претпоставимо да је клијент од банке узео кредит за куповину стана, али је дошло до краха клијента, и преостали дуг у тренутку краха износи* 200000 *долара. У том тренутку банка ради заплену стана (што представља депозит) и може га продати по нето цени од* 160000 *(укључујући и трошкове које се односе на откуп). Тада је удео гибитка услед неиспуњења обавеза једнак*

$$LDG = (200000 - 160000)/200000 = 0.2,$$

односно 20%, а стопа спаса је једнака $RR = 1 - LDG = 0.8$*, односно 80%.*

1.3.5 Очекивани губитак

Очекивани губитак (Expected Loss - EL) је средња вредност губитка ако се деси догађај неиспуњења обавезе током временског интервала $[0, T]$, где је T тренутак догађаја неиспуњења обавеза. Ако претпоставимо да је дата променљива губитка L:

$$L = \begin{cases} X, & \text{десио се догађај неиспуњења обавеза } E; \\ 0, & \text{иначе.} \end{cases}$$

тада очекивани губитак износи $EL = P(E) \cdot E(X)$.

Како би се заштитили од могућег неврaћања позајмица, банке често користе неку врсту осигурања од могућег губитка, како за лица која су

врло ризична, тако и за мање ризична. Наплаћивањем одговарајуће премије ризика (износ преко вредности кредита - камата) за сваки кредит и прикупљање ових премија ризика на интерном банковном рачуну, названом резервисања за очекиване губитке, створиће капитални јастук за покривање губитака који настају због неисплаћених кредита.

На финансијском тржишту је уведен додатни концепт приликом издавања кредита који има за циљ да се већина кредита отплати током времена, као и да се смањи износ неотплаћеног износа код кредита који нису враћени. Са друге стране, кредити су обично подржани хипотеком, чија се вредност мења временом у односу на неизмирену вредност кредита.

Банка сваком клијенту додељује одређену вероватноћу краха (default probability - PD), удео губитка која се назива губитак у случају неплаћања (loss given default - LGD), удео изложености кредита за који се очекује да неће бити враћен у случају неизвршења обавеза, и изложеност (exposure At Default -EAD) износ који ће бити изгубљен у одређеном временском периоду. На основу претходног, очекивана вредност губитака (EL) било ког дужника представља очекивану вредност променљиве губитка и износи:

$$EL = EAD \times LGD \times PD, P(D) = PD,$$

где је L губитак, D представља догађај да дужник не врати кредит у одређеном временском периоду, и $P(D)$ је вероватноћа овог догађаја [4].

1.4 Модели кредитног ризика

Будући да успешност пословања банке непосредно зависи од способности предвиђања и квалификације ризика, за њу је врло важно исправно проценити кредитну способност особа које аплицирају за позајмицу. Пре појаве модела кредитног ризика и развоја одговарајућих технологија потребних за њихову имплементацију, таква се процена заснивала искључиво на искуству и субјективном осећају банкарског аналитичара. Резултати таквих одлука често нису били задовољавајући. Последица је била да су кредити одобравани клијентима који су касније запали у потешкоће с отплатом зајма, док су одбијани захтеви оних који су били у могућности извршавати своје обавезе.

Модели кредитног ризика развијени су са циљем утврђивања вероватноће да клијент, с одређеним карактеристикама, неће бити у могућности да изврши своје обавезе.

С временом модели кредитног ризика постали су снажна подршка банкама у управљању кредитним ризиком. Према Хенд и Хенлеју[7] кредитни ризик је термин који се користи за описивање формалних статистичких метода класификације апликаната за кредит (ризичних и мање ризичних). Дакле, кредитни ризик се своди на проблем класификације, где су улазни подаци они који се односе на подносиоце захтева за кредит.

При креирању модела кредитног ризика потребно је бити свестан чињенице да се не може тачно класификовати сваки анализирани клијент. Савршена класификација је немогућа, будући да и добри и лоши клијенти понекад имају исте или сличне карактеристике. Због тога се при моделирању настоји одредити правило које као резултат даје најмањи могући број нетачних класификација [4].

Статистичка теорија нуди различите методе за оцењивање кредитног ризика. У овом раду пажњу ћемо посветити моделима логистичке регресије, као и моделима стабла одлучивања.

Регресија се користи за описивање и предвиђање зависне променљиве у односу на скуп независних променљивих. Како се бинарне променљиве појављују код кредитног ризика (на пример да ли клијент успешно испунио своје обавезе, или је дошло до краха), логистичка регресија је нашла примену у моделирању овог ризика. На основу ње се може проценити да ли ће дужник у потпуности испунити своје обазеве, или ће доћи до догађаја неиспуњења обавеза.

Стабло одлука је познато и као стабло класификације. Стабла представљају моделе који се састоје из скупа "ако-онда" услова дељења код случајева класификације на две или више различитих група. У случају бинарне класификације, сваки чвор стабла је додељен правилу одлучивања који описује узорак и дели га на две подгрупе. Тада се

[7] Хенлеј (W. E. Henley) и Хенд (D. J. Hand) су у раду "К најближих суседних класификатора за процену кредитног ризика дужника" ("A k-Nearest-Neighbour Classifier for Assessing Consumer Credit Risk") описали њихову методу за процену кредитног ризика на основу раздвајања скупа података у различите класе и рачунање удаљености између суседних класификатора.

процес посматрања развија наниже преко стабла у складу са правилом доношења одлука све до крајњег чвора у разгранатој шеми стабла, који тада представља класификацију овог стабла одлука [8].

У овом раду описаћемо детаљније ове статистичке методе и њихове примене у подручју управљања кредитним ризиком.

Поглавље 2

Логистичка регресија

Логистичка расподела је непрекидна расподела вероватноће чија је функција расподеле:

$$F(x) = \frac{1}{1 + e^{\frac{x-m}{s}}}, x \in \mathbb{R}.$$

У овом поглављу описаћемо настанак и развој логистичке расподеле, као и њене најважније особине.

2.1 Развој логистичке функције током времена

Логистчка функција је настала у 19. веку за потребе моделовања раста различитих популација. Различити истраживачи су се још током 18. века бавили проучавањем и предвиђањем раста људске популације у некој земљи.

Томас Малтус (1776-1834), економиста из Енглеске, је у свом раду "An essay on the principle of population as it affects the future improvement of society" из 1789. године изложио своје гледиште да се са повећањем броја становника повећава и количина произведених ресурса, хране и слично, али ово повећање расте аритметичком прогресијом, док раст броја становника прати геометријску прогресију. После одређеног броја година, ресурса ће бити мање, а становника који ће их користити више,

па ће тако завладати оскудице. Ово стање ће се временом погоршавати и добило је назив - *демографска (Малтусова) катастрофа*. Дошли су до закључка да је једини начин да се избегне или одложи катастрофа смањење прираштаја, што се може постићи повећањем смртности - намерно изазваним ратовима, болестима, оскудицама, или ограниченим рађањем.

Овај проблем се своди на проучавање неке количине $W(t)$ (на пример величина људске популације у временском тренутку t) и њеног прираштаја у јединици времена који ћемо обележити са $W'(t)$:

$$W'(t) = \frac{dW(t)}{dt}. \tag{2.1}$$

На примеру популације, пребројавањем долазимо до податка да у неком тренутку t_0 на Земљи живи $W(0)$ становника. Популација у следећем тренутку је сразмерна популацији у претходном јер раст становништва прати геометријску прогресију, односно $W(1) = rW(0)$, где је r параметар који описује нето прираштај становништва и може се добити из постојећих података.

Ако са γ означимо константну брзину рађања у јединици времена по јединки (стопа наталитета), а са δ константну брзина умирања у јединици времена по јединки (стопа морталитета), тада важи да је константан прираштај $\beta = \gamma - \delta$.

Ако је са $W(t)$ означен број јединки у тренутку t, онда је он после неког временског интервала Δt једнак

$$W(t + \Delta t) = W(t) + \beta W(t)\Delta t.$$

Видимо да је раст сразмеран постојећој популацији и времену, односно да постоји нека константа β таква за коју важи:

$$W'(t) = \beta W(t), \tag{2.2}$$
$$W(0) = W_0, \tag{2.3}$$
$$\beta = \frac{W'(t)}{W(t)}. \tag{2.4}$$

Решавањем ових диференцијалних једначина добијамо:

$$\frac{dW(t)}{W(t)} = \beta dt,$$
$$ln|W(t)| = \beta t + c,$$
$$W(t) = e^{\beta t}e^{c}.$$

Решење једначине (2.1) је:

$$W(t) = W_0 e^{\beta t}.$$

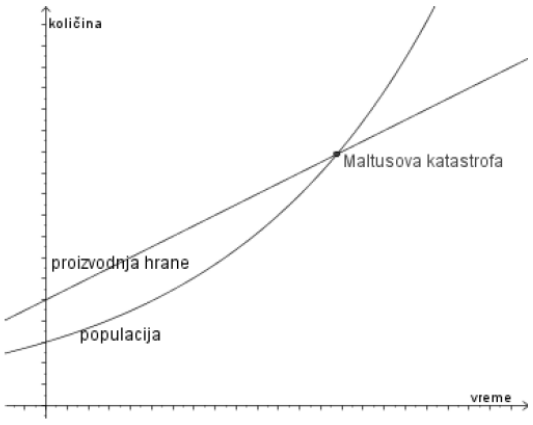
Па се долази до закључка да је раст популације експоненцијалан, односно да постоји нека константа A за коју важи:

$$W(t) = Ae^{\beta t},$$

где се за A често узима величина популације у почетном тренутку посматрања $W(0)$

$$W(0) = W_0 = Ae^0 = A.$$

Овај модел се показао као добар при проучавању младих популација, као што је на пример популација Сједињених Америчких Држава у првим деценијама по њиховом настанку, и назива се основни (Малтусов) популациони модел.



Слика 2.1: Основни Малтусов модел[1]

[1]Томас Малтус (Thomas Maltus) је овај модел описао у свом раду "An Essey on the Principle of Population", детаљније се може видети у [6].

Међутим, белгијски математичар и астроном Алфонс Кетле и његов млађи сарадник математичар Пјер-Франсоа Верхулст су приметили да овакво решење после неког времена доводи до нереалних процена и да би требало ограничити прираштај популације на неки начин. Они су сматрали да ниједна средина не може на себи да одржава неограничен број јединки, односно да раст популације треба ограничити до неке максималне фиксне вредности карактеристичне за систем који се посматра, односно до неког максималног носивог капацитета средине. Ограничени ресурси успоравају раст популације и популација тежи ка граничном засићену. Такође линеарне брзине рађања и умирања нису константе, и смањују брзину рађања, а увећавају брзину умирања са растом популације.

Они су у једначину (2.2) додали елемент $\phi(W(t))$ који представља отпор популације према даљем расту у тренутку t:

$$W'(t) = \beta W(t) - \phi(W(t)).$$

Верхулст је затим експериментисао са различитим облицима за $\phi(W(t))$ и дошао на идеју да уведе константу ω која би представља горњу границу засићености за W. Прираштај популације би тада био пропорционалан тренутној величини, али и њеном простору за даљи раст $\omega - W(t)$:

$$W'(t) = \beta W(t)(\omega - W(t)).$$

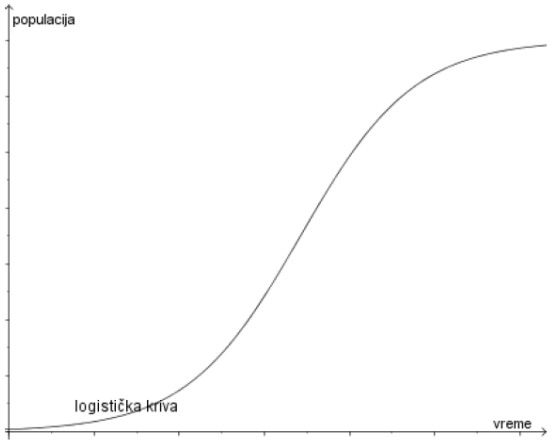
Увођењем смене $P(t) = \frac{W(t)}{\omega}$ добијамо следећу диференцијалну једначину:

$$P'(t) = \beta P(t)(1 - P(t)),$$

а њено решење је облика:

$$P(t) = \frac{e^{\alpha+\beta t}}{1 + e^{\alpha+\beta t}}.$$

Крива $P(t)$ има S-облик и Верхулст ју је назвао логистичком кривом (Слика 2.2). Овај модел је бољи него Малтусов модел, али има недостатке јер нису узети у обзир и многи спољашњи утицаји.



Слика 2.2: Верхулстова логистичка крива[2]

Логистичка расподела је поново откривена од стране Рејмонда Перла и Ловела Рида 1920. године. Они нису били упознати са Верхулстовим радом, али су сами дошли до сличних закључака приликом посматрања популације Сједињених Држава. После објављивања њихових радова, логистичка функција је примењивана за предвиђање величина различитих људских и животињских популација, међутим већина тадашњих математичара је давала већи значај пробит моделу који се базирао на нормалној расподели него логистичком.

Тек захваљујући развоју рачунарства у другој половини 20. века логистичка расподела стиче широку популарност. Њена предност у односу на пробит модел била је у једноставнијем облику и повољним аналитичким својствима који је чине много погоднијом за израчунавање уз помоћ различитих алгоритама.

Данас је логистичка расподела најпознатија по својој примени у моделима логистичке регресије и неких врста неуронских мрежа.

Дефиниција 2.1.1. *Стандардна логистичка функција (сигмоид крива), која је још позната и под називом основна логистичка функција, једнака је:*

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

[2]Детаљан опис модела се може видети у [7].

Ова једначина се добија као решење диференцијалне једначине првог реда:

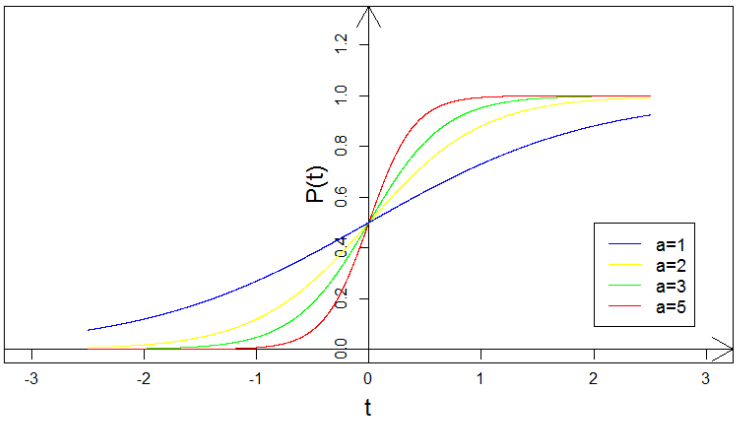
$$\frac{dP}{dt} = P(1-P),$$
$$P(0) = \frac{1}{2}.$$

Дефиниција 2.1.2. *Логистичка функција је строго растућа функција која се може приказати и у следећем облику:*

$$P(t) = \frac{1}{1+e^{-at}}, \tag{2.5}$$

где је a параметар нагиба сигмоидне функције.

Мењајући вредност параметра a, добијају се различити облици, што је приказано на Слици 2.3.



Слика 2.3: Стандардна логистичка функција[3]

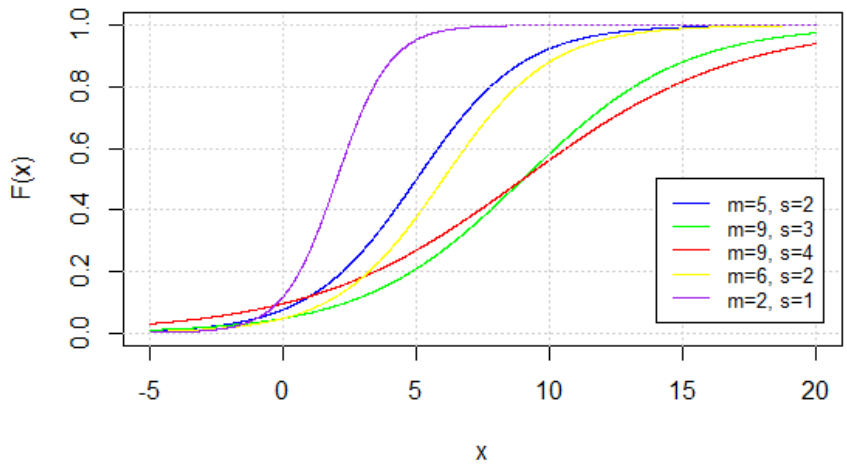
[3]Графикон је креиран у програмском језику R за различите вредности параметра a логистичке фунцкије $F(x) = \frac{1}{1+e^{-ax}}$, код се може видету у прилогу Б.1.

Логистичка расподела је симетрична расподела тешких репова.

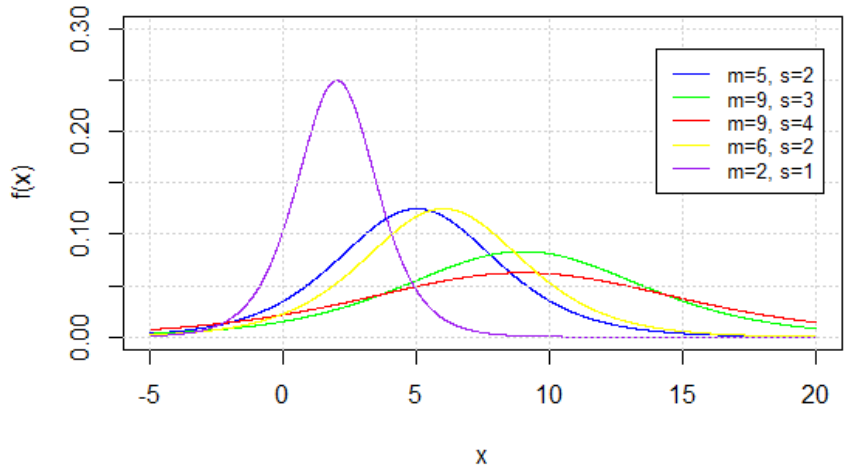
Дефиниција 2.1.3. *Нека је X случајна променљива са логистичком расподелом. Тада X има следећу функцију расподеле и густину расподеле:*

$$F(x) = \frac{1}{1 + e^{-\frac{x-m}{s}}}, s > 0, m \in R, x \in R,$$
$$f(x) = \frac{e^{-\frac{x-m}{s}}}{s(1 + e^{-\frac{x-m}{s}})^2}, s > 0, m \in R, x \in R.$$

На сликама (2.4.) и (2.5.) приказане су функције и густине логистичке расподеле за различите вредности параметара s и m:



Слика 2.4: Функција логистичке расподеле[3]



Слика 2.5: Густина логистичке расподеле[3]

Неке основне особине логистичке расподеле су:

- Очекивање: $E(X) = m$,
- Медијана: $\mu = m$,
- Мод: $mod = m$,
- Дисперзија: $D(X) = \frac{s^2\pi^2}{3}$,
- Коефицијент симетрије: $\gamma_1 = 0$,
- Коефицијент спљоштености: $\gamma_2 = \frac{6}{5}$.

2.2 Логистичка регресија

Логистичка регресија је један од модела уопштене линеарне регресије у коме зависна променљива узима само две вредности (бинарна променљива), док независне променљиве могу бити нумеричке, категоријске или њихова комбинација. Логистички регресиони модел се назива још и бинарни логистички регресиони модел (Binary Logistic Regression Model). Зависна променљива се кодира тако што се једном исходу додељује 1, а другом 0, при чему је свеједно који се исход кодира јединицом, а који нулом.

2.2.1 Основни модел

Логистичка (logit) трансформација се користи за предвиђање вероватноће наступања појаве која је кодирана јединицом.

Нека су дате n независне случајне промењљиве $X_1, ..., X_n$ на основу које треба предвидети вредности за Y и нека Y узима само две различите вредности, $G = \{0, 1\}$. Уместо директног предвиђања којој ће класи припадати Y, идеја логистичке регресије је оцењивање вероватноће да Y припадне свакој од класа, ако је вредност за $X_1, ..., X_n$ позната. Дакле, треба проценити следеће вероватноће:

$$P\{Y = 1 \mid X_1, ..., X_n\}, P\{Y = 0 \mid X_1, ..., X_n\}.$$

Означимо са:

$$P(X) = P\{Y = 1|X_1, ..., X_n\}.$$

Проблем се своди на оцењивање вредности $P(X)$. Како $P(X)$ представља неку вероватноћу, потребно је да функција којом се ова вредност моделира буде непрекидна, монотона и да узима вредности између 0 и 1.

Тврђење 2.2.1. *За креирање логистичко регресионог модела користи се логистичка функција следећег облика:*

$$P(X) = \frac{e^{\beta_0+\sum_{k=1}^{n}\beta_k X_k}}{1 + e^{\beta_0+\sum_{k=1}^{n}\beta_k X_k}}, \beta_0, \beta_1, ..., \beta_n \in R, \beta_1 \neq 0, ..., \beta_n \neq 0, \tag{2.6}$$

за коју важи да је непрекидна, монотона и да узима вредности између 0 *и* 1.

Очигледно је да функција (2.6) испуњава услове модела за $P(X)$. Једноставним трансформацијама из (2.6) добијамо следећу једнакост:

$$\frac{P(X)}{1 - P(X)} = e^{\beta_0+\sum_{k=1}^{n}\beta_k X_k}. \tag{2.7}$$

Израз $\frac{P(X)}{1-P(X)}$ се назива квотом и може да узме било коју вредност између 0 и ∞. Квоте се чешће од вероватноћа користе у моделовањима кредитног ризика: вредности близу 0 одговарају веома малим шансама да дужник не испуни своје обавезе и зато што је вредност квоте већа, то је већа и шанса негативног исхода, тј да се догоди догађај краха.

Применом природних логаритама на обе стране једначине (2.7) добијамо:

$$ln\Big(\frac{P(X)}{1 - P(X)}\Big) = \beta_0 + \sum_{k=1}^{n}\beta_k X_k. \tag{2.8}$$

Лева страна једначине (2.8) се назива логит трансформацијом од $P(X)$. Када вредност $P(X)$ припада интервалу $[0, 1]$, вредност логит функције се креће од $-\infty$ до $+\infty$.

Код логистичке регресије вредност зависне променљиве Y за дато X се може изразити као $Y|X = P(X) + \varepsilon$. Најчешћи је случај да променљива ε има нормалну расподелу са математичким очекивањем које је једнак нули и константном варијансом. Међутим, пошто је случајна променљива Y бинарна, ε има бинарну расподелу. Из чињенице да је $Y|X - P(X) = \varepsilon$ следи да променљива ε узима вредност $1 - P(X)$ са вероватноћом $P(X)$ (када је $Y = 1$) и вредност $-P(X)$ са вероватноћом $1 - P(X)$ (када је $Y = 0$) [14]. Дакле, променљива ε има следећу расподелу:

$$\varepsilon : \begin{pmatrix} 1 - P(X) & -P(X) \\ P(X) & 1 - P(X) \end{pmatrix},$$

а њено математичко очекивање и дисперзија износе

$$E(\epsilon) = 0,$$
$$D(\varepsilon) = P(X)(1 - P(X)).$$

2.2.2 Квота догађаја

Посматрајмо бинарну променљиву која узима вредност 1 ако се посматрани догађај десио и вредност 0 ако се догађај није десио. Средња вредност променљиве представља вероватноћу остварених догађаја тј. $\mu = P\{Y = 1\}$. На пример, ако је $\mu = 0.83$, тада се од свих посматраних догађаја остварило њих 83%.

Дефиниција 2.2.1. *Квота догађаја (odds) представља количник вероватноће да се догађај десио и вероватноће да се догађај није десио. Ако са $P(X)$ означимо вероватноћу да се догађај десио, тада квота догађаја износи:*

$$odds = \frac{P(X)}{1 - P(X)}.$$

На пример, ако је вероватноћа да се догађај деси 0.83, тада је квота догађаја $odds = \frac{0.83}{0.17} = 4.88$, што значи да догађај има 4.88 пута веће шансе да се деси него да се не деси. Ако је позната квота догађаја, вероватноћа догађаја се може израчунати као

$$P(X) = \frac{odds}{1 - odds}.$$

У случају да је вероватноћа остварења догађаја 0.5, тада је квота догађаја 1 и једнака је квоти комплементарног догађаја. Што је квота неког догађаја већа, већа је и вероватноћа да се тај догађај деси, па на тај начин можемо видети да ли је већа вероватноћа да се деси неки догађај или њему комплементарни догађај.

2.3 Параметри логистичке регресије

Посматрајмо логистички модел у коме је случајна променљива једнодимензиона:

$$P(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}. \tag{2.9}$$

Коефицијент β_1 се назива коефицијент нагиба и указује на стопу раста или опадања функције $P(X)$ у зависности од тога да ли је β_1 позитивна или негативна константа. Ако је $\beta_1 = 0$, тада је $P(X)$ константна за све вредности X, па крива логистичке регресије прелази у хоризонталну праву. Нагиб тангенте логистичке криве је дат изразом $\beta_1 P(X)(1 - P(X))$. Ако је, на пример, $P(X) = 0.5$ тангента логистичке криве има нагиб $0.25\beta_1$, а ако је $P(X) = 0.9$ или $P(X) = 0.1$, тај нагиб износи $0.09\beta_1$. Нагиб се приближава вредности 0, како се $P(X)$ приближава 0 или 1. Највећи нагиб тангенте се постиже када је $P(X) = 0.5$.

2.3.1 Оцењивање параметара

Модел логистичке регресије (2.6) зависи од параметара $\beta_0, \beta_1, \beta_2, ..., \beta_n$ које је потребно оценити из узорка. Приликом оцењивања параметара логистичке регресије користи се метод максималне веродостојности.

Овај метод даје вредности параметрима $\beta_1, \beta_2, ..., \beta_n$ које максимизирају вероватноћу добијања регистрованог скупа података.

Дефиниција 2.3.1. *Функција веродостојности*

$$L(\beta) = L(\beta_0, \beta_1, \beta_2, ..., \beta_n)$$

је функција непознатих параметара $\beta_0, \beta_1, \beta_2, ..., \beta_n$.

Под условом да је ова функција диференцијабилна, тражимо вредности параметара које су решење једначине

$$\frac{dL(\beta)}{d\beta} = 0.$$

Прелазимо на логаритам функције веродостојности, па ћемо посматрати фунцкију $\ln L(\beta)$.

Нека је дат узорак обима n регистрованих вредности парова $(\mathbf{X}_i, Y_i)$, где је $\mathbf{X}_i = (X_{1i}, X_{2i}, ..., X_{ni})$, а Y_i узима вредност 0 и 1 за свако $i = 1, 2, ..., n$. Формирамо функцију максималне веродостојности параметара на основу овог узорка:

$$L(\beta_0, \beta_1, ..., \beta_n) = \prod_{i=1}^{n} P(\mathbf{X}_i)^{Y_i}(1 - P(\mathbf{X}_i))^{1-Y_i}. \tag{2.10}$$

Применом логаритма на ову једначину и затим враћањем трансформације из (2.7) добијамо:

$$\begin{aligned}
\ln L(\beta_0, \beta_1, ..., \beta_n) &= \sum_{i=1}^{n} \ln(P(\mathbf{X}_i))^{Y_i} + \sum_{i=1}^{n} \ln(1 - P(\mathbf{X}_i))^{1-Y_i} \\
&= \sum_{i=1}^{n} [Y_i \ln(P(\mathbf{X}_i)) + (1 - Y_i) \ln(1 - P(\mathbf{X}_i))] \\
&= \sum_{i=1}^{n} Y_i \ln \frac{P(\mathbf{X}_i)}{1 - P(\mathbf{X}_i)} + \sum_{i=1}^{n} \ln(1 - P(\mathbf{X}_i)) \\
&= \sum_{i=1}^{n} Y_i \mathbf{X}_i^T \boldsymbol{\beta} - \sum_{i=1}^{n} \ln(1 + e^{\mathbf{X}_i^T \boldsymbol{\beta}}).
\end{aligned}$$

Оцене $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_n$ параметара $\beta_0, \beta_1, ..., \beta_n$ се добијају као решења следећег система једначина:

$$\frac{\partial \ln L(\beta_0, \beta_1, ..., \beta_n)}{\partial \beta_0} = 0, \tag{2.11}$$

$$\frac{\partial \ln L(\beta_0, \beta_1, ..., \beta_n)}{\partial \beta_1} = 0, \tag{2.12}$$

$$\cdots \tag{2.13}$$

$$\frac{\partial \ln L(\beta_0, \beta_1, ..., \beta_n)}{\partial \beta_n} = 0. \tag{2.14}$$

Сада ћемо навести теорему која се односи на функције веродостојности и која нам гарантује потребан и довољан услов да је статистика довољна:

Теорема 2.1. *Статистика $T_n = T_n(X)$ је довољна статистика за параметар β ако и само ако функција веродостојности $L(x; \beta)$ може бити представљена у облику $L(x; \beta) = g(T_n(x), \beta)h(x)$, где функција $h(x)$ не зависи од параметра β.*

Доказ теореме се може видети у [9].

У наставку навешћемо и неке основне теореме које се односе на оцене добијене методом максималне веродостојности.

Теорема 2.2. *Нека је T_n ефикасна оцена параметра β. Тада је T_n једино решење једначине веродостојности.*

Теорема 2.3. *Ако је T_n довољна статистика за параметар β, онда је свако решење једначине веродостојности функција од T_n.*

Теорема 2.4. *Ако је $\hat{\beta}$ оцена по методи максималне веродостојности за параметар β и $q(\beta)$ непрекидна функција која има инверзну функцију, тада је $q(\hat{\beta})$ оцена по методи максималне веродостојности за $q(\beta)$.*

Докази ових теорема се могу видети у [9] и [10].

2.3.2 Тестирање значајности параметара

Након оцењивања параметара вишеструке логистичке регресије, потребно је наћи најбољи модел, односно модел који најбоље описује зависну променљиву. Ово укључује формулисање и тестирање статистичких хипотеза за одређивање да ли значајно утичу на понашање зависне променљиве. Тестира се хипотеза

H_0: променљива (неколико променљивих) није значајна против алтернативе

H_1: променљива (неколико променљивих) је значајна.

Овим тестирањем поредимо регистроване вредности резултујуће променљиве са вредностима добијене помоћу два модела (где један од модела садржи а други не садржи променљиву чија се значајност тестира). Ако су предвиђене вредности на основу модела који садржи ту променљиву боље или тачније него вредности на основу модела који не садржи ту променљиву, тада је променљива у моделу значајна и модел који садржи ту променљиву се сматра бољим од модела који је не садржи. У наставку ћемо описати неке од начина за тестирање значајности параметара.

Тест количника веродостојности

Поређење регистроване и оцењене вредности добијене из модела који садржи независну променљиву и модела који је не садржи је базирано на логаритму функције веродостојности. При томе се сматра да је

регистрована вредност зависне променљиве она оцењена вредност која се добија из потпуног модела[4].

Када искористимо функцију макисмалне веродостојности из једнакости (2.10) добијамо:

$$D = -2\ln\frac{Lf}{Lp} = -2\sum_{i=1}^{n}(Y_i\ln\frac{\hat{Y}_i}{Y_i} + (1-Y_i)\ln\frac{1-\hat{Y}_i}{1-Y_i}),$$

где је $\hat{Y}_i = \hat{Y}_i(X_i)$, статистика D се назива одступање или девијација *(deviance)*, Lf функција веродостојности оцењеног модела, а Lp је функција веродостојности потпуног модела, док се израз $\frac{Lf}{Lp}$ назива количник веродостојности.

При испитивању значајности независне променљиве посматрају се модели са и без независне променљиве. Означимо са G промену која настаје у D приликом укључивања независне променљиве, односно:

$$\begin{aligned}
G =& D(\text{модел без независне променљиве}) \\
&- D(\text{модел са независном променљивом}) \\
=& -2\ln\frac{\left(\frac{\sum Y_i}{n}\right)^{\sum Y_i}\left(\frac{n-\sum Y_i}{n}\right)^{n-\sum Y_i}}{\prod_{i=1}^{n}\hat{Y}_i^{Y_i}(1-\hat{Y}_i)^{1-Y_i}} \\
=& 2\sum_{i=1}^{n}\left(Y_i\ln\hat{Y}_i + (1-Y_i)\ln\hat{Y}_i\right) \\
&- 2\left(\sum Y_i\ln\sum Y_i + (n-\sum Y_i)\ln(n-\sum Y_i) - \sum Y_i\ln\sum Y_i)\right).
\end{aligned}$$

Теорема 2.5. *Претпоставимо да оцењен модел садржи предикторе* $X_1, ..., X_k, k \in \{1, 2..., n\}$, *док потпуни модел садржи предикторе* $X_1, ..., X_n$, *тада тест статистика*

$$G = -2\ln\frac{\left(\frac{\sum Y_i}{n}\right)^{\sum Y_i}\left(\frac{n-\sum Y_i}{n}\right)^{n-\sum Y_i}}{\prod_{i=1}^{n}\hat{Y}_i^{Y_i}(1-\hat{Y}_i)^{1-Y_i}}$$

има χ^2_{n-k} *расподелу.*

Доказ се може видети у [14].

[4]Засићен, потпун или комплетан модел *(saturated model)* је онај модел који садржи онолико много параметара колико има регистрованих вредности, тј. n. Најједноставнији пример потпуног модела је модел просте линеарне регресије који има само две тачке ($n = 2$).

Валдов тест

Један од начина тестирања значајности параметара је Валдов тест за једнодимензиону логистичку регресију, који тестира следеће хипотезе:

$$H_0 : \beta_1 = 0,$$
$$H_1 : \beta_1 \neq 0.$$

Теорема 2.6. *Валдова тест-статистика z^* која при важењу H_0 има нормалну расподелу $\mathcal{N}(0,1)$ је једнака:*

$$z^* = \frac{\hat{\beta}_1}{\hat{\sigma}(\hat{\beta}_1)},$$

где $\hat{\sigma}(\hat{\beta}_1)$ представља оцену стандардне девијације параметра β_1.

Уколико са $\mathbf{H}$ обележимо матрицу других извода функције $lnL(\beta_0, \beta_1)$ дефинисану на следећи начин:

$$\mathbf{H} = [h_{ij}]_{2\times 2}, i, j \in 0, 1,$$
$$h_{ij} = \frac{\partial^2 \ln L(\beta_0, \beta_1)}{\partial\beta_i \partial\beta_j}, i, j \in 0, 1.$$

тада се оцена за дисперзију оцене $\hat{\beta}_1$ може добити преко детерминанте претходне једнакости:

$$\hat{\sigma}^2(\hat{\beta}_1) = (\mid -h_{ij} \mid_{\beta_1=\hat{\beta_1}})^{-1},$$

а одатле се наравно добија оцена за стандардну девијацију.

Валдов тест може да буде и једнострани и двострани, у зависности од природе конкретног проблема. На пример, у двостраној варијанти Валдовог теста, доношење одлуке о одбацивању или прихватању хипотезе H_0 за задатим прагом значајности α се врши на следећи начин:

$$\mid z^* \mid \leq z(1 - \frac{\alpha}{2}) \Rightarrow \text{прихватамо} \quad H_0,$$
$$\mid z^* \mid > z(1 - \frac{\alpha}{2}) \Rightarrow \text{одбацујемо} \quad H_0,$$

где је z инверзна функција нормалне расподеле $N(0,1)$.

Када је у питању вишеструка логистичка регресија тестира се хипотеза да су p коефицијената уз предикторе једнаки нули, при чему

важи $p < n$ где је n величина узорка. Претпоставимо да је дат случајан вектор $X = \{X_1, X_2, ..., X_n\}$ са n променљивих, тест статистика је:

$$W = \hat{\boldsymbol{\beta}}^T (Var(\hat{\boldsymbol{\beta}}))^{-1} \hat{\boldsymbol{\beta}},$$

где је $\hat{\boldsymbol{\beta}}$ оцењен вектор која се састоји од оцењених коефицијената $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_n$, $\hat{\boldsymbol{\beta}}^T$ је транспонован вектор вектора $\hat{\boldsymbol{\beta}}$ и $Var(\hat{\boldsymbol{\beta}})$ матрица оцена стандардне девијације.

Теорема 2.7. *Тест статистика*

$$W = \hat{\boldsymbol{\beta}}^T (Var(\hat{\boldsymbol{\beta}}))^{-1} \hat{\boldsymbol{\beta}}$$

под претпоставком да су сви p коефицијената уз предикторе једнаки 0, има χ^2_{p-1} расподелу.

Доказ се може наћи у [13].

Бонферонијева корекција

Вишеструка поређења се јављају када статистичка анализа укључује вишеструке статистичке тестове. Што је више различитих модела, то је већи број закључака донетих на основу резултата добијених из модела, па је самим тим и вероватноћа да се донесе погрешан закључак већа. Развијено је неколико статистичких техника, које директно упоређују нивое значајности за појединачна и вишеструка поређења. Ове технике захтевају и строжији праг значајности, како би се донели исправни закључци на основу добијених резултата. У наставку ћемо описати један од начина корекције прага значајности Бонферонијеву корекцију.

Бонферонијева корекција надокнађује могућности грешке, тако што се свака хипотеза тестира са нивооом значајности $\frac{\alpha'}{m}$, где α' представља грешку друге врсте, а m је укупан број тестираних хипотеза.

Дефиниција 2.3.2. *Нека је дат скуп хипотеза $H_1, H_2, .., H_m$ и нека су $p_1, p_2, ..., p_m$ одговарајуће p вредности. Нека је m укупан број нултих хипотеза и m_0 број тачних нултих хипотеза. Укупан број погрешних одлука (Familiwise error rate) је вероватноћа одбацивања најмање једне тачне нулте хипотезе H_i, односно грешка друге врсте α'. Бонферонијева корекција одбацује нулту хипотезу за свако $p_i \leq \frac{\alpha'}{m}, i \in \{1, 2..., m\}$, где је укупан број погрешних одлука $\leq \alpha'$.*

Ако су сви тестови извођени са нивоом значајности α тада је свеукупна вероватноћа прављења бар једног нетачног одбацивања, односно вероватноћа грешке друге врсте већа од α и њена вредност је обично непозната. Свакако може бити показано да када год се изводи скуп m тестова, сваки са нивоом значајности α, тада је α' највише $1-(1-\alpha)m$. Како се m_0 повећава вероватноћа грешке може постати неприхватљиво велика. Да би смо надоместили тај проблем, најбоље је спроводити само оне тестове који су од стварног интереса. За одговорно спровођење теста, одабраћемо неку разумно малу горњу границу b за вероватноћу прављења најмање једног нетачног одбацивања. Тада спроводимо сваки тест са нивоом значајности $\frac{m_0}{m}$, где m означава број изведених тестова.

Пример 2. *Ако желимо да α' буде највише* 0.1 *и изводимо све могуће тестове за $k=5$ група, спровешћемо свако наше двоструко поређење са нивоом значајности $\alpha = 0.1/10 = 0.01$.*

2.3.3 Интервали поверења за параметре

Из решења једначина (2.11) и (2.12) директно следе једначине за израчунавање $100(1-\alpha)\%$ интервала поверења за параметре $\beta_0, \beta_1, ..., \beta_n$:

$$
\begin{aligned}
I_{\beta_0} &= (\hat{\beta}_0 - z_{1-\frac{\alpha}{2}}\hat{\sigma}^2(\hat{\beta}_0), \hat{\beta}_0 + z_{1-\frac{\alpha}{2}}\hat{\sigma}^2(\hat{\beta}_0)),\\
I_{\beta_1} &= (\hat{\beta}_1 - z_{1-\frac{\alpha}{2}}\hat{\sigma}^2(\hat{\beta}_1), \hat{\beta}_1 + z_{1-\frac{\alpha}{2}}\hat{\sigma}^2(\hat{\beta}_1)),\\
&\dots\\
I_{\beta_n} &= (\hat{\beta}_n - z_{1-\frac{\alpha}{2}}\hat{\sigma}^2(\hat{\beta}_n), \hat{\beta}_n + z_{1-\frac{\alpha}{2}}\hat{\sigma}^2(\hat{\beta}_n)),
\end{aligned}
$$

где је $\hat{\sigma}^2(\hat{\beta}_i), i \in \{0,1,...,n\}$ оцена стандардне грешке одговарајућег параметра, а $z_{1-\frac{\alpha}{2}}$ таблична вредност стандардне нормалне расподеле за коју важи

$$P(|\hat{\beta}_i| \leq z_{1-\frac{\alpha}{2}}) = 1-\alpha, i \in \{0,1,...,n\}.$$

2.4 Предвиђање

Када су параметри модела оцењени, оцена вредности $P(X)$ се једноставно добија њиховим убацивањем у формулу (2.6):

$$\hat{P}(X) = \frac{e^{\hat{\beta}_0 + \sum_{k=1}^{n}\hat{\beta}_k X_k}}{1 + e^{\hat{\beta}_0 + \sum_{k=1}^{n}\hat{\beta}_k X_k}}.$$

Класификација променљиве Y се затим врши на основу $\hat{P}(X)$:

$$\hat{Y} = \begin{cases} 0, & \hat{P}(X) < q; \\ 1, & \hat{P}(X) > q. \end{cases}$$

где је $q \in (0, 1)$ унапред одређена константа. Скуп тачака

$$D = \{x \mid \hat{p}(X) = q\}$$

се назива границом одлуке класификатора. Тачке из D могу се произвољно прикључити и једној и другој класи. У пракси се оне најчешће стављају у "позитивну" класу $Y = 1$, односно класификација се врши по следећем правилу:

$$\hat{Y} = \begin{cases} 0, & \hat{P}(X) < q; \\ 1, & \hat{P}(X) \geq q. \end{cases}$$

Стандардна вредност за q је $\frac{1}{2}$, али постоје и случајеви у којима се узимају друге вредности. То се обично ради у ситуацијама када грешка приликом класификације не носи исте последице за сваку погрешну класу или када су класе значајно различитих величина.

2.5 Процена слагања модела са подацима

Када смо креирали модел, односно, подразумевамо да модел садржи оне променљиве које су значајне, занима нас колико ефикасно наш модел описује резултујућу (зависну) променљиву, како одредити њихове перформансе и како међусобно поредити моделе.

Постоје разне методе за израчунавање перформасни, у наставку ћемо описати неке од њих, као што су матрица класификације *(Confusion matrix)*, *ROC* крива и Џинијев *(Gini)* коефицијент.

2.5.1 Идентификација аутлајера међу подацима

Аутлајер *(outlier)* можемо схватити као вредност из узорка која знатно одступа од осталих вредности. Иако се аутлајери често сматрају

грешком, они могу носити важну информацију. Неоткривени аутлајери могу водити ка погрешној спецификацији модела, пристрасној оцени параметара и нетачним резултатима. Стога је важно тачно идентификовати аутлајере пре моделовања и анализе.

Постоји велики број метода идентификације и откривања аутлајера. Ми ћемо у наставку описати графички приказ аутлајера (помоћу правоугаоних дијаграма) и тест Куково растојање.

Један од најчешће коришћених начина утврђивања да ли је нешто аутлајер или не је Куково растојање *(Cook's distance)*. Подаци који имају велико Куково растојање сматрају се значајним за даљу анализу.

Теорема 2.8. *Куково растојање је тест који се користи за утврђивање и идентификацију података који знатно одступају од осталих, и израчунава се:*

$$CD_i = \frac{(\hat{\boldsymbol{\beta}}^{-i} - \hat{\boldsymbol{\beta}})^T (X^T V X)(\hat{\boldsymbol{\beta}}^{-i} - \hat{\boldsymbol{\beta}})}{(p+1)\hat{\sigma}^2}, i \in \{1, 2, ..., n\},$$

где је $\hat{\boldsymbol{\beta}}^{-i}$ *оцена вектора* $\boldsymbol{\beta}$ *без* i*-те опсервације, а* $\hat{\sigma}$ *је позитиван корен средњеквадратне грешке, односно важи:*

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

Уколико је Куково растојање

$$CD_i > \frac{4}{n - (p+1)},$$

тада се i-та опсервација може сматрати аутлајером.

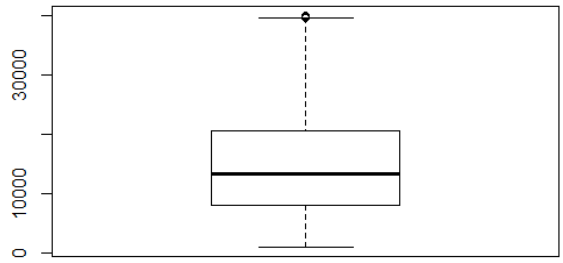
Правоугаони дијаграм *(box-plot)* представља начин графичког приказивања података погодан за прелиминарна поређења са симетричним расподелама, посебно нормалном. Овај дијаграм се може добити тако што се на изабраној оси одреде тачке које одговарају узорачкој медијани и квартилима q_1 и q_3, затим се рачунају унутрашње границе

f_1 и f_3 и спољашње границе F_1 и F_3 на следећи начин:

$$f_1 = q_1 - 1.5(q_3 - q_1),$$
$$f_3 = q_1 + 1.5(q_3 - q_1),$$
$$F_1 = q_1 - 3(q_3 - q_1),$$
$$F_3 = q_1 + 3(q_3 - q_1).$$

И онда се одређује a_1-најмањи међу елементима узорка који је већи од f_1 и a_3-највећи међу елементима узорка који је мањи од f_3.

Дијаграм се састоји од правоугаоника чија је једна страна паралелна изабраној оси и једнака одсечку (q_1, q_3). Димензија друге стране се бира произвољно. У правоугаоник се уцртава права линија која одговара узорачкој медијани me. Ако је та линија близу средине правоугаоника, расподела обележја на узорку би могла бити нека симетрична расподела, у супротном је у питању асиметрична расподела (на слици 2.7. је приказан пример једног дијаграма).



Слика 2.6: Правоугаони дијаграм

2.5.2 Класификација

Нека су дати случајни вектори $X = \{X_1, X_2, ..., X_N\}$, при чему случајне променљиве $X_1, X_2, ..., X_N$ могу бити квантитативне (дискретне или непрекидне) или квалитативне, и случајна променљива Y која је квалитативног типа, узима вредности из скупа $G = \{G_1, G_2, ..., G_M\}$ и зависна

је од случајног вектора X. Нека је познато n реализација ових случајних величина:

$$\{x_{11}, x_{21}, ..., x_{N1}, y_1\}, \{x_{12}, x_{22}, ..., x_{N2}, y_2\}, ..., \{x_{1n}, x_{2n}, ..., x_{Nn}, y_n\}.$$

Овакву класу проблема који се баве предвиђањем вредности зависне променљиве Y на основу вредности предиктора X и неког скупа података на коме су вредности и за X и за Y познате у статистици називамо проблемима класификације. У складу са тим, модел за предвиђање вредности применљиве Y се назива класификатором, а чланови скупа G класама.

Класификациони модели имају велику примену у науци и индустрији. Они играју кључну улогу у системима у којима је потребно брзо, поуздано и аутоматски донети неку одлуку или сортирати неке објекте. Неки од примера употребе класификације у пракси су:

- Давање дијагнозе за неког пацијента на основу уочених симптома.
- Детекција нежељених порука у електронској пошти.
- Препознавање рукописа са дигиталне фотографије.
- Одлучивање да ли неком кориснику банке треба одобрити финансијски кредит.

2.5.3 Мера успеха класификатора

Сврха сваког класификатора је да са што већим успехом предвиди вредности променљиве Y, односно сврста Y у одговарајућу класу. Постоји више различитих мера за поређење успеха класификатора.

Све ове мере се рачунају на основу тзв. "матрице класификатора" *(confusion matrix)* која у колонама садржи стварне вредности променљиве Y (стварне класе), а у редовима вредности променљиве Y оцењене од стране класификатора. Оваква матрица се назива и "матрица забуне", и представља добар начин да се прикаже који део случајева је исправно, а који погрешно класификован. Матрица класификације се најчешће рачуна за случајеве када постоје само две класе, на пример "позитивну" класу $Y = 0$ и негативну класу $Y = 1$.

Дефиниција 2.5.1. *Матрица класификације (confusion matrix) представља матрицу погрешних и тачних класификација података. Са матрице можемо прочитати следеће вредности:*

- *Број случајева који су исправно класификовани као позитивни (true positive). Ознака за ову вредност је TP.*
- *Број случајева који су погрешно класификовани као позитивни (false positive). Ознака за ову вредност је FP.*
- *Број случајева који су исправно класификовани као негативни (true negative). Ознака за ову вредност је TN.*
- *Број случајева који су погрешно класификовани као негативни (false negative). Ознака за ову вредност је FN.*

Сама матрица ће изгедати:

		Стварне класе	
		0	1
Оцењене класе	0	FN	TN
	1	FP	TP

Табела 2.1: Матрица класификације

На основу вредности из матрице могу да се израчунају различите мере успеха класификатора. Споменућемо неке од основних, које ће касније бити коришћене у практичном делу.

Дефиниција 2.5.2. *Мере успеха класификатора су:*

- *Тачност (accuracy) је основна мера која се увек израчунава. Она представља удео успешно класификованих примера. Ознака за ову вредност је ACC. Израчунава се по следећој формули:*

$$ACC = \frac{TP + TN}{TP + FP + TN + FN};$$

- *Прецизност (precision - positive predictive value) је удео успешних класификација међу примерима који су класификовани као позитивни. Ознака за ову вредност је PPV. Израчунава се по следећој формули:*

$$PPV = \frac{TP}{TP + FP};$$

- *Сензитивност (sensitivity - true positive rate) је удео успешних класификација међу примерима који су позитивни у стварности. Ознака за ову вредност је TPR. Израчунава се по следећој формули:*
$$TPR = \frac{TP}{TP + FN};$$

- *Специфичност (specificity - true negative rate) је удео неуспешних класификација међу примерима који су негативни у стварности. Ознака за ову вредност је TNR. Израчунава се по следећој формули:*
$$TNR = \frac{TN}{TN + FP};$$

- *$F1$ метрика је хармонијска средина између прецизности и сензитивности. Израчунава се по следећој формули:*
$$F1 = 2 \cdot \frac{PPV \cdot TPR}{PPV + TPR} = 2 \cdot \frac{TP}{2TP + FP + FN}.$$

Избор одређене мере као метрике успеха се врши на основу конкретних особина задатка који класификатор треба да врши у пракси. У неким случајевима је битно само што више инстанци класификовати на прави начин, у другима је важно избећи погрешно процењивање неке од класа јер оно носи велике последице итд.

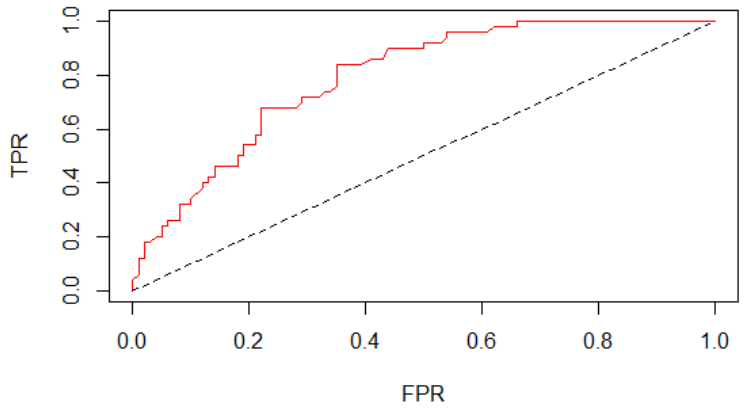
Један од начина за приказ резултата оцењеног логистичког регресионог модела је помоћу претходно дефинисане матрице класификације. Да бисмо креирали ову матрицу предвиђених вредности из нашег модела, за оцењен параметар насупрот тачној вредности, морамо прво дефинисати ниво одлучивања *(cut-off value)* са којим ћемо поредити сваку оцењену вредност. Најчешће вредност за овај параметар је 0.5.

2.5.4 *ROC* крива

ROC крива (Receiver Operating Characteristic Curve) је графичка техника која је више од 30 година веома популарна посебно у лабараторијској медицини. Примена ове технике ја започела током Другог светског рата за евалуацију лажно позитивних и стварно позитивних сигнала на екрану радара. Касније је адаптирана од стране радиолога и лабараторијских научника за евалуацију осетљивости и специфичности

медицинских одређивања при различитим нивоима одлучивања.

Када се сензитивност и специфичност теста израчунају за читав низ нивоа вероватноће, нивоа одлучивања, могуће је конструисати *ROC* криву која повезује сензитивност (вероватноћу тачног детектовања присуства особине) и специфичност, (вероватноћу нетачног детектовања присуства особине). Свака тачка *ROC* криве представља уређени пар (сензитивност, 1- специфичност) који одговара појединачном нивоу одлучивања.



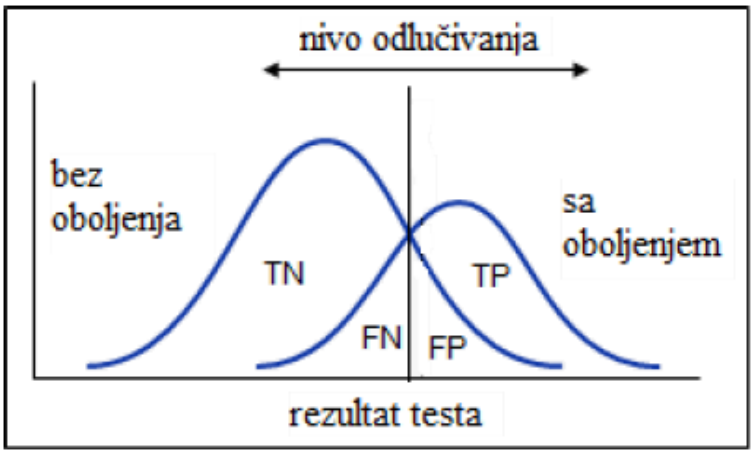
Слика 2.7: *ROC* крива[5]

ROC крива које се одликује комплетним раздвајањем (нема преклапања расподеле резултата две групе) пролази кроз горњи леви угао где стварно позитивни удео износи 1,0 односно осетљивост 100%, а лажно позитивни удео 0, односно 1-специфичност. Крива за тест код кога нема раздвајања (идентична расподела резултата две групе) је дијагонална линија од доњег левог угла до горњег десног угла. Већина *ROC* кривих се налази између ове две крајности и квалитативно гледано она која је ближа горњем левом углу указује на тест са већом тачношћу. Уколико је више *ROC* кривих приказано на једном дијаграму она која се налази изнад и на лево у односу на *ROC* криву са којом се пореди указује на тест са већом посматраном тачношћу. Релативни положај две или више

[5]На графику је црвеном бојом приказана *ROC* крива, на x-оси је приказана мера 1-специфичност FPR, а на y-оси је приказана мера сензитивност TPR, графикон је креиран у програмском језику *(R)*, код се може видету у прилогу Б.1.

кривих омогућава квалитативно поређење више тестова.

Пример 3. *Посматрајмо резултате одређеног теста у две популације. Нека је једна популација са обољењем, и другу без обољења, ретко ћемо добити перфектно раздвајање између ове две групе. Уместо тога расподела резултата теста ће се преклапати, као што је приказано на Слици 2.9, а детаљнија анализа се може видети у [15].*



Слика 2.8: Расподела популација са и без обољења[6]

За сваку могућу критичну вредност коју смо изабрали да раздваја две популације, постојаће неки случајеви са обољењем који су коректно класификовани као позитивни (TP = true positive fraction), али ће неки случајеви са обољењем бити класификовани као негативни, то јест лажно негативни (FN = false negative fraction). Са друге стране, неки случајеви без обољења ће бити коректно класификовани као негативни (TN = true negative fraction), док ће неки случајеви без обољења бити класификовани као позитивни, тј. лажно позитивни (FP = false positive fraction), што је приказано у Табели матрица класификације, табела 2.1.

Површина испод *ROC* криве, која се креће од нуле до један, је мера способности модела у раздвајању субјеката који су искусили догађај који се посматра у односу на оне који нису. Површина испод *ROC* криве, у ознаци *AUC* (The Area Under the Curve), је прихваћена традиционална

[6]Детаљнији опис примера се може видети у [15]

изведена мера за ROC криву.

Као опште правило, користимо следеће:

- $AUC = 0.5$ - нема раздвајања,
- $0.5 \leq AUC < 0.7$ - лоше раздвајање,
- $0.7 \leq AUC < 0.8$ - прихватљиво раздвајање,
- $0.8 \leq AUC < 0.9$ - одлично раздвајање,
- $AUC \geq 0.9$ - изванредно раздвајање.

Још једна од мера која је користи за процену слагања моделе са подацима је Ђинијев *(Gini)* коефицијент. Када смо одредили ROC криву (слика 2.9), и означили простор између дијалонале и ROC криве са DR, а простор између ROC криве $x-$осе са AUC, тада можемо одредити Ђинијев коефицијент, и он је једнак:

$$Gini = \frac{DR}{2} = 2(AUC - 0.5) = 2AUC - 1.$$

Вредност овог коефицијента је између 0 и 1. У случају када су резултати добијеног модела добри, тада је вредност $Gini$ коефицијента прилижно једнака 1, а када су лоши резултати тада је приближно једнака 0.

Поглавље 3

Стабло одлучивања

Стандардни модели подразумевају параметарске моделе, међутим, за моделе кредитног ризика могуће је користити и непараметарске моделе, као што је стабло одлучивања. У наставку описаћемо конструкцију стабла одлука, и касније на конкретном примеру креирати модел стабла одлучивања и искорисити за поређење са логистичко регресионим моделом.

Овај метод има широку примену у развоју модела кредитног ризика. У пракси је познат и под називом дрво расподеле или стабло класификације, и представља модел који се састоје из скупа "ако-онда" услова дељења (класификације) на две или више различитих група. Процес израде стабла је: избор једне променљиве која "најбоље" раздваја податке у две (или више) подгрупе, након тога процес поделе се понавља рекурзивно све док чвор не достигне минималну величину (изабрани минални број података за последњи чвор) или се резултати модела не могу побољшати. У зависности од изабраног алгоритма, циљ је категоризација нових променљивих и рекатегоризација већ постојећих категоричких променљивих. Једна од кључних разлика у односу на параметарске моделе је та да су све променљиве третиране као категоричке променљиве.

Кораци алгоритма за конструкцију стабла су:

1. Израда стабла одлучивања рекурзивним бинарним дељењем скупа података, заустављајући се тек кад сваки крајњи чвор има мање од задатог броја података;

2. Скраћивање (подрезивање стабла) како бисмо добили низ најбољих подстабала;

3. Одабир подстабла које производи најмању грешку, односно избор оптималног параметра комплексности.

Стабло се састоји од чворова и грана. Гране повезују родитељске чворове *(parent node)* са дечијим чворовима *(child node)*. Код стабла постоје три врсте чворова:

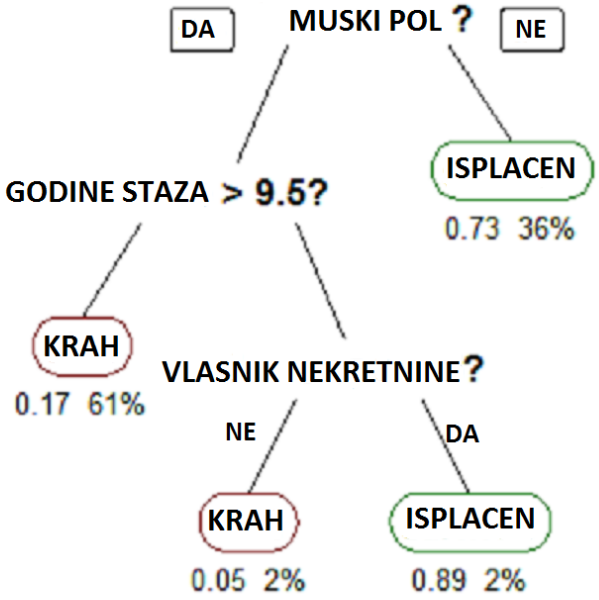
- почетни чвор (назива се и корени чвор и представља почетни чвор у стаблу, којем не претходи ниједан чвор),
- крајњи чвор (њиме се завршава одређена грана стабла, и представљају сва могућа решења задатог проблема),
- чвор одлуке (дефинише одређени критеријум у облику вредности атрибута из којег излазе гране које задовољавају одређене вредности тог атрибута).

Основне предности методе стабла одлучивања су: могућност генерисања разумљивих модела, јасна важност појединих атрибута за конкретни проблем и широка доступност софтверских решења.

Недостатак стабала одлучивања је њихова нестабилност јер мала одступања у узорку података могу имати велике варијације у додељеним класификацијама, прорачуни могу постати јако комплексни, нарочито ако су многе вредности непоуздане и/или ако је много позитивних исхода.

Пример 4. *На примеру модела кредитног ризика, претпоставка примене овог стабла одлука је поседовање базе података апликаната за кредит, који су описани са n атрибута $x_1, x_2, ..., x_n$. Подносиоци захтева подељени су у два подскупа, и означени су: добри (неризични) и лоши (ризични) клијенти. Циљ модела кредитног ризика јесте проналазак класификатора (атрибута) који најбоље раздваја узорак добрих клијената од узорка лоших клијената. Алгоритам почиње чвором који садржи узорке добрих и лоших клијената, након чега се проналазе сви могући исходи с циљем добијања најкориснијег атрибута x и одговарајуће граничне вредности c која најбоље врши раздвајање узорака добрих и лоших клијената. Подаци се деле према свим могућим критеријима у две гране. При томе се изабере критеријум који податке дели у подскупове које су хомогенији од почетног скупа података.*

Процедура се наставља док се не достигне минимална величина чвора (један од примера стабла је приказан на слици 3.1.).



Слика 3.1: Пример стабла одлучивања[1]

Стабла одлучивања је модел која се користи у случајевима класификацијских и предикцијских проблема. Постоји велики број алгоритама[2] који се користе за креирање стабла одлучивања. У наставку ћемо описати један од алгоритама класификације (Classification And Regression Tree - CART).

3.1 Израда стабла одлучивања

CART (Classification And Regression Tree) алгоритам, који је представио Бреиман (Leo Breiman), може се користити за израду модела

[1]Пример стабла одлука је креиран у програмском језику R на основу података примера који је описан у четвртом поглављу.

[2]Неки од алгоритама су: ID3, CART, C4.5, CHAID, MARS.

стабла одлучивања на основу класификације или регресионе анализе [21]. У наставаку преставићемо модел који се заснива на класификацији.

Претпоставимо да имамо n података (записа) и укупно C класа. Модел стабла одлучивања поделиће ове податке у k коначних група, где је свакој од ових група додељена оцењена класа. Можемо дефинисати:

- $\pi_i, i \in 1, 2..., C$ је вероватноћа класе $C_i, i \in 1, 2..., C$;
- $L(i, j), i, j \in 1, 2..., C$ је матрица губитака за погрешно класификовање класе i као j, важи $L(i, i) = 0$;
- A је чвор у стаблу;
- $\tau(x)$ је стварна класа записа x, где је x вредност вектора $X = (X_1, ..., X_m)$ који представља вектор предиктора $X_1, ..., X_m$;
- $\tau(A)$ је класа додељена чвору A, где је A крајњи чвор у стаблу;
- n_i, n_{iA} број података који припадају класи $i, i \in 1, 2..., C$, односно број података који припадају класи i у чвору A.

Дефиниција 3.1.1. *Вероватноћа у чвору A и вероватноћа класе i при услову да припада чвору A, редом, су једнаке:*

$$P(A) = \sum_{i=1}^{C} \pi_i P\{x \in A, |\tau(x) = i\},$$

$$P(i|A) = P\{\tau(x) = i | x \in A\} = \pi_i \frac{P\{x \in A | \tau(x) = i\}}{P\{x \in A\}},$$

где је π_i вероватноћа класе i, односно $\pi_i = P\{\tau(x) = i\}$.

Већина параметара се оцењује на основу података, па су оцене претходних вероватноћа једнаке:

$$\hat{P}(A) = \sum_{i=1}^{C} \pi_i \frac{n_{iA}}{n_i}, \tag{3.1}$$

$$\hat{P}(i|A) = \pi_i \frac{\frac{n_{iA}}{n_i}}{\sum_{i=1}^{C} \pi_i \frac{n_{iA}}{n_i}}. \tag{3.2}$$

Дефиниција 3.1.2. *Сложеност креирања чвора A и сложеност креирања модела (стабла) T су једнаке:*

$$R(A) = \sum_{i=1}^{C} P(i|A)L(i, \tau(A)),$$

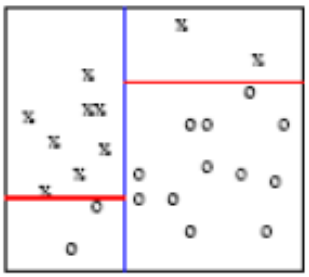
$$R(T) = \sum_{j=1}^{k} P(A_j)R(A_j),$$

где је $\tau(A)$ изабран тако да сложеност доведе на најмањи могућ ниво, а A_j су одговарајући чворови у стаблу.

Можемо прећи на конструкцију стабла, односно избор одговарајућих атрибута (класе) у сваком чвору за креирање стабла.

3.1.1 Избор атрибута за креирање стабла

Главни циљ исправне селекције атрибута је изабрати подскуп улазних атрибута како би се елиминисали атрибути који нису релевантни и који не дају предиктивну информацију, са циљем постизања високе тачности класификације (Ramaswami i Bhaskaran, 2009.). Ако желимо да поделимо скуп D у мање делове, идеално би било да сваки подскуп буде чист (све инстанце у једној партицији да припадају истој класи, као на слици 3.1).



Слика 3.2: Подела података у групе са истим класама[3]

Када је у питању класификациони алгоритам (CART) стабло је увек бинарно и сваки чвор има тачно две гране. Овај алгоритам рекурзивно

[3]На слици је приказан пример поделе скупа података на хомонеге податке, детаљније се може видети у [22].

дели почетни скуп у подскупове са истим вредностима циљног атрибута (исте класе).

Дефиниција 3.1.3. *Мера која се корсити за утврђивање "повољности" грањања је:*

$$\Phi(A|i) = 2P(A_L)P(A_R)\sum_{j=1}^{C}|P(j|A_L) - P(j|A_R)|, \tag{3.3}$$

где је $\Phi(A|i)$ *мера повољности гранања за кандидат гранања класу* i *чвора* A*.*

A_L и A_R су леви односно десни потомак (дечији чвор) чвора A, и C представља број класа, а вероватноће $P(A_L), P(A_R), P(j|A_L), P(j|A_L)$ су приказане у једнакостима (3.1) и (3.2). Оптимално гранање је оно са максималном вредности $\Phi(A|i)$ за сва могућа гранања за чвор A, односно одређујемо максималну вредност функције $\max_i \Phi(A|i)$ за свако $i \in \{1, 2..., C\}$.

На основу претходног можемо закључити да $\Phi(A|i)$ расте када обе компоненте производа расту $2P(A_L)P(A_R)$ и $\sum_{j=1}^{C}|P(j|A_L) - P(j|A_R)|$. Такође, компонента $2P(A_L)P(A_R)$ има максималну вредност када су оба потомка исте величине (имају исту заступљеност), и тада она износи $0.5 \cdot 0.5 = 0.25$. Ако означимо са $Q(i|A)$ другу компоненту, њена максимална вредност је када су све инстанце чвора потомка потпуно униформне (чисте).

Поступак максимизације $\Phi(A|i)$ се понавља за сваки наредни чвор, док се не креирају сва могућа грањања или се не достигне задати минимални број података у крајњем чвору. Када се исцрпе сва могућа гранања генерисано је пуно стабло.

Један од индекса који служи за мерење квалитета поделе је Џинијев коефицијент. Користи се за тестирање сваког појединог раздвајања и мерења хомогености података. Џинијев коефицијент често се описује као мера "чистоће" чвора. Чистоћа чвора представља оне чворове у којима је велики проценат инстанци који припадају истој класи. Мала вредност овог коефицијента указује на "чисте" чворове. На пример нека је S атрибут дискретне вредности који има n различитих вредности $\{s_1, s_2, ..., s_n\}$ у скупу D. Како би се опредили за најбољи бинарни прелом атрибута S, испитујемо све могуће подскупове које се могу

формирати користећи вредност атрибута S.

За проверу слагања модела са подацима користе се и методе које су описане у поглављу 2.

3.2 Скраћивање стабла одлука

Када смо креирали потупно стабло (стабло са свим могућим поделама), постоји могућност да је креирано стабло исувише комплексно. Често нису сви крајњи чворови стабла хомогени, што оставља одређени степен грешке приликом класификације. Метода скраћивања стабла *(pruning)* решава проблем преклапања (нехомогености) података.

Подрезана верзија је мање комплексна, и лакша је за разумевање. Обично је бржа и боља при класификацији података који се користе за тестирање, него неподрезана стабла. Постоје два приступа за скраћивање стабала: подрезивање и надрезивање[4] стабла.

При приступу подрезивања стабала, стабло се скраћује на начин да се донесе одлука да ли ће се дељење наставити на одређеном чвору или не. Након заустављања, чвор постаје лист (крајњи чвор). Код израде стабла, мере као што су статистичка значајност, ђинијев коефицијент и друге могу се користити за процену квалитета поделе.

Нека су дати $A_1, A_2, ..., A_k$ крајњи чворови креираног стабла T. Дефинишемо:

- $|T|$ укупан број крајњих чворова у стаблу T,
- $R(T) = \sum_{i=1}^{k} P(A_i)R(A_i)$ је сложеност креирања стабла T.

Нека је α неки број између 0 и ∞ који представља сложеност додавања новог предиктора моделу, и назива се параметар комплексности *(complexity parameter)*.

Дефиниција 3.2.1. *Сложеност креирања стабла T_α (cost-complexity pruning measure) је мера*

$$R(T_\alpha) = R(T) + \alpha|T|,$$

[4]Надрезивање стабла је процес надоградње стабла, додавање нових чворова, овај метод нећемо разматрати у раду, детаљније се може видет и [17].

где је T_α подстабло потпуног модела који има најмању вредност сложености.

Можемо приметити да је T_0 потпун модел (са свим предикторима), док је T_∞ модел без иједног грањања.

Тврђење 3.2.1. *Важе следећа тврђења:*

- *Ако су T_1 и T_2 подстабла стабла T, при чему важи $R_\alpha(T_1) = R_\alpha(T_2)$, тада важи да је стабло T_1 подстабло стабла T_2 или је стабло T_2 подстабло стабла T_1, па је $|T_1| < |T_2|$ или $|T_2| < |T_1|$;*
- *Ако је $\alpha > \beta$ тада је $T_\alpha = T_\beta$ или је T_α подстабло стабла T_β;*
- *За дате вредности параметара комплексности $\alpha_1, \alpha_2, ..., \alpha_m$, $T_{\alpha_1}, T_{\alpha_2}, ..., T_{\alpha_m}$ и $R(T_{\alpha_1}), R(T_{\alpha_m}), ..., R(T_{\alpha_m})$ се могу ефикасно израчунати.*

Доказ претходног тврђења се може видети у [21].

На основу претходног, можемо дефинисати скраћено стабло T_α као најмање стабло T за које важи да је $R(T_\alpha)$ минимално, односно да је сложеност креирања стабла минимална:

$$R(T_\alpha) = \min_{T_\alpha \subseteq T} R(T).$$

3.3 Избор оптималног параметра комплексности

За најбољи избор скраћеног постабла, односно најбољи избор вредности α, користи се крос-валидација *(cross-validation)*.

Групишемо све могуће вредности α у m инетервала, где је $m < |T|$:

$$\begin{aligned} I_1 &= [0, \alpha_1], \\ I_2 &= (\alpha_1, \alpha_2], \\ &\ldots \\ I_m &= (\alpha_{m-1}, \infty]. \end{aligned}$$

Поступак крос-валидације је:

1. Оцењује се комплетан модел тако што се израчунају вредности:

$$\begin{aligned}\beta_1 &= 0,\\ \beta_2 &= \sqrt{\alpha_1\alpha_2},\\ &\dots\\ \beta_{m-1} &= \sqrt{\alpha_{m-2}\alpha_{m-1}},\\ \beta_m &= \infty,\end{aligned}$$

   где је свако β_i вредност за одговарајући интервал I_i.

2. Подаци се поделе у s група $G_1, G_2, ..., G_s$ сваки величине $\frac{s}{n}$, и за сваку групу појединачно се понавља поступак:

   - Оцењује се потупун модел на основу свих вредност осим вредност из групе G_i и креирају се стабла $T_{\beta_1}, T_{\beta_2}, ..., T_{\beta_m}$ за овај редуковани скуп података,
   - Израчуна се оцењена класа за свако G_i на основу сваког модела $T_{\beta_j}, 1 \leq j \leq m$,
   - На основу претодног израчуна се сложеност креирања модела за сваки од креираних модела (стабла).

3. Сумирају се вредности у групи G_i како бисмо добили оцењени ризик за свако β_j. Стабло T_β је најбоље одабрано (скраћено) подстабло за одговарајућу вредност β за коју је израчунат параметар комплексности најмањи.

У неким моделима свака од група G_i садржи само једну опсервацију, док је за креирање модела стабла одлучивања ова подела исувише компликована, уобичајна подела која се користи је $s = 10$ група.

Поглавље 4

Примена модела кредитног ризика

У овом поглављу представићемо претходно описане моделе на примеру кредитног ризика на основу података клијената који су аплицирали за кредит преко сајта *LendingClub*. Ова компанија се бави повезива-њем клијената (или фирми) која аплицирају за кредит (позајмицу) и инвеститора (особа или компанија) који су спремни да инвестирају, са циљем остварења добити повратом инвестиције. Детаљнији опис података се може видети у [19], а програмски код коришћен приликом израде модела се налази у прилогу Б овог рада.

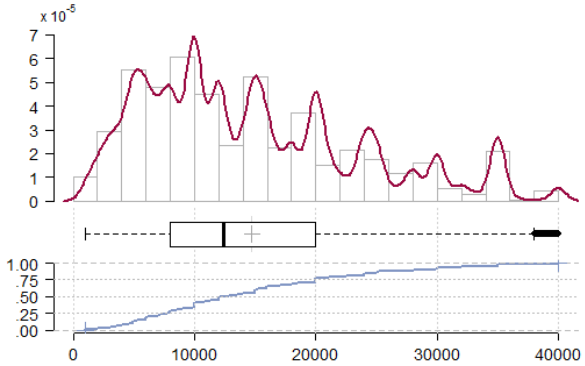
Подаци се састоје од издатих кредита током 2016, који укључују тренутне статусе отплате кредита (тренутни, касни са отплатом, неисплаћен, цео износ отплаћен итд).

Састоји се од 25 променљивих са укупно 434407 записа. У табели у прилогу А се могу видети описи променљивих које ћемо користити у истраживању. Детаљнији опис података се може видети [19].

Сваки од записа садржи детаљне информације о клијенту, као на пример: износ кредита, износ рате, каматна стопа, преостали износ за отплату, године стажа запосленог, поседовање некретнине, држава клијента, ранг клијента, итд.

Износи позајмица (кредита) се углавном између 8000$ и 20000$, док

је мали број клијената аплицирао за кредит који је мањи и већи од ових сума, што се може видети на слици 4.1.

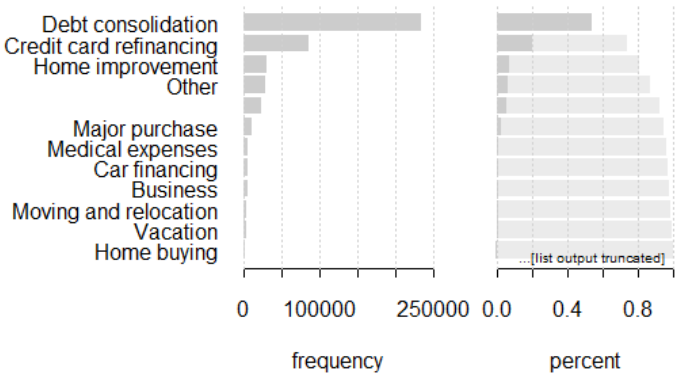


Слика 4.1: Расподела променљиве износ кредита[1]

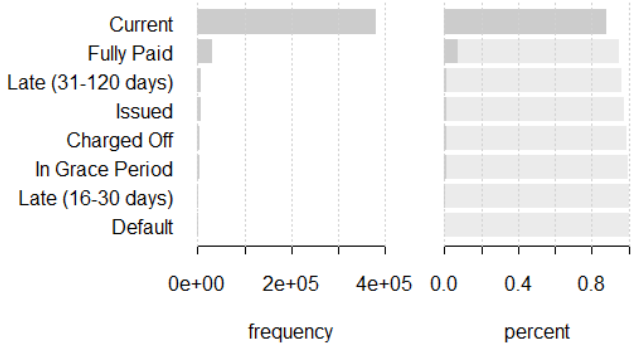
Када је у питању сврха аплицирања за кредит, можемо приметити да је највећи број клијената аплицирао са потребом рефинансирања тренутних кредита, око 57% укупног броја клијената, док је око 21% аплицирао за рефинансирање кредитних картица, ту су још укључени и кредити за здравствене сврхе, куповину аутомобила, кућа, одмор... Графички приказ се може видети на слици 4.2, док се детаљн број по категоријама и проценат може видети у прилогу Б.

Такође, можемо приметити да постоје различити статуси кредита. Највећи број кредита је тренутно активан (у току) 87%, отплаћених је приближно 7%, и аплицираних 2%, осталих 4% чине неотплаћени и кредити са кашњењем, детаљнији подаци се могу видети на слици 4.3.

[1]На графикону су приказани густина расподеле променљиве износ кредита, емпиријска функција расподеле и правоугаони (box-plot) дијаграм, користећи функцију *Desc* у програмском језику *R*, детаљнији опис се може видети у [20].



Слика 4.2: Опис кредита[2]



Слика 4.3: Тренутни статуси кредита[3]

[2]На графикону су приказане фреквенције, проценти и кумулативни проценти промењиве опис кредита по категоријама, користећи функцију *Desc* у програмском језику *R*, детаљнији опис се може видети у [20].

[3]На графикону су приказане фреквенције, проценти и кумулативни проценти промењиве статус кредита, детаљнији опис се може видети у [20].

Клијенти се могу груписати у различите категорије, тј рангирати, на основу претходне историје података о њиховом аплицирању. Рангирање клијената је урађено на основу историјских података о њима[4] (података из претходних година, од почетка 2015. године) са циљем даље анализе критичних и мање критичних клијената, детаљније се може видети [19]. Најкритичније особе (особе које неће бити у могућности да отплате кредит) су означене са G, а најмање критичне са A. Можемо приметити да највећи број клијената припада групи C, B и A око 78%, детаљна расподела ранга клијената на основу статуса кредита се може видети у табели 4.1.

Ранг клијента	Статус кредита		
	0	1	Тотал
A	4876	814	5690
	85.7%	14.3%	12.1%
B	8498	2983	11481
	74%	26%	29.9%
C	9049	5038	14087
	64.2%	35.8%	29.9%
D	4541	3703	8244
	55.1%	44.9%	17.5%
E	2311	2533	4844
	47.7%	52.3%	10.3%
F	899	1249	2148
	41.9%	58.1%	4.6%
G	253	434	687
	36.8%	63.2%	1.5%
Тотал	30427	16754	47181

Табела 4.1: Матрица класификације[5]

На основу информација статуса кредита *(loan_status)* можемо урадити груписање клијената у две групе на основу питања да ли су своје обавезе извршили на време или је дошло до краха. На основу тога кредите који имају статус *Full Paid* ћемо ознчитити са 0, (односно *Survived*),

[4]Опис процеса доделе ранга клијентима се може видети у првом поглављу рада. Пошто је факторска променљива, у програмском језику *R* се разматра као посебна категорија променљивих, код за трансформацију ове променљиве се може видети у прилогу Б.2.

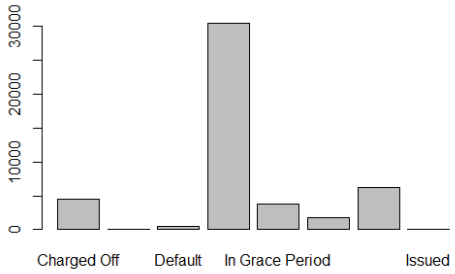
[5]Резултати у табели су приказани на основу променљивих статус кредита и ранг кредита, код за израду табеле се може видети у прилогу Б.2.

и они представљају дужнике коју су отплатили цео износ кредита. У другу групу ћемо оставити кредите са статусима *Charged_off, Defaulted, Late (31-120 days), Late (16-30 days)*, и означићемо их са 1 (односно *Defaulted*), они представљају кредите за који се догодио догађај краха, или касне са отплатом. На основу претходног, можемо да формирамо нову променљиву:

$$Y = \begin{cases} 0, & \text{ако променљива статус кредита узима вредност } \textit{Full Paid}; \\ & \text{ако променљива статус кредита узима вредности:} \\ 1, & \textit{Charged\_off, Defaulted, Late (31-120 days), Late (16-30 days).} \end{cases} \tag{4.1}$$

коју ћемо користити као зависну променљиву приликом израде модела. Кредите који су у току (са статусима *Issued, Current*) нећемо користити у анализи и изради модела.

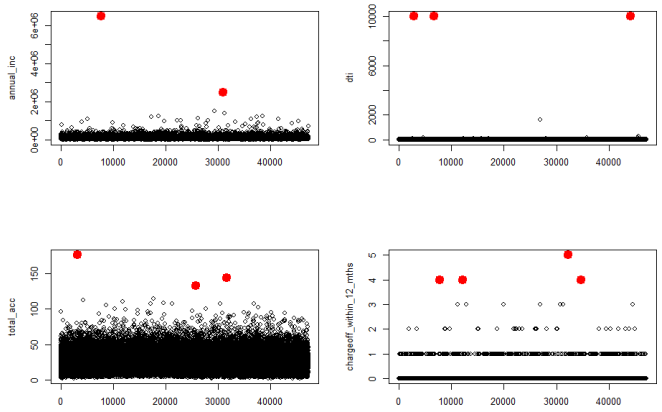
Након издвојених података можемо приметити да имамо укупно 47181 записа, а расподела по статусима се може видети на слици 4.4.



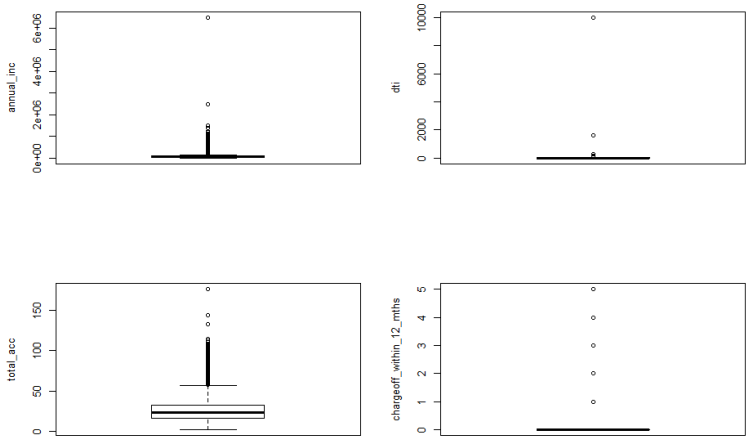
Слика 4.4: Расподела статуса кредита након издвојених података[6]

Када су у питању аутлајери *(outliers)*, можемо приметити да постоје поједини записи који драстично одступају од осталих вредности, ови записи припадају променљивима *annual_inc, dti, total_acc* и *charge-off_within_12_mths* (слика 4.5.). Како бисмо утврдили да ли су наведени аутлајери представљају грешке у подацима, искористићемо тест Кукуво растојање за детекцију аутлајера и графички приказ података помоћу правоугаоних *(box-plot)* дијаграма.

[6]Приказ расподеле променљиве статус кредита, након груписања клијента, издвојених статуса *Full Paid, Charged_off, Defaulted, Late (31-120 days), Late (16-30 days).*



Слика 4.5: Присуство аутлајера међу подацима[7]



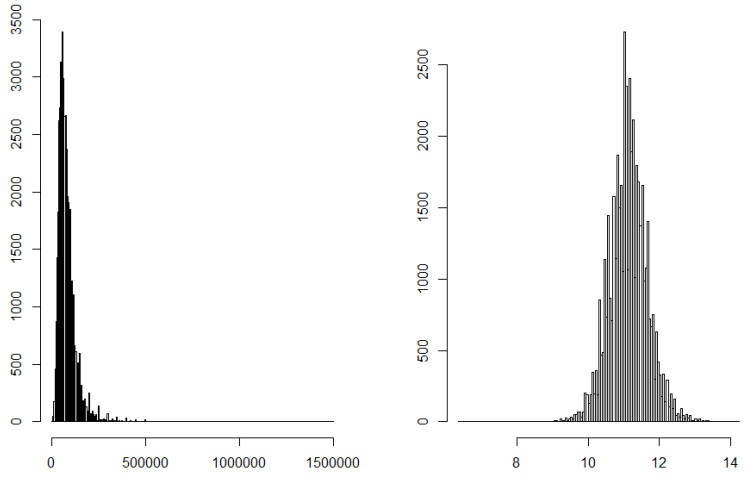
Слика 4.6: Провера аутлајера[8]

[7]Графички приказ аутлајера за промењиве *annual_inc, dti, total_acc* и *charge-off_within_12_mths*, где су на y-оси приказане вредности промењивих, а на x-оси индекс, односно редни број записа, код за израду графикона се може видети у прилогу Б.2.

[8]Графички приказ аутлајера помоћу правоугаоних *(box-plot)* дијаграма

На основу дијаграма са слике 4.6. правоугаоних *(box-plot)* дијаграма можемо закључити да претходно наведени подаци јесу аутлајери. За проверу искористићемо и тест Куково растојање (описан у поглављу 2), који ћемо применити директно на модел логистичке регресије, а након тога донети закључак у аутлајерима.

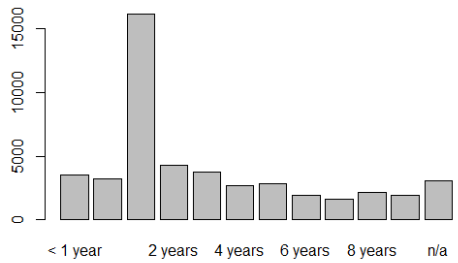
Када је у питању променљива која се односи на годишње приходе клијената можемо приметити да су износи изузетно велики (у хиљадама, може се видети на слици 4.7. - леви графикон), па ћемо уместо стварних вредности користи логаритамске вредности (слика 4.7. - десни графикон).



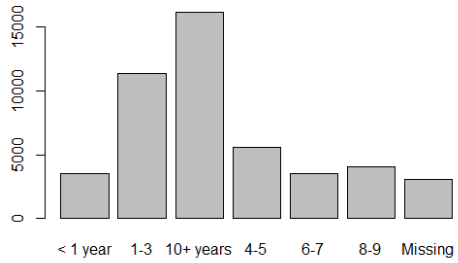
Слика 4.7: Нова расподела статуса кредита

Такође, код процената (колоне *int_rate* и *revol_util*) користићемо децималне бројеве, претходно помножене са 100.

Када су у питању недостајуће вредности (Na's), приметили смо да је код променљиве дужина радног стажа *emp_length* постоји одређен број записа који немају овај податак (може се видети на слици 4.8.). Зато ћемо вредности ове променљиве груписати у одређене категорије, тако да једна од група садржи само недостајуће вредности (слика 4.9.). Поред ове променљиве, и променљива *revol_util* је садржала одређене недостајуће податке, па смо ове вредности заменили медијаном ових података (може се видети у прилогу Б.2.).

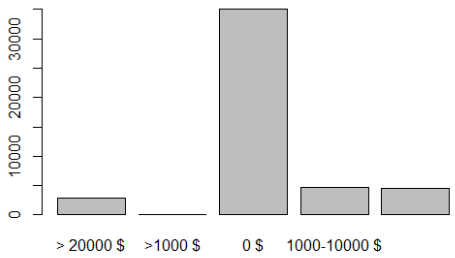


Слика 4.8: Расподела дужине радног стажа клијента



Слика 4.9: Нова расподела дужине радног стажа клијента

Променљива преостали износ кредита *(out_prncp)* садржи вредност 0 уколико је отплаћен цео износ кредита, а уколико је дошло од краха садржи износ који није отплаћен. На основу ове променљиве креираћемо нову категоричку променљиву која ће садржати вредности ове променљиве подељене у категорије (графички приказ може се видети на слици 4.10).



Слика 4.10: Расподела променљиве преостали износ за отплату кредита

Када смо податке средили, можемо прећи на креирање модела. Први модел који ћемо искористити је логистичко регресиони модел (описан у поглављу 2).

Пре него што креирамо модел, важно је поделити податке у два скупа, први скуп који ће садржати податке на основу којих ћемо тренирати модел *(train)* скуп података, и други за тестирање модела *(test)* скуп података, што се може видети у прилогу Б.2. Скуп података којим ћемо тренирати модел садржаће 70% оригиналног скупа, док ће подаци за тестирање садржати 30%.

Желимо да тестирамо да ли ће клијенти исплатити износ кредита у потпуности (преживети) или неће бити у могућности да исплате до краја (десити се догађај краха). За креирање логистичко регресионог модела користићемо зависну променљиву која узима две вредности 0 (преживео) и 1(дошло је до краха):

$$Y = \begin{cases} 0, & \text{преживео;} \\ 1, & \text{десио се догађај краха.} \end{cases} \tag{4.2}$$

где смо ову променљиву претходно формирали на основу променљиве статус кредита (једнакост (4.1)), и променљиву ћемо назвати *Defaulted*. Логистичко регресиони модел ће имати облик:

$$P\{Y = 1|X_1, ..., X_k\} = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_k X_k}}{1 + e^{\beta_0+\beta_1 X_1+...+\beta_k X_k}},$$

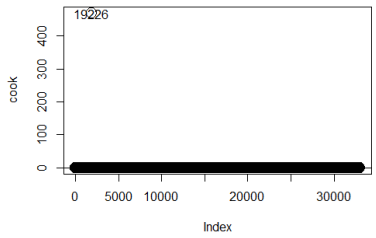
где су променљиве $X_1, ..., X_k$ износ кредита, износ рате, каматна стопа, преостали износ, итд (списак свих променљивих из модела се може

видети у прилогу Ц.1), и укупан број ових променљивих је $k = 20$, а зависна променљива Y је претходно дефинисана променљива *Defaulted* (једнакост 4.2).

У програмском језику *R* користићемо функцкију *glm* за креирање логистичко регресионог модела. Модел ћемо креирати на основу већег броја променљивих који су описане у прилогу 1, што се може видети у прилогу 2. Променљиве које нећемо укључити у модел су годишњи приход *(annual_inc)*, пошто је ова вредност укључена у логаритамску вредност променљиве *log_annual_inc*, као и одређене категоричке променљиве (на пример *grade, emp_length*) пошто смо формирали њима аналогне променљиве. Креирани модел се може видети у наставку:

```
glm(formula = defaulted ~ log_annual_income + term + int_rate + installment +
home_ownership + purpose + dti + inq_last_6mths + open_acc + revol_bal +
revol_util + tot_cur_bal + chargeoff_within_12_mths + emp_length_new +
delinq_2yrs + total_acc + out_prncp_cat + total_rec_int +
verification_status + grade_ord, family = "binomial",
data = train)
```

Статистике креираног модела се могу видети у прилогу Ц.1, и на основу ових резултата (p-вредности додељене променљивима) закључујемо да су одређене променљиве значајне (као што су *int_rate, installment, home_ownership*).
Пре него што почнемо са анализом модела, проверићемо присуство аутлајера. У претходном делу смо утврдили присуство на основу графичког приказа вредности самих променљивих појединачно, а сада ћемо искористити тест заснован на кукувом растојању, за проверу да ли постоје аутлајери у креираном моделу. На слици 4.11. можемо уочити једну вредност која знатно одступа од осталих, па можемо рећи да је она утицајна вредност у претходно креираном логистичко регресионом моделу.



Слика 4.11: Утицајне вредности у моделу на основу теста Куково растојање[10]

[10]На x-оси графика су приказани индекси, односно редни број записа међу подацима означене са *Index*, а на y-оси је вредност мере Куково растојање означене са *cook*.

Запис на који се односи ова утицајна вредност се може видети у прилогу Б.2. Ову вредност ћемо искључити из скупа података приликом креирања новог модела, који ћемо користити за тестирање података.

Када смо креирали модел, желимо да проверимо да ли су све променљиве у моделу значајне и да ли можемо побољшати резултате добијеног модела. За проверу значајности параметара искористићемо Валдов тест (описан у другом поглављу), којим ћемо тестирати значајност сваке појединачне променљиве. У наставку можемо видети резултате Валдових тестова значајности предиктора *log_annual_income* и *out_prncp_cat*:

```
Wald test for log_annual_income
in glm(formula = defaulted ~ log_annual_income + term + int_rate +
installment + home_ownership + purpose + dti + inq_last_6mths +
open_acc + revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths +
emp_length_new + purpose + delinq_2yrs + total_acc + out_prncp_cat +
total_rec_int + verification_status, family = "binomial",
data = train)
F = 10.27854 on 1 and 32986 df: p= 0.0013471

Wald test for out_prncp_cat
in glm(formula = defaulted ~ log_annual_income + term + int_rate +
installment + home_ownership + purpose + dti + inq_last_6mths +
open_acc + revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths +
emp_length_new + purpose + delinq_2yrs + total_acc + out_prncp_cat +
total_rec_int + verification_status, family = "binomial",
data = train)
F = 0.009294951 on 4 and 32986 df: p= 0.99983
```

Како смо користи велики број Валдових тестова за тестирање значајности предикора, укупно 20 тестираних хипотеза (за сваки од предиктора), вероватноћа доношења погрешних закључака постаје већа. Искористићемо Бонферонијеву корекцију (описану у другом поглављу) како бисмо смањили могућност грешке. Желимо да грешка друге врсте буде највише $\alpha' = 0.1$, и како је укупан број тестираних хипотеза $m = 20$, спровешћемо свако поређење са нивоом значајности $\alpha = \frac{0.1}{20} = 0.005$, односно нулту хипотезу ћемо одбацити ако је p вредност теста мања од 0.005.

На основу добијених резултата можемо закључити да је предиктор *log_annual_income* значајан (p-вредност теста је $p = 0.0013471$ па одбацујемо нулту хипотезу H_0 : променљива није значајна) са нивоом значајности $\alpha = 0.005$. Са истим нивоом значајности предиктор *out_prncp_cat* није појединачно значајан ($p = 0.99983$, па се прихвата нулта хипотеза). Детаљни резултати тестирања уз помоћ Валдовог теста су приказани у прилогу Ц.2, а променљиве које су појединачно значајне су *log_annual_income* и *open_acc*.

Са циљем побољшања резултате модела, искористићемо тест коли-

чника максималне веродостојности (описан у поглављу 2) како бисмо упоредили међусобно моделе, и одлучили који предикори јесу, а који нису значајни.

Искористићемо резултате добијене Валдовим тестом и Бонферонијеву корекцију, и креирати нове моделе који неће садржати предикторе који нису значајни на основу Валдовог теста и на основу значајности предиктора из првог иницијалног модел.

Прво ћемо тестирати да су сви коефициијенти уз предикторе једнаки нули, и резултати добијеног теста су:

```
Analysis of Deviance Table

Model 1: defaulted ~ 1
Model 2: defaulted ~ log_annual_income + term + int_rate + installment +
home_ownership + purpose + dti + inq_last_6mths + open_acc +
revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths +
emp_length_new + delinq_2yrs + total_acc + out_prncp_cat +
total_rec_int + verification_status + grade_ord
Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     33026      42971
2     32980      16900 46    26072 < 2.2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1          1
```

На основу добијених резултата, и p-вредности теста која је мања од $2.2e^{-16}$, одбацујемо нулту хипотезу да су сви коефицијенти једнаки нули.

Када смо утврдили да нису сви коефијенти једнаки нули, желимо да проверимо да ли су коефицијенти уз одређене предикторе једнаки нули. На пример, желимо да тестирамо да ли су коефицијенти уз предикторе *verification_status* и *chargeoff_within_12_mths* једнаки нули, добијени резултати су:

```
Analysis of Deviance Table

Model 1: defaulted ~ log_annual_income + term + int_rate + installment +
home_ownership + purpose + dti + inq_last_6mths + open_acc +
revol_bal + revol_util + tot_cur_bal + emp_length_new + delinq_2yrs +
total_acc + out_prncp_cat + total_rec_int + grade_ord
Model 2: defaulted ~ log_annual_income + term + int_rate + installment +
home_ownership + purpose + dti + inq_last_6mths + open_acc +
revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths +
emp_length_new + delinq_2yrs + total_acc + out_prncp_cat +
total_rec_int + verification_status + grade_ord
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     32983      16906
2     32980      16900  3   6.2803  0.09874 .
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1          1
```

На основу резултата, можемо закључити да са нивоом значајности 0.005 прихватамо нулту хипотезу да ови предиктори нису значајни.

Желимо да упоредимо и утицај парова предиктора у конкретном моделу. Испитаћемо утицај пара предиктора *out_prncp_cat* и *revol_bal*, и резултати су:

```
Analysis of Deviance Table

Model 1: defaulted ~ out_prncp_cat + revol_bal
Model 2: defaulted ~ log_annual_income + term + int_rate + installment +
home_ownership + purpose + dti + inq_last_6mths + open_acc +
revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths +
emp_length_new + delinq_2yrs + total_acc + out_prncp_cat +
```

```
total_rec_int + verification_status + grade_ord
Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     33021      18933
2     32980      16900 41   2033.3 < 2.2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1          1
```

На основу p-вредности која је мања од $2.2e^{-16}$ одбацујемо нулту хипотезу да су заједно једнаки нули, и закључујемо да су у пару значајни.

На основу резултата основног модела (модел *modLog* приказан у прилогу Ц.1.), можемо приметити да се значајност предиктора *"Missing"* и *"10+ years"* знатно разликује од осталих предиктора промењиве *emp_length_new*, па желимо да проверимо да ли можемо побољшати резултате модела на основу прегруписања категорија ове промењиве. Формираћемо нову промењиву:

$$emp\_length\_bon = \begin{cases} Missing, & \text{ако је } emp\_length\_new = \text{"Missing"}, \\ 10 + years & \text{ако је } emp\_length\_new = \text{"10+ years"}, \\ Others, & \text{остале вредности ове промењиве.} \end{cases}$$

и ову промењиву ћемо искористити приликом поређења модела како бисмо утврили да ли су добијени резултати модела побољшани.

Нов тест који желимо да проверимо је утицај ново креиране категоричке промењиве *emp_length_bon* у односу на утицај промењиве *emp_length_new*, и добијени резултати су:

```
Analysis of Deviance Table

Model 1: defaulted ~ emp_length_bon
Model 2: defaulted ~ emp_length_new
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     33024      42869
2     33020      42843  4   25.789 3.49e-05 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1          1
```

На основу мале p-вредности која износи $3.49e^{-05}$ одбацујемо нулту хипотезу и можемо закључити да је модел бољи са предиктором *emp_length_new*.

Сви креирани модели са одређеним избаченим предикторима, као и резултати теста максималне веродостојности се могу видети у прилогу Ц.3, а као крајњи резултат добили смо нов модел, где смо искључили предикторе *verification_status, chargeoff_within_12_mths, dti i revol_bal* и он изгледа (резултати модела се могу видети у прилогу Ц.1.):

```
glm(formula = defaulted ~ log_annual_income + term + int_rate +
installment + home_ownership + purpose + inq_last_6mths +
open_acc + revol_util + tot_cur_bal + emp_length_new + delinq_2yrs +
total_acc + out_prncp_cat + total_rec_int + grade_ord, family = "binomial",
data = train)
```

Резултати поређења овог модела са избаченим предикторима *verification_status, chargeoff_within_12_mths, dti i revol_bal* у односу на почетни модел су(користећи тест количника веродостојности):

```
Analysis of Deviance Table

Model 1: defaulted ~ log_annual_income + term + int_rate + installment +
home_ownership + purpose + inq_last_6mths + open_acc + revol_util +
tot_cur_bal + emp_length_new + delinq_2yrs + total_acc +
out_prncp_cat + total_rec_int + grade_ord
Model 2: defaulted ~ log_annual_income + term + int_rate + installment +
home_ownership + purpose + dti + inq_last_6mths + open_acc +
revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths +
emp_length_new + delinq_2yrs + total_acc + out_prncp_cat +
total_rec_int + verification_status + grade_ord
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     32985      16906
2     32980      16900  5   6.3658   0.2722
```

На основу p-вредности која је једнака 0.2722 прихватамо нулту хипотезу, и закључујемо су коефицијенти уз предикторе *verification_status, chargeoff_within_12_mths, dti i revol_bal* једнаки нули, и да ови предиктори нису значајни.

Функција *glm* враћа оцене параметара $\beta_0, \beta_1, ..., \beta_{17}$ у логистичко регресионом моделу, и оне износе:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	23.1611	229.0902	0.1011	0.9195
log_annual_income	−0.1396	0.0509	−2.7433	0.0061**
term 60 months	0.2159	0.0519	4.1593	0.0000***
int_rate	−0.1664	0.0155	−10.7203	0.0000***
installment	0.0011	0.0001	11.8098	0.0000***
home_ownershipOWN	0.1872	0.0642	2.9154	0.0036**
home_ownershipRENT	0.4372	0.0516	8.4785	0.0000***
home_ownershipANY	−17.5983	7499.8684	−0.0023	0.9981
purposecredit_card	0.3621	0.2391	1.5148	0.1298
purposedebt_consolidation	0.4613	0.2349	1.9634	0.0496*
purposehome_improvement	0.7079	0.2444	2.8966	0.0038**
purposehouse	−0.2614	0.3346	−0.7813	0.4346
purposemajor_purchase	0.3916	0.2645	1.4802	0.1388
purposemedical	0.7941	0.2812	2.8238	0.0047**
purposemoving	0.5543	0.2998	1.8489	0.0645.
purposeother	0.6267	0.2433	2.5758	0.0100*
purposerenewable_energy	1.1725	0.5517	2.1253	0.0336*
purposesmall_business	0.8782	0.2865	3.0653	0.0022**
purposevacation	0.4319	0.3300	1.3085	0.1907
inq_last_6mths	0.1101	0.0210	5.2567	0.0000***
open_acc	0.0207	0.0050	4.1461	0.0000***
revol_util	0.0105	0.0009	12.0315	0.0000***
tot_cur_bal	0.0000	0.0000	−5.3923	0.0000***
emp_length_new1−3	−0.0809	0.0822	−0.9840	0.3251
emp_length_new10+ years	−0.2607	0.0815	−3.1978	0.0014**
emp_length_new4−5	−0.0258	0.0916	−0.2821	0.7779
emp_length_new6−7	−0.1296	0.1029	−1.2588	0.2081
emp_length_new8−9	−0.0492	0.0982	−0.5013	0.6161
emp_length_newMissing	0.4489	0.1005	4.4679	0.0000***
delinq_2yrs	0.0795	0.0204	3.8978	0.0001***
total_acc	−0.0076	0.0024	−3.1827	0.0015**
out_prncp_cat>1000 $	−0.0929	1673.6975	−0.0001	1.0000
out_prncp_cat0 $	−21.7145	229.0892	−0.0948	0.9245
out_prncp_cat1000−10000 $	0.0417	289.9220	0.0001	0.9999
out_prncp_cat10000−20000 $	−0.0365	293.9903	−0.0001	0.9999
total_rec_int	−0.0004	0.0000	−13.3632	0.0000***
grade_ord.L	5.2861	0.3152	16.7718	0.0000***
grade_ord.Q	−0.2074	0.0864	−2.4014	0.0163*
grade_ord.C	−0.2052	0.0757	−2.7109	0.0067**
grade_ord^4	0.0362	0.0650	0.5567	0.5777
grade_ord^5	0.0195	0.0539	0.3621	0.7173
grade_ord^6	0.0222	0.0457	0.4851	0.6276

Сада можемо прећи на тестирање података[11]. Креираћемо нову

[11]Користићемо фунцкију predict у R-у за предвиђање података

променљиву predicted у скупу података тест (прилог Ц.2.).

На основу добијених резултата можемо израчунати матрицу класификације *(confusion matrix)*, и добијене вредности су приказане у табели испод (4.2.), где је Def зависна променљива, а $\hat{Def}$ оцењена вредност.

		Стварне класе - *(Def)*	
		0	1
Оцењене класе - $(\hat{Def})$	0	9073	1331
	1	55	3695

Табела 4.2: Матрица класификације

На основу ове матрице можемо израчунати и одређене мере (описане у поглављу 2), и добијамо да су њихове вредности[12]:

$$ACC = 0.9021, \tag{4.3}$$
$$TPR = 0.9940, \tag{4.4}$$
$$TNR = 0.7352, \tag{4.5}$$
$$PPV = 0.8721, \tag{4.6}$$
$$F1 = 2 \cdot \frac{PPV \cdot TPR}{PPV + TPR} = 0.9291. \tag{4.7}$$

Како су ови резултати добри (на основу описа мера у поглављу 2), можемо закључити да овај модел логистичке регресије добро описује податке.

Још један од начина провере слагања модела је и *ROC* крива. Конструкција ове криве се може видети у прилогу 2, а њен изглед је приказан на слици 4.12.

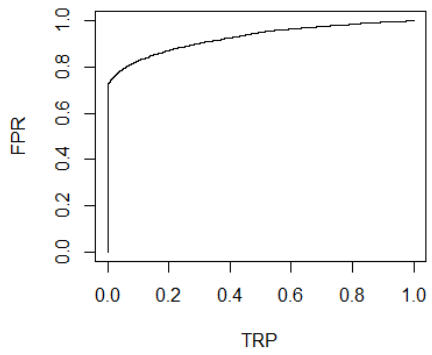
На основу *ROC* криве можемо одредити вредност мере *AUC (Area Under the Curve)* и Ђинијев коефицијент:

$$AUC = 0.9269, \tag{4.8}$$
$$Gini = 0.8539. \tag{4.9}$$

.

На основу описа ових мера у поглављу 2, закључујемо да су резултати добијеног модела добри, и да је урађено одлично раздвајање.

[12]Мере се могу израчунати на основу описаних формула у поглављу 2, а ми смо користили уграђену функцију у R-у *ConfusionMatrix*, и добијени резултати се могу видети у прилогу Б.2.



Слика 4.12: *ROC* крива

Желимо да упоредимо резултате добијеног модела и са другим моделом, и да онда на основу добијених резултата међусобно упоредимо моделе. Искористићемо модел стабла одлучивања (описан у трећем поглављу рада).

Стабло одлука *(Decision Tree)* је добар метод класификације података, и релативно је једноставан за имплементацију. За потребе креирања стабла одлука, потребно је користити одговарајућу зависну променљиву, тј класификовати клијенте у две групе:

- Ризични клијенти (код којих постоји могућност да не испуне своје обавезе до краја),
- Сигурни клијенти (клијенти који ће своје обавезе испунити у потпуности).

Ову променљиву смо дефинисали приликом креирања модела логистичке регресије (променљива Y из једнакости (4.1)), па ћемо њу искоритити приликом креирања и овог модела.

Стабло одлучивања се креира на основу искључиво категоричких променљивих, па ћемо за израду овог модела искористити тренутне категоричке променљиве као предикторе: *grade, home_ownership, term, emp_length_new, loan_status, home_ownership, verification_status.* Искористићемо и непрекидне променљиве, које ћемо поделити у одговарајуће

интервале и на основу којих ћемо креирати одговарајуће нове категоричке променљиве (то су променљиве: *loan_amnt_cat, annual_inc_cat, total_acc_cat, int_rate_cat, installment_cat, revol_bal_cat, out_prncp_cat*, а интервали на основу којих су подељени се могу видети у прилогу Б.3.)

Недостатак стабала одлучивања је њихова нестабилност јер мале флуктуације у узорку података могу резултирати великим варијацијама у додељеним класификацијама, прорачуни могу постати јако комплексни, нарочито ако су многе вредности непоуздане и/или ако је много позитивних исхода.

Можемо приметити, да међу подацима има много више сигурних кредита него ризичних, а несразмеран број података може довести до погрешних закључака и прорачуни могу бити јако комплексни (може се видети у трећем поглављу приликом описа недостатака стабла одлучивања), па ћемо због тога издвојити исти број сигурних записа колико имамо и ризичних.

Фокусираћемо се на бинарно стабло одлука, и зато нам је потребно да наше променљиве трансформишемо у бинарне променљиве.

На пример, оригинална променљива ранг клијента *(grade)* је категоричка променљива и узима вредности A,B,C...:

	grade
2	B
27	C
61	C
77	A
86	C
92	C

Слика 4.13: Вредности променљиве ранг клијента[13]

[13]У табели су приказане оригиналне вредности за променљиву ранг за одређене записе, тако на пример у другом запису променљива ранг узима вредност B.

Након трансформације вредности променљиве ранг у бинарне вредности, на основу атрибута које је узимала оригинална променљива, креиране су нове променљиве које узимају вредности 0 и 1, као на слици 4.14.

loansDS$gradeA	loansDS$gradeB	loansDS$gradeC	loansDS$gradeD	loansDS$gradeE	loansDS$gradeF	loansDS$gradeG
0	1	0	0	0	0	0
0	0	1	0	0	0	0
0	0	1	0	0	0	0
1	0	0	0	0	0	0
0	0	1	0	0	0	0
0	0	1	0	0	0	0

Слика 4.14: Бинарне вредности променљиве ранг клијента

На овај начин ћемо трансформисати све променљиве које се налазе у нашем скупу података (код трансформације се може видети у прилогу 2).

Када смо прилагодили податке за израду модела, можемо издвојити податке које ћемо користити за тренирање модела, и касније за тестирање добијених резултате, 50% података користићемо за тренирање и 50% за тестирање модела (прилог 2).

Сада када имамо издвојене податке за тренирање модела, можемо креирати стабло одлука на основу ових података. За креирање стабла користићемо фунцкију *rpart* у R-у, и резултати добијеног модела су:

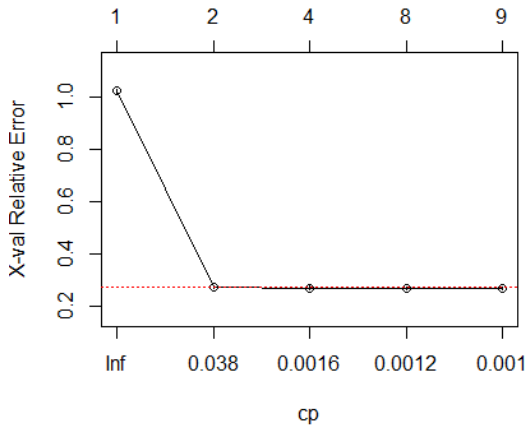
```
Classification tree:
rpart(formula = safe ~ ., data = train_data, method = "class")

Variables actually used in tree construction:
[1] loansDS$annual_inc_cat10000-100000 $ loansDS$emp_lengthMissing
[3] loansDS$gradeD                       loansDS$gradeG
[5] loansDS$home_ownershipMORTGAGE       loansDS$home_ownershipOWN
[7] loansDS$int_rate_cat20-30 $          loansDS$out_prncp_cat0 $

Root node error: 8377/16754 = 0.5

n= 16754

CP nsplit rel error  xerror      xstd
1 0.7274681      0   1.00000 1.02507 0.0077233
2 0.0019697      1   0.27253 0.27253 0.0053010
3 0.0012733      3   0.26859 0.26967 0.0052774
4 0.0010744      7   0.26334 0.26680 0.0052536
5 0.0010000      8   0.26227 0.26681 0.0052534
```

Графички приказ стабла одлучивања се може видети у прилогу Б.3.

Када смо креирали стабло, желимо да проверимо да ли можемо побољшати резултате стабла (скратити стабло[14]). За овај метод, потребно је одредити граничну вредност c за коју је најмања вредност грешке. Ова вредност се може прочитати на основу резултата добијеног модела, и она износи $c = 0.001$ на основу најмање вредности грешке

[14]Метод скраћивање стабла *(pruning)* је описан у поглављу 3 овог рада.

($xerror = 0.26227$), што се може видети и на основу графичког приказа са слике 4.15.

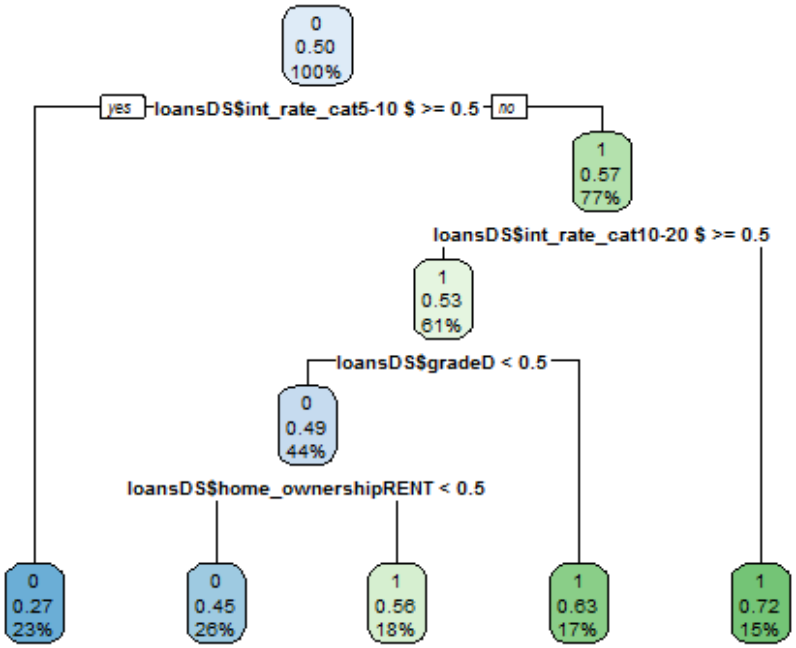


Слика 4.15: Граниче вредности у односу на вредност грешке

Ову граничну вредност ћемо искористити за скраћивање стабла, и ново креирано стабло се може видети на слици 4.16.

На основу чворова стабла, можемо закључити да је 23% укупног броја клијената који су узели кредит са каматном стопом од 5-10% ће испунити своје обавезе (отплатити цео износ кредита у договореном временском периоду), и да је вероватноћа овог догађаја једнака 0.27. А када је у питању догађај неиспуњења обавеза, 15% укупног броја клијената су узели кредит са каматном стопом мањом од 5% и већом од 20% са вероватноћом 0.72 неће бити у могућности да отплате кредит, као и 17% клијената који су узели кредит са каматном стопом од 10-20% и не припадају рангу D. И на основу овог модела стабла одлучивања, тестираћемо податке[15].

[15]Користићемо функцију *predict* у R-у, код се може видети у прилогу Б.3.



Слика 4.16: Графички приказ стабла одлучивања

На основу добијених резултата, креирамо матрицу класификације[16], и резултати су приказани у табели 4.3.

		Стварне класе - (Def)	
		0	1
Оцењене класе - $(\hat{Def})$	0	8110	1984
	1	267	6393

Табела 4.3: Матрица класификације

[16]Матрицу ћемо креирати на основу функције *ConfusionMatrix* у програмском језику R.

На основу резултата матрице можемо израчунати и мере (описане у поглављу 2), и резулатати су:

$$ACC = 0.8656, \tag{4.10}$$
$$TPR = 0.9681, \tag{4.11}$$
$$TNR = 0.7632, \tag{4.12}$$
$$PPV = 0.8034, \tag{4.13}$$
$$F1 = 2 \cdot \frac{PPV \cdot TPR}{PPV + TPR} = 0.8781. \tag{4.14}$$

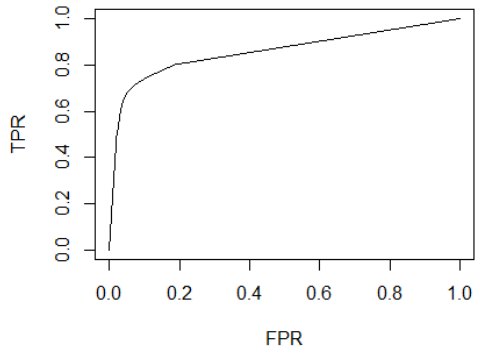
Како су ови резултати добри (на основу описа мера у поглављу 2), можемо закључити да креирани модел стабла одлучивања добро описује податке.

Искористићемо и мере AUC и Ђинијев коефицијент за проверу модела, и вредности ових коефицијената су:

$$AUC = 0.8656, \tag{4.15}$$
$$Gini = 0.7312. \tag{4.16}$$

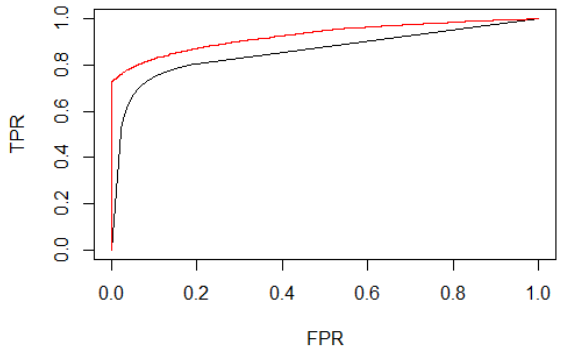
Такође, на основу ових вредности (описа параметара у другом поглављу) можемо закључити да модел добро описује податке.

Одговарајућа ROC крива се може видети на слици 4.17.



Слика 4.17: ROC крива

Када упоредимо резултате добијене логистичко регресионим моделом и модела стабла одлуке (на основу резултата приказаних у једнакостима (4.1)-(4.6) за логистичко регресиони модел, и једнакостима (4.8)-(4.14) за модел стабла одлучивања) можемо закључити да оба модела добро описују податке, али је за нијансу бољи логистичко регресиони модел, што се може видети и на основу поређења положаја *ROC* кривих на слици (4.18).



Слика 4.18: Поређење *ROC* кривих[17]

[17]На графику је црвеном бојом приказана *ROC* крива креирана на основу логистичко регресионог модела, а црном бојом је креирана на основу стабла одлука.

Закључак

Погрешна процена кредитне способности клијената има за последицу губитке, који могу изазвати велике проблеме. С једне стране одобравање кредита клијенту који неће бити у могућности да испуни своје обавезе у потпуности директан је трошак банке, а са друге стране, неодобравање кредита добром клијенту умањује финанцијски учинак. Стога је питање процене кредитне способности клијента потребно посветити посебну пажњу. Као што је у раду напоменуто, као подршка финансијским институцијама, развијени су и користе се бројни модели кредитног ризика.

У овом раду су приказани логистичко регресиони модели и модели стабла одлучивања, теоријски и примена са конкретним подацима. Резултати анализе показали су да се важним предиктором може сматрати променљива ранг клијената, каматна стопа, преостали износ за отплату кредита, па тако клијенти који су узели кредит са мањом каматном стопом, као и клијенти рангирани са оценом A са великом вероватноћом ће отплатити цео износ кредита на време, док за клијенте рангиране испод оцене D и са великим преосталим износом за отплату кредита постоји велика могућност да неће вратити кредит на време, као и утицај поседовања некретнина на саму отплату кредита. Иако је тек након имплементације у пракси могуће извршити адекватну валидацију модела, на основу наведеног може се закључити да примена модела доприноси побољшању управљања кредитним ризиком, што доприноси бољим анализама, и редукцију кредитног ризика.

Литература

[1] W. Schoutens, J. Cariboni (2009): Levy processes in credit risk

[2] V. Matić (2009): ASB Banking risk 13: Basell II-The Standardised approach to credit risk

[3] V. Matić (2011): BAZEL III - IZMENJENI KONCEPT KAPITALA, Udruženje banaka Srbije

[4] V. Misić (2012): Internal models for the credit risk management

[5] Wikipedia: Credit Risk, Exposure at default, Probability at deafult, Expected Loss

[6] T.Maltus (1798): An Essey on the Principle of Population, Paul's churchyard

[7] P. C. Thompson, M. Carlson (2010): A Dissertation Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy

[8] Lj. Kvesić (2013): Primena stabla odlučivanja u kreditnom skoringu, Ekonomski vjesnik, 382-390.

[9] П. Младеновић (2008): Вероватноћа и статистика, Математички факултет, Београд

[10] В. Јевремовић, Ј. Малишић (2002): Статистичке методе у метеорологији и инжињерству

[11] G. Grozdić (2011): Primenjena logistička regresija

[12] J. S. Crammer (2002): The Origins of Logistic Regression

[13] BIOST 515: Estimation and hypothesis testing for logistic regression, Lecture 13

[14] D. W. Hosmer, S. Lemeshow (2000): Applied Logistic Regression

[15] MEDCALC (1993-2017): ROC curve analysis

[16] J. J. Faraway (2006): Extending the Linear Model with R, New York

[17] D. T. Larose (2005): Discoovering knowledge in data, Canada

[18] D. T. Larose (2006): Data mining methods and models, Canada

[19] Lending Club Corporation: Lending Club Statistics

[20] A. Signorell (2017): Tools for Descriptive Statistics, Package 'DescTools', CRAN

[21] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone (1983) Classification and Regression Trees, Wadsworth

[22] T. M. Therneau, E. J. Atkinson (2017): An Introduction to Recursive Partitioning Using the RPART Routines

Прилози

Прилог А

Опис података

Редни број	Назив променљиве	Опис	Вредност
1	addr_state	Дрзава клијента	AK,AL,AR,..
2	annual_inc	Годишњи приход клијента	Износ у доларима $
3	chargeoff_within_12_mths	Број наплата у последњих 12 месеци	Цео број
4	delinq_2yrs	Укупан број касњења већа од 30 дана у току две године	Цео број
5	dti	Удео месецног износа кредита у односу на укупан износкључујући хипотеке	Проценат
6	out_prncp	Преостали износ главнице кредита за отплату	Износ у доларима $
7	emp_length	Године стажа	Категорије, мање од 1 год, 1 год, 2 год ...
8	home_ownership	Поседовање некретнине	any, mortgage, own, rent, none

Редни број	Назив променљиве	Опис	Вредност
9	id	Редни број	Цео број
10	inq_last_6mths	Број пријава за кредит у последњих шест месеци	Цео број
11	installment	Рата кредита	Износ у доларима $
12	int_rate	Каматна стопа	Проценат
13	loan_amnt	Укупан износ позајмице/кредита	Износ у доларима $
14	loan_status	Статус кредита	Current, Default, Fully Paid...
15	member_id	Редни број клијента	Цео број
16	open_acc	Број отворених рачуна клијента	Цео број
17	purpose	Сврха аплицирања за кредит	Слободан опис
18	revol_bal	Укупно износ кредитних картица клијента	Износ у доларима $
19	revol_util	Износ који клијент користи на свим кредитним картицама	Износ у доларима $
20	grade	Рангирање клијента	A,B,C....
21	tot_cur_bal	Тренутни салто клијетна	Износ у доларима $

Редни број	Назив променљиве	Опис	Вредност
22	total_acc	Укупан број рачуна	Цео Број
23	verification_status	Статус потврде	Not, Source ili Verified
24	term	Број месеци за отплату кредита	Цео број
25	total_rec_int	Укупна отплаћена камата кредита до данас	Износ у доларима $

Прилог Б

Програмски код

Б.1 Слике коришћене у раду

```
> # sigmoid funkcija:
> sigmoid <- function(a,x){
+   1 / ( 1 + exp(-a* x))
+ }
> x <- seq(-2.5, 2.5, 0.001)
>
> plot(x, sigmoid(1,x),xlim=c(-3,3), ylim = c(0, 1.3)
>       , col='blue',xlab = '',ylab = '',type="l",axes=F)
> u <- par("usr")
>
> axis(1, pos=0,outer = TRUE)
> axis(2, pos=0)
> arrows(u[1], 0, u[2], 0, code = 2, xpd = TRUE)
> arrows(0, u[3], 0, u[4], code = 2, xpd = TRUE)
>
> title(xlab="t", line=2, cex.lab=1.5, mgp=c(8,5,0))
> title(ylab="P(t)", line=-10, cex.lab=1.5)
>
> lines(x, x/x+0.02, ylim = c(0, 1.3), col='black')
> lines(x, x+0.5, ylim = c(0, 1.3), col='black')
> lines(x, sigmoid(2,x), ylim = c(0, 1), col='yellow')
> lines(x, sigmoid(3,x), ylim = c(0, 1), col='green')
> lines(x, sigmoid(5,x), ylim = c(0, 1), col='red')
> lines(x, sigmoid(1,x), ylim = c(0, 1), col='blue')
>
> # logisticka funkcija raspodele
> library(stats)
>
> x<-seq(-5,20,0.01)
>
> plot(x,plogis(x,location=5,scale=2, lower.tail = TRUE
+       , log.p = FALSE),ylim=c(0,1),xlim=c(-5,20)
+       , col='blue',xlab = 'x',ylab = 'F(x)',type="l")
>
> lines(x,plogis(x,location=9,scale=3, lower.tail = TRUE
+               , log.p = FALSE), col='green')
> lines(x,plogis(x,location=9,scale=4, lower.tail = TRUE
+               , log.p = FALSE), col='red')
> lines(x,plogis(x,location=6,scale=2, lower.tail = TRUE
+               , log.p = FALSE), col='yellow')
> lines(x,plogis(x,location=2,scale=1, lower.tail = TRUE
+               , log.p = FALSE), col='purple')
> grid()
>
> legend(14, 0.5, legend=c("m=5,␣s=2","m=9,␣s=3","m=9,␣s=4"
+               ,"m=6,␣s=2","m=2,␣s=1"),
+        col=c("blue", "green", "red", "yellow","purple")
+               , lty=1, cex=0.8)
>
> # logisticka gustina raspodele
> x<-seq(-5,20,0.01)
>
```

```
> plot(x,dlogis(x,location=5,scale=2, log =FALSE)
+          , ylim=c(0,0.3), xlim=c(-5,20)
+       , col='blue',xlab = 'x',ylab = 'f(x)',type="l")
> lines(x,dlogis(x,location=9,scale=3, log = FALSE)
+                , col='green')
> lines(x,dlogis(x,location=9,scale=4, log = FALSE)
+                , col='red')
> lines(x,dlogis(x,location=6,scale=2, log = FALSE)
+                , col='yellow')
> lines(x,dlogis(x,location=2,scale=1, log = FALSE)
+                , col='purple')
> grid()
>
> legend(14, 0.28, legend=c("m=5,␣s=2","m=9,␣s=3","m=9,␣s=4"
+  ,"m=6,␣s=2","m=2,␣s=1"),col=c("blue", "green", "red",
+  "yellow","purple"), lty=1, cex=0.8)
>
> # ROC kriva
> library(ROCR)
> iris$isv <- as.numeric(iris$Species == "versicolor")
> mod <- glm(isv~Sepal.Length+Sepal.Width, data=iris
+         , family="binomial")
> pred1 <- prediction(predict(mod), iris$isv)
> perf1 <- performance(pred1,"tpr","fpr")
>
> plot(perf1,xlab='FPR',ylab='TPR',type="l",col="red")
> x<-seq(0,1,0.01)
> lines(x, x, ylim = c(0, 1.3), col='black',type="l"
+       , lty=2, lwd=1.4)
```

Б.2 Примена модела

Б.2.1 Основне статистике података

```
> # preuzimanje podataka za 2016 godinu:
> loan_data_2016Q1 <- read.csv("D:/Master_rad_izrada/Primeri/loan_data/LoanStats_2016Q1.csv")
> loan_data_2016Q2 <- read.csv("D:/Master_rad_izrada/Primeri/loan_data/LoanStats_2016Q2.csv")
> loan_data_2016Q3 <- read.csv("D:/Master_rad_izrada/Primeri/loan_data/LoanStats_2016Q3.csv")
> loan_data_2016Q4 <- read.csv("D:/Master_rad_izrada/Primeri/loan_data/LoanStats_2016Q4.csv")
> # nadovezivanje podataka - jedan skup podataka:
> loan_data <- rbind(loan_data_2016Q1,loan_data_2016Q2,loan_data_2016Q3,loan_data_2016Q4)
> #statistike:
> summary(loan_data)
id              member_id           loan_amnt          term           int_rate
Min.   :   55716   Min.   :   113909   Min.   : 1000   36 months:323495   Min.
:0.0532
1st Qu.:74685647   1st Qu.: 80062544   1st Qu.: 8000   60 months:110912
1st Qu.:0.0949
Median :81454189   Median : 87190177   Median :12400
Median :0.1199
Mean   :81944688   Mean   : 87851154   Mean   :14734                      Mean
:0.1304
3rd Qu.:90137900   3rd Qu.: 96588878   3rd Qu.:20000
3rd Qu.:0.1559
Max.   :96453160   Max.   :103570872   Max.   :40000                      Max.
:0.3099

installment      grade        sub_grade                 emp_title         emp_length
Min.   :  30.12   A: 70847   C1     : 31576                    : 28487   10+ years:149972
1st Qu.: 247.22   B:134512   B5     : 31339   Teacher         :  8248   2 years  : 39601
Median : 375.63   C:132178   B4     : 29053   Manager         :  7403   3 years  : 34734
Mean   : 444.12   D: 59178   C2     : 26596   Owner           :  5007   < 1 year : 31918
3rd Qu.: 592.39   E: 25807   B3     : 25898   Registered Nurse:  3381   1 year   : 29156
Max.   :1584.90   F:  9334   C4     : 25878   (Other)         :381880   n/a      : 28214
G:  2551   (Other):264067   NA's   :     1   (Other)  :120812
home_ownership     annual_inc          verification_status       loan_status
MORTGAGE:211516   Min.   :      0   Not Verified   :133828   Current            :381178
OWN     : 53037   1st Qu.:  48000   Source Verified:174294   Fully Paid         : 30427
RENT    :169744   Median :  67000   Verified       :126285   Late (31-120 days):  6280
ANY     :   110   Mean   :  79498                            Issued             :  6048
3rd Qu.:  95000                              Charged Off       :  4567
Max.   :9573072                              In Grace Period   :  3748
(Other)            :  2159
purpose                         title           zip_code       addr_state
debt_consolidation:248899   Debt consolidation     :234563   750xx  :  4571   CA
: 57888
credit_card       : 91609   Credit card refinancing: 86183   945xx  :  4475   TX
: 37036
```

```
home_improvement  : 31182   Home improvement        : 29884   112xx  :  4416   NY
: 35505
other             : 28469   Other                  : 27745   606xx  :  4104   FL
: 31727
major_purchase    : 10406                          : 23173   300xx  :  3959   IL
: 17715
medical           :  5440   Major purchase         : 10031   331xx  :  3715   NJ
: 15891
(Other)           : 18402   (Other)                : 22828   (Other):409167
(Other):238645
dti           delinq_2yrs       inq_last_6mths      open_acc        revol_bal
revol_util
Min.   : -1.00   Min.   : 0.0000   Min.   :0.0000   Min.   : 1.00   Min.   :       0
Min.   :0.0000
1st Qu.:  12.31   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.: 8.00   1st Qu.:    6003
1st Qu.:0.3240
Median :  18.22   Median : 0.0000   Median :0.0000   Median :11.00   Median :   11203
Median :0.5040
Mean   :  20.33   Mean   : 0.3606   Mean   :0.5614   Mean   :11.88   Mean   :   16943
Mean   :0.5068
3rd Qu.:  24.87   3rd Qu.: 0.0000   3rd Qu.:1.0000   3rd Qu.:15.00   3rd Qu.:   20064
3rd Qu.:0.6900
Max.   :9999.00   Max.   :29.0000   Max.   :5.0000   Max.   :97.00   Max.   :1044210
Max.   :1.7200
NA's   :1                                                  NA's   :264
total_acc       out_prncp       total_pymnt     total_rec_prncp    total_rec_int
last_pymnt_amnt
Min.   :  2.00   Min.   :    0   Min.   :    0   Min.   :    0.0   Min.   :   0.0
Min.   :    0.0
1st Qu.: 16.00   1st Qu.: 5246   1st Qu.: 1015   1st Qu.:  649.4   1st Qu.: 258.2
1st Qu.:  250.5
Median : 23.00   Median :10003   Median : 2327   Median : 1484.5   Median : 593.0
Median :  392.8
Mean   : 24.52   Mean   :11763   Mean   : 3732   Mean   : 2813.8   Mean   : 915.9
Mean   : 1316.8
3rd Qu.: 31.00   3rd Qu.:16989   3rd Qu.: 4549   3rd Qu.: 3004.2   3rd Qu.:1225.4
3rd Qu.:  654.9
Max.   :176.00   Max.   :40000   Max.   :46886   Max.   :40000.0   Max.   :9459.1
Max.   :42148.5

tot_cur_bal       chargeoff_within_12_mths
Min.   :      0   Min.   :0.000000
1st Qu.:  30328   1st Qu.:0.000000
Median :  81448   Median :0.000000
Mean   : 143451   Mean   :0.009074
3rd Qu.: 212890   3rd Qu.:0.000000
Max.   :5445012   Max.   :9.000000

> str(loan_data)
'data.frame':   434407 obs. of  32 variables:
$ id                        : int  74523825 75993583 76022756 75800404 75933549 76041549 ...
$ member_id                 : int  79900601 81484367 81513504 81268205 81424322 81532277
...
$ loan_amnt                 : int  12000 22000 25000 12500 24000 16800 30000 35000 5625 10550 ...
$ term                      : Factor w/ 2 levels " 36 months"," 60 months": 2 1 2 1 2 2 2 2 1 2 ...
$ int_rate                  : num  0.1147 0.0649 0.1299 0.1299 0.1953 ...
$ installment               : num  264 674 569 421 630 ...
$ grade                     : Factor w/ 7 levels "A","B","C","D",..: 2 1 3 3 4 4 2 4 4 3 ...
$ sub_grade                 : Factor w/ 35 levels "A1","A2","A3",..: 10 2 12 12 20 17 9 18 18 15 ...
$ emp_title                 : Factor w/ 121664 levels "","
Assembler",..: 41347 21257 ...
$ emp_length                : Factor w/ 12 levels "< 1 year","1 year",..: 3 3 3 7 2 3 4 3 3 3 ...
$ home_ownership            : Factor w/ 4 levels "MORTGAGE","OWN",..: 3 1 3 3 2 3 1 1 1 3 ...
$ annual_inc                : num  30000 134000 138000 55000 135000 ...
$ verification_status       : Factor w/ 3 levels "Not Verified",..: 3 3 1 3 3 2 3 3 3 1 ...
$ loan_status               : Factor w/ 8 levels "Charged Off",..: 2 2 2 2 2 2 2 2 2 2 ...
$ purpose                   : Factor w/ 13 levels "car","credit_card",..: 3 3 3 3 2 3 3 3 4 3 ...
$ title                     : Factor w/ 14 levels "","Business",..: 6 6 6 6 5 6 6 6 9 6 ...
$ zip_code                  : Factor w/ 911 levels "007xx","008xx",..: 723 319 70 873 251 ...
$ addr_state                : Factor w/ 50 levels "AK","AL","AR",..: 6 10 31 47 27 15 43 14 37 10 ...
$ dti                       : num  40.8 26.3 12.4 33.7 20.2 ...
$ delinq_2yrs               : int  0 0 0 0 0 0 0 0 0 3 ...
$ inq_last_6mths            : int  1 1 1 0 0 0 0 0 0 0 ...
$ open_acc                  : int  29 20 9 11 14 14 20 23 7 10 ...
$ revol_bal                 : int  23705 60963 8577 18170 25934 9906 26495 41602 1773 3594 ...
$ revol_util                : num  0.167 0.67 0.371 0.733 0.679 0.416 0.532 0.63 0.269 0.313 ...
$ total_acc                 : int  43 34 24 18 19 35 43 26 14 20 ...
$ out_prncp                 : num  10606 16894 22198 9813 21704 ...
$ total_pymnt               : num  2358 6064 5082 3772 5795 ...
$ total_rec_prncp           : num  1394 5106 2802 2687 2296 ...
$ total_rec_int             : num  965 957 2280 1086 3498 ...
$ last_pymnt_amnt           : num  264 674 569 421 630 ...
$ tot_cur_bal               : int  36850 405151 11565 46568 118667 51348 256353 310728 ...
```

```
$ chargeoff_within_12_mths: int  0 0 0 0 0 0 0 0 0 0 ...
> library(DescTools)
Warning message:
package   DescTools   was built under R version 3.3.3
> # raspodela sl promenljive loan_amount
> Desc(loan_data$loan_amnt, main = "", plotit = TRUE)
──────────────────────────────────────────────────────────────────────────────


length          n        NAs      unique          0s        mean      meanCI
434'407    434'407          0       1'533           0   14'734.04   14'707.30
100.0%       0.0%                   0.0%               14'760.78

.05        .10        .25      median         .75         .90         .95
3'200.00   5'000.00   8'000.00   12'400.00   20'000.00   28'300.00   34'850.00

range          sd      vcoef         mad         IQR        skew        kurt
39'000.00   8'991.62       0.61   8'895.60   12'000.00        0.75       −0.19

lowest : 1'000 (1'975), 1'025 (5), 1'050 (10), 1'075 (7), 1'100 (45)
highest: 39'900 (4), 39'925 (3), 39'950, 39'975 (4), 40'000 (3'432)

> # promenljiva purpose
> Desc(loan_data$purpose, main = "", plotit = TRUE)
──────────────────────────────────────────────────────────────────────────────

length       n      NAs  unique  levels   dupes
434'407 434'407        0      13      13       y
100.0%     0.0%

level      freq    perc  cumfreq  cumperc
1    debt_consolidation   248'899   57.3%   248'899    57.3%
2           credit_card    91'609   21.1%   340'508    78.4%
3      home_improvement    31'182    7.2%   371'690    85.6%
4                 other    28'469    6.6%   400'159    92.1%
5         major_purchase    10'406    2.4%   410'565    94.5%
6               medical     5'440    1.3%   416'005    95.8%
7                   car     4'813    1.1%   420'818    96.9%
8         small_business     4'790    1.1%   425'608    98.0%
9               vacation     3'262    0.8%   428'870    98.7%
10                moving     3'229    0.7%   432'099    99.5%
11                 house     2'002    0.5%   434'101    99.9%
12     renewable_energy       304    0.1%   434'405   100.0%
... etc.
[list output truncated]

> # izdvajanje podskupa podataka
> # full_paid -> 1
> # charged_off, defaulted, late (31−120 days), Late (16−30 days) -> 0
> loans <- subset(loan_data, loan_status %in% c("Charged Off","Default","Fully Paid"
+                                               ,"In Grace Period"
+                                               ,"Late (16−30 days)","Late (31−120 days)"))
> # ordered promenljive u R−u
> loans$grade_ord <- ordered(loans$grade)
> # log vredbost kolone annual_income
> loans$log_annual_income <- log(loans$annual_inc)
> loans$log_annual_income[which(loans$log_annual_income == "−Inf")] <- 0
> loans$emp_length_new <- rep(NA, length(loans$emp_length)) # nova kolona
> # inicijalizacija, grupisanje vrednosti u kategorije
> loans$emp_length_new[which(loans$emp_length == "< 1 year")] <- "< 1 year"
> loans$emp_length_new[which(loans$emp_length %in% c("1 year","2 years","3 years"))] <- "1−3"
> loans$emp_length_new[which(loans$emp_length %in% c("4 years","5 years"))] <- "4−5"
> loans$emp_length_new[which(loans$emp_length %in% c("6 years","7 years"))] <- "6−7"
> loans$emp_length_new[which(loans$emp_length %in% c("8 years","9 years"))] <- "8−9"
> loans$emp_length_new[which(loans$emp_length == "10+ years")] <- "10+ years"
> loans$emp_length_new[which(loans$emp_length == "n/a")] <- "Missing"
> loans$emp_length_new= as.factor(loans$emp_length_new)
> summary(loans$emp_length_new)
< 1 year       1−3 10+ years       4−5       6−7       8−9   Missing
3509     11319     16106      5561      3538      4092      3056
> # revol_util − Na's
> loans$revol_util[which(is.na(loans$revol_util))] <- median(loans$revol_util,na.rm=TRUE)
> # nova kategor prom − za status kredita, vrednosti 0 i 1
> loans$defaulted <- 1 # inicijalizacija
> index_fp <- which(loans$loan_status == "Fully Paid")
> loans$defaulted[index_fp] <- 0
> # int rate i revol_util, pomnozene sa 100
> loans$int_rate <- loans$int_rate *100
> loans$revol_util <- loans$revol_util * 100
```

Б.2.2 Логистичко регресиони модел

```
> library(caTools)
> # podela podataka - train and test:
> spl = sample.split(loans$defaulted, 0.7)
> train = subset(loans, spl == TRUE)
> test = subset(loans, spl == FALSE)
>
> modLog = glm(defaulted ~ log_annual_income + term + int_rate + installment
+              + home_ownership + purpose + dti + inq_last_6mths + open_acc
+              + revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths
+              + emp_length_new + purpose + delinq_2yrs + total_acc
+              + out_prncp_cat + total_rec_int + verification_status + grade_ord
+              , data=train, family="binomial")
> # Kukovo rastojanje:
> cook <- cooks.distance(modLog)
> plot(cook, cex=2, main="",xlab = "",ylab = "") # plot cook's distance
> abline(h = 4*mean(cook, na.rm=T)) # add cutoff line
> text(x=1:length(cook)+1, y=cook
+      , labels=ifelse(cook>4*mean(cook, na.rm=T)
+                      ,names(cook),"")) # add labels
> influential <- as.numeric(names(cook)
+                           [(cook > 4*mean(cook, na.rm=T))]) # influential row numbers
> head(loans[influential, ])
loan_amnt       term int_rate installment grade     emp_title emp_length home_ownership annual_inc
1245      6000  36 months     9.16      191.25     B medical sales  10+ years
MORTGAGE     150000
verification_status loan_status             purpose addr_state
dti delinq_2yrs inq_last_6mths
1245             Verified Charged Off debt_consolidation         NY 4.46
2                1
open_acc revol_bal revol_util total_acc out_prncp total_rec_int last_pymnt_amnt tot_cur_bal
1245        9      3402       27.9        13          0        170.39
191.25       30805
chargeoff_within_12_mths grade_ord log_annual_income emp_length_new defaulted new_loan_status
1245                        0         B          11.91839      10+ years
1       Defaulted
out_prncp_cat total_rec_prncp_cat loan_amnt_cat total_acc_cat
annual_inc_cat int_rate_cat
1245             0 $              >1000 $     < 10000 $
10-50 100000-1000000 $       5-10 $
installment_cat revol_bal_cat
1245       100-500 $      < 5000 $
> # Waldov test
> library(survey)
> regTermTest(modLog, "term")
Wald test for term
in glm(formula = defaulted ~ log_annual_income + term + int_rate +
installment + home_ownership + purpose + dti + inq_last_6mths +
open_acc + revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths +
emp_length_new + purpose + delinq_2yrs + total_acc + out_prncp_cat +
total_rec_int + verification_status + grade_ord, family = "binomial",
data = train)
F = 24.07312  on  1  and  32980  df: p= 9.3188e-07
> regTermTest(modLog, "int_rate")
Wald test for int_rate
in glm(formula = defaulted ~ log_annual_income + term + int_rate +
installment + home_ownership + purpose + dti + inq_last_6mths +
open_acc + revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths +
emp_length_new + purpose + delinq_2yrs + total_acc + out_prncp_cat +
total_rec_int + verification_status + grade_ord, family = "binomial",
data = train)
F = 116.3649  on  1  and  32980  df: p= < 2.22e-16
> regTermTest(modLog, "installment")
Wald test for installment
in glm(formula = defaulted ~ log_annual_income + term + int_rate +
installment + home_ownership + purpose + dti + inq_last_6mths +
open_acc + revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths +
emp_length_new + purpose + delinq_2yrs + total_acc + out_prncp_cat +
total_rec_int + verification_status + grade_ord, family = "binomial",
data = train)
F = 134.8402  on  1  and  32980  df: p= < 2.22e-16
> # konacan model:
> modlog_new = glm(defaulted ~ log_annual_income + term + int_rate + installment
+                  + home_ownership + purpose + inq_last_6mths + open_acc
+                  + revol_util + tot_cur_bal
+                  + emp_length_new + purpose + delinq_2yrs + total_acc
+                  + out_prncp_cat + total_rec_int + grade_ord
+                  , data=train[-influential, ], family="binomial")
> #koeficijenti modela:
> round(summary(modlog_new)$coef,4)
Estimate Std. Error  z value Pr(>|z|)
```

```
(Intercept)                    23.2737   230.6399    0.1009   0.9196
log_annual_income              -0.1491     0.0512   -2.9141   0.0036
term 60 months                  0.2166     0.0522    4.1494   0.0000
int_rate                       -0.1664     0.0156  -10.6687   0.0000
installment                     0.0011     0.0001   11.8522   0.0000
home_ownershipOWN               0.1890     0.0645    2.9305   0.0034
home_ownershipRENT              0.4398     0.0518    8.4905   0.0000
home_ownershipANY             -17.5972  7496.4990   -0.0023   0.9981
purposecredit_card              0.3514     0.2393    1.4683   0.1420
purposedebt_consolidation       0.4503     0.2352    1.9144   0.0556
purposehome_improvement         0.6870     0.2448    2.8066   0.0050
purposehouse                   -0.2638     0.3349   -0.7879   0.4308
purposemajor_purchase           0.3819     0.2652    1.4403   0.1498
purposemedical                  0.7687     0.2822    2.7238   0.0065
purposemoving                   0.5198     0.3014    1.7244   0.0846
purposeother                    0.6219     0.2436    2.5527   0.0107
purposerenewable_energy         1.1645     0.5518    2.1104   0.0348
purposesmall_business           0.8629     0.2876    3.0006   0.0027
purposevacation                 0.4301     0.3304    1.3018   0.1930
inq_last_6mths                  0.1112     0.0210    5.2846   0.0000
open_acc                        0.0213     0.0050    4.2386   0.0000
revol_util                      0.0104     0.0009   11.9358   0.0000
tot_cur_bal                     0.0000     0.0000   -5.3473   0.0000
emp_length_new1-3              -0.0877     0.0825   -1.0629   0.2878
emp_length_new10+ years        -0.2649     0.0818   -3.2396   0.0012
emp_length_new4-5              -0.0285     0.0919   -0.3097   0.7568
emp_length_new6-7              -0.1430     0.1035   -1.3824   0.1669
emp_length_new8-9              -0.0500     0.0985   -0.5078   0.6116
emp_length_newMissing           0.4485     0.1008    4.4477   0.0000
delinq_2yrs                     0.0830     0.0204    4.0598   0.0000
total_acc                      -0.0078     0.0024   -3.2429   0.0012
out_prncp_cat>1000 $           -0.0935  1673.9169   -0.0001   1.0000
out_prncp_cat0 $              -21.7124   230.6390   -0.0941   0.9250
out_prncp_cat1000-10000 $       0.0389   291.7211    0.0001   0.9999
out_prncp_cat10000-20000 $     -0.0377   296.0230   -0.0001   0.9999
total_rec_int                  -0.0004     0.0000  -13.2591   0.0000
grade_ord.L                     5.2765     0.3166   16.6652   0.0000
grade_ord.Q                    -0.1977     0.0865   -2.2849   0.0223
grade_ord.C                    -0.2000     0.0759   -2.6366   0.0084
grade_ord^4                     0.0360     0.0652    0.5516   0.5812
grade_ord^5                     0.0165     0.0541    0.3048   0.7605
grade_ord^6                     0.0199     0.0459    0.4343   0.6641
> #koeficijenti modela:
> round(summary(modlog_new)$coef,4)
Estimate Std. Error  z value Pr(>|z|)
(Intercept)                    23.2737   230.6399    0.1009   0.9196
log_annual_income              -0.1491     0.0512   -2.9141   0.0036
term 60 months                  0.2166     0.0522    4.1494   0.0000
int_rate                       -0.1664     0.0156  -10.6687   0.0000
installment                     0.0011     0.0001   11.8522   0.0000
home_ownershipOWN               0.1890     0.0645    2.9305   0.0034
home_ownershipRENT              0.4398     0.0518    8.4905   0.0000
home_ownershipANY             -17.5972  7496.4990   -0.0023   0.9981
purposecredit_card              0.3514     0.2393    1.4683   0.1420
purposedebt_consolidation       0.4503     0.2352    1.9144   0.0556
purposehome_improvement         0.6870     0.2448    2.8066   0.0050
purposehouse                   -0.2638     0.3349   -0.7879   0.4308
purposemajor_purchase           0.3819     0.2652    1.4403   0.1498
purposemedical                  0.7687     0.2822    2.7238   0.0065
purposemoving                   0.5198     0.3014    1.7244   0.0846
purposeother                    0.6219     0.2436    2.5527   0.0107
purposerenewable_energy         1.1645     0.5518    2.1104   0.0348
purposesmall_business           0.8629     0.2876    3.0006   0.0027
purposevacation                 0.4301     0.3304    1.3018   0.1930
inq_last_6mths                  0.1112     0.0210    5.2846   0.0000
open_acc                        0.0213     0.0050    4.2386   0.0000
revol_util                      0.0104     0.0009   11.9358   0.0000
tot_cur_bal                     0.0000     0.0000   -5.3473   0.0000
emp_length_new1-3              -0.0877     0.0825   -1.0629   0.2878
emp_length_new10+ years        -0.2649     0.0818   -3.2396   0.0012
emp_length_new4-5              -0.0285     0.0919   -0.3097   0.7568
emp_length_new6-7              -0.1430     0.1035   -1.3824   0.1669
emp_length_new8-9              -0.0500     0.0985   -0.5078   0.6116
emp_length_newMissing           0.4485     0.1008    4.4477   0.0000
delinq_2yrs                     0.0830     0.0204    4.0598   0.0000
total_acc                      -0.0078     0.0024   -3.2429   0.0012
out_prncp_cat>1000 $           -0.0935  1673.9169   -0.0001   1.0000
out_prncp_cat0 $              -21.7124   230.6390   -0.0941   0.9250
out_prncp_cat1000-10000 $       0.0389   291.7211    0.0001   0.9999
out_prncp_cat10000-20000 $     -0.0377   296.0230   -0.0001   0.9999
total_rec_int                  -0.0004     0.0000  -13.2591   0.0000
grade_ord.L                     5.2765     0.3166   16.6652   0.0000
grade_ord.Q                    -0.1977     0.0865   -2.2849   0.0223
```

```
grade_ord.C                  -0.2000     0.0759   -2.6366    0.0084
grade_ord^4                   0.0360     0.0652    0.5516    0.5812
grade_ord^5                   0.0165     0.0541    0.3048    0.7605
grade_ord^6                   0.0199     0.0459    0.4343    0.6641
> cm < - confusionMatrix(as.numeric(test$predicted >= 0.5), test$defaulted)
> cm
Confusion Matrix and Statistics

Reference
Prediction    0    1
0 9075 1330
1   53 3696

Accuracy : 0.9023
95% CI : (0.8973, 0.9071)
No Information Rate : 0.6449
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.7737
Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.9942
Specificity : 0.7354
Pos Pred Value : 0.8722
Neg Pred Value : 0.9859
Prevalence : 0.6449
Detection Rate : 0.6412
Detection Prevalence : 0.7351
Balanced Accuracy : 0.8648

'Positive' Class : 0

> #Test Area Under the Curve (AUC), i ROC kriva
> library(ROCR)
> library(ROCR)
> p <- predict(modlog_new, newdata=test, type="response")
> pr <- prediction(p, test$defaulted)
> prf <- performance(pr, measure = "tpr", x.measure = "fpr")
> plot(prf,xlab = "TRP",ylab="FPR")
>
> auc <- performance(pr, measure = "auc")
> auc <- auc@y.values[[1]]
> auc # AUC vrednost
[1] 0.9269694
> # gini mera:
> gini <- 2* auc -1
> gini
[1] 0.8539388
```

Б.2.3 Модели стабла одлука

```
> # kreiranje novih kategorickih promenljivih:
> loans$loan_amnt_cat <- rep(NA, length(loans$loan_amnt)) # nova kolona
> loans$loan_amnt_cat[which(loans$loan_amnt < 10000)] <- "<_10000_$"
> loans$loan_amnt_cat[which(loans$loan_amnt >= 10000 & loans$loan_amnt
>                                                    < 20000)] <- "10000-20000_$"
> loans$loan_amnt_cat[which(loans$loan_amnt >= 20000 & loans$loan_amnt
>                                                         < 30000)] <- "20000-30000_$"
> loans$loan_amnt_cat[which(loans$loan_amnt >= 30000 & loans$loan_amnt <= 40000)]
>                                                         <- "30000-40000_$"
> loans$loan_amnt_cat= as.factor(loans$loan_amnt_cat)
> summary(loans$loan_amnt_cat)
< 10000 $ 10000-20000 $ 20000-30000 $ 30000-40000 $
15595          17111           9452           5023
> loans$total_acc_cat <- rep(NA, length(loans$total_acc)) # nova kolona
> loans$total_acc_cat[which(loans$total_acc < 10)] <- "<_10"
> loans$total_acc_cat[which(loans$total_acc >= 10 & loans$total_acc < 50)] <- "10-50"
> loans$total_acc_cat[which(loans$total_acc >= 50 & loans$total_acc < 100)] <- "50-100"
> loans$total_acc_cat[which(loans$total_acc >= 100)] <- ">100"
> loans$total_acc_cat= as.factor(loans$total_acc_cat)
> summary(loans$total_acc_cat)
< 10    >100  10-50 50-100
2658      14   42375    2134
> loans$annual_inc_cat <- rep(NA, length(loans$annual_inc)) # nova kolona
> loans$annual_inc_cat[which(loans$annual_inc < 1000)] <- "<_1000_$"
> loans$annual_inc_cat[which(loans$annual_inc >= 1000 & loans$annual_inc < 10000)]
>                                                    <- "1000-10000_$"
> loans$annual_inc_cat[which(loans$annual_inc >= 10000 & loans$annual_inc < 100000)]
>                                                    <- "10000-100000_$"
> loans$annual_inc_cat[which(loans$annual_inc >= 100000 & loans$annual_inc <= 1000000)]
>                                                    <- "100000-1000000_$"
```

```
> loans$annual_inc_cat[which(loans$annual_inc > 1000000)] <- ">_1000000_$"
> loans$annual_inc_cat= as.factor(loans$annual_inc_cat)
> summary(loans$annual_inc_cat)
< 1000 $       > 1000000 $      1000-10000 $    10000-100000 $ 100000-1000000 $
4                 11                40             36308             10818
> loans$int_rate_cat <- rep(NA, length(loans$int_rate)) # nova kolona
> loans$int_rate_cat[which(loans$int_rate < 5)] <- "<_5_$"
> loans$int_rate_cat[which(loans$int_rate >= 5 & loans$int_rate < 10)] <- "5-10_$"
> loans$int_rate_cat[which(loans$int_rate >= 10 & loans$int_rate < 20)] <- "10-20_$"
> loans$int_rate_cat[which(loans$int_rate >= 20 & loans$int_rate <= 30)] <- "20-30_$"
> loans$int_rate_cat[which(loans$int_rate > 30)] <- ">_30_$"
> loans$int_rate_cat= as.factor(loans$int_rate_cat)
> summary(loans$int_rate_cat)
> 30 $ 10-20 $ 20-30 $  5-10 $
63    29075     6839    11204
> loans$installment_cat <- rep(NA, length(loans$installment)) # nova kolona
> loans$installment_cat[which(loans$installment < 50)] <- "<_50_$"
> loans$installment_cat[which(loans$installment >= 50 & loans$installment < 100)]
>                                                 <- "50-100_$"
> loans$installment_cat[which(loans$installment >= 100 & loans$installment < 500)]
>                                                 <- "100-500_$"
> loans$installment_cat[which(loans$installment >= 500 & loans$installment <= 1000)]
>                                                 <- "500-1000_$"
> loans$installment_cat[which(loans$installment > 1000)] <- ">_1000_$"
> loans$installment_cat= as.factor(loans$installment_cat)
> summary(loans$installment_cat)
< 50 $    > 1000 $   100-500 $    50-100 $ 500-1000 $
480         2945       27408        1687       14661
> loans$revol_bal_cat <- rep(NA, length(loans$revol_bal)) # nova kolona
> loans$revol_bal_cat[which(loans$revol_bal < 5000)] <- "<_5000_$"
> loans$revol_bal_cat[which(loans$revol_bal >= 5000 & loans$revol_bal < 10000)]
>                                                 <- "5000-10000_$"
> loans$revol_bal_cat[which(loans$revol_bal >= 10000 & loans$revol_bal < 20000)]
>                                                 <- "10000-20000_$"
> loans$revol_bal_cat[which(loans$revol_bal >= 20000 & loans$revol_bal <= 50000)]
>                                                 <- "20000-50000_$"
> loans$revol_bal_cat[which(loans$revol_bal > 50000)] <- ">_50000_$"
> loans$revol_bal_cat= as.factor(loans$revol_bal_cat)
> summary(loans$revol_bal_cat)
< 5000 $      > 50000 $ 10000-20000 $ 20000-50000 $   5000-10000 $
10920           1965          13196           9637          11463
> library(dplyr)
> loans_dt <- select(loans, grade_ord, term, home_ownership, emp_length_new
+                    ,loan_status, defaulted, verification_status, purpose
+                    ,loan_amnt_cat, annual_inc_cat, total_acc_cat, int_rate_cat
+                    ,installment_cat, revol_bal_cat, out_prncp_cat)
> safe_loans  <- loans_dt[loans_dt$loan_status == 'Fully_Paid',]
> risky_loans <- loans_dt[loans_dt$loan_status %in% c("Charged_Off","Default"
+                                                     ,"In_Grace_Period"
+
,"Late_(16-30_days)","Late_(31-120_days)"),]
> # izdvajamo isti broj redova rizicnih i sigurnih kredita
> safe_loans <- safe_loans[1:nrow(risky_loans),]
> # Spajamo podatke:
> loansDS <- rbind(safe_loans, risky_loans)
> str(loansDS)
'data.frame':   33508 obs. of  15 variables:
$ grade_ord          : Ord.factor w/ 7 levels "A"<"B"<"C"<"D" <..: 4 3 1 4 2 1 1 2 3 2 ...
$ term               : Factor w/ 2 levels "_36_months","_60_months": 1 1 1 1 1 1 1 1 1 1 ...
$ home_ownership     : Factor w/ 4 levels "MORTGAGE","OWN",..: 3 1 2 1 2 3 3 2 2 1 ...
$ emp_length_new     : Factor w/ 7 levels "<_1_year","1-3",..: 3 3 7 4 3 6 2 7 6 2 ...
$ loan_status        : Factor w/ 8 levels "Charged_Off",..: 4 4 4 4 4 4 4 4 4 4 ...
$ defaulted          : num  0 0 0 0 0 0 0 0 0 0 ...
$ verification_status: Factor w/ 3 levels "Not_Verified",..: 2 1 1 3 1 1 2 1 1 2 ...
$ purpose            : Factor w/ 13 levels "car","credit_card",..: 2 3 2 2 3 3 4 3 3 2 ...
$ loan_amnt_cat      : Factor w/ 4 levels "<_10000_$","10000-20000_$",..: 1 1 3 4 3 2 1 1 2 1 ...
$ annual_inc_cat     : Factor w/ 5 levels "<_1000_$",">_1000000_$",..: 4 4 5 5 4 5 4 4 4 4 ...
$ total_acc_cat      : Factor w/ 4 levels "<_10",">100",..: 3 3 4 4 3 3 3 3 3 3 ...
$ int_rate_cat       : Factor w/ 4 levels ">_30_$","10-20_$",..: 2 2 4 2 2 4 4 2 2 4 ...
$ installment_cat    : Factor w/ 5 levels "<_50_$",">_1000_$",..: 3 3 5 2 5 3 3 3 3 3 ...
$ revol_bal_cat      : Factor w/ 5 levels "<_5000_$",">_50000_$",..: 5 3 4 2 3 5 5 4 3 5 ...
$ out_prncp_cat      : Factor w/ 5 levels ">_20000_$",">1000_$",..: 3 3 3 3 3 3 3 3 3 3 ...
> # Transformacija podataka u binarne - binarno stablo odluka
> library(knitr)
> # provera originalnih i transformisanih podataka za promenljivu rang
> knitr::kable(head(loansDS[1]))


|    |grade_ord |
|:---|:---------|
|18  |D         |
|20  |C         |
|58  |A         |
```

```
|69 |D          |
|73 |B          |
|79 |A          |
> knitr::kable(head(model.matrix(~loansDS$grade - 1)))


| loansDS$gradeA| loansDS$gradeB| loansDS$gradeC| loansDS$gradeD|
loansDS$gradeE| loansDS$gradeF| loansDS$gradeG|
|--------------:|--------------:|--------------:|--------------:
|--------------:|--------------:|--------------:|
|              0|              0|              0|              1|
0|              0|              0|
|              0|              0|              1|              0|
0|              0|              0|
|              1|              0|              0|              0|
0|              0|              0|
|              0|              0|              0|              1|
0|              0|              0|
|              0|              1|              0|              0|
0|              0|              0|
|              1|              0|              0|              0|
0|              0|              0|
> # Transformacija svih podataka:
> loans.data <- data.frame(safe = loansDS$defaulted)
> loans.data <- cbind(loans.data, model.matrix(~loansDS$grade - 1))
> loans.data <- cbind(loans.data, model.matrix(~loansDS$term - 1))
> loans.data <- cbind(loans.data, model.matrix(~loansDS$home_ownership - 1))
> loans.data <- cbind(loans.data, model.matrix(~loansDS$verification_status - 1))
> loans.data <- cbind(loans.data, model.matrix(~loansDS$purpose - 1))
> loans.data <- cbind(loans.data, model.matrix(~loansDS$emp_length - 1))
> loans.data <- cbind(loans.data, model.matrix(~loansDS$loan_amnt_cat - 1))
> loans.data <- cbind(loans.data, model.matrix(~loansDS$total_acc_cat - 1))
> loans.data <- cbind(loans.data, model.matrix(~loansDS$annual_inc_cat - 1))
> loans.data <- cbind(loans.data, model.matrix(~loansDS$installment_cat - 1))
> loans.data <- cbind(loans.data, model.matrix(~loansDS$int_rate_cat - 1))
> loans.data <- cbind(loans.data, model.matrix(~loansDS$revol_bal_cat - 1))
> loans.data <- cbind(loans.data, model.matrix(~loansDS$out_prncp_cat - 1))
> library(caret)
> library(rpart)
> library(rpart.plot)
> library(caTools)
> # podela podataka - train and test:
> spl = sample.split(loans.data$safe, 0.5)
> train_data = subset(loans.data, spl == TRUE)
> test_data = subset(loans.data, spl == FALSE)
> # izrada stabla
> rpart_tree <- rpart(safe ~ ., train_data, method = 'class',cp = 0.001)
> printcp(rpart_tree) # ispis rezultata

Classification tree:
rpart(formula = safe ~ ., data = train_data, method = "class",
cp = 0.001)

Variables actually used in tree construction:
[1] loansDS$gradeD                  loansDS$gradeE
loansDS$home_ownershipMORTGAGE
[4] loansDS$int_rate_cat20-30 $     loansDS$out_prncp_cat0 $        loansDS$term 36 months

Root node error: 8377/16754 = 0.5

n= 16754

CP nsplit rel error  xerror      xstd
1 0.7216187      0   1.00000 1.00752 0.0077255
2 0.0034022      1   0.27838 0.27838 0.0053485
3 0.0014325      3   0.27158 0.27468 0.0053185
4 0.0010000      6   0.26728 0.27182 0.0052951
> plotcp(rpart_tree,col = 2,upper = "size") # vizuelni prikaz, cp = 0.01
> rpart.plot(rpart_tree) # iscrtavanje stabla
> pfit<- prune(rpart_tree, cp= 0.0016) #  za odgovarajucu vrednost cp
> rpart.plot(pfit)
> # predikcja
> rpart_pred <- predict(rpart_tree, test_data, type = 'class')
> test_data$rpart_pred<- predict(rpart_tree, test_data, type = 'class')
> test_data$rpart_pred <- as.numeric(test_data$rpart_pred)
> # matrica klasifikacije
> confusionMatrix(rpart_pred, test_data$safe)
Confusion Matrix and Statistics

Reference
Prediction    0    1
0 8172 1966
1  205 6411
```

```
Accuracy : 0.8704
95% CI : (0.8652, 0.8755)
No Information Rate : 0.5
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.7408
Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.9755
Specificity : 0.7653
Pos Pred Value : 0.8061
Neg Pred Value : 0.9690
Prevalence : 0.5000
Detection Rate : 0.4878
Detection Prevalence : 0.6051
Balanced Accuracy : 0.8704

'Positive' Class : 0

> library(ROCR)
> #Test Area Under the Curve (AUC), i ROC kriva
> pred = prediction(test_data$rpart_pred, test_data$safe)
> as.numeric(performance(pred, "auc")@y.values) # AUC vrednost
[1] 0.870419
> ROCRperf = performance(pred, "tpr", "fpr")
> plot(ROCRperf, xlab="FPR", ylab="TPR")
```

Прилог Ц

Логистичко регресиони модел

Ц.1 Креирани модели

```
> modLog = glm(defaulted ~ log_annual_income + term + int_rate + installment
+              +  home_ownership + purpose + dti + inq_last_6mths + open_acc
+              + revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths
+              + emp_length_new + purpose + delinq_2yrs + total_acc
+              + out_prncp_cat + total_rec_int + verification_status + grade_ord
+              , data=train, family="binomial")
> summary(modLog)

Call:
glm(formula = defaulted ~ log_annual_income + term + int_rate +
installment + home_ownership + purpose + dti + inq_last_6mths +
open_acc + revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths +
emp_length_new + purpose + delinq_2yrs + total_acc + out_prncp_cat +
total_rec_int + verification_status + grade_ord, family = "binomial",
data = train)

Deviance Residuals:
Min         1Q     Median         3Q        Max
-1.57666  -0.49004  -0.32202   0.00007   2.94067

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)                          2.315e+01  2.290e+02   0.101 0.919506
log_annual_income                   -1.341e-01  5.345e-02  -2.510 0.012082 *
term 60 months                       2.241e-01  5.240e-02   4.278 1.89e-05 ***
int_rate                            -1.688e-01  1.561e-02 -10.818  < 2e-16 ***
installment                          1.079e-03  9.443e-05  11.428  < 2e-16 ***
home_ownershipOWN                    1.886e-01  6.439e-02   2.929 0.003405 **
home_ownershipRENT                   4.357e-01  5.175e-02   8.419  < 2e-16 ***
home_ownershipANY                   -1.763e+01  7.485e+03  -0.002 0.998121
purposecredit_card                   3.571e-01  2.390e-01   1.494 0.135123
purposedebt_consolidation            4.523e-01  2.349e-01   1.925 0.054201 .
purposehome_improvement              7.052e-01  2.444e-01   2.886 0.003901 **
purposehouse                        -2.756e-01  3.348e-01  -0.823 0.410421
purposemajor_purchase                3.861e-01  2.645e-01   1.459 0.144458
purposemedical                       7.881e-01  2.813e-01   2.802 0.005078 **
purposemoving                        5.273e-01  3.001e-01   1.757 0.078927 .
purposeother                         6.152e-01  2.434e-01   2.528 0.011479 *
purposerenewable_energy              1.137e+00  5.519e-01   2.061 0.039336 *
purposesmall_business                8.775e-01  2.864e-01   3.063 0.002189 **
purposevacation                      4.182e-01  3.302e-01   1.266 0.205360
dti                                 -3.551e-05  2.098e-04  -0.169 0.865582
inq_last_6mths                       1.085e-01  2.097e-02   5.175 2.28e-07 ***
open_acc                             2.106e-02  5.055e-03   4.167 3.09e-05 ***
revol_bal                           -2.037e-07  1.217e-06  -0.167 0.867045
revol_util                           1.040e-02  9.136e-04  11.384  < 2e-16 ***
tot_cur_bal                         -1.018e-06  1.981e-07  -5.140 2.74e-07 ***
chargeoff_within_12_mths             1.777e-01  1.623e-01   1.095 0.273496
emp_length_new1-3                   -8.091e-02  8.231e-02  -0.983 0.325642
emp_length_new10+ years             -2.640e-01  8.172e-02  -3.230 0.001236 **
emp_length_new4-5                   -2.825e-02  9.178e-02  -0.308 0.758258
emp_length_new6-7                   -1.328e-01  1.031e-01  -1.288 0.197578
```

```
emp_length_new8-9                      -5.098e-02  9.834e-02  -0.518 0.604157
emp_length_newMissing                   4.169e-01  1.017e-01   4.100 4.13e-05 ***
delinq_2yrs                             7.560e-02  2.082e-02   3.630 0.000283 ***
total_acc                              -7.820e-03  2.390e-03  -3.272 0.001069 **
out_prncp_cat>1000 $                   -7.969e-02  1.672e+03   0.000 0.999962
out_prncp_cat0 $                       -2.172e+01  2.290e+02  -0.095 0.924459
out_prncp_cat1000-10000 $               3.867e-02  2.899e+02   0.000 0.999894
out_prncp_cat10000-20000 $             -4.165e-02  2.939e+02   0.000 0.999887
total_rec_int                          -4.172e-04  3.117e-05 -13.385  < 2e-16 ***
verification_statusSource Verified     -1.247e-02  5.356e-02  -0.233 0.815910
verification_statusVerified             9.111e-02  5.668e-02   1.607 0.107984
grade_ord.L                             5.303e+00  3.155e-01  16.811  < 2e-16 ***
grade_ord.Q                            -2.108e-01  8.657e-02  -2.435 0.014883 *
grade_ord.C                            -2.067e-01  7.576e-02  -2.728 0.006366 **
grade_ord^4                             3.684e-02  6.504e-02   0.566 0.571091
grade_ord^5                             2.333e-02  5.397e-02   0.432 0.665540
grade_ord^6                             2.209e-02  4.576e-02   0.483 0.629308
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 42971  on 33026  degrees of freedom
Residual deviance: 16900  on 32980  degrees of freedom
AIC: 16994

Number of Fisher Scoring iterations: 18
>
>
> modlog_new = glm(defaulted ~ log_annual_income + term + int_rate + installment
+                  + home_ownership + purpose + inq_last_6mths + open_acc
+                  + revol_util + tot_cur_bal
+                  + emp_length_new + purpose + delinq_2yrs + total_acc
+                  + out_prncp_cat + total_rec_int + grade_ord
+                  , data=train[-influential, ], family="binomial")
> summary(modlog_new)

Call:
glm(formula = defaulted ~ log_annual_income + term + int_rate +
installment + home_ownership + purpose + inq_last_6mths +
open_acc + revol_util + tot_cur_bal + emp_length_new + purpose +
delinq_2yrs + total_acc + out_prncp_cat + total_rec_int +
grade_ord, family = "binomial", data = train[-influential,
])

Deviance Residuals:
Min         1Q     Median         3Q        Max
-1.60062  -0.49062  -0.32233   0.00007   2.91236

Coefficients:
Estimate       Std. Error z value             Pr(>|z|)
(Intercept)                  23.1618455024  229.0903860042   0.101                0.91947
log_annual_income            -0.1396447511    0.0508737497  -2.745
0.00605 **
term 60 months                0.2157396575    0.0519101106   4.156
0.0000323834 ***
int_rate                     -0.1663560331    0.0155171764 -10.721 < 0.0000000000000002 ***
installment                   0.0010954113    0.0000927464
11.811 < 0.0000000000000002 ***
home_ownershipOWN             0.1870316136    0.0642137478   2.913
0.00358 **
home_ownershipRENT            0.4369501819    0.0515639943
8.474 < 0.0000000000000002 ***
home_ownershipANY           -17.5981122411 7499.9262476062  -0.002                0.99813
purposecredit_card            0.3623887365    0.2390501297   1.516                0.12953
purposedebt_consolidation     0.4612088765    0.2349421094   1.963
0.04964 *
purposehome_improvement       0.7078066822    0.2443823577   2.896
0.00378 **
purposehouse                 -0.2614093527    0.3345615314  -0.781                0.43460
purposemajor_purchase         0.3915722203    0.2645388276   1.480                0.13882
purposemedical                0.7941552100    0.2812281723   2.824
0.00474 **
purposemoving                 0.5543595039    0.2998259825   1.849
0.06447 .
purposeother                  0.6267053133    0.2433208333   2.576
0.01001 *
purposerenewable_energy       1.1725286523    0.5517076704   2.125
0.03356 *
purposesmall_business         0.8781026988    0.2864984348   3.065
0.00218 **
purposevacation               0.4318615522    0.3300308857   1.309                0.19069
inq_last_6mths                0.1100885289    0.0209512821   5.255
```

```
0.0000001484 ***
open_acc                          0.0206933903    0.0049924531   4.145
0.0000339911 ***
revol_util                        0.0104827711    0.0008711522
12.033 < 0.0000000000000002 ***
tot_cur_bal                      -0.0000010253    0.0000001901  -5.394
0.0000000688 ***
emp_length_new1-3                -0.0807185466    0.0822493345  -0.981              0.32640
emp_length_new10+ years          -0.2606887746    0.0815134642  -3.198
0.00138 **
emp_length_new4-5                -0.0258516232    0.0916318998  -0.282              0.77785
emp_length_new6-7                -0.1295664290    0.1029189896  -1.259              0.20806
emp_length_new8-9                -0.0492441015    0.0981943845  -0.501              0.61602
emp_length_newMissing             0.4487930192    0.1004623073   4.467
0.0000079221 ***
delinq_2yrs                       0.0794954524    0.0204009568   3.897
0.0000975311 ***
total_acc                        -0.0075851638    0.0023846653  -3.181
0.00147 **
out_prncp_cat>1000 $             -0.0926946482 1673.7028664652   0.000              0.99996
out_prncp_cat0 $                -21.7142468934  229.0894536204  -0.095              0.92449
out_prncp_cat1000-10000 $         0.0419108566  289.9220491956   0.000              0.99988
out_prncp_cat10000-20000 $       -0.0363563969  293.9903879719   0.000              0.99990
total_rec_int                    -0.0004159230    0.0000311346 -13.359 < 0.0000000000000002 ***
grade_ord.L                       5.2861868293    0.3151714156
16.772 < 0.0000000000000002 ***
grade_ord.Q                      -0.2075454882    0.0863795482  -2.403
0.01627 *
grade_ord.C                      -0.2052576788    0.0757099951  -2.711
0.00671 **
grade_ord^4                       0.0363037569    0.0650054865   0.558              0.57652
grade_ord^5                       0.0195202607    0.0539206829   0.362              0.71734
grade_ord^6                       0.0220066220    0.0456832330   0.482              0.63000
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 42970  on 33025  degrees of freedom
Residual deviance: 16906  on 32984  degrees of freedom
AIC: 16990

Number of Fisher Scoring iterations: 18
```

Ц.2 Валдов тест

```
> library(survey)
> # znacajni parametri:
> regTermTest(modLog, "log_annual_income")
Wald test for log_annual_income
in glm(formula = defaulted ~ log_annual_income + term + int_rate +
installment + home_ownership + purpose + dti + inq_last_6mths +
open_acc + revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths +
emp_length_new + purpose + delinq_2yrs + total_acc + out_prncp_cat +
total_rec_int + verification_status, family = "binomial",
data = train)
F =  10.27854  on  1  and  32986  df: p= 0.0013471
> regTermTest(modLog, "open_acc")
Wald test for open_acc
in glm(formula = defaulted ~ log_annual_income + term + int_rate +
installment + home_ownership + purpose + dti + inq_last_6mths +
open_acc + revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths +
emp_length_new + purpose + delinq_2yrs + total_acc + out_prncp_cat +
total_rec_int + verification_status, family = "binomial",
data = train)
F =  17.0196  on  1  and  32986  df: p= 3.7087e-05
> # parametri koji nisu znacajni:
> regTermTest(modLog, "revol_bal")
Wald test for revol_bal
in glm(formula = defaulted ~ log_annual_income + term + int_rate +
installment + home_ownership + purpose + dti + inq_last_6mths +
open_acc + revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths +
emp_length_new + purpose + delinq_2yrs + total_acc + out_prncp_cat +
total_rec_int + verification_status, family = "binomial",
data = train)
F =  0.04145856  on  1  and  32986  df: p= 0.83866
> regTermTest(modLog, "chargeoff_within_12_mths")
Wald test for chargeoff_within_12_mths
in glm(formula = defaulted ~ log_annual_income + term + int_rate +
installment + home_ownership + purpose + dti + inq_last_6mths +
```

```
open_acc + revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths +
emp_length_new + purpose + delinq_2yrs + total_acc + out_prncp_cat +
total_rec_int + verification_status, family = "binomial",
data = train)
F =  1.360891  on  1  and  32986  df: p= 0.24339
> regTermTest(modLog, "out_prncp_cat")
Wald test for out_prncp_cat
in glm(formula = defaulted ~ log_annual_income + term + int_rate +
installment + home_ownership + purpose + dti + inq_last_6mths +
open_acc + revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths +
emp_length_new + purpose + delinq_2yrs + total_acc + out_prncp_cat +
total_rec_int + verification_status, family = "binomial",
data = train)
F =  0.009294951  on  4  and  32986  df: p= 0.99983
```

Ц.3 Тест количника максималне веродостојности

```
> # model bez prediktora
> anova(update(modLog, ~1),
+       modLog,
+       test="Chisq")
Analysis of Deviance Table

Model 1: defaulted ~ 1
Model 2: defaulted ~ log_annual_income + term + int_rate + installment +
home_ownership + purpose + dti + inq_last_6mths + open_acc +
revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths +
emp_length_new + delinq_2yrs + total_acc + out_prncp_cat +
total_rec_int + verification_status + grade_ord
Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     33026      42971
2     32980      16900 46    26072 < 2.2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
> # odbacujemo nultu hipotezu
> # testiranje parova prediktora:
> modLog2 = glm(defaulted ~ out_prncp_cat + revol_bal
+               , data=train, family="binomial")
> anova(modLog2,modLog, test ="Chisq")
Analysis of Deviance Table

Model 1: defaulted ~ out_prncp_cat + revol_bal
Model 2: defaulted ~ log_annual_income + term + int_rate + installment +
home_ownership + purpose + dti + inq_last_6mths + open_acc +
revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths +
emp_length_new + delinq_2yrs + total_acc + out_prncp_cat +
total_rec_int + verification_status + grade_ord
Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     33021      18933
2     32980      16900 41   2033.3 < 2.2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
> # odbacujemo nultu hipotezu
> # izbacen prediktor verification_status
> modLog3 = glm(defaulted ~ log_annual_income + term + int_rate + installment
+               + home_ownership + purpose + dti + inq_last_6mths + open_acc
+               + revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths
+               + emp_length_new + delinq_2yrs + total_acc
+               + out_prncp_cat + total_rec_int + grade_ord
+               , data=train, family="binomial")
> anova(modLog3, modLog, test ="Chisq")
Analysis of Deviance Table

Model 1: defaulted ~ log_annual_income + term + int_rate + installment +
home_ownership + purpose + dti + inq_last_6mths + open_acc +
revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths +
emp_length_new + delinq_2yrs + total_acc + out_prncp_cat +
total_rec_int + grade_ord
Model 2: defaulted ~ log_annual_income + term + int_rate + installment +
home_ownership + purpose + dti + inq_last_6mths + open_acc +
revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths +
emp_length_new + delinq_2yrs + total_acc + out_prncp_cat +
total_rec_int + verification_status + grade_ord
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     32982      16905
2     32980      16900  2   5.1242  0.07714 .
---
```

```
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
> # prihvata se nulta hipoteza, p = 0.07714 za prag znacajnosti 0.005
> # izbaceni prediktori verification_status i chargeoff_within_12_mths
> modLog4 = glm(defaulted ~ log_annual_income + term + int_rate + installment
+               +  home_ownership + purpose + dti + inq_last_6mths + open_acc
+               + revol_bal + revol_util + tot_cur_bal
+               + emp_length_new + delinq_2yrs + total_acc
+               + out_prncp_cat + total_rec_int + grade_ord
+               , data=train, family="binomial")
> anova(modLog4, modLog, test ="Chisq")
Analysis of Deviance Table

Model 1: defaulted ~ log_annual_income + term + int_rate + installment +
home_ownership + purpose + dti + inq_last_6mths + open_acc +
revol_bal + revol_util + tot_cur_bal + emp_length_new + delinq_2yrs +
total_acc + out_prncp_cat + total_rec_int + grade_ord
Model 2: defaulted ~ log_annual_income + term + int_rate + installment +
home_ownership + purpose + dti + inq_last_6mths + open_acc +
revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths +
emp_length_new + delinq_2yrs + total_acc + out_prncp_cat +
total_rec_int + verification_status + grade_ord
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     32983      16906
2     32980      16900  3   6.2803  0.09874 .
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
> # prihvata se nulta hipoteza, p = 0.09874 za prag znacajnosti 0.005
> # izbaceni prediktori verification_status, chargeoff_within_12_mths i dti
> modLog5 = glm(defaulted ~ log_annual_income + term + int_rate + installment
+               + home_ownership + purpose + inq_last_6mths + open_acc
+               + revol_bal + revol_util + tot_cur_bal
+               + emp_length_new + delinq_2yrs + total_acc
+               + out_prncp_cat + total_rec_int + grade_ord
+               , data=train, family="binomial")
> anova(modLog5, modLog, test ="Chisq")
Analysis of Deviance Table

Model 1: defaulted ~ log_annual_income + term + int_rate + installment +
home_ownership + purpose + inq_last_6mths + open_acc + revol_bal +
revol_util + tot_cur_bal + emp_length_new + delinq_2yrs +
total_acc + out_prncp_cat + total_rec_int + grade_ord
Model 2: defaulted ~ log_annual_income + term + int_rate + installment +
home_ownership + purpose + dti + inq_last_6mths + open_acc +
revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths +
emp_length_new + delinq_2yrs + total_acc + out_prncp_cat +
total_rec_int + verification_status + grade_ord
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     32984      16906
2     32980      16900  4   6.3335   0.1756
> # prihvata se nulta hipoteza, p = 0.1756 za prag znacajnosti 0.1
> # izbaceni prediktori verification_status, chargeoff_within_12_mths, dti i revol_bal
> modLog6 = glm(defaulted ~ log_annual_income + term + int_rate + installment
+               + home_ownership + purpose + inq_last_6mths + open_acc
+               + revol_util + tot_cur_bal
+               + emp_length_new + delinq_2yrs + total_acc
+               + out_prncp_cat + total_rec_int + grade_ord
+               , data=train, family="binomial")
> anova(modLog6, modLog, test ="Chisq")
Analysis of Deviance Table

Model 1: defaulted ~ log_annual_income + term + int_rate + installment +
home_ownership + purpose + inq_last_6mths + open_acc + revol_util +
tot_cur_bal + emp_length_new + delinq_2yrs + total_acc +
out_prncp_cat + total_rec_int + grade_ord
Model 2: defaulted ~ log_annual_income + term + int_rate + installment +
home_ownership + purpose + dti + inq_last_6mths + open_acc +
revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths +
emp_length_new + delinq_2yrs + total_acc + out_prncp_cat +
total_rec_int + verification_status + grade_ord
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     32985      16906
2     32980      16900  5   6.3658   0.2722
> # prihvata se nulta hipoteza, p = 0.2722 za prag znacajnosti 0.1
> # zamena sa purpose_bon
> modLog7 = glm(defaulted ~ log_annual_income + term + int_rate + installment
+               + home_ownership + purpose_bon + inq_last_6mths + open_acc
+               + revol_util + tot_cur_bal
+               + emp_length_new + delinq_2yrs + total_acc
+               + out_prncp_cat + total_rec_int + grade_ord
+               , data=train, family="binomial")
> anova(modLog7, modLog, test ="Chisq")
Analysis of Deviance Table
```

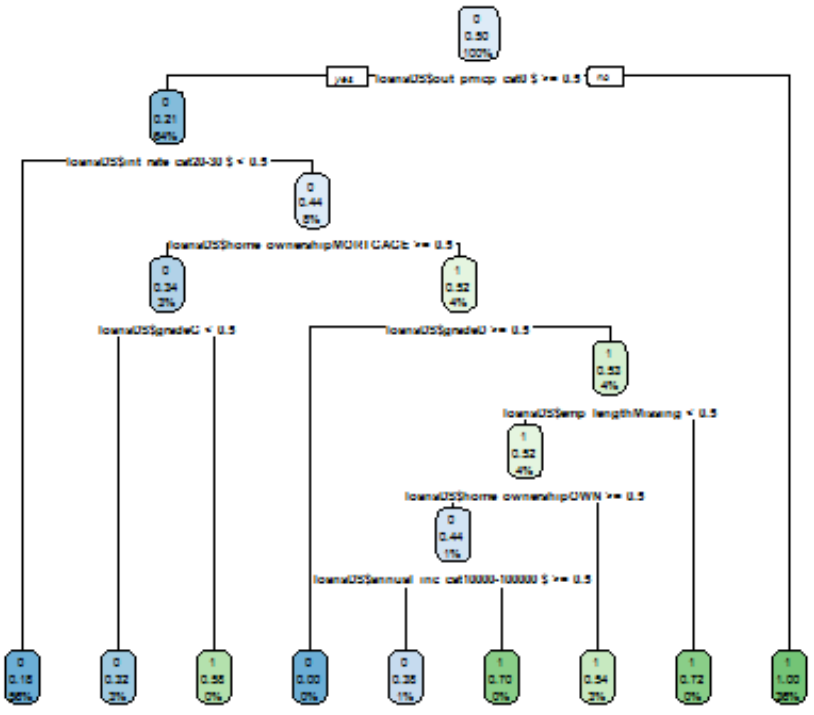
```
Model 1: defaulted ~ log_annual_income + term + int_rate + installment +
home_ownership + purpose_bon + inq_last_6mths + open_acc +
revol_util + tot_cur_bal + emp_length_new + delinq_2yrs +
total_acc + out_prncp_cat + total_rec_int + grade_ord
Model 2: defaulted ~ log_annual_income + term + int_rate + installment +
home_ownership + purpose + dti + inq_last_6mths + open_acc +
revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths +
emp_length_new + delinq_2yrs + total_acc + out_prncp_cat +
total_rec_int + verification_status + grade_ord
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     32993      16932
2     32980      16900 13   32.581 0.001969 **
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
> # odbacuje se nulta hipoteza
>
> # zamena sa emp_length_bon
> modLog8 = glm(defaulted ~ log_annual_income + term + int_rate + installment
+               + home_ownership + purpose + inq_last_6mths + open_acc
+               + revol_util + tot_cur_bal
+               + emp_length_bon + delinq_2yrs + total_acc
+               + out_prncp_cat + total_rec_int + grade_ord
+               , data=train, family="binomial")
> anova(modLog8, modLog, test ="Chisq")
Analysis of Deviance Table

Model 1: defaulted ~ log_annual_income + term + int_rate + installment +
home_ownership + purpose + inq_last_6mths + open_acc + revol_util +
tot_cur_bal + emp_length_bon + delinq_2yrs + total_acc +
out_prncp_cat + total_rec_int + grade_ord
Model 2: defaulted ~ log_annual_income + term + int_rate + installment +
home_ownership + purpose + dti + inq_last_6mths + open_acc +
revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths +
emp_length_new + delinq_2yrs + total_acc + out_prncp_cat +
total_rec_int + verification_status + grade_ord
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     32989      16908
2     32980      16900  9   8.6232   0.4728
> # prihvara se nulta hipoteza
> # testiranje znacajnosti prediktora emp_length_bon u odnosu na prediktor
> # emp_length_new
> modLog_b = glm(defaulted ~ emp_length_bon
+               , data=train, family="binomial")
> modLog_n = glm(defaulted ~ emp_length_new
+               , data=train, family="binomial")
> anova(modLog_b,modLog_n,test="Chisq")
Analysis of Deviance Table

Model 1: defaulted ~ emp_length_bon
Model 2: defaulted ~ emp_length_new
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     33024      42869
2     33020      42843  4   25.789 3.49e-05 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
> # odbacuje se nulta hipoteza
> # zamena sa grade_bon
> modLog9 = glm(defaulted ~ log_annual_income + term + int_rate + installment
+               + home_ownership + purpose + inq_last_6mths + open_acc
+               + revol_util + tot_cur_bal
+               + emp_length_bon + delinq_2yrs + total_acc
+               + out_prncp_cat + total_rec_int + grade_bon
+               , data=train, family="binomial")
> anova(modLog9, modLog, test ="Chisq")
Analysis of Deviance Table

Model 1: defaulted ~ log_annual_income + term + int_rate + installment +
home_ownership + purpose + inq_last_6mths + open_acc + revol_util +
tot_cur_bal + emp_length_bon + delinq_2yrs + total_acc +
out_prncp_cat + total_rec_int + grade_bon
Model 2: defaulted ~ log_annual_income + term + int_rate + installment +
home_ownership + purpose + dti + inq_last_6mths + open_acc +
revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths +
emp_length_new + delinq_2yrs + total_acc + out_prncp_cat +
total_rec_int + verification_status + grade_ord
Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     32994      17222
2     32980      16900 14   321.76 < 2.2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
> # odbacuje se nulta hipoteza
>
> # zamena sa out_prncp_bon
```

```
> modLog10 = glm(defaulted ~ log_annual_income + term + int_rate + installment
+                + home_ownership + purpose + inq_last_6mths + open_acc
+                + revol_util + tot_cur_bal
+                + emp_length_bon + delinq_2yrs + total_acc
+                + out_prncp_bon + total_rec_int + grade_ord
+                , data=train, family="binomial")
> anova(modLog10, modLog, test ="Chisq")
Analysis of Deviance Table

Model 1: defaulted ~ log_annual_income + term + int_rate + installment +
home_ownership + purpose + inq_last_6mths + open_acc + revol_util +
tot_cur_bal + emp_length_bon + delinq_2yrs + total_acc +
out_prncp_bon + total_rec_int + grade_ord
Model 2: defaulted ~ log_annual_income + term + int_rate + installment +
home_ownership + purpose + dti + inq_last_6mths + open_acc +
revol_bal + revol_util + tot_cur_bal + chargeoff_within_12_mths +
emp_length_new + delinq_2yrs + total_acc + out_prncp_cat +
total_rec_int + verification_status + grade_ord
Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     32991      22462
2     32980      16900 11   5561.8 < 2.2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
> # odbacuje se nulta hipoteza
```

Прилог Д

Стабло одлучивања

Д.1 Модел стабло одлука



Слика Д.1: Иницијлално стабло одлука

Д.2 Детаљне статистике чворова стабла одлуке

```
> summary(rpart_tree)
Call:
rpart(formula = safe ~ ., data = train_data, method = "class",
cp = 0.001)
n= 16754

CP nsplit rel error    xerror        xstd
1 0.721618718      0 1.0000000 1.0075206 0.007725533
2 0.003402173      1 0.2783813 0.2783813 0.005348466
3 0.001432494      3 0.2715769 0.2746807 0.005318505
4 0.001000000      6 0.2672795 0.2718157 0.005295086

Variable importance
loansDS$out_prncp_cat0 $  loansDS$out_prncp_cat1000-10000 $ loansDS$out_prncp_cat10000-20000 $
48                                19                                18
loansDS$out_prncp_cat> 20000 $        loansDS$int_rate_cat20-30 $
loansDS$gradeF
11                                 2                                 1
loansDS$gradeE
1

Node number 1: 16754 observations,    complexity param=0.7216187
predicted class=0  expected loss=0.5  P(node) =1
class counts:  8377  8377
probabilities: 0.500 0.500
left son=2 (10709 obs) right son=3 (6045 obs)
Primary splits:
loansDS$out_prncp_cat0 $           < 0.5 to the right, improve=4728.6360, (0 missing)
loansDS$out_prncp_cat1000-10000 $  < 0.5 to the left,  improve=1375.5020, (0 missing)
loansDS$out_prncp_cat10000-20000 $ < 0.5 to the left,  improve=1289.5240, (0 missing)
loansDS$out_prncp_cat> 20000 $     < 0.5 to the left,  improve= 767.4004, (0 missing)
loansDS$int_rate_cat5-10 $         < 0.5 to the right, improve= 528.6396, (0 missing)
Surrogate splits:
loansDS$out_prncp_cat1000-10000 $  < 0.5 to the left,  agree=0.780, adj=0.391, (0 split)
loansDS$out_prncp_cat10000-20000 $ < 0.5 to the left,  agree=0.773, adj=0.370, (0 split)
loansDS$out_prncp_cat> 20000 $     < 0.5 to the left,  agree=0.723, adj=0.233, (0 split)
loansDS$out_prncp_cat>1000 $       < 0.5 to the left,  agree=0.642, adj=0.007, (0 split)
loansDS$purposesmall_business      < 0.5 to the left,  agree=0.640, adj=0.003, (0 split)

Node number 2: 10709 observations,    complexity param=0.003402173
predicted class=0  expected loss=0.2177608  P(node) =0.6391906
class counts:  8377  2332
probabilities: 0.782 0.218
left son=4 (9350 obs) right son=5 (1359 obs)
Primary splits:
loansDS$int_rate_cat20-30 $ < 0.5 to the left,  improve=180.30630, (0 missing)
loansDS$int_rate_cat5-10 $  < 0.5 to the right, improve=152.94560, (0 missing)
loansDS$gradeA              < 0.5 to the right, improve= 84.62551, (0 missing)
loansDS$gradeB              < 0.5 to the right, improve= 74.43885, (0 missing)
loansDS$gradeE              < 0.5 to the left,  improve= 72.47467, (0 missing)
Surrogate splits:
loansDS$gradeF < 0.5 to the left,  agree=0.917, adj=0.348, (0 split)
loansDS$gradeE < 0.5 to the left,  agree=0.910, adj=0.287, (0 split)
loansDS$gradeG < 0.5 to the left,  agree=0.886, adj=0.104, (0 split)

Node number 3: 6045 observations
predicted class=1  expected loss=0  P(node) =0.3608094
class counts:     0  6045
probabilities: 0.000 1.000

Node number 4: 9350 observations
predicted class=0  expected loss=0.1827807  P(node) =0.5580757
class counts:  7641  1709
probabilities: 0.817 0.183

Node number 5: 1359 observations,    complexity param=0.003402173
predicted class=0  expected loss=0.4584253  P(node) =0.08111496
class counts:   736   623
probabilities: 0.542 0.458
left son=10 (608 obs) right son=11 (751 obs)
Primary splits:
loansDS$home_ownershipMORTGAGE < 0.5 to the right, improve=21.231620, (0 missing)
loansDS$home_ownershipRENT     < 0.5 to the left,  improve=15.289870, (0 missing)
loansDS$term 36 months         < 0.5 to the left,  improve= 8.534720, (0 missing)
loansDS$term 60 months         < 0.5 to the right, improve= 8.534720, (0 missing)
loansDS$emp_length10+ years    < 0.5 to the right, improve= 6.937574, (0 missing)
Surrogate splits:
```

```
loansDS$home_ownershipRENT              < 0.5 to the left,
agree=0.857, adj=0.679, (0 split)
loansDS$emp_length10+ years             < 0.5 to the right, agree=0.603, adj=0.113, (0 split)
loansDS$home_ownershipOWN               < 0.5 to the left,
agree=0.591, adj=0.086, (0 split)
loansDS$revol_bal_cat20000-50000 $      < 0.5 to the right, agree=0.582, adj=0.066, (0 split)
loansDS$annual_inc_cat100000-1000000 $ < 0.5 to the right, agree=0.578, adj=0.056, (0 split)

Node number 10: 608 observations
predicted class=0  expected loss=0.3601974  P(node) =0.03628984
class counts:   389   219
probabilities: 0.640 0.360

Node number 11: 751 observations,    complexity param=0.001432494
predicted class=1  expected loss=0.4620506  P(node) =0.04482512
class counts:   347   404
probabilities: 0.462 0.538
left son=22 (7 obs) right son=23 (744 obs)
Primary splits:
loansDS$gradeD                  < 0.5 to the right, improve=4.089572, (0 missing)
loansDS$gradeE                  < 0.5 to the right, improve=3.738918, (0 missing)
loansDS$revol_bal_cat< 5000 $   < 0.5 to the left,  improve=3.233276, (0 missing)
loansDS$gradeF                  < 0.5 to the left,  improve=2.283683, (0 missing)
loansDS$installment_cat50-100 $ < 0.5 to the right, improve=2.205932, (0 missing)

Node number 22: 7 observations
predicted class=0  expected loss=0  P(node) =0.0004178107
class counts:     7     0
probabilities: 1.000 0.000

Node number 23: 744 observations,    complexity param=0.001432494
predicted class=1  expected loss=0.4569892  P(node) =0.04440731
class counts:   340   404
probabilities: 0.457 0.543
left son=46 (388 obs) right son=47 (356 obs)
Primary splits:
loansDS$gradeE                < 0.5 to the right, improve=4.610685, (0 missing)
loansDS$revol_bal_cat< 5000 $ < 0.5 to the left,  improve=3.241338, (0 missing)
loansDS$purposemajor_purchase < 0.5 to the right, improve=1.900147, (0 missing)
loansDS$gradeF                < 0.5 to the left,  improve=1.877615, (0 missing)
loansDS$emp_length< 1 year    < 0.5 to the left,  improve=1.859323, (0 missing)
Surrogate splits:
loansDS$gradeF          < 0.5 to the left,  agree=0.876, adj=0.742, (0 split)
loansDS$gradeG          < 0.5 to the left,  agree=0.645, adj=0.258, (0 split)
loansDS$term 36 months < 0.5 to the right, agree=0.558, adj=0.076, (0 split)
loansDS$term 60 months < 0.5 to the left,  agree=0.558, adj=0.076, (0 split)
loansDS$emp_length1-3  < 0.5 to the left,  agree=0.556, adj=0.073, (0 split)

Node number 46: 388 observations,    complexity param=0.001432494
predicted class=0  expected loss=0.4896907  P(node) =0.02315865
class counts:   198   190
probabilities: 0.510 0.490
left son=92 (215 obs) right son=93 (173 obs)
Primary splits:
loansDS$term 36 months      < 0.5 to the left,  improve=3.147906, (0 missing)
loansDS$term 60 months      < 0.5 to the right, improve=3.147906, (0 missing)
loansDS$emp_length10+ years < 0.5 to the right, improve=2.526974, (0 missing)
loansDS$emp_lengthMissing   < 0.5 to the left,  improve=2.292238, (0 missing)
loansDS$emp_length< 1 year  < 0.5 to the left,  improve=1.913308, (0 missing)
Surrogate splits:
loansDS$term 60 months            < 0.5 to the right, agree=1.000, adj=1.000, (0 split)
loansDS$loan_amnt_cat< 10000 $    < 0.5 to the left,  agree=0.799, adj=0.549, (0 split)
loansDS$installment_cat500-1000 $ < 0.5 to the right, agree=0.668, adj=0.254, (0 split)
loansDS$revol_bal_cat< 5000 $     < 0.5 to the left,  agree=0.647, adj=0.208, (0 split)
loansDS$installment_cat100-500 $  < 0.5 to the left,  agree=0.606, adj=0.116, (0 split)

Node number 47: 356 observations
predicted class=1  expected loss=0.3988764  P(node) =0.02124866
class counts:   142   214
probabilities: 0.399 0.601

Node number 92: 215 observations
predicted class=0  expected loss=0.4325581  P(node) =0.01283276
class counts:   122    93
probabilities: 0.567 0.433

Node number 93: 173 observations
predicted class=1  expected loss=0.4393064  P(node) =0.01032589
class counts:    76    97
probabilities: 0.439 0.561
```

Биографија

Ана Томић рођена је 30. априла 1991. године у Ћуприји, живи у Параћину, где 2006. године завршава основну школу као носилац Вукове дипломе. Након завршене основне школе, уписује Гимназију у Параћину, коју завршава 2010. године. Исте године уписује основне академске студије на Математичком факултету Универзитета у Београду, смер Статистика, финансијка и актуарска математика. Основне студије завршава 2014. године звањем "Дипломирани математичар" и просечном оценом 9,02. Исте године уписује мастер академске студије на истом смеру, Математичког факултета у Београду.

Радно искуство запичиње јануара 2015. године у компанији "GoPro, d.o.o." на позицији "Програмер пословних решења", где се сусреће са нових технологијама, упознавање различитих пословних процеса, тимски рад на различитим пројектима, вођење пројеката. Након годину дана, радно искуство наставља на позицији "Пројектант система пословне интелигенције" где се додатно усавршава у познавању процеса, изради прилагођених решења у конкретним компанијама, и додатним изазовима. Крајем октобра 2016. године радно искуство наставља у компанији "Phoenix pharma, d.o.o." на истој позицији, где се сусреће са новим пословним процесима приликом рада на пројектима.

Течно говори енглески језик, а служи се руским. Интересовања су јој читање класика, путовања и учење страних језика.