Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Tarpusavio skolinimo platformos fizinių asmenų kreditingumo rizikos vertinimas

Baigiamasis magistro studijų projektas

Justina Laškovaitė-Kolinienė

Projekto autorė

Doc. dr. Mindaugas Kavaliauskas

Vadovas

Doc. dr. Lina Sinevičienė

Vadovė

Kaunas, 2025

Kauno technologijos universitetas
Matematikos ir gamtos mokslų fakultetas

Tarpusavio skolinimo platformos fizinių asmenų kreditingumo rizikos vertinimas

Baigiamasis magistro studijų projektas
Didžiųjų verslo duomenų analitika (6213AX001)

Justina Laškovaitė-Kolinienė
Projekto autorė

Doc. dr. Mindaugas Kavaliauskas
Vadovas

Doc. dr. Lina Sinevičienė
Vadovė

Prof. dr. Evaldas Vaičiukynas
Recenzentas

Prof. dr. Aušrinė Lakštutienė
Recenzentė

Kaunas, 2025

Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Justina Laškovaitė-Kolinienė

Tarpusavio skolinimo platformos fizinių asmenų kreditingumo rizikos vertinimas

Akademinio sąžiningumo deklaracija

Patvirtinu, kad:

1. baigiamąjį projektą parengiau savarankiškai ir sąžiningai, nepažeisdama(s) kitų asmenų autoriaus ar kitų teisių, laikydamasi(s) Lietuvos Respublikos autorių teisių ir gretutinių teisių įstatymo nuostatų, Kauno technologijos universiteto (toliau – Universitetas) intelektinės nuosavybės valdymo ir perdavimo nuostatų bei Universiteto akademinės etikos kodekse nustatytų etikos reikalavimų;

2. baigiamajame projekte visi pateikti duomenys ir tyrimų rezultatai yra teisingi ir gauti teisėtai, nei viena šio projekto dalis nėra plagijuota nuo jokių spausdintinių ar elektroninių šaltinių, visos baigiamojo projekto tekste pateiktos citatos ir nuorodos yra nurodytos literatūros sąraše;

3. įstatymų nenumatytų piniginių sumų už baigiamąjį projektą ar jo dalis niekam nesu mokėjęs (-usi);

4. suprantu, kad išaiškėjus nesąžiningumo ar kitų asmenų teisių pažeidimo faktui, man bus taikomos akademinės nuobaudos pagal Universitete galiojančią tvarką ir būsiu pašalinta(s) iš Universiteto, o baigiamasis projektas gali būti pateiktas Akademinės etikos ir procedūrų kontrolieriaus tarnybai nagrinėjant galimą akademinės etikos pažeidimą.

Justina Laškovaitė-Kolinienė

Patvirtinta elektroniniu būdu

Laškovaitė-Kolinienė, Justina. Tarpusavio skolinimo platformos fizinių asmenų kreditingumo rizikos vertinimas. Magistro studijų baigiamasis projektas / vadovas doc. dr. Mindaugas Kavaliauskas, vadovė doc. dr. Lina Sinevičienė; Kauno technologijos universitetas, Matematikos ir gamtos mokslų fakultetas.

Studijų kryptis ir sritis (studijų krypčių grupė): Taikomoji matematika (Matematikos mokslai).

Reikšminiai žodžiai: kreditingumo rizikos vertinimas, tarpusavio skolinimas, mašininis mokymasis, klasifikavimas.

Kaunas, 2025. 58 p.

Santrauka

Lietuvoje vis augant fizinių asmenų paskolų portfeliams, auga ir tarpusavio skolinimo platformų populiarumas. Tiek besiskolinančiųjų tarpe, tiek siekiančių paskolinti tarpe. Kadangi tarpusavio skolinimo platformose gali investuoti ir neprofesionalūs investuotojai, labai svarbu, kad būtų teisingai nustatoma skolininko kredito rizika, kuri tiesiogiai susijusi su investuotojų nuostoliais.

Šiame darbe nagrinėjamas kreditingumo rizikos vertinimas tarpusavio skolinimo platformose. Pirmoje darbo dalyje atlikta mokslinės literatūros analizė, kurioje aptartos kreditingumo rizikos vertinimo metodikos, galimi duomenų šaltiniai bei dažniausiai tyrimams naudojami duomenų rinkiniai per kintamųjų prizmę. Antroje darbo dalyje pristatyta tyrimo metodologija – aptarti pasirinkti mašininio mokymosi algoritmai bei jų veikimo principai. Atliktos literatūros analizės pagrindu pasirinkti penki mašininio mokymosi algoritmai: logistinės regresijos, sprendimų medžio, atsitiktinio miško, XGBoost bei LightGBM metodai. Tyrimo rezultatų dalyje analizuojamam vienos, Lietuvoje veikiančios, tarpusavio skolinimo platformos paskolų duomenų rinkiniui pritaikyti visi penki klasifikavimo metodai. Metodų efektyvumas vertintas pagal AUC metriką. Rezultatai parodė, jog geriausiai mokius ir nemokius skolininkus klasifikuojantis metodas buvo XGBoost algoritmas, kurio AUC siekė 0,812. Prastesnius rezultatus parodė LightGBM, logistinės regresijos, sprendimų medžio ir atsitiktinio miško metodai. Taip pat, gavus rezultatus identifikuoti reikšmingiausią įtaką darantys kintamieji bei atliktas rezultatų palyginimas su analizuotoje literatūroje pasiektais rezultatais.

Laškovaitė-Kolinienė, Justina. Creditworthiness Risk Assessment of Individual Borrowers on Peer-to-Peer Lending Platforms. Master's Final Degree Project / supervisor doc. dr. Mindaugas Kavaliauskas, supervisor doc. dr. Lina Sinevičienė; Faculty of Mathematics and Natural Sciences, Kaunas University of Technology.

Study field and area (study field group): Applied Mathematics (Mathematical Sciences).

Keywords: creditworthiness risk assessment, peer-to-peer lending, machine learning, classification. Kaunas, 2025. 58 p.

Summary

As loan portfolios of individuals in Lithuania continue to grow, the popularity of peer-to-peer lending platforms is growing as well. Both among borrowers and those seeking to borrow. As peer-to-peer lending platforms are also open to investment by retail investors, it is essential that the credit risk of the borrower, which is directly linked to investor losses, is correctly identified.

This paper focuses on the assessment of creditworthiness risk on peer-to-peer lending platforms. The first part of the paper analyses the academic literature, discussing methodologies for credit risk assessment, available data sources and commonly used datasets for research through the prism of variables. The second part of the paper presents the research methodology by discussing the selected machine learning algorithms and their operational principles. Five machine learning algorithms were selected on the basis of the literature analysis: logistic regression, decision tree, random forest, XGBoost and LightGBM methods. In the results part of the study, all five classification methods were applied to the loan dataset of a peer-to-peer lending platform operating in Lithuania. The performance of the methods was evaluated using the AUC metric. The results showed that the best method for classifying both solvent and insolvent borrowers was the XGBoost algorithm with an AUC of 0.812. The LightGBM, logistic regression, decision tree and random forest methods performed worse. The results were also used to identify the most significant influencing variables and to compare the results with those obtained in the literature.

Turinys

Lentelių sąrašas

Paveikslų sąrašas

Santrumpų ir terminų sąrašas

Santrumpos:

BVP – bendrasis vidaus produktas;

BDAR – bendrasis duomenų apsaugos reglamentas;

DTI – įsipareigojimų ir pajamų santykis (angl. *debt-to-income ratio*);

PVI – pirkimo vadybininkų indeksai (angl. *Purchasing Managers Index);*

Fintech – technologijomis pagrįstos finansinės inovacijos (angl. *FinTech*);

ECB – Europos Centrinis bankas.

Įvadas

Paskolos yra daugelio lietuvių neatsiejama kasdienybės dalis. Žmonės skolinasi beveik viskam, pradedant mažais pirkiniais, kelionėmis, paskolomis studijų kainai padengti, baigiant didžiausiais pirkiniais – būstais ar prabangiais automobiliais.

Kreditų davėjai nuolat stengiasi prisitaikyti prie besikeičiančių rinkos sąlygų ir siekia atliepti klientų poreikį. Todėl rinkoje yra įvairių pasiskolinimo galimybių: pirkimas išsimokėtinai, paskola, paskola su įmokų atidėjimu, paskola įkeičiant nekilnojamąjį turtą ir kt. Taip pat egzistuoja paskolų tarpininkų platformų, kuriai skolininkas gali atiduoti savo asmeninius duomenis, o platforma kreipiasi į finansines paslaugas teikiančias įmones ir už klientą surenka jam pasiūlymus, taip leidžiant klientui išsirinkti pačias palankiausias skolinimo sąlygas sugaištant mažiau savo brangaus laiko. Tuo pačiu, tai įpareigoja įmones teikti rinkoje konkurencingus pasiūlymus.

Taigi galime pastebėti, kad klientui ypatingai svarbus ir atsakymo greitis. Tam, kad atsakymą galima būtų pateikti greitai, kredito davėjai turi automatizuoti savo vidinius kreditingumo rizikos vertinimo procesus ir pritaikyti juos taip, kad atlikus teisingą vertinimą, atsakymą dėl paskolos suteikimo klientas gautų beveik iš karto. Čia labai svarbus ne tik greitis, bet ir tikslumas. Nes nuo teisingo rizikos įvertinimo priklauso ne tik įmonės pelnas, bet ir įmonės reputacija.

Viena iš kredito davėjų rūšių yra tarpusavio skolinimas. Jis turi būti gerokai atsakingesnis už kitų kredito davėjų rūšis, nes asmenims, besikreipiantiems dėl paskolų, yra skolinamas ne tik įmonės kapitalas. Į tokias paskolas gali investuoti ir nepatyrę investuotojai. Tai yra alternatyvi investicijų rūšis, kuri susijusi su didesne rizika. Todėl labai svarbu užtikrinti ne tik besiskolinančiųjų, bet ir investuojančiųjų į paskolas interesus. O tai kasdienybėje neįsivaizduojama be mašininio mokymosi algoritmų pagalbos.

Mašininio mokymosi algoritmai padeda teisingai įvertinti klientų kreditingumo riziką, pateikiant šią informaciją investuotojams lengvai ir suprantamai. Tuomet investuotojai gali atsirinkti ar norėtų savo pinigus skolinti rizikingesniems, ar mažiau rizikingiems skolininkams. Atitinkamai, rizikingesni klientai yra vertinami aukštesne grąža, o mažiau rizikingi klientai generuotų mažesnę grąžą.

Vis dėl to, nors tarpusavio skolinimo platformos ir veikia technologijomis grįstoje aplinkoje, pažangūs mašininio mokymosi algoritmai dar nedrąsiai įtraukiami į tarpusavio skolinimo platformų kreditingumo rizikos vertinimą. Iš analizuotos literatūros matyti, kad daugelyje tyrimų taikytini labiau tradiciniai metodai, tokie kaip logistinės regresijos algoritmas. Nepaisant to, kad ne visada pasiekia tiksliausius rezultatus. Todėl aktualu tirti, kaip įvairūs mašininio mokymosi algoritmai gali pagerinti kreditingumo rizikos vertinimą tarpusavio skolinimo platformoms.

Lietuvoje tarpusavio skolinimas yra griežtai reglamentuotas. Todėl yra žinoma, kokią informaciją, prieš suteikiant paskolą besikreipiantiems būtina gauti tam, kad įmonė įvertintų kliento kreditingumo riziką. Visos, Lietuvoje veikiančios platformos, apie skolininką turi tą pačią informaciją, nes ji yra gaunama iš valstybinių registrų. Tačiau kiekviena platforma turi galimybę rinkti papildomą informaciją tiek, kiek jai leidžia bendrasis duomenų apsaugos reglamentas (BDAR), nepažeidžiant kliento teisių, vadovaujantis duomenų rinkimo, saugojimo ir valdymo principais. Žinoma, visos platformos turi tą patį uždavinį – kuo tiksliau atlikti klientų kreditingumo rizikos vertinimą. Taigi,

šio darbo metu atliktą tyrimą galėtų pritaikyti bet kuri, Lietuvoje veikianti, tarpusavio skolinimo platforma.

Darbo problema – šiais laikais alternatyvūs skolinimosi šaltiniai darosi vis populiaresni dėl komercinių bankų aukštų reikalavimų ir griežtų finansavimo galimybių bei kreditų sąlygų. Tad Lietuvoje veikiančios tarpusavio skolinimo platformos patiria didelį augimą. Tačiau vis dar išlieka iššūkių tiksliai įvertinant kiekvieno kliento nemokumo riziką ir užtikrinant efektyvų rizikos valdymą.

Darbo tikslas – išrinkti tiksliausią metodą, kuris leistų pasiekti didesnį tikslumą vertinant klientų kreditingumą tarpusavio skolinimo platformose.

Darbo uždaviniai:

1. Identifikuoti kredito rizikos nustatymo metodus, atlikus mokslinės literatūros apžvalgą.
2. Pasirinkti mašininio mokymosi metodus, tinkamus fizinių asmenų kreditingumo rizikos vertinimui.
3. Pasirinkti kintamuosius, tinkamus fizinių asmenų kreditingumo rizikos vertinimo tyrimui, remiantis moksline literatūra.
4. Sukurti kreditingumo rizikos vertinimo modelius, naudojant pasirinktus mašininio mokymosi algoritmus turimam duomenų rinkiniui.
5. Įvertinti bei palyginti metodų prognozavimo efektyvumą, taikant praktikoje naudojamus vertinimo kriterijus.
6. Pateikti išvadas bei rekomendacijas apie metodų praktinį taikymą kreditingumo rizikos vertinimui tarpusavio skolinimo platformose.

1. Literatūros apžvalga

1.1. Kreditingumo rizikos vertinimo mokslinių tyrimų apžvalga

Kreditingumo rizikos vertinimas yra svarbus kredito rizikos valdymo elementas. [2] Kredito rizika iš esmės yra rizika, kad asmuo, gavęs kreditą jo negrąžins, taip įmonė ar fizinis asmuo, suteikęs kreditą, gali patirti papildomų išlaidų arba net patirti nuostolį. Jei asmuo negrąžina paskolos dalies, ar nesumoka palūkanų, pusė, suteikusi kreditą, ne tik neuždirba pajamų, bet dar ir patiria nuostolį.

1.1.1. Kreditingumo rizikos vertinimo problematika ir aktualumas

Kredito rizika yra neatsiejama nuo bankų veiklos, nes bankai skolina savo kapitalą tam, kad sugeneruotų turtą, tačiau yra veikiami nemokumo rizikos. Kuomet rinkos svyravimai yra reguliuojami ir prognozuojami, šios rizikos ir nuostoliai yra valdomi. Kredito rizika iš esmės daro įtaką tiek šalies, tiek visuomenės stabilumui. Todėl siekiant ją kuo labiau sumažinti, būtina aktyviai vertinti kreditingumo rizikas ir diegti efektyvius rizikos vertinimo metodus. Rizika kyla ne tik bankų kapitalui, bet ir bendram ekonomikos stabilumui. Teisingai vertinant kredito rizikas, pasiekiamas ekonominis šalies stabilumas. [3]

Vis dėl to kreditingumo rizikos vertinimas labiausiai aktualus bankams ar kitoms finansinėms institucijoms, kurios teikia finansines paslaugas. Tokios įstaigos yra labiausiai suinteresuotos kaip įmanoma labiau minimizuoti savo nuostolį ir maksimizuoti gaunamą pelną. O tam svarbu gebėti tinkamai vertinti riziką kiekvieno kliento lygmenyje. Tad galime teigti, jog kredito rizikos valdymas yra viena iš reikšmingiausių finansinių institucijų užduotis. [4]

Kreditingumo rizikos vertinimas iš esmės yra pinigų praradimo tikimybės apskaičiavimas. Šiais laikais duomenų kiekiai yra smarkiai išaugę, todėl šis uždavinys ne toks ir paprastas. Svarbu atsirinkti kokie duomenys yra tikrai svarbūs bei aktualūs ir kokią įtaką jie daro. Todėl technologijų integracija finansų įstaigose yra privaloma siekiant kuo tiksliau prognozuoti riziką. [5] Kreditingumo rizikos vertinimo metu yra nustatoma ar asmuo gali gauti kreditą ir kokios bus suteikto kredito sąlygos, įskaitant ir palūkanų normą. Vertinimui pasitelkiama labai daug informacijos, tokios kaip skolininko kredito istorija, prašomos paskolos suma, prašomos paskolos terminas, turimi asmens įsipareigojimai, esami įsipareigojimų padelsimai ir kita prieinama informacija. [6]

Metodai, taikomi kreditingumo rizikos nustatymui, su laiku keitėsi, nuo paprastų iki labai pažangių. Tačiau jų visų veiklos principas yra analogiškas. Pirmiausia duomenis reikia surinkti, tuomet visa surinkta informacija kruopščiai įvertinama ir nustatomas kreditingumo rizikos balas arba kredito nemokumo tikimybė. Svarbu pabrėžti, kad prognozės tikslumas priklauso ne tik nuo teisingai pasirinkto metodo, tačiau ir nuo duomenų tikslumo bei apimties. Kuo daugiau duomenų yra įtraukiama į modelį, tuo jo prognozės, tikėtina, bus tikslesnės. [7]

Mašininio mokymosi algoritmai iš esmės keičia finansų institucijų procesus, susijusius su kreditingumo rizikos vertinimu. Anksčiau naudoti metodai, dažniausiai identifikuodavo tiesines priklausomybes duomenyse, tačiau ne visada tarp duomenų jos egzistuoja. Tuo tarpu mašininio mokymosi algoritmai geba apdoroti dideles ir įvairias duomenų imtis bei modeliuoti sudėtingas, netiesines priklausomybes. Šių metodų pagalba dabar pavyksta pagerinti kreditingumo rizikos nustatymo tikslumą bei netgi lankstumą. [8]

Kreditingumo rizikos vertinimas yra sudėtingas procesas, susiduriantis su įvairiais iššūkiais, kurie gali daryti įtaką rizikos nustatymui [9]:

a) **Duomenų kokybė ir prieinamumas** – surinkus nepakankamai duomenų, arba surinkus netikslius duomenis, kreditingumo rizika bus įvertinta, tačiau ji bus nepatikima, o tai gali sąlygoti klaidingus sprendimus. Taip pat dėl duomenų apsaugos kyla apribojimai naudoti jautrius duomenis, tokius kaip socialinių tinklų informacija, elektroninės prekybos informacija.
b) **Reguliaciniai reikalavimai** – finansų sektorius yra griežtai reguliuojamas ir prižiūrimas institucijų, taip pat ir kreditingumo rizikos vertinimas, kuris privalo atitikti tarptautinius bei šalies reikalavimus. Tarpusavio skolinimo platformos dar dažnai susiduria su neaiškia ar nepakankamai reglamentuota reguliacine aplinka.
c) **Ekonominis nepastovumas** – ekonominio nuosmukio metu daugėja nemokių klientų, o tai yra įtakota makroekonominių veiksnių, tokių kaip infliacijos, palūkanų normų augimas.
d) **Pasaulinis tarpusavio ryšys** – bankai bei finansų institucijos dažnai veikia keliose rinkose. Todėl ir rizikos vertinimas turi apimti skirtingas jurisdikcijas bei ekonomines sąlygas.
e) **Netradicinių investuotojų augimas** – fintech įmonės, tokios kaip tarpusavio skolinimo platformos ar kriptovaliutų rinkos su alternatyviomis investicijomis, kelia iššūkių. Tokioms įmonėms tradiciniai kredito rizikos nustatymo metodai gali būti neveiksmingi. Taip pat, alternatyvias investicijas atlikti yra gerokai lengviau nepatyrusiems investuotojams, kas gali lemti per didelę prisiimtą riziką.
f) **Žmogiškieji veiksniai (klaidos**) – kreditingumo rizikos vertinimas neįsivaizduojamas be žmogaus dalyvavimo procese, o tai reiškia, kad žmonės gali suklysti ar net tyčia manipuliuoti duomenimis. Taip pat nors kreditingumo rizikos vertinimo algoritmai ir automatizuojami, parenkant tinkamus parametrus būtinas žmogaus dalyvavimas, o neteisingai parinkus parametrus ar neteisingai suprogramavus algoritmus, finansų institucijos gali turėti neigiamų pasekmių.

1.2. Tarpusavio skolinimo platformos

Šiomis dienomis skolinimasis užima labai reikšmingą vietą žmonių gyvenime. Dažniausiai žmonės pildo paskolų paraiškas siekdami padidinti savo perkamąją galią. [10] Tačiau skolinimosi priežastys gali būti labai įvairios – nuo ilgai planuotų didesnių pirkinių (būsto ar automobilio) iki labai netikėtai patirtų medicininių išlaidų. Gyventojai siekia gauti prekę ar paslaugą šiandien, tačiau už ją susimokėti ateityje, kartu sumokėdami ir atitinkamą kainą.

Finansinių institucijų, teikiančių vartojimo kreditus šalies gyventojams, konkurencija Lietuvoje yra labai didelė. Todėl gyventojai, nedidelių sumų vartojimo paskolų atveju, priprato paskolos paraišką užpildyti patogiai ir su minimaliu įsitraukimu, atsakymą dėl paskolos gavimo gauti greitai, o pinigus savo sąskaitoje, daugumoje atveju, matyti dar tą pačią dieną. Vadinasi, siekdamos nepralaimėti konkurencinės kovos, visos finansų bendrovės privalo įdėti didelį indėlį į atsakymo paskolos gavėjui suteikimo greitį. O tai reiškia, kad didžioji dauguma sprendimų priėmimų turėtų būti automatizuojami.

Ypač svarbu yra kuo tiksliau prognozuoti paskolos nemokumo tikimybę, kad pagal tai finansų institucija įsivertintų ar klientas atitinka įmonės rizikos lygį, ir ar yra kredituotinas. Netikslios prognozės gali lemti finansų institucijos pelno dalies praradimą. Tad geriausia tokius sprendimus priimti remiantis didžiaisiais duomenimis bei matematiniais modeliais. [10]

Vienas finansinių institucijų tipų, veikiančių Lietuvoje, yra tarpusavio skolinimo platformos. Šios platformos yra internetinės platformos, kurios tiesiai sujungia besiskolinančius asmenis su asmenimis, siekiančiais investuoti. [11]

1 pav. Supaprastintas tarpusavio skolinimo platformų veikimo modelis [12]

Pirmosios tarpusavio skolinimo platformos buvo 2005 m. Jungtinėje Karalystėje įkurta „Zopa“ bei 2006 m. Jungtinėse Amerikos Valstijose įkurta „Prosper“. [13] Platformos vaidmuo svarbus, nes pasitelkiant technologijas vertinama paskolos gavėjo rizika, randami investuotojai kiekvienai paskolai, atliekamas paskolos finansavimo procesas. Tuo pačiu platformos privalo užtikrinti reguliacinę atitiktį. [14] Nors šio modelio pirminė idėja buvo sukurti lengvą investavimo įrankį fiziniams asmenims, tačiau ilgainiui tai išsiplėtė dar plačiau, apimant vis daugiau sričių. JAV rinkoje į tarpusavio skolinimo veiklą yra įtraukti ir instituciniai investuotojai, tokie kaip rizikos draudimo fondai, privataus kapitalo fondai investiciniai bankai ir kt. [15]

Įprastai komerciniai bankai orientuoti teikti įvairias finansines paslaugas – nuo kasdienių mokėjimų, banko kortelių iki indėlių ar kitų investavimo įrankių. Tarpusavio skolinimo platforma apsiriboja tik paskolų teikimu. Vidiniai platformų procesai yra iš dalies arba visiškai automatizuoti, kiek tai galima pasiekti. O tai savo ruožtu mažina finansavimo laiko kaštus. [14] Tuomet šis skolinimosi būdas, dėl greito atsakymo gavimo, tampa patrauklesnis vartotojui.

Nors tarpusavio skolinimo platformos pasaulyje veikia jau daugiau nei du dešimtmečius, Lietuvoje sukaupta 11 metų patirtis. Pirmoji platforma įsteigta 2014 m. ir kiekvienais metais platformų vartojimo kredito portfeliai stabiliai auga. O 2024 m. pabaigos duomenimis, tarpusavio skolinimo operatoriaus licenciją turėjo 4 bendrovės, iš kurių 3 aktyviai vykdo veiklą. [1] Tad nors lietuviai turi plačią kreditų davėjų pasirinkimo galimybę – komercinius bankus, kredito unijas, užsienio banko filialus ir kitas vartojimo kredito davėjų licencijas turinčias įmones [16] – jie vis labiau linkę pasitikėti tarpusavio skolinimo platformų veikla.

2 pav. Lietuvoje suteiktų vartojimo kreditų sumos (mln. Eur) [16]

Tokių platformų populiarumo augimui įtakos turi tiek vidiniai, tiek išoriniai veiksniai. Vidiniai veiksniai yra susiję su pačios platformos technologine pažanga. Platformos sukuria vartotojui draugišką ir lengvai valdomą aplinką, kur jis nesunkiai supildo paskolos paraišką, pasirašo sutartį ir tuomet labai lengvai gali vykdyti mokėjimus. Taigi palengvintas procesas ne tik pasidaro patrauklus vartotojui, bet dar ir yra labai spartus. [17] Klientui nebereikia laukti dienų ar savaičių tam, kad sužinotų savo pasiskolinimo galimybes. Dažniausiai atsakymas yra pateikiamas iš karto, dar vartotojui neuždarius internetinio puslapio lango, arba per keletą valandų, jei reikalingas kreditingumo rizikos vertinimo skyriaus vadovo įsikišimas.

Išoriniai veiksniai yra susiję ne tik su ekonominiais veiksniais, kurie tiek komerciniams bankams, tiek tarpusavio skolinimo platformoms yra tokie patys ir daro vienodą įtaką, bet ir su teisiniu reguliavimu Lietuvoje. Esminis skirtumas yra tas, kad tarpusavio skolinimo platformos veikia pagal Vartojimo kredito įstatymą, o komerciniai bankai patenka į Bankų įstatymo reguliavimo sritį. Tai juos įpareigoja laikytis rizikos valdymo reikalavimų ir turėti finansinių atsargų, siekiant apsaugoti indėlininkų lėšas. [17]

Labai svarbu paminėti, kad dalyvavimas tarpusavio skolinimo platformų veikloje yra susijęs su rizika, ypač dėl potencialių paskolų gavėjų nemokumo. Kadangi šis alternatyvus skolinimo šaltinis siūlo didesnę grąžą, natūralu, jog ir siūloma investavimo rizika yra aukštesnė. [18] Siekiant mažinti šią riziką, platformos skiria didelį dėmesį teisingam paskolų klasifikavimui ir mokumo vertinimui. Tačiau galutinį sprendimą dėl investavimo priima pats investuotojas. Tik jis yra atsakingas už savo pasirinkimą dėl investavimo. Todėl platformos atskleidžia investuotojams daug papildomos informacijos ne tik apie paskolą, bet ir apie paskolos gavėją.

Paskolų mokumas priklauso tiek nuo ekonominių veiksnių, tiek nuo skolininkų aplinkybių pasikeitimo. Investavimo metu, investuotojai mato fiksuotą dabartinę skolininko situacija, kuri ateityje nebūtinai gali keistis į gerąją pusę. Asmuo gali susidurti su ilgalaikiais sunkumais grąžinti paskolą. Taip pat žinoma, kad nors aukštą mokumo įvertinimą turintys klientai gali rinktis iš kur pasiskolinti geresnėmis sąlygomis, tačiau žemesnio kreditingumo klientai įprastai pasirinkimo neturi, arba turi gerokai siauresnį. [19] O tai leidžia suprasti, kad tarpusavio skolinimo platformų portfeliai yra rizikingesni nei, pavyzdžiui, tradicinių komercinių bankų, nes siūlo lankstesnes sąlygas.

Siekdamos kuo tiksliau nustatyti kreditingumo riziką, tarpusavio skolinimo platformos kruopščiai surenka bei patikrina duomenis apie paskolos gavėją. Tuomet taiko įvairius matematinius metodus ir algoritmus paskolų klasifikavimui ir kreditingumo rizikos vertinimui bei siekiant nustatyti ar paskolos gavėjas yra krediuotinas. [20] Tai apima statistinius modelius, mašininio mokymo algoritmus ir kitus matematinio modeliavimo būdus.

1.3. Kreditingumo rizikos vertinimui naudojamų duomenų ypatybės

Dėl informacinių technologijų plėtros, didieji duomenys žengia didelį žingsnį ir daro perversmą įvairiose srityse. Iš esmės tai yra informacijos visuma, kurią kompiuteriu galima greitai surinkti, apdoroti, saugoti, analizuoti bei valdyti. Tokių duomenų analizė leidžia didelius duomenų srautus, pritaikius įvairius matematinius modelius, panaudoti skirtingiems prognozavimo tikslams. Tarp jų ir kredito rizikos prognozavimui. [21]

Įprastiniai duomenys apie kredito paraiškos teikėją dažniausiai susideda iš jo asmeninės informacijos, tokios kaip amžius, lytis, gyvenamoji vieta. Tačiau duomenų kiekiai auga itin sparčiai, šiais laikais, kredito bendrovės surenka labai daug informacijos apie kredito paraiškos teikėją. Tad ją reikia mokėti teisingai panaudoti. Ypač svarbu tuomet, jei asmuo neturi jokios kredito istorijos duomenų. Tam galima panaudoti bet kokią prieinamą informaciją, pavyzdžiui, surinkti duomenis iš finansinių institucijų apie asmens lėšų judėjimą sąskaitose. Taip galima suprasti kaip asmuo valdo savo finansus, kokie asmens apsipirkinėjimo internetu įpročiai, kokios pagrindinės išlaidų grupės, apskritai ar geba tvarkytis su savo išlaidomis bei finansiniais įsipareigojimais. [22] Be kita ko, didieji duomenys padeda identifikuoti ir finansinio sukčiavimo atvejus. Gebant identifikuoti tokius atvejus, galima užkirsti kelią naujų atvejų atsiradimui, o kartu ir išvengti papildomų išlaidų. [23]

Kredito rizikos vertinimui tradiciškai pasitelkiama istorinė informacija apie skolininko gebėjimus tvarkytis su finansiniais įsipareigojimais, įskaitant tiek dabartinius, tiek pasibaigusius įsipareigojimus. O tai tiesiogiai atspindi skolininko gebėjimą grąžinti paskolą laiku ateityje. Tad ši informacija yra nepamainoma prognozuojant kredito gavėjo nemokumą. Taip pat visada atsižvelgiama į tai, ką klientas nurodo vertindamas savo pajamas ar finansinius įsipareigojimus. Faktinė informacija sulyginama su tuo, ką nurodo pats klientas. Pastebėta įdomi išvada – skolininkai, kurie sąmoningai nurodo didesnes pajamas, nei faktinės, yra labiau linkę tapti nemokiais, naujos paskolos atžvilgiu. [24]

Duomenis, tinkančius kredito rizikos vertinimui, galima surinkti bei analizuoti ir iš tokių socialinių tinklų kaip „Linkedin“, kur galima skaičiavimams pasitelkti ne tik asmens užimamų pareigų informaciją, bet ir analizuoti jo profesinius ryšius. Egzistuoja veikiančių platformų, kurios duomenis renka ne tik iš „Linkedin“, bet ir iš „Facebook“ ar „Twitter“ socialinių tinklų. Tokiu būdu galima įvertinti asmens ryšius su rizikos grupėms priskirtais asmenimis, ar net analizuoti asmens intelektinį lygį pagal jo socialiniuose tinkluose viešai paskelbtų pranešimų vartojamą žodyną bei padarytas rašybos ir skyrybos klaidas. [25]

Egzistuoja ir kita medalio pusė, kredito rizikos vertinimui pasitelkiant socialinių tinklų informaciją. Kyla rizika, kad pasiskolinti siekiantys asmenys sąmoningai keis savo informaciją taip, kad keltų didesnį pasitikėjimą. Pavyzdžiui, didintų savo ryšių sąrašą, ieškodami patikimų kontaktų. Tai galėtų būti valstybės tarnautojų profiliai. [25]

Tikriausiai niekas neprieštarautų, jog didieji duomenys, vertinant kredito riziką, sukuria labai didelę vertę. Gebėjimas pasinaudoti didžiaisiais duomenimis gali suteikti rimtą konkurencinį pranašumą. Tačiau tai tuo pačiu neapsieina ir be iššūkių. Generuojami duomenų kiekiai iš įvairių šaltinių gali būti ypač dideli. Taip tampa vis sunkiau užtikrinti, kad jie yra teisingi ar laiku pateikiami. Be kita ko, labai svarbu užtikrinti duomenų, iš skirtingų šaltinių, suderinamumą bei efektyvų apdorojimą. O taip pat yra ir etikos bei saugumo užtikrinimo klausimai, kai jautrūs duomenys yra gaunami iš netradicinių kanalų, pavyzdžiui, socialinių tinklų. [9]

Pastebėta, jog susidaro įspūdis, kad kuo daugiau duomenų surenkama, tuo tikslesnį prognozavimo modelį galima sukurti, bet tai nėra tiesa. Klaidinga manyti, kad tradicinio mašininio mokymosi modelių efektyvumas bus geresnis dėl didesnio įvesties duomenų kiekio. Nors mašininio mokymosi algoritmai yra neatsiejama didžiųjų duomenų analizės dalis, tačiau per daug duomenų gali gerokai prailginti algoritmų veikimo trukmę. Todėl svarbu teisingai pasirinkti, kokie duomenys bus reikšmingi tyrimui. [26]

Alternatyvūs duomenys suprantami kaip duomenys, kurie nedalyvauja tradiciniame kredito rizikos vertinimo procese. Dažniausiai kalbama apie socialinių tinklų istorinę informaciją, naršymo istoriją, elektroninių parduotuvių apsipirkimo istoriją, telefono naršymo istoriją, skaitmeninį pėdsaką ir kt. Tokia informacija leidžia suprasti skolininko elgseną, jo finansinius ar apsipirkimo internete įpročius. Tuomet, neturint jokios kredito istorijos duomenų, yra galimybė susidaryti skolininko paveikslą ir nuspėti jo elgseną su naujai atsiradusiu finansiniu įsipareigojimu. [27]

Kredito rizikos vertinimui naudojami alternatyvūs duomenų šaltiniai leidžia pasiekti aktualią ir pačią naujausią informaciją vos per kelias sekundes. O atliekant ilgalaikę stebėseną galima pastebėti blogėjančią asmens finansinę situaciją dar prieš susiduriant su rimtomis nemokumo problemomis. Tai gali padėti finansinėms institucijoms, gavus neigiamus signalus apie skolininką, sumažinti jo kredito limitus, keisti paskolos sąlygas ar net imtis griežtesnių veiksmų. Tai finansų įstaigoms leidžia būti pasiruošus staigiems skolininko kreditingumo pokyčiams. [28]

Alternatyvių duomenų naudojimas atneša naudos diskriminavimo kontekste. Jei tradiciniam rizikos vertinimui beveik visada yra įtraukiama informacija apie asmens amžių, lytį, tautybę ar pilietybę, tai naudojant alternatyvius duomenis galima šios informacijos nenaudoti. Dėl šių priežasčių į vertinimo procesą galima įtraukti tik tuos duomenis, kurie tiksliai prognozuotų kredito riziką.

Nepaisant to, kad alternatyvūs duomenys pastaruoju metu pradėti itin plačiai naudoti vertinant kredito riziką, svarbu atkreipti dėmesį, kad tai yra labai nauja sritis, kuri dar neturi aiškaus reguliavimo ir teisinės bazės, kuri apibrėžtų kaip šiuos duomenis naudoti. Tad ši sritis dar yra labai miglota ir atsakingos šalies institucijos turėtų reglamentuoti alternatyvių duomenų naudojimą įstatymiškai. Alternatyvių duomenų šaltinių naudojimas kredito rizikos vertinimo tikslais turėtų būti smarkiai apribotas, kad niekaip nepažeistų skolininko interesų. [29]

Apibendrinant galima teigti, kad duomenys yra esminis veiksnys siekiant vertinti klientų kreditingumo riziką. Jie gali būti įvairiausi, pradedant asmenine informacija, baigiant didelės apimties duomenų rinkiniais, gaunamais iš viešai prieinamų šaltinių. Šiuolaikinėje visuomenėje, kuomet visi naudojasi skaitmeninėmis technologijomis, kiekvienas asmuo sukuria savo skaitmeninį pėdsaką. O informacija, kurią galima pasiekti viešai, gali būti labai naudinga net ir vertinant fizinio asmens kreditingumo riziką. Iš analizuotų šaltinių matyti, kad kai kurios, užsienyje veikiančios, tarpusavio skolinimo platformos panaudoja socialinių tinklų informaciją vertinant asmens kreditingumo riziką.

1.4. Kreditingumo rizikos vertinimo praktikoje taikomi metodai

Duomenų atsirinkimo svarba yra neabejotina, tačiau ne ką mažiau svarbu ir teisingai pasirinkti efektyvius kreditingumo rizikos vertinimo metodus. Šiame skyrelyje apžvelgsime mokslinėje literatūroje analizuotų tarpusavio skolinimo platformų kreditingumo rizikos vertinimo praktikas bei vyraujančias tendencijas.

Tarpusavio skolinimo platforma „Lending Club“, veikianti Jungtinėse Amerikos Valstijose, savo veiklą pradėjo 2006 m. ir sėkmingai veikia iki dabar. Platforma turi sukaupusi didelę klientų bei paskolų duomenų bazę, kuri yra lengvai prieinama viešuose šaltiniuose. Dėl šios priežasties, atliekant mokslinius tyrimus, gana dažnai remiamasi „Lending Club“ platformos duomenimis analizuojant kreditingumo rizikos prognozavimo metodus. Tyrėjai pasirenka skirtingus laikotarpius bei metodus, todėl ir jų pateikiamos išvados apie metodų efektyvumą bei tinkamumą gali skirtis priklausomai nuo duomenų imties.

Mokslinėje literatūroje gausu tyrimų, analizuojant paskolų rizikas, atliktų naudojant viešai prieinamus „Lending Club“ platformos duomenis. Nors autoriai gana dažnai pasirenka panašius metodus, tačiau išvados dėl efektyviausių metodų nebūtinai sutampa. Toliau apžvelkime keletą tokių tyrimų.

Autoriai P. Teply’is ir M. Polena [30] savo tyrime analizavo „Lending Club“ platformos 2009 – 2013 m. laikotarpio duomenis ir lygino 10–ies skirtingų metodų efektyvumą, įskaitant logistinę regresiją, dirbtinius neuroninius tinklus ir linijinę diskriminacinę analizę. Vertinimui panaudoję 6 skirtingų matų rinkinį išreitingavo kiekvieną metodą pagal efektyvumą. O galutinis rezultatas parodė, kad jų imčiai geriausiai tiko logistinės regresijos metodas. Autorių teigimu, tai parodo, jog kreditų duomenų rinkiniai gali būti tiesiškai atskiriami.

Panašias išvadas pateikė ir W. Zhang’as su kolegomis [31], kurie analizavo dviejų tarpusavio skolinimo platformų „Lending Club“ ir „Renrendai“ duomenis, apimančius 2007 – 2014 m. laikotarpį. Tyrėjai pastebėjo, jog iš 8 bandytų metodų, abiem duomenų imtims labiausiai pasiteisino logistinės regresijos modelis. Analizuojant dvi atskiras platformas, veikiančias visiškai skirtingose rinkose ir abiem duomenų imtims pritaikant vienodų metodų rinkinį, gautas vieningas rezultatas.

Analizuojant labai mažą duomenų imtį galima pasiekti netikėtų rezultatų. Štai autorius L. Zhu’is su komanda [32] savo tyrime analizuodami, palyginus, mažą duomenų imtį – paskolas, suteiktas per 2019 m. pirmąjį ketvirtį, gavo visai kitą rezultatą. Tyrime taikydami atsitiktinio miško (angl. *random forest*), sprendimų medžio, atraminių vektorių klasifikatoriaus (angl. *support vector machine*), logistinės regresijos metodus, pagal gautus rezultatus nustatė, jog tiksliausiai reikšmes prognozavo atsitiktinio miško metodas. Pastarasis metodas yra geras tuo, jog geba greitai apdoroti itin didelius duomenų kiekius. Kita vertus, autoriaus V. Padimi’s ir komandos tyrimas parodė, jog analizuojant labai didelę imtį ir lyginant penkių skirtingų algoritmų efektyvumą, geriausius rezultatus pademonstruoja atsitiktinio miško algoritmas. Verta paminėti, jog komanda analizavo Europoje veikiančios platformos „Bondora“ duomenis. [33]

R. Sifrain’as (2023), analizavęs pakankamai neseną „Lending Club“ platformos duomenų imtį pastebėjo, kad logistinė regresija nedavė tiksliausio rezultato. Tyrimui buvo pasitelkti 2017–2018 metų duomenys, o analizėje panaudoti logistinės regresijos, atsitiktinio miško bei dirbtinių neuroninių tinklų modeliai parodė, jog tiksliausiai duomenis prognozavo dirbtinių neuroninių tinklų modelis.

[34] Dirbtiniai neuroniniai tinklai geba aptikti netiesinius ryšius, o tai parodo, kad duomenyse atpažinta netiesinė priklausomybė. Tačiau dirbtiniai neuroniniai tinklai praktikoje nėra dažnai naudojami dėl sudėtingo paaiškinamumo. Tarpusavio skolinimo platformos privalo laikytis bendrojo duomenų apsaugos reglamento (BDAR) reikalavimų, o jie suteikia teisę į paaiškinimą – kuomet vartotojai gali kreiptis paaiškinimo dėl procesų, darančių įtaką sprendimo priėmimui. [35]

Pateikti pavyzdžiai gerai atspindi, kad nors naudoti tos pačios platformos duomenys, tyrėjams pasirinkus skirtingų laikotarpių duomenis, analizuojamos skirtingos imtys. Todėl ir taikomų metodų rezultatyvumas skiriasi. Tad nors dažniausiai „Lending Club“ platformos duomenims tinkamas buvo logistinės regresijos modelis, matyti, kad viename straipsnyje sėkmingiausias buvo atsitiktinio miško metodas, o kitos analizės geriausiu rezultatu tapo dirbtinių neuroninių tinklų metodas.

Toliau analizuojami moksliniai straipsniai, kuriuose tiriami kitų tarpusavio platformų duomenys kreditingumo rizikos nustatymo tikslais. Štai Kinijos tyrėjų komanda analizavusi Kinijoje veikiančios tarpusavio skolinimo platformos duomenis ir lyginusi 10–ies metodų veikimo tikslumą nustatė, kad tiksliausias metodas, vertinant kreditingumo riziką, buvo atsitiktinio miško metodas. Analizuojamų metodų sąraše buvo ir gerai žinomi logistinės regresijos, neuroninių tinklų, k–artimiausių kaimynų, atraminių vektorių klasifikatorių metodai. [36] Kita tyrėjų komanda, analizavusi dviejų didžiausių Kinijos tarpusavio skolinimo platformų „Renrendai“ ir „Paipaidai“ duomenis, tarpusavyje lygino tik 4 metodų veiksmingumą – logistinės regresijos metodą lygino su atsitiktinio miško, k–artimiausių kaimynų bei atraminių vektorių klasifikatoriaus metodais. Abiem duomenų imtims panaudoti tie patys algoritmai parodė, kad tiksliausiai nemokumo riziką nustato atsitiktinio miško metodas ir jis buvo ženkliai pranašesnis už logistinės regresijos metodą. [37]

Analogišką rezultatą pasiekė ir kita tyrėjų komanda, tyrusi neįvardintos Kinijos tarpusavio skolinimo platformos duomenis. Jie lygino 6 skirtingus metodus, tarp kurių buvo ir atsitiktinio miško metodas, neuroninių tinklų algoritmas, k–artimiausių kaimynų metodas, atraminių vektorių klasifikatorių metodas, tačiau neįtraukė logistinės regresijos metodo. Tyrimo išvados parodė, kad sėkmingiausias metodas buvo atsitiktinio miško algoritmas, su kuriuo pavyko pasiekti itin aukštą efektyvumą – AUC reikšmė buvo net 0,97. [38]

Kinijos mokslininkų komanda nagrinėjo tarpusavio skolinimo platformos, veikiančios Jungtinėse Amerikos Valstijose, duomenis. Platformos pavadinimas šaltinyje nebuvo atskleistas. Autoriai teigia, kad buvo pasirinkti 4 patys geriausi metodai, tai, žinoma, logistinės regresijos metodas, sprendimų medžių metodas bei du mašininio mokymosi algoritmai XGBoost ir LightGBM. Pastarieji du metodai parodė pačius tiksliausius rezultatus. [39] Šie metodai naudoja sprendimų medžius, kas leidžia atlikti itin tikslų duomenų skaidymą siekiant išgauti tiksliausią informaciją iš duomenų. [40] O kadangi abu algoritmai pasiekė labai panašų tikslumą, renkantis galutinį variantą rekomenduotina atsižvelgti į algoritmo vykdymo greitį. Tuo tarpu, kita tyrėjų iš Kinijos grupė pasirinko naudoti XGBoost, logistinės regresijos ir sprendimų medžių metodus. Jiems pavyko pasiekti labai gerų rezultatų su XGBoost metodu ir pagrįsti, jog šis metodas yra tinkamas, siekiant išspręsti klasifikavimo uždavinį. [41]

Dažniausiai literatūroje palyginus abu, LightGBM ir XGBoost, algoritmus pastebima, kad LightGBM algoritmas pateikia tikslesnius rezultatus. Taip yra todėl, kad LightGBM metodas yra greitesnis, naudoja mažiau atminties resursų bei lengviau dirba su didžiaisiais duomenimis. [42] Labai svarbi LightGBM metodo savybė yra ta, jog jis grįstas sprendimų medžiais. Įprastai algoritmai medžius

augina horizontalia kryptimi, tačiau šis mašininio mokymosi algoritmas veikia vertikaliai. Algoritmas augimui pasirenka lapą su didžiausiais delta nuostoliais, o tai padeda sumažinti nuostolius vėlesnėse iteracijose. Visa tai leidžia pasiekti daug didesnį tikslumą, lyginant su kitais algoritmais. [43]

Klientų kreditingumo rizikos vertinimo uždaviniuose lyginant mašininio mokymosi metodų LightGBM ir XGBoost efektyvumą su kitų algoritmų efektyvumu, labai dažnai šie du algoritmai parodo itin gerus rezultatus. G. S. Alzamora's su komanda tirdami Peru mikrokreditus taip pat įrodė, jog LightGBM metodas yra tiksliausias, nors tyrime buvo panaudota net 10 skirtingų metodų, tarp kurių naudoti ir logistinės regresijos bei XGBoost metodai. [44]

Pastaraisiais metais vis daugėja tyrimų, kurie į tiriamų metodų sąrašą įtraukia ir LightGBM algoritmą. Pastebima, kad šis algoritmas puikiai atlieka klasifikavimo užduotį. Nepaisant to, jog kiti gerai žinomi algoritmai pateikia labai panašius rezultatus, būtina svarstyti paties tiksliausio algoritmo integraciją. Nes tai ne tik leis tiksliau numatyti tikimybę ar klientas taps nemokus, tačiau ir padės įmonei uždirbti daugiau pajamų ar net padidinti investuotojų pasitikėjimą platforma. Mokslininkų P. Ko'o ir kt. komanda, analizavusi didelės tarpusavio skolinimo platformos, veikiančios JAV, duomenis įrodė, kad pasirinkus LightGBM metodą nemokumui nustatyti, platforma uždirbtų papildomai 24 milijonus JAV dolerių pajamų per vienerių metų laikotarpį. [45]

Nors vyrauja nuomonė, kad logistinės regresijos modelis yra tinkamiausias metodas rizikai vertinti dėl aukšto tikslumo lygio ir lengvos interpretacijos, [46] kredito rizikos vertinimui naudojami metodai yra labai įvairūs. Jie nuolatos tobulinami, siekiant užtikrinti didesnį tikslumą. Tradicinis logistinės regresijos metodas ilgą laiką buvo vertinamas kaip patikimiausias, bet vis dažniau pastebima, kad pirmenybė teikiama mašininio mokymosi algoritmams – atsitiktinio miško, LightGBM, XGBoost ar neuroninių tinklų metodams. Nepaisant to, labai svarbu yra užtikrinti prognozių tikslumą ir nuolat stebėti metodų elgseną keičiantis duomenims.

1 lentelė. Efektyviausi kreditingumo rizikos vertinimo metodai, remiantis mokslinės literatūros analize

Autorius	**Metodai ir jų tikslumo įvertis AUC/tikslumas (angl. *accuracy*)**	**Tyrimo metu naudotų metodų kiekis**	**Tyrimo imtis**
P. Teply'is (2020) [30]	Logistinė regresija – AUC 0,698	Lyginti 10 metodų rezultatai	JAV tarpusavio skolinimo platforma „Lending Club"
W. Zhang'as (2020) [31]	Logistinė regresija – AUC 0,700	Lyginti 8 metodų rezultatai	JAV tarpusavio skolinimo platforma „Lending Club"
	Logistinė regresija – AUC 0,918	Lyginti 8 metodų rezultatai	Kinijos tarpusavio skolinimo platforma „Renrendai"
L. Zhu'is (2019) [32]	Atsitiktinio miško metodas – AUC 0,983	Lyginti 4 metodų rezultatai	JAV tarpusavio skolinimo platforma „Lending Club"
V. Padimi's (2022) [33]	Atsitiktinio miško metodas – ACC 0,946	Lyginti 5 metodų rezultatai	Estijos tarpusavio skolinimo platforma „Bondora"
R. Sifrain'as (2023) [34]	Logistinė regresija – Accuracy 0,899	Lyginti 3 metodų rezultatai	JAV tarpusavio skolinimo platforma „Lending Club"

Autorius	Metodai ir jų tikslumo įvertis AUC/tikslumas (angl. *accuracy*)	Tyrimo metu naudotų metodų kiekis	Tyrimo imtis
W. Yin'as (2023) [36]	Atsitiktinio miško metodas – preciziškumas (angl. *precision*) 0,810	Lyginti 10 metodų rezultatai	Kinijos tarpusavio skolinimo platforma
Y. Liu'is (2022) [37]	Atsitiktinio miško metodas – AUC 0,966	Lyginti 4 metodų rezultatai	Kinijos tarpusavio skolinimo platforma „Renrendai"
	Atsitiktinio miško metodas – AUC 0,836	Lyginti 4 metodų rezultatai	Kinijos tarpusavio skolinimo platforma „Paipaidai"
H. Wang'as (2021) [38]	Atsitiktinio miško metodas – AUC 0,970	Lyginti 6 metodų rezultatai	Kinijos tarpusavio skolinimo platforma
X. Zhu'is (2023) [39]	LightGBM – AUC 0,721	Lyginti 4 metodų rezultatai	JAV tarpusavio skolinimo platforma
Z. Li'is (2021) [41]	XGBoost – AUC 0,977	Lyginti 3 metodų rezultatai	JAV tarpusavio skolinimo platforma „Lending Club"
G. S. Alzamora's (2022) [44]	LightGBM – AUC 0,962	Lyginti 10 metodų rezultatai	Peru mikrofinansų įmonė
P. Ko'as (2022) [45]	LightGBM – AUC 0,749	Lyginti 8 metodų rezultatai	JAV tarpusavio skolinimo platforma

Apibendrinant, mokslinėje literatūroje gausiai nagrinėjami įvairūs kreditingumo rizikos vertinimo metodai, taikomi tarpusavio skolinimo platformose suteikiamoms paskoloms. Nors ilgą laiką logistinės regresijos modelis buvo laikomas geriausiai atspindinčiu tiesines duomenų priklausomybes, pastebima, kad sudėtingesni mašininio mokymosi algoritmai įgauna populiarumą dėl aukštesnio kreditingumo rizikos vertinimo efektyvumo. Taip yra todėl, kad duomenų priklausomybės nebėra tiesinės – egzistuoja sudėtingesni ryšiai. Tokius ryšius vis sėkmingiau atpažįsta pažangūs LightGBM ar XGBoost metodai. Šie metodai geba ne tik pastebėti sudėtingus ryšius, bet ir greitai apdoroti didelius duomenų kiekius. Nors metodų efektyvumas yra labai svarbus siekiant atlikti kreditingumo rizikos vertinimą, ne visada galima pasirinkti geriausiai klasifikuojančius metodus dėl jų sudėtingo modelių paaiškinamumo. Kai kuriais atvejais lengvas paaiškinamumas yra svarbiau už modelio efektyvumą, tad kreditingumo rizikos vertinimui pasirenkami klasikiniai metodai.

1.5. Kreditingumo rizikos vertinimo praktikoje naudojamų kintamųjų apžvalga

Dauguma autorių sutaria, jog paskolos įsipareigojimo nevykdymo veiksnius galima suskirstyti į keturias grupes:

- informacija apie paskolą,
- asmens kredito istorija,
- paskolos gavėjo asmeninė informacija,
- makroekonominiai veiksniai.

Informacija apie paskolą dažnai yra svarbus kintamasis susijęs su paskolos nemokumo tikimybe. Tai gali būti tokia informacija kaip paskolos suma, terminas, tikslas ar net palūkanų norma. Pasak analizės autorių, labai dažnai randamas neigiamas ryšys tarp paskolos sumos ir paskolos nemokumo

tikimybės. Tai reiškia, kad kuo didesnė paskolos suma, tuo didesnė tikimybė, kad asmuo nesugebės įvykdyti įsipareigojimų laiku. Tačiau tai iš dalies yra pateisinama tuo, kad kuo didesnė besiskolinama suma, tuo ilgesnis būna paskolos periodas. [47]

Asmens kredito istorija vienas iš esminių rodiklių įrodančių asmenų gebėjimą vykdyti įsipareigojimus laiku. Tai apima visą istorinę informaciją apie turėtus įsipareigojimus, jų mokėjimus, tame tarpe ir buvusius įsipareigojimų pradelsimus. Pasitelkę įvairius įsipareigojimų vykdymo registrus galime įvertinti tikimybę, kad asmuo bus linkęs mokėti paskolą laiku. Atsakingi skolininkai yra suinteresuoti išlaikyti gerą kreditingumo istoriją, kad ateityje, prireikus, galėtų lengvai ir geresnėmis sąlygomis pasiskolinti. [48] Įprastai kredito istorija yra laikoma vertingu informacijos šaltiniu, nes norint asmeniui išlaikyti aukštą kreditingumą, reikia įdėti nemažai pastangų ir disciplinos. [49]

Prie paskolos gavėjo asmeninės informacijos dažniausiai yra priskiriama tokia informacija kaip amžius, lytis, asmeninės pajamos, turimi finansiniai įsipareigojimai, šeiminis statusas, darbinė patirtis ir kt. Dažnu atveju pastebima, kad asmens pajamos, amžius, šeiminis statusas turi reikšmingą įtaką nustatant tikimybę, jog asmuo susidurs su sunkumais vykdant įsipareigojimus. [48] Štai X. Tang'as su komanda savo tyrime nustatė, kad lytis turi reikšmingą vaidmenį, nustatant nemokumo tikimybę, jie teigia, kad vyrai bus labiau linkę įsipareigojimus vykdyti laiku. [49]

Makroekonominiai veiksniai svarbūs ne tik ekonominės situacijos prognozavimui, tačiau ir skolininkų elgsenos prognozavimui. Vyrauja nuomonė, kad makroekonominiai veiksniai ar regioninė aplinka daro stiprią įtaką skolininkų mokumui. [50] Infliacija ir palūkanų normų didėjimas tiesiogiai veikia visus gyventojus. [48] Auganti infliacija bei kylančios palūkanų normos mažina skolininkų liekamų pajamų dalį, o tai reiškia, kad asmenys gali susidurti su trumpalaikio ar ilgalaikio mokumo problemomis. [51] O tai sąlygotų išaugusi nemokumą.

Pasak kitų autorių, analizavusių istorinę kelerių metų paskolų informaciją, makroekonominiai veiksniai daro reikšmingą įtaką klientų mokumui. Nors pagrindiniai kintamieji vis vien išlieka paskolos charakteristikos, asmeninė faktinė bei istorinė skolininko elgsenos informacija, tyrimas įrodė, kad šalies nedarbo lygio augimas ir BVP augimas yra reikšmingi rodikliai. Abu faktoriai daro stiprią neigiamą įtaką mokumui. [52]

Be šių pagrindinių kintamųjų tyrėjai analizuoja ir kitokius duomenis, tarp kurių ir skolininko elgsena tarpusavio skolinimo platformoje. Analizei pasitelkti paskutinio prisijungimo duomenys, kurie parodė, kad tai yra labai svarbus rodiklis. Kuo daugiau laiko klientas yra praleidęs platformoje, tuo labiau tikėtina, kad kliento nemokumo tikimybė bus žemesnė. Tuo pačiu nustatyta, kad aktyvumas platformoje priklauso nuo skolininko kredito istorijos. Kuo geresnė kreditingumo istorija, tuo daugiau laiko klientas linkęs praleisti platformos savitarnoje ir tuo didesnė tikimybė, kad klientas yra labiau linkęs laiku grąžinti paskolą. [49]

Ir nors nepriklausomų kintamųjų įvairovė yra labai plati, ji priklauso nuo tarpusavio skolinimo platformų surenkamos informacijos. Tad kiekvienos platformos rezultatai apie didžiausią įtaką darančius veiksnius gali skirtis. Tyrėjas R. Sifrain'as išanalizavo, jog pajamų ir įsipareigojimų santykis (DTI), metinės paskolos gavėjo pajamos bei paskolos suma daro didžiausią įtaką nemokumui. [34]

X. Zhu’is ir kiti atlikę tyrimą konstatuoja, kad didžiausią įtaką nemokumo rodikliui daro paskolos trukmė, paskolos reitingas (suteikiamas atsižvelgiant į paskolos gavėjo įsipareigojimų istoriją), faktas apie nekilnojamojo turto turėjimą, paskolos suma, paskolos mėnesio įmokos suma, įsipareigojimų ir pajamų santykis, paskolos gavėjo kreditingumo reitingas. [39] . Panašius rezultatus pateikė ir autoriai S. Zhao’as su J. Zou’is, pastebėję, kad didžiausią įtaką kreditingumo rizikos vertinime turi paskolos charakteristikos: suma, trukmė, palūkanų norma ir skolininko pajamos. Pabrėžiama, kad terminas daro neigiamą įtaką, o tai reiškia, kad paskolų gavėjai, turintys trumpesnio termino paskolas, yra linkę labiau nevykdyti įsipareigojimų, nei ilgesnio termino paskolų turėtojai. Tuo pačiu, paskolos suma bei palūkanų norma turi teigiamą įtaką nemokumo rodikliui – joms didėjant, didėja tikimybė, jog skolininkas susidurs su nemokumo problemomis. [53] K. Ho’as su komanda taip pat nustatė, kad didžiausią neigiamą įtaką daro paskolos trukmė ir paskolos suma. [54]

Tačiau X. Ma’s ir bendražygių analizė rodo, kad didžiausią įtaką darančius rodiklius galima būtų išrikiuoti paeiliui, rikiuojant juos nuo svarbiausio iki mažiau svarbaus: paskolos charakteristikos, asmens finansinė informacija, kredito istorijos informacija ir demografinė skolininko informacija. Teigiama, jog tarp labiausiai nereikšmingų kintamųjų buvo metinės asmens pajamos, turimų mėnesinių įsipareigojimų suma ir net DTI santykis. [42]

Kitame atliktame tyrime atvirkščiai pabrėžiama, jog didžiausią įtaką darantys kintamieji yra asmeninė skolininko informacija: skolininko amžius, lytis, šeiminė padėtis, gyvenamoji vieta (nuosavas/nuomojamas butas/namas), o taip pat skolininko kredito istorija. [55] Dar vienas tyrimas taip pat pabrėžia, kad tarp reikšmingiausią įtaką darančių kintamųjų yra asmens išsilavinimas bei būsto ir automobilio nuosavybės faktas. Taip pat be šių kintamųjų straipsnyje išskirta paskolos trukmė, paskolos palūkanų norma bei bankų indėlių norma. [56] Kitame šaltinyje teigiama, kad didžiausią įtaką daro kredito istorijos informacija ir prabrėžiama, kad būtent trukmė nuo pirmosios paskolos ar kreditinės kortelės pradžios datos. Kuo ilgesnis terminas, tuo skolininko nemokumo tikimybė yra žemesnė. O taip pat tarp svarbių kintamųjų yra ir DTI santykis, metinės skolininko pajamos bei paskolos suma. [57]

Lentelėje žemiau pateikiami anksčiau apžvelgtų tyrimų reikšmingiausi kintamieji, nurodant kokią įtaką jie darė priklausomam kintamajam, kuris, tiriamais atvejais, buvo nemokumo tikimybė. Taip pat paminėtas ir tyrimo kontekstas bei šalis, kurios duomenų pagrindų atliktas tyrimas.

2 lentelė. Reikšmingiausių kintamųjų ir jų įtakos santrauka iš analizuotų literatūros šaltinių

Autorius	**Reikšmingiausi kintamieji**	**Įtaka nemokumo vertinimui**	**Tyrimo imtis**
X. Tang’as (2023) [49]	1. Skolininko įsitraukimas platformoje 2. Paskolos suma 3. Paskolos terminas 4. Amžius 5. Lytis (moteriška) 6. Šeiminis statusas (susituokę) 7. Registracijos trukmė platformoje	1. Neigiama įtaka 2. Neigiama įtaka 3. Teigiama įtaka 4. Neigiama įtaka 5. Neigiama įtaka 6. Neigiama įtaka 7. Neigiama įtaka	Kelių Kinijos komercinių bankų duomenys
Q. Xu’is (2024) [48]	1. Visa mokėtina paskolos suma 2. Pirkimo vadybininkų indeksai (PVI) 3. Paskolos terminas 4. Paskolos suma 5. Bendra pasiskolinta suma	1. Teigiama įtaka 2. Neigiama įtaka 3. Neigiama įtaka 4. Neigiama įtaka 5. Neigiama įtaka	Kinijos tarpusavio skolinimo platforma „Renrendai“

Autorius	Reikšmingiausi kintamieji	Įtaka nemokumo vertinimui	Tyrimo imtis
E. Baumohl'as (2024) [50]	1. Paskolos suma 2. Paskolos terminas 3. Palūkanų norma 4. Nedarbo lygis	1. Teigiama įtaka 2. Teigiama įtaka 3. Teigiama įtaka 4. Neigiama įtaka	JAV veikiančios tarpusavio skolinimo platformos „Lending Club“, „Prosper“ ir „Zopa“
R. Sifrain'as (2023) [34]	1. Paskolos suma 2. DTI santykis 3. Metinės pajamos 4. Bendra negrąžintų paskolų suma 5. Darbo trukmė 6. Kredito užklausų skaičius per 6 mėn. 7. Vieši įrašai apie nepatikimumą	1. Teigiama įtaka 2. Teigiama įtaka 3. Neigiama įtaka 4. Neigiama įtaka 5. Teigiama įtaka 6. Teigiama įtaka 7. Teigiama įtaka	JAV tarpusavio skolinimo platforma „Lending Club“
X. Zhu'is (2023) [39]	1. Paskolos trukmė 2. Paskolos reitingas 3. NT nuosavybės statusas 4. DTI santykis 5. Metinės pajamos 6. FICO balas	1. Teigiama įtaka 2. Teigiama įtaka 3. Neigiama įtaka 4. Teigiama įtaka 5. Neigiama įtaka 6. Neigiama įtaka	JAV tarpusavio skolinimo platforma
S. Zhao'as (2021) [53]	1. Paskolos palūkanų norma 2. Paskolos trukmė 3. Pajamos 4. Paskolos suma 5. Bendra (mokėtina) palūkanų suma	1. Teigiama įtaka 2. Neigiama įtaka 3. Neigiama įtaka 4. Neigiama įtaka 5. Teigiama įtaka	JAV tarpusavio skolinimo platforma
K. Ho'as (2024) [54]	1. Paskolos trukmė 2. Paskolos suma 3. Išsilavinimo lygis 4. Skolininko kredito reitingas 5. Amžius 6. NT nuosavybės statusas	1. Teigiama įtaka 2. Teigiama įtaka 3. Neigiama įtaka 4. Neigiama įtaka 5. Neigiama įtaka 6. Neigiama įtaka	Kinijos tarpusavio skolinimo platforma „Renrendai“
K. Wang'as (2022) [55]	1. Amžius 2. Lytis 3. Šeiminis statusas 4. NT nuosavybės statusas 5. Skolininko kredito istorija 6. Paskolos trukmė	1. Neigiama įtaka 2. Nenurodyta 3. Nenurodyta 4. Neigiama įtaka 5. Neigiama įtaka 6. Neigiama įtaka	Kinijos bankas
Y. Wu'is (2021) [56]	1. Skolininko kredito reitingas 2. Paskolos palūkanų norma 3. Paskolos trukmė 4. DTI santykis 5. Išsilavinimas 6. Lytis	1. Neigiama įtaka 2. Teigiama įtaka 3. Teigiama įtaka 4. Teigiama įtaka 5. Neigiama įtaka 6. Nenurodyta	Kinijos tarpusavio skolinimo platforma „Renrendai“
A. Perrotta [75]	1. Metinės pajamos 2. DTI santykis 3. Kredito linijos limito panaudojimas 4. Paskolos trukmė	1. Neigiama įtaka 2. Teigiama įtaka 3. Teigiama įtaka 4. Teigiama įtaka	JAV tarpusavio skolinimo platforma „Lending Club“

Apibendrinant galima sakyti, kad vieningai reikšmingų kintamųjų nėra. Nors dažnai literatūroje pasikartojimų ir pastebima, pavyzdžiui paskolos charakteristikų atvejų (paskolos suma, terminas, palūkanų norma), tačiau jų nemokumui daroma įtaka gali skirtis. Vis dėl to dažniau didesnė suma, aukštesnės paskolos palūkanų norma ar ilgesnė trukmė daro neigiamą įtaką skolininko mokumui, nes

tai sukelia didesnį finansinį spaudimą skolininkui. Nemaža dalis tyrimų patvirtina, jog paskolos gavėjo kreditingumo istorija taip pat yra labai svarbus rodiklis. Jis leidžia įvertinti istorinę skolininko informaciją apie praeityje patirtus sunkumus su įsipareigojimų vykdymu bei lengviau prognozuoti elgseną ateityje. Taip pat yra išskirti ir kiti finansiniai asmens faktoriai – pajamos, įsipareigojimai, DTI santykis. Bei asmeninė informacija – amžius, lytis, šeiminis statusas, išsilavinimas. Pastarasis kintamasis taip pat dažnai minimas šaltiniuose kaip reikšmingą įtaką nemokumo tikimybei darantis veiksnys.

1.6. Praktikoje taikomų modelių įvertinimo technikų apžvalga

Neabejojama, jog kiekvieno sukurto modelio efektyvumas turi būti įvertintas tam, kad galima būtų pasirinkti sėkmingiausiai klasifikuojantį modelį. O norit palyginti modelius tarpusavyje privalu naudoti tą pačią metriką kiekvienam modeliui atskirai. Gavus reikšmes jos yra palyginamos ir iš tiriamų metodų gausos yra išrenkami sėkmingiausiai klasifikuojantys metodai. [4] Taip pat atkreipiamas dėmesys į tai, kad specifinės metrikos leidžia palyginti modelį patį su savimi. Norint pasiekti kuo aukštesnio tikslumo, modeliuose būtina parinkti hiperparametrus. Tad, įverčių metrikos padeda suprasti, kurie modelio hiperparametrai yra teisingiausi tiriamai duomenų imčiai. [31]

Įverčių metrikas naudoti būtina ne tik norint palyginti metodų efektyvumą tarpusavyje, bet ir todėl, jog taip padeda suprasti ar modelis nepersimokė analizuodamas apmokymo imtį. Dažniausiai, norint įvertinti klasifikavimo modelio efektyvumą yra naudojamos kelios metrikos. Viena jų yra ROC kreivė. Tai dvimatė kreivė, kuri braižoma pagal du rodiklius: tikrų teigiamų atvejų rodiklis (angl. *true positive rate*) ir klaidingų teigiamų atvejų rodiklis (angl. *false positive rate*). O su šia kreive glaudžiai susijusi metrika, kuri apibendrina ROC kreivės veikimą – tai plotas po kreive, vadinamas AUC metrika. [20] Kuo reikšmė artimesnė 1, tuo tiksliau modelis klasifikuoja imties elementus.

Labai dažnai pasitaikanti klasifikavimo modelių efektyvumo vertinimo metrika yra sumaišymo matrica ir iš jos išplaukiantys įvairūs matai. Šis metodas yra populiarus dėl to, jog objektyviai vertina klasifikavimo metodų veikimą. [77] O iš apskaičiuotos matricos dar galima gauti iš išvestinių rodiklių, kurie leidžia daryti išvadas apie vertinamų modelių efektyvumą bei palyginti modelius tarpusavyje. Išvestinės metrikos taip pat labai svarbios tuomet, jei duomenų klasės yra nesubalansuotos, nes gali parodyti kaip efektyviai kreditingumo rizikos vertinimo modelis geba identifikuoti mažesnę, nemokių klientų, klasę.

Kai kurie tyrėjai naudoja ir su finansiniais ištekliais susijusias modelių vertinimo metrikas. Vis dažniau pasitaikanti tikėtino maksimalaus pelno EMP (angl. *expected maximum profit*) metrika. Ši metodo efektyvumą parodanti metrika vertinama palankiai dėl to, kad ji randa pusiausvyrą tarp iš vienos paskolos gaunamų pajamų ir tikėtinų nuostolių. [76] Toks modelio efektyvumo įvertinimas gali būti tinkamesnis kreditingumo rizikos vertinime dėl to, kad įvertina pelną. O tai parodo tiesioginę įtaką įmonės veiklai.

Apibendrinant galima teigti, jog kredito rizikos vertinimo modelių efektyvumo įvertinimas yra neatsiejama proceso dalis. Tai leidžia ne tik išrinkti efektyviausiai veikiantį metodą iš taikytų metodų gausos, bet ir teisingai parinkti kreditingumo rizikos vertinimo modelio hiperparametrus, siekiant geresnio paties metodo veikimo. Literatūroje populiariausios yra ROC kreivės bei AUC metrikos, tačiau vis dažniau pasitaiko įvairesnių sprendimų, kurie leidžia įvertinti priimamų sprendimų naudą pinigine verte. Tad dažniausiai rekomenduojama neapsistoti ties viena efektyvumo vertinimo metrika, bet vertinti keletą jų ir tik tada teikti apibendrintas išvadas.

1.7. Apibendrinimas

Kredito rizika ir kreditingumo rizikos vertinimas yra neatsiejama šių dienų kasdienybė komercinių bankų ir finansų institucijų veikloje, o tinkamas kreditingumo rizikos vertinimas leidžia nustatyti ar skolininkas gebės atsakingai vykdyti prisiimtus finansinius įsipareigojimus. Galima teigti, kad kreditingumo rizikos valdymas yra svarbus tik finansų institucijai, nes jis daro tiesioginę įtaką institucijos pelningumui. Tačiau tai tiesiogiai siejasi ir su šalies ekonomika, nes finansų institucijoms gebant teisingai valdyti riziką, nekyla grėsmių ekonomikos stabilumui. Taip pat ir atvirkščiai, išoriniai makroekonominiai veiksniai daro įtaką kreditingumo rizikos vertinimui. Pavyzdžiui, kylančios palūkanų normos ar augantys infliacijos rodikliai, daro neigiamą įtaką skolininkų mokumui.

Kreditingumo rizikos vertinimui būdingi didieji duomenys. Pastebima, kad tradiciniai metodai nebėra tokie sėkmingi, nes tarp duomenų išnyksta tiesinės priklausomybės ir randami sudėtingesni ryšiai. Įtraukiant didžiuosius duomenis į kreditingumo rizikos vertinimo modelį susiduriama su įvairiais iššūkiais. Tad nors dideli duomenų kiekiai iš įvairių duomenų šaltinių leidžia sukurti tikslesnį rizikos prognozavimo modelį, labai svarbu nepamirši šalyje galiojančių BDAR nuostatų, nes tai susiję su konfidencialumu, ir užtikrinti duomenų apsaugą.

Daugumoje analizuotų straipsnių tarp sėkmingiausių metodų atsiduria logistinės regresijos, atsitiktinių miškų ar kiti mašininio mokymosi metodai. Tačiau net ir naudojant tą patį duomenų rinkinį, metodo sėkmė priklauso nuo pasirinktų kintamųjų. Visuose tyrimuose matoma, kad tarp didžiausią įtaką darančių kintamųjų buvo asmeninė paskolos gavėjo informacija, tokia kaip amžius, lytis, šeiminė padėtis. Taip pat paskolos grąžinimo tikimybei įtaką daro ir pasirinkta paskolos suma, paskolos palūkanų norma, mėnesio įmoka bei paskolos terminas. Nepaisant to, labai svarbi yra besiskolinančiojo elgsena praeityje, nes visais atvejais į modelius įtraukiamas paskolos gavėjo kredito reitingas arba atskira informacija apie praeityje buvusius įsipareigojimų pradelsimus.

Galima pastebėti, kad kai kurie kintamieji Kinijoje ir Jungtinėse Amerikos Valstijose yra reikšmingi, tačiau Lietuvoje jie neturės jokios įtakos modelio kūrimui. Pavyzdžiui, viskas, kas susiję su analizuojamoje šalyje pakeistu reguliavimu. Lietuvoje reguliavimas nuo pat pirmosios platformos įkūrimo buvo griežtas ir reglamentuotas. Todėl tarpusavio skolinimo platformos neturi duomenų apie suteiktas paskolas, laikotarpiu, kai nebuvo jokio reguliavimo. Tad kiekviena platforma gali rinkti skirtingus duomenis apie skolininką ir iš jų gauti skirtingą naudą. O Lietuvoje veikiančios tarpusavio skolinimo platformos apie skolininkus surenka daugiau mažiau tokią pačią informaciją, nes ji pateikiama iš bendrų, privalomų registrų. Taip pat ne visą informaciją galima patikrinti ir tenka pasitikėti tuo, ką nurodo pats paskolos gavėjas, pildydamas paskolos paraišką.

2. Metodologija

Apskaičiuoti įsipareigojimų nevykdymo rodiklius yra labai svarbu siekiant įvertinti ar suteikus paskolą, kredito įstaiga patirs pelną, ar nuostolį. Ir nors dauguma paskolų ir yra grąžinamos sėkmingai, t. y. grąžinamos pagal sutartinį paskolos grafiką arba anksčiau numatyto termino, tačiau yra skolininkų, kurie finansinių įsipareigojimų nevykdo laiku. [53] O tai yra verslo rizika. Tad, svarbu, kad prieš išduodant paskolą, skolintojas gebėtų įvertinti kreditingumo riziką, t. y. tikimybę, kad paskolos gavėjas nevykdys įsipareigojimų laiku.

Mašininis mokymasis tai yra mokslo sritis, kuri tirdama didelius duomenų masyvus, ieško juose sąryšių, struktūros panašumų ir sukonstruoja klasifikavimo arba prognozavimo modelius. Tai galima daryti arba prižiūrimo mokymosi, arba neprižiūrimo mokymosi metodais. Šiais laikais dažniausiai naudojami prižiūrimo mokymosi metodai. Tokiems metodams svarbu, kad duomenys turėtų aiškiai apibrėžtą išvestį, būtų padalinti į mokymosi ir testavimo imtis.

Prižiūrimo mokymosi atvejais svarbu, kad algoritmui pateikiama mokymosi duomenų imtis būtų sužymėti (arba su etiketėmis). Taigi algoritmai yra mokomi su duomenimis, kur aiškiai apibrėžti tiek įvesties, tiek išvesties parametrai. [58]

2.1. Logistinė regresija

Logistinės regresijos modelis yra bene pats populiariausias pasirinkimas, norint apskaičiuoti ar paskolos gavėjas nustos mokėti įsipareigojimą, ar vykdys jį laiku. [59] Tikriausiai dėl to, jog tai yra metodas, kurį labai lengva realizuoti ir kuris pasiekia labai gerą našumą tiesiškai atskirtomis klasėmis. Šis metodas yra vienas iš esminių metodų. Svarbiausias jo privalumas yra tai, kad jis gali būti naudojamas tiek klasifikavimui, tiek klasių tikimybei įvertinti, nes yra susietas su logistinių duomenų paskirstymu. [60]

Logistinė regresija sprendžia klasifikavimo problemas ir padeda nustatyti į kurią klasę geriausia priskirti naują objektą. [61] Regresijos tikslas – rasti tinkamiausią modelį, kuris paaiškintų ryšį tarp priklausomų kintamųjų ir kai kurių regresorių (požymių). [62]

Dvinarės logistinės regresijos modelis yra geras tuo, kad priklausomas kintamasis gali įgyti tik dvi reikšmes – 0 arba 1. Įgyjamos reikšmės priklauso nuo atliekamo tyrimo. O šiuo atveju, laikome, kad 1 reiškia, jog paskolos gavėjas nevykdo įsipareigojimo.

Logistinės regresijos formulė:

$$\frac{P(Y_i=1)}{P(Y_i=0)} = log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k, \quad (1)$$

kur:

$p_i \in [0; 1]$ – tikimybė, kad priklausomas kintamasis įgyja reikšmę 1;

$x_1, x_2, \dots, x_k$ – regresoriai.

Logistinės regresijos privalumai:

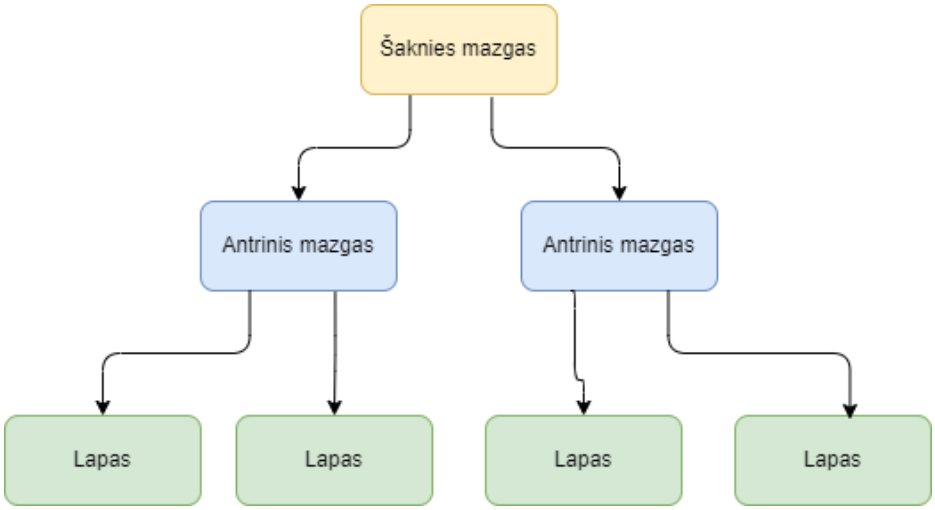
- nesudėtingas algoritmo diegimas ir interpretavimas;
- greitai apdoroja didelius duomenų bei regresorių kiekius;
- geba apdoroti tiek diskrečius, tiek tolydžius duomenis.

2.2. Sprendimų medžiai

Sprendimų medžiai išsiskiria kaip vienas iš labiausiai mėgstamų interpretuojamo mašininio mokymosi metodų. Be atskiro metodo panaudojimo, sprendimų medžiai sudaro pagrindą daugeliui sudėtingesnių mašininio mokymosi algoritmų, įskaitant tokius, kaip atsitiktiniai miškai.

Vėlyvais 1970–aisiais, ankstyvais 1980–aisiais mašininio mokymosi tyrėjas J. R. Quinlan'as sukūrė sprendimų medžių algoritmą, geriau žinomą ID3 pavadinimu. Tyrinėdamas toliau, autorius patobulino algoritmą, sukurdamas naują, pavadinimu C4.5. Tačiau beveik tuo pačiu metu, 1984 m., grupė statistikų, L. Breiman'as, J. Friedman'as, R. Olshen'as ir C. Stone'as, išleido knygą „Klasifikavimas ir regresijos medžiai“ (*Classification and Regression Trees CART*), kurioje kruopščiai aprašė dvinarių sprendimų medžių vystymąsį. Ir nors ID3 bei CART algoritmai buvo sukurti atskirai, jie yra gana panašūs savo veikimo principu. [63]

Nepaisant to, kad spendimų medžiai turi ne vieną išreiškimo galimybę, dažniausiai sutinkamas – dvejetainis sprendimų medžių tipas. Dvejetainiai sprendimų medžiai konstruojami iš viršaus į apačią, kur pagrindinis (šaknies) mazgas, atvaizduojantis visą mokomąjį duomenų rinkinį, turi tiksliai dvi šakas, vedančias į antrinius mazgus. Taip stebėjimai nukreipiami į kairįjį arba dešinįjį antrinį mazgą pagal šakojimosi taisyklę. Mazgai, kurie nebeturi šakų, vadinami lapais. O kiekvienas lapas priskiriamas klasei k, kad bet koks stebėjimas, nukreiptas į tą lapą, būtų klasifikuojamas kaip priklausantis klasei k. [64] Sprendimų medžio struktūra pateikta žemiau esančiame paveiksle.



3 pav. Sprendimų medžio struktūra

Vienas iš galimų scenarijų kuomet algoritmas sustoja – kai antrinių mazgų daugiau nebepavyksta padalinti. Tai reiškia, kad dalinant antrinį mazgą abu nauji antriniai mazgai yra priskiriami tokioms pačioms klasėm. Taip pat gali pasitaikyti, jog imtyje nebeliko laisvų atributų, kuriais galima būtų skaidyti antrinius mazgus, nes visi jie panaudoti ankstesniuose sub–medžio antriniuose mazguose. Tačiau egzistuoja ir tokia galimybė, kai paprasčiausiai duomenų rinkinyje nebelieka duomenų. Visais atvejais paskutinis antrinis mazgas yra paverčiamas lapu. [62]
Nors sprendimų medžiai gali apdoroti didelius atributų bei duomenų kiekius, tikslas yra kaip įmanoma teisingiau sukurti homogeniškus duomenų pogrupius. Ypač svarbu teisingai pasirinkti atributus, įvertinant skirtingų atributų naudingumą skaidant duomenų rinkinį. O tokiai užduočiai atlikti pasitarnauja atributų atrankos priemonės, kurių šis algoritmas turi ne vieną.

Pirmasis metodas – entropija. Entropija – tai neapibrėžtumo arba grynumo matas, kur kuo aukštesnis matas, tuo grynumas yra žemesnis. Matas apskaičiuojamas taip:

$$E(S) = \sum_{i=1}^{c} -P_i \log_2(P_i), \quad (2)$$

kur *c* yra klasių skaičius, P_i yra nenulinė klasės imtyje tikimybė.

Antrasis yra informacijos išlošio metodas. Konstruojant sprendimų medį svarbi užduotis pirmiausia parinkti tokį atributą, kuris turi kuo didesnį informacijos išlošį ir kuo mažesnę entropiją. Informacijos išlošis parodo kaip gerai atributas padalina mokymosi duomenų imtį į duomenims priskirtas klases. Rodiklis apskaičiuojamas taip:

$$IG(X,Y) = E(Y) - E(Y,X) \quad (3)$$

Paskutinis metodas yra Gini indeksas. Priešingai informacijos išlošio rodikliui, Gini indeksas pirmenybę teikia didesniems duomenų skaidiniams. O turint nepriekaištingai klasifikuotą pavyzdį, Gini indeksas bus lygus 0. Šio rodiklio formulė atrodo taip:

$$Gini = 1 - \sum_{i=1}^{c}(P_i)^2 = 1 - (P(class\ A)^2 + P(class\ B)^2 + \cdots + P(class\ N)^2), \quad (4)$$

Kur P_i tikimybė, kad elementas bus priskirtas tam tikrai klasei. [65]

Sprendimų medžių privalumai:

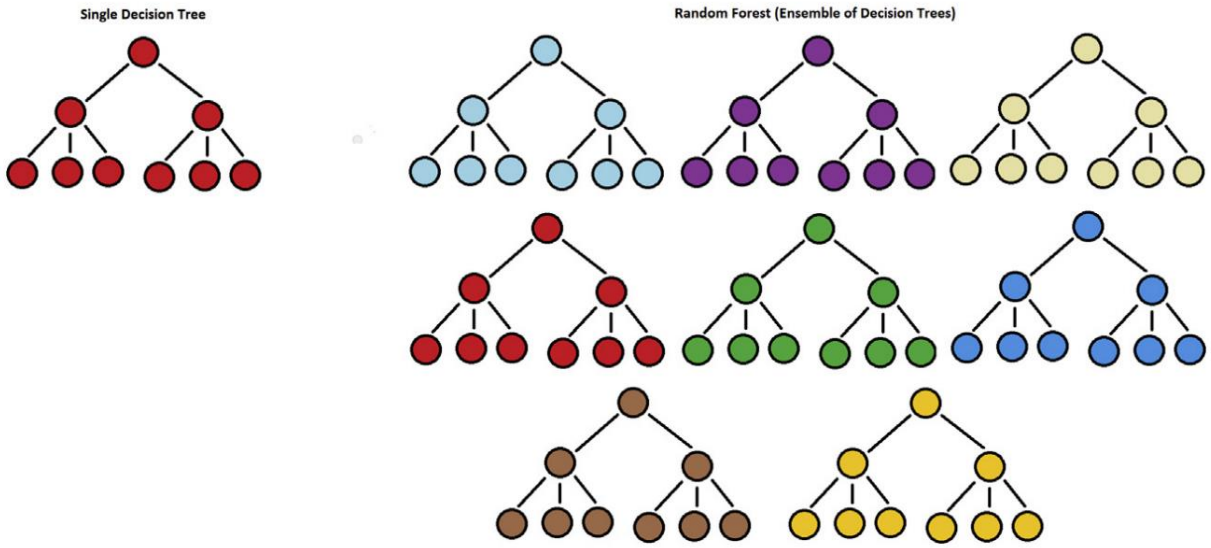
- algoritmo struktūrą lengvai galima vizualizuoti. Ji yra lengvai suprantama bei interpretuojama;
- išskirtys nedaro didelės įtakos algoritmo rezultatams;
- geba susitvarkyti su kategoriniais kintamaisiais;
- reikalingas minimalus duomenų paruošimas;
- algoritmas nereikalauja duomenų normalizavimo mastelio keitimo.

2.3. Atsitiktinio miško metodas

Atsitiktinio miško metodas yra grįstas klasifikavimo ir sprendimų medžių naudojimu. Atsitiktinis miškas konstruojamas naudojant daugybę vieną nuo kito nepriklausančių sprendimų medžių. Medžiams priskiriami atsitiktiniai požymiai, o rezultatas yra klasė (arba vidurkis), kurią priskyrė dauguma atsitiktinio miško medžių atskirai.

Daugybės sprendimų medžių apjungimas į vieną modelį yra atsitiktinio miško pagrindas. Prognozė, gauta iš vieno sprendimų medžio, gali būti netiksli. Tačiau sujungus daug medžių, prognozė bus artima vidurkio reikšmei. Atsitiktinio miško metodas dažniausiai yra tikslesnis nei sprendimų medžių metodas, nes iš daugiau prognozių įgyjama daug daugiau informacijos.

Sprendžiant regresijos užduotį, atsitiktiniai miškai panaudoja sprendimų medžių vidurkį savo galutinei prognozei. Tačiau, kaip jau žinoma, atsitiktinio miško metodas geba spręsti ir klasifikavimo problemas. O prognozuojamą klasę priskiria naudojantis daugumos balsais. [65]



4 pav. Sprendimų medžio ir atsitiktinio miško pavyzdys [65]

Atsitiktinių miškų privalumai:

- metodas sugeba apdoroti dideles duomenų imtis ir geba pasiekti didelį tikslumą lyginant su kitais algoritmais;
- metodas sėkmingai susitvarko su klasių disbalansu duomenų rinkiniuose;
- sugeba susitvarkyti su trūkstamais duomenimis ir pasiekia didelį tikslumą net tuo atveju, jei imtyje yra trūkstama didelė dalis duomenų; [32]
- puikiai prisitaiko prie įvairių duomenų tipų. Geba apdoroti tiek diskrečius, tiek tolydžius duomenis;
- metodas atsparus išskirtims, nes atsitiktinėmis dalimis renkasi duomenis ir iš jų konstruoja medžius. Taigi, net jeigu keletas medžių yra netikslūs dėl išskirčių daromos įtakos, prognozės yra grindžiamos daugybės medžių rezultatais, kas sumažina išskirčių įtaką ir padidina patikimumą. [66]

2.4. Extreme Gradient Boosting (XGBoost)

XGBoost metodas buvo sukurtas mažiau nei prieš dešimtmetį. 2016 m. T. Chen'as sukūrė iteracinį sprendimų medžių algoritmą su keletu sprendimų medžių. Kur kiekvienas sprendimų medis yra apmokomas iš liekanų, kurios lieka nuo praeitų sprendimų medžių. Metodas visai skiriasi nuo atsitiktinių miškų metodų, kur rezultatas būna nugalėtojas gavęs, daugumos balsus. XGBoost algoritmo išvestis yra rezultatų suma:

$$\hat{y}_i = \sum_{k=1}^{n} f_k(x_i), f_k \in F, \tag{5}$$

kur F reiškia regresijos medžių erdvę, f_k reiškia medį, tuomet $f_k(x_i)$ yra k–tojo medžio rezultatas, o $\hat{y}_i$ prognozuojama i–tojo elemento x_i reikšmė.

XGBoost optimizavimo uždavinio tikslo funkcija yra:

$$Obj(\theta) = L(\theta) + \Omega(\theta), \quad (6)$$

kur $L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i)$ yra nuostolio funkcija, $\hat{y}_i$ prognozė, y_i tikslas, o $\Omega(\theta) = \sum_{k=1}^{K} \Omega(f_k)$ yra bauda už modelio sudėtingumą.

Tuomet modelis yra apmokamas adityviniu būdu. Laikome, kad $\hat{y}_i$ yra prognozuojama *i*–tojo elemento prognozė ties *t*–taja iteracija, tuomet $\hat{y}_i$ gali būti išreiškiamas:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i). \quad (7)$$

Šioje situacijoje, jis minimizuoja tikslo funkciją:

$$Obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t). \quad (8)$$

Toliau naudojama antros eilės aproksimacija, siekiant greitai optimizuoti:

$$Obj^{(t)} = \sum_{i=1}^{n} \left(l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right) + \Omega(f_t), \quad (9)$$

kur g_i ir h_i yra nuostolio funkcijos, atitinkamai pirmos ir antros eilės, gradiento statistika. [67]

XGBoost metodo privalumai:

- XGBoost algoritmas yra ypatingo tikslumo ir labai dažnai pasiekia didesnį tikslumą nei kiti mašininio mokymo algoritmai;
- modelis sėkmingai susitvarko su netiesinėmis priklausomybėmis duomenyse;
- algoritmas savyje turi integruotas L1 (Lasso) ir L2 (Ridge) reguliarizacijas, kurios reikalingos norint išvengti modelio persimokymo;
- sėkmingai susitvarko su trūkstamais duomenimis imtyse;
- labai efektyvus dideliems duomenų rinkiniams, nes palaiko lygiagretųjį apdorojimą.
- metodas yra ypač našus, nes geba tinkamai panaudoti kompiuterio atmintį. [68]

2.5. Light Gradient Boosting Machine (LightGBM)

2017 m. LightGBM metodą pirmą kartą pasiūlė „Microsoft“ tyrėjų komanda. Jie siekė patobulinti labai populiarų XGBoost metodą. Tyrėjų tikslas buvo sukurti greitai besimokantį algoritmą su dar didesniu tikslumu ir dar mažesniu kompiuterio atminties panaudojimu, kuris būtų sunkiai persimokantis. Taigi šis algoritmas yra panašus į XGBoost metodą ir taip pat naudoja sprendimų medžių ansamblio modelį. Modelio esmė: sulig kiekviena iteracija neigiamas nuostolio funkcijos gradientas yra naudojamas kaip aproksimacija esamam sprendimų medžiui ir taip sukuriamas naujas sprendimų medis. [69]

Metodas yra plačiai naudojamas sprendžiant regresijos ar klasifikavimo problemas. Jis sujungia kelis silpnus algoritmus, sukurdamas vieną stiprų mokymosi modelį. Sustiprinimo algoritmai didina neteisingai klasifikuotų elementų svorius bei mažina teisingai klasifikuotų elementų svorius tam, kad neteisingai klasifikuotiems elementas būtų skiriamas didesnis dėmesys kitame mokymosi etape. Finale visi mašininio mokymosi modeliai yra tiesiškai derinami, o kombinuoto modelio svoriai koreguojami pagal klasifikatoriaus klaidų koeficientą. Pagrindinę mintį galima pateikti šia lygtimi: [70]

$$f(x) = \sum_{q=1}^{Q} \alpha_q T(x, \theta_q), \quad (10)$$

kur *f(x)* mokymo pavyzdį atitinkanti tikslinė vertė, *Q* yra pradinių besimokančiųjų skaičius, α_q yra *q*–tojo pagrindinio besimokančiojo svorio koeficientas, *x* yra mokymo pavyzdys, θ_q yra besimokančiojo klasifikavimo parametras, o $T(x, \theta_q)$ yra *q*–tasis bazinis mokinys, dalyvaujantis mokyme.

Apsibrėžus modelio nuostolio funkciją ir mokymosi imtį, mokymosi procesas keičiamas optimizavimo problema, siekiant minimizuoti nuostolio funkciją. Tuomet tikslo funkcija atrodo taip:

$$\arg min \sum_{h=1}^{H} L(y_h, f(x_h)), \quad (11)$$

kur *H* mėginių skaičius, *h* yra imties indeksas, y_h yra tikroji duomenų vertė $f(x_h)$ yra tikslinė vertė, atitinkanti *h*–tąjį imties elementą, o $L(y_h, f(x_h))$ yra *h*–tojo elemento nuostolio funkcijos reikšmė.

LightGBM metodas taiko dvi pagrindines ypatybes:

1) Leaf–wise medžių auginimo metodą, kuris remiasi lapų auginimo strategija, bet ne vienodo lygio auginimo strategija. Todėl metodas užtikrina efektyvų skaičiavimą ir išvengia persimokymo, užtikrindamas minimalų lapų duomenų kiekį ir sprendimų medžio gylį.

2) Histograma grįstą sprendimų medžio algoritmą, kuris renkantis atributus pereina ir randa optimalų padalijimo tašką. Taip optimizuojamos atminties ir skaičiavimo sąnaudos. Todėl metodas treniruojasi itin efektyviai, kai yra dirbama su dideliais duomenų masyvais. [70]

LightGBM metodo privalumai:

- dėl naudojamos leaf–wise technikos metodas yra žinomas dėl savo greito apmokymo;
- dažnai pasiekia geresnį tikslumą, nei bet kurie kiti mašininio mokymosi algoritmai;
- geba naudoti labai mažą kiekį kompiuterio atminties;
- be didelių problemų susitvarko su ypač dideliais duomenų srautais. [71]

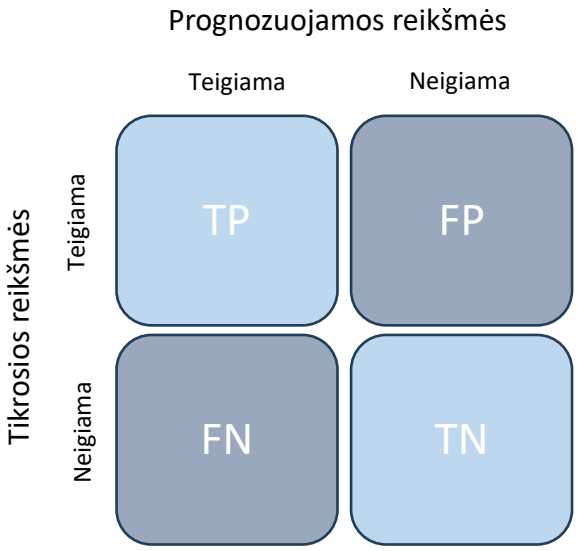
2.6. Modelių įvertinimo technikos

2.6.1. Sumaišymo matrica

Sumaišymo matrica naudojama siekiant įvertinti kiek mašininio mokymo algoritmas teisingai elementams priskyrė klases, naudojant klasifikavimo algoritmus. O taip pat galime sakyti, kad sumaišymo matrica yra teisingai priskirtų ir klaidingai priskirtų klasių suvestinė.

Sumaišymo matricoje stulpeliuose pateiktos tikros elementų klasės, o eilutėse priskirtos klasės, kur:

- TP (angl. *True Positive*): reikšmė buvo klasifikuojama kaip teigiama, o modelis taip pat priskyrė teigiamą klasę.
- FP (angl. *False Positive*): reikšmė klasifikuojama kaip neigiama, tačiau modelis priskyrė teigiamą klasę. Ši klaida vadinama I rūšies klaida.
- FN (angl. *False Negative*): reikšmė klasifikuojama kaip teigiama, o modelis priskyrė neigiamą klasę. Ši klaida vadinama II rūšies klaida.
- TN (angl. *True Negative*): reikšmė klasifikuojama kaip neigiama, o modelis taip pat priskyrė neigiamą reikšmę.



5 pav. Sumaišymo matricos vaizdavimas

Iš sumaišymo matricos gali būti apskaičiuojami įvairūs rodikliai. [72]

Tikslumo rodiklis apskaičiuoja kokia elementų dalis, procentiškai, buvo klasifikuota teisingai. Kuo didesnė tikslumo reikšmė, tuo modelis geba tiksliau klasifikuoti duomenis. Tačiau tikslumas nėra pati geriausia metrika, jei duomenų klasės yra nesubalansuotos.

$$Tikslumas = accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{12}$$

Jautrumas parodo procentinę dalį teigiamų elementų, kurie buvo teisingai klasifikuoti. Reikšmė svyruoja tarp 0 ir 1. Jei reikšmė yra artimesnė 1, tai parodo, kad modelis geba teisingai identifikuoti 1 klasės reikšmes.

$$Jautrumas = sensitivity = \frac{TP}{TP+FN} \tag{13}$$

Specifiškumas parodo kokia procentinė dalis neigiamų elementų buvo klasifikuota teisingai. Reikšmė svyruoja tarp 0 ir 1. Jei reikšmė yra artimesnė 1, tai parodo, kad modelis geba teisingai identifikuoti 0 klasės reikšmes.

$$Specifiškumas = specificity = \frac{TN}{TN+FN} \tag{14}$$

Preciziškumas parodo kokia procentinė dalis teigiamų elementų buvo klasifikuota teisingai. Reikšmė svyruoja tarp 0 ir 1. Jei reikšmė yra artimesnė 1, tai parodo, kad modelis geba teisingai identifikuoti visus duomenis kaip teigiamą klasę.

$$Preciziškumas = precision = \frac{TP}{TP+FP} \tag{15}$$

Klaidingo klasifikavimo rodiklis parodo kokia dalis elementų buvo klasifikuota neteisingai. Kuo mažesnė rodiklio reikšmė, tuo modelis geba tiksliau klasifikuoti duomenis.

$$Klaidos\ rodiklis = error\ rate = \frac{FP+FN}{TP+TN+FP+FN} = 1 - accuracy \tag{16}$$

F1 metrika iš esmės apjungia jautrumo ir preciziškumo metrikas ir pateikia jų subalansuotą įvertį. F1 metrika apskaičiuota vertinant harmoninį šių dviejų metrikų vidurkį. Šis įvertis geras tuo, kad leidžia įvertinti klasifikavimo tikslumą, duomenų imtyje turint nesubalansuotas klases.

$$F1\ metrika = F1 - score = \frac{2 \times Preciziškumas \times Jautrumas}{Preciziškumas + Jautrumas} \quad (17)$$

2.6.2. ROC kreivė

ROC kreivės yra puiki technika, siekiant vizualiai įvertinti klasifikatoriaus efektyvumą. Iš esmės, kreivė atvaizduoja ryšį tarp teisingų teigiamų klasifikacijų ir klaidingų teigiamų klasifikacijų. Arba kitaip tariant, atvaizduoja ryšį tarp jautrumo ir specifiškumo matų.

ROC grafikai yra dvimačiai, kur:

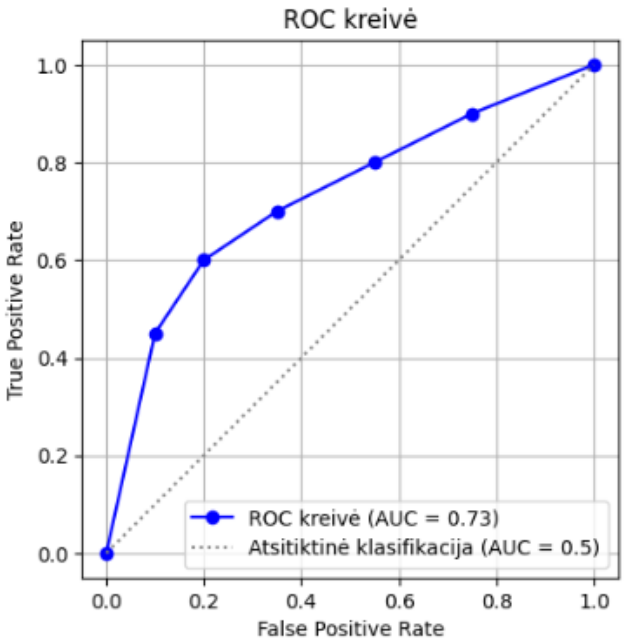
- *y* ašyje atvaizduojamas teisingų teigiamų atvejų rodiklis (angl. *true positive rate*). Rodiklio formulė atitinka jau anksčiau aptarto jautrumo formulę:

$$TPR = Jautrumas = \frac{TP}{TP+FN} \quad (18)$$

- *x* ašyje atvaizduojamas klaidingai teigiamų atvejų rodiklis (angl. *false positive rate*). Rodiklį galima apskaičiuoti neigiamus atvejus, kurie buvo klaidingai klasifikuoti kaip teigiami atvejai padalinus iš visų neigiamų atvejų:

$$FPR = \frac{FP}{FP+TN} \quad (19)$$

ROC kreivės yra geros tuo, jog jų pagalba labai lengva lyginti skirtingus klasifikatorius pagal jų veikimą ir visa tai pateikiama vizualiai – lengva ir aiškia forma. O siekiant ROC kreivę paversti skaitiniu rodikliu, naudojamas AUC (angl. *area under the ROC curve*) matas. Jis apskaičiuoja koks plotas yra po ROC kreive. Kadangi grafikas yra vienetinio kvadrato, nuo (0,0) iki (1,1), AUC reikšmė irgi bus tarp 0 ir 1. [73]



6 pav. ROC kreivės pavyzdys

Taigi, AUC matas yra tarsi ROC kreivės apibendrinimas, kur:

- AUC = 0,5 rodo atsitiktinį klasifikavimą. AUC negali įgyti mažesnės reikšmės.
- AUC = 1,0 rodo puikų klasifikavimą. Tačiau pasiekus tokį rezultatą būtina kritiškai įvertinti ar modelis nėra persimokęs.

3. Rezultatai

Šioje projekto dalyje aptariamas pasirinktas duomenų rinkinys bei pritaikyti matematiniai metodai, klientų kreditingumui įvertinti. Taip pat aptariami gauti rezultatai, modelių rezultatų įvertinimai, pateikiami patarimai.

3.1. Naudota programinė įranga

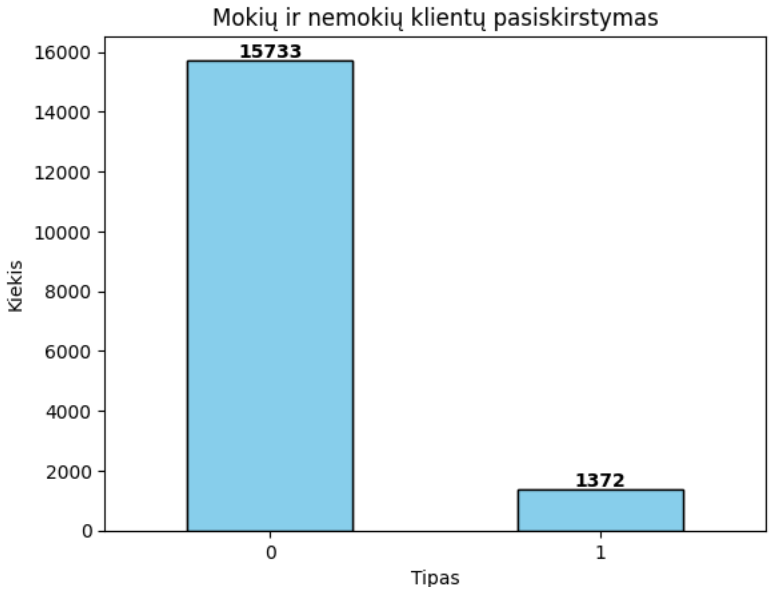
Tyrimui pasitelkta „Python“ programavimo kalba, 3.11.9 versija. Kodo rašymui naudota atvirojo kodo programa Visual Studio Code, 1.99.3 versija. Duomenų analizės ir prognozavimo užduotims atlikti buvo naudotos šios bibliotekos:

- *Scikit–learn* – mašininio mokymosi biblioteka, naudojama statistiniam modeliavimui, ar klasifikavimo, regresijos, klasterizavimo bei dimensijos mažinimo užduotims atlikti. Darbe biblioteka naudota logistinės regresijos, sprendimų medžio ir atsitiktinių medžių metodams.
- *LighGBM* – mašininio mokymosi biblioteka, naudota LightGBM metodui.
- *XGBoost* – mašininio mokymosi biblioteka, naudota XGBoost metodui.

3.2. Duomenų žvalgomoji analizė

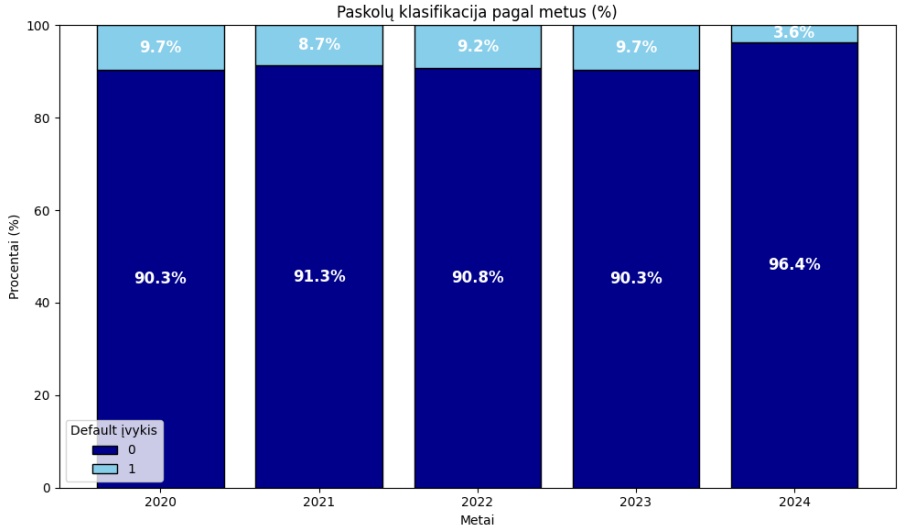
Tyrimui naudoti vienos, Lietuvoje veikiančios, tarpusavio skolinimo platformos fizinių asmenų paskolų duomenys. Duomenų imtį sudaro paskolų, išduotų per 4,5 metų laikotarpį, informacija. Naudoti duomenys yra įvairiapusiai, įskaitant skolininko asmeninę informaciją (amžių, lytį ir kt.), paskolos atributų informaciją (paskolos sumą, paskolos palūkanų normą ir kt.), o taip pat ir skolininko kredito istorijos duomenis (turimų įsipareigojimų, skolų informaciją ir kt.). Duomenys apima 17 105 unikalių paskolų.

Sprendžiant klasifikavimo uždavinį, priklausomu kintamuoju pasirinktas mokumas. Šiuo atveju, klientas laikomas nemokiu, jei paskolos mokėjimo metu su fiziniu asmeniu buvo nutraukta pasirašyta sutartis. Mokios paskolos (su požymiu 0) yra 15 733 atvejais, o nemokios paskolos (su požymiu 1) apima 1 372 atvejus. Matoma, kad priklausomas kintamasis yra labai nesubalansuotas – 91,98 % mokių paskolų ir 8,02% nemokių paskolų.



7 pav. Priklausomo kintamojo klasių pasiskirstymo atvaizdavimas

Duomenyse yra pilni 2020 – 2023 metų duomenys. O 2024 metų duomenys yra tik pusės metų, nes sutartys įprastai nutraukiamos tik po 3–6 įmokų vėlavimo. Tačiau reikėtų atkreipti dėmesį, kad ne visos 2024 metų sutartys jau yra nutrauktos. Dalis tarpusavio skolinimo platformos paskolų šiuo metu moka įmokas laiku ir tik ateityje susidurs su mokumo problemomis, o dalis nebemoka įmokų laiku, bet dar per anksti nutraukti sutartis ir yra vykdomos kitos išieškojimo procedūros.



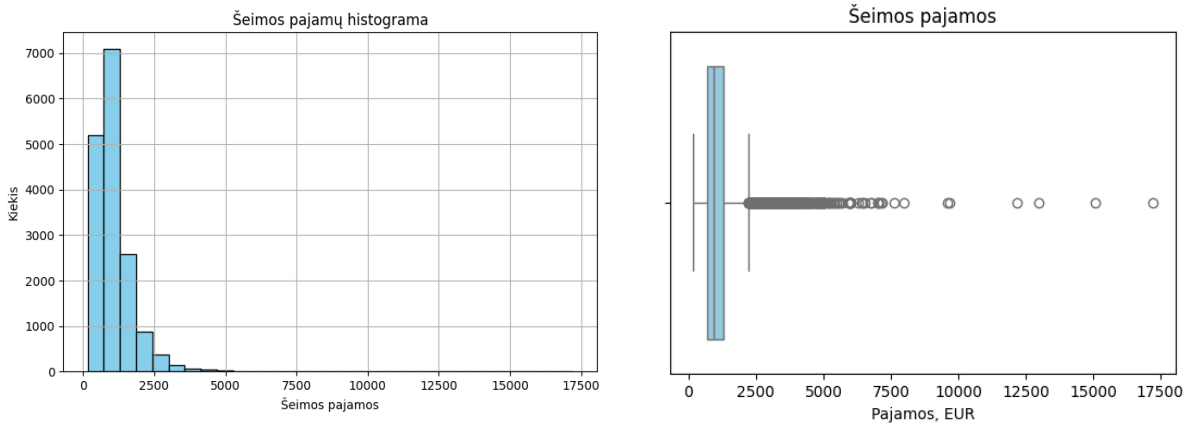
8 pav. Priklausomo kintamojo klasių pasiskirstymo atvaizdavimas grupuojant pagal metus

3.3. Duomenų paruošimas

Duomenų paruošimą modeliavimo algoritmams galima būtų išskirstyti į keletą etapų. Kurie apimtų, kategorinių kintamųjų identifikavimą ir paruošimą, tolydžiųjų kintamųjų identifikavimą ir paruošimą, tuščių reikšmių problemų sprendimą, multikolinearumo problemą ir kt.

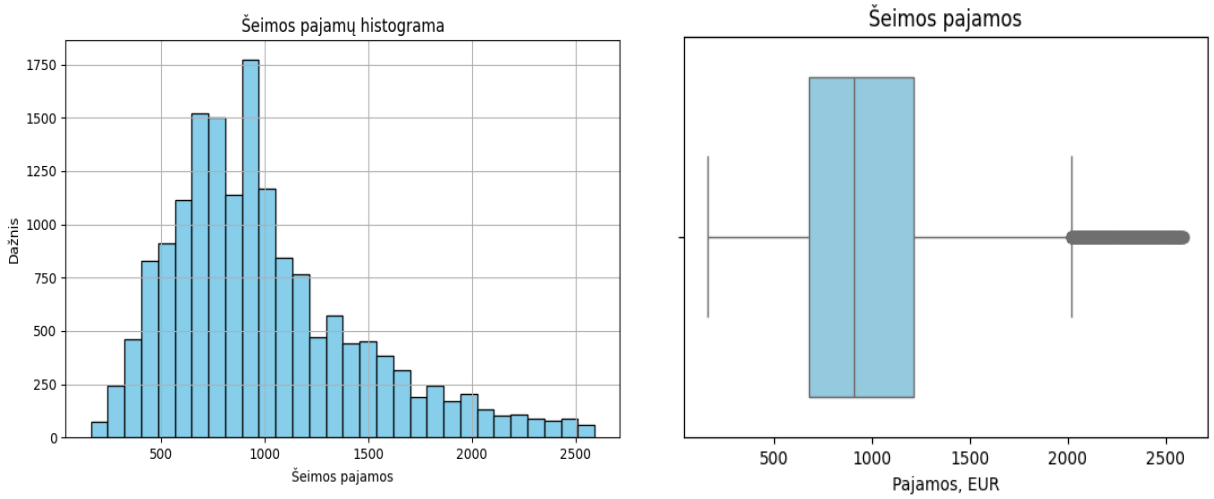
Pirmiausia, žinoma, nei vienas modelis neveiks teisingai, jeigu duomenyse bus tuščių reikšmių. Patikrinus, paaiškėjo, kad tokių įrašų nėra daug, tad visos duomenų eilutės buvo pašalintos iš imties. Taip pat sutvarkyti trupmeniniai skaičiai, panaikinti identifikaciniai (ID) ir tuščių reikšmių stulpeliai.

Toliau analizuoti tolydieji kintamieji. Pašalintos išskirtys, pakeičiant jas medianos reikšmėmis. Žemiau pateiktame paveikslėlyje pavaizduotas šeimos pajamų kintamojo pasiskirstymas (histogramoje kairėje), o dešinėje esančioje stačiakampėje diagramoje pavaizduotas pasiskirstymas apink vidurkį bei išskirtys.



9 pav. Šeimos pajamų kintamasis prieš išskirčių panaikinimą

Pastebima, jog po išskirčių panaikinimo duomenyse vis dar yra išskirčių. Tačiau daugiau išskirčių šalinimo procesas nebebus kartojamas, siekiant per daug nesusintetinti turimo kintamojo.



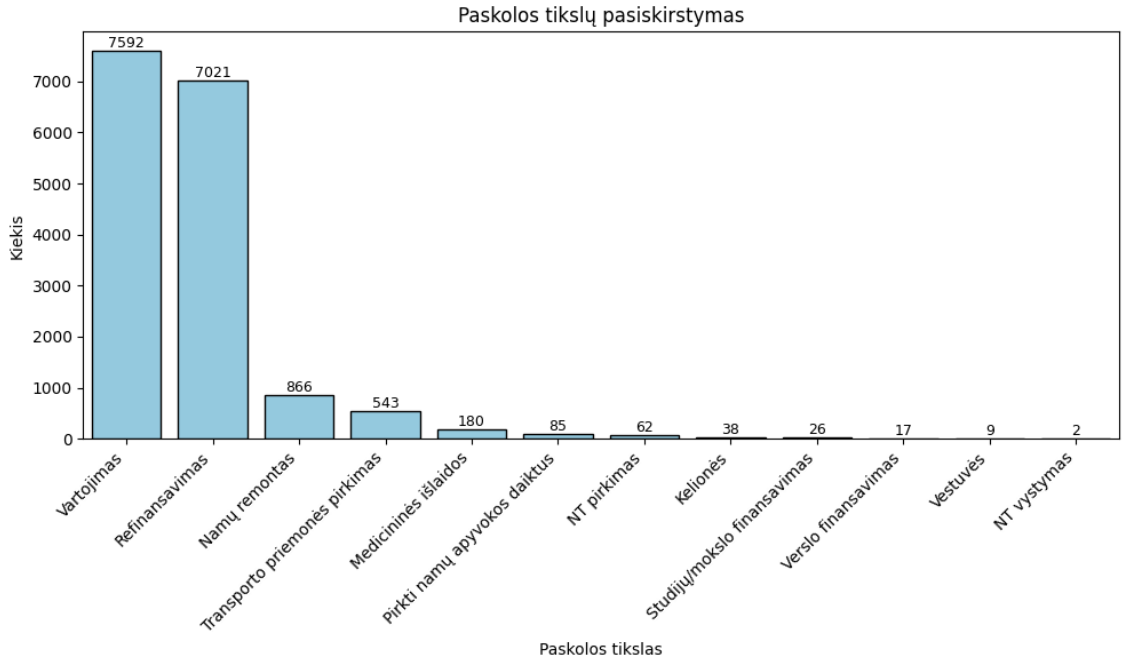
10 pav. Šeimos pajamų kintamasis po išskirčių panaikinimo

Sekanti užduotis buvo susitvarkyti su kategoriniais kintamaisiais. Jų imtyje buvo daugiausiai. Kintamieji, kurie turėjo tik dvi reikšmes, buvo koduojami 0, 1. Tačiau kintamieji, kurie turėjo daugiau nei 2 galimas reikšmes, buvo skaidomi į atskirus kintamuosius. Nauji kintamieji gali įgyti tik dvi reikšmes – 0 arba 1. Žemiau pateiktoje lentelėje yra kintamojo nt_nuosavybe išskaidymo į keletą stulpelių pavyzdys.

3 lentelė. Kategorinio kintamojo su daugiau nei 2 kategorijomis dekodavimo pavyzdys

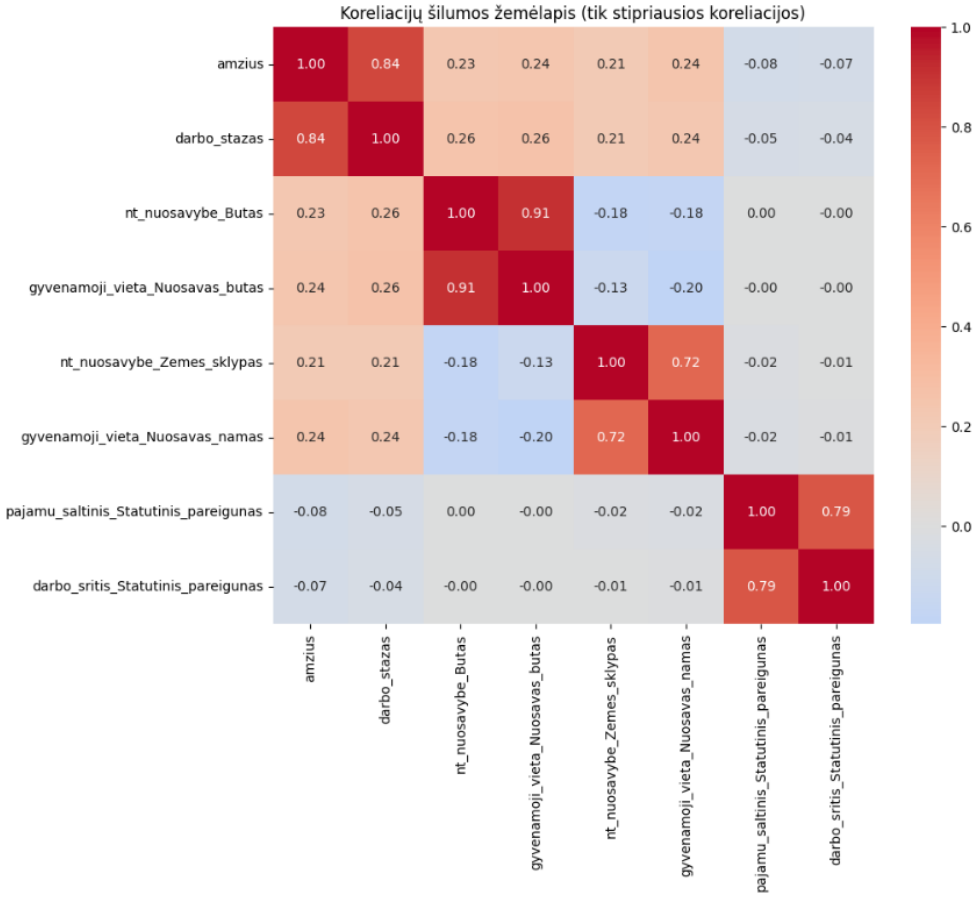
Stulpelio nt_nuosavybe galimos reikšmės	**Naujo stulpelio pavadinimas**	**Galimos reikšmės naujame stulpelyje**
Neturi	nt_nuosavybe_Neturi	0/1
Butas	nt_nuosavybe_Butas	0/1
Kita	nt_nuosavybe_Kita	0/1
Namas	nt_nuosavybe_Namas	0/1
Zemes_sklypas	nt_nuosavybe_Zemes_sklypas	0/1

Pastebėta, kad kai kurie kintamieji negali būti skaidomi tokiu būdu. Pavyzdžiui, kintamasis paskolos_tikslas gali įgyti net 12 skirtingų reikšmių. Tačiau išskaidžius šį kintamąjį didžiąją dalį naujų kintamųjų reikėtų panaikinti, kaip duomenų triukšmą, paliekant tik 5–6 kintamuosius. O taip pat kyla ir kita problema. Išskaidžius į atskirus stulpelius, stulpelis paskolos_tikslas_Vartojimas ir stulpelis paskolos_tikslas_Refinansavimas pradeda tarpusavyje labai stipriai koreliuoti ir kai kuriems algoritmams tai daro didelę neigiamą įtaką. Tad siekiant panaikinti multikolinearumo problemą, būtina pašalinti vieną kintamąjį. O tai reiškia labai didelį informacijos praradimą. Todėl šį stulpelį rekomenduotina koduoti skaitinėmis reikšmėmis: 0, 1, 2, ..., 10, 11.



11 pav. Kintamojo paskolos_tikslas pasiskirstymas pagal galimas įgyti reikšmes

Kai kurie algoritmai negeba tvarkytis su stipriomis koreliacijomis tarp kintamųjų. Tad kitu etapu įvertiname koreliacijas tarp kintamųjų ir panaikiname multikolinearumo problemą. Žemiau esančiame paveiksle pavaizduotas koreliacijų šilumos žemėlapis, vaizduojantis tik pačias stipriausias koreliacijas, kurios yra didesnės už 0.7 arba mažesnės už –0.7.



12 pav. Pačios stipriausios koreliacijos tarp kintamųjų

Abiejų koreliuojančių kintamųjų pašalinti nebūtina. Iš imties išimama tik po vieną iš stipriai koreliuojančios poros.

Taip pat pastebėta, kad nors kintamieji nebesusiduria su stipriomis tarpusavio koreliacijomis, logistinės regresijos algoritmas nesusitvarko su duomenimis. Taip gali būti dėl to, kad duomenyse vis dar liko stiprių tiesinių priklausomybių. Todėl siekiant patikrinti priklausomybių egzistavimą tarp kintamųjų, buvo pasitelktas dar vienas matas – dispersijos padidėjimo daugiklis VIF (angl. *variance inflation factor*). Daugiklis iš esmės parodo kiek koeficientų dispersija padidėja dėl kintamųjų tarpusavio koreliacijos. VIF formulė:

$$VIF_i = \frac{1}{1-R_i^2} \tag{20}$$

kur R_i^2 – yra determinacijos koeficientas, kuris rodo kiek stiprus yra tiesinis ryšys tarp vieno nepriklausomo kintamojo lyginant su visais kitais.

VIF reikšmių paaiškinimas:

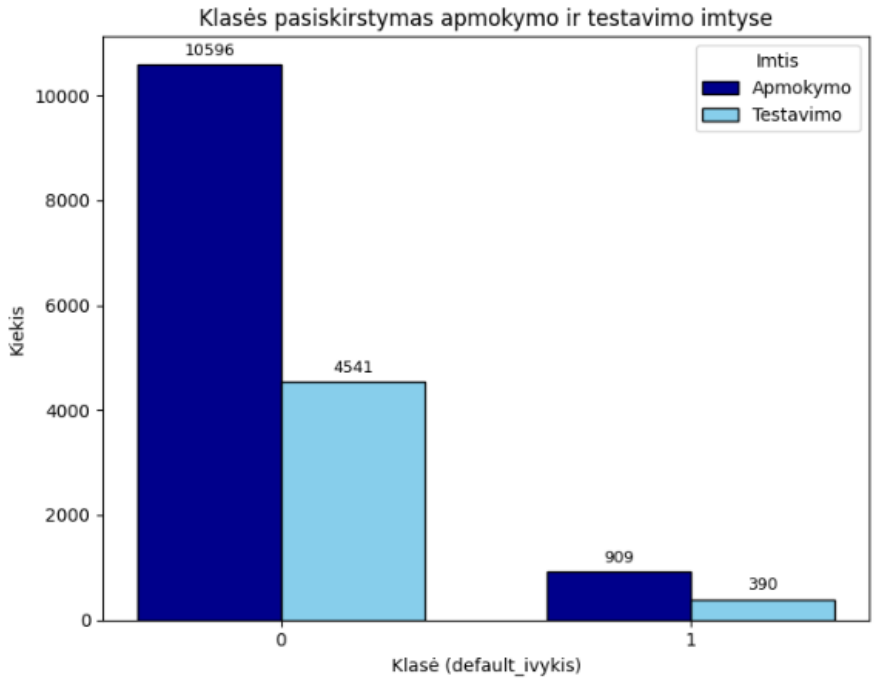
- VIF = 1 – rodo, kad jokio multikolinearumo su tiriamu kintamuoju neidentifikuota,
- VIF = 4 – rodo, kad ryšys yra stipresnis ir gali būti rastas multikolinearumas, tačiau reikalingas detalesnis pasigilinimas,
- VIF > 10 – rodo, kad koeficientas yra labai aukštas ir kitų kintamųjų ryšyje su šiuo kintamuoju yra identifikuojama stipri tiesinė priklausomybė. [74]

Taigi, po šio etapo buvo pašalinti dar keli kintamieji, kurie turėjo stiprų tiesinį tarpusavio ryšį ir būtų galėję trukdyti pasiekti gerų rezultatų su logistinės regresijos modeliu.

3.4. Modelių apmokymas

Modelių apmokymui reikėjo pasidalinti imtį į dvi dalis – testavimo ir apmokymo imtis. Toks padalijimas būtinas siekiant objektyviai įvertinti modelio tikslumą bei jo gebėjimą apibendrinti informaciją apie nematytus duomenis. Apmokymo imtis naudojama modelio parametrų nustatymui, o testavimo – modelio kokybės įvertinimui.

Pasirinkta, kad testavimo imtį sudarys 30 % visos pradinės imties, o apmokymo imtį sudarys 70 % visos imties duomenų. Visos turimų 16 436 duomenų eilutės buvo padalintos į 11 505 eilutes apmokymo imčiai ir 4 931 eilutę testavimo imčiai. Žemiau pateiktame paveiksle atvaizduojamas imčių pasiskirstymas pagal priklausomo kintamojo reikšmes.



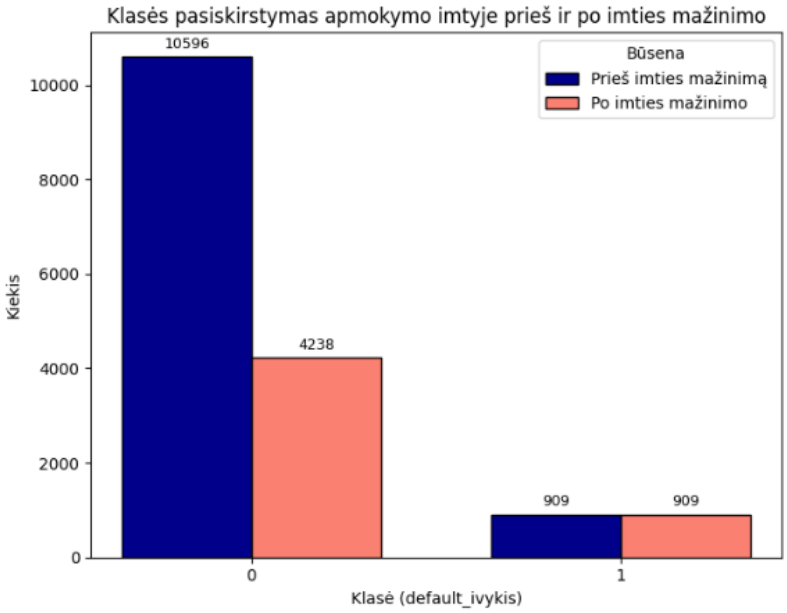
13 pav. Apmokymo ir testavimo imčių pasiskirstymas

Iš pateiktos informacijos matoma, kad 0 ir 1 klasių pasiskirstymas yra labai netolygus. Apmokymo imtyje 1 klasės atvejai sudaro tik 7,9 % visų duomenų eilučių. O tuo tarpu 0 klasės atvejų yra didžioji dauguma, net 92,1 %. Dėl šios priežasties labai sudėtinga teisingai apmokyti algoritmus. Apmokius, jie nesugeba identifikuoti 1 klasės objektų ir dėl to nepasiekiamas aukštas modelio tikslumas.

Apmokymo imtį galima nežymiai pakoreguoti panaudojus imties mažinimo techniką. Jos pagalba pagal pasirinktą koeficientą iš imties bus pašalinta dalis reikšmių. Imties duomenys šalinami atsitiktine tvarka. Svarbu šią užduotį atlikti tik apmokymo imčiai. Testavimo imtis turi būti nepaliesta ir niekaip nekeista.

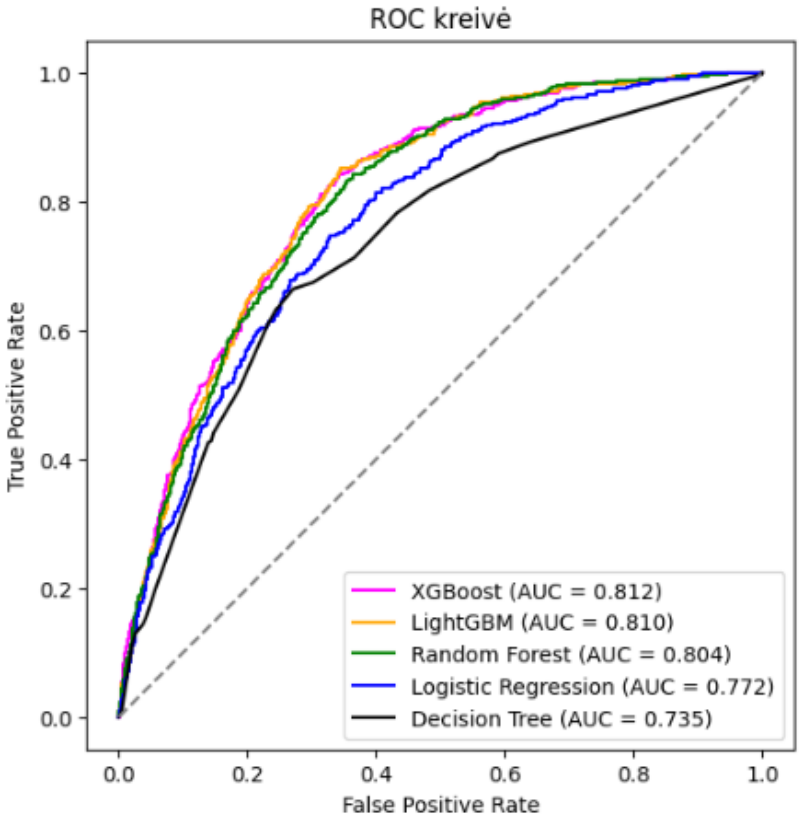
Šiuo atveju, pasirinkta, kad apmokymo imtyje turi likti tik 40% visų nulinės klasės elementų. Tokiu būdu sumažiname apmokymo imties dydį bei galime pasiekti tikslesnių rezultatų. Žemiau pateiktame paveiksle pavaizduota kaip apmokymo imtis atrodė prieš imties mažinimo procedūrą bei po imties mažinimo procedūros atlikimo.

Prieš imties mažinimą 0 klasės elementų buvo 10 596 vienetai, o po imties mažinimo 0 klasės elementų liko 4 238 vienetai. Panaikinta net 60 % visų apmokymo imtyje esančių 0 klasės elementų. Tačiau 1 klasės elementai nebuvo keisti. Taigi, sumažinus apmokymo imtį, nors ir nežymiai, bet buvo pagerintas visų modelių efektyvumas.



14 pav. Apmokymo imtis prieš ir po imties mažinimo procedūros

Remiantis atlikta literatūros analize buvo atsirinkti penki modeliai: logistinės regresijos, sprendimų medžio, atsitiktinio miško, LightGBM ir XGBoost. Toliau visi šie modeliai buvo apmokomi ir testuojami su atsirinktomis tarpusavio skolinimo platformos fizinių asmenų paskolų duomenų imtimis. Kiekvienam modeliui suformuojant sumaišymo matricas, apskaičiuojant AUC metrikas ir nubraižant ROC kreives. Žemiau esančiame paveiksle pateikiama bendra visų analizuotų modelių ROC kreivių ir AUC reikšmių suvestinė.



15 pav. ROC kreivė su galutiniais rezultatais

Iš gautų rezultatų matoma, kad pats geriausias AUC rezultatas buvo pasiektas su XGBoost modeliu. Tačiau antroje vietoje likęs LightGBM metodas atsilieka labai nežymiai. Todėl tikrai galima teigti, jog nors XGBoost metodas šį kartą ir buvo pats tiksliausias, bet abu metodai yra efektyvūs. Tuo pačiu matoma, kad logistinės regresijos metodas nebuvo pats prasčiausias iš visų analizuotų.

4 lentelė. Atrinktų modelių AUC rezultatų suvestinė

Metodas	**Pasiektas AUC**
XGBoost	0,812
LightGBM	0,810
Atsitiktinis miškas	0,804
Logistinė regresija	0,772
Sprendimų medis	0,735

Paanalizuokime XGBoost algoritmo sumaišymo matricą. Modelis teisingai klasifikavo mokius klientus 4 008 atvejais, o nemokius klientus 190 atvejų. Neteisingai klasifikavo mokų klientą kaip nemokų 533 atvejais bei atvirkščiai, 200 atvejų nemokūs klientai buvo palaikyti mokiais.

		Prognozuojama klasė	
		0	**1**
Tikroji klasė	**0**	TN = 4008	FP = 533
	1	FN = 200	TP = 190

Iš šios matricos apsiskaičiuojame skirtingas XGBoost modelio metrikas.

Metrika	**Įvertis**
Tikslumas (angl. *accuracy*)	85,13 %
Jautrumas (angl. *sensitivity/recall*)	48,72 %
Specifiškumas (angl. *specificity*)	88,27 %
Preciziškumas (angl. *precision*)	26,28 %
F1 įvertis (ang. *F1 score*)	33,55 %

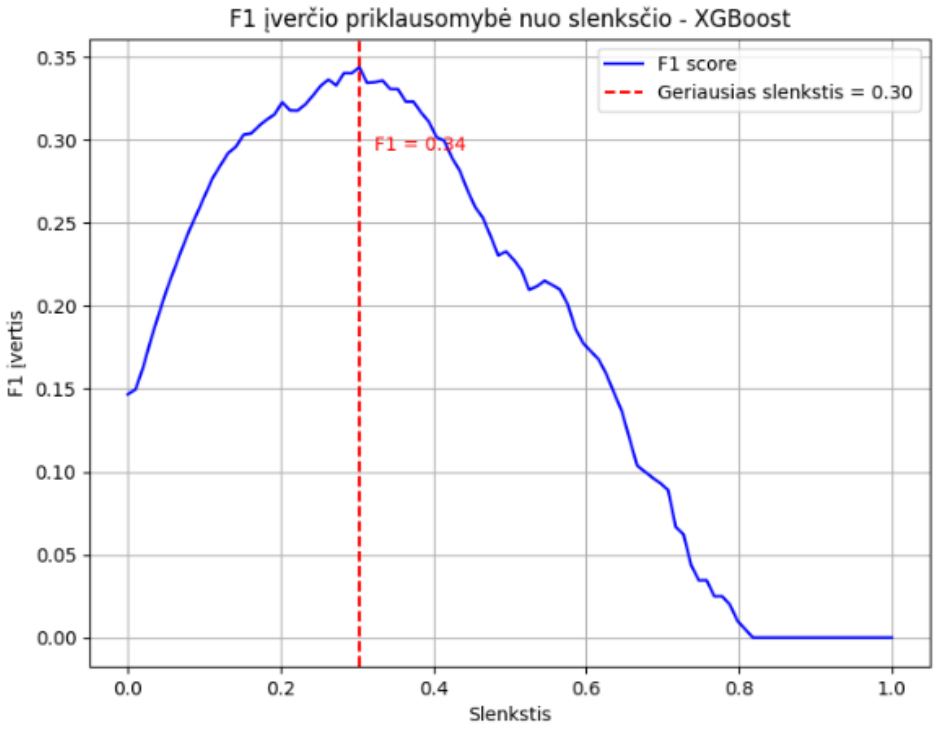
- Tikslumas parodo, kad 85,13 % visų atvejų buvo klasifikuoti teisingai. Tačiau tai nėra itin informatyvi metrika dėl klasių disbalanso. Ir nors atrodo, kaip labai aukštas rodiklis, tai ne visai tiesa.
- Jautrumas 48,72 % parodo, kad tik pusė 1 klasės elementų (nemokių paskolų) buvo klasifikuoti teisingai, kaip 1 klasės objektai. Kitaip tariant, maždaug kas antra nemoki paskola yra klaidingai praleidžiama kaip moki.
- Specifiškumas 88,27 % parodo, kad modelis tikrai gerai identifikuoja visus teigiamus atvejus. Vadinasi, šis modelis su 88,27 % visų tikrai mokių klientų identifikuoja kaip tikrai mokius.

- Preciziškumas 26,28 % parodo, kad ketvirtadalis prognozuotų teigiamos klasės elementų buvo teisingai identifikuoti kaip teigiama klasė. Kitaip tariant, tik kas ketvirtas modelio prognozuotas kaip nemokus klientas iš tiesų toks ir buvo.
- F1 įvertis 33,55 % yra subalansuotas jautrumo ir preciziškumo vidurkis, kuris atspindi kaip modelis geba aptikti nemokius klientus. Iš šio įverčio matyti, kad modelis ne pilnai atpažįsta nemokius klientus – tiek praleidžia tikrų atvejų, tiek suklysta pažymint neteisingai.

Taigi, modelis pasiekia tikrai aukštą tikslumą ir gan tiksliai atlieka klientų kreditingumo vertinimo užduotį. Deja, dėl klasių disbalanso, jo nemokėjimas atpažinti nemokius fizinius asmenis yra gana ribotas. Tad nors modelis gerai atpažįsta mokius klientus, jautrumas ir preciziškumas pabrėžia, jog nemokius klientus identifikuoja kebliau – tikrai nemokūs dažnai lieka nepastebėti arba mokūs klientai yra klaidingai priskiriami nemokiems.

Dar vienas iššūkis, su kuriuo susidurta, buvo prognozavimo slenksčio pasirinkimas. Visi modeliai pirmiausia suskaičiuoja tikimybes, o tada pagal pasirinktą slenkstį priskiria paskolą prie mokių, arba prie nemokių paskolų klasės. Įprastai modeliuojant šis slenkstis būna 0,5. Tačiau turimų duomenų atveju, 0,5 gali pasirodyti per daug grubus. Kadangi 0 ir 1 klasės yra stipriai nesubalansuotos, būtina pasirinkti žemesnį slenkstį. Dažnai pasirenkama 0,2 – 0,3 slenkstis. Todėl ir šiame tyrime, visiems modeliams buvo taikytas 0,3 slenkstis.

Šį pasirinkimą galima patikrinti išsispausdinus F1 įverčio ir klasifikavimo slenksčių (angl. *threshold*) grafiką. XGBoost modelio atveju jis pateiktas paveiksle žemiau. Grafikas parodo, kad geriausias klasifikavimo slenksčio pasirinkimas yra 0,30, prie šio slenksčio, F1 įvertis pasiekia aukščiausią įvertį = 0,34. Tai apskaičiuota aukščiau.



16 pav. XGBoost modelio F1 ir klasių prognozavimo slenksčio santykis

Toliau pateiktas XGBoost metodo reikšmingiausių kintamųjų sąrašas. Apžvelkime detaliau kiekvieną iš pagrindinių:

- **Paskutine_skola_pries** – parodo prieš kiek dienų asmuo yra turėjęs ir ar aplamai yra turėjęs viešai registruotų skolų. Jei asmuo niekada neturėjo skolos, buvo koduojama „99999“ reikšmė. Tai parodo, kad asmenys, niekada neturėję skolų arba turėję labai seniai (kuo didesnė reikšmė) yra susiję su mažesne kredito rizika. Ir atvirkščiai, kai šio kintamojo reikšmė žema, tai susiję su didesne kredito rizika.
- **Bvkkmn** – šis kintamasis rodo bendros vartojimo kredito kainos metinę normą. O shap analizė parodo, kad kuo didesnis šis rodiklis, tuo asmuo bus labiau linkęs susidurti su nemokumo problemomis.
- **Suma** – kintamasis parodo prašomos paskolos sumą. Kuo mažesnė paskolos suma, tuo didesnė tikimybė, kad klientas bus mokus ir atvirkščiai – kuo didesnė paskolos suma, tuo didesnė tikimybė, kad klientas bus nemokus.
- **Metai** – kintamasis parodo kuriais metais buvo suteikta paskola. Reikšmės nuo 2020 iki 2024. Iš analizės matoma, kad naujesnės paskolos yra linkusios būti mokesnėmis, nei seniau suteiktos paskolos.
- **Trukme_esamoje_darbovieteje** – kintamasis, kuris parodo kiek laiko asmuo dirba darbovietėje, kurios pajamos buvo vertintos paskolos išdavimo metu. Kadangi tai yra kategorinis kintamasis, jis buvo koduotas. Tad labai svarbu atkreipti dėmesį į kodavimą, kuris pateiktas lentelėje žemiau. Matyti, kad kuo ilgesnis darbo stažas, tuo didesnė ir reikšmė. Vadinasi, šio kintamojo svarbą interpretuojame taip, kad klientai ilgiau dirbantys dabartinėje darbovietėje yra mažiau rizikingesni, lyginant su tais, kurie dirba trumpiau.

5 lentelė. Kintamojo trukme_esamoje_darbovieteje kodavimas

Reikšmė	**Užkoduota**
1–4 mėn.	0
4–12 mėn.	1
1–2 m.	2
3–4 m.	3
5+ m.	4

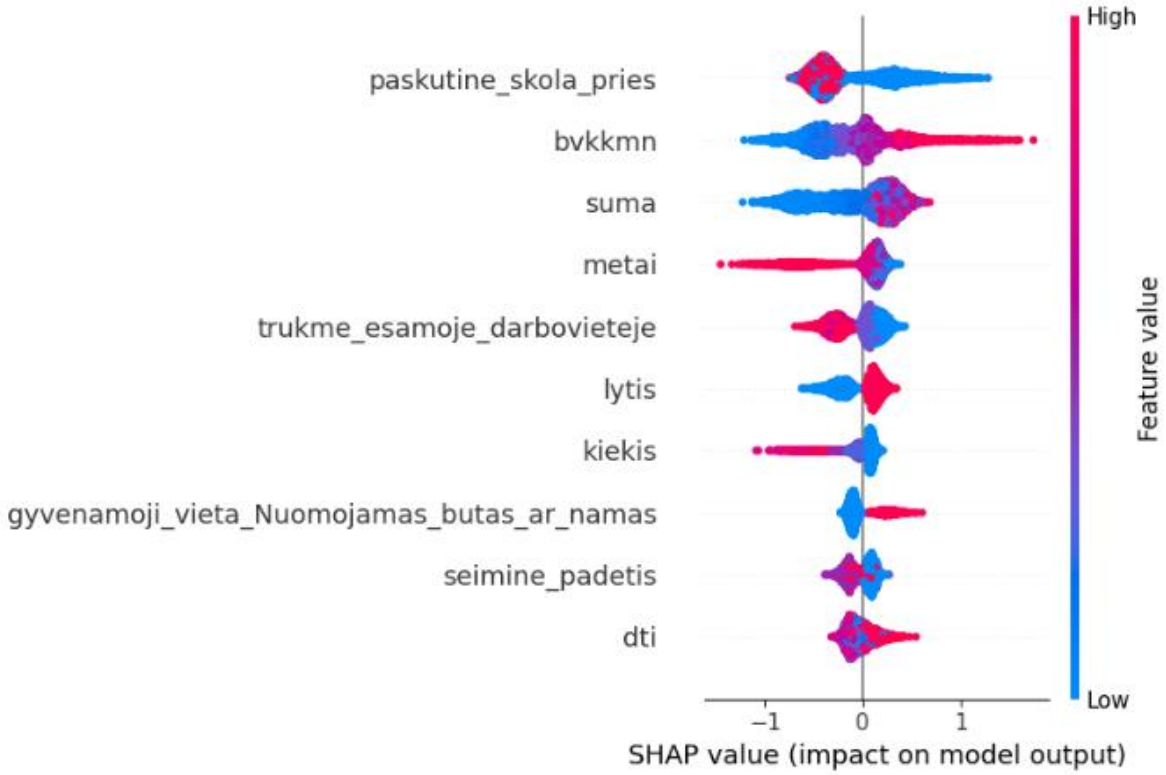
- **Lytis** – šis kintamasis taip pat koduotas. Moteris yra 0, vyras 1. Tad analizė parodo, kad vyrai yra šiek tiek rizikingesni už moteris.
- **Kiekis** – kintamasis parodo kiek asmuo yra turėjęs paskolų platformoje prieš skolinantis. Iš pateikto grafiko matoma, kad kuo didesnis turėtų paskolų skaičius, tuo didesnė kreditingumo rizika.
- **Gyvenamoji_vieta_Nuomojamas_butas_ar_**namas – šis kategorinis kintamasis yra sukurtas iš kintamojo gyvenamoji_vieta ir koduotas reikšmėmis 0 ir 1. Analizė parodo, kad jeigu asmuo gyvena nuomojamame name ar nuomojamame bute, jis bus labiau linkęs susidurti su mokumo problemomis.
- **Seimine_padetis** – tai yra kategorinis kintamasis, kuris buvo užkoduotas skaitinėmis reikšmėmis. Shap interpretacija šiuo atveju yra ganėtinai sudėtinga. Pagal grafiką matoma, kad kuo didesnė įgyjama reikšmė, tuo kreditingumo rizika bus aukštesnė. Tačiau žemiau pateiktoje lentelėje matoma kaip reikėtų iškoduoti užkoduotas reikšmes ir ši interpretacija yra visiškai netinkama. Galima tik atsargiai teigti, kad nesusituokę arba susituokę dažniau bus

siejami su mažesne rizika, o išsiskyrę, našliai arba partnerystėje esantys asmenys gali būti (bet nebūtinai) siejami su didesne rizika.

6 lentelė. Kintamojo seimine_padetis kodavimas

Reikšmė	**Užkoduota**
Nesusituokęs	0
Susituokęs	1
Išsiskyręs	2
Našlys	3
Partnerystė	4

- **DTI** – tai tolydusis kintamasis, parodantis koks yra asmens įsipareigojimų ir pajamų santykis paskolos suteikimo metu. Iš pateikto paveikslo matoma, kad kuo aukštesnė reikšmė, tuo labiau klientai linkę tapti nemokiai ir atvirkščiai, kuo mažesnė reikšmė, tuo labiau klientai bus linkę vykdyti įsipareigojimus laiku. Nors reikšmės yra ganėtinai išsibarsčiusios, tendencija vis tiek ryški.

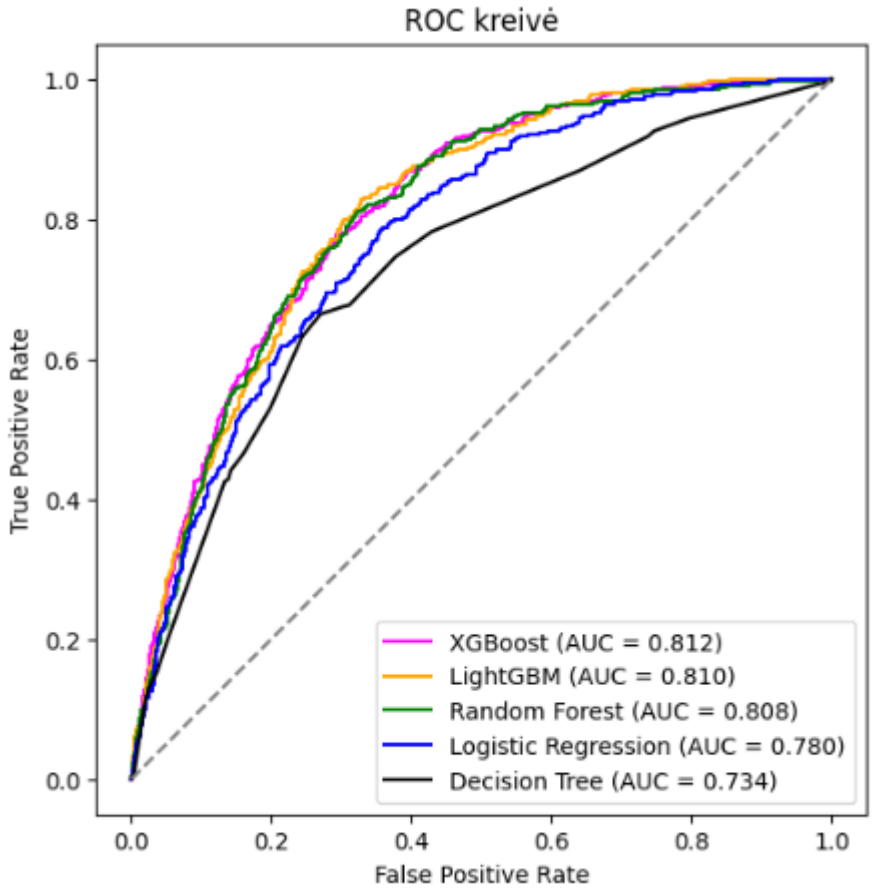


17 pav. XGBost metodo reikšmingiausių kintamųjų sąrašas

Atliekant tyrimą, papildomai norima patikrinti, kokią įtaką modelių efektyvumui padarytų prie kintamųjų pridėti keli makroekonominiai kintamieji, atliekant fizinių asmenų kreditingumo rizikos vertinimo uždavinį. Pasirinkti keturi makroekonominiai rodikliai: šalies BVP rodiklis, nedarbo lygis, vidutinė metinė infliacija ir Europos Centrinio banko palūkanų normos. Kiekvieną kintamąjį verta patikslinti atskirai:

- šalies BVP indeksas, kuris parodo BVP vertės pasikeitimą, lyginant su praėjusių metų tuo pačiu laikotarpiu, išreiškiant indeksu. Duomenys yra ketvirtiniai, tad siekiant išgauti mėnesio duomenis buvo panaudota linijinė interpoliacija.
- Šalies nedarbo lygis, kuris taip pat yra ketvirtinis. Norint gauti mėnesio duomenis, panaudota linijinė interpoliacija. Vertinti vyrai ir moterys kartu pagal visas amžiaus grupes ir neišskiriant miesto ar kaimo.
- Vidutinė metinė infliacija, apskaičiuota pagal vartotojų kainų indeksą. Pateikti duomenys mėnesiniai, kur mėnesio kainos lyginamos su to paties laikotarpio kainomis ankstesniais metais.
- Europos Centrinio banko (ECB) palūkanų norma, kuri yra viena iš trijų ECB palūkanų normų. ECB jas nustato su kainų stabilumo palaikymo tikslu euro zonoje. Duomenys yra dienos tikslumo, todėl pasirinkta imti paskutinės mėnesio dienos palūkanų normą.

Žemiau pateiktoje ROC kreivių diagramoje galime matyti modelių rezultatus. Matome, kad pačių geriausių algoritmų patikslinti nebepavyko. Tačiau atsitiktinio miško ir logistinės regresijos metodai nežymiai pagerėjo. Atsitiktinio miško metodo AUC reikšmė pakilo nuo 0,804 iki 0,808, o logistinės regresijos metodo AUC reikšmė pakilo nuo 0,772 iki 0,780.



18 pav. Modelių rezultatų palyginimas prie kintamųjų įtraukus makroekonominius rodiklius

Galima matyti, kad modelių efektyvumas gali pagerėti pridėjus makroekonominius kintamuosius, tačiau šiame tyrime jie sutapatinti ne visai korektiškai. Dabar tapatinimui parinkti paskolos išdavimo ir ekonominio rodiklio laikotarpiai. Bet siūlytina tapatinimui pasirinkti nemokumo įvykio datą, nes

tai parodytų korektiškesnį ryšį. Deja, turimame fizinių asmenų paskolų duomenų rinkinyje tokios informacijos nėra.

3.5. Rezultatų palyginimas

Analizuotos tarpusavio skolinimo platformos fizinių asmenų kreditingumo rizikos vertinimo metu sudaryto metodo efektyvumą galima palyginti su mokslinėje literatūroje aprašytų kreditingumo rizikos metodų vertinimo efektyvumu. Į lentelę, esančią žemiau, perkeltas sukurto modelio ir analizuotos literatūros šaliniuose tirtų metodų efektyvumas.

7 lentelė. XGBoost modelio rezultato palyginimas su literatūroje minimais modelių efektyvumo rezultatais

Autorius	Metodas lyderis	AUC	Tikslumas	Preciziškumas
Tyrimo metu analizuotas XGBoost metodas	XGBoost	0,812	85,13 %	26,28 %
P. Teply'is (2020)	Logistinė regresija	0,698		
W. Zhang'as (2020)	Logistinė regresija	0,700		
W. Zhang'as (2020)	Logistinė regresija	0,918		
L. Zhu'is (2019)	Atsitiktinio miško metodas	0,983		
V. Padimi's (2022)	Atsitiktinio miško metodas	-	94,60 %	
R. Sifrain'as (2023)	Logistinė regresija	-	89,90 %	
W. Yin'as (2023)	Atsitiktinio miško metodas	-		81,00 %
Y. Liu'is (2022)	Atsitiktinio miško metodas	0,966		
	Atsitiktinio miško metodas	0,836		
H. Wanga (2021)	Atsitiktinio miško metodas	0,970		
X. Zhu'is (2023)	LightGBM	0,721		
Z. Li'us (2021)	XGBoost	0,977		
G. Sotomayor'as (2022)	LightGBM	0,962		
P. Ko'as (2022)	LightGBM	0,749		

Iš lentelėje pateiktų rezultatų matoma, kad sukurtas XGBoost metodas nepasiekia didžiausio tikslumo, lyginant su aukščiausias AUC reikšmes įgijusiais pavyzdžiais iš analizuotos literatūros. Yra atvejų, kai atsitiktinio miško, XGBoost ar LightGBM metodai pasiekia ypatingai aukštas AUC reikšmes. Nepaisant to, sukurtas modelis beveik visada pasiekia didesnį tikslumą už literatūroje analizuotus logistinės regresijos metodus.

Visgi vertėtų atkreipti dėmesį į tai, kad analizuoti skirtingos apimties, skirtingų platformų fizinių asmenų duomenys, kurių laikmečiai gali skirtis. Taip pat yra žinoma, kad kai kurie tyrėjai savo duomenų analizėje naudojo ekonominius rodiklius, tokius kaip BVP, šalies nedarbo lygis, PVI ir kitus. Šių duomenų į savo tyrimą įtraukti nerekomenduotina, nes tyrimo tikslas yra ne išsiaiškinti nemokumo priežastis, kurios darė įtaką praeityje, bet prognozuoti kliento kreditingumo riziką sprendimo dėl paskolos suteikimo metu. Pažymėtina, kad Lietuvoje tokie svarbūs makroekonominiai

rodikliai nėra pasiekiami realiu laiku ar net su nedideliu atsilikimu. Dėl šių priežasčių tiesioginis rezultatų palyginimas su literatūroje tirtais metodais nėra galimas. Nepaisant to, vis dėl to sukurtas XGBoost klientų kreditingumo rizikos vertinimo metodas rodo pakankamai gerą tikslumą, nors ir turi potencialo tobulinimui. Taip pat pabrėžtina, kad galimas tolimesnis tyrimas, ieškant teisingų būdų klientų kreditingumo rizikos vertinimui panaudoti Lietuvoje skaičiuojamus makroekonominius rodiklius.

Toliau analizuojamas tyrimo metu tarpusavio skolinimo platformos fizinių asmenų duomenims pripažinto geriausio kreditingumo rizikos vertinimo metodo kintamųjų reikšmingumas. Palyginti XGBoost metodo reikšmingi kintamieji su literatūroje sutinkamais reikšmingais kintamaisiais. Dėmesys atkreiptas į atitikmenis literatūroje ir jų daromą įtaką. Duomenys pateikti lentelėje žemiau.

8 lentelė. XGBoost modelio reikšmingų kintamųjų palyginimas su literatūroje minimais reikšmingais kintamaisiais

XGBoost modelyje svarbus kintamasis	**Kintamojo įtaka nemokumui XGBoost modelyje**	**Literatūroje**	**Pastaba**
Paskutine_skola_pries	Neigiama (kuo naujesnė skola, tuo nemokumo tikimybė aukštesnė)	Konkretus kintamasis neminimas, tačiau lyginant su kredito istorijos kintamuoju, įtaka tokia pati	Siejama su asmens kredito istorija
Bvkkmn	Teigiama (kuo Bvkkmn reikšmė aukštesnė, tuo nemokumo tikimybė aukštesnė)	Literatūroje minimas dažnai, įtaka teigiama	Siejama su bendra grąžintina suma ar mokėtina palūkanų suma
Suma	Teigiama (kuo paskolos suma didesnė, tuo nemokumo tikimybė aukštesnė)	Labai dažnai minimas literatūroje, įtaka teigiama	Labai dažnai pasitaikantis kintamasis
Metai	Neigiama (naujesnių metų paskolos susijusios su mažesne rizika)	Literatūroje nėra minima	Galimai susijęs su makroekonominėmis sąlygomis
Trukme_esamoje_darbovieteje	Neigiama (kuo ilgesnis darbo stažas dabartinėje darbovietėje, tuo mažesnė nemokumo rizika)	Literatūroje pasitaiko, įtaka teigiama	Rečiau pasitaikantis – galimas atitikmuo bendras darbo stažas
Lytis	Vyrai šiek tiek rizikingesni	Dažnai pasitaiko literatūroje	Įtaka analogiška
Kiekis	Teigiama (kuo daugiau paskolų turėjęs, tuo didesnė nemokumo rizika)	Pasitaiko literatūroje, teigiama	Literatūroje sutinkamas ganėtinai dažnai
Gyvenamoji_vieta_Nuomojamas_butas_ar_namas	Teigiama (besinuomojantys nekilnojamąjį turtą yra su aukštesne nemokumo tikimybe)	Literatūroje turintys nuosavą NT daro neigiamą įtaką nemokumo tikimybei	Literatūroje galėtų būti NT nuosavybės atitikmuo
Seimine_padetis	Priklauso nuo kategorijos, bet iš duomenų matosi, kad nesusituokę ar susituokę labiau susiję su mažesne rizika	Literatūroje sutinkamas tarp reikšmingą įtaką darančių kintamųjų.	Sudėtingesnis įverčio komentavimas dėl kategorinio kintamojo kodavimo

XGBoost modelyje svarbus kintamasis	Kintamojo įtaka nemokumui XGBoost modelyje	Literatūroje	Pastaba
DTI	Teigiama (kuo aukštesnis pajamų ir įsipareigojimų santykis, tuo aukštesnė nemokumo tikimybė)	Literatūroje minimas dažnai, teigiama įtaka	Vienas iš dažniausiai sutinkamų kintamųjų literatūroje

Išanalizavus lentelėje pateiktus duomenis ir atlikus reikšmingų kintamųjų palyginimą, galima teigti, kad iš esmės dalis kintamųjų yra svarbiausi tiek sukurtame XGBoost modelyje, tiek analizuotoje literatūroje, o jų daromos įtakos analogiškos. Pavyzdžiui paskolos sumos, DTI santykio, lyties, šeiminės padėties kintamieji – visi šie kintamieji yra dažnai literatūroje nustatomi kaip darantys reikšmingą įtaką tarpusavio skolinimo platformos fizinių asmenų kreditingumo rizikos vertinime. Taip pat kredito istorijos informacija literatūroje dažnai paminima kaip reikšmingas kintamasis, šiuo atveju ir esamame tyrime pirmąją reikšmingumo poziciją užima kredito istorijos veiksnys, kintamojo paskutine_skola_pries pavidalu. Tačiau yra ir tokių kintamųjų, kurie būdingi tik šiam sukurtam modeliui, pavyzdžiui, metai, kada buvo suteikta paskola. Literatūroje nėra tiesioginės priklausomybės nuo paskolos suteikimo laikotarpio. Tačiau tai galėtų būti sietina su tam tikrais makroekonominiais veiksniais.

Taigi, apibendrinant lyginant tarpusavio skolinimo platformos fizinių asmenų kreditingumo rizikos vertinimui sukurto XGBoost modelio efektyvumą su mokslinėje literatūroje minimais kredito rizikos vertinimo modeliais ir jų efektyvumu, galime teigti, kad sukurtas XGBoost modelis parodė aukštesnį tikslumą nei tradiciniai logistinės regresijos metodai. Svarbu paminėti, kad skirtinguose šaltiniuose yra analizuoti skirtingų laikotarpių ir skirtingos apimties duomenų rinkiniai, todėl modelių tiesiogiai palyginti negalime. Vis dėl to, tikima, kad sukurtas XGBoost metodas turi potencialo tobulinimui.

Vertinant reikšmingus tarpusavio skolinimo platformos fizinių asmenų duomenų rinkinio kintamuosius, kuriuos sukurtas XGBoost modelis pristatė kaip reikšmingiausius ir lyginant juos su literatūroje minimais reikšmingą įtaką darančiais kintamaisiais pastebėta, kad didžioji dalis kintamųjų ir literatūroje įvardijami kaip reikšmingą įtaką darantys. Ypač išsiskiria paskolos suma, kredito kaina (bvkkmn), DTI santykis, lytis ir šeiminė padėtis. Tačiau yra ir kintamųjų, kurie nėra dažnai sutinkami literatūroje, bet jų įtaka buvo reikšminga sukurtam modeliui. Tai kintamieji tokie kaip: paskolos suteikimo metai, nuomojama gyvenamoji vieta. Visa tai parodo, kad kiekvienos tarpusavio skolinimo platformos vertinami fizinių asmenų duomenų rinkinių kintamieji, dažniausiai yra linkę sutapti, bet kiekviena platforma gali atrasti reikšmingos informacijos savo renkamuose duomenyse.

Išvados

1. Atlikus tarpusavio skolinimo platformų kreditingumo rizikos vertinimo mokslinės literatūros apžvalgą pastebėta, kad nors ši tema ir analizuojama ganėtinai seniai, ji vis dar nepraranda populiarumo ir kelia daug diskusijų. Tyrėjai nesutaria kurie algoritmai yra patys tinkamiausi ir tiksliausiai įvertina kreditingumo riziką. Tačiau galima buvo išskirti keletą dažniausiai minimų algoritmų, identifikuotų kaip reikšmingiausi kreditingumo rizikos vertinimo uždavinių sprendime. Nors logistinės regresijos modelis yra labiau tradicinis, jis vis dar nepraranda savo tikslumo tam tikrais atvejais. Nepaisant to, kad modernesni ir sudėtingesnes duomenų priklausomybes randantys algoritmai dažniau literatūroje įvardinami kaip geriausi, pasitaikydavo ir šaltinių, kur logistinės regresijos modelis tiksliausiai įvertindavo kreditingumo riziką tarpusavio skolinimo platformų fizinių asmenų paskolų rinkiniams.

2. Kreditingumo rizikos vertinimo klasifikavimo uždavinį nutarta spręsti su klasikiniu logistinės regresijos metodu bei įvairiais sprendimų medžiais grįstais algoritmais. Tai sprendimų medžio, atsitiktinio miško, LightGBM bei XGBoost metodai. Pastarieji du algoritmai dažnai literatūroje minimi kaip didelio našumo gradientinio stiprinimo metodai.

3. Panaudoti vienos, Lietuvoje veikiančios, tarpusavio skolinimo platformos fiziniams asmenims suteiktų paskolų duomenys. Kintamieji parinkti pagal literatūroje apžvelgtus. Pasirinkti kintamieji iš šių duomenų grupių: informacija apie paskolą, asmens kredito istorija, paskolos gavėjo asmeninė informacija.

4. Panaudojus skirtingus klasifikavimo metodus turimam duomenų rinkiniui, nustatyta, jog efektyviausiai kreditingumo rizikos vertinime buvo LightGBM bei XGBoost metodai. Tačiau logistinės regresijos metodas, nors ir demonstravo prastesnį rezultatą, tačiau nebuvo prasčiausias pasirinkimas. Tai parodo, kad pasirinktos tarpusavio skolinimo platformos fizinių asmenų paskolų duomenyse egzistuoja netiesiniai ryšiai.

5. Atlikus modelių įvertinimą pagal AUC metriką, nustatyta, kad geriausias metodas klasifikuojantis turimą tarpusavio skolinimo platformos fiziniams asmenims suteiktų paskolų duomenų rinkinį buvo XGBoost algoritmas, pasiekęs AUC = 0,812. Nors LightGBM metodas parodė prastesnį rezultatą, tačiau AUC = 0,810 įvertis rodo, kad šis metodas taip pat pasiekė aukštą rezultatą, kuris nuo XGBoost rezultato skiriasi labai minimaliai.

6. Tarpusavyje lyginant kreditingumo rizikos vertinimui naudotus metodus, pastebėta literatūroje apžvelgta tendencija, kuri atkreipia dėmesį į LightGBM ir XGBoost metodų efektyvumą. Šie metodai lengviau geba apdoroti netvarkingus duomenis ir jie daro ne tokią didelę įtaką modelio efektyvumui, lyginant su kitais modeliais. Atliktame tyrime šie du metodai taip pat buvo efektyviausi analizuotos tarpusavio skolinimo platformos fizinių asmenų paskolų duomenims. Nagrinėjant tarpusavio skolinimo platformų kreditingumo rizikos vertinimo literatūrą pastebėta, jog prognozavimui naudojami praeities makroekonominiai rodikliai, tokie kaip BVP, šalies nedarbo lygis ar kt. Šiame tyrime makroekonominių rodiklių nerekomenduotina naudoti, nes Lietuvoje jų negalime apskaičiuoti realiu laiku. Tikėtina, kad į duomenų rinkinį įtraukus makroekonominius rodiklius, modelio tikslumas išaugtų. Tačiau sprendžiant šį uždavinį to daryti

nerekomenduojama, nes modelis reikalingas tam, kad vertintų fizinio asmens kreditingumo riziką realiu laiku, kai asmuo teikia prašymą paskolai gauti tarpusavio skolinimo platformoje. Kita vertus, XGBoost prognozuoja nemokumą pakankamai tiksliai, todėl šį modelį rekomenduotina naudoti tarpusavio skolinimo platformoms, norinčioms atlikti fizinių asmenų kreditingumo rizikos vertinimą. Tad galime teigti, kad tyrimo rezultatai gali būti tiesiogiai pritaikomi tarpusavio skolinimo platformose, siekiant priimti sprendimus susijusius su kreditingumo rizikos vertinimu.

Tolesniuose tyrimuose siūloma:

- į tyrimą įsitraukti daugiau duomenų, kuriuos tarpusavio skolinimo platformos gali surinkti apie fizinį asmenį, pvz. išsamesnę kredito istoriją, elgsenos tarpusavio skolinimo platformoje duomenis ar net viešai prieinamą socialinių tinklų informaciją.
- vertinti modelio patikimumą ilguoju laikotarpiu. Svarbu po kurio laiko peržiūrėti ar modelis išlaiko tokį patį prognozavimo efektyvumą, nes keičiantis ekonominėms sąlygoms gali keistis ir asmenų elgsena. Tuo pačiu, gali keistis ir tiriami kintamieji.

Literatūros sąrašas

1. Lietuvos bankas. (2025, vasario 12). Finansų rinkos dalyviai. https://www.lb.lt/lt/finansu-rinku-dalyviai?type=30&market=1
2. Xiaoming, Z. ir Lean, Y. (2024) Consumer credit risk assessment: A review from the state-of-the-art classification algorithms, data traits, and learning methods. *Expert Systems with Applications, vol. 237* (2024), pp. 121484. https://doi.org/10.1016/j.eswa.2023.121484. Prieigos data 2025-03-11
3. Shen, C. Ir Wu, J. (2025) Research on credit risk of listed companies: a hybrid model based on TCN and DilateFormer. *Sci Rep 15, 2599* https://doi-org.ezproxy.ktu.edu/10.1038/s41598-025-86371-7. Prieigos data 2025-03-29
4. Wen, H., Sui, X., ir Lu, S. (2021). Study on Effect of Consumer Information in Personal Credit Risk Evaluation. *Complexity*, *2022*(1), 7340010. https://doi.org/10.1155/2022/7340010. Prieigos data 2025-03-12
5. Amarnadh, V. ir Moparthi, N. R. (2023). Comprehensive review of different artificial intelligence-based methods for credit risk assessment in data science. *Intelligent Decision Technologies*. https://doi.org/10.3233/IDT-230190. Prieigos data 2025-03-12
6. Bello, O., Folorunso, A., Ejiofor, O. ir kiti. (2023) Machine Learning Algorithms for Credit Risk Assessment: An Economic and Financial Analysis . *International Journal of Management Technology*, 10(1), 85–109. https://doi.org/10.37745/ijmt.2013/vol10n1109133. Prieigos data 2025-03-13
7. Lee, C. M., Delgado Fernandez, J., S., Rieger, A. ir Fridgen, G. (2023). Federated Learning for Credit Risk Assessment. *Proceedings of the 56th Hawaii International Conference on System Sciences*, Maui, HI, USA, 3–6 January 2023. University of Hawaii. ISBN 978-0-9981331-6-4. 10.24251/HICSS.2023.048. Prieigos data 2025-03-13
8. Saeed, S., Daniel, O., Salam, T. ir Olaoye, G. (2024). Machine learning for credit risk assessment and scoring. *ResearchGate*. Prieigos data 2025-03-29
9. Nahar, J., Rahaman, M. A., Alauddin, M. ir Rozony, F. Z. (2024). Big Data in Credit Risk Management: A Systematic Review Of Transformative Practices And Future Directions. *International Journal of Management Information Systems and Data Science,* [S.l.], v. 1, n. 4, p. 68–79, 2024. 10.62304/ijmisds.v1i04.196. Prieigos data 2025-03-15
10. Wang. Q. (2022). Research on the Method of Predicting Consumer Financial Loan Default Based on the Big Data Model. *Wireless Communications and Mobile Computing*, *2022*(1), 3786707. 10.1155/2022/3786707. Prieigos data 2025-03-11
11. Jiang, J., Liao, L., Wang, Z. ir Zhang, X. (2021). Government Affiliation and Peer-To-Peer Lending Platforms in China. *Journal of Empirical Finance*, *62*, 87-106. https://doi.org/10.1016/j.jempfin.2021.02.004. Prieigos data 2025-02-17
12. Nguyen, T. T. A., Pham, T. M. H., Vu, T. L. T. (2021). Default In The Us Peer-To-Peer Market With Covid-19 Pandemic Update: An Empirical Analysis From Lending Club Platform. *International journal of entrepreneurship*, vol. 25 (2021), nr. 7, pp. 1–19. 1939-467-25-7-586. Prieigos data 2025-02-18
13. A. Basha, S., Elgammal, M. M. ir Abuzayed, B. M. (2021). Online peer-to-peer lending: A review of the literature. *Electronic Commerce Research and Applications*, *48*, 101069. https://doi.org/10.1016/j.elerap.2021.101069. Prieigos data 2025-02-16
14. Munmun, M., Zhang, D. ir Luo, C.C. (2023). Peer-to-Peer Lending Performance Improvement: Learn from Lean Principles. *International Journal of Business and Management*, [S.l.], v. 19, n. 1, p. 101–115, 2024. DOI: 10.5539/ijbm.v19n1p101. Prieigos data 2025-02-17

15. Patwardhan, A. (2017). Peer-To-Peer Lending. *Handbook of Blockchain, Digital Finance, and Inclusion, Volume 1*, 389-418. https://doi.org/10.1016/B978-0-12-810441-5.00018-X. Prieigos data 2025-02-17
16. Waliszewski, K. ir Gebski, L. (2024). FinTech lenders on the consumer finance market in Central and Eastern Europe. *Przegląd Wschodnioeuropejski*, [S.l.], v. 15, n. 1, p. 81–96, 2024. DOI: 10.31648/pw.10180. Prieigos data 2025-02-17
17. Taujanskaitė, K. ir Milčius, E. (2022). Accelerated Growth of Peer-to-Peer Lending and Its Impact on the Consumer Credit Market: Evidence from Lithuania. Economies, 10(9), 210. https://doi.org/10.3390/economies10090210. Prieigos data 2025-03-13
18. Gunita M. (2022). Financial platforms as alternative financial instrument to crediting in Europe. *Proceedings of the 28th International Scientific Conference: Research for Rural Development 2022. Jelgava: Latvia University of Life Sciences and Technologies, 2022* (pp. 202–209). 10.22616/rrd.28.2022.029. Prieigos data 2025-02-12
19. Maloney, D. D., Hong, S. ir Nag, B. (2024). Loan Pricing in Peer-to-Peer Lending. *Journal of Risk and Financial Management*, *17*(8), 331. https://doi.org/10.3390/jrfm17080331. Prieigos data 2025-02-18
20. Vinod, K. L., Subramanyam, N., Keerthana S. ir kiti.(2016). Credit Risk Analysis in Peer-to-Peer Lending System. *Proceedings of the 2016 IEEE International Conference on Knowledge Engineering and Applications (ICKEA)*, 2016, p. 1–6. DOI: 10.1109/ICKEA.2016.7803017. Prieigos data 2025-02-18
21. Du, G., Liu, Z., ir Lu, H. (2021). Application of innovative risk early warning mode under big data technology in Internet credit financial risk assessment. *Journal of Computational and Applied Mathematics*, *386*, 113260. https://doi.org/10.1016/j.cam.2020.113260. Prieigos data 2025-03-14
22. Lin, M., ir Chen, J. (2022). Research on Credit Big Data Algorithm Based on Logistic Regression. *Procedia Computer Science*, *228*, 511-518. https://doi.org/10.1016/j.procs.2023.11.058. Prieigos data 2025-03-13
23. Murugan, K., Selvakumar, V. ir Venkatesh P. (2023). The big data analytics and its effectiveness on bank financial risk management. *Proceedings of the 2023 6th International Conference on Recent Trends in Advance Computing (ICRTAC)*, 2023, p. 1–6. DOI: 10.1109/ICRTAC59277.2023.10480831. Prieigos data 2025-03-16
24. Jiang, J., Liao, L., Lu, X., Wang, Z., ir Xiang, H. (2021). Deciphering big data in consumer credit evaluation. *Journal of Empirical Finance*, *62*, 28-45. https://doi.org/10.1016/j.jempfin.2021.01.009. Prieigos data 2025-03-15
25. Sadok, H., Sakka, F. ir El Maknouzi, M. E. H. (2022). Artificial intelligence and bank credit analysis: A review. *Cogent Economics & Finance*, *10*(1). https://doi.org/10.1080/23322039.2021.2023262. Prieigos data 2025-03-22
26. Wen, C., Yang, J., Gan, L., & Pan, Y. (2021). Big data driven Internet of Things for credit evaluation and early warning in finance. *Future Generation Computer Systems*, *124*, 295-307. https://doi.org/10.1016/j.future.2021.06.003. Prieigos data 2025-03-14
27. Yan, G. (2024). Research on the Application of Alternative Data in Credit Risk Management. *Highlights in Business, Economics and Management*, 40, 1156–1160. DOI: 10.54097/vn32pp64. Prieigos data 2025-03-30
28. Oye, E., Mattews, A., Peace, P., ir Andews M. (2025). Real-Time Credit Risk Monitoring with AI-Generated Insights. https://www.researchgate.net/publication/388675025_Real-Time_Credit_Risk_Monitoring_with_AI-Generated_Insights. Prieigos data 2025-03-30

29. Chen, S., Wu, R., ir Yin Z. (2023) The Application of Alternative Data in Personal Consumption Credit. *Advances in Economics, Management and Political Sciences*, [S.l.], v. 52, p. 212–216, 2023. DOI: 10.54254/2754-1169/52/20230720. Prieigos data 2025-03-30
30. Teply, P., ir Polena, M. (2019). Best classification algorithms in peer-to-peer lending. *The North American Journal of Economics and Finance*, *51*, 100904. https://doi.org/10.1016/j.najef.2019.01.001. Prieigos data 2025-02-14
31. Zhang, W., Wang, C., Zhang, Y., ir Wang, J. (2020). Credit risk evaluation model with textual features from loan descriptions for P2P lending. *Electronic Commerce Research and Applications*, *42*, 100989. https://doi.org/10.1016/j.elerap.2020.100989. Prieigos data 2025-02-15
32. Zhu, L., Qiu, D., Ergu, D., Ying, C., ir Liu, K. (2018). A study on predicting loan default based on the random forest algorithm. *Procedia Computer Science*, *162*, 503-513. https://doi.org/10.1016/j.procs.2019.12.017. Prieigos data 2025-02-15
33. Padimi, V., Sravan, V., ir Ningombam, D. D. (2022). Applying Machine Learning Techniques To Maximize The Performance of Loan Default Prediction. *International Journal of Computer Applications*, [S.l.], v. 178, n. 16, p. 1–7, 2023. DOI: 10.54216/JNFS.020204. Prieigos data 2025-03-22
34. Sifrain, R. (2023). Predictive Analysis of Default Risk in Peer-to-Peer Lending Platforms: Empirical Evidence from LendingClub *Journal of Financial Risk Management*, 12(1), 28–49. https://doi.org/10.4236/jfrm.2023.121003. Prieigos data 2025-02-18
35. Lyócsa, Š., Vašaničová, P., Hadji Misheva, B., ir Vateha, M. D. (2022). Default or profit scoring credit systems? Evidence from European and US peer-to-peer lending markets. *Financial Innovation*, *8*(1), 1-21. https://doi.org/10.1186/s40854-022-00338-5. Prieigos data 2025-02-19
36. Yin, W., Kirkulak-Uludag, B., Zhu, D., ir Zhou, Z. (2023). Stacking ensemble method for personal credit risk assessment in Peer-to-Peer lending. *Applied Soft Computing*, *142*, 110302. https://doi.org/10.1016/j.asoc.2023.110302. Prieigos data 2025-04-01
37. Liu, Y., Yang, M., Wang, Y., Li, Y., Xiong, T., ir Li, A. (2021). Applying machine learning algorithms to predict default probability in the online credit market: Evidence from China. *International Review of Financial Analysis*, *79*, 101971. https://doi.org/10.1016/j.irfa.2021.101971. Prieigos data 2025-04-01
38. Wang, H., Chen, W., & Da, F. (2021). Zhima Credit Score in Default Prediction for Personal Loans. *Procedia Computer Science*, *199*, 1478-1482. https://doi.org/10.1016/j.procs.2022.01.188. Prieigos data 2025-04-01
39. Zhu, X., Chu, Q., Song, X., Hu, P., ir Peng, L. (2023). Explainable prediction of loan default based on machine learning models. *Data Science and Management*, *6*(3), 123-133. https://doi.org/10.1016/j.dsm.2023.04.003. Prieigos data 2025-02-14
40. Alshboul, O., Almasabha, G., Shehadeh, A., & Al-Shboul, K. (2024). A comparative study of LightGBM, XGBoost, and GEP models in shear strength management of SFRC-SBWS. *Structures*, *61*, 106009. https://doi.org/10.1016/j.istruc.2024.106009. Prieigos data 2025-02-19
41. Li, Z., Li, S., Li, Z., Gao, H. (2021) Application of XGBoost in P2P Default Prediction. *Journal of Financial Risk Management*, 2023, 12(1), p. 28–49. DOI: 10.4236/jfrm.2023.121003. Prieigos data. Prieigos data 2025-04-30
42. Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., ir Niu, X. (2018). Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications*, *31*, 24-39. https://doi.org/10.1016/j.elerap.2018.08.002. Prieigos data 2025-02-16

43. Ponsam, J. G., Bella Gracia, S. V. J., Geetha, G., Karpaselvi, S., ir Nimala, K. (2021). Credit Risk Analysis using LightGBM and a comparative study of popular algorithms. *2021 4th International Conference on Computing and Communications Technologies (ICCCT)*, Chennai, India, 2021, pp. 634-641, doi: 10.1109/ICCCT53315.2021.9711896. Prieigos data 2025-02-20
44. Alzamora, G. S., Aceituno-Rojo, M. R., ir Condori-Alejo, H. I. (2021). An Assertive Machine Learning Model for Rural Micro Credit Assessment in Peru. *Procedia Computer Science*, *202*, 301-306. https://doi.org/10.1016/j.procs.2022.04.040. Prieigos data 2025-02-16
45. Ko, P., Lin, P., Do, H., ir Huang, Y. (2022). P2P Lending Default Prediction Based on AI and Statistical Models. *Entropy*, *24*(6), 801. https://doi.org/10.3390/e24060801. Prieigos datar 2025-03-11
46. Dumitrescu, E., Hué, S., Hurlin, C., ir Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, *297*(3), 1178-1192. https://doi.org/10.1016/j.ejor.2021.06.053. Prieigos data 2025-04-01
47. Pang, P. S., Hou, X., ir Xia, L. (2021). Borrowers' credit quality scoring model and applications, with default discriminant analysis based on the extreme learning machine. *Technological Forecasting and Social Change*, *165*, 120462. https://doi.org/10.1016/j.techfore.2020.120462. Prieigos data 2025-02-20
48. Xu, Q., Liu, C., Luo, J., ir Liu, F. (2024). Using machine learning to investigate the determinants of loan default in P2P lending: Are there differences between before and during COVID-19? *Pacific-Basin Finance Journal*, *88*, 102550. https://doi.org/10.1016/j.pacfin.2024.102550. Prieigos data 2025-02-20
49. Tang, X., Zhu, J., He, M., ir Feng, C. (2023). How can we learn from a borrower's online behaviors? The signal effect of a borrower's platform involvement on its credit risk. *Electronic Commerce Research and Applications*, *59*, 101272. https://doi.org/10.1016/j.elerap.2023.101272. Prieigos data 2025-02-15
50. Baumöhl, E., Lyócsa, Š., ir Vašaničová, P. (2024). Macroeconomic environment and the future performance of loans: Evidence from three peer-to-peer platforms. *International Review of Financial Analysis*, *95*, 103416. https://doi.org/10.1016/j.irfa.2024.103416. Prieigos data 2025-03-10
51. Nigmonov, A., Shams, S., & Alam, K. (2021). Macroeconomic determinants of loan defaults: Evidence from the U.S. Peer-to-peer lending market. *Research in International Business and Finance*, *59*, 101516. https://doi.org/10.1016/j.ribaf.2021.101516. Prieigos data 2025-02-15
52. Avgeri, E., ir Psillaki, M. (2023). Factors determining default in P2P lending (2023 *Journal of Economic Studies*, 2024, t. 51, nr. 4, p. 823–840. DOI: 10.1108/JES-07-2023-0376. Prieigos data 2025-03-10
53. Zhao, S., ir Zou, J. (2021). Predicting Loan Defaults Using Logistic Regression *Journal of Student Research*, 2021, t. 10, nr. 1. DOI: 10.47611/jsrhs.v10i1.1326. Prieigos data 2025-03-11
54. Ho, K., Gu, Y., Yan, C., ir Gozgor, G. (2024). Peer effects in the online peer-to-peer lending market: Ex-ante selection and ex-post learning. *International Review of Financial Analysis*, *92*, 103056. https://doi.org/10.1016/j.irfa.2023.103056. Prieigos data 2025-03-12
55. Wang, K., Li, M., Cheng, J., Zhou, X., ir Li, G. (2021). Research on personal credit risk evaluation based on XGBoost. *Procedia Computer Science*, *199*, 1128-1135. https://doi.org/10.1016/j.procs.2022.01.143. Prieigos data 2025-02-14
56. Wu, Y., ir Zhang, T. (2021). Can credit ratings predict defaults in peer-to-peer online lending? Evidence from a Chinese platform. *Finance Research Letters*, *40*, 101724. https://doi.org/10.1016/j.frl.2020.101724. Prieigos data 2025-02-15

57. Li, K., Zhou, F., Li, Z., Yao, X., ir Zhang, Y. (2021). Predicting loss given default using post-default information. *Knowledge-Based Systems*, *224*, 107068. https://doi.org/10.1016/j.knosys.2021.107068. Prieigos data 2025-03-11
58. Zahi, S., ir Achchab, B. (2019). Modeling car loan prepayment using supervised machine learning. *Procedia Computer Science*, *170*, 1128-1133. https://doi.org/10.1016/j.procs.2020.03.055. Prieigos data 2024-12-01
59. Barbaglia, L., Manzan, S., ir Tosetti, E. (2023). Forecasting Loan Default in Europe with Machine Learning. *Journal of Financial Econometrics*, 21(2), 569-596. https://doi.org/10.1093/jjfinec/nbab010. Prieigos data 2024-12-03
60. Bartosik, A., ir Whittingham, H. (2020). Evaluating safety and toxicity. *The Era of Artificial Intelligence, Machine Learning, and Data Science in the Pharmaceutical Industry*, 119-137. https://doi.org/10.1016/B978-0-12-820045-2.00008-8. Prieigos data 2024-12-03
61. Edgar, T. W., ir Manz, D. O. (2016). Exploratory Study. *Research Methods for Cyber Security*, 95-130. https://doi.org/10.1016/B978-0-12-805349-2.00004-2. Prieigos data 2024-12-03
62. Giannetto, C., Alibrandi, A., Zirilli, A., ir Lanfranchi, M. (2016). Egg consumption among young people: A study through the application of the logistic regression model. *American Journal of Applied Sciences*, 13(6), 697–707. DOI: 10.3844/ajassp.2016.697.707. Prieigos data 2024-12-10
63. Han, J., Kamber, M., ir Pei, J. (2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann Publishers. ISBN 978-0-12-381479-1. Prieiga per https://www.sciencedirect.com/book/9780123814791/data-mining-concepts-and-techniques. P rieigos data 2024-12-15
64. Boutilier, J., Michini, C. Ir Zhou, Z. Optimal multivariate decision trees. *Constraints* 28, 549–577 (2023). https://doi.org/10.1007/s10601-023-09367-y. Prieigos data 2024-10-20
65. Belyadi, H., ir Haghighat, A. (2020). Supervised learning. *Machine Learning Guide for Oil and Gas Using Python*, 169-295. https://doi.org/10.1016/B978-0-12-821929-4.00004-4. Prieigos data 2025-01-03
66. Zhang, Q. (2023). *Loan Risk Prediction based on Random Forest Model*. Advances in Economics, Management and Political Sciences, 5, 20220082. https://doi.org/10.21203/rs.3.rs-3094217/v1. Prieiga per: https://assets-eu.researchsquare.com/files/rs-3094217/v1_covered_7a61aec7-8467-4a73-a1c0-93ea978b29b1.pdf. Prieigos data 2025-01-03
67. Wang, Y., Pan, Z., Zheng, J. ir kiti. (2019). A hybrid ensemble method for pulsar candidate classification. Astrophysics and Space Science, 364(8), 139. DOI: 10.1007/s10509-019-3602-4. Prieigos data 2024-12-01
68. Dhaliwal, S. S., Nahid, A., & Abbas, R. (2018). Effective Intrusion Detection System Using XGBoost. *Information*, *9*(7), 149. https://doi.org/10.3390/info9070149. Prieigos data 2024-10-20
69. Shu, S., Sun, Y., ir Zhou, Y. (2022). Research on Factor Identification and Contribution of Bank Nonperforming Loans Based on Lightgbm Algorithm. *BCP Business & Management*, 30, 209–217. DOI: 10.54691/bcpbm.v30i.2434. Prieigos data 2024-10-21
70. Li, S., Jin, N., Dogani, A., Yang, Y., Zhang, M., ir Gu, X. (2024). Enhancing LightGBM for Industrial Fault Warning: An Innovative Hybrid Algorithm. *Processes*, 12(1), 221. DOI: 10.3390/pr12010221. Prieigos data 2024-10-21
71. Ghourabi, A. (2022). A Security Model Based on LightGBM and Transformer to Protect Healthcare Systems From Cyberattacks. *IEEE Access*, vol. 10, pp. 48890–48903. DOI: 10.1109/ACCESS.2022.3172432. Prieigos data 2024-10-23

72. Sirenden, B. H., Mursanto, P., ir Wijonarko, S. (2023). Dynamic texture analysis using Temporal Gray scale Pattern Image for water surface velocity measurement. *Image and Vision Computing*, *137*, 104749. https://doi.org/10.1016/j.imavis.2023.104749. Prieigos data 2025-04-25
73. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010. Prieigos data 2025-04-30
74. Salmerón-Gómez, R., García-García, C.B. & García-Pérez, J. A Redefined Variance Inflation Factor: Overcoming the Limitations of the Variance Inflation Factor. Comput Econ 65, 337–363 (2025). https://doi-org.ezproxy.ktu.edu/10.1007/s10614-024-10575-8. Prieigos data 2025-05-09
75. Perrota, A., ir Bliatsios, G. (2021). Why segmentation matters: a Machine Learning approach for predicting loan defaults in the Peer-to-Peer (P2P) Financial Ecosystem. *Risk Management Magazine*, 16(2), 35–49. DOI: 10.47473/2020rmm0089. Prieigos data 2025-02-15
76. Bravo, C. (2012). Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research 238 (2014) 505–513*. DOI: http://dx.doi.org/10.1016/j.ejor.2014.04.001. Prieigos data: 2025-05-16
77. Wang, Y., Jia, Y., Tian, Y., & Xiao, J. (2022). Deep reinforcement learning with the confusion-matrix-based dynamic reward function for customer credit scoring. *Expert Systems With Applications*, *200*, 117013. https://doi.org/10.1016/j.eswa.2022.117013. Prieigos data: 2025-05-16