Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Kredito rizikos modeliavimas investavimui sutelktinio finansavimo platformoje

Magistro studijų baigiamasis projektas

Projektą parengė

Indrė Balytė-Zykė

Projektui vadovavo

Prof. dr. Evaldas VAIČIUKYNAS

Doc. dr. Asta Daunorienė

Kaunas, 2026

Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Kredito rizikos modeliavimas investavimui sutelktinio finansavimo platformoje

Magistro studijų baigiamasis projektas

Didžiųjų verslo duomenų analitika (6213AX001)

Projektą parengė

Indrė Balytė-Zykė

Projektui vadovavo

Prof. dr. Evaldas Vaičiukynas
Doc. dr. Asta Daunorienė

Projektą recenzavo

Dr. Paulius Danėnas
Doc. Dr. Lina Sinevičienė

Kaunas, 2026

Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Indrė Balytė-Zykė

Kredito rizikos modeliavimas investavimui sutelktinio finansavimo platformoje

Akademinio sąžiningumo deklaracija

Patvirtinu, kad:

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

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

3. baigiamajame projekte tinkamai laikiausi asmens duomenų apsaugos reikalavimų, nenaudojau neskelbtinų ar konfidencialių duomenų be teisėto pagrindo, o jei juos naudojau, jie yra tinkamai nuasmeninti;

4. jei rengiant baigiamąjį projektą naudojausi dirbtinio intelekto (toliau – DI) ar kitais automatizuotais įrankiais, juos taikiau pagal Universitete nustatytą tvarką, nepažeisdama akademinio sąžiningumo principų;

5. nesumokėjau ir nesu įsipareigojusi mokėti jokių įstatymų nenumatytų piniginių sumų už baigiamąjį projektą ar jo dalis jokiam fiziniam ar juridiniam asmeniui;

6. suprantu, kad išaiškėjus akademinio nesąžiningumo ar kitų asmenų teisių pažeidimo faktui, man bus taikoma atsakomybė pagal Universitete nustatytą tvarką ir galiu būti pašalinta iš Universiteto; akademinio nesąžiningumo atvejis gali būti nagrinėjamas ir po studijų baigimo, inicijuojant kvalifikacinio laipsnio atšaukimo procedūrą.

Indrė Balytė-Zykė. Kredito rizikos modeliavimas investavimui sutelktinio finansavimo platformoje. Magistro studijų baigiamasis projektas / vadovas prof. dr. Evaldas Vaičiukynas, Informatikos fakultetas, vadovė doc. dr. Asta Daunorienė, Ekonomikos ir verslo fakultetas, Kauno technologijos universitetas

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

Reikšminiai žodžiai: mašininis mokymasis, atsitiktiniai miškai, CatBoost, sutelktinis finansavimas, EstateGuru, kredito rizika

Kaunas, 2026. 81 p.

Santrauka

Šiame magistro darbe nagrinėjami kredito rizikos vertinimo metodai sutelktinio finansavimo platformai „EstateGuru“, orientuojantis į nekilnojamojo turto paskolų segmentą. Darbo aktualumą lemia spartus P2P skolinimo rinkos augimas ir poreikis tiksliau vertinti paskolų nemokumo riziką, siekiant apsaugoti investuotojų kapitalą bei priimti efektyvesnius investavimo sprendimus.

Darbo tikslas – sukurti ir palyginti kredito rizikos vertinimo modelius „EstateGuru“ platformos duomenims, integruoti tiek paskolų charakteristikas, tiek makroekonominius rodiklius bei įvertinti jų pritaikomumą investavimo strategijų formavimui.

Tyrimo metu buvo surinkti ir parengti „EstateGuru“ platformos paskolų duomenys, atliktas duomenų valymas, transformacijos ir požymių inžinerija. Į modelius papildomai integruoti makroekonominiai rodikliai: infliacija, palūkanų normos, BVP bei nekilnojamojo turto kainų indeksų pokyčiai. Kredito rizikos vertinimui buvo taikyti logistinė regresija, atraminių vektorių mašina (SVM), atsitiktiniai miškai bei *CatBoost* modeliai. Modelių sėkmingumas lygintas naudojant detekcijos kreives (ROC, DET, preciziškumo-jautrumo) ir įvairius tikslumo įverčius.

Tyrimo rezultatai parodė, kad geriausius prognozavimo rezultatus pasiekė atsitiktinių miškų modelis, kuris efektyviausiai identifikavo rizikingas paskolas. Kintamųjų svarbos analizė parodė, kad didžiausią įtaką modelio prognozėms turėjo turto vertė, refinansavimo požymis, paskolos ir turto vertės santykis (LTV), šalies indikatorius bei dalis makroekonominių rodiklių. Papildoma SHAP analizė atskleidė kintamųjų poveikio kryptį ir parodė, kad didesnė turto vertė buvo siejama su mažesne paskolos nemokumo tikimybe.

Remiantis gautais kredito rizikos modeliais buvo suformuotos investavimo strategijos, leidžiančios filtruoti didesnės rizikos paskolas ir optimizuoti rizikos bei grąžos santykį. Tyrimo rezultatai parodė, kad pažangių mašininio mokymosi metodų bei makroekonominių rodiklių integravimas gali pagerinti kredito rizikos vertinimo kokybę P2P nekilnojamojo turto finansavimo rinkoje.

Indrė Balytė-Zykė. Credit Risk Modeling for Investing Through Crowdfunding Platform. Master‘s Final Project / supervisor prof. Evaldas Vaičiukynas; Faculty of Informatics, supervisor doc. dr. Asta Daunorienė; Faculty of Economics and Business, Kaunas University of Technology

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

Keywords: Machine Learning, Random Forest, CatBoost, Peer-to-Peer Lending, Credit Risk, EstateGuru

Kaunas, 2026. 81 p.

Summary

This master’s final project examines credit risk assessment methods for the “EstateGuru“ crowdfunding platform, focusing on the segment of real estate loans. The relevance of the study is driven by the rapid growth of the peer-to-peer (P2P) lending market and the increasing need for accurate default risk evaluation in order to protect investors’ capital and optimize investment decisions.

The aim of the paper is to develop and compare credit risk assessment models for “EstateGuru“ platform data by integrating both loan-specific characteristics and macroeconomic indicators, as well as to evaluate their applicability for investment strategy formation.

During the research, loan-level data from the “EstateGuru“ platform were collected and prepared including data cleaning, transformation, and feature engineering procedures. Macroeconomic indicators were additionally integrated into the models, including inflation, interest rates, gross domestic product, and changes in real estate price indexes. Credit risk assessment was performed using logistic regression, Support Vector Machine (SVM), Random Forest, and CatBoost models. Model performance was evaluated using detection curves (ROC, DET, and Precision-Recall curves) together with various classification accuracy metrics.

The results of the study showed that the Random Forest model achieved the best predictive performance and most effectively identified high-risk loans. Variable importance analysis revealed that property value, refinancing indicator, loan-to-value ratio (LTV), country indicator, and several macroeconomic variables had the greatest influence on model predictions. Additional SHAP analysis provided insights into the direction of variable effects and showed that higher property values were associated with a lower probability of loan default.

Based on the developed credit risk models, investment strategies were constructed to filter higher-risk loans and optimize the risk-return trade-off. The findings suggest that the integration of advanced machine learning methods and macroeconomic indicators can improve the quality of credit risk assessment in the P2P real estate financing market.

Turinys

Lentelių sąrašas

Paveikslų sąrašas

Įvadas

Pastaraisiais metais sutelktinio finansavimo rinka sparčiai plečiasi tiek Lietuvoje, tiek kitose Europos šalyse. Augant skolinimo platformų (angl. peer-to-peer, P2P) populiarumui, didėja investavimo galimybės, tačiau kartu auga ir kredito rizikos valdymo svarba. Investuotojams tampa aktualu tiksliai įvertinti paskolų riziką bei priimti pagrįstus investavimo sprendimus.

Nekilnojamojo turto finansavimo segmentas išsiskiria didesnėmis investicijų sumomis, ilgesniu projektų įgyvendinimo laikotarpiu ir jautrumu ekonominiams pokyčiams. Dėl šios priežasties kredito rizikos vertinimas šioje rinkoje tampa ypač svarbus. Netiksliai įvertinta paskolos rizika gali lemti reikšmingus investuotojų nuostolius bei sumažinti platformų patikimumą.

Tradiciniai kredito rizikos vertinimo metodai ne visada geba tiksliai įvertinti sudėtingus ryšius tarp paskolų charakteristikų ir paskolos nemokumo tikimybės. Skolinimo platformų (toliau – P2P) rinkose dažnai susiduriama su heterogeniškais (įvairiais) duomenimis, ribota istorinių duomenų apimtimi bei nestandartinėmis paskolų sąlygomis. Todėl vis didesnis dėmesys skiriamas pažangiems mašininio mokymosi (angl. machine learning, toliau – ML) metodams, kurie leidžia efektyviau analizuoti didelius duomenų kiekius ir identifikuoti sudėtingas priklausomybes tarp kintamųjų.

Mokslinėje literatūroje kredito rizikos vertinimui vis dažniau taikomi atsitiktinių miškų, gradientinio pastiprinimo, atraminių vektorių mašinos bei kiti ML metodai. Tyrimai rodo, kad šie algoritmai dažnai pasižymi geresniu prognozavimo tikslumu nei tradiciniai statistiniai modeliai. Taip pat vis daugiau dėmesio skiriama geresnei modelių interpretacijai, kintamųjų svarbos analizei ir makroekonominių rodiklių integravimui į rizikos vertinimo procesą.

Baltijos regiono P2P rinka išsiskiria aktyvia finansinių technologijų (toliau – fintech) sektoriaus plėtra ir sparčiai augančiu nekilnojamojo turto finansavimo segmentu. Viena žinomiausių šio regiono platformų yra „EstateGuru“, kuri specializuojasi nekilnojamuoju turtu užtikrintų paskolų finansavime. Nepaisant rinkos augimo, mokslinių tyrimų, orientuotų į Baltijos regiono P2P nekilnojamojo turto paskolų rizikos modeliavimą, vis dar yra palyginti nedaug.

Tyrimo objektas – paskolų rizikos vertinimo metodai sutelktinio finansavimo rinkoje, orientuojantis į nekilnojamojo turto trumpalaikių paskolų segmentą.

Darbo tikslas – įvertinti paskolų kredito riziką naudojant sutelktinio finansavimo platformos duomenis ir, remiantis modelio prognozėmis, sudaryti investavimo strategijas.

Tiriamojo darbo uždaviniai:

1. atlikti mokslinės literatūros analizę siekiant apibrėžti kredito rizikos vertinimo metodus ir investavimo strategijas sutelktinio finansavimo rinkoje;

2. identifikuoti ir suformuoti reikšmingus kredito rizikos vertinimo požymius, taikant požymių inžinerijos metodus, naudojant sutelktinio finansavimo platformos „EstateGuru“ duomenis;
3. integruoti makroekonominius rodiklius į kredito rizikos vertinimo modelius;
4. pritaikyti ir palyginti tradicinius statistinius bei šiuolaikinius mašininio mokymosi metodus kredito rizikai vertinti;
5. sukurti ir įvertinti investavimo strategijas, pagrįstas sukurtais kredito rizikos vertinimo modeliais, bei suformuluoti rekomendacijas investuotojams dėl rizikos valdymo ir investicijų optimizavimo sutelktinio finansavimo platformose.

Tyrimo metodai – mokslinių straipsnių analizė, antrinių duomenų analizė.

1. Literatūros analizė

Moksliniuose straipsniuose dažniausiai išskiriami keli pagrindiniai sutelktinio finansavimo tipai:

- P2P skolinimas (*angl. peer-to-peer lending*) – investuotojai skolina pinigus fiziniams ar juridiniams asmenims su palūkanomis;
- nuosavybės finansavimas (*angl. equity crowdfunding*) – investuotojai įsigyja dalį įmonės akcijų;
- atlygio pagrindu suteikiamas finansavimas, kai rėmėjai gauna produktą ar paslaugą mainais už paramą;
- paramos pagrindu suteikiamas finansavimas, kai lėšos skiriamos be atlygio, pvz., labdarai.

Ypatingas dėmesys tyrimuose skiriamas P2P skolinimo segmentui, kuris laikomas viena iš brandžiausių ir labiausiai reguliuojamų sutelktinio finansavimo formų. Sutelktinio finansavimo rinka vis labiau auga [1], todėl daugėja ir akademinių tyrimų, siekiančių tobulinti kredito rizikos vertinimo metodus. Literatūroje išryškėja skirtingi požiūriai, apimantys tiek klasikinę statistinę analizę, tiek pažangius ML algoritmus, struktūrinius duomenų modelius, tekstinę informaciją ir net investuotojų elgseną.

Šioje dalyje analizuojami straipsniai, tiesiogiai susiję su kredito rizikos vertinimu sutelktiniame finansavime. Straipsniai, kuriuose naudojami kredito vertinimo metodai, pateikiami 1 lentelėje. Analizuojant taikomus algoritmus, galima įžvelgti tam tikrą metodologinę dinamiką. Ankstyvieji tyrimai (2012–2016 m.) [2] [3] dažniausiai remiasi klasikiniais statistiniais modeliais: logistine regresija, kartais papildoma *Probit* modelio aproksimacija ar *Lasso* reguliarizacija. Tokie metodai orientuojasi į paprastą binarinį klasifikavimą, bandydami išskirti kreditavimo atvejus į rizikingus ir nerizikingus.

Nuo maždaug 2016 metų literatūroje ryškėja ML metodų dominavimas, ypač atsitiktiniai miškai (*angl. Random Forest*), atraminių vektorių mašina (*angl. Support Vector Machines*, toliau – SVM), gradientinio pastiprinimo metodas (*angl. Gradient Boosting*) bei *XGBoost* ir *LightGBM* modeliai. Šie metodai leidžia pasiekti aukštesnį prognozavimo tikslumą, ypač dirbant su dideliais ir heterogeniškais duomenų rinkiniais. Fridmanas ir Džinas [4], Malekipirbazaris ir Aksakalis [5], Lua Ti Trinh [6] bei Bėsensas ir Smedsas [7] empiriškai įrodė, kad ansamblių ir pastiprinimo metodai reikšmingai lenkia tradicinius modelius pagal klasifikavimo tikslumą.

Naujausiuose straipsniuose (2020–2023 m.) vis dažniau aptinkami specializuoti sprendimai:

- giliojo mokymosi pagrindu sukurtas išgyvenamumo analizės modelis (*angl. DeepSurv*);
- natūraliosios kalbos apdorojimas (NLP: BERT), skirtas paskolų aprašymams;

- tinklo topologijos analizė, kuria galima vertinti ryšių ir tarpusavio ryšio parametrus: tarpininkavimą ir centralizaciją.

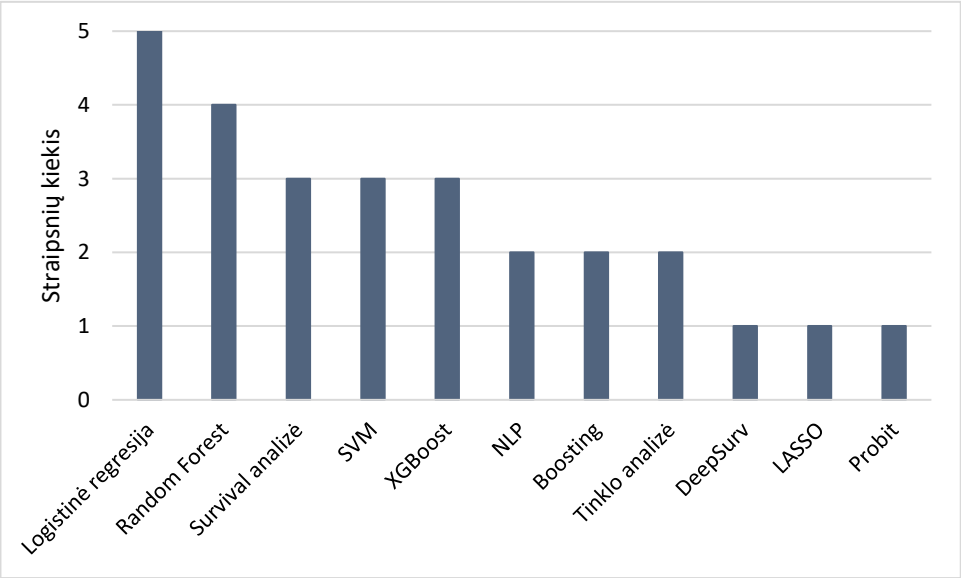
Tokie metodų pokyčiai atspindi tendenciją eksperimentuoti su vis daugiau algoritmų. Taip pat tai rodo didėjantį dėmesį duomenų įvairovei – nuo struktūrinių požymių iki tekstinių ir tinklinių duomenų.

1 lentelė. Straipsniai, kuriuose analizuojami kredito rizikos modeliai sutelktiniame finansavime

Metai	Autorius, Universitetas	Straipsnio pavadinimas	Naudotas metodas
2012	Michels, University of Michigan	Do unverifiable disclosures matter? Evidence from peer-to-peer lending	Probit modelis, regresija
2015	Duarte, Siegel & Young, University of Texas at Austin	Trust and credit: the role of appearance in peer-to-peer lending	Logistinė regresija
2015	Malekipirbazari, Aksakalli, Sabanci University	Risk assessment in social lending via random forests	Random Forest, detekcija
2016	Serrano-Cinca & Gutiérrez-Nieto, University of Zaragoza	Predicting default in P2P using a lending-based scoring system	Logistinė regresija, LASSO
2017	Freedman & Jin, Harvard University	Learning from peer-to-peer lending	Random Forest, SVM
2017	Xia, Liu et al., UST China & IBM Research	A Boosted Decision Tree Approach Using Survival Analysis	Boosted Trees + survival modeliai
2019	Zhang, Zhou et al., University of Texas at Dallas	Detecting Fraud in P2P Lending Using Supervised Learning	NLP, SVM, XGBoost (detekcija)
2020	Bart Baesens, Kristien Smedts, University of Southampton, United Kingdom	Boosting credit risk models	Pastiprinimo metodai (XGBoost, LightGBM)
2020	Chen, Wang et al., Carnegie Mellon University	DeepSurv: Applying Deep Neural Networks to Survival Analysis	DeepSurv (giliojo mokymosi i6gyvenamumo analizė)
2021	Barbara Dom, Ferenc Illes, Tímea Olvedi, Corvinus University of Budapest	Peer-to-peer lending: Legal loan sharking or altruistic investment? Analyzing	Kredito rizikos analizė, logistine regresija

		platform investments from a credit risk perspective	
2021	Zeng, Shen et al., ETH Zurich	Survival Analysis for Default Prediction in P2P Lending	Cox modelis, Kaplan-Meier analizė
2022	Yiting Liu, Lennart John Baals, Jörg Osterrieder, Branka Hadji-Misheva, Bern University ,Switzerland University of Twente, the Netherlands	Leveraging network topology for credit risk assessment in P2P lending: A comparative study under the lens of machine learning	Mašininis mokymasis (SVM, Random Forest, Gradient Boosting), tinklo analizė (betweenness, centrality)
2022	Lua Thi Trinh, Vietnam National University	A comparative analysis of consumer credit risk models in Peer-to-Peer Lending	Logistinė regresija, Random Forest, XGBoost
2023	Yan Wang, Xuelei Sherry Ni, Kennesaw State University	Improving Investment Suggestions for Peer-to-Peer (P2P) Lending via Integrating Credit Scoring into Profit Scoring	Kreditų reitingavimas (logistinė regresija) + pelno modeliavimo algoritmai

1 pav. pateikiamas nagrinėtų metodų populiarumas tyrimuose. Tai santykinai atspindi, kokie metodai taikomi nagrinėjant kredito riziką sutelktiniame finansavime. Nors pirmoje vietoje vis dar yra logistinė regresija, tačiau ML metodai dominuoja tarp populiariausių metodų.



1 pav. Algoritmų populiarumas tyrimuose

1.1. Kredito rizikos sampratų palyginimas

Mokslinėje literatūroje P2P kredito rizika suprantama įvairiai, priklausomai nuo tyrimo tikslo ir metodų. Tradiciškai kredito rizika apibrėžiama kaip tikimybė, kad skolininkas neįvykdys finansinių įsipareigojimų. Tokią sampratą naudoja dauguma klasikinių tyrimų, ypač kai

modeliuojama binarinė klasifikacija (grąžinta / negrąžinta paskola). Šį požiūrį taiko ir Seranas Sinka ir Gutjeresas. Jie nagrinėja P2P paskolų nevykdymo tendencijas, naudodami logistinę regresiją ir *Lasso* metodus.

P2P skolinimo kontekste svarbi ne tik binarinė rizika, t. y. ar paskola bus nevykdoma, bet ir platesnis rizikos vertinimas. Mokslinėje literatūroje vis dažniau analizuojamas ir laiko aspektas – ne tik ar paskola taps nemoki, bet ir kada tai gali įvykti. Tokiems tyrimams taikomi išgyvenamumo analizės metodai.

Kita rizikos samprata siekia įvardyti ne tik skolininko nemokumą, bet ir finansinius padarinius investuotojui. Tai apima investicinę grąžą, pelną ir nuostolį. Šis aspektas akcentuojamas literatūroje, kuri sujungia kredito rizikos įvertinimą su investavimo sprendimais. Pavyzdžiui, Vangas ir Ni [8] siūlo integruotą metodą, kuriame kredito reitingai sujungiami su pelno modeliavimo algoritmais. Taip siekiama ne tik prognozuoti riziką, bet ir optimizuoti investavimo grąžą.

Naujausi tyrimai rizikos vertinimą plečia įtraukdami ne tik finansinius rodiklius, bet ir papildomus informacijos šaltinius, pavyzdžiui, tekstinę informaciją ar ryšius tarp skolininkų. Sansas-Gvereras ir Arojas [9] parodė, kad natūralios kalbos apdorojimo metodai gali analizuoti paskolų aprašymus. Tuo tarpu Liu ir kt. [6] taiko tinklo analizės metodus ir nustato, kad ryšiai tarp paskolų bei skolininkų taip pat gali suteikti svarbios informacijos prognozuojant kredito riziką.

1.2. P2P skolinimo rinkos raida Baltijos regione

Pastaraisiais metais tarpusavio skolinimo ir sutelktinio finansavimo rinka Europoje demonstravo spartų augimą. Baltijos šalys tapo vienu aktyviausių alternatyvaus finansavimo regionų. Lietuva, Latvija ir Estija išsiskiria ne tik dideliu *fintech* sektoriaus augimu, bet ir palankia reguliacine aplinka bei aukštu skaitmenizavimo lygiu. Dėl šių priežasčių regione susiformavo stipri P2P platformų ekosistema, orientuota tiek į vartojimo kreditus, tiek į nekilnojamuoju turtu užtikrintas paskolas.

Kembridžo alternatyvių finansų centro ataskaitose [10] pažymima, kad Baltijos šalys pagal alternatyvaus finansavimo apimtis vienam gyventojui ilgą laiką buvo tarp lyderių Europoje. Regiono rinkos plėtrą skatino šie veiksniai:

- ribotas tradicinių bankų finansavimas smulkesniems projektams;
- spartus *fintech* sektoriaus augimas;
- investuotojų susidomėjimas didesnę grąžą siūlančiomis alternatyviomis investicijomis.

Baltijos regione veikia nemažai tarptautiniu mastu žinomų P2P ir sutelktinio finansavimo platformų. Viena didžiausių vartojimo paskolų platformų Europoje yra Latvijoje įkurta „Mintos“. Ji investuotojams suteikia galimybę investuoti į įvairių paskolų iniciatorių paskolas iš skirtingų šalių. Estijoje įkurta „Bondora“ specializuojasi vartojimo paskolų rinkoje bei automatizuotose investavimo strategijose. Tuo tarpu „EstateGuru“ orientuojasi į

nekilnojamuoju turtu užtikrintas paskolas ir išsiskiria tuo, kad finansuoja NT vystymo projektus Baltijos bei kitose Europos šalyse [11].

Lietuvoje taip pat veikia aktyvi P2P rinka [12]. Tarp žinomiausių platformų galima išskirti „Paskolų klubą“, „Finbee“, „HeavyFinance“ ir „Profitus“. Šios platformos veikia skirtinguose segmentuose: nuo vartojimo paskolų iki žemės ūkio ar nekilnojamojo turto finansavimo. Lietuvos bankas pažymi, kad alternatyvaus finansavimo sektorius Lietuvoje išlieka vienu sparčiausiai augančių *fintech* segmentų regione.

Moksliniuose tyrimuose Baltijos regiono P2P rinka dažniausiai nagrinėjama fintech plėtros, kredito rizikos ir investavimo požiūriu. Navickas ir Gudaitis [13] pažymi, kad Lietuva tapo viena pažangiausių *fintech* jurisdikcijų regione.

Didėjant palūkanų normoms ir griežtėjant bankų skolinimui, P2P platformų svarba dar labiau išaugo. Jos tapo svarbiu finansavimo šaltiniu smulkesniems verslo ir NT projektams. Investuotojams tai suteikė galimybę siekti didesnės grąžos, tačiau kartu padidino kredito rizikos valdymo svarbą.

Todėl kredito rizikos modeliavimas P2P rinkoje tampa itin aktualus tiek investuotojams, tiek platformoms. Tikslūs modeliai padeda efektyviau nustatyti rizikingas paskolas, optimizuoti investavimo sprendimus ir mažinti galimus nuostolius. Baltijos regionas šiuo požiūriu yra patraukli tyrimų aplinka, nes čia derinami spartus *fintech* augimas, alternatyvus NT finansavimas ir pažangių ML metodų taikymas kredito rizikos vertinime.

1.3. Tradiciniai statistiniai metodai

Tradiciniai statistiniai metodai (pvz., logistinė regresija) yra vieni iš pirmųjų, naudotų P2P paskolų rizikos prognozavimui. Šie modeliai pasižymi paprastumu ir aiškia interpretacija. Todėl dažnai naudojami kaip atskaitos taškas (*angl. baseline*) sudėtingesnių metodų vertinimui [14]. Statistiniai modeliai leidžia integruoti pagrindinius finansinius rodiklius ir, remiantis jais, apskaičiuoti tikimybę, kad paskola nebus grąžinta. Modeliai dažnai buvo naudojami tuose darbuose, kuriuose siekiama sudaryti įvertinimo (*angl. scoring*) sistemas arba analizuoti platformos naudotojų duomenis. Nors šie metodai tinkami paprastesnėms analizėms, jų apribojimas yra gebėjimas apdoroti tik linijines [15] sąveikas tarp kintamųjų. Vienas iš pirmųjų sistemingų bandymų taikyti klasikinius modelius P2P aplinkoje buvo atliktas Serano-Sinkos ir Gutjereso-Njeto [16]. Jie pasiūlė įvertinimo sistemą, paremtą paskolos dydžiu, trukme, kredito balu ir skolos/pajamų santykiu. Taip pat sukūrė paprastą ir aiškų rizikos įvertinimo modelį bei naudojo klasikinius rodiklius: skolos santykį, pajamas, paskolos sumą. Tai leido investuotojui greitai įvertinti riziką. Tuo tarpu Mičelsas [3] atskleidė, kad net subjektyvūs veiksniai (pvz. savanoriškai pateikta informacija paskolos aprašyme) gali turėti reikšmingą statistinį ryšį su paskolos rezultatu. Taip pat nustatė, kad skaidrumas reikšmingai siejasi su paskolos sėkme ir siūlė įtraukti aprašymus kaip papildomus rizikos indikatorius. Nors tokie modeliai paprasti ir interpretuojami, jų gebėjimas apdoroti sudėtingas nestruktūruotas sąveikas yra ribotas, ypač lyginant su pažangesniais ML metodais.

1.4. Mašininio mokymosi metodai

ML metodai suteikia daugiau galimybių apdoroti didelius ir sudėtingus P2P duomenų rinkinius [15]. Jie geba automatiškai aptikti paslėptus ryšius tarp kintamųjų, o tai ypač svarbu esant dideliam kiekiui tiek struktūrinių, tiek elgsenos duomenų. Dažniausiai naudojami algoritmai – tai atsitiktiniai miškai, *XGBoost*, *Boosted Trees* ir SVM. Šie metodai leidžia pasiekti aukštą prognozavimo tikslumą, ypač kai naudojami realūs P2P platformų duomenys. Šie modeliai taip pat tinka situacijoms, kur reikia klasifikuoti paskolas į rizikingas ir nerizikingas, optimizuoti paskirstymą pagal investavimo prioritetus ar vertinti paskolos patikimumą. Tačiau jų trūkumas yra mažesnis interpretavimo aiškumas, dėl kurio kartais investuotojui ar platformai sunku paaiškinti sprendimo logiką [17] [18]. Nepaisant to, jie tampa pagrindine priemone rizikos klasifikavimui, ypač kai derinami su kitais metodais. Fridmanas ir Džinas [4], naudodami *Prosper* platformos duomenis, parodė, kad šie metodai pasižymi ženkliai aukštesniu prognozavimo tikslumu nei tradiciniai modeliai. Jų naudojami atsitiktiniai miškai ir SVM parodė, kad ML gerokai lenkia klasikinius metodus prognozės tikslumu. Lua Ti Trinh [19] atliko skirtingų metodų palyginimą, o jo tyrimo rezultatai parodė, kad *XGBoost* lenkia kitas alternatyvas tiek tikslumu, tiek generalizacija. Autorius palygino kelis modelius: logistinę regresiją, atsitiktinius miškus, *XGBoost*. *XGBoost* pasirodė tiksliausias modelis paskolų prognozei. Malekipirbazaris ir Aksakalis [5] taip pat pasitelkė atsitiktinių miškų algoritmą, kuris efektyviai aptiko rizikingus paskolų modelius socialinio skolinimo platformoje.

1.5. Kiti metodai

1.5.1. Socialiniai, informaciniai ir netekstiniai veiksniai

Be finansinių kintamųjų, vis daugiau tyrimų analizuoja papildomą, dažnai neformalų turinį, kuris gali daryti įtaką paskolos sėkmei. Tai apima tiek skolininko pateikiamus aprašymus, tiek profilio informaciją, kuri nėra struktūriškai tikrinama. Tokie veiksniai kaip savanoriškai pateikta informacija, skolininko motyvacija, rizikos tolerancija, pasitikėjimas platforma ar net vertybinės nuostatos (pvz., altruizmas) daro įtaką jų sprendimams. Net jo vizualinis pasirodymas profilio nuotraukoje tampa svarbiu rizikos indikatoriumi. Tyrimai rodo, kad investuotojai dažnai reaguoja ne tik į kiekybinius rodiklius, bet ir į socialinius, emocinius ar net vizualinius aspektus. Nors tokius veiksnius sunkiau susisteminti, jų įtraukimas gali praturtinti rizikos vertinimą, ypač jei jie derinami su klasikine analize ar NLP metodais. Tokie duomenys ypač svarbūs platformoms, kurios orientuotos į vartotojo pasitikėjimą ir socialinį ryšį. Elgsenos ir socialiniai aspektai, kurie tradiciškai buvo mažai vertinami kredito analizėje, P2P kontekste įgauna vis didesnę reikšmę. Duartas, Žigelis ir Jangas [20] nustatė, kad skolininko profilio išvaizda (pvz., nuotrauka) gali turėti įtakos investuotojų pasitikėjimui ir sprendimui finansuoti paskolą. Tai rodo, kad net vizualūs signalai gali veikti investavimo sprendimus, formuodami papildomą suvokiamos rizikos aspektą. Barbara Dom [21] su kolegomis išskyrė dar vieną svarbią temą – investuotojų motyvaciją. Tyrinėjo investuotojų motyvus (pelnas ar altruizmas) P2P platformose ir parodė, kad elgesys priklauso nuo investuotojo tipo. Tyrimas parodė, kad kai kurie investuotojai vadovaujasi ne tik pelno

siekimu, bet ir altruistinėmis vertybėmis, o tai gali turėti įtakos jų požiūriui į riziką. Taigi rizikos vertinimas gali priklausyti nuo socialinio konteksto.

1.5.2. Natūralios kalbos apdorojimas ir nestandartiniai duomenys

Vienas iš naujausių metodologinių posūkių P2P tyrimuose – natūralios kalbos apdorojimo (NLP) taikymas [9]. Tyrėjai analizuoja paskolų aprašymus, vartotojų žinutes ar komentarus, kad iš jų išgautų semantinius rizikos požymius. Naudojami modernūs kalbos modeliai, tokie kaip *Bert* ar kiti *transformer* tipo algoritmai. Jie leidžia automatizuotai identifikuoti tekstuose paslėptą rizikos informaciją (pvz., emocinį toną, subjektyvumą ar neaiškumą). Šie metodai svarbūs tuo, kad leidžia apjungti tradicinę duomenų analizę su tekstine, dažnai subjektyvia informacija, kuri anksčiau likdavo nepanaudota. Jau anksčiau minėti NLP metodai pasižymi aukštu efektyvumu, tačiau reikalauja daug duomenų ir skaičiavimo išteklių, todėl dažnai taikomi kartu su ML algoritmais, kurie gali efektyviai apdoroti gautas savybes. Plėtojantis NLP technologijoms, atsirado galimybė naudoti paskolų aprašymus rizikos vertinime. Sansas-Gvereras ir Arojas [9] pasitelkė didžiuosius kalbos modelius LLM. Jie analizavo paskolų aprašymus P2P kontekste ir nustatė, kad tekstuose slypi reikšmingi rizikos indikatoriai. Tekstiniai duomenys padėjo geriau prognozuoti paskolų nevykdymą. Tai leidžia analizuoti ne tik kiekybinius duomenis, bet ir kalbines subtilybes, kurios anksčiau nebuvo sistemingai įtraukiamos į vertinimo modelius. Panašiai Džangas ir kt. [22] parodė, kad tekstiniai ir elgsenos duomenys padeda aptikti sukčiavimo atvejus. Tokių metodų pritaikymas gali žymiai padidinti platformos saugumą bei sumažinti investuotojų nuostolius.

1.5.3. Struktūros analizė

Dar viena inovatyvi kryptis – struktūrinių duomenų (tinklo) analizė. Šiuo atveju P2P platformos duomenys analizuojami kaip grafas: paskolos, skolininkai ir investuotojai suprantami kaip mazgai, o jų sąveikos – kaip ryšiai. Tinklo analizės metodai, tokie kaip „centriškumas“, artimumas, jungiamumas, padeda nustatyti, kurie vartotojai turi strateginę poziciją tinkle ir kaip tai siejasi su rizika. Tokie modeliai leidžia ne tik įvertinti atskirų paskolų riziką, bet ir suprasti, kaip ji plinta per sistemą. Nors tokia analizė reikalauja papildomų duomenų apie paskolų sąsajas, šie duomenys papildo rizikos vertinimą. Tinklo analizės metodai siūlo dar vieną pažangų rizikos vertinimo kampą. Liu ir kt. [6] pasiūlė naudoti tinklo analizę (*angl. social network analysis*) vertinant kredito riziką. Jie analizavo skolininkų ir paskolų tinklus kaip grafus, kuriuose galima analizuoti tarpusavio ryšius. Naudojant tokias tinklo charakteristikas autoriai parodė, kad rizikos modeliai tampa jautresni sisteminiams ryšiams, kurie gali būti pasislėpę klasikinėje analizėje. Ši metodika ypač naudinga platformose, kuriose skolininkai ar investuotojai tarpusavyje susiję.

1.5.4. Modelių kalibravimas ir išgyvenamumo analizė

Išgyvenamumo analizė tampa vis svarbesne kryptimi P2P skolinimo kontekste. Ji leidžia ne tik prognozuoti, ar paskola bus neįvykdyta, bet ir kada tai gali įvykti. Tokie modeliai kaip *Cox* regresija ar giliojo mokymosi metodai *DeepSurv* padeda sukurti laiko dimensiją turinčius rizikos modelius. Ši informacija naudinga tiek rizikos mažinimui, tiek investavimo laikotarpių planavimui. Taip pat kalibravimas leidžia modelius pritaikyti specifinėms platformoms ar

rinkos sąlygoms. Tai ypač svarbu, kai dirbama su nestandartinėmis paskolomis ar besikeičiančiais investuotojų lūkesčiais. Zengas ir kt. [1] pasiūlė taikyti *Kaplan–Meier ir Cox* modelius laiko iki paskolos nevykdymo analizei, o Čia [23] apjungė pastiprinimo algoritmus su išgyvenamumo modeliais. Dar labiau pažengęs pavyzdys – Čen ir kt. [24] *DeepSurv* modelis. Jo sukurtas modelis geba prognozuoti paskolos „gyvavimo“ laiką ir automatiškai mokytis iš duomenų be išankstinių prielaidų. Tokie modeliai tampa svarbūs platformoms, kurios siekia ne tik identifikuoti riziką, bet ir valdyti laikinius investavimo sprendimus.

1.6. Makroekonominiai veiksniai

Nors dauguma tyrimų orientuojasi į vidinius duomenis (individualias paskolas, skolininko savybes), vis daugiau dėmesio skiriama ir makroekonominiams veiksniams finansų srityje [25]. Pagrindiniai makroekonomikos rodikliai – palūkanų normos, infliacija, nedarbo lygis, BVP augimas. Jie gali daryti įtaką viso paskolų portfelio rizikai. Kitaip tariant, ekonomikos ciklo svyravimai paveikia daugelio paskolų grąžinimą vienu metu. Dėl šios priežasties investuotojams svarbu įvertinti ne tik kiekvienos paskolos ypatybes, bet ir bendrą ekonominį foną.

Makroekonominiai veiksniai ypač išryškėja išgyvenamumo (*angl. survival*) analizės kontekste, Zengas ir kt. [1], taikydamas išgyvenamumo analizę, parodė, kad laikui iki paskolos nevykdymo prognozuoti didelę įtaką turi ir makroekonominės sąlygos. Tuo tarpu Čenas [24] pažymi, kad jei prie paskolų duomenų būtų įtraukti makroekonominiai rodikliai, prognozių tikslumas galėtų dar labiau padidėti. Šie darbai pabrėžia, kad rizikos modeliavimo tikslumas ir portfelio stabilumas pagerėja, kai kartu su mikrolygmens duomenimis vertinami ir makrolygmens rodikliai.

Makroekonominiai rodikliai padeda investuotojams ir portfelio valdytojams suprasti bendrą rizikos kontekstą. Pavyzdžiui, palūkanų normų augimas padidina paskolų įmokas, todėl gali išaugti nevykdomų paskolų dalis. Infliacija mažina realią skolininkų perkamąją galią. Dėl to jiems sunkiau vykdyti įsipareigojimus. Nedarbo lygio kilimas rodo augančią ekonominę įtampą, kuri gali neigiamai paveikti daugelio skolininkų galimybes grąžinti paskolas. Aidemiro [26] tyrime nustatyta, kad staigūs makroekonominės aplinkos pablogėjimai (pvz. valiutos nuvertėjimas ar nedarbo augimas) ženkliai padidina kreditų portfelio riziką. Taigi įtraukdami makroekonominius duomenis tyrėjai gali dinamiškiau įvertinti paskolas. Modeliai realiu laiku atsižvelgia į ekonomikos ciklo fazę, o investuotojai gali geriau numatyti galimus nuostolių šuolius ekonomikos nuosmukio ar staigaus palūkanų normų pakilimo atvejais.

Įvairūs modeliai integruoja makroekonomikos rodiklius. Be išgyvenamumo analizės, makroekonominiai duomenys naudojami ir kredito rizikos portfelio streso testavime, kur vertinama, kaip portfelis atlaikytų ekonominius sukrėtimus. Tai leidžia modeliuoti blogiausius scenarijus ir laiku imtis priemonių rizikai mažinti. Pavyzdžiui, bankų sektoriuje IFRS 9 metodikoje kreditų rizikos vertinimas visuomet derinamas su skirtingais ekonomikos raidos scenarijais, kad būtų apskaičiuojami tikėtini nuostoliai recesijos ar krizės metu. Sugozu [27] atlikta analizė parodė, kad aukštas BVP augimas mažina bendrą kredito riziką, o didesnė infliacija – priešingai, padidina paskolų negrąžinimo tikimybę. Šie rezultatai atskleidžia, kaip

makroekonominiai rodikliai gali būti panaudojami įspėjant apie rizikos pokyčius. Tai gali padėti investuotojams lanksčiau valdyti paskolų portfelį, atsižvelgiant į platesnį ekonomikos kontekstą.

Makroekonominių rodiklių integravimas į kredito rizikos modelius ir investavimo strategijas suteikia svarbios informacijos apie sistemines rizikas. Tai leidžia investuotojams ir rizikos valdytojams suprasti, kaip jų portfelis gali elgtis ekonomikos pakilimo ar nuosmukio laikotarpiais. Naujausi moksliniai darbai rodo, kad derinant mikrolygmens duomenis su makroekonominiais rodikliais galima tiksliau prognozuoti portfelio veiklos rezultatus. Taip pat tai leidžia geriau subalansuoti riziką bei grąžą įvairiomis sąlygomis.

Šiame magistro darbo skyriuje aptartos įžvalgos pabrėžia, kad modernūs investavimo modeliai turėtų būti grįsti duomenimis ir įtraukti makroekonominius aspektus, siekiant didesnio investicijų stabilumo.

1.7. Modelių vertinimo metrikų analizė

Literatūroje kredito rizikos modelių vertinimo metrikos atspindi įvairius požiūrius į modelio veikimą ir jo taikymą praktikoje. Tyrimuose, kur naudojami klasikiniai statistiniai metodai (pvz., logistinė regresija), vertinimo metrikos tradiciškai apima tikslumą, jautrumą, specifiškumą ir AUC (plotas po ROC kreive). Seranas-Sinka ir Gutjeresas-Njetas [38] bei Barbara Dom ir kt. (2021) [21] naudoja šias metrikas vertindami binarinio klasifikavimo modelių gebėjimą atskirti patikimas paskolas nuo rizikingų.

Dalis naujesnių tyrimų, orientuotų į investavimo sprendimus, papildomai į vertinimą įtraukia investuotojo grąžos rodiklius ar rizikos-grąžos santykio metrikas. Jos nėra tiesiogiai klasifikavimo dalis, bet parodo praktinę modelio naudą investavimui [8]. Tokios metrikos gali apimti:

- portfelio grąžą;
- nuostolių sumažėjimą;
- *Sharpe* santykis ar kitus portfelio efektyvumo rodiklius.

1.8. Klasių disbalansas kredito rizikos modeliavime

Vienas iš pagrindinių iššūkių kredito rizikos modeliavime yra klasių disbalansas (*angl. class imbalance*). Ši problema atsiranda tuomet, kai viena klasė duomenų rinkinyje reikšmingai dominuoja prieš kitą. Kredito rizikos vertinimo uždaviniuose dažniausiai didžiąją duomenų dalį sudaro mokios paskolos, o nemokios paskolos sudaro tik nedidelę stebėjimų dalį. Dėl šios priežasties modeliai gali būti linkę prognozuoti dominuojančią klasę ir nepakankamai tiksliai identifikuoti rizikingas paskolas.

Hė ir Garsija [14] pažymi, kad klasių disbalansas yra viena svarbiausių problemų taikant ML metodus finansiniams duomenims. Autoriai pabrėžia, kad tradiciniai klasifikavimo algoritmai dažnai optimizuojami siekiant bendro tikslumo (*angl. accuracy),* todėl nesubalansuotuose duomenyse gali pasiekti aukštą tikslumą net ir prastai aptikdami mažumos klasę.

Kredito rizikos kontekste mažumos klasė dažniausiai yra svarbiausia [14], nes būtent nemokios paskolos sukelia didžiausius finansinius nuostolius investuotojams ar finansų institucijoms. Braunas ir Miusas [28] savo tyrime pabrėžia, kad kredito rizikos modeliuose svarbu ne tik bendras klasifikavimo tikslumas, bet ir modelio gebėjimas identifikuoti problemines paskolas. Todėl literatūroje vis dažniau naudojamos papildomos vertinimo metrikos, tokios kaip:

- ROC AUC;
- preciziškumo–jautrumo kreivės (*angl. PR curve*);
- F1 rodiklis;
- jautrumas (*angl. Recall)*;
- *kapa* koeficientas.

Sanas [29] pažymi, kad nesubalansuoti duomenys gali lemti modelio šališkumą daugumos klasės atžvilgiu, todėl kredito rizikos tyrimuose dažnai taikomi specialūs disbalanso mažinimo metodai. Literatūroje dažniausiai išskiriamos trys pagrindinės strategijos:

- duomenų balansavimas;
- algoritmų modifikavimas;
- tinkamų vertinimo metrikų parinkimas.

Viena dažniausiai taikomų strategijų – duomenų balansavimas keičiant klasių proporcijas. Čavla [30] pasiūlė SMOTE (*angl. Synthetic Minority Over-sampling Technique*) metodą, kuris generuoja sintetinius mažumos klasės stebėjimus ir padeda sumažinti modelio šališkumą. Šis metodas plačiai taikomas kredito rizikos modeliavime bei kituose finansinių sukčiavimų ar anomalijų identifikavimo uždaviniuose.

Kita tyrimų kryptis orientuojasi į algoritmus, kurie geriau prisitaiko prie nesubalansuotų duomenų. Lesmanas ir kt. [31] atliko plataus masto kredito rizikos modelių palyginimą ir nustatė, kad ansamblių metodai, tokie kaip atsitiktiniai miškai ar gradientinio pastiprinimo (*angl. Gradient Boosting*) metodai, dažnai efektyviau identifikuoja mažumos klasę nei tradiciniai statistiniai modeliai. Tokie algoritmai geba geriau modeliuoti sudėtingas netiesines sąveikas bei yra mažiau jautrūs duomenų disbalansui.

Literatūroje taip pat pabrėžiama, kad esant nesubalansuotiems duomenims ROC kreivė ne visada pakankamai atspindi modelio veikimą. Saitas ir Rėmsmajeris [16] teigia, kad preciziškumo-jautrumo kreivė dažnai yra informatyvesnė vertinant retos klasės aptikimą. Ji geriau parodo modelio gebėjimą identifikuoti mažumos klasę neprarandant prognozių tikslumo.

P2P kreditavimo platformose klasių disbalanso problema ypač aktuali, nes nemokios paskolos sudaro santykinai mažą dalį visų paskolų. Dėl šios priežasties kredito rizikos modeliai turi būti vertinami ne tik pagal bendrą tikslumą, bet ir pagal jų gebėjimą aptikti rizikingas paskolas bei sumažinti investuotojų nuostolius.

1.9. SHAP analizė kredito rizikos modeliuose

Pastaraisiais metais kredito rizikos tyrimuose vis didesnis dėmesys skiriamas ne tik modelių prognozavimo tikslumui, bet ir jų geresnei interpretacijai. Ši tendencija ypač išryškėjo pradėjus plačiau taikyti sudėtingus ML algoritmus, tokius kaip atsitiktiniai miškai, *XGBoost* ar *CatBoost*, kurių sprendimų logika dažnai laikoma „juodąja dėže“ (*angl. black-box models*) [6]. Dėl šios priežasties literatūroje vis labiau populiarėja kintamųjų svarbos paaiškinimo metodai, tarp kurių vienas plačiausiai naudojamų yra *SHAP* (*angl. SHapley Additive exPlanations*) analizė.

Liundbergas ir Li [32] pasiūlė *SHAP* metodą kaip vieningą modelių interpretavimo sistemą, leidžiančią paaiškinti sudėtingų ML modelių prognozes. Autoriai pabrėžia, kad SHAP analizė leidžia įvertinti individualų kiekvieno požymio poveikį prognozuojamam rezultatui. Taip pat užtikrina nuoseklų požymių svarbos vertinimą skirtinguose modeliuose.

Finansų ir kredito rizikos srityje galimybė interpretuoti modelius tampa ypač svarbi dėl reguliacinių reikalavimų ir poreikio pagrįsti automatizuotus sprendimus. Bušmanas [17] pažymi, kad finansų sektoriuje nepakanka vien aukšto prognozavimo tikslumo. Būtina suprasti, kokie veiksniai lemia modelio sprendimus. Autoriai pabrėžia, kad SHAP analizė leidžia identifikuoti svarbiausius kredito rizikos veiksnius.

Panašias išvadas pateikia Adas [33]. Jis analizavo ML metodų taikymą kredito rizikos modeliavime. Tyrime akcentuojama, kad pažangūs ML algoritmai geba aptikti sudėtingas netiesines sąveikas tarp požymių, tačiau jų interpretavimas tampa sudėtingesnis nei tradicinių statistinių modelių. *SHAP* analizė šiuo atveju leidžia ne tik nustatyti svarbiausius požymius, bet ir interpretuoti jų poveikio kryptį.

Literatūroje *SHAP* analizė dažnai naudojama vertinant finansinių bei makroekonominių rodiklių poveikį paskolų nemokumo prognozėms. Tyrimai rodo, kad tokie veiksniai kaip paskolos ir turto vertės santykis (LTV), paskolos suma, pajamų lygis, užstato vertė ar palūkanų norma dažnai turi reikšmingą įtaką kredito rizikai. Taip pat *SHAP* metodas leidžia nustatyti, kaip šių požymių poveikis skiriasi tarp atskirų stebėjimų. Dėl to tampa įmanoma atlikti ne tik globalią, bet ir individualią paskolų analizę.

Molnaras [18] pažymi, kad *SHAP* metodas ypač svarbus praktiniuose taikymuose, kai modelio rezultatus reikia interpretuoti ne tik duomenų analitikams, bet ir verslo sprendimų priėmėjams ar investuotojams. Todėl *SHAP* analizė tampa svarbi priemonė kredito rizikos modelių skaidrumui ir patikimumui didinti.

1.10. Investavimo strategijų modeliavimas

Pastaraisiais metais mokslinėje literatūroje vis daugiau dėmesio skiriama duomenimis grįstoms investavimo strategijoms P2P rinkose. Šiuolaikiniai tyrimai siekia sujungti kredito rizikos modelius su portfelio valdymo metodais. Pagrindinis tikslas – pasinaudoti pažangiomis rizikos prognozėmis ir taip formuoti efektyvesnį investicijų portfelį. Šiame skyriuje aptariamos naujausios įžvalgos apie tokias integruotas investavimo strategijas.

Kredito reitingų integravimas į investavimo sprendimus. Pastaruoju metu pasirodė tyrimų, nagrinėjančių, kaip P2P platformose taikomą kredito rizikos vertinimą paversti tiesioginiu investavimo įrankiu. Šiuose darbuose pabrėžiama, kad rizikos prognozės gali būti naudojamos ne tik paskolų klasifikavimui, bet ir investuotojams priimant sprendimus.

Janas Vangas ir Šiuėlei Šeri Ni [8] parodė, kad investavimo sprendimai neturėtų būti grindžiami vien paskolos palūkanų norma. Į sprendimų priėmimą svarbu įtraukti ir prognozuotą paskolos nemokumo tikimybę. Autorių rezultatai rodo, kad toks požiūris pagerina rizikos ir grąžos santykį. Tyrime kredito reitingavimas, paremtas logistine regresija, buvo sujungtas su pelningumo prognozavimo modeliais. Pasiūlytas integruotas metodas leido optimizuoti portfelio grąžą ir sumažinti galimų nuostolių tikimybę. Gauti rezultatai rodo, kad kredito rizikos modelių integravimas į investavimo procesą leidžia pasiekti stabilesnę portfelio grąžą. Tai ypač aktualu tais atvejais, kai investuotojas vertina paskolos nemokumo tikimybę (PD) kartu su palūkanų rodikliais.

Duomenų analizė ir portfelio optimizavimas. Naujieji tyrimai, tęsiantys klasikinės portfelio teorijos tradiciją, pabrėžia rizikos ir grąžos balansavimo svarbą. Šis principas aktualus ir P2P platformose. Investuotojas turi vertinti ne tik atskiras paskolas, bet visą portfelį.

Diversifikacijos principas, iškeltas Markovitso [34] portfelio teorijoje, išlieka svarbus ir šiandien. Skirtingo rizikingumo paskolų derinimas leidžia valdyti bendrą portfelio riziką. Tai ypač svarbu P2P investavime, kur paskolos pasižymi dideliu heterogeniškumu.

Tuo tarpu Šarpas [35] primena, kad didesnė rizika turi būti kompensuojama didesne tikėtina grąža. Ši logika aiškiai pritaikoma ir P2P investavime. Investuotojai turėtų vertinti ne tik paskolos palūkanų normą, bet ir galimą nuostolį.

Kredito rizikos literatūroje taip pat dažnai naudojama tikėtino nuostolio (*angl. Expected Loss*) samprata [36], kuri apskaičiuojama kaip PD ir LGD sandauga:

$$EL = EAD \times PD \times LGD \qquad (1.10.1)$$

čia:

EAD – investuojama suma;

PD – paskolos nemokumo tikimybė;

LGD – nuostolis nemokumo atveju.

Šis rodiklis leidžia įvertinti, ar siūloma palūkanų norma pakankamai kompensuoja prisiimamą kredito riziką. Kitaip tariant, investuotojas turi įsitikinti, kad didėjanti rizika, išreikšta PD ir LGD reikšmėmis, yra atlyginama didesne grąža.

Daugelyje P2P platformų paskolos yra neužtikrintos, neturinčios užstato ar garantijų, todėl nesėkmės atveju investuotojas praranda visą investuotą sumą. Literatūroje pabrėžiama, kad konservatyvus požiūris, jog įsipareigojimų nevykdymo atveju prarandama visa investuota suma (t. y. LGD = 1 arba 100 %), yra pagrįstas, nes tikėtinas atgavimas beveik nulinis [37].

P2P paskolų atveju ribotas išieškojimas lemia, kad investuotojai žvelgia į blogiausią scenarijų, skaičiuodami tikėtiną nuostolį [24]. Dėl to tikėtinas nuostolis apskaičiuojamas paprastai: $EL = EAD \times PD$. Taigi laikoma, kad nuostolio dydis nemokumo atveju yra 100 % [2].

Kredito rizikos modeliai kaip investavimo atrankos filtras. Duomenų mokslu ir ML algoritmais grįstos metodikos leidžia dinamiškai filtruoti rizikingas paskolas. Tai ypač aktualu P2P platformose. Jose investuotojai susiduria su dideliu paskolų kiekiu ir nevienodu rizikos lygiu. Taip pat informacija apie skolininkus dažnai būna ribota.

Reaguojant į šiuos iššūkius, nauji tyrimai taiko mašininio mokymosi metodus kredito rizikai prognozuoti. Šie modeliai naudojami tiesiogiai investavimo sprendimų palaikymui. Pavyzdžiui, Liu ir kt. darbe buvo pritaikyta tinklo analizė P2P kreditų rizikai vertinti. Tai padėjo investuotojams geriau identifikuoti sistemines rizikas paskolų portfelyje.

Inovatyvūs sprendimai ir teorijų adaptacijos. Klasikiniai kredito rizikos modeliai išlieka svarbūs ir šiuolaikiniame kontekste. Mertono struktūrinis kredito rizikos modelis [38] parodė, kad didėjant nemokumo tikimybei investuotojai reikalauja didesnių palūkanų. Šis principas šiandien pritaikomas rizikos vertinimui P2P platformose.

Kiti tyrėjai, pavyzdžiui, Seranas-Sinka, nagrinėja, kaip tradiciniai rizikos rodikliai integruojami į investavimo algoritmus. Tokie rodikliai apima kredito reitingus ar skolininko finansinius duomenis. Naujausi darbai pabrėžia holistinį požiūrį į investavimą. Investavimo modeliai turėtų apimti ne tik mikrolygmens informaciją apie skolininką, bet prireikus ir makroekonominius veiksnius. Tai leidžia kurti portfelio strategijas, atsparesnes ekonomikos ciklams.

Pastaraisiais metais akademinė literatūra rodo aiškią kryptį. P2P investavimo strategijos tampa vis labiau analitinės ir paremtos duomenimis. Tradiciniai investavimo principai, tokie kaip rizikos ir grąžos kompromisas, diversifikacija ar rizikos premija, nėra atmetami. Priešingai, jie papildomi šiuolaikinėmis kredito rizikos prognozėmis ir portfelio optimizavimo algoritmais.

Dėl to investuotojai gali geriau subalansuoti savo portfelį. Tokios strategijos gali sumažinti galimus nuostolius ir padidinti grąžos stabilumą. Šios įžvalgos sudaro teorinį pagrindą šiame magistro darbe nagrinėjamoms investavimo strategijoms. Jos remiasi kredito rizikos prognozių integravimu į sprendimų priėmimą ir pažangiais paskolų atrankos metodais.

1.11. Išvados

Atlikta literatūros analizė parodė, kad kredito rizikos vertinimas P2P skolinimo platformose tampa vis sudėtingesne ir daugiasluoksne tyrimų sritimi. Ankstyvuosiuose tyrimuose dominavo tradiciniai statistiniai metodai, tačiau naujesniuose darbuose vis dažniau taikomi pažangūs ML algoritmai. Literatūra rodo, kad šie metodai dažniausiai pasižymi geresniu prognozavimo tikslumu nei klasikiniai statistiniai modeliai.

Moksliniuose tyrimuose taip pat ryškėja tendencija plėsti naudojamų duomenų spektrą. Be tradicinių finansinių rodiklių, vis dažniau analizuojami tekstiniai duomenys, socialiniai signalai, tinklo struktūros bei makroekonominiai veiksniai. NLP metodai ir tinklo analizė leidžia identifikuoti papildomus rizikos požymius, kurie nėra matomi struktūriniuose finansiniuose duomenyse. Tuo tarpu makroekonominių rodiklių integravimas suteikia galimybę vertinti sisteminius rinkos pokyčius ir jų poveikį paskolų portfeliui.

Literatūroje taip pat pabrėžiama modelių interpretavimo svarba. Didėjant sudėtingų ML algoritmų taikymui, vis aktualesni tampa kintamųjų paaiškinimo metodai, tokie kaip SHAP analizė, leidžianti interpretuoti modelių sprendimus bei nustatyti svarbiausius kredito rizikos veiksnius.

Kredito rizikos tyrimuose itin svarbi išlieka klasių disbalanso problema, nes nemokios paskolos dažniausiai sudaro santykinai nedidelę duomenų dalį. Dėl šios priežasties literatūroje rekomenduojama naudoti ne tik bendro tikslumo rodiklius, bet ir papildomas vertinimo metrikas, tokias kaip ROC AUC ar preciziškumo-jautrumo kreives.

Literatūros analizė parodė, kad Baltijos šalių P2P rinka išsiskiria *fintech* plėtra, aktyvia nekilnojamojo turto finansavimo platformų veikla ir augančiu ML metodų taikymu kredito rizikos vertinimo srityje. Nepaisant to, mokslinių tyrimų, orientuotų būtent į Baltijos regiono nekilnojamojo turto P2P platformas, vis dar yra nedaug.

Apžvelgta literatūra pagrindžia šio magistro darbo aktualumą. Ji taip pat atskleidžia poreikį kurti kredito rizikos modelius, kuriuose būtų naudojami tiek paskolų charakteristikos kintamieji, tiek makroekonominiai rodikliai. Taip užtikrinama geresnė modelių interpretacija ir praktinis pritaikymas investuotojams realiose P2P platformose, tokiose kaip „EstateGuru“.

2. Metodologija

2.1. Duomenų rinkinys

Duomenų rinkinį sudaro 7562 eilutės, kur kiekviena eilutė skirta vienai paskolai. Pradinė duomenų aibė (Priedas 1), paimta iš „EstateGuru“ svetainės, buvo pakoreguota:

- atsisakyta kintamųjų, kurie yra žymi paskolos rezultatą, tam kad būtų išvengta duomenų nutekėjimo (*angl. Data lekeage*): *Actual Return, Principal paid, Interest paid, Bonus paid (borrower), Penalty paid, Indemnity Paid, Total Delay Days, Number of payments delayed, Prolonged, Outstanding Principal, Recovered Principal, Written Off* (1 priedas).
- atsisakyta datos tipo kintamųjų;
- įtraukti makroekonominiai rodikliai;
- suformuotas tikslo kintamasis: ar paskola yra moki;
- pridėtas laiko kintamasis – tam, kad paskolos būtų nuosekliai priskirtos laiko periodams kredito rizikos modeliams taikyti.

Suformuoti tokie stulpeliai:

Loan code - unikalus identifikatorius paskolai „EstateGuru“ platformoje;

Country - kategorinis kintamasis, nusakantis šalį, kurioje yra skolininkas / nekilnojamas turtas pagal „EstateGuru“;

YearMonth – kintamasis, sudarytas remiantis ankstesne iš datų, kai paskola buvo visiškai investuota (*angl. Fully Invested Date*) ir kai paskola pradėjo rinkti investicijas (*angl. Funded Date*) reikšmių. Investavimo data suapvalinama iki mėnesio lygmens (metai-mėnuo). Skirtas surūšiuoti paskolas pagal laiką;

Interest Rate – skaitinis kintamasis – palūkanų norma, kurią moka skolininkas;

Schedule Type – kategorinis kintamasis - mokėjimų tvarkaraštis/grafikas;

Loan Type – kategorinis kintamasis – paskolos tipas;

Property Category – kategorinis kintamasis – įkeičiamo turto (*angl. collateral*) tipo kategorija;

LTV – *Loan-to-Value* santykis – paskolos suma palyginus su turto verte;

Loan Period – paskolos laikotarpis – per kiek laiko skolininkas turi grąžinti paskolą;

Funded Total Amount - skaitinis kintamasis - iš viso paskolai surinkta (finansuota) pinigų suma per „EstateGuru“ investuotojus;

diff – skaitinis kintamasis, sudaromas apskaičiuojant laiko skirtumą tarp datos, kai paskola pradėjo rinkti investicijas (*angl. Funded Date*), ir datos, kai paskola buvo visiškai investuota

(*angl. Fully Invested Date*). Išreikštas valandomis kaip skirtumas tarp šių datų bendros trukmės sekundėmis;

Syndication_total_hours – skaitinis kintamasis, apskaičiuojamas paverčiant „*Syndication Period* „kintamąjį į laiko trukmės formatą ir išreiškiant bendrą sindikavimo laikotarpio trukmę valandomis;

Property Value – skaitinis kintamasis – nekilnojamojo turto vertė, pagal kurią „EstateGuru“ vertina užstatą (*angl. collateral*);

Final Rank – kategorinis kintamasis, sujungtas iš *First Ranking* – pirmoji hipoteka ir *Second Ranking* – antro rango hipoteka;

Stage Exists – kategorinis kintamasis, nusakantis, ar paskola turi kelis etapus;

Refinance – kategorinis kintamasis, nusakantis, ar paskola buvo refinansuota (pvz., pakeistas finansavimo šaltinis, pratęstas terminas).

Makroekonominiai rodikliai. Norint, jog sukurtas modelis geriau atspindėtų realią rinkos situaciją bei pagerintų modelio tikslumą ir bendrą analizę, į duomenų rinkinį įtraukti išoriniai veiksniai, kurie gali turėti įtakos paskolų rizikai. Duomenys paimti iš Eurostat svetainės. Dauguma rodiklių naudojami ne tiesiogiai, o paskaičiuoti jų pokyčiai. Tokios modifikacijos pasirinktos siekiant sumažinti momentinių svyravimų įtaką ir užtikrinti, kad modeliuose būtų užfiksuotas ciklinis bei struktūrinis makroekonomikos rodiklių judėjimas, aktualus kredito rizikos vertinimui. Pokyčiai paskaičiuoti remiantis šiomis formulėmis:

mėnesinis pokytis (*MoM – Month-over-Month*):

$$MoM_t = \left(\frac{X_t - X_{t-1}}{X_{t-1}}\right) \times 100 \qquad (2.1.1)$$

ketvirtinis pokytis (*QoQ – Quarter-over-Quarter*):

$$QoQ_t = \left(\frac{X_t - X_{t-3}}{X_{t-3}}\right) \times 100 \qquad (2.1.2)$$

metinis pokytis (*YoY – Year-over-Year*):

$$YoY_t = \left(\frac{X_t - X_{t-12}}{X_{t-12}}\right) \times 100 \qquad (2.1.3)$$

skirtumas (*angl. difference*) tarp laikotarpių:

$$diff_k(X)_t = X_t - X_{t-k}\ , \qquad (2.1.4)$$

čia

X_t – bet kuris ekonominis rodiklis laiko momentu *t*;

k – laiko poslinkis mėnesiais.

BVP – bendrasis vidaus produktas – tai vienas svarbiausių makroekonominių rodiklių, parodančių bendrą šalies ūkio sukuriamą vertę. BVP apskaičiuojamas sudedant visų per tam tikrą laikotarpį šalyje pagamintų prekių ir suteiktų paslaugų galutinę vertę rinkos kainomis. Formaliai nacionalinėse sąskaitose BVP formulė apibrėžiama kaip sukurta pridėtinė vertė plius produktų mokesčiai minus subsidijos. Eurostat skelbia BVP rodiklius pašalinus kainų įtaką ir pakoregavus sezoniškai bei kalendoriškai, jog duomenys būtų palyginami tarp laikotarpių. Praktiškai BVP naudojamas ekonomikos dydžiui ir augimui įvertinti. Kredito rizikos modeliavimui BVP rodiklis įtrauktas šiais kintamaisiais *YoY, QoQ, MoQ.* Taip pat papildomai įtraukiamas produkcijos atotrūkis (*angl. output gap*), gautas taikant Hodrick-Prescott filtrą logaritminiam BVP, kaip ciklinė jo komponentė:

$$OutputGap_t = c_t \qquad (2.1.5)$$

c_t – ciklinė komponentė;

τ_t – ilgalaikė BVP komponentė;

$$y_t = \log(BVP_t) = \tau_t + c_t \ . \qquad (2.1.6)$$

HPI – būsto kainų indeksas (*angl. House Price Index*) – rodiklis, atspindintis gyvenamojo nekilnojamojo turto kainų pokyčius. Eurostat HPI apskaičiuojamas stebint, kaip kinta būstų kainos laikui bėgant, lyginant su nustatytu baziniu laikotarpiu. Standartiškai skelbiamas indeksas su baziniais metais (pvz., 2015 m. = 100) – jis parodo, kiek procentų pasikeitė vidutinės būsto kainos nuo bazinių metų lygio Euribor (3 mėn. palūkanų norma). Darbe rodiklis buvo transformuotas ir modifikuotas siekiant geriau įvertinti nekilnojamojo turto kainų dinamikos poveikį paskolų kredito rizikai, naudojamos šios modifikacijos: *YoY*, $diff_3$, $diff_{k12}$.

HICP – infliacija (vartotojų kainų indeksas) apibūdina bendrojo prekių ir paslaugų kainų lygio kilimą ekonomikoje. ES infliacijai matuoti naudojamas suderintas vartotojų kainų indeksas (SVKI) (*angl. Harmonised Index of Consumer Prices*) – Eurostato apskaičiuojamas vartotojų kainų indeksas, leidžiantis palyginti infliaciją tarp šalių. Darbe naudojamos šios infliacijos rodiklio modifikacijos: *YoY*, $diff_3$, $diff_{k1}$.

Euribor3 – tai 3 mėnesių Europos tarpbankinė siūloma palūkanų norma (*angl. 3-month Euribor*), plačiai naudojamas finansinis rodiklis, rodantis vidutinę palūkanų normą, už kurią didieji euro zonos bankai sutinka skolinti vieni kitiems eurus 3 mėnesių terminui. Darbe naudojama ši rodiklio modifikacija: $diff_1$.

PI – gamybos indeksas (*angl. production index*) parodo metinį produkcijos augimą / kritimą. Tai vienas iš aiškiausių verslo ciklo indikatorių, kuris padeda vertinti, ar ekonominė aplinka gerėja ar blogėja. Kredito rizikos požiūriu naudingas, nes ekonomikos lėtėjimas ar kritimas dažnai blogina paskolų portfelio kokybę. Darbe naudojama ši rodiklio modifikacija: *YoY.*

2.2. Tikslo kintamojo apibrėžimas

Šiame tyrime tikslo (angl. target) kintamasis apibrėžiamas kaip dvejetainis kintamasis, nusakantis paskolos grąžinimo kokybę. Pagrindinis tikslas – identifikuoti nemokias paskolas, todėl tikslo kintamasis konstruojamas remiantis paskolos galutiniu arba esamu statusu „EstateGuru“ platformos duomenyse.

Formaliai „*default*“ (nevykdymo) būsena platformoje siejama su paskolos sutarties nutraukimu ir išieškojimo proceso inicijavimu (statusas „*Defaulted*“). Tačiau analizuojamame duomenų rinkinyje šis statusas nėra naudojamas. Dėl šios priežasties nemokumo įvykis formuojamas remiantis alternatyviais paskolos būsenos statusais, kurie atspindi reikšmingus nukrypimus nuo sutartinių įsipareigojimų vykdymo.

Tikslo kintamasis apibrėžiamas taip:

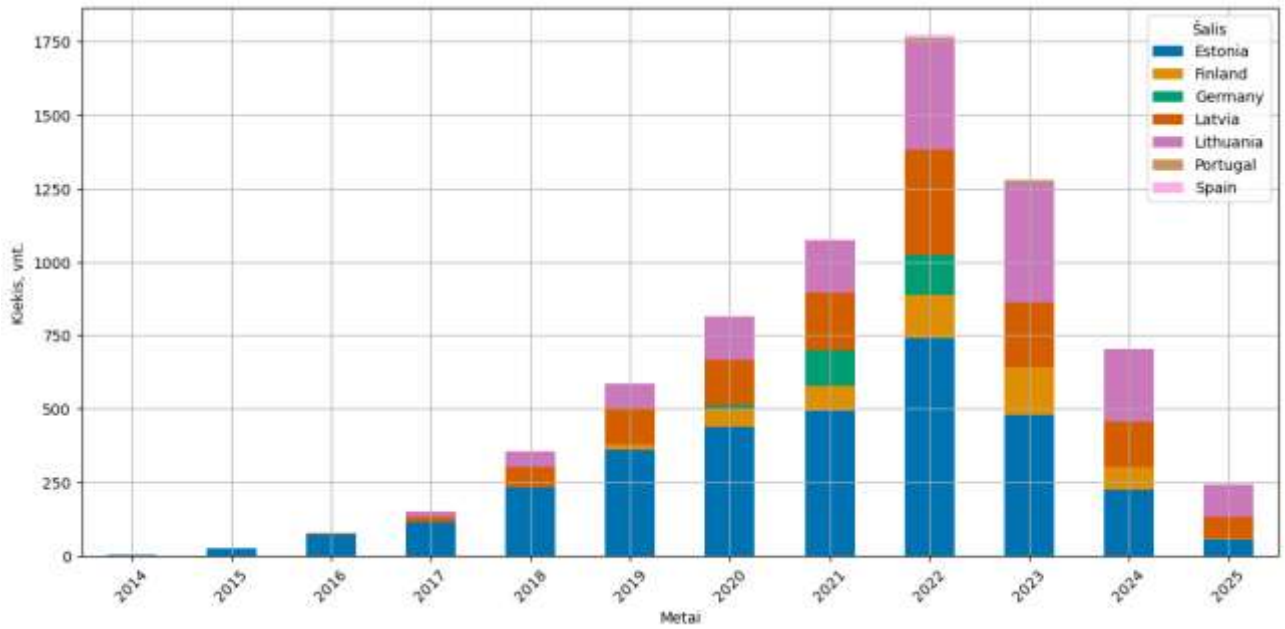
- Y = 0 (moki paskola), jei paskolos statusas yra „Grąžinta“ (angl. Repaid), t. y. skolininkas įvykdė savo įsipareigojimus pagal sutartį;
- Y = 1 (nemoki paskola), jei paskolos statusas atitinka bent vieną iš šių kategorijų: „Vėluoja“ (angl. Late), „Išieškota“ (angl. Recovered), „Nurašyta“ (angl. In Default), „Iš dalies grąžinta nesumokėta suma“ arba „Iš dalies grąžinta sumokėta suma“ (angl. Partially recovered). Šie statusai apima tiek aktyvų įsipareigojimų nevykdymą (vėlavimą), tiek jau realizuotus kredito nuostolius ar dalinį jų susigrąžinimą.

Į modelio konstravimą neįtraukiamos paskolos su statusais „Išmokėta“ (angl. Funded), „Investuota“ (angl. Fully Invested) ir „Atidaryta“ (angl. Open), kadangi šios paskolos yra aktyvios ir jų galutinė grąžinimo būsena dar nėra žinoma analizės momentu.

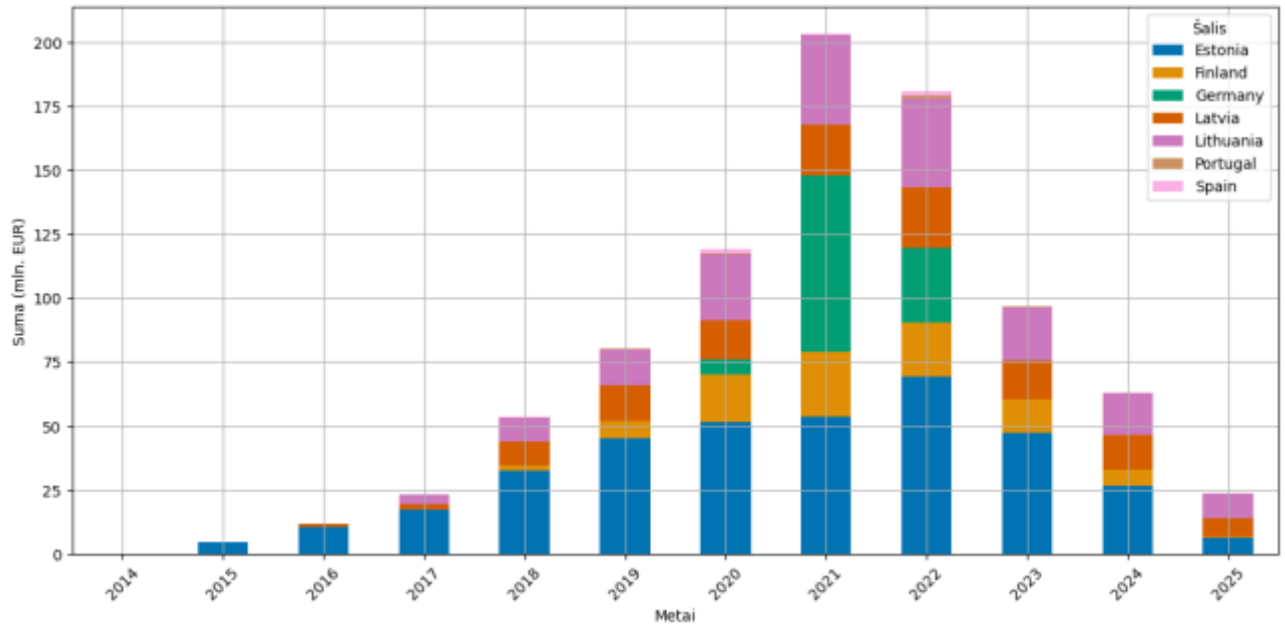
2.3. Duomenų žvalgomoji analizė

Šiame etape atliekama aprašomoji duomenų analizė, kurios tikslas – identifikuoti pagrindinius dėsningumus paskolų duomenims bei preliminariai įvertinti kintamųjų ryšį su kredito rizika.

Paskolų kiekis ir finansavimo apimtys augo nuo 2017 m., piką pasiekdamos 2021–2022 m., po kurių 2023–2025 m. stebimas aiškus aktyvumo sumažėjimas (2 pav. ir 3 pav.). Didžiausią paskolų skaičių ir sumą generavo Estija ir Lietuva, kurios tapo pagrindinėmis platformos rinkomis viso analizuojamo laikotarpio metu. Tai rodo stiprią geografinę koncentraciją ir ciklišką paskolų rinkos vystymąsi.

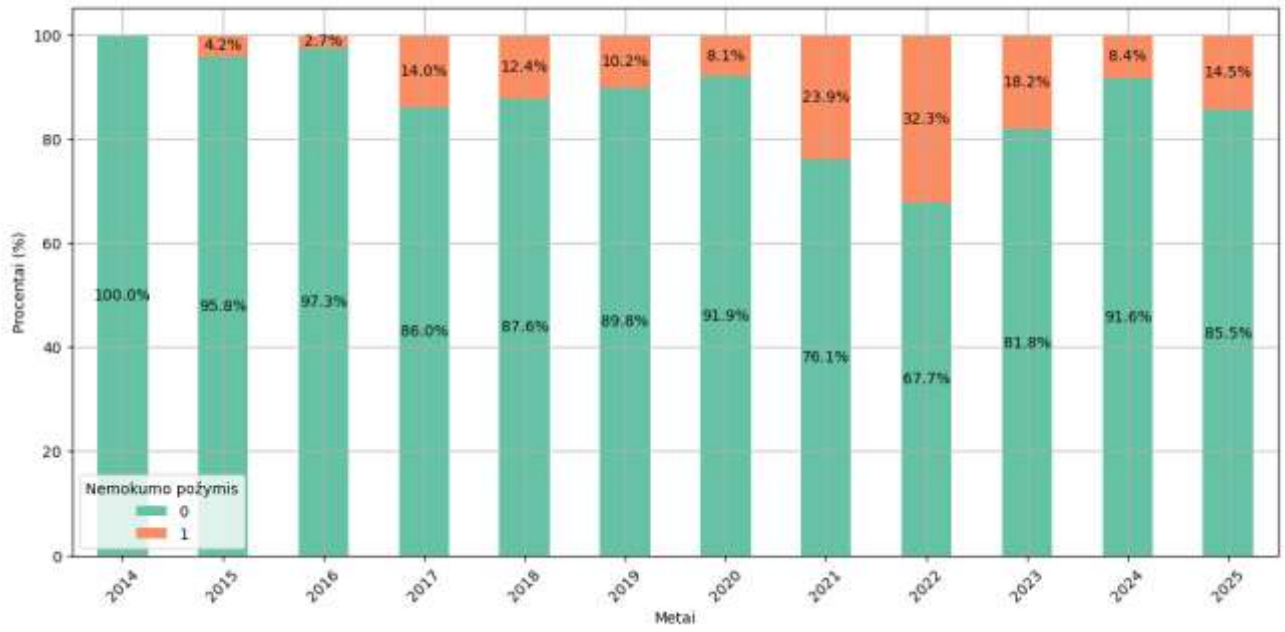


2 pav. Paskolų kiekis pagal metus ir šalis

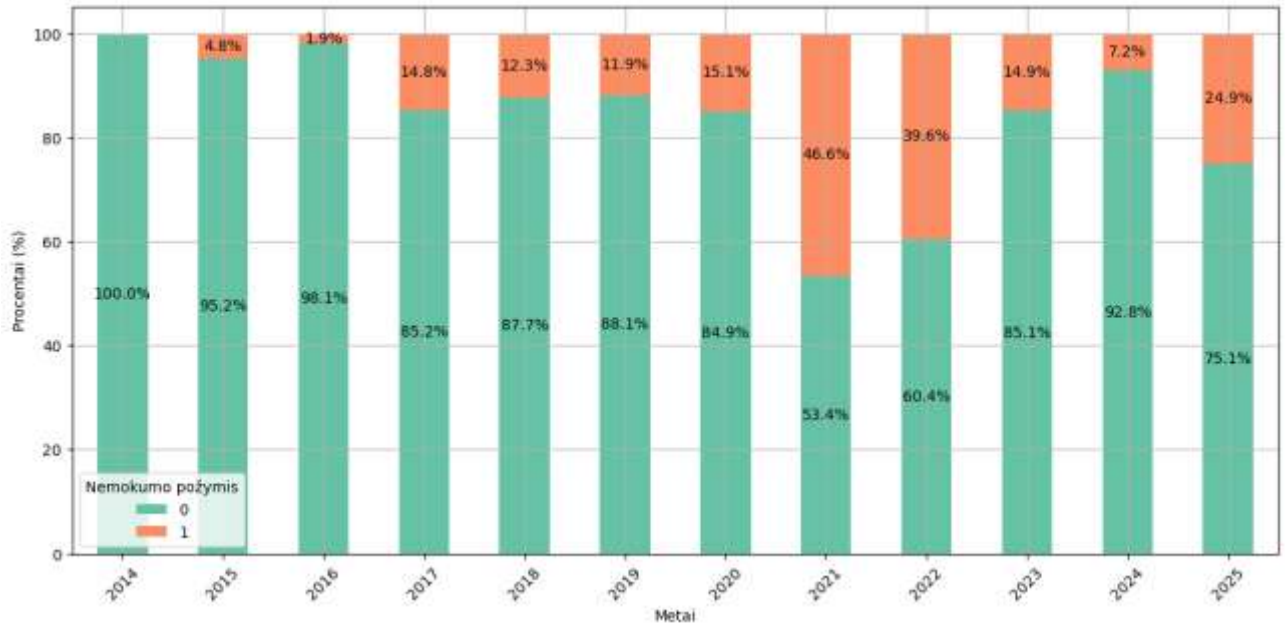


3 pav. Paskolų suma pagal metus ir šalis

Nemokių paskolų dalis reikšmingai padidėjo 2021–2022 m. (4 pav., 5 pav.), kai tiek pagal paskolų kiekį, tiek pagal sumą blogų paskolų dalis buvo didžiausia visame laikotarpyje. Pagal sumą šis efektas yra dar ryškesnis, kas leidžia daryti ir prielaidą, jog nemokumas buvo susijęs ir su didesnės vertės projektais. Vėlesniais metais (2023–2024 m.) nemokių paskolų dalis vėl sumažėjo, tačiau išlieka didesnė nei ankstyvaisiais metais.

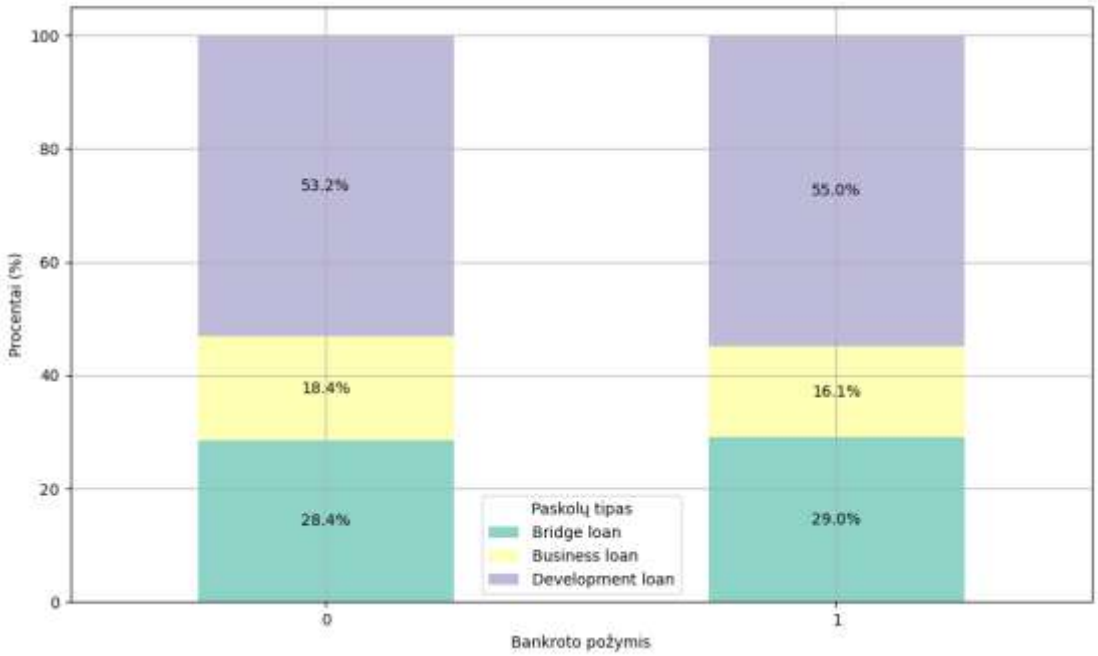


4 pav. Paskolų kiekis pagal metus ir nemokumo požymį



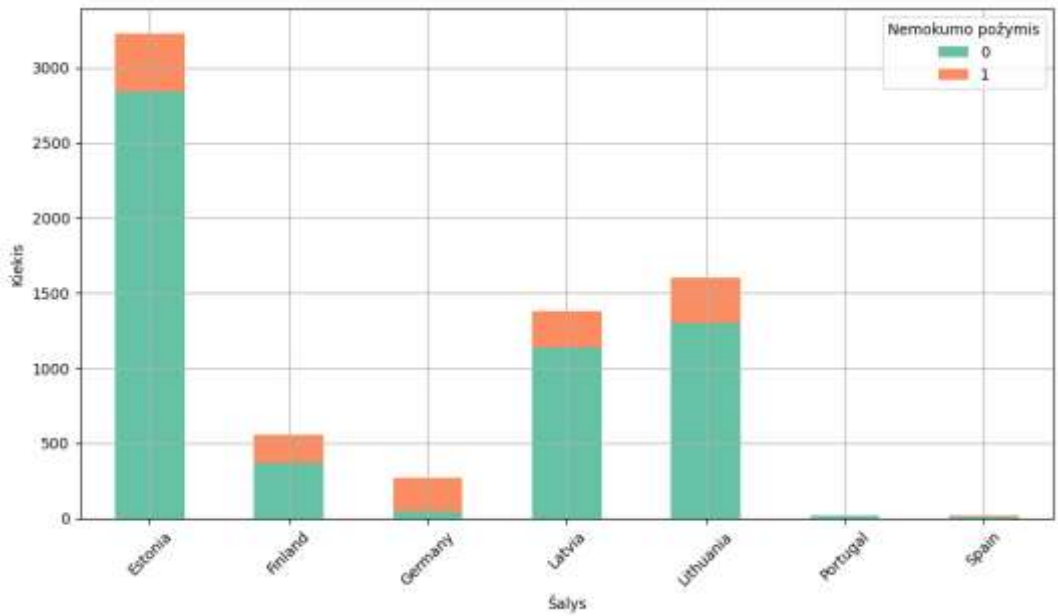
5 pav. Paskolų suma pagal metus ir nemokumo požymį

Analizuojant paskolų tipus pagal nemokumo statusą (6 pav.), pastebima, kad tipų pasiskirstymas tarp mokių ir nemokių paskolų yra labai panašus. Vystymo (*angl. Development*) paskolos sudaro didžiausią dalį abiejose grupėse, o tarpinė paskola (*angl. Bridge*) ir verslo (*angl. Business*) paskolų proporcijos reikšmingai nesiskiria. Tai rodo, kad vien paskolos tipas savaime nėra pakankamas nemokumo rizikos diskriminatorius.



6 pav. Paskolų tipai pagal nemokumo požymį

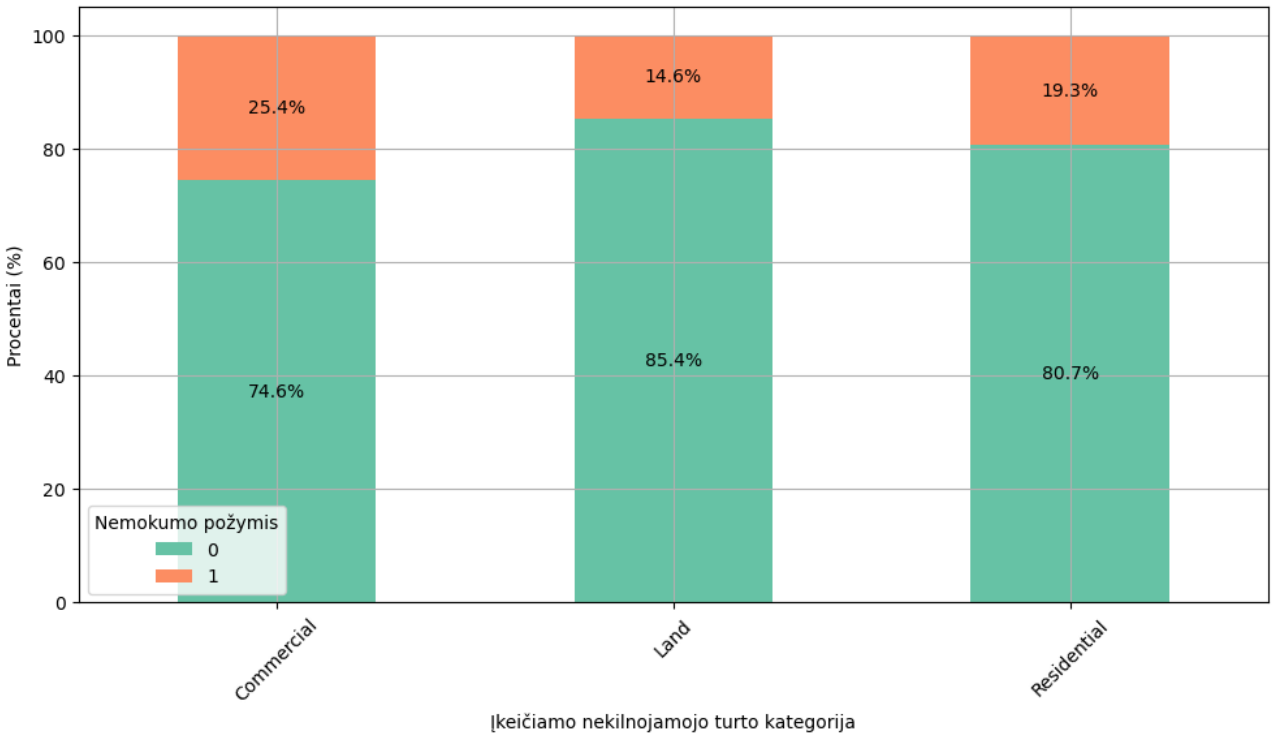
Didžiausias absoliutus nemokių paskolų skaičius fiksuojamas šalyse, kuriose yra didžiausias bendras paskolų kiekis – Estijoje, Lietuvoje ir Latvijoje (7 pav.). Tuo tarpu kitose šalyse (Vokietijoje, Suomijoje, Portugalijoje, Ispanijoje) nemokių paskolų skaičius yra ženkliai mažesnis, kas iš dalies atspindi mažesnį bendrą paskolų aktyvumą. Tai rodo, kad nemokumo koncentracija glaudžiai susijusi su rinkos dydžiu.



7 pav. Nemokumo pasiskirstymas pagal šalis

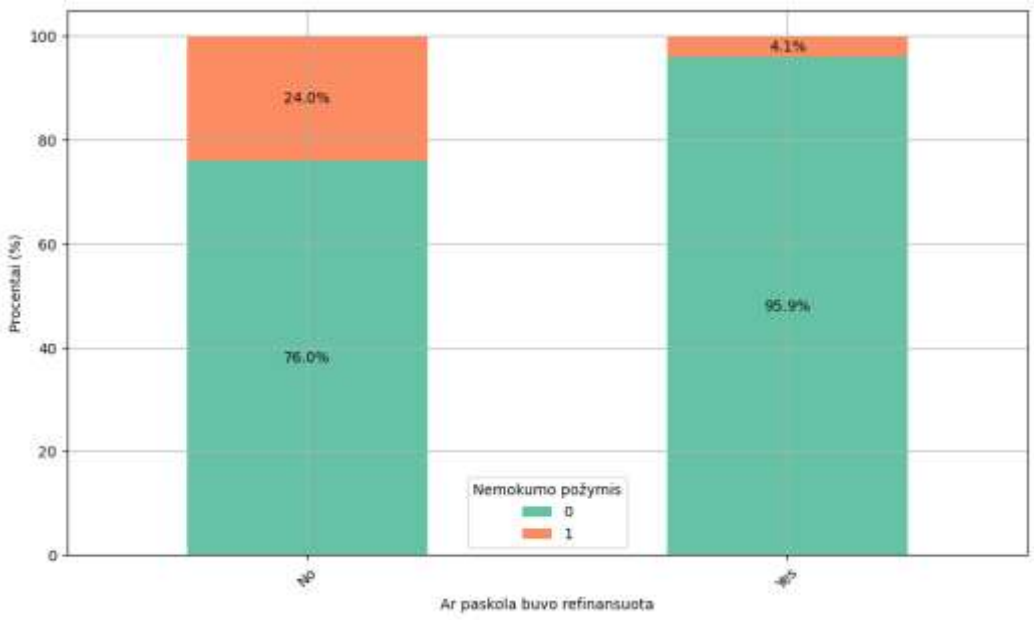
Paskoloms, kurių įkeistas turtas yra komercinis (*angl. Commercial*) nekilnojamas turtas, būdinga didžiausia nemokių paskolų dalis, palyginti su gyvenamosios paskirties (*angl. Residential*) ir žemės (*angl. Land*) tipo įkeistu turtu (8 pav.). Žemės tipo įkeistas turtas

pasižymi mažiausia nemokių paskolų dalimi. Ši analizė leidžia daryti prielaidą, kad įkeičiamo nekilnojamojo turto kategorija gali būti reikšmingas kredito rizikos veiksnys.



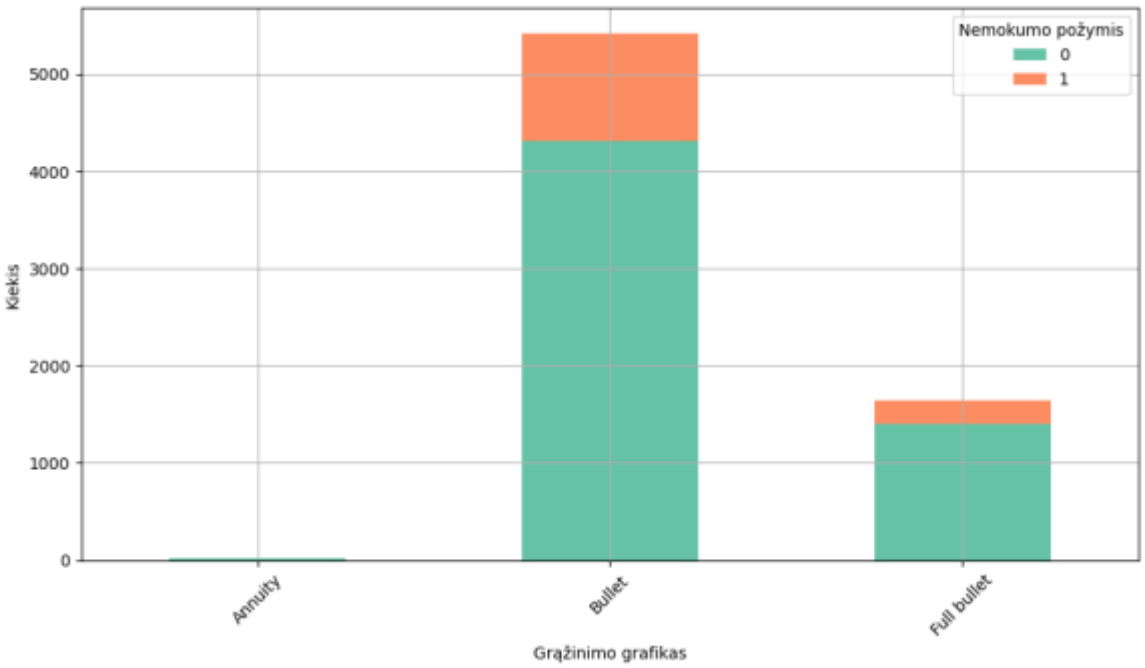
8 pav. Nemokumo pasiskirstymas pagal įkeičiamo nekilnojamojo turto kategoriją

Refinansuotos paskolos pasižymi žymiai mažesne nemokių paskolų dalimi nei nerefinansuotos paskolos (9 pav.). Tai gali indikuoti, kad refinansavimas veikia kaip riziką mažinantis mechanizmas, susijęs su geresniais projekto finansiniais rodikliais arba aktyviu paskolos restruktūrizavimu. Šis požymis turi aiškią diskriminacinę galią.

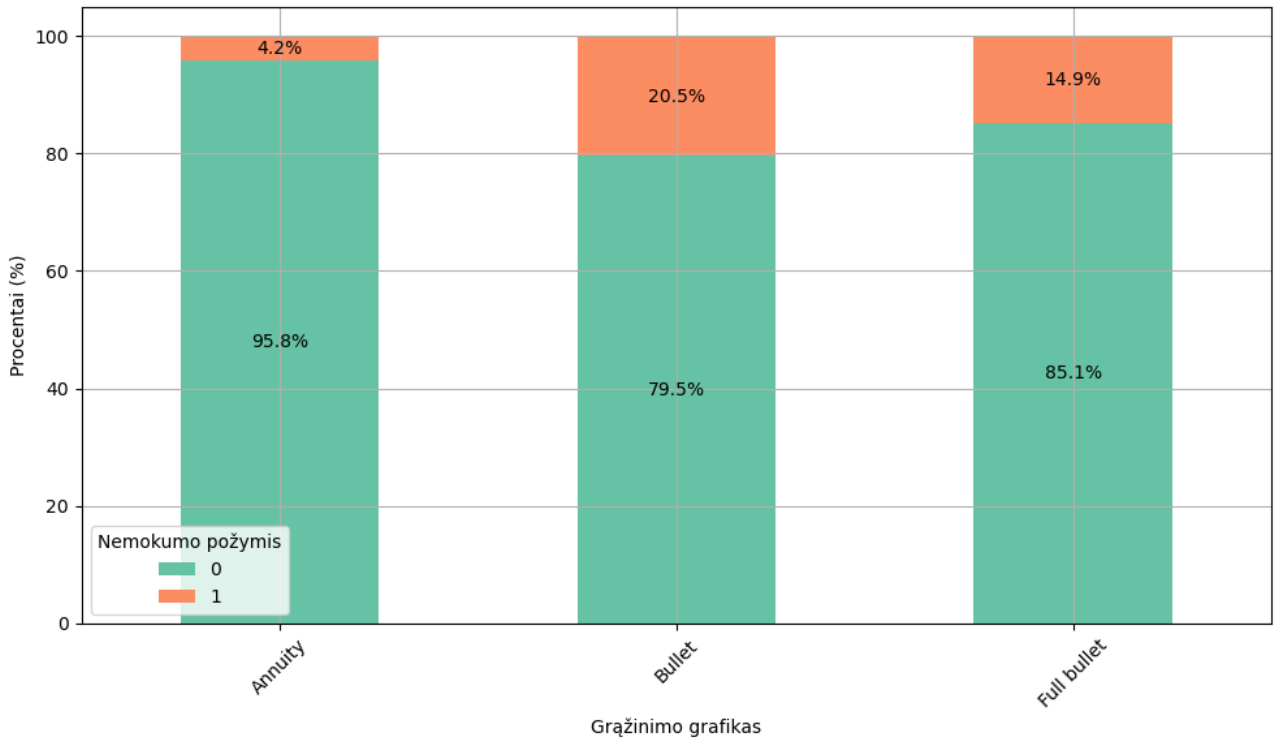


9 pav. Nemokumo pasiskirstymas pagal refinansavimą

Termino pabaigoje grąžinamų (*angl. Bullet*) tipo paskolos pasižymi didžiausia nemokių paskolų dalimi tiek absoliučiais skaičiais (10 pav.), tiek procentine išraiška (11 pav.). Anuitetinių paskolų nemokumo dalis yra mažiausia, o visos paskolos suma grąžinama termino pabaigoje tipas užima tarpinę poziciją. Tai rodo, kad grąžinimo grafiko struktūra gali būti svarbus kredito rizikos veiksnys.



10 pav. Nemokumo pasiskirstymas pagal grąžinimo grafiko tipą

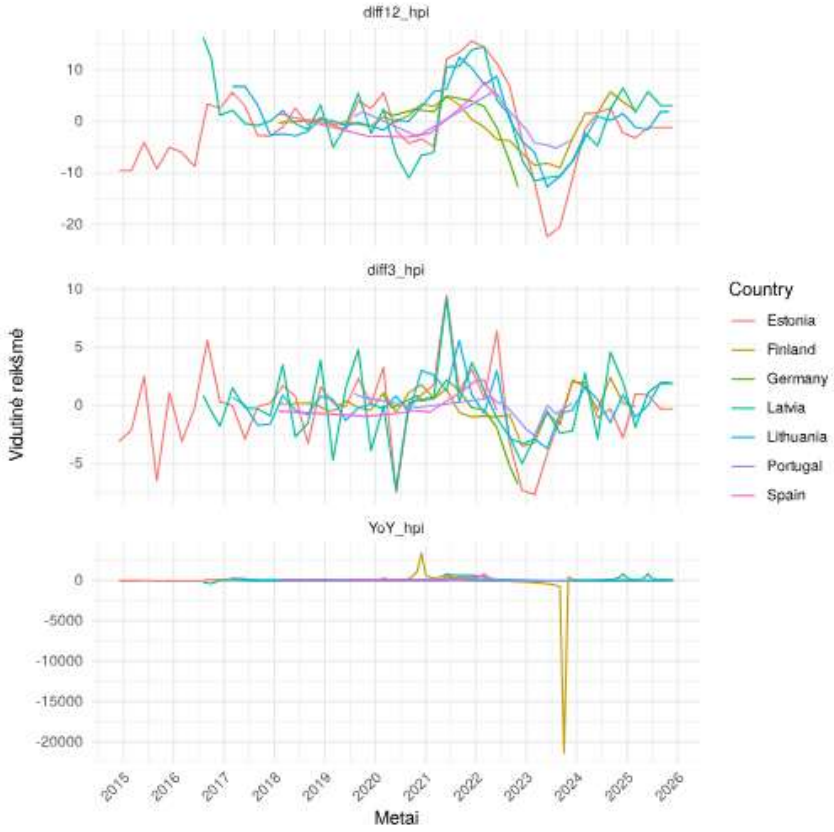


11 pav. Nemokumo pasiskirstymas pagal grąžinimo grafiko tipą, %

Iš duomenų aprašomosios analizės matome, kad kai kurie kintamieji turi daugiau įtakos nemokumo faktui, tačiau kai kurie pasiskirstę labai panašiai. Vis dėlto yra svarbu, kad kai kurie kintamieji gali turėti įtakos ne atskirai, o kartu su kitu kintamuoju. Tam, kad būtų nustatyta, kas lemia paskolų nemokumą, ir yra atliekamas kredito rizikos modeliavimas, remiantis ML algoritmais.

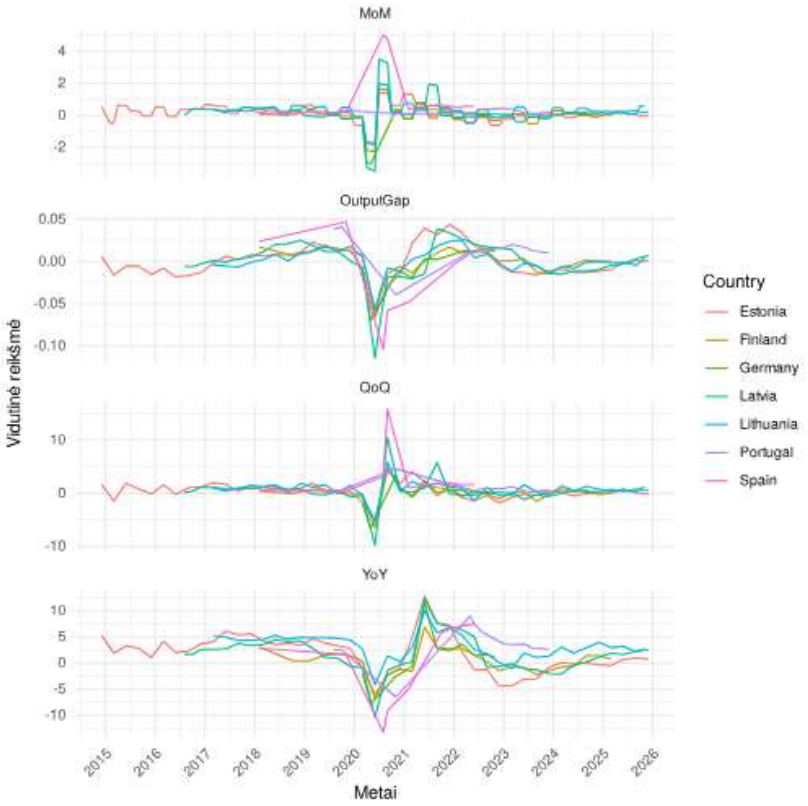
Makroekonominių rodiklių analizė. Šiame skyriuje analizuojami pagrindiniai makroekonominiai rodikliai – nekilnojamojo turto kainų indeksas (HPI), bendrasis vidaus produktas (BVP), EURIBOR3 palūkanų norma ir infliacija – siekiant įvertinti jų dinamiką skirtingose šalyse ir galimą ryšį su kredito rizika.

HPI rodikliai (12 pav.) atskleidžia ryškų nekilnojamojo turto kainų augimą 2020–2022 m., po kurio seka staigus kritimas 2022–2023 m. laikotarpiu. Baltijos šalys pasižymi didesniais svyravimais, kas rodo didesnį rinkos jautrumą ekonominiams ciklams. Ši dinamika leidžia tikėtis reikšmingo HPI poveikio kredito rizikai.



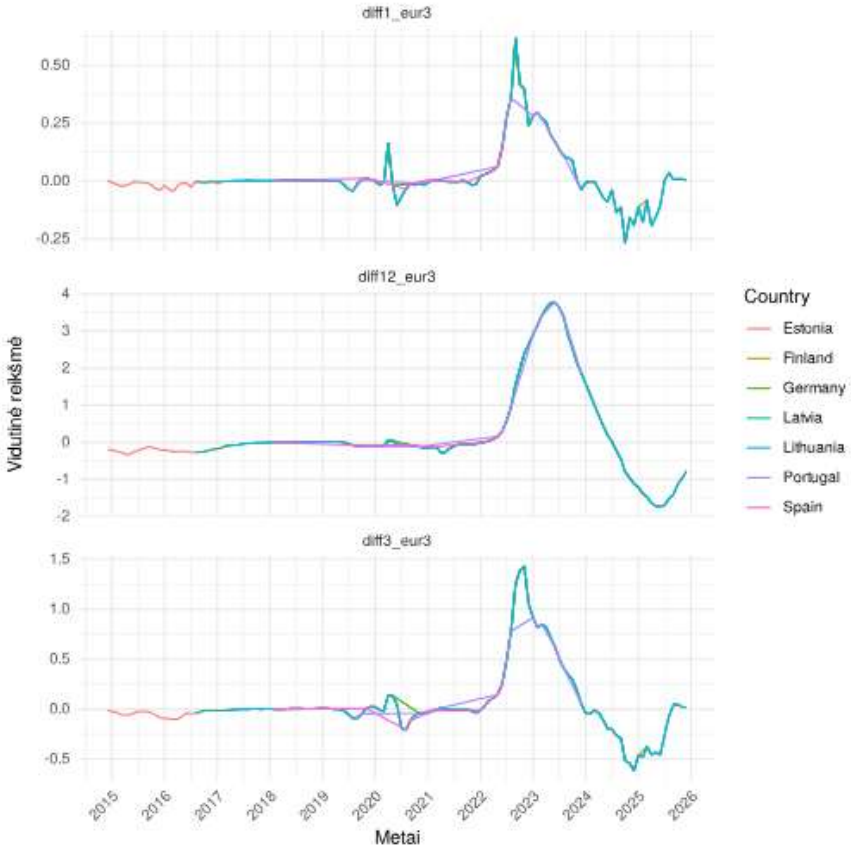
12 pav. HPI rodiklių kreivės pagal šalį

BVP rodikliai (13 pav.) rodo aiškų ekonominį kritimą 2020 m. ir stiprų atsigavimą 2021 m., kurį lydėjo teigiamas produkcijos atotrūkis. Vėlesniu laikotarpiu ekonomikos augimas stabilizuojasi, o skirtingų šalių rodikliai tampa panašesni. Tai rodo, kad ekonominiai ciklai yra sinchroniški tarp šalių.



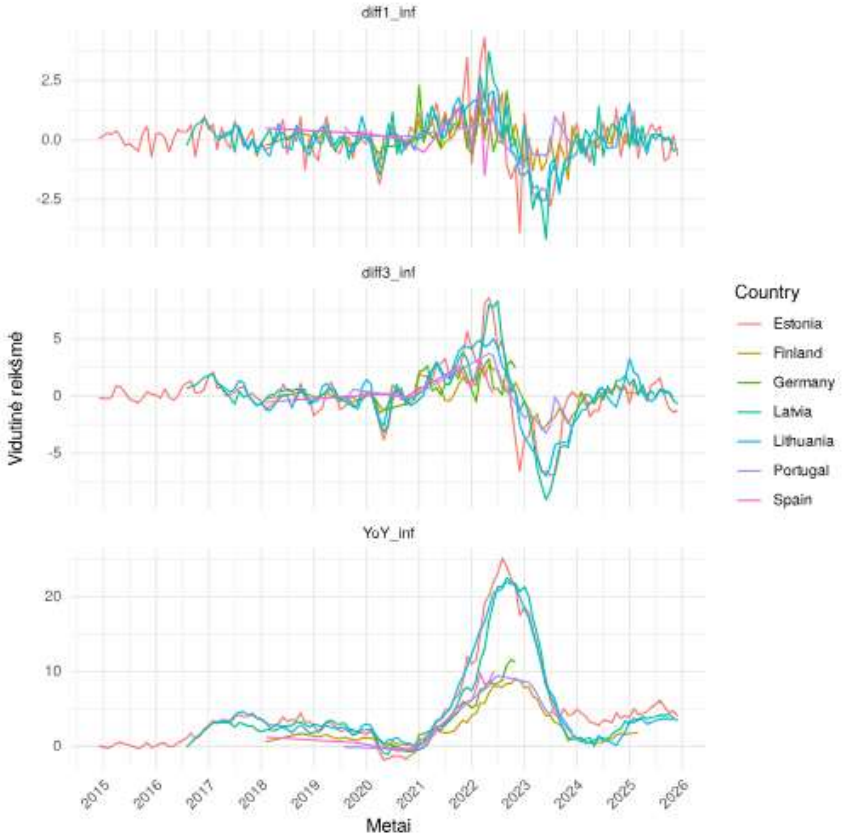
13 pav. BVP rodiklių kreivės pagal šalį

EURIBOR3 (14 pav.) ilgą laiką išliko žemas, tačiau nuo 2022 m. stebimas staigus palūkanų normų augimas, vėliau pereinantis į mažėjimo fazę. Šis pokytis yra vienas ryškiausių per visą laikotarpį ir yra vienodas visose šalyse. Tai indikuoja stiprų pinigų politikos poveikį paskolų sąlygoms.



14 pav. Euribor3 rodiklių kreivės pagal šalį

Infliacija reikšmingai išaugo 2021–2022 m., ypač Baltijos šalyse, ir vėliau sumažėjo (15 pav.). Tarp šalių skirtumai nėra dideli, tačiau svyravimų amplitudė skiriasi. Infliacijos dinamika atspindi bendrą makroekonominį nestabilumą nagrinėjamu laikotarpiu.



15 pav. Infliacijos rodiklių kreivės pagal šalį

Iš makroekonomikos rodiklių analizės, matome, kad 2020-2022 m. stebimi šuoliai visuose rodikliuose. Verta atkreipti dėmesį, kad makroekonomikos kintamųjų pasiskirstymai apmokymo ir testavimo laikotarpiuose skiriasi: 2014-2023 m. apmokymo imtis, nuo 2023 m. testavimo imtis. Vadinasi, modeliai apmokomi su ekstremaliomis sąlygomis, o testavimo duomenyse ekonominė būsena yra stabilesnė. Toks laikotarpio pasiskirstymas gali sukelti duomenų pasiskirstymo pokytį (*angl. distribution shift),* todėl modelio veikimas testavimo imtyje gali skirtis nuo mokymo rezultatų, ypač esant mažesniam makroekonominių svyravimų lygiui. Modelis gali pervertinti riziką testavimo laikotarpiu ir tampa konservatyvesnis, t. y. per daug atsargus ir linkęs priskirti didesnes PD vertes.

2.4. Logistinė regresija

Logistinė regresija – tai statistinis klasifikavimo metodas, kuris prognozuoja binarinio (dvejetainio) atsako kintamojo tikimybę pagal vieną ar daugiau nepriklausomųjų kintamųjų [39]. Tai regresijos metodas, naudojamas klasifikacijai, todėl dažnai vadinamas generalizuotu linijiniu modeliu (GLM) su logistine jungtimi (*angl. logit link function*). Modelio tikslas – įvertinti tikimybę, kad stebėjimas priklauso taikinio klasei (pvz., paskola nebus išmokėta).

Tikimybė modeliuojama per logistinės regresijos lygtį:

$$P(Y = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}}, \qquad (2.4.1)$$

kur

- $P(Y = 1|X)$ – įvykio tikimybė;
- X = ($x_1, x_2, \ldots, x_k$) – nepriklausomi kintamieji;
- $\beta_0, \beta_1, \ldots, \beta_k$– modelio koeficientai.

Ši funkcija užtikrina, kad modelio reikšmės būtų ribotos intervale (0,1), t. y. atitiktų tikimybes.

Tikimybė gali būti perrašyta kaip logit funkcija – logaritminė tikimybės ir atvirkštinės tikimybės (*angl. odds*) santykio transformacija:

$$\log it(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_j \ , \qquad (2.4.2)$$

čia:

- p – atsitiktinio įvykio tikimybė;
- $\frac{p}{1-p}$ – galimybės (*angl. odds*) santykis, t. y. tikimybės p ir priešingos tikimybės (1-p) santykis;
- β_0 – laisvasis narys;
- β_i – koeficientai, įvertinantys nepriklausomų kintamųjų poveikį;
- X_i – nepriklausomi kintamieji.

Ši formulė leidžia suprasti, kaip kiekvienas požymis proporcingai veikia įvykio tikimybes, t. y., kiek kartų padidėja ar sumažėja tikimybė esant vieno vieneto pokyčiui kintamajame.

Tikimybė apskaičiuojama pagal *logit* funkciją:

$$p = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_k X_k)}} \qquad (2.4.3)$$

Logistinės regresijos koeficientai ***β*** įvertinami tikėtinumo metodu. Logaritminė tikėtinumo funkcija:

$$l(\beta) = \sum_{i=1}^{n} [y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)]; \qquad (2.4.4)$$

čia

- $y_i \in \{0,1\}$ – stebėtas priklausomojo kintamojo rezultatas (klasė);
- $p_i = P(Y_i = 1 \,|X_i)$ – modelio prognozuojama įvykio tikimybė, kad klasė yra taikinio (Y = 1).

Klasifikavimo sprendimo priėmimas. Nors modelis pateikia p̂ = P(Y = 1|X), klasifikacijai būtina pasirinkti sprendimo slenkstį (*angl. threshold*) ***θ***, kuriuo remiantis nepriklausomas kintamasis bus priskirtas 1 arba 0 klasei. Jam esant, kiekvienam stebėjimui priimamas klasifikavimo sprendimas:

$$\hat{y} = \begin{cases} 1, \ jei \ \hat{p} \geq \theta \\ 0, \ jei \ \hat{p} < \theta \end{cases} \qquad (2.4.5)$$

Pagal nutylėjimą dažnai naudojamas $\theta = 0,5$, tačiau tai nėra optimali reikšmė visose situacijose. Slenkstis gali būti adaptuojamas atsižvelgiant į:

- duomenų klasių disbalansą, kai viena klasė pastebimai retesnė;
- klaidų kainą ar pasekmes, t. y. kai klaidingai priskirta klasė sukelia nevienodą riziką;
- norimą optimizuoti metriką;
- verslo taisykles, pavyzdžiui, finansų sektoriuje ribos gali būti griežtesnės siekiant sumažinti nuostolius.

Vienas iš logistinės regresijos privalumų – lengva interpretacija:

- $\beta i > 0$: didėjant x_i , tikimybė, kad Y = 1, auga;
- $\beta i < 0$: didėjant x_i, tikimybė mažėja;
- e^{β_i} : tikimybių santykis (*angl. odds ratio*) – kiek kartų padidėja tikimybė, kai klasė yra taikinio (Y = 1) padidinus x_j vienu vienetu.

Logistinė regresija yra vienas iš pagrindinių metodų naudojamų kredito rizikos vertinime (*angl. Credit scoring*). Ji taikoma vertinant paskolos gavėjo tikimybę sulaukti paskolos išmokėjimo (arba nemokumo). Šis metodas leidžia modeliuoti tikimybes ir derinti sprendimus su verslo logika. Taip pat pateikia interpretacinius rodiklius, pvz., svarbiausius kintamuosius, jų įtaką, tikimybių santykius.

2.5. *Elastic Net* reguliarizacija

Reguliarizacija yra statistinis metodas, skirtas sumažinti modelio sudėtingumą ir pagerinti jo gebėjimą generalizuoti. Tai svarbu tais atvejais, kai:

- daug požymių (didelis p);
- mažai stebėjimų (mažas n);
- yra stiprių tarpusavio koreliacijų tarp požymių.

Tokiose situacijose tradicinė logistinė regresija be reguliarizacijos gali sukelti persimokymą (*angl. overfitting*) ir menkai identifikuoti reikšmingus kintamuosius.

Elastic Net – tai reguliarizacijos metodas, skirtas kintamųjų atrankai ir modelio reguliavimui [40]. Jis ypač naudingas, kai kintamųjų (p) yra daugiau nei stebėjimų (n), arba kai tarp kintamųjų yra stipri koreliacija. Šis metodas apjungia *Lasso* (L1) ir *Ridge* (L2) regresijos privalumus.

Ridge (L2) metodas sumažina koeficientų dydžius tam, kad sumažintų dispersiją:

$$\hat{\beta}_{ridge} = \arg\min_{\beta} \quad \left[-\ell(\beta) + \lambda \sum_{j=1}^{p} \beta_j^2 \right]. \qquad (2.5.1)$$

Savybės:

- maži koeficientai, nė vienas netampa tiksliai nuliu;
- skiria dideles baudas dideliems parametrams;
- stabilus esant kolinearumui;
- nenaudoja požymių atrankos.

Lasso (L1) metodas padeda atrinkti svarbiausius kintamuosius, skatindamas kai kurių koeficientų tapimą lygiems nuliui:

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \left[-\ell(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right]. \qquad (2.5.2)$$

Savybės:

- atliekamas parametrų parinkimas;
- daugelis koeficientų lygūs 0;
- nestabilus, kai požymiai koreliuoti;
- gali pasirinkti tik ne daugiau nei *n* kintamųjų.

Kai taikome logistinę regresiją, modelio parametrai β (koeficientai) paprastai parenkami taip, kad būtų padidinta logaritminė tikimybė (*angl. log-likelihood*) pagal stebėtus duomenis:

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \ln(\pi_i) + (1 - y_i) \ln(1 - \pi_i)], \pi_i = \frac{1}{1+\exp(-x_i^{\top}\beta)}. \quad (2.5.3)$$

Tačiau be jokių apribojimų modelis linkęs persimokyti, ypač kai požymių daug, jie koreliuoti ar duomenys triukšmingi. Todėl į optimizavimo uždavinį įtraukiama baudos funkcija (*angl. penalty function*), kuri reguliuoja parametrų dydį. *Elastic Net* atveju naudojama dviejų baudų suma:

- $\lambda_1 \| \beta \|_1 \rightarrow$ L1 norma (*Lasso* dalis);
- $\lambda_2 \| \beta \|_2^2 \rightarrow$ L2 norma (*Ridge* dalis).

Elastic Net metodas jungia abiejų metodų savybes. Logistinėje regresijoje *Elastic Net* bauda integruojama į logaritminės tikimybės optimizavimą (*angl. log-likelihood*):

$$\hat{\beta} = \arg\min_{\beta} \{ -l(\beta) + \lambda_2 \| \beta \|_2^2 + \lambda_1 \| \beta \|_1 \}; \qquad (2.5.4)$$

čia

λ_1 – *Lasso* (L1) baudos koeficientas;

λ_2 – *Ridge* (L2) baudos koeficientas;

β – regresijos koeficientai.

Elastic Net bauda yra L1 ir L2 baudos kombinacija: L1 bauda skatina parametrų vektoriaus retumą, kitaip tariant – daugelio regresijos koeficientų reikšmės tampa lygios nuliui, o modelyje išlaikomi tik statistiškai reikšmingi kintamieji. Taip pat šis metodas palaiko vadinamąjį grupinį efektą, kai stipriai koreliuojantys požymiai linkę būti atrenkami kartu ir

priskiriant panašius koeficientus, taip išlaikant jų struktūrinį vientisumą modelyje. Dėl L2 komponentės šis metodas sumažina koeficientų dispersiją ir padidina parametrų stabilumą, todėl rezultatai tampa mažiau jautrūs duomenų pokyčiams ir multikolinearumui.

Parametrų parinkimas. *Elastic Net* metodui tenka parinkti du reguliavimo parametrus:

- λ_1 – L1 baudos stiprumas;
- λ_2 – L2 baudos stiprumas.

Kartais literatūroje jie parametrizuojami alternatyviai:

- $\alpha = \lambda_1 + \lambda_2$–bendras baudos dydis;
- $\gamma = \frac{\lambda_1}{\lambda_1+\lambda_2}$–L1 dalies proporcija.

Šį parametrizavimą taip pat rekomenduoja Zou ir Hastis [40], tai padeda labiau interpretuoti grupinį efektą bei elgesį ribinėse situacijose.

Optimalūs $\lambda_1$ir λ_2 parenkami empiriškai, paprastai naudojant K – dalių kryžminę patikrą.

Bendras optimizavimo principas:

$$(\hat{\lambda}_1, \hat{\lambda}_2) = \arg\min_{\lambda_1,\lambda_2}[CV_{loss(\lambda_1,\lambda_2)}], \qquad (2.5.5)$$

čia CV_{loss} – kryžminės patikros validavimo klaida, kuri logistinei regresijai gali būti matuojama:

- neigiamu $\ell(\beta)$;
- AUC kriterijumi;
- klasifikavimo klaida.

Zou ir Hastis pabrėžia, kad *Elastic Net* ypač naudingas, kai atliekamas modelio pasirinkimas, todėl reguliavimo parametrai gali būti pasirenkami taip, kad:

- būtų minimizuojama validavimo klaida;
- kartu kontroliuojamas kintamųjų atrankos kiekis.

2.6. Atsitiktinio miško modeliai

Atsitiktiniai miškai yra priskiriami metodams “ansambliams” (*angl. ensemble methods)* ir yra skirti tiek klasifikavimo, tiek regresijos uždaviniams spręsti. Metodą pasiūlė Leo Breimanas [41], išplėsdamas *bagging* idėją, papildydamas ją atsitiktiniu požymių parinkimu kiekviename sprendimų medžio šakojimosi taške. Vienai apmokymo imčiai sudaromas ne vienas medis, o daug sprendimų medžių (pvz. 100 ar 1000). Tam, kad galėtume sudaryti medžius, kurie nebūtų identiški, medžių kūrimo metu naudojamas atsitiktinumas (*angl. randomization*).

Apibrėžimas. Atsitiktinis miškas – tai klasifikatorius, sudarytas iš daugybės medžio struktūros klasifikatorių. Kiekvienas medis auginamas pagal atsitiktinį vektorių Θ_k, kur Θ_k , k = *1, …, L*, yra nepriklausomi ir vienodai pasiskirstę. Kiekvienas medis „balsuoja“ už populiariausią klasę, kai pateikiamas įėjimas *x*.

Pagal šį apibrėžimą, atsitiktinis miškas gali būti kuriamas imant mėginius iš požymių rinkinio, iš duomenų rinkinio arba tiesiog atsitiktinai keičiant kai kuriuos medžio parametrus. Bet koks šių įvairovės šaltinių derinys taip pat sudarys atsitiktinį mišką. Pavyzdžiui, galima sudaryti imtis ir iš požymių rinkinio, ir iš duomenų rinkinio.

Atsitiktinių miškų algoritmas apima šiuos žingsnius:

1. imties su grąžinimu *(angl. bootstrap)* generavimas: iš pradinės duomenų aibės generuojama *B* atsitiktinių imčių su grąžinimu. Kiekviena imtis naudojama atskiram medžiui T_b mokyti;
2. atsitiktinis požymių parinkimas: kiekviename medžio mazge (*ang. split*) ne visi požymiai yra tiriami – vietoj to, atsitiktinai pasirenkamas požymių poaibis (iš *p* galimų), paprastai dydžio $m \ll p$, ir iš jų pasirenkamas geriausias skaidymo kriterijus;
3. medžių formavimas: kiekvienas medis auginamas iki galo (be genėjimo), naudojant tik jam priskirtą duomenų ir požymių pogrupį. Medžiai yra nepriklausomi vienas nuo kito, nes kiekvienas gauna skirtingą duomenų ir požymių kombinaciją;
4. agregavimas:
   - Klasifikavimo atveju naudojamas daugumos balsavimas:
     $\dot{C_{RF}}(x) = mode\{T_b(x)\}_{b-1}^{B}$ (2.6.1)
   - Regresijos atveju – vidurkio skaičiavimas:
     $f_{RF}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$; (2.6.2)

čia:

x – naujas stebėjimas (įrašas), kurio prognozė skaičiuojama. Tai vektorius su požymių reikšmėmis;

T_b *(x)* – *b*-tasis sprendimų medis taiko savo taisykles įrašui *x* ir pateikia prognozę;

b – indeksas, žymintis konkretų medį ansamblyje (nuo 1 iki B);

B – iš viso sukurtų sprendimų medžių skaičius atsitiktiniame miške;

mode – dažniausiai pasitaikanti reikšmė tarp visų medžių prognozių (klasifikavime – balsavimas).

Vienas esminių rezultatų – modelio generalizavimo paklaida (*angl. generalization error*) yra ribojama, kai tik medžių skaičius $B \to \infty$.

Klasifikavimo atveju paklaidos viršutinė riba priklauso nuo:

- stiprumo (*angl. strength)* – tai kiek vidutiniškai geri yra individualūs medžiai;

- koreliacijos tarp medžių – norima, kad medžiai būtų kuo mažiau koreliuoti.

Viršutinė generalizavimo paklaidos riba įvertinama taip:

$$PE \leq \frac{\rho(1-s^2)}{s^2}, \qquad (2.6.3)$$

čia:

- *PE* – klasifikatoriaus tikroji generalizavimo paklaida;
- *s* – stiprumas: tikimybė, kad medis klasifikuos teisingai (daugiau nei atsitiktinai);
- *ρ* – vidutinė koreliacija tarp sprendimų medžių.

Vienas iš atsitiktinių miškų privalumų – gebėjimas įvertinti kiekvieno požymio įtaką modelio sprendimams. Tam yra keletas metrikų:

- vidutinis tikslumo sumažėjimas (*angl. Mean Decrease in Accuracy*) – apskaičiuojama, kiek suprastėja modelio tikslumas, atsitiktinai sumaišant (*angl. permutation*) konkretaus požymio reikšmes;
- vidutinis Gini indekso sumažėjimas (*angl. Mean Decrease in Gini*) – įvertina, kiek kiekvienas požymis sumažina Gini negrynumo indeksą per visus sprendimų medžių skaidymus (*angl. split*). Kuo didesnė reikšmė, tuo požymis laikomas svarbesniu.

Vienas iš svarbių atsitiktinių miškų metodo privalumų yra tai, kad nereikia atskirai skirstyti duomenų į mokymo ir testavimo imtis, nes modelis savaime įvertina savo prognozavimo paklaidą. Tam jis naudoja integruotą paklaidos įsivertinimo mechanizmą, kai stebėjimai yra „už imties ribų“ (*angl. Out-of-Bag, OOB*). Kiekvienas medis mokomas su atsitiktine imtimi su grąžinimu (*angl. bootstrap*), vidutiniškai apie 36,8 % pradinės duomenų aibės įrašų nepatenka į tą imtį. Tokie įrašai vadinami „Out-of-Bag“ stebėjimais.

Breimanas siūlo *OOB* stebėjimus naudoti testavimui kiekvienam medžiui atskirai:

- kadangi medis šių įrašų nematė mokymosi metu, dėl to jų prognozės atspindi realų modelio generalizavimo gebėjimą;
- kiekvienam įrašui galima surinkti prognozes iš tų medžių, kuriems tas įrašas buvo *OOB* ir apskaičiuoti bendrą klaidą.

OOB paklaidos skaičiavimas:

1. kiekvienas įrašas x_i turi po kelis medžius, kuriems jis buvo OOB;
2. gaunamos šių medžių prognozės $T_b(x_i)$;
3. agreguojama (daugumos balsavimas arba vidurkis);
4. skaičiuojama paklaida lyginant su tikrąja reikšme y_i;
5. paklaidos vidurkis per visus įrašus bus OOB klaida.

2.7. Atraminių vektorių mašina

Atraminių vektorių mašina (*SVM*) – tai prižiūrimojo mokymosi metodas, plačiai taikomas dviejų klasių klasifikavimo uždaviniams spręsti. Pagrindinė algoritmo idėja – rasti tokią sprendimo ribą, kuri maksimaliai atskirtų skirtingoms klasėms priklausančius stebėjimus. Ši riba apibrėžiama kaip hiperplokštuma, kurios atstumas iki artimiausių abiejų klasių taškų (vadinamoji paraštė) yra didžiausias. Tokia maksimali paraštė leidžia pagerinti modelio gebėjimą apibendrinti ir sumažinti persimokymo riziką [42].

Kietos paraštės (*angl. hard-margin*) SVM taikomas tada, kai duomenys yra tiesiškai atskiriami. Tuomet mokymo imtis (x_i, y_i), kur $x_i \in \mathbb{R}^m$ yra požymių vektorius, o $y_i \in \{-1, +1\}$ – klasės požymis.

Siekiama rasti hiperplokštumą, kuri apibrėžiama lygtimi:

$$w^T x + b = 0; \qquad (2.7.1)$$

čia w yra svorių vektorius, nusakantis hiperplokštumos orientaciją požymių erdvėje,

b – poslinkio parametras, apibrėžiantis hiperplokštumos padėtį koordinačių sistemos atžvilgiu. Jis leidžia perkelti sprendimo ribą taip, kad būtų užtikrintas optimalus klasių atskyrimas, išlaikant tą pačią hiperplokštumos orientaciją, kurią apibrėžia svorių vektorius.

Lygtis tenkintų sąlygą:

$$y_i(w^T x_i + b) \geq 1, \forall i \ . \qquad (2.7.2)$$

Optimalus sprendinys gaunamas sprendžiant kvadratinio programavimo uždavinį

$$\min_{w,b} \frac{1}{2}\|w\|^2, \qquad (2.7.3)$$

esant minėtiems apribojimams. Ši optimizavimo problema yra išgaubta, todėl turi globalų vienareikšmį sprendinį. Stebėjimai, kuriems tenkinama lygybė:

$$y_i(w^T x_i + b) = 1, \qquad (2.7.4)$$

vadinami atraminiais vektoriais. Būtent jie nulemia sprendimo ribos padėtį.

Praktikoje duomenys dažnai nėra idealiai atskiriami, todėl taikomas minkštos paraštės (*angl. soft-margin*) SVM. Šiuo atveju įvedami laisvumo kintamieji $\xi_i \geq 0$, leidžiantys daliai stebėjimų pažeisti atskyrimo sąlygą:

$$y_i(w^T x_i + b) \geq 1 - \xi_i. \qquad (2.7.5)$$

Optimizavimo kriterijus tampa

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{M} \xi_i, \qquad (2.7.6)$$

čia C > 0 yra reguliarizavimo parametras, valdantis kompromisą tarp plačios paraštės ir klasifikavimų klaidų. Dažniausiai praktikoje naudojamas L1 minkštos paraštės SVM. Šio uždavinio dualioji forma parodo, kad sprendimo funkcija gali būti užrašyta per atraminius vektorius:

$$D(x) = \sum_{i \in S} \alpha_i y_i x_i^T x + b, \qquad (2.7.7)$$

kur $\alpha_i > 0$ yra Lagrange'o daugikliai, o *S* – atraminių vektorių aibė. Tai reiškia, kad modelio sudėtingumas priklauso ne nuo visų mokymo taškų, bet tik nuo atraminių vektorių skaičiaus.

Netiesiniams klasifikavimo uždaviniams spręsti SVM naudoja vaizdavimą į aukštesnės dimensijos požymių erdves ir vadinamąją branduolio gudrybę (*angl. kernel trick*). Vietoj tiesioginio atvaizdo $\phi(x)$ skaičiavimo naudojama branduolio funkcija:

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j), \qquad (2.7.8)$$

kuri leidžia dirbti aukštesnėje erdvėje implicitiniu būdu. Branduolio funkcija turi tenkinti *Mercer* sąlygą – būti simetriška ir pusiau teigiamai apibrėžta, kad egzistuotų atitinkama požymių erdvė.

Dažniausiai naudojami šie branduoliai.

- tiesinis branduolys taikomas, kai duomenys pakankamai gerai atskiriami pradinėje erdvėje:

$$K(x, z) = x^T z \ ; \qquad (2.7.9)$$

- polinominis branduolys leidžia modeliuoti nelinijines sprendimo ribas, priklausomai nuo polinomo laipsnio *d:*

$$K(x, z) = (x^T z + 1)^d \ ; \qquad (2.7.10)$$

- radialinės bazinės funkcijos (RBF) branduolys yra universalus ir plačiai taikomas praktikoje, nes gali aproksimuoti sudėtingos formos sprendimo ribas; parametro *σ* reikšmė nusako poveikio lokalumą:

$$K(x, z) = e^{(-\sigma \|x - z\|^2)} \ . \qquad (2.7.11)$$

Algoritmas veikia tiksliausiai, kai randami optimalūs parametrai *σ* ir *C*. Tam taikoma tinklelio paieška (*angl. grid search*) kartu su 5 dalių kryžmine validacija, kai visos iš anksto apibrėžtos SVM su *RBF* branduoliu hiperparametrų kombinacijos nuosekliai įvertinamos naudojant tik mokymo duomenis. Hiperparametrų tinklelis sudaromas iš branduolio pločio parametro *σ*, kuris kontroliuoja RBF branduolio lokalumą, ir reguliarizavimo parametro *C*, nusakančio kompromisą tarp paraštės pločio ir klasifikavimo klaidų baudos. Galutinė hiperparametrų pora parenkama kaip ta, kuri pasiekia mažiausią vidutinį *logLoss* per visus validacijos kartus. Šis metodas yra deterministinis ir garantuoja geriausio sprendinio parinkimą iš nurodyto tinklelio, tačiau yra skaičiavimo požiūriu brangesnis nei alternatyvūs metodai.

2.8. CatBoost modelis

CatBoost yra gradientinio pastiprinimo sprendimų medžių algoritmas, skirtas klasifikavimo, regresijos ir rangavimo uždaviniams spręsti [43]. Algoritmas priklauso pastiprinimo (*angl. boosting)* metodų grupei, kai galutinis modelis sudaromas nuosekliai jungiant daug silpnesnių modelių, dažniausiai sprendimų medžių. Kiekvienas naujas medis mokomas taip, kad sumažintų ankstesnių medžių padarytas klaidas.

CatBoost modelis optimizuoja pasirinktą nuostolių funkciją, pavyzdžiui, klasifikavimo atveju gali būti naudojama logaritminė nuostolių funkcija, o regresijos atveju – kvadratinė paklaida. Kiekviename iteracijos žingsnyje modelis apskaičiuoja nuostolių funkcijos gradientą ir pagal jį konstruoja naują medį, kuris koreguoja ankstesnio modelio prognozes.

Vienas pagrindinių *CatBoost* privalumų yra efektyvus kategorinių kintamųjų apdorojimas. Skirtingai nei kai kurie kiti gradientinio stiprinimo algoritmai, *CatBoost* nereikalauja išankstinio kategorinių kintamųjų kodavimo. Vietoje to taikomi specialūs tikslinės statistikos kodavimo metodai, mažinantys duomenų nutekėjimo ir persimokymo riziką. Taip pat *CatBoost* naudoja simetrinius sprendimų medžius, todėl modelis dažnai pasižymi geru tikslumo ir skaičiavimo efektyvumo balansu [44].

CatBoost modelio veikimas priklauso nuo pasirinktų hiperparametrų. Svarbiausi iš jų yra šie:

- *iterations* – sprendimų medžių skaičius modelyje. Didesnė reikšmė leidžia modeliui mokytis sudėtingesnių priklausomybių, tačiau gali padidinti persimokymo riziką;
- *learning_rate* – mokymosi greitis, nusakantis, kokią įtaką kiekvienas naujas medis turi galutinei prognozei. Mažesnė reikšmė paprastai reikalauja didesnio iteracijų skaičiaus, tačiau gali pagerinti modelio generalizaciją;
- *depth* – sprendimų medžio gylis. Didesnis gylis leidžia modeliui aptikti sudėtingesnes sąveikas tarp požymių, tačiau taip pat gali didinti persimokymo tikimybę;
- *l2_leaf_reg* – L2 reguliarizacijos koeficientas medžių lapų reikšmėms. Didesnė reikšmė stiprina reguliarizaciją ir gali sumažinti persimokymą;
- *loss_function* – optimizuojama nuostolių funkcija. Ji parenkama pagal uždavinio tipą, pavyzdžiui, *logloss* binarinei klasifikacijai, *MultiClass* daugiaklasei klasifikacijai arba RMSE regresijai;
- *eval_metric* – metrika, pagal kurią vertinamas modelio veikimas validavimo imtyje. Ji nebūtinai sutampa su nuostolių funkcija;
- *random_strength* – atsitiktinumo lygis, taikomas medžių konstravimo metu. Šis parametras gali padėti sumažinti persimokymą;
- *bagging_temperature* – parametras, kontroliuojantis objektų atrankos atsitiktinumą mokymo metu. Didesnė reikšmė didina atsitiktinumą;
- *border_count* – skaitinių kintamųjų diskretizavimo ribų skaičius. Šis parametras turi įtakos tam, kaip tiksliai skaitiniai kintamieji suskaidomi į intervalus;
- *early_stopping_rounds* – ankstyvo stabdymo parametras. Jeigu validavimo metrika nepagerėja per nurodytą iteracijų skaičių, apmokymas sustabdomas, taip sumažinant persimokymo riziką.

CatBoost modelio hiperparametrų optimizavimui gali būti taikomi skirtingi būdai, vienas iš jų *Optuna* biblioteka. Optimizavimo metu apibrėžiama tikslo funkcija, kurioje modelis mokomas su skirtingomis hiperparametrų reikšmėmis, o jų kokybė vertinama pagal pasirinktą validavimo metriką, pvz. AUC. *Optuna* yra automatinio hiperparametrų optimizavimo priemonė, naudojanti *Bajeso* optimizavimo principu pagrįstą TPE algoritmu, kuris naujus parametrų derinius parenka atsižvelgdamas į ankstesnių bandymų rezultatus. Toks metodas leidžia efektyviau ištirti hiperparametrų erdvę ir sumažinti skaičiavimo sąnaudas palyginus su pilna tinklelio paieška.

2.9. Sumaišymo matrica

Kuriant kredito rizikos modelius, kurių tikslas prognozuoti, ar paskolos gavėjas laiku grąžins paskolą arba ar paskola bus išmokėta, taikomi binariniai klasifikavimo modeliai (pvz., atsitiktiniai miškai, logistinė regresija). Tokie modeliai priskiria kiekvieną paskolos paraišką vienai iš dviejų klasių, dažniausiai:

- 0 – moki paskola (įsipareigojimų investuotojams įvykdymas);
- 1 – nemoki paskola (įsipareigojimų investuotojams neįvykdymas).

Modelio veikimo kokybei įvertinti naudojami keli klasifikavimo tikslumo rodikliai. Vienas iš esminių ir plačiausiai taikomų – sumaišymo matrica (*angl. confusion matrix*), kuri leidžia matyti ne tik bendrą tikslumą, bet ir specifines modelio klaidas, svarbias kredito rizikos valdymo kontekste.

Sumaišymo matrica (2 lentelė) yra lentelė (jei prognozuojamas kintamasis įgyja daugiau nei 2 reikšmes, atitinkamai stulpelių ir eilučių kiekis pasipildys kiekvienai klasei), kurioje lyginamos modelio prognozės su tikrosiomis klasėmis. Tipiškai matrica atrodo taip:

2 lentelė. Sumaišymo matrica

	Prognozuota reikšmė: 0	Prognozuota reikšmė: 1
Realybė: įgyjama reikšmė 0	TN – Tikri neigiami	FP – klaidingai teigiami
Realybė: įgyjama reikšmė 1	FN – klaidingai neigiami	TP – tikri teigiami

- TP (True Positives) – atvejai, kai modelis teisingai identifikavo nemokią paskolą;
- TN (True Negatives) – atvejai, kai modelis teisingai identifikavo mokią paskolą;
- FP (False Positives) – atvejai, kai modelis klaidingai prognozavo nemokią paskolą, kuri realiai buvo moki (vadinamas tipo I klaida);
- FN (False Negatives) – atvejai, kai modelis klaidingai priskyrė mokumo statusą paskolai, kuri realiai nebuvo nemoki (tipo II klaida).

Remiantis sumaišymo matricos elementais, skaičiuojamos šios pagrindinės modelio vertinimo metrikos:

Tikslumas (*angl. Accuracy*) parodo, kokia dalis prognozių yra teisingos:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.9.1)$$

Atkūrimo metrika – jautrumas (*angl. Sensitivity arba Recall*) parodo, kokią dalį teigiamų atvejų modelis teisingai klasifikavo:

$$Recall = \frac{TP}{TP+FN} \quad (2.9.2)$$

Specifiškumas (*angl. Specificity*) parodo, kokią dalį neigiamų atvejų modelis teisingai klasifikavo kaip neigiamus:

$$Specificity = \frac{TN}{TN+FP} \quad (2.9.3)$$

Preciziškumas (*angl. Precision*) parodo, kokia dalis prognozuotų teigiamų atvejų iš tikrųjų yra teisingai teigiami:

$$\Pr ecision = \frac{TP}{TP+FP} \quad (2.9.4)$$

F1 metrika (*angl. F1 Score*) – harmoninis vidurkis tarp jautrumo ir preciziškumo:

$$F1 = 2 * \frac{\Pr ecision \cdot Recall}{\Pr ecision + Recall} \quad (2.9.5)$$

Kredito rizikos vertinimo kontekste kritiškiausia yra II tipo klaida (False Negative), kai nemoki paskola klaidingai klasifikuojama kaip išmokėta. Tokia klaida lemia šias neigiamas pasekmes:

- tiesioginiai finansiniai nuostoliai investuotojams;
- padidėja nemokumo tikimybė PD (*angl. Probability of default*);
- gali nukentėti platformos reputacija ir investuotojų pasitikėjimas;
- sutelktinio finansavimo atveju – nuostoliai išskaidomi daugeliui investuotojų.

Todėl modelių vertinime ypač didelis dėmesys skiriamas atkūrimo metrikai, arba jautrumui, (*angl. recall*) ir nemokių paskolų atpažinimo gebėjimui, o ne vien bendram tikslumui.

Cohen'o kappa (κ) [45] yra statistinis rodiklis, naudojamas įvertinti dviejų klasifikacijų (ar vertintojų) suderinamumą, atsižvelgiant į atsitiktinio sutapimo tikimybę. Skirtingai nei paprastas tikslumas (*angl. accuracy*), kapa koreguoja rezultatą pagal tai, kiek sutapimų galėtų atsirasti vien dėl atsitiktinumo.

Kapa rodiklis apskaičiuojamas pagal formulę:

$$\kappa = \frac{p_0 - p_e}{1 - p_e}; \quad (2.9.6)$$

čia

- p_0 – faktinis stebėtas sutapimo lygis;

- p_e – tikėtinas sutapimo lygis atsitiktinai.

Kapa reikšmė svyruoja nuo −1 iki 1:

- $\kappa = 1$ – tobulas sutapimas;
- $\kappa = 0$ – sutapimas atitinka atsitiktinumą;
- $\kappa = 1$ – blogesnis nei atsitiktinis sutapimas;

Praktikoje dažnai naudojama interpretacija:

0,00–0,2 – silpnas suderinamumas,

0,21–0,40 – patenkinamas suderinamumas,

0,41–0,60 – vidutinis suderinamumas,

0,61–0,80 – geras suderinamumas,

0,81–1.00 – labai geras suderinamumas.

Kapa rodiklis plačiai taikomas klasifikavimo modelių vertinime, ypač kai klasės yra nesubalansuotos, nes leidžia objektyviau įvertinti modelio veikimą nei vien tik tikslumo metrika.

2.10. Detekcijos kreivės

2.10.1. ROC kreivė

ROC kreivė yra naudingas vizualus įrankis, leidžiantis palyginti du klasifikavimo modelius. ROC kreivės [46] kilusios iš signalų aptikimo teorijos, kuri buvo sukurta Antrojo pasaulinio karo metu analizuoti radarų vaizdus. Ši kreivė konkrečiam modeliui parodo kompromisą tarp tikro teigiamo rodiklio (TPR) ir klaidingo teigiamo rodiklio (FPR) (16 pav.). TPR – tai dalis pozityvių (pvz., „taip“, “1”) objektų, kurie teisingai atpažinti modelio; FPR – tai dalis negatyvių (pvz., „ne“, “0”) objektų, kurie klaidingai priskirti pozityviai klasei.

Jei TP, FP, P ir N yra tikrųjų teigiamų, klaidingų teigiamų, pozityvių ir negatyvių objektų skaičius, tuomet:

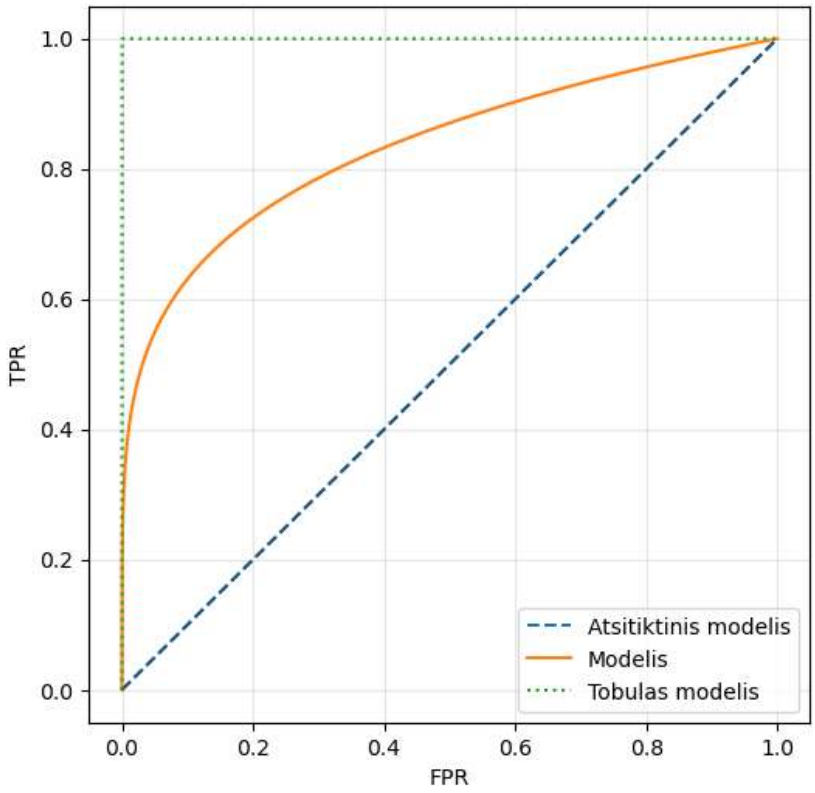
- $TPR = \frac{TP}{P}$ – tai jautrumas (*angl. sensitivity*); (2.10.1.1)
- $FPR = \frac{FP}{N}$ – tai specifiškumas (*angl. specifity*). (2.10.1.2)

Dviejų klasių problemai ROC kreivė leidžia vizualizuoti kompromisą tarp modelio gebėjimo teisingai atpažinti pozityvius atvejus ir klaidingai pozityvius atvejus, keičiant klasifikavimo slenkstį. Bet koks TPR padidėjimas vyksta FPR padidėjimo sąskaita.

Norint nubraižyti ROC kreivę konkrečiam klasifikavimo modeliui M, modelis turi grąžinti prognozuojamos klasės tikimybę kiekvienam testuojamam objektui. Tuomet objektai surikiuojami pagal tikimybę priklausyti taikinio klasei tokiu būdu: didžiausia tikimybė viršuje, mažiausia apačioje.

ROC kreivės braižymo principas:

- vertikalioje ašyje – TPR, horizontalioje – FPR;
- pradedama nuo apatinio kairiojo kampo (kur TPR = 0, FPR = 0);
- einama per surikiuotus objektus: jei objektas teisingai priskirtas taikinio klasei (TP), TPR didėja – grafike judama aukštyn; jei klaidingai priskirtas taikinio klasei (FP), FPR didėja – grafike judama į dešinę;
- procesas kartojamas per visus testuojamus objektus, kiekvieną kartą judant aukštyn (TP) arba į dešinę (FP).



16 pav. ROC kreivė

Plotas po ROC kreive (AUC) yra modelio tikslumo matas. AUC (*angl. Area Under the Curve)* – tai klasifikavimo modelio kokybės metrika, kuri įvertina modelio gebėjimą atskirti klases, nepriklausomai nuo pasirinkto klasifikavimo slenksčio (*angl. threshold*). AUC nurodo tikimybę, kad modelis atsitiktinai pasirinktai nemokiai paskolai priskirs didesnę riziką nei mokiai. AUC įgyja reikšmes nuo 0 iki 1, kuo didesnė AUC reikšmė, tuo didesnė tikimybė, kad modelis teisingai atskiria išmokėtą nuo neišmokėtos paskolos pagal prognozuojamą rizikos tikimybę.

2.10.2. DET kreivė

DET kreivė (*angl. Detection Error Tradeoff curve*) yra klasifikavimo modelių vertinimo metodas, naudojamas analizuoti kompromisą tarp klaidingai teigiamų ir klaidingai neigiamų klasifikavimo rezultatų [47]. Skirtingai nei ROC kreivė, DET kreivėje vaizduojami (17 pav.) du klaidų rodikliai: klaidingai teigiamų rezultatų dažnis FPR ir klaidingai neigiamų rezultatų dažnis FNR.

Klaidingai neigiamų rezultatų dažnis:

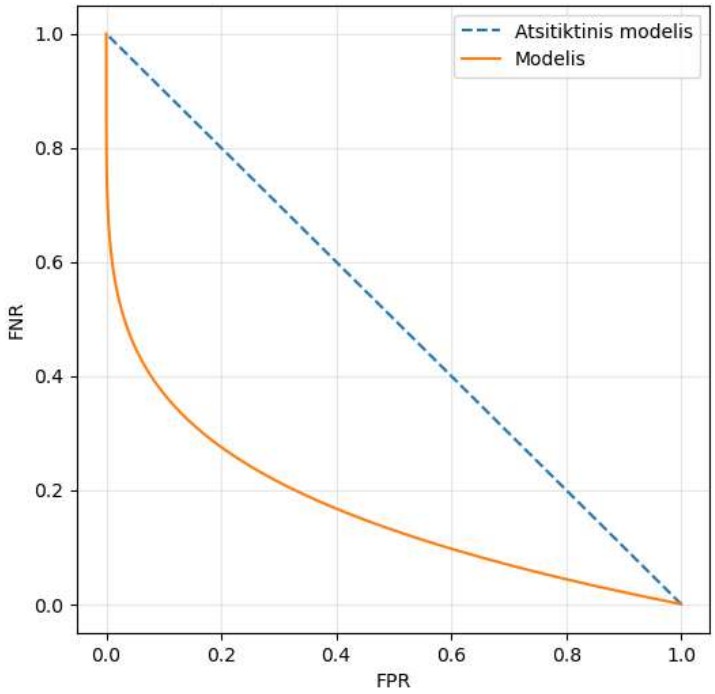
$$FNR = \frac{FN}{FN+TP} \qquad (2.10.2.1) \qquad (33)$$

DET kreivė parodo, kaip keičiasi šie rodikliai keičiant klasifikavimo slenkstį. Geresnis modelis grafike yra arčiau apatinio kairiojo kampo, nes tai reiškia mažesnes abiejų klaidų reikšmes [46].

Kredito rizikos vertinimo kontekste DET kreivė leidžia įvertinti kompromisą tarp rizikingų paskolų neaptikimo ir saugių paskolų klaidingo priskyrimo rizikingoms. Jei teigiama klasė reiškia nemokią paskolą, tuomet:

- FNR rodo, kokia dalis nemokių paskolų nebuvo aptikta modelio;
- FPR rodo, kokia dalis mokių paskolų buvo klaidingai pažymėtos kaip rizikingos.

Todėl DET kreivė leidžia pasirinkti klasifikavimo slenkstį pagal investuotojo rizikos toleranciją bei padeda išsamiau įvertinti modelio veikimą nei vien tik bendras tikslumas ar ROC AUC rodiklis.

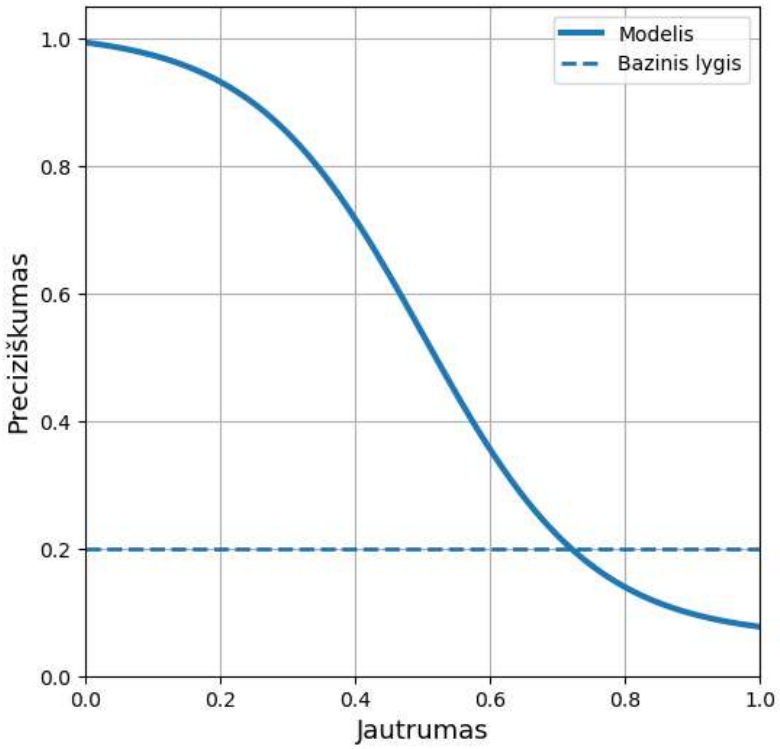


17 pav. DET kreivės pavyzdys

2.10.3. Preciziškumo-jautrumo kreivė

Preciziškumo-jautrumo (PR) kreivė yra klasifikavimo modelių vertinimo metodas, naudojamas analizuoti kompromisą tarp tikslumo ir jautrumo keičiant klasifikavimo slenkstį [37]. Ši kreivė ypač naudinga esant klasių disbalansui, kai viena klasė pasitaiko gerokai rečiau nei kita, todėl dažnai taikoma kredito rizikos vertinimo uždaviniuose.

Preciziškumo-jautrumo kreivė (18 pav.) vaizduoja šių dviejų rodiklių santykį keičiant klasifikavimo slenkstį. Geresnis modelis pasižymi kreive, esančia arčiau viršutinio dešiniojo kampo, nes tai reiškia didesnį tikslumą ir jautrumą.



18 pav. Preciziškumo-jautrumo kreivės pavyzdys

Kredito rizikos vertinimo kontekste PR kreivė leidžia įvertinti modelio gebėjimą aptikti nemokias paskolas kartu sumažinant klaidingų perspėjimų skaičių. Jei teigiama klasė reiškia nemokią paskolą, didelis jautrumas rodo, kad modelis sugeba aptikti didžiąją dalį rizikingų paskolų, o didelis preciziškumas reiškia, kad dauguma kaip rizikingų pažymėtų paskolų iš tiesų yra probleminės.

PR kreivė dažnai yra informatyvesnė nei ROC kreivė, kai dirbama su nesubalansuotais duomenimis. Ji geriau atspindi modelio veikimą retos klasės atžvilgiu [16].

2.11. SHAP analizė

SHAP metodas remiasi žaidimų teorijoje naudojamomis *Shapley* reikšmėmis, kurios leidžia įvertinti kiekvieno kintamojo indėlį į galutinę modelio prognozę [32]. Pagrindinė idėja –

kiekvienam požymiui priskirti „indėlį“ į prognozuojamą rezultatą, įvertinant visas galimas požymių kombinacijas.

Bendra SHAP modelio forma:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i \ , \qquad (2.11.1)$$

čia:

$f(x)$ – modelio prognozė;

ϕ_0 – bazinė modelio prognozė;

ϕ_i – i-tojo požymio SHAP reikšmė;

M – požymių skaičius.

Teigiama SHAP reikšmė rodo, kad požymis didina modelio prognozuojamą tikimybę priklausyti teigiamai klasei, o neigiama – mažina. Kredito rizikos vertinimo kontekste tai leidžia interpretuoti, kurie veiksniai didina arba mažina paskolos nemokumo tikimybę.

SHAP analizė turi kelis svarbius privalumus:

- leidžia interpretuoti sudėtingus ML modelius;
- suteikia tiek globalų, tiek lokalinį modelio paaiškinimą;
- leidžia nustatyti ne tik požymio svarbą, bet ir poveikio kryptį;
- padeda identifikuoti rizikos veiksnius bei modelio sprendimų logiką.

2.12. Investavimo strategijos

Investavimo strategijų pelningumo skaičiavimuose darbe vartojami šie kredito rizikos terminai ir formulės:

PD (*angl. Probability of default)* - sukurto modelio prognozuojama tikimybė, kad paskola bus nemoki.

EAD (*angl. Exposure at default)* - investuota suma.

$$Tikėtinas\ pelnas\ = EAD\ \times \frac{Palūkanų\ norma}{100} \times (1 - PD) \qquad (2.12.1)$$

$$Tikėtinas\ nuostolis\ = EAD \times PD \qquad (2.12.2)$$

$$Tikėtinas\ Net\ pelnas\ = Tikėtinas\ pelnas - Tikėtinas\ nuostolis \qquad (2.12.3)$$

$$Tikėtina\ Net\ grąža\ = \frac{Tikėtinas\ Net\ pelnas}{EAD}\ \times 100 \qquad (2.12.4)$$

$$Realizuotas\ Net\ pelnas\ = \begin{cases} -EAD & , kai\ paskola\ nemoki \\ EAD \times \frac{Palūkanų\ norma}{100} & , kai\ paskola\ moki \end{cases} \qquad (2.12.5)$$

$$Realizuota\ grąža\ = \frac{Realizuotas\ Net\ pelnas}{EAD} \times 100 \qquad (2.12.6)$$

Remiantis literatūra suformuojamos strategijos, kad rizika būtų subalansuota su grąža. Sukuriama konservatyvi strategija, atitinkanti Markovitso [34] idėją, kai mažos rizikos turtas portfelyje sukuria didesnį stabilų pelną. Taip pat įtraukiama agresyvi strategija - kraštutinumo pavyzdys, kai rizika negauna pakankamos kompensacijos.

Žemiau pateikiami kiekvienos strategijos apibrėžimai, parametrai ir taisyklės.

Visuotinio investavimo strategija. Visuotinio investavimo strategija reiškia, jog investuojama ta pati suma į visų kategorijų paskolas be jokių papildomų filtrų pagal riziką ar kitus kriterijus. Kitaip tariant, tai „rinkos portfelis“ – investuotojas paskirsto lėšas po visą platformos paskolų spektrą, proporcingai pasiūlai (arba vienodomis dalimis). Šios strategijos esmė – maksimali diversifikacija ir rizikos išskaidymas. Ji tinka pasyviam investuotojui, kuris neturi aiškios nuomonės dėl rizikos, nori gauti vidutinę rinkos grąžą ir išvengti didelių nukrypimų. Šioje strategijoje netaikomas joks PD slenkstis, į visas paskolas investuojama vienodai, todėl toks portfelis apims ir saugias paskolas, ir rizikingas.

Konservatyvi strategija. Ši investavimo strategija orientuojasi į maksimalią kapitalo apsaugą. Investuojama tik į tas paskolas, kurias modelis įvertina kaip saugiausias – PD, kurios yra mažesnės už pirmojo kvantilio reikšmes. Tokia strategija tinkama rizikos vengiančiam investuotojui, kuris nori užtikrinto pelno, tačiau ši strategija pasižymi tuo, kad joje gali būti mažai paskolų, į kurias galima investuoti.

Rizikos-grąžos optimizavimo strategija. Ši strategija siekia optimalios rizikos ir pelno pusiausvyros – investuotojas renkasi paskolas ne aklai pagal žemiausią PD ar aukščiausias palūkanas, o pagal kriterijų, kad tikėtinas pelningumas yra pakankamas. Šiame darbe apibrėžta strategija remiasi taisykle: paskolos palūkanų norma ≥ PD.

Agresyvi strategija. Agresyvi strategija reiškia investavimą vien tik į aukščiausio pajamingumo paskolas, t. y., turinčias didžiausias palūkanas, ignoruojant jų rizikos rodiklius. Tokios paskolos platformoje paprastai turi dviženklę pačią didžiausią palūkanų normą (pvz., 12–14 %), kas reiškia, kad skolininkas yra arba prastesnės kredito kokybės, arba projektas labai rizikingas. Šiame darbe apibrėžta strategija remiasi taisykle: paskolos palūkanų norma > 11%.

Svertinė atrankos strategija. Papildomai suformuota svertinė investavimo strategija, kurioje atrenkamos tik aukštas palūkanas turinčios paskolos, o investicijų svoriai priskiriami atvirkščiai proporcingai nemokumo tikimybei. Tokia strategija leidžia koncentruoti investicijas į mažiausios rizikos paskolas, išlaikant aukštą nominalų palūkanų lygį. Svertinė strategija formuojama derinant du kriterijus: aukštą nominalią palūkanų normą ir mažą prognozuojamą nemokumo tikimybę. Pirmiausia į strategijos imtį atrenkamos tik tos paskolos, kurių palūkanų norma yra didesnė nei 11 %. Taip siekiama išlaikyti aukštesnį potencialų pajamingumą. Iš šios paskolų grupės paskolos surikiuojamos pagal prognozuotą nemokumo tikimybę PD didėjimo tvarka. strategija formuojama derinant du kriterijus: aukštą

nominalią palūkanų normą ir mažą prognozuojamą nemokumo tikimybę. Tokiu būdu strategija neinvestuoja į visas aukštų palūkanų paskolas, o atrenka tik tas, kurios pasižymi santykinai mažiausia kredito rizika.

Atrinktam pogrupiui taikoma svertinė schema, kur kiekvienos paskolos svoris apskaičiuojamas kaip:

$$w_i = \frac{1-PD_i}{\sum_{j=1}^{n}(1-PD_j)}; \quad (2.12.7)$$

čia PD_i – i-tosios paskolos nemokumo tikimybė, o w_i – jai priskirtas svoris.

$$\sum w_i = 1 \quad (2.12.8)$$

2.13. Programinė įranga

Tyrimo duomenų parengimas, modeliavimas, rezultatų analizė ir vizualizavimas buvo atlikti naudojant *R* (versija 4.5.1) ir *Python* (versija 3.12.9) programavimo kalbas. Skirtingos programinės aplinkos pasirinktos atsižvelgiant į taikomų metodų specifiką ir naudojamų bibliotekų funkcionalumą.

R programavimo aplinka buvo naudojama logistinės regresijos, SVM, atsitiktinių miškų modelių kūrimui, kintamųjų svarbos analizei bei daliai modelių vertinimo procedūrų. Pagrindiniai R paketai: *caret, glmnet, pRoc, randomForest, tuneRanger.*

Python programavimo kalba buvo naudojama *CatBoost* modelio kūrimui, aprašomajai duomenų analizei, rezultatų vizualizavimui bei investavimo strategijų formavimui ir vertinimui. *Python* aplinkoje naudotos šios bibliotekos: *pandas, matplotlib, catboost ir optuna.*

Abi programavimo kalbos buvo naudojamos KTU internetinėje programavimo ir duomenų analizės aplinkoje, paremta *JupyterLab* platforma. Ji leidžia programuoti, vykdyti kodą ir analizuoti duomenis tiesiog naršyklėje, nereikalaujant papildomo programinės įrangos diegimo kompiuteryje.

3. Kredito rizikos modeliavimas

Kredito rizikos modeliavimas „EstateGuru“ duomenims atliekamas taikant 4 algoritmus:

1. logistinė regresija su ElasticNet reguliarizacija;
2. SVM;
3. atsitiktiniai miškai;
4. CatBoost algoritmas.

Kadangi duomenys yra suskirstyti nuo seniausių iki naujausių paskolų, dėl to pirmiausia buvo atskirtas mokymo ir testavimo rinkinys pagal chronologinę seką (70 % senesnių duomenų naudota mokymui, 30 % naujesnių – testavimui). Vėliau šis datos stulpelis yra pašalinamas iš duomenų rinkinio.

3.1. Logistinės regresijos modelis

Logistinės regresijos modelio su *elastic net* reguliarizacija parametrai buvo optimizuojami taikant 5 dalių kryžminę validaciją. Parametro *alpha* paieškai naudota transformuota seka intervale [0;1], leidžianti detaliau įvertinti skirtingus L1 ir L2 reguliacijos derinius. Nustatyti optimalūs parametrai: *alpha* = 0,512, rodantis subalansuotą L1 ir L2 reguliaciją, ir *lambda* = 0,00052, kuris indikuoja santykinai silpną reguliacijos lygį.

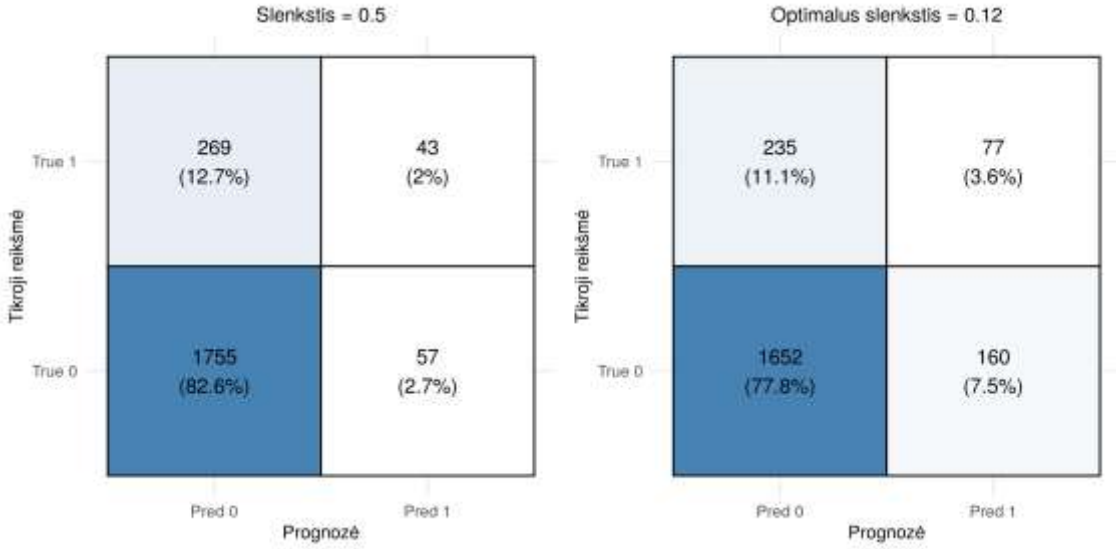
Modelio diskriminacinė geba yra vidutinė (AUC = 0,717). Naudojant optimizuotą 0,122 slenkstį, tikslumas (0,814) panašus kaip ir su slenksčiu 0,5, tačiau modelio gebėjimas aptikti teigiamą klasę vis tiek nėra aukštas (jautrumas = 0,247), o F1 metrikos reikšmė (0,28) ir *kapa* rodiklis (0,176) rodo prastą klasifikavimo balansą.

3 lentelė. Logistinės regresijos modelio pagrindinės metrikos

Metrika	**Slenkstis 0,5**	**Optimalus slenkstis 0,122**
AUC	0,717	0,717
Tikslumas	0,847	0,814
Preciziškumas	0,430	0,325
Jautrumas	0,138	0,247
F1 įvertis	0,209	0,281
Kapa rodiklis	0,148	0,176
Bazinis tikslumas	0,853	0,853

Sumaišymo matricų (19 pav.) analizė patvirtina, kad modelis yra stipriai orientuotas į dominuojančią klasę: net sumažinus slenkstį, teisingai identifikuotų rizikingų paskolų dalis išlieka labai maža, o didžioji jų dalis priskiriama saugių paskolų kategorijai.

Apibendrinant, logistinės regresijos modelis šioje užduotyje nėra tinkamas kredito rizikos prognozavimui, nes nepajėgia efektyviai identifikuoti rizikingų paskolų.



19 pav. Logistinės regresijos sumaišymo matrica skirtingiems slenksčiams

3.2. SVM modelis

Atraminių vektorių mašinos modeliui taikoma tinklelio paieška optimaliems parametrams rasti. Kiekviena (*σ, C*) kombinacija vertinama 5 kartus, taikant kryžminę validaciją, o jos našumas apskaičiuojamas pagal logaritminio nuostolio (*angl. logloss*) metriką, kuri įvertina prognozuojamų tikimybių kokybę. Gauti šie optimalūs parametrai: σ = 0,0884 ir C = 4. Atlikus keletą bandymų su skirtingais branduoliais, pasirinktas radialinės bazinės funkcijos (RBF) branduolys pagal aukščiausią AUC reikšmę.

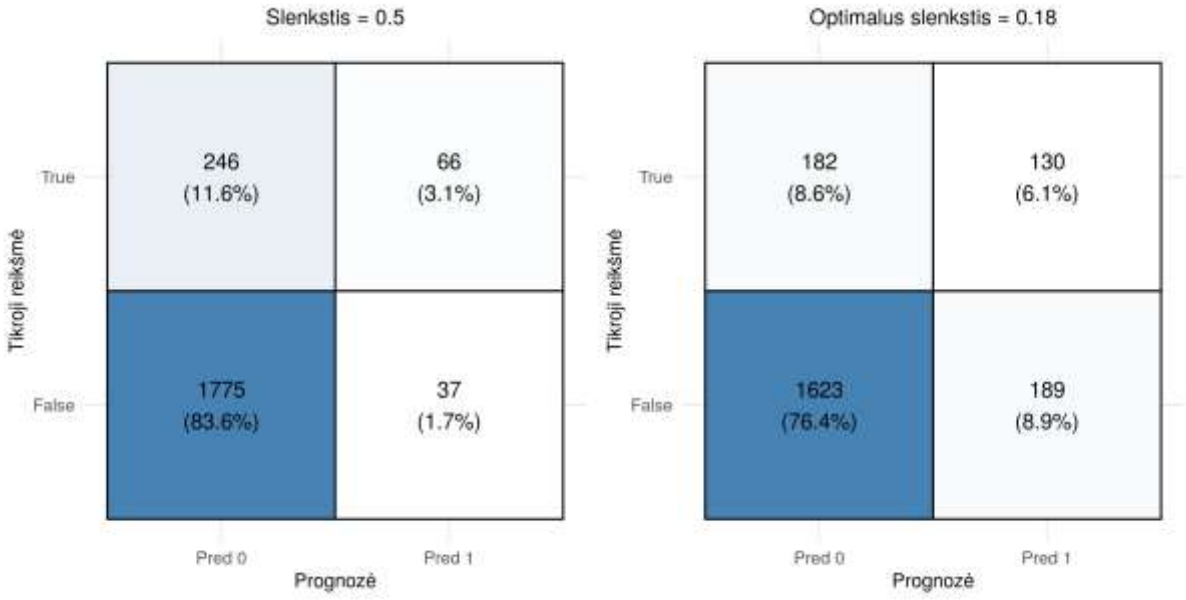
SVM modelis (4 lentelė) pasižymi vidutiniu diskriminaciniu gebėjimu (AUC = 0,703), tačiau esant standartiniam 0,5 slenksčiui jis yra konservatyvus – pasiekiamas aukštas tikslumas (0,867), bet labai žemas jautrumas (0,212), todėl daug teigiamų atvejų lieka neaptikti. Sumažinus slenkstį iki optimalaus (0,171), jautrumas reikšmingai pagerėja (iki 0,417), tačiau sumažėja preciziškumas ir bendras tikslumas. Apskritai modelis rodo kompromisą tarp klaidingų teigiamų ir klaidingų neigiamų prognozių, o F1 metrikos ir *kapa* rodiklio pagerėjimas rodo labiau subalansuotą veikimą pasirinkus optimalų slenkstį.

4 lentelė. SVM modelio pagrindinės metrikos

Metrika	**Slenkstis 0,5**	**Optimalus slenkstis 0,171**
AUC	0,703	0,703
Tikslumas	0,867	0,825
Preciziškumas	0,641	0,408
Jautrumas	0,212	0,417

F1 įvertis	0,318	0,412
Kapa rodiklis	0,264	0,309
Bazinis tikslumas	0,853	0,853

Iš sumaišymo matricos (20 pav.) matyti, kad esant slenksčiui 0,5, modelis teisingai identifikuoja daugumą neigiamų atvejų (1775), tačiau aptinka palyginti mažai teigiamų (66), todėl jautrumas yra žemas. Sumažinus slenkstį iki 0,18, padidėja teisingai nustatytų teigiamų atvejų skaičius (130), tačiau kartu išauga klaidingų teigiamų prognozių skaičius (189).



20 pav. SVM sumaišymo matrica skirtingiems slenksčiams

3.3. Atsitiktiniai miškai

Atsitiktinių miškų modelio parametrų optimizavimui buvo taikytas *tuneRanger* algoritmas, naudojant 5 dalių kryžminę validaciją (*angl. 5-fold cross-validation*). Optimizavimo metu siekta nustatyti geriausią parametrų kombinaciją, minimizuojančią logaritminę nuostolio funkciją (*angl. logloss*).

Atlikus hiperparametrų paiešką, modelis buvo papildomai pertreniruotas naudojant mokymo duomenų aibę, sudarančią 70 % visos imties. Pakartotinai taikant *tuneRanger* metodą, buvo nustatyti optimalūs modelio parametrai: *mtry* = 12, *min.node.size* = 4 ir *sample.fraction* = 0,659. Gauto modelio prognozavimo tikslumą rodančios nuostolio funkcijos reikšmė: 0,2187. Šis rezultatas rodo pakankamai gerą modelio gebėjimą tiksliai įvertinti kredito riziką, atsižvelgiant į prognozuojamų tikimybių kokybę.

Parametras *mtry* = 12 nurodo, kad kiekvieno medžio skaidymo metu atsitiktinai parenkama 12 kintamųjų iš viso požymių rinkinio. Didesnė *mtry* reikšmė paprastai didina modelio gebėjimą išnaudoti informatyvius kintamuosius.

Parametras *min.node.size* = 4 apibrėžia mažiausią stebėjimų skaičių terminaliniuose medžio mazguose. Santykinai maža šio parametro reikšmė rodo, kad modelis leidžia formuoti gana giliai išsišakojusius medžius, kurie gali užfiksuoti sudėtingus netiesinius ryšius tarp kintamųjų.

Parametras *sample.fraction* = 0,659 reiškia, kad kiekvienas medis mokomas naudojant apie 65,9 % atsitiktinai atrinktų stebėjimų iš mokymo aibės.

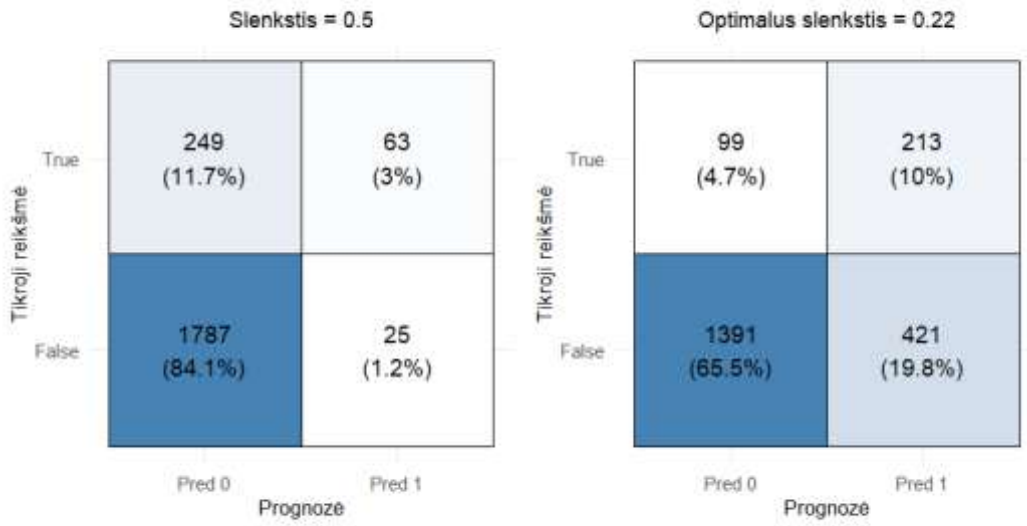
Apibendrinant, parinkti hiperparametrai rodo balansą tarp modelio sudėtingumo ir generalizacijos gebėjimo: modelis yra pakankamai lankstus užfiksuoti sudėtingus ryšius.

Modelio diskriminacinė geba (5 lentelė) yra pakankamai gera (AUC = 0,82), tačiau rezultatų analizė rodo, kad klasifikavimo slenksčio pasirinkimas turi esminę įtaką praktiniam pritaikomumui. Naudojant optimizuotą 0,225 slenkstį, ženkliai pagerėja rizikingų paskolų identifikavimas (jautrumas = 0,683), nors sumažėja bendras tikslumas ir preciziškumas. Nepaisant to, bendras modelio balansas gerėja (F1 įvertis = 0,450; kapa rodiklis = 0,316), o tai rodo didesnį tinkamumą kredito rizikos vertinimo uždaviniui. Atsižvelgiant į tai, kad svarbiau sumažinti neaptiktos rizikos tikimybę, mažesnis klasifikavimo slenkstis laikytinas labiau pagrįstu sprendimu praktiniame kontekste.

5 lentelė. Atsitiktinių miškų modelio pagrindinės metrikos

Metrika	**Slenkstis 0,5**	**Optimalus slenkstis 0,22**
AUC	0,820	0,820
Tikslumas	0,871	0,755
Preciziškumas	0,716	0,336
Jautrumas	0,202	0,683
F1 įvertis	0,315	0,450
Kapa rodiklis	0,268	0,316
Bazinis tikslumas	0,853	0,853

Sumaišymo matricų analizė (21 pav.) patvirtina, kad esant 0,5 slenksčiui modelis labai retai identifikuoja rizikingas paskolas (tik 25 teisingai klasifikuoti teigiami atvejai), tuo tarpu didžioji dalis jų priskiriama saugių paskolų klasei. Tai rodo konservatyvų modelio elgesį ir paaiškina žemą jautrumo reikšmę. Tuo tarpu sumažinus slenkstį iki 0,22, teisingai identifikuotų rizikingų paskolų skaičius ženkliai padidėja (iki 421 atvejo), tačiau kartu išauga klaidingai pažymėtų saugių paskolų skaičius. Šis pokytis atspindi sąmoningą kompromisą tarp klaidų tipų, prioritetą teikiant rizikos aptikimui.



21 pav. Atsitiktiniai miškų sumaišymo matrica skirtingiems slenksčiams

3.4. CatBoost modelis

CatBoost algoritmo optimalių parametrų paieškai buvo pritaikytas *Optuna* metodas, gauti parametrai:

depth = 10 – gana gilūs medžiai, vadinasi, leidžia modeliuoti sudėtingas sąveikas tarp kintamųjų. Toks parametro dydis rodo, kad ryšiai nėra tiesiniai;

learning_rate = 0,06798962421591129, vidutinė reikšmė rodo, kad modelis mokosi stabiliai, nėra didelių svyravimų;

Reguliarizacijos parametrai *l2_leaf_reg* = 1,6305687346221471 rodo vidutinę reguliarizacija, o *random_strength* = 2,454599737656805, mažina modelio persimokymą;

grow_policy = *Lossguide* – naudojami asimetriniai medžiai;

bootstrap_type = *Bayesian;*

bagging_temperature = 0,6118528947223795 atsitiktinumo kontroliavimas;

iterations = 2000, išbandyta pakankamas kiekis iteracijų, kad būtų gautas optimalus parametrų sąrašas;

eval_metric = AUC metrika, kurios maksimali reikšmė nulemia parametrų pasirinkimą;

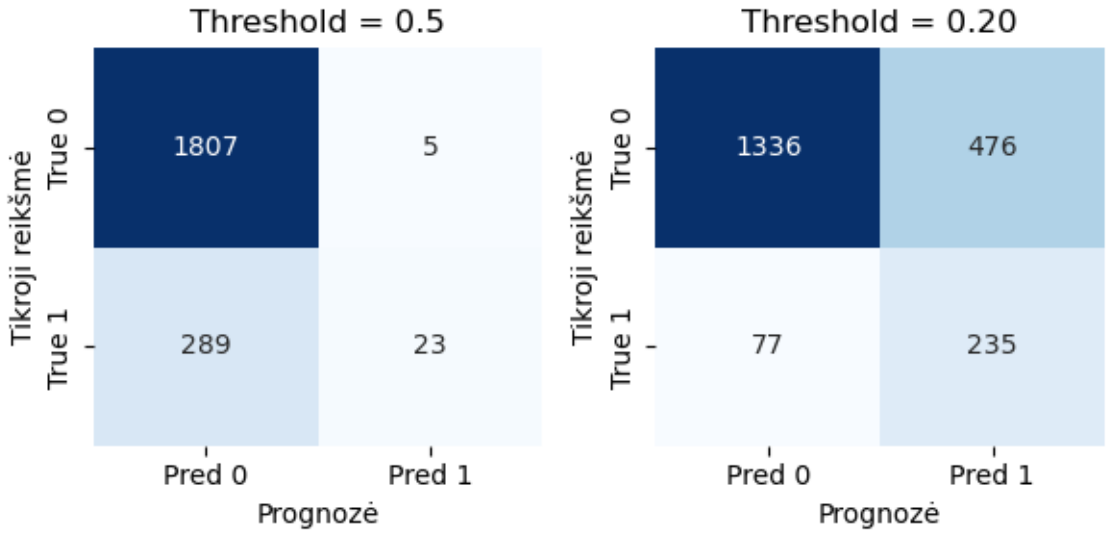
custom_metric = *logloss* – papildoma metrika su AUC optimizavimu.

Modeliui pritaikius optimalius parametrus gautos tokios metrikos reikšmės:

6 lentelė. CatBoost modelio pagrindinės metrikos

Metrika	Slenkstis 0,5	Optimalus slenkstis 0,2
AUC	0,802	0,802
Tikslumas	0,862	0,740
Preciziškumas	0,821	0,331
Jautrumas	0,074	0,753
F1 įvertis	0,135	0,459
Kapa rodiklis	0,114	0,321
Bazinis tikslumas	0,853	0,853

Modelio diskriminacinis gebėjimas atskirti klases yra pakankamai aukštas: 80,2 %. Didžiausią įtaką praktiniam pritaikymui turi pasirinktas optimalus slenkstis. Naudojant standartinį 0,5 slenkstį modelis pasižymi labai mažu jautrumu (jautrumas = 0,074), todėl didžioji dalis nemokių paskolų lieka neidentifikuota. Tuo tarpu sumažinus slenkstį iki 0,2, reikšmingai pagerėja modelio gebėjimas aptikti rizikingas paskolas (jautrumas = 0,753), taip pat padidėja F1 įvertis (0,459) ir *kapa* rodiklis (0,321), kas rodo geresnį bendrą modelio veikimą. Nors sumažėja tikslumas ir preciziškumas, toks kompromisas yra pagrįstas, nes leidžia efektyviau valdyti kredito riziką. Todėl galima daryti išvadą, kad optimalus slenkstis yra esminis veiksnys, lemiantis modelio naudą praktiniuose investavimo sprendimuose.



22 pav. CatBoost sumaišymo matrica skirtingiems slenksčiams

Sumaišymo matricų (22 pav.) analizė parodo esminį skirtumą tarp standartinio ir optimalaus slenksčio. Naudojant 0,5 slenkstį modelis labai retai prognozuoja nemokumą (tik 23 teisingi atvejai), todėl dauguma nemokių paskolų (289) lieka neidentifikuotos. Tuo tarpu sumažinus slenkstį iki 0,2, ženkliai padidėja teisingai atpažintų nemokių paskolų skaičius (235), nors kartu išauga klaidingai pažymėtų mokių paskolų skaičius (476). Tai rodo, kad mažesnis

slenkstis leidžia efektyviau identifikuoti rizikingas paskolas, kas yra svarbiau kredito rizikos valdymo ir investavimo kontekste, net jei padidėja klaidingų signalų kiekis.

3.5. Modelių palyginimas

Iš modelių rezultatų matome, kad tradicinis modelis (logaritminė regresija su *ElasticNet* reguliarizacija) stipriai nusileidžia modernesniems modeliams *CatBoost* ir atsitiktiniams miškams. Pastarieji modeliai demonstruoja gana panašius rezultatus. SVM šiuo atveju veikia panašiai kaip logistinė regresija.

Siekiant įvertinti sukurtų kredito rizikos modelių prognozavimo kokybę, buvo atliktas keturių modelių palyginimas: *CatBoost*, atsitiktiniai miškai (RF), logistinė regresija (Logit) ir branduolio SVM. Vertinimui naudotos dvi pagrindinės metrikos: ROC AUC ir preciziškumo-jautrumo AUC (PR AUC), taip pat analizuotos ROC, preciziškumo-jautrumo ir DET kreivės.

Lyginant modelių metrikas (7 lentelė), atsitiktiniai miškai ir *CatBoost* išsiskiria ir demonstruoja labai panašius rezultatus. Nors *Catboost* jautrumas yra šiek tiek aukštesnis, tačiau atsitiktiniai miškai turi didesnę bendro tikslumo ir preciziškumo metriką. Vis dėlto ši lentelė neduoda aiškaus atsakymo, kuris modelis geriausiai prognozuoja. Taip pat svarbu atkreipti dėmesį, kad nei vienas modelis su optimaliais slenksčiais nepasiekė to lygio, kai prognozuoja geriau nei atsitiktinis spėjimas: tikslumas < bazinis tikslumas visiems modeliams. Taip pat parametro *kapa* rodiklis pasiekė tik patenkinamą lygį (< 0,4). Galime daryti išvadą, kad tikslaus modelio nepavyko sukurti nei su vienu algoritmu, tačiau tai netrukdo rasti geriausią iš turimų ir naudoti modelio tikimybes.

7 lentelė. Visų modelių pagrindinės metrikos

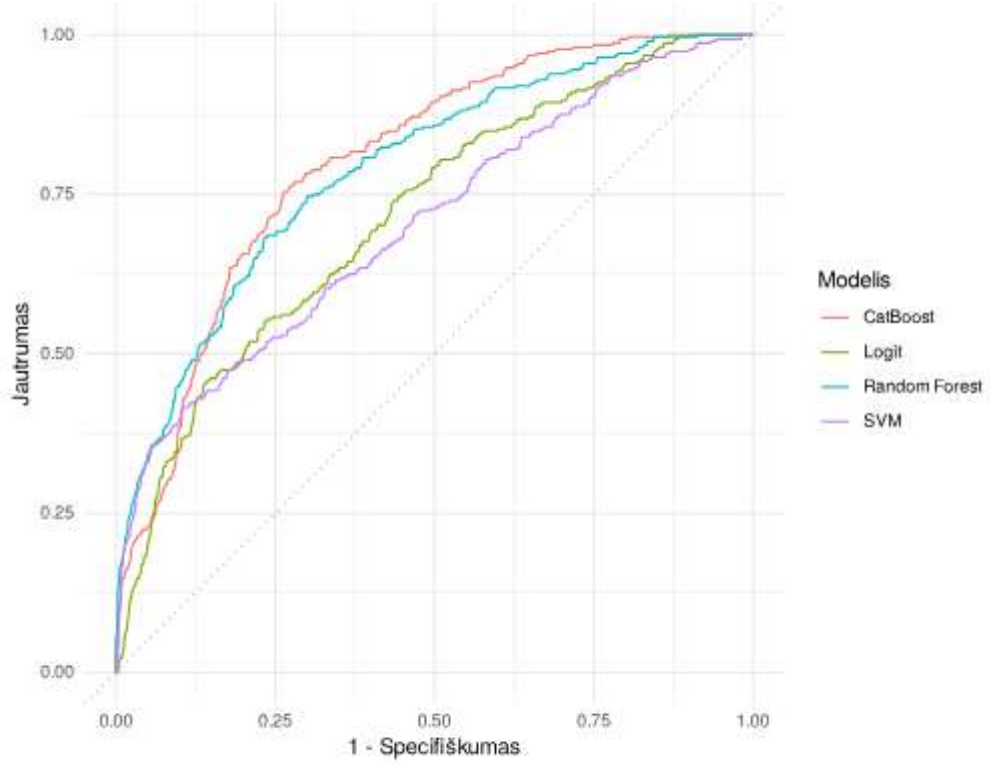
Metrika\Modelis	**Logit**	**SVM**	**RF**	**CatBoost**
Optimalus slenkstis	0,12	0,17	0,22	0,20
AUC	0,717	0,703	0,82	0,802
Tikslumas	0,814	0,825	0,755	0,74
Preciziškumas	0,325	0,408	0,336	0,331
Jautrumas	0,247	0,417	0,683	0,753
F1 įvertis	0,281	0,412	0,450	0,459
Kapa rodiklis	0,176	0,309	0,316	0,321
Bazinis tikslumas	0,853	0,853	0,853	0,853

Iš 8 lentelės matome, kad atsitiktinių miškų algoritmas turi geriausius tiek AUC (bendras atskyrimas tarp klasių), tiek PR AUC (atskyrimas tarp klasių, kai duomenys nesubalansuoti).

8 lentelė. Modelių AUC ir PER AUC

Algoritmas	**AUC**	**PR AUC**
Atsitiktiniai miškai	0,820	0,528
CatBoost	0,802	0,432
Logistinė regresija	0,717	0,306
SVM	0,703	0,386

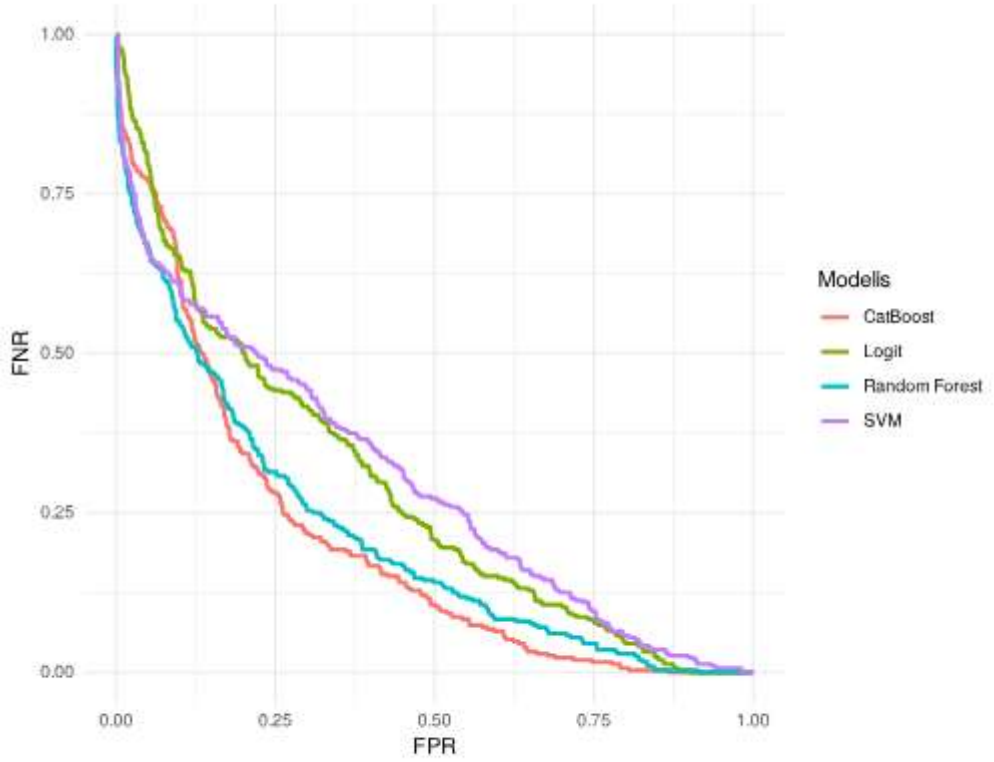
Iš ROC kreivių grafiko (20 pav.) atsitiktinių miškų ir *CatBoost* modeliai pasižymi geriausia diskriminacine geba, kas patvirtina aukštesnes AUC reikšmes. Abu modeliai efektyviai atskiria mokias ir nemokias paskolas visame slenksčių intervale, todėl sudėtinga iš akies nustatyti, kuris modelis yra patikimesnis.



23 pav. ROC kreivės

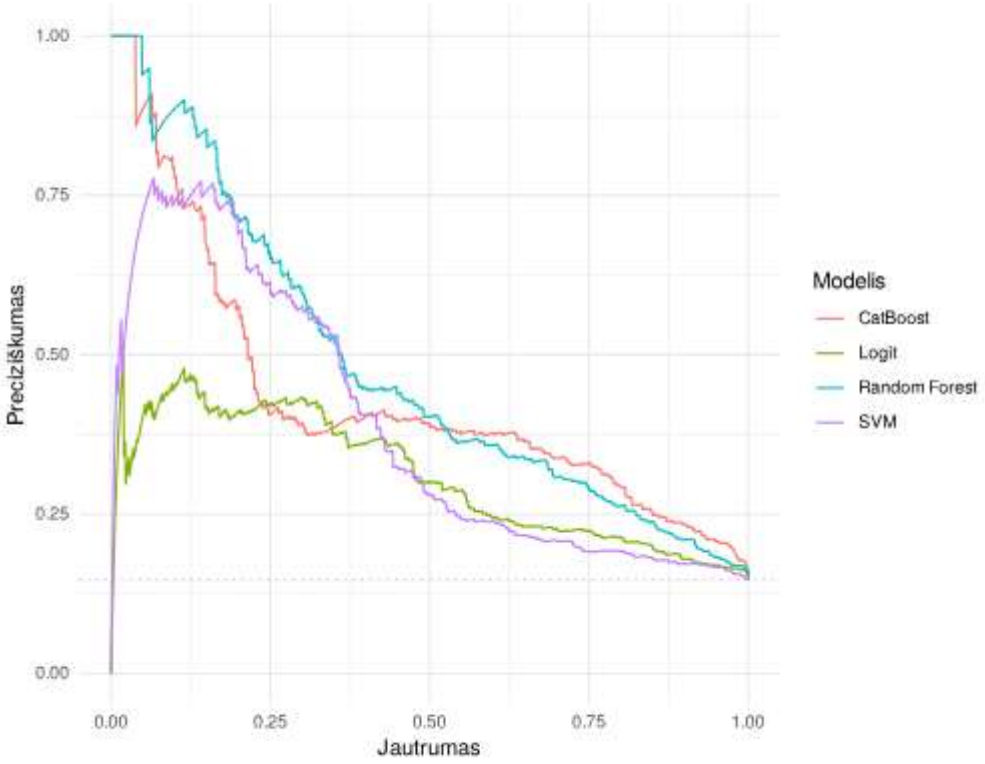
DET kreivių palyginimas rodo, kad *CatBoost* modelis pasiekia geriausią klaidų kompromisą, nes jo kreivė artimiausia apatiniam kairiajam grafiko kampui. Ji išlieka žemiausia plačiame FPR intervale, demonstruodama stabilų veikimą. Atsitiktiniai miškai nedaug atsilieka, tuo

tarpu SVM modelio pranašumas pastebimas tik prie itin mažo FPR – didėjant FPR jo klaidų rodikliai staigiai blogėja, tai rodo ribotą stabilumą.



24 pav. DET kreivės

Preciziškumo-jautrumo kreivė turimiems duomenims yra svarbiausia, kadangi prognozuojamo kintamojo klasės yra nesubalansuotos. Iš grafiko (25 pav.) aiškiai matoma, kad atsitiktiniai miškai dominuoja. *CatBoost* pradžioje taip pat rodo gerus rezultatus, bet didėjant jautrumo metrikos reikšmei, preciziškumas krenta, o atsitiktiniai miškai išlaiko aukštesnį preciziškumą prie didesnio jautrumo.

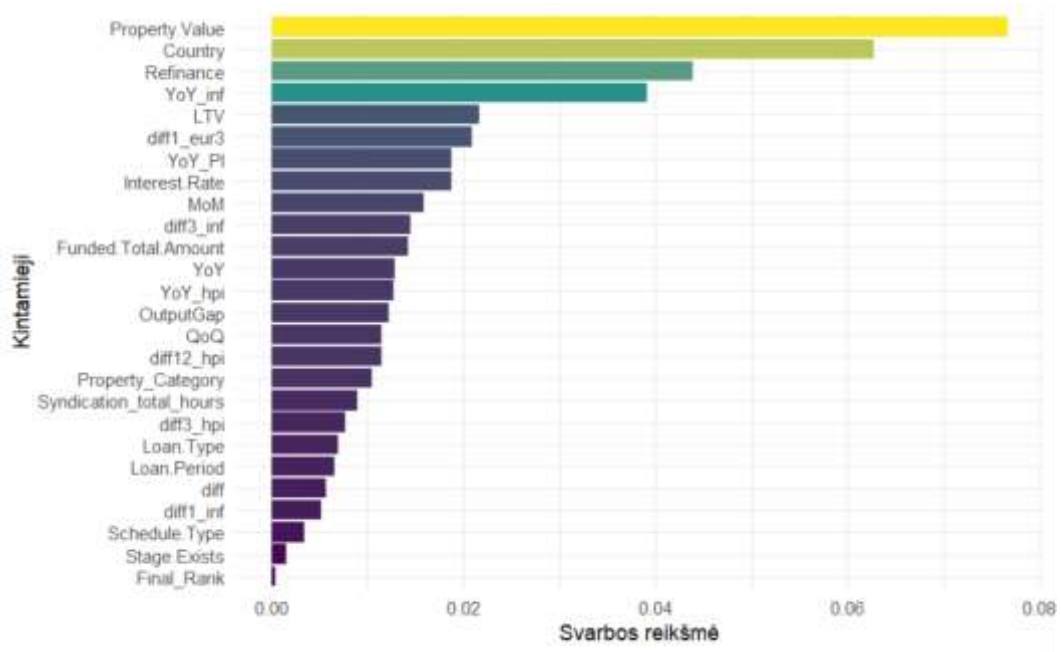


25 pav. PR kreivės

ROC ir DET kreivių analizė parodė, kad atsitiktinių miškų ir *CatBoost* modeliai pasižymi panašia diskriminacine geba. Tačiau preciziškumo-jautrumo kreivė atskleidė reikšmingus skirtumus – atsitiktinių miškų modelis išlaiko aukštesnį tikslumą didėjant preciziškumui, todėl geriau identifikuoja nemokias paskolas nesubalansuotų duomenų sąlygomis. Tai leidžia teigti, kad atsitiktiniai miškai yra tinkamiausias modelis praktiniam taikymui, kai svarbu efektyviai aptikti nemokias paskolas.

3.6. Kintamųjų svarbos analizė

Atsitiktiniai miškai pasirinkti kaip optimalus algoritmas, todėl būtent šiam algoritmui atliekama kintamųjų svarbos analizė. Kintamųjų svarba buvo vertinama taikant permutacinės svarbos metodą. Šis metodas įvertina, kiek pablogėja modelio prognozavimo tikslumas, kai konkretaus kintamojo reikšmės atsitiktinai permaišomos. Didesnė svarbos reikšmė rodo didesnį kintamojo indėlį į modelio prognozavimo gebėjimą.



26 pav. Kintamųjų svarba atsitiktinių miškų modelyje

Permutacinės svarbos analizė testavimo duomenų rinkinyje parodė, kad svarbiausi modelio kintamieji (26 pav.) yra susiję su užstato verte, geografine rinka bei makroekonominiais veiksniais. Didžiausią įtaką turi turto vertė (*Property Value*). Tai leidžia teigti, kad turto vertė yra vienas svarbiausių veiksnių, lemiančių modelio sprendimus ir prognozavimo tikslumą. Antroje vietoje pagal svarbą yra šalies indikatorius (*Country*), rodantis, kad geografiniai bei skirtingų rinkų ypatumai taip pat reikšmingai prisideda prie prognozių formavimo. Tarp svarbiausių kintamųjų taip pat išsiskiria refinansavimo požymis (*Refinance*) bei paskolos ir turto vertės santykis (*LTV*). Šie rezultatai rodo, kad modelis jautriai reaguoja į paskolos charakteristikas.

Reikšmingą vaidmenį atlieka makroekonominiai rodikliai, tokie kaip infliacija (*YoT_inf*), palūkanų normų pokyčiai (*diff1_eur3*), BVP (*MoM*) bei gamybos indeksas (*YoY*), leidžiantys modeliui atsižvelgti į ekonominio ciklo svyravimus. Šis faktas leidžia manyti, kad paskolų rinka stipriai susijusi su ekonomine situacija šalyse. Sprendimas įtraukti į modelį makroekonominius rodiklius pagrįstas.

Siekiant interpretuoti modelio sprendimus, taikoma SHAP analizė, leidžianti įvertinti ne tik kiekvieno kintamojo įtaką, bet ir poveikio kryptį. Šiame darbe SHAP analizė taikoma atsitiktinių miškų modeliui siekiant interpretuoti svarbiausių kintamųjų poveikį kredito rizikos prognozėms bei įvertinti, kaip finansiniai ir makroekonominiai rodikliai veikia paskolos nemokumo tikimybę. Skirtingai nei tradiciniai kintamųjų svarbos įverčiai, SHAP reikšmės leidžia nustatyti, ar konkretus kintamasis didina ar mažina nemokumo tikimybę. Kai kurių kintamųjų SHAP reikšmės neigiamos, o kitų – teigiamos. Teigiama SHAP reikšmė reiškia, kad kintamasis didina modelio prognozuojamą tikimybę priklausyti taikinio klasei (paskola yra nemoki), o neigiama reikšmė – mažina. Absoliuti SHAP reikšmė atspindi poveikio stiprumą, todėl didesnė absoliuti reikšmė rodo didesnį kintamojo indėlį į modelio prognozes.

SHAP analizės rezultatai (27 pav.) parodė, kad didžiausią poveikį atsitiktinių miškų modelio prognozėms turi turto vertė (*Property Value*), refinansavimo požymis (*Refinance*), paskolos ir turto vertės santykis (LTV), bendra finansavimo suma (*Funded Total Amount*) ir įkeisto turto kategorija (*Property Category*).

Neigiamos SHAP reikšmės rodo, kad didesnė turto vertė ir LTV mažina paskolos nemokumo tikimybę. Tai neatitinka teorijos, kur didesnis LTV turėtų didinti kredito riziką. Modelyje nustatytas neigiamas ryšys gali būti paaiškinamas platformos specifika: galbūt didesnės reikšmės LTV paskolas platforma suteikia tik atitinkant tam tikrus saugiklius, todėl realiuose duomenyse tokios paskolos rečiau nevykdomos – modelis tai galėjo išmokti kaip saugesnių paskolų požymį.

Teigiamos SHAP reikšmės rodo, kad refinansavimo atveju didėja paskolos nemokumo tikimybė, nors iš duomenų analizės rezultatų buvo priešingai. Procentiškai daugiau nerefinansuotų paskolų buvo nemokios nei refinansuotų paskolų (9 pav.)

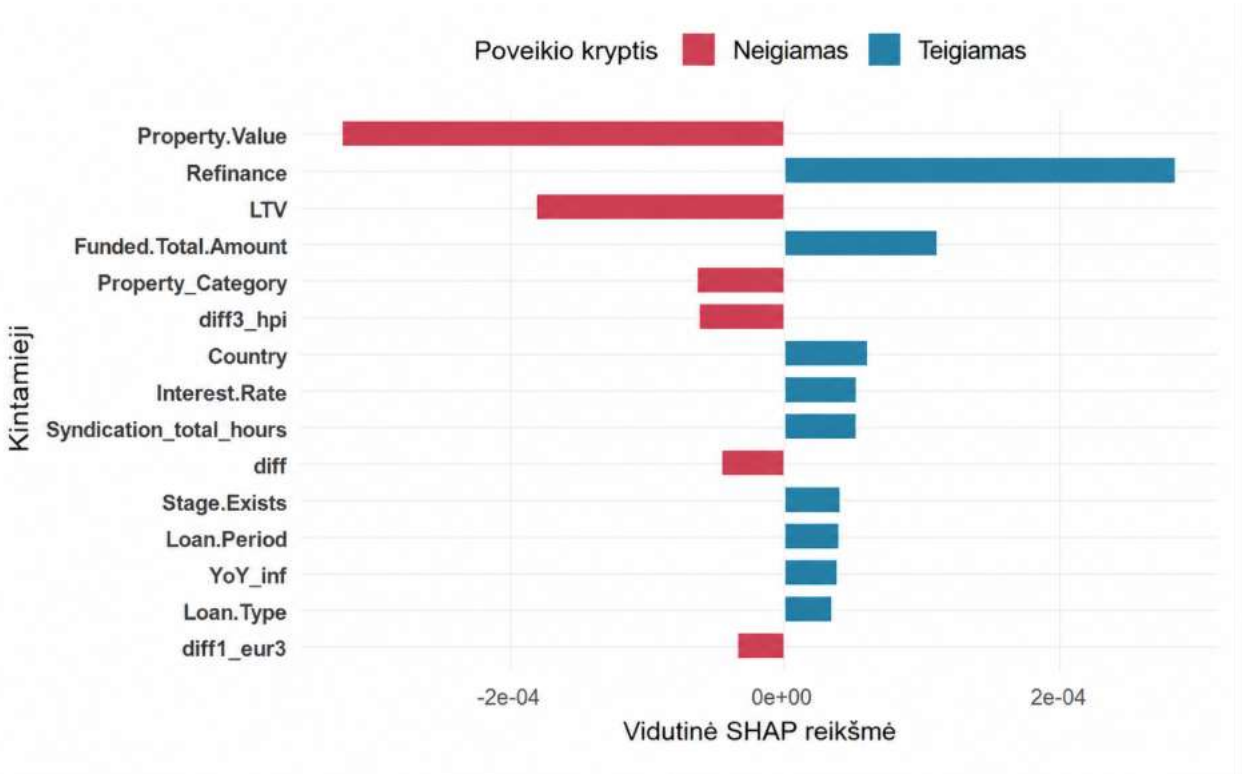
Reikšmingą poveikį taip pat turi įkeisto turto kategorija (*Property Category*). SHAP reikšmės rodo, kad tam tikros turto kategorijos yra susijusios su mažesne paskolos nemokumo rizika. Iš aprašomosios analizės rezultatų buvo nustatyta, kad komercinio nekilnojamojo įkeisto turto paskoloms būdinga didžiausia nemokių paskolų dalis (8 pav.).

Kintamojo finansuota paskolos suma (*Funded Total Amount*) teigiamos SHAP reikšmės rodo, kad didesnė suteikta paskolos suma didina nemokumo tikimybę. Tai ekonomiškai yra logiška, nes didesnė paskolos suma didina riziką.

Augantis būsto kainų indeksas (diff3_hpi) turi neigiamą poveikį nemokumo rizikai. Kai didėja būsto kainų indeksas, didėja nekilnojamojo turto (užstato) vertė. Tuomet gerėja skolininko finansinė padėtis ir stiprėja paskata vykdyti įsipareigojimus, todėl paskolos nemokumo rizika mažėja.

Makroekonominiai rodikliai, tokie kaip infliacija (*YoY_inf*) bei palūkanų norma (*diff1_eur3*), taip pat prisidėjo prie prognozių formavimo. Jų poveikis yra mažesnis nei pagrindinių finansinių kintamųjų. Tai rodo, kad ekonominė aplinka turi papildomą, tačiau ne dominuojantį vaidmenį modelio veikime.

Bendra SHAP analizė patvirtina ankstesnius kintamųjų svarbos rezultatus ir parodo, kad modelio prognozes labiausiai lemia turto charakteristikos, paskolos parametrai bei dalis makroekonominių rodiklių. Modelio vidinė svarba ir SHAP reikšmės papildo viena kitą. Pirmoji leidžia suprasti modelio struktūrą, o antroji – realų kintamųjų poveikį prognozėms. Vis dėlto ekonominiu požiūriu toks rezultatas reikalauja atsargaus interpretavimo: modelis identifikavo statistinį ryšį, bet tai nebūtinai reiškia priežastinę priklausomybę.



27 pav. TOP 15 kintamųjų pagal SHAP reikšmes

3.7. Investavimo strategijos

Antrinė magistro darbo dalis – investavimo strategijų kūrimas ir testavimas pasitelkiant rizikos modelį. Atlikus modelių analizę nustatyta, kad geriausiai veikė atsitiktinių miškų modelis. Strategijoms kurti panaudotos atsitiktinių miškų modelio prognozės (tikimybės, kad paskola bus nemoki), su kuriomis simuliuojama, kaip investuotojas galėtų paskirstyti lėšas ir kokią grąžą bei riziką gauti. Investavimo strategijų vertinimas atliekamas naudojant testavimo imtį, siekiant objektyviai įvertinti modelio taikymą naujiems, anksčiau nematytiems duomenims.

Remiantis ankstesne duomenų analize, tikimybių pasiskirstymas buvo suskirstytas į šias investavimo strategijas:

- visuotinio investavimo strategija;
- konservatyvi strategija;
- rizikos-grąžos optimizavimo strategija;
- agresyvi strategija;
- svertinė atrankos strategija.

Šiame darbe daroma konservatyvi prielaida, kad įsipareigojimų nevykdymo atveju investuotojas praranda visą investuotą sumą, todėl LGD laikomas lygiu 1. Dėl šios priežasties tikėtinas nuostolis apskaičiuojamas kaip *EAD×PD*. Ši konservatyvi prielaida sutampa su finansinio reguliavimo logika: kai nėra duomenų apie atgavimą arba paskola neturi užtikrinimo priemonių, taikoma blogiausio atvejo LGD = 100 % (IFRS 9, 2014). Taigi

šiame darbe pasirinktas atsargus vertinimas, pagrįstas praktika ir literatūros įžvalgomis (žr. 1.10 skyrelį), kad P2P sektoriuje nemokumo atveju investuotojas praranda visą investuotą sumą [8].

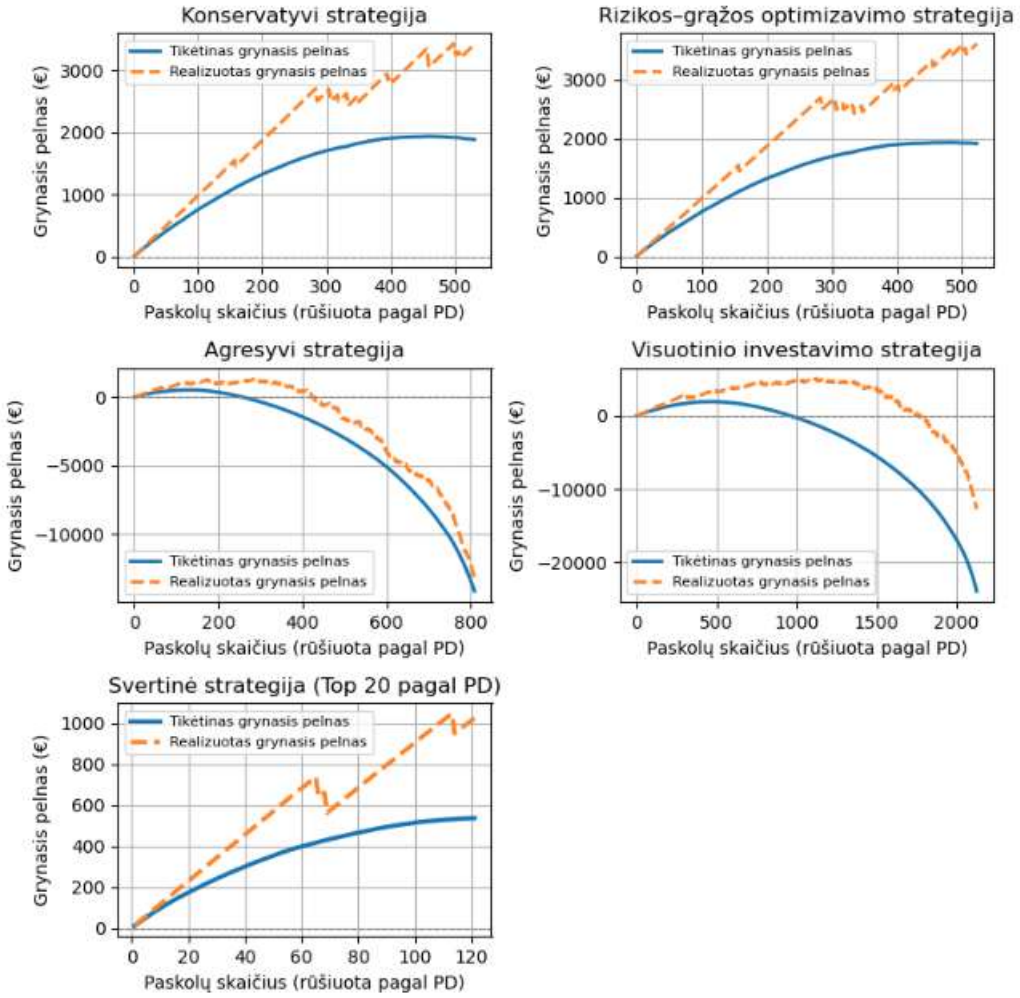
9 lentelė. Strategijų statistika: tikėtinas, realizuotas ir net pelnas, nuostolis ir grąža (%)

Strategija	Paskolų kiekis	Vidutinė PD	Tikėtinas pelnas	Tikėtinas Nuostolis	Tikėtinas Net pelnas	Tikėtina grąža, %	Realizuotas pelnas	Realizuotas Nuostolis	Realizuotas Net pelnas	Realizuota grąža, %
Svertinė (Top 15 %)	121	6,06	10,49	6,06	4,43	4,43	10,90	2,44	8,46	8,46
Rizikos-grąžos optimizavimo	522	5,81	9,49	5,81	3,68	3,68	9,80	2,87	6,93	6,93
Konservatyvi	530	5.82	9,37	5,82	3,55	3,55	9,65	3,21	6,44	6,44
Visuotinė	2124	19,50	8,24	19,50	–11,25	–11,25	8,70	1,69	–5,99	–5,99
Agresyvi	809	25,87	8,41	25,87	–17,46	–17,46	8,51	24,72	–16,21	–16,21

9 lentelėje pateikti strategijų taikymo rezultatai, suskaičiuoti remiantis 2.12.1-2.12.6 formulėmis. Investuojama suma EAD = 100, todėl šiuo atveju net pelnas sutampa su grąža. Pateikti skaičiai rodo vidutinę vieno paskolos vertę portfelyje. Remiantis kiekvienos strategijos statistika, galime atlikti analizę, kuri strategija yra pelningiausia, ir kokia rizika investuotojams, renkantis vieną iš strategijų. Analizuojant paskolų kiekius ir vidutinę nemokumo tikimybę (PD), matyti aiškūs skirtumai tarp strategijų. Svertinė (Top 15 %) strategija apima mažiausią paskolų skaičių (121), tačiau pasižymi santykinai žema nemokumo tikimybe, kas rodo griežtą atranką ir orientaciją į kokybiškiausias paskolas. Tuo tarpu rizikos-grąžos optimizavimo ir konservatyvi strategijos apima panašų ir gerokai didesnį paskolų kiekį (atitinkamai 522 ir 530), tačiau nemokumo tikimybė yra panaši (5,81 ir 5,82), kas leidžia užtikrinti diversifikaciją neprisiimant reikšmingai didesnės kredito rizikos.

Visuotinė ir agresyvi strategijos pasižymi ženkliai didesniu paskolų skaičiumi bei aukšta nemokumo tikimybės verte, ypač agresyvios strategijos atveju, kur stebima didžiausia nemokumo tikimybė. Tai rodo, kad didinant aprėptį ir mažinant atrankos griežtumą, į portfelį įtraukiama daugiau rizikingų paskolų.

Detalesni strategijų palyginimai pagal pelningumo ir grąžos rodiklius pateikiami grafiškai.



28 pav. Strategijų tikėtino ir realizuoto pelno kreivės

Iš 28 pav. matyti, kad žemos rizikos segmentas yra patikimas ir generuoja stabilų pelną. Konservatyvi strategija generuoja stabilų ir nuoseklų pelną, pasižymi maža rizika ir nedideliu rezultatų svyravimu. Tai rodo, kad modelis patikimai identifikuoja mažos rizikos paskolas. Taip pat realizuotas pelnas viršija tikėtiną, kas leidžia daryti išvadą, jog modelis yra konservatyvus ir linkęs pervertinti riziką.

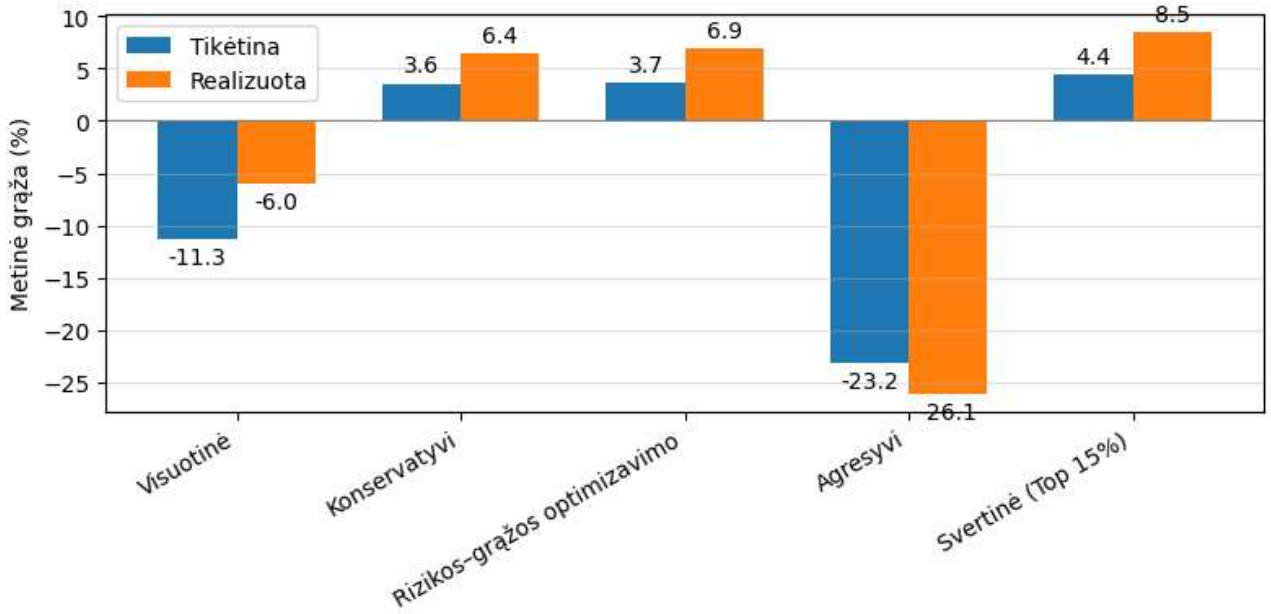
Rizikos-grąžos optimizavimo strategija pasižymi geriausiais rezultatais tarp visų nagrinėtų strategijų. Tai patvirtina, kad investicijų atranka pagal sąlygą $r \geq PD$ leidžia efektyviai eliminuoti nepelningas paskolas ir maksimaliai išnaudoti modelio teikiamą informaciją. Strategija užtikrina gerą balansą tarp rizikos ir grąžos.

Agresyvi strategija, paremta vien tik aukštomis palūkanomis, generuoja reikšmingus nuostolius. Tai rodo, kad didesnė palūkanų norma nekompensuoja padidėjusios kredito rizikos. Rezultatai patvirtina, kad ignoruojant PD rodiklį investavimo sprendimuose, portfelio veikimas tampa neefektyvus.

Visuotinio investavimo strategija, kai investuojama į visas paskolas be atrankos, lemia reikšmingus nuostolius. Tai rodo, kad vien diversifikacija nėra pakankama rizikos valdymo priemonė. Rezultatai pabrėžia būtinybę taikyti kredito rizikos vertinimą investavimo sprendimuose.

Svertinė strategija, paremta mažos rizikos paskolų prioretizavimu ir investicijų paskirstymu pagal PD, generuoja stabilų ir reikšmingą pelną.

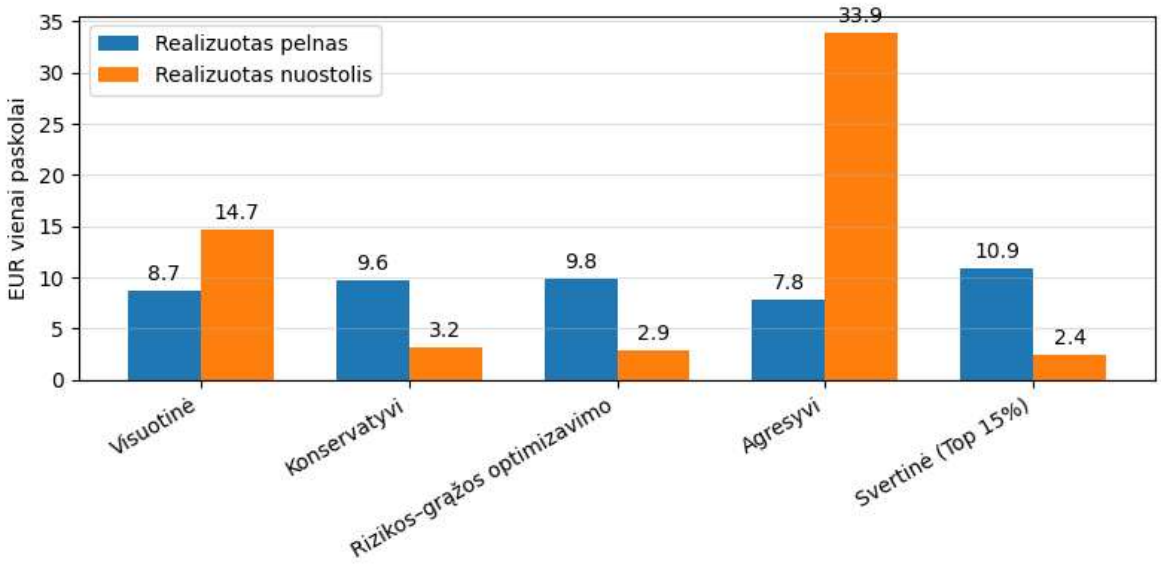
Gauti rezultatai (29 pav.) rodo aiškią priklausomybę tarp kredito rizikos ir investicinės grąžos. Dar kartą matyti, kad strategijos, paremtos PD įvertinimu, reikšmingai pranoksta atsitiktinį ar tik palūkanomis grįstą investavimą. Tai patvirtina, kad kredito rizikos modelis sukuria pridėtinę vertę investavimo sprendimams. Ypač efektyvi pasirodė svertinė investavimo strategija, kuri leidžia eliminuoti neigiamo tikėtino pelningumo paskolas. Tuo tarpu agresyvi bei visuotinio investavimo strategijos generuoja nuostolius, patvirtindamos, kad aukšta nominali grąža nėra pakankama kompensacija už padidėjusią kredito riziką.



29 pav. Tikėtina ir realizuota grąža (%) pagal strategijas

Pastebima, kad daugeliu atvejų realizuota grąža viršija tikėtiną, kas rodo, jog modelis linkęs pervertinti nemokumo tikimybę. Toks konservatyvus įvertinimas yra palankus rizikos valdymo požiūriu, nes leidžia išvengti pernelyg optimistinių investicinių sprendimų.

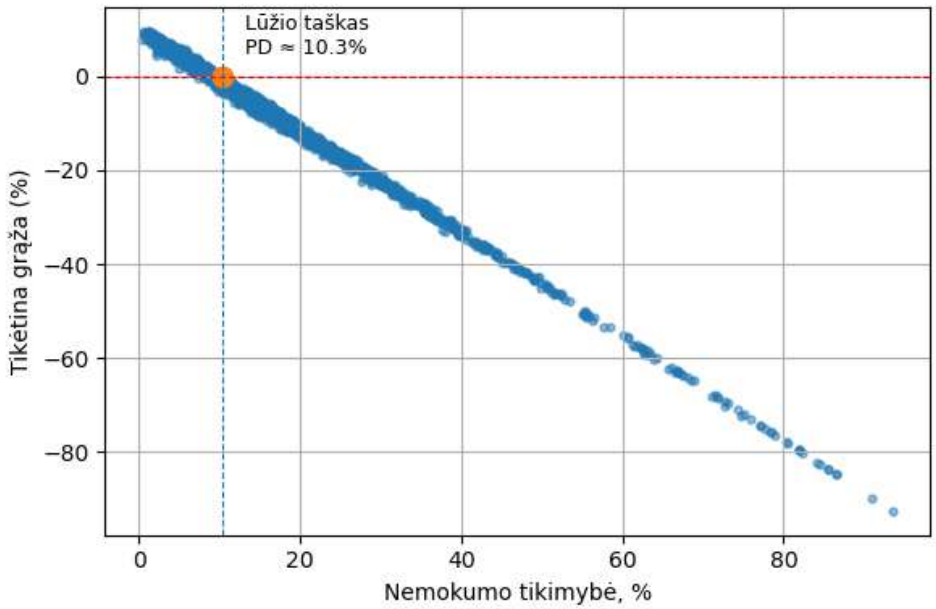
Rizikos diferencijavimas per svorius leidžia efektyviau išnaudoti modelio informaciją nei paprasta atranka.



30 pav. Realizuotas pelnas ir nuostolis pagal strategijas

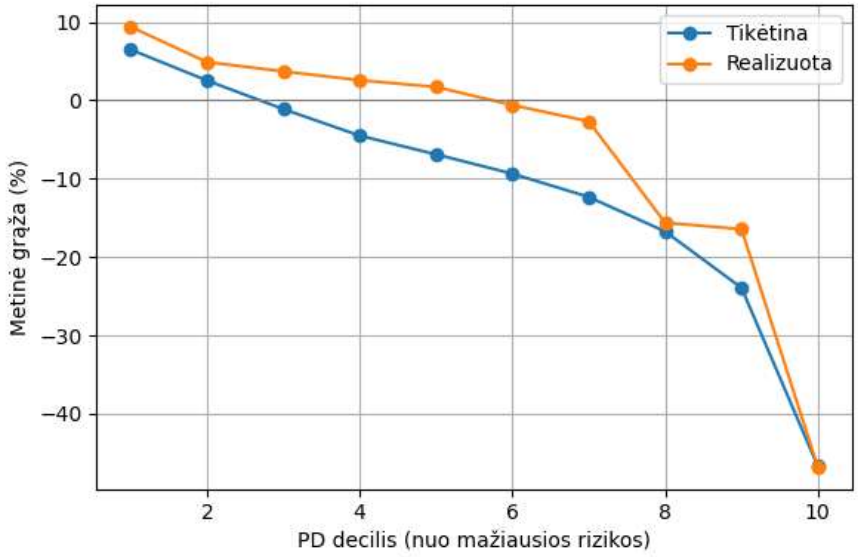
Iš 30 pav. matyti, kad strategijų skirtumus daugiausia lemia nuostolių valdymas, o ne pelno didinimas. Visų strategijų realizuotas pelnas yra gana panašus, tačiau nuostoliai labai stipriai skiriasi. Agresyvios strategijos nuostolis yra net 33,9 palyginus su svertinės strategijos tik 2,4. Konservatyvi strategija taip pat turi mažą nuostolį, tačiau jos esmė imti tik paskolas su mažiausiomis PD reikšmėmis. Svertinė strategija turi didžiausią pelną ir atitinkamai mažiausią nuostolį, vadinasi, svorių taikymas leidžia ne tik sumažinti riziką, bet ir padidinti pelningumą. Rizikos grąžos strategija turi panašų, šiek tiek mažesnį pelną ir didesnį nuostolį nei svertinė, tačiau tai kontroliuojama rizika, su mažais nuostoliais.

Analizuojant realizuoto pelno ir nuostolio komponentes nustatyta, kad pagrindinis investavimo strategijų efektyvumą lemiantis veiksnys yra ne gaunamų palūkanų dydis, o patiriamų nuostolių kontrolė. Strategijos, kurios efektyviai riboja nemokumo atvejus, generuoja teigiamą grąžą net ir esant panašiam palūkanų lygiui. Tuo tarpu aukštos grąžos paskolos pasižymi neproporcingai dideliais nuostoliais, kurie eliminuoja visą uždirbtą pelną.



31 pav. Paskolos tikėtinos grąžos priklausomybė nuo nemokumo tikimybės

31 pav. pateikta tikėtinos grynosios grąžos priklausomybė nuo nemokumo tikimybės. Matyti aiški neigiama priklausomybė – didėjant PD, tikėtina grąža nuosekliai mažėja. Grafike taip pat identifikuojamas lūžio taškas, kuriame grąža tampa neigiama, t. y. investicija tampa ekonomiškai nepatraukli. Šis rezultatas patvirtina, kad ne visos aukštų palūkanų paskolos yra pelningos, nes didesnė rizika ne visada yra pakankamai kompensuojama didesne palūkanų norma. Todėl investavimo sprendimuose būtina atsižvelgti į PD rodiklį, o ne vien į nominalią grąžą.



32 pav. Tikėtina ir realizuota grąža pagal nemokumo tikimybės decilius

Modelis teisingai reitinguoja paskolas pagal riziką – aukštesni PD deciliai atitinka blogesnius investicinius rezultatus (32 pav.). Mažiausios rizikos paskolos generuoja stabilų ir prognozuojamą pelną. Matyti, kad egzistuoja PD riba, nuo kurios investicijos tampa

nuostolingos. Riba apibrėžiama, kai kreivės kerta x ašį ties nuliu. Visuose deciliuose realizuota grąža > tikėtina, vadinasi, modelis sistemingai pervertina PD. Tokį modelį galime apibūdinti kaip konservatyvų.

Pritaikę keletą skirtingų strategijų, galime daryti išvadą, kad net ir vidutinio tikslumo modelis gali būti labai naudingas, jei jis teisingai reitinguoja riziką. Pateikta investavimo strategijų analizė patvirtina, kad esminis veiksnys, lemiantis portfelio rezultatą, yra ne maksimalios grąžos siekimas, bet efektyvus kredito rizikos valdymas. Strategijos, kurios remiasi modelio prognozuota nemokumo tikimybe ir taiko atrankos ar svorių paskirstymo principus, leidžia suformuoti stabilesnius ir kokybiškesnius paskolų portfelius.

Tuo tarpu investavimas be aiškios rizikos kontrolės, ypač orientuojantis į aukštesnes palūkanas, lemia didesnę nemokumo tikimybę ir prastesnius rezultatus. Tai rodo, kad kredito rizikos modelio integravimas į investavimo sprendimus yra būtina sąlyga siekiant ilgalaikio investavimo efektyvumo.

Išvados ir rekomendacijos

1. Atlikta mokslinės literatūros analizė parodė, kad kredito rizikos vertinime P2P rinkoje tradicinius statistinius metodus vis dažniau papildo pažangūs mašininio mokymosi algoritmai. Pastarieji lenkia tradicinius statistinius modelius prognozių tikslumu. Taip pat pabrėžiama svarba integruoti modelių rezultatus į investavimo sprendimus, taip siekiant pagerinti rizikos ir grąžos balansą investuotojų portfeliuose. Nustatyta, kad kredito rizikos vertinimui vis svarbesni tampa makroekonominiai rodikliai bei modelių interpretavimo metodai.
2. Sutelktinio finansavimo platformos „EstateGuru“ duomenų analizė su požymių inžinerija leido identifikuoti reikšmingiausius kredito rizikos veiksnius. Kintamųjų svarbos ir SHAP analizė parodė, kad didžiausią įtaką kredito rizikos prognozėms turėjo turto vertė, refinansavimo požymis, paskolos ir turto vertės santykis (LTV), šalies indikatorius bei dalis makroekonominių rodiklių. SHAP metodas leido interpretuoti ne tik požymių svarbą, bet ir jų poveikio kryptį, atskleidžiant, kad didesnė turto vertė buvo siejama su mažesne paskolos nemokumo tikimybe.
3. Makroekonominiai rodikliai padidino modelio informatyvumą ir tikslumą. Į modelius įtraukus rodiklius (pvz., infliaciją, palūkanų normas, NT kainų indeksą), pastebėtas bendras klasifikavimo kokybės pagerėjimas. Kintamųjų svarbos analizė atskleidė, kad be paskolų charakteristikų (pvz., LTV, turto vertės), dalis makroekonominių veiksnių patenka tarp reikšmingiausių požymių modelio prognozėms. Tai rodo, kad modelis su makroekonomikos kintamaisiais geriau atspindi sisteminę riziką ir ekonominių ciklų poveikį skolininkų nemokumo rizikai.
4. Tradicinių statistinių ir ML metodų palyginimas patvirtino, kad geriausi prognozavimo rezultatai pasiekti pritaikius atsitiktinių miškų modelį. Jis pranoko kitus modelius diskriminacinės gebos ir nesubalansuotų duomenų vertinimo požiūriu (ROC AUC ≈ 0,82 ir PR AUC ≈ 0,53, palyginti su ≤ 0,80 ROC AUC ir žemesnėmis PR AUC alternatyviuose modeliuose). Be to, atsitiktiniai miškai demonstravo aukščiausią jautrumą (~ 68 %), kas rodo jog efektyviausiai identifikuoja rizikingas paskolas. Pažymėtina, kad atsitiktiniai miškai pasižymėjo „konservatyvumu“ – modelio optimalaus klasifikavimo slenkstis (PD ~22 %) buvo aukštesnis nei kitų modelių. Todėl šis algoritmas griežčiau atrenka rizikingas paskolas, taip sumažindamas klaidingai „geromis“ priskiriamų rizikingų paskolų tikimybę. Toks požiūris investavimo kontekste yra pagrįstas, nes kredito rizikos vertinime svarbiau laiku aptikti potencialiai nemokias paskolas nei maksimaliai padidinti bendrą klasifikavimo tikslumą. Konservatyvus modelio veikimas leido efektyviau filtruoti aukštesnės rizikos paskolas ir prisidėjo prie stabilesnės rizikos.
5. Investavimo strategijų kūrimo ir vertinimo rezultatai parodė, kad kredito rizikos modelių prognozės veiksmingai pritaikomos investiciniams sprendimams. Remiantis atsitiktinių miškų modelio nemokumo tikimybių prognozėmis, buvo suformuotos penkios investavimo strategijos (visuotinė, konservatyvi, rizikos-grąžos optimizavimo, agresyvi ir svertinė atranka) ir atlikti jų rizikos-grąžos vertinimai. Lyginamoji analizė atskleidė, kad portfeliai su rizikos modelio filtrais reikšmingai lenkia nefiltruotą investavimą: svertinės atrankos strategija pasiekė aukščiausią tikėtiną ir realizuotą grąžą, viršydama tiek grynai

konservatyvų portfelį, tiek agresyvų (rizikingiausių paskolų) variantą. Taikant rizikos modelio grįstas investavimo taisykles, pavyko optimizuoti rizikos ir grąžos pusiausvyrą. Investuotojas gali gauti didesnę grąžą išlaikydamas žemą rizikos lygį, lyginant su portfeliais, formuojamais vien pagal žemiausią riziką arba vien pagal aukščiausias palūkanas. **Rekomendacija praktikai:** investuotojams P2P platformose verta diversifikuoti portfelius, naudoti kredito rizikos modelių išvadas paskolų atrankai ir derinti rizikos bei pelningumo kriterijus – tuomet gerėja portfelio stabilumas ir rizikos koreguota grąža.

Rekomendacijos. Rekomenduojama integruoti makroekonominius rodiklius į kredito rizikos vertinimo modelius, siekiant atsižvelgti į ekonominių ciklų svyravimų poveikį paskolų nemokumo tikimybei. Tyrimo rezultatai parodė, kad tikslinga taikyti atsitiktinių miškų modelį kredito rizikos vertinime. Atsižvelgiant į šio modelio polinkį konservatyviai vertinti riziką, būtina užtikrinti tinkamą rezultatų interpretavimą. Galiausiai investavimo sprendimus rekomenduojama grįsti prognozuotomis nemokumo tikimybėmis ir taikyti svertinę investavimo strategiją, kurios rizikos–grąžos santykis buvo geriausias.

Literatūros sąrašas

1. ZENG, Ming, SHEN, Yu ir kt. Survival analysis for default prediction in P2P lending. European Journal of Operational Research, 2021, vol. 293, no. 2.
2. SERRANO-CINCA, Carlos ir GUTIÉRREZ-NIETO, Begoña. Predicting default in P2P lending using a lending-based scoring system. Decision Support Systems, 2016, vol. 89, p. 1–10.
3. MICHELS, Jared. Do unverifiable disclosures matter? Evidence from peer-to-peer lending. The Accounting Review, 2012, vol. 87, no. 4, p. 1385–1413.
4. FREEDMAN, Seth ir JIN, Ginger Zhe. Learning from peer-to-peer lending. Journal of Economic Perspectives, 2017, vol. 31, no. 3, p. 25–44.
5. MALEKIPIRBAZARI, Milad ir AKSAKALLI, Vural. Risk assessment in social lending via random forests. Expert Systems with Applications, 2015, vol. 42, no. 10, p. 4621–4631.
6. LIU, Yiting, BAALS, Lennart John, OSTERRIEDER, Jörg ir HADJI-MISHEVA, Branka. Leveraging network topology for credit risk assessment in P2P lending: A comparative study under the lens of machine learning. Expert Systems with Applications, 2022, vol. 203.
7. BAESENS, Bart ir SMEDTS, Kristien. Boosting credit risk models. Journal of Credit Risk, 2020, vol. 16, no. 4.
8. WANG, Yan ir NI, Xuelei Sherry. Improving investment suggestions for peer-to-peer (P2P) lending via integrating credit scoring into profit scoring. Journal of Risk and Financial Management, 2023, vol. 16, no. 2.
9. SANZ-GUERRERO, Mario, ARROYO, Javier ir CINCA, Carlos. Credit risk meets large language models: Building a risk indicator from loan descriptions in P2P lending. Expert Systems with Applications, 2023.
10. ZIEGLER, Tania et al. The European Alternative Finance Benchmarking Report. Cambridge Centre for Alternative Finance, 2021.
11. ESTATEGURU. Annual Report. Tallinn: EstateGuru, įvairūs metai.
12. LIETUVOS BANKAS. Sutelktinio finansavimo ir tarpusavio skolinimo rinkos apžvalga. Vilnius: Lietuvos bankas.
13. NAVICKAS, Valentinas ir GUDAITIS, Tomas. Crowdfunding in Lithuania: Development trends and challenges. Entrepreneurship and Sustainability Issues, 2019.
14. HE, Haibo ir GARCIA, Edwardo A. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 2009, vol. 21, no. 9, p. 1263–1284.
15. HAN, Jiawei, KAMBER, Micheline ir PEI, Jian. Data Mining: Concepts and Techniques. 3rd ed. Burlington: Morgan Kaufmann, 2011. ISBN 978-0-12-381479-1.
16. SAITO, Takaya ir REHMSMEIER, Marc. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE, 2015, vol. 10, no. 3.

17. BUSSMANN, Niklas, GIUDICI, Paolo, MARINELLI, Dimitri ir PAPENBROCK, Jochen. Explainable AI in credit risk management. Computational Economics, 2021, vol. 57, no. 1, p. 203–216.
18. MOLNAR, Christoph. Interpretable Machine Learning. 2nd ed. 2022. Prieiga per internetą: https://christophm.github.io/interpretable-ml-book/
19. TRINH, Lua Thi. A comparative analysis of consumer credit risk models in Peer-to-Peer lending. Cogent Economics & Finance, 2022, vol. 10, no. 1.
20. DUARTE, Jefferson, SIEGEL, Stephan ir YOUNG, Lance. Trust and credit: The role of appearance in peer-to-peer lending. Review of Financial Studies, 2012, vol. 25, no. 8, p. 2455–2484.
21. DOM, Barbara, ILLES, Ferenc ir OLVEDI, Tímea. Peer-to-peer lending: Legal loan sharking or altruistic investment? Analyzing platform investments from a credit risk perspective. Society and Economy, 2021, vol. 43, no. 3.
22. ZHANG, Yuchen, ZHOU, Yao ir kt. Detecting fraud in P2P lending using supervised learning. Decision Support Systems, 2019, vol. 125.
23. XIA, Yufei, LIU, Chuanyou ir kt. A boosted decision tree approach using survival analysis for loan default prediction. Knowledge-Based Systems, 2017, vol. 134, p. 183–192.
24. CHEN, Ricky T. Q., WANG, Yifei ir kt. DeepSurv: Applying deep neural networks to survival analysis. BMC Medical Research Methodology, 2020, vol. 20, no. 1.
25. HODRICK, Robert J. ir PRESCOTT, Edward C. Postwar U.S. business cycles: An empirical investigation. Journal of Money, Credit and Banking, 1997, vol. 29, no. 1, p. 1–16.
26. AYDEMIR, R. et al. 2023. Assessing Credit Risk in Different Banking Systems under Macroeconomic Shocks.
27. IBRAHIM HALIL SUGOZU, Can Verberi ir Sema Yasar. Machine learning approaches to credit risk: Evaluating Turkish participation and conventional banks. Borsa Istanbul Review, 2025, t. 25, nr. 3, p. 497–512.
28. BROWN, Ian ir MUES, Christophe. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 2012, vol. 39, no. 3, p. 3446–3453.
29. SUN, Yanmin, WONG, Andrew K. C. ir KHAN, Mohamed S. M. Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence, 2007, vol. 23, no. 4, p. 687–719.
30. CHAWLA, Nitesh V. et al. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, vol. 16, p. 321–357.
31. LESSMANN, Stefan et al. Benchmarking state-of-the-art classification algorithms for credit scoring. European Journal of Operational Research, 2015, vol. 247, no. 1, p. 124–136.
32. LUNDBERG, Scott M. ir LEE, Su-In. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 2017, vol. 30.
33. ADDO, Peter Martey, GUEGAN, Dominique ir HASSANI, Bertrand. Credit risk analysis using machine and deep learning models. Risks, 2018, vol. 6, no. 2.

34. MARKOWITZ, Harry. Portfolio selection. The Journal of Finance, 1952, vol. 7, no. 1, p. 77–91.
35. SHARPE, William F. Capital asset prices: A theory of market equilibrium under conditions of risk. The Journal of Finance, 1964, vol. 19, no. 3, p. 425–442.
36. INVESTOPEDIA. Expected return [interaktyvus]. [žiūrėta 2026-05-01]. Prieiga per internetą: https://www.investopedia.com/terms/e/expectedreturn.asp
37. DAVIS, Jesse ir GOADRICH, Mark. The relationship between Precision–Recall and ROC curves. Proceedings of the 23rd International Conference on Machine Learning, 2006, p. 233–240.
38. MERTON, Robert C. On the pricing of corporate debt: The risk structure of interest rates. The Journal of Finance, 1974, vol. 29, no. 2, p. 449–470.
39. HILBE, Joseph M. Logistic Regression Models. Boca Raton: Chapman and Hall/CRC, 2009. ISBN 978-1-4200-7575-5.
40. ZOU, Hui ir HASTIE, Trevor. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005, vol. 67, no. 2, p. 301–320.
41. BREIMAN, Leo. Random forests. Berkeley: Statistics Department, University of California, 2001.
42. ABE, Shigeo. Support Vector Machines for Pattern Classification. 2nd ed. London: Springer, 2010. ISBN 978-1-84996-097-7.
43. CATBOOST. CatBoost documentation [interaktyvus]. [žiūrėta 2026-04-07]. Prieiga per internetą: https://catboost.ai/docs/en/
44. PROKHORENKOVA, Liudmila, GUSEV, Gleb, VOROBEV, Aleksandr, DOROGUSH, Anna Veronika ir GULIN, Andrey. CatBoost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 2018, vol. 31.
45. COHEN, Jacob. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, vol. 20, no. 1, p. 37–46.
46. FAWCETT, Tom. An introduction to ROC analysis. Pattern Recognition Letters, 2006, vol. 27, no. 8, p. 861–874.
47. MARTIN, Alvin et al. The DET curve in assessment of detection task performance. Proceedings of Eurospeech, 1997.

Priedai

1 priedas. „EstateGuru“ svetainėje pateikiamų duomenų aprašymas

Loan code - unikalus identifikatorius paskolai „EstateGuru“ platformoje.
Country - šalis, kurioje yra skolininkas / nekilnojamas turtas pagal „EstateGuru“ .
Status - paskolos būklė, gali įgyti šias reikšmes: „Funded“, „Repaid“, „Late“, „Recovered“, „In Default“, „Partially“, „Recovered“, „Fully Invested“
Interest Rate - palūkanų norma, kurią moka skolininkas. Tai svarbu investuotojams.
Schedule Type - mokėjimų tvarkaraštis/grafikas: gali įgyti šias reikšmes: „Full bullet“, „Bullet“, „Annuity“.
Loan Type - paskolos tipas: „development loan“, „bridge loan“, „business loan“. „EstateGuru“ platformoje yra skirtingi paskolų tipai.
Property Type - užstato („collateral“) tipas – nekilnojamojo turto tipas: gyvenamasis, komercinis, statomas turtas ir t.t.
Stage Number - jeigu paskola yra „stage loan“ (etapinė paskola), tai etapo numeris (pvz., etapas 1, 2 …). „EstateGuru“ naudoja etapines paskolas, kai statomas objektas.
LTV - Loan-to-Value santykis – paskolos suma palyginus su turto verte. „EstateGuru“ maksimalus LTV yra ~75 %. intercom-help.eu
Loan Period - paskolos laikotarpis – per kiek laiko skolininkas turi grąžinti paskolą (pvz., mėnesiai ar metai).
Funded Total Amount - iš viso paskolai surinkta (finansuota) pinigų suma per „EstateGuru“ investuotojus.
Currency - Paskolos valiuta (pvz., EUR).
Fully Invested Date - data, kada paskola buvo visiškai investuota (kai nebeliko laisvų investavimo „vietų“ paskoloje).
Funded Date - data, kai prasidėjo finansavimas – kai paskola pradėjo rinkti investicijas.
Next Payment Date - sekanti data, kada turi būti mokėjimas (palūkanos / pagrindinė dalis), pagal paskolos grafiką.

Initial Maturity Date - pradinė paskolos terminuotės data – kada pagal sutartį paskola turėtų pasibaigti, jeigu viskas vyktų kaip planuota.
Maturity Date - faktinė data, kada paskola bus grąžinta / pasibaigs pagal pakeitimus ar realią būklę.
Repaid Date - data, kada paskola iš tikrųjų buvo pilnai grąžinta (jei tai jau įvyko).
Syndication Period - laikotarpis, per kurį paskola buvo sindikuojama, t. y. finansuojama kelių investuotojų.
Actual Return - faktinė grąža investuotojui, apskaičiuota remiantis gautomis palūkanomis, premijomis, galimais nuostoliais ir pan. „EstateGuru“ apskaičiavimas:
Principal paid - suma, kurią investuotojai jau gavo atgal iš pagrindinės paskolos sumos („principal“), t. y. skolininko grąžinta dalis.
Interest paid - iš viso sumokėta palūkanų investuotojams.
Bonus paid (borrower) - papildomas „bonusas“, kurį skolininkas sumokėjo investuotojams (jei „EstateGuru“ tokį moka).
Penalty paid - baudos, kurias skolininkas sumokėjo dėl vėlavimo ar kitų sutartinių pažeidimų.

Indemnity Paid - kompensacija (indemnity), kurią skolininkas gali mokėti už tam tikrus įsipareigojimus.
Property Value - nekilnojamojo turto vertė, pagal kurią „EstateGuru“ vertina užstatą (collateral).
First Ranking - pirmoji hipoteka („first ranking mortgage“) – suma, kurią pirmojo rango hipoteka dengia. „EstateGuru“ sako, kad daug paskolų yra pirmojo rango.
Second Ranking - antro rango hipoteka („second ranking mortgage“) – papildoma hipoteka, jei yra antras hipoteko davėjas.
Total Delay Days - iš viso dienų, kiek paskola vėlavo mokėjimais („delayed“), per visą paskolos gyvavimo laiką.
Number of payments delayed - kiek kartų paskolos mokėjimai buvo atidėti / vėlavo (pvz., kiek įmokų vėlavo).
Prolonged - ar paskola buvo prailginta („extension“) – taip / ne. „EstateGuru“ leidžia prailginti paskolas, jei reikia daugiau laiko.
Stage Exists - ar paskola turi kelis etapus (stages), t. y. ar tai yra „stage loan“.
Refinance - ar paskola buvo refinansuota (pvz., pakeistas finansavimo šaltinis, pratęstas terminas).
Outstanding Principal - dabartinė (likusi) nepilnai grąžinta pagrindinė paskolos suma; investuotojams dar nepriskirta grąža iš pagrindinės skolos dalies.
Recovered Principal - pagrindinė paskolos dalis, kuri jau buvo susigrąžinta iš vėluojančių ar nutrauktų paskolų (pvz., per atgavimo („recovery“) procesą).
Written Off - paskolų suma, kuri buvo „nurašyta“ („written off“) – manoma, kad jos nebus pilnai grąžintos.