KAUNO TECHNOLOGIJOS UNIVERSITETAS

MATEMATIKOS IR GAMTOS MOKSLŲ FAKULTETAS

Ieva Palevičiūtė

FIZINIŲ ASMENŲ PASKOLŲ GRĄŽINIMO GALIMYBIŲ VERTINIMAS: TARPUSAVIO SKOLINIMOSI PLATFORMŲ ATVEJIS

Baigiamasis magistro projektas

Vadovai
Doc. dr. Aura Drakšaitė
Doc. dr. Evaldas Vaičiukynas

KAUNAS, 2017

KAUNO TECHNOLOGIJOS UNIVERSITETAS

MATEMATIKOS IR GAMTOS MOKSLŲ FAKULTETAS

FIZINIŲ ASMENŲ PASKOLŲ GRĄŽINIMO GALIMYBIŲ VERTINIMAS: TARPUSAVIO SKOLINIMOSI PLATFORMŲ ATVEJIS

Baigiamasis magistro projektas

Didžiųjų verslo duomenų analitika (621G12002)

Vadovai

(parašas) Doc. dr. Aura Drakšaitė

(data)

(parašas) Doc. dr. Evaldas Vaičiukynas

(data)

Recenzentai

(parašas).

(data)

(parašas)

(data)

Projektą atliko

(parašas) Ieva Palevičiūtė

(data)

KAUNAS, 2017

KAUNO TECHNOLOGIJOS UNIVERSITETAS

Matematikos ir gamtos mokslų fakultetas
(Fakultetas)

Ieva Palevičiūtė
(Studento vardas, pavardė)

Didžiųjų verslo duomenų analitika, 621G12002
(Studijų programos pavadinimas, kodas)

„FIZINIŲ ASMENŲ PASKOLŲ GRĄŽINIMO GALIMYBIŲ VERTINIMAS: TARPUSAVIO SKOLINIMOSI PLATFORMŲ ATVEJIS“

AKADEMINIO SĄŽININGUMO DEKLARACIJA

20 ___ m. _____ ____ d.
Kaunas

Patvirtinu, kad mano, **Ievos Palevičiūtės**, baigiamasis projektas tema „FIZINIŲ ASMENŲ PASKOLŲ GRĄŽINIMO GALIMYBIŲ VERTINIMAS: TARPUSAVIO SKOLINIMOSI PLATFORMŲ ATVEJIS“ yra parašytas visiškai savarankiškai ir visi pateikti duomenys ar tyrimų rezultatai yra teisingi ir gauti sąžiningai. Šiame darbe nei viena dalis nėra plagijuota nuo jokių spausdintinių ar internetinių šaltinių, visos kitų šaltinių tiesioginės ir netiesioginės citatos nurodytos literatūros nuorodose. Įstatymų nenumatytų piniginių sumų už šį darbą niekam nesu mokėjus.

Aš suprantu, kad išaiškėjus nesąžiningumo faktui, man bus taikomos nuobaudos, remiantis Kauno technologijos universitete galiojančia tvarka.

_____ _____
(vardą ir pavardę įrašyti ranka) *(parašas)*

TURINYS

Ieva, Palevičiūtė. FIZINIŲ ASMENŲ PASKOLŲ GRĄŽINIMO GALIMYBIŲ VERTINIMAS: TARPUSAVIO SKOLINIMOSI PLATFORMŲ ATVEJIS. Magistro baigiamasis projektas / vadovai doc. Evaldas Vaičiukynas; doc. Aura Drakšaitė. Kauno technologijos universitetas, Matematikos ir gamtos mokslų fakultetas.

Mokslo kryptis ir sritis: Didžiųjų verslo duomenų analitika:

Reikšminiai žodžiai: Tarpusavio skolinimas, skaitmeninis mokymas, duomenų mokslas.

Kaunas, 2017. 58 p.

SANTRAUKA

Šio darbo tikslas įvertinti fizinių asmenų paskolų grąžinimo galimybes tarpusavio skolinimosi platformų atveju. Norint pasiekti šį tikslą pirmiausia yra apžvelgiamos tarpusavio skolinimosi platformų tendencijos, apžvelgiami esami tyrimai siekiant įvertinti galimybes grąžinti paskolas, atliekama tinkamų šiai problemai skaitmeninio mokymosi metodų apžvalga. Tiriamojoje dalyje atliekama duomenų aprašomoji analizė naudojantis „Bondora“ ir „Lending Club“ tarpusavio skolinimo platformų duomenis, šiems duomenų rinkiniams pritaikomi skaitmeninio mokymo metodai siekiant nustatyti ar vartotojas grąžins paskolą.

Duomenų rinkiniams buvo pritaikyti šie skaitmeninio mokymosi metodai – logistinė regresija, vidurkinis perceptronas, atraminiai vektoriai, gilieji atraminiai vektoriai, sustiprintų sprendimų medis, sprendimų miškas, sprendimų džiunglės, Bajeso taškas. Tyrimo metu buvo nustatyta, kad tiksliausią prognozę ar klientas grąžins paskolą gauname naudodamiesi sustiprintų sprendimų medžio metodą.

Palevičiūtė, Ieva. NATURAL PERSONS LOAN REPAYMENT CAPABILITY ASSESSMENT: THE CASE OF PEER-TO-PEER LENDING PLATFORMS: Master‘s thesis / supervisors assoc. prof. Evaldas Vaičiukynas, assoc. prof. Aura Draksaitė. The Faculty of Mathematics and Natural Science, Kaunas University of Technology.

Research area and field: Business Big Data Analytics

Key words: peer to peer lending, machine learning, data science.

Kaunas, 2017. 58 p.

SUMMARY

The aim of this thesis is to evaluate natural persons’ loan repayment capability in case of peer to peer lending platforms. In order to reach this goal, general tendencies of peer to peer lending platforms were overviewed, the analysis of existing literature on capabilities to repay loans was carried on, and machine learning models that can be used in solving this problem were overviewed. In the research part of the thesis, explanatory analysis of peer to peer lending platforms „Bondora“ and „Lending Club“ datasets was carried on. There were machine learning algorithms applied to evaluate if a person can repay the loan or not.

The following machine learning methods were applied for the datasets – logistic regression, averaged perceptron, support vector machine, deep supports vector machine, boosted decision tree, decision forest, decision jungle, Bayes Point machine, neural nets. According to the research that was carried on, the most accurate prediction for both datasets was reached using boosted decision tree model.

1. ĮŽANGA

Tarpusavio skolinimosi platformų kūrimasis, sudėtingas paskolų gavimo procesas, didelės palūkanos bankuose paskatino vartotojus paskolas imti naudojantis tarpusavio skolinimosi platformomis. Investuotojai ieškodami palankių investavimo sąlygų atrado nerizikingą būdą gauti dideles palūkanas investuojant į paskolas. Tačiau geras paskolą imančių fizinių asmenų kreditingumo balas ne visuomet lemia tai, kad paskola bus grąžinta. Tarpusavio skolinimosi sistemų kūrėjai vis dažniau pastebi, kad paskolas imantiems klientams nustatytas kredito balas ne visuomet atitinka tikimybę grąžinti paskolą.

Tobulėjant skaitmeninio mokymosi algoritmams ir populiarėjant įžvalgų ir prognozavimo analitikai, tarpusavio skolinimosi platformų kūrėjai ir į paskolas investuojantys investuotojai gali sumažinti paskolų suteikimo asmenims, kurie negrąžina paskolų kiekį. Tai galima pasiekti pritaikius klasifikavimo metodus bandant prognozuoti paskolos baigtį (ar paskola bus grąžinta ar ne). Duomenys, kuriuos kaupia tarpusavio skolinimo platformos per paskolos išdavimo procesą, yra beribis įžvalgų šaltinis. Sukūrus strategiją, kurioje būtų atsižvelgiama į kredito balą, palūkanų normą ir naudojantis prognozuojamais paskolos rezultatais, būtų galima išvengti daugybės nuostolingų investicijų.

Gebėjimas prognozuoti paskolos grąžinimo statusą būtų naudingas tiek investuotojams, nenorintiems prarasti savo investuotų pinigų, tiek verslininkams, kuriantiems tarpusavio skolinimo paskolų platformas. Platformos, kuriose didžiąją dalį sudaro klientai , kurie negrąžina pinigų, praranda investuotojus ir pelną.

1.1.Darbo problema

Kaip įvertinti fizinių asmenų paskolų grąžinimo galimybes tarpusavio skolinimosi platformų atveju.

1.2.Darbo tikslas

Įvertinti fizinių asmenų paskolų grąžinimo galimybes tarpusavio skolinimosi platformų „Bondora“ ir „Lending Club“ atveju.

1.3.Darbo uždaviniai

- Naudojantis literatūros šaltiniais atlikti esamų tarpusavio skolinimosi platformų analizę.

- Atlikti esamų mokslinių tyrimų analizę nustatant dažniausiai taikomus metodus kredito rizikai modeliuoti bei globaliai sprendžiamas problemas, susijusias su paskolomis.
- Atlikti naudojamų skaitmeninio mokymo algoritmų tinkamų paskolų grąžinimo galimybėms vertinti apžvalgą.
- Atlikti aprašomąją duomenų analizę analizuojamų tarpusavio skolinimosi platformų atveju.
- Įvertinti fizinių asmenų paskolų grąžinimo galimybes, pasitelkiant skaitmeninio mokymo algoritmus.

2. LITERATŪROS APŽVALGA

2.1.Tarpusavio skolinimas ir jo istorija

Tarpusavio skolinimo platformos dar kitaip vadinamos P2P skolinimo platformomis (*angl. Peer to Peer Lending platforms*) – tai skaitmeninės platformos, naudojamos skolinti pinigams, kuomet lėšos yra skolinamos tiesiogiai tiems asmenims, kurie siekia gauti paskolą. Skolinimo procesas yra kontroliuojamas tarpininko, skaitmeninės platformos administratoriaus, kuriam atitenka administracinis paskolos mokestis.

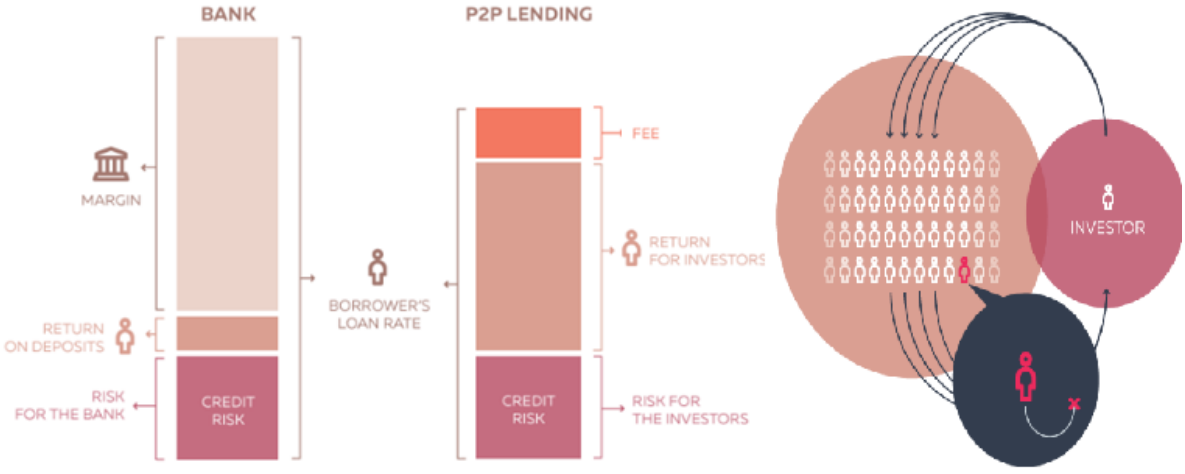
Ryanas Lichtenwaldas," Lend Academy" žurnalistas savo straipsnyje „The History of Peer to Peer Lending" teigia jog tarpusavio skolinimas tapo populiarus jau 1700 metais, kuomet Jonathanas Swiftas pradėjo skolinti nedideles pinigų sumas turintiems poreikį. 18 ir 19 amžiuje tarpusavio skolinimas tapo vienu populiariausių pinigų skolinimo metodu Europoje, tačiau jo populiarumas sumažėjo 20 amžiuje, padidėjus bankų įtakai. Skaitmeninės tarpusavio skolinimo platformos pradėjo populiarėti tik nuo 2005 metų, kuomet Didžiojoje Britanijoje buvo sukurta pirmoji skaitmeninė tarpusavio skolinimosi platforma „Zopa". Tačiau didžiausias tarpusavio skolinimosi platformų populiarumas buvo pasiektas tik 2008 metais, kuomet bankrutavus „Lehman Brothers" įmonei vartotojai prarado pasitikėjimą finansinėmis institucijomis ir pradėjo ieškoti alternatyvių, patikimesnių būdų investuoti pinigus. Netrukus tarpusavio skolinimo platformos išpopuliarėjo Europoje, Amerikoje, bei kituose žemynuose 1].

Johnas Carsonas savo straipsnyje „Peer to Peer Lending Sites → 24 of the World's Best" atliko skirtingų tarpusavio skolinimo sistemų analizę ir įvertino jas žemynų mastu. Jo teigimu, globaliu mastu geriausiai įvertinta tarpusavio skolinimo sistema pasaulyje yra „Bitbond", kurioje naudojama bitkoinų mokėjimo sistema. „Bitbond" yra didžiausias palūkanas siūlanti platforma už nulinius mokesčius. Jungtinėse Amerikos valstijose pirmauja „Lending Club" skolinimo platforma, kurioje patyrę investuotojai geba gauti 13,3 % grąžą. Tuo tarpu Europoje pagal populiarumą pirmaują „Bondora" skolinimo platforma, kurioje investuotojui Clausui Lehmannui pavyko gauti 26 % grąžą. Jungtinėje Karalystėje pirmą vietą užima pionierė „Zopa", kurioje pasiekiama 9 % grąža 2].

Vartotojai yra linkę imti paskolas tarpusavio skolinimo sistemose. Tamara Wilhite savo straipsnyje „Why People Use Peer to Peer Lending" išskiria toliau aprašomas priežastis tarpusavio skolinimo platformų naudojimui. Daugeliui tai yra vienintelė galimybė gauti paskolą, turint žemą kredito balą ar vos tik pradėjus gyventi finansinį gyvenimą.

Taipogi tarpusavio skolinimosi sistemose siūlomos kur kas mažesnės palūkanų normos, nei esamos kredito kortelėse, dėl to jaunajai kartai, norinčiai gauti pinigų greitai ir lengvai, tai yra vienas patogiausių ir pigiai atsieinančių būdų gauti paskolą. Tarpusavio skolinimasis šiuo metu yra vienintelė reali galimybė gauti paskolą asmenims, patyrusiems bankrotą ir ieškantiems greitų, smulkaus kiekio paskolų. Skaitmenizavus tarpusavio skolinimą dabar vieno mygtuko paspaudimu galima grąžinti paskolas visiems asmenims vienu kartu, vietoj grąžinimo pinigus kiekvienam asmeniui atskirai [3].

Tarpusavio skolinimas yra naudingas tiek investuotojams (skolintojams), tiek pinigus besiskolinantiems asmenims. Loic Le Pichoux „Klear“ skolinimo platformos bendrasavininko ir vadovo pateiktoje diagramoje (1 pav.) matomas tarpusavio skolinimo platformų pranašumas lyginant su finansinėmis institucijomis. Nors kredito rizika tiek investuotojams, tiek bankui išlieka panaši, tačiau grąža, kuri yra gaunama investuojant į paskolas, bankui yra net kelis kartus mažesnė už grąžą, kuri gaunama investuojant į paskolas tarpusavio skolinimo platformose. Kadangi skolinimo platformos administratorius pasiima kur kas mažesnį pelną, gautą nuo paskolos nei bankas, besiskolinantis asmuo gali skolintis pinigus su kur kas mažesnėmis palūkanomis [4].



1 pav. Tarpusavio skolinimosi principas [4]

Tarpusavio skolinimo platformos taip pat yra pranašios tuo, kad investuojamą pinigų sumą galima diversifikuoti skolinant nedideles pinigų sumas į daug paskolų. Diversifikuodami savo portfelį skolintojai yra užtikrinti, jog rizika prarasti pinigus yra kur kas mažesnė, nei skolinti didelę pinigų sumą vienam asmeniui. Taipogi skolinant tarpusavio skolinimo sistemose iš karto galima pasirinkti portfelių rizikos matą (labai rizikingą, vidutinės rizikos, nerizikingą), pagal kurį norima investuoti. Gabesniems investuotojams suteikiama galimybė pasinaudoti aplikacijų programavimo sąsaja (*angl. API*), esančia platformose, kuriose galima nustatyti norimus skolinimo kriterijus, investuoti

automatiškai neprisijungiant prie sistemos, investuoti naudojantis iš anksto suprogramuota skolinimo strategija 4].

2.2.Tarpusavio skolinimosi platforma „Bondora“

Tiriamojo darbo analizei buvo pasirinkti populiariausios Europoje tarpusavio skolinimo platformos „Bondora“ duomenys. Ši tarpusavio skolinimo platforma yra viena sparčiausiai augančių pasaulyje. Įmonė buvo įkurta 2009 metais Estijoje, šalyje su aukštu išsilavinimo lygiu, kur 99% banko operacijų atliekama internetu. Tuometinis įmonės pavaidinimas buvo „isePankur“. Ši įmonė leidžia skolintis pinigus gyventojams, kilusiems iš Estijos, Ispanijos, Slovakijos, Suomijos. Iki šiol įmonė jau buvo suteikusi 30 mln. Eur. vertės paskolų ir išmokėjusi 4 mln. Eur. palūkanų investuotojams. Pagal „InvestmentHaking“ straipsnį „Bondora High Yield Peer To Peer Lending Reviewed“, vidutinis „Bondora“ investuotojas, naudodamasis sistema, gauna 20 % grąžą [5,6,7].

Pagrindinį pajamų šaltinį „Bondora“ skolinimo platforma generuoja operuodama neefektyvioje kredito biržoje. Didžioji dalis bankų Estijoje apmokestina paskolas 20–29 % palūkanomis, tuo tarpu „Bondora“ skolinimosi platforma siūlydama savo klientams žemesnes palūkanas ir aukštas grąžas, pritraukia tiek investuotojus, tiek skolininkus [5,6,7].

Tačiau aukštos palūkanos rodo, kad didžioji dalis investicijų yra rizikingos. Kuo žemesnis asmens kredito balas, tuo aukštesnės palūkanos siūlomos. Investuotojai investuodami į paskolas su didžiausiomis palūkanomis, neretai praranda pinigus [6,7].

2.3.Tarpusavio skolinimosi platforma „Lending Club“

„Lending Club“ yra didžiausia tarpusavio skolinimo platforma, kurioje teikiamos tiek asmeninės tiek verslo paskolos. „Lending Club“ savo veiklą pradėjo 2007 metais San Franciske, Kalifornijoje. Iki 2017 kovo 31 dienos įmonė yra išdavusi virš 26,6 milijardų dolerių paskolų. Įmonė teikia paskolas tik Amerikoje gyvenantiems asmenims[8].

Didžioji dalis paskolų, kurios buvo paimtos „Lending Club“ vartotojų buvo tam, kad kompensuoti savo buvusias skolas ar kredito korteles. Tai rodo, kad „Lending Club“ siūlo kur kas mažesnes palūkanas, nei kitos finansinės institucijos. Vidutinė grąža, kurią investuotojas gauna iš paskolų yra 5-7% priklausomai nuo portfelio rizikingumo. Galutinė grąža, kurią gauna vartotojas yra kur kas mažesnė nei palūkanos, dėl investavimo mokesčio [8,9].

Palūkanos gaunamos naudojantis „Lending Club“ tarpusavio skolinimosi platforma yra gerokai mažesnės lyginant su „Bondora“, tačiau „Bondora“ tarpusavio

skolinimosi platformos klientai yra gerokai rizikingesni, kur kas didesnė dalis klientų nevykdo įsipareigojimų[9].

2.4.Galimybių grąžinti paskolą vertinimas

Kredito riziką galima apibrėžti kaip tikimybę, kad nebus įvykdyti įsipareigojimai pagal iš anksto nustatytas sąlygas. Pagrindinis kredito rizikos suvaldymo tikslas yra maksimizuoti koreguotą rizikos santykį išlaikant riziką priimtinuose parametruose [10].

Prieš atsirandant skaitmeninio mokymo algoritmams tikimybė, kad paskola nebus grąžinta būdavo vertinama naudojantis šiais metodais:

1. Sujungimo metodu (*angl. pooling approach*) – apskaičiuojamas istorinė tikimybė neįvykdyti įsipareigojimų. Tikimybė suskirstoma į skirtingas grupes ir kiekvienai grupei priskiriamos tam tikros charakteristikos, kurios priklauso tai tikimybių grupei. Atlikti skaičiavus skirtus šiam metodui įprasta taikyti dvi technikas – kochortas ir trukme paremtus metodus [11].
2. Statistiniu metodu – dažniausiai naudojami statistiniai metodai apskaičiuoti įsipareigojimų nevykdymo tikimybę yra regresija, kur rezultatas apribojamas 0 ir 1 arba logistinė regresija [11].
3. Struktūrinis metodas – dažniausiai naudojama korporacijoms, kur teigiama, kad įmonė neįvykdys savo įsipareigojimų, jei turimų aktyvų kiekis yra mažesnis nei įmonės paskola. Taip yra dėl to, kad tokiu būdų turimų turtų dydis tampa neigiamas (aktyvų vertė = turto vertė +įsipareigojimų vertė). Šiam metodui atlikti dažniausiai naudojami Mertono arba KMV modeliai [11].

Skaitmeninio mokymo algoritmai siekiant nustatyti ar vartotojas grąžins paskolą įvairiuose tyrimuose taikomi jau virš 10 metų. Bart Baesansas dar 2003 metais savo daktaro gynimo dizertacijoje „Developing Intelligent Systems for Credit Scoring Using Machine Learning Techniques“ pateikė pavyzdžius, kaip būtų galima pritaikyti klasifikavimą vertinant galimybę grąžinti paskolą. Tyrimo metu buvo panaudota logistinė regresija, diskriminantinė analizė, tiesinis programavimas, Bajeso tinklai, patiklus Bajesas, sprendimų medis, neuroniniai tinklai, mažiausių kvadratų atraminių vektorių metodas [12].

Peng Goh Chwee savo tyrime „Credit Scoring Using Data Mining Techniques“ 2004 metais siekdamas įvertinti skolintojo kreditingumą pasinaudojo Logistinės regresijos, neuroninių tinklų ir sprendimų medžio metodais [13]. Monika Szczerba, Andrzejus Ciemskis savo tyrime „Credit Risk Handling in Telecomunication Sector“ 2009 metais pritaikė sprendimų medžio metodą [14]. Iainas Brownas ir Christophas Muesas savo tyrime

„An experimental comparison of classification algorithms for imbalanced credit scoring data sets“ 2012 metais pritaikė logistinės regresijos, neuroninių tinklų, tiesinės ir kvadratinės diskriminantinės analizės, mažiausių kvadratų atraminių vektorių, sprendimų medžių, artimiausių kaimynų, atsitiktinio miško ir gradientinio sustiprinimo metodus. Tyrimo metu nustatė, kad geriausiai veikė gradientinio sustiprinimo ir atsitiktinio miško metodai [15]. 2016 metais Mahero Alarajo ir Maysamo F. Abbodo atliktame tyrime „A new hybrid ensemble credit scoring model based on classifiers consensus system approach“ buvo panaudoti neuroninių tinklų, atraminių vektorių, sprendimų medžių, atsitiktinio miško ir patiklaus Bajeso metodai. Tyrimo metu metodai panaudoti naudojant vokiečių, australų, japonų, iraniečių, lenkų, jordeniečių ir San Diego. Metodams įvertinti buvo naudotos šios tikslumo metrikos – tikslumas, AUC, H-įvertis, Brierio įvertis [16]. 2016 metais Aboobyda Jafaras Hamidas ir Tarigas Mohammedas Ahmedas savo tyrime „Developing Prediction Model Of Loan Risk in Banks Using Data Mining“ panaudojo sprendimų miško, Bajeso tinklų, patiklujį Bajeso metodus. Jų įvertinimu geriausiai veikė atsitiktinio miško algoritmas [17].

1 lentelė. Tyrimai, kuriuose buvo naudoti skaitmeniniai algoritmai skirti galimybei grąžinti paskolą įvertinti.

Metai	**Tyrėjai**	**Metodai**	**Publikacijos pavadinimas**
2003	Bart Baesens	Logistinė regresija, diskriminantinė analizė, tiesinis programavimas, Bajeso tinklai, patiklus Bajesas, sprendimų medžiai, neuroniniai tinklai, mažiausių kvadratų atraminių vektorių metodas	Developing Intelligent Systems for Credit Scoring Using Machine Learning Techniques
2004	Chwee Peng Goh	Logistinė regresija, neuroniniai tinklai, sprendimų medžiai	Credit Scoring Using Data Mining Techniques
2009	Monika Szczerba, Andrzej Ciemski	Sprendimų medžiai	Credit Risk Handling in Telecomunication Sector
2012	Iain Brown, Chritophe Mues	Logistinė regresija, Neuroniniai tinklai, tiesinė ir kvadratinė diskriminantinė analizė, mažiausių kvadratų atraminiai vektoriai, sprendimų medžiai, artimiausias kaimynas, atsitiktinis miškas, gradientinis sustiprinimas	An experimental comparison of classification algorithms for imbalanced credit scoring data sets
2016	Maher Ala’raj, Maysam F. Abbod	Neuroniniai tinklai, atraminiai vektoriai, sprendimų medžiai, atsitiktinis miškas, patiklus Bajesas	A new hybrid ensemble credit scoring model based on classifiers consensus system approach
2016	Aboobyda Jafar Hamid, Tarig Mohammed Ahmed	Sprendimų medžiai, Bajeso tinklų, patiklus Bajesas	Developing Prediction Model Of Loan Risk In Banks Using Data Mining

Kaip matoma iš 1 lentelės dažniausiai naudojami metodai įvertinti galimybę grąžinti paskolą yra sprendimų medžiai, tačiau taip pat labai populiaru naudoti neuroninių tinklų metodus, atsitiktinio miško metodus, Bajeso algoritmus, atraminių vektorių metodus.

Atlikus literatūros apžvalgą buvo nustatyta, kad investuojant į paskolas naudojantis tarpusavio skolinimosi platformomis galima gauti didesnes palūkanas su mažesne rizika, nei investuojant į bankus. Tačiau investuojant rizika vis tiek išlieka ir ją svarbu išmatuoti naudojantis matematiniais modeliais. Atlikus mokslinių tyrimų analizę buvo nustatyta, kad veiksmingiausias būdas rizikai išmatuoti yra modelio apmokymas naudojantis klasifikavimo metodais.

3. MEDŽIAGOS IR TYRIMŲ METODAI

3.1. Z įverčio transformacija

Tiriant duomenis dažnai susiduriama su skalės problema, kuomet vieno kintamojo dydis yra neaplyginimai didesnis nei kito (pvz. vieno duomenų stulpelio reikšmės svyruoja tarp 1000-100000 kito stulpelio reikšmės svyruoja tarp 0 – 1). Duomenų standartizavimas yra metodas skirtas norint panaikinti skirtumą tarp stulpelių, kurių duomenys skiriasi dideliu atstumu.Vienas iš duomenų standartizavimo būdų yra paversti duomenis į z įvertį. Naudojant šią duomenų transformaciją visi duomenys yra paverčiami šia forma:

$$z = \frac{x - mean(x)}{stdev(x)} \qquad (1)$$

Šiai standartizacijai vidurkis ir standartinis nuokrypis yra apskaičiuojami kiekvienam stulpeliui atskirai. Naudojamas populiacijos standartinis nuokrypis [18].

3.2.Klasifikavimas

Klasifikavimo problema aprašoma kaip kintamojo y_i prognozavimas naudojantis duotais kintamaisiais x_i[19]. Duomenų klasifikavimas tyrime bus naudojamas paskolos grąžinimo ir negrąžinimo statusui prognozuoti.

3.2.1. Logistinė regresija

Logistinės regresijos atveju modeliuojama tikimybė, kad Y turi reikšmę 1, kaip kintamų dydžių $X_1, X_2, \dots, X_p$ funkcija. Pažymėsime sąlyginę tikimybę atpažinti $Y = 1$ suteikiant kintamų dydžių reikšmes $x_1, x_2, \dots x_p$ pagal

$$\pi(x_1, x_2, \dots x_p) \qquad (2)$$

Tuomet kuriamas modelis aprašantis π reikšmę yra lygus:

$$\text{logit}\, p = \log \frac{p}{1-p} \qquad (3)$$

Taip, kad būtų gaunamas logistinės regresijos modelis

$$\text{logit}\, \pi(x_1, x_2, \dots x_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \qquad (4)$$

Šiuo atveju modelio parametrų reikšmės rodo įtaką tikimybei atpažinti $Y = 1$ išreikštą logaritminėje skalėje [20].

3.2.2. Vidurkinis perceptronas

Vidurkinis perceptronas (*angl. avaraged perceptron*) yra laikomas vienu paprasčiau skaitmeninio mokymo metodų tačiau gebantis išgauti gerą modelio tikslumą. Modelį pirmasis apibūdino Michaelas Collinsas naudodamasis paslėptaisiais Markovo metodais [21].

Apmokymo fazėje duotiems apmokymo pavyzdžiams su atsitiktiniais pradiniais svorių vektoriais w, duomenų pavyzdžiai yra iteraciškai apmokomi. Tegul y yra teisingų klasifakotorių seka visiems x, o y' prognozuotoji klasifikatorių seka. Jei modelis prognozuoja neteisingai $y \neq y'$, tuomet svorių vektorius yra pakeičiamas į:

$$w = w + \phi(x, y) - \phi(x, y') \qquad (5)$$

Galutinis svoris gautas perceptrono algoritme yra visų atsiradusių vektorių vidurkis pagal išgyvenimo ilgį (pvz. teisingai atspėtų kintamųjų kiekį prieš padarant klaidą ir vektoriui pasikeičiant [22].

3.2.3. Atraminiai vektoriai

Atraminių vektorių klasifikatorius *(angl. SVM – support vector machine)* suformuoja hiperplokšumų rinkinį begalinėje dimensinėje erdvėje, kuris duomenis padalina į dvi dalis [23].

Tarkime turima apmokymo duomenų imtis $\{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\}, y_i \epsilon \{-1,1\}, x_i \epsilon R^d$, kur x_i kintamasis priklausantis nuo y_i. l reprezentuoja apmokymų duomenų imties elementų kiekį. Tiesinis SVM klasifikatorius randa optimalią skiriančiąją parašę išspręsdamas optimizaciją minimizuodamas

$$\left\{\frac{1}{2}|w|^2 + C\sum_{i=1}^{l}\varepsilon_i\right\}, \varepsilon_i \geq 0 \qquad (6)$$

Jei

$$y_i(w^T x_i + b) \geq 1 - \varepsilon_i, i = 1,2, \dots, l \qquad (7)$$

, kur C yra baudos vertė, ε_i teigiami laisvi parametrai, w yra normalusis vektorius, b skaliarinis kiekis. Minimumo radimo problema gali būti sprendžiama naudojant Lagranžo daugiklį $\propto_i$, kuris įgyja optimalią reikšmė pagal Kuhno Tuckerio sąlygą. Jei $\propto_i > 0$, tuomet atitinkami duomenys x_i vadinami atraminiais vektoriai ir tuomet tiesinė diskriminantinė funkcija gali būti išreikšta su optimaliais hiperplokštumos parametrai w ir b šia lygtimi:

$$f(x) = sgn\left(\sum_{i=1}^{l} \propto_i y_i x_i^T x + b\right) \qquad (8)$$

Lygtis transformuojama dualią formą be ribojimų maksimizuojant:

$$max\left\{\sum_{i=1}^{l} \propto_i - \frac{1}{2}\sum_{i=1}^{l} \propto_i \propto_j y_i y_j x_i x_j\right\}, \qquad (9)$$

$$C \geq \propto_i \geq 0, \qquad i = 1, \dots, l, \quad \sum_{i=1}^{l} \propto_i y_i. \qquad (10)$$

Lygtį galima išspręsti naudojant kvadratinio programavimo technikas ir stacionariąsias Kuhno Tuchkerio sąlygas. Sprendinys $\boldsymbol{W}$ gali būti išreiškiamas kaip tiesinė apmokymo vektorių kombinacija, o $\boldsymbol{b}$ išreiškiamas kaip atraminių vektorių suma:

$$\boldsymbol{W} = \sum_{i=1}^{l} \propto_i y_i x_i, \qquad (11)$$

$$\boldsymbol{b} = \frac{1}{N_{SV}} \sum_{i=1}^{N_{SV}} (W x_i - y_i), \qquad (12)$$

Kur N_{SV} yra atraminių vektorių kiekis.

Tiesinis SVM gali būti išplėstas į netiesinį modelį perkeliant x_i į ypatybių erdvę $\theta(x_i)$, kur $x_i^T x$ atvaizduojama $\theta(x_i)^T \theta(x_i)$ forma ypatybių erdvėje. Tokiu būdu, netiesinė diskriminantinė funkcija atvaizduojama tokiu formatu:

$$f(x) = sgn\left(\sum_{i=1}^{l} \propto_i y_i K(x_i, x) + b\right); \qquad (13)$$

Kur $K(x_i, x) = \langle \theta(x_i), \theta(x) \rangle$ ir $K(x_i, x)$ yra branduolio funkcija. Dažnai naudojama branduolio funkcija yra spindulio bazių funkcija *(angl. radial basis function)* dar žinoma kaip Gauso branduolys, dėl jos tikslaus ir patikimo atlikimo, kuris gali būti apibūdintas kaip

$$K(x_i, x) = \exp(-\gamma \|x_i - x\|)^2. \qquad (14)$$

γ yra iš anksto apibūdintas sklandumo parametras, kuris kontroliuoja branduolio spindulio pagrindo funkcijos plotį, kuris aprašomas šia formule:

$$max\left\{\sum_{i=1}^{l} \propto_i - \frac{1}{2}\sum_{i=1}^{l} \propto_i \propto_j y_i y_j \exp(-\gamma\|x_i - x\|)^2\right\}, \qquad (15)$$

$$C \geq \propto_i \geq 0, \qquad i = 1, \dots, l, \quad \sum_{i=1}^{l} \propto_i y_i. \qquad (16), [23]$$

3.2.4. Gilieji atraminiai vektoriai

Giliųjų atraminių vektorių klasifikatorius *(angl. DSVM – deep support vector machine)* tai metodas, kuriam naudojama lokali giliojo branduolio optimizacija. Naudojant optimizaciją pagreitinamas netiesinis SVM klasifikatorius neprarandant tikslumo. Modelyje apibendrinamas daugialypis branduolio apmokymas įterpiant medžiu pirmykštę funkciją. Tokiu būdu prognozavime atsikratoma prognozavimo sąnaudų mažinant atraminių vektorių kiekį ir įterpiant medžio struktūros ypatybes taip pagreitinant prognozavimą eksponentiškai.

Modelyje naudojama ši optimizacijos funkcija, kurią išvedė Cijo Jose, Prasoonas Goyalas, Parvas Aggrwalas, Manikas Varma savo tyrime „Local Deep Kernel Learning for Efficient Non-linear SVM Prediction“.

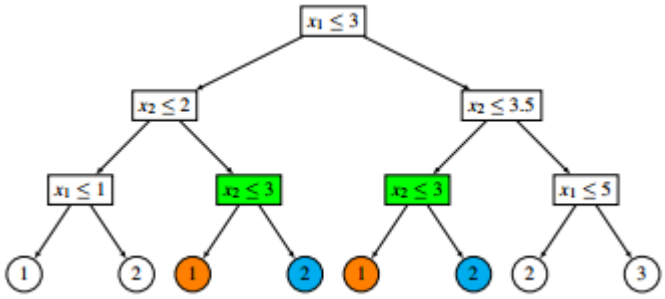
$$\min_{W,\theta,\theta'} P(W, \theta, \theta') = \frac{\lambda_W}{2} Tr(W^t W) + \frac{\lambda_\theta}{2} Tr(\theta^t, \theta) + \frac{\lambda_{\theta^t}}{2} Tr(\theta'^t, \theta') + \sum_{i=1}^{N} L(y_i \phi_L^t(x_i) W^t x_i) \qquad (17)$$

, kur $L = \max(0,1 - y_i \phi_L^t(x_i) W^t x_i)$, ϕ_L - lapų mazgų kiekis [24].

3.2.5. Sustiprintas sprendimų medis

Sustiprintas sprendimų medis *(angl. boosted decision tree)* laikomas vienu efektyvesnių klasifikavimo metodų. Gradientinis sustiprinimas laikomas įprastine technika, kuri gali būti pritaikoma prastai apmokytiems rezultatams. Dažniausiai gradientinis sustiprinimas naudojamas medžiuose. Gradientinį sustiprinimą pristatė Jerome H. Friedmanas savo straipsnyje „Greedy Function Approximation: A Gradient Boosting Machine“[25].

Sprendimų medis 2 pav. tai metodas naudojantis grafą paremtą sprendimų modeliavimu įtraukiant galimus scenarijus ir galutinius rezultatus.



2 pav. Sprendimų medis [29]

Medžių atveju modelio rezultatai galima aprašyti taip:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F, \qquad (18)$$

kur K yra medžių kiekis, f yra funkcija funkcinėje erdvėje F, kur tikslo funkcija gali būti aprašoma kaip:

$$obj(\theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \qquad (19)$$

Kur l yra apmokymo praradimo funkcija, o Ω reguliarizacijos sąvoka.

Siekiant susistiprinti įprastų medžių algoritmą stiprinama ši tikslo funkcija:

$$obj = \sum_{i}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{t} \Omega(f_t) \qquad (20)$$

Siekiant patikslinti modelio rezultatus apmokant medžių algoritmą pridedama po vieną naują medį:

$$\hat{y}_i^{(0)} = 0 \qquad (21)$$

$$\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \qquad (22)$$

...

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \qquad (23)$$

Renkantis medį kurį norima pridėti tikrinama, kuris iš jų optimizuoja tikslo funkciją pagal MSE žingsnyje t galima aprašyti tokiu būdu:

$$obj^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t), \qquad (24)$$

Kur g_i ir h_i aprašomi kaip:

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) \qquad (25)$$

$$g_i = \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) \qquad (26)$$

Šia formule aprašomas optimizacijos tikslas naujam medžiui. Tokiu būdu nuostolio funkcija tampa priklausoma, tik nuo 2 kintamųjų g_i ir h_i. Aprašyti reguliarizacijai reikia aprašyti medžio sudėtingumą $\Omega(f)$, tarkime, kad medis $f(x)$ gali būti aprašomas tokiu būdu:

$$f_t(x) = w_{q(x)}, w \in R^T, q: R^d \to \{1,2,\dots,T\}, \qquad (27)$$

Kur w yra lapų įverčių vektorius, q funkcija priskirianti kiekvieną duomenų tašką atitinkamam lapui, o T lapų skaičius. Tokiu būdu naudojant gradientinį sustiprinimą kompleksiškumas aprašomas kaip:

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 \qquad (28)$$

Tokiu būdu tikslo funkciją galimą aprašyti kaip t medį naudojantis šia lygybe:

$$obj^{(t)} \approx \sum_{i=1}^{n} \left[g_i w_{q(x_i)} + \frac{1}{2} h_i w^2_{q(x_i)} \right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2$$

$$= \sum_{j=1}^{T} \left[\left(\sum_{i \epsilon I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \epsilon I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T, \qquad (29)$$

Kur $I_j = \{i | q(x_i) = j\}$ yra indeksų rinkinys priskirtas j lapui. Šią formulę galima suspausti naudojantis, kad $G_j = \sum_{i \epsilon I_j} g_i$, o $H_j = \sum_{i \epsilon I_j} h_i$:

$$obj^{(t)} = \sum_{j=1}^{T} \left[G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2 \right] + \gamma T \qquad (30)$$

Šioje lygtyje w_j yra nepriklausomi vienas kitam, o $G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2$ dėmuo turi kvadratinę formą todėl geriausias w_j duotai struktūrai $q(x)$ ir geriausiai tikslo redukcijai gali būti aprašomas taip:

$$w_j^* = -\frac{G_j}{H_j + \lambda} \qquad (31)$$

$$obj^{(t)} = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T \qquad (32)$$

Tokiu būdu paskutinė lygybė apskaičiuoja kiek gera yra medžio struktūra $q(x)$. Žinodami kiek geras yra medis patį geriausią medžio variantą randame naudodamiesi šia naudos lygybe:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} + \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \qquad (33)$$

Tokiu būdu galima įvertinti medį. Jei $Gain < \gamma$, tuomet modelis veikia geriau nepridėjus papildomos šakos [25,26].

3.2.6. Sprendimų miškas

Sprendimų miškas, tai medžių klasifikatorių kombinacija, taip kad kiekvienas medis priklauso nuo atsitiktinio vektoriaus reikšmių, kurios yra atrinktos nepriklausomai ir naudojant tą patį pasiskirstymą visiems medžiams miške. Miško algoritmą pirmieji sudarė Leo Breimanas and Adele Cutlerė [27].

Sprendimų miško sudarymo algoritmas yra artimas savirankos (imčių su pasikartojimu) agregavimo *(angl. bagging)* algoritmui. Tegul N yra stebėjimų kiekis, o klasifikatorius dvinaris kintamasis. Tuomet sprendimų medžių sudarymo algoritmas vykdomas tolesne seka:

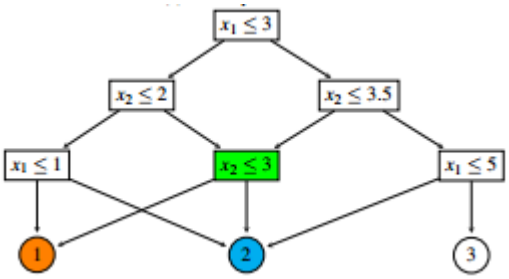
1. Pasirenkamas atsitiktinis duomenų kiekis N su duomenų pakeitimu (bootstrap algoritmas).
2. Pasirenkamas atsitiktinis duomenų pavyzdys be pakeistų klasifikatorių.
3. Išskaidomi duomenys naudojant klasifikatorius pasirinktus 2 žingsnyje.
4. Pakartojami 2 ir 3 žingsniai kiekvienam duomenų išskaidymui, kol medis tampa norimo dydžio. Kiekvienas medis yra pagaminamas iš atsitiktinio duomenų pavyzdžio ir kiekvienas ir kiekviename žingsnyje naudojami atsitiktiniai klasifikatorių pavyzdžiai.
5. Įvertinama klaida naudojama naudojamiems duomenims, kur nustatomi kintamieji nepatekę į medžio apmokymo imtį *(angl. out of bag).* Išlaikoma klasė priskirta kiekvieno stebėjimo metu su kiekvieno stebėjimo prognozuojamomis vertėmis.
6. Pakartojami žingsniai 1-5 nustatytam kiekiui.

7. Kiekvienam duomenų rinkinio stebėjimui apskaičiuojamas medžių kiekis, kuris yra klasifikuojamas vienoje kategorijoje per medžių kiekį.

8. Galutinei kategorijai priskiriamas kiekvienas stebėjimas pagal daugumos balsus lyginant su medžių rinkiniai. Jei daugiau nei 50% kartų per medžių rinkinius duotam klasifikatoriui buvo priskirta reikšmė 1, tuomet tai tampa klasifikuota modelio reikšme [27,28].

3.2.7. Sprendimų džiunglės

Sprendimų džiunglės yra metodas apjungiantis sprendimų nukreiptuosius alifatinius grafus *(angl. DAG – directed acyclic graph)* ir sprendimų medžius. Toliau atvaizduotame paveikslėlyje matome, kad nukreiptasis alifatinis grafas gali turėti daugiau nei vieną viršūnę lyginant su sprendimų medžiu (3 pav.)



3 pav. Sprendimų nukreiptasis alifatinis grafas [29]

Nukreiptasis alifatinis grafas yra įsišaknijęs nukreiptasis alifatinis grafas $G = (V, E)$, kurio visi nukreiptieji takai nuo šaknies mazgo r iki mazgo v yra tokių pat ilgių ir jų mazgai $v \in V$ yra papildyti šiais atributais:

- Jei v yra lapo mazgas, tuomet v yra susieta su klasių histograma h_v (funkcija $h: \{1, \dots, C\} \to \mathbb{R}$ su $\sum_{i=1}^{C} h(i) = 1$)
- Jei v nėra lapų mazgas (pavyzdžiui šakninis mazgas arba vidinis mazgas), tuomet v yra papildoma sutvarkytų elementų rinkiniu:

$$(d_v, \theta_v, l_v, r_v) \epsilon \{1, \dots, n\} \times \mathbb{R} \times \{w \in V: (v, w) \in E\}^2, \qquad (34)$$

kur d_v – ypatybių dimensija. θ_v – slenkstis, l_v – kairysis vaiko mazgas, r_v – dešinysis vaiko mazgas. Apmokant sprendimų nukreiptąjį alifatinį grafą ieškoma optimalių parametrų $(d_v, \theta_v, l_v, r_v)$ kiekvienam mazgui v. Lyginant su medžiu apmokyti nukreiptąjį alifatinį grafą yra sunkiau, nes parametrai l_v ir r_v (grafo struktūra) medžiuose yra fiksuoti.

Tarkime klasifikatorius šioje klasifikavimo problemoje aprašomas kaip $f_G: \mathbb{R}^n \to \{1, \dots, C\}$. Tuomet klasių histograma aprašoma tokiu būdu:

- Jei v yra lapų mazgas, tuomet $\hat{h}_v(x) = h_v$
- Jei v nėra lapų mazgas, tuomet

$$\hat{h}_v(x) = \begin{cases} \hat{h}_{l_v} & jei\ x_{d_v} \leq \theta_v \\ \hat{h}_{r_v} & jei\ x_{d_v} > \theta_v \end{cases} \qquad (35)$$

Klasifikatorius f_G tokiu atveju aprašomas kaip

$$f_G: \mathbb{R}^n \to \{1, \dots, C\}, x \to \underset{c=1,\dots,C}{\arg\max} \left(\hat{h}_v(x)\right)(c), \qquad (36)$$

Kur $r \in V$ yra G šakninis mazgas.

Atsitiktinis sprendimų nukreiptasis alifatinis grafas yra sprendimų nukreiptasis alifatinis grafas, kurio atributai Θ_v kiekvienam vidiniam mazgui $v \in V$ yra nepriklausomai ir identiškai pasiskirstę atsitiktiniai kintamieji.

Tuomet sprendimų džiunglės yra $J = (G_1, \dots, G_m)$ yra m atsitikinių sprendimų nukreiptųjų alifatinių grafų rinkinys $G_1, \dots, G_m$. Pagrindinį džiunglių klasifikatorių f_J galima aprašyti toliau pateikta formuluote:

$$f_J: \mathbb{R}^n \to \{1, \dots, C\}, x \to \underset{c=1,\dots,C}{\arg\max} \sum_{i=1}^{m} \mathbb{I}\left(c, f_{G_i}(\boldsymbol{x})\right), \qquad (37)$$

Kur $\mathbb{I}$ yra rodiklio funkcija:

$$\mathbb{I}(x, y) = \begin{cases} 1, jei\ x = y \\ 0, kitu\ atveju \end{cases} \qquad (38), [29,30]$$

3.2.8. Bajeso taškas

Bajeso taško klasifikatorius *(angl. Bayes Point machine*) yra vienas iš Bajeso klasifikavimo metodų. Bajeso taško algoritmą galima aprašyti šiuo būdu:

Tarkime turime fiksuota hipotezių erdvę $H \subseteq X^Y$ ir fiksuotą nuostolį $l: Y \times Y \to R^+$. Tuomet bet kuriems dviems matams P_X ir P_H Bajeso taško algoritmas $\mathcal{A}_{bp}$ gali būti aprašomas kaip

$$\mathcal{A}_{bp} \coloneqq \underset{h \in H}{\arg\min}\, E_X\left[E_{H|Z^m=z}\left[l\big(h(X), H(X)\big)\right]\right], \qquad (39)$$

kur kiekvienam apmokymo pavyzdžiui $z \in Z^m$ Bajeso taško algoritmas pasirenka klasifikatorių $h_{bp} \coloneqq \mathcal{A}_{bp}(z) \in H$, kuris imituoja geriausią Bajeso klasifikavimo strategiją pagal vidurkį atsitiktinai pasirinkdamas testavimo taškus. Klasifikatorius $\mathcal{A}_{bp}(z)$ vadinamas Bajeso tašku [31].

3.3.Klasifikavimo į dvi klases (detekcijos) sėkmingumo vertinimas

Klasifikavimo metodų vertinimo metrikos tyrime bus naudojamos prognozuotų kintamųjų tikslumui įvertinti. Dvimačių klasifikavimo metodų prognozavimo rezultatus yra geriausia vertinti remiantis sumaišymo matrica (*angl. Confusion matrix*), kaip parodyta lentelėje. Sumaišymo matrica parodo duomenų kiekį, kuriam buvo teisingai prognozuota teigiama klasė, teisingai prognozuota neigiama klasė, neteisingai prognozuota teigiama klasė, bei neteisingai prognozuota neigiama klasė.

2 lentelė. Sumaišymo matrica dvimatei klasifikacijai vertinti.

	Tikra teigiama klasė	**Tikra neigiama klasė**
Prognozuota teigiama klasė	Teisinga teigiama *(**TP** – True positive)*	Neteisinga teigiama *(**FP** – False positive)*
Prognozuota neigiama klasė	Neteisinga neigiama *(**FP** – False positive)*	Teisinga neigiama *(**TN** – True negative)*

Klasifikavimo metodų prognozavimo rezultatams vertinti dažniausiai naudojamos šios metrikos:

Tikslumas *(acc)* - *(angl. Accuracy)* matuoja teisingai prognozuotų klasių kiekio santykį su visa klasifikavimo imtimi.

$$acc = \frac{tp + tn}{tp + fp + tn + fn} \qquad (40)$$

Preciziškumas *(p)* – *(angl. Precision)* matuoja teisingai prognozuotų teigiamos klasės reikšmių santykį su visa teigiamos klasės apimtimi.

$$p = \frac{tp}{tp + fp} \qquad (41)$$

Atkūrimas *(r)* – *(angl. Recall)* matuoja teisingai prognozuotų teigiamos klasės reikšmių santykį su teisingai klasifikuota imtimi.

$$r = \frac{tp}{tp + tn} \qquad (42)$$

Jautrumas *(TPR)* – *(angl. Sensitivity / true positive rate)* matuoja teisingai prognozuotų teigiamos klasės reikšmių santykį su teigiama klase.

$$TPR = \frac{tp}{tp + fn} \qquad (43)$$

Specifiškumas *(SPC) – (angl. Specificity / true negative rate)* matuoja teisingai prognozuotų neigiamų klasės reikšmių santykį su neigiama klase.

$$SPC = \frac{tn}{fp + tn} \qquad (44)$$

F1 įvertis *(FS) – (angl.F1 Score, F-Score, F-Measure)* matuoja harmoninį vidurkį tarp atkūrimo ir preciziškumo reikšmių. Matas svyruoja nuo 0 iki 1, idealiu atveju metrikos reikšmė yra 1.

$$FS = \frac{2 * p * r}{p + r} \qquad (45)$$

AUC – *(angl. area under the ROC curve (*erdvė po ROC kreive*))* šis prognozavimo kokybės matas laikomas vienu taisyklingiausiu vertinant modelio kokybę. Priešingai nei modelių kokybės matai minėti prieš tai šis matas apibrėžia bendrą klasifikatoriaus kokybę. Matą galima apskaičiuoti naudojantis formule:

$$AUC = \frac{S_p - \frac{n_p(n_n + 1)}{2}}{n_p * n_n} \qquad (46)$$

Kur S_p – visų teigiamos klasės pavyzdžių suma, n_p ir n_n – teigiami ir neigiami pavyzdžiai.[34]

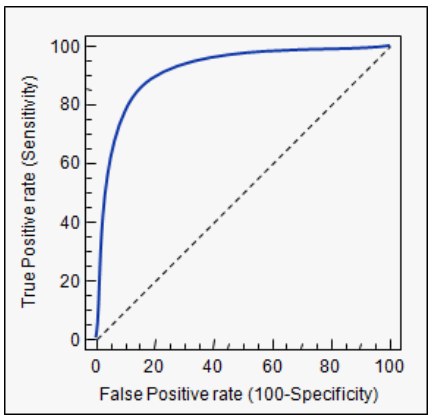
Thomas G. Tape savo knygoje „Interpreting Diagnostic Test“ pateikia šiuos AUC reikšmių vertinimus:

- 0.90-1 = Puikus (A)
- 0.80-0.90 = Geras (B)
- 0.70-0.80 = Patenkinamas (C)
- 0.60-0.70 = Prastas (D)
- 0.50-0.60 = Blogas (F) [33]

Vertinant tiek teorine, tiek empirine prasme AUC matas buvo pripažintas geresniu nei tikslumo matas, vertinant klasifikatoriaus tikslumą, tačiau šis matas turi didelė skaičiavimo kainą [34].

Taipogi klasifikavimo tikslumui nustatyti yra populiaru naudoti šias kreives:

ROC (*angl. Receive Operating Characteristic curve*) kreivė 4 pav. atvaizduojanti teisingai prognozuojamų teigiamų reikšmių ir neteisingai prognozuojamų teigiamų reikšmių santykį skirtinguose imties taškuose.

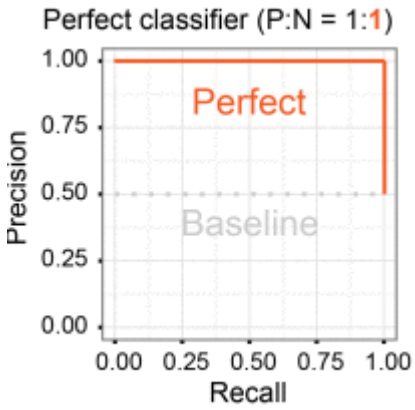


4 pav. ROC kreivė [34]

ROC kreivės savybės:

- Atvaizduoja sąryšį tarp jautrumu ir specifiškumo. (Didėjant jautrumui mažės specifiškumas).
- Kuo arčiau kreivė kairios sienos ir viršutinės sienos ROC erdvėje, tuo gauti klasifikavimo rezultatai yra tikslesni.
- Kuo arčiau kreivė pasiekia 45 laipsnių įstrižainę, tuo rezultatai gautų klasifikavimo rezultatų tikslumas yra mažesnis. [33]

PR kreivė – atvaizduoja preciziškumo ir atkūrimo santykį. Kreivė savo logika yra itin artima ROC kreivei. Idealiu atveju kreivė turėtų būti sudaryta iš dviejų tiesių linijų sienos viršuje ir dešiniame kampe 5 pav.



5 pav. PR kreivė [35]

PR kreivės savybės:

- Interpoliacija tarp dviejų preciziškumo – atkūrimo taškų yra netiesinė.
- Santykis tarp teigiamų ir neigiamų reikšmių apibūdina pradinę kreivę.

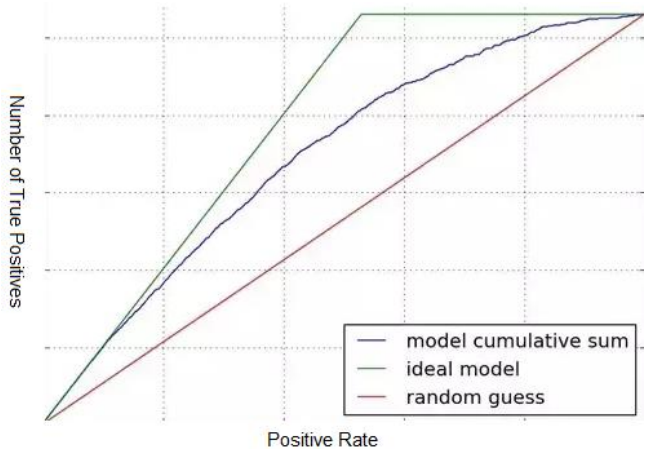
PR kreivė dažniau yra naudojama siekiant išsiaiškinti kiek reikšmingi rezultatai yra lyginant su atsitiktinių spėjimu, o ROC kreivės rezultatai dažniau naudojami siekianti

įvertinti klasifikatoriaus tikėtiną atlikimą apskritai, neatsižvelgiant į atsitiktinio spėjimo tikimybes [35].

LIFT kreivė – atvaizduoja visų teisingai prognozuotų teigiamų reikšmių santykį su teigiamai spėtų reikšmių santykiu.

$$x = \frac{tp + fp}{p + n}, \qquad y = TP \qquad (47)$$

6 pav. atvaizduota LIFT kreivė tobulu atveju, atsitiktinio spėjimo atveju ir atsitiktinio modelio atveju [36].



6 pav. Pav. Lift kreivė [36]

4. EMPIRINIS TYRIMAS

4.1.Pradiniai duomenų rinkiniai

Duomenų analizei atlikti buvo naudojami 2 skirtingi duomenų rinkiniai. Pirmajame duomenų rinkinyje buvo naudojami vieši tarpusavio skolinimo platformos „Bondora“ duomenys, gauti iš https://www.bondora.com/en/public-reports svetainės. Analizuojamų duomenų rinkinys apima duomenis sukauptus nuo 2015-03-19 iki 2017-02-01.

Pradiniame duomenų rinkinyje iš viso buvo 32369 eilutės ir 111 stulpelių. Duomenų rinkinį sudarė asmeniniai duomenys apie skolintoją (pvz. amžius, lytis, šalis, uždarbis, t. t.), metaduomenys, skirti aprašyti paskolos įvykį (pvz. paskolos į sistemą įkėlimo data, vartotojo id, paskolos id, t. t.), paskolos grąžinimo vertinimai, remiantis „Bondora“ svetaine (pvz. paskolos negrąžinimo tikimybė, kreditingumo balas, t. t.), duomenys apie įvykius po paskolos paėmimo (pvz. įmokų grąžinimo laikotarpiai, grąžintos sumos, perskaičiuotosios palūkanos t. t.).

Antrajame duomenų rinkinyje buvo naudojami Amerikoje pirmaujančios skolinimo platformos „Lending Club“ duomenys, gauti iš https://www.lendingclub.com/info/download-data.action svetainės. Analizuojamų duomenų rinkinys apima duomenis sukauptus nuo 2007 gruodžio mėnesio iki 2016 metų pabaigos.

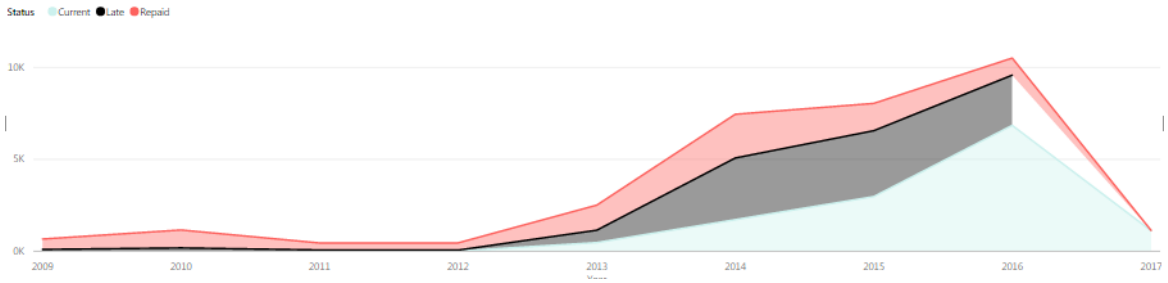
Pradiniame duomenų rinkinyje iš viso buvo 602590 eilutės ir 129 stulpeliai. Duomenų rinkinyje yra kur kas mažiau duomenų apie darbuotojo išsilavinimą, darbo pobūdį nei „Bondora“ duomenų rinkinyje. Taip pat kur kas mažiau duomenų apie paskolą ėmusio asmens elgesį po paskolos paėmimo. Itin daug stulpelių neturinčių nė vienos reikšmės arba turinčių tik vieną unikalią reikšmę. Tačiau „Lending Club“ turi kur kas daugiau informacijos apie asmens kredito istoriją ir patikimumą (pvz. pirmos kredito operacijos atlikimo data, vieši įrašai). Galime netgi matyti tekstą, kuriuo buvo kreipiamasi dėl paskolos.

4.2.Aprašomoji duomenų analizė

4.2.1. Tarpusavio skolinimo platformos „Bondora“ atvejis

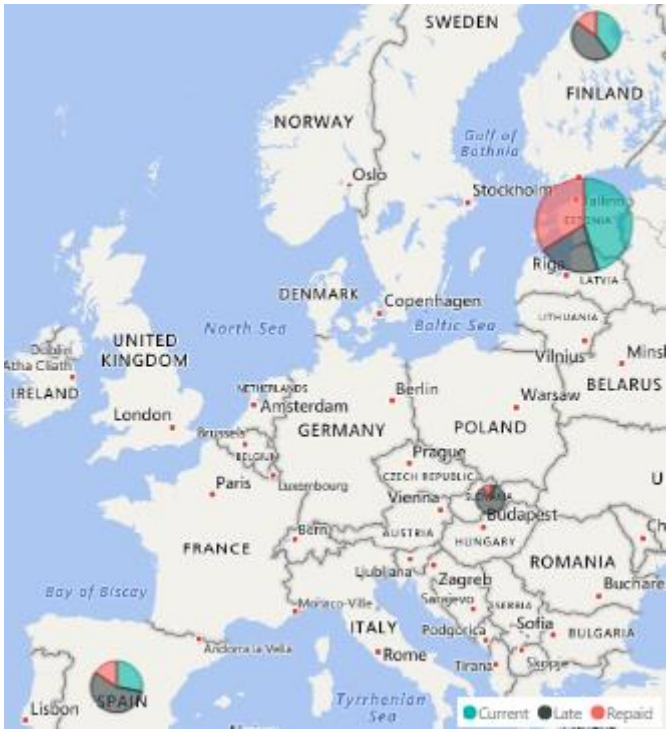
Atlikus „Bondora“ skolinimosi platformos paskolų kiekio priklausomybę nuo laiko (7 pav) buvo pastebėta, kad paskolų kiekis pradėjo augti 2012 metais, kiekvienais sekančiais metais įgydamas vis didesnę vertę (2012 metais ~500 paskolų, 2013 metais ~ 2000 paskolų,2014 metais ~ 7000 paskolų, 2015 ~8000 paskolų, 2016 ~ 10000 paskolų).

Pastebime, kad iki 2012 metų vartotojai turėjo apytiksliai 40% daugiau grąžintų paskolų nei vėluojančių, tačiau su kiekvienais tolesniais metais negrąžintų paskolų kiekis viršydavo grąžintų paskolų kiekį apytiksliai 30%.



7 pav. Paskolų ir jų statusų pokytis laike. Tarpusavio skolinimosi platformos „Bondora“ atvejis

Iš 8 pav. matome, kad daugiausiai klientų tarpusavio skolinimo platformos „Bondora“ tarpe yra Estijos piliečių (18671 klientai), antroje vietoje Ispanija (6471 klientai), trečioje vietoje Suomija (5600 klientai), ketvirtoje vietoje Slovakija (283 klientai).

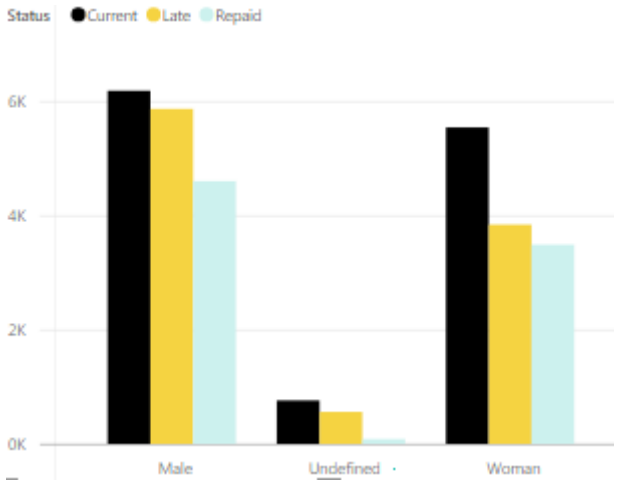


8 pav. Paskolų ir statusų pasiskirstymas skirtingose šalyse. Tarpusavio skolinimosi platformos „Bondora“ atvejis

Įvertinus statuso reikšmes skirtingoms šalims matome, kad santykinai daugiausiai vėluojančių skolų turi Slovakijos klientai. Antroje vietoje pagal skolų negrąžinimą yra Ispanija. Santykinai daugiausiai skolų grąžino Estijos klientai, antroje vietoje Suomijos klientai.

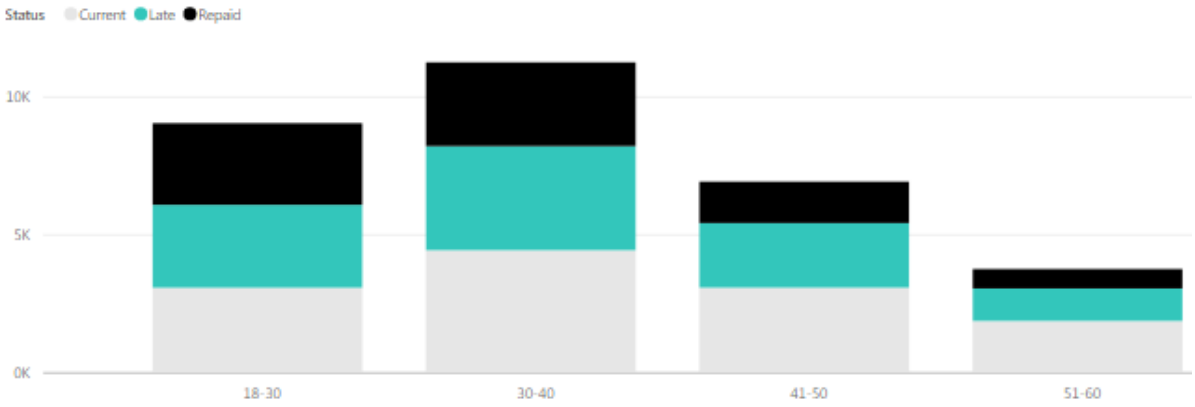
Įvertinus lyčių pasiskirstymą (9 pav.) pastebėta, kad vyrai ima kur kas daugiau paskolų negu moterys. Taip pat matome, kad vyrai yra kur kas labiau linkę negrąžinti

paskolos negu moterys. Pastebėta, kad jei asmuo nenori nurodyti savo lyties, tuomet tikimybė, kad jis negrąžins paskolos yra labai didelė.



9 pav. Paskolų ir jų statusų pasiskirstymas tarp skirtingų lyčių. Tarpusavio skolinimosi platformos „Bondora“ atvejis

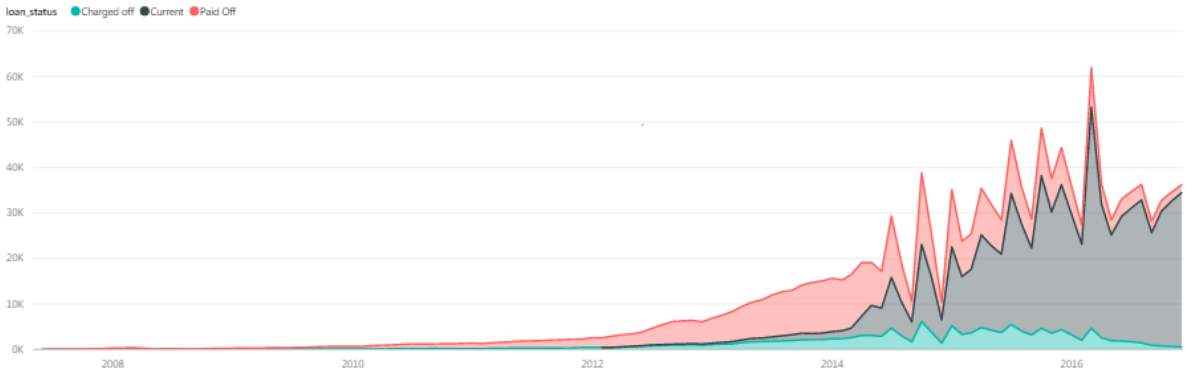
Įvertinus tarpusavio skolinimosi platformos „Bondora“ klientų amžių (10 pav.) buvo nustatyta, kad pagrindiniai platformos klientai yra 30 – 40 metų asmenys. Pastebėta, kad negrąžintų paskolų santykis yra didėjantis didėjant kliento amžiui.



10 pav. Paskolų ir jų statusų pasiskirstymas tarp skirtingų amžiaus grupių. Tarpusavio skolinimosi platformos „Bondora“ atvejis

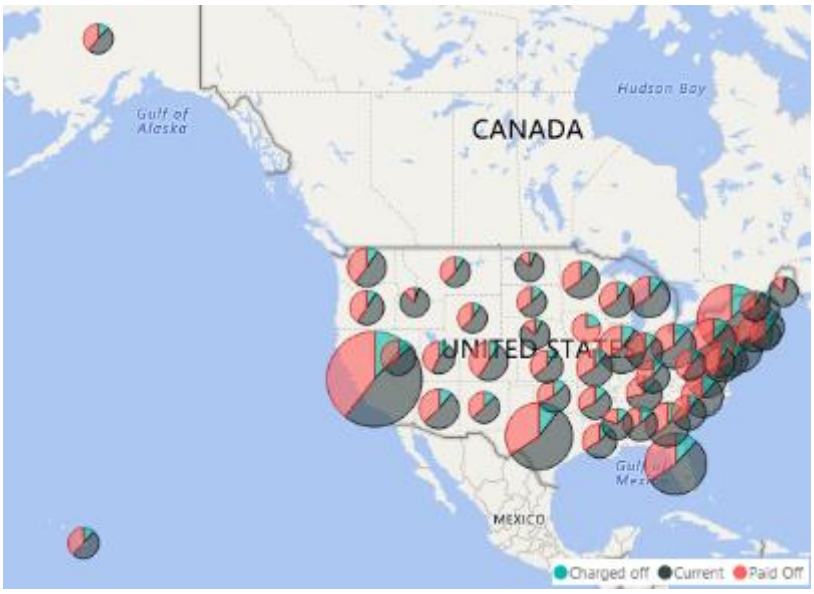
4.2.2. Tarpusavio skolinimo platformos „Lending Club“ atvejis

Atlikus paskolų kiekio analizę nuo laiko (11 pav.) pastebime jog paslaugų kiekis ženkliai pradėjo augti nuo 2012 metų. Didžiausias paskolų kiekis buvo išduotas 2016 metų gruodžio mėnesį. Taipogi pastebime, kad didžiausią dalį sudaro esamos skolos, tačiau įvertinus išmokėtas ir vėluojančias paskolas, pastebime, kad išmokėtų paskolų kiekis yra kur kas didesnis nei neišmokėtų paskolų kiekis.



11 pav. Paskolų ir jų statusų pokytis laike. Tarpusavio skolinimosi platformos „Lending Club“ atvejis

Įvertinus duomenis pagal valstijas (12 pav.) buvo nustatyta, kad daugiausia paskolą ėmusių asmenų yra Kalifornijos valstijoje. Šiek tiek atsilieka, tačiau taip pat didelį kiekį paskolų yra išdavusios šios valstijos – Niujorkas, Teksasas, Florida. Nors grąžintų paskolų santykis kiekvienoje valstijoje unikalus, nepaisant geografinės vietovės, visose valstijose yra didesnis išmokėtų paskolų kiekis, nei neišmokėtų paskolų kiekis.



12 pav. Paskolų ir statusų pasiskirstymas skirtingose šalyse. Tarpusavio skolinimosi platformos „Lending Club“ atvejis

Buvo pastebėta, kad tarpusavio skolinimo platforma „Lending Club“ nekaupia duomenų apie paskolą ėmusio asmens lytį, amžių, santuokinį statusą ar kitas asmenines detales. Tai gali būti nulemta dėl labai stiprios antidiskriminacinės politikos šalyje.

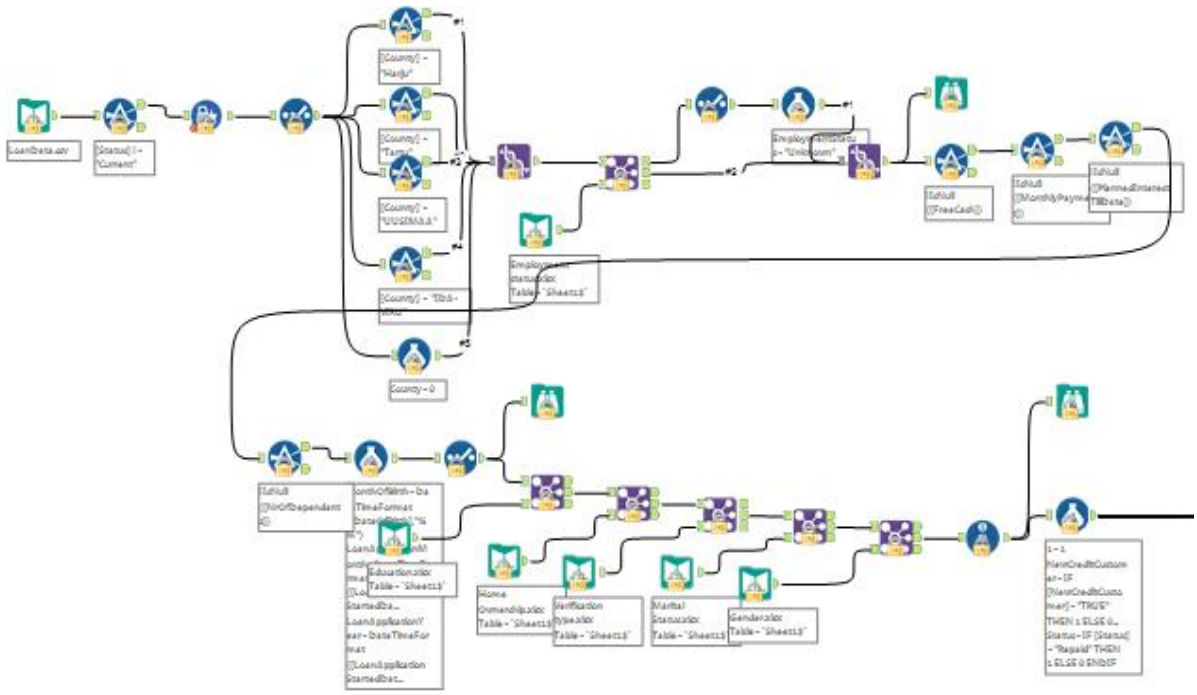
4.3.Duomenų paruošimas klasifikavimui

Tam, kad būtų galima klasifikuoti duomenis, buvo reikalinga atlikti duomenų paruošimą. Norint taikyti įprastinius klasifikavimo metodus negali būti tuščių eilučių, duomenys turi būti skaitmeninio ar kategorinio tipo. Papildomai norėta sukurti naujus

kintamuosius iš jau turimų kintamųjų. Duomenų tvarkymas buvo atliekamas naudojant „Alteryx“ programinę įrangą.

„Alteryx“ programinė įranga yra viena pirmaujančių priemonių, skirta savitarnos analitikai. Šis įrankis yra ypač galingas, gebantis apdoroti itin didelius duomenų kiekius bei suteikiantis galimybę įžvelgti gilesnes įžvalgas iš duomenų per labai trumpą laiko tarpą. Programinė įranga skirta tiek duomenų valymui, tvarkymui, paruošimui, tiek išsamioms analizėms bei jų išvadoms gauti. Programine įranga itin lengva naudotis. Šį platforma turi galimybę analizuoti duomenis tiek esančius serveriuose, tiek debesyse, tiek paprastose lentelėse. Duomenys taip pat paprastai gali būti įrašomi į serverius ar debesis, su programa lengvai galima kurti raportus. Programinė įranga savyje turi paketus, kuriais gali aprašyti skaitmeninio mokymo algoritmus, taip pat „Alteryx“ programinė įranga turi tiesioginę sąsają su „R“ programine įranga savo viduje, kurią galima panaudoti algoritmų išplėtimui ar norimų rezultatų gavybai [37].

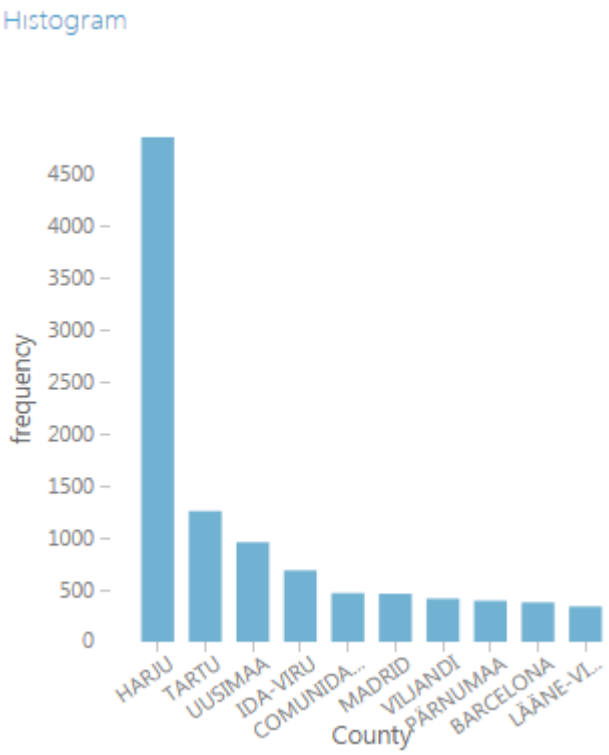
Toliau pavaizduotame paveikslėlyje (13 pav.) atvaizduota procedūra, kuri buvo naudota atlikti duomenų paruošimui.



13 pav. Tarpusavio skolinimo platformos „Bondora“ duomenų paruošimo procesas

Tarpusavio skolinimo platformos „Bondora“ duomenims klasifikuoti pasirinktas klasifikatorius statusas, kuris apibūdina ar paskola bus grąžinta laiku, ar ją grąžinti bus vėluojama. Duomenys, kurių statusas teigia, kad paskola esama, buvo pašalinti, nes klasifikavimui nesuteikia realaus rezultato apie įsipareigojimų vykdymą. Taip buvo prarasta apytiksliai 50 % duomenų. Taip pat buvo pašalinti visi stulpeliai, kurie nėra pateikiami

investuotojams prieš paskolą, išskyrus klasifikatorių statusas. Kadangi savivaldybės kintamųjų buvo per daug, kad juos būtų galima paversti faktoriais, buvo atrinkta 10 pačių populiariausių savivaldybių ir jos buvo pavaizduotos histograma (14 pav.). Iš šių savivaldybių atrinktos 4 pačios populiariausios, o likusioms priskirtas pavadinimas „Kita“. Kategoriniai kintamieji pradiniame duomenų rinkinyje buvo pavaizduoti iš eilės einančiais skaičiais. Kadangi buvo norėta juos palikti kategoriniais kintamaisiais, jų reikšmės buvo importuotos naudojantis išoriniais failais ir sujungiamos su kintamaisiais. Skaitinio tipo stulpeliai, kuriuose trūko didelės duomenų apimties, buvo pašalinti. Eilutės, priklausiusios skaitinio tipo stulpeliams, kuriuose trūko didelės duomenų apimties, buvo pašalintos. Taip pat iš esamų kintamųjų buvo sukurti nauji kintamieji, tokie kaip gimimo mėnesis, paskolos prašymo mėnesis, paskolos prašymo metai, paskolos prašymo valanda, paskolos prašymo savaitės diena.

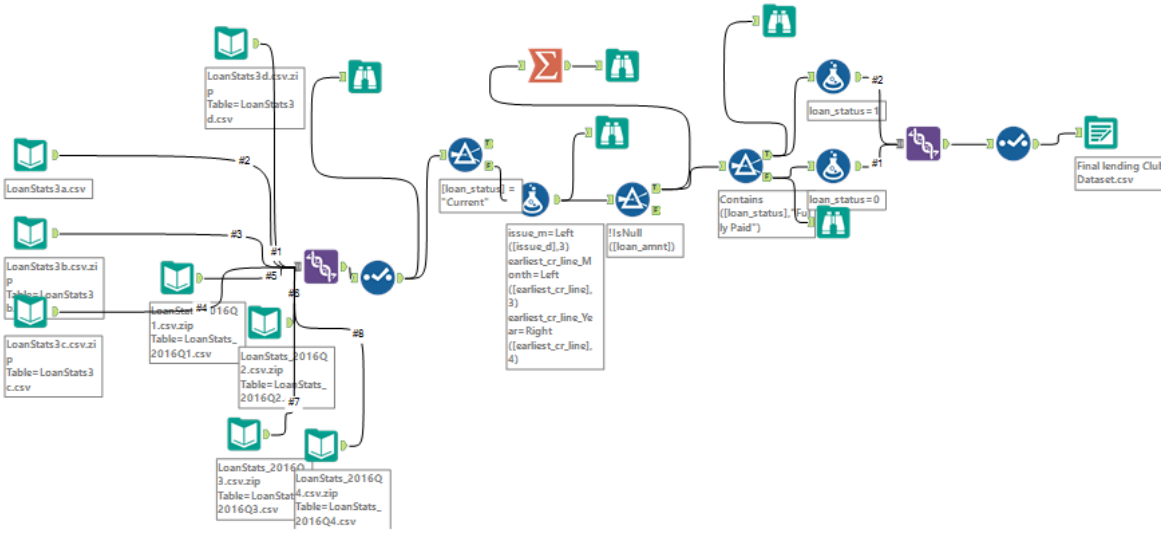


14 pav. Savivaldybių histograma

Atlikus duomenų tvarkymą, duomenų rinkinyje liko 13661 eilutės ir 44 stulpeliai. Galutinis duomenų rinkinys bus naudojamas duomenų klasifikavimui atlikti.

Siekiant atlikti duomenų paruošimą naudojantis „Lending Club“ tarpusavio skolinimo sistemos duomenis reikėjo įdėti mažiau pastangų (15 pav.). Atskirų metų duomenys buvo atskiruose duomenų rinkiniuose todėl reikėjo juos apjungti. Tuomet buvo pašalinti stulpeliai, kuriuose trūko visų arba labai daug reikšmių. Taip pat pašalinti tie stulpeliai, kurie gali būti turimi tik po suteiktos paskolos. Klasifikavimui pasirinktas

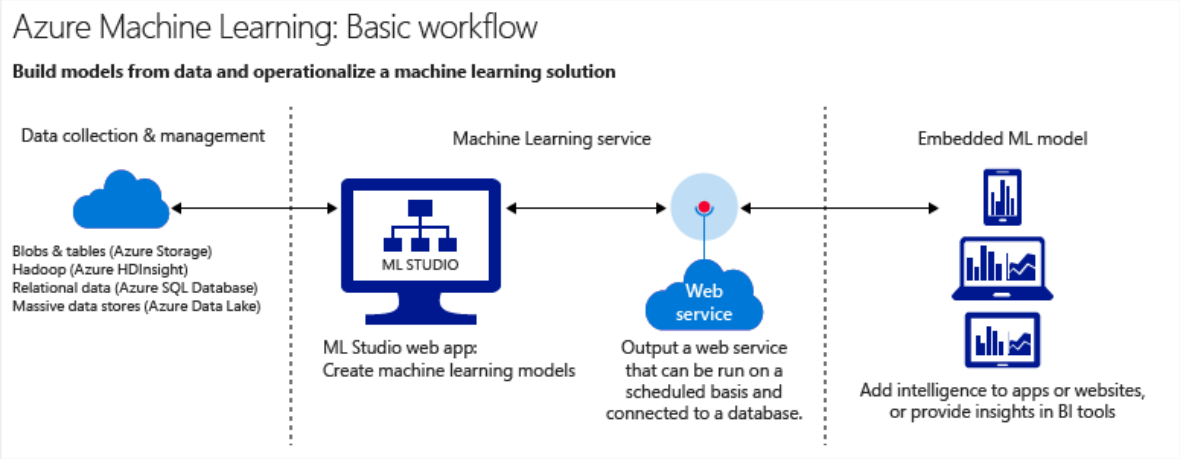
kintamasis paskolos statusas. Buvo pašalinti visi kintamieji, kurie reprezentavo vis dar esamas paskolas. Sudaryti nauji kintamieji, tokie kaip paskolos paėmimo mėnuo, pirmo kredito įrašo metai, pirmo kredito įrašo mėnuo. Deja didesnio tikslumo iš datų gauti nebuvo galima, nes naudojamos datos formatas yra metai ir diena. Duomenų stulpelis paskolos statusas buvo atskirtas ir kintamieji, kurių statusas „Fully Paid“ priskirti vienetui, tie kintamieji kurių statusas rodė kitą reikšmę, kuri rodė paskolos nesumokėjimą priskirti 0.



15 pav. .Tarpusavio skolinimo platformos „Lending Club“ duomenų paruošimo procesas

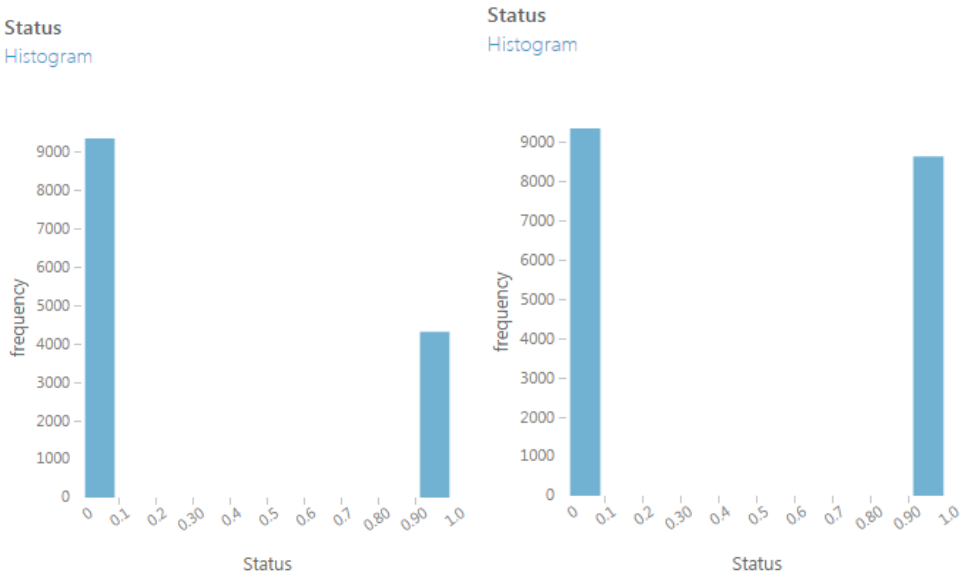
Atlikus duomenų tvarkymą, duomenų rinkinyje liko 566980 eilutės ir 26 stulpeliai. Galutinis duomenų rinkinys bus naudojamas duomenų klasifikavimui atlikti.

Pirminio bandymo metu buvo bandoma atlikti modelių analizę naudojantis „R“ programinę įrangą. Tačiau pastoviai buvo susiduriama su atminties paskirstymo klaidomis *“Error: cannot allocate vector of size n Mb”* arba tekdavo laukti labai ilgą laiką paprastoms operacijoms atlikti. Dėl šios priežasties buvo ieškoma būdu atlikti analizę greitesniu būdu. Tokiu būdu buvo atrasta „Azure Machine Learning Studio“. „Azure Machine Learning Studio“ (16 pav.) leidžia duomenis patalpinti debesyje ir atlikti mašininio mokymosi algoritmus naudojantis internetine aplikacija. Didžiausias programinės įrangos pliusas yra galimybė dislokuoti modelį taip nesunkiai jį susiejant su aplikacijų programinėmis sąsajomis (API) [38].

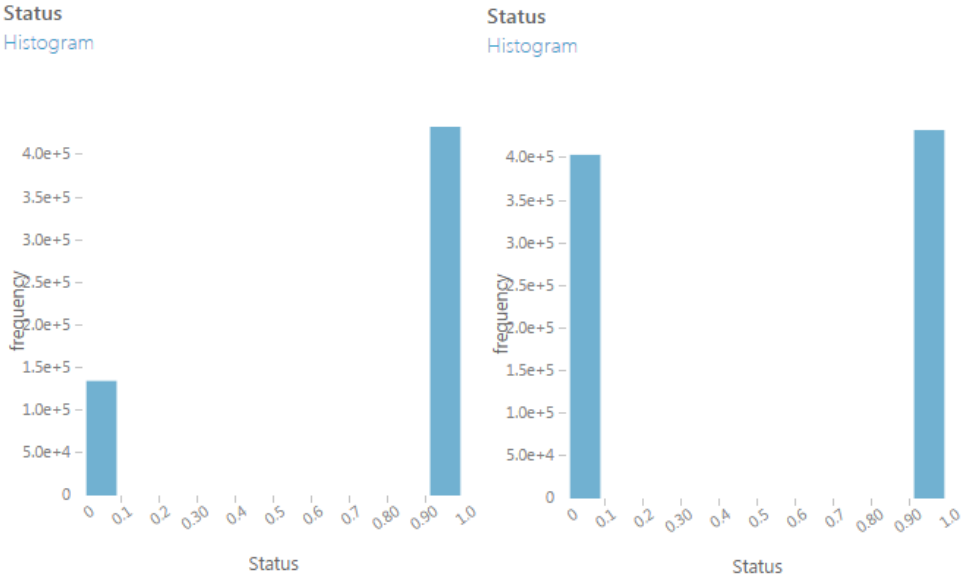


16 pav. „Microsoft Azure Machine Learning Studio“ veiklos diagrama. [38]

Likęs duomenų paruošimas buvo atliekamas „Microsoft Azure Machine Learning Studio“ programoje. Visų pirma duomenys buvo normalizuoti naudojantis zScore transformacijos metodu. Kategoriniuose stulpeliuose rastos tuščios reikšmės buvo pakeičiamos stulpelio modos reikšme (dažniausiai pasikartojančia požymio reikšme imtyje), tuo tarpu skaitmenines vertes turintys stulpeliai buvo paverčiami į stulpelio medianą (skaičių aibės vidurinį (centrinį) skaičių, kuris aibę dalija į dvi dalis). Abiejų duomenų imtyse buvo pastebimas disbalansas. „Bondora“ skolinimo platformos atveju disbalansas buvo mažesnis nei „Lending Club“ tačiau duomenis vis tiek reikėjo tvarkyti. Duomenų disbalansui panaikinti buvo naudojama sintetinė mažumų pernaudojimo technika SMOTE (*angl. Synthetic Minority Oversampling Technique)*. SMOTE realizacijai atlikti iš anksto reikia nusistatyti tolesnius parametrus: SMOTE procentą, kur 0% reikštų nepakitusį .imties didį. 100% - imtis padidinta mažesnės klasės imties dydžiu; artimiausių kaimynų kiekį – kuris nurodo ypatybių erdvės dydį, kurias naudoja sudarydamas naujas eilutes. Kuo didesnis artimiausių kaimynų kiekis, tuo įvairesnės ypatybės gaunamos. Naudodami mažą artimiausių kaimynų kiekį turime ypatybes panašiausias į originalų duomenų rinkinį [40]. „Bondora atveju“ naudojamas 100% SMOTE procentas ir 5 artimiausi kaimynai. Kaip matome (17 pav) prieš panaudojus SMOTE techniką mažesniojoje klasėje buvo 31,59% duomenų, o ją atlikus duomenų kiekis mažesniojoje klasėje tapo 48,01%. Tokiu būdu duomenų kiekis padidėjo iki 17975. „Lending Club“ atveju naudojamas 200% SMOTE procentas. Kaip matome (18 pav.) prieš panaudojant SMOTE techniką mažesniojoje klasėje buvo 23,7% duomenų, o po SMOTE technikos panaudojimo mažesnioji klasė sudarė 48,3% duomenų. Galutinį „Lending Club“ duomenų rinkinį sudarė 835922 eilutės. Taipogi, buvo pastebėta, kad duomenims panaudojus SMOTE techniką modelių tikslumo įvertis AUC žymiai pagerėjo.



17 pav. Status kintamojo histograma prieš ir po panaudojant SMOTE technika. „Bondora“ atvejis



18 pav. Status kintamojo histograma prieš ir po panaudojant SMOTE technika. „Lending Club“ atvejis

Stebėdami klasių disbalansą galime daryti išvadą, kad asmenys imantys paskolas naudojantis tarpusavio skolinimosi platforma „Bondora“ linkę jų negrąžinti, tuo tarpu asmenys besinaudojantys „Lending Club“ tarpusavio skolinimosi platforma paskolų negrąžina tik 24% atvejų.

Modeliams testuoti abu duomenų rinkiniai buvo išskaidyti į apmokymo ir testavimo imtis. Apmokymo imtį sudarė 70% duomenų tuo tarpu testavimo imtį sudarė 30% duomenų. Duomenų skaidymas buvo atliktas atsitiktinai parenkant duomenis iš duomenų imties. Imtis išskaidyta naudojant suratifikavimą, tokiu atveju užtikrinant, kad abiejų klasių duomenų kiekis apmokymo imtyje yra apyilgis. „Bondora“ atveju apmokymo imtyje buvo 12583 eilutės iš kurių 6542 neigiamos eilutės (paskola negrąžinta) ir 6041

teigiamos eilutės (paskola grąžinta), o testavimo imtyje viso buvo 5392 eilutės, iš kurių 2803 neigiamos eilutės ir 2589 teigiamos eilutės. Tuo tarpu „Lending Club“ apmokymo imtyje buvo 585145 eilutės iš kurių 282389 neigiamos eilutės ir 302756 teigiamos eilutės, o testavimo imtyje 250777 eilutės iš kurių 121024 neigiamos reikšmės ir 129753 teigiamos reikšmės.

Modelių tikslumui pagerinti buvo naudojamas modelio parametrų derinimas. Pasirinktas atsitiktinės imties (*angl. random sweep*) metodas, kuris laikomas tikslesniu nei gardelių paieškos metodas [39]. Buvo nustatyta 60 iteracijų, kurių metu modulis atsitiktinai parenka parametrų reikšmes, kurių metu ieškomas modelis turintis aukščiausią AUC parametrą.

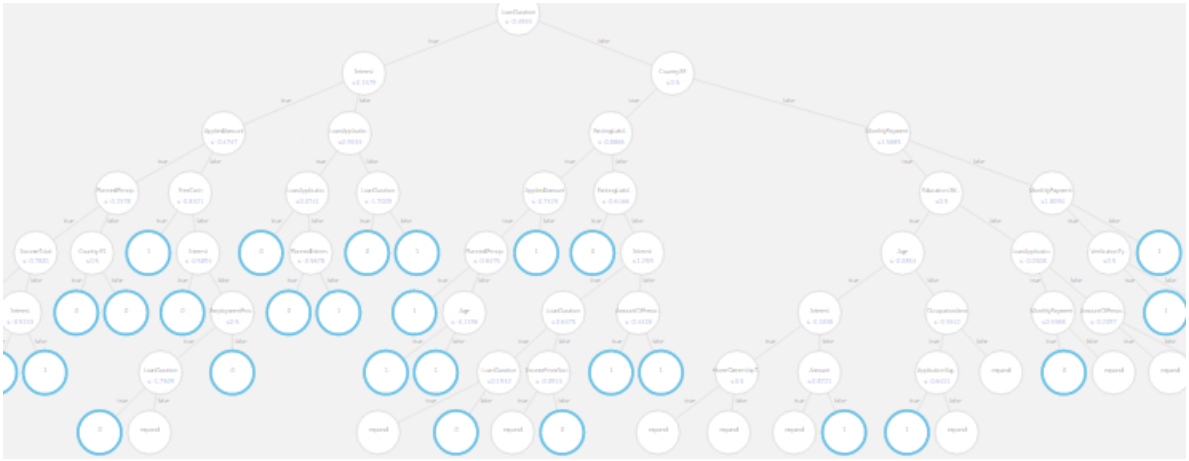
4.4.Modelių generavimas. Tarpusavio skolinimo platformos „Bondora“ duomenų atvejis

Duomenims klasifikuoti buvo išbandyti 8 skirtingi klasifikavimo metodai. Prognozuojamas kintamasis yra statusas. Prognozei pasirinktas slenkstis 0,5. Geriausi rezultatai buvo gauti duomenis klasifikuojant naudojant sustiprinto sprendimų medžio (*angl. boosted decision tree*) klasifikatorių.

4.4.1. Sustiprintas sprendimų medis

Naudojant modelio parametrų derinimo algoritmą buvo nustatyta, kad sustiprintų medžių modelis geriausiai veikia naudojant šiuos parametrus:

- Lapų kiekis: 71
- Mažiausias atskirų lapų atvejų kiekis: 47
- Mokymosi greitis: 0,091365
- Medžių kiekis: 407

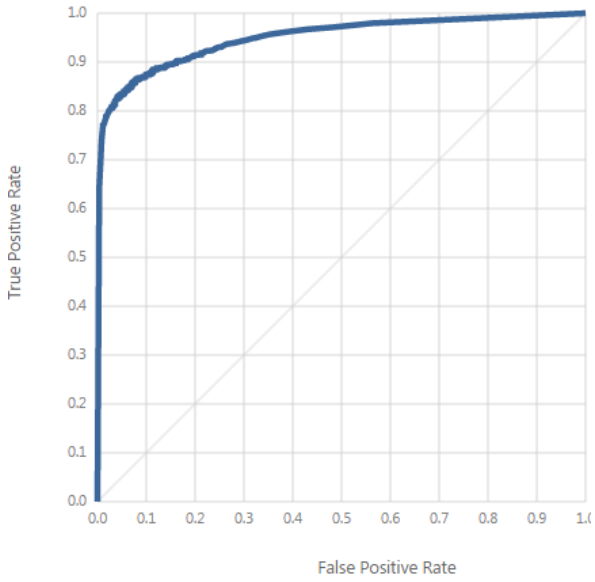


19 pav. Skolinimo platformos „Bondora“ sprendimų medis

Naudojantis šiuo klasifikatoriumi sukuriami scenarijai, kur tam tikrų kintamųjų visuma lemia asmens tikimybę grąžinti paskolą arba vėluoti ją grąžinti. 19 pav. pateikta viršutinė medžio dalis vaizduojanti galimus scenarijus paskolą grąžinti laiku (1) arba negrąžinti laiku (0).

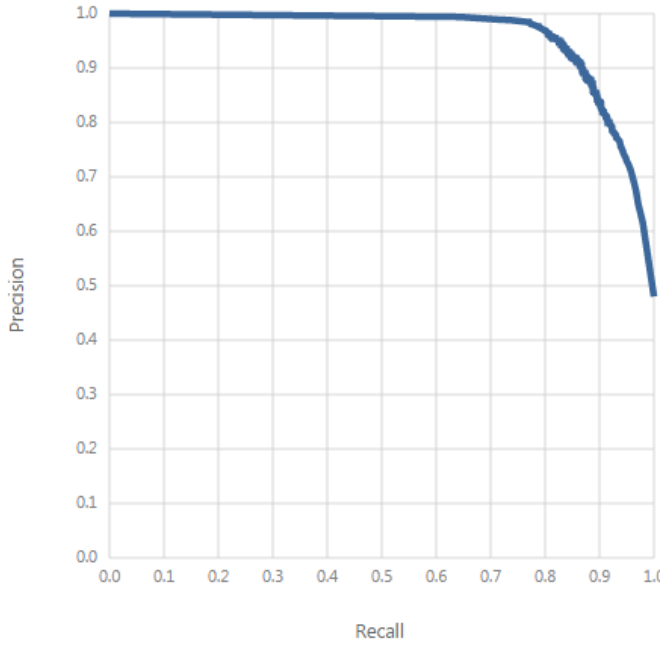
AUC įvertis lygus 0.95 yra itin artimas vienetui, o tai rodo jog sudarytas modelis yra labai aukšto tikslumo. 89% reikšmių buvo prognozuotos taisyklingai, tuo tarpu, kad paskola bus grąžinta laiku buvo atspėta 89,8% lyginant su visomis šios klasės reikšmėmis.

Prognozavimo rezultatų gerumą taip matome ir iš ROC kreivės 20 pav.. Kreivė beveik sutampa su kairiąja ir viršutinėmis sienomis, kas rodo puikius modelio prognozavimo rezultatus.



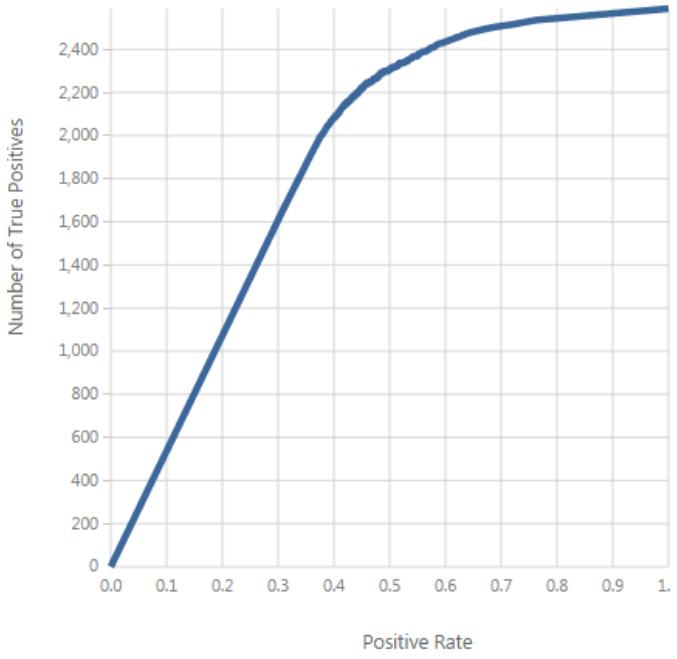
20 pav. Sustiprinto sprendimų medžio ROC kreivė skolinimosi platformos „Bondora“ atvejis

Modelio rezultatų tikslumu įsitikiname ir matydami PR kreivę 21 pav., kuri beveik sutampa su viršutine grafiko siena ir yra labai artima dešiniajai grafiko sienai.



21 pav. Sustiprinto sprendimų medžio PR kreivė skolinimosi platformos „Bondora“ atvejis

Sustiprinto sprendimų medžio LIFT kreivė taip pat yra artima idealiai LIFT kreivei (22 pav.).



22 pav. Sustiprinto sprendimų medžio LIFT kreivė

4.4.2. Logistinė regresija

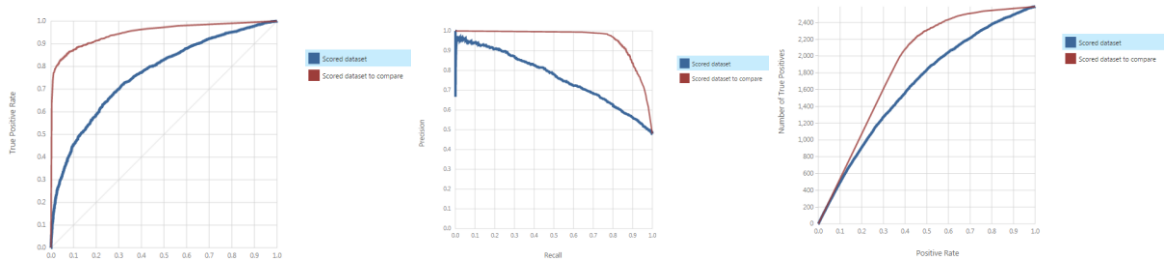
Naudojant logistinę regresiją buvo gautas žymiai mažesnis modelio tikslumas. Taikant modelio parametrų derinimo algoritmą buvo nustatyta, jog modelis geriausiai veikia su šiais parametrais:

- Optimizacijos tolerancija: 1.28624663E-06

- L1 svoris: 0.0462255
- L2 svoris: 0.0701681
- Atminties dydis 27

Nors modelio tikslumas yra patenkinamas, tačiau žymiai žemesnis lyginant su kitais modeliais. AUC lygus 0.767, kuris priskiriamas patenkinamam tikslumui, iš viso teisingai klasifikuotos 70,2% reikšmių lyginant su visa imtimi.

Rezultatų tikslumą matome ir iš ROC, PR, LIFT kreivių 23 pav.. Matome, kad lyginant su sustiprintu sprendimų medžiu kreivės yra kur kas žemiau, kas rodo mažesnį tikslumą, tačiau kreivės yra aukščiau pradinės linijos, kas žymį modelio gebėjimą prognozuoti reikšmes geriau nei atsitiktinio sprendimo metu.



23 pav. Logistinės regresijos ROC, PR, LIFT kreivės lyginant su sustiprinto sprendimų medžio ROC, PR, LIFT kreivėmis skolinimosi platformos „Bondora“ atvejis

Apžvelgiant modelio rezultatus matome, kad didžiausią teigiamą įtaką dėl paskolos grąžinimo laiku turi šie kintamieji – pajamos gaunamos iš socialinio aprūpinimo, turimų vaikų kiekis lygus dviem, prieš tai turėtų paskolų kiekis, gaunamos pajamos, įvairios užimamos pareigos, kurių pavadinimai duomenų rinkinyje yra įvesti vietine kalba Didžiausią neigiamą įtaką daro šie kintamieji – laisvi pinigai, mėnesinis paskolos mokestis, įvairios užimamos pareigos, kurių pavadinimai duomenų rinkinyje yra įvesti vietine kalba, palūkanos, paskolos terminas

4.4.3. Atraminiai vektoriai

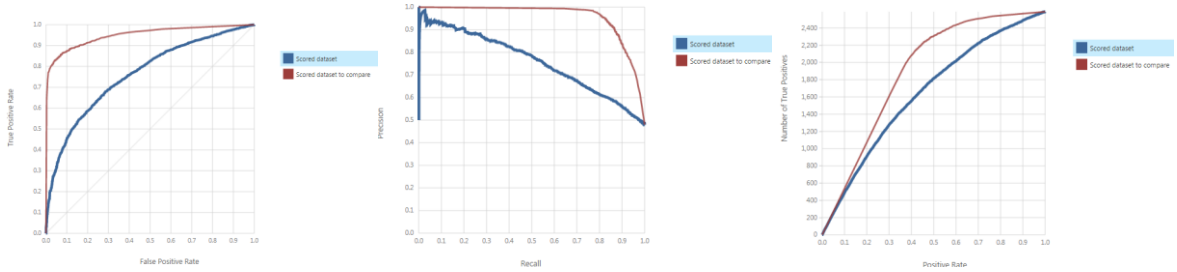
Naudojant atraminių vektorių *(angl. SVM – support vector machine)* klasifikavimo modelį pasiektas patenkinamas tikslumas. Pritaikius modelio parametrų derinimo algoritmą buvo nustatyta, kad modelis geriausiai veikia naudojant šias parametrų reikšmes:

- Iteracijų kiekis: 97
- Lambda: 0.000186474339

Kadangi rezultatų tikslumo matas AUC lygus 0.761, galime daryti išvadą, kad atraminių vektorių klasifikatorius reikšmės prognozuoja patenkinamai ir rezultatų tikslumas

yra mažesnis net ir už logistinės regresijos rezultatus. Atraminių vektorių metodas gebėjo teisingai prognozuoti klasę 69,7% kintamųjų.

Iš ROC, PR, LIFT kreivių 24 pav. matome, kad kreivės yra kur kas žemiau už sustiprinto sprendimų medžio kreives, kas žymi kur kas prastesnį atraminių vektorių tikslumą, tačiau kadangi kreivė yra aukščiau pradinės linijos kreivės, tai rodo, kad modelis prognozuoja reikšmes geriau nei atsitiktinis klasės spėjimas.



24 pav. Atraminių vektorių ROC, PR, LIFT kreivės lyginant su sustiprinto sprendimų medžio ROC, PR, LIFT kreivėmis skolinimosi platformos „Bondora“ atvejis

Modelis parodė, kad didžiausią teigiamą įtaką paskolos grąžinimui laiku daro šie kintamieji: prieš tai turėtos paskolų kiekis, pajamos gaunamos iš socialinio aprūpinimo, prašytas paskolos dydis, turimų vaikų kiekis lygus 2, asmuo gyvena Estijoje, įvairios pareigybės užrašytos vietine kalba (tačiau dauguma jų lengvai išsiverčia į direktorių, savininką). Kintamieji, kurie daro didžiausią paskolos vėlavimui: laisvi pinigai, mėnesinės įmokos už paskolą, paskolos trukmė, amžius, palūkanos, asmens gyvenamoji šalis Ispanija, gautas paskolos dydis, nedirbantis asmuo, įvairios profesijos įrašytos vietinė kalba..

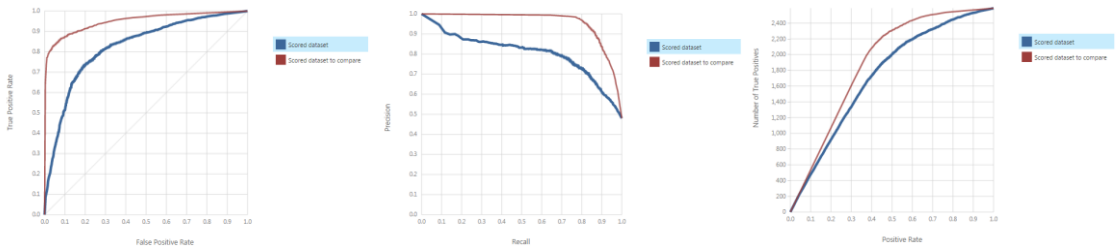
4.4.4. Gilieji atraminiai vektoriai

Taikant gilųjį atraminių vektorių klasifikatorių (*angl. DSVM – deep support vector machine*) buvo gautas geras tikslumas, tačiau tikslumas žemesnis nei naudojant sustiprinto sprendimų medžio klasifikatorių. Naudojant modelio parametrų derinimo algoritmą buvo nustatytą, jog modelio tikslumas yra geriausias naudojant šiuos parametrus:

- Medžių gylis: 6
- LambdaW: 0.0748782456
- Lambda Theta: 0.0279957131
- Sigma: 0.6682039
- Iteracijų kiekis: 10803

Modelio AUC lygus 0.825, tai rodo, jog modelio tikslumas yra geras, tačiau ne puikus. Tikslumo matas lygus 0.763, kuris rodo, jog teisingai buvo prognozuota 76,3% reikšmių.

Gilusis atraminių vektorių metodas gerokai aplenkė paprastąjį atraminių vektorių metodą, tačiau nepralenkė sustiprinto sprendimų medžio metodo. Kaip matome iš ROC, PR, LIFT kreivių 25 pav., jos yra kur arčiau sustiprintų medžių vektoriaus, nei prieš tai buvę du modeliai, tačiau šio modelio kreivės vis tiek yra žemiau nei sustiprintų medžių vektoriaus.



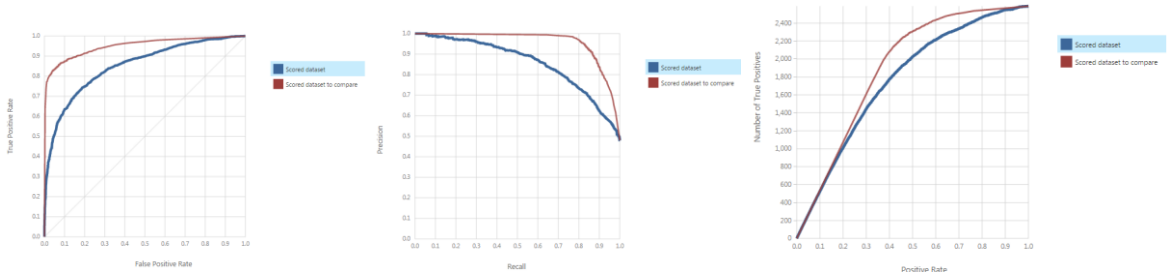
25 pav. Giliųjų atraminių vektorių ROC, PR, LIFT kreivės lyginant su sustiprinto sprendimų medžio ROC, PR, LIFT kreivėmis skolinimosi platformos „Bondora“ atvejis

4.4.5. Sprendimų džiunglės

Atlikus modelio parametrų derinimą, buvo nustatyta, kad naudojant sprendimų džiunglių klasifikatorių gaunamos tiksliausios prognozės naudojant šiuos parametrus:

- Optimizacijos žingsnių kiekis: 1771
- Didžiausiais plotis: 300
- Didžiausias gylis: 78
- Bendras elementų kiekis: 14

Modeliui naudotas *bagging* pernaudojimo metodas. Modelio tikslumas labai geras, AUC mato vertė lygi 0.851. Naudojant sprendimų džiunglių klasifikatorių teisingai prognozuota 77.9% reikšmių. Geras modelio tikslumas pastebimas ir iš ROC, PR, LIFT kreivių (26 pav.). Matome, kad kreivės nėra per daug nutolusios nuo geriausiai prognozavusio sustiprinto medžio algoritmo. Taipogi, buvo pastebėta, kad šis modelis reikalavo daugiausiai resursų ir užtruko ilgiausiai skaičiuojant rezultatus.



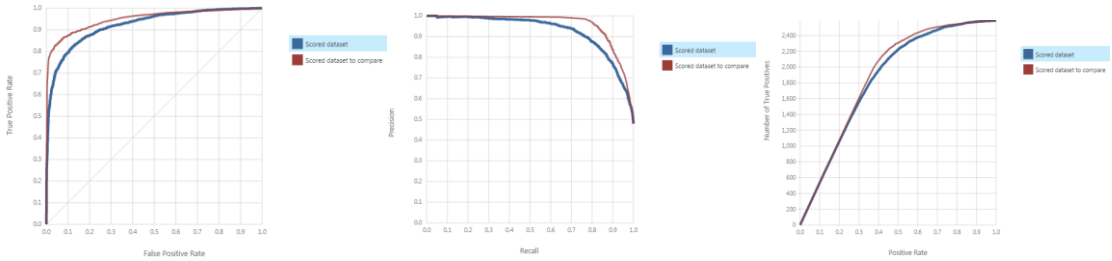
26 pav. Sprendimų džiunglių ROC, PR, LIFT kreivės lyginant su sustiprinto sprendimų medžio ROC, PR, LIFT kreivėmis skolinimosi platformos „Bondora“ atvejis

4.4.6. Sprendimų miškas

Atlikus metodo parametrų derinimo algoritmą, buvo nustatyta, kad modelis pasiekia geriausią tikslumą naudojant šiuos parametrus:

- Mažiausias pavyzdžių kiekis per lapo mazgą: 1
- Atsitiktinių padalijimų kiekis per mazgą: 141
- Didžiausias sprendimų medžio gylis: 16
- Medžių kiekis: 27

Atsižvelgiant į tikslumo matą AUC, kuris siekia 0.922 galime teikti jog klasifikavimo rezultatai yra puikūs. Naudojant šį klasifikatorių teisingai prognozuota 87,4% reikšmių. Sprendimų miško rezultatai buvo artimiausi sustiprintų sprendimų medžio rezultatams, kaip matome iš ROC, PR, LIFT kreivių (27 pav), abiejų metodų rezultatų kreivės yra itin artimos viena kitai, tačiau sprendimų miškas nedaug nusileidžia sustiprintų sprendimų medžių algoritmui.



27 pav. Sprendimų miško ROC, PR, LIFT kreivės lyginant su sustiprinto sprendimų medžio ROC, PR, LIFT kreivėmis skolinimosi platformos „Bondora“ atvejis

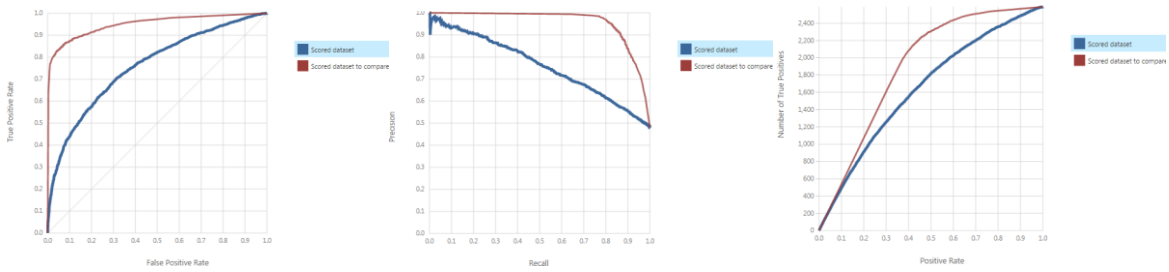
4.4.7. Vidurkinis perceptronas

Atlikus modelio parametrų derinimo algoritmą, buvo nustatyta, kad šis modelis geriausiai veikia naudodamas šiuos parametrus:

- Partijos dydis: 256
- Pradinis mokymosi greitis: 0.6334839
- Mokymosi greitis skylant eksponentei: 0.5
- Vidutinis svoris: 0.5
- Tolerancija: 1E-05
- Didžiausias iteracijų kiekis 8

Sprendžiant iš tikslumo mato AUC, kuris lygus 0.759 šio modelio tikslumas yra patenkinamas. Naudojant šį modelį buvo atspėta 69.2% imties reikšmių. Tai, kad modelio

tikslumas neprilygsta sustiprinto sprendimų medžių tikslumui galime matyti ir iš ROC, PR, LIFT kreivių (28 pav.).



28 pav. Vidurkinio perceptrono ROC, PR, LIFT kreivės lyginant su sustiprinto sprendimų medžio ROC, PR, LIFT kreivėmis skolinimosi platformos „Bondora“ atvejis

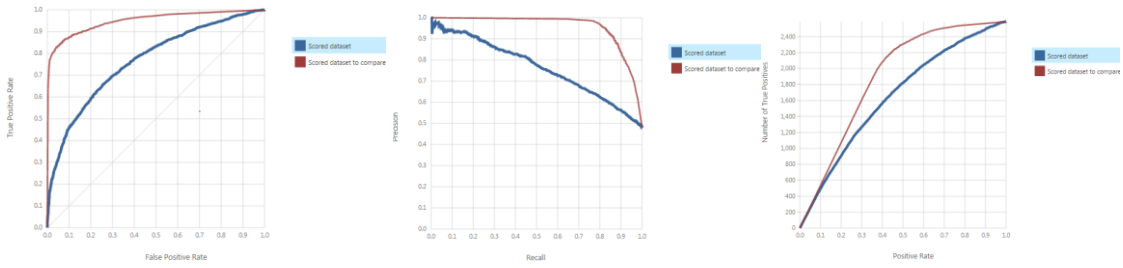
Sudarydami modelį matome, kad didžiausią teigiamą įtaką grąžinti paskolą turi šie kintamieji: prieš tai buvusių paskolų kiekis, pajamos gaunamos iš socialinio aprūpinimo, 2 vaikų turėjimas, prašyta pinigų suma, kilmės šalis Estija ir kiti kintamieji. Tuo tarpu didžiausią neigiamą įtaką daro šie kintamieji: Laisvų pinigų kiekis, paskolos terminas, palūkanų dydis. Amžius, įkaitinis namo turėjimo statusas, buvimas našle, tam tikrų pareigų turėjimas (kintamasis įvestas vietine kalba).

4.4.8. Bajeso taškas

Atlikus modelio parametrų derinimą Bajeso taško klasifikatoriui *(angl. Bayes Point machine)* buvo nustatyta, kad modelis gauna tiksliausius rezultatus naudojant šiuos parametrus:

- Atsitiktinė sėkla 2342
- Apmokymo iteracijų kiekis 30

Sprendžiant iš AUC rodiklio, kuris yra lygus 0.766 modelio tikslumas yra patenkinamas. Modelis teisingai prognozavo 70.3% reikšmių. Atsižvelgus į modelio ROC, PR, LIFT kreives pastebime, kad metodas nutolės nuo geriausiai veikiančio sustiprintų medžių metodo, tačiau veikia geriau nei atsitiktinis spėjimas(29 pav.).



29 pav. Bajeso taško ROC, PR, LIFT kreivės lyginant su sustiprinto sprendimų medžio ROC, PR, LIFT kreivėmis skolinimosi platformos „Bondora“ atvejis

Atlikus modelio analizę, pamatyta, kad modeliui didžiausią įtaką tam tikrų pareigybių turėjimas, kurių pavadinimai įvesti vietine kalba. Skirtingoms pareigoms priskirti skirtingi parametrai. Pastebėta, kad didžiausią teigiamą naudą grąžinti paskolą be pareigų turėjimo turi šie kintamieji: 2 vaikų turėjimas, gyventi Estijoje, patvirtinimo dėl pajamų ir išlaidų turėjimas, pajamų iš tiesioginio darbdavio turėjimas. Tuo tarpu didžiausią neigiamą naudą paskolos gražinimui be užimamų pareigų turėjo šie kintamieji: Būti iš Viru savivaldybės, gyventi su tėvais, gyventi jungtiniame nuomojame bute, dirbti mažiau nei 2 metus, būti nuomininku neapstatytame name, paskolos laikas, būti namų šeimininke.

Tarpusavio skolinimo platformos „Bondora“ duomenų klasifikavimo rezultatai buvo įvertinti naudojantis AUC ir išrikiuoti pagal tikslumą 3 lentelėje. Matome, kad puikūs rezultatai buvo gauti naudojantis sustiprinto sprendimo medžio ir sprendimo miško klasifikatoriumi. Geri klasifikavimo tikslumo rezultatai gauti naudojantis sprendimų džiunglių, neuroninių tinklų, giliųjų atraminių vektorių klasifikatoriais. Tuo tarpu patenkinami rezultatai gauti naudojantis Logistine regresija, Bajeseso klasifikatoriumi, atraminiais vektoriais ir vidurkinio perceptrono metodais. Pastarojo metodo prognozė buvo mažiausiai tiksli.

3 lentelė. Modeliai pagal tikslumą. Tarpusavio skolinimo platformos „Bondora“ atvejis.

Modelio pavadinimas	AUC	TP	TN	FP	FN	Tikslu-mas	Preciziš-kumas	Atkū-rimas	F1 Įvertis	Jautru-mas	Speci-fiškumas
Sustiprintas sprendimų medis	0,95	2249	2548	255	340	0,89	0,898	0,869	0,883	0,869	0,909
Sprendimų miškas	0,922	2079	2503	300	510	0,85	0,874	0,803	0,837	0,803	0,893
Sprendimų džiunglės	0,851	1851	2347	456	738	0,779	0,802	0,715	0,756	0,715	0,837
Neuroninis tinklas	0,826	2076	2061	742	513	0,767	0,737	0,802	0,768	0,802	0,735
Gilieji atraminiai vektoriai (DSVM)	0,825	1978	2138	665	611	0,763	0,748	0,764	0,756	0,764	0,763
Logistinė regresija	0,767	1700	2087	716	889	0,702	0,704	0,657	0,679	0,657	0,745
Bajaso taškas	0,766	1698	2090	713	891	0,703	0,704	0,656	0,679	0,656	0,746
Atraminiai vektoriai (SVM)	0,761	1685	2070	733	904	0,696	0,697	0,651	0,673	0,651	0,738

Vidurkinis perceptronas	0,759	1672	2057	746	917	0,692	0,691	0,646	0,668	0,646	0,734

Kadangi didžiausia rizika yra investuoti į paskolą, kuri nebus grąžinta dėl to svarbiausia atsižvelgti, kuris modelis turi mažiausiai reikšmių, kurioms buvo neteisingai priskirta teigiama reikšmė (paskola grąžinta). Šiuo atveju sustiprintų medžių metodas turi mažiausiai neteisingai prognozuotų teigiamų ir neteisingai prognozuotų neigiamų reikšmių. Tai rodo, kad pasirinkus sustiprintų medžių metodą bus mažiausia rizika investuoti į paskolą, kuri nebus grąžinta ir mažiausia rizika prarasti paskolą, kuri būtų grąžinama.

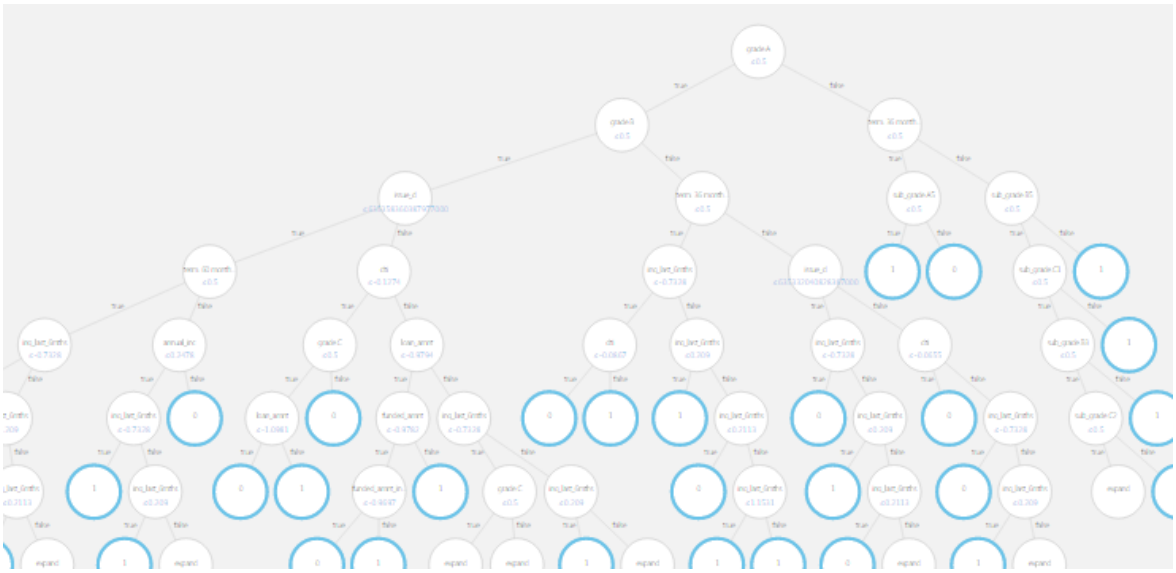
4.5.Modelių generavimas. Tarpusavio skolinimo platformos „Lending Club“ duomenų atvejis

Kaip ir tarpusavio skolinimo platformos „Bondora“, modelių generavimui naudojant „Lending Club“ duomenims klasifikuoti buvo išbandyti 8 skirtingi klasifikavimo metodai. Modeliams nustatytas tas pats kintamasis (statusas), pasirinktas tas pats slenkstis (0,5), derinant modelio parametrus ir pasirinkta 60 iteracijų geriausiems parametrams atrinkti.

4.5.1. Sustiprintas sprendimų medis

Naudojant modelio parametrų derinimo algoritmą buvo nustatyta, kad sustiprintų medžių metodas geriausia veikia naudojantis šiais parametrais:

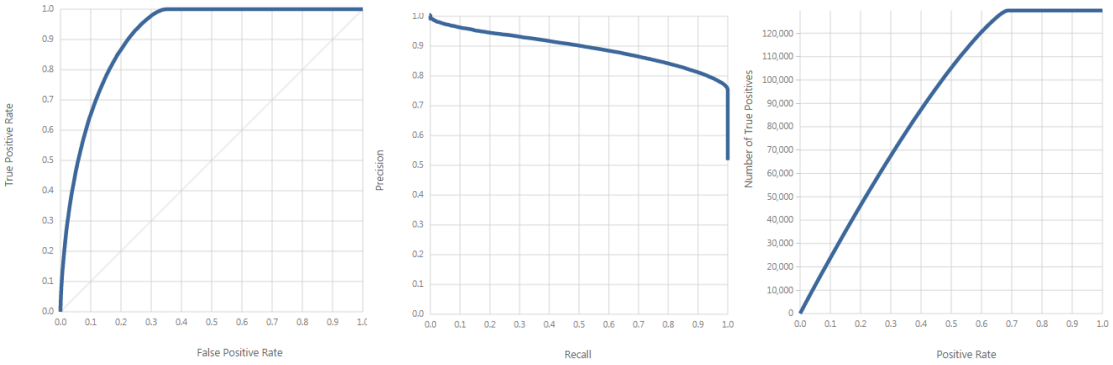
- Lapų kiekis: 71
- Mažiausias atskirų lapų atvejų kiekis: 47
- Mokymosi greitis: 0,091365
- Medžių kiekis: 407



30 pav. Sustiprintas sprendimų medis tarpusavio skolinimosi platformos „Lending Club“ atvejis

31 pav. atvaizduotas viršutinė sustiprinto sprendimų medžio dalis. Remiantis šiuo medžiu, kai kredito balas yra A ir paskola imama 36 mėnesiams, tuomet beveik visais atvejais paskola bus grąžinta. Tačiau kitos medžio šakos rodo, kad kredito balas nėra viską lemiantis, ir esant mažesniems kredito balams, tačiau turint kita ypatybes, taip pat užtikrinamas paskolos grąžinimas. Kaip, kas pastebėjome „Bondora“ atveju, labai didelę įtaką tikimybei grąžinti paskolą turi paskolos terminas ir paskolos dydis. Dažnais atvejais pakanka tik kredito balo, paskolos termino ir paskolos dydžio apibrėžti ar asmuo grąžins paskolą.

Kaip ir „Bondora“ atveju modelis buvo tiksliausias naudojant sustiprintų medžių metodą. Naudojant šį metodą buvo pasiektas 0.912 AUC. Modelis teisingai prognozavo 84,2% reikšmių. Modelio tikslumu galime įsitikinti ir stebėdami ROC, PR, LIFT kreives (32 pav.), kurios rodo labai gerą tikslumą.



31 pav. Sustiprinto sprendimų medžio ROC, PR, LIFT kreivės lyginant su sustiprinto sprendimų medžio ROC, PR, LIFT kreivėmis skolinimosi platformos „Lending Club“ atvejis

4.5.2. Logistinė regresija

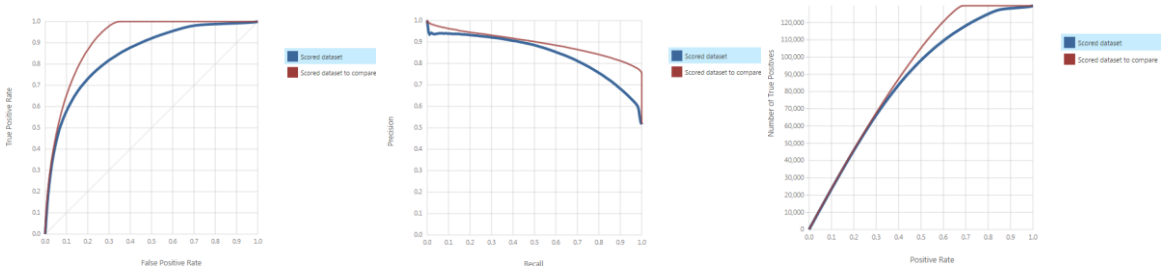
Buvo nustatyta, kad logistinės regresijos metodas geriausiai veikia naudojantis šiais parametrais:

- Optimizacijos tolerancija: 1.15527178E-06
- L1 svoris: 0.02400378
- L2 svoris: 0.0101877479
- Atminties dydis: 45

Buvo pastebėta, kad naudojantis logistinės regresijos metodu didžiausią neigiamą įtaką grąžinti paskolą daro šie kintamieji: įvairios profesijos, tokios kaip pataisos namų prižiūrėtojas, infrastruktūros operacijų vadybininkas, pirmosios kredito operacijos atlikimo data.. Tuo tarpu didžiausia teigiamą įtaką grąžinti paskolą turėjo šie kintamieji – metinės pajamos, pirmosios kredito operacijos atlikimo metai, buvimas teisėju, vyriausiuoju

patarėju, seržantu, dirbti tokiose įmonėse kaip „International Paper“, „Cognizant Technology Solutions“, „Capital One“ banke.

Naudojantis logistinės regresijos metodu buvo gautas geras tikslumas. Pasiektas 0.845 AUC, 76.5 % reikšmių buvo prognozuotos teisingai. Geru tikslumu galima įsitikinti ir iš ROC, PR, LIFT kreivių (33 pav.). Kreivės artimos geriausio modelio tikslumui.



32 pav. Logistinės regresijos ROC, PR, LIFT kreivės lyginant su sustiprinto sprendimų medžio ROC, PR, LIFT kreivėmis skolinimosi platformos „Lending Club“ atvejis

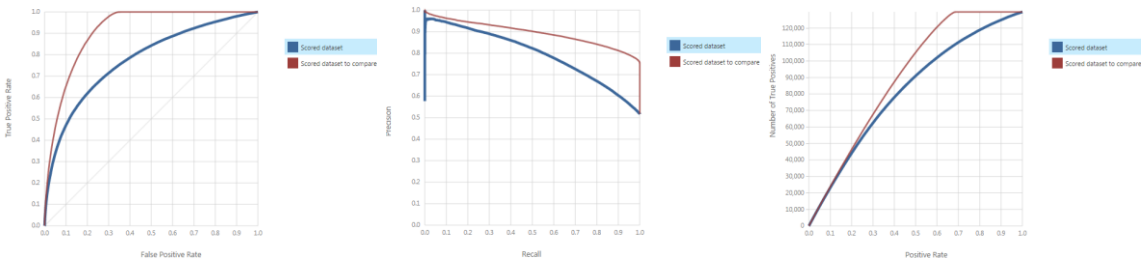
4.5.3. Vidurkinis perceptronas

Atlikus modelio parametrų derinimo algoritmą, buvo nustatyta, kad šis modelis geriausiai veikia naudodamas šiuos parametrus:

- Partijos dydis: 256
- Pradinis mokymosi greitis: 0.851337135
- Mokymosi greitis skylant eksponentei: 0.5
- Vidutinis svoris: 0.5
- Tolerancija: 1E-05
- Didžiausias iteracijų kiekis 9

Pagal vidurkinio perceptrono nustatytus ypatybių svorius matome, kad didžiausią įtaką grąžinti paskolą laiku turi šie kintamieji: metinės pajamos, viešų neigiamų įrašų kiekis, dirbti vyriausiuoju programuotoju, nuovados šerifu, vyriausiuoju mokesčių vadybininku, reaktoriaus operatoriumi, seržantu, projektų analitiku. Tuo tarpu didžiausią įtaką negrąžinti paskolos turėjo šie kintamieji: paskolos dydis, turimų kredito kortelių kiekis, dirbti namų šeimininke, langų dėjimo darbuotoju, terapeutu, turėti darbovietę tokią kaip „Walmart“, „CMC Capital Steel“, „Prometric“.

Naudojantis vidurkinio perceptrono metodu buvo pasiektas patenkinamas tikslumas. AUC siekia 0.777 tuo tarpu teisingai prognozuotos 70,7% reikšmių. Patenkinamas tikslumas matomas ir iš ROC, PR, LIFT kreivių (34 pav.), kur matome, kad modelis nors ir veikia geriau nei atsitiktinis spėjimas, tačiau yra per daug nutolęs nuo geriausio modelio.



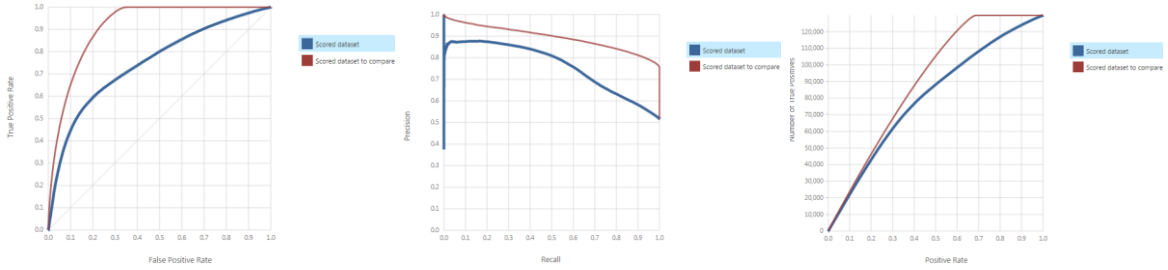
33 pav. Vidurkinio perceptrono ROC, PR, LIFT kreivės lyginant su sustiprinto sprendimų medžio ROC, PR, LIFT kreivėmis skolinimosi platformos „Lending Club“ atvejis

4.5.4. Atraminiai vektoriai

- Iteracijų kiekis: 42
- Lambda: 1.215461E-05

Stebint modelio ypatybių svorius buvo pastebėta, kad didžiausią įtaką naudojantis šiuo modeliu grąžinti paskolą daro šie kintamieji – metinės pajamos, įžeidžiamų viešųjų įrašų kiekis, dirbti federaliniu agentu, specialiuoju agentu, seržantu, finansų analitiku, turimų kredito kortelių kiekis, palūkanų norma 6%. Tuo tarpu didžiausią įtaką negrąžinti paskolos turi šie kintamieji – paskolos kiekis, paskolos išdavimo data, dirbti įmonėse „G4S Secure Solutions“, „Walmart“, dirbti transporto prižiūrėtoju, turėti palūkanų normą 12,92%, dirbti sertifikuota sesele.

Modelio tikslumas patenkinamas. AUC siekia 0.749. Modelis teisingai prognozuoja 68.1% reikšmių. Iš ROC, PR, LIFT kreivių (35 pav.) taip pat matoma, kad tikslumas nors ir veikia geriau nei atsitiktinis spėjimas, tačiau gerokai nusileidžia geriausiam metodui.



34 pav. Atraminių vektorių ROC, PR, LIFT kreivės lyginant su sustiprinto sprendimų medžio ROC, PR, LIFT kreivėmis skolinimosi platformos „Lending Club“ atvejis

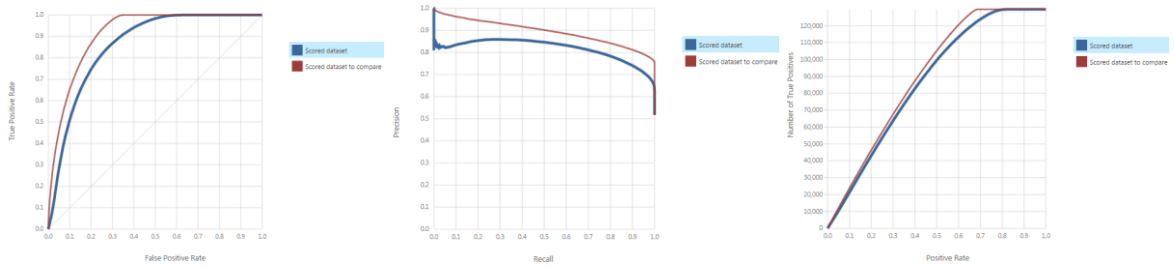
4.5.5. Gilieji atraminiai vektoriai

Naudojant parametrų derinimo algoritmą buvo nustatyta, kad modelis geriausiai veikia naudojantis šiais parametrais:

- Medžių gylis: 6
- LambdaW: 0.00144066708

- Lambda Theta: 0.00245464034
- Lambda Theta Prime: 0.0308579933
- Sigma: 0.343650043
- Iteracijų kiekis: 13126

Modelis turi gerą tikslumą. AUC siekia 0.859, teisingai prognozuojama 78.7% reikšmių. Modelis itin artimas geriausiam (sustiprintų medžių) modeliui. Tai galima matyti ir iš ROC, PR, LIFT kreivių (36 pav.). Tačiau šis modelis kur kas geriau prognozuoja neigiamas reikšmes, nei teigiamas.

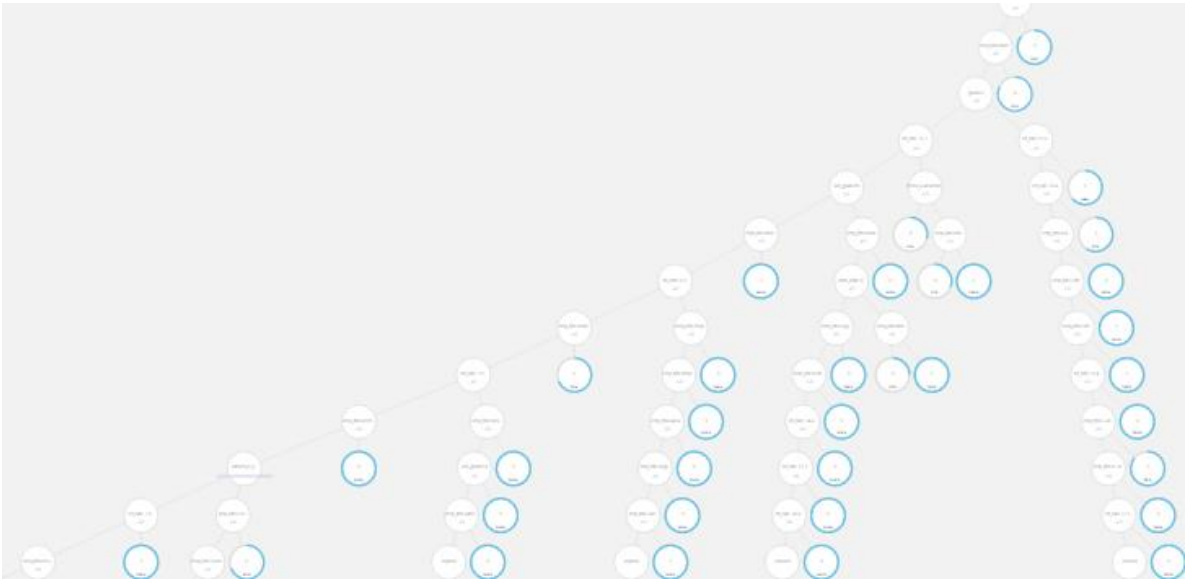


35 pav. Giliųjų atraminių vektorių ROC, PR, LIFT kreivės lyginant su sustiprinto sprendimų medžio ROC, PR, LIFT kreivėmis skolinimosi platformos „Lending Club“ atvejis

4.5.6. Sprendimų miškas

Atlikus metodo parametrų derinimo algoritmą, buvo nustatyta, kad modelis pasiekia geriausią tikslumą naudojant šiuos parametrus:

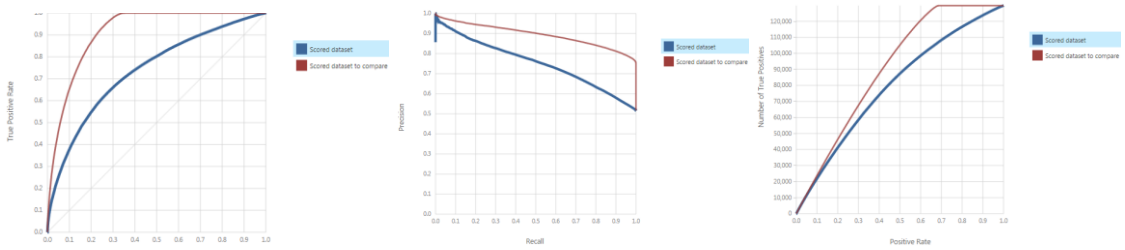
- Mažiausias pavyzdžių kiekis per lapo mazgą: 1
- Atsitiktinių padalijimų kiekis per mazgą: 421
- Didžiausias sprendimų medžio gylis: 41
- Medžių kiekis: 14



36 pav. Sprendimų miško viršutinė dalis skolinimosi platformos „Lending Club“ atveju

Peržiūrėjus mišką geriausią sprendimų miško grafą 37 pav., buvo pastebėta, kad modeliui užtenka vos kelių verčių žinoti atsakymą ar asmuo grąžins paskolą ar ne. Asmuo dirbantis valytoju arba remonto darbuotoju iškarto buvo priskiriamas asmenims, kurie negrąžina paskolų. Tolesni duomenys skirstomi į asmenis turinčius D kredito balą ir jo neturinčius. Buvo pastebėta, kad nors žmonės ir turi prastą kredito balą, tačiau nemaža dalis jų vis tiek galutiniame rezultate buvo nustatyti, kaip grąžinantys paskolą. Jei asmuo turi žemą kredito balą (D) tačiau, jis dirba programų palaikytoju arba vartotojų atstovu, arba jo palūkanų norma yra 15,9 remiantis sprendimų mišku jis vis tiek grąžins paskolą. Nors modelio rezultatai patenkinami, tačiau tokiu medžiu yra sunku pasitikėti, nes jo rezultatus lemia per mažai kintamųjų, tai rodo, kad modelis gali būti permokytas pateiktiems duomenims arba 60 iteracijų modelio parametrų derinimui buvo per mažai.

Dėl modelio netinkamumo įsitikinama ir žiūrint į modelio vertinimo rezultatus. Modelis turi patenkinamą tikslumą, nors įprastai miško algoritmas pasižymi itin geru tikslumu. AUC siekia 0.736, modelis teisingai atspėja 66.9% reikšmių. Modelio silpnumas matomas ir iš AUC, PR, LIFT kreivių (38 pav.). Modelis neprilygsta geriausiam sustiprintų medžių modeliui ir nors ir atsiplėšia nuo atsitiktinio spėjimo.



37 pav. Sprendimų miško ROC, PR, LIFT kreivės lyginant su sustiprinto sprendimų medžio ROC, PR, LIFT kreivėmis skolinimosi platformos „Lending Club“ atvejis

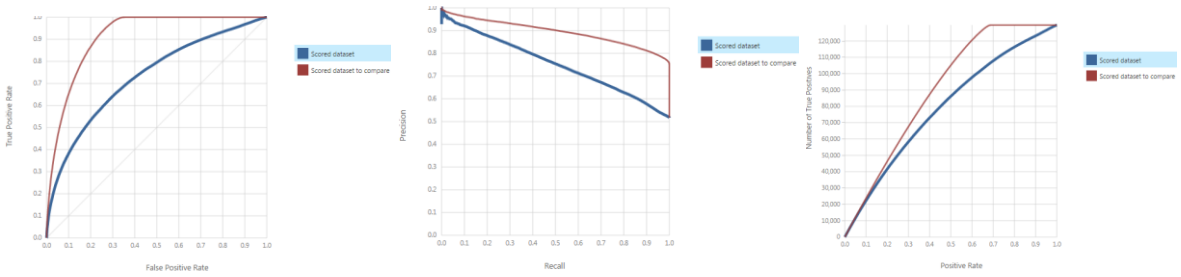
4.5.7. Sprendimų džiunglės

Atlikus modelio parametrų derinimą, buvo nustatyta, kad naudojant sprendimų džiunglių klasifikatorių gaunamos tiksliausios prognozės naudojant šiuos parametrus:

- Optimizacijos žingsnių kiekis: 14875
- Didžiausiais plotis: 232
- Didžiausias gylis: 81
- Bendras elementų kiekis: 31

Modeliui naudojamas *Bagging* pernaudojimo metodas. Modelio tikslumas patenkinamas, AUC tikslumo matas lygus 0.73. Modelis teisingai prognozuoja 66.1%

reikšmių. Pagal ROC, PR, LIFT(38 pav.) kreives taip pat matome, kad modelis labai atsilieka nuo geriausio sustiprinto medžio modelio, tačiau prognozuoja geriau nei atsitiktinis spėjimas.



38 pav. Sprendimų džiunglių ROC, PR, LIFT kreivės lyginant su sustiprinto sprendimų medžio ROC, PR, LIFT kreivėmis skolinimosi platformos „Lending Club“ atvejis

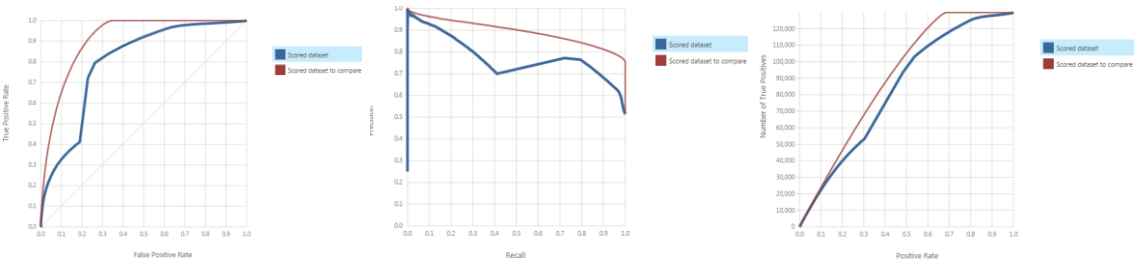
4.5.8. Bajeso taškas

Atlikus modelio parametrų derinimą Bajeso taško klasifikatoriui buvo nustatyta, kad modelis gauna tiksliausius rezultatus naudojant šiuos parametrus:

- Atsitiktinė sėkla 2342
- Apmokymo iteracijų kiekis 30

Pagal nustatytus modelio ypatybių svorius buvo nustatyta, kad didžiausi tiek neigiami, tiek teigiami svoriai buvo priskirti pareigybėms arba darbovietėms. Kiti kintamieji turėjo žymiai mažesnius svorius.

Modelio tikslumas yra patenkinamas. AUC tikslumo matas siekia 0.797. Modelis taisyklingai prognozuoja 74.6% reikšmių. Modelis nemažą dalį teigiamų kintamųjų prognozuoja neigiamai, tačiau šis modelis turi mažiausią klaidingai teigiamai prognozuotų reikšmių vertę lyginant su kitais modeliais. Sprendžiant iš modelio AUC, PR, LIFT kreivių 39 pav. matoma, kad modelis yra silpnesnis už geriausiai prognozuojantį sustiprinto medžio modelį, taip pat jų tikslumas labai svyruojantis.



39 pav. Bajeso taško ROC, PR, LIFT kreivės lyginant su sustiprinto sprendimų medžio ROC, PR, LIFT kreivėmis skolinimosi platformos „Lending Club“ atvejis

4 lentelėje rezultatai išrikiuoti pagal tikslumo matą AUC. Matome, kad geriausius modelio rezulatatus gauname naudodamiesi sustiprintu sprendimų medžio metodu. Geri

rezultatai gaunami naudojant giliuosius atraminius vektorius ir logistinę regresiją. Patenkinami rezultatai gaunami naudojantis Bajeso tašku, vidurkiniu perceptronu, atraminiais vektoriais, sprendimų mišku ir sprendimų džiunglėmis.

4 lentelė. Modeliai pagal tikslumą. Tarpusavio skolinimo platformos „Lending Club“ atvejis.

Modelio pavadinimas	AUC	TP	TN	FP	FN	Tikslumas	Preciziškumas	Atkūrimas	F1 įvertis	Jautrumas	Specifiškumas
Sustiprintas sprendimų medis	0,912	117775	93290	27734	11978	0,842	0,809	0,908	0,856	0,908	0,771
Gilieji atraminiai vektoriai (DSVM)	0,859	11270	84593	36431	17048	0,787	0,756	0,869	0,808	0,869	0,699
Logistinė regresija	0,845	99785	92058	28966	29968	0,765	0,775	0,769	0,772	0,769	0,761
Bajeso taškas	0,797	93657	93341	27683	36096	0,746	0,772	0,722	0,746	0,722	0,771
Vidurkinis perceptronas	0,777	92187	85234	35790	37566	0,707	0,72	0,71	0,715	0,710	0,704
Atraminiai vektoriai (SVM)	0,749	90438	80392	40632	39315	0,681	0,69	0,697	0,693	0,697	0,664
Sprendimų miškas	0,736	98194	69638	51386	31559	0,669	0,656	0,757	0,703	0,757	0,575
Sprendimų džiunglės	0,73	98371	67470	53554	31382	0,661	0,647	0,758	0,698	0,758	0,557

Didžiausia rizika yra patiriama investuojant į tas paskolas, kurios yra negrąžinamos, todėl labai svarbu atsižvelgti, kad naudojamas metodas turėtų kiek įmanoma mažiau neteisingai prognozuojamų teigiamų reikšmių. Matome, kad šiuo atžvilgiu, mažiausiai kartų buvo suklysta neteisingai priskyrus teigiamą reikšmę naudojantis Bajeso taško klasifikatoriumi. Tačiau skirtumas, lyginant su sustiprintų medžių metodu, yra labai nežymus, o žinant, kad sustiprintas medžių metodas kur kas geriau prognozuoja ir neigiamas reikšmes, naudingiau būtų rinktis sustiprinto medžio klasifikatorių.

4.5.9. Rekomendacijos

Kadangi didžiausias palūkanas galima gauti investuojant į „Bondora“ skolinimosi platformoje esančias paskolas ir naudojantis sustiprintų medžių metodu galima labai tiksliai prognozuoti, ar paskola bus grąžinta ar ne, rekomenduoti investuoti būtent į šią platformą. Remiantis geografine vietove, patartina iš anksto nusistatyti investavimą į Estijos.

Siekiant mažiau pelningų, tačiau užtikrintų rezultatų, būtų galima didinti slenksčio dydį, užtikrinant, kad tikimybė, kad asmuo grąžins paskolą yra pakankamai didelė, tokiu būdu prarandant rizikingas paskolas, tačiau užtikrinant sėkmingas investicijas. Rizikuojant, tačiau siekiant gauti didesnes pajamas, rekomenduotina naudojant sustiprintų medžių metodą rinktis tik tas paskolas, kurių palūkanos yra didelės.

Siekiant užtikrinti didesnį patikimumą vartotojams, kurie naudojasi tarpusavio skolinimosi platformomis, platformų kūrėjai turėtų naudoti skaitmeninio mokymo metodus savo svetainėse, taip siekdami, kad jų svetainėje nebūtų suteikiama paskolų, kurios bus negrąžintos.

Siekiant tikslesnių modelių rezultatų būtų galima padidinti iteracijų kiekį parametrų derinimo algoritme. Taipogi detektoriui būtų galima parinkti įvairų veikimo tašką *(angl. operating point),* randant optimalų slenkstį išėjimų reikšmėms atskirti į klases. Siekiant parinkti optimalų tašką galima naudoti įvairius kriterijus, pavyzdžiui R paketas OptimalCutpoints siūlo bent 30 skirtingų kriterijų optimaliam taškui rasti. Siekiant rasti EER veikimo tašką galimą naudotis SpEqualSe paketu. Kredito rizikos vertinimo uždaviniui labiausiai rekomenduojama EMP *(angl. expected maximum profit)* kriterijus.

5. IŠVADOS

Pastebėta, kad vis daugiau vartotojų, norėdami įsigyti paskolą, renkasi tarpusavio skolinimosi platformas vietoje finansinių institucijų. nustatyta, kad Amerikoje populiariausia tarpusavio skolinimosi platforma yra „Lending Club“, Europoje „Bondora“, Jungtinėje karalystėje „Zopa“, pasaulyje „Bitbond“.

Įvertinta, kad paskolų sektoriuje skaitmeniniai algoritmai apytiksliai pradėti tirti prieš 15 metų. Dažniausiai ir populiariausiai naudojami sprendimų medžių, atsitiktinių miškų, neuroninių tinklų algoritmai.

Tyrimui pasirinkti dviejų skirtingų skolinimo platformų duomenys („Lending Club“ ir „Bondora“). Atlikus aprašomąją duomenų analizę buvo nustatyta, kad „Lending Club“ skolinimosi platforma yra išdavusi daugiau nei dešimt kartų daugiau paskolų, nei „Bondora“. Taip pat pastebėta, kad sėkmingi „Bondora“ klientai gauna dvigubai didesnes palūkanas, tačiau skolintojai kur kas rečiau grąžina skolas.

Duomenys buvo klasifikuoti naudojantis 8 skirtingais metodais (logistine regresija, vidurkiniu perceptronu, atraminiais vektoriais, giliaisiais atraminiais vektoriais, sustiprintais sprendimų medžiais, sprendimų mišku, sprendimų džiunglėmis, Bajeso tašku) siekiant išsiaiškinti ar asmuo grąžins paskolą ar ne. Buvo nustatyta, kad geriausią tikslumą su abejais duomenų rinkiniais pavyko gauti naudojantis sustiprintu medžių metodu („Bondora“ atveju pasiektas 0.95 AUC, 0.89 tikslumas, 0.898 preciziškumas, 0.869 atkūrimas, 0.883 F1 įvertis, „Lending Club“ atveju 0.912 AUC, 0.842 tikslumas, 0.809 preciziškumas, 0.908 atkūrimas, 0.856 F1 įvertis).

Buvo nustatyta, kad rekomenduotina investuoti į „Bondora“ skolinimosi platformoje esančias paskolas dėl platformoje pateikiamų aukštų palūkanų, Pritaikius sustiprintų medžių algoritmą, atmetus klientus, kuriems prognozuojamas paskolos negrąžinimo statusas bus išvengta rizikingų paskolų ir taip užtikrinama didžiausia investicinė grąža.

LITERATŪRA:

[1] LICHTENWALD, Ryan. The History of Peer to Peer Lending [interaktyvus] 2014 Prieiga per http://peersociallending.com/news/history-peer-peer-lending/

[2] CARSON, John. Peer to Peer Lending Sites → 24 of the World's Best [interaktyvus]. 2017. Prieiga per http://peersociallending.com/investing/peer-to-peer-lending-sites-16-of-the-worlds-best/

[3] WILHITE, Tamara. Why People Use Peer to Peer Lending [interaktyvus]. 2016. Prieiga per https://hubpages.com/money/Why-people-use-peer-to-peer-lending

[4] PICHOUX, Le Loic. TOP 10 Everything about Peer to Peer Lending [interaktyvus]. 2017. Prieiga per https://www.klearlending.com/en/Blog/Articles/p2p-investing

[5] INVESTMENT HACKING. Bondora high yield peer to peer lending reviewed [interaktyvus]. 2014. Prieiga per http://investmenthacking.com/p2p-lending/bondora-high-yield-peer-to-peer-lending-reviewed/

[6] FAULKNER, NEIL. Bondora Makes it Easy to Compare Borrowers Across Europe [interaktyvus]. 2014. Prieiga per https://www.4thway.co.uk/news/bondora-makes-it-easy-to-compare-borrowers-across-europe/

[7] BONDORA viešasis tinklalapis [interaktyvus]. Prieiga per https://www.bondora.com

[8] LENDING CLUB viešasis tinklapis [interaktyvus]. Prieiga per https://www.lendingclub.com/

[9] MARTUCCI, Brian. Lending Club Review – Peer-to-Peer (P2P) Financial Lending [interaktyvus]. Prieiga per http://www.moneycrashers.com/lending-club-review-peer-to-peer-lending/

[10] BANKŲ PRIEŽIŪROS BASEL KOMITETAS. Principles for the Management of Credit Risk [interaktyvus]. 2000. Prieiga per http://www.bis.org/publ/bcbs75.pdf

[11] CRISIL GLOBAL RESEARCH & ANALYTICS. Credit Risk Estimation Techniques [interaktyvus]. Prieiga per https://www.crisil.com/Crisil/pdf/global-offshoring/Credit_Risk_Estimation_Techniques.pdf

[12] BAESENS, Bart. Developing Intelligent Systems for Credit Scoring Using Machine Learning Techniques [interaktyvus]. 2003. Priega per http://www.dataminingapps.com/wp-content/uploads/2015/04/Phd-Bart-Baesens.pdf

[13] CHWEE, Peng Goh. Credit Scoring Using Data Mining Techniques. 2004

[14] SZCZERBA, Monika, CIEMSKI, Andrzej. Credit Risk Handling in Telecomunication Sector. 2009

[15] BROWN, Iain, MUES ,Christoph. An experimental comparison of classification algorithms for imbalanced credit scoring data sets.2012

[16] ALARAJ, Maher, ABBOD, Maysam F. A new hybrid ensemble credit scoring model based on classifiers consensus system approach. 2016

[17] HAMID, Aboobyda Jafar, AHMED Tarig Mohammed. Developing Prediction Model Of Loan Risk In Banks Using Data Mining [interaktyvus]. 2016. Prieiga per http://aircconline.com/mlaij/V3N1/3116mlaij01.pdf

[18] ABDI, Herve. Normalizing Data [interaktyvus]. 2010. Prieiga per https://www.utdallas.edu/~herve/abdi-Normalizing2010-pretty.pdf

[19] MICHIE, D., SPIEGELHALTER, D.J., TAYLOR, C.C. Machine Learning, Neural and Statistical Classification [interaktyvus]. 1994. Prieiga per https://www1.maths.leeds.ac.uk/~charles/statlog/whole.pdf

[20] THE INTERNATIONAL SCHOOL OF QUANTITATIVE RESEARCH. Logistic [interaktyvus]. Prieiga per https://www.isqr.uni-freiburg.de/logistic.pdf

[21] COLLINS, Michael. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms [interaktyvus] 2002. Prieiga per http://www.aclweb.org/anthology/W02-1001

[22] NGUYEN, Nam, GUO, Yunsong. Comparisons of Sequence Labeling Algorithms and Extensions [interaktyvus]. Prieiga per http://www.machinelearning.org/proceedings/icml2007/papers/206.pdf

[23] CHAO, Feng-Chih, HORNG, Ming – Huwi. The Construction of Support Vector Machine Classifier Using the Firefly Algorithm [interaktyvus].2014. Prieiga per https://www.hindawi.com/journals/cin/2015/212719/

[24] JOSE, Cijo, GOYAL, Prasoon, AGGRWAL, Parv, VARMA, Manik. Local Deep Kernel Learning for Efficient Non-linear SVM Prediction [interaktyvus]. Prieiga per http://manikvarma.org/pubs/jose13.pdf

[25] FRIEDMAN, Jerome H. Greedy Function Approximation: A Gradient Boosting Machine [interaktyvus]. 1999. Prieiga per https://statweb.stanford.edu/~jhf/ftp/trebst.pdf

[26] INTRODUCTION TO BOOSTED TREES [interaktyvus]. Prieiga per http://xgboost.readthedocs.io/en/latest/model.html

[27] BREIMAN, Leo. Random Forests [interaktyvus]. 2001. Prieiga per https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

[28] WALKER, Michael. Random Forests Algorithm [interaktyvus]. 2013. Prieiga per http://www.datasciencecentral.com/profiles/blogs/random-forests-algorithm

[29] SHOTTON, Jamie, SHARP, Toby, KOHLI, Pushmeet, NOWOZIN, Sebastian, WINN, John, CRIMINISI, Antonio. Decision Jungles: Compac and Rich Models for Classification [interaktyvus]. Prieiga per https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/DecisionJunglesNIPS2013Supplementary.pdf

[30] POHLEN, Tobias. Decision Jungles [interaktyvus]. 2014. Prieiga per http://geekstack.net/resources/public/downloads/tobias_pohlen_decision_jungles.pdf

[31] HERBRICH, Ralf, GRAEPEL, Thore, CMPBELL, Colun. Bayes Point Machines [interaktyvus]. Prieiga per http://www.jmlr.org/papers/volume1/herbrich01a/herbrich01a.pdf

[32] REBY, David, LEK Sovan, DIMOPOULOS, Ioannis, JOACHIM, Jean, LAUGA, Jacques, AULAGNIER, Stephane. Artificial neural networks as a classification method in the behavioural sciences [interaktyvus]. 1996. Priega per http://www.lifesci.sussex.ac.uk/cmvcr/Publications_files/rebyproc.pdf

[33] THOMAS, G., TAPE, MD. Interpreting Diagnostic Tests [interaktyvus] Prieiga per http://gim.unmc.edu/dxtests/Default.htm

[34] HOSSIN, M., SULAIMAN, M.N. A Review On Evaluation Metrics For Data Classification Evaluations [interaktyvus]. 2015. Prieiga per http://aircconline.com/ijdkp/V5N2/5215ijdkp01.pdf

[35] SAITO, Takaya. Introduction to the precision-recall plot [interaktyvus]. Prieiga per https://classeval.wordpress.com/introduction/introduction-to-the-precision-recall-plot/

[36] VUK, Miha, CURK, Tomaž, ROC Curve, Lift Chart and Calibration Plot [interaktyvus]. 2006. Priega per http://www.stat-d.si/mz/mz3.1/vuk.pdf

[37] BROOKS, Steve. How close can Alteryx and Microsoft get? [interaktyvus]. 2016. Prieiga per https://www.enterprisetimes.co.uk/2016/03/22/close-can-alteryx-microsoft-get/

[38] INTRODUCTION TO AZURE MACHINE LEARNING IN THE CLOUD [interaktyvus]. 2017. Prieiga per https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-what-is-machine-learning

[39] BERGSTRA James, BENGIO Youshua. Random Search for Hyper-Parameter Optimization [interaktyvus]. 2012. Prieiga per: http://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf

[40] CHAWLA, Nitesh V., BOWYER, Kevin W., HALL, Lawrence O. KEGELMEYER, W. Philip. SMOTE: Synthetic Minority Over-sampling Technique [interaktyvus]. Prieiga per: https://www.jair.org/media/953/live-953-2037-jair.pdf