Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Mašininio mokymosi metodų pritaikymas klientų kreditingumui vertinti

Baigiamasis magistro studijų projektas

Kamilė Baltrušonė

Projekto autorė

Prof. dr. Robertas Alzbutas

Vadovas

Doc. dr. Šviesa Leitonienė

Vadovė

Kaunas, 2020

Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Mašininio mokymosi metodų pritaikymas klientų kreditingumui vertinti

Baigiamasis magistro studijų projektas

Didžiųjų verslo duomenų analitika (kodas: 6213AX001)

Kamilė Baltrušonė

Projekto autorė

Prof. dr. Robertas Alzbutas

Vadovas

Doc. dr. Šviesa Leitonienė

Vadovė

Lekt. dr. Lina Dindienė

Recenzentė

Doc. dr. Kristina Kundelienė

Recenzentė

Kaunas, 2020

Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Kamilė Baltrušonė

Mašininio mokymosi metodų pritaikymas klientų kreditingumui vertinti

Akademinio sąžiningumo deklaracija

Patvirtinu, kad mano, Kamilės Baltrušonės, baigiamasis projektas tema „Mašininio mokymosi metodų pritaikymas klientų kreditingumui vertinti“ yra parašytas visiškai savarankiškai ir visi pateikti duomenys ar tyrimų rezultatai yra teisingi ir gauti sąžiningai. Šiame darbe nei viena dalis nėra plagijuota nuo jokių spausdintinių ar internetinių šaltinių, visos kitų šaltinių tiesioginės ir netiesioginės citatos nurodytos literatūros nuorodose. Įstatymų nenumatytų piniginių sumų už šį darbą niekam nesu mokėjęs.

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

(vardą ir pavardę įrašyti ranka)	(parašas)

Baltrušonė, Kamilė. Mašininio mokymosi metodų pritaikymas klientų kreditingumui vertinti. Magistro baigiamasis projektas / vadovai: prof. dr. Robertas Alzbutas, doc. dr. Šviesa Leitonienė; Kauno technologijos universitetas, Matematikos ir gamtos mokslų fakultetas.

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

Reikšminiai žodžiai: kreditingumo vertinimas, klasterizavimas, klasifikavimas.

Kaunas, 2020. 80 p.

Santrauka

Didmeniniams pirkėjams teikiama apmokėjimo atidėjimo galimybė t.y. suteikiamas prekybos kreditas turi didelę įtaką tiek ilgalaikiui bendradarbiavimui tarp įmonių, tiek abiejų įmonių apyvartinių lėšų valdymo grandinėje. Kadangi įmonėje netgi iki 25 proc. trumpalaikio turto gali būti sudaryta iš tokio išduodamo prekybos kredito, jo išdavimas turi būti kontroliuojamas ir kiekvienas pirkėjas turi būti vertinamas tiek prieš pirmosios sąskaitos išrašymą, tiek ir viso tolimesnio bendradarbiavimo metu. Toks pirkėjo kreditingumo vertinimas gali būti atliekamas pagal iš išorinių šaltinių gauta informacija, tačiau toks vertinimas gali būti atnaujinamas geriausiu atveju kas pusmetį ir yra mokamas. Tuo tarpu pačios įmonės viduje yra kaupiama pirkėjo mokumo istorija, pagal kurią galima matyti pirkėjo einamąją padėtį ir atnaujinti jo kreditingumo vertinimą einamuoju momentu.

Darbo siekis yra įmonių verslo kreditingumui vertinti pritaikyti mašininio mokymosi metodus siekiant: sukurti priemonę, kuria naudojantis būtų vykdomas pirkėjų klasterizavimas bei patikslintas ekspertinis kreditingumo vertinimas. Vėliau naudojantis šia informacija, nauji pirkėjai ar pirkėjai su pasikeitusiais verslo duomenimis automatiškai būtų klasifikuojami į skirtingus kreditingumo lygmenis. Tyrimo metu buvo atliktas pirkėjų klasterizavimas remiantis jų 2018 metų mokumo istorija naudojant *K-vidurkių metodą*. 38,99 proc. klasterizavimo rezultatų atvejai buvo patikslinti vykdant ekspertinį kreditingumo vertinimą. Tuomet, buvo atliktas mašininis apmokymas ir palyginti automatinio klasifikavimo rezultatai gauti *Atraminių vektorių su radialine branduolio funkcija* ir A*tsitiktinių miškų metodais* bei nustatyta, jog taikant pastarąjį metodą galima teisingai įvertinti kreditingumą ir suklasifikuoti net 90 proc. naujų pirkėjų, turint informaciją tik apie jų mokumą. Ištyrus klasifikavimo rezultatų tikslumo priklausomybę nuo kintamųjų, buvo atrinkti šeši kintamieji, kurie turi daugiausiai įtakos klasifikatoriaus rezultatui.

Baltrušonė, Kamilė. Machine learning methods application for assessment of customer creditworthiness. Master's Final Degree Project / supervisors: Prof. Dr. Robertas Alzbutas and Assoc. Prof. Dr. Šviesa Leitonienė; Faculty of Mathematic and Natural sciences, Kaunas University of Technology.

Study field and area (study field group): Mathematical sciences.

Keywords: creditworthiness assessment, clustering, classification.

Kaunas, 2020. 80 pages.

Summary

The choice of deferring payment i.e. granting trade credit to customers has a significant impact on both long-term relationship between companies and the working capital management chain of both companies. Because up to 25 percent of short-term assets may consist of such trade credit, the granting of this credit must be controlled, and each customer must be assessed both before the first invoice is issued and throughout further cooperation. Such assessment of the buyer's creditworthiness may be made based on information obtained from external sources, but this information can be updated semi-annually and must be is paid for. Meanwhile, the customer's solvency history is accumulated within the company itself, which can be used to assess the buyer' s current situation and update his creditworthiness assessment at the current moment.

The aim of the work is to apply machine learning methods to assess the creditworthiness of companies' businesses. First, create a tool for clustering buyers and refine expert credit assessment. Using this information, new buyers or buyers with changed solvency data would be automatically classified into different levels of creditworthiness. During the study, buyers were clustered based on their 2018 solvency history using the *K-mean method*. 38,99 percent of the cases were revised with an expert creditworthiness assessment. Then, machine learning was performed using Support vector with radial basis function and Random forest methods and their results were compared and was found that the latter can correctly assess creditworthiness up to 90 percent of new buyers with information only on their solvency. After examining the dependence of the accuracy of the classification results by the variables, six of them were selected, which were considered to have the most significant influence in the accuracy of the prediction.

Turinys

Lentelių sąrašas

Paveikslų sąrašas

Įvadas

Temos aktualumas. Kiekvieno naujo kliento patikimumą, įmonė tikrina naudojantis išorinių tiekėjų pagalba. Šie, turėdami informaciją apie kliento mokumo istoriją, balansus ir kitus duomenis suteikia jam kreditingumo įvertinimą. Už šios informacijos suteikimą, įmonės dažniausiai turi sumokėti. Tačiau bendradarbiaujant su klientu yra kaupiama informacija apie jam išrašytas sąskaitas ir jų apmokėjimus. Įmonė naudojantis šiais duomenimis gali susikurti savo kreditingumo vertinimo sistemą. Tokios sistemos didžiausi privalumai yra jog reikalinga informacija t.y. mokumo istorija yra prieinama be papildomų kaštų ir sistemos atnaujinimas vyksta pastoviai, kai, tuo tarpu, išoriniai tiekėjai gali atnaujintą informaciją pateikti kas pusmetį ar dar rečiau. Esami tyrimai patvirtino, jog mokumo informacijos įtraukimas, kreditingumo vertinimo modelio tikslumą pagerina, tačiau yra tik keli tyrimai, kurie vertina kliento kreditingumą remiantis tik mokumo informacija.

Tyrimo problema. Kliento mokumo istorijos įtraukimas į kreditingumo modelio sudarymą, jo tikslumą pakelia iki 13 proc. ir nors modeliai, sudaryti pagrinde iš mokumo informacijos pasiekia bendrą tikslumą iki 98 proc., jie klasifikuoja tik iki 50 proc. teisingai neigiamo kreditingumo vertinimo klientus. Taip pat, literatūroje aptariami yra modelių sudarymai, kai klientai jau yra priskirti kreditingumo grupėms, tačiau kaip šį priskyrimą atlikti, informacijos nėra. Dažniausiai yra pasikliaujama specialistų vertinimu. Atliekamas tyrimas užpildys spragą, kaip klientams priskirti kreditingumo vertinimą ir pagal tai sudaryti modelį, kuris sprendžiant iš mokumo informacijos atskirtų, kurie klientai yra didesnio rizikingumo.

Tyrimo objektas. Įmonių pardavimo sąskaitų apmokėjimo istorija ir verslo kreditingumas bei rizikingumas, suteikiant kreditą.

Tyrimo tikslas. Pritaikant mašininio mokymosi metodus ir naudojantis įmonių mokumo istoriją sudaryti verslo kreditingumo vertinimo modelį ir atlikti jo tyrimą.

Tyrimo uždaviniai:

1. Atlikti literatūros analizę ir išsiaiškinti įmonių kreditingumo vertinimo svarbą bei kokie metodai dažniausiai taikomi vykdant kreditingumo vertinimą;
2. Atlikti įmonių klasterizavimą, remiantis pardavimo sąskaitų faktūrų apmokėjimo istorija ir klasterizavimo patikslinimą, atsižvelgiant į ekspertinę informaciją;
3. Sudaryti klientų rizikingumo vertinimo modelį pritaikant mašininio mokymo metodus ir įvertinti modelio tikslumą;
4. Išanalizavus klasifikavimo modelio rezultatus, nustatyti, kurie kintamieji turi daugiausiai įtakos pirkėjo kreditingumo vertinimui.

1. Literatūros apžvalga

Šioje dalyje bus apžvelgiami įmonių, įsigyjančių prekes, piniginio ir prekybos kreditų tipai ir jų tarpusavio ryšys, analizuojami prekybos kredito teikiami privalumai iš tiekėjo, pirkėjo ir finansinių institucijų pusės; taip pat, kodėl toks kredito tipas yra rizikingas tiekėjui ir kodėl klientų kreditingumo vertinimas yra svarbiausia klientų skolų valdymo dalis. Papildomai bus apžvelgta, kokie metodai yra taikomi kintamųjų atrinkimui, modelio sudarymui ir į kokias problemas reikia atkreipti dėmesį sudarinėjant kreditingumo vertinimo modelius.

1.1. Kreditų tipai: piniginis kreditas ir prekybos kreditas

Įmonei perkant prekes, už jas galima atsiskaityti įvairiais būdais: mokėjimu prieš gaunant prekes ar iškarto po gavimo, gaunant kreditą iš finansinės įstaigos ar pačio prekių tiekėjo. Galima išskirti du pagrindinius trumpalaikių kreditų tipus:

- Piniginis kreditas (kitaip vadinamas bankiniu kreditu). Tai finansinės įstaigos, dažniausiai banko, teikiamas trumpalaikis kreditas su nustatyta palūkanų norma ir grąžinimo terminu, kuris yra skirtas atsiskaityti už konkrečias įsigytas prekes ar paslaugas.
- Prekybos kreditas. Tai tiekėjo suteikiamas kreditas, išreikštas per galimybę sumokėti už išrašytą sąskaitą faktūrą vėliau, t. y., suteikiamas tikslus mokėjimo atidėjimo terminas. Tokiu atveju į pardavimo kainą būna įskaičiuotos atidėjimo palūkanos ir dažnai nurodoma, jog sumokėjus anksčiau bus pritaikoma iš anksto sutarta nuolaida.

Įvairiuose šaltiniuose šie du kreditų tipai yra apibūdinami kaip papildantys arba pakeičiantys vienas kitą. Šį ryšį ir vienos kredito rūšies poveikį kitai galima apžvelgti / analizuoti įvairiais aspektais. Vieni tyrimai rodo, jog ryšys priklauso nuo įmonės apyvartinių lėšų lygio. Kai jis vidutinis, šios kreditų rūšys laikomi pakaitalais. Tačiau apyvartinėms lėšoms mažėjant, jie yra kombinuojami norint gauti geriausias kredito sąlygas ir didžiausią kredito sumą [1]. Įtakos šių kreditų tipų tarpusavio ryšiui turi ir valstybės teisinės sistemos lygis. Kuo jis didesnis, tuo abu kreditai yra išduodami dažniau, nes finansinės ir nefinansinės įstaigos patiria mažesnę riziką. Tačiau esant silpnesniam teisinės sistemos lygiui, būtent prekybos kreditas turi pirmenybę, nes įmonė, kaip tiekėjas, turi didesnę galią prieš pirkėją (gali nutraukti prekių tiekimą, padidinti kainą ir t. t.), lyginant su finansinėmis įstaigomis. Tačiau minėti kreditų tipai taip pat yra laikomi pakaitalais, t. y., yra susieti neigiama priklausomybę, bet jos stiprumas priklauso vėlgi nuo valstybės teisinės sistemos lygio [2] ir kitų faktorių, tokių kaip įmonės gyvavimo trukmės ir tipas. Pavyzdžiui, gamybinės įmonės turi didesnę tikimybę gauti piniginį kreditą, negu ne gamybinės. Įmonių kredito iš finansinių įstaigų gavimo tikimybė didėja su jau išduotais tokio tipo kreditais, o pačioje pradžioje didelę įtaką išdavimui turi prekybos kredito istorija. Dėl to galima teigti, jog jie papildo vienas kitą, bet didelę įtaką tam turi ir galimybės gauti abiejų tipų kreditus [3]. Nepaisant koks ryšys yra tarp šių dviejų kreditų rūšių ir kokios yra to priežastys, jie yra laikomi pačiais svarbiausiais išorinių lėšų šaltiniais mažose įmonėse [2].

Apibendrinant galima remtis [4] autorių padaryta išvada, jog įmonei neturint finansinių trikdžių gauti piniginį kreditą, piniginis ir prekybos kreditai yra pakaitalai. Tačiau kai tokios galimybės nėra, kreditai irgi laikomi pakaitalais, nes įmonės turi laviruoti ir pasirinkti tokią kombinaciją, kuri būtų balansas tarp gaunamos pinigų sumos ir jos kainos t.y., abejais atvejais, šie kreditų tipai yra pakaitai, bet to priežastis skiriasi.

1.2. Prekybos kreditas ir jo privalumai

Prekybos kredito suteikimo išraiška yra nustatymas atidėti apmokėjimą sutartam terminui. Terminai gali varijuoti nuo kelių iki keliasdešimt dienų. Taip pat yra įprasta susitarti dėl nuolaidų, kurios bus taikomos, jeigu klientas sumokės visą arba dalį sumos anksčiau nurodyto termino arba net ir prieš sąskaitos išrašymą. Nesumokėję sąskaitos iš anksto ir negavę nuolaidos, pirkėjai sumoka skirtumą, kuris gali būti laikomas palūkanomis [5]. Tačiau, lyginant su piniginio kredito palūkanomis, jos yra žymiai didesnės ir kai kurie autoriai teigia, jog jų norma siekia net iki 40 proc. Nepaisant tokio tipo kredito brangumo, skaičiuojama, jog už maždaug 40 proc. tarptautinių pardavimų yra atsiskaitoma būtent tokiu būdu, t. y., atidedant mokėjimą. Tokį didelį šio tipo kredito vartojimo lygį galima paaiškinti prekybos kredito suteikiamais privalumais, kuriuos galima suskirstyti į tris grupes: pardavimo rizikos sumažinimas, ryšių su pirkėju palaikymas ir įmonių augimas.

Pirmasis yra susijęs su pačios transakcijos rizika, kuri kyla iš kelių neapibrėžtumų ir jų sumažinimu. Daugiausia literatūroje yra teigiama, jog suteiktas prekybos kreditas pirkėjui signalizuoja apie gerą prekių kokybę [1][4][6]. O kokybei neatitikus reikalavimų, pirkėjas gali grąžinti prekes, kol dar nebus už jas mokėjęs. Taip papildomai sutaupoma ir lėšų pervedimo grandinėje, t. y., nereikia atlikti papildomų mokėjimo operacijų. Prekių kokybei atitikus standartus yra ir kitų pliusų, tokių kaip piniginių lėšų srautų planavimas abejoms sandėrio pusėms [4], t. y., tiekėjas žino, kada gaus lėšas ir taip gali planuotis savo lėšų judėjimą, o pirkėjas turi galimybę vietoje kelių apmokėjimų atlikti tik vieną. Šalia prekių kokybės garantijos, prekybos kreditas padeda užtikrinti ir prekių pristatymą [7]. Tai yra itin aktualu vykdant tarptautinius pardavimus, kai tiekėjas yra atsakingas prekių pristatymą. Taip pat suteiktas atidėjimo terminas bendruoju atveju mažina pardavimo kainą, nes nereikalingas papildomas tarpininkas, t. y., finansinė institucija, o to teigiamas aspektas padidėja, jei pirkėjas turi finansinių sunkumų [7].

Antroji privalumų grupė yra susijusi su santykių tarp tiekėjo ir pirkėjo užmezgimu ir vystymu. Tiekėjo suteikiamas prekybos kreditas yra vienas iš įrankių diferencijuoti klientus [1][8], t. y., nekeičiant pardavimo kainos suteikti ilgesnį atidėjimo periodą. Tokiu būdu nuolaida suteikiama per ilgesnį atidėjimo terminą, nes laikui bėgant pinigai nuvertėja. Tiekėjo įmonei suteikiamas kreditas taip pat turi pliusų, nes suteikdamas kreditą, tiekėjas turi didesnę įtaką pirkėjui ir tą gali panaudoti įvairiais būdais [2]. Tarkim, gali lengviau derėtis dėl didesnio parduodamų prekių kiekio mainais į ilgesnį atidėjimo terminą [1][3]. Taip tiekėjas pakelia savo pardavimų kiekį bei apyvartą. O jeigu sektoriuje yra ganėtinai mažas pasirinkimas tiekėjų, suteikiamas apmokėjimo atidėjimas yra būdas prisitraukti naujų klientų [6]. Nauda iš suteikiamo kredito gali būti ir ilgalaikė, t. y., tiekėjas, suteikdamas pirkėjui mokėjimo atidėjimą, gali padėti jam sunkiu periodu [3][8] ir taip sustiprinti bendradarbiavimą ateityje.

Galiausiai, prekybos kredito išdavimas yra susijęs su nauda, kurią tiesiogiai patiria įmonės. Pirkėjo įmonei suteiktas prekybos kreditas tampa teigiamu indikatoriumi dėl jo mokumo ir taip signalizuoja finansinėms įstaigoms, jog pirkėjas yra mokus ir jo šansai gauti piniginį kreditą didėja [1][3][6]. Tai padeda ir finansinėms įstaigoms, kurios tiesiogiai gali ir nedalyvauti pardavimų operacijoje, nes įmonių suteikiamų kreditų istorija padeda sumažinti informacijos asimetriją apie pirkėjo kreditingumą [2][3][4]. Taip suteikiamas prekybos kreditas prisideda prie įmonių augimo proceso [9]. Tiekėjas, turėdamas klientų mokumo istoriją, gali pagal ją koreguoti savo santykius su jais ir spręsti dėl tolimesnio bendravimo ir kredito suteikimo [1][8]. Pardavimų metu surinkta mokumo informacija taip pat yra pigesnė negu iš kitų šaltinių gaunama finansinė informacija apie pirkėją, nes

yra renkama kartu su kita pardavimo informacija ir už ją nereikia papildomai mokėti [2][8]. Taip pat ši renkama informacija yra einamoji ir tai suteikia jai privalumą, nes finansinės institucijos turi priėjimą tik prie metinių, geriausiu atveju, ketvirtinių duomenų ir todėl informaciją gali atnaujinti ganėtinai retai [8]. Tyrimai rodo, jog įmonės, kurios duoda mokėjimo atidėjimus, taip pat yra pelningesnės [3] ir tai yra teigiamas faktorius ir joms pačioms gauti išorinį finansavimą.

Apibendrinant galima teigti, jog prekybos kredito išdavimo priežasčių yra tikrai ne viena ir jis yra naudingas visoms šalims ir net išorinėms finansinėms institucijoms. Tačiau reikia suprasti, kada jis yra patrauklesnis už piniginį kreditą. Kai prekybos ir piniginis kreditas yra pakaitalai, įmonių pasirinkimą lemia tokie faktoriai:

- Tiekėjo suteikiamas prekybos kreditas turi didesnį limitą negu finansinės įstaigos suteiktas piniginis [1] ir šis limitas didėja tęsiantis bendradarbiavimui tarp įmonių [6].
- Šalyse, kuriose silpnai išsivysčiusios finansinės institucijos, įmonės labiau pasikliauja prekybos kreditu, nes jis gaunamas lengviau [1][2].
- Prekybos kredito išdavimo procesas yra greitesnis negu piniginio [2]. Tą galima paaiškinti tuo, jog prekybos kreditas yra tik dalis pardavimo proceso, o piniginis reikalauja sudaryti atskirą sutartį su finansine institucija ir todėl užtrunka ilgiau.
- Prekybos kredito procesas yra lengvesnis, t. y., tiekėjai yra labiau linkę suteikti kreditą negu finansinės įstaigos. Tai yra ypač naudinga, kai pirkėjas nebegali gauti piniginio kredito [4].
- Prekybos kreditą išduodančios įmonės turi daugiau galimybių apriboti kliento veiklą, kai jis nesilaiko savo įsipareigojimų lyginant su finansinėmis institucijomis. Tai ypač pasireiškia, jei klientas turi daug atsargų, nes tada yra galimybė atpirkti parduotas prekes ar nusipirkti kitų ir taip atlikti skolų sudengimus t. y., sumažinti tarpusavio skolą be mokėjimų [2][6][8].
- Prekybos kreditas taip pat padeda lyginti rinkoje esančių prekių finansų lėšas [6]. Įmonės, kurios yra stabilesnės ir turi didesnį kiekį vidinių lėšų, gali padėti toms, kurios neturi ir pati rinka.
- Suteikdamas prekybos kreditą, t. y., mokėjimo atidėjimą, tiekėjas atskiria pinigų ir atsargų perdavimo procesus. Dėl to abudu juos galima lengviau valdyti, nes jie nepriklausomi [8][7].

Įdomu paanalizuoti ir kokios įmonės yra labiau linkę imti prekybos kreditą. Buvo minėta, jog tarptautinės įmonės didelę dalį savo pardavimų finansuoja būtent prekybos kreditu, tačiau yra ir kitų priežasčių. Įmonės, kurios turi finansinių sunkumų, labiau linkę skolintis iš kitų įmonių negu finansinių institucijų [6], tačiau pastarosios ir sunkiau gauna piniginius kreditus. Tokia pati situacija yra ir su naujomis įmonėmis – joms finansinės įstaigos nėra linkusios suteikti kreditus dėl ribotos finansinės informacijos. Geri santykiai su tiekėjais yra pradedami kurti ir kaupiama teigiama mokumo istorija šį ryšį stiprina ir padeda įmonėms augti.

Šalia visų išvardintų teigiamų aspektų, prekybos kreditas turi būti išduodamas atsargiai. Tai būtina dėl to, jog net iki 25 proc. [5] trumpalaikio turto yra sudaryta iš klientų skolų. Jei ši dalis nebus valdoma efektyviai, skolos gali būti įšaldytos ilgam periodui, pinigų srautai tapti nenuspėjamais ir įmonės lėšos tapti nepastovios.

1.3. Klientų skolų valdymas

Kalbant apie įmonių, kaip pirkėjų, skolų valdymą, dažniausiai turima omenyje dvi sudedamąsias dalis, t. y., kredito išdavimo politiką ir kredito monitoringo politiką. Pirmoji apima visas sąlygas iki

prekybos kredito išdavimo, o antroji – po kredito išdavimo. Tačiau prieš šių procesų sudėliojimą, įmonės dar turi nuspręsti, kurie iš jų bus vykdomi įmonės viduje, o kurie atiduoti išoriniams pirkėjų skolų valdymo paslaugų tiekėjams.

Išorės pirkėjų skolų valdymo paslaugų tiekėjai gali pasiūlyti paslaugas visuose klientų skolų valdymo lygmenyse. Pirmiausia, galima iš įvairių institucijų nusipirkti informaciją apie kliento kreditingumo reitingą ir naudojantis juo spręsti ar išduoti prekybos kreditą, ar ne. Toliau, išdavus kreditą galima pasinaudoti faktoringo galimybe ir taip perkelti didžiąją atsiskaitymo rizikos dalį kitai įmonei, kuriai pirkėjas turės sumokėti. Yra specializuotų įmonių, iš kurių galima pirkti ir skolų monitoringo, ir išieškojimo paslaugas. Ir jeigu vis dėlto visas skolų valdymas yra atliekamas įmonės viduje, yra galimybė apdrausti skolas ir klientui nevykdant įsipareigojimų atgauti dalį investicijų.

Kiek įmonės yra pasiruošusios ir turi noro atlikti visą skolų valdymo procesą pačioje įmonėje, priklauso nuo kelių faktorių. Tos įmonės, kurios naudoja mokėjimo atidėjimo terminą kaip kainos diferencijavimo įrankį, t. y., nesant galimybei keisti kainos, manipuliuoja mokėjimo atidėjimo sąlygomis, yra labiau linkę vykdyti pirkėjų skolų valdymą savo viduje. Tą galima pagrįsti tuo, jog pačiai įmonei valdant kredito išdavimo procesą, jis gali būti lankstesnis ir lengviau personalizuojamas pagal kiekvieną klientą. Senesnės įmonės yra labiau linkę skolų valdymą atlikti viduje dėl turimos sukauptos informacijos apie klientus, tačiau atsižvelgus į parduodamas prekes, t. y., kai į prekybos kreditą yra žiūrima kaip į kokybės rodiklį, gaunama priešinga išvada – kad įmonės linkusios atiduoti pirkėjų skolų administravimą išorės paslaugų tiekėjams. O nepaisant to, kokia įmonė išduoda daug prekybos kreditų įvairiems klientams, didesnė tikimybė, jog skolų administravimas bus atiduotas išorei [10].

Įmonei susidėliojus pirkėjų skolų valdymo proceso strategiją, t. y., kiek bus atliekama viduje, kiek bus atiduota išorės įmonėms, galima prieiti prie pačio pirkėjų skolų valdymo ir čia galima išskirti tris dideles dalis:

- Prekybos kredito išdavimo politika, į kurią įeina kliento finansinės būsenos patikrinimas, kreditingumo įvertinimas remiantis turimais duomenimis apie jo finansinę ir nefinansinę būseną; nurodymas, kada kreditas gali būti išduodamas, o kada ne. Nusprendus suteikti kreditą, reikia įvertinti, koks kredito limitas gali būti suteikimas. Pasikeitus išoriniams faktoriams ar kliento situacijai, reikia peržiūrėti visas sąlygas ir remiantis naujausia informacija jas atitinkamai pakeisti.
- Sąskaitų išrašymo politika apima taisykles, kada turi būti išrašytos sąskaitos, nes nuo tos dienos ir yra skaičiuojamas sutartas apmokėjimo atidėjimas. Reikia užtikrinti, jog pardavimo sąskaita faktūra, pagal kurią išduodamas prekinis kreditas, pasiektų klientą.
- Prekybos kredito surinkimo politika, kuri yra susijusi su jau išduotų kreditų sekimu, identifikavimu padidinto rizikingumo klientus, kurie delsia apmokėti. Reikia nustatyti, kokie veiksmai ir po kiek dienų vėlavimo turi būti atlikti, pakartoti, kada skolas atiduoti išieškojimui.

Išduodant prekybos kreditus yra svarbu suprasti ir tinkamai nustatyti, kokie pastovūs ir kintami kaštai yra patiriami, kai pirkėjui yra suteikiamas kreditas, nes be akivaizdžios rizikos, kad pirkėjas neįvykdys savo įsipareigojimų, reikia žinoti ir kiek vidinių arba išorinių resursų reikės skirti nuo kredito išdavimo iki jo apmokėjimo arba nurašymo [11].

Pirkėjų skolų valdymo proceso pradžia yra kredito suteikimas, todėl ši vieta yra pati svarbiausia visame procese ir dėl to įvertinti kliento kreditingumą yra itin svarbus uždavinys.

1.4. Kliento kreditingumo vertinimas

Kliento kreditingumo vertinimas apibrėžiamas labai paprastai – tam tikros rizikos laipsnio klientui priskyrimas, remiantis apie jį turima informacija. Rizikos grupės gali būti kelių tipų, t. y., galima priskirti balą penkiabalėje, dešimtbalėje arba kokioje kitoje skalėje, gali būti suskirstyti į grupes pagal rizikingumą – mažas, vidutinis ir didelis, arba pagal turimą informaciją gali būti priimtas sprendimas ar suteikti kreditą, ar ne. Tačiau, nors šis uždavinys iš esmės yra paprastas, bandant jį realizuoti, jis tampa labai sudėtingas [12].

Pirmiausia galima išskirti kreditingumo vertinimus pagal tai, kada jie yra atliekami bendradarbiavimo su klientu eigoje. Dar prieš pradedant bendradarbiauti su įmone, turi būti patikrinamas jos patikimumas ir kreditingumas t.y., ar verta pradėti su šia įmone dirbti. Pradėjus bendradarbiavimą kreditingumo vertinimas yra atliekamas nustatymui kokia yra jo elgsena t.y. ar laikomasi įsipareigojimų, kaip dažnai atliekami mokėjimai ir ar yra jų vėlavimai ir t.t. Po šio įvertinimo eina klientų segmentavimas ir pagal tai sprendimas kaip dažnai reikia atlikti pakartotinius vertinimus, kas kiek laiko reikia peržiūrėti bendradarbiavimo politika ir pan. Ir paskutinis vertinimas yra nustatyti ar klientas nevykdo apgavysčių [13]. Daugiausia literatūroje yra kalbama apie patį pirmąjį kreditingumo vertinimą, nes jis yra naudojamas kai sprendžiama ar suteikti klientui kreditą ar ne. Tačiau nors tai yra pagrindinis vertinimo tikslas, sudarius gerą modelį jį galima taikyti ir pastoviam monitoringui, kurio metu galima pastebėti įtartiną kliento elgseną ir taip iš anksto pasiruošti veiksmų planą klientui pradėjus nesilaikyti savo įsipareigojimų.

Modelio sudarymas prasideda nuo duomenų surinkimo. Duomenys gali būti skirstomi į finansinius, nefinansinius ir išorinius. Finansiniai faktoriai apima įmonės veiklos rezultatus, kurie gaunami iš balansų ar įvairių ataskaitų. Nefinansiniais rodikliais gali būti laikoma mokumo istorija, informacija apie darbuotojus, vadovybę, gyvavimo trukmę, klientų atsiliepimus ir pan. Išoriniai faktoriai apibūna ekonominę padėtį pasaulyje, šalyje, sektoriuje ir t.t. Yra atlikta tyrimų, jog nefinansinių duomenų pridėjimas į modelį gali pagerinti jo tikslumą. [14] autoriai argumentuoja, kad jų pagalba galima pagerinti modelio tikslumą iki 13 proc. Taip pat, nefinansinius duomenis galima dažnai atnaujinti be papildomų išlaidų taip gerinant modelio tikslumą. Dėl šios priežasties jie yra labai vertingi.

Turint tiek daug informacijos yra labai svarbu atrinkti, kuri yra svarbi ir naudinga sudarant modelį. Tam yra atliekama kintamųjų atranka ir literatūroje dažniausiai yra išskiriami trys konkrečios priežastys kodėl tai daroma:

- Modelio paprastumas. Sudarinėjant modelius, vienas iš tikslų yra jog jis būtų suprantamas ir logiškai paaiškinamas. Kai į modelį bus įtraukti tik tie kintamieji, kurie logiškai bus surišti su modelio rezultatais, juos bus lengviau paaiškinti ir suprasti.
- Modelio sudarymo laikas. Kadangi modeliams naudojami sudėtingi modeliai, kuriuose atliekami įvairūs skaičiavimai, kuo mažiau kintamųjų yra paduodama į modelį, tuo jis greičiau bus sukuriamas, testuojamas ir patikrinimas. Tai ypač aktualu kai kalbama apie didelio kiekio duomenų aprojimą.
- Modelio patikimumas. Kuo daugiau į modelį įeina kintamųjų, tuo jame gali atsirasti daugiau triukšmo dėl jų tarpusavio priklausomybių. Taip pat, modelio tikslumas gali būti mažesnis, nes mažai informacijos nešantys kintamieji gali paslėpti tikrais svarbią informaciją. Arba

modelis gali persimokyti ir prisitaikyti prie labai unikalių variantų ir todėl, nors gerai pasirodys su tęstine imtimi, paėmus kitus duomenis jos tikslumas gali būti žymiai prastesnis.
- Modelio brangumas. Kadangi didžioji dalis kintamųjų, kurie yra naudojami kreditingumo įvertinimo modeliuose yra finansiniai, juos įmonės turi pirkti iš kitų tiekėjų ir reguliariai šią informaciją atnaujinti. Dėl šios priežasties, sumažinus kintamųjų skaičių, įmonės sutaupo kaštų modelio sudarymui.

Literatūroje kintamųjų atrinkimo metodai yra suskirstyti į tris grupes:

- Filtravimo metodai. Šio grupės metodai yra taikomi atskirai nuo pačio modelio sudarymo t.y. pirmiausia atliekamas kintamųjų atrinkimas ir tik tada yra kuriamas modelis. Taip pat, jų esmė yra, kad visi kintamieji yra įvertinami naudojant pasirinktą rodiklį, išrikiuojami ir paimamas nustatytas kiekis kintamųjų. Tokie metodai, bendruoju atveju, duoda prastesnius rezultatus, bet yra labai greitai atliekami, o tai yra pliusas dirbant su dideliu kiekiu kintamųjų ir įrašų. Į tokią grupę įeina daug metodų. [13] autorius siūlo naudoti kintamųjų klasterizavimą ir rezultatus lygina naudojant F_1 rodiklį. [15] autoriai naudoja kelis metodus, tokius kaip *Pagrindinių komponenčių analizė*, *Genetinius algoritmus* ir kt. iš kurių geriausias parenkamas *Pagrindinių komponenčių metodas*. Būtent šį metodą rekomenduoja ir [16] autorius, kuris parodo, jog naudojant jį galima sumažinti kintamųjų kiekį apie 60 proc. O kadangi pati metodo idėja yra surasti svarbiausius komponentus, metodas gali būti pritaikytas bet kokiam duomenų rinkiniui.
- Auginimo metodai. Į šiuos metodus patenka tie, kuriuose geriausių kintamųjų grupė yra auginama nuo nulio arba iš jos yra vis atimama po vieną kintamąjį. Abejais atvejais, kiekviename žingsnyje yra išrenkamas geriausiai kriterijus atitinkantis kintamasis. Į šią kategoriją papuola [17] aprašytas metodas, kuris yra pagrįstas *Genetiniu algoritmu* su dviem tikslo funkcijomis – kintamųjų kiekio minimizavimu ir tikėtinu didžiausio pelno maksimizavimu. Pateikti rezultatai rodo, jog toks kintamųjų parinkimas rezultatais prilygsta gerumu ir kitiems nagrinėtiems, o įgyvendinamas yra per trumpesnį laiko tarpą.
- Jungtiniai metodai. Šie atrankos metodai yra modelio kūrimo dalis t.y. modelio sudaryme ir klasifikatoriaus apmokyme yra taip pat atsižvelgiama į kintamuosius ir jie parenkami tokie, kurie neša daugiausiai informacijos ir geriausiai tinka modeliui. Šių atrankos metodų didžiausia problema ir yra, jog jie negali būti taikomi atskirai nuo pačio klasifikatoriaus.

Tačiau taip pat yra svarbu ir sutvarkyti pačius duomenys. Šį darbą galima išskirti į tris dalis: valymą, transformavimą ir normalizavimą. Pirmasis yra reikalingas, kad į modelį būtų įtraukta tik teisinga informacija. Šioje dalyje yra labai svarbu suprasti ir turėti gilesnes žinias apie pačius duomenis arba bendrauti su specialistu, kuris gali duoti šią informaciją. Transformavimo dalis yra svarbi, nes skirtingi modeliai turi skirtingus reikalavimus t.y. jiems duomenis reikia pritaikyti. Dėl to, taikant įvairius metodus gali reikėti duomenis transformuoti kelis kartus ir turėti kelis atskirus duomenų rinkinius. Normalizavimas yra būtinas žingsnis tada, kai kintamųjų reikšmės yra labai plataus diapazono. Prataikius normalizavimą modelis bus stabilesnis ir rezultatus bus lengviau interpretuoti. Po duomenų atrinkimo ir sutvarkymo seka modelio sudarymas.

Literatūroje kreditingumo vertinimas yra įvardinamas kaip klasifikavimo uždavinys t.y. tikslas yra pagal turimus duomenis apie klientą priskirti jį į vieną iš klasių ir taip nustatyti ar jis grąžins suteiktą kreditą ar ne. Kadangi kreditų išdavimas yra plačiai taikomas dalykas tiek finansinėse tiek nefinansinėse įstaigoje, šis uždavinys yra aktualus, nepaisant koks yra išduodamo kredito tipas.

Sudarant modelius, reikia atsižvelgti į du dalykus – modelio tikslumą ir jo kainą. Pastaroji gali būti sprendžiama pagal duomenų gavimo kaštus, skaičiavimo laiką ir reikalingus resursus. Dėl to ir yra tiek daug skirtingų modelių ir literatūroje galima rasti įvairius priėjimus prie modelio kūrimo.

Ankščiau buvo taikomi atskiri klasifikatoriai, tokie kaip *Logistinė regresija*, *Dispersinė analizė*. Paskui buvo pradėti taikyti mašininio mokymo metodai – *Dirbtiniai neuroniniai tinklai, Sprendimų medžiai, Atraminių vektorių, Bajeso* ir kt. metodai. Tačiau šiuo metu literatūroje galima matyti krypimą į bendrą metodų naudojimą, tarkim sudarius pavienius klasifikatorius visų jų rezultatai yra sujungiami suteikiant tam tikrus svorius [18]. [13] autoriai sujungia kelias variacijas *Neuroninių tinklų, Naivųjį Bajeso* ir *Sprendimų medžių* metodus. Atsiranda ir tokių autorių kaip [15], kurie siūlo visai kitą priėjimą prie klasifikavimo uždavinio ir naudoja entropijos matą nustatyti ar naujas įrašas yra panašus į esamus jau ar ne ir taip vykdo klasifikavimą. Minėti autoriai naudoja Paprastąją ir Bendrąją Shannon entropijas ir gautais rezultatais nenusileidžia tipiniams klasifikatoriams. Tačiau yra ir tokių, kaip [19] kurie naudoja metodus po vieną, bet kiekvienam duomenų rinkiniui siūlo išbandyti visus su maksimaliai tinkamai parinktais nustatymais ir tada išsirinkti geriausiai tinkamą.

Literatūroje daug atvejų, kai naudojami *Atraminių vektorių metodai* ir jų variacijos. Pats metodas, nekombinuojant jo su kitais turi daug skirtingų pritaikymo būdų – gali būti naudojamos tiesinės ir netiesinė branduolio funkcijos [20], paraštės būtų minkštos arba ne. [21] autoriai teigia, jog kai duomenų kiekis didelis, geriausiai veikia metodas su tiesine branduolio funkcija. O [22] autorius naudoja *Atraminių vektorių metodą* sujungdamas jį su kintamųjų atranka ir parodo, jog sumažinus kintamųjų kiekį su minėtų klasifikatoriumi gaunami rezultatai yra tenkinamo tikslumo, bet su mažomis laiko ir papildomų išlaidų sąnaudomis. Tokį patį priėjimą siūlo ir [23] [24] [25] autoriai. O [26] autorius siūlo šį metodą naudoti kartu su klasterizavimu t.y. pirmiau klientus klasterizuoti ir kiekvieno jo viduje taikyti atraminių vektorių metodą. [27] autoriai siūlo jį naudoti kaip pagrindą klasifikatoriams ir tada naudoti *Gilaus suvokimo tinklo* (angl. *Deep belief network*) pagrindo ansamblio metodą su pakartotinu atrinkimu. [28] autoriai naudoja klasifikatorių sujungę su duomenų tvarkymo metodais sudaro 16 sluoksnių *Genetinį Kaskadinį* klasifikatorių ansamblį.

Viena iš priežasčių kodėl *Atraminių vektorių metodas* dažnai naudojamas tyrimuose yra tai, jog jis neturi kitų metodų silpnybių, tokių kaip lokalių sprendinių radimo, persimokymo ar ilgo skaičiavimo laiko. Pastarasis yra ypač svarbus modelio sudarymui. [29] autoriai palygina *Atraminių vektorių metodą su Gradientiniu metodu, Ekstremaliomis mokymosi mašinomis* (angl. *Extreme Learning Machine*) ir viena iš jų variacijų - *Inkrementinėmis mokymosi mašinomis* (angl. *Incremental Extreme Learning Machine*). Rezultatai parodė, jog *Atraminių vektorių metodas* pasirodė prasčiau, negu minėtos, tačiau autorius pabrėžia jog jo gerumą galima pagerinti derinant parametrus. Tačiau yra ir tokių tyrimų, kuriuose šis metodas savo gerumu pralenkė ir *Logistinę regresiją*, bei *Neuroninius tinklus* [30]. Beje šio metodo variacijos irgi yra plačiai minimos literatūroje. Pvz. [31] autorius siūlo naudoti *Kaskadinius Koreliacinius Neuroninius tinklus*.

Nors ir iš pirmo požiūrio yra pereinama prie mašininio mokymo metodų naudojimo, toliau naudojami ir seniau žinomi metodai, tokie kaip *Logistinė regresija* [32] ir jos variacijos. [33] autoriai siūlo naudoti grupinę *Lasso Logistinę Regresiją*, kuri pagal atliktą tyrimus geriau pasirodė negu *Logistinės regresijos* su įprastais kintamųjų parinkimo metodais. Taip pat yra ir tokių mokslininkų kaip [34], kurie siūlo kreditingumą vertinti dvejais etapais. Pirmuoju įmones suskirstyti į dvi klases pagal tikimybę jog jie atiduos paskolą ir tada tiems, kurie neatiduos (kuriuos klasifikatorius priskyrė prie rizikingų) nustatyti kokia su bus prarandama tokiu atveju. Abejose stadijose naudojama dvidešimt

įvairių klasifikatorių ir jų junginių. Tas daroma, norint sumažinti kiekvieno iš metodo trūkumus ir sujungiant visus rezultatus gauti kiek įmanoma tiksliausią rezultatą.

Taip pat galima pažiūrėti į skirtingų klasifikatorių naudojimą iš kitos pusės – kaip nustatyti kurie klasifikatoriai turi būti naudojami, jog jų bendras rezultatas būtų geriausias. Tokią problemą tiria [35] autorius ir siūlo atlinkti klasifikatorių atrandą panašiai kaip atliekama ir kintamųjų atranka t.y. pridedami skirtingi klasifikatoriai tol, kol modelio tikslumas pasiekiamas maksimalus.

Ir nors klasifikatoriaus parinkimas yra labai svarbi tyrimų dalis, taip pat reikia atsižvelgti ir į kitas klasifikavimo uždavinio keblias vietas. Sprendžiant šį uždavinį, viena iš problemų, su kuriomis yra susiduriama yra vadinama „šalta pradžia“ t.y. norint atlikti klasifikavimą reikia turėti jau suklasifikuotus duomenis klasifikatoriaus apmokymui. Tačiau klientų, kurie yra klasifikuojami kaip „blogieji“ t.y. tie, kurie nesilaikė savo įsipareigojimų yra mažai. Jeigu tokių klientų iš vis nėra, gaunama minėta problema, o atsiradus tokiems duomenims, susiduriama su kita problema – klasių disbalansu. Pirmąją t.y. „šalto starto“ pradžią yra siūloma spręsti įvairiais būdais ir vienas iš jų pateiktas [15] autorių yra po kiekvieno naujo elemento pridėjimo tikrinti visų duomenų entropijos pasikeitimą t.y. jeigu naujas elementas stipriai skiriasi nuo jau esamų, priimti jog jis yra nepatikimas. Tokiu būdu galima atlikti klasifikavimą turint tik vienos klasės duomenis. Antrąją problemą – klasių disbalansą galima spręsti irgi keliais būdais. Populiariausi yra „didžiosios“ klasės sumažinimas iki reikiamo dydžio, tačiau dėl to gali atsirasti informacijos praradimas arba „mažosios“ klasės dirbtinis padidinimas atsitiktinai imant įrašus kelis kartus arba taikant funkcijas, kurios sugeneruoja panašius duomenis, bet šio būdo problema, jog metodas gali persimokyti t.y. prisitaikyti prie pasikartojančių įrašų. Jog būtų sumažinami abiejų priėjimų trūkumai, modeliuose jie taikomi kaip kombinacija. Kitas būdas išspręsti šią problemą yra naudoti modelyje klasių svorius t.y. pagal klasių dydžius priskirti svorį ir taip pritaikyti modelį nepersimokymui [36]. Šalia šių problemų ir kitos – duomenų stoka, kuri atsiranda dėl vidinių duomenų stokos su kuriais galima būtų apmokyti, testuoti ir validuoti modelį. O norint gauti išorinius duomenis, už juos reikia mokėti. Tai tiesiogiai susisieja su tuo, kodėl yra verta vykdyti kintamųjų atranką prieš sudarant modelį. Kadangi modelis yra sudaromas pagal jau turimus duomenis, jis labai sunkiai atpažįsta naujas tendencijas ir todėl reikia laiko kol prie jų prisitaiko. Taip pat šį procesą t.y. klientų elgsenos pokyti yra sunku pamatyti iš modelio rezultatų. Pasirinkus taikyti mašininio mokymosi metodus, reiki atkreipti dėmesį, jog finansinės įstaigos išduodančios kreditus kitoms įmonėms, yra pagal įstatymą įpareigotos tiksliai įvardinti vertinimo kriterijus ir todėl naudojami modeliai turi būti aiškiai interpretuojami. Dėl šios priežasties populiarėja taisyklių išgavimo metodai, kurie naudojami šalia modelių [37], tačiau vietoje naudojimo vieno metodo nustatyti kreditingumo vertinimą ir kito naudojimo taisyklėms išgauti galima naudoti sprendimų medžius [38][39], kurie nenusileidžia savo tikslumu ir kitiems siūlomiems metodams.

Kadangi daugiausia literatūroje yra kalbama apie klasifikavimo uždavinį, modelio gerumas yra skaičiuojamas naudojant tikslumo, specifiškumo ir jautrumo rodiklius. Pastarieji du yra naudojami dėl to, jog vien tik tikslumas neparodo klasifikatoriaus gerumo, reikia atsižvelgti į kiekvieną iš klasių taip pat. Esant daugiau negu vienai klasei, ypač naudinga suformuoti ir sumaišymo matricą, kuri atspindi koks procentas kiekvienos klasės buvo suklasifikuotas teisingai.

Visi naudojami metodai turi ganėtinai aukštus tikslumus t.y. nuo 80 proc. ir aukštesnius. Tačiau atsižvelgus į tai, jog kreditų išdavimas yra finansinių įstaigų pagrindinė veikla ir jų poreikis yra labai didelis, netgi 1 proc. tikslumo pakėlimas įmonei gali sutaupyti labai dideles sumas. Lygiai tokia pati

logika gali būti pritaikyti ir ne finansinėms įstaigoms, nes klientams prekinis kreditas yra suteikiamas ir skolos atsiradusios iš to turi generuoti kuo mažiau praradimų.

Apibendrinant galima sakyti, jog lengvai suprantamas ir pritaikomas, tikslus ir korektiškas kreditingumo vertinimas yra vienas svarbiausių įrankių įmonėms, kurių pagrindinė veikla yra kreditų išdavimas. Tačiau šis įrankis yra itin svarbus ir įmonėms, kurio kreditą išduoda kitokiu pavidalu t.y. prekiniu kreditu arba paprastai sakant – apmokėjimo atidėjimu.

1.5. Tyrimo poreikis

Atliekant projektą „Prekybinių procesų rizikos valdymas pasitelkiant didžiuosius duomenis ir dirbtinį intelektą“ vienas iš tikslų yra sujungti kiek įmanoma daugiau informacijos apie klientus ir nustatyti jų kreditingumo vertinimą. Į duomenų šaltinius yra įtraukta ir pirkėjo mokumo informacija. Šių duomenų didžiausias pliusas yra, jog jie yra generuojami kiekvieno pardavimo metu ir atspinti einamąją kliento būseną t.y. naudojant juos galima pamatyti kliento elgsenos pasikeitimus iš karto. To reikia, nes perkant informaciją iš kitų šaltinių, ji atspindi kliento finansinę būklę praeitame periode ir atnaujinama yra sąlyginai retai. Dėl šios priežasties svarbu suformuoti kliento kreditingumo vertinimą tik iš mokumo informacijos ir toliau jį sujunti su vertinimais iš kitų duomenų. Projekto vykdymo metu, nėra sukurto ir pritaikymo kreditingumo vertinimo modelio paremto turimais vidiniams duomenimis.. Iš to ir kilo tyrimo poreikis t.y. sukurti kreditingumo vertinimo modelį remiantis klientų mokumo istorija.

Atliekant literatūros analizę buvo rastas straipsnis [40], kurios autoriai naudojo vidinius nefinansinius duomenis sudarant klasifikatorių kreditingumo įvertinimui nustatyti. Jame buvo paminėta, jog buvo atliekamas klientų įvertinimas ir taip jiems priskiriamos klasės ir tada apmokamas klasifikatorius naudojant šias klases, tačiau nieko detalaus apie patį įvertinimą nebuvo užsiminta. Todėl atliekamas tyrimas bus atliekamas panašiai kaip ir [40] autorių, tačiau daugiau dėmesio bus skiriama ir klientų pradiniam segmentavimui į grupes pagal kurias vėliau bus formuojamas klasifikatorius. Papildomai, po literatūros apžvalgos buvo nuspręsta pritaikysi Atraminių Vektorių metodo atmainą ir palyginti šio metodo tikslumą su Atsitiktinių medžių metodu. Pastarasis metodas taip pat yra naudojamas ir išsiaiškinti kokia informacija turi daugiausiai įtakos pirkėjų vertinime.

1.6. Apibendrinimas

Šioje dalyje buvo apžvelgti kreditų tipai, išanalizuoti jų tarpusavio ryšio tipai. Remiantis literatūra, apibendrinti prekinio kredito teikiami privalumai. Taip pat, apžvelgti taikomi metodai nustatyti kliento kreditingumo lygį ir aptartą šio vertinimo reikalingumas. Padaryta išvada, jog turint duomenis apie klientų mokumą galiam apgerinti kreditingumo vertinimo gerumą, bet taip pat galima šiuos duomenis naudoti ir atskiro vertinimo sudarymui. Toks vertinimas ir bus sudarinėjamas, pirmiausia atlikus klientų segmentavimą į grupes ir pagal jas apmokant klasifikatorius.

2. Tyrimo metodai

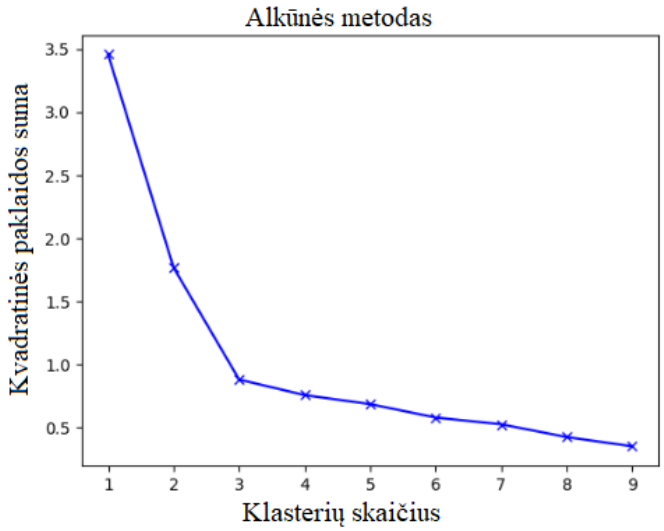
Šioje dalyje yra apžvelgiami metodai yra taikomi tyrime ir kaip jie vertinami. Aprašymai pateikiama remiantis [41] ir [42] šaltiniais.

2.1. Klasterizavimas

Darbe yra naudojamas *K-vidurkių* klasterizavimo metodas. Šis metodas priskiriamas skaidymo algoritmams. Tarkime, jog duomenų rinkinį D sudaro n objektų. K-vidurkių metodo idėja yra kiekvieną iš n objektų priskirti vienam iš k klasterių t.y. C_1, C_2,..., C_k taip, kad $C_i \subset D$ ir $C_i \cap C_j = \emptyset$, kai $1 \leq i, j \leq k$ ir $i \neq j$. Tikslas yra suskaidyti duomenų rinkinį taip, kad objektai klasterių viduje būtų panašūs vienas į kitą ir nepanašūs į objektus kituose klasteriuose. Norint tai įgyvendinti reikia apsibrėžti kiekvieno klasterio C_i $(1 \leq i \leq k)$ centrą kaip c_i ir tada objekto atstumas nuo šio centro gali būti žymimas $dist(p, c_i)$, kur $p \in C_i$ ir $dist(x, y)$ yra Euklidinis atstumas tarp x ir y objektų. Tada klasterizavimo rezultato gerumas gali būti matuojamas dydžiu E, kuris apibrėžiamas (2.1.1) formule:

$$E = \sum_{i=1}^{k} \sum_{p \epsilon C_i} dist(p, c_i)^2 \qquad (2.1.1)$$

čia: E – yra visų duomenų rinkinio objektų kvadratinės paklaidos suma, p – taškas apibūdinantis objektą, c_i – klasterio C_i centrą atitinkantis taškas. Nagrinėjant šį dydį galima įvertinti ir koks yra optimalus klasterių skaičius. Tyrime naudojamas metodas vadinamas Alkūnės metodu ir taikant jį reikia atlikti klasterizavimą vis didinant klasterių skaičių ir apskaičiuoti objektų kvadratinės paklaidos sumą kiekvieną kartą. Tada šio dydžio priklausomybę nuo klasterių yra atvaizduojama grafiku (pvz. pateiktas 2.1.1 pav.) ir jame yra randamas klasterių skaičius, nuo kurio toliau didinant jį, kvadratinės paklaidos suma reikšmingai nebemažėja.



2.1.1 pav. Alkūnės metodo taikymo metu gauto grafiko pavyzdys

K-vidurkių klasterizavimo metodo didžiausi privalumai yra, jog jis lengvai suprantamas bei įgyvendinamas ir todėl gali būti pritaikomas ir dideliems duomenų rinkiniams. Didžiausi šio metodo trūkumai yra, jog vartotojas turi nurodyti klasterių skaičių ir jis yra priklausomas nuo duomenų skalių t.y. kadangi atstumui skaičiuoti yra naudojamas Euklidinis atstumas, esant skirtingų skalių kintamiesiems, vieni gali turėti daugiau įtakos klasterių sudarymui negu kiti. Dėl to tyrime klasterizavimas bus atliekamas etapais:

- Duomenų rinkinių paruošimas. Kadangi duomenų rinkinyje yra įvairių skalių duomenų, klasterizavimui naudojamos bus ir duomenų modifikacijos – pradiniai duomenys bus normalizuojami; pradiniuose duomenyse bus panaikinamos išskirtys ir tada šis duomenų rinkinys normalizuojamas. Tokiu būdu klasterizavime bus naudojami keturi duomenų rinkiniai – originalus ir trys jo modifikacijos;
- Klasterių skaičiaus nustatymas. Kiekvienam duomenų rinkiniui, naudojant Alkūnės metodą, nustatomas optimalus klasterių skaičius;
- Klasterių interpretacija. Iš visų sudarytų klasterių, pagal jų centrus yra išrenkamas tas, kuris duomenis geriausiai ir logiškiausiai atskiria į grupes.

Po geriausio rezultato parinkimo, pirkėjų vertinimas buvo pakoreguotas atsižvelgiant į informaciją, kuri yra žinoma specialistui dirbančiam su pirkėjais ir tolimesni klasifikavimo metodai pritaikyti pakoreguotam rezultatui.

2.2. Klasifikavimas

Klasifikavimui atlikta yra naudojami du metodai – *Atraminių vektorių* ir *Atsitiktinio miško* metodai. Siekiant įvertinti ir pagerinti jų gautus rezultatus duomenys yra atskiriami į mokymo ir testavimo imtis naudojant kryžminį patikrinimą (angl. *Cross Validation*) ir tada klasifikatoriaus sudarymą galima suskaidyti į etapus:

1) Duomenų imtis atsitiktinai suskaidoma į *k* dalių;
2) Paimama viena dalis duomenų ir ji laikoma testavimo imtimi, o visos kitos – mokymo;
3) Sudaromas klasifikatorius su mokymo imtimi;
4) Atliekamas prognozavimas testavimo imčiai;
5) Veiksmai 2-4 yra kartojami kiekvienai iš *k* duomenų dalių.

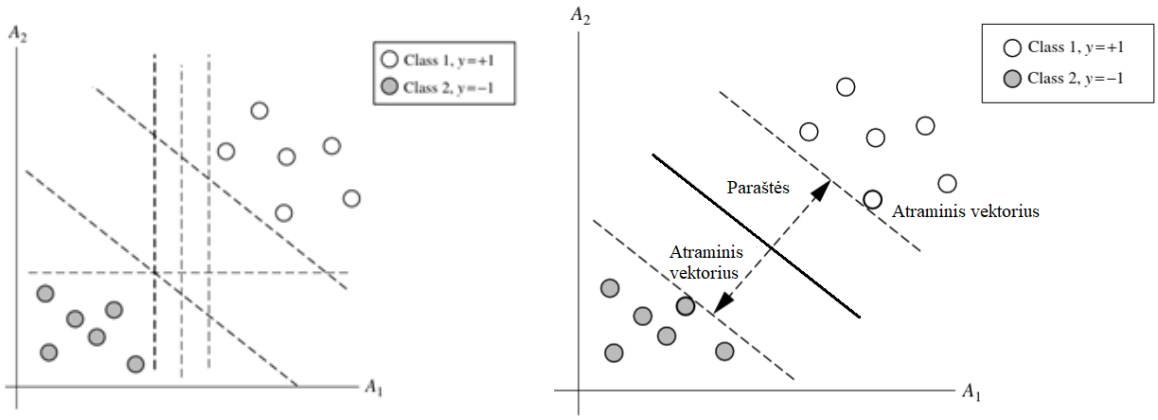
Tokio tipo duomenų padalinimas leidžia sudaryti kelis klasifikatorius ir taip sumažina persimokymo tikimybę, taip pat klasifikatoriaus rezultatus leidžia interpretuoti naudojant visus duomenis.

2.2.1. Atraminių vektorių metodas (SVM)

Atraminių vektorių metodo pagrindinis tikslas yra objektus į klases atskirti hiperplokštuma, kurios nustatymui yra naudojami atraminiai vektoriai. Tarkim, jog duomenų rinkinį *D* sudaro *n* objektų, kuriuos galima įvardinti taip: $\{(\boldsymbol{x_1}, y_1),\dots,(\boldsymbol{x_n}, y_n)\}$, čia $\boldsymbol{x_i}$ yra d-matis vektorius sudarytas iš kintamųjų apibūdinančių objektą, o y_i yra kintamasis, kuris nurodo stebėjimo $\boldsymbol{x_i}$ klasę. Dėl paprastumo sakykite, jog stebėjimas gali turėti dvi skirtingas klases ir tada y_i galima apibrėžti taip:

$$y_i = \begin{cases} +1, \ kai\ \boldsymbol{x_i}\ priklauso\ pirmai\ klasei; \\ -1, \ kai\ \boldsymbol{x_i}\ priklauso\ antrai\ klasei. \end{cases} \quad (2.2.1.1)$$

Perkėlus $\boldsymbol{x_i}$ į dvimatę erdvę, galima lengvai įsivaizduoti kintamųjų atskyrimus į klases ir tokių atskyrimų galima surasti ne vieną (žr. 2.2.1.1 pav. kairėje). Jog būtų galima apibrėžti vienintelį klasių atskyrimą, yra įvedamos paraštės (angl. *margins*) ir atraminiai vektoriai (angl. *support vectors*). Paraštės yra atstumas nuo hiperplokštuma iki skirtingų klasių objektų, o atraminai vektoriai yra ant paraštės krašto esantys objektai (žr. 2.2.1.1 pav. dešinėje).



2.2.1.1 pav. Objektų klasių atskyrimų pavyzdžiai (kairėje) ir objektų atskyrimas įvedus paraštes ir atraminių vektorių apibrėžimus (dešinėje)

Klases atskirianti hiperplokštuma užrašoma (2.2.1.2) formule:

$$\boldsymbol{W} * \boldsymbol{X} + b = 0 \tag{2.2.1.2}$$

čia: *W* yra d-matis svorio vektorius $W = \{w_1, w_2, \ldots, w_d\}$, *d* yra objektą apibūdinančių kintamųjų kiekis, *b* yra skaliaras. Metodo tikslas yra atskirti objektus taip, kad paraštės dydis būtų didžiausias ir taip yra, kai paraštės dydis yra $\frac{2}{\|W\|}$, kur $\|W\|$ yra Euklidinė *W* norma. Pertvarkius (2.2.1.2) lygybę SVM metodą galima aprašyti kaip sprendimo funkciją su (2.2.1.3) formule. Į ją įstačius objektą apibūdinančių kintamųjų reikšmes, pagal gauto rezultato ženklą, galima spręsti, kurioje hiperplokštumos pusėje yra objektas.

$$D(\boldsymbol{X^T}) = \sum_{i \in S} \alpha_i y_i \boldsymbol{X_i}\, \boldsymbol{X^T} + b \tag{2.2.1.3}$$

čia: *D* yra atskyrimo taisyklė, *S* yra atraminių vektorių skaičius, α_i ir b yra koeficientai, kurie yra parenkami optimizuojant skaičiavimus metodo viduje, y_i yra X_i atraminio vektoriaus klasė, X^T yra objekto kintamųjų reikšmių vektorius, kuriam reikia nustatyti klasę.

Dažniausiai praktikoje susiduriama su duomenimis, kurių neįmanoma atskirti į klases be persidengimų, todėl (2.2.1.3) formulė yra pakoreguojama ir vietoje X_i ir X^T sandaugos yra naudojama kita funkcija, kuri vadinama branduolio (angl. *kernel*) funkcija, kuri leidžia identifikuoti netiesinį objektų pasiskirstymą. Plačiausiai yra naudojamos trys branduolio funkcijos: polinominė, Gauso radialinė ir sigmoidinė. Tyrime yra naudojama Gauso radialinė branduolio funkcija (angl. *Gaussian radial basis function kernel*), kuri yra aprašoma (2.2.1.4) formule:

$$K(X_i, X_j) = e^{-\gamma \|X_i - X_j\|^2} \tag{2.2.1.4}$$

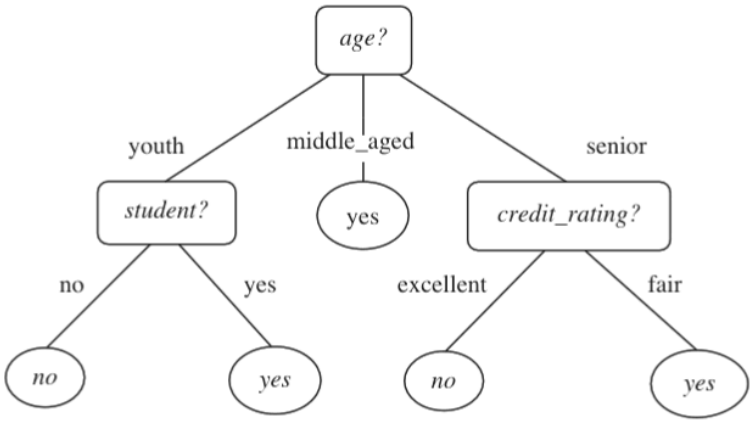
Papildomai metode dar yra įvedama ir bauda, kuri leidžia objektams persidengti t.y. papulti į paraštes ar net į kitos klasės erdvę – ši korekcija vadinama minkštos paraštės. Naudojant branduolio funkciją ir minkštas paraštes metode, leidžia geriau pritaikyti metodą tik kiekvieną kartą reikia abiejų parametrų, – *C* (angl. *cost*) ir γ (angl. *gamma*), reikšmes pritaikyti duomenims.

2.2.2. Atsitiktinio miško metodas (RF)

Atsitiktinio miško metodo esmė yra sudaryti daug Sprendimų medžių ir juos sujungiant gauti vieną bendrą rezultatą. Dėl to, pirmiausia apibrėšime Sprendimų medžio sudarymą. Šis klasifikavimo metodas taip vadinamas dėl to, jog jo vizualizuojamas sprendinys primena medžio struktūrą, kurioje galima išskirti tokias dalis:

- Šaknis (angl. *root*) yra medžio grafo viršūnė;
- Vidinis mazgas (angl. *internal node*), kuriame yra aprašoma tam tikrą sąlygą;
- Šaka (angl. *branch*), kuri atvaizduoja sąlygą tenkinančius duomenis;
- Lapas (angl. *leaf node*), kuris atvaizduoja objektam priskirtą klasę;

Sprendimų medžio pavyzdys pateiktas 2.2.2.1 pav.



2.2.2.1 pav. Sprendimų medžio vizualizuojamo sprendinio pavyzdys

Sprendimų medžio sudarymo svarbiausia dalis yra atrinkti kuris kintamasis turi būti naudojamas atskyrimui ir su kokia sąlyga. Tam yra naudojami rodikliai, nurodantys kiek informacijos bus įgytą atlikus atskyrimą. Dažniausiai sutinkami informacijos vertinimo rodikliai yra informacijos gavimas (angl. *information gain*), informacijos gavimo santykis (angl. *gain ration*)
Abiem rodikliams yra naudojamas entropijos matas (angl. *entropy*), kuri skaičiuojama (2.2.2.1) formule:

$$E(D) = -\sum_{i=1}^{m} p_i log_2(p_i) \qquad (2.2.2.1)$$

čia: *E(D)* – entropija duomenų rinkiniui *D*, *m* – klasių skaičius, p_i neneigiama tikimybė, jog objektas priklauso klasei C_i. Tarkim, jog duomenys yra suskaidomi į v dalių ir duomenų rinkinys *D* dalinamas į *v* dalių $\{D_1, D_2, ..., D_v\}$. Atlikus tokį skaidymą galima pritaikyti (2.2.2.1) formulę ir apskaičiuoti entropiją po suskaidymo, kuri išreiškiama formule (2.2.2.2):

$$E_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * E(D_j) \qquad (2.2.2.2)$$

Čia: E_A yra entropija atlikus suskaidymą *A*, kai duomenys *D* suskaidyti į v poaibių. Tada informacijos gavimas yra skirtumas tarp pradinės ir po atskyrimo *A* gautos entropijos t.y. (2.3.2.3) formulė:

$$Gain(A) = E(D) - E_A(D) \qquad (2.2.2.3)$$

Informacijos gavimas apskaičiuojamas panaudojus informacijos gavimo (2.2.2.3) formulę ir dar papildomai įvertinus naujai sudarytų D_j poaibiuose esančių objektų skaičių (2.2.2.4 formulė):

$$GainRatio(A) = \frac{Gain(A)}{Split_A(D)} = \frac{Gain(A)}{-\sum_{j=1}^{v}\frac{|D_j|}{|D|} * log_2\left(\frac{|D_j|}{|D|}\right)} \quad (2.2.2.4)$$

Naudojant apibrėžtus matus, kiekviename *Sprendimų medžio* sudarymo žingsnyje yra patikrinama visi galima kintamųjų skaidymo variantai ir tas skaidymas, kuris pagal pasirinktą matą geriausiai atskiria objektus į klases yra atliekamas. Toliau, jau atskirtuose poaibiuose yra kartojamas šis veiksmas ir toks skaidymas vyksta tol, kol visi medžio lapai yra sudaryti tiki š vienai klasei priklausančių objektų arba yra tenkinama skaidymo sustojimo sąlyga.
Atsitiktinio miško idėja yra sugeneruoti nurodytą skaičių *Atsitiktinių medžių* imant vis kitą, atsitiktinai sugeneruotą, duomenų rinkinio *D* poaibį. Tada remiantis visų jų rezultatais, objektui priskirti tokią klasę, kurią nurodė didžioji Sprendimų medžių dalis.

2.2.3. Metodų tikslumo vertinimas

Svarbi metodų ypatybė yra jų pagalba gautų rezultatų tikslumo vertinimas. Klasterizavimo metodai tyrime yra vertinami pagal tai, kaip gerai atskiria pirkėjus, atsižvelgiant į klasterių centrus. Klasifikavimo metodų tikslumo vertinimui atlikti galima rasti daug rodiklių. Kadangi tyrime pirkėjai yra suskirstyti į daugiau negu vieną klasę, pirmiausia bus naudojama sumaišymo matrica (angl. *confusion matrix*) (žr. 2.2.3.1 lentele), kuri nurodo kiek objektų buvo teisingai suklasifikuota ir yra naudojama tolimesniems rodikliams apskaičiuoti (lent. Pirkėjus skirstant į tris klases pagal jų kreditingumo vertinimą – geras, vidutinis ir blogas (angl. *good, average* ir *bad*) ir šios klasės yra žymimos atitinkamai *G, A* ir *B* raidėmis. Tos reikšmės, kurios buvo suklasifikuoto teisingai yra žymimos *TG, TA* ir *TB,* kurie atitinka *True Good, True Average* ir *True Bad*. Reikšmės, kurios buvo blogai suklasifikuotos žymimos *False Good, False Average* ir *False Bad* (sutrumpinti žymėjimai *FG, FA* ir *FB*).

2.2.3.1 lentelė. Sumaišymo lentelė

	Prognozuojama reikšmė – G	Prognozuojama reikšmė – A	Prognozuojama reikšmė – B
Stebėta reikšmė – G	TG	FA	FB
Stebėta reikšmė – A	FG	TA	FB
Stebėta reikšmė – B	FG	FA	TB

Pagal šią lentelę skaičiuojamas bendras metodo tikslumas (angl. *accuracy*) pagal (2.2.3.1) formule:

$$Tikslumas = \frac{TG+TA+TB}{TG+FG+TA+FA+TB+FB} \quad (2.2.3.1)$$

Šio rodiklio tikslas yra nurodyti, koks bendras kiekis objektų buvo suklasifikuotas teisingai. Kadangi tyrime yra trys atskirtos klasės, svarbu yra žinoti ir kiek teisingai yra suklasifikuota kiekvienos iš jų objektų, nes modelio tikslumas to nenurodo. Rodiklis, kuris parodo kiekvienos klasės tikslumą yra vadinamas jautrumu (angl. *sensitivity*) ir kiekvienai iš klasių yra skaičiuojamas pagal atitinkama (2.2.3.2) formulę. Šis rodiklis rodo tikimybę, jog objektai priklausantis tam tikrai klasei bus suklasifikuojamas teisingai.

$$Jautrumas(G) = \frac{TG}{TG+FA+FB};\ Jautrumas(A) = \frac{TA}{TA+FG+FB};\ Jautrumas(B) = \frac{TB}{TB+FB+FA}$$ (2.2.3.2)

Kitas rodiklis, kuris yra naudojamas vadinamas precizija (angl. *precision*) ir jis nurodo, kokia tikimybė, jog suklasifikuota reikšmė iš tikro priklauso šiai klasei ir skaičiavimui yra naudojama atitinkama (2.2.3.3) formulė:

$$Precizija(G) = \frac{TG}{TG+FG};\ Precizija(A) = \frac{TA}{TA+FA};\ Precizija(B) = \frac{TB}{TB+FB} \quad (2.2.3.3)$$

Jautrumas ir precizija parodo kaip tiksliai buvo suklasifikuota kiekviena iš grupių. Norint apjungti šiuos rodiklius, skaičiuojamas bendras rodiklis F1 kiekvienai klasei pagal (2.2.3.4) formule ir kuo jis didesnis, tuo geriau yra klasifikuojami klasės objektai:

$$F1 = 2 * \frac{precizija * jautrumas}{precizija + jautrumas} \quad (2.2.3.4)$$

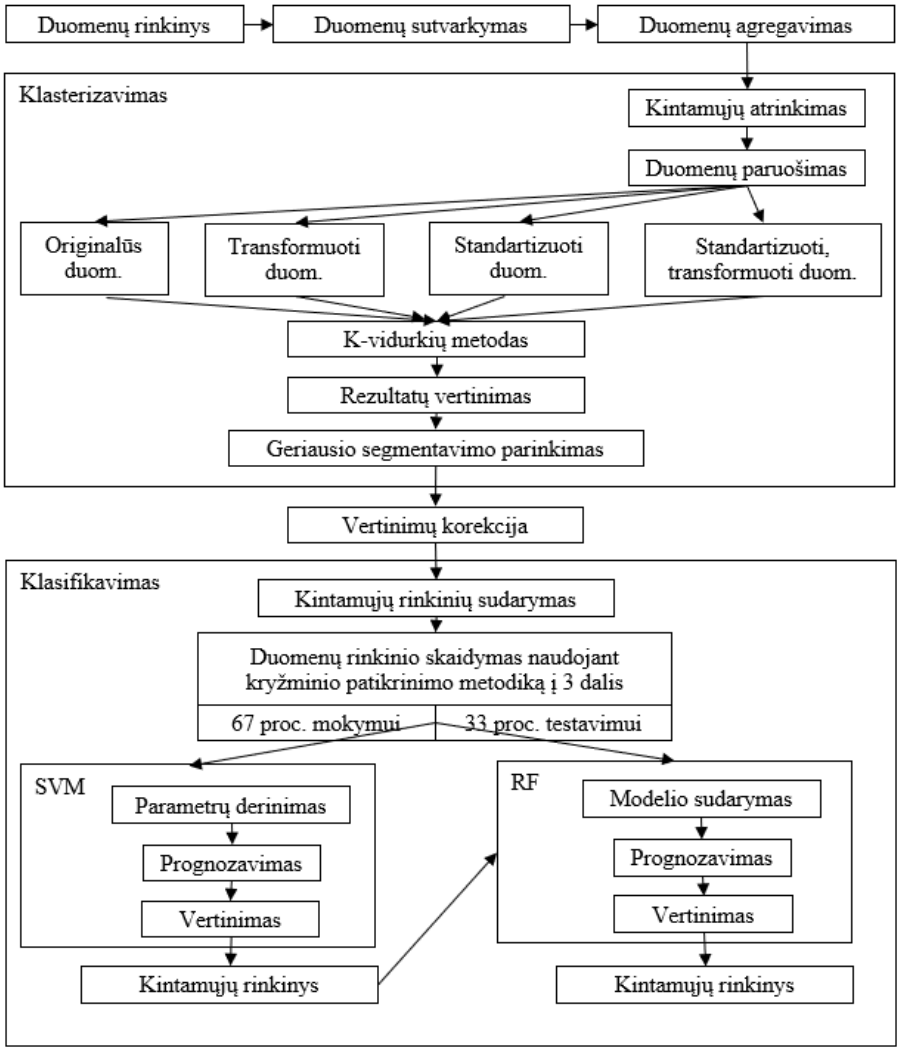
Ir tada naudojant kiekvienos klasės F1 rodiklį, suskaičiuojamas bendras F1 rodiklis visam modeliui naudojant paprasto vidurkio formulę. Naudojant visus šiuos rodiklius, bus apskaičiuojamas sudaryto klasifikatoriau gerumas.

3. Tyrimo eiga ir rezultatai

Šioje dalyje apžvelgiama iš kur yra gauti duomenys ir kokie jie yra. Atsilikus jų pradinę žvalgomąją analizę, duomenys agreguojami klientų lygyje ir gauti rezultatai taip pat aptariami. Toliau atliekamas klasterizavimas, kurio tikslas yra suskirstyti klientus pagal jų kreditingumo vertinimą. Turint šią informaciją yra kuriamas klientų klasifikavimo modelis.

3.1. Tyrimo eiga

Pradinis duomenų rinkinys, su informacija apie pirkėjų mokumą sąskaitų lygyje, yra apžvelgiamas ir sutvarkomas. Informacija yra agreguojama pirkėjo lygyje ir tuo metu apskaičiuojami rodikliai, apibūdinantys pirkėjo mokumą. Toliau tyrimas vyksta pagal 3.1.1 paveiksle pateiktą schemą.



3.1.1 pav. Tyrimo atlikimo schema

Visą tyrimo schema galima suskirstyti į dvi dalis: klasterizavimą ir klasifikavimą. Pirmosios tikslas yra automatizuotai sugrupuoti pirkėjus pagal jų mokumo informaciją. Tam, kad būtų lengviau interpretuoti klasterizavimo rezultatus, kintamieji yra suskaidomi į keturias grupes ir iš kiekvienos yra išrenkami kintamieji, kurie grupės viduje turi mažiausią priklausomybę nuo likusiųjų. Kadangi kintamųjų reikšmės yra įvairių skalių, todėl yra sudaromos trys papildomos duomenų variacijos. Pirmoji gaunama pradinius duomenys standartizuojant, antroji – pradinius duomenis transformuojant

bei pašalinant išskirtis ir paskutinė, prieš gautus rezultatus tada transformuojant. Tada atliekama klasterizavimas naudojant K-vidurkių metodą. Gauti rezultatai yra interpretuojami ir pagal juos yra parenkamas vienas variantas, geriausiai atspindintis pirkėjų kreditingumo vertinimą. Remiantis gautais rezultatais, pirkėjai yra vertinami ir jų įvertinimas pakoreguojamas.
Toliau seka pirkėjų klasifikavimo modelio sudarymas naudojant jiems priskirtas kreditingumo vertinimo grupes. Modelio sudarymui pirmiau yra naudojamas Atraminių vektorių metodas su radialine branduolio funkcija. Prieš jo naudojimą, kintamųjų priklausomybė yra apžvelgiama ir sudaromos įvairios kintamųjų kombinacijos ir visoms joms pritaikomas minėtas metodas ir modelio parametrai suderinami kiekvienam rinkiniui atskirai. Gavus rezultatus, jie palyginami ir išrenkamas geriausią rezultatą davęs kintamųjų rinkinys. Jam tada pritaikomas Atsitiktinio miško metodas ir pagal jį nustatomas kintamųjų svarbumas rezultatams. Metodas vis naudojamas, pašalinant nereikalingus kintamuosius, kol gaunamas geriausias rezultatas su mažiausiai kintamųjų.

3.2. Modelio realizacija

Tyrimo metu duomenų apdorojimas, modelių sudarymas ir vertinimas yra įgyvendinamas naudojant R programavimo kalbą. Ji pasirinkta dėl to, jog yra atviro kodo ir jai sukurti paketai suteikia galimybes visus reikiamus veiksmus atlikti vienoje vietoje. Taip pat, ši programavimo kalba yra laikoma viena geriausių, kalbant apie duomenų analitika. Realizacijos kodas yra pateikiamas 7 priede ir yra išskaidytas į tris atskirtus programų kodus, kuri skirti duomenų apžvalgai (1 programa), klasterizavimo modelių sudarymui (2 programa) ir klasifikavimo modelių sudarymui (3 programa).

Programose yra naudojami tokie paketai: *readr, pastecs, dplyr, ggpubr, corrplot, Hmisc, factoextra, pracma, caret, e1071, tidyr, randomForest, randomForestExplainer, readxl*.

3.3. Duomenys

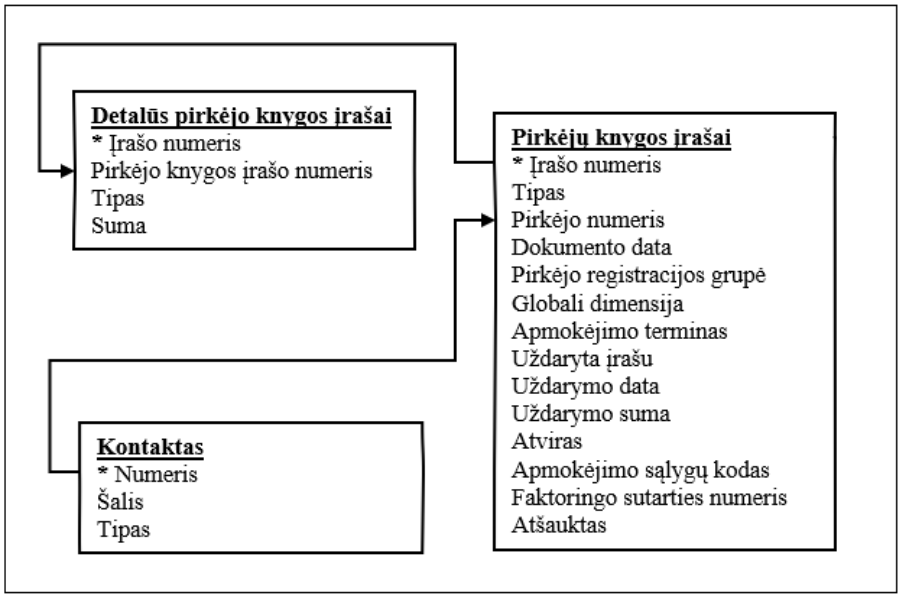
Tyrimo duomenys yra siejami su UAB „Idėjų valda", kuri bendradarbiauja su Kauno technologijos universitetu, vykdant projektą „Prekybinių procesų rizikos valdymas pasitelkiant didžiuosius duomenis ir dirbtinį intelektą“. Šiame projekte yra naudojama trečiųjų šalių turima klientų mokumo informacija. Duomenys yra nuo 2015 metų pradžios iki 2019 metų pabaigos. Ši informacija yra gaunama kaip pardavimų vykdymo šalutinis rezultatas ir yra saugoma vienos duomenų bazės keliose lentose (bendruoju atveju, informacijos apie padavimus ir mokėjimus yra saugojama ir projekte naudojama daugiau, o detaliau apžvelgiama tik ta informacija, kuri naudojama baigiamajame projekte).

Naudojami duomenys yra saugomi trijose lentose:

- “*Kontaktas”*. Šioje lentoje saugoma bendrinė informacija apie pirkėją. Kadangi yra naudojami nuasmeninti duomenys, iš šios lentos buvo paimti keli atributai, kurie apibūdina, bet tiksliai neidentifikuoja pirkėjo.
- *“Pirkėjų knygos įrašai”*. Šioje lentoje yra informacija apie visas kliento transakcijas (pardavimus, mokėjimus, sudengimus ir t.t.). Kadangi reikalinga tik pardavimo sąskaitų mokumo informacija nurodytu periodu, buvo naudojamas filtras įrašo tipui, dokumento datai ir požymiui *“Atšauktas”* (atšaukti dokumentai nėra naudojami projekte).
- *“Detalūs pirkėjų knygos įrašai”.* Šioje lentoje yra informacija apie transakcijų sumas vietine valiuta t.y. Eurais. Kadangi aktualu yra pardavimo sąskaitų pradinės sumos, naudojamas filtras įrašo tipui.

Šios lentos tarpusavyje yra susijusios toki ryšiais (žr. 3.3.1 pav.):

- *„Kontaktas“* įrašas gali turėti daug *„Pirkėjų knygos įrašų“* įrašų (ryšio tipas: vienas su daug) ir šios lentos yra sujungtos atributais *„Numeris“* (pirminis *„Kontaktas“* lentos raktas) ir *„Pirkėjo numeris“* (antrinis *„Pirkėjų knygos įrašai“* lentos raktas);
- *„Pirkėjo knygos įrašai“* įrašas gali turėti daug *„Detalūs pirkėjo knygos įrašai“* įrašų (ryšio tipas: vienas su daug) ir šios lentos yra sujungtos atributais *„Įrašo numeris“* (pirminis *„Pirkėjų knygos įrašai“* lentos rantas) ir *„Pirkėjo knygos įrašo numeris“* (antrinis *„Detalūs pirkėjo knygos įrašai“* lentos raktas).



3.3.1 pav. Duomenų saugojimo schema

Naudojant aprašytus ryšius informacija buvo apjungta į vieną duomenų rinkinį, kuriame informacija yra pardavimo sąskaitos faktūros lygyje.

Duomenų rinkinys buvo papildomai apdorojamas: pakeičiamas datų saugojimo formatas, duomenų bazėje naudojami reikšmių kodai pakeičiami į atitinkamas reikšmes, netinkami įrašai ištrinami (remiantis specialisto vertinimu). Atlikus šį apdorojimą, buvo atrinkti įrašai apie 2018 metais išrašytas pardavimų sąskaitas, kurie naudojami baigiamajame projekte.

3.4. Duomenų apžvalga

Pradiniame duomenų rinkinyje yra 48 262 įrašai apie per 2018 metus pirkėjams išrašytas sąskaitas ir jų apmokėjimą. Šią informaciją sudaro 16 kintamųjų, kurių aprašymas pateiktas 3.4.1 lentelėje.

3.4.1 lentelė. Pradiniame duomenų rinkinyje esantys kintamieji

Pavadinimas duomenų rinkinyje	**Lietuviškas pavadinimas**	**Aprašymas**
Entry_no	Įrašo numeris	Įrašo identifikatorius, pagal kurį galima atskirti, jog įrašai yra apie pardavimus iš skirtingų įmonių
Customer_no	Pirkėjo numeris	Pirkėjo identifikatorius, kuris neturi jokios informacijos apie patį pirkėją

Country	Šalis	Vardų skalės kintamasis, nurodantis pirkėjo šalį
Customer_type	Pirkėjo tipas	Vardų skalės kintamasis, kuris nurodo kliento tipą pagal jo registracijos kodą
Customer_group	Pirkėjo registracijos grupė	Vardų skalės kintamasis su keturiomis reikšmėmis, kurios identifikuoja kokiai registracijos grupei priklauso klientas
Document_date	Dokumento data	Datos tipo kintamasis (formatas YYYY-MM-DD), nurodantis kada buvo išrašyta pardavimo sąskaita faktūra
Payment_code	Apmokėjimo sąlygų kodas	Vardų skalės kintamasis, kuris nurodo kokios apmokėjimo sąlygos buvo pritaikytos sąskaitoje
Due_date	Apmokėjimo terminas	Datos tipo kintamasis (formatas YYYY-MM-DD), nurodantis iki kada pirkėjas turi apmokėti pardavimo sąskaitą faktūrą
Invoiced_amount	Sąskaitos suma	Kiekybinis kintamasis, kuris parodo pardavimo sąskaitos sumą vietine valiuta t.y. eurais
Dimension	Dimensija	Vardų skalės kintamasis, kuris nurodo kokio tipo prekės buvo parduotos
Factoring	Faktoringas	Binarinis kintamasis: 1 – sąskaitai taikomas faktoringas, 0 – sąskaitai netaikomas faktoringas
Debt_default	Skola nurašyta	Binarinis kintamasis: 1 – skola nurašyta kaip beviltiška, 0 – skola buvo sumokėta arba dar laukiama mokėjimo
Open	Atviras	Binarinis kintamasis: 1 – skola neapmokėta t.y. sąskaita dar atvira, 0 – skola sumokėta t.y. sąskaita uždaryta
Closed_by_entry	Uždaryta įrašu	Įrašo, kuriuo buvo galutinai padengta t.y. apmokėta sąskaita, numeris
Closed_at_date	Uždarymo data	Datos tipo kintamasis (formatas YYYY-MM-DD), nurodantis kada buvo galutinai apmokėta sąskaita
Closed_by_amount	Uždarymo suma	Kiekybinis kintamasis, kuris parodo kokia suma buvo galutinai apmokėta sąskaita. Suma išreiškiama eurais

Kadangi šie duomenys bus agreguojami, jie nėra koreguojami daugiau negu buvo duomenų paėmimo metu. Bendrai juos peržiūrėjus galima išskirti jog didžioji dalis t.y. 83,9 proc. sąskaitų yra išrašoma Lietuvos pirkėjams (jie, pagal grupes yra išskirti į juridinius ir fizinius asmenis), tačiau šios sąskaitos sudaro tik 53,9 proc. visų išrašytų sąskaitų sumos. Tai reiškia, jog Lietuvos pirkėjams išrašomos sąskaitos yra mažesnių sumų lyginant su Trečiųjų šalių pirkėjams išrašytoms sąskaitoms. Kitų Europos Sąjungos šalių pirkėjams rašomos sąskaitos taip pat yra didesnių sumų lyginant su Lietuvos pirkėjams išrašomomis ir nors sąskaitų kiekis sudaro tik 15,1 proc. nuo visų išrašytų sąskaitų, jų suma sudaro beveik pusę sąskaitų bendros sumos (žr. 3.4.2 lentelę).

3.4.2 lentelė. Išrašytų pardavimų sąskaitų duomenys pagal klientų grupes

Klientų grupė	**Dokumentų kiekis, vnt.**	**Dokumentų kiekis, proc.**	**Dokumentų bendra suma, Eur**	**Dokumentų bendra suma, proc.**	**Vid. dokumentų suma, Eur**
EUROPOS	7 298	15,12	106 321 424	41,49	14 568,57
FIZINIAI	3 223	6,68	4 375 746	1,71	1 357,66
LIETUVOS	37 246	77,17	133 730 942	52,19	3 590,48
UZSIENIO	495	1,03	11 815 706	4,61	23 870,11
	48 262	**100,00**	**256 243 818**	**100,00**	**5 309,43**

Iš Europos Sąjungos pirkėjams išrašytų sąskaitų daugiausia atitenka Lenkijos ir Latvijos šalių pirkėjams, o imant TOP penkių Europos Sąjungos šalių pirkėjams išrašomas sąskaitas, jos sudaro apie 63 proc. visų užsienio šalims išrašomų sąskaitų (žr. 3.4.3 lentelę). Iš viso sąskaitos išrašytos klientams iš 37 skirtingų šalių. Tai rodo, jog prekės yra parduodamos į daug įvairių šalių, tačiau prekių eksportas sudaro mažą dalį visų pardavimų.

3.4.3 lentelė. Pardavimų sąskaitų, išrašytų ne Lietuvos pirkėjams, kiekiai

Šalies kodas	**Dokumentų kiekis, vnt.**	**Dokumentų kiekis, proc.**
PL	1 415	18,15
LV	1 372	17,59
IT	831	10,66
EE	747	9,58
DE	582	7,47

Didžiajai daliai išrašytoms sąskaitoms yra taikomas apmokėjimo atidėjimas. 29,61 proc. atvejų yra skiriamas vieno mėnesio atidėjimas, 10,82 proc. – einamasis mėnesis ir sekantis mėnesis. Kiti apmokėjimo terminai varijuoja, nors išlieka prisirišimas prie mėnesio – pusę mėnesio, pusantro mėnesio, du ar trys mėnesiai ir t.t. 9,17 proc. pardavimų yra su išankstiniu apmokėjimu ir tik 5,28 proc. sąskaitų yra apmokėta grynais. Tai parodo, jog didžioji dalis pinigų, gaunamų iš pardavimų yra gaunama po tam tikro laiko t.y. ne iš karto. Pažiūrėjus kaip yra apmokėtos sąskaitos (žr. 3.4.4 lentelę), matoma jog tik labai maža dalis skolų buvo nurašytos kaip beviltiškos. Daugiau negu pusę sąskaitų buvo apmokėtos iš anksto arba tiksliai termino metu. Likusioji dalis sąskaitų yra apmokamos po termino ir tai sudaro net 45,22 proc. visų išrašytų sąskaitų.

3.4.4 lentelė. Pardavimų sąskaitų apmokėjimo tendencijos

Apmokėjimo laikas	**Sąskaitų kiekis, vnt.**	**Sąskaitų kiekis, proc.**
Iki termino	17 667	36,61
Lygiai su terminu	8 675	17,98
Vėlavo iki 30 d.	17 835	36,95
Vėlavo nuo 30 iki 60 d.	3 121	6,47
Vėlavo daugiau negu 60 d.	870	1,80
Nurašyta skola	78	0,16
Neapmokėti dokumentai	16	0,03

Apžvelgus visus atributus, jie agreguojami pirkėjų lygyje ir iš viso išvesta daugiau negu 25 atributų, kurie apibūdina pirkėjų pirkimo tendencijas. Pilnas atributų sąrašas pateiktas 1 priede, o jų charakteristikos – 2 priede. Visi jie buvo suskirstyti į penkias grupes, jog būtų lengviau atskirti kokią informaciją apie pirkėją jie teikia. Pagal kiekvieną iš jų trumpai apžvelgiami pirkėjai.

Pirmoji kintamųjų grupė apibūdina pačius pirkėjus ir net 81,2 proc. jų yra Lietuvos juridiniai (55,7 proc.) ir fiziniai (25,5 proc.) asmenys. Kitų šalių pirkėjai sudaro 18,8 proc. visų ir nors prekiaujama yra su 36 skirtingų šalių pirkėjais, daugiausiai išsiskiria Latvijos ir Italijos pirkėjų kiekis (žr. 3.4.1 pav.)

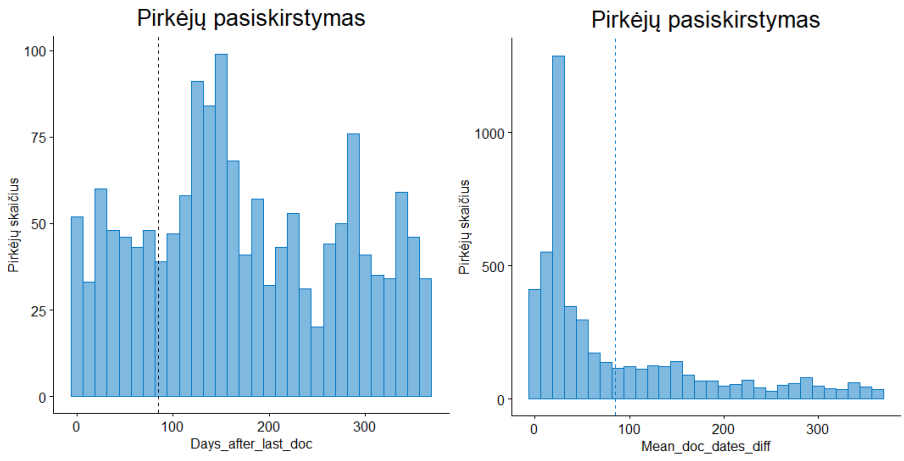
3.4.1 pav. Pirkėjų pasiskirstymas pagal šalis (neįskaitant Lietuvos)

Taip pat pirkėjai retai kada perka kelių pramonės šakų prekes ir dažniausiai bendradarbiauja tik su viena iš nagrinėjamų įmonių (žr. 3.4.5 lentelę).

3.4.5 lentelė. Pirkėjų pasiskirstymas pagal pirkimo tendencijas

Bendradarbiavimo tipas ir perkami produktai	**Pirkėjų kiekis, Vnt.**	**Pirkėjų kiekis, Proc.**
Bendradarbiauja viena įmone ir perka vienos pramonės šakos prekes	4343	89,42
Bendradarbiauja viena įmone, bet perka kelių pramonės šakų prekes	269	5,54
Bendradarbiauja su keliomis įmonės, bet perka vienos pramonės šakos prekes	54	1,11
Bendradarbiauja su keliomis įmonėmis ir perka kelių pramonės šakų prekes	191	3,93

Sekanti kintamųjų grupė apibūdina pirkėjui išrašytų sąskaitų kiekį, per kokį laikotarpį jos išrašytos ir jų dažnumą ir kada buvo paskutinė išrašyta sąskaita. Šioje grupėje yra kintamųjų, kurie turėjo tuščių reikšmių, kurios buvo užpildytos (detali informacija pateikta 3 priede). Iš pirkėjų, net 31,13 proc. turi tik vieną sąskaitą. Įvertinus, jog vidutinis dienų skirtumas tarp sąskaitų yra 85 ir tai kaip seniai buvo išrašyta sąskaita, iš 1512 pirkėjų, 341 galima laikyti naujais (žr. 3.4.2 pav. kairėje).



3.4.2 pav. Dienų, nuo paskutinės išrašyto sąskaitos (kairėje) ir vidutinių dienų skaičiaus tarp sąskaitų (dešinėje) pasiskirstymas

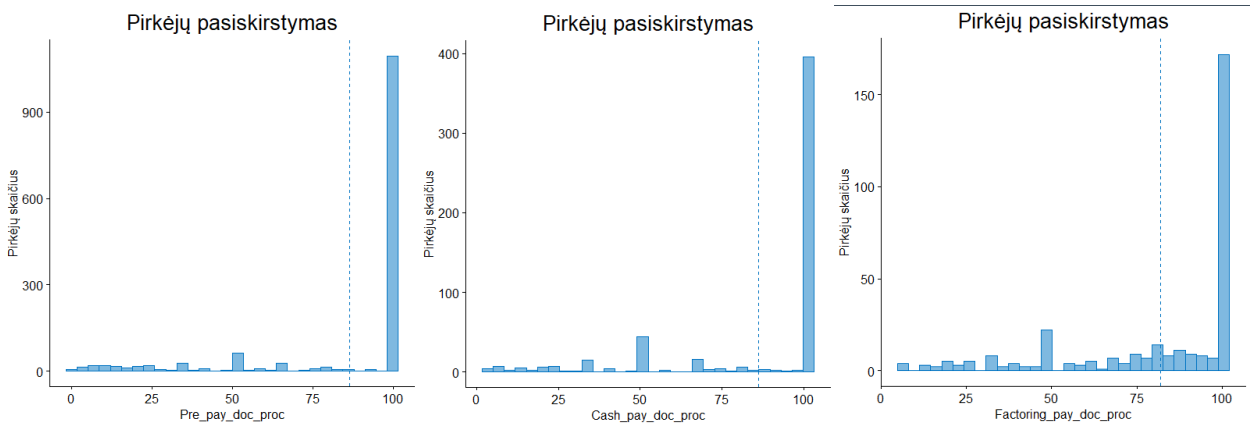
Dažniausiai, pirkėjų vidutiniai periodai tarp sąskaitų yra kalendorinis mėnesis (žr. 3.4.2 pav. dešinėje), o vidutiniškai su pirkėjų bendradarbiaujama 5 mėnesius. Tai rodo, kad didesnė dalis pirkėjų yra orientuota į ilgalaikę partnerystę.

Trečioji kintamųjų grupė yra susijusi su sąskaitų apmokėjimo sąlygomis ir parodo kiek procentų sąskaitų kiekvienas pirkėjas apmokėjo iš anksto, grynais, kiek sąskaitų turėjo faktoringo sutartis ir koks vidutiniškai yra apmokėjimo atidėjimo laikotarpis ir kaip jis varijuoja. Pirkėjai, kurie atsiskaito už visas savo sąskaitas išankstiniu apmokėjimu arba grynais vidutiniškai atlieka mažai pirkimų, bet sudaro apie 30 proc. visų pirkėjų. Atskirai išskirti yra ir tiek pirkėjai, kurie turi faktoringo sutartis, nes didžiąją dalį jų skolos sumoka tretieji asmenys ir taip jų rizikingumas stipriai sumažėja. Tokios sutartys yra sudaromos pirkėjams, su kuriais bendradarbiavimas yra ilgalaikis ir tą patvirtina faktas, jog vidutiniškai tokie pirkėjai turi po 18 sąskaitų (žr. 3.4.6 lentelę).

3.4.6 lentelė. Pirkėjų, kurie atsiskaito be atidėjimų kiekis

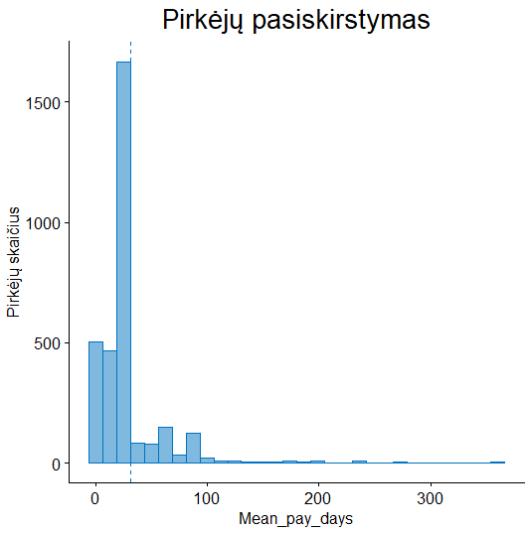
Apmokėjimo būdas	**Pirkėjai, kurie atsiskaitė bent vieną kartą**	**Pirkėjai, kurie atsiskaitė už visas sąskaitas**	**Vidutinis pirkėjų sąskaitų skaičius**
Išankstinis apmokėjimas	1415	1096	2,30
Apmokėjimas grynais	537	395	2,07
Faktoringo sutartis	331	170	17,75

Įdomus faktas yra, jog dažniausiai pirkėjai kurie naudoja minėtus atsiskaitymo būdus, juos taiko visiems savo pirkimams t.y. yra mažai pirkėjų, kurie kombinuoja šiuos apmokėjimo būdus su apmokėjimų atidėjimais (žr. 3.4.3 pav.)



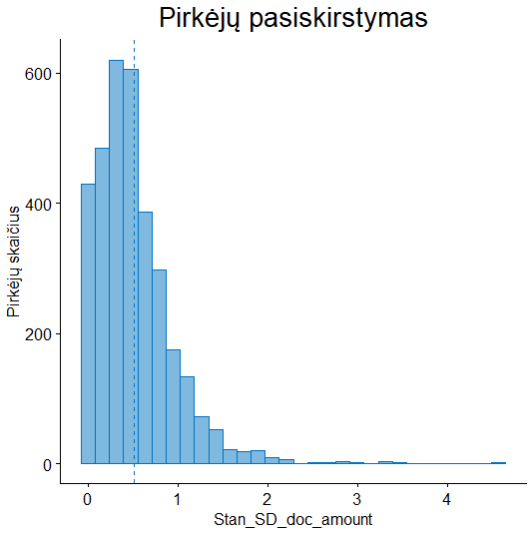
3.4.3 pav. Pirkėjų sąskaitų kiekis, už kurį atsiskaitoma be apmokėjimo atidėjimo

Likusių pirkėjų, kuriems yra taikomi apmokėjimų atidėjimai, dažniausiai apmokėjimo atidėjimas yra vienas mėnesis (žr. 3.4.4 pav.). Taip pat, suteiktas apmokėjimo atidėjimo terminas labai mažai kinta t.y. pirkėjui apmokėjimo atidėjimo terminas keičiasi, bet labai nežymiai – vidutinė apmokėjimo termino variacija yra apie 5 dienos.



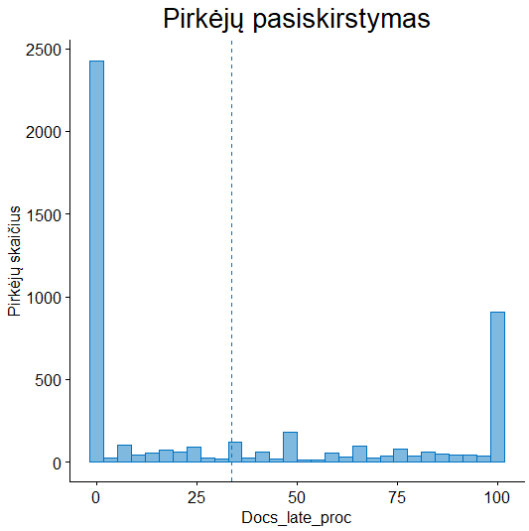
3.4.4 pav. Pirkėjų vidutiniai apmokėjimų atidėjimo terminai

Priešpaskutinė kintamųjų grupė skirta apibūdinti pirkėjų sąskaitų sumas ir jų variaciją. Iš karto matosi, jog pirkėjų vidutinės sąskaitų sumos labai stipriai varijuoja. Kaip pavyzdys, tik 495 pirkėjų vidutinė sąskaitų suma yra virš 10 000 Eur. O didžiosios dalies, net 3491 pirkėjų, sąskaitų sumos vidurkis yra žemiau negu 2 500 Eur. Tai rodo, jog didžioji dalis pirkėjų yra linkusi pirkti mažomis sumomis. Pagal sąskaitų sumų variacijos koeficientą, matoma, jog didžiosios dalies pirkėjų sąskaitų sumos varijuoja iki pusės vidurkio sumos (iš šio grafiko yra pašalinti pirkėjai, su vien sąskaita) (žr. 3.4.5 pav.). Tai reiškia, jog pirkėjai perka ganėtinai pastoviomis sumomis.



3.4.5 pav. Pirkėjų sąskaitų sumų variacijos koeficiento pasiskirstymas

Paskutinioji kintamųjų dalis yra susijusi su faktiniu sąskaitų apmokėjimu. Pagal juos galima pasakyti, jog tik 328 pirkėjai vidutiniškai vėluoja apmokėti sąskaitas daugiau negu 30 dienų. Tačiau pirkėjų, kurie virš 50 proc. sąskaitų vėluoja apmokėti yra 1531 (žr. 3.4.6 pav.). Tai parodo, jog pirkėjų mokėjimo laikai yra nepastovūs ir tą patvirtina sklaida apie vidutinį apmokėjimo atlikimą, kuri yra apie 8 dienas.



3.4.6 pav. Pirkėjų sąskaitų dalis, už kurias jie vėluoja sumokėti

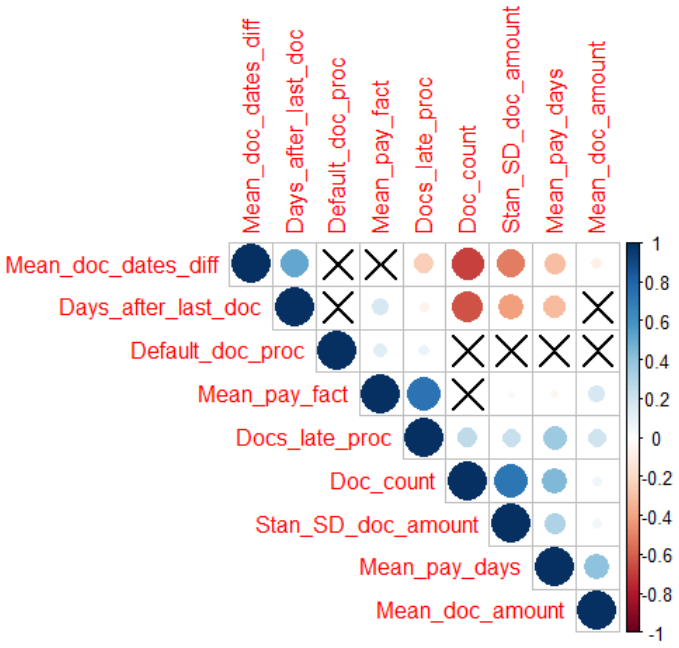
Duomenų rinkinyje yra 15 pirkėjų, kurie nėra apmokėję dalies arba visų sąskaitų, ir 27 pirkėjai, kurių skolų dalis yra nurašyta kaip beviltiška ir bendra šių nurašymų suma yra virš 500 tūkst. Eur.

3.5. Klasterizavimas

Po duomenų agregavimo, kiekvieną klientą apibūdina 27 kintamieji. Kadangi klasterizavimui naudojamas yra *K-vidurkių metodas*, kokybiniai kintamieji negali būti panaudojami, o kiekybinių reikia sumažinti – taip segmentavimas bus lengviau interpretuojamas. Tam iš visų keturių kiekybinių kintamųjų grupių išrenkami yra 9, kurie geriausiai apibūdina tam tikrus aspektus apie pirkėją ir taip pat tarpusavyje minėtose kintamųjų grupėse yra mažiausiai priklausomi tarpusavyje:

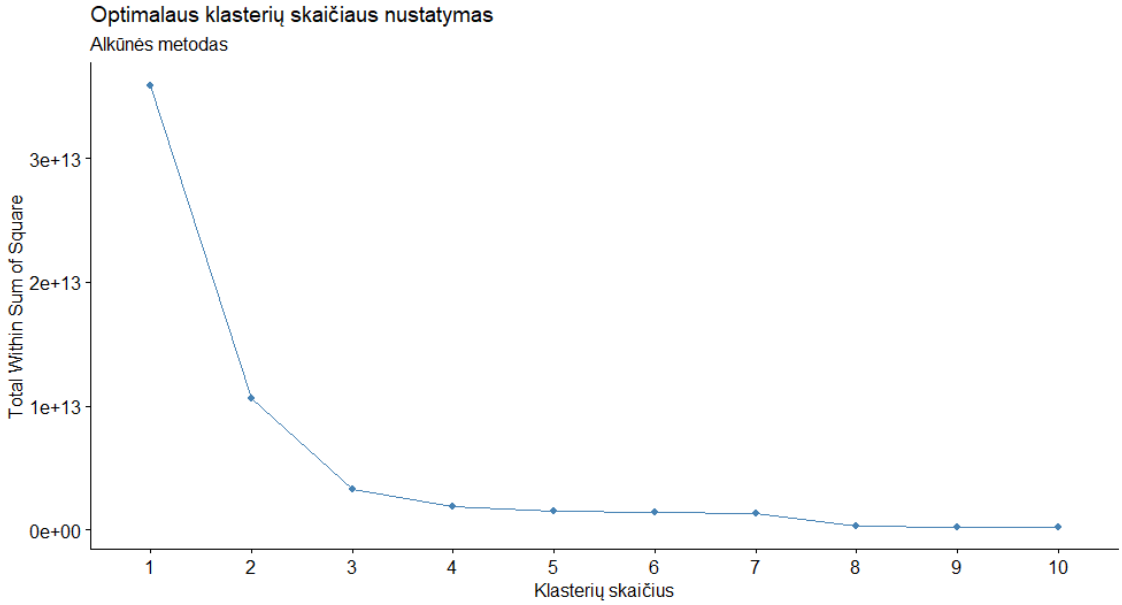
- *„Mean_doc_dates_diff“, „Days_after_last_doc“* ir *„Doc_count“* yra susiję sąskaitų dažnumu, bet apibūdina skirtingus to aspektus – kaip dažnai pirkėjas atlieka pirkimus, kada buvo jo paskutinis pirkimas ir kiek pirkimų jis iš viso atliko;
- *„Mean_pay_days“* – nurodo koks vidutinis apmokėjimo atidėjimas taikomas yra pirkėjui;
- *„Stan_SD_doc_amount“* ir *„Mean_doc_amount“* apibūdina kaip stipriai varijuoja pirkėjui išrašomų sąskaitų suma ir koks yra pirkėjo sąskaitų sumos vidutinė suma, tačiau šie du kintamieji yra nepriklausomi;
- *„Default_doc_proc“, „Mean_pay_fact“* ir *„Docs_late_proc“* yra susiję su pirkėjo nurašytomis skolomis (nors šis kintamasis turi labai mažai nenulinių reikšmių, jis yra būtinas), kiek vidutiniškai pirkėjas vėluoja apmokėti sąskaitas ir kiek procentų sąskaitų vėluoja.

Atrinkus kintamuosius, jiems buvo apskaičiuoti Spearmano koreliacijos koeficientai (nes jie nėra pasiskirstę pagal normalųjį skirstinį). Rezultatuose matoma, jog didžioji dalis kintamųjų yra susiję (žr. 3.5.1 pav.), tačiau kadangi jie yra išvestiniai ir trūkstamos reikšmės buvo pakeistos kombinuojant kitas - to buvo tikėtasi ir stiprią priklausomybę turinčių kintamųjų yra mažuma. O ryšys, kurį su 90 proc. tikimybę galima laikyti nereikšmingu yra pažymėtas X.



3.5.1 pav. Kintamųjų, naudojamų K-vidurkių metode, tarpusavio priklausomybė

Toliau pasinaudojus Alkūnės metodu buvo nuspręsta, jog optimalus klasterių skaičius yra trys (žr. 3.5.2 pav.) ir tai sutampa su tikslu suskirstyti pirkėjus į trijų lygių kreditingumo įvertinimą. Tačiau taip pat palyginimui atliekamas klasterizavimas ir į keturis klasterius.



3.5.2 pav. Alkūnės metodo vizualizacija optimalaus klasterių skaičiui nustatyti

Kadangi *K-vidurkių metodas* yra imlus išskirtims, metodas buvo pritaikytas keturiems skirtingiems duomenų rinkiniams. Pirmasis buvo originalus, antrasis - sukurtas standartizuojant šias reikšmes, siekiant panaikinti skirtingų kintamųjų skalių įtaką klasterizavimui. Paskui reikšmės buvo transformuotos taip, kad kiekvienas kintamasis turėtų kuo mažiau išskirčių:

- *„Mean_doc_dates_diff“, „Days_after_last_doc“, „Stan_SD_doc_amount“, „Default_doc_proc“* ir *„Docs_late_proc“* kintamųjų reikšmėms buvo ištraukta kvadratinė šaknis;

- *„Doc_count“, „Mean_pay_days“* ir *„Mean_doc_amount“* kintamųjų reikšmės buvo logoritmuojamos;
- *„Mean_pay_fact“* kintamojo reikšmėms ištraukta trečio laipsnio šaknis.

Po transformacijų išskirtys buvo pakeistos atitinkamai 5 ir 95 kvantiliais. Taip buvo gautas trečias duomenų rinkinys klasterizavimui. Ketvirtasis rinkinys buvo standartizuotos trečiojo reikšmės.

Atlikus klasterizavimą, visi rezultatai buvo analizuojami ir klasteriai interpretuojami. Iš visų (visi klasterizavimų rezultatai pateikti 4 priede) segmentavimu geriausiu išrinktas klasterizavimas, gautas standartizavus originalias kintamųjų reikšmes (žr. 3.5.1 lentelę)

3.5.1 lentelė. Sudarytų pirkėjų klasterių centrai

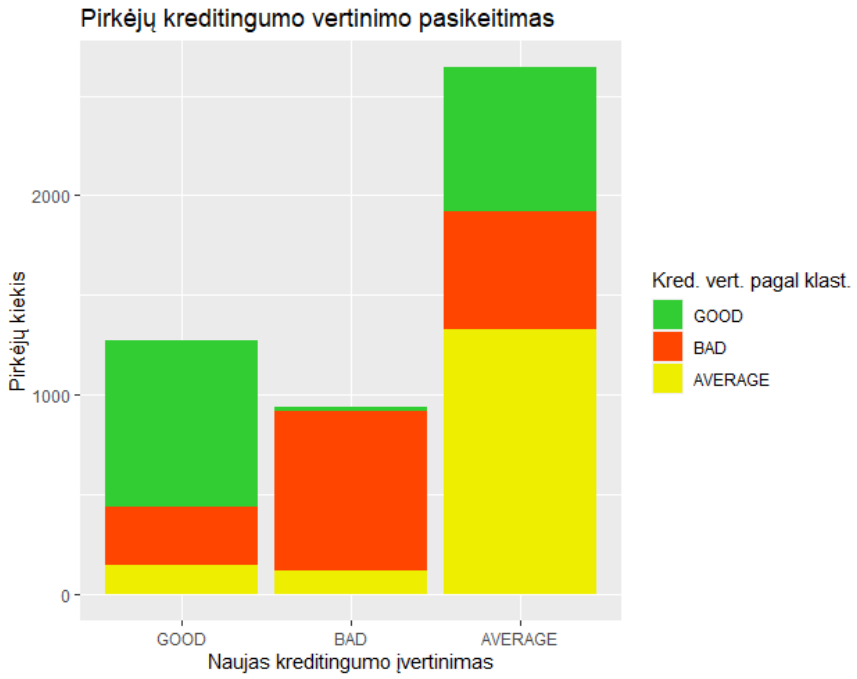
Nr.	Mean_doc_dates_diff	Doc_count	Days_after_last_doc	Mean_pay_days	Stan_SD_doc_amount	Mean_doc_amount	Default_doc_proc	Mean_pay_fact	Docs_late_proc	n
1	34,71	14,26	21,17	24,19	0,44	4543,43	0,01	-8,45	7,72	1585
3	177,23	1,70	201,00	13,49	0,13	5239,77	0,25	3,30	11,94	1593
2	44,39	13,66	79,80	36.95	0,48	11292,82	0,53	18,77	78,68	1679

Šių klasterių apibūdinimas:

- Gero kreditingumo vertinimo pirkėjai (1 klasteris). Labiausiai iš kitų, šie pirkėjai išsiskiria tuo, jog sąskaitas apmoka iki apmokėjimo termino ir yra pastovūs. Jie atlieka pirkimus vidutiniškai kas mėnesį ir todėl jų sąskaitų sumos yra net žemiau vidutinio sąskaitos dydžio.
- Vidutinio kreditingumo vertinimo pirkėjai (3 klasteris). Šie pirkėjai sąskaitas apmoka po apmokėjimo termino praėjus kelioms dienoms, perka retai ir jiems taikomas apmokėjimo atidėjimas yra trumpas. Tai gali būti indikacija, jog šie pirkėjai atsiskaito išankstiniais apmokėjimais.
- Blogo kreditingumo vertinimo pirkėjai (2 klasteris). Nors šioje grupėje esantys pirkėjai gali būti laikomi pastoviais ir reguliariais, jų apmokėjimai dažniausiai vėluoja iki mėnesio po termino. Tačiau jie perka didelėmis sumomis ir todėl galima spręsti, jog apmokėjimų vėlavimas dėl to yra toleruojamas.

Sudarytas pirkėjų vertinimas buvo koreguojamas, remiantis ekspertinėmis žiniomis, kurio buvo sukauptos bendradarbiavimo metu ir neatsispindi turimuose duomenyse. Vertinant pirkėjus buvo pasiremta klasterizavimo rezultatais ir įvertinimai pakeisti 1894 pirkėjams. Tai sudaro 38,99 proc. visų pirkėjų. Po pervertinimo, pirkėjų išsidėstymas pasikeitė (žr. 3.5.3 pav.) keliais aspektais. Pirmiausia klasių dinamika pasikeitė t.y. atsirado nežymus klasių disbalansas ir daugiausiai pirkėjų įgavo vidutinį kreditingumo vertinimą. Į šią grupę buvo pridėti 725 pirkėjai, kuriems buvo priskirtas geras kreditingumo vertinimas ir 592 pirkėjai, kuriems buvo priskirtas blogas kreditingumo vertinimas.

Toliau tyrime bus naudojami kreditingumo vertinimų angliški pavadinimai – *good, average* ir *bad*.



3.5.3 pav. Pirkėjų kreditingumo vertinimo grupių pakeitimas

Pažiūrėjus kaip pasikeitė grupių centrai (vertinant tuos pačius kintamuosius, kurie buvo naudojami klasterizavime) bendros tendencijos išliko nepasikeitę (žr. 3.5.2 lentelę), tik centrų reikšmės labiau atsiskyrė.

3.5.2 lentelė. Patikslintų pirkėjų grupių, pagal kreditingumo vertinimą, centrai

Nr.	Mean_ doc_ dates_ diff	Doc_ count	Days_ after_ last_doc	Mea_ pay_ days	Stan_ SD_ doc_ amount	Mean_ doc_ amount	Default_ doc_ proc	Mean_ pay_ fact	Docs_ late_ proc	n
1	41,77	15,91	48,13	36,17	0,469	8544.87	0	-8,44	17,37	1274
3	110,09	7,17	122,35	15,26	0,268	6746,98	0	-0,81	22,76	2642
2	72,13	9,63	109,64	37,68	0,418	6160,73	1,38	38,57	86,16	941

Tolimesniame tyrime yra naudojamas patikslintas pirkėjų įvertinimas ir pagal jį atliekamas klasifikavimas naudojant visus kiekybinius kintamuosius.

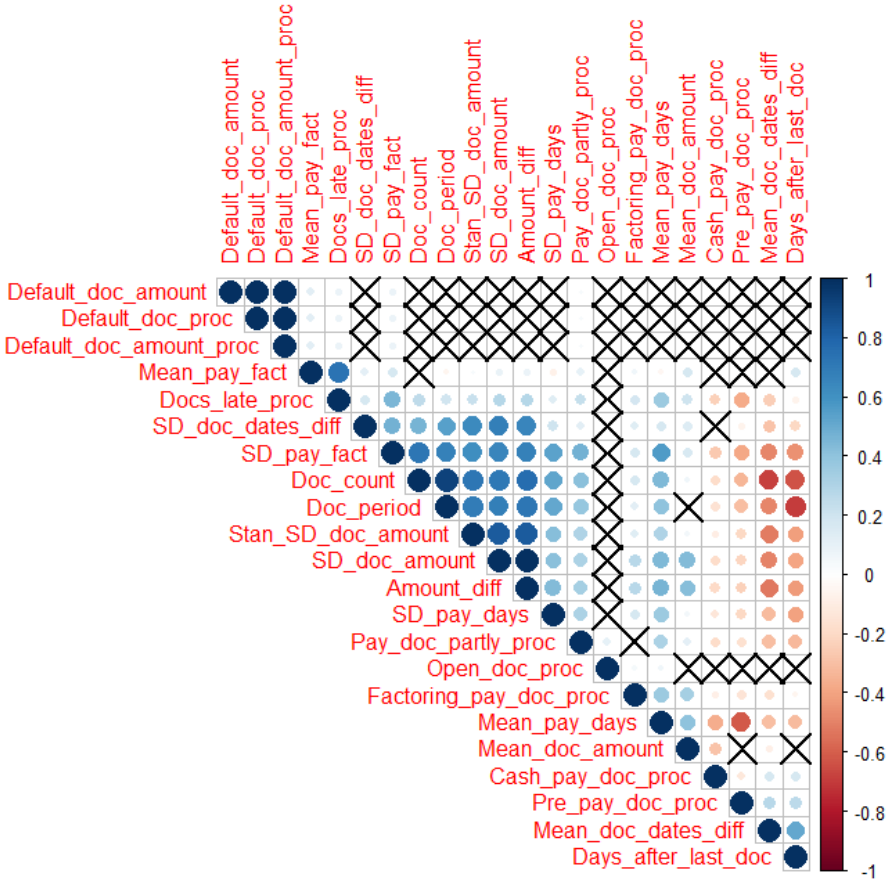
3.6. Klasifikavimas

3.6.1. Atraminių vektorių metodo realizacija

Klasifikavimui pirmiausia atlikti naudojamas *Atraminių vektorių metodas* ir jame galima naudoti tik kiekybinius kintamuosius. Pirmiausia pažiūrima kaip jie yra priklausomi nuo vienas kito (žr. 3.6.1.1 pav.) ir padaromos kelios išvados:

- *„Default_doc_proc“, „Default_doc_amount“* ir *„Default_doc_amount_proc“* kintamieji gali būti tarpusavyje keičiami, nes jiems paskaičiuotas Spearmano koreliacijos koeficientas yra lygus vienetui. Tokia pati situacija yra ir su kintamaisiais *„SD_doc_amount“* ir *„Amount_diff“*. Todėl modeliuose bus naudojami tik po vieną kintamąjį iš šių grupių: *„Default_doc_proc“* ir *„SD_doc_amount“;*

- *„Mean_pay_fact“* ir *„Docs_late_proc“*, bei *„Doc_count“*, *„Doc_period“*, *„Stan_SD_doc_amount“* ir *„SD_pay_fact“* sudaro dvi stipriai susijusių kintamųjų grupes, todėl į modeliuose bus įtraukiamos įvairios šių kintamųjų kombinacijos;
- Likę kintamieji gali būti laikomi silpnai susiję ir todėl į visus modelius bus įtraukiami.



3.6.1.1 pav. Kiekybinių kintamųjų tarpusavio priklausomybės

Sudėliojus visas galimas kintamųjų kombinacijas, buvo sudaryti 47 kintamųjų rinkiniai, kuriuose kintamųjų skaičius svyruoja nuo 13 iki 19 ir visiems jiems buvo pritaikytas *Atraminių Vektorių metodas su Radialine branduolio funkcija*. Visų modelių vertinimai (tikslumas, F1 kiekvienai klasei ir apibendrintas F1 rodiklis) pateikti 5 priede.

Geriausia kombinacija pagal sukurtus modelius susideda iš 18 kintamųjų: *„Default_doc_proc“*, *„SD_doc_amount“*, *„Mean_doc_dates_diff“*, *„SD_doc_dates_diff“*, *„Days_after_last_doc“*, *„Pre_pay_doc_proc“*, *„Cash_pay_doc_proc“*, *„Factoring_pay_doc_proc“*, *„Mean_pay_days“*, *„SD_pay_days“*, *„Mean_doc_amount“*, *„Pay_doc_partly_proc“*, *„Open_doc_proc“*, *„Docs_late_proc“*, *„Mean_pay_fact“*, *„Doc_count „,* *„Doc_period“*, *„SD_pay_fact“*. Pagal juos sukurtas modelis yra pirmas pagal tikslumą, modelio F1 rodiklį, F1 rodiklį *GOOD* ir *AVERAGE* klasėms. Ir 6 pagal gerumą pagal F1 rodiklį *BAD* klasei. Tačiau pastarasis nuo geriausio tik 0,006 ir modelis, kuris yra geriausias pagal minėtą rodiklį, pagal visus kitus yra 17 vietoje (žr. 3.6.1.1 lentelę). Geriausio modelio sumaišymo matrica pateikta 3.6.1.2 lentelėje.

3.6.1.1 lentelė. Geriausio modelio pagal, keturis rodiklius, palyginimas su geriausiu pagal vieną

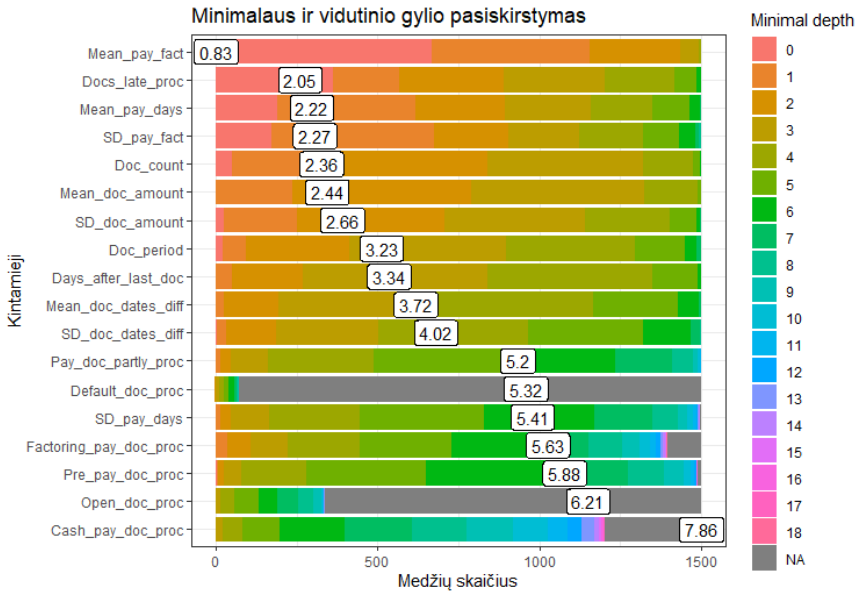
Modelio Nr.	Modelio Nr. pagal rodiklius	Tikslumas	F1_GOOD	F1_AVERAGE	F1_BAD	Modelio F1
34-19	1	0,825	0,742	0,850	0,866	0,819
13-16	17	0,801	0,699	0,824	0,871	0,798

3.6.1.2 lentelė. Geriausio SVM modelio sumaišymo matrica

	Prognozuojama reikšmė – G	Prognozuojama reikšmė – A	Prognozuojama reikšmė – B
Stebėta reikšmė – G	930	326	18
Stebėta reikšmė – A	277	2281	84
Stebėta reikšmė – B	26	119	796

3.6.2. Atsitiktinio miško metodo realizacija

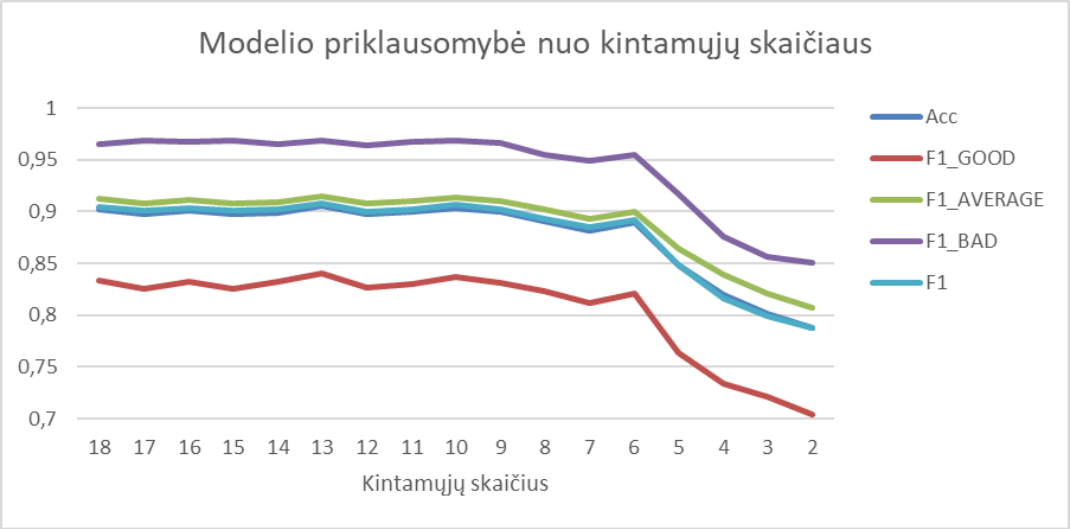
Toliau atrinktus kintamuosius naudojame ir *Atsitiktinio miško* klasifikatoriui sudaryti. Pirmiausia sudarome vieną modelį su visais kintamaisiais ir nustatome, kurie iš jų yra svarbiausi. Iš rezultatų (žr. 3.6.2.1 pav.) galima sakyti, jog pats svarbiausias kintamasis klasifikatorius yra kaip pirkėjas apmoka sąskaitas. Pagal gautus kintamųjų svarbumus, sudėliojame eilę, kaip jie yra išimamai iš modelio sudarymo. Kadangi turime šiokį tokį klasių disbalansą, šalia paprasto *Atsitiktinio miško*, dar naudojame ir dvi jo modifikacijas su mokymosi bauda ir klasių sulyginimu.



3.6.2.1 pav. Kintamųjų svarbumas sudarytuose Sprendimų medžiuose

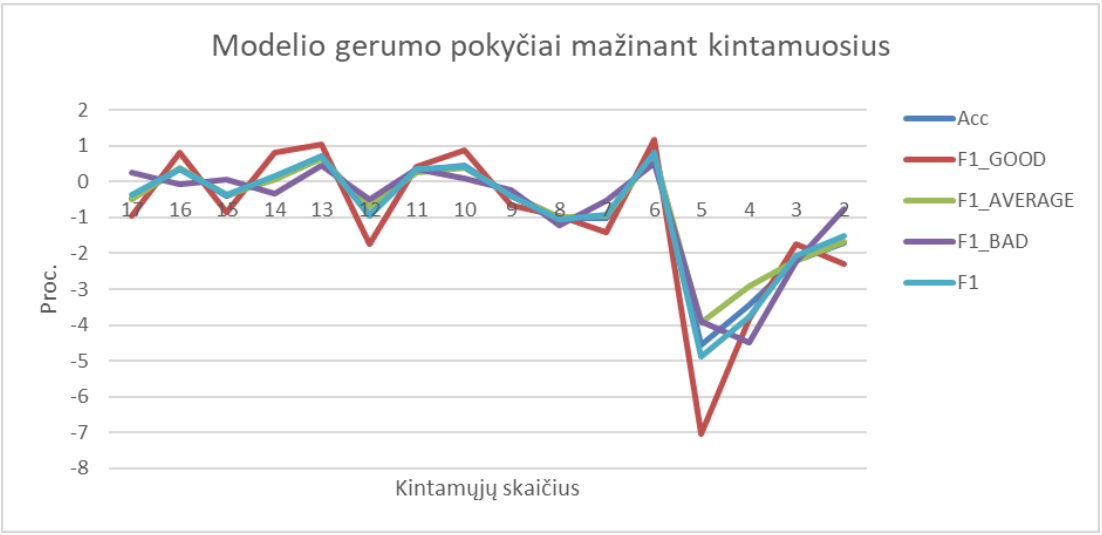
Palyginus visų modelių vertinimo rodiklius (vidų modelių rezultatai pateikti 6 priede), geriausi modeliai buvo sukurti naudojant paprastą *Atsitiktinio miško metodą*. Taip pat iš rodiklių palyginimo (žr. 3.6.2.2 pav.), jog modelio tikslumas nemažėja taip stipriai mažinant kintamųjų kiekį ir ryškesnis skirtumas pasirodo tik kai į modelį įtraukiami penki ir mažiau kintamųjų. Įdomūs rezultatai gauti klasių klasifikavimo lygmenyje. Geriausiai t.y. iki beveik 97 proc. tikslumu yra klasifikuojami blogo

kreditingumo vertinimo pirkėjai. Tai yra labai geras rezultatas, nes būtent šie pirkėjai gali atnešti daugiausia žalos.



3.6.2.2 pav. Modelio parametrų priklausomybė nuo kintamųjų kiekio

Jog būtų lengviau nuspręsti, kiek kintamųjų yra geriausiai įtraukti į modelį, apskaičiuojama modelių rodiklių skirtumus mažinant kintamųjų skaičių (žr. 3.6.2.3 pav.) ir tada ryškiai pasimato, jog su 6 kintamaisiais modelį sudaryti yra geriausiai, nes su daugiau kintamųjų rodikliai žymiai nesiskiria, o mažinant dar daugiau – pradeda pastebimai mažėti.



3.6.2.3 pav. Modelio parametrų pokyčių priklausomybė nuo kintamųjų kiekio

Šeši kintamieji, kurie įeina į modelį yra *„Mean_pay_days“, „Mean_doc_amount“, „Docs_late_proc“, „Mean_pay_fact“, „Doc_count“, „SD_pay_fact“* ir su jais sudaryto modelio sumaišymo matrica (žr. 3.6.2.1 lentelę) rodo, jog gauti rezultatai yra geresni negu taikant Atsitiktinių vektorių metodą.

3.6.2.1 lentelė. Atsitiktinio miško modelio su šešiais kintamaisiais sumaišymo matrica

	Prognozuojama reikšmė – G	Prognozuojama reikšmė – A	Prognozuojama reikšmė – B
Stebėta reikšmė – G	1062	212	0
Stebėta reikšmė – A	240	2364	38
Stebėta reikšmė – B	11	36	894

Apibendrinant galima pasakyti, jog 89 proc. tikslumu, galima įvertinti pirkėjo kreditingumą iš jo mokumo istorijos apskaičiavus:

- Vidutinę jam išrašytų pardavimų sąskaitų sumą ir kiek iš viso jam buvo išrašyta sąskaitų;
- Vidutinį jam taikomo apmokėjimo atidėjimo terminą dienomis;
- Vidutinį jo vėlavimą apmokėti sąskaitas dienomis, kaip šis skaičius varijuoja ir kokį procentą visų sąskaitų jis yra vėlavęs apmokėti.

Pirkėjų grupių vidutinės šių kintamųjų reikšmės pateiktos 3.6.2.2 lentelėje.

3.6.2.2 lentelė. Pirkėjų grupių centrai pagal į klasifikavimo modelį įtrauktus kintamuosius

Pirkėjų kreditingumo vertinimas	**Vidutinė sąskaitų suma, Eur**	**Sąskaitų kiekis, vnt.**	**Vidutinis apmokėjimo atidėjimas, d.**	**Vidutinis vėlavimas apmokėti sąskaitas, d.**	**Mokėjimų atlikimo nepastovumas, d.**	**Pavėluotai apmokėtų sąskaitų kiekis, proc.**
Geras	8544,87	15,91	36,17	-8,44	6,09	17,37
Vidutinis	6746,98	7,16	15,26	-0,81	3,89	22,77
Blogas	6160,73	9,63	37,68	38,57	21,01	86,16

Labiausiai joje krenta į akis, jog blogo kreditingumo vertinimo pirkėjai yra labai nepastovūs t.y. jų apmokėjimo įpročiai pastoviai keičiasi. Tai gali būti laikoma pirmąją indikacija, jog į pirkėją reikia atidžiau pažiūrėti.

4. Išvados

1. Atlikus literatūros analizę nustatyta, jog didmeninių pirkėjų kreditingumo vertinimas yra svarbus ne tik finansinėms įmonėms, kurių pagrindinė veikla yra kreditų išdavimas. Vertinimas turi būti atliekamas įvairaus verslo įmonėse, jeigu jų klientams yra taikomas apmokėjimo atidėjimas t.y. suteikiamas prekybos kreditas. Taip pat nustatyta, jog tokios įmonės turi informacijos apie pirkėjų mokumą, kuri gali būti panaudojama kreditingumo vertimui sudaryti ir jį nuolatos atnaujinti.
2. Atlikus įmonių klasterizavimą, naudojant K-vidurkių metodą, jos buvo suskirstytos į tris grupes – gero, vidutinio ir blogo kreditingumo klasterius. Netgi 61,01 proc. įmonių priskyrimas klasteriams, taikant šį metodą, eksperto buvo įvertintas kaip teisingai atliktas. Likusiųjų įmonių klasterizavimas buvo patikslintas.
3. Pritaikius du metodus, – Atraminių vektorių ir Atsitiktinio miško, buvo sudarytas klasifikavimo modelis su jo variacijomis: pirmąjį metodą pritaikius nustatyti, kurie kiekybiniai kintamieji galėtų būti įtraukti į modelį, o antrąjį – nustatyti jų svarbumas. Tuomet klasifikavimas buvo atliktas palikus tik svarbiausius kintamuosius. Remiantis sudarytų modeliu, buvo pasiektas iki 90 proc. klasifikavimo tikslumas.
4. Išanalizavus atlikto klasifikavimo rezultatus ir jo tikslumą, buvo gauta, jog, atsižvelgiant į įmonių mokumo informaciją, svarbiausiais galima laikyti šešis kintamuosius: sąskaitų kiekį, vidutinę jų sumą ir vidutinį jų apmokėjimo atidėjimo laikotarpį, vidutinį vėlavimą apmokėti sąskaitas ir kaip šis vidurkis kinta, bei pavėluotai apmokėtų sąskaitų kiekį.

Literatūros sąrašas

1. Cai, G., Chen, X. and Xiao, Z. The Roles of Bank and Trade Credits: Theoretical Analysis and Empirical Evidence. *Production and Operations Management*, 2014, 23 (4), 583-598. ISSN 1059-1478.
2. Palacín-Sánchez, M., Canto-Cuevas, F. and di-Pietro, F. Trade credit versus bank credit: a simultaneous analysis in European SMEs. *Small Business Economics*, 2019, 53, 1079-1096.
3. Andrieu G., Staglianò R. and Zwan P.W. van der, Bank debt and trade credit for SMEs in Europe: firm-, industry-, and country-level determinants, *Small Business Economics*, 2017, 51(1), 245-264.
4. Engemann M., Eck K. and Schnitzer M. Trade Credits and Bank Credits in International Trade: Substitutes or Complements? *The World Economy*, 2014, 37 (11), 1507-1540.
5. Aleknevičienė V. *Įmonės finansų valdymas*, Kaunas: Spalvų Kraitė, 2011, ISBN 9789955921042.
6. Fisman R. and Love I. Trade Credit, Financial Intermediary Development, and Industry Growth. *The Journal Of Finance*, 2003, 58 (1), 353-374.
7. Ferris J. S. A Transactions Theory of Trade Credit Use. *The Quarterly Journal of Economics*, 1981, 96 (2), 243-270
8. Petersen M. A. and Rajan R. G. Trade Credit: Theories And Evidence. *Review of Financial Studies*, 1997, 10 (3), 661-691.
9. Du J., Lu Y. and Tao Z. Bank loans vs. trade credit: Evidence from China. *Economics of Transition*, 2012, 20 (3), 457–480.
10. Asselbergh G. Strategic Approach on Organizing Accounts Receivable Management: Some Empirical Evidence. *Journal of Management and Governance*, 1999, 3, 1-29.
11. Scherr F. C. Optimal Trade Credit Limits. *Financial Management*, 1996, 25 (1), 71-85.
12. Eddy Y. L. and Bakar E. M. N. E. A. Credit scoring models: Techniques and issues. *Journal of Advanced Research in Business and Management Studies*, 2019, ISSN: 2462-1935.
13. Tripathia D., Edlaa D. R., Kuppilia V., Bablania A. and Dharavathb R. Credit Scoring Model based on Weighted Voting and Cluster based Feature Selection. *Procedia Computer Science*, 2018, 132, 22-31
14. Altman E. I., Sabato G. and Wilson N. The Value Of Non-Financial Information In SME Risk Management, *SSRN Electronic Journal*, 2008, 6 (2).
15. Carta S., Ferreira A., Reforgiato Recupero D., Saia M. and Saia R. A combined entropy-based approach for a proactive credit scoring. *Engineering Applications of Artificial Intelligence*, 2020, 87.
16. Chen K. H. and Shimerda T. A. An Empirical Analysis of Useful Financial Ratios. *Financial Management*, 1981, 10 (1), 51-60.
17. Kozodoi N., Lessmann S., Papakonstantinou K., Gatsoulis Y. and Baesens B. A multi-objective approach for profit-driven feature selection in credit Scoring. *Decision Support Systems*, 2019, 120, 106-117.
18. Pławiak P., Abdar M., Pławiak J., Makarenkov V. and Acharya R. DGHNL: A new deep genetic hierarchical network of learners for prediction of credit scoring. *Information Sciences*, 2020, 516, 401-418.
19. Koutanaei F. N., Sajedi H. And Khanbabaei M. A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring. *Journal of Retailing and Consumer Services*, 2015, 27, 11-23.

20. Tian Y., Yong Z. and Luo J. A new approach for reject inference in credit scoring using kernel-free fuzzy quadratic surface support vector machines. *Applied Soft Computing Journal*, 2018, 73, 96-105.
21. Danenas P. ir Garsva G. Selection of Support Vector Machines based classifiers for credit risk domain. *Expert Systems with Applications*, 2015, 42 (6), 3194-3204.
22. Maldonado S., Perez J. and Bravo C. Cost-based feature selection for Support Vector Machines – An application in credit scoring. *European Journal of Operational Research*, 2017, 261 (2), 656-665.
23. Long Chen F. and Chia Li F. Combination of feature selection approaches with SVM in credit scoring. *Expert Systems with Applications*, 2010, 37 (7), 4902-4909.
24. Danenas P., Garsva G. ir Gudas S. Credit Risk Evaluation Model Development Using Support Vector Based Classifiers. *Procedia Computer Science*, 2011, 4, 1699-1707.
25. Maldonado S., Bravo C., Lopez J., Perez J., Integrated framework for profit-based feature selection and SVM classification in credit scoring. *Decision Support Systems*, 2017, 104, 113-121.
26. Harris T. Credit scoring using the clustered support vector machine. *Expert Systems with Applications*, 2015, 42 (2), 741-750.
27. Yu l., Zhou R., Tang L. and Chen R. A DBN-based resampling SVM ensemble learning paradigm
28. for credit classification with imbalanced data. *Applied Soft Computing*, 2018, 69, 192-202.
29. Pławiak P., Abdar M., Pławiak J. and Acharya R. Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring. *Applied Soft Computing*, 2019, 84.
30. Zhong H., Miao C., Shen Z. and Feng Y. Comparing the learning effectiveness of BP, ELM, I-ELM, and SVM for corporate credit ratings. Neurocomputing, 2014, 128, 285-295.
31. Ping Y. and Yongheng L. Neighborhood rough set and SVM based hybrid credit scoring classifier. *Expert Systems with Applications*, 2011, 38(9), 11300-11304.
32. Abdou H. A., Dongmo Tsafack M. D., Ntim C. G., Baker R. D., Predicting creditworthiness in retail banking with limited scoring data. *Knowledge-Based Systems*, 2016, 103, 89-103.
33. Kanapickienė R. ir Špicas R. Credit Risk Assessment Model for Small and Micro-Enterprises: The Case of Lithuania. *Risks*, 2019 7 (2), 1-23, ISSN 2227-9091.
34. Chen H., Xiang Y., The Study of Credit Scoring Model Based on Group Lasso. *Procedia Computer Science*, 2017, 122, 677-684.
35. Papouskova M., Hajek P., Two-stage consumer credit risk modelling using heterogeneous ensemble. *Learning Decision Support Systems*, 2019, 118, 33-45.
36. Handhika T., Achmad Fahrurozi A., Zen R. I. M., Lestari D. P, Murni I. S., Modified Average of the Base-Level Models in the Hill-Climbing Bagged Ensemble Selection Algorithm for Credit Scoring, *Procedia Computer Science*, 2019, 157, 229-237.
37. Xiao J., Zhou X., Zhong Y., Xie L., Gu X. and Liu D. Cost-sensitive semi-supervised selective ensemble model for customer credit scoring. *Knowledge-Based Systems*, 2020, 189.
38. Hayashi Y. Application of a rule extraction algorithm family based on the Re-RX algorithm to financial credit risk assessment from a Pareto optimal perspective. *Operations Research Perspectives*, 2016, 3, 32-42.
39. Chia Chang Y., Hu Chang K., Hsuan Chu H. and Ing Tong L., Establishing decision tree-based short-term default credit risk assessment models. *Communications in Statistics - Theory and Methods*, 2016, 45 (23), 6803-6815, ISSN: 0361-0926.

40. He F. and Chen X. Credit networks and systemic risk of Chinese local financing platforms: Too central or too big to fail? – based on different credit correlations using hierarchical Methods. *Physica A: Statistical Mechanics and its Applications*, 2016, 461, 158-170.
41. Karan M. B., Ulucan A. and Kaya M. Credit risk estimation using payment history data: a comparative study of Turkish retail stores. *Central European Journal of Operations Research*, 2012, 21 (2).
42. Omran M., Engelbrecht A. and Salman. K. A. An overview of clustering methods. *Intelligent Data Analysis*, 2007, 11(6), 583-605.
43. Han J., Kamber M. and Pei J. *Data Mining: Concepts and Techniques*, 3rd ed. Elsevier Science & Technology, 2011, ISBN 9780123814807

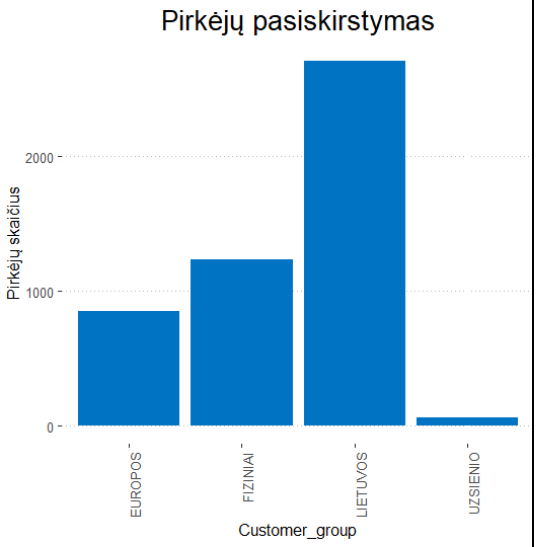
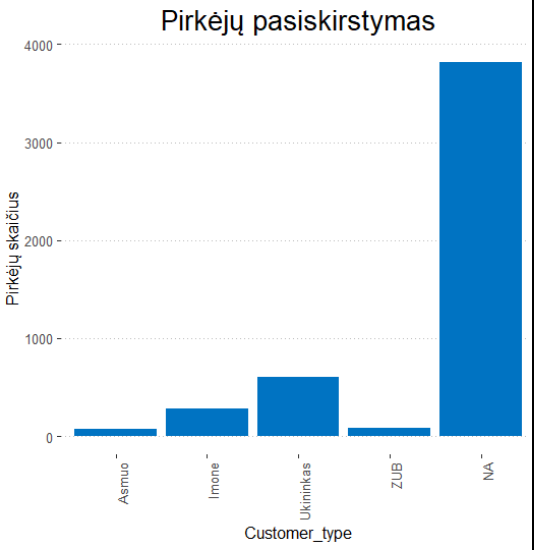
Priedai

1 priedas. Skaičiavimuose naudojamų kintamųjų apibūdinimas

Pavadinimas duomenų rinkinyje	**Lietuviškas pavadinimas**	**Aprašymas**
Customer_no	Kliento numeris	Skirtas identifikuoti kiekvieną klientą
Customer_group	Pirkėjo registracijos grupė	Vardų skalės kintamasis su keturiomis reikšmėmis, kurios identifikuoja kokiai registracijos grupei priklauso klientas
Customer_type	Pirkėjo tipas	Vardų skalės kintamasis, kuris nurodo kliento tipą
Country	Šalis	Vardų skalės kintamasis, nurodantis pirkėjo šalies kodą
Purch_from_multi	Pirkimai iš kelių įmonių	Binarinis kintamasis. Reikšmės: 0 – bendradarbiauja su viena įmone, 1 – bendrauja su keliomis įmonėmis
Purch_multi_product	Pirkimai skirtingų prekių	Binarinis kintamasis. Reikšmės: 0 – perka tik vienos pramonės šakos prekes, 1 – perka kelių pramonės šakų prekes
Doc_count	Dokumentų kiekis	Kiekybinis kintamasis. Reikšmės yra sveikieji skaičiai, kurie nurodo kiek kiekvienam klientui yra išrašyta sąskaitų
Mean_doc_dates_diff	Vidutinis tarpas tarp sąskaitų	Kiekybinis kintamasis. Reikšmės yra sveikieji skaičiai ir nurodo koks vidutinis tarpas tarp sąskaitų. Jeigu reikšmės nėra, klientui buvo išrašyta tik viena sąskaita
SD_doc_dates_diff	Standartinis nuokrypis tarp sąskaitų	Kiekybinis kintamasis. Reikšmės sveikieji skaičiai ir nurodo kaip varijuoja sąskaitų išrašymų tarpai
Doc_period	Dokumentų rašymo periodas	Kiekybinis kintamasis. Reikšmės yra sveikieji skaičiai, kurie parodo kiek dienų buvo bendrauta su klientu
Days_after_last_doc	Dienų skaičius nuo paskutinės sąskaitos	Kiekybinis kintamasis. Reikšmės yra sveikieji skaičiai, kurie parodo kiek dienų praėjo nuo paskutinės sąskaitos (skaičiuojant nuo 2018 metų pabaigos)
Pre_pay_doc_proc	Iš anksto apmokėtų dokumentų procentas	Kiekybinis kintamasis. Reikšmės yra dviejų šimtųjų tikslumu ir nurodo kiek procentų sąskaitų klientas apmoka išankstiniu apmokėjimu
Cash_pay_doc_proc	Grynais pinigais apmokėtų dokumentų procentas	Kiekybinis kintamasis. Reikšmės yra dviejų šimtųjų tikslumu ir nurodo kiek procentų sąskaitų klientas apmoka grynais
Factoring_pay_doc_proc	Faktoringo sąskaitų procentas	Kiekybinis kintamasis. Reikšmės yra dviejų šimtųjų tikslumu ir nurodo kiek procentų kliento sąskaitų yra pagal faktoringo sutartį
Mean_pay_days	Vidutinis apmokėjimo atidėjimo laikas	Kiekybinis kintamasis. Reikšmės yra sveikieji skaičiai ir nurodo koks yra vidutinis kliento sąskaitų apmokėjimo atidėjimo periodas
SD_pay_days	Standartinis nuokrypis tarp apmokėjimo atidėjimo laikų	Kiekybinis kintamasis. Reikšmės yra sveikieji skaičiai ir nurodo kaip varijuoja sąskaitų apmokėjimo atidėjimo periodai
Mean_doc_amount	Vidutinė sąskaitos suma	Kiekybinis kintamasis. Reikšmės yra dviejų šimtųjų tikslumu ir nurodo vidutinę kliento sąskaitų sumą
SD_doc_amount	Standartinis nuokrypis tarp sąskaitų sumos	Kiekybinis kintamasis. Reikšmės yra dviejų šimtųjų tikslumu ir nurodo kiek varijuoja klientui išrašomų sąskaitų suma
Stan_SD_doc_amount	Sąskaitų sumos variacijos koeficientas	Kiekybinis kintamasis. Reikšmės yra dviejų šimtųjų tikslumu ir nurodo kiek varijuoja klientui išrašomų sąskaitų suma atsižvelgiant į sąskaitų sumą

Amount_diff	Sąskaitų sumų intervalas	Kiekybinis kintamasis. Reikšmės yra dviejų šimtųjų tikslumu ir nurodo skirtumą tarp didžiausios ir mažiausios sumų sąskaitų
Mean_pay_fact	Vidutinis apmokėjimo vėlavimas	Kiekybinis kintamasis. Reikšmės yra sveikieji skaičiai, kurie nurodo kiek vidutiniškai klientas vėlavo apmokėti sąskaitas
SD_pay_fact	Standartinis nuokrypis nuo vidutinio vėlavimo	Kiekybinis kintamasis. Reikšmės yra sveikieji skaičiai, kurie nurodo kiek varijuoja kliento vėlavimas apmokėti sąskaitas
Docs_late_proc	Vėluotų mokėjimų procentas	Kiekybinis kintamasis. Reikšmės yra dviejų šimtųjų tikslumu ir nurodo kiek procentų sąskaitų klientas vėlavo apmokėti
Pay_doc_partly_proc	Dalinai apmokėtų sąskaitų procentas	Kiekybinis kintamasis. Reikšmės yra dviejų šimtųjų tikslumu ir nurodo kiek sąskaitų klientas apmokėjo per kelis kartus
Open_doc_proc	Atvirų dokumentų procentas	Kiekybinis kintamasis. Reikšmės yra dviejų šimtųjų tikslumu ir nurodo kiek klientas turi dar neapmokėtų sąskaitų
Default_doc_proc	Nurašytų dokumentų procentas	Kiekybinis kintamasis. Reikšmės yra dviejų šimtųjų tikslumu ir nurodo kiek nurašyti dokumentai sudaro visų dokumentų
Default_doc_amount	Nurašytos skolos dydis	Kiekybinis kintamasis. Reikšmės yra dviejų šimtųjų tikslumu ir nurodo kokios sumos sąskaitos buvo nurašytos kaip beviltiškos
Deafult_doc_amount_proc	Nurašytos skolos dydis nuo visos sąskaitų sumos	Kiekybinis kintamasis. Reikšmės yra dviejų šimtųjų tikslumu ir nurodo kokia dalis pirkėjo sąskaitų sumos buvo nurašyta procentais

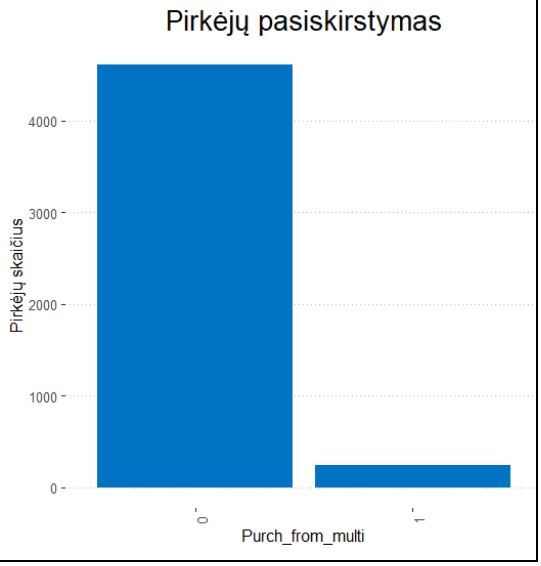
2 priedas. Pirkėją apibūdinančių atributų pagrindinės charakteristikos

<table>
<tr><th>Pavadinimas</th><th>Charakteristikos</th><th>Pasiskirstymas</th></tr>
<tr><td colspan="3">Pirmoji grupė – pirkėją apibūdinantys atributai</td></tr>
<tr><td>Customer_group</td><td>
<table>
<tr><th>Grupės</th><th>Kiekis, vnt.</th><th>Kiekis, proc.</th></tr>
<tr><td>EUROPOS</td><td>855</td><td>17,6</td></tr>
<tr><td>FIZINIAI</td><td>1237</td><td>25,5</td></tr>
<tr><td>LIETUVOS</td><td>2705</td><td>55,7</td></tr>
<tr><td>UZSIENIO</td><td>60</td><td>1,2</td></tr>
</table>
</td><td> 
</td></tr>
<tr><td>Customer_type</td><td>
<table>
<tr><th>Grupės</th><th>Kiekis, vnt.</th><th>Kiekis, proc.</th></tr>
<tr><td>Asmuo</td><td>74</td><td>1,5</td></tr>
<tr><td>Įmone</td><td>281</td><td>5,8</td></tr>
<tr><td>Ukininkas</td><td>599</td><td>12,3</td></tr>
<tr><td>ZUB</td><td>83</td><td>1,7</td></tr>
<tr><td>NA</td><td>3820</td><td>78,7</td></tr>
</table>
</td><td> 
</td></tr>
<tr><td>Country</td><td>
<table>
<tr><th>Grupės</th><th>Kiekis, vnt.</th><th>Kiekis, proc.</th></tr>
<tr><td>LT</td><td>3941</td><td>81,14</td></tr>
<tr><td>LV</td><td>160</td><td>3,29</td></tr>
<tr><td>IT</td><td>152</td><td>3,13</td></tr>
<tr><td>PL</td><td>78</td><td>1,61</td></tr>
<tr><td>EE</td><td>75</td><td>1,54</td></tr>
<tr><td>RO</td><td>61</td><td>1,26</td></tr>
<tr><td>DE</td><td>54</td><td>1,11</td></tr>
<tr><td>BY</td><td>37</td><td>0,76</td></tr>
<tr><td>CZ</td><td>33</td><td>0,68</td></tr>
<tr><td>HU</td><td>33</td><td>0,68</td></tr>
<tr><td>FR</td><td>31</td><td>0,64</td></tr>
</table>
</td><td> 
</td></tr>
</table>

ES	30	0,62	
FI	30	0,62	
HR	30	0,62	
BG	22	0,45	
SK	12	0,25	
DK	10	0,21	
GR	9	0,19	
SE	9	0,19	
SI	8	0,16	
NL	6	0,12	
NO	5	0,10	
RU	5	0,10	
AT	3	0,06	
GB	3	0,06	
IE	3	0,06	
UA	3	0,06	
BA	2	0,04	
BE	2	0,04	
EG	2	0,04	
RS	2	0,04	
CH	1	0,02	
KR	1	0,02	
ME	1	0,02	
PT	1	0,02	
TM	1	0,02	
UK	1	0,02	

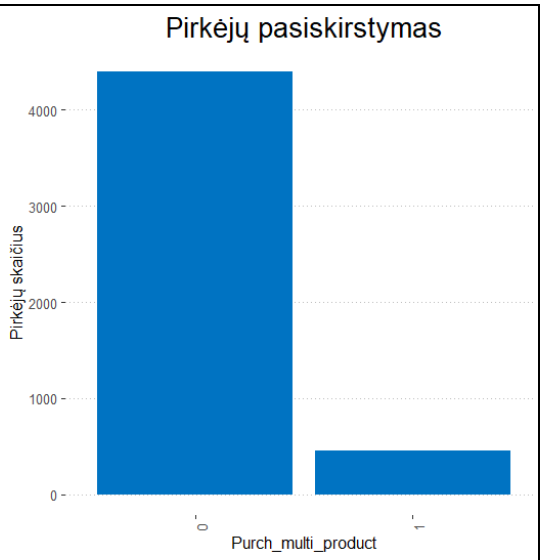
Purch_from_multi

Grupės	**Kiekis, vnt.**	**Kiekis, proc.**
0	4612	95,0
1	245	5,0



Purch_multi_product

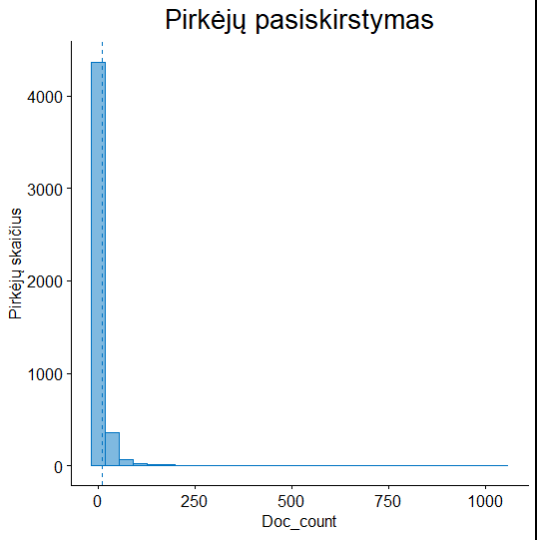
Grupės	Kiekis, vnt.	Kiekis, proc.
0	4397	90,5
1	460	9,5



Antroji grupė – pirkėjams išrašytų sąskaitų dažnumą apibūdinantys atributai

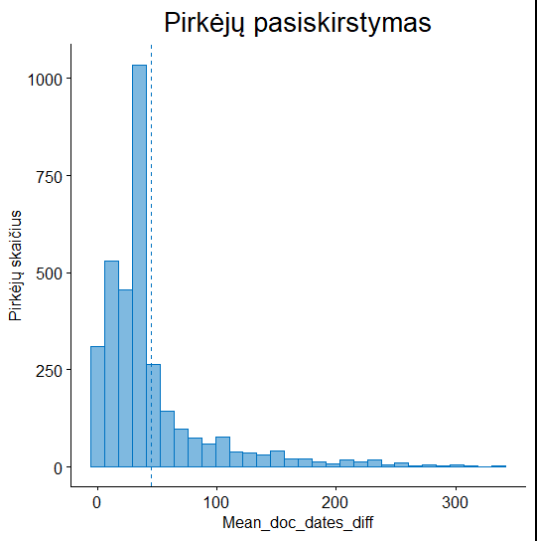
Doc_count

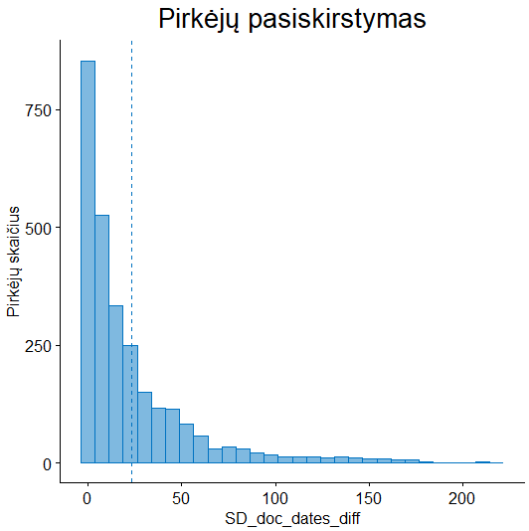
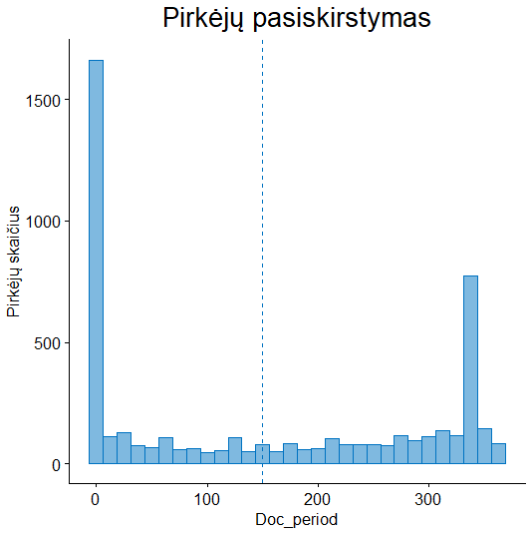
Reikšmių kiekis, Vnt.	4857
Nulinės reikšmės, Vnt.	0
Tuščios reikšmės, Vnt.	0
Minimali reikšmė	1,00
Maksimali reikšmė	1041,00
Mediana	3,00
Vidurkis	9,94
Standartinė vidurkio paklaida	0,45
Vidurkio 0,95 pasikliovimo intervalas	0,89
Dispersija	993,58
Standartinis nuokrypis	31,52
Variacijos koeficientas	3,17



Mean_doc_dates_diff

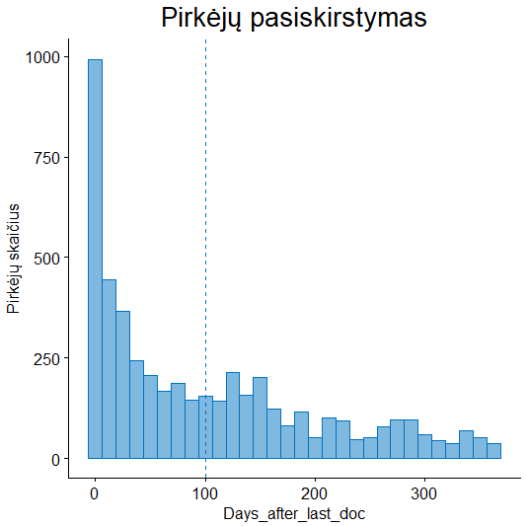
Reikšmių kiekis, Vnt.	3345
Nulinės reikšmės, Vnt.	48
Tuščios reikšmės, Vnt.	1512
Minimali reikšmė	0,00
Maksimali reikšmė	336,00
Mediana	30,00
Vidurkis	44,42
Standartinė vidurkio paklaida	0,85
Vidurkio 0,95 pasikliovimo intervalas	1,67
Dispersija	2415,88



	Standartinis nuokrypis	49,15
	Variacijos koeficientas	1,11
SD_doc_dates_diff	Reikšmių kiekis, Vnt.	2721
	Nulinės reikšmės, Vnt.	21
	Tuščios reikšmės, Vnt.	2136
	Minimali reikšmė	0,00
	Maksimali reikšmė	218,00
	Mediana	11,00
	Vidurkis	23,15
	Standartinė vidurkio paklaida	0,60
	Vidurkio 0,95 pasikliovimo intervalas	1,18
	Dispersija	983,60
	Standartinis nuokrypis	31,36
	Variacijos koeficientas	1,36
		
Doc_period	Reikšmių kiekis, Vnt.	4857
	Nulinės reikšmės, Vnt.	1553
	Tuščios reikšmės, Vnt.	0
	Minimali reikšmė	0,00
	Maksimali reikšmė	363,00
	Mediana	122,00
	Vidurkis	149,80
	Standartinė vidurkio paklaida	2,04
	Vidurkio 0,95 pasikliovimo intervalas	3,99
	Dispersija	20114,22
	Standartinis nuokrypis	141,82
	Variacijos koeficientas	0,95
		

Days_after_last_doc

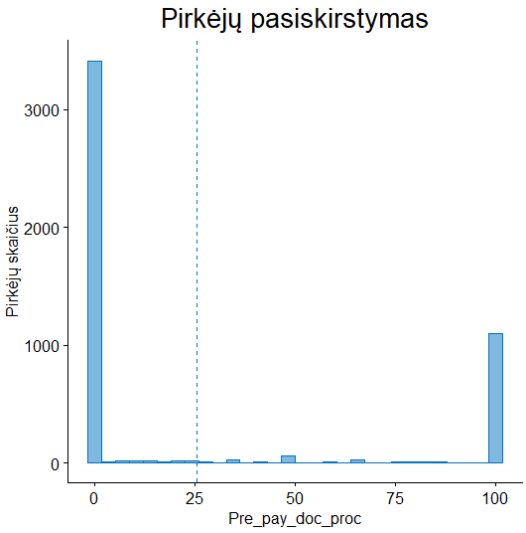
Reikšmių kiekis, Vnt.	4857
Nulinės reikšmės, Vnt.	882
Tuščios reikšmės, Vnt.	0
Minimali reikšmė	0,00
Maksimali reikšmė	363,00
Mediana	69,00
Vidurkis	100,42
Standartinė vidurkio paklaida	1,45
Vidurkio 0,95 pasikliovimo intervalas	2,84
Dispersija	10193,24
Standartinis nuokrypis	100,96
Variacijos koeficientas	1,01



Trečioji grupė – pirkėjams taikomas apmokėjimo sąlygas apibūdinantys atributai

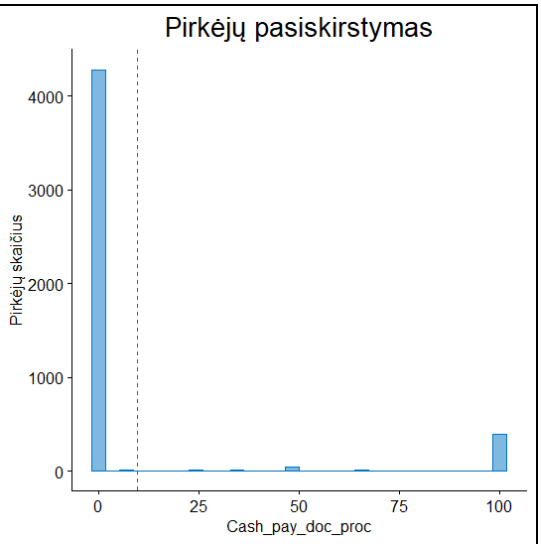
Pre_pay_doc_proc

Reikšmių kiekis, Vnt.	4857
Nulinės reikšmės, Vnt.	3442
Tuščios reikšmės, Vnt.	0
Minimali reikšmė	0,00
Maksimali reikšmė	100,00
Mediana	0,00
Vidurkis	25,17
Standartinė vidurkio paklaida	0,60
Vidurkio 0,95 pasikliovimo intervalas	1,18
Dispersija	1768,14
Standartinis nuokrypis	42,15
Variacijos koeficientas	1,67



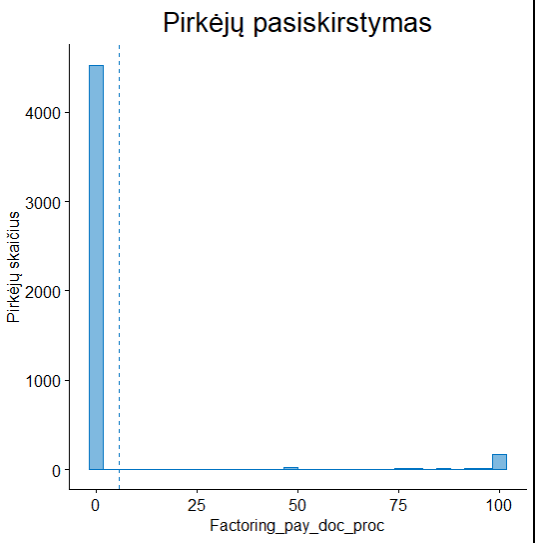
Cash_pay_doc_proc

Reikšmių kiekis, Vnt.	4857
Nulinės reikšmės, Vnt.	4320
Tuščios reikšmės, Vnt.	0
Minimali reikšmė	0,00
Maksimali reikšmė	100,00
Mediana	0,00
Vidurkis	9,51
Standartinė vidurkio paklaida	0,41
Vidurkio 0,95 pasikliovimo intervalas	0,80
Dispersija	805,04
Standartinis nuokrypis	28,47
Variacijos koeficientas	2,98



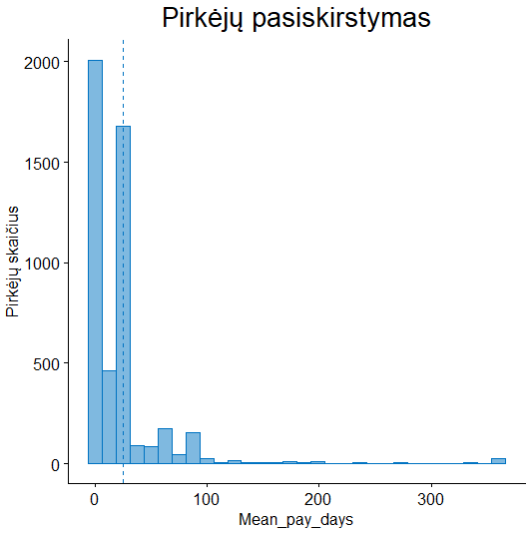
Factoring_pay_doc_proc

Reikšmių kiekis, Vnt.	4857
Nulinės reikšmės, Vnt.	4526
Tuščios reikšmės, Vnt.	0
Minimali reikšmė	0,00
Maksimali reikšmė	100,00
Mediana	0,00
Vidurkis	5,58
Standartinė vidurkio paklaida	0,31
Vidurkio 0,95 pasikliovimo intervalas	0,61
Dispersija	471,54
Standartinis nuokrypis	21,72
Variacijos koeficientas	3,89



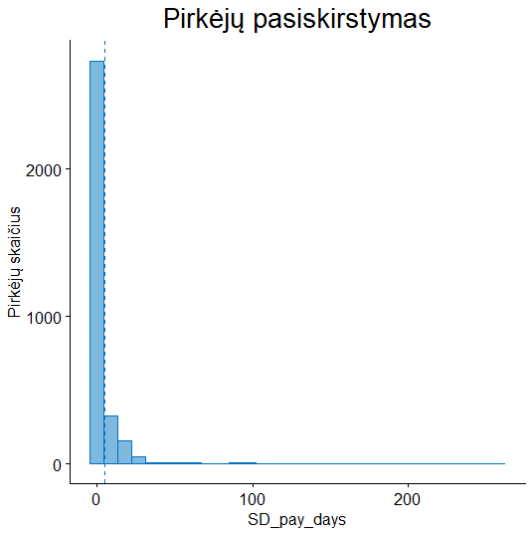
Mean_pay_days

Reikšmių kiekis, Vnt.	4857
Nulinės reikšmės, Vnt.	1806
Tuščios reikšmės, Vnt.	0
Minimali reikšmė	0
Maksimali reikšmė	360,00
Mediana	16,00
Vidurkis	25,09
Standartinė vidurkio paklaida	0,61
Vidurkio 0,95 pasikliovimo intervalas	1,20
Dispersija	1811,73
Standartinis nuokrypis	42,56
Variacijos koeficientas	1,70



SD_pay_days

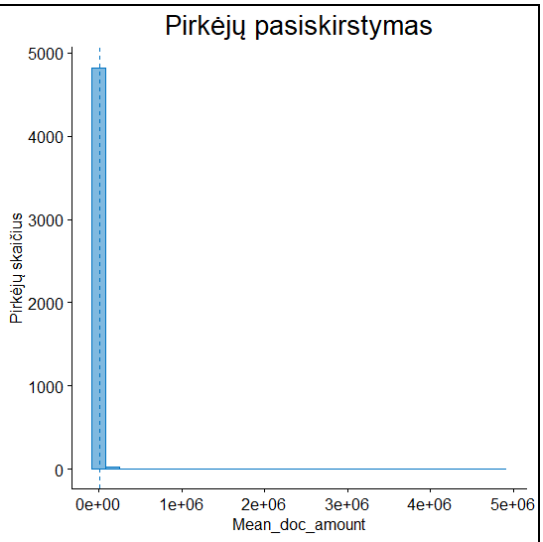
Reikšmių kiekis, Vnt.	3345
Nulinės reikšmės, Vnt.	1979
Tuščios reikšmės, Vnt.	1512
Minimali reikšmė	0,00
Maksimali reikšmė	258,00
Mediana	0,00
Vidurkis	4,59
Standartinė vidurkio paklaida	0,31
Vidurkio 0,95 pasikliovimo intervalas	0,60
Dispersija	318,08
Standartinis nuokrypis	17,83
Variacijos koeficientas	3,88



Ketvirtoji grupė – pirkėjams išrašytų sąskaitų sumas apibūdinantys atributai

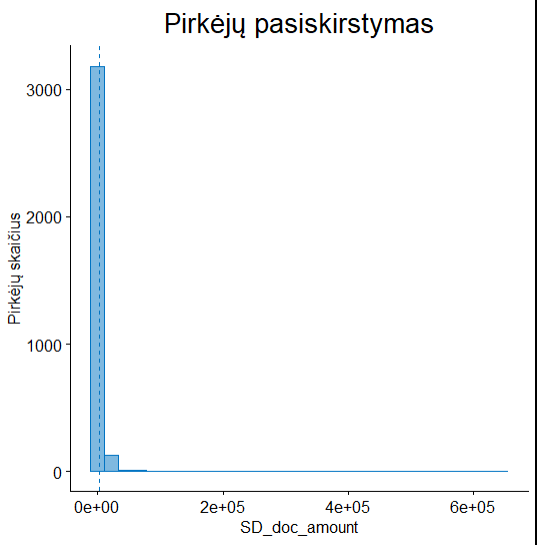
Mean_doc_amount

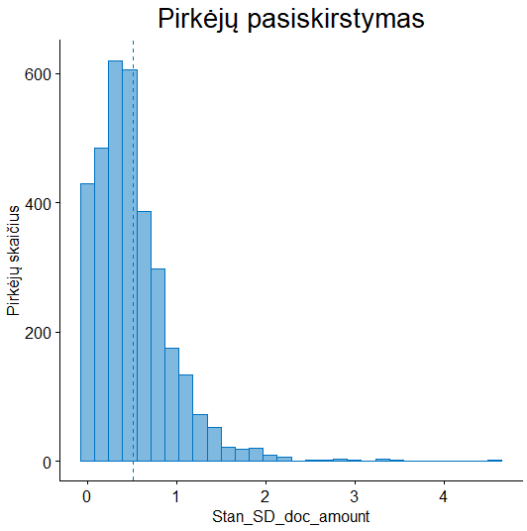
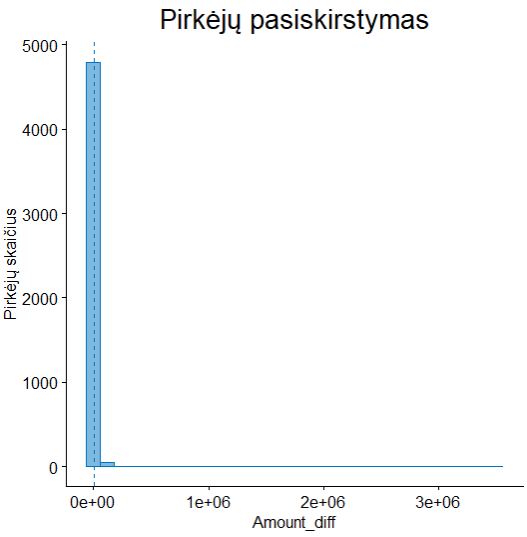
Reikšmių kiekis, Vnt.	4857
Nulinės reikšmės, Vnt.	0
Tuščios reikšmės, Vnt.	0
Minimali reikšmė	0,30
Maksimali reikšmė	4830500,00
Mediana	674,96
Vidurkis	7104,99
Standartinė vidurkio paklaida	1234,05
Vidurkio 0,95 pasikliovimo intervalas	2419,29
Dispersija	7396605000
Standartinis nuokrypis	86003,51
Variacijos koeficientas	12,10



SD_doc_amount

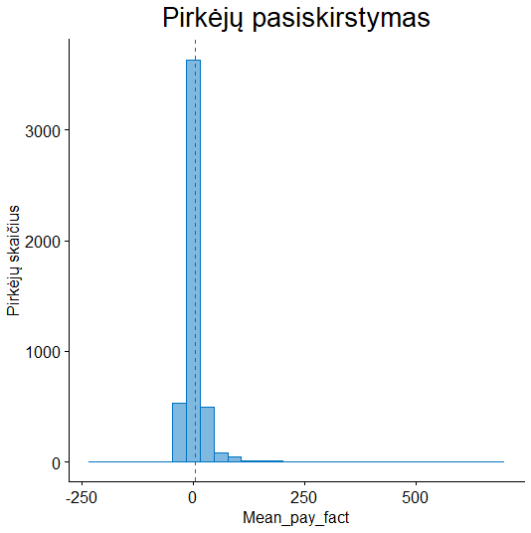
Reikšmių kiekis, Vnt.	3345
Nulinės reikšmės, Vnt.	112
Tuščios reikšmės, Vnt.	1512
Minimali reikšmė	0,00
Maksimali reikšmė	644872,90
Mediana	227,14
Vidurkis	3001,17
Standartinė vidurkio paklaida	358,44
Vidurkio 0,95 pasikliovimo intervalas	702,77
Dispersija	429751200,00
Standartinis nuokrypis	20730,44
Variacijos koeficientas	6,91



<table>
<tr>
<td>Stan_SD_doc_amount</td>
<td>
<table>
<tr><td>Reikšmių kiekis, Vnt.</td><td>3345</td></tr>
<tr><td>Nulinės reikšmės, Vnt.</td><td>138</td></tr>
<tr><td>Tuščios reikšmės, Vnt.</td><td>1512</td></tr>
<tr><td>Minimali reikšmė</td><td>0,00</td></tr>
<tr><td>Maksimali reikšmė</td><td>4,57</td></tr>
<tr><td>Mediana</td><td>0,43</td></tr>
<tr><td>Vidurkis</td><td>0,51</td></tr>
<tr><td>Standartinė vidurkio paklaida</td><td>0,01</td></tr>
<tr><td>Vidurkio 0,95 pasikliovimo intervalas</td><td>0,01</td></tr>
<tr><td>Dispersija</td><td>0,18</td></tr>
<tr><td>Standartinis nuokrypis</td><td>0,42</td></tr>
<tr><td>Variacijos koeficientas</td><td>0,84</td></tr>
</table>
</td>
<td> 
</td>
</tr>
<tr>
<td>Amount_diff</td>
<td>
<table>
<tr><td>Reikšmių kiekis, Vnt.</td><td>4857</td></tr>
<tr><td>Nulinės reikšmės, Vnt.</td><td>1624</td></tr>
<tr><td>Tuščios reikšmės, Vnt.</td><td>0</td></tr>
<tr><td>Minimali reikšmė</td><td>0,00</td></tr>
<tr><td>Maksimali reikšmė</td><td>3497601</td></tr>
<tr><td>Mediana</td><td>161,41</td></tr>
<tr><td>Vidurkis</td><td>6252,44</td></tr>
<tr><td>Standartinė vidurkio paklaida</td><td>977,54</td></tr>
<tr><td>Vidurkio 0,95 pasikliovimo intervalas</td><td>1916,42</td></tr>
<tr><td>Dispersija</td><td>4641269000</td></tr>
<tr><td>Standartinis nuokrypis</td><td>68126,86</td></tr>
<tr><td>Variacijos koeficientas</td><td>10,90</td></tr>
</table>
</td>
<td> 
</td>
</tr>
<tr>
<td colspan="3">Penktoji grupė – pirkėjų faktinio apmokėjimo tendencijas apibūdinantys atributai</td>
</tr>
</table>

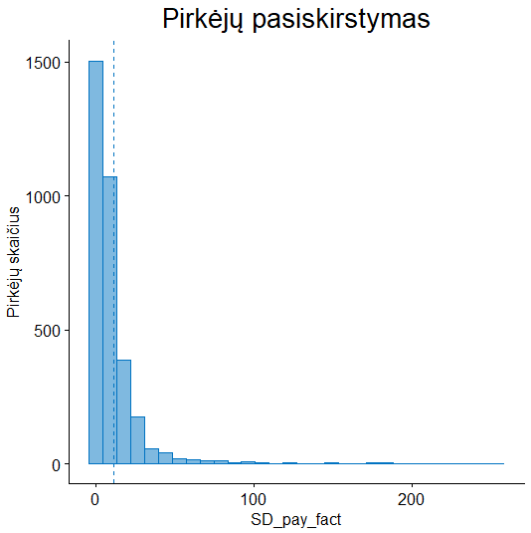
Mean_pay_fact

Reikšmių kiekis, Vnt.	4850
Nulinės reikšmės, Vnt.	1573
Tuščios reikšmės, Vnt.	7
Minimali reikšmė	-215,00
Maksimali reikšmė	685,00
Mediana	0,00
Vidurkis	4,76
Standartinė vidurkio paklaida	0,51
Vidurkio 0,95 pasikliovimo intervalas	1,00
Dispersija	1253,46
Standartinis nuokrypis	35,40
Variacijos koeficientas	7,43



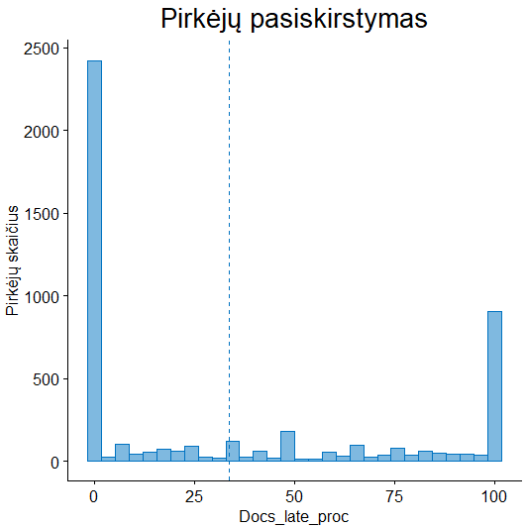
SD_pay_fact

Reikšmių kiekis, Vnt.	3343
Nulinės reikšmės, Vnt.	600
Tuščios reikšmės, Vnt.	1514
Minimali reikšmė	0,00
Maksimali reikšmė	254
Mediana	5,00
Vidurkis	11,32
Standartinė vidurkio paklaida	0,38
Vidurkio 0,95 pasikliovimo intervalas	0,75
Dispersija	494,72
Standartinis nuokrypis	22,24
Variacijos koeficientas	1,97



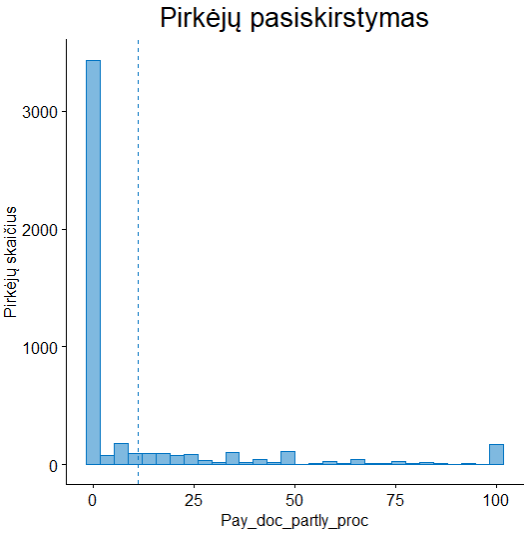
Docs_late_proc

Reikšmių kiekis, Vnt.	4850
Nulinės reikšmės, Vnt.	2421
Tuščios reikšmės, Vnt.	7
Minimali reikšmė	0,00
Maksimali reikšmė	100,00
Mediana	2,04
Vidurkis	33,64
Standartinė vidurkio paklaida	0,59
Vidurkio 0,95 pasikliovimo intervalas	1,15
Dispersija	1681,32
Standartinis nuokrypis	41,00
Variacijos koeficientas	1,22



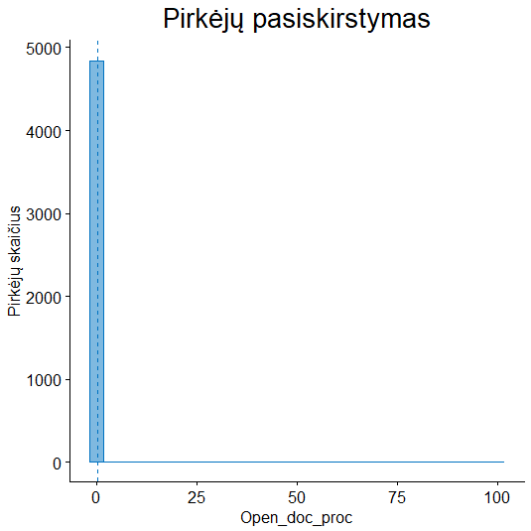
Pay_doc_partly_proc

Reikšmių kiekis, Vnt.	4841
Nulinės reikšmės, Vnt.	3413
Tuščios reikšmės, Vnt.	16
Minimali reikšmė	0,00
Maksimali reikšmė	100,00
Mediana	0,00
Vidurkis	10,94
Standartinė vidurkio paklaida	0,34
Vidurkio 0,95 pasikliovimo intervalas	0,67
Dispersija	572,59
Standartinis nuokrypis	23,93
Variacijos koeficientas	2,19



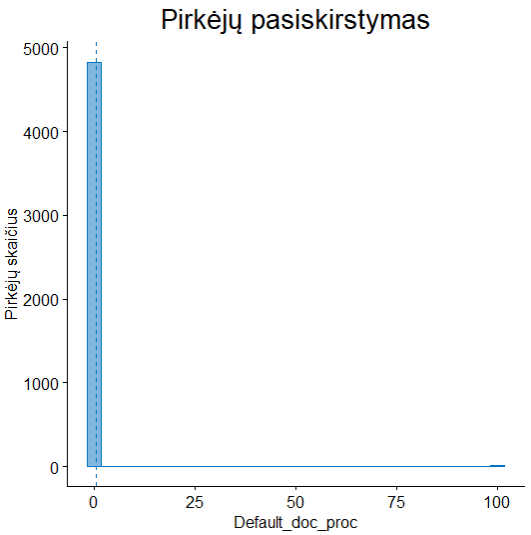
Open_doc_proc

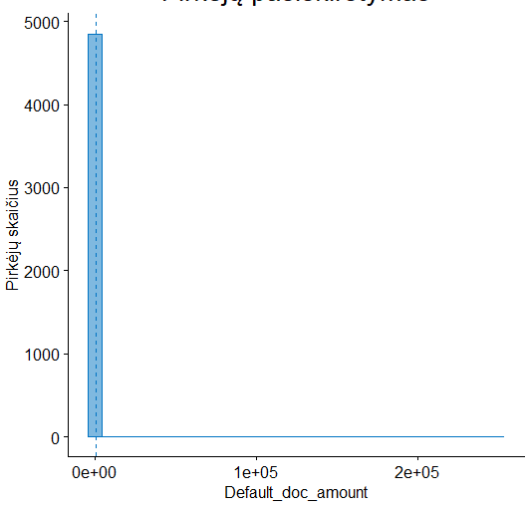
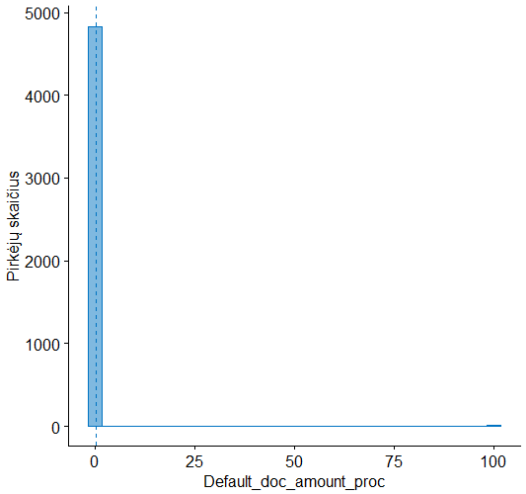
Reikšmių kiekis, Vnt.	4857
Nulinės reikšmės, Vnt.	4842
Tuščios reikšmės, Vnt.	0
Minimali reikšmė	0,00
Maksimali reikšmė	100,00
Mediana	0,00
Vidurkis	0,20
Standartinė vidurkio paklaida	0,06
Vidurkio 0,95 pasikliovimo intervalas	0,12
Dispersija	16,73
Standartinis nuokrypis	4,09
Variacijos koeficientas	20,67



Default_doc_proc

Reikšmių kiekis, Vnt.	4857
Nulinės reikšmės, Vnt.	4830
Tuščios reikšmės, Vnt.	0
Minimali reikšmė	0,00
Maksimali reikšmė	100,00
Mediana	0,00
Vidurkis	0,27
Standartinė vidurkio paklaida	0,07
Vidurkio 0,95 pasikliovimo intervalas	0,13
Dispersija	21,29
Standartinis nuokrypis	4,61
Variacijos koeficientas	17,28



Default_doc_amount		Pirkėjų pasiskirstymas 
Default_doc_amount_proc		Pirkėjų pasiskirstymas 

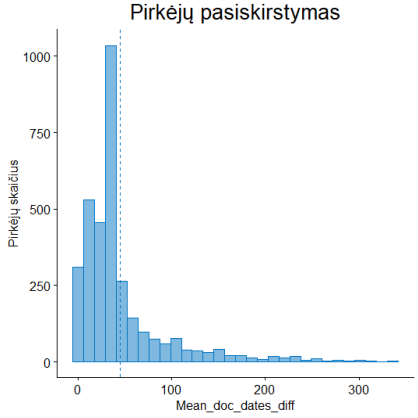
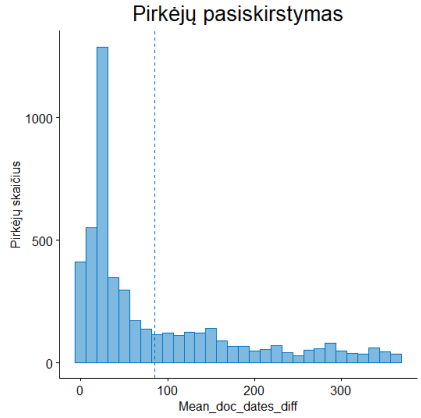
Reikšmių kiekis, Vnt.	4857
Nulinės reikšmės, Vnt.	4830
Tuščios reikšmės, Vnt.	0
Minimali reikšmė	0,00
Maksimali reikšmė	249377,77
Mediana	0,00
Vidurkis	109,10
Standartinė vidurkio paklaida	62,79
Vidurkio 0,95 pasikliovimo intervalas	123,10
Dispersija	19149800
Standartinis nuokrypis	4376,05
Variacijos koeficientas	40,11

Reikšmių kiekis, Vnt.	4857
Nulinės reikšmės, Vnt.	4830
Tuščios reikšmės, Vnt.	0
Minimali reikšmė	0,00
Maksimali reikšmė	100,00
Mediana	0,00
Vidurkis	0,26
Standartinė vidurkio paklaida	0,07
Vidurkio 0,95 pasikliovimo intervalas	0,13
Dispersija	21,19
Standartinis nuokrypis	4,60
Variacijos koeficientas	18,00

3 priedas. Kintamųjų trūkstamų reikšmių užpildymas

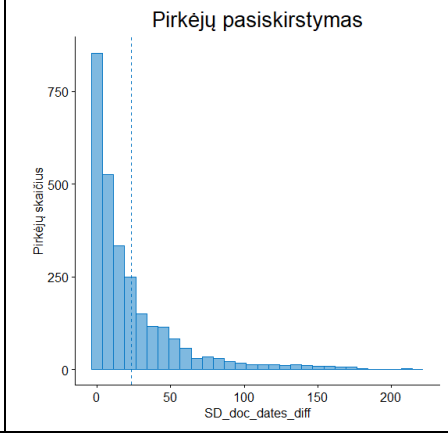
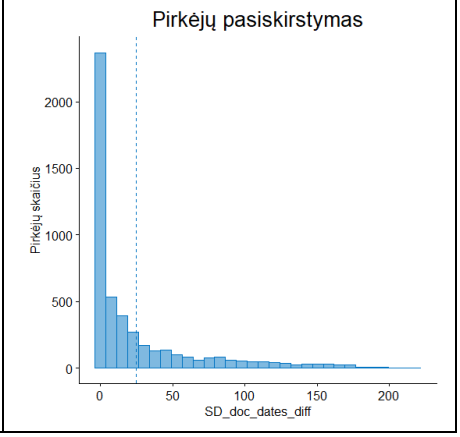
„Mean_doc_dates_diff“. Šis kintamasis parodo vidutinį skirtumą, tarp pirkėjui išrašytų sąskaitų, dienomis. Todėl, kai klientas turi vieną sąskaitą – reikšmė neegzistuoja. Tuščios reikšmės buvo pakeistos kintamojo „Days_after_last_doc“ reikšmėmis. Kadangi mažiau negu pusę pirkėjų, turinčių vieną sąskaitą, galima laikyti naujais, po reikšmių pakeitimo vidutiniai tarpai tarp sąskaitų išaugo (žr. 1 lentelę).

1 lentelė. Kintamojo "Mean_doc_dates_diff" charakteristikų pasikeitimas po tuščių reikšmių supildymo

Charakteristika	**Prieš**	**Po**
Reikšmių kiekis, Vnt.	3 345	4 857
Nulinės reikšmės, Vnt.	48	88
Tuščios reikšmės, Vnt.	1 512	0
Minimali reikšmė	0,00	0
Maksimali reikšmė	336,00	363
Mediana	30,00	38
Vidurkis	44,42	84,82
Standartinė vidurkio paklaida	0,85	1,33
Vidurkio 0,95 pasikliovimo intervalas	1,67	2,61
Dispersija	2 415,88	8 597,58
Standartinis nuokrypis	49,15	92,72
Variacijos koeficientas	1,11	1,09
Pasiskirstymas		

„SD_doc_dates_diff“. Šis kintamasis rodo kiek gali keistis kintamojo *„Mean_doc_dates_diff“* reikšmės ir yra apskaičiuojamas mažiausiai iš trijų skaičių, todėl pirkėjams, turintiems nedaugiau 2 sąskaitų, charakteristika neegzistuoja. Tuščių reikšmių užpildymas buvo atliekamas atsižvelgiant į dokumentų kiekį t.y. pirkėjams, turintiems vieną dokumentą, reikšmė buvo pakeista į 0, tuo tarpu pirkėjams, turintiems 2 sąskaitas buvo panaudotas *„Days_after_last_doc“* kintamasis ir apskaičiuota reikšmių sklaida tarp jo ir *„Mean_doc_dates_diff“* reikšmių. Kaip pasikeitė kintamojo pagrindinės charakteristikos ir histograma pateikiama 2 lentelėje.

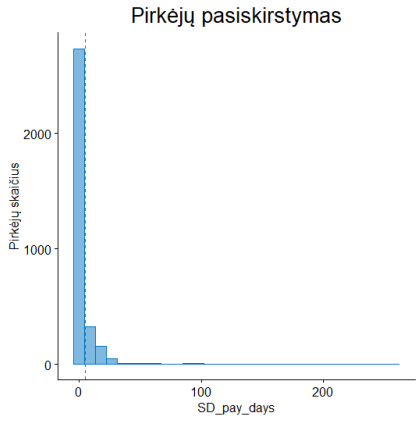
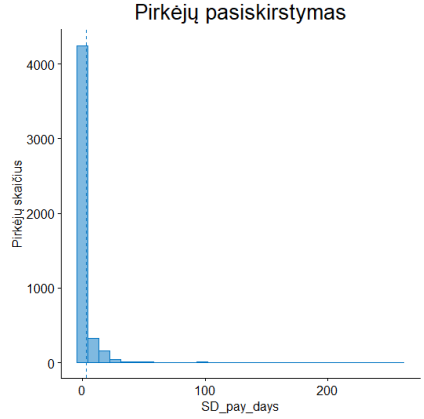
2 lentelė. Kintamojo „SD_doc_dates_diff " charakteristikų pasikeitimas po tuščių reikšmių supildymo

Charakteristika	**Prieš**	**Po**
Reikšmių kiekis, Vnt.	2 721	4 857
Nulinės reikšmės, Vnt.	21	1 534
Tuščios reikšmės, Vnt.	2 136	0
Minimali reikšmė	0,00	0,00
Maksimali reikšmė	218,00	218,00
Mediana	11,00	5,00
Vidurkis	23,15	24,43
Standartinė vidurkio paklaida	0,60	0,57
Vidurkio 0,95 pasikliovimo intervalas	1,18	1,11
Dispersija	983,60	1 570,28
Standartinis nuokrypis	31,36	39,63
Variacijos koeficientas	1,36	1,62
Pasiskirstymas		

„SD_pay_days“. Šis kintamasis parodo kokia yra reikšmių sklaida apie kiekvieno pirkėjo vidutinį sąskaitų apmokėjimo terminą. Kadangi šis parametras negali būti apskaičiuotas tik iš vieno dokumento, visas tokias reikšmes pakeičiame nulinėmis. Kintamojo pagrindinių charakteristikų ir pasiskirstymo pasikeitimas parodytas 3 lentelėje.

3 lentelė. Kintamojo „SD_pay_days“ charakteristikų pasikeitimas po tuščių reikšmių supildymo

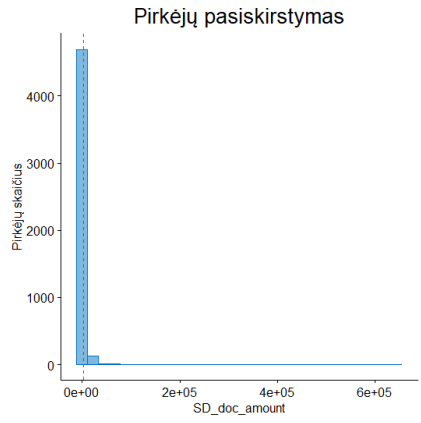
Charakteristika	**Prieš**	**Po**
Reikšmių kiekis, Vnt.	3 345	4 857
Nulinės reikšmės, Vnt.	1 979	3 491
Tuščios reikšmės, Vnt.	1 512	0
Minimali reikšmė	0,00	0
Maksimali reikšmė	258,00	258,00
Mediana	0,00	0
Vidurkis	4,59	3,16
Standartinė vidurkio paklaida	0,31	0,21

Vidurkio 0,95 pasikliovimo intervalas	0,60	0,42
Dispersija	318,08	223,56
Standartinis nuokrypis	17,83	14,95
Variacijos koeficientas	3,88	4,73
Pasiskirstymas		

„SD_doc_amount“ kintamasis parodo kokia sklaida yra apie vidutinę pirkėjų sąskaitų sumą ir negali būti apskaičiuotas kai pirkėjas turi vieną sąskaitą. Kadangi tokių pirkėjų sąskaitų sumas galima laikyti pastoviomis (nes jų yra tik viena), todėl visoms tuščios reikšmėms priskiriame nulines reikšmes. Kaip pasikeitė kintamojo charakteristikos pateikiama 4 lentelėje.

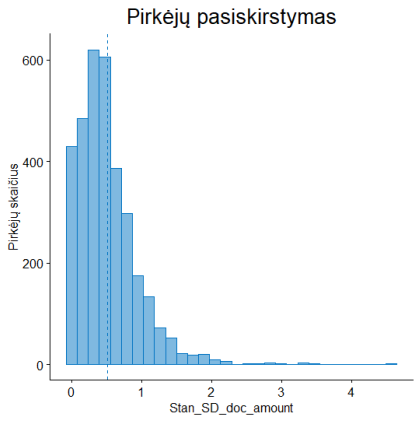
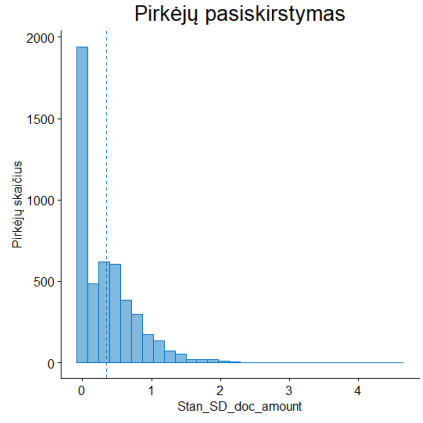
4 lentelė. Kintamojo „SD_doc_amount“ charakteristikų pasikeitimas po tuščių reikšmių supildymo

Charakteristika	**Prieš**	**Po**
Reikšmių kiekis, Vnt.	3 345	4 857
Nulinės reikšmės, Vnt.	112	1 624
Tuščios reikšmės, Vnt.	1512	0
Minimali reikšmė	0,00	0,00
Maksimali reikšmė	644 872,90	644 872,90
Mediana	227,14	60,45
Vidurkis	3 001,17	2 066,90
Standartinė vidurkio paklaida	358,44	247,65
Vidurkio 0,95 pasikliovimo intervalas	702,77	485,50
Dispersija	429 751 200,00	297 872 111,00
Standartinis nuokrypis	20 730,44	17 258,97
Variacijos koeficientas	6,91	8,35

Pasiskirstymas		

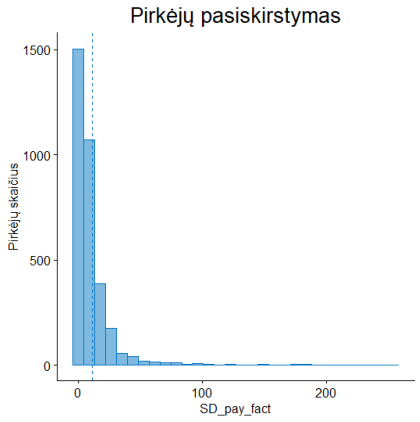
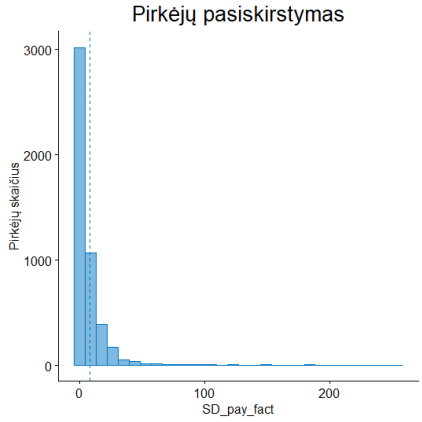
„Stan_SD_doc_amount“ parodo pirkėjų vidutinių sąskaitų sumų variaciją įvertinus ir patį vidurkį t.y. parodo kiek kartų sumos gali skirtis nuo vidurkio. Kadangi jis apskaičiuojamas padalinant standartinį nuokrypį iš vidurkio, jis neegzistuoja tada kaip standartinis nuokrypis neegzistuoja t.y. kai pirkėjas turi vieną sąskaitą. Todėl šio kintamojo tuščios reikšmės pakeičiamos nulinėmis ir kaip pasikeičia kintamojo charakteristikos pateikiama 5 lentelėje.

5 lentelė. Kintamojo „Stan_SD_doc_amount“ charakteristikų pasikeitimas po tuščių reikšmių supildymo

Charakteristika	**Prieš**	**Po**
Reikšmių kiekis, Vnt.	3 345	4 857
Nulinės reikšmės, Vnt.	138	1 650
Tuščios reikšmės, Vnt.	1 512	0
Minimali reikšmė	0,00	0,00
Maksimali reikšmė	4,57	4,57
Mediana	0,43	0,24
Vidurkis	0,51	0,35
Standartinė vidurkio paklaida	0,01	0,01
Vidurkio 0,95 pasikliovimo intervalas	0,01	0,01
Dispersija	0,18	0,18
Standartinis nuokrypis	0,42	0,41
Variacijos koeficientas	0,84	1,21
Pasiskirstymas		

„Mean_pay_fact“ parodo vidutinį faktinį apmokėjimą t.y. kiek dienų vidutiniškai skyrėsi sąskaitų apmokėjimas nuo apmokėjimo termino. Kadangi duomenų rinkinyje yra 7 pirkėjai, kurie yra neapmokėję visų sąskaitų, jiems šio kintamojo reikšmės negalima apskaičiuoti. Todėl yra apskaičiuojamas skirtumas tarp sąskaitų apmokėjimo termino ir 2018 metų galo, kuris nurodys ar pirkėjai vėluoja susimokėti ar ne. Kadangi 7 pirkėjai sudaro 0,14 proc. duomenų, charakteristikos nesikeičia.
„SD_pay_fact“ parodo pirkėjų atliktų mokėjimų vėlavimo sklaidą ir negalima būti apskaičiuojamas 1514 pirkėjų, nes jie turi po vieną sąskaitą arba neturi apmokėtų sąskaitų. Visais atvejais pakeičiame tuščias reikšmes į nulines reikšmes. Kaip tai pakeičia kintamojo pasiskirstymą galima matyti 6 lentelėje.

6 lentelė. Kintamojo „SD_pay_fact“ charakteristikų pasikeitimas po tuščių reikšmių supildymo

Charakteristika	**Prieš**	**Po**
Reikšmių kiekis, Vnt.	3 343	4 857
Nulinės reikšmės, Vnt.	600	2 114
Tuščios reikšmės, Vnt.	1 514	0
Minimali reikšmė	0,00	0,00
Maksimali reikšmė	254	254
Mediana	5,00	2,00
Vidurkis	11,32	7,79
Standartinė vidurkio paklaida	0,38	0,28
Vidurkio 0,95 pasikliovimo intervalas	0,75	0,54
Dispersija	494,72	367,96
Standartinis nuokrypis	22,24	19,18
Variacijos koeficientas	1,97	2,46
Pasiskirstymas		

„Docs_late_proc“ parodo kiek procentais pirkėjas sąskaitų vėlavo apmokėti ir kadangi 7 pirkėjai nėra apmokėję visų savo sąskaitų, jiems neįmanoma nustatyti šio kintamojo reikšmės taip, kaip kitiems įrašams. Šioje vietoje pasinaudoti galima pakoreguota kintamojo „Mean_pay_fact“ reikšme. Jeigu pirkėjas nevėluoja apmokėti sąskaitų, jam bus priskirta nulinė reikšmė. Visais kitais atvejais, bus nurodyta, jo vėluoja 100 proc. sąskaitų. Kadangi 7 pirkėjai sudaro 0,14 proc. duomenų, charakteristikos nesikeičia.

„Pay_doc_partly_proc“ kintamasis parodo kiek procentų sąskaitų pirkėjas apmoka dalimis. Šios reikšmės negalima apskaičiuoti 16 pirkėjų, nes jie nėra apmokėję sąskaitų arba visų sąskaitų sumos buvo nurašytos kaip beviltiškos skolos. Todėl visos tuščios reikšmės yra pakeičiamos į nulines reikšmes. Kadangi 16 pirkėjai sudaro 0,33 proc. duomenų, charakteristikos nesikeičia.

4 priedas. K-vidurkių metodu gautų klasterių centrai

Klasterių centrai gauti naudojant originalius duomenis:

Nr.	Mean_doc_ dates_diff	Doc_ count	Days_after_ last_doc	Mean_ pay_ days	Stan_SD_ doc_amount	Mean_doc_ amount	Default_ doc_proc	Mean_pay_ fact	Docs_late_ proc	n
1	84,80	9,95	100,00	25,10	0,35	4704,00	0,27	4,81	33,60	4851
2	108,00	1,00	108,00	180,00	0,00	4830500,00	0,00	5,00	100,00	1
3	84,60	1,40	92,20	7,00	0,01	1371953,00	0,00	6,80	33,30	5

Nr.	Mean_doc_ dates_diff	Doc_ count	Days_after_ last_doc	Mean_ pay_ days	Stan_SD_ doc_amount	Mean_doc_ amount	Default_ doc_proc	Mean_pay_ fact	Docs_late_ proc	n
1	38,40	6,08	35,80	26,70	0,64	347591,00	0,00	3,50	39,80	12
2	84,90	9,96	101,00	25,10	0,35	3854,00	0,27	4,82	33,60	4839
3	108,00	1,00	108,00	180,00	0,00	4830500,00	0,00	5,00	100,00	1
4	84,60	1,40	92,20	7,00	0,01	1371953,00	0,00	6,80	33,30	5

Klasterių centrai gauti naudojant standartizuotus duomenis:

Nr.	Mean_doc_ dates_diff	Doc_ count	Days_after_ last_doc	Mean_ pay_ days	Stan_SD_ doc_amount	Mean_doc_ amount	Default_ doc_proc	Mean_pay_ fact	Docs_late_ proc	n
1	34,70	14,30	21,20	24,20	0,44	4543,00	0,01	-8,45	7,72	1585
2	44,40	13,70	79,80	37,00	0,48	11293,00	0,53	18,80	78,70	1679
3	177,00	1,70	201,00	13,50	0,13	5240,00	0,25	3,30	11,90	1593

Nr.	Mean_doc_ dates_diff	Doc_ count	Days_after_ last_doc	Mean_ pay_ days	Stan_SD_ doc_amount	Mean_doc_ amount	Default_ doc_proc	Mean_pay_ fact	Docs_late_ proc	n
1	81,20	3,14	119,00	11,50	0,27	5806,00	0,01	-3,93	1,03	1243
2	51,00	9,43	92,50	38,30	0,45	11683,00	0,59	19,50	80,50	1456
3	264,00	1,06	259,00	17,00	0,02	4784,00	0,53	9,84	21,30	755
4	26,50	21,30	6,59	27,80	0,50	4754,00	0,02	-5,41	20,60	1403

Klasterių centrai gauti naudojant transformuotus duomenis be išskirčių:

Nr.	Mean_doc_ dates_diff	Doc_ count	Days_after_ last_doc	Mean_ pay_ days	Stan_SD_ doc_amount	Mean_doc_ amount	Default_ doc_proc	Mean_pay_ fact	Docs_late_ proc	n
1	45,10	18,00	75,50	41,60	0,51	11495,00	0,82	22,20	82,90	1580
2	245,00	1,12	241,00	14,70	0,02	6705,00	0,00	3,79	17,80	901
3	50,60	7,90	63,50	18,00	0,37	4338,00	0,00	-6,37	6,89	2376

Nr.	Mean_doc_ dates_diff	Doc_ count	Days_after_ last_doc	Mean_ pay_ days	Stan_SD_ doc_amount	Mean_doc_ amount	Default_ doc_proc	Mean_pay_ fact	Docs_late_ proc	n
1	49,10	8,14	60,80	17,20	0,38	4459,00	0,00	-6,12	7,12	2327
2	45,90	18,50	75,90	30,80	0,51	8468,00	0,86	23,60	84,10	1510
3	84,10	3,48	136,00	258,00	0,25	54158,00	0,00	-5,49	46,10	99
4	239,00	1,14	237,00	10,60	0,02	6497,00	0,00	2,70	16,50	921

Klasterių centrai gauti naudojant standartizuotus, transformuotus duomenis be išskirčių:

Nr.	Mean_doc_ dates_diff	Doc_ count	Days_after_ last_doc	Mean_ pay_ days	Stan_SD_ doc_amount	Mean_doc_ amount	Default_ doc_proc	Mean_pay_ fact	Docs_late_ proc	n
1	40,20	15,70	72,80	38,90	0,52	12625,00	0,50	20,20	77,70	1585
2	165,00	1,46	173,00	7,23	0,09	5270,00	0,28	4,32	15,70	1821
3	32,40	14,20	39,70	32,40	0,49	3378,00	0,00	-11,40	8,00	1451

Nr.	Mean_doc_ dates_diff	Doc_ count	Days_after_ last_doc	Mean_ pay_ days	Stan_SD_ doc_amount	Mean_doc_ amount	Default_ doc_proc	Mean_pay_ fact	Docs_late_ proc	n
1	34,10	13,00	29,60	21,00	0,50	325,00	0,00	-10,50	5,80	1005
2	36,40	14,20	72,70	47,30	0,46	12419,00	0,00	-9,66	14,10	642
3	173,00	1,35	178,00	7,63	0,06	5376,00	0,29	4,78	17,10	1705
4	39,30	15,80	71,90	38,10	0,53	11325,00	0,53	21,30	79,30	1505

5 priedas. SVM metodu sudarytų modelių palyginimai

Modelio Nr.	Tikslumas	F1_GOOD	F1_AVERAGE	F1_BAD	Modelio F1
34-19	**0,825**	**0,742**	**0,850**	**0,866**	**0,819**
41-18	0,817	0,737	0,842	0,857	0,812
43-17	0,817	0,740	0,845	0,842	0,809
44-17	0,815	0,741	0,841	0,841	0,808
1-20	0,814	0,737	0,838	0,853	0,809
46-17	0,814	0,733	0,843	0,843	0,806
3-18	0,814	0,727	0,837	0,865	0,810
35-19	0,814	0,738	0,839	0,847	0,808
33-19	0,813	0,741	0,841	0,830	0,804
39-18	0,812	0,730	0,839	0,842	0,804
37-18	0,811	0,739	0,838	0,832	0,803
36-19	0,810	0,726	0,836	0,851	0,804
31-16	0,809	0,726	0,837	0,843	0,802
38-18	0,808	0,732	0,834	0,839	0,802
47-17	0,808	0,725	0,837	0,839	0,800
10-17	0,806	0,719	0,828	0,866	0,804
4-18	0,806	0,720	0,829	0,861	0,803
42-18	0,804	0,719	0,833	0,835	0,796
40-18	0,804	0,725	0,831	0,832	0,796
8-17	0,803	0,702	0,829	0,861	0,797
5-18	0,803	0,702	0,827	0,866	0,798
7-17	0,801	0,702	0,825	0,866	0,798
45-17	0,801	0,714	0,826	0,848	0,796
13-16	0,801	0,699	0,824	0,871	0,798
15-16	0,800	0,700	0,825	0,861	0,795
11-17	0,799	0,696	0,824	0,867	0,796
2-18	0,799	0,696	0,823	0,870	0,796
12-16	0,798	0,698	0,824	0,863	0,795
6-17	0,796	0,697	0,822	0,852	0,790
9-17	0,789	0,686	0,814	0,856	0,785
14-16	0,788	0,684	0,812	0,859	0,785
25-17	0,774	0,718	0,806	0,762	0,762
18-18	0,774	0,728	0,806	0,745	0,760
16-20	0,772	0,719	0,803	0,753	0,759
23-17	0,769	0,708	0,803	0,751	0,754
19-18	0,767	0,715	0,801	0,743	0,753
20-18	0,767	0,707	0,800	0,751	0,753

30-16	0,763	0,703	0,799	0,744	0,748
26-17	0,760	0,699	0,793	0,748	0,747
27-16	0,754	0,712	0,794	0,701	0,736
28-16	0,753	0,708	0,792	0,706	0,735
24-17	0,748	0,708	0,783	0,703	0,731
21-17	0,748	0,706	0,785	0,699	0,730
22-17	0,747	0,700	0,783	0,713	0,732
29-16	0,742	0,683	0,782	0,713	0,726
17-18	0,736	0,684	0,777	0,691	0,717
32-14	0,648	0,566	0,745	0,479	0,597

6 priedas. RF metodu sudarytų modelių palyginimai

Modelio Nr.	Tikslumas	F1_GOOD	F1_AVERAGE	F1_BAD	Modelio F1
RF-1	0,902	0,834	0,913	0,965	0,904
RF-csl-1	0,890	0,835	0,897	0,954	0,895
RF-md-1	0,891	0,827	0,902	0,947	0,892
RF-2	0,898	0,826	0,908	0,968	0,901
RF-csl-2	0,885	0,828	0,892	0,949	0,890
RF-md-2	0,884	0,816	0,896	0,943	0,885
RF-3	0,901	0,832	0,911	0,967	0,904
RF-csl-3	0,885	0,826	0,892	0,952	0,890
RF-md-3	0,886	0,820	0,898	0,940	0,886
RF-4	0,898	0,825	0,908	0,968	0,901
RF-csl-4	0,889	0,832	0,897	0,951	0,893
RF-md-4	0,889	0,823	0,901	0,948	0,891
RF-5	0,899	0,832	0,909	0,965	0,902
RF-csl-5	0,891	0,833	0,898	0,958	0,896
RF-md-5	0,887	0,820	0,899	0,945	0,888
RF-6	0,905	0,841	0,914	0,969	0,908
RF-csl-6	0,889	0,834	0,896	0,951	0,894
RF-md-6	0,891	0,826	0,902	0,947	0,892
RF-7	0,897	0,826	0,908	0,964	0,900
RF-csl-7	0,887	0,828	0,894	0,956	0,893
RF-md-7	0,890	0,827	0,901	0,945	0,891
RF-8	0,900	0,830	0,910	0,968	0,903
RF-csl-8	0,887	0,830	0,894	0,952	0,892
RF-md-8	0,891	0,826	0,902	0,949	0,892
RF-9	0,904	0,837	0,914	0,968	0,906
RF-csl-9	0,895	0,841	0,902	0,958	0,900
RF-md-9	0,888	0,825	0,899	0,944	0,889
RF-10	0,900	0,831	0,911	0,966	0,903
RF-csl-10	0,891	0,837	0,898	0,950	0,895
RF-md-10	0,891	0,828	0,903	0,946	0,892
RF-11	0,891	0,823	0,902	0,955	0,893
RF-csl-11	0,876	0,822	0,882	0,939	0,881
RF-md-11	0,879	0,810	0,891	0,942	0,881
RF-12	0,882	0,812	0,893	0,950	0,885
RF-csl-12	0,872	0,817	0,879	0,935	0,877
RF-md-12	0,882	0,813	0,894	0,945	0,884
RF-13	**0,889**	**0,821**	**0,900**	**0,955**	**0,892**

RF-csl-13	0,879	0,827	0,885	0,936	0,883
RF-md-13	0,884	0,815	0,895	0,947	0,886
RF-14	0,849	0,763	0,864	0,917	0,848
RF-csl-14	0,838	0,764	0,848	0,912	0,841
RF-md-14	0,855	0,777	0,869	0,920	0,855
RF-15	0,820	0,734	0,839	0,876	0,817
RF-csl-15	0,802	0,733	0,816	0,857	0,802
RF-md-15	0,818	0,737	0,839	0,867	0,814
RF-16	0,802	0,721	0,821	0,857	0,800
RF-csl-16	0,795	0,737	0,801	0,856	0,798
RF-md-16	0,799	0,711	0,819	0,857	0,796
RF-17	0,788	0,705	0,807	0,850	0,787
RF-csl-17	0,773	0,701	0,785	0,846	0,777
RF-md-17	0,787	0,708	0,806	0,848	0,787

7 priedas. Programos tekstas

1 programa - Duomenų apžvalga.R

```
library(readr)
library(pastecs)
library(dplyr)
library(ggpubr)

# Duomenų įkėlimas
Data <- read_delim("DATA.csv", ";", escape_double=FALSE,
   col_types=cols(Closed_at_date=col_date(format="%Y-%m-%d"),
   Closed_by_entry=col_character(),Customer_no=col_character(),
    Customer_type=col_character(), Debt_default=col_character()
   Document_date=col_date(format="%Y-%m-%d"), Due_date=col_date(format="%Y-%m-%d"),
   Factoring=col_character(), Open=col_character()), trim_ws=TRUE)

# Pakeičiami kintamųjų tipai
Data$Debt_default <- as.factor(Data$Debt_default)
Data$Customer_no <- as.factor(Data$Customer_no)
Data$Customer_group <- as.factor(Data$Customer_group)
Data$Dimension <- as.factor(Data$Dimension)
Data$Open <- as.factor(Data$Open)
Data$Payment_code <- as.factor(Data$Payment_code)
Data$Factoring <- as.factor(Data$Factoring)
Data$Country <- as.factor(Data$Country)
Data$Customer_type <- as.factor(Data$Customer_type)

# Pradinių duomenų analizė
Data %>% group_by(Customer_group) %>% summarize(Dok_kiekis=n(),
   Dok_suma=sum(Invoice_amount), Vid_dok_suma=Dok_suma/Dok_kiekis) %>%
   mutate(Dok_proc=Dok_kiekis/sum(Dok_kiekis)*100,
   Dok_suma_proc=Dok_suma/sum(Dok_suma)*100)
Data %>% group_by(Country) %>% filter(Country!="LT") %>% summarise(Dok_kiekis=n()) %>%
   mutate(Dok_proc=Dok_kiekis/sum(Dok_kiekis)*100) %>%
   top_n(5, Dok_proc) %>% arrange(desc(Dok_proc))
Data %>% select(Country) %>% distinct() %>% summarise(n=n())
Data %>% group_by(Payment_code) %>% summarise(n=n()) %>% mutate(proc = n/sum(n)*100) %>%
   top_n(30) %>% arrange(desc(proc))
Data %>% mutate(Late=as.double(Closed_at_date-Due_date)) %>% select(Late, Debt_default,
   Open) %>% summarise(Defaulted=sum(Debt_default==1), Unpayed=sum(Open==1),
   Payed=sum(Debt_default==0 & Open==0), Payed_upfront=sum(!is.na(Late) & Debt_default==0
   & Open==0 & Late<0), Payed_ontime=sum(!is.na(Late) & Debt_default==0 & Open==0 &
   Late==0), Payed_late30=sum(!is.na(Late) & Debt_default==0 & Open==0 & Late>0 &
   Late<30), Payed_late60=sum(!is.na(Late) & Debt_default==0 & Open==0 & Late>=30 &
   Late<60), Payed_latemore=sum(!is.na(Late) & Debt_default==0 & Open==0 & Late>=60))

# Duomenų agregavimas
# Customer_group, Customer_type, Country
Data_2 <- Data %>% group_by(Customer_no) %>% select(Customer_no, Customer_group,
   Customer_type, Country) %>% distinct()
# Purch_from_multi
temp <- Data
temp$Entry_no <- gsub("\\d","",temp$Entry_no)
temp <- temp %>% group_by(Customer_no, Entry_no) %>% summarise(n=n()) %>%
   group_by(Customer_no) %>% tally()
temp$n[temp$n == 1] <- 0
temp$n[temp$n != 0] <- 1
colnames(temp)[2] <- "Purch_from_multi"
Data_2 <- merge(Data_2, temp, by="Customer_no", all.x=TRUE)
rm(temp)
```

```
# Purch_multi_product
temp <- Data %>% group_by(Customer_no, Dimension) %>% summarize(n=n()) %>%
   group_by(Customer_no) %>% tally()
temp$n[temp$n == 1] <- 0
temp$n[temp$n != 0] <- 1
colnames(temp)[2] <- "Purch_multi_product"
Data_2 <- merge(Data_2, temp, by="Customer_no", all.x=TRUE)
rm(temp)
# Doc_count, Mean_doc_dates_diff, SD_doc_dates_diff, Doc_period, Days_after_last_doc
temp <- Data %>% group_by(Customer_no) %>% arrange(Document_date) %>%
   summarise(Doc_count=n(),
   Mean_doc_dates_diff=round(mean(as.double(diff.Date(Document_date))),0),
   SD_doc_dates_diff=round(sd(diff.Date(Document_date)),0),
   Doc_period=as.double(max(Document_date)-min(Document_date)),
   Days_after_last_doc=as.double(as.Date("2018-12-31", format="%Y-%m-%d")-
   max(Document_date)))
Data_2 <- merge(Data_2, temp, by="Customer_no", all.x=TRUE)
rm(temp)
# Pre_pay_doc_proc, Cash_pay_doc_proc, Factoring_pay_doc_proc, Mean_pay_days, SD_pay_days
temp <- Data %>% group_by(Customer_no) %>%
   summarise(Pre_pay_doc_proc=round(sum(Payment_code=="ISANKSTINIS")/n()*100,2),
   Cash_pay_doc_proc=round(sum(Payment_code=="GRYNAIS")/n()*100,2),
   Factoring_pay_doc_proc=round(sum(Factoring==1)/n()*100,2),
   Mean_pay_days=round(mean(as.double(Due_date-Document_date)),0),
   SD_pay_days=round(sd(Due_date-Document_date),0))
Data_2 <- merge(Data_2, temp, by="Customer_no", all.x=TRUE)
rm(temp)
# Mean_doc_amount, SD_doc_amount, Stan_SD_doc_amount, Amount_diff
temp <- Data %>% group_by(Customer_no) %>% arrange(Invoice_amount) %>%
   summarise(Mean_doc_amount=round(mean(Invoice_amount),2),
   SD_doc_amount=round(sd(Invoice_amount),2),
   Stan_SD_doc_amount=round(SD_doc_amount/Mean_doc_amount,2),
   Amount_diff=max(Invoice_amount)-min(Invoice_amount))
Data_2 <- merge(Data_2, temp, by="Customer_no", all.x=TRUE)
rm(temp)
# Mean_pay_fact, SD_pay_fact, Docs_late_proc
   temp <- Data %>% group_by(Customer_no) %>% mutate(payment=as.double(Closed_at_date-
   Due_date)) %>% filter(!is.na(payment)) %>%
   summarise(Mean_pay_fact=round(mean(payment),0), SD_pay_fact=round(sd(payment),0),
   Docs_late_proc=round(sum(payment>0)/n()*100,2))
Data_2 <- merge(Data_2, temp, by="Customer_no", all.x=TRUE)
rm(temp)
# Pay_doc_partly_proc
temp <- Data %>% mutate(diff=round(Invoice_amount - Closed_by_amount,2)) %>%
   filter(Closed_by_entry!=0, Debt_default==0) %>% select(Customer_no, diff) %>%
   group_by(Customer_no) %>% summarize(Pay_doc_partly_proc=round(sum(diff!=0)/n()*100,2))
Data_2 <- merge(Data_2, temp, by="Customer_no", all.x=TRUE)
rm(temp)
# Open_doc_proc, Default_doc_proc, Default_doc_amount, Default_doc_amount_proc
temp <- Data %>% group_by(Customer_no) %>%
   summarise(Open_doc_proc=round(sum(Open==1)/n()*100,2),
   Default_doc_proc=round(sum(Debt_default==1)/n()*100,2),
   Default_doc_amount=sum(Invoice_amount[which(Debt_default==1)]),
   Default_doc_amount_proc=round(Default_doc_amount/sum(Invoice_amount)*100,2))
Data_2 <- merge(Data_2, temp, by="Customer_no", all.x=TRUE)
rm(temp)

# Agreguotų duomenų apžvalga (bendros charakteristikos)
str(Data_2[,2:6])
Data_2$Purch_from_multi <- as.factor(Data_2$Purch_from_multi)
Data_2$Purch_multi_product <- as.factor(Data_2$Purch_multi_product)
```

```
for(i in 2:6) {
  name <- colnames(Data_2)[i]
  print(Data_2 %>% group_by(Data_2[,i]) %>% summarise(Kiekis=n()) %>%
     mutate(Proc=round(Kiekis/sum(Kiekis)*100,2)))
  print(ggplot(Data_2, aes(Data_2[,i])) +  geom_bar(fill="#0073C2FF") +  theme_pubclean()
     + theme(axis.text.x = element_text(angle=90, hjust=1)) +
     ggtitle("Pirkėjų pasiskirstymas") + xlab(name) + ylab("Pirkėjų skaičius") +
     theme(plot.title=element_text(size=20, hjust=0.5)))
}

temp <- NULL
for(i in 7:28) {
  name <- colnames(Data_2)[i]
  temp <- rbind(temp, as.data.frame(t(round(stat.desc(Data_2[, i]), 2))))
  print(gghistogram(Data_2, x=name, add="mean", fill="#0073C2FF", color="#0073C2FF",
     main="Pirkėjų pasiskirstymas") +
     xlab(name) + ylab("Pirkėjų skaičius") + theme(plot.title=element_text(size=20,
     hjust=0.5)))
}

# Tuščių reikšmių užpildymas
Data_2$Mean_doc_dates_diff[is.na(Data_2$Mean_doc_dates_diff)] <-
   Data_2$Days_after_last_doc[is.na(Data_2$Mean_doc_dates_diff)]
Data_2$SD_doc_dates_diff[Data_2$Doc_count==1] <- 0
temp <- Data_2 %>% filter(Doc_count==2) %>% group_by(Customer_no) %>% select(Customer_no,
   Days_after_last_doc, Mean_doc_dates_diff) %>%
   mutate(f=0, s=Mean_doc_dates_diff, t=Days_after_last_doc) %>%
   summarise(sd=sd(c(f,s,t)))
Data_2$SD_doc_dates_diff[Data_2$Doc_count==2] <- temp$sd
rm(temp)
Data_2$SD_pay_days[is.na(Data_2$SD_pay_days)] <- 0
Data_2$SD_doc_amount[is.na(Data_2$SD_doc_amount)] <- 0
Data_2$Stan_SD_doc_amount[is.na(Data_2$Stan_SD_doc_amount)] <- 0
Data_2$Mean_pay_fact[Data_2$Customer_no==14084] <- 51
Data_2$Mean_pay_fact[Data_2$Customer_no==38409] <- -259
Data_2$Mean_pay_fact[Data_2$Customer_no==57164] <- 91
Data_2$Mean_pay_fact[Data_2$Customer_no==59294] <- 123
Data_2$Mean_pay_fact[Data_2$Customer_no==59539] <- 273
Data_2$Mean_pay_fact[Data_2$Customer_no==60439] <- 30
Data_2$Mean_pay_fact[Data_2$Customer_no==61409] <- -20
Data_2$SD_pay_fact[is.na(Data_2$SD_pay_fact)] <- 0
Data_2$Docs_late_proc[is.na(Data_2$Docs_late_proc) & Data_2$Mean_pay_fact<0] <- 100
Data_2$Docs_late_proc[is.na(Data_2$Docs_late_proc) & Data_2$Mean_pay_fact>=0] <- 0
Data_2$Pay_doc_partly_proc[is.na(Data_2$Pay_doc_partly_proc)] <- 0
```

2 programa - Klasterizavimas.R

```
library(corrplot)
library(Hmisc)
library(factoextra)
library(pracma)
library(dplyr)

outliers <- function(x){
  qnt <- quantile(x, probs=c(.25, .75), na.rm = T)
  caps <- quantile(x, probs=c(.05, .95), na.rm = T)
  H <- 1.5 * IQR(x, na.rm=T)
  x[x < (qnt[1]-H)] <- caps[1]
  x[x > (qnt[2]+H)] <- caps[2]
  return(x)
}
```

```
for(i in 7:28){
  print(shapiro.test(Data_2[,i]))
}
# Visi ne normaliojo pasiskirstymo, todel bus naudojamas Spearmano kor. koef.
temp <- rcorr(as.matrix(Data_2[,7:11]), type="spearman")
corrplot(temp$r, type="upper", order="hclust",p.mat=temp$P, sig.level = 0.01)
# Is sios grupes paimsim kintamuosiu, SD_doc_dates_diff ir Doc_count
temp <- rcorr(as.matrix(Data_2[,12:16]), type="spearman")
corrplot(temp$r, type="upper", order="hclust",p.mat=temp$P, sig.level = 0.01)
# Is sios grupes paimam Mean_pay_days
temp <- rcorr(as.matrix(Data_2[,17:20]), type="spearman")
corrplot(temp$r, type="upper", order="hclust",p.mat=temp$P, sig.level = 0.01)
#  Is sios grupes paimama Dtan_SD_doc_amount ir Mean_doc_amount
temp <- rcorr(as.matrix(Data_2[,21:28]), type="spearman")
corrplot(temp$r, type="upper", order="hclust",p.mat=temp$P, sig.level = 0.01)
#  Is sios grupes paimam Defaut_doc_proc ir Mean_pay_fact ir Docs_late_proc
Data_3 <- Data_2 %>% select(Customer_no, Mean_doc_dates_diff, Days_after_last_doc,
Doc_count, Mean_pay_days, Stan_SD_doc_amount, Mean_doc_amount,
                            Default_doc_proc, Mean_pay_fact, Docs_late_proc)
temp <- rcorr(as.matrix(Data_3[,2:10]), type="spearman")
corrplot(temp$r, type="upper", order="hclust",p.mat=temp$P, sig.level = 0.01)
rm(temp)

# Standartizuojam
Data_3.stand <- bind_cols(as.data.frame(Data_3[,1]), as.data.frame(scale(Data_3[,2:10])))
colnames(Data_3.stand)[1] <- "Customer_no"
# transformuojam informacija ir ismetam isskirtis
Data_3$Mean_doc_dates_diff_out <- sqrt(Data_3$Mean_doc_dates_diff)
Data_3$Days_after_last_doc_out <- sqrt(Data_3$Days_after_last_doc)
Data_3$Doc_count_out <- outliers(log(Data_3$Doc_count))
Data_3$Mean_pay_days_out <- outliers(log(Data_3$Mean_pay_days+1))
Data_3$Stan_SD_doc_amount_out <- outliers(sqrt(Data_3$Stan_SD_doc_amount))
Data_3$Mean_doc_amount_out <- outliers(log(Data_3$Mean_doc_amount))
Data_3$Default_doc_proc_out <- sqrt(Data_3$Default_doc_proc)
Data_3$Mean_pay_fact_out <- outliers(nthroot(Data_3$Mean_pay_fact,3))
Data_3$Docs_late_proc_out <- sqrt(Data_3$Docs_late_proc)
Data_3.stand <- bind_cols(Data_3.stand, as.data.frame(scale(Data_3[,11:19])))

# Alkunes metodas
fviz_nbclust(Data_3[,2:10], kmeans, method="wss", k.max=10) + labs(subtitle="Alkūnės
metodas", title="Optimalaus klasterių skaičiaus nustatymas", x="Klasterių skaičius")

# Klasterizavimas
clustering_1 <- kmeans(Data_3[,2:10], 3)
clustering_2 <- kmeans(Data_3[,2:10], 4)
clustering_3 <- kmeans(Data_3[,11:19], 3)
clustering_4 <- kmeans(Data_3[,11:19], 4)
clustering_5 <- kmeans(Data_3.stand[,2:10], 3)
clustering_6 <- kmeans(Data_3.stand[,2:10], 4)
clustering_7 <- kmeans(Data_3.stand[,11:19], 3)
clustering_8 <- kmeans(Data_3.stand[,11:19], 4)

# Klasteriu vidurkiu apskaiciavimas
Data_3$clustering_1 <- clustering_1$cluster
Data_3$clustering_2 <- clustering_2$cluster
Data_3$clustering_3 <- clustering_3$cluster
Data_3$clustering_4 <- clustering_4$cluster
Data_3$clustering_5 <- clustering_5$cluster
Data_3$clustering_6 <- clustering_6$cluster
Data_3$clustering_7 <- clustering_7$cluster
Data_3$clustering_8 <- clustering_8$cluster
```

```
for (i in 20:27) {
Data_3 %>% group_by(Data_3[,i]) %>%
   summarise(Mean_doc_dates_diff=mean(Mean_doc_dates_diff), Doc_count=mean(Doc_count),
   Days_after_last_doc=mean(Days_after_last_doc), Mean_pay_days=mean(Mean_pay_days),
   Stan_SD_doc_amount=mean(Stan_SD_doc_amount), Mean_doc_amount=mean(Mean_doc_amount),
   Default_doc_proc=mean(Default_doc_proc), Mean_pay_fact=mean(Mean_pay_fact),
   Docs_late_proc=mean(Docs_late_proc), n=n())
}

Data_2$Rating <- Data_3[,22]

# istrinam nereikalingus duomenis
rm(clustering_1)
rm(clustering_2)
rm(clustering_3)
rm(clustering_4)
rm(clustering_5)
rm(clustering_6)
rm(clustering_7)
rm(clustering_8)
rm(Data_3)
rm(Data_3.stand)
rm(temp)
```

3 programa – Klasifikavimas.R

```
library(readr)
library(dplyr)
library(ggpubr)
library(corrplot)
library(Hmisc)
library(caret)
library(e1071)
library(tidyr)
library(randomForest)
library(randomForestExplainer)
library(readxl)
library(writexl)

temp <- read_delim("Vertinimas.csv", ";", escape_double=FALSE,
   col_types=cols(Customer_no=col_character()), trim_ws=TRUE)
Data_2 <- merge(Data_2, temp, by="Customer_no", all.x=TRUE)
temp <- read_delim("Klasteriai_original.csv", ";", escape_double=FALSE,
   col_types=cols(Customer_no=col_character()), trim_ws=TRUE)
Data_2 <- merge(Data_2, temp, by="Customer_no", all.x=TRUE)
rm(temp)
Data_2 %>% group_by(`Rating NEW`) %>%
   summarise(Mean_doc_dates_diff=mean(Mean_doc_dates_diff), Doc_count=mean(Doc_count),
   Days_after_last_doc=mean(Days_after_last_doc), Mean_pay_days=mean(Mean_pay_days),
   Stan_SD_doc_amount=mean(Stan_SD_doc_amount), Mean_doc_amount=mean(Mean_doc_amount),
   Default_doc_proc=mean(Default_doc_proc), Mean_pay_fact=mean(Mean_pay_fact),
   Docs_late_proc=mean(Docs_late_proc), n=n())
Data_2$`Rating NEW` <- as.factor(Data_2$`Rating NEW`)
Data_2$Rating <- as.factor(Data_2$Rating)
Data_2$`Rating NEW` <- factor(Data_2$`Rating NEW`,labels = c("GOOD", "BAD","AVERAGE"))
Data_2$Rating <- factor(Data_2$Rating,labels = c("GOOD", "BAD","AVERAGE"))

ggplot(Data_2, aes(`Rating NEW`, fill = Rating)) + labs(fill="Kred. vert. pagal klast.")
   + geom_bar() + labs(title = "Pirkėjų kreditingumo vertinimo pasikeitimas", x = "Naujas
   kreditingumo įvertinimas", y = "Pirkėjų kiekis") + scale_fill_manual(values = c("GOOD"
```

```
   = "limegreen", "BAD" = "orangered", "AVERAGE" = "yellow2"))
ggplot(Data_2, aes(`Rating NEW`, fill = Customer_group)) + geom_bar() + labs(title =
   "Pirkėjų pasiskirstymas (pagal specialisto vertinimą)", x = "Kreditingumo
   įvertinimas", y = "Pirkėjo reg. grupė")
ggplot(Data_2, aes(`Rating NEW`, fill = Country)) + geom_bar(position="fill") +
   labs(title = "Pirkėjų pasiskirstymas (pagal specialisto vertinimą)", x = "Kreditingumo
   įvertinimas", y = "Pirkėjo reg. grupė")

View(Data_2 %>% select(-c(Customer_no, Customer_group, Customer_type, Country,
   Purch_from_multi, Purch_multi_product, `Rating NEW`)) %>% group_by(Rating) %>%
   summarise_all("mean"))
View(Data_2 %>% select(-c(Customer_no, Customer_group, Customer_type, Country,
   Purch_from_multi, Purch_multi_product, Rating)) %>% group_by(`Rating NEW`) %>%
   summarise_all("mean"))

Data_2 <- Data_2[,-30]

temp <- rcorr(as.matrix(Data_2[,7:28]), type="spearman")
corrplot(temp$r, type="upper", order="hclust",p.mat=temp$P, sig.level = 0.01)
temp$r

# modeliai <- read_excel("modeliams2.xlsx")
# modeliai <- read_excel("modeliams.xlsx")
pavadinimai <- NULL
myResults <- NULL

for(t in 1:47){
  vardai_is_eiles <- unlist(strsplit(as.character(modeliai[t,2]), split=" "))
  vardai_is_eiles[length(vardai_is_eiles)+1] <- "Rating NEW"

  SVM_data <- Data_2 %>% select(vardai_is_eiles)
  ilgis <- length(vardai_is_eiles)
  no <- t
  pavadinimai <- rbind(pavadinimai, data.frame(vardai_is_eiles, ilgis, no))

  k <- 3
  stulp_Y <- "Rating NEW"
  index_Y <- which(colnames(SVM_data) %in% stulp_Y)

  SVM_in <- SVM_data[,colnames(SVM_data) %in% pavadinimai$vardai[pavadinimai$no==t]]
  print(dim(SVM_in))
  stulp_Y <- "Rating NEW"
  index_Y <- which(colnames(SVM_in) %in% stulp_Y)
  Folds <- createFolds(SVM_in[, index_Y], k)
  variables <- paste(first(pavadinimai$no[pavadinimai$no==t]),
first(pavadinimai$ilgis[pavadinimai$no==t]), sep="-")
  Start <- Sys.time()
  for (i in 1:k)
  {
    test_index <- Folds[[i]]
    extra <- as.logical(rep(1, 1, nrow(SVM_in)))
    extra[test_index] <- FALSE
    train_index <- which(extra)
    Y <- SVM_in[train_index, index_Y]
    target <- SVM_in[test_index, index_Y]

    cat(sprintf("\nCV fold %d out of %d / SVM\n", i, k))
    train <- SVM_in[train_index, ]
    test <- SVM_in[-train_index, ]
    tunefit <- tune.svm(`Rating NEW` ~ ., data = train, cost = 10^(-3:3), gamma = 10^(-
        3:3))
```

```
    summary(tunefit)
    SVM_model <- tunefit$best.model
    gamma <- rep(tunefit$best.parameters[,1], length(target))
    cost <- rep(tunefit$best.parameters[,2], length(target))
    score <- predict(SVM_model, test)
    myResults <- rbind(myResults, data.frame(variables, test_index, gamma, cost, score,
        target))
    rm(SVM_model)
  }
  End <- Sys.time()
  print (End-Start)
}

myModels <- levels(as.factor(myResults[, "variables"]))
myScores <- spread(myResults, variables, score)

scores <- NULL
for (variables in myModels)
{
  print(variables)
  confusionMatrix <- caret::confusionMatrix(as.factor(myScores[,variables]),
     as.factor(myScores$target))
  print(confusionMatrix)
  Acc <- as.numeric(confusionMatrix$overall['Accuracy'])
  Sens_GOOD <- as.numeric(confusionMatrix$byClass[1,'Sensitivity'])
  Sens_AVERAGE <- as.numeric(confusionMatrix$byClass[3,'Sensitivity'])
  Sens_BAD <- as.numeric(confusionMatrix$byClass[2,'Sensitivity'])
  Prec_GOOD <- as.numeric(confusionMatrix$byClass[1,'Pos Pred Value'])
  Prec_AVERAGE <- as.numeric(confusionMatrix$byClass[3,'Pos Pred Value'])
  Prec_BAD <- as.numeric(confusionMatrix$byClass[2,'Pos Pred Value'])
  F1_GOOD <- 2*(Sens_GOOD*Prec_GOOD)/(Sens_GOOD+Prec_GOOD)
  F1_AVERAGE <- 2*(Sens_AVERAGE*Prec_AVERAGE)/(Sens_AVERAGE+Prec_AVERAGE)
  F1_BAD <- 2*(Sens_BAD*Prec_BAD)/(Sens_BAD+Prec_BAD)
  F1 <- (F1_GOOD+F1_AVERAGE+F1_BAD)/3
  scores <- rbind(scores, data.frame(variables, Acc, Sens_GOOD, Sens_AVERAGE, Sens_BAD,
     Prec_GOOD, Prec_AVERAGE, Prec_BAD, F1_GOOD, F1_AVERAGE, F1_BAD, F1))
}

# write_xlsx(x = SVM_results, path = "myResults.xlsx", col_names = TRUE)
# write_xlsx(x = scores, path = "scoresRF.xlsx", col_names = TRUE)
# write_xlsx(x = RF_results, path = "RF_results.xlsx", col_names = TRUE)

# RF:
set.seed(28)
k <- 3
ntree <- 500
ptree <- 150
myResults <- NULL
# min_depth_frame <- NULL
for (t in 1:17) {
  vardai_is_eiles <- unlist(strsplit(as.character(modeliai[t,2]), split=" "))
  vardai_is_eiles[length(vardai_is_eiles)+1] <- "Rating NEW"

  RF_data <- Data_2 %>% select(vardai_is_eiles)
  mtry <- floor(sqrt(ncol(RF_data)-1))

  stulp_Y <- "Rating NEW"
  index_Y <- which(colnames(RF_data) %in% stulp_Y)
  Folds <- createFolds(RF_data[, index_Y], k)

  for (i in 1:k)
  {
```

```
    test_index <- Folds[[i]]
    extra <- as.logical(rep(1, 1, nrow(RF_data)))
    extra[test_index] <- FALSE
    train_index <- which(extra)
    Y <- RF_data[train_index, index_Y]
    target <- RF_data[test_index, index_Y]

    cat(sprintf("\nCV fold %d out of %d / Random Forest. %d\n", i, k, t))
    rf_model <- tuneRF(RF_data[train_index, -index_Y], Y, mtryStart = mtry, ntreeTry =
   ntree,stepFactor = 2, improve = 0.01, plot = FALSE, doBest = T,
   strata = Y, replace = FALSE, importance = FALSE, do.trace = ptree)
    print(rf_model)
    model <- rep(paste("RF", t, sep="-"), length(target))
    score <- predict(rf_model, RF_data[test_index,-index_Y])
    myResults <- rbind(myResults, data.frame(test_index, model, score, target))
    # min_depth_frame <- rbind(min_depth_frame,
        data.frame(min_depth_distribution(rf_model)))
    rm(rf_model)

    cat(sprintf("\nCV fold %d out of %d / Random Forest - cost sensitive learning. %d\n",
        i, k, t))
    model_classwt <- prop.table(table(Y))
    rf_model <- tuneRF(RF_data[train_index, -index_Y], Y, mtryStart = mtry, ntreeTry =
        ntree, stepFactor = 2, improve = 0.01, plot = FALSE, doBest = T,
        classwt = model_classwt, cutoff = model_classwt,
        strata = Y, replace = FALSE, importance = FALSE, do.trace = ptree)
    print(rf_model)
    model <- rep(paste("RF-csl", t, sep="-"), length(target))
    score <- predict(rf_model, RF_data[test_index,-index_Y])
    myResults <- rbind(myResults,data.frame(test_index, model, score, target))
    rm(rf_model)

    cat(sprintf("\nCV fold %d out of %d / Random Forest - majority downsampling. %d\n",
        i, k, t))
    model_classwt <- prop.table(table(Y))
    rf_model <- tuneRF(RF_data[train_index, -index_Y], Y, mtryStart = mtry, ntreeTry =
        ntree, stepFactor = 2, improve = 0.01, plot = FALSE, doBest = T,
        sampsize = round(min(table(Y)*0.99)),
        strata = Y, replace = FALSE, importance = FALSE, do.trace = ptree)
    print(rf_model)
    model <- rep(paste("RF-md", t, sep="-"), length(target))
    score <- predict(rf_model,RF_data[test_index,-index_Y])
    myResults <- rbind(myResults,data.frame(test_index, model, score, target))
    rm(rf_model)
  }
}

RF_results <- myResults
myModels <- levels(myResults[, "model"])
myScores <- spread(myResults, model, score)

scores <- NULL
for (model in myModels)
{
  print(model)
  confusionMatrix <- caret::confusionMatrix(as.factor(myScores[,model]),
     as.factor(myScores$target))
  print(confusionMatrix)
  Acc <- as.numeric(confusionMatrix$overall['Accuracy'])
  Sens_GOOD <- as.numeric(confusionMatrix$byClass[1,'Sensitivity'])
  Sens_AVERAGE <- as.numeric(confusionMatrix$byClass[3,'Sensitivity'])
  Sens_BAD <- as.numeric(confusionMatrix$byClass[2,'Sensitivity'])
```

```
  Prec_GOOD <- as.numeric(confusionMatrix$byClass[1,'Pos Pred Value'])
  Prec_AVERAGE <- as.numeric(confusionMatrix$byClass[3,'Pos Pred Value'])
  Prec_BAD <- as.numeric(confusionMatrix$byClass[2,'Pos Pred Value'])
  F1_GOOD <- 2*(Sens_GOOD*Prec_GOOD)/(Sens_GOOD+Prec_GOOD)
  F1_AVERAGE <- 2*(Sens_AVERAGE*Prec_AVERAGE)/(Sens_AVERAGE+Prec_AVERAGE)
  F1_BAD <- 2*(Sens_BAD*Prec_BAD)/(Sens_BAD+Prec_BAD)
  F1 <- (F1_GOOD+F1_AVERAGE+F1_BAD)/3
  scores <- rbind(scores, data.frame(model, Acc, Sens_GOOD, Sens_AVERAGE, Sens_BAD,
     Prec_GOOD, Prec_AVERAGE, Prec_BAD, F1_GOOD, F1_AVERAGE, F1_BAD, F1))
}

plot_min_depth_distribution(min_depth_frame, mean_sample = "relevant_trees", k = 18, main
  = "Minimalaus ir vidutinio gylio pasiskirstymas") + labs(fill="Min. gylis", x =
  "Kintamieji", y = "Medžių skaičius")
info <- Data_2 %>% group_by(`Rating NEW`) %>% summarise(m1 = mean(Mean_doc_amount),
  m2=mean(Doc_count), m3=mean(Mean_pay_days), m4=mean(Mean_pay_fact),
  m5=mean(SD_pay_fact), m6=mean(Docs_late_proc))
```