



VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
FUNDAMENTINIŲ MOKSLŲ FAKULTETAS
MATEMATINĖS STATISTIKOS KATEDRA

Aurelija Jankauskė

**KREDITO ĮSIPAREIGOJIMŲ NEVYKDYMO
MODELIAVIMAS IR PROGNOZAVIMAS**
Credit default modeling and prediction

Baigiamasis magistro darbas

Taikomosios statistikos studijų programa, valstybinis kodas 6211AX009
Statistinių metodų finansuose ir ekonomikoje specializacija
Statistikos studijų kryptis

Vilnius, 2019

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
FUNDAMENTINIŲ MOKSLŲ FAKULTETAS
MATEMATINĖS STATISTIKOS KATEDRA

TVIRTINU
Katedros vedėjas

(Parašas)

doc. dr. Kazimieras Padvelskis

(Vardas, pavardė)

(Data)

Aurelija Jankauskė

**KREDITO ĮSIPAREIGOJIMŲ NEVYKDYMO
MODELIAVIMAS IR PROGNOZAVIMAS**
Credit default modeling and prediction

Baigiamasis magistro darbas

Taikomosios statistikos studijų programa, valstybinis kodas 6211AX009
Statistinių metodų finansuose ir ekonomikoje specializacija
Statistikos studijų kryptis

Vadovas

dr. Nomedą Bratėikovienė

(Pedagoginis vardas, vardas, pavardė)

(Parašas)

(Data)

Lietuvių kalbos konsultantas

dr. Aušra Žemienė

(Pedagoginis vardas, vardas, pavardė)

(Parašas)

(Data)

Vilnius, 2019

Vilniaus Gedimino technikos universitetas Fundamentinių mokslų fakultetas Matematinės statistikos katedra	ISBN Egz. sk. Data-.....-.....
Antrosios pakopos studijų Taikomosios statistikos programos magistro baigiamasis darbas	
Pavadinimas	Kredito įsipareigojimų nevykdymo modeliavimas ir prognozavimas
Autorius	Aurelija Jankauskė
Vadovas	Nomeda Bratčikovienė
Kalba: lietuvių	
Anotacija Finansų institucijos, prieš priimdamos sprendimą ar suteikti paskolą, vertina kredito riziką. Atsiradus tarpusavio skolinimo platformoms, rizika egzistuoja ne tik bankams, bet ir privatiems asmenims. Čia iškyla problema - ar paprastas žmogus, investuodamas tokioje platformoje, supranta riziką ir moka ją įvertinti? Analizuojant tarpusavio skolinimo platformos duomenis ir pritaikant regresinę analizę bei duomenų koreliacijas, atrenkami tinkamiausi kintamieji. Kredito įsipareigojimų nevykdymo modeliai sudaromi naudojant logistinę regresiją ir dirbtinius neuroninius tinklus. Ieškant geriausio modelio pastebima, kad naudojamas 10 metų laikotarpis yra per ilgas, ir optimaliu parenkamas 3 metų intervalas. Sudarytų modelių rezultatų tikslumas yra panašus, tačiau logistinę regresiją išrenkame geriausiu modeliu dėl jos paprastumo.	
Prasminiai žodžiai: Kredito įsipareigojimų nevykdymas, kredito rizika, tarpusavio skolinimas, logistinė regresija, neuroniniai tinklai.	

Vilnius Gediminas Technical University
Faculty of Fundamental Sciences
Department of Mathematical Statistics

ISBN ISSN
Copies No.
Date-.....-.....

Master Degree Studies **Applied Statistics** study programme Master Graduation Thesis

Title **Credit default modeling and prediction**

Author **Aurelija Jankauskė**

Academic supervisor **Nomeda Bratčikvienė**

Thesis language: Lithuanian

Annotation

Finance institutions assess credit risk before taking a decision to grant a loan. After introduction of peer-to-peer lending platforms, not only banks, but also individuals hold credit risk. The problem is - does ordinary person investing in such platform understands the risk and knows how to evaluate it? When analysing data from peer-to-peer lending platform, regression analysis and data correlations are used to select most suitable variables. Afterwards, credit default models are created using logistic regression and artificial neural networks. When looking for the best model, it is noticeable that the 10 year period is too long and the 3 year interval is chosen as an optimal one. The accuracy of both models is similar, but logistic regression is chosen as the best model because of its simplicity.

Keywords: Credit default, credit risk, peer-to-peer lending, logistic regression, neural network.

Turinys

ĮVADAS	6
1. LITERATŪROS IR MODELIAVIMO METODŲ APŽVALGA	8
1.1. Tarpusavio skolinimas	8
1.2. Kredito rizikos vertinimas	8
1.3. Metodai	10
1.3.1. Logistinė regresija	10
1.3.2. Neuroniniai tinklai	11
1.4. Modelių tikslumo įverčiai	14
1.4.1. Vidutinė kvadratinė paklaida	14
1.4.2. Plotas po ROC kreive	14
1.4.3. Somers' D įvertis	14
2. DUOMENŲ TYRYBA	16
2.1. Duomenų apžvalga	16
2.1.1. Paskolos statusas	16
2.1.2. Darbas	17
2.1.3. Pajamos	17
2.1.4. Paskolos įrašo komentaras	17
2.1.5. Skolų ir pajamų santykis	17
2.2. Kategorinių kintamųjų analizė	18
2.2.1. Gyvenamojo būsto tipas	18
2.2.2. Paskolos tikslas	19
2.2.3. Paskolos reitingas	20
2.2.4. Valstijos	21
2.3. Reikšmingų kintamųjų atrinkimas regresinei analizei	25
2.4. Modelio kintamųjų sąrašas	27
3. MODELIAVIMAS IR PROGNOZAVIMAS	28
3.1. Logistinė regresija	28
3.2. Neuroniniai tinklai	33
IŠVADOS IR REZULTATAI	41
LITERATŪRA	42
PRIEDAI	43
PRIEDAS A: REGRESINĖS ANALIZĖS REZULTATAI	43

Paveikslų sąrašas

1 pav. Taisyklėmis pagrįstos sistemos rezultatai	9
2 pav. Tiesinės regresijos rezultatai	10
3 pav. Netiesinės regresijos rezultatai	10
4 pav. Paprastas dirbtinis neuronas	12
5 pav. Paprastas dviejų sluoksnių neuroninis tinklas	12
6 pav. Bendras neurono grafinis atvaizdavimas	13
7 pav. ROC kreivės pavyzdys	14
8 pav. Paskolos tikslo suskirstymas k-vidurkių metodu	20
9 pav. Reitingo suskirstymas k-vidurkių metodu	21
10 pav. Blogų klientų dalis (didžiausia – raudona, mažiausia – žalia)	23
11 pav. Vergovė Šiaurės Amerikoje 1861 metais	23
12 pav. K-vidurkių metodo pritaikymas valstijoms	24
13 pav. Valstijų suskirtymas k-vidurkių metodu	24
14 pav. Logistinės regresijos modelis iš atrinktų kintamųjų	28
15 pav. Logistinės regresijos modelio kintamųjų VIF reikšmės	29
16 pav. ROC kreivė logistinės regresijos modeliui	29
17 pav. Logistinės regresijos modelis iš atrinktų kintamųjų	31
18 pav. ROC kreivė logistinės regresijos modeliui 2015–2017 metams (Validavimo duomenys)	32
19 pav. Neuroninių tinklų 2 lygių po 20 neuronų modelio rezultatai	33
20 pav. Neuroninių tinklų 2 lygių po 20 neuronų 10 apmokymų istorija	34
21 pav. Neuroninių tinklų 2 lygių po 20 neuronų 50 apmokymų istorija	34
22 pav. Neuroninių tinklų 2 lygių po 20 neuronų apmokymų istorija su 5 imčių kryžmine patikra	35
23 pav. Neuroninių tinklų 2 lygių po 20 neuronų apmokymų istorija su 3 imčių kryžmine patikra	35
24 pav. Neuroninio tinklo [50, 50, 50] su 50 ciklų apmokymo procesas	37
25 pav. Neuroninio tinklo [50, 50, 50] ROC kreivė	38
26 pav. Neuroninio tinklo [50, 50, 50] svarbiausi kintamieji	38
27 pav. Neuroninio tinklo 2015–2017 metų duomenims svarbiausi kintamieji	39

Lentelių sąrašas

1 lentelė. Paskolų skaičius pagal statusą	17
2 lentelė. Blogų klientų dalis pagal būsto tipą	18
3 lentelė. Blogų klientų dalis pagal paskolos tipą	19
4 lentelė. Blogų klientų dalis pagal reitingą	20
5 lentelė. Blogų klientų dalis pagal Valstiją	22
6 lentelė. Koreliacija tarp skaitinių rodiklių	25
7 lentelė. Geriausių kintamųjų sąrašas	26
8 lentelė. Logistinės regresijos rezultatai skirtingiems laiko intervalams	30
9 lentelė. Duomenų rinkinių dydžiai pagal naudojamą laikotarpį	31
10 lentelė. 2015–2017 metų logistinės regresijos modelio rezultatai testavimo ir validavimo imtims	32
11 lentelė. Neuroninių tinklų modeliai 2 lygių po 20 neuronų	36
12 lentelė. Skirtingų neuroninių modelių rezultatai, pirmieji bandymai	36
13 lentelė. Skirtingų neuroninių modelių rezultatai, antrieji bandymai	37
14 lentelė. Neuroninių tinklų rezultatai skirtingiems laiko intervalams	39
15 lentelė. 2015–2017 metų neuroninių tinklų modelio rezultatai testavimo ir validavimo imtims	40
16 lentelė. Geriausi logistinės regresijos ir neuroninių tinklų modeliai	40
17 lentelė. Kintamųjų sąrašas su regresinės analizės rezultatais	43
18 lentelė. Labiausiai koreliuoti kintamieji	44

ĮVADAS

Nuo civilizacijos pradžios buvo praktikuojamas skolinimas su iš anksto sutarta nauda skolintojui, kuri atitiko dabartines palūkanas. Mesopotamijoje, 2000 metų prieš Kristų, ūkininkai skolindavo grūdus žinodami, kad pasėjus jų užauga daugiau. Pradėjus naudoti sidabrą kaip atsiskaitymo priemonę buvo įvestas Hammurabi kodas, kuriame apibrėžta sidabro kaina ir jo palūkanos, o 400 metų prieš Kristų įvestas užstatas, dėl kurio sumažinama rizika (Amery, 2018).

Einant laikui skolinimą tarp turtingųjų ir vargšų pakeitė finansų institucijos. Joms greičiau ir lengvinant kredito išdavimo procesus, skolinimasis tapo toks dažnas ir įprastas, kad retas žmogus neturi jokių finansinių įsipareigojimų. Tai gali būti būsto paskola, automobilio ilgalaikė nuoma, greitas kreditas ar tiesiog kreditinė kortelė su limitu. Pirkti daiktus neturint jiems pinigų tapo taip lengva, kad net būtina technika perkama su išsimokėjimu. Tai parodo finansinių įmonių prisitaikymą prie žmogaus noro turėti viską čia ir dabar. Renkantis tarp ilgo taupymo ir paskolos, didelė dalis pasirenka antrą variantą, tačiau ne kiekvienam norinčiam turėtų būti suteikiamas kreditas. Skolinant pinigus yra rizika, kad skolininkas nesugebės jų grąžinti. Šioje srityje bankai yra sukaupę didelę patirtį ir prieš priimdami sprendimus ar duoti paskolą, vertina kredito riziką.

Daugėjant finansinių produktų paklausai, atsiranda alternatyvių būdų skolintis. Vienas iš jų – tarpusavio skolinimo platformos. Šiose platformose skolintojai yra ne bankai, o privatūs asmenys, norintys investuoti savo pinigus. Yra net kelios priežastys, kodėl šios platformos yra sėkmingos. Pirma, bankų sukelta 2008 metų ekonominė krizė sumažino žmonių pasitikėjimą finansinėmis įmonėmis. Tokiu atveju, tarpusavio skolinimosi platformos organizuodamos paskolas tarp privačių asmenų atrodo kaip vienintelis galimas paskolos šaltinis. Antra, bankai dažnai yra konservatyvūs priimant sprendimus, todėl dalis prašančių paskolos, jos negavę ieško kitų alternatyvų. Trečia, tarpusavio skolinimo platformoje palūkanos yra mažesnės, nei greitųjų kreditų įmonėse.

Naudojant tarpusavio skolinimo platformas kredito rizika pereina nuo bankų ant žmonių. Čia iškyla problema – ar paprastas žmogus, investuodamas tokioje platformoje, supranta riziką ir moka ją įvertinti? Kaip sumažinti riziką iš duomenų gaunamų tarpusavio skolinimo platformoje?

Šiame darbe yra sujungiamos dvi temos: investavimas tarpusavio skolinimo platformose ir rizikos valdymas prognozuojant kredito įsipareigojimų nevykdymą. Pasitelkiant bankiniuose sektoriuose naudojamus statistinius metodus, sukuriama modeliai skaičiuojantys tikimybę, kad paskola nebus grąžinta laiku. **Darbo tikslas** yra identifikuoti kintamuosius, turinčius stipriausias sąsajas su kredito įsipareigojimų nevykdymu ir palyginti logistinės regresijos ir dirbtinių neuroninių tinklų tikslumą. Tyrimo metu sprendžiami šie uždaviniai:

- Atliekama tarpusavio skolinimo platformos paskolų rinkinio duomenų analizė ir identifikuojama reikšminga informacija.
- Sukuriamas logistinės regresijos modelis kredito rizikos vertinimui.
- Sukuriamas dirbtinių neuroninių tinklų modelis kredito rizikos vertinimui.
- Atrenkamas geriausias modelis.

Darbas sudarytas iš teorinės dalies, duomenų tyrybos ir modeliavimo. Teorinėje dalyje apžvelgiami prognozavimui naudojami modeliai, o duomenų tyryboje nagrinėjami analizei naudojami Jungtinių Amerikos Valstijų tarpusavio skolinimo įmonės *Lending Club* duomenys. Pasitelkiant loginį mąstymą, regresinę analizę ir duomenų koreliaciją, atrenkami tinkamiausi

kintamieji modeliams kurti. Modeliuojant duomenis naudojama logistinė regresija ir dirbtiniai neuroniniai tinklai. Abiem metodais ieškoma tiksliau skaičiuojančių modelių, o geriausi logistinės regresijos ir dirbtinių neuroninių tinklų modeliai yra lyginami tarpusavyje.

1. LITERATŪROS IR MODELIAVIMO METODŲ APŽVALGA

1.1. Tarpusavio skolinimas

Tarpusavio skolinimas yra paskolų teikimo būdas, kuriame privatūs asmenys skolina pinigus nepažįstamiems asmenims per tam skirtą platformą. Šio finansavimo tipo išskirtinis bruožas yra bankų nedalyvavimas procese. Pirmoji tokio tipo įmonė buvo įkurta 2005 metais Jungtinėje Karalystėje, o apie 2009 metus šis finansavimo tipas plačiai paplito po visą pasaulį.

Lietuvoje taip pat veikia tarpusavio skolinimo platformos. Viena jų yra 2010 metais įkurta įmonė „SAVY“. Pagal šios įmonės duomenis, ji turi virš 24 tūkst. investuotojų, o skolininkams teikia apie 24,4 % vidutinę metinę palūkanų normą. Nors tokia palūkanų norma yra gan didelė lyginant su bankų palūkanų norma (apie 10 %), tačiau greitųjų kreditų įmonėse metinė palūkanų norma gali siekti net 70 %. Pradedant investuoti per „SAVY“ tarpusavio skolinimo platformą, investuotojas gali pats pasirinkti į kokio rizikingumo paskolas nori investuoti, kai tikėtinos palūkanos yra atitinkamai nuo 8 % iki 75 %.

Tarpusavio skolinimo platformų naudotojai yra dviejų tipų – skolininkai ir skolintojai. Dažna priežastis imti paskolą per tarpusavio skolinimo platformą yra bankų neprieinamumas, kai tuo tarpu skolintojai yra žmonės norintys investuoti pinigus su didesne investicine grąža. Norintys pasiskolinti žmonės dalinasi asmenine ir finansine informacija su potencialiais skolintojais, o šie gali patys nuspręsti ar įsitraukti į konkretų sandorį. Pagal taisykles, vienas investuotojas gali prie sandorio prisidėti tik maža dalimi, o visa paskola susideda iš daug investuotojų (Galloway, 2009/2010).

Bankai priimdami sprendimus ar duoti klientui paskola ar ne, naudoja statistinius modelius rizikos vertinimui, tuo tarpu investuotojai tarpusavio skolinimo platformoje turi tik konkretaus asmens duomenis ir jokių įrankių rizikai prognozuoti. Šiuo atveju, alternatyva galėtų būti asmeninė informacija ir socialiniai tinklai. Tarkime, jeigu kas nors iš pažįstamų žmonių gali laiduoti dėl norinčio pasiskolinti patikimumo, tuomet rizika sumažėja (Lin, 2009).

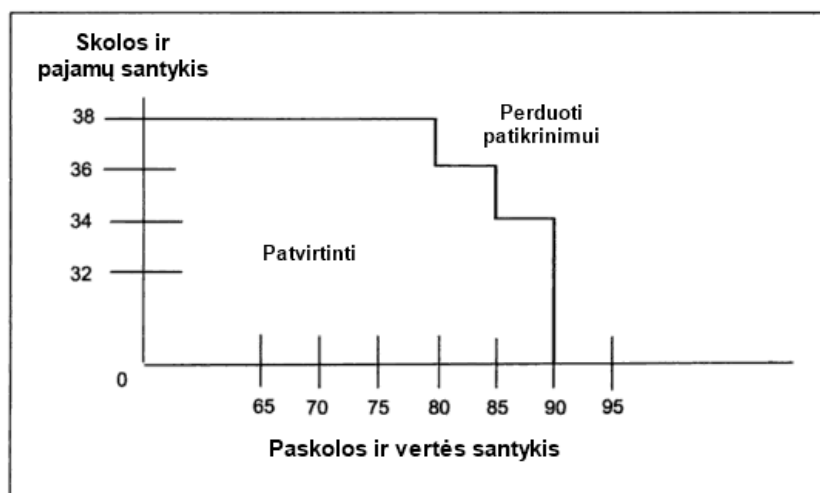
Kredito riziką galima paskaičiuoti naudojant panašius rodiklius į naudojamus bankuose, tačiau tyrimai rodo, kad investuotojai atsižvelgia ir į paprastesnius dalykus kaip žmogaus išvaizda. Taip pat pastebėta, kad priklausomai nuo kredito reitingo, atsižvelgiama į skirtingą informaciją priimant sprendimą dėl paskolos suteikimo. Pavyzdžiui, jeigu žmogus turi aukštesnį kredito reitingą, tuomet yra stebimas šio skolos ir pajamų santykis, kai tuo tarpu žemą reitingą turintiems skolininkams tikslesnę informaciją suteikia detalės apie įmokų vėlavimus (Iyer, Khwaja, Luttmer ir Shue, 2010).

1.2. Kredito rizikos vertinimas

Bankų teikiami produktai, pavyzdžiui kreditinės kortelės, vartojimo paskolos ir būsto paskolos, yra įprasta kiekvienam vartotojui, tačiau ne tik vartotojai renkasi banką, bet ir bankai renkasi klientus. Šioje dalyje svarbų vaidmenį turi kredito rizika. Lengviausias ir visiems bankams prieinamas būdas įvertinti riziką yra gauti kliento kredito reitingą kreipiantis į kreditų biurą. Kreditų biuro pavyzdys gali būti „FICO“ – įmonė įsikūrusi Jungtinėse Amerikos Valstijose, arba „Credit Info“ veikianti Lietuvoje. Jie turi informaciją, apie visus finansinius įsipareigojimus nepriklausomai nuo finansų institucijos, todėl gali pateikti tikslų kiekvieno asmens reitingą. Nepaisant to, kredito riziką gali skaičiuoti ir patys bankai. Šio tipo rizika yra labai svarbi stengiantis išvengti klientų, vėluojančių grąžinti paskolą arba jos išvis negrąžinančių. Kaip ir kitų rizikingose srityse dirbančių įmonių, taip ir bankų tikslas yra pasiekti kuo didesnę pelną prisiimant kuo mažesnę riziką.

Nors kreditų biurai neturi duomenų apie konkrečią paskolą, asmens pajamos, turimos skolos santykinė išraiška ar net ekonominiai duomenys gali būti naudingi prognozuojant kredito riziką. Nors kredito biurai turi visų egzistuojančių paskolų informaciją, bankai yra pranašesni turėdami kliento mokėjimų istoriją. William M. Makuch vardiya skirtingas reitingų rūšis, iš kurių verta paminėti elgsenos reitingą (angl. „Behavioral score“) ir paraiškos reitingą (angl. „Application score“). Paraiškos reitingas yra naudojamas sprendžiant paraiškos patvirtinimo klausimą ir nustatant įkainius, nes jame naudojami būtent iš paraiškos gaunami duomenys, o elgsenos reitingas yra skaičiuojamas pasinaudojant turima istorine kliento informacija. Elgsenos reitingo tikslumas aplenkia paraiškos reitingą praėjus vidutiniškai dvylikai mėnesių po kredito suteikimo (Makuch, 2000).

Kuriant kredito reitingavimo modelius svarbu ne tik koki statistikinį metodą naudoti, bet ir kintamųjų ar duomenų rinkinio dydžio pasirinkimas. Pasak John M. L. Gruenstein, yra keturi pagrindiniai modelio variantai – taisyklėmis pagrįsta sistema, tiesinė regresija, netiesinė regresija ir dirbtiniai neuroniniai tinklai. Taisyklėmis pagrįsta sistema yra sudaryta iš viena kitą papildančių taisyklių. Kuo daugiau taisyklių, tuo daugiau skirtingos informacijos apimama, tuo lankstesnė yra sistema. Tokio modelio priimtų sprendimų iliustracija yra pavaizduota 1 paveiksle. Atsižvelgiant į bendros skolos ir pajamų santykį bei paskolos ir užstatomo turto vertės santykį, nusprendžiama ar paskolą galima patvirtinti.



1 pav. Taisyklėmis pagrįstos sistemos rezultatai

Taisyklėmis pagrįstos sistemos rezultatus lengva paaiškinti ir modelis gali būti gana lankstus pridėjus daugiau taisyklių. Kuo sudėtingesnės taisyklės, tuo labiau „kampuota“ būtų linija tarp patvirtintų ir perduotų patikrinimui atsakymų. Vis dėlto, jeigu ribą tarp patvirtintų ir atmestų paraiškų siekiama ištiesinti, paprasčiau tiesiog naudoti tiesinę regresiją.

Tokio modelio lygtis:

$$P = a + b * LTV + c * DTI$$

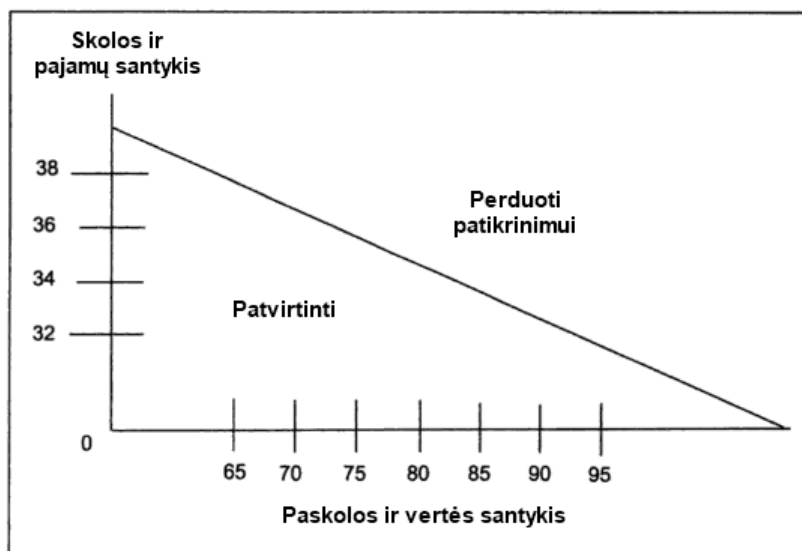
Čia LTV – paskola vertei;

DTI – skolos ir pajamų santykis;

a, b, c – modelio koeficientai.

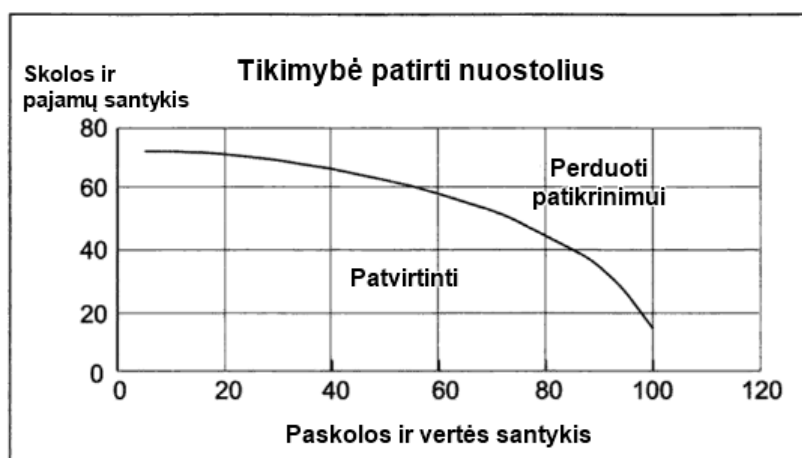
Kaip ir taisyklėmis pagrįstoje sistemoje, taip ir tiesinėje regresijoje kuo daugiau kintamųjų įtraukiama, tuo tikslesnius rezultatus galima gauti, tokiu atveju linija tarp patvirtintų ir atmestų paraiškų yra tiesė (2 pav.).

Vis dėlto sunku įsivaizduoti, kad tiesė atitiktų situaciją realybėje, todėl svarstyтина ir netiesinė regresija. Netiesinė regresija gali įgyti bet kokios kreivinės funkcijos formą, kuriai galima



2 pav. Tiesinės regresijos rezultatai

parašyti funkciją. Iš 3 paveikslo matome, kad šio modelio duoti rezultatai patvirtinti ar atmesti paskolą yra tarp taisyklėmis pagrįstos sistemos ir tiesinės regresijos.



3 pav. Netiesinės regresijos rezultatai

Kredito rizikos modeliuose kaip netiesinė regresija dažnai naudojama logistinė regresija.

Dar vienas minėtas būdas prognozuoti kredito riziką yra naudojant neuroninius tinklus. Šis metodas buvo įkvėptas jungčių tarp žmogaus smegenyse esančių neuronų. Vienas neuronas statistiniame modelyje paima svorinę kintamųjų sumą ir pritaikęs aktyvavimo funkciją, transformuoja ją į skaičių intervale tarp nulio ir vienetų. Neuroninis tinklas turi vieną ar daugiau neuronų sluoksnių, kur išvestys iš vieno sluoksnio yra įvestys kitam sluoksniui. Šis statistinis metodas yra ypatingai lankstus, tačiau gali tapti gana sudėtingas didėjant neuronų sluoksnių skaičiui (Gruenstein, 2001).

1.3. Metodai

1.3.1. Logistinė regresija

Logistinė regresija kitaip dar vadinama logistiniu modeliu analizuoja ryšį tarp nepriklausomų kintamųjų ir priklausomo kintamojo. Yra dvi logistinės regresijos rūšys – dvinarė ir daugianarė. Dvinarė logistinė regresija naudojama tuomet, kai priklausomą kintamąjį galima

suskirstyti į dvi kategorijas. Kai kategorijų yra daugiau nei dvi, naudojama daugianarė logistinė regresija (Park, 2013). Logistinės regresijos patogumas yra tai, kad gautos vertės atitinka tikimybės apibrėžimą ir yra galimybė prognozuoti dvinarius priklausomus kintamuosius.

Logistinės regresijos pagrindą sudaro tikimybių santykio natūrinis logaritmas. Tarkime, jog tikimybė, kad asmuo negrąžins paskolos yra lygi p , o grąžinimo tikimybė $1 - p$. Tuomet logistinį modelį galima apibrėžti sekančia lygybe.

$$\text{logit}(Y) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta X \quad (1)$$

Čia Y yra prognozuojamas įvykis, X – nepriklausomas kintamasis, p – tikimybė, kad įvykis Y įvyks, o α ir β yra modelio koeficientai.

Iš šios lygybės galime išvesti tikimybę, kad įvykis Y įvyks. Mūsų atveju, tai būtų tikimybė, kad asmuo negrąžins paskolos.

$$p = P(Y = 1|X = x) = \frac{e^{\alpha + \beta X}}{1 + e^{\alpha + \beta X}} \quad (2)$$

Identifikuojant ar aiškinamasis kintamasis yra reikšmingas yra naudojama β koeficiento vertė ir nulinė hipotezė. Ši hipotezė teigia, kad β lygi nuliui. Tokiu atveju sakoma, kad aiškinamasis kintamasis prie β koeficiento yra nereikšmingas. Atmetant nulinę hipotezę leidžiama manyti, kad ryšys tarp X ir natūrinio logaritmo iš Y egzistuoja (Peng, Lee ir Ingersoll, 2002).

Pridedant daugiau kintamųjų į lygtį gautume:

$$\text{logit}(Y) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \quad (3)$$

$$p = P(Y = 1|X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) = \frac{e^{\alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k}}{1 + e^{\alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k}} \quad (4)$$

Dėl naudojamos netiesinės logistinės transformacijos, logistinė regresija veikia su netiesiniais ryšiais tarp priklausomojo ir aiškinančiųjų kintamųjų. Be kita ko, nepriklausomi kintamieji gali būti net tik skaitiniai, bet ir kategoriniai, jeigu jų reikšmės pateikiamos kaip atskiri kintamieji (Park, 2013).

Vis dėlto yra ir keletas reikalavimų, kurie turi būti išpildyti. Pirma, priklausomas kintamasis turi būti diskretus. Dažniausiai naudojami priklausomi kintamieji gali įgyti tik dvi reikšmes. Antra, kadangi logistinė regresija skaičiuoja tikimybę $P(Y = 1)$, prognozuojamas kintamasis turi įgyti reikšmes 0 ir 1. Trečia, modelį turi sudaryti tik reikšmingi kintamieji. Ketvirta, modelis turi neturėti multikolinerumo. Galiausiai, nors nepriklausomi kintamieji neturi būti tiesiškai susiję su priklausomu kintamuoju, tačiau jie turi būti tiesiškai susiję su įvykio tikimybių santykio logaritmu (Park, 2013).

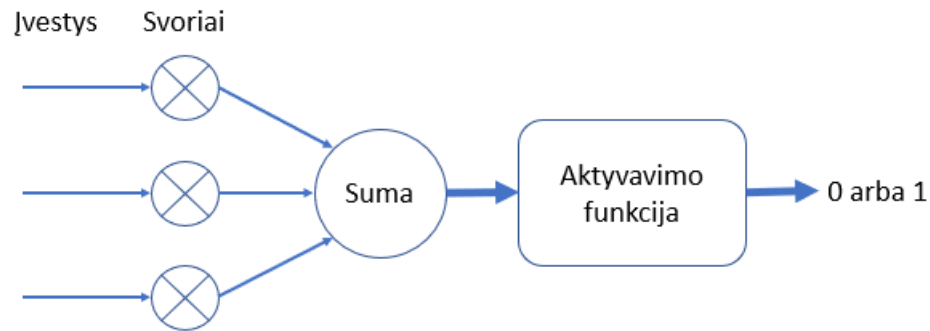
1.3.2. Neuroniniai tinklai

Kompiuterinės technologijos palengvina žmonių gyvenimą daugumoje aspektų, tačiau žmogus vis dar pirmauja savo gebėjimais. Žinoma, atliekant sudėtingus matematinius skaičiavimus kompiuteris greičiau atlieka matematinius skaičiavimus nei žmogus, tačiau formų ar rašto atpažinimas visai kas kita. Tokia užduotis yra lengva pradinųjų klasių mokiniui, tačiau yra sunki, o kartais net ir neįmanoma kompiuteriui.

Dirbtinių neuroninių tinklų idėja kilo bandant atkartoti smegenų veikimo principus. Pasak Simon Haykin, dirbtinis neuroninis tinklas yra milžiniškas lygiagrečiai paskirstytas procesorius, sudarytas iš mažų nesudėtingų apdorojimo elementų ir turintis polinkį kaupti žinias ir paruošti jas naudojimui. Dirbtinis neuroninis tinklas panašus į smegenis tuo, kad informacija įsisavinama

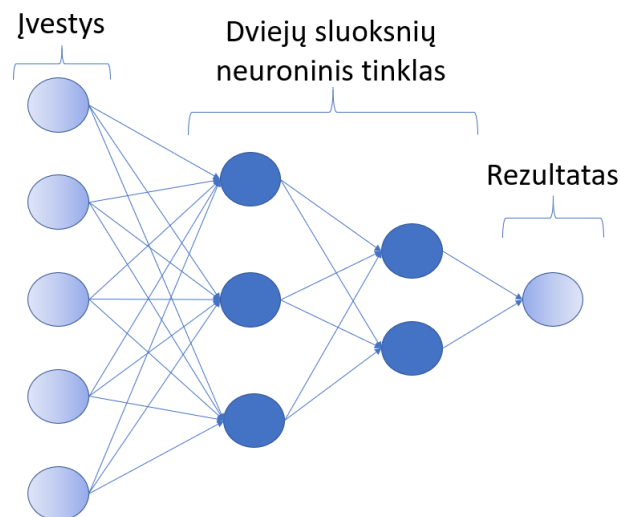
iš aplinkos mokymosi procesu, o ryšiai tarp neuronų yra naudojami informacijai kaupti (Haykin, 2009).

Modelio įvestys yra sujungtos su neuronais, o šie apdoroja gautą informaciją. Paprastas dirbtinio neurono pavyzdys yra pateiktas 4 paveiksle. Modelio įvestims yra priskiriami svoriai, o neuronas sumuoja gautas vertes. Tuomet aktyvavimo funkcija įvertina gauto dydžio reikšmingumą ir perduoda rezultatą. Paprasčiausia ir seniausia naudojama aktyvavimo funkcija lygina gautą vertę su ribine verte ir duoda rezultatą nulis arba vienas. Angliškai tokio pobūdžio aktyvavimo funkcija vadinama TLU – Threshold Logic Unit (Gurney, 1997).



4 pav. Paprastas dirbtinis neuronas

Dirbtinis neuroninis tinklas gali būti sudarytas iš daug tokių neuronų ir jų sluoksnių. Neuronai, apdoroję informaciją, siunčia rezultatus kitame sluoksnyje esantiems neuronams pakartotinai pritaikant svorius, sumavimą ir aktyvavimo funkciją. Ši procedūra yra kartojama tol kol pereinami visi neuronų sluoksniai ir pateikiamas rezultatas. Taip veikiantys neuroniniai tinklai angliškai vadinami „Feedforward neural networks“ ir yra dažniausiai naudojami. Paprastas tokio tinklo pavyzdys yra penktame paveiksle. Jis atvaizduoja neuroninį tinklą su 5 aiškinamaisiais kintamaisiais, 2 neuronų sluoksniais po 3 ir 2 neuronus ir viena rezultato reikšmę.

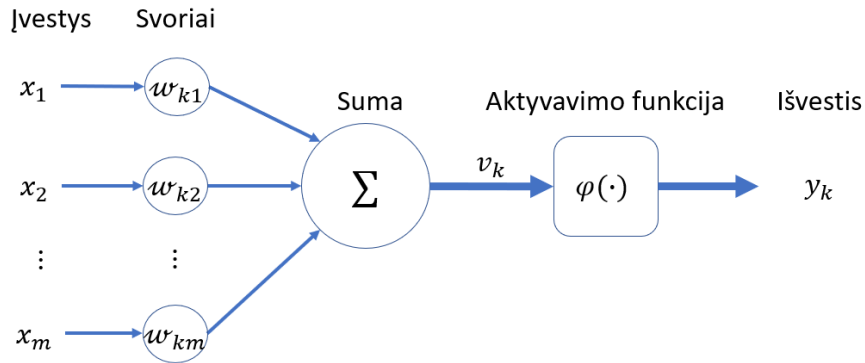


5 pav. Paprastas dviejų sluoksnių neuroninis tinklas

Kita dirbtinių neuroninių tinklų rūšis prieš pateikiant rezultatus, patikrina ar jie yra tikslūs. Jei nustatoma, kad rezultatai neatitinka reikalavimų, modelis paleidžiamas dar kartą išlaikant įgytą informaciją. Tai kartojama tol kol rezultatai atitinka reikalavimus. Tokie neuroniniai tinklai vadinami pasikartojančiais (angl. „Recurrent neural networks“).

Bet kurio neuroninio tinklo apmokymas vyksta koreguojant svoriams priskiriamas vertes. Naudojant apmokymo duomenų rinkinį, neuroninis tinklas apdoroja įvestis ir palygina gauta rezultata su tikrąja reikšme. Jeigu atsakymai nesutampa, svoriai koreguojami. Procedūra kartojama kitoms įvestims tol kol pereinamas visas apmokymo duomenų rinkinys. Pasikartojančių neuroninių tinklų atveju, šis ciklas gali būti kartojamas, jei rezultatai nėra pakankamai tikslūs.

Žemiau atvaizduotas dirbtinis neuronas yra k-tojo neurono pavyzdys. Įvestys gali būti modelio kintamieji arba ankstesniame sluoksnyje buvusių neuronų išvestys. Čia $x_1, x_2, \dots, x_m$ yra įvestys, $w_{k1}, w_{k2}, \dots, w_{km}$ yra įvestims priskiriami svoriai, v_k yra sumavimo rezultatas, $\varphi(\cdot)$ yra aktyvacijos funkcija, y_k išvestis (Haykin, 2009).



6 pav. Bendras neurono grafinis atvaizdavimas

Paveiksle atvaizduota neuroną galime išreikšti pora matematinių formulių:

$$v_k = \sum_{j=1}^m (w_{kj} x_j) \quad (5)$$

$$y_k = \varphi(v_k) \quad (6)$$

Aktyvavimo funkcija $\varphi(v)$ nusako neurono išvestį. Pagrindinės aktyvavimo funkcijos yra ribinė ir sigmoidinė (Haykin, 2009).

Ribinės funkcijos pavyzdys:

$$\varphi(v) = \begin{cases} 1 & \text{if } v \geq 0 \\ 0 & \text{if } v < 0 \end{cases} \quad (7)$$

Sigmoidinės funkcijos pavyzdys gali būti logistinė funkcija:

$$\varphi(v) = \frac{1}{1 + \exp(-av)} \quad (8)$$

Kuriant dirbtinį neuroninį tinklą nebūtina dėti papildomų pastangų atrenkant tinkamiausius kintamuosius ar peržvelgiant jų tarpusavio koreliacijas. Modelis pats atlieka darbą pasiimdamas tik reikiamą informaciją, nes nereikšmingiems kintamiesiems priskiriami maži svoriai.. Deja, negalima to paties pasakyti apie tinkamų modelio parametrų parinkimą. Taikant dirbtinius neuroninius tinklus nėra greito būto parinkti geriausią sluoksnių ar neuronų skaičių – tai atliekama tik atliekant daug iteracijų.

1.4. Modelių tikslumo įverčiai

Modelių tikslumą vertinsime trimis būdais – paklaidomis, AUC ir Somers’ D.

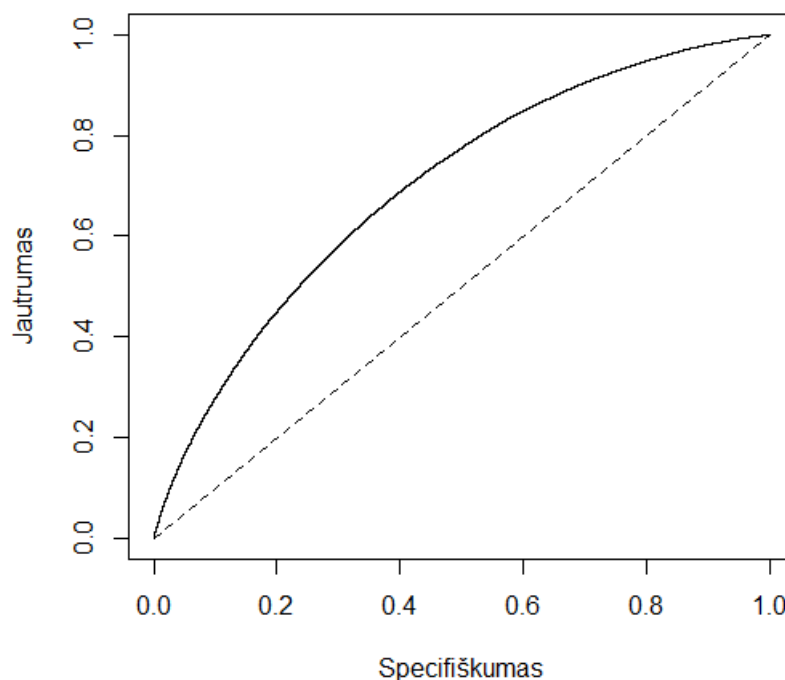
1.4.1. Vidutinė kvadratinė paklaida

Vienas iš būdų vertinti tikslumą yra įvertinti kokio dydžio paklaidos lieka sudarius prognozę. Tam naudosime vidutinės kvadratinės paklaidas (trumpiau vadinamas MSE). Jeigu Y yra prognozuojamo kintamojo vektorius, o $\hat{Y}$ – jo įvertinys, tuomet MSE galima apibūdinti formule:

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (9)$$

1.4.2. Plotas po ROC kreive

ROC kreivė vaizduoja modelio gebėjimą prognozuoti rezultatus. Brūkšninė linija centre rodo koks būtų atsitiktinio spėjimo rezultatas. Taigi kuo ROC kreivė yra aukščiau, tuo prognozė yra tikslesnė (7 pav.). Plotas po kreive vadinamas AUC yra modelio tikslumo įvertinys. Kuo jis didesnis, tuo modelis yra geresnis.



7 pav. ROC kreivės pavyzdys

1.4.3. Somers’ D įvertinys

Somers’ D įvertis yra pavadintas Roberto H. Somerso garbei, kuris pirmasis pasiūlė šį tikslumo matą 1962 metais. Šis įvertis vertina ryšį tarp kintamųjų porų ir įgyja reikšmes tarp 1 ir -1. Vienetas reiškia, kad visos poros įvertintos teisingai, o -1 kad visos poros įvertintos priešingai.

Jeigu prognozuojamas ir aiškinamasis kintamasis įgyja tik reikmes 0 arba 1, tuomet Somers’ D lygus tikimybių skirtumui:

$$Somers'D = P(Y = 1|X = 1) - P(Y = 1|X = 0) \quad (10)$$

Jeigu aiškinamasis kintamasis yra diskretus, tuomet kintamųjų reikšmės yra išrikiuojamos į eilę ir nagrinėjamas porų suderinamumas. Suderintos poros yra tokios, kurios išrikiuotos ta pačia eilės tvarka. Nesuderintos yra tos, kurios yra išrikiuotos priešinga eilės tvarka.

$$Somers'D = \frac{N_C - N_D}{N_C + N_D + N_T} \quad (11)$$

Čia N_C yra suderintos poros, N_D – nesuderintos poros, o N_T – nei suderintos, nei nesuderintos.

2. DUOMENŲ TYRYBA

2.1. Duomenų apžvalga

Analizei naudojame Jungtinių Amerikos Valstijų (JAV) tarpusavio skolinimo įmonės *Lending Club* duomenis nuo 2007 metų iki 2017 metų. Šie duomenys yra viešai teikiami įmonės internetinėje svetainėje. Duomenų rinkinį sudaro įrašai apie išduotas paskolas, detali informacija apie skolininką, ir paskolos statusas. Šis duomenų rinkinys turi 142 kintamuosius ir daugiau nei milijoną septynis šimtus tūkstančių įrašų. Iš turimų 142 kintamųjų, ne visi duomenys yra mums naudingi, todėl analizei atsirenkame tik dalį kintamųjų.

Kintamuosius atrenkame atsižvelgdami į savybes išvardintas žemiau.

- *Trūkstamų reikšmių skaičius* Iškeliame klausimą – ar galime logiškai paaiškinti trūkstamus duomenis ir juos užpildyti tinkamomis reikšmėmis? Kai kurie kintamieji turi labai nedaug reikšmių, todėl naudos iš jų būtų mažai. Bandant užpildyti kitomis reikšmėmis, didelė dalis duomenų būtų ne tikri, o simuliuoti. Kintamuosius, kurie turi mažiau nei 60 % reikšmių ir jų negalime logiškai pakeisti nuliais, atmetame.
- *Kintamojo reikšmių apipavidalinimas*. Peržvelgiame kintamojo duomenis, verčių dydžius ir unikalias reikšmes. Čia ieškome neįprastų reikšmių, netinkančių modeliui. Jeigu kintamasis yra kategorinis ir turi didelį skaičių unikalių reikšmių, tuomet norint jį naudoti, kintamojo reikšmes reikėtų sugrupuoti. Jei sunku tai padaryti dėl didelio unikalių reikšmių skaičiaus, kintamąjį atmetame. Vienas iš tokių kintamųjų yra darbovietės pavadinimas.
- *Duomenys, kurie gaunami tik išdavus paskolą*. Išdavus paskolą, skolininko elgsena stebima skaičiuojant vėlavimus ir panašią informaciją. Kadangi mūsų tikslas prognozuoti ar žmogus vykdys paskolos įsipareigojimus, modeliui kurti apsiribosime duomenimis, kurie turimi dar nepriėmus sprendimo ar skolinti pinigų. Neįtraukiant duomenų naudojamų endogeniniam kintamajam sukurti, visus likusius kintamuosius, kuriems informacija surinkama tik po sprendimo priėmimo, pašaliname iš analizės.

Padarius pradinę duomenų apžvalgą ir atmetus kintamuosius pagal aukščiau išvardintas taisykles, identifikuojame kintamuosius, kurių trūkstamos reikšmės rodo įvykio nebuvimą. Tarkime, jeigu sąskaitų skaičiaus nėra, suprantame, kad asmuo neturi sąskaitų. Tokias trūkstamas reikšmes užpildome nuliais.

Po pradinio duomenų apdorojimo, pereinama prie informacijos pritaikymo. Kai kurie duomenys negali būti panaudojami be papildomų pakeitimų, todėl perdarome tokius kintamuosius į naujus, kad galėtume pritaikyti juos analizėje.

2.1.1. Paskolos statusas

Paskola gali būti apmokėta, egzistuojanti, vėluojanti, nevykdoma arba neapmokėta. Šios kategorijos yra suskirstytos plačiau apibūdžiant vėlavimo trukmę, tačiau mums nereikia tokios detalios informacijos ir apsiribojame prieš tai išvardintomis keturiomis kategorijomis. Vėluojančios paskolos yra tos, kurias vėluojama apmokėti nuo 1 iki 120 dienų. Jeigu vėluojama apmokėti daugiau nei 120 dienų, tuomet paskola vadina nevykdoma (angl. „Defaulted“). Neapmokėta (angl. „Charged off“) yra tokia paskola, kurios skoliniojas nebelaukia ir perduoda skolų išieškotojams. Dažniausiai paskola tampa neapmokėta praėjus 30 dienų po to, kai ji tampa nevykdoma, tačiau tai gali nutikti ir anksčiau, tarkim jei skolininkas paskelbė bankrotą. Lentelėje apačioje turime skaičius kiek kokių paskolų yra duomenų rinkinyje.

1 lentelė. Paskolų skaičius pagal statusą

	value	count
1	Apmoketa	1031295
2	Egzistuojanti	451136
3	Neapmoketa	260453
4	Nevykdoma	33
5	Veluojanti	22509

Šiame darbe prognozuojamos neapmokėtos ir nevykdomos paskolos. Kaip matome, tokių įrašų yra daugiau nei 260 tūkstančių ir jie sudaro 14,75 % visų paskolų (1 lentelė).

2.1.2. Darbas

Duomenų rinkinyje yra duoti du kintamieji apie skolininko darbą – darbovietės pavadinimas ir darbo stažas. Darbovietės pavadinimas yra kategorinis kintamasis turintis daugybę unikalų reikšmių. Deja, mes neturime būdo jas suskirstyti į grupes, tačiau pastebime trūkstamų reikšmių skaičių – 6,36 % žmonių nėra įrašę darbovietės pavadinimo. Tikėtina, kad šie žmonės neturi darbo ir tai gali būti svarbus veiksnys prognozuojant kredito riziką. Pasinaudojant šia informacija sukuriame kintamąjį su reikšmėmis „1“, jeigu trūksta darbovietės pavadinimo ir „0“ jeigu darbovietės pavadinimas įrašytas. Šį kintamąjį pavadinsime „unemployed“.

Skaitinio kintamojo rodančio darbo stažą reikšmės yra natūralieji skaičiai nuo 0 iki 10. Taip parodomas dirbtų metų skaičius, kur nulis reiškia mažiau nei metus, o 10 – dešimt ir daugiau metų. Šis kintamasis turi 5,9 % tuščių reikšmių, kurias užpildome nuliais.

2.1.3. Pajamos

Vertinant paskolą būtina atsižvelgti į pajamas ir teisingai jas įvertinti. Paskola gali būti paimta ir vieno ir dviejų asmenų. Jeigu paskolą ima vienas asmuo, jo pajamos bus naudojamos analizei, tačiau, jei yra du asmenys, naudojame bendras jų pajamas. Šiam kintamajam turime tik 4 įrašus su trūkstamomis reikšmėmis ir darant prielaidą, kad šie asmenys neturi pajamų, trūkstamas reikšmes užpildome nuliais.

Papildomai dar turime kintamąjį parodantį ar pajamų šaltinis buvo patvirtintas. Iš kategorinio kintamojo sukuriame binominį pavadinimu „verified_flag“, kuris lygus vienetui, jei duomenys patvirtinti ir nuliui kitu atveju. Šis kintamasis išskiria nepatikimą informaciją apie pajamas.

2.1.4. Paskolos įrašo komentaras

Asmenys, prašydami jiems suteikti paskolą, gali palikti komentarą su detaliu aprašymu, kokių tikslų paskola yra reikalinga. Potencialūs skolininkai šią informaciją mato, ir gali nuspręsti ar priežastis tinkama paskolai suteikti. Mes neturime galimybės sugrupuoti komentarus, tačiau pastebime, kad komentarai yra neprivalomi ir juos rašo tik 7 % besiskolinančių asmenų. Pasinaudojame šia informacija sukurdami kintamąjį „desc_flag“ lygu vienetui, jei komentaras įrašytas ir nuliui, jei komentaro nėra.

2.1.5. Skolų ir pajamų santykis

Priimant sprendimus dėl skolinimo, finansinės institucijos atsižvelgia į skolos ir pajamų santykį (angl. „DTI – Debt to Income“). Šis santykis yra apskaičiuojamas sudedant visų paskolų mėnesines įmokas ir padalinant jas iš mėnesinių pajamų. Kuo didesnis skolų ir pajamų santykis, tuo rizikingiau yra skolinti asmeniui, nes gali būti, kad skolos našta taps per didelė. Dažniausiai

saugiu skolos ir pajamų santykiu yra laikomas 0,4 arba kitaip tariant 40 % gaunamų pajamų išleidžiama skoloms padengti. Šiame duomenų rinkinyje turime skolos ir pajamų santykį išreikštą procentais. Dviejų paskolos pareiškėjų atveju turime skolos ir pajamų santykį kiekvienam iš jų atskirai. Norėdami įtraukti į analizę abu asmenis, sukuriame kintamąjį „dti_max“, kuris yra lygus maximumui šių dviejų reikšmių.

2.2. Kategorinių kintamųjų analizė

Turimų duomenų rinkinys labai platus, tačiau ne visi duomenys yra tinkami analizei. Kaip jau minėta, dalis kintamųjų turi informaciją gautą jau išdavus paskolą ir stebint skolininko elgseną. Taip pat yra duomenų, kurių negalime pritaikyti analizei dėl jų struktūros arba trūkstamų reikšmių skaičiaus. Tokius kintamuosius atmetus turime duomenų rinkinį, kurį analizuojame detaliau. Čia prieiname prie kitos svarbios dalies – kategorinių kintamųjų. Po duomenų valymo ir atrinkimo, turime 4 kategorinius kintamuosius: valstiją iš kur skolininkas yra kilęs, paskolos tikslą, gyvenamojo būsto tipą ir paskolos reitingą, suteiktą tarpusavio skolinimo įmonės „Lending club“. Šie kintamieji gali turėti naudingos informacijos, padedančios nuspėti ar paskola bus išmokėta, tačiau naudoti juos tiesiogiai su Logistine regresija yra nepatogu ir netgi neįmanoma su neuroninių tinklų modeliu. Dėl šios priežasties analizuojamos turimos reikšmės norint sugrupuoti ir sukurti atitinkamus faktorius.

2.2.1. Gyvenamojo būsto tipas

Remiantis bendra nuomone, gyvenamojo būsto tipas gali būti reikšmingas finansinio stabilumo indikatorius. Asmenys, turintys savo būstą, turi daugiau finansinės laisvės, nes sutaupo pinigų neturėdami mokėti nuomos. Tuo tarpu būsto paskolas turintys asmenys rodo gebėjimą būti finansiškai atsakingais ir mokėti paskolos įmokas.

Duomenų rinkinyje turime penkis gyvenamojo būsto tipus: nuosavas būstas („Own“), būsto paskola („Mortgage“), nuomojamas būstas („Rent“), kita („Other“) ir nėra gyvenamosios vietos („None“). Lentelėje apačioje turime duomenis apie blogų klientų dalies pasiskirstymą pagal būsto tipą.

2 lentelė. Blogų klientų dalis pagal būsto tipą

home_ownership	kiekis	dalys_blogu	vid_pajamos
RENT	699523	0.1718928	67011.98
OWN	192929	0.1484328	73816.40
NONE	54	0.1481481	60433.91
OTHER	690	0.1318841	80275.27
MORTGAGE	872230	0.1278413	90047.36

Iš duomenų matome, kad kiekviena grupė skiriasi nuo kitų ir geriausi skolininkai yra tie, kurie turi būsto paskolas arba pažymi gyvenamąjį būstą kaip „Kita“. Tuo tarpu blogiausi klientai yra tie, kurie nuomojasi būstą. Iš antros lentelės vienintelė neįprasta informacija yra apie neturinčius būsto skolininkus. Į šią grupę patenkantys asmenys paskolas grąžina pakankamai gerai – negrąžintų paskolų dalis sudaro mažiau nei 15 %. Nors jų dalis labai maža, tačiau tikėtina, kad tai yra jauni žmonės gyvenantys su tėvais. Tuomet jie neturi mokėti nei būsto paskolos, nei nuomos, bet išgyvena ir grąžina paskolas turėdami žemiausias pajamas.

Dar vienas pastebėjimas, kad blogų klientų dalis turi priešingą ryšį su vidutinėmis pajamomis. Grupės turinčios daugiausiai blogų klientų vidutinės pajamos yra mažiausios ir tokia tendencija išlieka per kitas grupes išskyrus jau aptartus žmones neturinčius gyvenamojo būsto („None“).

Nors grupės „Kita“ („Other“) gyvenamasis būstas yra neaiškus, tačiau tokių klientų vos 690. Lyginant su visu duomenų rinkiniu viršijančiu vieną su puse milijono, tokie klientai nesudaro net vieno procento turimų duomenų. Yra tikimybė, kad šios grupės turima informacija yra atsitiktinė ir iš tiesų neturi jokių išskirtinių savybių, tačiau gali būti ir tai, kad asmenys pažymintys gyvenamąjį būstą „Kita“ iš tiesų turi daugiau nei vieną būstą ir yra sutrikę, ką turėtų pažymėti. Tokios situacijos pavyzdys galėtų būti, žmogus turintis du būstus – vieną nuosavą ir vieną su paskola.

Remiantis analizuotais duomenimis nusprendžiame nejungti būsto tipų į grupes, o sukurti faktorius kiekvienam iš būsto tipų. Nauji kintamieji bus naudojami modeliuose vietoje turimo kategorinio gyvenamojo būsto kintamojo.

2.2.2. Paskolos tikslas

Žemiau esančioje lentelėje turime suskirstymą pagal paskolos tikslą. Lentelėje pateikiamas paskolų kiekis, procentas blogų klientų kiekvienoje grupėje ir vidutinės tos grupės pajamos per metus.

3 lentelė. Blogų klientų dalis pagal paskolos tipą

purpose	kiekis	dalys_blogu	vid_pajamos
small_business	20106	0.2223714	94245.12
educational	424	0.2075472	51280.18
renewable_energy	1170	0.1846154	75549.61
moving	12288	0.1728516	71428.28
house	8706	0.1631059	85237.59
debt_consolidation	1018235	0.1578064	78106.03
medical	20866	0.1533116	75812.29
other	104422	0.1488001	73611.66
vacation	12024	0.1374750	70324.10
major_purchase	38823	0.1326276	80414.51
home_improvement	117709	0.1269402	94403.75
credit_card	389269	0.1249727	78863.92
wedding	2350	0.1238298	69575.55
car	19034	0.1083850	71044.70

Pastebime, kad rizikingiausios yra paskolos imtos mažiems verslams (angl. „small business“) arba mokymosi išlaidoms padengti (angl. „educational“). Tuo tarpu saugiausios yra paskolos imamos mašinoms įsigyti (angl. „car“). Taip pat akivaizdu, kad vidutinės pajamos nelabai turi įtaką ar skolininkas gražins paskolą. Tai parodo paskolos mažiems verslams ir namų renovacijoms (angl. „home improvement“). Šiose kategorijose vidutinės metinės pajamos beveik vienodos – 94 tūkstančiai per metus, tačiau paskolose mažiems verslams pasitaiko beveik dvigubai daugiau blogų klientų – net 22,2 %, kai tuo tarpu namų renovacijai imantys paskolas jų nesugeba gražinti tik 12,7 % atvejų.

Apskaičiavus blogų paskolų dalį visoms paskolos tikslo grupėms ir pritaikius K-vidurkių (angl. k-means) metodą, optimaliu randame skirstymą į 3 grupes.

Iš 8-to paveikslą matome, kaip ir buvo pastebėta, blogiausios yra paskolos mažiems verslams ir mokslams. Jos priskirtos trečiai grupei ir turi 21,5 % blogų paskolų. Trečios sąrašė pagal neišmokėjimą yra paskolos atsinaujinantiems energijos šaltiniams. Nors jos ir nepateko į blogiausių paskolų grupę, tačiau akivaizdu, kad turėdamos 18,4 % negražinimą, nėra labai saugios. Nors atsinaujinanti energija yra pigesnė, tačiau įranga tokiai energijai išgauti yra labai brangi ir atsiperka tik per ilgą laikotarpį. Kadangi „Lending club“ išduoda paskolas ne ilgesniam nei 6 metų periodui, tikėtina, kad skolininkai pervertina kiek daug jie sutaupys naudodami atsinaujinančią energiją.

k-means clustering with 3 clusters of sizes 6, 6, 2

Cluster means:

```

dalis_blogu
1  0.1257050
2  0.1634152
3  0.2149593

```

Clustering vector:

```
[1] 3 3 2 2 2 2 2 2 1 1 1 1 1 1
```

within cluster sum of squares by cluster:

```
[1] 0.0004920169 0.0008857319 0.0001098794
(between_SS / total_SS = 89.6 %)
```

	purpose	purpose_gr
small_business		3
educational		3
renewable_energy		2
moving		2
house		2
debt_consolidation		2
medical		2
other		2
vacation		1
major_purchase		1
home_improvement		1
credit_card		1
wedding		1
car		1

8 pav. Paskolos tikslo suskirstymas k-vidurkių metodu

Geriausia yra pirmoji grupė turinti tik 12,6 % blogų paskolų. Šiai grupei priskiriamos paskolos mašinoms (angl. „car“), vestuvėms (angl. „wedding“), būsto renovacijai (angl. „home improvement“), dideliems pirkiniais (angl. „major purchase“) ir atostogoms (angl. „vacation“). Sudarytas grupės panaudosime sukurdami faktorius paskolos tikslui apibrėžti. Faktorių naudojimas leis mums atskirti grupes su mažiausia ir didžiausia blogų skolininkų dalimi nuo vidutiniškai gerų paskolų.

2.2.3. Paskolos reitingas

Tarpusavio skolinimo platforma „Lending club“ kartu su įprastiniais duomenimis pateikia ir skolininko reitingą, kurį jie patys apskaičiuoja. Reitingas („Grade“) yra suskirstytas į 7 grupes nuo A iki G, kur A yra geriausias, o G yra blogiausias. Lentelėje apačioje matome, kad A reitingas turi tik 4,6 % blogų skolininkų, kai tuo tarpu G reitingas turi net 39,8 % tokių skolininkų. Taip pat, verta pastebėti, kad blogu klientų dalis didėja blogėjant reitingui su kiekviena grupe, o tai įrodo, kad skolininkai įvertinami teisingai.

4 lentelė. Blogų klientų dalis pagal reitingą

grade	kiekis	dalis_blogu	vid_pajamos
G	11497	0.39801687	77738.04
F	38625	0.37139159	74121.16
E	116681	0.30267139	73024.14
D	255378	0.23011379	72295.64
C	523203	0.15874144	76031.33
B	522192	0.09742777	79950.28
A	297850	0.04550277	92158.02

Šiame suskirstyme tik A reitingą turintys asmenys išsiskiria pagal vidutines pajamas. Jų metinis atlyginimas siekia net 92 000 dolerių per metus. Kitose grupėse tendencijos nematyti ir vidutinės pajamos išsidėsčiusios nuo 72 000 iki 80 000 dolerių per metus.

Norėdami sumažinti grupių skaičių pritaikome K-vidurkių metodą skirstymui į grupes. Tinkamiausias grupių skaičius vėlgi yra 3, nes taip mes galime atskirti blogiausias, vidutines ir geriausias paskolas. Gautus rezultatus galime apžiūrėti žemiau esančiame 9 paveiksle.

```
k-means clustering with 3 clusters of sizes 2, 2, 3

cluster means:
  dalis_blogu
1  0.07146527
2  0.19442762
3  0.35735995

clustering vector:
[1] 3 3 3 2 2 1 1

within cluster sum of squares by cluster:
[1] 0.001348103 0.002547006 0.004840711
(between_SS / total_SS = 92.1 %)

      grade grade_gr
      G         3
      F         3
      E         3
      D         2
      C         2
      B         1
      A         1
```

9 pav. Reitingo suskirstymas k-vidurkių metodu

Matome, kad pirma grupė turi 7,1 % blogų paskolų, antra 19,4 %, o trečia net 35,7 % blogų paskolų. Grupėms sudaryti sujungiami A ir B reitingai, C ir D reitingai, o blogiausiais laikomi E, F ir G reitingai.

Grupes panaudojame sukurdami atitinkamus reitingo faktorius, kuriuos naudosime modelių sudaryme.

2.2.4. Valstijos

Dar vienas tekstinis kintamasis – Valstijos. Jis suteikia informaciją apie skolininko gyvenamąją vietą, t.y. valstiją, o tai gali mums parodyti ar tam tikrose valstijose žmonės yra mažiau mokūs nei kitose. Sąraše turime 50 valstijų ir JAV sostinę Vašingtoną. Kadangi sąrašas labai ilgas, tai žemiau esančioje lentelėje pavaizduotos tik tos valstijos, kuriose išduota daugiau nei 10 tūkstančių paskolų. Kaip ir prieš tai analizuotiems kintamiesiems apskaičiuojame išduotų paskolų skaičių, jose esančių blogų paskolų dalį ir vidutines metines pajamas.

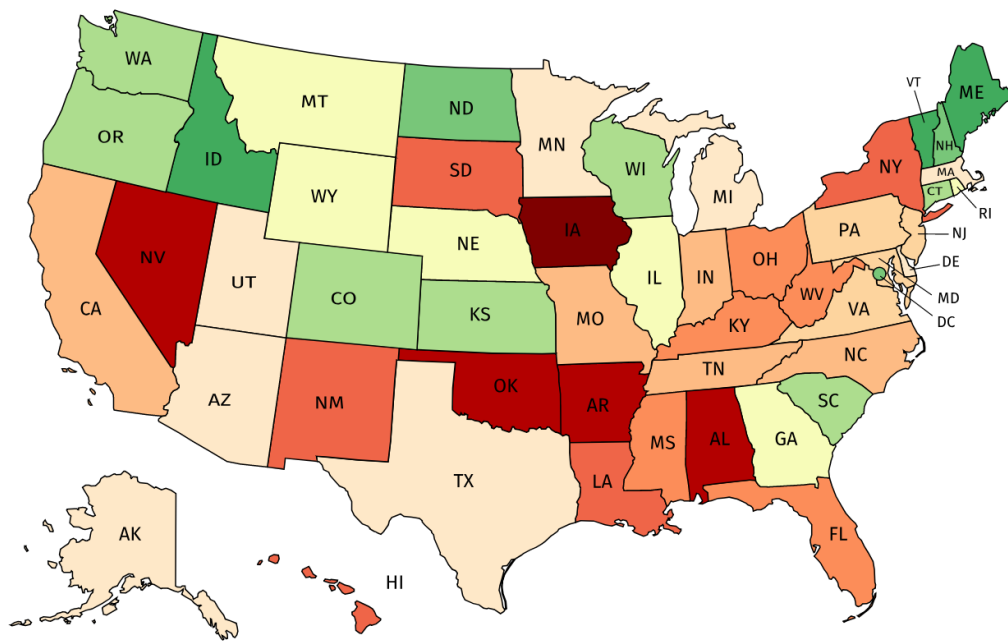
Paveiksle matome, kad dauguma valstijų yra panašios, tačiau turime kelias, kurios išsiskiria. Didžiausią portfelių blogų paskolų turi Arkanzasas (trumpinys AR) ir Alabama (trump. AL). Mažiausią Oregonas (trump. OR) ir Šiaurės Karolina (trump. SC). Taip pat pastebime, kad ir čia pajamos nerodo jokios įtakos blogų paskolų skaičiui. Arkanzase ir Oregone vidutinės metinės pajamos yra panašios – apie 70 000 dolerių. tačiau Arkanzasas turi 17,7 % blogų paskolų, o tuo tarpu Oregone blogų paskolų yra tik 11,1 %.

Toks blogų klientų dalies skirtumas tarp valstijų verčia susimąstyti kokios šio skirtumo priežastys. Vienas iš spėjimų yra gyventojų mentalitetas „gyventi ne pagal savo kišenę“. Galimas tokio požiūrio susiformavimas turi turėti savo šaknis. Žemiau turime atvaizduotus žemėlapius

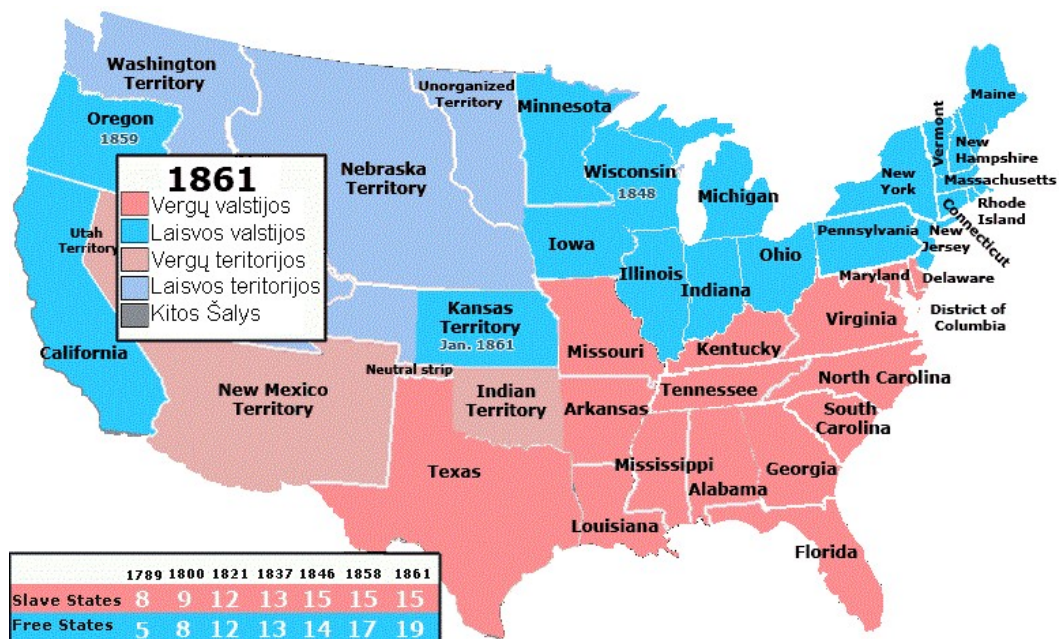
5 lentelė. Blogų klientų dalis pagal Valstiją

	addr_state	kiekis	dalis_blogu	vid_pajamos
1:	AR	13264	0.1769451	69839.63
2:	AL	21680	0.1761070	71830.26
3:	OK	16094	0.1734187	74158.79
4:	NV	25042	0.1707132	74550.43
5:	LA	20564	0.1698599	77263.06
6:	NY	146794	0.1596727	81820.26
7:	FL	124604	0.1590800	74172.71
8:	MO	28225	0.1554650	71940.28
9:	TN	27291	0.1553992	72068.71
10:	IN	29138	0.1549523	72396.00
11:	NC	49099	0.1542394	74910.18
12:	KY	17086	0.1540442	70971.05
13:	MD	41953	0.1536005	87036.18
14:	NJ	65324	0.1529759	91186.11
15:	PA	60940	0.1513620	74979.55
16:	CA	247266	0.1509832	84462.62
17:	MI	46175	0.1499729	72487.28
18:	AZ	41628	0.1494667	75778.08
19:	OH	59095	0.1489635	70470.82
20:	VA	49994	0.1482778	87158.82
21:	MN	31426	0.1465665	73990.65
22:	TX	145434	0.1455506	84298.78
23:	UT	11833	0.1412998	76815.08
24:	MA	40597	0.1405769	83925.85
25:	WI	23317	0.1338080	70813.07
26:	GA	57800	0.1335294	78724.84
27:	IL	71499	0.1280298	81143.69
28:	WA	36780	0.1217238	78394.26
29:	KS	15175	0.1208567	73684.63
30:	CT	27831	0.1205490	88098.45
31:	CO	37071	0.1198241	77936.66
32:	SC	21582	0.1163933	74449.09
33:	OR	20756	0.1108595	70035.05

rodančius blogų klientų pasiskirstymą ir Šiaurės Amerikos teritorijas, kuriose 1861 metais buvo legali vergija.



10 pav. Blogų klientų dalis (didžiausia – raudona, mažiausia – žalia)



11 pav. Vergovė Šiaurės Amerikoje 1861 metais

Nors ryšys nėra akivaizdus, tačiau galima pastebėti, kad valstijose, esančiose JAV pietuose, blogų klientų dalis yra didesnė. Tai gali būti susiję su tuo, kad šiose valstijose vergija buvo legali ir pasekmės yra vis dar jaučiamos šiomis dienomis.

Pastebint, kad valstijų duomenys turi mums naudingos informacijos, norisi jas įtraukti sudarant prognozavimo modelį. Kadangi iš viso rinkinyje yra 50 valstijų ir vienas miestas, reikia jas suskirstyti į grupes arba naudoti faktorius kelioms svarbiausioms valstijoms. Apskaičiavus blogų paskolų dalį visoms valstijoms ir pritaikius K-vidurkių metodą, randame optimalų skirstymą į 3 grupes.

24

2.3. Reikšmingų kintamųjų atrinkimas regresinei analizei

Peržiūrėjus visus galimus kintamuosius ir sudarius norimus faktorius, turime 45 aiškinamuosius kintamuosius. Žinoma, dalis jų turi labai panašią informaciją arba neturi aiškinamosios galios. Nenorint apkrauti modelių nereikalinga informacija, reikia atrinkti pačius geriausius.

Pirmiausia peržvelgiame koreliacijas. Žemiau esančioje lentelėje turime koreliacijas tarp pagrindinių skaitinių reikšmių.

6 lentelė. Koreliacija tarp skaitinių rodiklių

	loan_amnt	int_rate	income	installment	bad_flag
loan_amnt	1.0000000	0.13424749	0.20222361	0.94670516	0.02535380
int_rate	0.1342475	1.00000000	-0.04640182	0.15263096	0.19191840
income	0.2022236	-0.04640182	1.00000000	0.19308454	-0.02728974
installment	0.9467052	0.15263096	0.19308454	1.00000000	0.02880549
bad_flag	0.0253538	0.19191840	-0.02728974	0.02880549	1.00000000

Pastebime, kad pajamos turi nors ir žemą, bet neigiamą koreliaciją su palūkanų norma ir blogų klientų indikatoriumi. Ši informacija sutampa su bendrai nusistovėjusia nuomone, kad mažiau uždirbantys asmenys yra rizikingesni. Be kita ko, paskolos dydis ir mėnesinė įmoka stipriai koreliuoja (0,95). Tai yra logiška, nes paskolos laikotarpis yra ribotas – 36 arba 60 mėnesių.

Kadangi prognozuojamas kintamasis yra kategorinis, koreliacijų negalima naudoti reikšmingumo nustatymui. Dėl šios priežasties sudaromos atskiros logistinės regresijos tarp prognozuojamo rodiklio („bad_flag“) ir kiekvieno turimo kintamojo.

Remiantis gautomis T statistikomis atrenkame geriausiai prognozuojančius rodiklius ir taip kintamųjų skaičių sumažiname iki 31. Be kita ko, patikrinus kintamųjų tarpusavio koreliacijas, išimame dar 7 rodiklius, kurie stipriai koreliuoja su likusiais kintamaisiais. Tarp pašalintų yra ir paskolos suma. Ši vertė stipriai susijusi su mėnesine įmoka ir pastaroji pagal T statistika prognozuoja šiek tiek geriau.

Žemiau esančioje lentelėje yra pavaizduoti 24 atrinkti kintamieji ir jų statistikos. Neigiamos T statistikos reikšmės rodo, kad ryšys tarp kintamųjų yra priešingas. Tai reiškia, kad žmogaus priklausymas „grade_geri“ grupei (reitingas A arba B) mažina tikimybę, kad paskola bus negrąžinta.

Lentelę su visais 45 kintamaisiais ir jų T statistikomis bei labiausiai koreliuoti kintamieji yra pateikti A priede.

	kintamasis	T statistika	P reikšmė
1:	term	131.6146	0
2:	int_rate	279.5582	0
3:	installment	48.94794	0
4:	inq_last_6mths	105.026	0
5:	open_acc	28.52878	5.150117e-179
6:	revol_bal	-32.69157	2.057634e-234
7:	acc_open_past_24mths	129.6827	0
8:	mort_acc	-69.65076	0
9:	num_actv_rev_tl	63.53672	0
10:	pub_rec_bankruptcies	35.82775	4.085409e-281
11:	unemployed	40.87152	0
12:	empl_years	-28.69741	4.109781e-181
13:	income	-76.84623	0
14:	verified_flag	103.9017	0
15:	dti_max	68.27278	0
16:	home_mortgage	-72.77642	0
17:	home_rent	73.69579	0
18:	state_geri	-45.24518	0
19:	state_blogi	24.16443	5.265703e-129
20:	purpose_geri	-57.3522	0
21:	purpose_blogi	29.96764	2.591965e-197
22:	grade_geri	-232.3779	0
23:	grade_blogi	205.6749	0
24:	joint_application	-33.457	2.03594e-245
	kintamasis	T statistika	P reikšmė

7 lentelė. Geriausių kintamųjų sąrašas

2.4. Modelio kintamųjų sąrašas

Žemiau turime aprašymus kintamųjų, kuriuos naudosime modeliavime.

- `bad_flag` – indikatorius, kad paskola yra nevykdoma arba neapmokėta;
- `term` – paskolos terminas (36 arba 60 mėnesių);
- `int_rate` – paskolos palūkanų norma;
- `installment` – mėnesinė įmoka, jei paskola bus suteikta;
- `inq_last_6mths` – kreditų paklausimų skaičius per pastaruosius 6 mėnesius neįtraukiant būsto ir transporto priemonių paskolų;
- `open_acc` – atidarytų sąskaitų kiekis;
- `revol_bal` – bendra kreditinių kortelių skola;
- `acc_open_past_24mths` – skaičius sandorių, atidarytų per paskutinius 24 mėnesius;
- `mort_acc` – būsto paskolų skaičius;
- `num_actv_rev_tl` – aktyvių banko kortelių su kredito limitu skaičius;
- `pub_rec_bankruptcies` – skaičius kiek kartų asmeniui buvo paskelbtas bankrotas;
- `unemployed` – indikatorius, jog asmuo yra galimai bedarbis. Nenurodė darbovietės;
- `empl_years` – darbo stažas metais;
- `income` – metinės pajamos (suma metinių pajamų, jeigu yra du skolininkai);
- `verified_flag` – indikatorius, ar skolininko pajamos arba pajamų šaltinis buvo patvirtinti;
- `dti_max` – skolos ir pajamų santykis (angl. „Debt to income“). Jeigu paskolos prašo du asmenys, imamas maksimumas tarp abiejų skolininkų;
- `home_mortgage` – gyvenamajam būstui yra paimta paskola;
- `home_rent` – gyvenamasis būstas yra nuomojamas;
- `joint_application` – indikatorius, rodantis ar paskola ima du asmenys;
- `state_geri` – valstijos, kuriose yra mažiausias blogų klientų procentas;
- `state_blogi` – valstija, kurioje gyvena daugiausiai blogų klientų;
- `purpose_geri` – paskolos tikslas priklauso grupei, turinčiai mažiausią blogų klientų kiekį;
- `purpose_blogi` – paskolos tikslas priklauso grupei, turinčiai didžiausią blogų klientų dalį;
- `grade_geri` – klasė, priskirta „Lending club“, yra A arba B;
- `grade_blogi` – klasė, priskirta „Lending club“, yra E, F arba G.

3. MODELIAVIMAS IR PROGNOZAVIMAS

3.1. Logistinė regresija

Paskolų nemokumą prognozuojame pasinaudodami logistine regresija. Sudėjus visus atrinktus kintamuosius į modelį, matome, kad jie visi yra reikšmingi.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0271  -0.6103  -0.4507  -0.3400   8.4904

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.182e+00  2.480e-02 -128.293 < 2e-16 ***
term          1.056e-02  2.799e-04  37.739 < 2e-16 ***
int_rate      5.177e-02  1.209e-03  42.808 < 2e-16 ***
installment   4.113e-04  1.308e-05  31.433 < 2e-16 ***
inq_last_6mths 9.836e-02  2.947e-03  33.378 < 2e-16 ***
open_acc     -9.860e-03  7.579e-04 -13.009 < 2e-16 ***
revol_bal    -1.567e-06  1.938e-07  -8.085 6.21e-16 ***
acc_open_past_24mths 5.179e-02  1.032e-03  50.196 < 2e-16 ***
mort_acc     -3.450e-02  1.839e-03 -18.762 < 2e-16 ***
num_actv_rev_tl 1.699e-02  1.107e-03  15.353 < 2e-16 ***
pub_rec_bankruptcies 4.558e-02  7.248e-03   6.290 3.18e-10 ***
unemployed    2.955e-01  1.179e-02  25.073 < 2e-16 ***
empl_years   -5.055e-03  8.191e-04  -6.172 6.74e-10 ***
income       -3.177e-06  8.692e-08 -36.547 < 2e-16 ***
verified_flag  1.917e-01  6.920e-03  27.696 < 2e-16 ***
dti_max       5.061e-03  2.642e-04  19.158 < 2e-16 ***
home_mortgage -7.591e-02  9.823e-03  -7.728 1.09e-14 ***
home_rent     1.448e-01  9.654e-03  15.003 < 2e-16 ***
state_geri    -2.371e-01  7.519e-03 -31.538 < 2e-16 ***
state_blogi    1.528e-01  1.137e-02  13.440 < 2e-16 ***
purpose_geri  -4.614e-02  6.379e-03  -7.233 4.71e-13 ***
purpose_blogi  2.980e-01  2.325e-02  12.815 < 2e-16 ***
grade_geri    -4.145e-01  9.556e-03 -43.378 < 2e-16 ***
grade_blogi    1.581e-01  1.178e-02  13.430 < 2e-16 ***
joint_application -6.065e-01  2.155e-02 -28.141 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 886618  on 1059254  degrees of freedom
Residual deviance: 820454  on 1059230  degrees of freedom
AIC: 820504
```

14 pav. Logistinės regresijos modelis iš atrinktų kintamųjų

Prisimenant logistinės regresijos teorijos dalyje modeliui vardintus reikalavimus, pastebime, kad dauguma jų išpildomi ruošiantis sudaryti logistinę regresiją. Sąlyga, reikalaujanti į modelį įtraukti tik reikšmingus kintamuosius, yra įvykdyta, nes visi kintamieji yra statistiškai reikšmingi (14 pav.). Kitu atveju, didesnę nei 0,05 p-reikšmę turintys kintamieji būtų šalinami iš modelio tol, kol liktų tik reikšmingi.

Kita svarbi sąlyga, multikolinearumo nebuvimas. Nors atrenkant kintamuosius buvo atsižvelgta į jų tarpusavio koreliacijas, tačiau modelis vis tiek gali būti multikolinearus, jeigu kolinearumas egzistuoja tarp trijų ar daugiau kintamųjų. Tai reikštų, kad aiškinamieji kintamieji turi pasikartojančią informaciją. Multikolinearumui patikrinti naudojame dispersijos mažėjimo daugiklį, sutrumpintai vadinama VIF. Jeigu jo reikšmė yra daugiau nei 4, tuomet

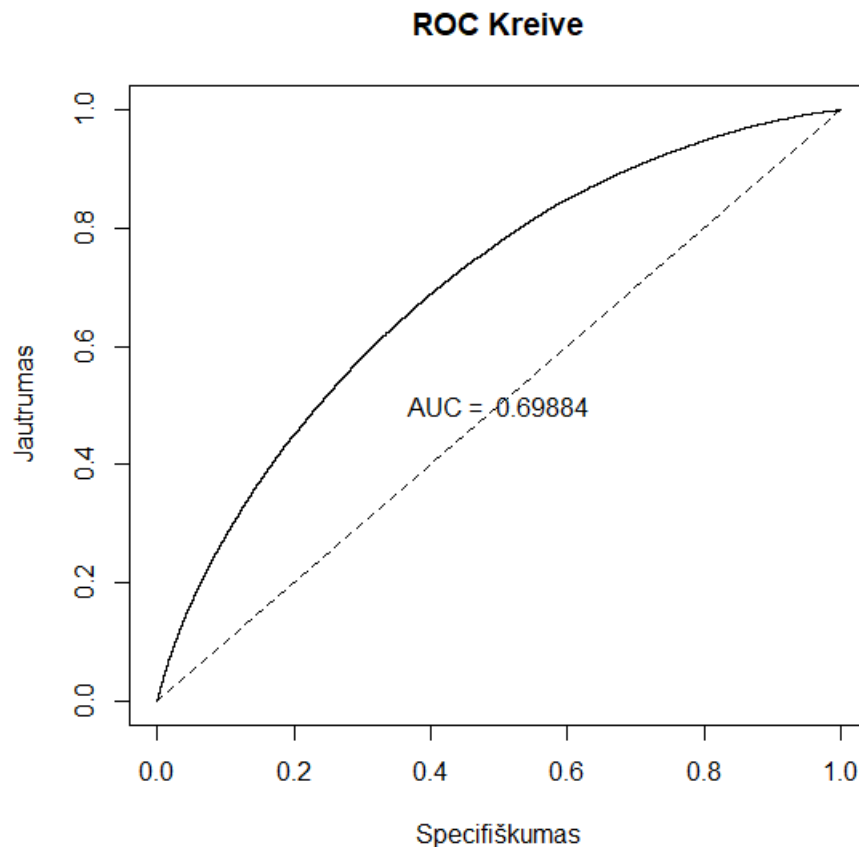
multikolinearumas egzistuoja, jei vertė daugiau nei 10, tuomet multikolinearumas yra per didelis ir turi būti šalinamas.

term	int_rate	installment
1.302258	4.444312	1.494098
inq_last_6mths	open_acc	revol_bal
1.121306	2.299617	1.477225
acc_open_past_24mths	mort_acc	num_actv_rev_tl
1.592496	1.442275	2.001414
pub_rec_bankruptcies	unemployed	empl_years
1.046746	1.227636	1.253835
income	verified_flag	dti_max
1.658149	1.077821	1.207144
home_mortgage	home_rent	state_geri
2.955713	2.865669	1.022443
state_blogi	purpose_geri	purpose_blogi
1.026723	1.038517	1.020999
grade_geri	grade_blogi	joint_application
2.298855	2.432160	1.113257

15 pav. Logistinės regresijos modelio kintamųjų VIF reikšmės

Iš paveikslo matome, kad visi kintamieji turi žemas VIF reikšmes ir tik palūkanų normos reikšmė lygi 4,4. Kadangi ji nėra labai aukšta ir vienintelė yra ant ribos, todėl nieko nekeičiame.

Modelio tikslumą stebime brėždami ROC kreivę. Gautas grafikas parodo modelio gebėjimą prognozuoti rezultatus. Brūkšninė linija centre rodo atsitiktinio spėjimo rezultatą, todėl kuo ROC kreivė yra aukščiau, tuo prognozė yra tikslesnė. Modelio tikslumui vertinti paskaičiuojame ir AUC reikšmę, kuri yra plotas po ROC kreive, vidutinę kvadratinę paklaidą, trumpai vadinamą MSE ir Somers' D įvertį. Gauname $AUC=0,699$, $MSE=0,118$, o Somers' $D = 0,398$.



16 pav. ROC kreivė logistinės regresijos modeliui

Modelio lygtis:

$$P = -3,182 + 0,01 * term + 0,052 * int_rate + 0,0004 * installment + 0,098 * inq_last_6mths - 0,01 * open_acc - 0,000002 * revol_bal + 0,051 * acc_open_past_24mths - 0,035 * mort_acc + 0,017 * num_actv_rev_tl + 0,046 * pub_rec_bankruptcies + 0,296 * unemployed - 0,005 * empl_years - 0,000003 * income + 0,192 * verified_flag + 0,005 * dti_max - 0,076 * home_mortgage + 0,145 * home_rent - 0,237 * state_geri + 0,152 * state_blogi - 0,046 * purpose_geri + 0,298 * purpose_blogi - 0,416 * grade_geri + 0,158 * grade_blogi - 0,607 * joint_application$$

Gautas modelis yra geriausias logistinės regresijos modelis naudotiems duomenis, tačiau nėra labai tikslus skaičiuodamas kredito įsipareigojimų nevykdymo tikimybes. Modelio netikslumo priežastys gali būti netinkamas modelio parinkimas, silpnas ryšys tarp priklausomo ir nepriklausomų kintamųjų arba netinkamas duomenų rinkinys. Nenorint keisti modelio ir žinant, kad atrinkti yra geriausi turimi kintamieji, analizuojame trečią priežastį.

Tarpusavio skolinimas yra naujovė finansų rinkoje lyginant su bankais. Analizuojama įmonė *Lending Club* egzistuoja vos daugiau nei dešimt metų. Jos klientų skaičius augo su kiekvienais metais ir tikėtina, kad pradėjus teikti tarpusavio skolinimo paslaugas pirmaisiais klientais tapo mažiau patikimi asmenys, kuriems niekas kitas neskolina pinigų. Praėjus daugiau laiko ir įgavus klientų pasitikėjimą, vis daugiau žmonių pasirenka skolintis naudojantis šią platformą ir rinktis ją vietoje bankų. Tokiu būdu daugiau patikimų skolininkų pradeda naudotis įmonės *Lending Club* tarpusavio skolinimo platforma.

Pritaikant šią informaciją duomenų rinkiniui, suprantame, kad paskolų rinkinys keičiasi einant laikui. Tokiu atveju, imtis iš naujesnių paskolų turėtų duoti tikslesnius rezultatus. Šią teoriją išbandome kurdami logistinės regresijos modelius mažinant įtraukiamų metų skaičių. Lentelėje apačioje galime apžvelgti rezultatus visiems logistinės regresijos modeliams skirtingiems laiko intervalams.

8 lentelė. Logistinės regresijos rezultatai skirtingiems laiko intervalams

Metams nuo:	MSE	AUC	Somers' D
2007	0,1176	0,6988	0,3981
2008	0,1176	0,6989	0,3988
2009	0,1175	0,6992	0,3984
2010	0,1175	0,6995	0,3993
2011	0,1175	0,6997	0,3995
2012	0,1175	0,7002	0,4002
2013	0,1171	0,7014	0,4023
2014	0,1164	0,7034	0,4065
2015	0,1130	0,7043	0,4090
2016	0,1012	0,7044	0,4087

Lengva pastebėti, kad imant vis naujesnius duomenis, modelio tikslumas didėja tiek pagal Somers' D, tiek pagal AUC reikšmes. Tiksliausias modelis yra sudarytas 2015–2017 metų duomenų rinkiniui. Kadangi didžiausias skaičius paskolų yra suteiktais pastaraisiais metais, prarandamas nedidelis kiekis įrašų. Apmokymo, testavimo ir validavimo duomenų rinkinių dydžiai yra parodyti lentelėje apačioje.

9 lentelė. Duomenų rinkinių dydžiai pagal naudojamą laikotarpį

Metai	Apmokymo	Testavimo	Validavimo
2007–2017	1 059 255	353 085	353 086
2015–2017	779 788	259 999	259 294
Dalis originalių duomenų %	73,6 %	73,6 %	73,4 %

Duomenų rinkinys 2015–2017 metams yra pakankamai didelis, o 8-toje lentelėje turimi modelio tikslumo įvertiniai rodo šį modelį esant tiksliausiu iš visų sudarytų logistinės regresijos modelių testavimo duomenų imčiai. Detalės apie šį modelį yra pateiktos 17-tame paveiksle.

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.9774	-0.5953	-0.4385	-0.3302	7.4856

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-3.049e+00	2.895e-02	-105.307	< 2e-16	***
term	7.379e-03	3.329e-04	22.163	< 2e-16	***
int_rate	4.287e-02	1.398e-03	30.669	< 2e-16	***
installment	4.606e-04	1.502e-05	30.673	< 2e-16	***
inq_last_6mths	9.259e-02	3.864e-03	23.961	< 2e-16	***
open_acc	-1.564e-02	9.150e-04	-17.097	< 2e-16	***
revol_bal	-2.184e-06	2.292e-07	-9.529	< 2e-16	***
acc_open_past_24mths	5.786e-02	1.222e-03	47.355	< 2e-16	***
mort_acc	-2.839e-02	2.276e-03	-12.470	< 2e-16	***
num_actv_rev_tl	2.830e-02	1.362e-03	20.774	< 2e-16	***
pub_rec_bankruptcies	7.034e-02	8.203e-03	8.575	< 2e-16	***
unemployed	2.930e-01	1.397e-02	20.973	< 2e-16	***
empl_years	-7.620e-03	9.748e-04	-7.817	5.40e-15	***
income	-2.691e-06	9.857e-08	-27.302	< 2e-16	***
verified_flag	2.371e-01	8.307e-03	28.538	< 2e-16	***
dti_max	4.782e-03	2.760e-04	17.328	< 2e-16	***
home_mortgage	-9.980e-02	1.142e-02	-8.741	< 2e-16	***
home_rent	1.526e-01	1.120e-02	13.616	< 2e-16	***
state_geri	-2.563e-01	8.974e-03	-28.564	< 2e-16	***
state_blogi	1.559e-01	1.338e-02	11.650	< 2e-16	***
purpose_geri	-2.653e-02	7.561e-03	-3.509	0.00045	***
purpose_blogi	1.865e-01	3.024e-02	6.168	6.91e-10	***
grade_geri	-4.761e-01	1.132e-02	-42.064	< 2e-16	***
grade_blogi	2.391e-01	1.451e-02	16.479	< 2e-16	***
joint_application	-5.568e-01	2.179e-02	-25.552	< 2e-16	***

 signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 634757 on 779787 degrees of freedom
 Residual deviance: 585559 on 779763 degrees of freedom
 AIC: 585609

17 pav. Logistinės regresijos modelis iš atrinktų kintamųjų

Modelio lygtis:

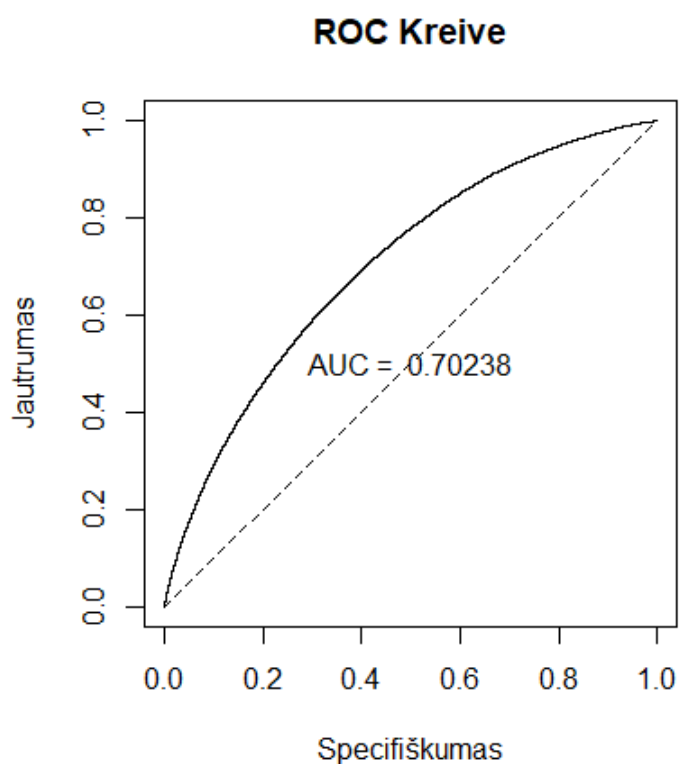
$$P = -3,049 + 0,007 * term + 0,043 * int_rate + 0,0004 * installment + 0,093 * inq_last_6mths - 0,02 * open_acc - 0,000002 * revol_bal + 0,058 * acc_open_past_24mths - 0,028 * mort_acc + 0,028 * num_actv_rev_tl + 0,07 * pub_rec_bankruptcies + 0,293 * unemployed - 0,007 * empl_years - 0,000003 * income + 0,237 * verified_flag + 0,005 * dti_max - 0,1 * home_mortgage + 0,153 * home_rent - 0,256 * state_geri + 0,156 * state_blogi - 0,027 *$$

$purpose_geri + 0,187 * purpose_blogi - 0,476 * grade_geri + 0,239 * grade_blogi - 0,557 * joint_application$

Norint sužinoti ar modelis yra toks pat tikslus ir analizei nenaudotiems duomenims pritaikome jį validacijos imčiai. Rezultatai lyginant testavimo ir validavimo imtis yra lentelėje apačioje, o ROC kreivė 18-tame paveiksle.

10 lentelė. 2015–2017 metų logistinės regresijos modelio rezultatai testavimo ir validavimo imtims

Duomenų imtis	MSE	AUC	Somers' D
Testavimas	0,1130	0,704	0,4090
Validavimas	0,1130	0,702	0,4054



18 pav. ROC kreivė logistinės regresijos modeliui 2015–2017 metams (Validavimo duomenys)

Iš rezultatų akivaizdu, kad modelis yra toks pat tikslus ir testavimo ir validavo imtims. Jį laikome geriausiu logistinės regresijos modeliu įmonės *Lending Club* kredito įsipareigojimų nevykdymo prognozavimui.

3.2. Neuroniniai tinklai

Kuriant neuroninių tinklų modelius, procesas yra kur kas sudėtingesnis nei naudojant logistinę regresiją. Be kintamųjų sąrašo reikia nustatyti neuronų ir jų sluoksnių skaičių, kuris randamas tik bandymų būdu. Dėl neuroninių tinklų sudėtingumo, pirmiausia išbandome mažesnius modelius stebėdami rezultatus.

Modeliui sukurti naudojamas H2O paketas, kuris geba apdoroti didelius duomenų rinkinius ir turi daugybę parametrų modeliams koreguoti. Į modelį dedami 24 atrinkti aiškinamieji kintamieji ir padaromas nedidelis 2 lygių neuroninis tinklas abiejuose lygiuose turintis po 20 neuronų. Numatytas apmokymų skaičius yra 10 kartų. Tai reiškia, kad modelis apmokomas 10 kartų stebint jo tikslumą po kiekvieno apmokymo.

```
H2ORegressionModel: deeplearning
Model ID: mod_1
Status of Neuron Layers: predicting bad_flag, regression, gaussian distribution,
Quadratic loss, 941 weights/biases, 17.4 KB, 10,596,234 training samples, mini-
batch size 1
```

	layer	units	type	dropout	l1	l2	mean_rate	rate_rms	momentum
1	1	24	Input	0.00 %	NA	NA	NA	NA	NA
2	2	20	Rectifier	0.00 %	0.000000	0.000000	0.004401	0.005784	0.000000
3	3	20	Rectifier	0.00 %	0.000000	0.000000	0.010366	0.016352	0.000000
4	4	1	Linear	NA	0.000000	0.000000	0.000779	0.000937	0.000000

```
mean_weight weight_rms mean_bias bias_rms
1 NA NA NA NA
2 -0.006222 0.706672 -1.083065 0.413410
3 -0.098763 0.228012 0.229760 0.400936
4 -0.208970 0.357195 0.856383 0.000000
```

```
H2ORegressionMetrics: deeplearning
** Reported on training data. **
** Metrics reported on temporary training frame with 10004 samples **

MSE: 0.1163246
RMSE: 0.3410639
MAE: 0.2264888
RMSLE: 0.2395413
Mean Residual Deviance : 0.1163246
```

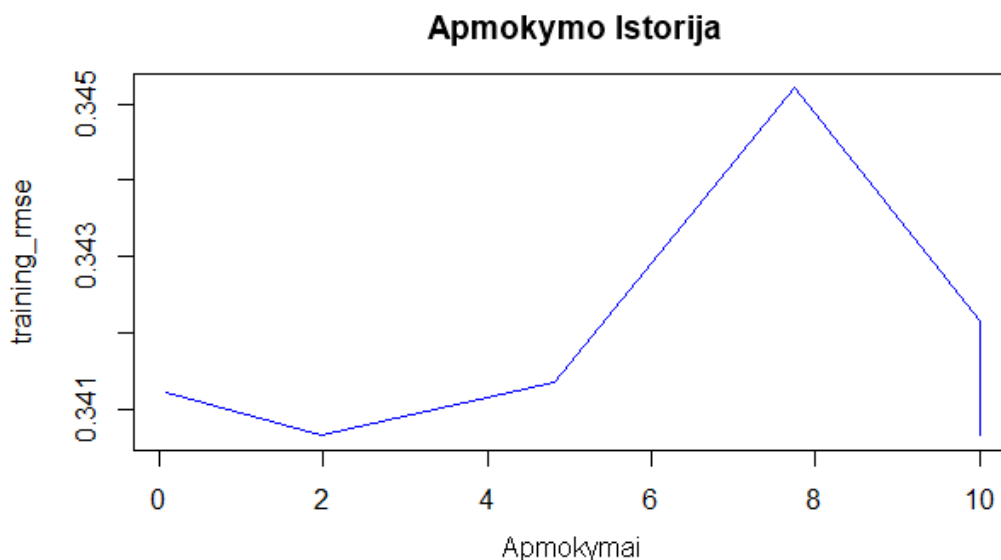
19 pav. Neuroninių tinklų 2 lygių po 20 neuronų modelio rezultatai

Modelio rezultatai pateikia informaciją apie modelio struktūrą ir paskaičiuotas paklaidų statistikas. Kadangi pastarosios yra apskaičiuotos naudojant apmokymo duomenis, to neužtenka tikslumui vertinti ir statistikos apskaičiuojamos testavimo duomenų imčiai. Šio modelio vidutinė standartinė paklaida yra lygi 0,1185, AUC yra 0,692, o Somers' D 0,383.

Norint surasti geriausią modelį pirmiausia pradedama nuo modelio parinkčių, kurios nekeičia neuronų kiekio ir tik atrinkus tinkamiausius parametrus ieškomas tinkamas neuronų ir jų sluoksnių skaičiaus.

Kuriant dirbtinį neuroninį tinklą galima pasirinkti kiek kartų modelis bus apmokomas. Po kiekvieno apmokymo yra stebimos modelio paklaidos ir jeigu modelio tikslumas pradeda blogėti ir nepakinta per tam tikrą apmokymų skaičių, apmokymas sustabdomas. Modelio tikslumas stebimas per vidutinius paklaidų nuokrypius ir kintamą vidurkį.

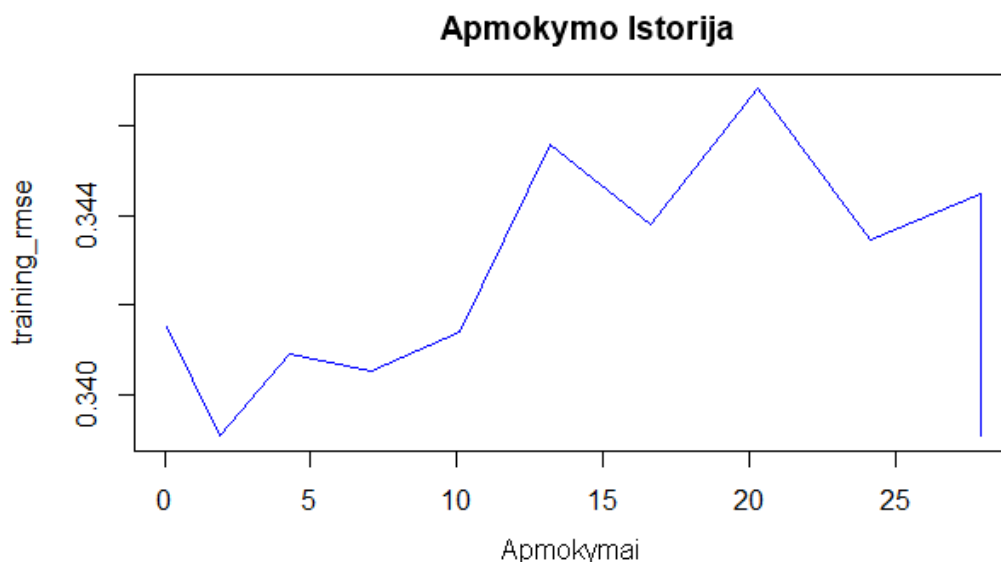
Sekančiame paveiksle pavaizduota prieš tai aptarto modelio apmokymo istorija. Grafike turime apmokymų skaičių ant x ašies ir šaknį iš vidutinės kvadratinės paklaidos (RMSE) ant y ašies. Kadangi nėra pasirinkimo stebėti MSE, todėl naudojamas RMSE tikslumui vertinti. Pastebime, kad modelio tikslumas buvo panašus pradedant apmokymų ciklą, tačiau pradėjo blogėti po penkių apmokymų.



20 pav. Neuroninių tinklų 2 lygių po 20 neuronų 10 apmokymų istorija

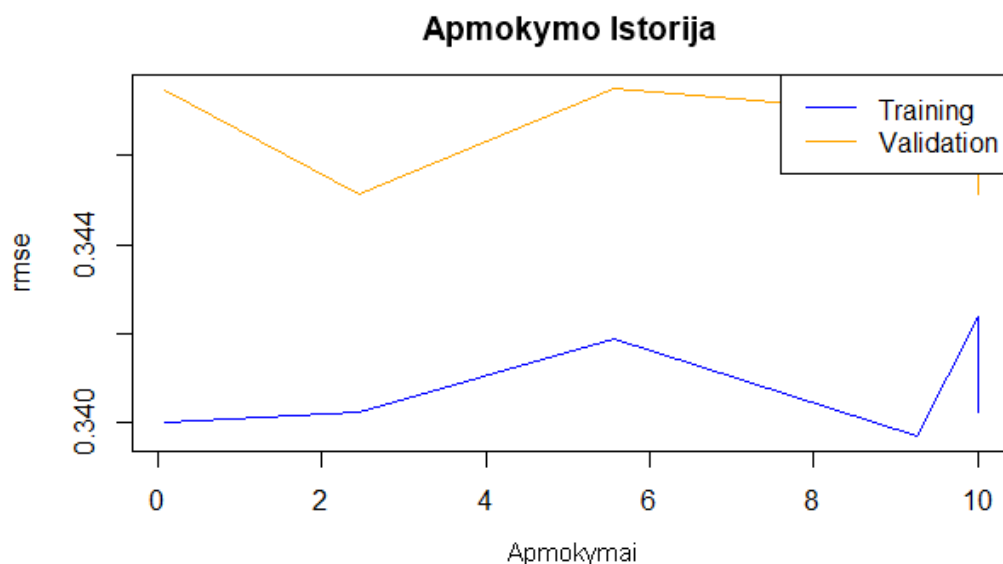
Iš pirmo žvilgsnio atrodo, kad optimalus apmokymų skaičius yra 5, tačiau penkių ir dešimties apmokymų paklaidos skiriasi nežymiai atsižvelgiant į skaitinę reikšmę. Tikėtina, kad atlikus daugiau apmokymų modelio paklaidos pradėtų mažėti, todėl sukuriamas kitas modelis apmokant 50 kartų. Šio modelio vidutinės kvadratinės paklaidos yra lygios 0,1182, AUC = 0,689, o Somers' D = 0,39.

Kaip matome iš 21 paveiksle esančio grafiko, modelis buvo sustabdytas ties 28 apmokymu, nes jo tikslumas tik blogėjo. Be kita ko, akivaizdu, kad 10 apmokymų yra pakankamas skaičius, todėl nuo šiol būtent tiek apmokymų bus naudojama kuriamiems modeliams. Tikslus apmokymų skaičių bus nuspręstas sudarant galutinį modelį.



21 pav. Neuroninių tinklų 2 lygių po 20 neuronų 50 apmokymų istorija

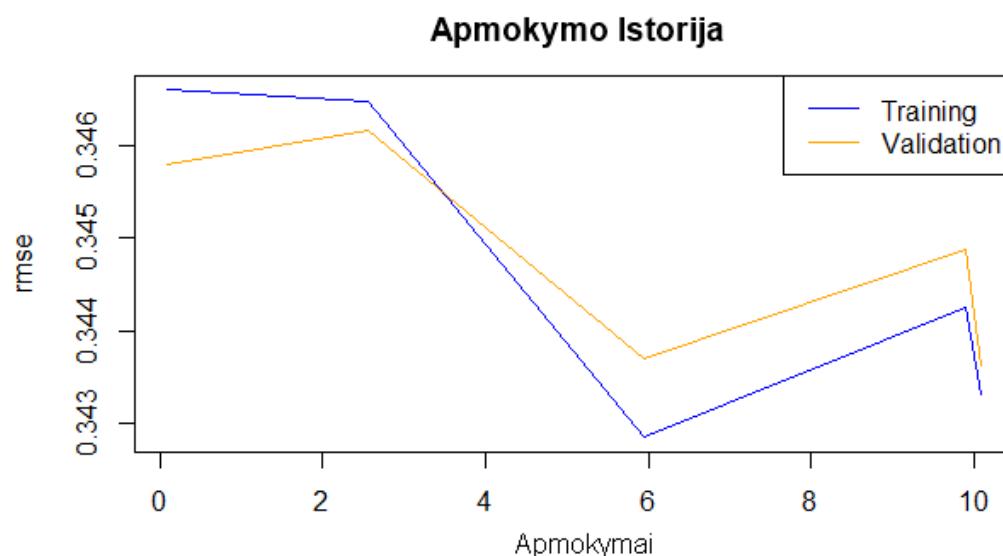
Kita parinktis, kurią galima naudoti neuroniniam tinklui, yra kryžminis patikrinimas. Šis metodas suskaido apmokymo aibę į poaibius, kur 80 % imties įrašų naudojami apmokymui, o rezultatai tikrinami su likusiais 20 % įrašų. Tokių modelių sukurama daugiau nei vienas ir kiekviename vis kiti 20 % apmokymo aibės duomenys naudojami testavimui. Galima pasirinkti norima skaičių modelių, tačiau dažniausiai naudojama nedaugiau 10. Galutinis modelis yra vidurkis visų kurtų modelių, kuris rengiamas naudojant visus apmokymo duomenis.



22 pav. Neuroninių tinklų 2 lygių po 20 neuronų apmokymų istorija su 5 imčių kryžmine patikra

Sukurtas modelis iš penkių kryžminės patikros modelių turi $MSE = 0,1189$, $AUC = 0,692$ ir Somers' $D = 0,385$. Naudojant kryžminį patikrinimą, apmokymo istorijos grafike brėžiami ir naudotų testavimo imčių rezultatai. Deja iš grafiko matosi, kad testavimui naudotų duomenų paklaidos yra didesnės už apmokymo duomenų paklaidas (22 pav.).

Pabandžius dešimties imčių kryžminę patikrą gauname panašius rezultatus, tačiau 3 imčių kryžminės patikros grafikas atrodo geriau (23 pav.)



23 pav. Neuroninių tinklų 2 lygių po 20 neuronų apmokymų istorija su 3 imčių kryžmine patikra

Visų bandymų rodiklių reikšmės yra pateiktos 11 lentelėje. Pastebime, kad kryžminis patikrinimas nepagerina modelio tikslumo, todėl jis nebenaudojamas. Kalbant apie apmokymų skaičių, grafikai rodo, kad jis turėtų būti žemesnis nei 10, tačiau Somers' D įvertis yra didžiausias, kai pasirenkame 50 apmokymų. Dėl šių prieštaringų rezultatų, neuronų skaičius renkamas naudojant 10 apmokymų, bet radus tinkamiausią rinkinį, modelis bus perskaičiuojamas didesniau apmokymų skaičiui.

Norint suprasti kokio tipo dirbtinis neuroninis tinklas geriausiai aprašo duomenis, išbandomi skirtingi modeliai keičiant lygių ir neuronų skaičių.

12-toje lentelėje turime rezultatus 13 neuroninių tinklų modelių su skirtingu sluoksnių ir neuronų skaičiumi. Čia lygių ir neuronų skaičius apibrėžtas laužtiniuose skliaustuose. Skaičiai

11 lentelė. Neuroninių tinklų modeliai 2 lygių po 20 neuronų

Modelio parametrai		Modelio rezultatai		
Kryžminė patikra	Apmokymų skaičius	MSE	AUC	Somers' D
-	10	0,1185	0,692	0,383
-	50	0,1182	0,689	0,390
5	10	0,1189	0,692	0,385
3	10	0,1188	0,688	0,372
10	10	0,1193	0,688	0,348

nurodo neuronų kiekį kiekviename lygmenyje, o skaičių kiekis atitinka lygių skaičių. Pavyzdžiui [20, 20, 20] apibrėžia 3 lygių neuroninį tinklą, kurio kiekvienas lygis turi po 20 neuronų.

12 lentelė. Skirtingų neuroninių modelių rezultatai, pirmieji bandymai

Modelio parametrai	Modelio rezultatai		
Lygių ir Neuronų skaičius	MSE	AUC	Somers' D
[20, 20]	0,1185	0,692	0,383
[20, 10]	0,1185	0,687	0,375
[50, 50]	0,1182	0,694	0,389
[100, 100]	0,1183	0,694	0,388
[200, 200]	0,1185	0,687	0,373
[300, 300]	0,1180	0,693	0,387
[20, 20, 20]	0,1177	0,695	0,389
[50, 50, 50]	0,1177	0,698	0,397
[100, 100, 100]	0,1179	0,695	0,389
[200, 200, 200]	0,1180	0,697	0,393
[20, 20, 20, 20]	0,1176	0,696	0,393
[200]	0,1181	0,694	0,387
[500]	0,1183	0,699	0,397

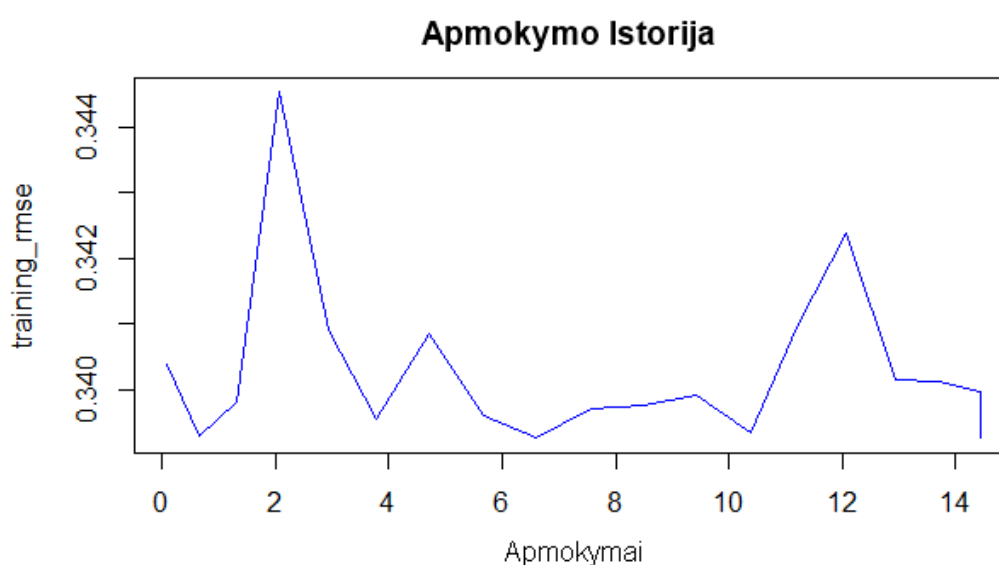
Iš 12 lentelės duomenų matome, kad mažiausios vidutinės kvadratinės paklaidos yra trijų ir keturių sluoksnių neuroniniuose tinkluose, o geriausi AUC ir Somers' D įvertiniai yra modeliuose [50, 50, 50] ir [500]. Taip pat galima pastebėti, kad modeliai su 50 neuronų sluoksnyje yra tinkamiausi kai naudojami 2 ar 3 jų sluoksniai. Vieno lygio modelis su 500 neuronų taip pat rodo gerus rezultatus, todėl išbandomi kiti modeliai keičiant sluoksnių skaičių ir išlaikant 50 ir 500 neuronų.

Tryliktoje lentelėje atvaizduoti naujų modelių rezultatai. Pastebime, kad bendrai jie yra geresni nei prieš tai buvę modeliai, nes visų Somers' D įvertinys yra aukštesnis nei 0,39, tačiau geriausias modelis su 5 sluoksniais po 50 neuronų yra toks pat geras kaip 11 lentelėje esantis modelis su 3 sluoksniais po 50 neuronų. Kadangi sudėtingesnis modelis neprideda vertės, pasirenkame paprastesnį modelį.

Atkartojamas geriausias modelis [50, 50, 50], tik šįkart vietoje 10 apmokymų naudojama 50. Šio modelio tikslumo įvertiniai šiek tiek žemesni: MSE=0,1177, AUC=0,696, o Somers' D=0,391. Peržiūrint apmokymų istoriją (24 pav.), pastebime, kad modelis buvo sustabdytas po 14 apmokymų ir mūsų anksčiau pasirinktas apmokymų skaičius 10 iš tiesų yra optimalus.

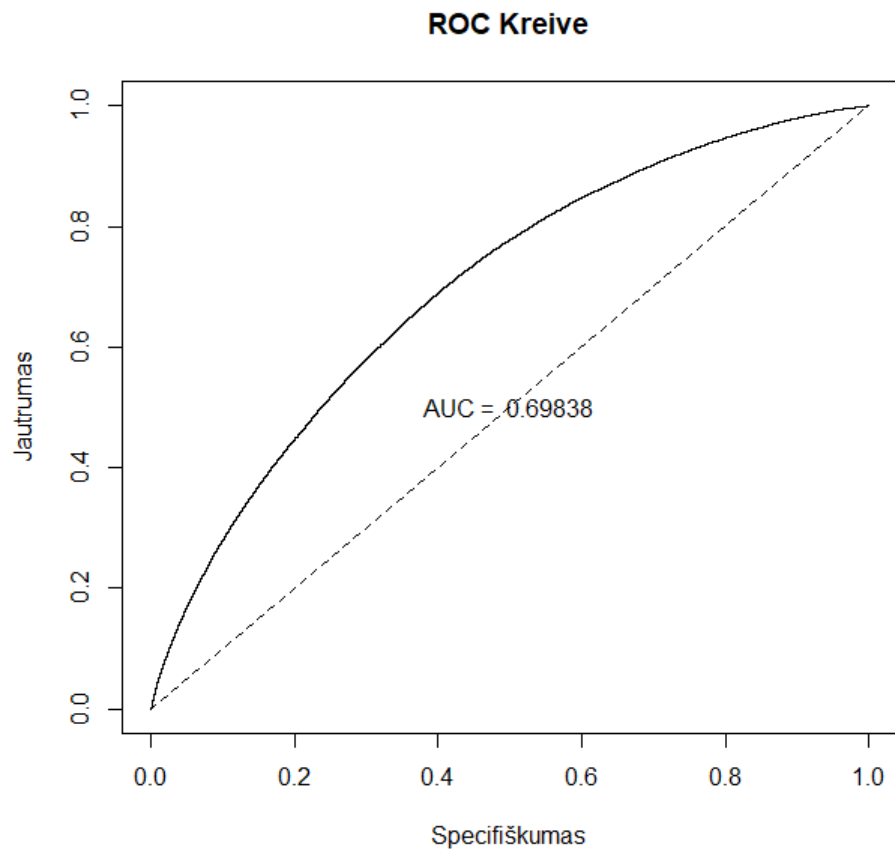
13 lentelė. Skirtingų neuroninių modelių rezultatai, antrieji bandymai

Modelio parametrai Lygių ir Neuronų skaičius	Modelio rezultatai		
	MSE	AUC	Somers' D
[50]	0,1179	0,695	0,390
[50, 50, 50, 50]	0,1178	0,697	0,391
[50, 50, 50, 50, 50]	0,1178	0,698	0,397
[500, 500]	0,1178	0,697	0,393
[500, 500, 500]	0,1178	0,697	0,395
[500, 500, 500, 500]	0,1178	0,698	0,395
[500, 50, 50]	0,1178	0,695	0,390
[500, 50, 50, 50]	0,1180	0,697	0,393
[500, 50, 50, 50, 50]	0,1178	0,697	0,395

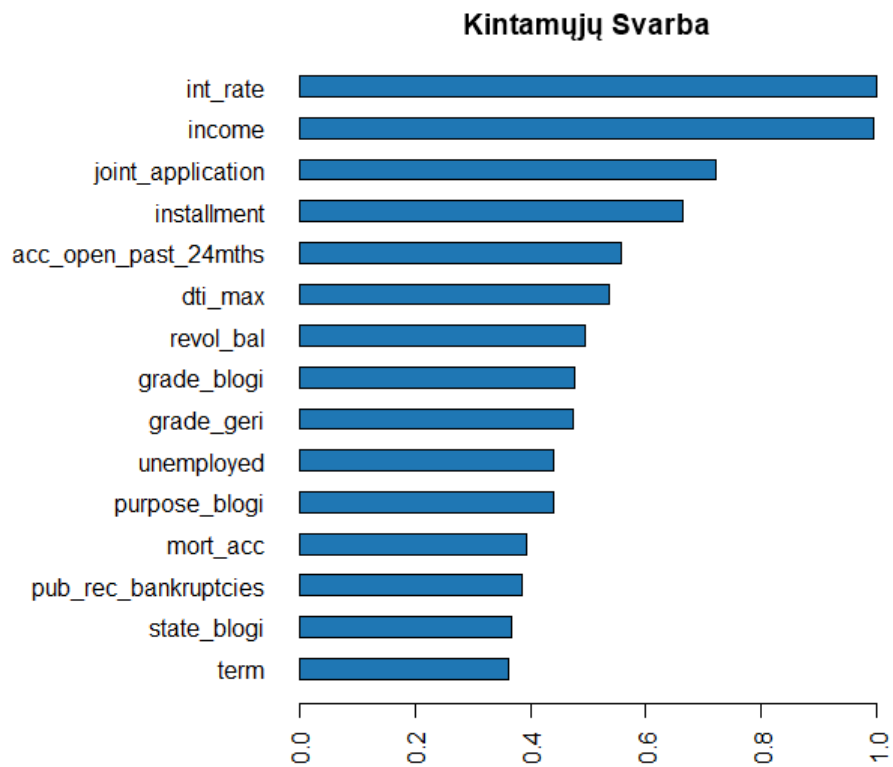


24 pav. Neuroninio tinklo [50, 50, 50] su 50 ciklų apmokymo procesas

Taigi, geriausiu modeliu parenkamas dirbtinis neuroninis tinklas turintis 3 lygius po 50 neuronų naudojant 10 apmokymo ciklų. Jo tikslumo įvertiniai yra geresni ar bent neblogesni už likusių modelių tiek 12-toje, tiek 13-toje lentelėse, o modelio sudėtingumas yra mažiausias iš panašų tikslumą turinčių modelių. Žemiau esančiuose paveiksluose turime šio modelio ROC kreivę ir geriausių kintamųjų sąrašą.



25 pav. Neuroninio tinklo [50, 50, 50] ROC kreivė



26 pav. Neuroninio tinklo [50, 50, 50] svarbiausi kintamieji

Dirbtiniame neuroniniame tinkle kintamieji yra išrikiuojami ir pateikiami pagal svarbą skaleje [0, 1]. Penkiolika svarbiausių kintamųjų pateikiami 26-tame paveiksle. Iš jo pastebima, kad

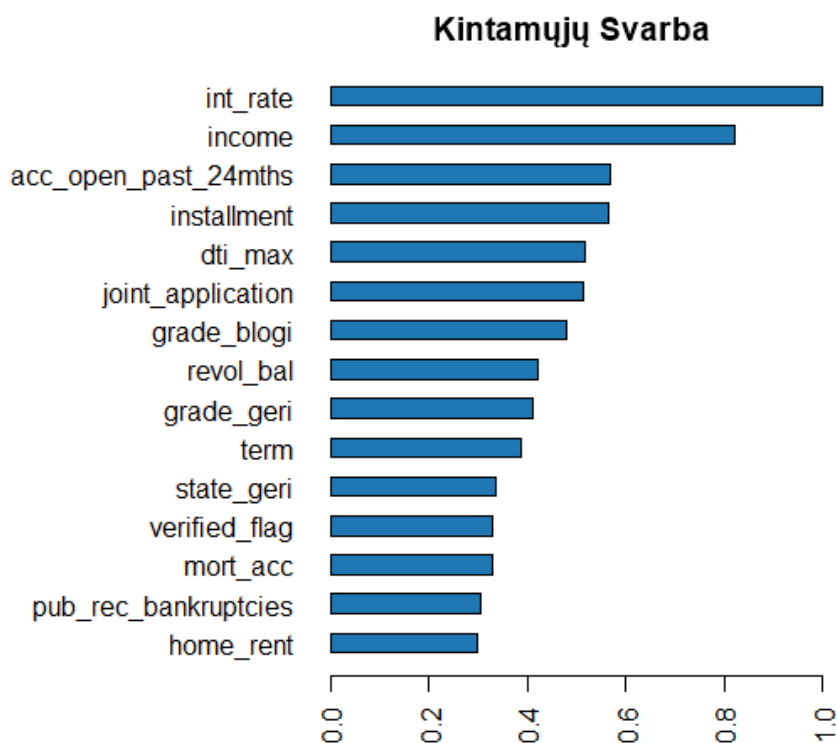
geriausiai negebėjimą gražinti paskolas prognozuoja palūkanų norma ir pajamos. Prisimenant regresijos analizės rezultatus, žinome, jog mažos pajamos ir didelės palūkanos yra susiję su didesne tikimybe, kad paskola bus negražinta. Palūkanų dydis priklauso ne nuo skolininko, o nuo tarpusavio skolinimo įmonės, todėl akivaizdu, kad jas nustatant atsižvelgiama į daugiau nei asmens reitingą, nes reitingas pagal svarbą yra įvertintas vidutiniškai.

Laikantis požiūrio, kad paskolų krepšelis su laiku keičiasi, skaičiuojamas atrinkto dirbtinio neuroninio modelio tikslumas mažinant naudojamų metų skaičių. Modelių rezultatai pateikti 14-toje lentelėje.

14 lentelė. Neuroninių tinklų rezultatai skirtingiems laiko intervalams

Metams nuo:	MSE	AUC	Somers' D
2007	0,1177	0,6984	0,3969
2008	0,1177	0,6983	0,3945
2009	0,1180	0,6955	0,3910
2010	0,1176	0,6975	0,3957
2011	0,1175	0,6994	0,3983
2012	0,1177	0,6959	0,3920
2013	0,1171	0,7000	0,3993
2014	0,1165	0,7007	0,4007
2015	0,1131	0,7027	0,4056
2016	0,1014	0,7007	0,4015

Vėlgi pasikartoja situacija, kai modeliai naujesniems metams yra tikslesni nei modelis visam laiko intervalui. Nors tikslumo rodiklių didėjimas nėra tolygus kaip stebėtas logistinės regresijos modeliams, tačiau jis yra akivaizdus. Geriausiu parenkame dirbtinių neuroninių tinklų modelį 2015–2017 metams, kurio svarbiausių kintamųjų sąrašą turime 27-tame paveiksle.



27 pav. Neuroninio tinklo 2015–2017 metų duomenims svarbiausi kintamieji

Galima pastebėti, kad sumažinus laiko intervalą, svarbiausiu kintamuoju tampa palūkanos, kai tuo tarpu 2007–2017 metų modelyje palūkanos turėjo tokią pat įtaką kaip ir pajamos. Kaip jau buvo kalbėta, palūkanas nustato tarpusavio skolinimo įmonė, todėl jos yra priklausomos nuo tam tikros informacijos apie skolininką. Vis dėlto, ryšiui gali daryti įtaką ir pats palūkanų dydis, lemiantis mažėjančią motyvaciją mokėti paskolą. Kuo palūkanos yra didesnės, tuo didesnę įmokos dalį sudaro pinigai nedengiantys paskolos, o tai didina skolos našta.

Dirbtinių neuroninių tinklų modelį pritaikome validavimo duomenų rinkiniui ir stebime prognozės tikslumą (15 lentelė).

15 lentelė. 2015–2017 metų neuroninių tinklų modelio rezultatai testavimo ir validavimo imtims

Duomenų imtis	MSE	AUC	Somers' D
Testavimas	0,1131	0,7027	0,4056
Validavimas	0,1130	0,7006	0,4018

AUC ir Somers' D reikšmės šiek tiek nukrenta, tačiau nežymiai. Kadangi tiksliausias modelis yra 3 sluoksnių po 50 neuronų, kai jis kuriamas įtraukiant tik paskutinius 3 metus (2015–2017), o modelio tikslumas nenukenčia validavimo duomenų rinkiniui, šį modelį laikome geriausiu dirbtinių neuroninių tinklų modeliu įmonės *Lending Club* kredito išipareigojimų nevykdymo prognozavimui.

Lyginant geriausią neuroninių tinklų modelį su logistinės regresijos rezultatais, matome, kad logistinė regresija prognozuoja nevykdomas paskolas šiek tiek tiksliau, bet visgi panašiai kaip neuroniniai tinklai (16 len.)

16 lentelė. Geriausi logistinės regresijos ir neuroninių tinklų modeliai

Modelis	MSE	AUC	Somers' D
Logistinė regresija	0,1130	0,702	0,4054
Neuroniniai tinklai	0,1130	0,701	0,4018

IŠVADOS IR REZULTATAI

Atlikus duomenų analizę nustatyta, kad vienos paskolos gali būti geresnės už kitas vertinant grąžintų paskolų dalį. Pajamos, skolos ir pajamų santykis ir net informacija apie gyvenamąjį būstą gali būti naudinga vertinant kredito riziką. Tačiau ne tik skolininko asmeninė informacija yra svarbi – rizika priklauso ir nuo paskolos parametrų. Jos tikslas, terminas ir paskirtų palūkanų dydis taip pat yra tarp kintamųjų, kurie daro reikšmingą įtaką. Ši ir kita informacija yra svarbi modeliuojant kredito įsipareigojimų nevykdymo prognozę, kuriai naudota logistinė regresija ir dirbtiniai neuroniniai tinklai.

Sudarant modelius nustatyta, kad logistinės regresijos ir neuroninių tinklų modelių rezultatai yra panašūs, tačiau logistinė regresija yra šiek tiek tikslesnė. Logistinės regresijos modelis, sukurtas 2007–2017 metų intervalui, turi AUC lygų 0,699 ir Somers' D 0,398. Atitinkamas neuroninių tinklų modelis turi AUC 0,698 ir Somers' D 0,397. Kadangi duomenų rinkinys yra net 10 metų intervalui, tikėtina, kad žmonių elgesys kinta ir senesni metai neatitinka dabartinės situacijos. Naudojant trejų metų intervalą (2015–2017), išlaikomi 73,6 % duomenų ir paliekamos labiausiai šias dienas atitinkančios paskolos. Pokytis turi teigiamą įtaką modelių tikslumui ir logistinės regresijos AUC reikšmė pakyla iki 0,702, o Somers' D 0,405, o neuroninių tinklų rezultatai atitinkamai yra 0,701 ir 0,402.

Lyginant modelių sudėtingumą akivaizdu, kad logistinė regresija yra lengviau paaiškinama ir jos sudarymas yra paprastesnis ir greitesnis. Randant geriausią logistinės regresijos modelį buvo sudaryti 10 modelių, kurių didžioji dalis buvo skirta tinkamam laiko intervalui nustatyti. Ieškant geriausio neuroninių tinklų modelio buvo sudaryti 37 modeliai, kurių didžiąją dalį sudaro skirtingų neuronų ir jų sluoksnių skaičiaus bandymai. Taigi galima teigti, jog ne tik dėl tikslumo, bet ir dėl paprastumo logistinės regresijos modelis yra tinkamesnis prognozuojant kredito įsipareigojimų tiesioginio skolinimo įmonei nevykdymą.

LITERATŪRA

1. Amery, A. (2018). A Brief History of Loans: Business Lending Through the Ages. Prieiga per internetą: <http://www.lending-express.com/blog/a-brief-history-of-loans-business-lending-through-the-ages/>
2. Makuch, William M. (2001). Scoring Applications. Iš Mays, Elizabeth, *Handbook of Credit Scoring* (p. 3-6). Čikaga: Glenlake Publishing Company.
3. Gruenstein, John M. L. (2001). Optimal Use of Statistical Techniques in Model Building. Iš Mays, Elizabeth, *Handbook of Credit Scoring* (p. 149–168). Čikaga: Glenlake Publishing Company.
4. Galloway, I. (2009/2010). Peer-to-Peer Lending and Community Development Finance. *Community Investments* 21(3), 18–19.
5. Lin, M. (2009, rugpjūtis). *Peer-to-Peer Lending: An Empirical Study*. Pranešimas konferencijoje: AMCIS 2009 Doctoral Consortium, San Franciskas, Kalifornija.
6. Iyer, R., Khwaja, A. I., Luttmer, E. F. P. ir Shue, K. (2010). Screening in New Credit Markets. Can Individual Lenders Infer Borrower Creditworthiness in Peer-to-Peer Lending?. *AFA Denver Meetings*, 1–27.
7. Park, H. (2013). An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain. *J Korean Acad Nurs* 43(2), 154-157. Prieiga per internetą: <http://synapse.koreamed.org/Synapse/Data/PDFData/0006JKAN/jkan-43-154.pdf>
8. Peng, C.-Y. J., Lee, K. L. ir Ingersoll G. M. (2002). An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research* 96(1), 3–5.
9. Gurney, K. (1997). *An Introduction to Neural Networks* (p. 14–15). Londonas: UCL Press.
10. Haykin, S. (2009). *Neural Networks and Learning Machines* (3-oji laida) (p. 1–14). Jungtinės Amerikos Valstijos: Pearson Education.
11. Hastie, T., Tibshirani, R. (2016). *SLDM IV: Deep Learning in H2O*. Prieiga per internetą: <http://htmlpreview.github.io/?https://github.com/ledell/sldm4-h2o/blob/master/sldm4-deeplearning-h2o.html>
12. Ling, C. X., Huang, J., Zhang, H. (2003). *AUC: a Statistically Consistent and more Discriminating Measure than Accuracy*. Prieiga per internetą: http://www.researchgate.net/publication/2855551_AUC_a_Statistically_Consistent_and_more_Discriminating_Measure_than_Accuracy
13. Somers, R. H. (1962). *A new asymmetric measure of association for ordinal variables*. American Sociological Review.

PRIEDAI

Priedas A: Regresinės analizės rezultatai

17 lentelė. Kintamųjų sąrašas su regresinės analizės rezultatais

	kintamasis	T statistika	P reikšmė
1:	loan_amnt	45.0419	0
2:	term	131.6146	0
3:	int_rate	279.5582	0
4:	installment	48.94794	0
5:	inq_last_6mths	105.026	0
6:	open_acc	28.52878	5.150117e-179
7:	revol_bal	-32.69157	2.057634e-234
8:	total_acc	9.699559	3.028037e-22
9:	acc_now_delinq	3.671592	0.0002410439
10:	inq_fi	7.457359	8.827402e-14
11:	inq_last_12m	11.09289	1.358304e-28
12:	acc_open_past_24mths	129.6827	0
13:	mort_acc	-69.65076	0
14:	num_actv_bc_tl	31.14139	6.633523e-213
15:	num_actv_rev_tl	63.53672	0
16:	num_bc_sats	10.68734	1.166663e-26
17:	num_bc_tl	3.959678	7.505076e-05
18:	num_il_tl	13.22103	6.63501e-40
19:	num_op_rev_tl	33.98805	3.345673e-253
20:	num_rev_accts	14.75968	2.665375e-49
21:	num_rev_tl_bal_gt_0	62.8756	0
22:	num_sats	25.52278	1.101377e-143
23:	pub_rec_bankruptcies	35.82775	4.085409e-281
24:	unemployed	40.87152	0
25:	empl_years	-28.69741	4.109781e-181
26:	income	-76.84623	0
27:	verified_flag	103.9017	0
28:	desc_flag	8.347777	6.955965e-17
29:	dti_max	68.27278	0
30:	home_own	1.160472	0.2458568
31:	home_mortgage	-72.77642	0
32:	home_rent	73.69579	0
33:	home_other	-1.159619	0.2462039
34:	home_none	0.01242454	0.9900869
35:	joint_application	-33.457	2.03594e-245
36:	payment_cash	-0.2398252	0.8104658
37:	state_geri	-45.24518	0
38:	state_blogi	24.16443	5.265703e-129
39:	state_vid	28.3466	9.217596e-177
40:	purpose_geri	-57.3522	0
41:	purpose_vid	50.04418	0
42:	purpose_blogi	29.96764	2.591965e-197
43:	grade_geri	-232.3779	0
44:	grade_vid	114.4463	0
45:	grade_blogi	205.6749	0
	kintamasis	T statistika	P reikšmė

18 lentelė. Labiausiai koreliuoti kintamieji

	loan_amnt	installment	open_acc	num_actv_bc_tl	
loan_amnt	1.000000000	0.946690948	0.188671764	0.204929442	
installment	0.946690948	1.000000000	0.176582262	0.204844624	
open_acc	0.188671764	0.176582262	1.000000000	0.539072609	
num_actv_bc_tl	0.204929442	0.204844624	0.539072609	1.000000000	
num_actv_rev_tl	0.166670384	0.170194769	0.645446020	0.830701949	
num_op_rev_tl	0.176449623	0.172200579	0.793597505	0.698654175	
num_rev_tl_bal_gt_0	0.165850178	0.169305426	0.646436639	0.824102661	
state_geri	0.001924659	-0.000623540	0.002679715	-0.005316302	
state_vid	0.004278550	0.007217187	0.008006126	0.016450000	
purpose_geri	-0.026944158	-0.037616393	0.003306848	0.031426342	
purpose_vid	0.024241100	0.032218869	0.002726634	-0.022680857	
grade_geri	-0.103220930	-0.096434226	0.005067204	-0.026500345	
grade_vid	0.028351403	0.014046192	-0.014510273	0.022215237	
	num_actv_rev_tl	num_op_rev_tl	num_rev_tl_bal_gt_0	state_geri	
loan_amnt	0.166670384	0.176449623	0.165850178	0.001924659	
installment	0.170194769	0.172200579	0.169305426	-0.000623540	
open_acc	0.645446020	0.793597505	0.646436639	0.002679715	
num_actv_bc_tl	0.830701949	0.698654175	0.824102661	-0.005316302	
num_actv_rev_tl	1.000000000	0.822364900	0.983719813	-0.008859723	
num_op_rev_tl	0.822364900	1.000000000	0.826479651	-0.012874817	
num_rev_tl_bal_gt_0	0.983719813	0.826479651	1.000000000	-0.009864459	
state_geri	-0.008859723	-0.012874817	-0.009864459	1.000000000	
state_vid	0.017896905	0.023543325	0.019089597	-0.844676656	
purpose_geri	-0.003227506	0.008706989	-0.003235963	0.006822024	
purpose_vid	0.014077220	0.001226910	0.014338627	-0.006302074	
grade_geri	-0.080489125	-0.002412150	-0.081352810	0.005882179	
grade_vid	0.058560039	-0.001850797	0.058953987	-0.004181200	
	state_vid	purpose_geri	purpose_vid	grade_geri	grade_vid
loan_amnt	0.0042785498	-0.026944158	0.024241100	-0.103220930	0.0283514028
installment	0.0072171873	-0.037616393	0.032218869	-0.096434226	0.0140461917
open_acc	0.0080061264	0.003306848	0.002726634	0.005067204	-0.0145102730
num_actv_bc_tl	0.0164500000	0.031426342	-0.022680857	-0.026500345	0.0222152370
num_actv_rev_tl	0.0178969048	-0.003227506	0.014077220	-0.080489125	0.0585600389
num_op_rev_tl	0.0235433254	0.008706989	0.001226910	-0.002412150	-0.0018507966
num_rev_tl_bal_gt_0	0.0190895970	-0.003235963	0.014338627	-0.081352810	0.0589539868
state_geri	-0.8446766562	0.006822024	-0.006302074	0.005882179	-0.0041812005
state_vid	1.0000000000	-0.004346721	0.003562111	0.001142082	-0.0005050039
purpose_geri	-0.0043467214	1.000000000	-0.974193909	0.142705923	-0.0914495327
purpose_vid	0.0035621106	-0.974193909	1.000000000	-0.132766358	0.0885125128
grade_geri	0.0011420817	0.142705923	-0.132766358	1.000000000	-0.8272587173
grade_vid	-0.0005050039	-0.091449533	0.088512513	-0.827258717	1.0000000000