Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Klientų lojalumo modeliavimas: detekcijos ir išlikimo analizės uždavinių palyginimas

Baigiamasis magistro projektas

Alvydas Misius
Projekto autorius

Doc. Evaldas Vaičiukynas
Vadovas
Doc. Aistė Dovalienė
Vadovė

Kaunas, 2019

Kauno technologijos universitetas
Matematikos ir gamtos mokslų fakultetas

Klientų lojalumo modeliavimas: detekcijos ir išlikimo analizės uždavinių palyginimas

Baigiamasis magistro projektas
Didžiųjų verslo duomenų analitika (621G12002)

Alvydas Misius
Projekto autorius

Doc. Evaldas Vaičiukynas
Vadovas
Doc. Aistė Dovalienė
Vadovė

Doc. Tomas Ruzgas
Recenzentas
Doc. Egidijus Rybakovas
Recenzentas

Kaunas, 2019

Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Alvydas Misius

Klientų lojalumo modeliavimas: detekcijos ir išlikimo analizės uždavinių palyginimas

Akademinio sąžiningumo deklaracija

Patvirtinu, kad mano, Alvydo Misiaus, baigiamasis projektas tema „Klientų lojalumo modeliavimas: detekcijos ir išliko analizės uždavinių palyginimas“ 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)

Turinys

Paveikslų sąrašas

Lentelių sąrašas

Misius, Alvydas. Klientų Lojalumo modeliavimas: detekcijos ir išlikimo analizės uždavinių palyginimas. Magistro baigiamasis projektas / vadovas doc. dr. Evaldas Vaičiukynas; doc. dr. Aistė Dovalienė; Kauno technologijos universitetas, Matematikos ir gamtos mokslų fakultetas.
Studijų kryptis ir sritis (studijų krypčių grupė): Taikomoji matematika (A02), Matematikos mokslai (A).
Reikšminiai žodžiai: klientų lojalumas, lojalumo detekcija, išlikimo analizė, logistinė regresija, atsitiktinis miškas.
Kaunas, 2019. 69 p.

Santrauka

Šiame darbe tirtas klientų lojalumas, remiantis įvairiais detekcijos ir išlikimo analizės metodais. Detekcija grindžiama abiejų klasių prognozavimo tikslumu. Fiksuojamos dvi skirtingos klasės – tai lojalūs ir nelojalūs klientai. Išbandomi penki detekcijos ir išlikimo modeliai: logistinė regresija, atsitiktiniai miškai, atraminių vektorių metodas, Cox proporcingumo rizikos modelis ir atsitiktiniai išlikimo miškai. Beveik visiems modeliais išbandomos daugumos mažinimo procedūros ir kiti modelių parametrų reguliarizacijos procesai. Daugumos mažinimo procedūra reiškia subalansuoto duomenų rinkinio sudarymą, kada suderinamas daugumos klasėje esančių klientų skaičius su atsitiktine imtimi iš daugumos klasės. Parametrų reguliarizacijos procesai susideda iš kiekvieno modelio kintamųjų įvertinimo pagal tam tikrus algoritmus.

Metodologijos pavyzdys pateikiamas dviejų duomenų rinkinių eksperimentu. Metodai apmokomi ir vėliau testuojami remiantis įmonių klientų duomenimis iš skirtingų verslo šakų: telekomunikacijų paslaugų ir „Premium“ klubo ypatingųjų paslaugų. Tyrimo metu naudota kryžminio patikrinimo procedūra ir rezultatai pateikiami kaip detekcijos sumaišymų matrica bei kiti įvairūs modelio klasifikavimo gerumo matai. Rezultatų apipavidalinimui pateikiamos modelių rezultatų kreivės: ROC, DET, preciziškumo – jautrumo ir pranašumo kreivės.

Rezultatai rodo, jog atsitiktiniai miškai kartu su duomenų rinkinio derinimu tiksliau klasifikuoja klientus nei kiti detekcijos ir išlikimo modeliai. Lyginant detekcijos ir išlikimo metodikas, detekcijos rezultatai suteikia tikslesnius prognozavimo rezultatus nei išlikimo metodai. Remiantis rezultatais, daroma prielaida, jog detekcijos modeliai turėtų būti naudojami klientų lojalumo prognozavimui.

Taip pat išbandomi du ribinės vertės nustatymo metodai, vienas iš jų kai ribinė vertė nustatoma kada fiksuojama lygių paklaidų vertė; kitas – tai tikėtino maksimalaus pelno atžvilgiu. Skirtingos ribinės vertės suteikia skirtingas sumaišymų matricas. Parodoma, jog finansinio pelno pagėrėjimas nėra susietas su modelio klasifikavimo tikslumo optimizavimu.

Misius, Alvydas. Modelling customer churn: comparison of detection and survival analysis methods. Master's Final Degree Project / supervisor assoc. prof. Evaldas Vaičiukynas; assoc. prof. Aistė Dovalienė; Faculty of Mathematics and Natural Sciences, Kaunas University of Technology.
Study field and area (study field group): Applied Mathematics (A02), Mathematical Sciences (A).
Keywords: customer loyalty, churn detection, survival analysis, logistic regression, random forest.
Kaunas, 2019. 69 pages.

Summary

In this paper, the detection of various churn and survival prediction models has been analysed. The detection is based on comparing the wellness of each group of prediction models. Five different detection and survival models were examined – logit, random forest, SVM, Cox proportional – hazards and random survival forest. Almost each were analysed with and without applying a majority down – sampling procedures with various parametric processes of regularization. Down – sampling stands for creation of balanced dataset by matching the number of samples in the majority class with the random sample from the same class. Parametric processes of regularization consists of each model's variable estimation.

An example of the methodology is provided with an experiment based on two datasets. Models were trained and tested by using customer data of two firms from different industries – the telecommunication service provider and extraordinary services of "Premium" club. Research was done by using cross – validation strategy. Results are supplied along with confusion matrixes and various detection measures of models fitness. Detector performance curves such as receiver operating characteristic, detector error trade off, precision – recall and lift plots also supplied.

The results shows that the random forest in combination with a balancing dataset procedures outperforms the other detection and survival methods. The outcomes also demonstrates that the ability to identify high risk customers with detection models is significantly better, than survival models for mentioned datasets. Research results indicates that churn models should be used for predicting customer behavior.

This research project is done by estimating two different methods for decision threshold – threshold in equal error rate, when sensitivity equals to specificity and threshold which optimizes expected maximum profit. Different thresholds gives different confusion matrixes. Analyzed data also concludes, that improvement in financial profit is not associated with an improvement in the number of churners targeted and optimizing the correctness of detection and survival models.

Įvadas

Tyrimų aktualumas. Klientų lojalumas – aktuali tema visoms verslo sritims. Tiriamos klientų elgsenos tendencijos rodo, jog klientai vis mažiau yra linkę būti ištikimi tam tikriems paslaugų tiekėjams ir vis lanksčiau žvelgia į paslaugų tiekėjų didėjantį konkurencingumą. Pavyzdžiui, apskaičiuojama, jog kas metus net 20 proc. vartotojų Amerikoje pakeičia savo kreditines korteles. Taip pat, apie 20 proc. – 38 proc. klientų kas metus keičia mobiliojo ryšio operatorius Europoje. Stebėta, jog naujų klientų įsigijimo kaštai vis auga, todėl klientų lojalumo valdymas tapo labai svarbi dilema. Pastebima, jog esant pakankamai lojalių klientų daliai, ryškėja verslo stabilaus augimo tendencija.

Nieko stebėtino, jog aukščiausio lygio vadovai pabrėžia klientų lojalumo svarbą. Klientų išlaikymas yra prioritetas rinkodaroje. Fiksuojama, jog įmonės vis didina savo biudžeto dalį, skirtą klientų išlaikymui. Svarbu paminėti, jog tame pačiame straipsnyje pabrėžiama statistinių įžvalgų svarba klientų lojalumo temoje. Verslo aplinkoje vis daugiau priimamų sprendimų grindžiami duomenų analize, modeliavimo metodų rezultatais ir hipotezių tikrinimu. Dėl tokios susidariusios situacijos, verslo plėtros vadovai patiria daugiau spaudimo bei dėmesio nei anksčiau. Klientų elgsenos modeliavimas, yra itin aktuali tema verslo sričiai. Norint išlaikyti populiarumą ir šiuolaikiškumą, įmonėms privaloma tirti savo verslo aplinką, ypač klientus ir jų elgseną.

Tyrimo metu panaudoti dviejų įmonių duomenys. Įmonės pasirinktos iš skirtingų sričių, atitinkamai skiriasi ir tam tikri kintamieji, kurie nusako klientą. Pirmasis aptariamas duomenų rinkinys yra telekomunikacijų kompanijos klientų duomenys. Iš viso, duomenų imtį sudaro 7043 klientai. Klientų naudojimosi įmonės paslaugomis vidutinė trukmė – 32 mėnesiai. Antrasis duomenų rinkinys priklauso „Premium“ ypatingųjų paslaugų klubui, kurio verslo veikla nėra detalizuojama. Duomenų imtį sudaro 10362 klientai. Pateikiami duomenys fiksuoti laiko intervale nuo 2006 iki 2013 metų. Klientų naudojimosi įmonės paslaugomis vidutinė trukmė – 24. Duomenų šaltinis – duomenų analizės ir mašininio mokymosi mokslininkų bendruomenės internetinė svetainė „Kaggle“[1].

Darbo objektas. Telekomunikacijų ir „Premium“ klubo įmonių klientai, kurie naudojasi įmonių teikiamomis paslaugomis.

Darbo tikslas. Padėti identifikuoti įmonės paslaugų ketinančius atsisakyti klientus dviem skirtingais modeliavimo būdais.

Darbo uždaviniai:

1. Aptarti literatūroje apžvelgiamą klientų lojalumą;
2. Pateikti klientų lojalumo publikuojamų tyrimų apžvalgą;
3. Aprašyti šiame darbe naudojamus lojalumo detekcijos ir išlikimo analizės metodus;
4. Atlikti pasirinktų duomenų pradinę analizę bei pritaikyti modeliavimo metodus;
5. Pateikti ir palyginti metodus remiantis gautais rezultatais.

Rengiant šį baigiamąjį magistro projektą, buvo naudojami lyginamasis, loginis, aprašomasis bei mokslinės literatūros analizavimo metodai. Tyrimo dalies uždaviniams atlikti buvo pasirinkta

[1] https://www.kaggle.com. Tinklalapis priklauso Google įmonei.

naudoti „R“[2] statistinio programavimo kalbą ir problemos sprendimus atitinkančius „R“ paketus bei jų funkcijas. Paketų sąrašas bei naudojimo prasmė aprašyti priede.

[2] „R“ programa https://www.r-project.org (naudota versija – 3.5.3).

1. Klientų lojalumo teorinė analizė

Paslaugų bei pagamintų galutinių produktų tiekėjai „verslas – klientui“ (angl. *Business – to – Customer*) aplinkoje privalo gerai suprasti savo klientų pobūdį, savybes bei poreikius. Žinant, jog poreikiai vis auga ir keičiasi, verslas privalo lanksčiai žiūrėti į permainas ir būti mobilus. Kliento sudominimas ir jo pasitenkinimas – bet kurios konkurencingos įmonės vienos iš pagrindinių siekiamybių. Todėl įmonės privalo didelį dėmesį skirti klientų išlaikymui ir jų lojalumo skatinimui (Hosseini, Maleki, & Gholamian, 2009). Taigi, pagrindinis įmonių tikslas – sukurti veiksmingą ir efektyvią metodiką, kuri būtų naudojama klientų lojalumo didinimui.

Pirmiausia reiktų apibrėžti sąlygas, pagal kurias klientas būtų vertinamas kaip nelojalus. Konkurencingoje kasdieninėje verslo aplinkoje susiduriama su problema, jog dauguma žmonių gali ir dažniausiai turi daugiau nei vieną vienos paslaugos tiekėją. Pavyzdžiui, kasdieninės bankininkystės srityje, klientas gali turėti einamąją sąskaitą viename banke ir būsto paskolą kitoje. Dauguma žmonių turi kelias einamąsias sąskaitas ir dažniausiai kai kurios būna nenaudojamos – vadinamosios „miegančios“ sąskaitos (Glady, Baesens, & Croux, 2008).

1.1. Klientų lojalumo išlaikymas ir išlaikymo strategijos

Šioje dalyje bus aptariama: kas yra klientų išlaikymo programa, kokie metodai taikomi ir kurie metodai efektyviausi. Klientų išlaikymas (angl. *Customer retention*) – veiklos ir veiksmai, kurių imasi įmonės, siekdamos padidinti savo klientų lojalumą (Galetto, 2015). Klientų išsaugojimo programos tikslas – padėti įmonėms išlaikyti kuo daugiau klientų, dažnai pasitelkiant klientų lojalumo iniciatyvas. Svarbu paminėti, jog klientų išlaikymo procesas prasideda nuo pirmojo kontakto su klientu, ir tęsiasi per visą laikotarpį, kol klientas naudojasi tam tikromis įmonės paslaugomis.

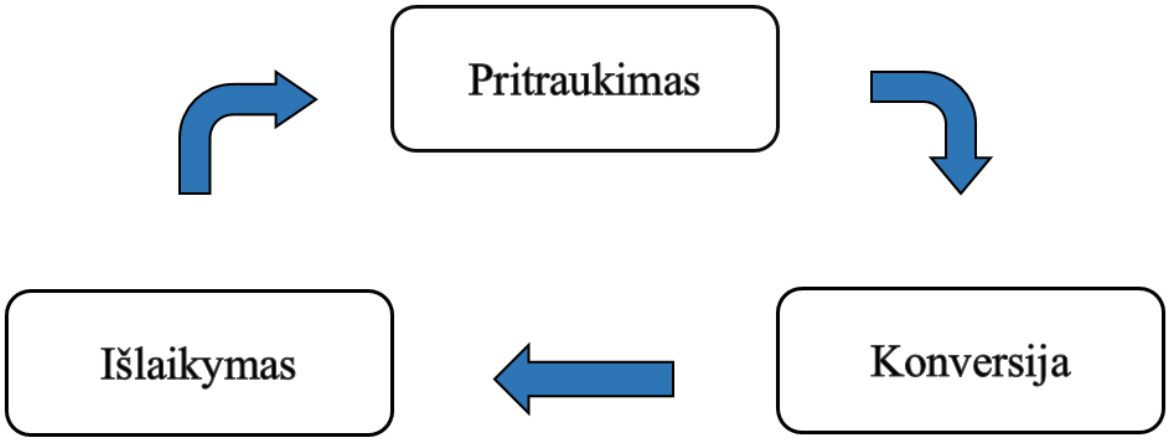
Kai kurios įmonės mano, jog naujų klientų pritraukimas ir sudominimas yra efektyvesnis būdas padidinti pajamas. Tačiau esamų klientų išlaikymo procesas dažnai būna greitesnis ir mažiau kaštų reikalaujantis procesas. Esamų klientų išlaikymas – pelningesnis procesas dėl šių priežasčių (Verbraken, 2013):

1. vidutiniškai net šešis kartus pigesnis procesas, nei naujų klientų pritraukimas;
2. pakankamai ilgą laiko tarpą esantys lojalūs klientai generuoja didesnį pelną, yra mažiau jautrūs pokyčiams, pasidaro pigiau juos aptarnauti bei jie gali tapti advokatais;
3. praradus klientus atsiranda papildomų išlaidų sumažėjus pardavimų kiekiui ir atsiradus poreikiui strategiškai persiorientuoti.

Ypač svarbu turėti lojalius klientus, kurie yra įsitikinę, jog įmonė jiems suteikia reikiamas paslaugas ir nesukelia abejonių ir poreikių keisti paslaugų šaltinio. Tokie klientai padeda įmonei reklamuotis, kada apie įmonės paslaugas kalbasi su kitais žmonėmis iš savo aplinkinių rato. Taip dalinantis savo patirtimi ir emocijomis apie įmonės teikiamas paslaugas, klientai daro savotišką reklamą įmonei. Tokie klientai vadinami advokatais.

1 pav. vaizduojamas supaprastintas klientų lojalumo skatinimo ciklas (Galetto, 2015). Pirmiausia klientai būna pritraukiami tam tikrais metodais (patys metodai aprašyti tolimesnėje šio poskyrio dalyje). Stengiamasi atkreipti dėmesį ir sudominti teikiamais pasiūlymais arba pasiūlymų pakeitimais. Tada susidomėję klientai būna pritraukiami ir ilgainiui tampa lojalūs klientai (įvyksta

konversija). Galiausiai klientų, kurie yra tapę lojaliais, elgseną, susijusią su įmonės teikiamomis paslaugomis, svarbu aktyviai stebėti ir atitinkamai reaguoti į pakitimus. Taip lojalūs klientai būna išlaikomi.



1 pav. Klientų lojalumo skatinimo ciklas

Aprašytas požiūris į klientą su lojalumo skatinimo ciklo elementais yra vadinamas santykių su klientais valdymas (angl. *Customer relationship management*) (toliau – CRM). CRM taip pat apibrėžiamas kaip procesų ir sistemų rinkinys, kuris padeda verslo strategijai kurti ilgalaikius ir pelningus santykius su konkrečiais klientais (Migueis, Van den Poel, Camanho, & Falcao e Cunha, 2012). Kiti šaltiniai CRM apibrėžia, kaip keturių elementų derinį: kliento identifikacija, pritraukimas, plėtojimasis ir išlaikymas (žiūrėti 2 pav.).

Klientų valdymo kontekste, duomenų analitikos metodai gali būti pritaikomi ir traktuojami, kaip verslą skatinantys procesai. Jie nukreipti į informacijos išgavimą apie klientą iš organizacijos turimų duomenų (Ngai, Xiu, & Chau, 2009). Kiekvienas iš keturių CRM elementų gali būti remiami įžvalgomis, padarytomis asociacijų taisyklių radimo, klasifikacijų, klasterizavimo, prognozavimo, regresijos, sekos atradimo ir vizualizacijos modeliais/metodais. Šiek tiek išsamiau aptariamas kiekvienas duomenų modeliavimo metodas:

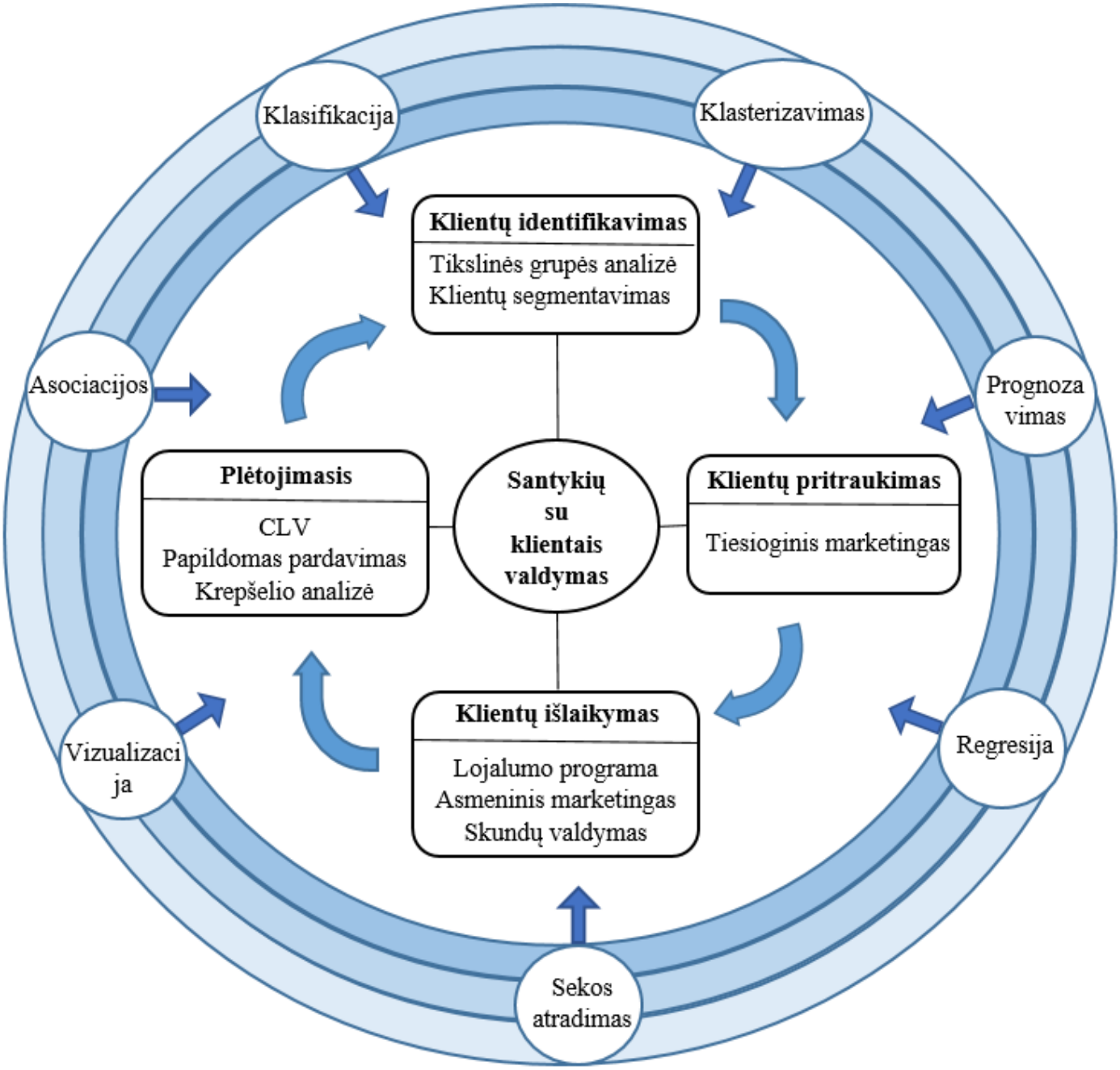
1. Asociacijos. Asociacijų metodo tikslas – nustatyti santykius tarp elementų, kurie egzistuoja kartu tam tikrame duomenų rinkinyje ir gali būti siejami tam tikrais požymiais. Kliento krepšelio analizė yra geriausias pavyzdys asociacijų modeliavimo metodo. Bendrųjų asociacijų modeliavimo priemonės yra statistiniai ir aprioriniai algoritmai.
2. Klasifikacija – tai viena iš populiariausių metodikų verslo duomenų analizėje. Juo siekiama sukurti modelį, pagal kurį būtų galima prognozuoti būsimą klientų elgseną, klasifikuojant duomenų bazės įrašus į keletą iš anksto nustatytų klasių, remiantis tam tikrais kriterijais. Klasifikacijai priskiriama detekcijos ir išlikimo modeliai. Šio darbo metu bus naudojami klientų klasifikavimo modeliai, todėl daugiausiai dėmesio bus skiriama šiai modeliavimo metodikai.
3. Klasterizavimas. Heterogeniškos populiacijos suskirstymas į homogeniškas grupes, pagal tam tikrus užfiksuojamus klientų elgsenos panašumus. Klasterizavimas skiriasi nuo klasifikavimo tuo, jog prieš pradedant analizę, nėra žinoma ir apibrėžta, kokie ir kiek klientų klasterių turi būti sudaroma.

4. Laiko eilučių prognozavimas remiasi turimais istoriniais duomenimis ir pagal užfiksuojamą populiacijos elgseną prognozuojamos būsimos ateities tendencijos. Rinkos paklausos prognozė – tipiškas prognozavimo modelio pavyzdys.
5. Regresija – tai statistinio įvertinimo metodas, kuris gali būti naudojamas numatyti kiekvienam duomenų objektui realiąją vertę ir svarbą, kada daromos bendrosios įžvalgos. Regresijos panaudojimas apima kreivių pritaikymą analizėje, prognozavimą, priežastinių ryšių modeliavimą ir mokslinių hipotezių testavimą apie kintamųjų tarpusavio ryšį.
6. Sekos atradimas apibūdinamas, kaip asociacijų ar nuolat besikartojančių elgsenos elementų atradimas tam tikrame laikotarpyje. Sekų pagrindinis tikslas – modeliuoti stebimo proceso sekas arba aptikti ir pranešti apie tendencijų nuokrypius laikotarpyje.
7. Vizualizacija itin reikalinga sudėtingų modelių išvesčių pateikimo supaprastinimui. Taip pat reikalinga, kad gauti rezultatai būtų aiškūs didžiajai daugumai vartotojų. Tai naudojama su kitais duomenų modeliavimo metodais, kad būtų galima geriau suprasti atliktus modelius ir lengviau interpretuoti rezultatus.

Kiekvienas iš išvardintų metodų turi savo stiprybių ir silpnybių tam tikrose srityse. Tačiau bet kokioje verslo situacijoje galima atrasti kelių modeliavimo metodų kombinaciją, kuri vienareikšmiškai pravers ir atlikus klientų tyrybą, padės išgauti vertingos informacijos. Kiekvienas iš modeliavimo metodų gali būti potencialiai panaudojamas bet kuriam iš CRM elementų. Šiek tiek išsamiau aptariama apie kiekvieną iš 2 pav. išvardintų verslo plėtros elementų (Ngai, Xiu, & Chau, 2009).

CRM prasideda nuo klientų identifikavimo. Tai dar vadinama klientų įsigijimo etapu. Šiame etape įmonė ieško savo tikslinės auditorijos, t. y. klientų, kurie tariamai atneš didžiausią pelną įmonei. Taip pat, šiame etape, yra atliekami pasitraukusių klientų tyrimai ir ieškoma metodų, kaip juos susigrąžinti. Klientų identifikavimo elementas apima tikslinės grupės analizę bei klientų segmentavimą. Tikslinės grupės analizės pagalba ieškoma pelningų klientų segmentų, analizuojant klientų pagrindines charakteristikas. Klientų segmentavimas apima visos klientų bazės suskirstymą į mažesnes klientų grupes, rinkodaroje dar vadinamais kanalais, kuriuos sudaro santykinai panašūs klientai kiekviename konkrečiame segmente.

Kitas elementas – klientų pritraukimas. Nustačius potencialių klientų segmentus, organizacijos gali nukreipti pastangas ir išteklius į tam tikrų klientų segmentų pritraukimą. Klientų pritraukimas dažniausiai siejamas su tiesioginiu marketingu, kuris yra klientų vartojimo skatinimo procesas. Pavyzdžiui siūlomos paskatos, kurių klientai galimai nesitiki. Norint užtikrinti tikslinės grupės klientų išlaikymą pravartu palaikyti ryšį su jais: atliktas trumpas skambutis (gali būti automatinis), išsiųstas laiškas. Taip pateikiami specialūs pasiūlymai, individualizuota reklama arba priminimai apie nepabaigtus pirkimus (Migueis et al., 2012).



2 pav. Duomenų gavybos metodų klasifikavimo sistema santykių su klientais valdyme

Pagrindinis CRM elementas – klientų išlaikymas. Klientų pasitenkinimas ir lūkesčių išpildymas, tai esminės klientų išlaikymo sąlygos. Išlaikymo elementas susideda iš lojalumo programų, asmeninio (angl. *One – to – One*) marketingo ir skundų valdymo. Visos šios sudedamosios dalys aprėpia daug skirtingų galimybių, kaip galima skatinti klientų išlaikymą. Toliau pateikiami klientų išlaikymo veiksmai, kurie sėkmingai naudojami šių laikų verslo plėtros planuose (Galetto, 2015):

- Nustatyti klientų lūkesčius. Nustatyti klientų lūkesčius kuo anksčiau ir juos šiek tiek paversti žemesniais, nei įmonės teikiamų paslaugų kokybė. Tokiu atveju pašalinamas neaiškumas dėl

teikiamų paslaugų kokybės ir užsitikrinama, jog įmonė visada įvykdys savo įsipareigojimus. Taip pat didėja galimybė teigiamai nustebinti vartotoją.

- Reikia tapti patikimu kliento patarėju. Įmonei yra svarbu turėti pakankamai kompetencijos savo verslo srityje, tam kad būtų galima tikėtis klientų pasitikėjimo ir taip didinti jų lojalumą.
- Privaloma kurti santykius, norint pasiekti klientų pasitikėjimą. Santykiai kuriami pritaikant bendras vertybes.
- Proaktyviai vykdyti klientų aptarnavimą. Tam tikrų paslaugų išankstinis įgyvendinimas padeda pašalinti problemas ir nesusipratimus dar prieš jiems įvykstant platesniu mastu.
- Patvaresniems ryšiams kurti padeda aktyvus dalyvavimas socialiniuose tinklalapiuose ir programėlėse, tokiose kaip „LinkedIn“, „Facebook“, „Instagram“ ir kt. Tokiais metodais klientams yra lengviau susisiekti su įmonės atstovais. Taip pat suteikiama erdvė ir galimybė pasidalinti savo patirtimi ir grįžtamuoju ryšiu su įmonė.
- Padaryti daugiau negu privaloma. Labai svarbu skirti papildomą dėmesį klientų poreikiams ir problemoms. Nors ir nedidelė asmeninio dėmesio išraiška gali atnešti daug vertės sukuriant glaudžius ryšius su klientais.
- Individualizuota paslauga teigiamai gerina klientų patirtį. Sulaukus išskirtinio aptarnavimo, klientui sustiprinamas ryšys su įmone ir įmonės prekės ženklu.
- Grįžtančiųjų klientų pakartotinai pardavimai. Dažnai pritraukus klientą ir jam atlikus pirminį pirkimą, dėmesys jam krenta. Svarbu padėkoti klientui už bendradarbiavimą ir įdiegti sistemą, pagal kurią klientas pakartotinai įsigytų įmonės prekių ar teikiamų paslaugų (Clay, 2017).
- Svarbu priimti neigiamą grįžtamąjį ryšį kaip padėką. Daugiau nei 9 iš 10 nepatenkintų klientų nesiskundžia, nesuteikia jokios informacijos įmonei ir galimybės pasitaisyti (Clay, 2017). Taigi, sulaukus, kad ir neigiamo grįžtamojo ryšio, įmonė gauna progą pasitaisyti ir tobulėti, geriau prisitaikydama prie klientų poreikių.
- Taip pat įmonė gali siūlyti specialius pasiūlymus jau modelio išskirtiems, galimai pasitraukiantiems, klientams. Tai gali būti speciali nuolaida kokiam pirkimui, papildoma dovana/priedas prie pirkinio. Klientai paprastai atsako teigiamai į tokį dėmesį, nes tai gali sukelti įvertinimo ir svarbumo jausmą (Migueis et al., 2012).

Galiausiai aptariamas paskutinis klientų valdymo proceso elementas – plėtojimasis. Tai apima nuolatinį sandorių intensyvumą, sandorių vertės ir kiekvieno iš klientų individualų pelningumo didinimą (Ngai et al., 2009). Plėtojimosi elementas apima CLV nustatymą, papildomus pardavimus (angl. *Up – selling*) ir krepšelių analizę. Papildomi pardavimai – tai skatinimo vartoti veikla, kuria siekiama padidinti bet kokių prekių ar paslaugų prekybą, glaudžiai susijusių su pagrindinėmis įmonės prekėmis ar paslaugomis. Krepšelių analizė tiria klientų prekes, ir bendrąsias pirkimo tendencijas, tam, kad galėtų pateikti racionalius papildomų prekių pasiūlymus pirkėjams.

Įmonės, kurios sukoncentruoja savo dėmesį į klientų išlaikymą ir pradeda klientų išlaikymo programą, dažnai atranda, jog tai – efektyvesnis procesas. Tai yra efektyvesnis procesas, nes paslaugos parduodamos klientams, kurie jau išreiškė susidomėjimą produktais anksčiau ir jų poreikiai geriau suprantami pačiai įmonei. Išlaikymas yra tvaresnis verslo modelis, ir tai – vienas iš pagrindinių įmonės sėkmės elementų, vedančių į tvarią plėtrą. Remiantis „Bain & Company“ atliktais tyrimais (Galetto, 2015), padidėjęs klientų išlaikymas 5 proc. gali padidinti pelną nuo 25 proc. iki 95 proc. Taip pat tyrimo rezultatai parodė, jog tikimybė, kad egzistuojantis klientas išliks lojalus yra 0,60 – 0,70, tuo tarpu kad naujas klientas išliks lojalus siekia tik 0,05 – 0,20.

1.2. Klientų pasitraukimo samprata

Pirminė klientų pasitraukimo samprata buvo grindžiama naudojamais produktais, vartotojų aktyvumu ir verslo logika paremtais, fiksuotais limitais. Buvo laikomasi metodikos, jeigu kliento aktyvumas nukrito žemiau nurodyto limito, tai sakoma, jog klientas nelojalus ir nebesinaudoja įmonės teikiamomis paslaugomis. Praktikoje yra susitinkama su tokios verslo logikos pavyzdžiais (Glady et al., 2008):

- klientai apibrėžiami, kaip pasitraukusiais, kada yra uždarę visas savo sąskaitas – būdingiausia finansines paslaugas teikiančioms įmonėms (Van den Poel & Lariviere, 2004);
- klientas, kurio pirkimo dažnumas yra retesnis nei visų likusių įmonės klientų pirkimo dažnumo vidurkis;
- privačių klientų bankinių paslaugų tiekėjų atžvilgiu, pasitraukęs klientas – tai klientas, kurio banko sąskaitoje yra mažiau disponuojamų pinigų nei tam tikra fiksuota suma.

Negalima kategoriškai teigti, jog išvardinti metodai yra visiškai nereikšmingi. Tačiau yra plėtojami klientų elgsenos vertinimo metodai, kurie turi kitokius principus ir padeda stebėti klientų veiklos raidą.

Pateikiamas pavyzdys, kodėl fiksuoto slenksčio principas ne visada yra tinkamas ir pritaikomas visai vartotojų populiacijai, kuri naudojasi tam tikros įmonės paslaugomis. Tarkim yra nustatyta tokia verslo taisyklė, jog visi klientai, kurie naudojasi įmonės paslaugomis arba apsiperka tam tikroje įstaigoje rečiau nei 5 kartus per metus yra laikomi nelojaliais. Esant tokiai taisyklei galima pateikti dviejų klientų visiškai skirtingą elgseną:

- Klientas paskutiniais metais pasinaudojo įmonės tiekiamomis paslaugomis tik 4 kartus, kada visada pasinaudodavo 5 kartus – toks klientas būtų laikomas nelojalus.
- Klientas, kuris kiekvienais metais vidutiniškai 100 kartų pasinaudodavo paslaugomis ir paskutiniais metais jo dažnumas nukrito iki 5 kartų. Toks klientas vis tiek būtų laikomas lojalus.

Toks pavyzdys parodo, jog tam tikro fiksuoto limito nustatytas ne visada yra tinkamas ir gerai parodo kliento elgsenos pokyčius. Visos verslo plėtros programos turėtų būti nukreiptos į klientus, kada klientai pozicionuojami, kaip pagrindinis išlaikymo objektas strategijoje. Kliento lojalumas turėtų būti nustatomas remiantis visa turima kliento elgsenos veikla. Taigi, galima teigti, jog naudingiau naudotis požiūriu, kuris nukreiptas į klientų elgsenos pokyčius.

Jeigu nustatomas toks limitas, kad klientas laikomas nelojalus tik tada, kada visai nebesinaudoja paslaugomis, tokiu atveju jau būtų per vėlu imtis išlaikymo veiksmų esamiems klientams. Tačiau kuriant modelį, įtraukiant tokius atvejus kaip nelojalius ir panaudojant naujiems klientams galima pastebėti nelojalius klientus. Galutinis tikslas – gauti pelną iš klientų, kurie yra linkę pasitraukti nuo įmonės, tačiau juos išsaugoti, pasitelkus tam tikrą išlaikymo programą. Galimas ateities uždarbis, arba kliento gyvavimo vertė yra svarbus indikatorius prieš pradedant vykdyti bet kokią klientų išlaikymo programą. Tai svarbu, nepaisant fakto, jog klientų potenciali ateities vertė nustatoma pagal įžvalgas, daromas iš klientų praeities elgsenos.

Klientas, kuris palieka įmonę, arba tam tikrą laiką nustoja naudotis įmonės teikiamomis paslaugomis, dažnai apibrėžiamas pagal jo istorinius duomenis, kurie susiję su išlaidomis. Tačiau

dviejų mokslininkų (Reinartz & Kumar, 2000) darbe, toks metodas buvo kritikuojamas ir pateikta pavyzdžių, jog pelnas ir kliento gyvavimo ciklas nebūtinai yra susiję. Norima pabrėžti, jog rinkodaros strategijose daugiausia dėmesio turėtų būti skiriama prognozuojamai finansinei grąžai. Šios rinkodaros pobūdžiui prognozuoti dar 1997 – 1998 metais buvo pateikta sistema, kurioje pirmą kartą buvo naudojama gyvavimo trukmės vertę (angl. *Customer lifetime value*) (toliau – CLV). Galiausiai buvo pateikiamos idėjos, jog įmonės pelnas, t. y. ir visos įmonės vertė, priklauso nuo visų klientų CLV sumos (Glady et al., 2008). Taigi, galima teigti, jog pirminė lojalumo samprata ir jo nustatymas buvo ganėtinai paprastas, tačiau per daug tiesmukiškas, kada nustatomi tam tikri, stacionarūs kliento elgsenos požymių rėžiai.

Tikslus tikimybės, jog klientas paliks įmonę, nustatymas yra svarbiausias elementas įvertinant CLV ir CRM. Kadangi CLV dažniausiai naudojamas klientų lojalumo modeliavime ir verslo plėtros strategijose, todėl modeliavimo tikslumas įgauna dar daugiau prasmės turint įtakos strateginių sprendimų priėmimui (Risselada, Verhoef, & Bijmolt, 2010).

Dažniausiai pasitraukiantis klientas nustatomas po fakto, jau po pasitraukimo nuo įmonės, sutarties nutraukimo ir pan. Toks klientų nustatymas turi trūkumą, jog iki pasitraukimo fakto, modeliavimui nesuteikia reikšmingos informacijos ir kartais netgi klaidina. Mažmeninės prekybos srityje, kada pelnas neskaičiuojamas kontraktais ir sutartimis, galimas kitoks pasitraukimo nustatyto metodas, dar vadinamas daliniu pasitraukimu (Migueis et al., 2012). Turint laiko eilutės duomenis, galima atlikti duomenų suskirstymą po tris mėnesius, t. y., ketvirčiais ir stebėti išleidžiamas kiekvieno iš klientų sumas. Jeigu kliento išleidžiama suma sumažėja iki 40 proc. praėjusio ketvirčio išleistos sumos, jis laikomas dalinai pasitraukęs ir nelojalus klientas.

1 lentelėje pateikiama dalinio pasitraukimo pavyzdžio skaitinė iliustracija. Pateikiami dviejų hipotetinių klientų lojalumo faktoriaus išvedimo pavyzdžiai pagal jų pirkimo kiekybinį pasiskirstymą finansine prasme. Iš pirmojo kliento pirkimų pasiskirstymo galime spręsti, jog klientas laikomas lojalus, nes nėra nei vieno ketvirčio, kuriame išleista vertė būtų mažesnė nei 40 proc. praėjusio ketvirčio vertės. Kalbant apie antrąjį klientą, matoma, jog jis laikomas nelojaliu klientu. Atsižvelgiant į trečiąjį ketvirtį, darome išvadą, kad visais vėlesniais laikotarpiais šis pirkėjas išleidžia mažiau nei 36,00 € (90,00 € * 0,4). Todėl darome prielaidą, kad šis klientas pasitraukė ketvirtąjį ketvirtį.

1 lentelė. Dalinio pasitraukimo pavyzdys

	1 ketvirtis	**2 ketvirtis**	**3 ketvirtis**	**4 ketvirtis**	**5 ketvirtis**
Lojalus klientas	100,00 €	80,00 €	90,00 €	40,00 €	60,00 €
Nelojalus klientas	100,00 €	80,00 €	90,00 €	**35,00 €**	**30,00 €**

Būtina priminti, jog pademonstruotas nelojalių klientų nustatymas yra pateikiamas kaip vienas iš keleto skirtingų būdų ir pritaikomas tik mažmeninės prekybos tinklams, kurie neįpareigoja kliento apribojimais ar sutartimis.

Taigi, buvo susipažinta su klientų išlaikymo programa, kaip tai įkomponuojama į santykių su klientais valdymą ir išvardinti pagrindiniai naudojami metodai. Sunku vienareikšmiškai teigti, koks metodas yra efektyviausias siekiant klientų išlaikymo. Tai labai priklauso nuo verslo krypties bei pačios įmonės verslo plėtros prioritetų. Tačiau, galima teigti, jog klientų išlaikymo programos gali būti pritaikomos bet kokiai verslo sričiai.

1.3. Lojalumo modeliavimo galimybių apžvalga ir įvertinimas

Ne paslaptis, jog lojalumo modeliavimas panaudoja įvairius statistinius modelius bei mašininio mokymosi (angl. *Machine learning*) metodus. Tradiciškai, lojalumo modeliavimas susideda ir dviejų etapų (Lemmens & Gupta, 2013):

1. statistinio modelio sudarymas, norint ištirti klientų lojalumą bei prognozuoti kiekvieno iš klientų tikimybę, jog klientas paliks įmonę arba nebesinaudos įmonės paslaugomis;
2. remiantis gautomis klientų lojalumo tikimybėmis, siūlomos įvairiausios paskatos turimiems klientams, kurie turi didžiausią polinkį pasitraukti.

Norint išplėsti pirmąjį etapą, būtina paminėti, jog vartotojų lojalumo tyrimui ir prognozei naudojami įvairiausi statistinio modeliavimo metodai, kurie minimi 1.1 poskyryje. Buvo atliktas lojalumo modeliavimo turnyras Duke universitete. Į turnyrą buvo pakviesti 44 dalyviai. Tyrimas atliktas norint surinkti daugiau informacijos apie vyraujančias lojalumo modeliavimo metodikas. Pusė respondentų buvo profesoriai ir kita pusė buvo praktikantai / statistinio modeliavimo naudotojai (Blattberg, Kim, & Neslin, 2008). Toliau pateikiamas naudotų modelių sąrašas:

- logistinė regresija;
- diskriminantinė analizė;
- Bayeso hierarchiniai modeliai;
- sprendimų medžiai;
- neuroniniai tinklai;
- atsitiktinis miškas;
- stochastinis savirankos agregavimo ir pastiprinimo metodas (angl. *Stochastic bagging and boosting method*);
- rizikos modeliai (angl. *Hazard models*).

Taip pat, dalis turimų duomenų buvo laiko eilutės, todėl buvo panaudotas modelis:

- paslėptieji Markovo modeliai.

Kaip matoma, buvo naudojamas platus skirtingų modelių diapazonas. Buvo naudoti parametriniai ir neparametriniai metodai, homoskedastiški ir heteroskedastiški kintamieji, laiko eilutės ir fiksuoto laiko duomenys. Tačiau, kad ir koks buvo didelis modelių pasirinkimas, kad ir kokiais įvairiausiais pjūviais duomenys buvo analizuojami, kiekvieno modelio tikslas buvo vienas – nuostolių funkcijos minimizavimas. Nuostolių funkcija buvo orientuota į tai, kaip sumažinti netinkamai klasifikuotų vartotojų procentinę dalį, t. y., sumažinti procentinę dalį lojalių klientų, kurie prognozuojami kaip nelojalūs ir nelojalių klientų dalį, kurie prognozuojami kaip lojalūs klientai. Galima į tai pažvelgti ir kitaip: buvo stengiamasi maksimizuoti procentinę dalį teisingai prognozuojamų lojalių ir nelojalių klientų dalį.

Toks uždavinio sprendimo požiūris yra visiškai statistiškai prasmingas ir teisingas, tačiau tokiu atveju prasilenkiama su įmonės poreikiais. Žvelgiant iš siekiamo pelno didinimo perspektyvos, ne kiekvienas klientas vienodai vertingas ir pelningas įmonės atžvilgiu (Lemmens & Gupta, 2013). Labai svarbu yra kliento vertinimas, atsižvelgiant į tai, kokį pelną įmonei jis gali suteikti. Taip pat, nustatomas koks kliento ilgaamžiškumas. Tai gali būti atliekama remiantis modeliavimo rezultatais ir rinkodaros įžvalgomis. Dėl šių priežasčių, galima teigti, jog pateiktas modeliavimo tikslas –

vykdant klientų išlaikymo programą, sumažinti procentinę dalį netinkamai klasifikuojamų vartotojų, nėra lygu pelno maksimizavimui. Toks požiūris tirtas ir kituose moksliniuose darbuose (Glady el al., 2008). Minėtame šaltinyje teigiama, jog vertėtų apsvarstyti standartinių statistinių metodų korektiškumą bei ieškoti alternatyvių skaičiavimo metodų. Taip pat galima modifikuoti esamus standartinius statistinius metodus pritaikant juos pelno maksimizavimo uždaviniui.

Aprašant antrąjį punktą, galima teigti, jog įmonei yra netikslinga investuoti į kiekvieną klientą, norint jį išsaugoti. Norint atrinki tik tam tikrą kiekį vartotojų iš populiacijos, reikia remtis duomenų segmentavimo ir klasterizavimo metodais. Tačiau, kaip jau buvo minėta anksčiau, tradicinis modeliavimo metodų požiūris neįvertina siekio maksimizuoti pelną. 2012 metais grupės mokslininkų (Verbeke, Dejarger, Martens, Hur, & Baesens, 2012) buvo pasiūlytas kitoks požiūris į šią dilemą. Pagrindinis klausimas išlieka, kokią klientų dalį yra tikslinga įtraukti į išlaikymo programą, norint išleisti kuo mažiau ir atitinkamai sulaukti kuo didesnio pelno. Taigi, norint pasirinkti, kokią dalį imtis turėtų sudaryti visos įmonės klientų populiacijos, buvo priimamas sprendimas remiantis keliais kriterijais:

- nustatant kliento vidutinę vertę (remiamasi CLV);
- nustatant vidutinį klientų atsako dažnį į išlaikymo programą.

Taip įvertinus modeliavimo rezultatus būtų galima nustatyti tikslinės klientų auditorijos dydį, kuris padėtų maksimizuoti įmonės pelną.

Apsvarsčius ir įvertinus klientų lojalumo detekciją ir įmonės pelno maksimizavimą, galima aptarti iš kokių elementų susideda pasiekiamas pelnas klientų išlaikymo programos vykdymo metu. Lentelėje (žiūrėti 2 lentelę) pateikiamas elementų sąrašas. Šis elementų sąrašas pirmą kartą aprašytas grupės mokslininkų (Neslin, Gupta, Kamakura, Lu, & Mason, 2006).

2 lentelė. Elementai, nuo kurių priklauso įmonės pelnas

	Aprašymas
1.	Kliento, kuris neįtraukiamas į išlaikymo programą, tolimesnė elgsena
2.	Kliento vertė įmonei
3.	Tikimybė, jog klientas atsakys teigiamai į išlaikymo programą ir toliau naudosis įmonės paslaugomis. Aprašomi klientai, kurie, pagal prognozavimą, nėra lojalūs
4.	Išlaikymo programos išlaidos klientui

Toliau aptariamas kiekvienas iš 2 lentelėje išvardintų punktų:

1. Pirmuoju elementu norėta pabrėžti, jog lojalių klientų tolimesnė elgsena yra labai svarbi. Nustatyta, jog investavimas į lojalius klientus nėra toks pelningas procesas, nei bandymas išlaikyti ne tokius lojalius klientus (Lemmens & Gupta, 2013).

2. Antrasis elementas. Išlaikymo programos finansinis pelnas priklauso nuo vertės, kurią kiekvienas klientas generuoja įmonei. Kitaip tariant, kliento galimybė būti pasirinktam priklauso ne tik nuo jo polinkio būti nelojaliam, tačiau ir nuo jo vertės įmonei. Ne visi klientai išleidžia tą pačią pinigų sumą, todėl prarandant didelę vertę turintį klientą yra didesnė finansinė žala nei prarasti mažos vertės klientą. Esant tokiam požiūriui kuo puikiausiai panaudojami detekcijos metodai remiantis pelnu, kurį kiekvienas klientas suteikia įmonei (Glady et al., 2008).

3. Kitas elementas – teigiamo kliento atsako į išlaikymo programą tikimybė. Išlaikymo programos finansinis pelnas priklauso nuo kliento atsako į išlaikymo veiksmą. Ne visi tiksliniai klientai būna sėkmingai išsaugoti net ir tada, kada įmonė juos ir nustato, kaip nelojalius klientus. Nepaisant to, jog įmonė rodo dėmesį ir investuoja į tam tikrų klientų išlaikymą, ne visada jie pasilieka, nors ir sulaukia paskatinimo. Nepaisant įmonės pastangų, kai kurie iš jų vis tiek nebesinaudoja įmonės paslaugomis ir pasitraukia. Dėl tokių priežasčių išlaikymo programa turi tiksliai apskaičiuoti tokias išlaidas ir atitinkamai suplanuoti išlaikymo programos planą, tikintis, jog klientai ne vien teigiamai reaguoja į išlaikymo planus.

4. Paskutinis elementas. Išlaikymo programos finansinis pelnas priklauso nuo programos išlaidų. Žinoma, įmonės pasiūlymai klientams suteikti tam tikros rūšies nuolaidas paslaugoms ar tiesiog suteikiamas kitoks dėmesys klientams, pačiai įmonei tai traktuojama, kaip papildomos išlaidos. Taigi, kuo daugiau klientų bus įtraukiami į išlaikymo programą, tuo daugiau bus pasiekta nelojalių klientų, tačiau imant per didelę imtį galima susidurti su didesnėmis išlaidomis nei pridedamu pelnu (Lemmens & Gupta, 2013).

Įgyvendinant 2 lentelėje esančius apibendrinamuosius žingsnius, verslo plėtros strategija jau įvykdyta. Turimas sukauptas žinias iš vykdomos išlaikymo programos, toliau gali būti taikomos modeliavimo tikslumui gerinti. Svarbu nustatyti, kiek laiko galima naudoti apskaičiuotą modelį (Risselada et al., 2010). Toks aktyvus modelio prižiūrėjimas, vertinimas ir tikslinimas padeda gerinti CLV prognozes, nes CLV priklauso nuo apskaičiuojamo lojalumo tikimybių. Tačiau iškyla klausimas, kiek laiko galima naujinti sukurtą modelį ir juo besąlygiškai naudotis taikant išlaikymo programas. Esamo modelio išlaikymas ir atnaujinimas taip pat reikalauja laiko ir finansinių išteklių. Tinkamų duomenų rinkimas, duomenų rinkinių paruošimas naudojimui ir modelio parametrų atnaujinimas gali būti itin sudėtingas procesas ir užtrukti pakankamai ilgai. Todėl yra svarbu atsižvelgti ir suderinti du aspektus: modelio tikslumą ir modelio sukūrimo efektyvumą. Norint padidinti modelio sukūrimo efektyvumą, tiriama, kaip modelis turi būti pritaikomas tiriant klientų tendencijas. Kitaip tariant, vykdomi modelio lankstumo tyrimai. Gaunamos įžvalgos iš tokių tyrimų padeda atrasti skirtumus tarp esamo modelio išlaikymo, tobulinimo ir naujo modelio sukūrimo.

Lyginant praėjusioje pastraipoje aptartą modelio pasirinkimą, pirmenybė teikiama detekcijos prognozavimo modeliui. Tačiau praktikoje matoma, jog dažniausiai detekcijos ir išlikimo modelių atnaujinimas ir priežiūra yra sudėtingas ir daug kaštų reikalaujantis procesas (Risselada et al., 2010). Svarbu paminėti, jog rinkos aplinka nuolat kinta. Globalios rinkos pokyčiai turi reikšmingos įtakos klientų elgsenai. Kadangi rinkos pokyčiai paprastai nėra įtraukiami į klientų detekcijos ar išlikimo modelius, sukurti modeliai nebetenka galios ir reikšmingumo. Dėl tokių priežasčių, daug dėmesio nėra skiriama ilgalaikiui modelių išlaikymui.

1.4. Optimalus tikslinės imties pasirinkimas

Šiame poskyryje pateikiamas tyrimas, aprašytas Harvardo verslo mokyklos atstovų (Lemmens & Gupta, 2013). Tyrimo metu analizuojami duomenys pateikti Duke universitete 2002 metais, kada buvo vykdomas detekcijos modeliavimo turnyras. Galiausiai pateikiamas detekcijos modelis, kuriame atsispindi visi toliau išvardinti elementai:

- suteikti kiekvienam klientui atitinkamą rangą pagal jo polinkį pasitraukti nuo įmonės;
- rodyti iniciatyvą ir vykdyti išlaikymo programą tam tikrai daliai klientų, kurie turi didžiausias rango reikšmes;

- vykdomas pasirinkimas, kuri dalis nuo visos klientų populiacijos bus įtraukiama į išlaikymo programą;
- atsižvelgti į tikimybę, jog klientas teigiamai sureaguos į išlaikymo programą ir išliks lojalus;
- atsižvelgti į kiekvieno iš klientų apskaičiuojamą CLV.

Atliktas tyrimas parodė, jog optimalus klientų rangavimas, atsižvelgiant į visus išvardintus elementus vidutiniškai padidina pelną 115 proc. palyginus su gaunamu pelnu iš tų pačių klientų atliekant kitokio pobūdžio verslo plėtros metodus. Panaudojama nuostolių funkcija, kuri akivaizdžiai orientuota į pelno optimizavimą. Pavyzdžiui, pritaikoma klientų išlaikymo programa, su aukščiau išvardintais modelio elementais, įmonei „Verizon Wireless“[3] atneštų 28 mln. pelno.

Tolimesnėje šio poskyrio dalyje trumpai aptariama atlikto tyrimo eiga, bei pateikiami tyrimo rezultatai (Lemmens & Gupta, 2013). Tyrinėjami duomenys buvo telekomunikacijų vartotojų duomenys, kurie naudojasi tos pačios įmonės paslaugomis 6 arba daugiau metų. Duomenys suskirstyti į tris imtis (apmokymo, vertinimo ir testavimo imtys). Kiekvienoje po 10000 vartotojų. Pirmoji, apmokymo imtis buvo subalansuota, t. y. 50 proc. lojalių ir 50 proc. nelojalių klientų. Vertinimo imtyje nelojalių klientų dalis siekė tik 1,68 proc., toks santykis atitinka visos turimos imties proporciją. Ši imtis buvo naudojama nustatyti tikslinės auditorijos optimalų dydį gaunant didžiausią pelną (b. atvejai 3 lentelėje). Galiausiai paskutinė – testavimo imtis turėjo 1,57 proc. nelojalių klientų ir šitos imties klientams buvo pritaikomi sudaryti modeliai ir pasirinktos tikslinės auditorijos dalys.

Modeliavimui buvo pasirinktas tikimybinis gradientinio didinimo modelis (angl. *Stochastic gradient boosting model*). Lentelėje pateikiami trys metodai, pagal kuriuos klientams suteikiami rangai. Nuo to priklausė, kurie klientai buvo įtraukiami į tikslinę auditoriją išlaiko programoje. Pirmasis metodas suteikia rangus klientams pagal nuostolių funkciją, kuri orientuota į kuo tikslesnį klientų prognozavimą (angl. *Accuracy and Precision*). Antrajame metode naudojama ta pati nuostolių funkcija orientuota į statistinį tikslumą, tačiau po modeliavimo klientams suteikiami nauji rangai remiantis modeliuotomis tikimybėmis, CLV ir išsaugojimo programos kaina. Trečiasis metodas gautas įtraukiant tikėtino pelno vertę į nuostolių funkciją modeliavimo metu.

Kaip ir anksčiau minėta, kiekvienam iš metodų pritaikomi fiksuotas ir optimalus tikslinės auditorijos dydžiai, parenkama atsakiusiųjų teigiamai į išlaikymo programą dalis ir skirtingos išlaikymo programos kainos. a. fiksuotas imties dydis pasirinktas 1,68 proc., nes tokia tikroji nelojalių klientų dalis tiriamoje duomenų imtyje. Ties kiekvienu iš šių parametrų skirtingų reikšmių kombinacijų apskaičiuotas tikėtinas pelnas, gaunamas iš 10000. 3 lentelėje pateikiami visi aptarti scenarijai ir apskaičiuotos tikėtinos pelno vertės:

[3] „Verizon Wireless“ – telekomunikacijų bendroje, teikianti belaidžių telekomunikacijų paslaugas ir produktus. Didžiausia telekomunikacijų tiekėja Jungtinėse Amerikos Valstijose.

3 lentelė. Išlaikymo programos pelno palyginimas skirtingiems modeliavimo metodams (Lemmens & Gupta, 2013)

		1. Neteisingo klasifikavimo nuostolis			**2. Neteisingo klasifikavimo nuostolis (suteikiant naujus rangus pagal pelną)**			**3. Kliento tikėtinas pelnas**		
		a. Fiksuotas dydis (1.68%)	b. Optimizuotas imties dydis		a. Fiksuotas dydis (1.68%)	b. Optimizuotas imties dydis		a. Fiksuotas dydis (1.68%)	b. Optimizuotas imties dydis	
Atsakiu-siųjų dalis	Išsaugojimo programos kaina per klientą	Išlaikymo pelnas ($)	Tikslinės aud. dalis (% nuo imties)	Išlaiky-mo pelnas ($)	Išlaikymo pelnas ($)	Tikslinės aud. dalis (% nuo imties)	Išlaiky-mo pelnas ($)	Išlaikymo pelnas ($)	Tikslinės aud. dalis (% nuo imties)	Išlaiky-mo pelnas ($)
10%	$70	-6339	0,00%	-	-6102	0,00%	-	-5573	0,00%	-
20%	$70	-5035	0,00%	-	-4217	0,00%	-	-3725	0,00%	-
30%	$70	-3730	0,41%	-487	-2333	0,00%	-	-1039	0,61%	**924**
40%	$70	-2425	0,81%	124	-448	0,00%	-	3266	1,81%	**3956**
50%	$70	-1121	0,81%	1036	1437	1,61%	134	5044	1,81%	**7027**
10%	$60	-5235	0,00%	-	-4956	0,00%	-	-4673	0,00%	-
20%	$60	-3919	0,00%	-	-3066	0,00%	-	-2170	0,61%	-514
30%	$60	-2602	0,41%	-220	-1177	0,00%	-	105	0,61%	**1335**
40%	$60	-1285	0,81%	664	713	1,61%	-316	3166	1,81%	**6513**
50%	$60	31	1,01%	1717	2603	1,61%	1253	7174	2,01%	**7432**
10%	$50	-4131	0,00%	-	-3810	0,00%	-	-3317	0,00%	-
20%	$50	-2803	0,00%	-	-1915	0,00%	-	-1443	0,61%	**499**
30%	$50	-1474	0,81%	273	-21	1,61%	-775	**2786**	1,61%	1494
40%	$50	-145	1,01%	1287	1874	1,61%	799	3532	2,01%	**7044**
50%	$50	1183	6,61%	3612	3769	1,61%	2372	**8330**	2,01%	7175
10%	$40	-3027	0,00%	-	-2664	0,00%	-	-2405	0,00%	-
20%	$40	-1687	0,41%	-127	-764	0,00%	-	102	0,61%	**906**
30%	$40	-346	1,01%	836	1135	1,61%	336	2985	2,21%	**3433**
40%	$40	995	6,61%	3026	3035	6,41%	2812	6704	2,01%	**8415**
50%	$40	2335	6,61%	8171	4935	8,01%	7474	**9486**	2,41%	9082
10%	$30	-1923	0,00%	-	-1518	0,00%	-	-1271	0,61%	**58**
20%	$30	-571	0,81%	386	387	1,61%	-134	1637	2,21%	**1821**
30%	$30	782	6,61%	2371	2291	6,41%	2154	4125	2,01%	**5356**
40%	$30	2135	6,61%	7551	4196	8,01%	7148	7850	2,21%	**9366**
50%	$30	3487	6,61%	12730	6101	8,01%	**13045**	9676	2,21%	11167
Vidutinis pelnas 10000 klientų		-1474		1718	-21		1452	2014		**3700**

Matoma, jog didžiausias pelnas gaunamas kada tikslinės imties dydis yra optimizuojamas ir pats klasifikavimo modelis apskaičiuoja klientų tikimybes remiantis pelnu grįsta nuostolių funkcija.

Pasiekiamas vidutiniškai daugiau nei dvigubai didesnis pelnas naudojant tokią metodiką nei klasifikuojant pagal statistinę nuostolių funkciją ir suteikiant naujus rangus klientams pagal paskaičiuotą tikėtiną pelną (2 metodas). Taigi, nors pelnu remtas klientų klasifikavimas nėra tiksliausias metodas atskirti nelojalius klientus nuo lojalių, tačiau yra pelningesnis pasirenkant tikslinę auditoriją ir apliekant klientų išlaikymo programą.

Taigi, galima teigti, jog atsižvelgus į klasifikavimo modelių rezultatus, fiksuojama, jog optimizuoti modelių rezultatus, klasifikavimo tikslumo atžvilgiu, nėra vienas ir tas pats, kas optimizuoti pelną iš išlaikomų klientų tikslinės grupės. Dėl to svarbu ištirti duomenis ir geriau suprasti klientų su kuriais bendradarbiaujama įpročius ir elgseną.

Ne visi duomenų modeliavimo metodai veikia vienodai efektyviai tam tikram duomenų rinkiniui, vieni algoritmai prisitaiko prie labai didelių duomenų rinkinių lanksčiau, kiti geresnius rezultatus parodo sąlyginai mažesniems rinkiniams. Taip pat pasirinkimui, kurie modeliai buvo išbandomi, turėjo įtakos modelių gaunamų rezultatų paaiškinamumas. Nors ir šio darbo metu tiriamos ne laiko eilutės, tačiau išlikimo modeliai gali kuo puikiausiai pasitarnauti klientų lojalumo tyrime.

1.5. Atliktų tyrimų apžvalga

4 lentelė. Susijusių detekcijos uždavinių publikacijos4 lentelėje matoma, jog klientų išlaikymo problema ir lojalumo prognozavimas yra aktuali įvairiuosiuose srityse. Taip pat matoma, jog klientų ir kintamųjų skaičius yra itin didelis, todėl yra galimos įvairios išankstinio apdorojimo procedūros, kurios padeda apdoroti duomenų imtį, sumažinti kintamųjų skaičių bei parinkti tam tikrą klientų dalį, kuri vertinama, kaip galimai svarbesnė įmonei. Išankstinis duomenų apdorojimas yra ne tik klientų segmentavimas, tačiau tai gali būti ir žvalgomosios analizės metu pastebėti nereikalingi kintamieji. Taip pat kintamieji, kurie turi tą pačią informacinę vertę, t. y. jie koreliuoti. Tai, kad duomenų rinkinyje užfiksuotų savybių skaičius yra sąlyginai didelis, nurodo, jog dažniausiai ne visos savybės yra informatyvios bei svarbios detekcijos modelių sudarymo procesui (Tsai & Lu, 2010). Kada duomenų rinkiniai pakankamai dideli, detekcijos modeliai dažniausiai nesuteikia geriausio prognozuojamo tikslumo, jeigu neatliekamas išankstinis duomenų apdorojimas.

Pateiktame susijusių tyrimų sąraše, detekcijos metodai daugiausiai grindžiami sprendimų medžiais, atsitiktiniais miškais bei neuroniniais tinklais. Iš statistinių regresinių modelių, populiariausia logistinė regresija. Taip pat, tarp pasirinkimų atsiranda ir atraminių vektorių metodas, tačiau dažniausiai jis neparodo geriausių prognozavimo rezultatų ir reikalauja itin daug resursų atliekamiems skaičiavimams. Tam tikri tyrimai orientuojasi į hibridinių modelių sudarymą, kurie grindžiami kelių detekcijos metodų mokymosi derinimu (Pendharkar, 2009) (Tsai & Lu, 2009). Apskritai, kelių metodų kombinacija derinama nuosekliai, t. y. pirmasis metodas būna panaudojamas išankstiniam duomenų apdorojimui – kintamųjų parinkimui, išskirčių atpažinimui ir kitai panašiai duomenų analizei. Tuo tarpu kitas metodas paremtas detekcijos procesu, kada jo apmokymo imtis gaunama iš pirmojo metodo išvestų, apdorotų duomenų.

Taip pat aptariami populiariausi tyrimų metodų įvertinimo būdai. Nuo pasirinkto detekcijos metodo įvertinimo būdo priklauso kiekvieno tyrimo galutinė išvada. Dažniausiai tyrimuose neatkreipiamas dėmesys į svarbumo pasiskirstymą tarp pirmos rūšies klaidos ir antros rūšies klaidos. Labai svarbu įvertinti, kiek nelojalių klientų yra neteisingai prognozuojamų, kaip lojalūs klientai, nes kiekvieno kliento lojalumo tyrimo vienas iš esminių siekių yra išsaugoti kuo daugiau nelojalių klientų. 4 lentelėje minimi tyrimai atsižvelgia į prognozavimo bendrąjį tikslumą (angl. *Accuracy*), kai

kur akcentuojamas ir pranašumo (angl. *lift*) rodiklis, t. y. atrasti kuo didesnę procentinę dalį nelojalių klientų su kuo mažesne pasirinkta išlaikymo programos tiksline klientų imtimi. Be to, didžioji dauguma tyrimų yra atliekami pasiremiant kryžminį patikrinimą (angl. *Cross – validation*). Tai reiškia, jog tyrimai atliekami keletą kartų su apmokymo ir tikrinimo (angl. *Train and test*) imtimis, naudojantis vis kita visos duomenų imties dalimi.

4 lentelė. Susijusių detekcijos uždavinių publikacijos

Šaltinis	**Naudoti metodai**	**Duomenų dydis**	**Industrijos sritis**	**Pasiektas tikslumas**
(Buckinx & Van den Poel, 2005)	Sprendimų medžiai, logistinė regresija, neuroniniai tinklai	158884 klientai ir 61 kintamasis	Mažmeninė prekyba	0,804
(Qi, Zhang, Zhang, & Shi, 2006)	Sprendimų medžiai, logistinė regresija	42000 klientų ir 93 kintamieji	Mobiliųjų telefonų telekomunikacijų	0,652
(Burez & Van den Poel, 2007)	Atsitiktiniai miškai, logistinė regresija	143198 klientai ir 81 kintamasis	Televizijos paslaugos	0,879
(Coussement & Van den Poel, 2008)	Atsitiktiniai miškai, logistinė regresija, atraminių vektorių metodas	62500 klientų ir 82 kintamieji	Abonementų duomenų bazė	0,891
(Anil Kumar & Ravi, 2008)	Atsitiktiniai miškai, logistinė regresija, atraminių vektorių metodas	90000 klientų ir 82 kintamieji	Laikraščio abonementai	0,956
(Glady, Baesens, & Croux, 2008)	Sprendimų medžiai, neuroninis tinklas, logistinė regresija	10000 klientų	Finansinės paslaugos	0,860
(Xie, Li, Ngai, & Ying, 2009)	Neuroniniai tinklai, sprendimų medžiai, atraminių vektorių metodas, atsitiktinai miškai	20000 klientų ir 27 kintamieji	Finansinės paslaugos	0,932
(Pendharkar, 2009)	Neuroninis tinklas	195956 klientai	Mobiliojo belaidžio ryšio paslaugos	0,974
(Tsai & Lu, 2009)	Hibridinis neuroninis tinklas	51306 klientai	Telekomunikacijų	0,943
(Yu, Guo, Guo, & Huang, 2011)	Neuroniniai tinklai, sprendimų medžiai, atraminių vektorių metodas	50000 klientų ir 27 kintamieji	Komercinė internetinė svetainė	0,891
(Chen, Fan, & Sun, 2012)	Įvairių branduolių atraminių vektorių metodas, neuroninis tinklas, tiesinė regresija, Cox regresija	8842 klientų ir 79 kintamieji	Telekomunikacijų ir kasdienių prekių parduotuvės	0,967

Aptarti publikuojami tyrimai, kuriuose analizuojami detekcijos metodai padedantys įvardinti įmonių klientų elgseną ir atskirti lojalius klientus nuo nelojalių. Skirtingi metodai, veikia skirtingai įvairių rūšių kintamuosius, vieniems metodams būtini skaitiniai kintamieji, kada metodo algoritmas netinkamai priima kategorinius kintamuosius. Tuo tarpu kiti metodai yra labiau automatizuoti, kada

patys savo algoritmuose turi parametrų derinimo metodų. Kitiems nėra automatizacijos principų, tokiu atveju reikia optimalių metodo rezultatų siekti rankiniu derinimo būdu.

2. Tyrimų metodai

Kaip ir buvo minėta, šio darbo tyrimo metu naudojamos detekcijos ir išlikimo modeliavimo metodikos ir jiems pritaikomi skirtingi modeliai. Šioje dalyje aptariami naudoti modeliai ir jų rezultatų vertinimui skirti gerumo matai.

2.1. Detekcijos uždavinys

Dažniausiai naudojami algoritmai klientų lojalumo detekcijos uždaviniuose yra logistinė regresija ir atsitiktiniai miškai (Risselada et al., 2010). Tyrimo metu buvo išmėginti minėtieji du modeliai bei atraminių vektorių modelis. Toliau pateikiamas kiekvieno iš modelių aprašymas.

2.1.1. Logistinė regresija

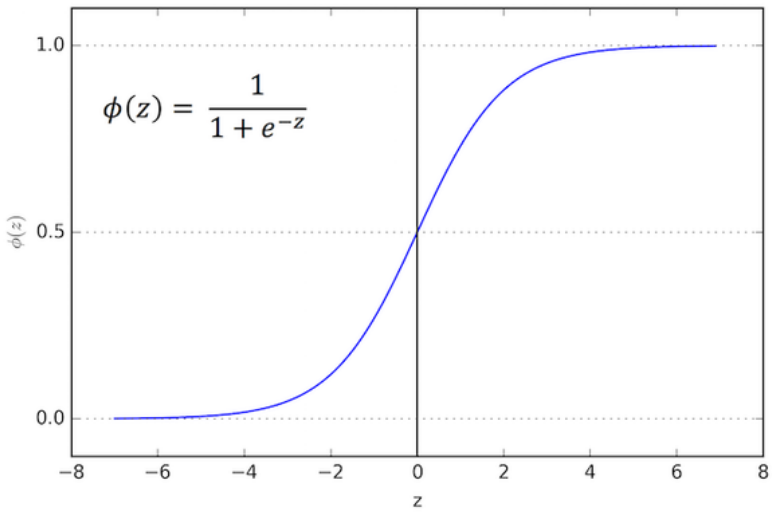
Vienas iš plačiausiai naudojamų ir lengviausiai interpretuojamų modelių yra logistinė funkcija. Tai matematinis modelis sudaromas ne pačiam priklausomam kintamajam o jo tikimybių santykio logaritmui (Čekanavičius & Murauskas, 2014):

$$ln\frac{P(Y=1)}{P(Y=0)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \quad (1)$$

Kitas modelio užrašymas atrodo taip:

$$P(Y=1) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}},$$
$$P(Y=0) = 1 - P(Y=1),$$

čia $z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots$, kur koeficientų $\beta_0, \beta_1, \beta_2, \beta_3, \ldots$ reikšmės nėra žinomos, todėl jų įverčiai yra apskaičiuojami logistinės regresijos proceso metu. O $X_1, X_2, \ldots$ yra turimų duomenų kintamieji, kurie apibūdina klientų charakteristiką. 3 pav. pateikiamas logistinės regresijos pavyzdys, kaip paskaičiuojama tikimybė, kuri priklauso nuo kintamojo z.



3 pav. Logistinės regresijos pavyzdys

Kintamųjų X_i gali būti daug ir skirtingų rūšių: įvairių intervalų skaitiniai arba kategoriniai kintamieji. Taip pat jų skalės gali labai skirtis – vienų kintamųjų reikšmės skaičiuojamos tūkstančiais o kitų vienetais, todėl negalima teigti, jog modeliui didžiausią įtaką turi tas kintamojo įvertis, kurio koeficientas absoliučiu didumu didžiausias (Čekanavičius & Murauskas, 2014). Norint palyginti skirtingų koeficientų įtaką modeliui, duomenis galima arba normalizuoti (stengiantis suvienodinti su

normaliuoju skirstiniu) arba pasinaudojus tam tikrais statistiniais testais, kurie pritaikyti įvertinti kiekvieno iš kintamųjų daromą įtaką modeliui atskirai.

Tyrimo metu išmėginta keletas logistinės regresijos modifikacijų. Buvo bandoma atsižvelgti į nevienodą kiekį skirtingų klasių klientų, suteikiant lojaliems ir nelojaliems atitinkamus svorius modeliuojant. Tai dar vadinama daugumo sumažinimo metodas (angl. *Majority down – sampling*). Siekiant, kad logistinė regresija gautų nepaslinktus įverčius, labiau tinkami skaitiniai kintamieji normalizuojami, kada iš skaitinių reikšmių atimamas kintamojo vidurkis ir dalinama iš standartinio nuokrypio. Keičiami ir kategoriniai kintamieji. Kategorinius kintamuosius, kurie turi daugiau nei 2 kategorijas būtina išskaidyti į binarinius kintamuosius, kitaip tariant sukurti fiktyvius kintamuosius (angl. *Dummy variables*). Taip pat logistinė regresija buvo sudaroma pasitelkiant „R“ programos dvi skirtingas funkcijas[4], kada viena iš jų suskaičiuoja visų kintamųjų įverčius klientų lojalumo atžvilgiu, o kita funkcija ieško optimalios kintamųjų kombinacijos prognozuojamo lojalumo tikslumo atžvilgiu.

2.1.2. Atsitiktiniai miškai

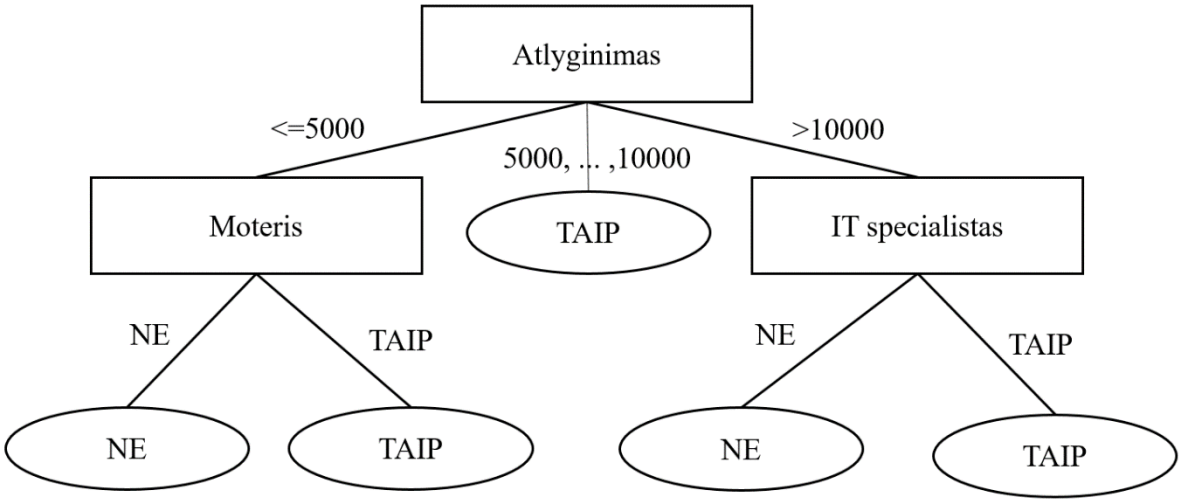
Kitas metodas yra atsitiktiniai miškai (angl. *Random forest*). Metodas priskiriamas „ensemble“ metodams. Šio metodo veikimo principas – sudaryti daug sprendimų medžių, paprastai nuo 100 iki 1000 sprendimų medžių iš apmokymo imties. Tam kad sudarytieji medžiai nebūtų identiški medžių kūrimo metu naudojami atsitiktinio pasirinkimo procesai:

- medžio sudarymui naudojama tik dalis atsitiktinai parinktų duomenų;
- kintamieji parenkami taip pat atsitiktinai, neprioretizuojant juos prieš sudarant kiekvieną medį.

Sprendimų medžio algoritmas yra linkęs persimokinti, t. y. prisitaiko prie apmokymo imties taip joje spėdamas lojalius ir nelojalius klientus su aukštu tikslumu, tačiau gaunamas mažesnis rezultatų tikslumas, kada pateikiama nauji duomenys apmokytam modeliui. Ši problema pakankamai sėkmingai yra sprendžiama su atsitiktinių miškų algoritmu, kada kiekvienam sprendimų medžiui yra paduodama vis kitokia apmokymo imties dalis. Galiausiai visiems sprendimų medžiams apsimokius galutinės medžių klasės parenkamos daugumos principu. Kartu tai leidžia nustatyti ir tikimybę su kuria klientas yra priskiriamas tam tikrai klasei.

Kadangi atsitiktiniai miškai naudoja sprendimų medžių algoritmą, aptariamas ir šis metodas. 4 pav. pavaizduojamas supaprastintas elementariausias sprendimų medžio pavyzdys, kuris klasifikuoja įmonės darbuotus pagal jų atlyginimą, lytį ir darbo poziciją (IT specialistas ar ne). Parinktos tam tikros kintamųjų reikšmių ribos, nuo kurių priklauso, kaip darbuotojai bus klasifikuojami.

[4] Naudotos glm ir cva.glmnet funkcijos logistinei regresijai įvertinti.



4 pav. Darbuotojų lojalumo sprendimų medžio pavyzdys

Renkantis kintamuosius, pagal kuriuos bus sudaromas sprendimų medis, svarbu įvertini padalinimo gerumo savybes, t. y. kaip gerai atskiriami lojalūs klientai nuo nelojalių. Tam naudojami įvairiausi grynumo (angl. *Purity*) matai. Aptariamas tik CART (angl. *Classification and regression tree*) algoritmas, nes juo remiantis buvo sudarinėjami sprendimų medžiai šio tiriamojo darbo detekcijos dalyje. CART algoritmas kintamojo parinkimui naudoja Gini indeksą, kuris matuoja imties klasių grynumą:

$$Gini(D) = \sum_{i=1}^{K} p_i(1 - p_i) = 1 - \sum_{i=1}^{K} p_i^2, \quad (2)$$

čia D – duomenų imtis, K – skirtingų klasių kiekis (detekcijos atveju tai lygu 2) ir p_i yra i – tosios klasės dažnis/tikimybė. Toliau apskaičiuojamas Gini indeksas kiekvienam iš kintamųjų. Tarkim, norima apskaičiuoti kintamojo A Gini indeksą, tada skaičiuojamas Gini indeksas kiekvienai kintamojo A kategorijai m.

$$Gini_A(D) = \sum_{j=1}^{m} \frac{|D_j|}{|D|} Gini(D_j), \quad (3)$$

kur D_j yra j – tosios kategorijos dalis imtyje D. Galiausiai kiekvienam iš kintamųjų paskaičiuojamas skirtumas nuo bendrojo Gini indekso:

$$\Delta Gini_A = Gini(D) - Gini_A(D). \quad (4)$$

Apskaičiuotas didžiausias Gini indeksas nurodo didžiausią kintamojo grynumą ir taip nustatomi kurie kintamieji yra vertingesni klasifikuojant lojalius ir nelojalius klientus.

Jeigu detekcijos uždavinio sprendimo duomenys būna skaitinių kintamųjų, tada parenkamos tam tikros skaitinės vertės (dažniausiai viena vertė), pagal kuria(s) visos kintamojo reikšmės priskiriamos tam tikrai kategorijai (žiūrėti 4 pav.). Norint surasti tą skaitinę vertę naudojami įvairūs metodai nuo paprasčiausių, kada skaidymui naudojama mediana, iki sudėtingesnių, kada papildomai išbandomi įvairiausi skaidymo variantai ir parenkama optimali reikšmė pagal tam tikrą pasirinktą kriterijų.

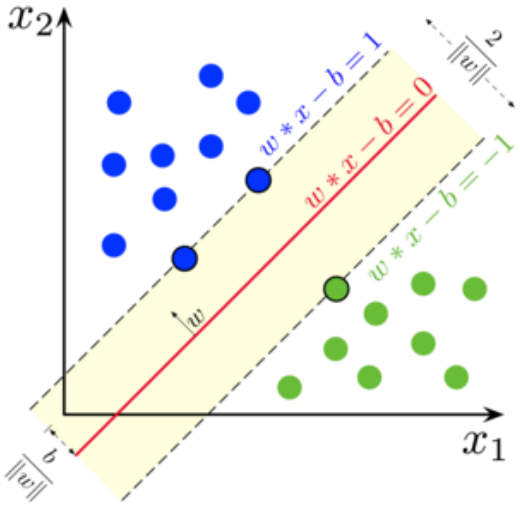
Kaip jau buvo minėta sprendimų medžiai yra lengvai interpretuojami bei pakankamai lengvai susitvarko su duomenų trūkstamomis reikšmėmis ir išskirtimis. Atsitiktiniams miškams neatliekamas duomenų standartizavimas, nes sudaromieji sprendimų medžiai kuo puikiausiai susidoroja su katorginiais kintamaisiais, o tuo pačiu ir skaitinius suskirsto į kategorijas pagal tam tikrą parenkamą kintamojo reikšmę. Taip pat klasių disbalansui sumažinti naudojamos modelių modifikacijos: daugumo sumažinimo metodas bei kaštams jautrus mokymasis (angl. *Cost sensitive learning*). Išmėgintos dvi skirtingos atsitiktinių miškų parametrų optimizavimo funkcijos, kurios pasitelkiant tam tikrus algoritmus ieško optimalių mazgų kiekio bei kintamųjų kiekio kuris paduodamas kiekvienam iš mazgu atsitiktinai.

Metodas taip pat gerai susitvarko su skaitiniais ir kategoriniais kintamaisiais. Dėl šių priežasčių naudojamo šio metodo modifikacija – atsitiktiniai miškai, kurie išsprendžia persimokinimo problemą pavieniuose sprendimo medžiuose.

2.1.3. Atraminiai vektoriai

Galiausi aprašomas paskutinis metodas detekcijos uždaviniui spręsti – atraminių vektorių metodas (angl. *Support vector machine*). Šio metodo idėja tokia, jog sprendžiant klasifikavimo uždavinį, apmokymo duomenų imtį teoriškai galima perskirti hiperplokštuma. Tada, atraminių vektorių metodas atranda hiperplokštumą, kuri suranda didžiausią atstumą nuo artimiausių abiejose klasėse esančių taškų (žiūrėti 5 pav.).

Vis dėlto realių duomenų atvejuose, dažniausiai neįmanoma surasti hiperplokštumos, kuri



5 pav. Atraminių vektorių metodo pavyzdys

visiškai atskirtų stebėjimus į atitinkamas grupes. Todėl metodas modifikuojamas, kada leidžiama peržengti hiperplokštumos ribas su tam tikra skaitine „bauda“ C. Tarkime, jog binarinio klasifikavimo kintamasis, ar lojalus klientas ar nelojalus, žymimas $y_i \in \{0, 1\}$, kur $i = 1, 2, \dots, N$ o $N-$ klientų skaičius. Klientus nusakantys kintamieji žymimi $x_i \in \mathbb{R}^p$. Metodo optimizavimo problema susiveda į Lagranžo funkcijos minimumo radimą:

$$\mathrm{L}(\alpha) = \frac{1}{2}\alpha' Q \alpha - e'\alpha, \qquad (5)$$

kur α tai vektorius, kuriam galioja savybės $0 \leq \alpha \leq C$ ir $y'\alpha = 0$. Vienetinis vektorius pažymėtas e. O Q yra matrica $Q_{ij} = y_i y_j K(x_i, x_j)$, kurioje $K(x_i, x_j) = \varphi(x_i)'\varphi(x_j)$ yra Kernelio funkcija. Dėl branduolio funkcijos taikymo, nereikia žinoti tikslios transformacijų funkcijos φ išraiškos.

Atraminių vektorių metodui pritaikomos įvairios branduolio K modifikacijos. Tyrimo metu naudotas tik tiesinis branduolys:

$$K(x_i, x_j) = x_i' x_j, \tag{6}$$

tokia branduolio modifikacija greičiausiai atlieka skaičiavimus ir rodo itin gerus detekcijos rezultatus pagal tikslumą ir jautrumą klientų lojalumo tyrimuose (Suchacka, Skolimowska – Kulig, & Potempa, 2015). Atraminių vektorių metodas labiausiai tinkamas su teisinio Kernel branduolio modifikacija. Dėl šios priežasties skaičiavimų metu buvo optimizuojamas tik baudos parametro reikšmė C. Metodo trūkumas, kad procesas ganėtinai lėtas lyginant su kitais naudotais modeliavimo metodais šio tyrimo metu.

2.2. Išlikimo analizė

Šiame poskyryje išsamiau aptariami išlikimo modeliavimo metodai, kurie buvo taikomi tyrimo metu. Išlikimo analizei tirti buvo naudojami Cox proporcingumo rizikos (angl. *Cox proportional – hazard*) modelis, atsitiktinių išlikimo medžių metodas bei keletas šių modelių modifikacijų, kurios aptariamos tolimesniuose poskyriuose.

2.2.1. Cox proporcingumo rizikos modelis

Anksčiau, kada išlikimo modeliai buvo sukurti, jie buvo naudojami žmonių įvairių ligų ir gydymo metodų efektyvumo tyrybai. Pradinė išliko modelio mintis buvo statistiškai nustatyti išgyvenimo tikimybę, fiksuojant stebėjimų informaciją tik apie tai, kiek laiko buvo išgyventa ir ar buvo išgyventa. Tačiau, išlikimo analizės metodai plačiai naudojami ir kitose srityse, kur galima pritaikyti klasifikavimo metodus. Daugeliu atveju pageidautina išlikimo modeliavimą susieti su kitais aiškinamaisiais kintamaisiais, tokiais kaip amžius, lytis, kliento išlaidos ir kiti kintamieji, kurie padidintų modeliavimo tikslumą. Tokiems uždaviniams spręsti naudojamas Cox proporcingumo rizikos modelis (Weathers & Cutler, 2017).

Cox modelis paprastai apibūdinamas kaip pusiau parametrinis, nes įverčiai, susijęs su aiškinamais kintamaisiais yra parametriniai, o išlikimo funkcijos įvertis gaunamas ne parametriniu metodu, taip pat nėra prielaidos apie pagrindinį išlikimo funkcijos pasiskirstymą. Modelio kintamieji gali būti realieji skaičiai arba kategoriniai. Tačiau, jei kategorinį kintamąjį sudaro daugiau nei 2 kategorijos, rekomenduojama konvertuoti į dvejetainių klasių kintamuosius (angl. *Dummy variables*), kad būtų galima atlikti regresiją.

Išlikimo analizėje svarbu atkreipti dėmesį į ryšį tarp kintamųjų ir išlikimo pasiskirstymo, ir tai galima padaryti nurodant log – rizikos modelį (Crumer, 2011). Pavyzdžiui, parametrų modelio rizikos funkcija, pagrįsta eksponentiniu pasiskirstymu, aprašoma taip:

$$\log h_i(t) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}, \tag{7}$$

ši lygybė ekvivalenti:

$$h_i(t) = \exp(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}), \tag{8}$$

čia i – tam tikras klientas, x_i – kintamųjų vektorius, apibūdinantis i – tąjį klientą, α – tai reikšmė, nusakanti pradinę modelio riziką ir β – įverčių vektorius, kuris nustatomas pagal dalinio tikėtinumo funkciją (angl. *Partial likelihood*). Cox modelyje β įverčiai yra gaunami paprastos regresijos parametriniu būdu, tačiau pradinė modelio rizika lieka neapibrėžta. Modelyje $\alpha(t) = \log h_0(t)$, kada visi kiti kintamieji prilyginami nuliui. Įstačius šią išraišką į lygtį (8) gauname:

$$h_i(t) = h_0(t)\exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}). \quad (9)$$

Cox modelis vadinamas pusiau parametriniu (angl. *semi – parametric*). Tarkime, jog imtyje turime visiškai skirtingus du stebėjimus i ir j, kurių kintamųjų vertės skirtingos. Apskaičiuojamos tiesinės prognozės šiems dviem stebėjimams $\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$ ir $\eta_j = \beta_1 x_{j1} + \beta_2 x_{j2} + \cdots + \beta_k x_{jk}$. Tada santykis tarp dviejų rizikos funkcijų gaunamas:

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t)\exp(\eta_i)}{h_0(t)\exp(\eta_j)} = \frac{\exp(\eta_i)}{\exp(\eta_j)} = \exp(\eta_i - \eta_j). \quad (10)$$

Kaip matoma, tokia išraiška nepriklausoma nuo laiko kintamojo. Taigi, rizikos funkcijos santykis yra proporcingas prognozuojamų kintamųjų įverčių skirtumams.

Taip pat, modelio rezultatams interpretuoti naudojamas log – rank testas. Šis testas lygina dviejų stebėjimų/klientų grupių išlikimų laikus. Apskaičiuojami tikrieji ir prognozuojami išlikimo dažniai skirtinguose laiko intervaluose. Cox modelis išpopuliarėjo dėl lankstaus panaudojimo ir paprasto interpretavimo. Modelis sukuriamas kaip paprastas regresinis modelis, reikalauja labai mažai laiko skaičiavimams ir pakankamai tiksliai nusako ryšius tarp prognozuojamų kintamųjų ir išlikimo laiko įvairiausiuose duomenų rinkiniuose.

Cox proporcingumo rizikos modelis atliktas taip pat keliais metodais. Panaudotos dvi skirtingos funkcijos[5], kurios optimizuoja kintamųjų įverčių reikšmes įvairiais metodais. Viena funkcija ieško $\lambda \in [0, 1]$ parametro reikšmių panaudojant kryžminio patikrinimo metodą, o kita – apskaičiuojama parametrų įverčius atsitiktinai nustačius fiksuotą kiekį skirtingų λ reikšmių. λ – tai elastinio tinklo (angl. Elastic net) reguliarizacijos parametras, kuris derina kintamųjų įverčių reikšmes.

2.2.2. Atsitiktinių išlikimo miškų metodas

Atsitiktinių išlikimo miškų (angl. *Random survival forest*) metodas yra paremtas tuo pačiu principu, kaip ir detekcijos dalyje, 2.1.2 poskyryje aprašytas atsitiktinių miškų metodas. Metodikos pasižymi tam tikrais skirtumais. Pirmiausia skiriasi išplėstu priklausomu kintamuoju – šalia binarinio fakto atsiranda ir skaitinis, laiko trukmės, aspektas už kiek laiko nuo požymių stebėjimo momento nutiko faktas. Taip pat skiriasi nuo detekcijos tuo, jog kiekvienam klientui yra suteikiama ne viena lojalumo tikimybė, o daug tikimybių išlikti lojaliam kiekvienu laiko momentu į ateitį. Tam, kad būtų sugeneruojamas išlikimo uždavimo sprendimas, atsitiktinių medžių metodas tikrinamas remiantis kitais statistiniais testais, labiau tinkančiais išlikimo analizei.

Atsitiktinių išlikimo miškų medžiai „auginami“ lygiai taip pat, kaip ir klasifikavimo bei regresijos medžiai atsitiktiniuose miškuose. Procesas prasideda šaknų mazge, medžio viršuje, apimančiame visus duomenis. Iš savirankos (angl. *bootstrap*) imties atsitiktinai pasirenkami p prognozavimo kintamieji ir naudojami šakninio mazgo padalinimui pagal iš anksto nustatytą išlikimo

[5] Panaudojama epsgo ir cva.glmnet funkcijos Cox proporcingumo rizikos modeliui

testą į du dukterinius mazgus. Kiekvienam iš dviejų dukterinių mazgų atsitiktinai pasirenkamas kitas p prognozuojamų kintamųjų rinkinys, naudojamas padalinti kiekvieną iš dviejų mazgų į du papildomus dukterinius mazgus ir taip toliau. Šis procesas suskirsto klientus, kurie yra panašūs pagal jų prognozes ir atskiria skirtingas požymius turinčius klientus.

Pagrindinis skirtumas – kintamųjų grynumo vertinimas sudarinėjant medžių mazgus. Kaip jau buvo aptarta, detekcijos atsitiktinių miškų gerumo matas yra Gini indeksas. Šiuo atveju, kintamųjų grynumas matuojamas log – rank skirstymo testu (Weathers & Cutler, 2017). Jų yra ir daugiau, tačiau tyrimo metu buvo naudojamas šis. Log – rank reikšmė apskaičiuojama pagal formulę:

$$L(x,c) = \frac{\sum_{i=1}^{N}\left(d_{i,1} - Y_{i,1}\frac{d_i}{Y_i}\right)}{\sqrt{\sum_{i=1}^{N}\frac{Y_{i,1}}{Y_i}\left(1-\frac{Y_{i,1}}{Y_i}\right)\left(\frac{Y_i - d_i}{Y_i - 1}\right)d_i}}, \quad (11)$$

čia c yra kintamojo x skiriamoji vertė, $d_{i,1}$– klientų skaičius patenkantis į dukterinį medžio mazgą laiko momentu t_i, $Y_{i,1}$ – klientai, kurie laiko momentu t_i buvo priskiriami prie nelojalių ir lojalių klientų dukteriniame mazge. Tikslas yra surasti c ir x, kurie suteiktų didžiausią log – rank testo vertę. Kitaip tariant, norima surasti tokius c^* ir x^*, kad $|L(x^*, c^*)| \geq |L(x,c)|$ kiekvienai galimai c ir x reikšmei. Šis procesas kartojamas kiekvienam medžio mazgui, kol pasiekiamas paskutinis mazgas.

Kaip ir atsitiktinių miškų metode, išlikimo miškų sudarymui naudota „ranger“ funkcija, kuri padėjo optimizuoti tuos pačius kintamuosius: medžio mazgų skaičių bei kintamųjų kiekį kuris paduodamas kiekvienam iš mazgų atsitiktinai. Taip pat modeliuojant buvo išmėginama ir daugumos mažinimo korekcija.

Būtina pabrėžti, jog tyrimo metu, taikant atsitiktinį išlikimo mišką, gautoms tikimybėms apskaičiuojamas vidurkis. Taip detekcijos ir išlikimo modeliavimo metodais gali būti palyginami tarpusavyje.

2.3. Modelių gerumo vertinimas

Šioje dalyje aprašyti gautų rezultatų gerumo vertinimo metodai. Buvo pasirinktas dvejopas modelių rezultatų vertinimas: statistiškai optimali detekcija (minimizuojant neteisingai atpažintų klientų kiekį) ir maksimizuojant pelną, sugeneruoto atliekant išlaikymo programą. Kadangi išlikimo analizės metodai galiausiai suvedami į tikimybes fiksuotu laiko momentu (prilyginami detekcijai), todėl detekcijos gerumo matai tinka abiem modeliavimo tipams.

2.3.1. Modelio rezultatų kreivės

Detekcijos tyrimo metu, gaunami modelių rezultatai, t. y. kiekvienam klientui suteikiama tam tikra tikimybė, jog jis taps lojaliu. Tikimybė dar nėra galutinis sprendimas, ar klientas yra lojalus ar ne, tam reikalinga pasirinkti tam tikrą reikšmę, kuri padalina klientus į lojalių ir nelojalių klientų klases. Yra keletas modelio tikslumą reprezentuojančių kreivių, kurios suteikiama informacijos apie modelio rezultatų efektyvų su bet kokia galima ribine verte.

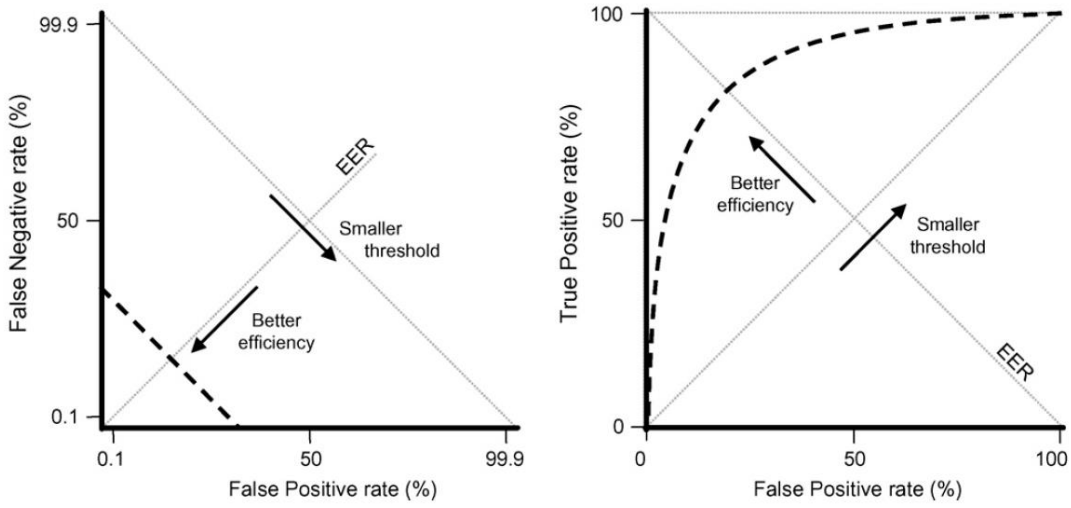
ROC kreivė (angl. *Receiver operating characteristic curve*) yra plačiai naudojama mašininio mokymosi (angl. *machine learning*) uždaviniuose. ROC kreivė – tai klasifikavimo charakteristikų grafinis atvaizdas su kintančia ribine verte t. Tai jautrumo ir 1 – specifiškumas diagrama (žiūrėti 6 pav.).

ROC kreivė nėra tiksliausias įrankis detektorių palyginimui, nes suteikia tik vizualų vaizdą. Ypač sunku lyginti du skirtingus modelius, kada ROC kreivės susikerta. Dėl to yra skaičiuojamas plotas po ROC kreive (angl. *Area under the ROC curve*) (toliau žymimas – AUC). AUC užfiksuoja modelių gerumą vienu skaičiumi, tačiau atsižvelgia į visą ribinių verčių t diapazoną. Pasiskirstymo funkcijų atžvilgiu, AUC apskaičiuojamas:

$$AUC = \int_0^1 F_0(t)\, dF_1(t). \tag{12}$$

Kuo didesnė AUC reikšmė, tuo geriau modelis klasifikuojama klientus. AUC statistiškai interpretuojama, kaip tikimybė, jog atsitiktinai pasirinktas nelojalus klientas iš imties turės mažesnį įvertinimą nei atsitiktinai pasirinktas lojalus klientas (Verbraken, 2013).

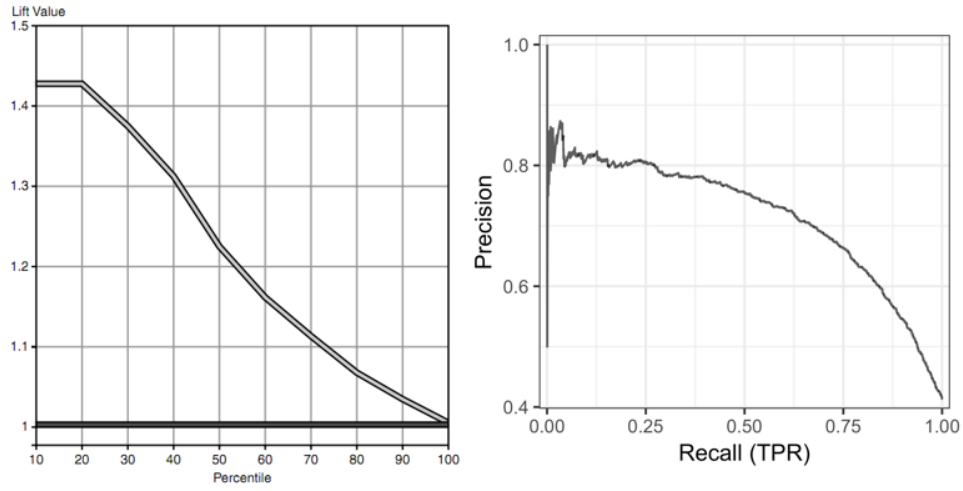
Taip pat naudojama detektoriaus paklaidų kompromiso (angl. *Detector error trade off*) kreivė (toliau DET kreivė). Jos x – ašyje vaizduojama klientų FP dalis procentais ir y – ašyje vaizduojama FN dalis. Lygių paklaidų linija (angl. *Equal error rate*) (toliau – EER) parodo, kokia kritinė reikšmė t turėtų būti naudojama, norint gauti vienodą tikslumą lojalių klientų atpažinimui ir nelojalių klientų atpažinimui. Kuo geresnis modelis, tuo kreivė yra arčiau kairiojo apatinio grafiko kampo.



6 pav. DET (kairė) ir ROC (dešinė) kreivių pavyzdys (Lechon, Llorente, Ruiz, & Vilda, 2006)

Tyrimo metu, taip pat naudojamos pranašumo ir preciziškumo – jautrumo (angl. *Precision – recall*) kreivės. Pranašumo kreivės x ašyje pateikiamas testavimo duomenų procentinė dalis, kuri galėtų būti įtraukiama į išlaikymo programą. O y ašyje fiksuojamas santykinis dydis, kiek kartų daugiau nelojalių klientų pasirinkta remiantis modeliu nei atsitiktine tvarka. Preciziškumo – jautrumo kreivė kartais naudojama, kaip alternatyva ROC kreivei detekcijos modeliams vertinti. Kreivė parodo kokia dalis iš tikro yra nelojalių klientų tarp modelio klasifikuojamų kaip nelojaliais prie tam tikro lojalių klientų detekcijos efektyvumo, kada pateikiamos visos galimos ribinės vertės. Aprašytoms dviem kreivėms pateikiami pavyzdžiai 7 pav. Pranašumo (kairė) ir preciziškumo – jautrumo (dešinė) kreivių pavyzdys.

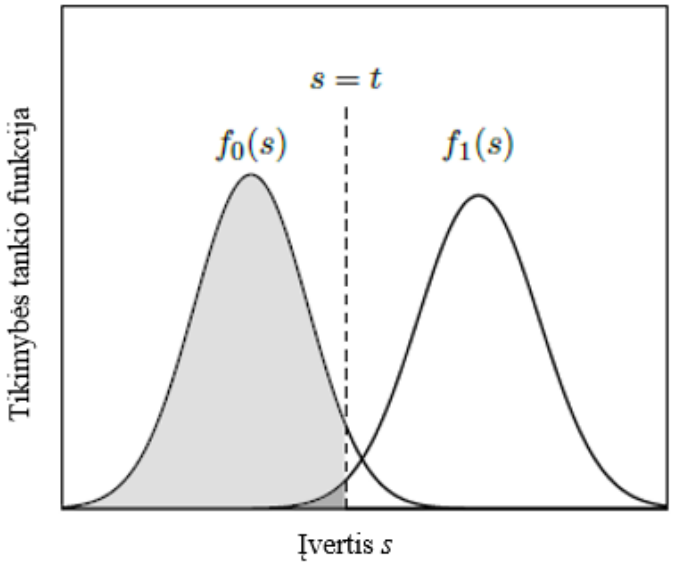
Visos keturios aprašytos kreivės naudojamos tyrimo metu rezultatų apipavidalinimui.



7 pav. Pranašumo (kairė) ir preciziškumo – jautrumo (dešinė) kreivių pavyzdys

2.3.2. Ribinė vertė

Šiame darbe daugiausiai dėmesio skiriama binarinėms klasifikavimo uždaviniams, t. y. detekcijai, kada kiekvienas klientas turi būti priskirtas tam tikrai klasei (lojalus arba nelojalus). Taip pat priskiriama reikšmė 0 arba 1. Būtina pabrėžti, jog 0 atitinką atvejį, kada klientas pasitraukia nuo įmonės ir 1 – kada klientas lojalus. Toks žymėjimas intuityvus ir susietas su modeliavimo tikslu – atpažinti pasitraukiantį klientą (angl. *Churn prediction*). Tikimybės, jog klientas patenka į vieną ar į kitą grupę žymimos π_0 ir π_1 atitinkamai.



8 pav. Modelio rezultatų pasiskirstymas ir detekcijos pavyzdys

8 pav. matoma modelio rezultatų pasiskirstymas ir detekcijos pavyzdys (Verbraken, 2013). Dažniausiai statistiniai modeliai suteikia tikimybės įvertį kiekvienam iš tirtų klientų. Tikimybė žymima $s \in \mathbb{R}$. Darome prielaida, jog nelojalūs klientai gauna žemesnį modelio įvertį, o lojalūs – aukštesnį įvertį. Toliau atliekama faktinė detekcija, t. y. kiekvienas klientas priskiriamas vienai iš dviejų grupių kada nustatoma ribinė vertė t (angl. *Threshold*). Visiems klientams, kurių įvertis

mažesnis nei nustatyta ribinė vertė, suteikiama reikšmė 0 (nelojalus). Kitiems klientams, kurių įvertis didesnis nei nustatyta ribinė vertė, suteikiama reikšmė 1 (lojalus).

Tarkime, jog funkcija $F_0(s)$ yra nelojalių klientų įverčių pasiskirstymo funkcija. Taip pat $F_1(s)$ – lojalių klientų pasiskirstymo funkcija. Analogiškai, $f_0(s)$ ir $f_1(s)$ yra nelojalių ir lojalių klientų modelio įverčių tankio funkcijos. Svarbu pabrėžti, jog kiekviena iš pasiskirstymo ir tankio funkcijų yra susieta su tam tikru modeliu arba kitaip tariant detektoriumi, kuris suteikia įverčius s. Kiekvienas klientas, kurio įverčio reikšmė mažesnė nei ribinė vertė $s < t$ klasifikuojamas, kaip nelojalus klientas. Pavyzdžiui, šviesiai pilka spalva po nelojalių klientų tankio funkcija rodo teisingai prognozuojami. Žvelgiant į tamsesnę pilką erdvę, matoma jog lojalūs klientai yra neteisingai klasifikuojami kaip nelojalūs (nes $s < t$). Taigi, kuo mažesnis tankio funkcijų persidengimas, tuo lengviau prognozuoti.

Visa tai, kas aprašoma remiantis 8 pav., apibendrinama sumaišymų matricos (angl. *Confusion matrix*) pagalba. Sumaišymų matrica pateikiama lentelės formą (žiūrėti 55 lentelė. Detekcijos rezultatų sumaišymų matrica lentelę). Lentelėje pateikiama paprasčiausia binarinio klasifikavimo rezultatų sumaišymų matrica, kurioje rodomi apibendrinti rezultatai gauti iš modelio skirstinių $f_0(s)$ ir $f_1(s)$.

5 lentelė. Detekcijos rezultatų sumaišymų matrica

	Tikroji kliento elgsena	
Modelio prognozuojama reikšmė	Nelojalus	Lojalus
Nelojalus	TP	FP
Lojalus	FN	TN

Trumpai aptariami lentelėje esantys trumpiniai. TP (angl. *True positive*) – tai tiriamos imties klientų kiekis, kurie iš tikro yra nelojalūs ir modelis tai sėkmingai nustato. Vėl gi, nereiktų susimaišyti, jog teigiamas rezultatas gaunamas kada yra nustatomas klientas, kuris palieka įmonę. TN (angl. *True negative*) – sėkmingai nustatyti lojalūs klientai. FP (angl. *False positive*) yra kiekis lojalių klientų, kurie modelio klasifikuojami kaip nelojalūs. Paskutinis FN (angl. *False negative*) – nelojalūs klientai, kurie modelio klasifikuojami kaip lojalūs. Toks klientų suskirstymo pateikimas yra supaprastintas.

Taip pat, kiekviena sumaišymų matricos dalis susijusi arba su išlaidomis arba pelnu. Išlaidos arba pelnas žymimas $c(k|l)$, kur k klasės klientas klasifikuojamas kaip l klasės klientas $k, l \in \{0, 1\}$. Paprastai pelnas arba išlaidos c yra skirtingos kiekvienai iš sumaišymų matricos dalių.

Pateikiamas išlaikymo programos kaštų pavyzdys. Jeigu modelis teisingai atpažįsta nelojalų klientą ir jį išsaugo, tada pasiekiama nauda lygi b_0. Žinoma, reikia atimti išlaikymo programos veiksmo kainą c^*. Klaidingai klasifikuojant lojalų klientą, patiriamos papildomos išlaidos c_1. Kada prognozuojama, jog klientas yra lojalus, nėra patiriama jokia žala ir nauda, nes jiems netaikoma išlaikymo programa ir tuo pačiu neskaičiuojamas pelnas gautas iš minėtų klientų. Pagrindiniai skaičiavimai akcentuojami į išlaikymo programos efektyvumą. Taigi apibrėžiami išlaidos ir pelnas:

$$\begin{aligned} c(0|0) &= b_0 - c^*; \\ c(0|1) &= c_1 + c^*; \end{aligned} \tag{13}$$

$$c(1|0) = c(1|1) = 0,$$

kur laikoma, jog $b_0,\ c^*$ ir $c_1 > 0$. Taip pat galima paminėti, jog pačio modelio sudarymas, duomenų surinkimas, paruošimas ir modelio priežiūra yra fiksuotos išlaidos ir tokios išlaidos neaptariamos, nes jos reikšmingai nekinta pasirenkant bet kokį klasifikavimo modelį.

Aptarti gerumo matai padeda surasti statistiškai geriausią modelį. Norint įvertinti klientų modelį pelno optimizavimo atžvilgiu reikia kitų išvestinių gerumo matų. Prieš tai minėti pelno ir išlaidų kintamieji b_0, c_1 ir c^*, kuriems nebuvo daromos jokios prielaidos, tik tai, jog jie visi turi būti teigiami. Šioje dalyje bus išsamiau aptarti šie kintamieji ir kaip tai susiveda į išlaikymo programos siekiamo pelno optimizavimą.

Pirmiausia aptariami pelno ir išlaidų fiksuoti dydžiai. Iš tikrųjų šie dydžiai negali būti fiksuoti, nes jei klientas iš tikro yra nelojalus ir jei detekcijos modelis nustato nelojalumą dar nėra garantijos, jog klientas iš tikro pasiliks įmonėje ar toliau tęs naudojimąsi įmonės paslaugomis. Taigi:

$$b_0 = \gamma(1-\delta)\cdot CLV;$$
$$c_1 = \ \delta \cdot CLV;$$
$$c^* = \phi \cdot CLV,$$

kur γ – dalis klientų teigiamai priėmusių išlaikymo programą, $\delta = c_1/CLV$ ir $\phi = c^*/CLV$. Dabar galima apibrėžti vidutinį klientų išlaikymo detekcijos pelną (angl. *The average classification profit for customer churn*):

$$P(t,\gamma,CLV,\delta,\phi) = CLV(\gamma(1-\delta)-\phi)\cdot\pi_0 F_0(t) - CLV(\delta+\phi)\cdot\pi_1 F_1(t). \quad (14)$$

Kiekviename tyrime egzistuoja neapibrėžtumas. Teigiamai atsakiusiųjų į išlaikymo programą dalis γ, greičiausiai sunkiausiai nuspėjamas kintamasis. Todėl dažniausiai jis laikomas atsitiktiniu kintamuoju ir todėl vidutiniame klientų išlaikymo detekcijos pelno apskaičiavime atsiranda neapibrėžtumas. Šiai problemai spręsti buvo išvestas tikėtino maksimalaus pelno kintamasis EMP (angl. *Expected maximum profit*):

$$EMP = \int_{\gamma} P\,(T,\gamma,CLV,\delta,\phi)\cdot h(\gamma)d\gamma, \quad (15)$$

čia T yra ribinė vertė, kuri priklauso nuo teigiamai atsakiusiųjų į išlaikymo programą klientų dalies γ, o $h(\gamma)$ – γ tankio funkcija. Remiantis EMP, randama ribinė vertė T nuo kurios ir priklauso kokia dalis klientų turėtų būti įtraukti į išlaikymo programą jog būtų pasiekiamas maksimalus pelnas.

Šie gerumo matai yra dažniausiai naudojami statistiniame modeliavime. Šio darbo tyrimų rezultatai taip pat bus nagrinėjami remiantis minėtais matais.

2.3.3. Detekcijos statistiniai gerumo matai

Toliau, aprašomi detekcijos gerumo matai, kurie dažniausiai naudojami detekcijos užduotiuose bei šio darbo tyrimų rezultatuose. Detekcijos modelių rezultatų apžvalgą galima rasti literatūros šaltinyje (Baldi, Brunak, Chauvin, Andersen, & Nielsen, 2000). Tikriausiai labiausiai populiarus gerumo matas yra tikslumas:

$$Tikslumas = \frac{TP+TN}{TP+TN+FP+FN}. \quad (16)$$

Tikslumas matuoja teisingai klasifikuojamų stebėjimų (šio darbo atžvilgiu – klientų) dalį. Tačiau tikslumo rezultatai labai priklauso nuo duomenų sudėties ir jo rezultatai gali klaidinti tam tikrais atvejais, kaip stebimas klasių disbalansas duomenyse ir pan. Jautrumas (angl. *Sensitivity*) ir specifiškumas (angl. *Specificity*) kiekvienas akcentuoja atitinkamai kaip efektyviai klasifikuojami lojalūs klientai ir nelojalūs:

$$Jautrumas = \frac{TP}{TP + FN}; \tag{17}$$

$$Specifiškumas = \frac{TN}{TN + FP}. \tag{18}$$

Kitas gerumo matas yra preciziškumas (angl. *Precision*). Jis matuoja kokia dalis iš tikro yra nelojalių klientų tarp modelio klasifikuojamų kaip nelojaliais:

$$Preciziškumas = \frac{TP}{TP + FP}. \tag{19}$$

Preciziškumas ir jautrumas panaudojamas F_β gerumo matui apskaičiuoti:

$$F_\beta = \frac{(1+\beta)^2 \times preciziškumas \times jautrumas}{\beta^2 \times preciziškumas + jautrumas}. \tag{20}$$

Šis matas atspindi modelio gerumą, kada fiksuojama svarba tarp jautrumo ir specifiškumo. Šių matų svarbą pažymį kintamasi β. Dažniausiai naudojamos β reikšmės yra 1, 2, ½.

Taip pat atsižvelgiama į gerumo matą kappa (angl. *Cohen's kappa coefficient*), kuris nurodo ne tik kaip tiksliai modelis prognozuoja, tačiau ir balansą tarp lojalių ir nelojalių klientų prognozių tikslumu. Teigiama, jog šis matas yra patikimesnis nei tikslumo matas, nes į apskaičiavimą yra įtraukta hipotetinė tikimybė, jog nuomonių susitarimas, t. y. prognozuotų ir realių reikšmių susitarimas įvyko atsitiktinai. Kappa užrašoma lygybe:

$$Kappa = 1 - \frac{1 - tikslumas}{1 - p_e}, \tag{21}$$

kur p_e – hipotetinė tikimybė, jog susitarimas atsitiktinis. Kuo kappa vertė arčiau vieneto tuo labiau subalansuoti gauti modelio rezultatai.

2.3.4. Klasių disbalansas ir jo eliminavimo būdai

Modeliavimo rezultatų tikslumui turi įtakos pati imties sudėtis, t. y. klasių balansas. Jeigu tiriamoje imtyje yra labai maža dalis nelojalių klientų, tam tikri modeliavimo metodai gali klaidingai arba visai neaptikti tinkamų kriterijų kodėl buvo norima pasitraukti nuo įmonės (Burez & Van den Poel, 2009). Pabrėžiamos šešios problemos, kurios išryškėja tiriant klientų duomenis, kuriuose fiksuojamas ryškus klasių disbalansas (Weiss, 2004):

1. Pasirenkami netinkami vertinimo matai. Dažnai modelių rezultatų vertinimo matai nėra pritaikyti įžvelgti išimtinius atvejus duomenyse ir vertina modelį pagal didžiosios duomenų dalies bendras savybes.
2. Duomenų trūkumas. Kada nelojalių klientų yra nedaug absoliutiniais dydžiais, sunku įžvelgti vyraujančias tendencijas, būdingas mažumai.
3. Santykinis duomenų trūkumas. Jeigu tiriama didelė įmonė, tada vartotojų kiekis absoliutiniais dydžiais yra tikrai pakankamas. Tačiau tada galima susidurti su problema, jog duomenų imtyje

neproporcingai mažai nelojalių klientų (1 proc. ar net mačiau). Tada dažniausiai naudojami modeliavimo metodai tampa neveiksmingi visos tiriamos imties mastu.
4. Dauguma lojalumo tyrimo modelių taiko skaidymo metodiką, kurios metu duomenys vis skaidomi į mažesnius poaibius ir taip stengiamasi įžvelgti vis mažesnes bendras tendencijas.
5. Netinkamas poslinkis į tendencijas. Didžioji dalis modeliavimo algoritmų turi tendenciją persimokinti turimai imčiai, todėl dažniausiai algoritmo mokymosi metu yra naudojami apribojimai, neleidžiantis prisitaikyti prie analizuojamos duomenų imties iki pačių smulkiausių detalių. Kada duomenyse fiksuojamas klasių disbalansas, apribotieji algoritmai, tikėtina, nesugebės apčiuopti duomenyse esančios mažumos tendencijų.
6. Triukšmas. Duomenyse esančios išskirtys gali reikšmingai paveikti modeliavimo algoritmą. Kada yra ryškus klasių disbalansas, pasidaro sunku įžvelgti skirtumus tarp mažumos tendencijų ir išimčių.

Galima naudoti kelis klasių disbalanso panaikinimo būdus. Vienas iš jų – atlikti modelio apmokymą subalansuotai duomenų imčiai. Taip būtų sudaroma sąlyginai dirbtinė duomenų imtis, kurioje apie 50 proc. duomenų būtų lojalių ir kita 50 proc. dalis – nelojalių klientų. Taip modelio apmokymo algoritmas vertintų vienodai reikšmingai ir vienos klasės ir kitos klasės požymius. Kitas būdas yra suteikti kiekvienam stebėjimui, t. y. klientui svorį prieš naudojant duomenis skaičiavimuose. Tam tikras svoris atitiktų klasės proporcinę dalį visoje duomenų imtyje. Tai tik pora metodų, kurie padeda išvengti aukščiau išvardintas problemas.

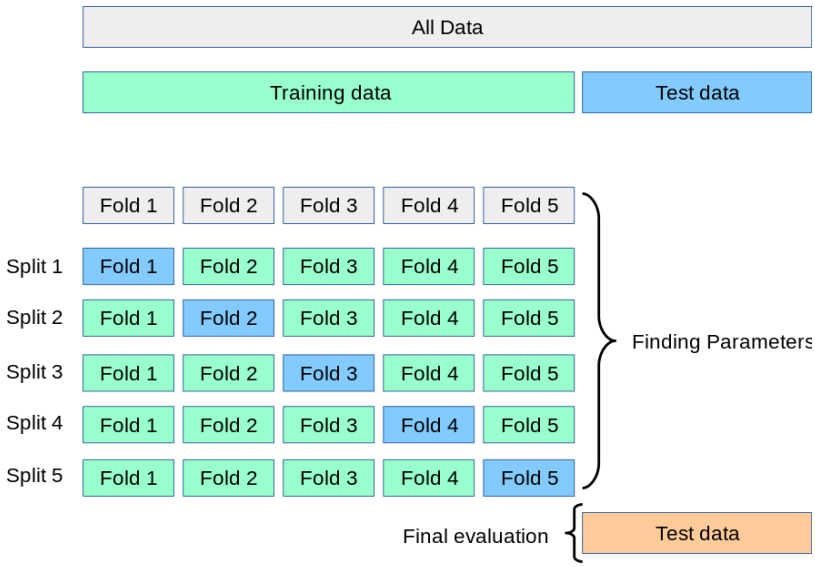
2.3.5. Kryžminis patikrinimas

Sudarinėjant modelius buvo naudojamas kryžminio patikrinimo metodas visai duomenų imčiai. Tai algoritmas padalinantis visą duomenų imtį į apmokymo ir testavimo imtis k kartų. Yra keletas priežasčių, kodėl buvo pasirinkta naudoti kryžminio patikrinimo metodą (Shulga, 2018):

- Panaudojama visa duomenų imtis modelių apmokymui. Žvelgiant iš didžiųjų duomenų perspektyvos, keletas tūkstančių eilučių nėra pakankamai didelė duomenų imtis. Paprastas padalinimas į apmokymo ir testavimo imtis sumažina modelio apsimokinimo galimybes.
- Kada sukurtiems penkiems skirtingiems modeliams naudojamas apmokymo algoritmas ir išbandomas dešimtyje skirtingų testų rinkinių, galima labiau įsitikinti algoritmo vientisu ir sėkmingu veikimu.

Dauguma mokymosi algoritmų reikalauja tam tikrų parametrų derinimo. Norima rasti optimalius metodų parametrus, kurie padaro modelį efektyvesnį prognozių atžvilgiu. Tokiu atveju, duomenų įvairovės ir gausos poreikis dar labiau išauga. Su kryžminio patikrinimo algoritmu mokymo imtis nepasidaro mažesnė nei visas duomenų rinkinys.

Toliau esančiame paveikslėlyje matomas duomenų kryžminių patikrinimo skaidymo pavyzdys, kai koeficientas $k = 5$ (žiūrėti 9 pav.). Jei $k = 5$, vadinasi sudaromi 5 modeliai su skirtingomis apmokymo ir tikrinimo imtimis. Labai dažnai reikalaujama, jog kryžminio patikrinimo metu skaidant duomenis, kiekvienoje duomenų imtyje būtų išlaikomas apytikslis priklausomų ir nepriklausomų kintamųjų klasių santykis.



9 pav. Duomenų skaidymo kryžminiu patikrinimu pavyzdys

3. Tyrimų rezultatai ir jų aptarimas

Prieš pradedant aptarinėti modelių gautus rezultatus bei interpretuojamas išvadas, būtina susipažinti su duomenimis. Todėl aptariamas klasių disbalansas, atliekama žvalgomoji analizė.

Detekcijos ir išlikimo analizės palyginimui buvo naudojami du skirtingi duomenų rinkiniai. Abu pritaikomi detekcijos ir išlikimo modeliams. Tam duomenų imtyje reikia turėti lojalumo kintamąjį ir klientų ilgaamžiškumą, kiek laiko klientai naudojasi įmonės paslaugomis.

Duomenų imtys yra paimtos iš skirtingų sričių: telekomunikacijų įmonės duomenys bei „Premium“ klubo įmonės klientų duomenys. Platesnė informacija nėra pateikiama, kokias paslaugas konkrečiai teikia įmonė, tačiau žvelgiant į duomenis galima daryti prielaidą, jog teikiamos išskirtinės paslaugos, už kurias samdomi agentai bei mokamos dideli metiniai mokesčiai – vidurkis lygus 178810,00 dolerių.

3.1. Žvalgomoji analizė

3.1.1. Telekomunikacijų įmonės duomenys

Šiame darbe analizuojami duomenys iš dviejų skirtingų verslo sričių. Abiejų duomenų rinkinių informacija nuasmeninta, suteikiant unikalius kodus, kiekvienam iš aprašomųjų asmenų: klientams bei agentams. Telekomunikacijų duomenų rinkinys susideda iš 7043 klientų bei 21 kintamojo, kurie vienaip ar kitaip apibūdina klientus (žiūrėti 6 lentelę). Duomenyse daugiausia yra kategorinių kintamųjų.

Telekomunikacijų duomenyse buvo nustatyta, jog 11 klientų turi trūkstamas reikšmes būtent klientų visų išlaidų kintamajame. Taip yra todėl, kad tie patys klientai įmonėje buvo trumpiau nei mėnesį, todėl duomenyse nėra užfiksuojamos jų išlaidos. Dėl šios priežasties visos išlaidos buvo prilygintos 0 ties 11 klientų. Laiko intervalo kintamasis (tenure) yra nuo 0 iki 72 mėnesių. Kadangi išlikimo modeliuose nepriimama informacija su išlikimo laikotarpiu lygiu 0, todėl 0 buvo pakeistas į pusę mėnesio trukmę – 0.5.

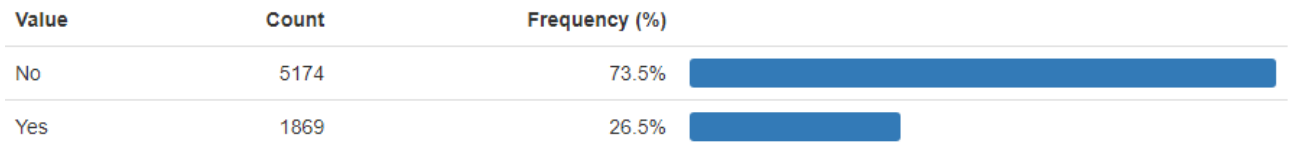
6 lentelė. Telekomunikacijų įmonės klientų duomenų struktūra

Kintamasis	**Tipas**	**Reikšmės**	**Aprašymas**
customerID	skaitinis	0002-ORFBO, 0003-MKNFE	Kliento identifikacinis numeris
SeniorCitizen	sveikasis skaičius	0, 1	Vyresniojo statusą turintis klientas
Partner	kategorinis	Yes, No	Partnerio statusą turintis klientas
Dependents	kategorinis	Yes, No	Ar klientas išlaikomas kito
tenure	sveikasis skaičius	1, 2, 3...	Paslaugų naudojimosi trukmė mėnesiais
PhoneService	kategorinis	Yes, No	Telefono paslaugos
MultipleLines	kategorinis	Yes, No, No phone service	Ar klientas naudojasi keliomis skirtingomis paslaugomis
InternetService	kategorinis	DSL, Fiber optic, No	Interneto paslaugų tipas
OnlineSecurity	kategorinis	Yes, No, No internet service	Interaktyvios apsaugos paslauga
OnlineBackup	kategorinis	Yes, No, No internet service	Atsarginės kopijos paslauga
DeviceProtection	kategorinis	Yes, No, No internet service	Prietaiso apsaugos paslauga

TechSupport	kategorinis	Yes, No, No internet service	Techninės priežiūros paslauga
StreamingTV	kategorinis	Yes, No, No internet service	Tiesioginės transliacijos paslauga
StreamingMovies	kategorinis	Yes, No, No internet service	Filmų nuomos paslauga
Contract	kategorinis	Month – to – month, One year, Two year	Sutarties tipas
PaperlessBilling	kategorinis	Yes, No	Sąskaitų pateikimas elektroniniu būdu ar fiziniu.
PaymentMethod	kategorinis	Bank transfer (automatic), Credit card (automatic), Electronic check Mailed check	Apmokėjimo būdas
MonthlyCharges	realusis skaičius	29,85; 56,95...	Mėnesinės kliento išlaidos
TotalCharges	realusis skaičius	29,85; 1889,50...	Visos kliento išlaidos
Churn	loginė reikšmė	Yes, No	Lojalus klientas ar ne
gender	kategorinis	Female, Male	Lytis

Taip pat atliekama tam tikrų kintamųjų kategorijų modifikacija. Septyniems kintamiesiems, kurie turi kategorijas „Yes“, „No“ ir „No phone service“ buvo pakeista paskutinioji kategorija į „No“. Buvo nustatyta, jog mėnesiniai mokesčiai su visomis klientų išlaidomis statistiškai reikšmingai koreliuoja (pagal Pearsono koreliacijos testą, nustatyta, jog koreliacija lygi 0,65 ir gaunama p – reikšmė $2{,}2*10^{-16}$, todėl atmetama nulinė hipotezė, jog koreliacija yra lygi 0). Dėl šios priežasties kintamasis Total Charges išimamas iš analizės, kaip ir klientų identifikacinis numeris.

Toliau peržvelgiama kiekvieno iš kategorinių kintamųjų struktūra. Pagrindinis prognozuojamas kintamas yra klientų „nubyrėjimas“. Suteiktas toks pavadinimas, nes „Yes“ reikšmė siejama su pasitraukiančiais ir nelojaliais klientais. Lojalių klientų yra 73 proc. ir nelojalių – 27 proc. (žiūrėti 10 pav.). Kaip matoma, klasės pasiskirsčiusios nevienodai, todėl esamas skirtumas gali turėti įtakos modelių kūrimo procese.

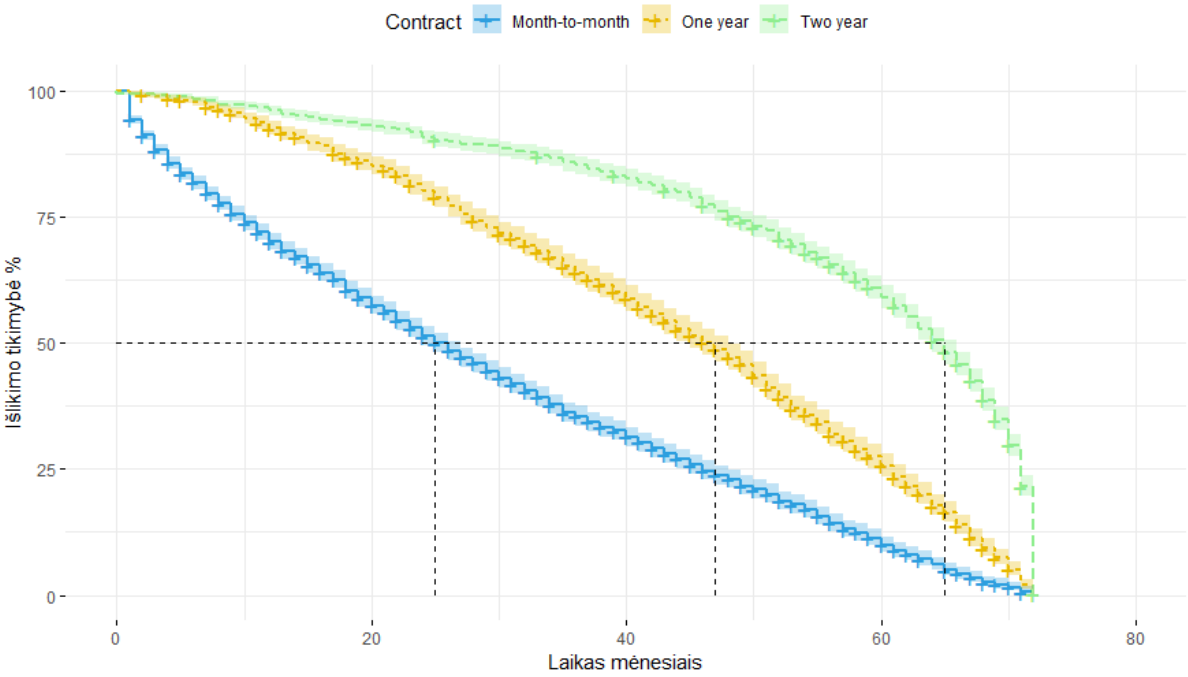


10 pav. Telekomunikacijų įmonės klientų lojalumo kintamojo struktūra

Didžiausi skirtumai tarp klasių pastebėti kintamuosiuose, kurie nusako ar klientas turi vyresniojo statusą ir ar klientas naudojasi telefoninėmis paslaugomis. Kategorinių kintamųjų pasiskirstymo histogramos ir skaitinių kintamųjų statistika pateikiami darbo priede.

Žvelgiant iš išlikimo analizės perspektyvos, galima aptarti, kokie kintamieji geriausiai atskiria lojalius nuo nelojalių klientų atsižvelgiant į klientų ilgaamžiškumą. Pagal Kaplan – Meier statistinį įvertinimą pastebėta, jog vienas iš labiausiai atskiriančių išlikimo tikimybę yra sutarties tipo kintamasis (žiūrėti 11 pav.). Intuityvu, jog klientai, turintys ilgalaikes sutartis yra labiau linkę būti lojaliais nei tie, kurie neįsipareigoja. Patikrinus šį kintamąjį su log – rank testu p – reikšmė lygi $2*10^{-16}$, todėl atmetama nulinė hipotezė, jog tarp 11 pav. 10 pav. Telekomunikacijų įmonės klientų lojalumo kintamojo struktūra esančių trijų kreivių nėra statistiškai reikšmingo skirtumo. Išlikimo linijos pasiekia tikimybę 0 viename taškę, tai praėjus 72 mėnesiams, nes būtent tokia yra maksimali

klientų paslaugų naudojimosi trukmė duomenų imtyje. Kiti kintamieji irgi pakankamai sėkmingai išskiria lojalius ir nelojalius klientus, todėl pradėjus modeliuoti yra pasirenkami visi pertvarkytos duomenų imties kintamieji. Kitų kintamųjų kreivės pridėtos priede.



11 pav. Išlikimo kreivė padalinta pagal klientų sutarčių tipą

3.1.2. „Premium“ klubo duomenys

Šiame poskyryje suteikiama pradinė informacija apie antrąjį „Premium“ klubo duomenų rinkinį, jo struktūra ir žvalgomoji analizė. Duomenų rinkinys sudarytas iš 10362 skirtingų klientų ir 15 kintamųjų. Kiekvienas kintamasis glaustai aprašytas 7 lentelėje.

Galima pastebėti, jog duomenų rinkinyje yra kitokie laiko kintamieji nei prieš tai esančiuose duomenyse. Iš pradžios ir pabaigos kintamųjų paskaičiuota paslaugų vartojimo trukmė mėnesiais. Tai atlikta tam, kad duomenims būtų pritaikomi išlikimo modeliams. Vėliausia kliento pasitraukimo data užfiksuota 2013–11–25, todėl nuspręsta šią datą naudoti kaip duomenų surinkimo datą, nuo kurios ir priklauso kiekvieno iš klientų paslaugų vartojimo trukmė. Sukurtasis laiko kintamasis yra nuo 0 iki 86 mėnesių. Kaip ir telekomunikacijų duomenims, taip ir šiuose duomenyse paslaugų vartojimo trukmė 0 buvo pakeistas į pusę mėnesio trukmę – 0.5.

Lyties ir šeimyninės padėties kintamieji turi nemažą dalį trūkstamų duomenų (6 proc. ir 25 proc. atitinkamai), todėl buvo pasirinkta klientų, su šiomis trūkstamomis reikšmėmis neišimti iš duomenų imties ir palikti, kaip atskirą kategoriją. Trūkstamų reikšmių turėjo ir klientų metinių pajamų kintamasis. Dėl kintamojo skirstinio asimetrijos (angl. *skewness*), buvo pasirinkta trūkstamas reikšmes pakeisti į medianą.

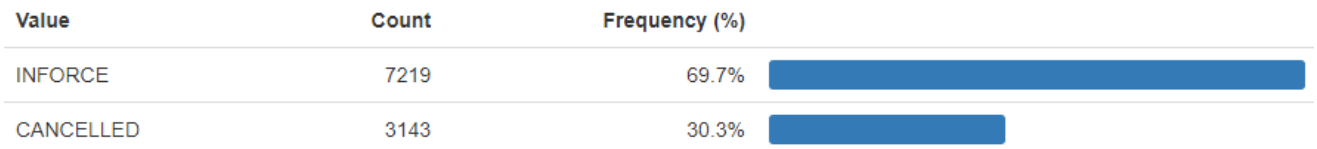
7 lentelė. „Premium“ klubo klientų duomenų struktūra

Kintamasis	**Tipas**	**Reikšmės**	**Aprašymas**
MEMBERSHIP_NUMBER	skaitinis	A00001, A00002	Kliento numeris
MEMBERSHIP_TERM_YEARS	sveikasis skaičius	12, 29...	Planuotas partnerystės terminas

ANNUAL_FEES	sveikasis skaičius	113125, 112220....	Metiniai klubo mokesčiai
MEMBER_MARITAL_STATUS	kategorinis	NA, D M S W	Šeimyninė padėtis
MEMBER_GENDER	kategorinis	NA, F, M	Lygis
MEMBER_ANNUAL_INCOME	sveikasis skaičius	25200000, 10339200	Kliento metinės pajamos
MEMBER_OCCUPATION_CD	sveikasis skaičius	1, 2, 3...	–
MEMBERSHIP_PACKAGE	kategorinis	TYPE-A, TYPE-B	Sutarties tipas
MEMBER_AGE_AT_ISSUE	sveikasis skaičius	27, 36...	Kliento amžius sutarties sudarymo metu
ADDITIONAL_MEMBERS	sveikasis skaičius	1, 2...	Papildomas klientų kiekis susijęs su pastaruoju
PAYMENT_MODE	kategorinis	Monthly, annually...	Apmokėjimo metodas
AGENT_CODE	kategorinis	1001155, 1002099...	Agento kodas
MEMBERSHIP_STATUS	kategorinis	CANCELLED, INFORCE	Lojalus ar nelojalus klientas
START_DATE..YYYYMMDD.	data	20060914...	Sutarties pasirašymo data
END_DATE...YYYYMMDD.	data	NA, 20090811...	Nutraukimo data

Į modeliavimo procesą nebuvo įtraukti identifikaciniai kintamieji kaip kliento numeris bei agento kodas. Nors pagal agento numerį buvo galima pamatyti skirtingus rezultatus, tačiau modeliavimui neparanku naudoti tokį kintamąjį, nes skirtingų agentų itin daug – 4317 skirtingų agentų kodų, kada klientų kiekis siekia 10362. tarp metinių mokesčių ir kliento metinių pajamų, koreliacijos koeficientas lygus 0,05 ir ranginės koreliacijos koeficientas pakankamai nedidelis, kad abu kintamieji būtų palikti modeliavime – 0,24.

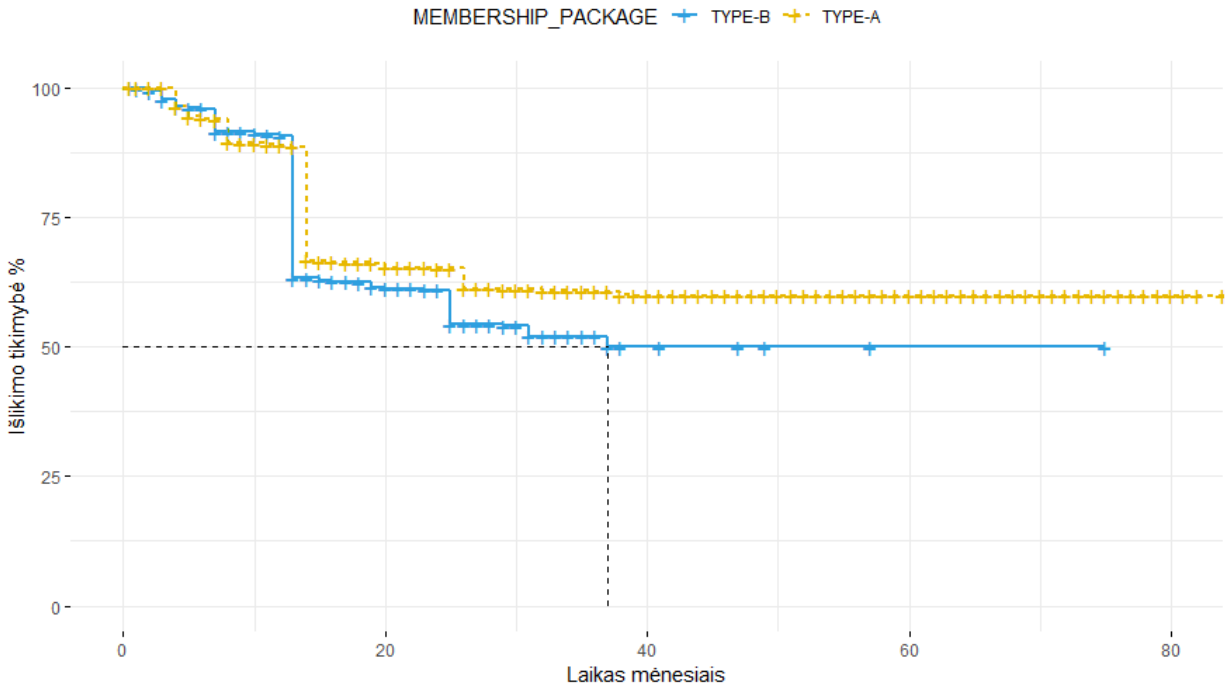
Šiuose duomenyse pagrindinis prognozuojamas kintamasis yra MEMBERSHIP_STATUS, kuris nurodo kiekvieno iš klientų padėtį, įmonės atžvilgiu. Dešimtame paveikslėlyje matoma, jog šiek tiek daugiau nei 30% klientų yra nutraukę sutartis, o likusieji vis dar yra aktyvūs klientai (žiūrėti 12 pav.). Kaip ir pirmoje duomenų imtyje, „Premium“ klubo duomenims nefiksuojamas klasių disbalansas. Taip pat pertvarkomas Lojalumo kintamasis MEMBERSHIP_STATUS pervadinant kategorijas pagal tai, ar klientas palikęs įmonę ar išlikęs lojaliu.



12 pav. „Premium“ klubo lojalumo kintamojo struktūra

Pastebima pasiskirstymo asimetrija klientų metinių pajamų kintamajam, t. y. pastebima nedidelė dalis klientų, kurių pajamos didesnės nei didžiosios daugumos. Kadangi tokia duomenų pateiktis yra realistiška, imties kintamieji daugiau nekoreguojami. Papildoma informacija apie kiekvieno iš naudojamų kintamųjų galima rasti antrame priede. Skaitiniams kintamiesiems pateikiama skirstinio statistika, o kategoriniams – histograma.

Pagal Kaplan – Meier statistinį įvertinimą nebuvo pastebėta, jog kategoriniai kintamieji labai drastiškai išskirtų lojalius ir nelojalius klientus. Kaip pavyzdys, pateikiama sutarties tipo kintamojo kreivė (žiūrėti 13 pav.).



13 pav. Išlikimo kreivė padalinta pagal sutarties tipą.

Grafike matoma, jog tie klientai, kurių sutarties tipas yra B tikimybė išlikti lojaliems yra mažesnė. Šiai klientų grupei, lojalumo tikimybė nukrenta iki 0,5 naudojantis įmonės paslaugomis mažiau nei 40 mėnesių. Atlikus log – rank testą sutarties tipui, gaunama p – reikšmė yra artima nuliui. Kadangi tai yra mažiau nei 0,05, todėl nulinė hipotezė yra atmetama, kuri teigia, jog sutarties tipo kintamasis neturi reikšmingos įtakos lojalumui ir kliento ilgaamžiškumui. Kitų kategorinių kintamųjų kreivės pateikiamos priede.

Sekančiuose poskyriuose aptariamas duomenų paruošimas modeliavimui, modeliavimo rezultatai bei detekcijos ir išlikimo modelių palyginimas. Analizuojami duomenų rinkiniai atskirai ir galiausiai pateikiamos bendros įžvalgos, tinkančios abiem duomenų rinkiniams.

3.2. Telekomunikacijų klientų lojalumo prognozavimas

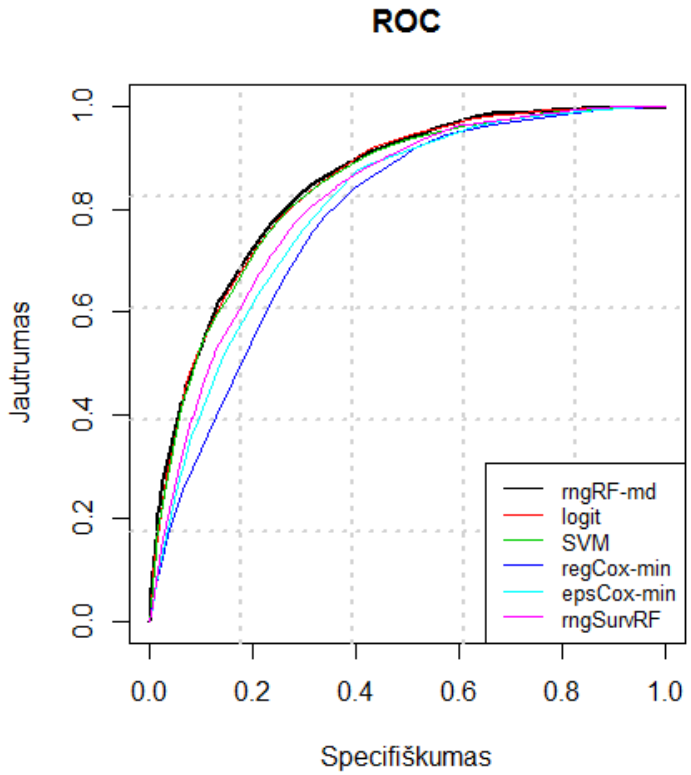
Visos duomenų modifikacijos jau aptartos duomenų žvalgomosios analizės dalyje. Primenama, jog šiame darbe naudoti modeliai:

- logistinė regresija;
- atsitiktiniai miškai;
- atraminių vektorių metodas su tiesiniu branduoliu;
- Cox proporcingumo rizikos modelis;
- atsitiktiniai išlikimo miškai.

Naudojamų modelių rezultatai vertinami keliais etapais: pirmiausia lyginami visi modeliai pagal prognozuojamus rezultatus kreivėmis. Po to patiekiama modeliavimo statistika. Pagal pateikiama informaciją, pasirenkamas vienas geriausias metodas iš detekcijos ir vienas geriausias iš

išlikimo metodas. Tada tolimesnis tyrimas su ribinių verčių nustatymu ir gerumo matų pateikimu vyksta su pasirinktais dviem modeliais.

Iš pradžių pasirenkama po vieną geriausią modelį iš detekcijos ir išlikimo modeliavimo sričių. Pasirinkimas vykdomas remiantis ROC kreive ir AUC plotu, DET kreive, preciziškumo – jautrumo kreive ir pranašumo kreive. ROC kreivė pateikia tikslesnius rezultatus, kada prognozuojamos klasės neturi disbalanso. Preciziškumo – jautrumo kreivė rodo tikslesnius rezultatus nei ROC kreivė, kada pastebimas bent nedidelis klasių disbalansas.

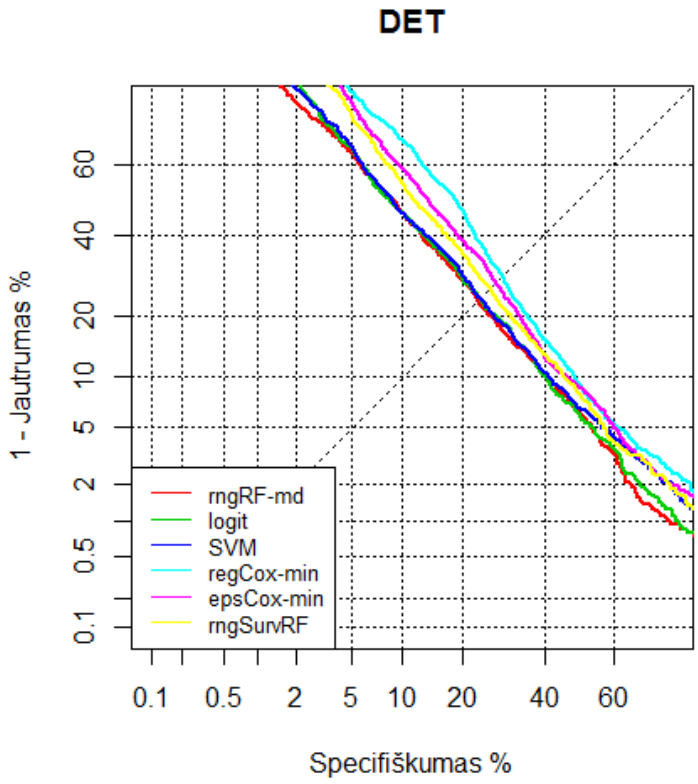


14 pav. Telekomunikacijų duomenų modelių ROC kreivė

ROC kreivėje matoma, jog geresnius rezultatus rodo detekcijos modeliai – detekcijos modelių kreivės yra arčiau viršutinio kairiojo kampo (žiūrėti 14 pav.). Iš ROC kreivės, vizualiai sunku pasakyti kuris detekcijos modelis geriausias, tačiau matoma, jog visi detekcijos lenkia išlikimo modelius. Tarp išlikimo modelių, geriausius rezultatus teikia atsitiktiniai išlikimo miškai. Iš visų modelių, blogiausiai pasirodė reguliarizuotas Cox proporcingumo rizikos modelis su reguliarizacijos parametru λ minimaliame testuotame taške. Žvelgiant į AUC ploto rezultatus, kurie pateikiami toliau esančioje 8 lentelėje, didžiausias AUC plotas fiksuojamas atsitiktinių miškų modelio ir lygus 0,849, tuo tarpu mažiausias plotas pateikiamas reguliarizuoto Cox modelio, kuris lygus 0,782.

Kaip ir prieš tai, taip ir DET grafike matoma tendencija panaši – nors ir nežymiai, tačiau atsitiktiniai miškai turi mažiausius paklaidų rodiklius. Lygių paklaidų lygyje, atsitiktinių miškų paklaidos lygios 23,18 proc. Po to seka kiti detekcijos modeliai ir šiek tiek prastesnius rezultatus fiksuoja išlikimo modeliai. Kreivė nurodo kaip pasiskirsto neteisingas lojalių ir nelojalių klientų nustatymas keičiant ribinę vertę. Iš kreivės matoma, jog atsitiktinių miškų modeliui, ne visose ribinėse vertėse tarp 0 ir 1 rodo mažiausias abiejų klasių paklaidas, t. y. atsitiktinių miškų kreivė susikerta su kitų modelių kreivėmis. Taip pat pastebima, jog visos kreivės yra ganėtinai tiesios ir

įstrižos. Tai parodo, jog keičiant ribinę vertę, vienos klasės paklaidos proporcingai mažėja, kada kitos klientų klasės paklaida didėja. Iš kreivių išsiskiria reguliarizuotas Cox proporcingumo rizikos



15 pav. Telekomunikacijų duomenų modelių DET kreivė

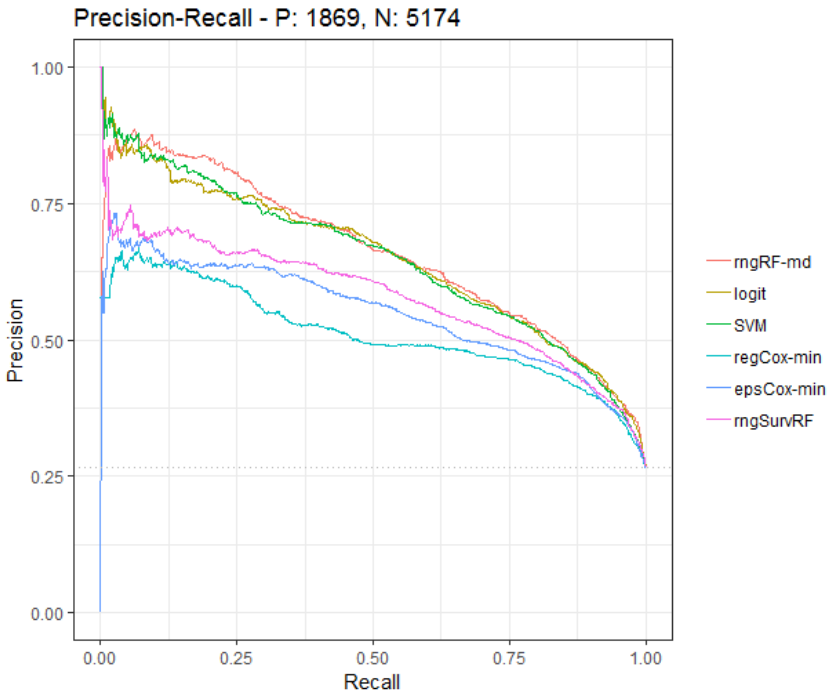
modelis. Jis beveik visose ribinės vertės taškuose turi prasčiausius paklaidos rodiklius. Lygių paklaidų lygyje, Cox modelio paklaidos lygios 28,80 proc.

Matoma, jog to paties Cox modelio rezultatai, gaunami kitos funkcijos ir kitokiu parametrų optimizavimo būdu, pastebimai lenkia reguliarizuotą Cox modelį (šviesiai mėlyna ir rožinės spalvos kreivės 15 pav.). Tačiau, jos išsiskiria, kada didesnis dėmesys yra skiriamas lojalių klientų tikslesniam nustatymui nei nelojalių, t. y. kai ribinė vertė t nustatoma tokia, kad nelojalių klientų tikslumo paklaida didesnė nei lojalių klientų paklaida.

Taip pat pagrindiniame tekste pateikiama bei aptariama preciziškumo – jautrumo kreivė (žiūrėti 16 pav.). Šiame grafike įžvelgiami didesni skirtumai tarp išbrėžtų kreivių. Pasirenkant ribinę vertę, kuri didesnį dėmesį skirtų preciziškumui tada akivaizdžiai geriausias metodas būtų atsitiktinų miškų. Kaip ir jau aptartose kreivėse, prasčiausius rezultatus parodo reguliarizuotas Cox proporcingumo rizikos modelis. Preciziškumo – jautrumo kreivės suteikia panašius rezultatus, jog geriau klientus klasifikuoja detekcijos modeliai nei išlikimo modeliai.

Jautrumui artėjant prie vieneto, preciziškumas artėja prie tikrosios nelojalių klientų dalies analizuojamoje duomenų imtyje. Taip yra todėl, jog kai jautrumas lygus vienetui, tada nelojalių klientų prognozavimas lojaliais lygus 0, o preciziškumas tiesiog apskaičiuoja santykį tarp esamų nelojalių klientų ir visų klientų.

Kadangi pranašumo kreivės suteikia panašias įžvalgas kaip ir kitos trys kreivės, todėl grafikas demonstruojamas priede.



16 pav. Telekomunikacijos duomenų modelių rezultatų preciziškumo – jautrumo grafikas, kur horizontalioje ašyje jautrumas, o vertikalioje – preciziškumas

Toliau pateikiama lentelė su telekomunikacijų įmonės modeliavimo rezultatais (žiūrėti 8 lentelę). Būtina paminėti, jog lentelėje aprašoma kiekvieno modelio tik geriausius rezultatus parodžiusi versija. Kadangi buvo naudojamas kryžminio patikrinimo metodas kur $k = 10$, todėl kiekvienas iš modelių buvo parametrizuojamas dešimt kartų, apmokymo ir testavimo imtys siekė viso duomenų rinkinio atitinkamai 90 proc. ir 10 proc. bei rezultatų pateikimui buvo naudojami visas duomenų rinkinys, nes dešimties patikrinimų metu, kiekvieno kliento duomenys buvo pateikiami testavimo imtyje.

8 lentelė. Telekomunikacijų įmonės modeliavimo rezultatai

Modeliai	**Detekcija**			**Išlikimas**		
	Logistinė regresija	**Atsitiktiniai miškai**	**Atraminiai vektoriai**	**Cox rizikos metodas**	**Cox rizikos metodas (epsgo)**	**Atsitiktiniai išlikimo miškai**
AUC	0,846	**0,849**	0,841	0,782	0,803	**0,817**
EER (%)	23,58	**23,18**	23,63	28,80	27,32	**25,88**
Mokymo laikas (min)	**0:01**	8:01	20:08	11:40	18:01	**0:50**

Kaip matoma lentelėje, didžiausią plotą po ROC kreive užfiksuoja atsitiktinių miškų modeliai. Tie patys modeliai fiksuoja geriausius rezultatus lygių paklaidų rodiklio EER atžvilgiu. Kaip ir galima tikėtis, elementariausia logistinė regresija, parodo išskirtinai gerus rezultatus laiko atžvilgiu. Kryžminio patikrinimo metodu, logistinę regresiją apskaičiuoti užtrunka vos sekundę. Tai vienas iš logistinės regresijos teigiamų bruožų, kuris išlaiko logistinę regresiją konkurencingą ir plačiai naudojamą įvairiausiose industrijose. Nors ir logistinė regresija nepralenkiama mokymosi trukme, pasirenkamas atsitiktinių miškų detekcijos metodas, kuris užfiksuoja geresnius tikslumo matus.

Išlikimo dalyje pasirenkamas atsitiktinių išlikimo miškų metodas, kadangi lentelėje pateiktiems kriterijams, šis metodas įgyja geriausius rezultatus.

Pasirinkus du geriausius modelius toliau surandamos ribinės vertės t pagal du rezultatų vertinimo požiūrius: statistiškai optimalus variantas, kada optimizuojamas lojalių ir nelojalių klientų klasių prognozavimo tikslumas; skaičiuojant pagal tikėtiną maksimalią pelno vertę EMP. Prieš randant tikėtiną maksimalią pelno vertę, privaloma nusistatyti trijų kintamųjų fiksuotas reikšmes: gyvavimo trukmės vertę CLV, išlaikymo programos kainą bei susisiekimo kainą. CLV pateikiamas kaip visų kliento išlaidų mediana, kuri yra lygi $1395,00. Išlaikymo programos išlaidos paskaičiuojamos vienam klientui padauginant numatytą mėnesinę nuolaidą $2,00 iš vidutinės klientų gyvavimo trukmės – 32 mėnesiai. Taip gaunama, jog išlaikymo programos išlaidos lygios $64,00, o susisiekimo su klientu kaina $4,00. Toliau pateikiami gerumo matai dviejų geriausių detekcijos ir išlikimo modelių pagal skirtingai parenkamas ribines vertes t. 9 lentelėje pateikiami modeliavimo gerumo matų rezultatai.

9 lentelė. Telekomunikacijų įmonės gerumo matai

Metodika	**Modelis**	**Ribinės vertė**	**Bendrasis Tikslumas**	**Jautrumas**	**Specifiškumas**	**Preciziškumas**	**EMP klientui**	**F_1**	**Kappa**
Detekcija	Atsitiktiniai miškai	EER	0,770	**0,769**	0,771	0,548	–	**0,640**	0,478
		EMP	**0,784**	0,709	0,812	0,576	$71,50	0,636	**0,485**
Išlikimas	Atsitiktiniai išlikimo miškai	EER	0,739	0,742	0,738	0,506	–	0,601	0,418
		EMP	0,762	0,205	**0,963**	**0,670**	**$75,16**	0,314	0,216

Iš lentelėje pateikiamų rezultatų negalima vienareikšmiškai nustatyti kuris modelis yra geriausias visų kriterijų atžvilgiu. Tačiau matoma, jog išlikimo modelis optimizuodamas tikėtiną maksimalų pelną nustato tokią ribinę vertę, kuri susikoncentruoja į nelojalių klientų klasę, o tuo tartu lojalių klientų nustatymo tikslumo rezultatai labai krenta. Aptariama situacija pateikiama sumaišymų matrica, kuri pateikiama priede.

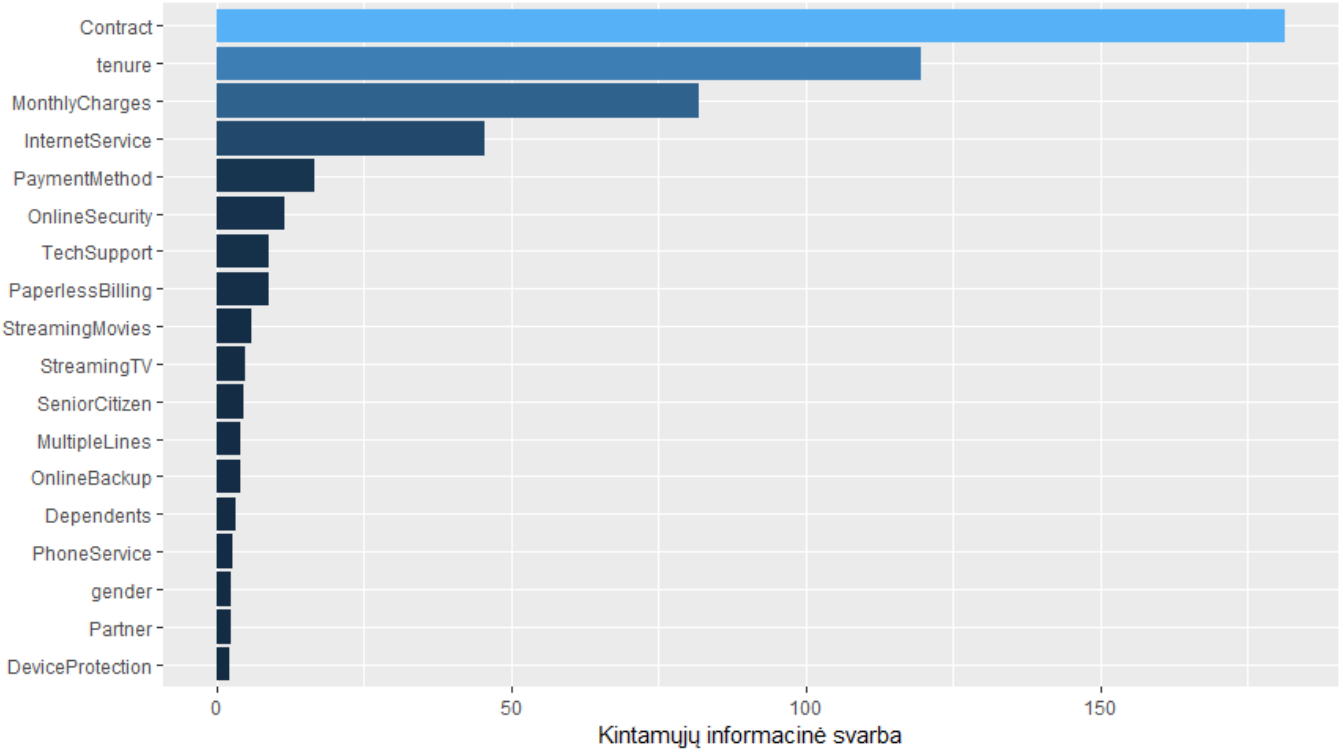
10 lentelė. Atsitiktinių miškų sumaišymo matrica, gaunama su ribine verte, kada ji optimizuojama remiantis tikėtinu maksimaliu pelnu

		Tikroji reikšmė			
		Nelojalus	Lojalus	Detekcijos suma	Preciziškumas
Detekcijos rezultatai	Nelojalus	1325	975	2300	57.609%
	Lojalus	544	4199	4743	88.53%
	Tikrųjų reikšmių suma	1869	5174	7043	
	Prisiminimas	70.894%	81.156%		
Bendrasis tikslumas:		78.432%			
Kappa:		0.485			

10 lentelėje pateikiama atsitiktinių miškų modelio prognozės sumaišymų matrica, kada ribinė vertė randama remiantis tikėtinu maksimaliu pelnu. Lentelėje matoma, jog nustatyta ribinė vertė suteikiama tokius rezultatus, jog apie 67 proc. klientų yra prognozuojami kaip lojalūs. Toks prognozuojamų lojalių klientų lygis sumažina jautrumo (dar vadinama prisiminimo) lygį. Gerumo

matų F_1 ir Kappa rodikliai pasikeičia nežymiai. Papildomai yra pateikiamos kitos sumaišymų matricos, gautos kiekvienu iš 9 lentelėje likusių modelių, priede.

Remiantis 8 lentelės rezultatais galima daryti prielaidą, jog analizuotai telekomunikacijų duomenų imčiai geresnius rezultatus parodė detekcijos modeliai. Būtina ištirti ar modeliavimo metodikų rezultatų skirtumai yra statistiškai reikšmingi[6]. Atlikus testą, gaunama p – reikšmė artima 0. Kadangi p – reikšmė yra mažesnė nei 0,05, todėl atmetama nulinė hipotezė, jog modelių tikslumas statistiškai reikšmingai nesiskiria.



17 pav. Geriausio modelio kintamųjų svarbumas telekomunikacijų duomenims

Taigi, geriausias modelis, tinkantis telekomunikacijų duomenų imties klientų lojalumui tirti yra detekcijos atsitiktinių miškų metodas. Šio metodo parametrai gaunami naudojant parametrų derinimo algoritmą, bei atsižvelgiant į nedidelį klasių disbalansą.

17 pav. pateikiama visų kintamųjų informacinė svarba sudarytojo modelio atžvilgiu. Kaip matoma, didžiausią įtaką atsitiktinio medžio sudarymui turi klientų sutarties tipas (Contract), paslaugų naudojimosi trukmė (tenure), mėnesinės išlaidos (MonthlyCharges) bei interneto paslaugų tipas (InternetService). Žvelgiant į modelį giliau, galima suteikti nelojaliam klientui būdingiausias kintamųjų kategorijas. Klientas, kuris paprastai yra nelojalus dažniausiai turi mėnesio trukmės sutartį, nesinaudoja interneto paslaugomis ir paslaugų naudojimo laikotarpis trumpesnis nei vidutinis. Remiantis modelio rezultatais, telekomunikacijų įmonei vertėtų labiausiai atkreipti dėmesį į klientus, kurie pasižymi šiais kriterijais.

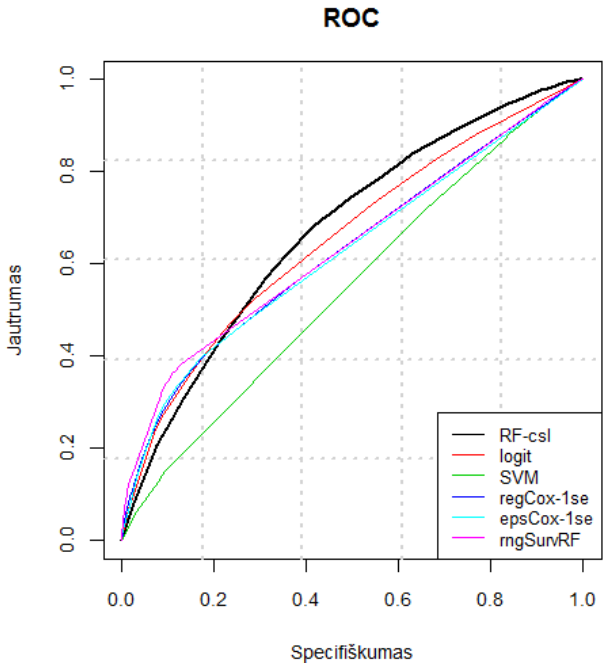
[6] Pasinaudojama internetinio tinklalapio http://www.biosoft.hacettepe.edu.tr/easyROC/ pateikiamais detekcijos tikslumo testais.

3.3. „Premium“ klubo klientų lojalumo prognozavimas

Kaip ir praėjusiame skyriuje, tokia pačia tvarka bus aptariama modeliavimo eiga ir rezultatai „Premium“ klubo klientų lojalumo duomenims. Nors ir modelio aiškinamųjų kintamųjų yra mažiau, tačiau modeliavimo metodai prisitaiko prie duomenų informacinių savybių bei sugeba parodyti gerus klientų lojalumo prognozavimo rezultatus.

Trumpai aptariamos modelių modifikacijos, kurios parodė geriausius ir tiksliausius rezultatus prognozuojant „Premium“ klubo duomenis:

- Detekcijos atsitiktinių miškų modelis parodė geriausius rezultatus, kada metodas buvo pakoreguotas, atsižvelgiant į nevienodą kiekį lojalių ir nelojalių klientų duomenų imtyje. Į tai atsižvelgiama panaudojus kaštams jautrų mokymąsi. Tai sprendimų medžių sudarymo modifikacija, kada kiekvieno medžio mazgo sudarymo metu atsižvelgiama į nurodytus lojalumo klasių kaštus.
- Kitas – logistinės regresijos modelis. Kaip ir telekomunikacijų, taip ir „Premium“ klubo duomenims pasirinktas greičiausias logistinės regresijos variantas, kuris sudaromas be papildomų įverčių derinimo algoritmų. Regresinis modelis analizuoja jau standartizuotus duomenis kartu su kategoriniams kintamiesiems sukurtais fiktyviais kintamaisiais.
- Atraminių vektorių metodui nebuvai taikomos algoritmo modifikacijos. Kaip jau buvo minėta 2.1.3 poskyryje, tik patikrinamos įvairios lygties baudos reikšmės C.
- Cox proporcingumo rizikos modelis. Elastinio tinklo reguliarizacijos parametras λ pasirenkamas ne mažiausias, tačiau modelio vieno paklaidų standartinio nuokrypio reikšmė. Su tokia λ reikšme, modeliui padedama išvengti persimokinimo. „Premium“ klubo duomenims, pasirenkama būtent vieno standartinio nuokrypio paklaidų reikšmė. Prie prognozuojamų rezultatų pateikiami du skirtingomis funkcijomis apskaičiuojami variantai.
- Reguliarizuotas atsitiktinių išlikimo miškų metodas. Tokia pati reguliarizacijos forma pateikia geriausius rezultatus, kaip ir telekomunikacijų įmonės duomenų atveju.

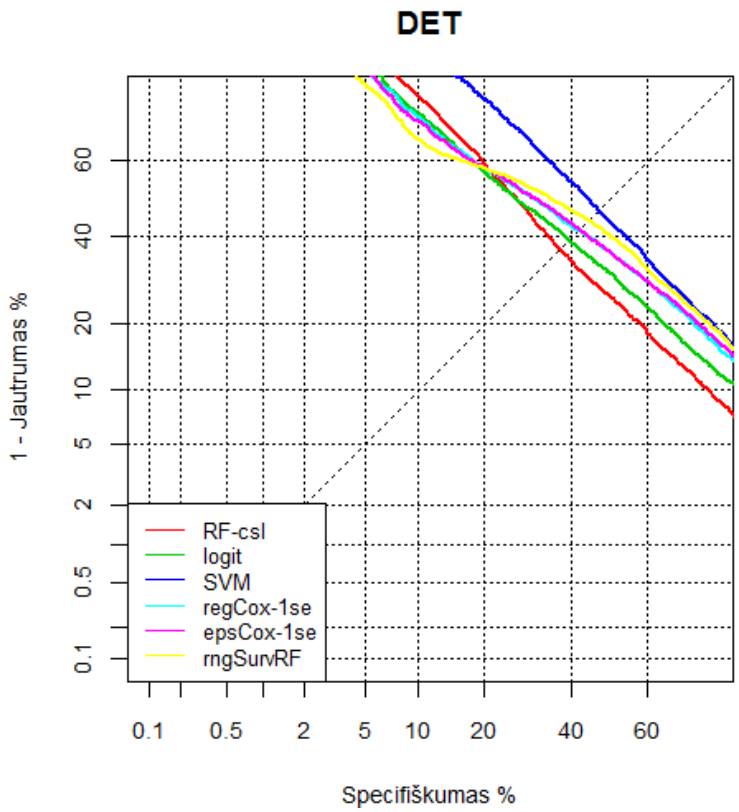


18 pav. „Premium“ klubo duomenų modelių ROC kreivės

Šiems duomenims panaudojama ta pati, anksčiau aprašyta procedūra. Pirmiausia aptariami gauti rezultatai grafiškai ir toliau aprašomi tikslūs skaitiniai rezultatai. Šiems duomenims, ROC kreivėje (žiūrėti 18 pav.) matomi šiek tiek didesni skirtumai tarp modelių rezultatų nei telekomunikacijų duomenų atžvilgiu.

Visi trys išlikimo modeliai tarpusavyje rodo labai panašius rezultatus ir net tam tikrais ribinės vertės atvejais lenkia net detekcijos modelius. Šiek tiek daugiau išsiskiria atsitiktinių miškų metodas su kaštams jautriu mokymosi algoritmu, kur AUC lygu 0,674. Matoma, jog besikeičiant ribinei vertei t atsitiktinių miškų modelio sprendimai nelojaliems klientams kategoriškai keičiasi, t. y. iki tam tikros ribinės vertės didžioji dauguma nelojalių klientų ir yra prognozuojami kaip nelojalūs. Po to (ribinei vertei pasikeitus), visi prognozavimo sprendimai pasikeičia ir modelis pradeda nustatyti visus klientus kaip nelojalius, todėl pradeda daugėti klientų FP dalyje – daugėja kiekis lojalių klientų, kurie modelio klasifikuojami kaip nelojalūs.

Logaritminė regresija nedaug skiriasi nuo atsitiktinių miškų metodo su AUC lygiu 0,655. Ganėtinai nuviliančiai atrodo atraminių vektorių metodas, kurio AUC – 0,549. Tai ganėtinai prastas rezultatas žinant, jog 0,5 žymi atsitiktinį klientų lojalumo spėjimą, nesinaudojant jokia papildoma informacija ir modeliais.



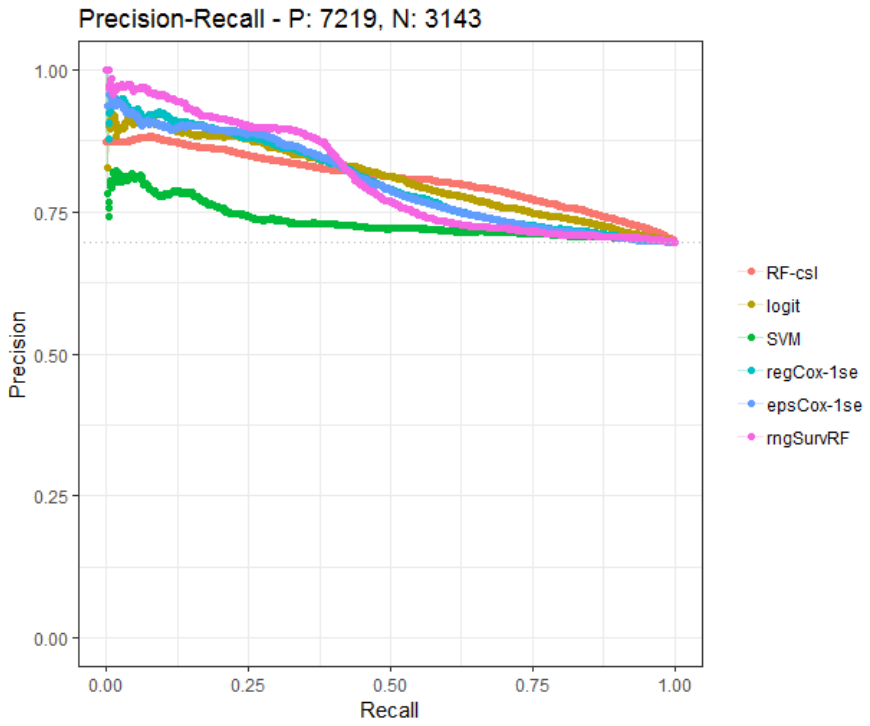
19 pav. „Premium“ duomenų modelių rezultatų DET kreivės

19 pav. pateikiamos modelių rezultatų DET kreivės. Kaip ir prieš tai, taip ir DET grafike matoma tendencija panaši – atsitiktiniai miškai turi mažiausius paklaidų rodiklius didžiojoje daugumoje ribinių verčių, po to seka logistinė regresija, visi išlikimo analizės modeliai ir galiausiai atraminių vektorių metodas. Kreivė nurodo kaip pasiskirsto neteisingas lojalių ir nelojalių klientų nustatytas keičiant ribinę vertę. Iš kreivės matoma, jog sumažinus ribinę vertę iki tam tikro lygio, atsitiktinių miškų modeliui rodo mažesnes abiejų klasių paklaidas. Padidinus ribinę vertę iki tam tikro

lygio, beveik visi modeliai apsiverčia ir mažiausias abiejų klasių paklaidas demonstruota atsitiktinių išlikimo miškų modelis. Vis dėlto, pasirenkant abiejų klasių lygių paklaidų lygį matoma, jog atsitiktinių miškų ERR lygus 36,82 proc. Kitų modelių EER didesnis, o atraminių vektorių metodo lygių paklaidų lygis didžiausias ir siekia net 46,91 proc.

Taip pat pastebima, jog visos kreivės yra ganėtinai tiesios ir įstrižos. Tai parodo, jog keičiant ribinę vertę, vienos klasės paklaidos proporcingai mažėja, kada kitos klientų klasės paklaida didėja. Iš kreivių išsiskiria atsitiktinių išlikimo medžių metodas, kuris pasiekia lygių paklaidų tašką ties 41,30 proc. Nors ir matomas kreivės išlinkimas ir mažiausių paklaidų rodymas ties tam tikromis ribinėmis vertėmis keičiasi neproporcingai, tačiau tai nepasiekia atsitiktinių miškų bendro paklaidų lygio.

Galiausiai aptariamas preciziškumo – jautrumo grafikas (žiūrėti 20 pav.). Matoma ganėtinai įdomi situacija, kada metodų gerumas keičiasi tarpusavyje su kintančia ribine verte. Jeigu jautrumas nustatomas apie 0,4 ir daugiau, tada geriausiai pasirodo atsitiktinių miškų metodas, tačiau, jeigu jautrumas gaunamas mažesni, tada modeliai tarpusavyje apsikeičia ir jau tada geriausius rezultatus demonstruoja atsitiktinių išlikimo miškų modelis. Tai galioja penkiems aprašomiems metodams. Iš jų išsiskiria atraminių vektorių metodas, kuris su bet kokia fiksuojama ribine verte rodo blogiausius rezultatus.



20 pav. „Premium“ klubo duomenų modelių rezultatų preciziškumo – jautrumo kreivės, kur horizontalioje ašyje jautrumas, o vertikalioje – preciziškumas

Kaip matoma 11 lentelėje, didžiausią plotą po ROC kreive užfiksuoja ir detekcijos ir išlikimo modelių tarpe atsitiktinių miškų metodai. Lygių paklaidų rodiklis EER rodo, jog atsitiktinių miškų metodui, pasirinkus tokią ribinę vertę, kuri pateiktų vienodą tikslumą abiejų klientų klasių atžvilgiu, lygių paklaidų vertė siektų 36,82 proc. Kitiems modeliams šis rodiklis didesnis, ypač atraminių vektorių modelo atveju, kur lygių paklaidų rodiklis siekia net 46,91 proc. Laiko atžvilgiu, logistinė regresija nepralenkiama kryžminio patikrinimu metu, kada $k = 10$ regresijai apskaičiuoti kintamųjų parametrų įverčius užtruko 3 sekundes. Atsitiktinių miškų modeliai taip pat pakankamai greitai parametrizuojami: detekcijos atsitiktiniai miškai užtruko šiek tiek daugiau nei 4 minutes, o atsitiktinių

išlikimo medžių algoritmas užtruko mažiau nei 2 minutes. Labiausiai laiko reikalaujantis procesas yra atraminių vektorių metodas. Jis užtruko daugiau nei 30 minučių.

11 lentelė. „Premium“ klubo duomenų modeliavimo rezultatai

Modeliai	**Detekcija**			**Išlikimas**		
	Atsitiktiniai miškai	**Logistinė regresija**	**Atraminiai vektoriai**	**Cox rizikos metodas**	**Cox rizikos metodas (epsgo)**	**Atsitiktiniai išlikimo miškai**
AUC	**0,674**	0,655	0,549	0,630	0,625	**0,638**
EER (%)	**36,82**	39,09	46,91	41,34	41,79	**41,30**
Mokymo laikas (min)	4:07	**0:03**	34:14	8:07	17:27	**1:37**

Taigi, atsižvelgiant į 11 lentelėje pateiktus rezultatus, pasirenkami atsitiktinių miškų modeliai. Kiekvienam iš modelių apskaičiuojamos dvi ribinės vertės, nuo kurių priklauso, kaip sumaišymų matricoje pasidalins klasifikuojami klientai. Kada kiekvienam iš metodų parenkamos dvi ribinės vertės, gaunami keturi modelių variantai. Kiekvieno iš modeliavimo varianto gerumo matai pateikiami 12 lentelėje.

Atsižvelgiant į tai, jog duomenyse pateikiama klientų finansinių išlaidų ir pajamų informacija, galima teigti, jog tokiai kompanijai riektų visai kitokios klientų išlaikymo programos nei tai daro telekomunikacijų įmonės. Kada klientų sutarčių kainos siekia milijoną ir daugiau, suprantama, jog išlaikymo programa turėtų suteikti dideles nuolaidas. Jei programos metu būtų siūlomos papildomos paslaugos ar nuolaidos už vidutiniškai $64,00, tada klientų nuomonės tai greičiausiai nepaveiktų. Dėl šių priežasčių, tikėtinam maksimaliam pelnui apskaičiuoti buvo pasirenkamos didesnės kaštų vertės. Klientų CLV apskaičiuota $118 000,00. Išlaikymo programos išlaidos paskaičiuojamos vienam klientui padauginant numatytą mėnesinę nuolaidą $600,00 iš vidutinės klientų gyvavimo trukmės – 24 mėnesiai. Taip gaunama, jog išlaikymo programos išlaidos lygios $14 400,00, o susisiekimo su klientu kaina $3 000,00. Pateikiant tokius parametrus ribinės vertės apskaičiavimo funkcijai, gaunama, jog atsitiktinių miškų modelio atveju, tikėtina maksimali kliento vertė siekia $6 426,00, o tuo tarpu atsitiktinių išlikimo miškų atveju – $5 866,00.

12 lentelė. „Premium“ klubo duomenų modeliavimo gerumo matai

Metodika	**Modelis**	**Ribinės vertė**	**Bendrasis Tikslumas**	**Jautrumas**	**Specifiškumas**	**Preciziškumas**	**EMP klientui**	**F_1**	**Kappa**
Detekcija	Atsitiktiniai miškai	EER	**0,630**	0,634	**0,629**	**0,426**	–	**0,703**	**0,231**
		EMP	0,610	0,671	0,583	0,412	**$6426,00**	0,676	0,216
Išlikimas	Atsitiktiniai išlikimo miškai	EER	0,558	0,556	0,559	0,354	–	0,638	0,099
		EMP	0,540	**0,709**	0,466	0,366	$5866,00	0,585	0,139

Lentelėje matoma, jog beveik visi modelių gerumo matai rodo, jog atsitiktinių miškų modelis yra geresnis. Ribinės vertės pakeitimas nuo vienodų detekcijos paklaidų suderinimo iki maksimalaus tikėtino pelno, pakeitė gerumo matus skirtingai. Atsitiktinių miškų atveju, tik nežymiai pasikeitė visi gerumo matai. Išlikimo miškų atveju, rezultatai labiau pasikeitė, tikslumas ir specifiškumas sumažėjo, atitinkamai ir gerumo matas F_1 sumažėjo.

Būtina patikrinti metodikų rezultatų skirtumą klasifikavimo tikslumo statistiniu testu. Atlikus testą, gaunama p – reikšmė artima 0. Kadangi p – reikšmė yra mažesnė nei 0,05, todėl atmetama nulinė hipotezė, jog modeliai rezultatai statistiškai reikšmingai nesiskiria. Galima teigti, jog „Premium“ klubo duomenų imčiai geresnius rezultatus pateikė detekcijos atsitiktinių miškų metodas.

Toliau pateikiama atsitiktinių miškų sumaišymų matrica, kada ribinė vertė nustatytą tikėtino maksimalaus pelno atžvilgiu (žiūrėti 13 lentelę). Kitų trijų atvejų, aprašytų 12 lentelėje, sumaišymų matricos pateikiamos priede. Žvelgiant į 13 lentelę pastebima, jog ir lojalūs ir nelojalūs klientai yra klasifikuojami didesniu nei 60 proc. tikslumu. Tai yra ganėtinai silpnas modelio detekcijos įvertinimas. Taip pat iš kappa koeficiento matoma, jog 0,216 yra tik pakankamas, šiek tiek geresnis nei nepatenkinamas lygis.

13 lentelė. Atsitiktinių miškų sumaišymo matrica su ribine verte, kada ji optimizuojama remiantis tikėtinu maksimaliu pelnu

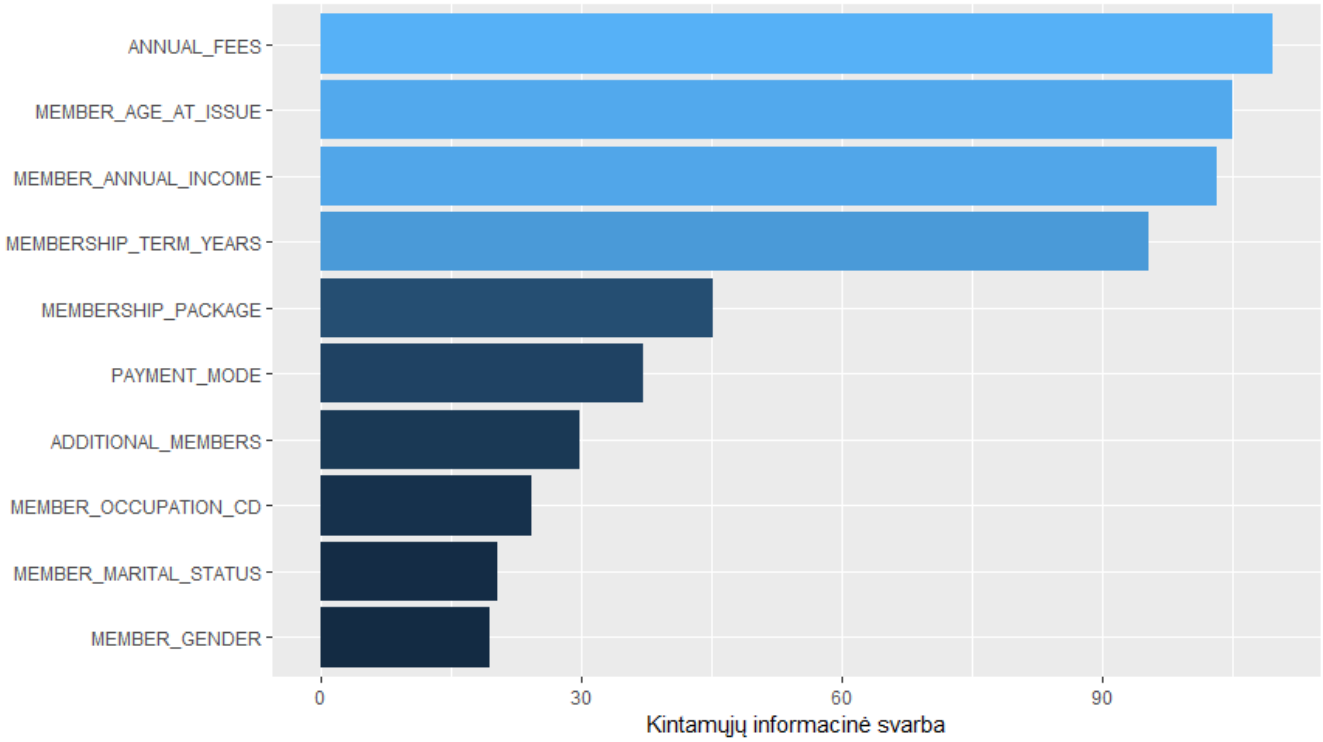
		Tikroji reikšmė			
		Nelojalus	Lojalus	Detekcijos suma	Preciziškumas
Detekcijos rezultatai	Nelojalus	2109	3010	5119	41.199%
	Lojalus	1034	4209	5243	80.278%
	Tikrųjų reikšmių suma	3143	7219	10362	
	Prisiminimas	67.101%	58.304%		
Bendrasis tikslumas:		60.973%			
Kappa:		0.216			

Galiausiai aptariamas atsitiktinių miškų modelis ir jame esančių kintamųjų svarba, arba kitaip tariant, informacinė vertė (žiūrėti 21 pav.). Pateikiama visų kintamųjų informacinė svarba sudarytojo modelio atžvilgiu, kurie surūšiuoti nuo svarbiausio viršuje iki mažiausiai įtakos turinčio klientų detekcijai apačioje. Sprendimų medžių sudarymui didžiausią įtaką turi metinių mokesčių dydis (ANNUAL_FEES), t. y. kokia suma nuo klientų yra nuskaitoma kiekvienais metais už tam tikras specializuotas paslaugas klientams. Po to svarbus yra kliento amžius sutarties sudarymo metu (MEMBER_AGE_AT_ISSUE), metinės kliento pajamos (MEMBER_ANNUAL_INCOME) bei planuotas sutarties galiojimo laikotarpis (MEMBERSHIP_TERM_YEARS). Taigi, atliktus papildomą kintamųjų analizę buvo ieškoma, su kuriomis reikšmėmis didėja tikimybė klientui atsiskirti. Didžiausią tikimybę palikti įmonę turi tie klientai, kurie:

- Yra mažiau apmokestinti. Klientai, kurių metiniai mokesčiai yra mažesni nei kintamojo mediana yra linkę dažniau atsiskirti nuo kompanijos.
- Klientai, kurie sudarymo sutarties metu buvo jaunesni nei 46 metai, taip pat labiau tikėtina jog paliks įmonę nei vyresniojo amžiaus klientai. Pabrėžiama, jog tam tikri klientai užfiksuoti garbaus amžiaus, todėl mirties atvejis nėra vertinamas kaip atsiskyrimas.
- Kliento metinės pajamos. Jos taip pat, kaip ir metiniai mokėjimai, paliekami modeliuose ir metoduose. Kai metinės pajamos yra mažesnės nei mediana, kliento palikimo tikimybė yra didesnė.

- Sutarties nustatytasis laikas. Klientai siekiantys trumpesnių sutarties laikotarpių yra labiau linkę palikti kompaniją.

Matoma, jog sprendimo medžių tikslumui vieno svarbiausio kintamojo nebūtų galima išskirti, nes, remiantis detekcijos metodu, tik tam tikrų kintamųjų kategorijų kombinacija praneša apie klientus galimai labiau linkusius atsiskirti. Tačiau, apibendrinant antrojo duomenų rinkinio analizę, matoma, jog yra pastebimas duomenų informatyvumo trūkumas. Tokia įžvalga daroma, atsižvelgiant į vidutinius gerumo matų rezultatus kiekvieno iš nagrinėtų metodų atveju.



21 pav. Geriausio modelio kintamųjų svarbumas „Premium“ duomenims

Apibendrinant abiejų duomenų rinkinių analizės rezultatus, matoma, jog nepralenkiamas yra atsitiktinių miškų metodas. Abiem duomenų rinkiniams, modelių daugumos mažinimo algoritmai suteikė reikšmingą rezultatų pagerinimo, nes duomenims būdingas skirtingų klasių disbalansas. Kaip ir tikėtasi, ribinės vertės nustatymas tikėtino didžiausio pelno atžvilgiu, fiksuoja prastesnius gerumo matų rezultatus. Tačiau, kaip jau buvo argumentuojama 1.4 poskyryje, toks ribinės vertės nustatymas yra prasmingesnis įmonės pelno atžvilgiu.

Kiekvieno kliento turimi duomenys yra vertingi, tačiau labiau vertėtų stengtis identifikuoti tuos, kurie yra linkę pasitraukti. Tokiu atveju daugiau žinoma apie nelojaliems klientams įprastą elgseną ir geriau atpažįstami jiems būdingi bruožai. Taigi, nors ir detekcijos metodai sėkmingiau atpažįsta nelojalius klientus nei išlikimo metodai, tačiau abi modeliavimo rūšys sėkmingai padeda identifikuoti klientus, kurie ketina atsisakyti nagrinėtų įmonių paslaugų.

Išvados

Remiantis išsikeltais uždaviniais šio darbo metu buvo prieita prie sekančių:

1. Literatūroje klientų lojalumas yra minimas kaip vienas iš svarbiausių kiekvienos organizacijos tikslų. Taip pat pastebima realių pavyzdžių, kada padidėjęs klientų lojalumas ženkliai padidina įmonės pelną. Todėl, norėdamos išlaikyti klientų lojalumą, įmonės yra suinteresuotos vykdyti klientų išsaugojimo programas. Tyrimai atskleidžia, jog tai – sėkmingai naudojama praktika įmonėse.
2. Informaciniuose šaltiniuose galima rasti nemažai publikuojamų tyrimų, kurie apžvelgia klientų lojalumą bei atlieka detekcijos modeliavimą. Nors literatūroje dažniausiai randama, jog detekcijos uždaviniai naudojami telekomunikacijų įmonėse, tačiau publikuojama lojalumo tyrimų ir kitose industrijų srityse. Priklausomai nuo analizuojamos įmonės, tiriamų klientų ir kintamųjų skaičius yra įvairus. Nors industrijų sritys ir duomenų dydis yra įvairus, tačiau tyrimuose naudojami metodai dažniausiai kartojasi. Populiariausi iš jų: Atsitiktiniai miškai, logistinė regresija bei neuroniniai tinklai.
3. Šiame darbe naudojami detekcijos ir išlikimo metodikos. Panaudoti detekcijos metodai: logistinė regresija, atsitiktiniai miškai ir atraminių vektorių metodas. Išlikimo metodai: Cox proporcingumo rizikos metodas bei atsitiktiniai išlikimo miškai. Beveik visiems metodams išbandomos daugumos mažinimo procedūros ir kiti modelių parametrų reguliarizacijos procesai.
4. Atlikus pradinę duomenų analizę pastebėta tam tikrų kintamųjų svarba. Pastebėtas požymis tinkantis abiem duomenų rinkiniams, jog klientai, kurių sutarties terminas yra ilgesnis – labiau linkę būti lojaliais kompanijai. Telekomunikacijų duomenims atsitiktinių miškų metodas fiksuoja 23,18 proc. lygių paklaidų vertę, kada atsitiktinių išlikimo miškų metodas – 25,88 proc. „Premium“ klubo duomenims atitinkamų metodų lygių paklaidų vertės siekia 36,82 proc. ir 41,30 proc.
5. Darbo metu tiriamoms duomenų imtims geresnius rezultatus parodė detekcijos modeliai. Nors ir išlikimo modelių gerumo matai nedaug atsiliko nuo detekcijos metodų, tačiau patikrinimus skirtumus testu, gaunamas statistiškai reikšmingas skirtumas. Patikrinus metodų adekvatumą ir tikslumą, atsitiktinių miškų metodas nustatytas kaip geriausiai prisitaikantis modelis prie skirtingų duomenų imčių.

Taigi, pagal keliamus uždavinius ir pasiektus rezultatus galima teigti, jog darbe panaudoti modeliavimo metodai sėkmingai padeda identifikuoti klientus, kurie ketina atsisakyti įmonės paslaugų.

Literatūros sąrašas

1. Anil Kumar, D., & Ravi, V. (2008). *Predicting credit card customer churn in banks using data mining.* Data Analysis Techniques and Strategies. doi:10.1504/IJDATS.2008.020020
2. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., & Nielsen, H. (2000). *Assessing the Accuracy of Prediction Algorithms for Classification: An Overview.* University Of California. Nuskaityta iš http://www.igb.uci.edu/~pfbaldi/publications/journals/predreview00.pdf
3. Blattberg, R. C., Kim, B. D., & Neslin, S. A. (2008). Database Marketing: Analyzing and Managing Customers. Springer.
4. Buckinx, W., & Van den Poel, D. (2005). *Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting.* European Journal of Operational Research. doi:10.1016/j.ejor.2003.12.010
5. Burez, J., & Van den Poel, D. (2007). *CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services.* Expert Systems with Applications. doi:10.1016/j.eswa.2005.11.037
6. Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 4626-4636. doi:10.1016/j.eswa.2008.05.027
7. Chen, Z. Y., Fan, Z. P., & Sun, M. (2012). *A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data.* European Journal of Operational Research. doi:10.1016/j.ejor.2012.06.040
8. Clay, T. (2017). *20 Customer Retention Strategies.* Nuskaityta iš Marketing Wizdom: https://marketingwizdom.com/strategies/retention-strategies
9. Coussement, K., & Van den Poel, D. (2008). *Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques.* Expert Systems with Applications. doi:10.1016/j.eswa.2006.09.038
10. Crumer, A. M. (2011). *Comparison Between Weibull and Cox Proportional Hazards Models.* Southeast Missouri State University. Nuskaityta iš https://krex.k-state.edu/dspace/bitstream/handle/2097/8787/AngelaCrumer2011.pdf
11. Čekanavičius, V., & Murauskas, G. (2014). *Taikomoji regresinė analizė socialiniuose tyrimuose.* Vilnius: Vilniaus universiteto leidykla. ISBN: 9786094593000
12. Forbes. (2011). *Bringing 20/20 Foresight to Marketing: CMOs Seek a Clearer Picture of the Customer.* Nuskaityta iš ftp://ftp.software.ibm.com/software/hk/pdf/product_tab_03_wp-forbes-bringing-foresight-to-marketing_4.pdf
13. Galetto, M. (2015). *What is Customer Retention?* Nuskaityta iš NG Data: https://www.ngdata.com/what-is-customer-retention/

14. Glady, N., Baesens, B., & Croux, C. (2008). Modeling churn using customer lifetime value. *European Journal of Operational Research*, 402-411. doi:10.1016/j.ejor.2008.06.027
15. Hosseini, S. M., Maleki, A., & Gholamian, M. R. (2009). *Cluster analysis using data mining approach to develop CRM methodology.* Tehran: Elsevier. doi:10.1016/j.eswa.2009.12.070
16. Yu, X., Guo, S., Guo, J., & Huang, X. (2011). *An extended support vector machine forecasting framework for customer churn in e-commerce.* Expert Systems with Applications. doi:10.1016/j.eswa.2010.07.049
17. Lechon, N. S., Llorente, J. I., Ruiz, V. O., & Vilda, P. G. (2006). *Methodological issues in the development of automatic systems for voice pathology detection.* Biomedical Signal Processing and Control. doi:10.1016/j.bspc.2006.06.003
18. Lemmens, A., & Gupta, S. (2013). Managing Churn to Maximize Profits. Harvard: Harvard Business School. Nuskaityta iš https://www.hbs.edu/faculty/Publication%20Files/14-020_3553a2f4-8c7b-44e6-9711-f75dd56f624e.pdf
19. Migueis, V. L., Van den Poel, D., Camanho, A. S., & Falcao e Cunha, J. (2012). Modeling partial customer churn: On the value of first product-category purchase sequences. *Expert Systems with Applications*, 11250–11256. doi:10.1016/j.eswa.2012.03.073
20. Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. *Journal of Marketing Research*, 204–211. Nuskaityta iš http://wak2.web.rice.edu/bio/My%20Reprints/Defection%20detection.pdf
21. Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: a literature review and classification. *Expert Systems with Applications*, 2592–2602. doi:10.1016/j.eswa.2008.02.021
22. Pendharkar, P. C. (2009). *Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services.* Expert Systems with Applications. doi:10.1016/j.eswa.2008.08.050
23. Qi, J., Zhang, Y., Zhang, Y., & Shi, S. (2006). *TreeLogit model for customer churn.* IEEE Asia-Pacific conference on services computing. doi:10.1109/APSCC.2006.111
24. Reinartz, J. W., & Kumar, V. (2000). *On the profitability of long-life customers in a non contractual setting: An empirical investigation and implications for marketing.* Fontainebleau: INSEAD. Nuskaityta iš https://flora.insead.edu/fichiersti_wp/inseadwp2000/2000-04.pdf
25. Risselada, H., Verhoef, P. C., & Bijmolt, T. H. (2010). Staying Power of Churn Prediction Models. *Journal of Interactive Marketing*, 198-208. doi:10.1016/j.intmar.2010.04.002

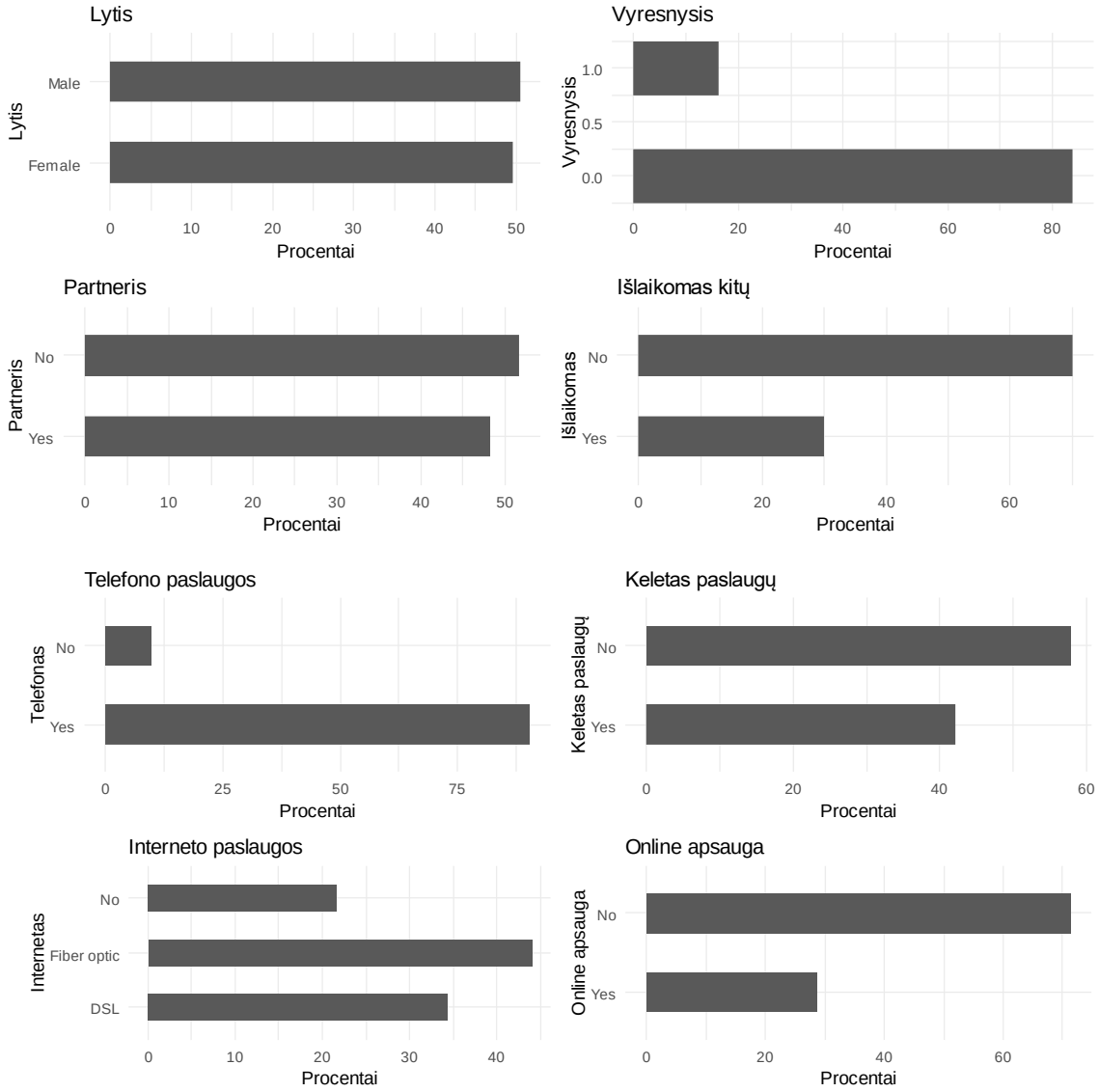
26. Shulga, D. (2018). *Towards Data Science*. Nuskaityta iš Medium: https://towardsdatascience.com/5-reasons-why-you-should-use-cross-validation-in-your-data-science-project-8163311a1e79
27. Suchacka, G., Skolimowska – Kulig, M., & Potempa, A. (2015). *Classification of E-Customer Sessions Based on Support Vector Machine.* Nuskaityta iš http://www.scs-europe.net/dlib/2015/ecms2015acceptedpapers/0594-dis_ECMS2015_0120.pdf
28. Tsai, C. F., & Lu, Y. H. (2009). *Customer churn prediction by hybrid neural.* Expert Systems with Applications. doi:10.1016/j.eswa.2009.05.032
29. Tsai, C. F., & Lu, Y. H. (2010). *Data Mining Techniques in Customer Churn Prediction.* Recent Patents on Computer Science.
30. Van den Poel, D., & Lariviere, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research*, 196-217. doi:10.1016/S0377-2217(03)00069-9
31. Verbeke, W., Dejarger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit. *European Journal of Operational Research*, 211-229. doi:10.1016/j.ejor.2011.09.031
32. Verbraken, T. (2013). *Thomas Verbraken, Business-Oriented Data Analytics: Theory and Case Studies.* Liuven university. Nuskaityta iš http://www.dataminingapps.com/wp-content/uploads/2015/04/PhDThesis-Thomas-Verbraken.pdf
33. Verbraken, T., Bravo, C., Weber, R., & Baesens, B. (2014). Development and application of consumer credit scoring models using. *European Journal of Operational Research*, 505-513. doi:10.1016/j.ejor.2014.04.001
34. Weathers, B., & Cutler, R. (2017). *Comparison of Survival Curves Between Cox Proportional Hazards, Random Forests, and Conditional Inference Forests in Survival Analysis.* Utah State University. Nuskaityta iš https://digitalcommons.usu.edu/cgi/viewcontent.cgi?article=1936&context=gradreports
35. Weiss, G. M. (2004). Mining with Rarity: A Unifying Framework. *AT&T Laboratories*, 7-19. Nuskaityta iš https://storm.cis.fordham.edu/gweiss/papers/sigkdd04.pdf
36. Xie, Y., Li, X., Ngai, E. W., & Ying, W. (2009). *Customer churn prediction using improved balanced random forests.* Expert Systems with Applications. doi:10.1016/j.eswa.2008.06.121

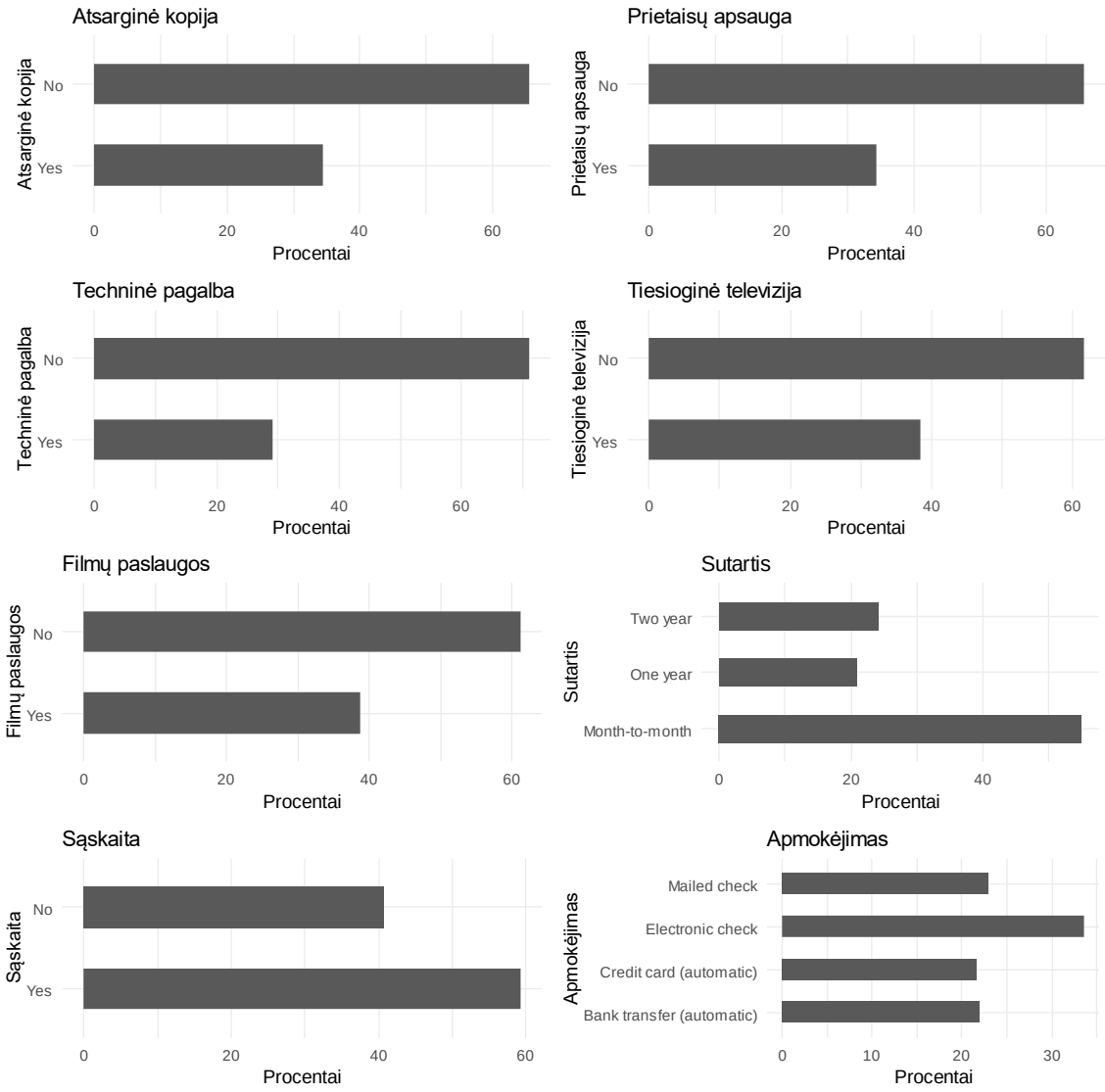
Priedai

1 priedas

Telekomunikacijų žvalgomosios analizės papildomos diagramos

Telekomunikacijų įmonės kategorinių kintamųjų pasiskirstymo histogramos.





Mėnesiniai mokesčiai:

Quantile statistics

Minimum	18.25
5-th percentile	19.65
Q1	35.5
Median	70.35
Q3	89.85
95-th percentile	107.4
Maximum	118.75
Range	100.5
Interquartile range	54.35

Descriptive statistics

Standard deviation	30.09
Coef of variation	0.46463
Kurtosis	-1.2573
Mean	64.762
MAD	26.222
Skewness	-0.22052
Sum	456120
Variance	905.41
Memory size	55.1 KiB

Paslaugų naudojimosis trukmė mėnesiais:

Quantile statistics

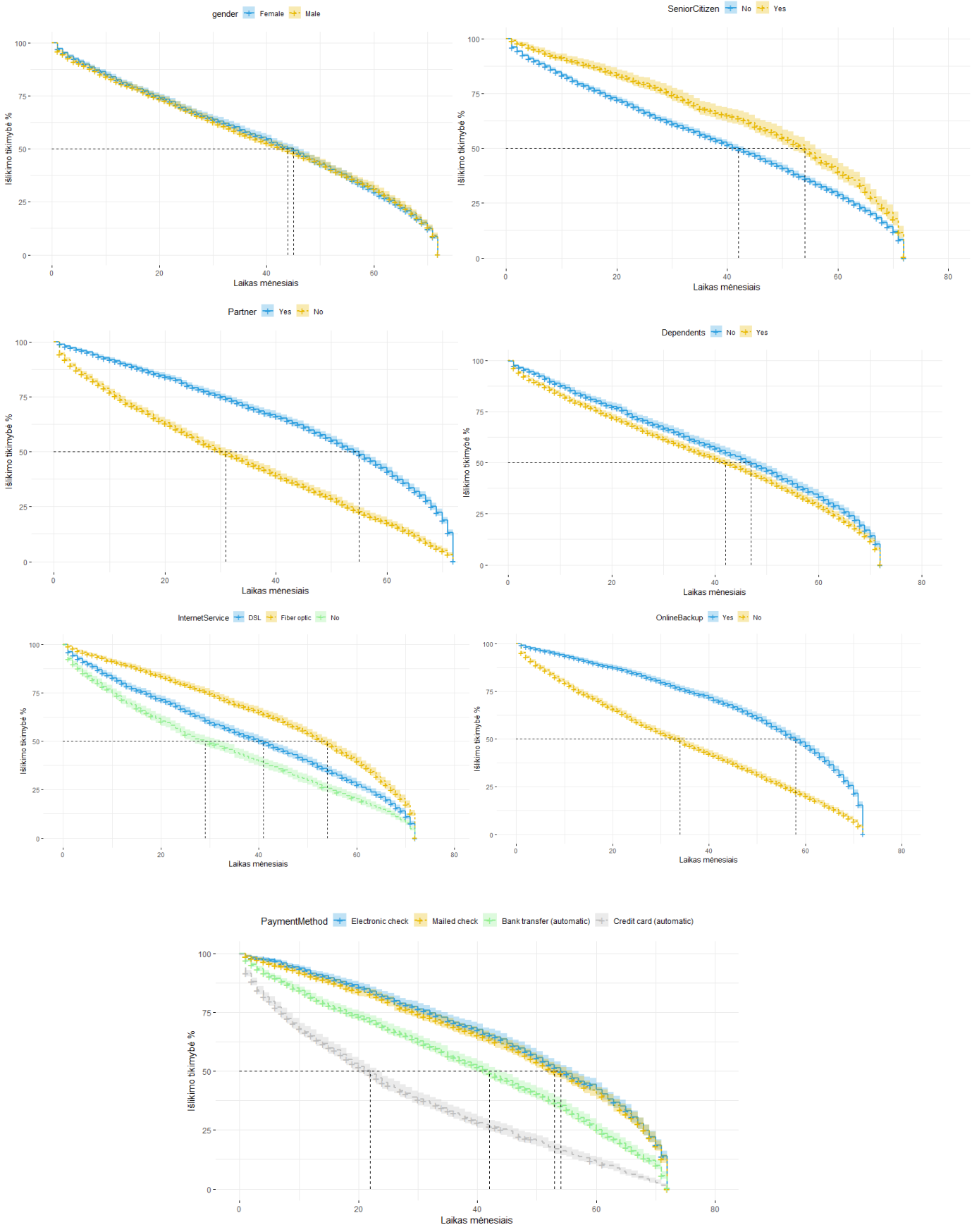
Minimum	0
5-th percentile	1
Q1	9
Median	29
Q3	55
95-th percentile	72
Maximum	72
Range	72
Interquartile range	46

Descriptive statistics

Standard deviation	24.559
Coef of variation	0.75868
Kurtosis	-1.3874
Mean	32.371
MAD	21.873
Skewness	0.23954
Sum	227990
Variance	603.17
Memory size	55.1 KiB

2 priedas

Kaplan – Meier kategorinių kintamųjų kreivės telekomunikacijų klubo duomenims



3 priedas

„Premium“ klubo žvalgomosios analizės santrumpa

Papildomas klientų kiekis susijęs su pastaruoju:

Value	Count	Frequency (%)
0	3050	29.4%
3	2488	24.0%
1	2478	23.9%
2	2346	22.6%

Klientų metinių pajamų statistika:

Quantile statistics

Minimum	0
5-th percentile	100000
Q1	100000
Median	118210
Q3	190000
95-th percentile	496500
Maximum	10100000
Range	10100000
Interquartile range	90000

Descriptive statistics

Standard deviation	268870
Coef of variation	1.5037
Kurtosis	585.22
Mean	178810
MAD	97689
Skewness	19.127
Sum	1852900000
Variance	72294000000
Memory size	81.0 KiB

Sutarties tipo histograma:

Value	Count	Frequency (%)
TYPE-B	6809	65.7%
TYPE-A	3553	34.3%

Planuoto partnerystės laiko kintamojo statistika:

Quantile statistics

Minimum	9
5-th percentile	12
Q1	12
Median	19
Q3	37
95-th percentile	82
Maximum	102
Range	93
Interquartile range	25

Descriptive statistics

Standard deviation	22.428
Coef of variation	0.76298
Kurtosis	1.602
Mean	29.395
MAD	17.267
Skewness	1.5696
Sum	304591
Variance	503
Memory size	81.0 KiB

Klientų amžiaus statistika:

Quantile statistics

Minimum	0
5-th percentile	25
Q1	37
Median	46
Q3	57
95-th percentile	70
Maximum	92
Range	92
Interquartile range	20

Descriptive statistics

Standard deviation	13.897
Coef of variation	0.29696
Kurtosis	-0.29568
Mean	46.798
MAD	11.293
Skewness	0.14222
Sum	484926
Variance	193.13
Memory size	81.0 KiB

Klientų metinių pajamų statistika:

Quantile statistics

Minimum	9996
5-th percentile	225000
Q1	400000
Median	550000
Q3	1000000
95-th percentile	3000000
Maximum	1000000000
Range	999990000
Interquartile range	600000

Descriptive statistics

Standard deviation	17572000
Coef of variation	12.109
Kurtosis	2700.5
Mean	1451100
MAD	1496600
Skewness	50.214
Sum	12491000000
Variance	308770000000000
Memory size	81.0 KiB

Lyties kintamojo histograma:

Value	Count	Frequency (%)
M	7330	70.7%
F	2421	23.4%
(Missing)	611	5.9%

Šeimyninės padėties histograma:

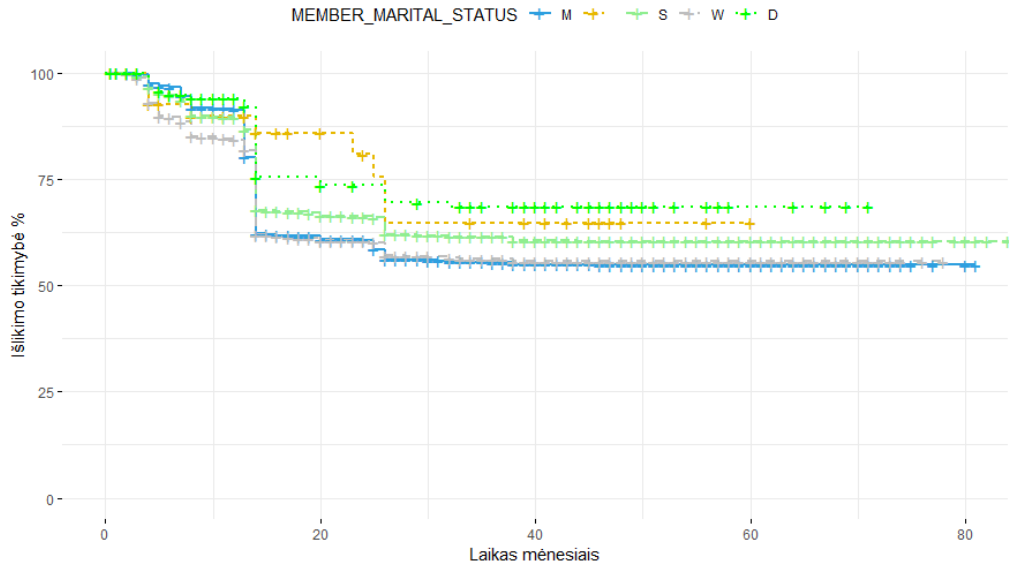
Value	Count	Frequency (%)
M	6430	62.1%
S	1144	11.0%
W	146	1.4%
D	45	0.4%
(Missing)	2597	25.1%

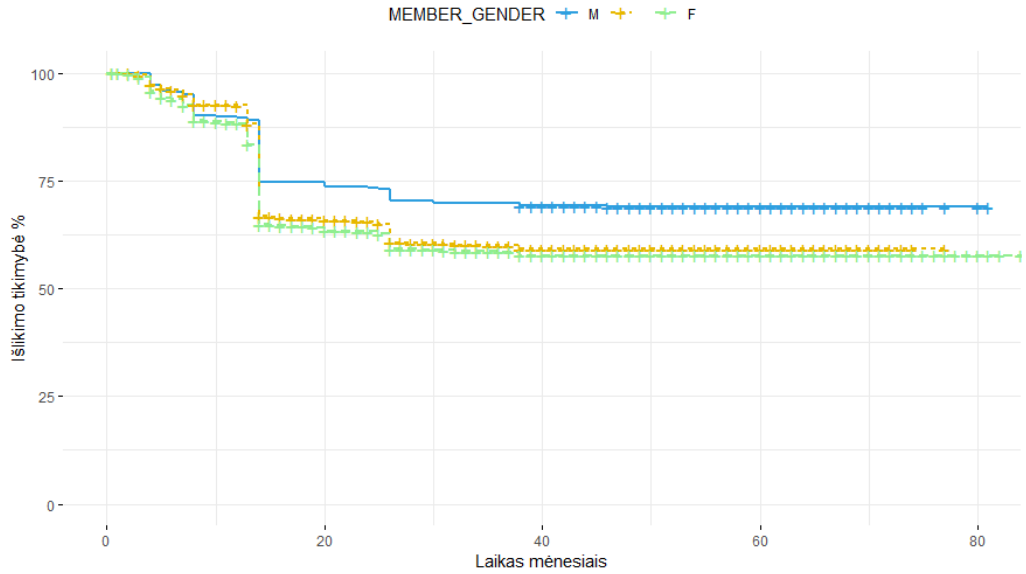
Apmokėjimo tipo histograma:

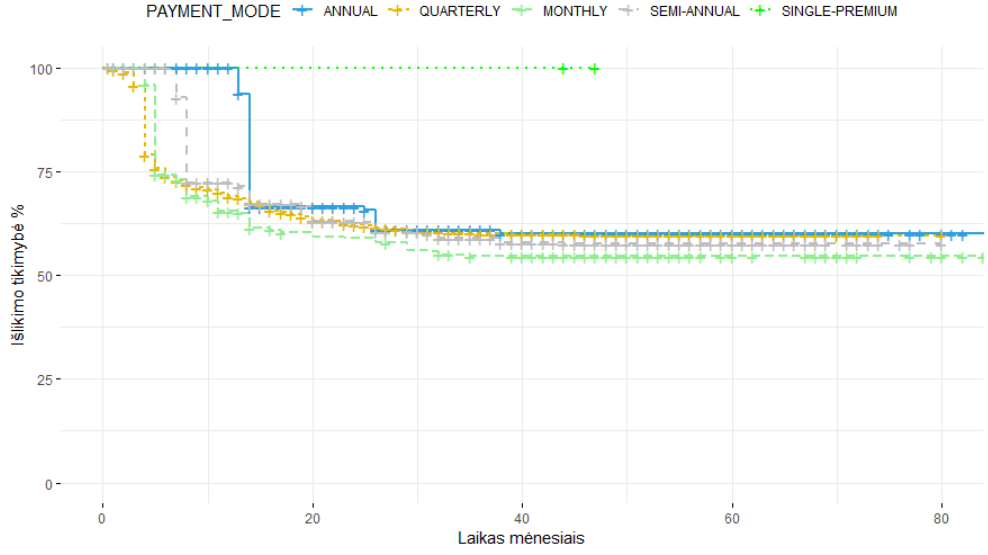
Value	Count	Frequency (%)
ANNUAL	6589	63.6%
MONTHLY	1881	18.2%
SEMI-ANNUAL	1493	14.4%
QUARTERLY	390	3.8%
SINGLE-PREMIUM	9	0.1%

4 priedas

Kaplan – Meier kategorinių kintamųjų kreivės „Premium“ klubo duomenims

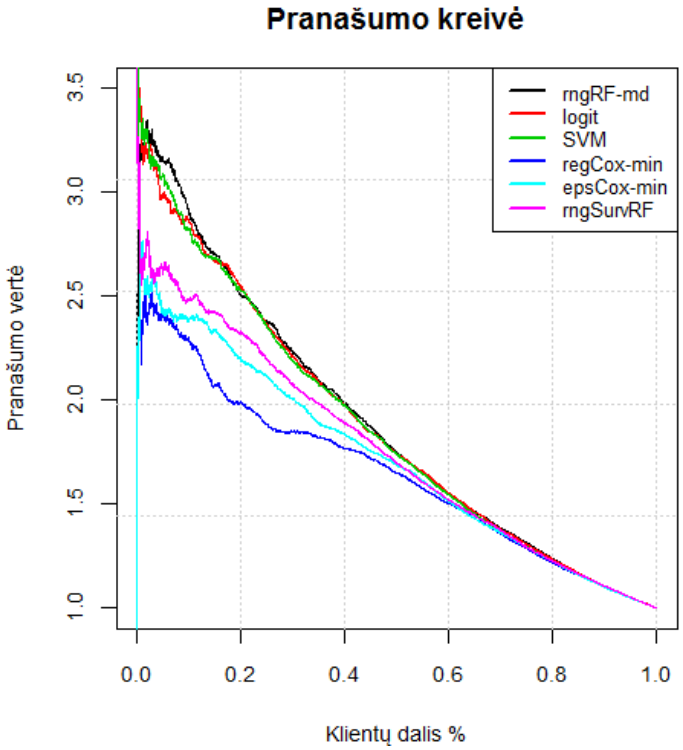






5 priedas

Telekomunikacijų duomenų panašumo kreivė



6 priedas

Telekomunikacijų duomenų sumaišymų matricos

Detekcijos atsitiktinių miškų sumaišymo matrica su ribine verte EER taške:

		Tikroji reikšmė			
		Nelojalus	Lojalus	Detekcijos suma	Preciziškumas
Detekcijos rezultatai	Nelojalus	1437	1186	2623	54.785%
	Lojalus	432	3988	4420	90.226%
	Tikrųjų reikšmių suma	1869	5174	7043	
	Prisiminimas	76.886%	77.078%		

Bendrasis tikslumas:	77.027%
Kappa:	0.478

Atsitiktinių išlikimo miškų sumaišymo matrica su ribine verte EER taške:

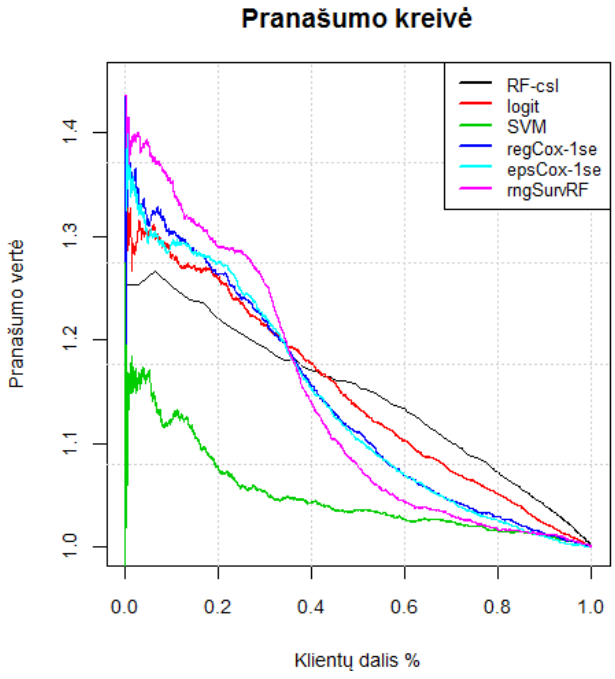
		Tikroji reikšmė			
		Nelojalus	Lojalus	Detekcijos suma	Preciziškumas
Detekcijos rezultatai	Nelojalus	1386	1354	2740	50.584%
	Lojalus	483	3820	4303	88.775%
	Tikrųjų reikšmių suma	1869	5174	7043	
	Prisiminimas	74.157%	73.831%		
Bendrasis tikslumas:		73.917%			
Kappa:		0.418			

Atsitiktinių išlikimo miškų sumaišymo matrica, gaunama su ribine verte, kada ji optimizuojama remiantis tikėtinu maksimaliu pelnu

		Tikroji reikšmė			
		Nelojalus	Lojalus	Detekcijos suma	Preciziškumas
Detekcijos rezultatai	Nelojalus	383	189	572	66.958%
	Lojalus	1486	4985	6471	77.036%
	Tikrųjų reikšmių suma	1869	5174	7043	
	Prisiminimas	20.492%	96.347%		
Bendrasis tikslumas:		76.218%			
Kappa:		0.216			

7 priedas

„Premium“ klubo duomenų pranašumo kreivės



8 priedas

„Premium“ klubo modelių sumaišymų matricos

Detekcijos atsitiktinių miškų sumaišymo matrica su ribine verte, kada lojalių klientų tikslumas sulyginamas su nelojalių klientų tikslumu:

		Tikroji reikšmė: Nelojalus	Tikroji reikšmė: Lojalus	Detekcijos suma	Preciziškumas
Detekcijos rezultatai	Nelojalus	1992	2680	4672	42.637%
	Lojalus	1151	4539	5690	79.772%
	Tikrųjų reikšmių suma	3143	7219	10362	
	Prisiminimas	63.379%	62.876%		

Bendrasis tikslumas:	63.028%
Kappa:	0.231

Atsitiktinių išlikimo miškų sumaišymo matrica su ribine verte EER taške:

		Tikroji reikšmė: Nelojalus	Tikroji reikšmė: Lojalus	Detekcijos suma	Preciziškumas
Detekcijos rezultatai	Nelojalus	1749	3186	4935	35.441%
	Lojalus	1394	4033	5427	74.314%
	Tikrųjų reikšmių suma	3143	7219	10362	
	Prisiminimas	55.647%	55.866%		

Bendrasis tikslumas:	55.8%
Kappa:	0.099

Atsitiktinių išlikimo miškų sumaišymo matrica su ribine verte, kuri optimizuojama remiantis tikėtinu maksimaliu pelnu:

		Tikroji reikšmė: Nelojalus	Tikroji reikšmė: Lojalus	Detekcijos suma	Preciziškumas
Detekcijos rezultatai	Nelojalus	2228	3854	6082	36.633%
	Lojalus	915	3365	4280	78.621%
	Tikrųjų reikšmių suma	3143	7219	10362	
	Prisiminimas	70.888%	46.613%		

Bendrasis tikslumas:	53.976%
Kappa:	0.138

9 priedas

Statistinio programavimo kalbos „R“ naudojamos bibliotekos

Biblioteka	Naudojimo paskirtis
c060	Reguliarizuoto Cox proporcingumo rizikos modelio sudarymui su epsgo funkcija

caret	Sumaišymų matricos sudarymui, kryžminiam patikrinimui, atraminių vektorių metodui, duomenų standartizavimui
doParallel	Skaičiavimo resursų optimizavimui
dplyr	Turimų duomenų manipuliacijai
e1071	Sumaišymų matricos sudarymui, kryžminiam patikrinimui, atraminių vektorių metodui, duomenų standartizavimui
EMP	Ribinės vertės suradimui tikėtino maksimalaus pelno taške
glmnet	Logistinei regresijai sudaryti
glmnetUtils	Elastinio tinklo reguliarizacijai išlikimo modeliuose
gridExtra	Atraminių vektorių metodo skirtingų kaštų verčių sudarymui
mlr	Metodų parametrų derinimui bei derinimo pasiruošimui
mlrMBO	Atsitiktinių išlikimo miškų parametrų derinimui
pacman	Kitų paketų patogesniam instaliavimui
plyr	Turimų duomenų manipuliacijai
precrec	Preciziškumo – jautrumo kreivės braižymui
PresenceAbsence	Ribinės vertės suradimui lygių paklaidų taške
randomForest	Atsitiktinių miškų metodo sudarymui
ranger	Detekcijos ir išlikimo atsitiktinių miškų sudarymui ir parametrų derinimui
ROC	ROC, DET kreivių braižymui
ROCR	Detekcijos gerumo matų rezultatų apskaičiavimui
survival	Išlikimo modelio objektui sukurti
survminer	Kaplan – Meier kreivių braižymui
tidyr	Turimų duomenų manipuliacijai
tuneRanger	Atsitiktinių miškų sudarymui ir parametrų derinimui