Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Pirkinių krepšelio analizė ir vartotojų pasitenkinimo prognozavimas

Baigiamasis magistro projektas

Marius Vadeika
Projekto autorius

Dr. Vilma Petrauskienė
Vadovė
Doc. dr. Aistė Dovalienė
Vadovė

Kaunas, 2018

Kauno technologijos universitetas
Matematikos ir gamtos mokslų fakultetas

Pirkinių krepšelio analizė ir vartotojų pasitenkinimo prognozavimas

Baigiamasis magistro projektas
Didžiųjų verslo duomenų analitika (621G12002)

Marius Vadeika
Projekto autorius

Dr. Vilma Petrauskienė
Vadovė
Doc. dr. Aistė Dovalienė
Vadovė

Dr. Lina Dindienė
Recenzentė
Doc. dr. Žaneta Piligrimienė
Recenzentė

Kaunas, 2018

Kauno technologijos universitetas
Matematikos ir gamtos mokslų fakultetas
Marius Vadeika

Pirkinių krepšelio analizė ir vartotojų pasitenkinimo prognozavimas

Akademinio sąžiningumo deklaracija

Patvirtinu, kad mano, Mariaus Vadeikos, baigiamasis projektas tema „Pirkinių krepšelio analizė ir vartotojų pasitenkinimo prognozavimas“ 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

Vadeika, Marius. Pirkinių krepšelio analizė ir vartotojų pasitenkinimo prognozavimas. Magistro baigiamasis projektas / vadovai dr. Vilma Petrauskienė ir doc. dr. Aistė Dovalienė; Kauno technologijos universitetas, Matematikos ir gamtos mokslų fakultetas.
Studijų kryptis ir sritis (studijų krypčių grupė): Fiziniai mokslai, Matematika (01P).
Reikšminiai žodžiai: susietumo taisyklės, nesubalansuotų klasių duomenų rinkinys, sprendimų medžiai, mašininio mokymosi algoritmai.
Kaunas, 2018. 63 p.

Santrauka

Pirkinių krepšelio analizė svarbi didelę produktų įvairovę turinčioms įmonėms prekių išdėstymo, rekomendavimo ir marketingo tikslais. Įmonėms, siekiančioms ilgalaikių santykių su klientais, jų lojalumo, labai svarbu gebėti suteikti išskirtinę patirtį vartotojui, kuris naudojasi jų paslaugomis. Darbo tikslas – išanalizuoti pirkinių krepšelius atrandant svarbius klientų įpročius bei spręsti sudėtingą klientų pasitenkinimo uždavinį sudarant kuo tikslesnius klientų prognozavimo modelius. Todėl šiame darbe buvo sprendžiami du atskiri uždaviniai panaudojant Kaggle varžybinius duomenis: susietumo taisyklių analizė (Instacart internetinės maisto prekių parduotuvės duomenys) ir klasifikavimas nesubalansuotų klasių atveju (banko Santander klientų duomenys).

Atliekant pirkinių krepšelio analizę buvo sugeneruotos susietumo taisyklės su populiariu apriori algoritmu nustatant *support* ir *confidence* slenksčius. Šiam duomenų rinkiniui buvo sudarytos susietumo taisyklės tiek prekių kategorijų, tiek produktų lygyje. Atrenkant įdomias susietumo taisykles buvo pastebėta, kad norint išvengti akivaizdžių taisyklių reikia naudoti žemesnes dažnumo reikšmes ir aukštas *lift* ir *conviction* matų reikšmes. Taisyklių atrinkimas užtrunka, nes nėra iš anksto aišku kokias *support* ir *confidence* reikšmes reikia parinkti įdomių taisyklių radimui. Atrinkus įdomesnes taisykles padaryta išvada, kad prekių kategorijų susietumo taisyklės gali būti pritaikytos prekių išdėstymo tikslais arba abstraktesniems pasiūlymams, o stiprios produktų susietumo taisyklės gali būti naudojamos konkrečios prekės pasiūlymui.

Tiriant banko Santander klientų pasitenkinimo duomenis su išskirčių aptikimo algoritmu buvo nustatyta, kad nepatenkinti vartotojai nėra išskirtiniai, dėl to juos sunkiau prognozuoti. Duomenų paruošimo žingsnyje banko duomenims buvo atlikta kintamųjų atranka, atsisakyta apie 31% kintamųjų nesumažinant vartotojų pasitenkinimo prognozavimo tikslumo. Po to ištirti imčių metodai ir apibendrinta, kad klasifikavimo tikslumą galima padidinti sumažinus klasių disbalansą duomenyse. Vartotojų pasitenkinimas prognozuotas apmokius tris populiarius mašininio mokymosi algoritmus, kurie remiasi sprendimų medžiais. Iš šių trijų modelių sudarytas daugumos balsų modelis, kuris teisingai suklasifikavo 42% nepatenkintų vartotojų.

Vadeika, Marius. Market Basket Analysis and Prediction of Customer Satisfaction / supervisors dr. Vilma Petrauskienė and dr. Aistė Dovalienė; Faculty of Mathematics and Natural Sciences, Kaunas University of Technology.
Study field and area (study field group): Physical Sciences, Mathematics.
Keywords: association rules, imbalanced classes data set, decision trees, machine learning algorithms.
Kaunas, 2018. 63 pages.

Summary

Market Basket Analysis is important for companies which sell a wide range of products for product placement, recommendation and marketing purposes. For companies, which aim for long-term relationships with their clients it is important to create an exclusive customer experience. The aim of this study is to find customer behavioral patterns by analyzing customer orders and to predict customer satisfaction by building as accurate as possible models. In this study, two tasks (which use Kaggle competition data) are solved: association rule analysis (Instacart data) and imbalanced classification task (Santander customer satisfaction data).

For Market Basket Analysis association rules were discovered with Apriori algorithm by setting support and confidence thresholds. Association rules were found for products and for product categories. It was concluded, that in order to find interesting rules, lower support values and higher lift and conviction values are required. Rule selection takes time, because it‘s not clear in advance what support and confidence thresholds to select. After selecting interesting rules it was concluded that association rules for product categories are more useful for product placement and abstract offers. On the other hand, strong association rules for products can be useful for suggesting specific products to clients.

While studying Santander bank customer satisfaction data with an anomaly detection algorithm, it was observed that unsatisfied customers are not very different from regular customers and therefore are hard to predict. Feature selection was done in the data preparation step and 31% of variables were removed without losing classification accuracy. Then, resampling methods were tested, and it was concluded that classification accuracy can be increased by reducing class imbalance. Customer satisfaction was predicted with three popular machine learning algorithms which use decision trees. From the three models a single majority vote model was built which correctly classified 42% of unsatisfied customers.

Santrumpos

PKA – pirkinių krepšelio analizė;

ST – susietumo taisyklės;

RS – rekomendavimo sistemos;

MAM – minimalios aprėpties medis;

MZŠK – maksimalus Z-įvertis tarp šešėlinių kintamųjų;

TP – Teisingas priėmimas (angl. *True Positive*);

TN – Teisingas atmetimas (angl. *True Negative*);

FP – Klaidingas priėmimas (angl. *False Positive*);

FN – Klaidingas atmetimas (angl. *False Negative*);

ROC – *Receiver Operating Characteristic* kreivė;

AUC – ploto po kreive tikslumo matas (angl. *Area Under Curve*);

RGF – *Regularized Greedy Forest* algoritmas;

XGBoost – *eXtreme Gradient Boosting* algoritmas.

Įvadas

Dažnai tokios kompanijos kaip mažmeninės prekybos parduotuvės susiduria su problema: kaip surasti prekes, kurios yra perkamos kartu. Bene vienintelis prieinamas informacijos šaltinis yra istoriniai transakciniai pardavimų duomenys. Pirkinių krepšelio analizė yra vienas iš populiariausių duomenų tyrybos metodų, analizuojant transakcijų duomenis. Šio metodo pagrindinė idėja – kad klientas išleistų daugiau pinigų, remiantis dviem principais: papildomas pardavimas (angl. *upsell*), kai perkami papildomi kiekiai to paties produkto ar jį papildančių ypatybių / prekių bei kryžminis pardavimas (angl. *cross-sell*), kai perkamos skirtingų kategorijų prekės. Todėl aktualu surasti tokius dažnus prekių derinius, kurie nėra akivaizdūs ir padeda atpažinti vartotojų įpročius. Tam tipiškai naudojamos susietumo taisyklės [1].

Augantis įmonių skaičius tam tikrose srityse (pvz. teikiančių telekomunikacijų paslaugas) išaugino konkurenciją bei klientų praradimo apimtis. Klientų praradimas dar gali būti apibrėžiamas kaip bet kokių transakcinių ryšių nutraukimas su kompanija. Vartotojai dažniausiai pasitraukia dėl nepasitenkinimo įmonės teikiamomis paslaugomis, kuris įprastai nėra vienkartinis. Kai kuriuos ryšius su klientu kompanija gali identifikuoti analizuodama savo kaupiamus duomenis. Pagal tai galima atpažinti dabartinę situaciją bei prognozuoti klientų praradimus. Suspėjus laiku identifikuoti aukštą pasitraukimo tikimybę turinčius klientus, kompanija gali įsiterpti ir juos išsaugoti [2].

Pirkinių krepšelio analizė svarbi didelę produktų įvairovę turinčioms įmonėms prekių išdėstymo, rekomendavimo ir marketingo tikslais. Įmonėms, siekiančioms ilgalaikių santykių su klientais, jų lojalumo, labai svarbu gebėti suteikti išskirtinę patirtį vartotojui, kuris naudojasi jų paslaugomis.

Pagrindinis darbo tikslas – išanalizuoti pirkinių krepšelius atrandant svarbius klientų įpročius bei spręsti sudėtingą klientų pasitenkinimo uždavinį sudarant kuo tikslesnius klientų prognozavimo modelius.

Todėl šiame darbe bus sprendžiami du atskiri uždaviniai panaudojant Kaggle varžybinius duomenis: susietumo taisyklių analizė (Instacart internetinės maisto prekių parduotuvės duomenys) ir klasifikavimas nesubalansuotų klasių atveju (banko Santander klientų duomenys).

1. Literatūros apžvalga

1.1. Pirkinių krepšelio optimizavimas ir vartotojų pasitenkinimo užtikrinimas kaip prioritetiniai ilgalaikių santykių su vartotojais plėtros sprendimai

Prieš kelis dešimtmečius verslo modeliai veikė pagal gamintojų sąlygas, nes pasiūla buvo mažesnė nei paklausa. Gamintojų visas dėmesys buvo sutelktas į gamybos apimčių didinimą, o ne į kokybę. Be to, produktai nepasižymėjo įvairove. Dabartiniais laikais gamintojai ir pardavėjai jau veikia visai pasikeitusioje aplinkoje. Dėl to, vartotojų identifikavimas, pritraukimas, išsaugojimas bei santykių vystymas tapo itin aktualūs kompanijoms. Šiuolaikiniuose verslo modeliuose vartotojų pasitenkinimas yra labai svarbus, nes pasiūla yra didesnė nei paklausa. Dabar klientas sprendžia kokius produktus gaminti, o gamintojas siekia užtikrinti kokybę. Prekių gyvenimo ciklas sutrumpėjo, o jų įvairovė išaugo. Todėl šiais laikais kompanijos yra priverstos siūlyti aukštesnės kokybės prekes ir paslaugas [3].

Vartotojai turi daugiau pasirinkimo nei bet kada anksčiau, o šiais laikais įmonių konkurentus pasiekti galima keliais pelės paspaudimais. Be klientų produktai nėra parduodami, o pajamos nėra uždirbamos. Neįtvirtinus klientų lojalumo net ir pelningas klientas atneš tik trumpalaikę naudą. Dėl šių priežasčių bankai automatizavo savo marketingo veiksmus ir pradėjo skaičiuoti individualaus kliento vertę, komunikacijų kompanijos stengiasi sumažinti vartotojų praradimus, o mažmeninės ir elektroninės prekybos parduotuvės stiprina ryšius su klientais lojalumo programomis.

Dar visai neseniai, verslai buvo labiau orientuoti į „ką parduoti“ vietoj „kas pirks“. Kitaip tariant, kompanijos stengėsi parduoti kuo daugiau prekių ir paslaugų nekreipdamos dėmesio į kas jas perka. Sėkmingi verslai supranta, kad individuali kliento patirtis užtikrina didelį vartotojų lojalumą. Tos kompanijos, kurios jau veikia su į klientus orientuotu požiūriu, žino, kad santykių su klientais valdymo dėka klientas jaučiasi asmeniškai susijęs bei turi malonesnę patirtį. Toks klientas noriai dalinasi savo teigiama kliento patirtimi su kitais [4].

Vienas iš pagrindinių santykių su klientais valdymo tikslų yra didinti bendrą klientų vertę. Klientų pasitenkinimas yra svarbus šiam tikslui pasiekti. Tačiau panašu, kad kompanijoms sunku užtikrinti klientų pasitenkinimą. Iš kliento pusės, 85% vartotojų teigia, kad kompanijos turėtų labiau pasistengti dėl klientų išsaugojimo. Iš kompanijos pusės, nors daugumos kompanijų aukščiausi vadovai teigia, kad klientų išsaugojimas turi aukštą prioritetą jų kompanijose, tačiau 49% pripažįsta esantys nepatenkinti esama situacija. Dėl to, kompanijos deda nemažai pastangų į asmeninius pasiūlymus bei klientų pasitenkinimo prognozavimą [5].

1.2. Pirkinių krepšelio analizė

Prieš kelis dešimtmečius prekybos tinklai ir parduotuvės pardavinėdavo savo prekes beveik nesiremdami transakcijų duomenimis kaip žinių šaltiniu. Per pastaruosius du dešimtmečius kompanijos pradėjo naudoti šiuos duomenis naudingai informacijai gauti. 1990–2000 metų laikotarpiu riboti kompiuteriniai resursai neleido efektyviai išgauti informacijos iš kasdien įvykstančių milijonų transakcijų. Tokiais atvejais buvo naudojami tik paprasti modeliai bei sumažinti duomenų rinkiniai. Šiuo laikotarpiu kompanijos pradėjo kaupti savo duomenų bazėse vis daugiau transakcinių, klientų, pardavimų ir kitokių duomenų. Jų nagrinėjimui buvo pasiūlytas Apriori algoritmas, kuris ir šiomis dienomis vis dar naudojamas kaip vienas pagrindinių įrankių pirkinių krepšelio analizei (PKA). Šiandien, lyginant su prieš tai minėtu laikotarpiu, skaičiavimo sistemos gerokai patobulėjo. Tiek techninė, tiek ir programinė įranga vystėsi įvairiose srityse: santykių su klientais valdymas, verslo valdymas, duomenų saugyklos bei kitokios sistemos. Saugomų duomenų vieta vis plėtėsi, o tuo tarpu buvo kuriami sudėtingi modeliai ir algoritmai duomenų pažinimui didelėse duomenų bazėse [1].

Pagrindinis PKA tikslas – nustatyti dažnų elementų rinkinius, o tam tipiškai naudojamos susietumo taisyklės (ST). ST analizėje yra taikoma daug įvairių taisyklių įdomumo matų. Tačiau M. Hahsleris ir K. Hornikas teigia, kad šie matai neatsižvelgia į tikimybines savybes. Todėl jie atlieka tyrimą apie patikimumo ir svarbos matus (angl. *confidence and lift*). Rezultatai atskleidė, kad patikimumas yra sistemingai veikiamas kairėje taisyklėje esančių elementų dažnumo, o svarbos matas prastai išfiltruoja atsitiktinį triukšmą transakciniuose duomenyse. Naudodami tikimybinę sistemą, autoriai sudarė du naujus susietumo matus: hiper-patikimumą ir hiper-svarbą (angl. *hyper-confidence and hyper-lift).* Naujieji susietumo matai pasižymi ženkliai geresniu filtravimu nei svarbos matas, kai šis nesusitvarko su klaidingomis taisyklėmis [6].

Vienas iš praktinių PKA panaudojimo būdų yra rekomendavimo sistemos sukūrimas. Tokia sistema dažniausiai vadovaujasi trimis žingsniais: duomenų apdorojimas, analizė ir rezultatų interpretavimas. Dėl to rekomendavimo sistemose yra taikomi įvairūs duomenų tyrybos metodai: panašumo matai, klasterizavimas, dimensijų mažinimas, klasifikavimas, susietumo taisyklės bei kt. ST yra intuityvus būdas rekomendavimui, kai duomenys yra susiję su transakcijomis. Nors ST yra glaudžiai susijusios su rekomendavimo sistemomis, tačiau jos nėra itin populiarios [7].

Vienas įdomiausių pastaruoju metu atliktų tyrimų PKA srityje buvo susietumo taisyklių išplėtimas su minimalios aprėpties medžiais (angl. *minimum spanning trees*). Šio metodo ypatybė – šakų atstumų minimizavimas. Minimalios aprėpties medžių (MAM) pavyzdys galėtų būti laidų kompanijos uždavinys: nutiesti laidus į visas gyvenvietes, minimizuojant sunaudoto laido ilgį. Iš analitiko perspektyvos, MAM leidžia atsakyti į tokį klausimą: kokiu būdu galima pasiekti produktą

B, kai yra žinoma, kad buvo pirktas produktas A? Pagrindinė šio metodo naujovė yra paprastas grafinis produktų grupių atvaizdavimas iš didelių duomenų kiekių. Šis metodas įgalina:

1) stipriausių priklausomybių tarp produktų atradimą (tose pačiose ar skirtingose kategorijose). Tai yra toks pats rezultatas, kaip ir atrinktos aukštos svarbos susietumo taisyklės;
2) pagrindinių produktų identifikavimas, kurie siejasi su kitomis produktų grupėmis, bei įtakos zonų atpažinimas minimalios aprėpties medžiuose. Tai leidžia naudoti šią metodiką „genėjant“ MAM tam, kad būtų atrenkamos ST;
3) naudojant produktų tinklą, jo sudėtingumą galima sumažinti sudarant minimalios aprėpties medį. Gautas MAM pasižymi analizuojamų produktų hierarchine struktūra;
4) autorių pateikta metodika leidžia tirti ryšį tarp prekių grupių, kurias galima įsigyti parduotuvėje ir tarp pagrindinių prekių grupių, kurias suformuoja pirkėjų įpročiai [8].

Nors susietumo taisyklės nėra plačiai naudojamos praktikoje, literatūroje galima atrasti įdomių pritaikymo atvejų. X. Amatriainas ir kiti bendraautoriai ištyrė, kad kai kurie susietumo taisyklių generavimo algoritmai yra tikslesni nei k–artimiausių kaimynų metodas (kNN). Autoriai pateikia pavyzdžių, kur susietumo taisyklės buvo sėkmingai pritaikytos praktikoje. Vienas jų yra tinklalapio personalizavimo sistema. Ši sistema identifikuoja vartotojų įpročius naudodama ST pagal pakartotines tinklalapio peržiūras. Toks metodas pranoksta kNN rekomendavimo sistemą tiek tikslumu, tiek ir aprėptimi.

Kitame pateikiamame pavyzdyje atliekama PKA: naudojamas Apriori algoritmas norint sudaryti panašumo matą tarp produktų. Tuo tarpu, dar vienu pavyzdžiu autoriai iliustruoja susietumo taisyklių pritaikymą, kai tikslas yra atpažinti kartu įvykstančias kritikas. Pavyzdžiui, įžvelgti vartotojo teikiamą pirmenybę tam tikrai rekomenduojamo produkto ypatybei.

Kitame efektyvaus pritaikymo pavyzdyje naudojamas naujas susietumo taisyklių algoritmas, kuriame parenkamas minimalus dažnumas, tam kad būtų atrenkamos tik svarbesnės taisyklės. Šios savo ruožtu yra atrenkamos tiek vartotojams, tiek ir produktams. Įvertintas tikslumas pranoksta koreliacijomis grįstas rekomendacijas ir prilygsta sudėtingesniam metodui – faktorizavimo ir dirbtinių neuroninių tinklų deriniui (angl. *ANN*) [7].

Taip pat literatūroje galima rasti pavyzdžių, kur PKA labiau taikoma ne metodų efektyvumui tirti, o vartotojų įpročiams atskleisti. M. M. Mostafas atliko PKA apie Kuveitą, kuris dar nėra taip gerai pažįstamas, lyginant su tradicinėmis Vakarų šalimis. Atlikus analizę, buvo pastebėta, kad PKA neturi konkrečios krypties, nes visi produktai krepšelyje yra laikomi vienu metu, tačiau atskleidžia, kurie produktai yra perkami kartu viename krepšelyje. Svarus argumentas, kuo PKA yra naudinga – rezultatai yra aiškiai interpretuojami ir juos lengva iškomunikuoti vadovams. Autorius teigia, kad naudojant PKA, parduotuvės gali nuspręsti kaip optimaliai pasiūlyti

specialius kombinuotų prekių pasiūlymus siekdamos didžiausio pelno. Kitas svarbus aspektas yra rezultatų panaudojimas efektyvesnei kainodarai ir reklamos strategijai.

PKA atskleidžia svarbius paslėptus pirkėjų įpročius perkant produktus. Analizės rezultatai atspindi tikrąją vartotojų elgseną, o ne deklaruojamą (pavyzdžiui, remiantis apklausomis). Taigi, Kuveito parduotuvės turėtų nuolatos rinkti pardavimų duomenis, kad galėtų greitai reaguoti į rinkos pokyčius. Autorius atlikdamas Kuveito PKA atrado, kad skirtingai nei Vakarų šalyse, tipinis Kuveito vartotojas neperka alkoholio arba kiaulienos. Tai yra susiję su religiniais įsitikinimais ir kultūriniais skirtumais. Tvirti įsitikinimai taip pat apriboja ir šių produktų reklamą [9].

PKA taip pat galima naudoti ir internetinėje parduotuvėje. X. Amatriaino ir kitų bendraautorių pavyzdyje susietumo taisyklės pritaikomos internetinės parduotuvės rekomendavimo sistemai kartu su sprendimų medžiais (angl. *decision trees)*. ST yra panaudojamos produktų susiejimui tarpusavyje. Tuomet rekomendacija yra parenkama pagal ST ir vartotojo poreikių sankirtą. Tai yra pasiekiama stebint skirtingus transakcinius duomenis, tokius kaip pirkimai, prekių įdėjimas į krepšelį bei prekių peržiūros. Iš šių duomenų sudaromos ST, kurioms yra suteikiami atitinkami svoriai, pvz., pirkimas yra laikomas svarbesniu nei paspaudimas [7].

M. A. Valle‘as ir kiti autoriai parodė, kad maisto produktų susietumo tinklas palengvina ST paiešką. Tai yra itin svarbu, nes ST yra įprastinis PKA įrankis, tačiau pasižymintis tokiais trūkumais kaip gausus sugeneruotų ST skaičius. Dėl to yra ne tik itin sudėtinga atrinkti pačias svarbiausias ST, bet tai užima ir daug laiko.

Autoriai taip pat atliko kelias praktines įžvalgas pagal savo sudarytus MAM. Viena jų – tos pačios kategorijos produktai tipiškai būna išsidėstę toje pačioje medžio šakoje. Tai reiškia, kad tam tikros kategorijos prekės pirkimas būna papildomas tos pačios kategorijos prekėmis. Tai galima paversti praktine nauda: atrandant stiprius ryšius tarp tos pačios kategorijos prekių bei panaudojant tai tiksliniams pasiūlymams ar reklamai. Antra, yra įmanoma atpažinti produktus, kurie sieja skirtingas produktų šakas. Atsižvelgiant į MAM hierarchinę struktūrą, produktai, kurie priklauso daugiau nei vienai šakai, gali būti aktualūs. Trečia, MAM paprastumas leidžia vertinti reikšmingesnę produktų sąryšių dalį, atsisakant tokių taisyklių, kurios turi didelę koreliaciją, bet tik atsitiktinai [8].

Atlikdamas Kuveito PKA, M. M. Mostafas analizavo transakcinius duomenis, norėdamas aptikti vienu metu įsigytus pirkinius. Tačiau, jis pabrėžia, kad PKA vis dar turi išspręsti nemažai iššūkių. Naudojant susietumo taisykles, nepavyks nustatyti ryšių, kai prekės perkamos kartu, tačiau ne tuo pačiu metu. Ateities tyrimai galėtų būti atliekami PKA pritaikomumui, kai kinta laikas, panaudojant kartu pirktų produktų tinklo analizę. Kitas svarbus susietumo taisyklių apribojimas yra didelis sugeneruotų taisyklių skaičius. Autorių analizėje buvo sugeneruota 1500 taisyklių, jas visas išanalizuoti yra sudėtinga. Tai išspręsti siūlo pasitelkiant „apvilkimo“ analizę (angl. *envelopment*

analysis), kuri atrinktas taisykles įvertina pagal keletą kriterijų, tam, kad jas reitinguotų efektyvumo ar pelno maksimizavimo atžvilgiu.

Kadangi PKA rezultatai priklauso nuo nustatytų minimalių dažnumo ir patikimumo reikšmių, autorius mano, kad jautrumo analizė turėtų būti taikoma, norint įvertinti rezultatų stabilumą. Jei bus pasirinktos žemos susietumo matų reikšmės, tuomet bus sugeneruotos klaidingos taisyklės, o jei aukštos – kai kurios prasmingos taisyklės liks neaptiktos [9].

Kai kurias susietumo taisyklių ydas gali padėti spręsti minimalios aprėpties medžiai. Jie gali būti laikomi kaip supaprastintas susietumo taisyklių perteikimas. Šis gali įveikti duomenų pertekliaus problemas, reikšmingai sumažinant ST paieškos erdvę. MAM taip pat veiksmingiau kontroliuoja klaidingai nustatytus sąryšius bei triukšmą. Vartotojų įpročių vaizdavimas minimalios aprėpties medžiais yra lengvai interpretuojamas bei suteikia galimybę atpažinti stiprius sąryšius tarp produktų. Dėl to šis metodas gali pasiteisinti kaip patrauklus įrankis rinkodaros vadybininkams. Apibendrinant, MAM koncentruojasi į ribotą kiekį prekių kategorijų, kurias vadybininkas gali panaudoti rinkodaros tikslais juos susiejant. Sudarant tų produktų poras, kurie yra vienas šalia kito MAM įtakos zonoje, yra galimybė išauginti jų pardavimus. Teikti jungtinius pasiūlymus galima produktams, esantiems šalia svarbaus medžio šakų susikirtimo taško.

Nepaisant MAM metodikos privalumų, beveik neegzistuoja literatūros, kurioje tai būtų pritaikoma verslo srityje. Kitose srityse MAM buvo pritaikyti apdorojant smegenų funkcijų vaizdus, klasterizuojant genų išraiškas, atliekant valiutų ir akcijų mainus [8].

1.3. Vartotojų pasitenkinimo prognozavimas

Santykių su klientais valdymas (angl. *Customer Relationship Management*) yra skirtas kurti, valdyti ir stiprinti santykius su klientais. Vienas pagrindinių santykių su klientais valdymo uždavinių yra išvengti klientų praradimo. Tai yra svarbu žinant, kad naujo vartotojo pritraukimas yra brangesnis nei esamo išsaugojimas (tyrimai rodo, jog kai kuriais atvejais net iki 20 kartų). Dinamiškos rinkos aplinkoje klientų praradimas gali pasireikšti žemu klientų pasitenkinimo lygiu, agresyviomis konkurentų strategijomis, naujų produktų pasiūla ir pan. Vartotojų praradimo modeliai stengiasi identifikuoti ankstyvus kliento praradimo signalus. Per pastarąjį dešimtmetį augo susidomėjimas bei buvo atliekami tokie tyrimai telekomunikacijų, bankininkystės, draudimo, nekilnojamo turto srityse ir net kompiuteriniuose žaidimuose [10].

Vartotojų pasitenkinimą taip pat tenka prognozuoti dirbant su didelėmis duomenų apimtimis. Analizuojant vienos didžiausių Kinijos telekomunikacijų kompanijos duomenis, buvo nustatyta, kad prognozių tikslumas ženkliai išauga remiantis trimis didžiųjų duomenų principais: apimtimi, greičiu ir įvairove (angl. *volume, velocity and variety*). Y. Huangas bei kiti bendraautoriai telekomunikacijų kompanijai padėjo vystyti klientų pasitraukimo sistemą, kuri naudoja didžiulius

apmokymo duomenų kiekius, įtraukia didelę požymių įvairovę bei suvaldo gausius naujų duomenų srautus.

Telekomunikacijų srityje dažniau prarandami klientai, kurie naudojasi išankstinio apmokėjimo planu. Laikoma, kad šių klientų pasitraukimo prognozavimas yra sudėtingesnis, todėl būtent jiems ir buvo sukurta klientų pasitraukimo sistema. Galutiniame rezultate iš milijonų aktyvių klientų ši sistema geba atrinkti aukštos pasitraukimo tikimybės išankstinio apmokėjimo klientų sąrašą. Šios sistemos rezultatas panaudojamas kartu su automatine vartotojų išsaugojimo kampanija ir tai stipriai prisideda prie klientų išlaikymo bei didesnės verslo vertės [11].

Daugumoje industrijų, pavyzdžiui telekomunikacijose, klientų praradimas yra retas atvejis. Kitaip tariant, prarandamų klientų skaičius yra gerokai mažesnis nei tų, kurie ir toliau naudojasi paslaugomis. Iš mašininio mokymosi uždavinio perspektyvos, vartotojų pasitenkinimas gali būti sprendžiamas kaip dichotominio kintamojo klasifikavimas. Šiame uždavinyje atsako kintamasis pasižymi nesubalansuotų klasių skirstiniu, kuriame mažumą sudaro paslaugomis nepatenkinti prarasti klientai.

Klasių disbalanso problema apsunkina standartinių klasifikavimo algoritmų taikymą. Nesubalansuotų klasių problema yra ta, kad klasifikatorius yra linkęs mažumos klasės stebinius neteisingai priskirti daugumos klasei. Kraštutiniais atvejais, klasifikatorius gali priskirti visus stebinius daugumos klasei. Taip jis pasiekia aukštą bendrą klasifikavimo tikslumą, tačiau nepriimtinai žemą tikslumą dominančios mažumos klasės atžvilgiu. Pavyzdžiui, jei modelis apmokytas duomenų rinkiniui, kuriame 1% sudaro mažumos klasę, tuomet 99% tikslumas gali būti pasiektas paprasčiausiai viską priskiriant daugumos klasei [12].

Nesubalansuotų duomenų prognozavimas taip pat aktualus tokiame uždavinyje kaip rizikos vertinimas. A. N. Haldankaras ir K. Bhowmickas tyrė banko klientų mokumą. Šiuo darbu autoriai atskleidė, kad prieš sudarant modelius yra labai pravartu atlikti požymių (kintamųjų) atranką. Tam jie naudojo Boruta algoritmą, kuris iteratyviai pašalina tokius požymius, kuriuos statistinis testas laiko mažiau reikšmingais nei atsitiktinumas. Šiuo metodu galima įvertinti kintamųjų svarbą bei atmesti neturinčius pridėtinės vertės požymius. Atlikus kintamųjų atranką buvo sudaromos unikalios *ensemble* metodų (*bagging* ir *boosting*) kombinacijos kartu naudojant slenksčio metodą. Pasirinkimas naudoti sluoksniuotas klasių imtis su kryžminiu patikrinimu padėjo užtikrinti gautų rezultatų teisingumą. Atlikus šiuos žingsnius buvo apibendrinta, jog pasirinktų metodų naudojimas padeda sumažinti klaidingą klientų klasifikavimą. Taip pat identifikuota, kad kliento mokumui didžiausią įtaką daro tokie veiksniai kaip kredito istorija, sąskaitos būklė, santaupų sąskaita, įsiskolinimai ir kiti veiksniai [13].

Norint spręsti nesubalansuotų imčių mokymosi problemą, per pastarąjį dešimtmetį buvo pasiūlyta įvairių metodų. Pasiūlyti sprendimai gali būti priskiriami dviems kategorijoms: duomenų

arba algoritmų lygio sprendimai. Algoritmų lygio sprendimai siekia pasiūlyti naujus arba adaptuoti jau esančius algoritmus nesubalansuotoms lygtims. Duomenų lygio sprendimai mėgina sumažinti klasių disbalanso efektą panaudojant imčių atrankos metodus duomenų paruošimui.

Algoritmų lygio sprendimai naudoja konkretų klasifikavimo algoritmą, kuris įprastai pasižymi naudojamo konteksto jautrumu. Norint sukurti algoritmą, reikia labai gerai išmanyti apie besimokančius algoritmus bei apie sritį, kurioje tai taikoma. Kita vertus, duomenų lygio sprendimai veikia kaip duomenų paruošimo žingsnis. Priimama prielaida, kad šių metodų taikymas yra nepriklausomas nuo vėliau naudojamo klasifikatoriaus. Todėl praktikoje duomenų lygio metodus taikyti yra paprasčiau. Nors literatūroje siūlomi duomenų lygio sprendimai pagerina klasifikavimo tikslumą, tačiau mokslo bendruomenei nepavyksta priimti bendros išvados, kuris metodas yra geriausias [12].

Mokslinėje literatūroje galima atrasti įvairių algoritmų klientų pasitenkinimo prognozavimui, tačiau T. Vafeiadis su kitais bendraautoriais išskiria kelis populiariausius:

- dirbtiniai neuroniniai tinklai (angl. *artificial neural networks*) – populiarus sprendimas sudėtingoms problemoms, klientų pasitenkinimo prognozavime galintis pranokti sprendimų medžius;
- atraminių vektorių metodas (angl. *support vector machines*) – suranda tarp dviejų klasių tokią sprendimo ribą, kad atstumas tarp klasių būtų didžiausias. Šis metodas vartotojų pasitenkinimo prognozavime yra vienas iš tiksliausių;
- sprendimų medžiai (angl. *decision trees*) – medžio struktūros algoritmas, kuris atvaizduoja sprendimus. Sprendimų medžiai nėra tinkamiausi atpažinti sudėtingiems netiesiniams sąryšiams, tačiau šiame uždavinyje jie pasiekia gerus rezultatus;
- logistinė regresija – paprastas tikimybinis klasifikavimo modelis. Įprastai naudojamas, kai duomenys yra tinkamai paruošti. Tokiu atveju logistinė regresija gali pasižymėti gerais rezultatais;
- naivus Bajeso algoritmas (angl. naive Bayes) – paprastas tikimybinis klasifikatorius, naudojantis Bajeso teoremą su nepriklausomumo prielaidomis. Literatūroje teigiama, kad kai kuriais atvejais gali prilygti sudėtingesniems algoritmams [10].

Aukšto efektyvumo klasifikavimo modelis nesubalansuotiems duomenims gali būti sudarytas naudojant populiarėjančius dirbtinius neuroninius tinklus (angl. *artificial neural networks*). Tai M. Fridrichas atskleidė tirdamas elektroninės prekybos duomenis. Pirmiausiai duomenys buvo padalinti į dvi dalis siekiant tinkamai įvertinti modelį. Pradinis duomenų apdorojimas atliktas standartizuojant duomenis su standartiniu z-įverčiu bei pritaikant pagrindinių komponenčių metodą (angl. *PCA*). Tuomet panaudotas dvisluoksnis dirbtinis neuroninis tinklas

kaip bazinis klasifikatorius, kurio parametrai optimizuoti naudojant genetinį algoritmą. Klasifikavimo modelis vertinamas įprastiniais tikslumo (angl. *accuracy*) bei ploto po kreive (angl. *AUC*) matais. Šiame tyrime buvo nustatyta, kad parametrų optimizavimas dirbtiniams neuroniniams tinklams padidina jų klasifikavimo tikslumą, kuris yra reikalingas norint teisingai nuspėti ir užtikrinti vartotojų pasitenkinimą [14].

Vienas iš originalesnių metodų vartotojų pasitenkinimui prognozuoti yra jonvabalio algoritmas (angl. *firefly algorithm*). Kaip ir dirbtiniai neuroniniai tinklai, šio algoritmo idėja yra įkvėpta gamtos. Jonvabalio algoritmas yra metaeuristinis, jis remiasi šių vabalų elgsena. Pagrindinė metodo idėja – jonvabaliai pritraukia kitų jonvabalių dėmesį spinduliuodami šviesą. Šviesos stiprumas smarkiai lemia jonvabalio patrauklumą. Algoritmas veikia šiomis prielaidomis:

- visi jonvabaliai yra vienalyčiai – kiekvienas jonvabalis gali pritraukti kitą;
- patrauklumas yra proporcingas jonvabalio spinduliuojamai šviesai;
- tarp dviejų jonvabalių ryškesnis pritrauks kitą;
- jei nei vienas jonvabalis nėra ryškesnis, tuomet vabalai juda atsitiktinai;
- optimizavimo problemai spręsti jonvabalio ryškumas yra susietas su tikslo funkcija.

Jonvabalio algoritmas gali efektyviai atrasti optimalius sprendinius. Tačiau kadangi kiekvieną jonvabalį reikia palyginti su visais kitais jonvabaliais, todėl vabalų skaičiui didėjant paieškos aibė bei skaičiavimai sparčiai išauga. Autoriai taip pat pasiūlė hibridinį jonvabalių algoritmą (angl. *hybrid firefly algorithm*) skaičiavimams pagreitinti. Šis atnaujintas algoritmas taip pat įtraukia jonvabalių poravimosi imitavimą. Eksperimentai parodė, kad hibridinis jonvabalių algoritmas yra ne tik ženkliai geriau optimizuotas nei pirminis jonvabalio algoritmas, tačiau prognozuojant vartotojų pasitenkinimą nenusileidžia ir kitiems, gamtos įkvėptiems algoritmams [2].

Vartotojų pasitenkinimui prognozuoti telekomunikacijų srityje T. Vafeiadis ir kt. bendraautoriai atliko išsamų populiariausių mašininio mokymosi algoritmų palyginimą. Pirmiausiai, visi modeliai buvo sudaryti ir įvertinti atliekant kryžminį patikrinimą (angl. *cross-validation*). Antrame žingsnyje buvo naudojamas modelių gerinimas (angl. *boosting*) jų tikslumui padidinti. Siekiant kiekvienam metodui nustatyti efektyviausias parametrų kombinacijas, buvo atliktos simuliacijos naudojant Monte Karlo metodą. Analizės rezultatai parodė, kad gerintos (angl. *boosted*) modelių versijos pranoksta pradinius modelius. Savo tyrimui autoriai pasirinko populiariausius mašininio mokymosi algoritmus, kurie pastaruoju metu yra naudojami telekomunikacijų kompanijų klientų pasitenkinimo prognozavimui. Lyginant pradinius modelius, gauta, kad dirbtiniai neuroniniai tinklai ir sprendimų medžiai klasifikavo klientus tiksliausiai (nuo jų nedaug atsiliko atraminių vektorių metodas). Tačiau atlikus modelių gerinimą, atraminių vektorių

metodas tapo tiksliausiu. Nesudėtingi modeliai (logistinė regresija ir Bajeso naivus algoritmas) pasirodė prognozavime neblogai, tačiau iš visų prasčiausiai [10].

Labiau kraštutiniais atvejais nesubalansuotų klasių uždavinys gali būti sprendžiamas kaip anomalijų aptikimas (angl. *anomaly detection*), pavyzdžiui identifikuojant kredito kortelių sukčiavimų atvejus. F. T. Liu kartu su kitais bendraautoriais pasiūlė naują metodą anomalijų aptikimui, kuris vadinamas izoliavimo mišku (angl. *isolation forest*). Dažnai tokie metodai pirmiausiai siekia atpažinti įprastus atvejus (daugumos klasę), o tada identifikuoti jai nebūdingus atvejus (mažumos klasę), kurie priskiriami anomalijoms. Vietoje tradicinio būdo, izoliavimo miškas leidžia iš karto išskirti netipinius atvejus. Šis metodas pasinaudoja pagrindiniais anomalijų požymiais – kelios ir išsiskiriančios. Pasinaudojant šiuo principu, izoliuojantis medis anomalijas išskiria arčiau medžio šaknies nei tai daro įprastų stebinių atveju. Ši unikali ypatybė izoliuojančiam miškui leidžia kurti dalinius modelius, kurie naudodami mažą apmokymo duomenų dalį pasižymi dideliu efektyvumu.

Izoliavimo mišką galima naudoti su didžiaisiais duomenimis: metodo skaičiavimo laiko augimas yra tiesinis, o atminties reikalavimai yra nedideli. Autorių empirinis metodo įvertinimas klasifikavimo uždavinyje atskleidė, kad izoliavimo miškas pranoksta tokius metodus kaip *ORCA, LOF* bei atsitiktinį mišką (angl. *Random Forest*) vertinant klasifikavimo tikslumą *AUC* bei skaičiavimo laiką. Labai didelių matavimų atvejais izoliavimo miškas pasižymi greitu ir tiksliu anomalijų atpažinimu. Tokiais atvejais atstumais besiremiantys metodai pasižymi mažesniu tikslumu arba reikalauja ženkliai didesnių laiko sąnaudų. Apibendrinant, izoliavimo miškas yra tikslus bei efektyvus prognozavimo metodas, skirtas spręsti nesubalansuotų klasių problemą, o metodo praktiškumą autoriai ištyrė ir pademonstravo su 12 skirtingų duomenų rinkinių [15].

B. Zhu ir kt. autoriai analizavo nesubalansuotų klasių problemą mažiau populiariu būdu. Dėmesys buvo kreipiamas ne į algoritmą, o į imties atrankos metodus, kurie yra taikomi nesubalansuotoms imtims. Palyginus metodus, buvo padarytos trys svarbios išvados. Pirma, klasių santykis 1:3 duomenyse yra priimtinas klientų praradimui prognozuoti. Buvo nustatyta, kad nėra statistiškai reikšmingo skirtumo tarp šio santykio ir subalansuotų klasių. Taigi, pakanka duomenis transformuoti taip, jog klasių disbalansas būtų mažesnis. Antra, imčių atrankos metodai yra pakankamai jautrūs skirtingiems jų vertinimo matams (angl. *metrics*). Taip pat pastebėta, jog vertinimo matams daroma įtaka priklauso ir nuo pasirinkto klasifikatoriaus. Kita vertus, atrankos metodų naudojimas su logistine regresija neturi jokios įtakos vertinimo matams. Trečia, naudoti atrankos metodai nepadėjo reikšmingai pagerinti vertinimo matų tokiems modeliams kaip atsitiktinis miškas (angl. *Random Forest*) bei atraminių vektorių metodas (angl. *Support Vector Machine*) [12].

2. Medžiagos ir tyrimų metodai

2.1. Susietumo taisyklės

Susietumo taisyklių tyryba yra naudinga apžvalginei analizei bei ryšiams tarp kintamųjų atpažinti. Šie gali būti produktai internetinėje parduotuvėje, ligos simptomai, raktažodžiai, demografinės charakteristikos ir kt. Tam kad analitikas atpažintų įdomias taisykles, yra naudojami įdomumo matai, kurių literatūroje galima rasti bent kelias dešimtis.

Dažnumas (angl. *support*) reiškia kaip dažnai elementų derinys pasitaiko duomenyse. Tai yra pagrindinis susietumo taisyklių matas, iš kurio yra sudaromi kiti, sudėtingesni matai. Dažnumas *A*, pasitaikantis transakcijose *T*, yra apibrėžiamas kaip proporcija transakcijų *t* duomenyse, kuriuose yra *A*. Jeigu *A* = {pienas, duona} pasitaiko 1 iš 20 transakcijų, tokiu atveju *A* dažnumas yra: supp(A) = 1/20 = 0,05.

$$supp(A) = \frac{|\{t \in T; A \subseteq t\}|}{|T|} \quad (1)$$

Patikimumas (angl. *confidence*) parodo kaip dažnai *B* pasitaiko transakcijose, kuriose yra *A*. Tai yra tas pats kaip sąlyginė tikimybė $P(B|A)$ ir kinta intervale nuo 0 iki 1. Pavyzdžiui taisyklė {pienas, duona} ⇒ {makaronai} turės patikimumą $conf(A \Rightarrow B) = \frac{0,03}{0,05} = 0,6$, vadinasi *B* galima rasti 60% atvejų, kai transakcijoje yra *A*.

$$conf(A \Rightarrow B) = \frac{supp(A \cup B)}{supp(A)} \quad (2)$$

Taisyklės su aukštu patikimumu gali įvykti atsitiktinai ir tai nėra pakankamas matas taisyklės įdomumui nusakyti. Tai priklauso nuo to, ar elementų rinkiniai yra statistiškai nepriklausomi. Taisyklės aktualumui yra dažnai naudojamas svarbos matas – (angl. *lift*). Svarba matuoja, kaip toli nuo nepriklausomumo yra *A* ir *B* bei kinta intervale $[0; \infty]$. Artimos vienetui reikšmės nusako, kad *A* ir *B* yra nepriklausomi ir taisyklė yra neįdomi. Reikšmės tolimos nuo vieneto parodo, kad *A* buvimas transakcijoje lemia ir *B* buvimą. Tęsiant prieš tai pateiktą pavyzdį, $lift(A \Rightarrow B) = \frac{0,6}{0,1} = 6$, vadinasi tokia taisyklė yra pakankamai įdomi.

$$lift(A \Rightarrow B) = \frac{conf(A \Rightarrow B)}{supp(B)} \quad (3)$$

Įsitikinimas (angl. *conviction*) kartais naudojamas kaip papildomas matas norint įveikti dažnumo ir svarbos matų silpnybes. Skirtingai nei svarba, įsitikinimas yra jautrus taisyklės krypčiai. Taigi,

$$conv(A \Rightarrow B) \neq conv(B \Rightarrow A) \quad (4)$$

Pavyzdžiui, {pienas, duona} ⇒ {makaronai} bus lygus $conv(A \Rightarrow B) = \frac{1-0,1}{1-0,6} = \frac{0,9}{0,4} = 2,25$. Įsitikinimo matas kinta intervale $[0,5; \infty]$ ir jam galioja toks pat principas, kad kuo tolimesnės reikšmės nuo vieneto, tuo taisyklė yra įdomesnė. Pateiktame pavyzdyje svarbos matas nors ir buvo didesnis (lygus 6), tačiau įsitikinimo matas pateikia santūresnį rezultatą (lygų 2,25). Įsitikinimas gali būti interpretuojams kaip santykis tarp tikėtino A dažnio, kuris įvyksta be B (kitaip tariant, kaip dažnai taisyklė klaidingai prognozuoja) jei A ir B buvo nepriklausomi ir padalinti iš klaidingų prognozių. Dėl to, įsitikinimas įgyja reikšmę 1, kai A ir B neturi bendrų elementų ir yra neapibrėžtas, kai taisyklė $A \Rightarrow B$ yra visuomet teisinga.

$$conv(A \Rightarrow B) = \frac{1 - supp(B)}{1 - conf(A \Rightarrow B)} \tag{5}$$

Šiame darbe elementų rinkiniu (angl. *itemset*) yra vadinamas vieno ir daugiau prekių (elementų) rinkinys, pavyzdžiui: {pienas, duona, sauskelnės}. Tada, susietumo taisyklėmis vadinama loginė implikacija $X \rightarrow Y$, kur X ir Y yra elementų rinkiniai. Pavyzdžiui: {pienas, sauskelnės} → {alus}. Turint transakcijų rinkinį T, susietumo taisyklių gavybos (angl. *mining*) tikslas yra rasti visas taisykles kurių dažnumas ≥ minimaliam dažnumo slenksčiui, o patikimumas ≥ minimaliam patikimumo slenksčiui.

Norint rasti susietumo taisykles dažnai naudojamas apriori algoritmas, kurio pagrindinis principas teigia: jei elementų rinkinys dažnas, tai visi jo poaibiai (angl. subset) bus taip pat dažni. Apriori principas galioja, dėl to, kad elementų rinkinio dažnumas niekada neviršija jo poaibių dažnumo (anti-monotoniška dažnumo savybė):

$$\forall X, Y\ (: X \subseteq Y) \Rightarrow supp(X) \geq supp(Y) \tag{6}$$

Apriori algoritmo veikimo eiga:

- pradeda nuo vienos prekės elementų rinkinių;
- praeina duomenis ir skaičiuoja dažnumus, randa visus vieno elemento rinkinius, kurie atitinka nustatytą dažnumo slenkstį;
- kombinuojami vienetiniai elementų rinkiniai, iš kurių sudaromi dviejų elementų rinkiniai;
- praeina duomenis ir skaičiuoja dažnumus, randa visus dviejų elementų rinkinius, kurie atitinka nustatytą dažnumo slenkstį;
- kombinuojami elementų rinkiniai iš kurių gaunami trijų elementų rinkiniai;
- ir t.t.

2.2. Metodika vartotojų pasitenkinimo prognozavimui

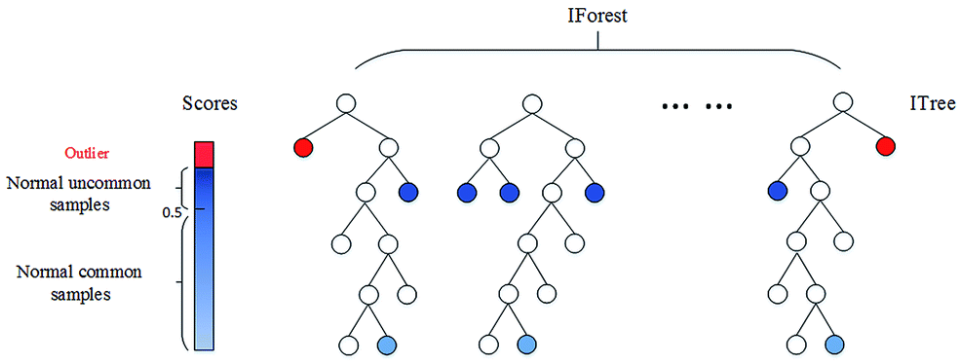
Šiame darbe vartotojų pasitenkinimui prognozuoti naudotos dvi pagrindinės metodikos: imčių ėmimo metodai klasių disbalanso sumažinimui ir mašininio mokymosi algoritmai klasifikavimo uždavinio prognozavimui.

2.2.1. Išskirčių radimas

Išskirčių (ar anomalijų) radimo metodai gali būti naudingi ir tokiuose nesubalansuotų klasių uždaviniuose kaip sukčiavimų aptikimas ar vartotojų pasitenkinimo prognozavimas. Izoliavimo miškas (angl. *Isolation Forest*) yra aktualus anomalijų aptikimo algoritmas šių dienų uždaviniuose, nes jis nenaudoja atstumų ar tankio matų ir dėl to jį galima pritaikyti didžiuosiuose duomenyse (skaičiavimų augimo laikas yra tiesinis). Šis metodas išskiria anomalijas, remdamasis pagrindinėmis anomalijų ypatybėmis:

- Mažumos klasė yra retai pasirodantis atvejis;
- Mažumos klasės stebiniai turi stipriai išsiskiriančias reikšmes iš įprastų atvejų.

Izoliavimo miške, medžio struktūra yra naudojama stebinių izoliavimui. Izoliavimo medis yra sudaromas atliekant atsitiktinį kintamojo padalinimą tarp jo minimalios ir maksimalios reikšmės. 1 pav. galima pamatyti, kad izoliavimo medžiuose išskirtimis yra laikomos reikšmės, kuriose medis nustoja augti arti savo šaknies. Reikšmių išskirtinumas yra matuojamas pagal medžio gylį. Pagal šiuos atstumus yra skaičiuojami visų izoliavimo medžių vidurkiai. Kai izoliavimo miške medžiai stebiniams kolektyviai nustato trumpesnius atstumus, tai tokie stebiniai yra laikomi anomalijomis.



1 pav. Izoliavimo miško struktūra

2.2.2. Kintamųjų atranka

Per pastarąjį dešimtmetį daugiamačiai duomenys su dideliais kiekiais kintamųjų tapo mašininio mokymosi kasdienybe. Norint išgauti naudingos informacijos iš tokių duomenų yra naudojami statistiniai metodai, kurie sumažina triukšmą arba perteklinius duomenis. Norint sudaryti

mašininio mokymosi modelį, dažnu atveju ne visi kintamieji yra reikalingi. Tam naudojama kintamųjų atranka. Atrinkus tik reikalingus kintamuosius modelio apmokymo laikas sutrumpėja, pats modelis supaprastinimas, o taip pat gali pagerėti modelio prognozavimo tikslumas. Šiame darbe naudojamas Boruta algoritmas iš R programos.

Boruta algoritmas vertina kintamųjų svarbą atsako kintamojo atžvilgiu naudodamas atsitiktinio miško (angl. *Random Forest*) klasifikavimo algoritmą. Šis metodas atlieka „iš viršaus į apačią“ kintamųjų svarbos paiešką lygindamas pradinių kintamųjų svarbą su galima atsitiktine svarba ir atmesdamas kintamuosius, kurie nustatyti kaip nereikšmingi.

Boruta iteratyviai palygina pradinių kintamųjų svarbą su šešėliniais kintamaisiais (kurių sumaišytos eilutės). Kintamieji, kurie turi reikšmingai prastesnę svarbą nei jų šešėliniai kintamieji yra pašalinami. Priešingu atveju, kai kintamieji yra ženkliai svarbesni nei jų šešėliai, jie yra laikomi reikšmingais. Šešėliniai kintamieji yra sukuriami iš naujo kiekvienos iteracijos metu. Algoritmas sustoja, kai lieka tik reikšmingi kintamieji arba pasiekiamas nustatytas maksimalus iteracijų skaičius. Antruoju atveju, kai kurie kintamieji lieka nepriskirti reikšmingiems ar atmestiems požymiams. Borutos algoritmo atliekami žingsniai:

1) Duomenų rinkinyje prie esamų kintamųjų pridedamos jų kopijos;
2) Kopijos yra sumaišomos, kad neliktų koreliacijos su atsako kintamuoju;
3) Apmokomi duomenys su *Random Forest* algoritmu ir gaunamas kiekvieno kintamojo Z-įvertis;
4) Randamas maksimalus Z-įvertis tarp šešėlinių kintamųjų (MZŠK), o pradiniai kintamieji kurie turi aukštesnį įvertį nei MZŠK yra pažymimi;
5) Kiekvienam kintamajam, kurio svarba liko neapibrėžta, atliekamas dvipusis lygybės testas su MZŠK;
6) Išmetami kintamieji, kurie turi reikšmingai mažesnę svarbą nei MZŠK;
7) Kintamieji kurie turi reikšmingai didesnę svarbą nei MZŠK yra laikomi svarbiais;
8) Pašalinami visi šešėliniai kintamieji;
9) Kartojami 1-8 žingsniai, kol svarba yra priskirta visiems kintamiesiems arba algoritmas pasiekė nustatytą iteracijų limitą.

2.2.3. Imties ėmimo metodai

Klasifikavimo uždaviniuose didelis skirtumas tarp atsako kintamojo Y klasių dažnių gali turėti neigiamą įtaką modelio apmokymui. Vienas iš būdų šiai problemai spręsti yra apmokymo imties klasių disbalanso sumažinimas imties metodais. Šiame darbe naudoti vieni iš populiariausių imčių ėmimo metodų:

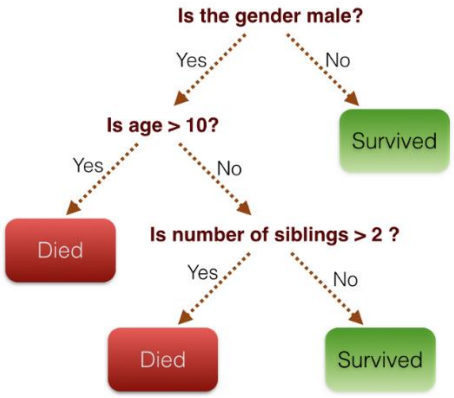
- *down-sampling:* atsitiktinai imama imtis kiekvienoje klasėje, kurios dydis yra lygus rečiausios klasės dydžiui duomenyse. Pavyzdžiui, jei klasė 1 pasikartoja 1 000 kartų, o klasė 2 pasikartoja 9 000 kartų, tai *down-sampling* metodas iš antros klasės atsitiktinai atrinks 1 000 stebinių. Tokiu atveju duomenų rinkinys sumažės nuo 10 000 iki 2 000 stebinių. Taikant šį metodą yra prarandama didelė dalis daugumos klasės stebinių, dėl to jis naudingas, kai duomenų rinkinys yra labai didelis;
- *up-sampling:* mažumos klasės stebiniai yra atsitiktinai pakartojami tiek kartų, kad mažumos klasės dydis susilygintų su daugumos klase. Šis metodas yra naudingas, kai duomenų rinkinys yra mažas. Verta pastebėti, kad atkartojant mažumos klasės stebinius išauga modelio persimokymo tikimybė;
- hibridiniai metodai: naudoja *down-sampling* metodą daugumos klasės sumažinimui ir susintetina naujus stebinius mažumos klasei. Darbe naudoti SMOTE ir ROSE hibridinai metodai.

Naudojant imties metodus klasių disbalanso mažinimui yra nerekomenduojama to daryti testavimo imčiai, nes ši imtis yra skirta įvertinti modelio gebėjimui klasifikuoti nematytus duomenis.

2.2.4. Sprendimų medžiai

Vartotojų pasitenkinimo prognozavimas šiame darbe yra klasifikavimo uždavinys su mokytoju (angl. *supervised classification*). Tai reiškia, kad mašininio mokymosi algoritmai yra apmokomi su duomenimis, kuriuose atsako kintamasis nurodo vartotojų pasitenkinimą. Tokių uždavinių sprendimui dažnai naudojami algoritmai, kurie remiasi sprendimų medžiais.

Medžiai turi daug gyvenimiškų analogijų, jie taip pat įkvėpė ir mašininio mokymosi sritį sprendžiant klasifikavimo bei regresijos uždavinius. Sprendimų analizėje galima pritaikyti sprendimų medį, kuris padeda atvaizduoti sprendimų atlikimo procesą (2 pav.). Šiame pavyzdyje sprendimų medis skirsto keleivius pagal kategorinius ir skaitinius kintamuosius. Pirmame susikirtimo taške keleiviai padalinami pagal kategorinį lyties kintamąjį siekiant nustatyti ar jiems pavyko išgyventi laivo sudužimą. Tolimesniems padalinimams naudojami skaitiniai kintamieji, tokie kaip amžius ir brolių bei seserų skaičius.



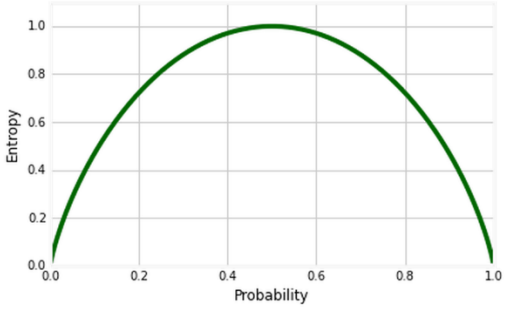
2 pav. Sprendimų medžio pavyzdys

Norint pasirinkti kurį požymį padalinti reikalingas padalinimo gerumo matas. Tam yra naudojami informacijos gavimo (angl. *information gain*) ir entropijos (angl. *entropy*) matai. Sprendimų medžiui užduodami klausimai turėtų suteikti daug informacijos apie medžio atliktas prognozes. Pavyzdžiui, jeigu paprastas „taip arba ne“ klausimas gali prognozuoti 99% tikslumu, tai toks klausimas leidžia „gauti“ daug informacijos apie duomenis. Norint išmatuoti informacijos gavimą yra naudojamas entropijos matas. Entropija matuoja duomenų neužtikrintumą (angl. *uncertainty*). Grįžtant prie ankstesnio pavyzdžio: jeigu kiekvienas keleivis išgyveno, tuomet entropija bus maža. Todėl siekiamybė yra taip padalinti duomenis, kad entropija būtų mažiausia. Tokiu atveju prognozės bus tiksliausios. Entropija yra apskaičiuojama pagal šią formulę:

$$H = -\sum p(x)\log(p(x)), \tag{7}$$

čia $p(x)$ – grupės priklausančios klasei procentas.

Entropija yra didelė kai kintamojo reikšmės pasiskirsčiusios tolygiai pagal klases ir maža, kai dauguma priklauso vienai klasei (pvz., „išgyveno“).



3 pav. Entropija ir tikimybė priklausyti klasei

Tai iliustruoja 3 pav.: jeigu kintamasis padalinamas su vienoda 50% tikimybe, tuomet entropija bus didžiausia. Kita vertus, entropija bus mažiausia, kai tikimybė priklausyti klasei bus

arba maža, arba didelė. Todėl sprendimų medis sudaromas taip, kad padalinimai minimizuotų entropiją. Tai nustato informacijos gavimas:

$$Gain(S, D) = H(S) - \sum_{V \in D} \frac{|V|}{|S|} H(V), \quad (8)$$

čia S – pradinė imtis, D – padalinta imtis, V – S poaibis.

Informacijos gavimo matą galima suprasti kaip duomenų entropiją prieš H(S) padalinimą iš jos atėmus svorinę entropijos padalinimų sumą.

Apžvelgus kaip sudaromas sprendimų medis reikia paminėti jo privalumus ir trūkumus. Sprendimų medžiai yra lengvai interpretuojami bei gali susitvarkyti su trūkstamomis reikšmėmis ir išskirtimis. Jie taip pat susitvarko su kategoriniais ir skaitiniais kintamaisiais bei su nereikšmingais požymiais. Tačiau pavieniai sprendimų medžiai gali nesunkiai persimokyti (angl. *overfit*). Todėl toliau bus apžvelgiami sudėtingesni algoritmai, kurie sprendimus priima sugeneruodami daug sprendimų medžių.

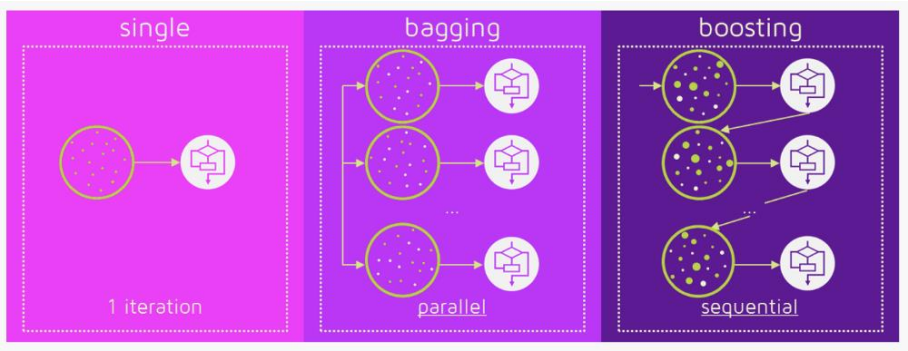
2.2.5. Ensemble metodai

Prognozuojant atsako kintamąjį su mašininio mokymosi algoritmais, pagrindiniai veiksniai, lemiantys skirtumą tarp tikrosios ir prognozuotos Y reikšmės yra: triukšmas, dispersija (angl. *variance*) bei poslinkis (angl. *bias*). Šiems veiksniams sumažinti yra naudojami vadinamieji „ansamblio“ (angl. *ensemble*) metodai. Šių metodų idėja yra tokia, kad naudojant daugybę skirtingų prediktorių to paties atsako kintamojo prognozei bus naudingiau nei tai atliktų bet kuris vienas prediktorius. „Ansamblio“ metodai yra meta algoritmai, kurie apjungia kelis mašininio mokymosi metodus į vieną prognozavimo modelį: *bagging* (sumažina dispersiją), *boosting* (sumažina poslinkį) bei *stacking* (padidina prognozavimo tikslumą).

Toliau šie metodai yra klasifikuojami į:

Bagging metodas

Taikant šį metodą yra sudaroma daug nepriklausomų prediktorių/modelių ir po to kombinuojami jų rezultatai naudojant vidurkius ar daugumos balsą. Tipiškai yra imamos atsitiktinės imtys kiekvienam modeliui, kad kiekvienas modelis skirtųsi nuo kitų. Kiekvienas stebinys turi vienodą tikimybę būti panaudotas kiekviename modelyje. Kadangi šis metodas paima daug nekoreliuotų modelių galutiniam modeliui, taip sumažėja modelio dispersija ir tuo pačiu paklaidos. *Bagging* metodas taip pat yra atsparus persimokymo problemai (4 pav.).



4 pav. *bagging* ir *boosting* metodai

***Boosting* metodas**

Šiuo atveju prediktoriai nėra sudaromi nepriklausomai, o iš eilės. Taigi, kiekvienas naujas prediktorius mokosi iš prieš tai buvusiojo klaidų. Dėl to stebiniai turi nevienodas galimybes būti įtraukti į tolimesnius modelius. Kadangi nauji prediktoriai mokosi iš ankstesnių, todėl prognozavimui reikia mažiau laiko/iteracijų. Tačiau čia iškyla naujas iššūkis: reikia atsargiai parinkti sustabdymo kriterijų, kad modelis nepersimokytų. Prediktoriai gali būti sprendimų medžiai, regresoriai, klasifikatoriai ir pan.

***Stacking* metodas**

Dauguma „ansamblio“ metodų naudoja vieną bazinį algoritmą, iš kurio sudaro homogeninį „ansamblį“. Tačiau yra tokių „ansamblio“ metodų, kurie naudoja heterogeninius algoritmus (skirtingų tipų). Tokiu atveju, norint pasiekti aukštą tikslumą, baziniai algoritmai privalo būti kuo tikslesni bei kiek įmanoma skirtingi.

Stacking (išvertus iš anglų k. krovimas) yra „ansamblio“ metodas, kuris apjungia klasifikavimo ar regresijos modelius naudodamas meta algoritmą. Iš pradžių apmokomi baziniai (pirmo lygio) algoritmai, o jų prognozės yra naudojamos antro lygio algoritmui apmokyti. Šis procesas gali būti kartojamas ir toliau, „kraunant“ ir apmokant algoritmus žemesnio lygio algoritmų prognozėmis.

Kadangi *stacking* metodu baziniai algoritmai yra naudojami kuo skirtingesni, todėl šis „ansamblis“ įprastai yra heterogeninis. Tokiose srityse, kur yra vertinamas aukščiausias tikslumas (pavyzdžiui *Kaggle* varžybos), *stacking* padeda pasiekti didesnį tikslumą nei individualus algoritmas bei sumažina persimokymo riziką. Tiesa, Kaggle varžyboms laimėti neretai naudojamos kelios dešimtys kruopščiai suderintų algoritmų, iš kurių yra sudaryti kelių lygių stacking „ansambliai“.

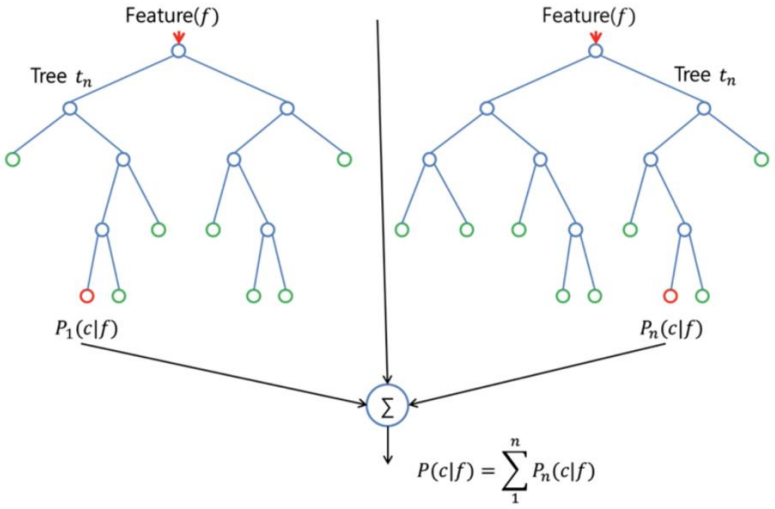
2.2.6. Mašininio mokymosi algoritmai

Toliau bus apžvelgiami darbe naudoti mašininio mokymo algoritmai klasifikavimo uždaviniuose. Visi naudoti algoritmai yra paremti sprendimų medžiais, nes tokius algoritmus yra paprasta pritaikyti, o jų tikslumas yra sunkiai pranokstamas.

Pasirinkti mašininio mokymosi algoritmai „su mokytoju“ – naudojami apmokymo duomenys su daug nepriklausomų kintamųjų x_i, tam, kad būtų prognozuojamas y_i.

Random Forest

Atsitiktinio miško (angl. *Random Forest*) idėja yra sugeneruoti daug sprendimų medžių ir juos apjungti siekiant gauti tikslesnius ir patikimesnius rezultatus (*bagging* metodas). Iš pradžių yra generuojami sprendimų medžiai: kiekvieno klasifikavimo medžio šakos pasibaigia atitinkamai parinkta klase. Kiekvienas medis atlieka klasifikavimą ir galutiniame rezultate balsuoja už klases. Iš medžių sudarytame miške kiekvienas stebinys yra klasifikuojamas pagal balsų daugumą. 5 pav. yra pateiktas atsitiktinio miško pavyzdys, kai yra du sprendimo medžiai.



5 pav. Atsitiktinio miško struktūra

Atsitiktinio miško algoritmas ima imtį su grąžinimu sprendimų medžių generavimui. Kadangi imtis imama su grąžinimu, tai didžioji duomenų dalis atsikartoja skirtinguose sprendimų medžiuose. Kita vertus, kiekvienas medis yra apmokomas tam tikra duomenų dalimi ir dėl šios savybės atsitiktinis miškas yra atsparus persimokymui.

Atsitiktinis miškas gali pateikti santykinę kintamųjų svarbą atsako kintamojo prognozei. Šis algoritmas yra paprastas ir patogus naudoti sprendžiant realius uždavinius. Jo numatytieji hiper-parametrai dažniausiai grąžina gerus rezultatus. Hiper-parametrų skaičius nėra didelis ir jų derinimui pirmiausiai reikia atkreipti dėmesį į medžių skaičių miške ir didžiausią sprendimo medžio gylį.

eXtreme Gradient Boosting

eXtreme Gradient Boosting, plačiau žinomas kaip *XGBoost*, yra 2014 metais pristatytas *boosting* tipo algoritmas, kuris greitai išpopuliarėjo ir dažnai naudojamas Kaggle varžyboms laimėti. *XGBoost* yra vienas tiksliausių mašininio mokymosi algoritmų, taip pat jis itin spartus, bei gerai apsaugotas nuo persimokymo. Toliau bus detaliau apžvelgiamos šio algoritmo ypatybės.

Pirmiausiai *XGBoost* algoritmui yra apibrėžiama tikslo funkcija ir po to ji optimizuojama. Tarkime, yra parenkama tokia tikslo funkcija (turinti apmokymo nuostolį ir reguliarizavimą (angl. *training loss and regularization*)):

$$obj(\theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{9}$$

Toliau reikia išsiaiškinti kokie yra medžių parametrai. Tam reikia surasti funkcijas f_i, iš kurių kiekviena pasižymi medžio struktūra. Tai yra labai sudėtinga padaryti visiems medžiams vienu metu, dėl to yra naudojama adityvi strategija: pridedant po naują medį yra pataisomos prieš tai buvusios klaidos su nauja išmokta informacija. Tada, prognozės reikšmė žingsnyje t užrašoma kaip $\hat{y}_i^{(0)}$ ir gaunama:

$$\begin{aligned}
\hat{y}_i^{(0)} &= 0 \\
\hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\
\hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\
&\dots \\
\hat{y}_t^{(t)} &= \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
\end{aligned} \tag{10}$$

Tuomet reikia pasirinkti koks medis reikalingas kiekviename žingsnyje. Tai bus toks medis, kuris optimizuoja tikslo funkciją:

$$obj^{(t)} = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{i=1}^{t} \Omega(f_i) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + const. \tag{11}$$

Kitas svarbus žingsnis yra reguliarizavimas – regresijos metodas, kuris sumažina persimokymo galimybę skirdamas baudas aukštesnio lygio polinomams (sudėtingesniems modeliams). Todėl būtina apibrėžti medžio sudėtingumą $\Omega(f)$. Prieš tai reikia pradėti nuo sprendimų medžio formulės:

$$f_t(x) = w_{q(x)}, w \in R^T, q: R^d \to \{1, 2, \dots, T\}, \tag{12}$$

čia $w-$ vektorius su lapų balais (svoriais), $q-$ funkcija, kuri atitinkamai priskiria reikšmes kiekvienam lapui, o T – yra lapų skaičius. Tada, *XGBoost* modelio sudėtingumas apibrėžiamas formule:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{13}$$

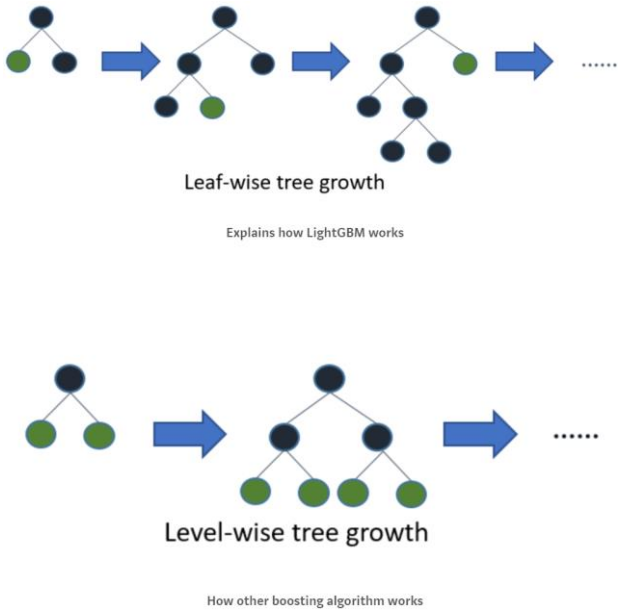
Apibendrinant, *XGBoost yra* gradientinis *boosting* metodas, kuris iš eilės sudeda silpnus klasifikatorius (ar regresorius) ir taiso prieš tai buvusių modelių klaidas. Tačiau, vietoj skirtingų svorių priskyrimo kiekvienos iteracijos iteracijos metu šis metodas apmoko modelį su naujomis ankstesnio modelio prognozių paklaidomis. Tuomet šis modelis minimizuoja pridėtos naujausios prognozės nuostolį. Todėl sudarant naujus sprendimo medžius modelis yra vis atnaujinamas naudojant gradientinio nusileidimo metodą. *XGBoost* taip pat turi reguliarizavimo narį savo optimizuojamoje tikslo funkcijoje, todėl šis metodas yra pakankamai atsparus persimokymo problemai.

XGBoost algoritmas priima tik skaitinius kintamuosius ir turi nemažai parametrų, kuriuos galima derinti, tačiau reikia atkreipti dėmesį, kad ir naudojant numatytus parametrus modelis tipiškai gerai klasifikuoja. R ir Python programose naudojami svarbiausi parametrai:

1) Bendrieji parametrai: kontroliuoja *booster* tipą, kuris lemia bendrą algoritmo funkcionavimą;
2) *Booster* parametrai: kontroliuoja pasirinkto *booster* veikimą. Pagrindiniai sprendimų medžių *booster* tipo parametrai yra iteracijų skaičius (klasifikavime tai yra panašu į auginamų medžių skaičių), mokymosi tempas (sprendimų medžio svoris), maksimalus medžio gylis (kuo didesnis, tuo modelis sudėtingesnis). Persimokymo prevencijai naudojami eilučių imties, stulpelių imties bei *gamma* (reguliarizavimo) parametrai;
3) Mokymosi užduoties parametrai: nustato ir įvertina pasirinkto *booster* mokymosi procesą apmokomiems duomenims. Yra galimybė naudoti nestandartinę tikslo funkciją ar paklaidų matus.

LightGBM

2017 metais Microsoft pristatė naują gradientinį *boosting* algoritmą *LightGBM*. Šis algoritmas augina sprendimų medžius lapų atžvilgiu (angl. *leaf-wise*), kai kiti *boosting* tipo algoritmai medžius augina medžio šakų lygio atžvilgiu (angl. *level-wise*). Tai schematiškai iliustruoja 6 pav.



6 pav. *LightGBM* medžių auginimo palyginimas su kitais *boosting* algoritmais

LightGBM naudoja naują metodą: gradientinį vienpusį imties ėmimą (angl. *Gradient-based One-Side Sampling* arba *GOSS)*. Šis metodas išfiltruoja stebinius ieškant kintamųjų padalinimo reikšmių. Naudojant *GOSS* yra priimama prielaida, kad apmokymo imties stebiniai su mažomis gradiento reikšmėmis turi mažesnes apmokymo paklaidas ir dėl to yra gerai apmokyti. *GOSS* pasilieka visus stebinius su didelėmis gradiento reikšmėmis ir ima stebinių su mažomis gradiento reikšmėmis atsitiktinę imtį. Norint išlaikyti tą patį duomenų skirstinį, skaičiuojant informacijos gavimą (angl. *information gain*) *GOSS* naudoja daugiklį stebiniams su mažais gradientais. Tokiu būdu *GOSS* pasiekia balansą tarp sumažinto stebinių skaičiaus bei sprendimų medžių tikslumo.

LightGBM tolydžių kintamųjų reikšmes histogramos principu suskirsto į diskrečias reikšmes. Taip paspartinamas algoritmo apmokymo laikas bei sumažinamas reikalingos atminties kiekis. Šis pranašumas leidžia apmokyti *LightGBM* algoritmą keletą kartų greičiau nei *XGBoost* (laiko kiekis priklauso nuo naudojamų hiper-parametrų). *LightGBM* iš viso turi virš 100 parametrų, kuriuos galima derinti apmokant algoritmą, bet didžiausią įtaką turi prieš tai minėti *XGBoost* parametrai kartu su bendro lapų skaičiaus medyje parametru. Verta pastebėti, kad *LightGBM* algoritmas nėra toks atsparus persimokymo problemai kaip *XGBoost*.

Regularized Greedy Forest

Nors gradientinių *boosting* algoritmų taikymas per pastarąjį dešimtmetį labai išpopuliarėjo praktikoje, tačiau jie pasižymi keliais trūkumais. R. Johnson su T. Zhang pasiūlė alternatyvą *boosting* algoritmams – reguliarizuotą godųjį mišką (angl. *Regularized Greedy Forest,* toliau *RGF*). Šio algoritmo idėja yra naudoti struktūrizuotą godaus principo paiešką, kuri yra apribojama reguliarizavimu. Galutiniame rezultate medžių „ansamblis“ pasižymi geru tikslumu bei yra sudaromas efektyviau nei tai daro gradientinio nusileidimo *boosting* algoritmai.

Vienas gradientinių *boosting* algoritmų trūkumų yra reguliarizavimo nebuvimas, todėl *RGF* algoritmas naudoja reguliarizavimo funkciją, kuri tiesiogiai priklauso nuo sudaryto miško struktūros. Kitą gradientinių algoritmų trūkumą autoriai įvardija kaip didelio medžių skaičiaus būtinybę tikslumui užtikrinti. Kai kuriais atvejais tai reikalauja daug kompiuterinių resursų. *RGF* algoritmas pakartotinai atlieka „godžią“ paiešką ir kartu optimizuoja visus koeficientus. Tai nesudaro persimokymo problemos, nes kartu yra naudojamas specialiai suformuluotas reguliarizavimas.

Apibendrinant, *RGF* algoritmas veikia tiesiogiai pagal sugeneruoto miško struktūrą (skirtingai nei gradientiniai *boosting* algoritmai). *RGF* algoritmas integruoja dvi pagrindines idėjas: reguliarizavimas pagal medžio struktūrą ir godaus principo integravimas generuojant sprendimų medžius [16].

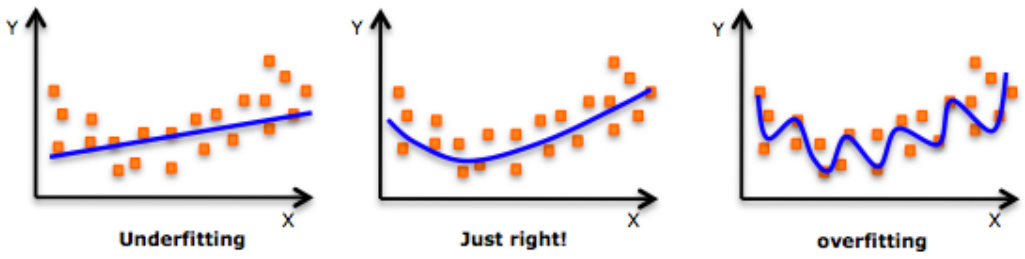
2.2.7. Algoritmų apmokymas ir modelio parametrų derinimas

Įprastai mašininio mokymosi uždaviniuose modeliai yra sudaromi duomenis padalinant į tris imtis:

- Apmokymo imtis (angl. *training set*): dažniausiai tai didžiausia imtis, kuri skirta apmokyti algoritmą su pasirinktais parametrais;
- Tikrinimo imtis (angl. *validation set*): skirta geriausių modelio parametrų parinkimui;
- Testavimo imtis (angl. *test set*): skirta tik įvertinti apmokyto modelio klasifikavimo rezultatus.

Šios imtys yra sudaromos dėl dviejų priežasčių: modelio parinkimas ir rezultatų įvertinimas. Norint parinkti modelį, reikia nustatyti optimalius jo parametrus konkrečiai klasifikavimo užduočiai. Šiame darbe naudojamiems *bagging* ir *boosting* algoritmams yra galimybė derinti įvairius parametrus. Norint įvertinti modelio gebėjimą prognozuoti yra naudojama testavimo imtis, o rezultatai vertinami pasirinktais tikslumo matais.

Apmokant modelį tenka susidurti su tokiomis problemomis kaip persimokymas (angl. *overfit*) ir nepakankamas mokymasis (angl. *underfit*). Abu atvejai yra nenaudingi siekiant sudaryti modelius, kurie gali tiksliai prognozuoti nematytus duomenis (7 pav.).



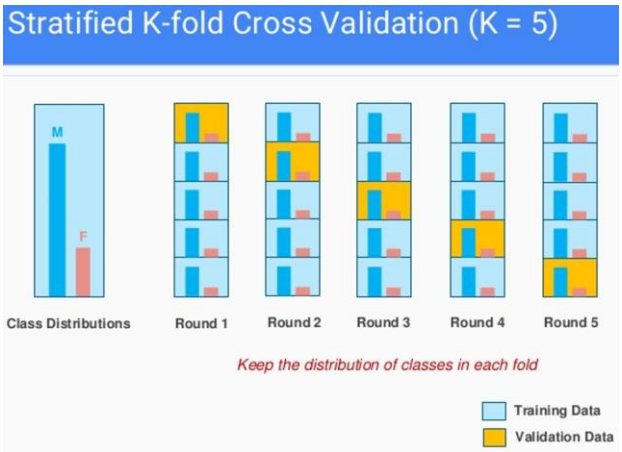
7 pav. Nepakankamas apmokymas, subalansuotas apmokymas ir persimokymas

Įprastai nepakankamas apmokymas būna tais atvejais, kai sudaromas labai paprastas modelis (nėra kintamųjų galinčių paaiškinti atsako kintamojo variaciją) arba kai parenkamas tiesinis modelis (pvz. tiesinė regresija) netiesiniams ryšiams modeliuoti.

Gerokai dažniau yra susiduriama su persimokymo problema. Taip nutinka, kai modelis yra per daug sudėtingas (pvz., turi per daug kintamųjų stebinių atžvilgiu). Toks modelis gerai aprašo apmokymo imties duomenis, tačiau dėl persimokymo jis išmoksta ne tik realius ryšius tarp kintamųjų bet ir triukšmą. Persimokymui išvengti dažnai duomenys padalinami į apmokymo ir testavimo imtis, o apmokymo imtyje dar atliekamas kryžminis patikrinimas.

Dažniausiai praktikoje yra naudojamas k-dalių kryžminis patikrinimas (angl. *k-fold cross validation*): apmokymo duomenys padalinami į *k* dalių. Tada, apmokymui naudojama *k* - 1 dalis, o patikrinimui paliekama likusi viena dalis. Šio metodo privalumas – visa apmokymo imtis yra išnaudojama apmokymui ir patikrinimui. Tokiu būdu galima turėti didesnę apmokymo imtį, nes nereikia turėti atskiros patikrinimo imties.

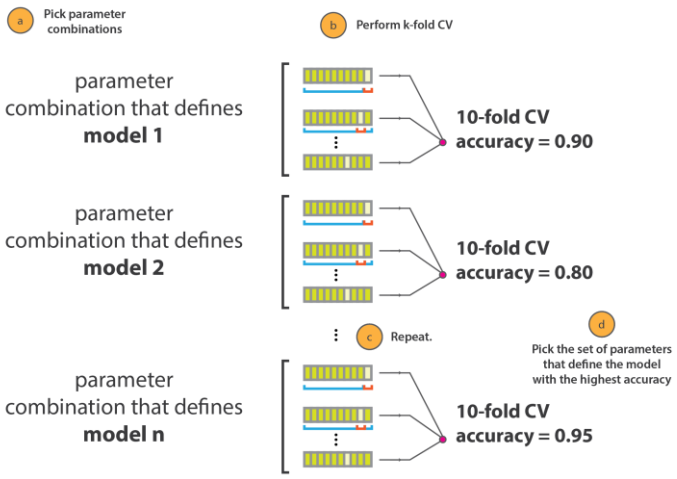
Kadangi šiame darbe naudoti duomenys pasižymi nesubalansuotomis klasėmis, todėl tinkamesnis yra sluoksniuotas k-dalių kryžminis patikrinimas (8 pav.). Pateiktame pavyzdyje yra *k* = 5 dalių kryžminis patikrinimas, vadinasi sudaromi 5 modeliai su skirtingomis apmokymo ir tikrinimo imtimis. Šis metodas papildomai įveda sąlygą, kad kiekvienoje dalyje turi būti išlaikomas apytikslis visos apmokymo imties klasių santykis.



8 pav. Sluoksniuotas k-dalių kryžminis patikrinimas

Kryžminis patikrinimas yra naudojamas kartu su parametrų derinimu. Šiame darbe parametrų derinimui buvo naudota tinklelio paieška (angl. *grid search*). Šiam metodui yra reikalinga nurodyti algoritmą bei parametrų sąrašą derinimui, pagal kurį yra sudaromos visos įmanomos modelių kombinacijos. Tada kiekviena parametrų kombinacija yra apmokoma ir tikrinama naudojant kryžminį patikrinimą. Iš gautų *k* modelių rezultatų apskaičiuojamas pasirinkto mato vidurkis: klasifikavimo atveju tai yra tikslumo matas, o regresijos uždavinyje paklaidos matas. Naudojamas matas yra geriausio modelio pasirinkimo kriterijus. Tai iliustruoja 9 pav., kuriame sudaroma *n*

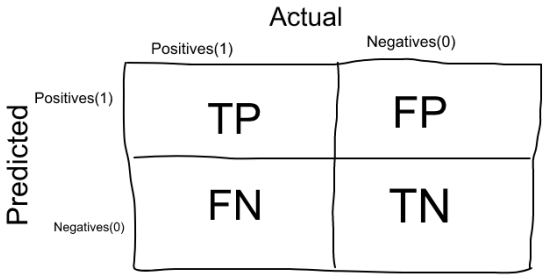
modelių, o kiekvienam modeliui yra atliekamas 10 dalių kryžminis patikrinimas. Pagal tikslumo matą pasirenkamas geriausias modelis, kurio realus gebėjimas prognozuoti nematytus duomenis vėliau turėtų būti įvertinamas su testavimo imtimi.



9 pav. Tinklelio paieška su kryžminiu patikrinimu

2.2.8. Klasifikavimo tikslumo matai

Klasifikavimo modelio tikslumas ir korektiškumas yra įvertinami prognozuotas reikšmes lyginant su tikrosiomis (kurios buvo atidėtos tikrinimo arba testavimo imtyje). Pagrindinis būdas klasifikavimo gerumui įvertinti yra sumaišymo matrica (10 pav.). Šis metodas gali būti naudojamas, kai yra klasifikuojamos dvi klasės ar daugiau. Pati sumaišymų matrica nėra naudojama kaip tikslumo matas, tačiau pagal ją yra apskaičiuojami beveik visi tikslumo matai.



10 pav. Sumaišymų matrica

Šiame darbe vartotojų pasitenkinimo duomenys suskirstyti į dvi klases: daugumos klasėje yra patenkinti paslaugomis klientai (pažymėti 0), o mažumos klasėje yra nepatenkinti klientai (pažymėti 1). Naudojant tai kaip pavyzdį, yra lengviau paaiškinti sumaišymų matricos elementus:

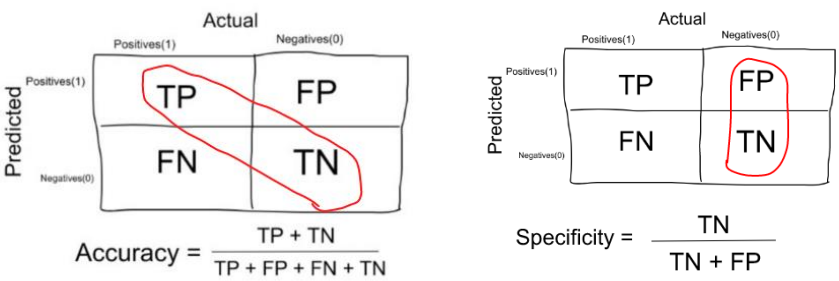
- Teisingas priėmimas (angl. *True Positive – TP*): atvejų skaičius, kai tikroji klasė ir prognozuojama klasė buvo lygios 1 (kai modelis teisingai prognozuoja, kad vartotojas yra nepatenkintas);

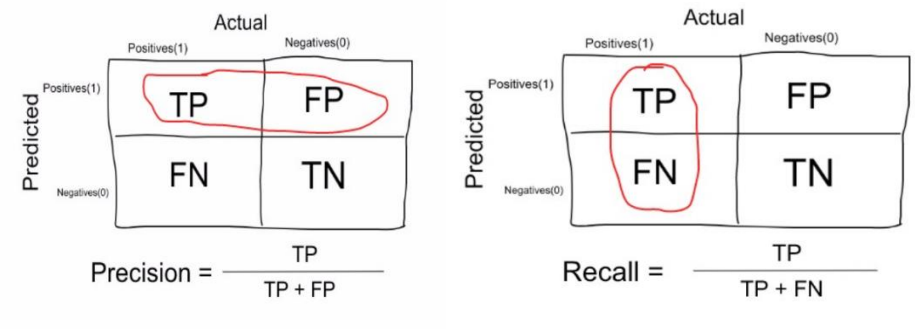
- Teisingas atmetimas (angl. *True Negative – TN*): atvejų skaičius, kai tikroji klasė ir prognozuojama klasė buvo lygios 0. Arba kiek kartų modelis teisingai prognozuoja daugumo klasę – 0;
- Klaidingas priėmimas (angl. *False Positive – FP*): atvejų skaičius, kai buvo prognozuota mažumos klasė (1), o tikroji klasė buvo daugumos (0). Ši klaida statistikoje dar žinoma kaip I rūšies klaida;
- Klaidingas atmetimas (angl. *False Negative – FN*): atvejų skaičius, kai buvo prognozuota daugumos klasė (0), o tikroji klasė buvo mažumos (1). Ši klaida statistikoje dar žinoma kaip II rūšies klaida.

Idealiu atveju sumaišymų matricoje turėtų nebūti klaidingų priėmimų ir klaidingų atmetimų. Ar labiau apsimoka minimizuoti *FP* ar *FN*, priklauso nuo uždavinio. Vartotojų pasitenkinimo atveju svarbesnis yra klaidingas atmetimas (II rūšies klaida), kai nepavyksta identifikuoti kliento, kuris nori atsisakyti paslaugų.

Pagal sumaišymo matricos elementus galima apskaičiuoti įvairius tikslumo matus modelių palyginimui (11 pav.). Dažnai naudojami matai:

- *accuracy*: šis matas yra tinkamas naudoti, kai klasės yra subalansuotos. Tačiau tokiame uždavinyje kaip vartotojų pasitenkinimo prognozavimas *accuracy* netinka. Pavyzdžiui, jei yra 5% stebinių mažumos klasėje, tai modelis pasieks 95% tikslumą visus stebinius klasifikuodamas į daugumos klasę;
- *specificity*: nusako kokia dalis vartotojų, kurie **nepriklausė** mažumos klasei buvo teisingai priskirti prie daugumos klasės;
- *precision*: nusako kokia dalis vartotojų, kurie buvo **prognozuoti** kaip nepatenkinti, iš tikrųjų priklausė mažumos klasei. *Precision* parodo kokia dalis buvo klaidingai priimta (*FP*);
- *recall*: nusako kokia dalis vartotojų, kurie **priklausė** mažumos klasei buvo algoritmo prognozuoti kaip nepatenkinti. *Recall* parodo kokia dalis buvo klaidingai atmesta (*FN*), o tai yra aktualu vartotojų pasitenkinimo uždavinyje.





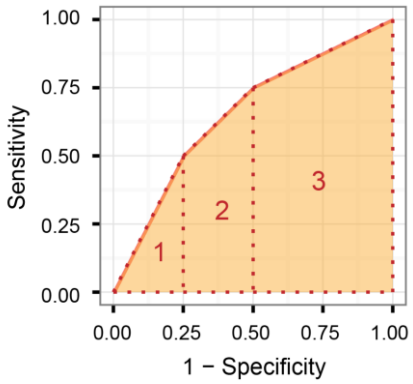
11 pav. *accuracy*, *specificity*, *precision* ir *recall* tikslumo matai

Kadangi *precision* atsižvelgia į *FP*, o *recall* į *FN*, tai paprasčiau būtų šiuos matus vertinti juos apjungus į vieną. Tam yra naudojamas *F1* matas, kuris yra harmoninis *precision* ir *recall* vidurkis:

$$F1 = \frac{2 * precision * recall}{(precision + recall)} \tag{14}$$

Harmoninis vidurkis ypatingas tuo, kad kai *precision* ir *recall* turi vienodas reikšmes, tuomet jis bus lygus aritmetiniam vidurkiui, o kai reikšmės skirtingos, tai *F1* bus artimesnis mažesnei reikšmei. *F1* tikslumo matą taip pat galima apskaičiuoti kiekvienai klasei atskirai. Tai yra labai aktualu, kai viena klasė yra laikoma svarbesne nei kita. Vartotojų pasitenkinimo prognozavimo atveju tai būtų nepatenkinti vartotojai (mažumos klasė).

Norint įvertinti kaip gerai modelis atskiria klases dažnai naudojama *ROC* (angl. *Receiver Operating Characteristic*) kreivė (12 pav.). Ši kreivė atvaizduoja: *x* ašyje 1 – *specificity*, o *y* ašyje *recall* (dar žinomą kaip *sensitivity*).



12 pav. Plotas po *ROC* kreive

Klasifikavimo uždaviniuose šią kreivę galima naudoti kaip tikslumo matą, apskaičiavus plotą po ROC kreive (angl. *Area Under Reveiver Operating Characteristc Curve* arba *AUC*). Šis matas parodo balansą tarp klasifikatoriaus gebėjimo priskirti klases į True Positive (teisingas priėmimas) ir False Positive (klaidingas priėmimas). Deja, *AUC* yra nejautri klasių disbalansui,

todėl yra mažiau vertinga vartotojų prognozavimo uždaviniui. *AUC* reikšmių interpretacija yra paprasta:

1 lentelė. *AUC* verčių interpretacija

***AUC* reikšmės**	**Klasifikavimo gerumas**
0,9-1	Puikus
0,8-0,9	Geras
0,7-0,8	Vidutinis
0,6-0,7	Prastas
0,5-0,6	Atsitiktinis

Dar vienas tikslumo matas yra *kappa* (angl. *Cohen's kappa coefficient*), kuris vertina dviejų nuomonių susitarimą (šiuo atveju prognozuotų ir realių reikšmių). Šis matas laikomas patikimesniu nei *accuracy* matas, nes *kappa* taip pat įtraukia hipotetinę tikimybę, kad susitarimas įvyko atsitiktinai:

$$\kappa \equiv \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - p_0}{1 - p_e}, \tag{15}$$

čia p_0 yra nuomonių tikslumas (*accuracy*), o p_e yra hipotetinė tikimybė, kad susitarimas atsitiktinis.

Nors teoriškai *kappa* kinta nuo -1 iki 1, tačiau praktikoje 0 reiškia, kad nuomonių sutapimas yra atsitiktinis, o 1 reiškia idealų susitarimą.

3. Tyrimų rezultatai ir jų aptarimas

3.1. Pirkinių krepšelio tyrimas

Pirkinių krepšelio analizei (PKA) atlikti bus tiriami Instacart kompanijos duomenys. Instacart siūlo internetinę paslaugą, kurią naudojantys vartotojai pildo virtualų prekių krepšelį per internetinę svetainę arba mobilią aplikaciją, o prekės yra pristatomos į jų namus greičiau nei per valandą. Ši kompanija vykdo veiklą JAV ir nuo įsikūrimo pradžios labai greitai įsitvirtino tuose miestuose, kuriuose ji siūlo maisto prekių internetu paslaugą. 2014 metais jau buvo pasiektos 100 mln. JAV dol. pajamos, o šiam verslui sėkmę žada ir įtakingi investuotojai.

Kuo Instacart save išskiria nuo savo pirmtako Webvan – jos verslo modeliu, kuris išnaudoja jau egzistuojančias maisto prekių parduotuves ir vežėjų transportą, vietoj nuosavos infrastruktūros kūrimo. Taigi, Instacart, išmoko savo pirmtakų pamokas, pasirinko saugesnį verslo modelį atsisakydama didžiausių išlaidų. Vietoj to, Instacart telkia visą savo dėmesį į užsakymų surinkimą bei pristatymą vartotojams.

Instacart anksti suprato, kad jie suteikia maisto prekių parduotuvėms inovatyvumo. Ši kompanija orientuojasi į didžiausią pelną nešančius pirkėjus, kurie yra linkę sumokėti daugiau už pridėtinį patogumą. Verta paminėti, kad šie duomenys nėra reprezentatyvūs ką apskirtai valgo amerikiečiai. Šie duomenys labiau atspindi ką valgo nedidelė amerikiečių dalis, kurie gali sau leisti šias paslaugas.

3.1.1. Žvalgomoji analizė

Šiame darbe analizuojami Instacart duomenys yra anoniminiai siekiant apsaugoti kompanijos klientų ir partnerių privatumą. Vienintelė prieinama informacija apie klientus yra užsakymų eiliškumas ir produktai, kurie yra užsakyme. Duomenų rinkinys susideda iš penkių lentelių, kuriose yra virš 30 mln. užsakymų iš daugiau nei 200 tūkst. klientų (2 lentelė). Apie kiekvieną klientą yra pateikiama tarp 4 ir 100 užsakymų išlaikant pirkinių seką užsakyme.

Duomenyse iš viso yra 21 prekių grupė, kurioje yra 134 prekių kategorijos, o unikalių produktų iš viso yra beveik 50 tūkst. Prekių grupes ir kategorijas yra patogu atvaizduoti su *treemap* diagrama (13 pav.). Šioje diagramoje prekių kategorijų dydis atspindi jose esančių produktų pirkimo dažnumą.

Instacart klientai dažniausiai perka vaisius ir daržoves, o taip pat dažnai perka tokius pieno produktus kaip jogurtas, sūris ir pienas. Galima pastebėti, kad Instacart internetinėje parduotuvėje yra nemaža paklausa sveikam maistui. Be jau minėtų vaisių ir daržovių, klientai perka soją be laktozės, granolos batonėlius, riešutus ir sėklas.

Kadangi duomenys neturi išsamių aprašymų, reikia remtis lentelių, kintamųjų vardais, bei įgyjamomis kintamųjų reikšmėmis. Pavyzdžiui, kintamasis „užsakymo savaitės diena“ įgyja reikšmes nuo 0 iki 6 ir nėra aišku, kuri tai yra savaitės diena (14 pav.). Galima tik spėti, kad dvi daugiau užsakymų turinčios dienos yra savaitgalis. Likusiomis savaitės dienomis užsakymų skaičius yra labai panašus.

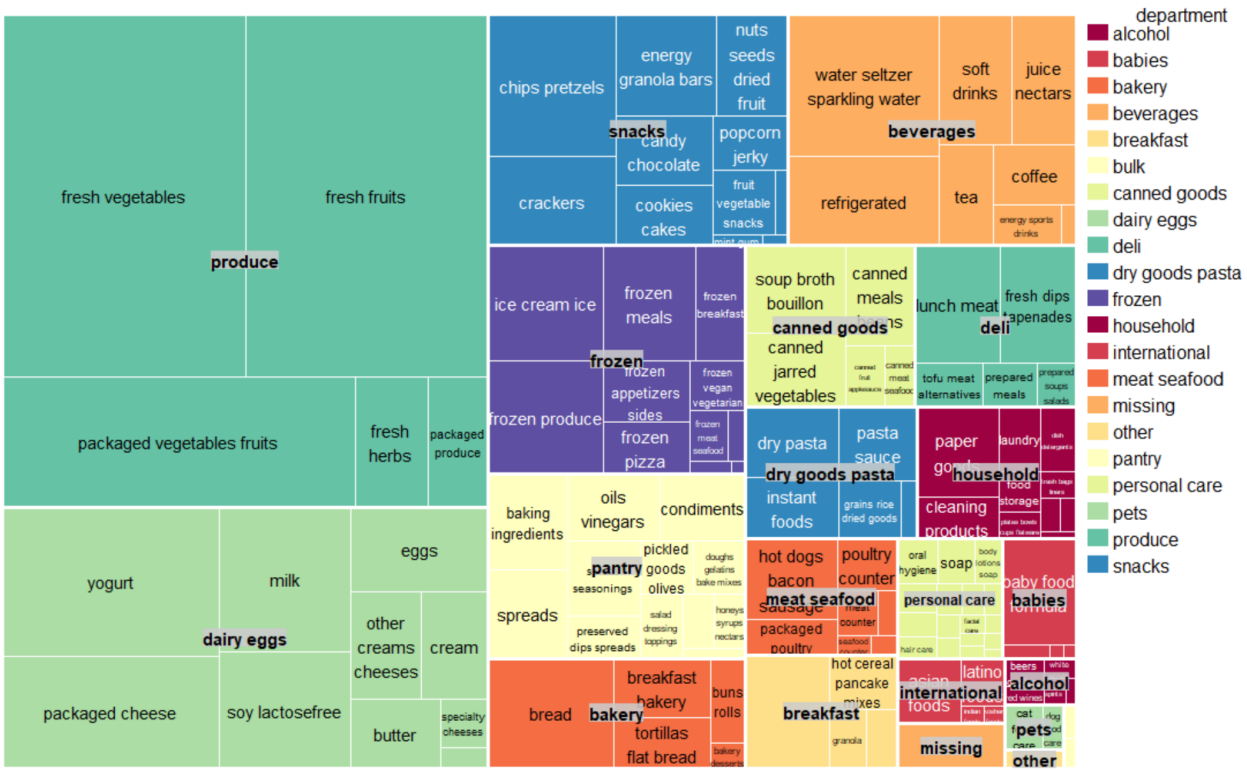
2 lentelė. Instacart duomenų struktūra

Lentelė	**Eil. sk.**	**Stulp. sk.**	**Lentelės pirmos trys eilutės**
Prekių grupės	21	2	department_id: int 1 2 3 department: chr "frozen" "other" "bakery"
Prekių kategorijos	134	2	aisle_id: int 1 2 3 aisle: chr "prepared soups salads" "specialty cheeses" "energy granola bars"
Produktai	49 688	4	product_id: int 1 2 3 product_name: chr "Chocolate Sandwich Cookies" "All-Seasons Salt" "Robust Golden Unsweetened Oolong Tea" aisle_id: int 61 104 94 department_id: int 19 13 7
Užsakymai	34 221 083	7	order_id: int 2539329 2398795 473747 user_id: int 1 1 1 eval_set: chr "prior" "prior" "prior" order_number: int 1 2 3 order_dow: int 2 3 3 order_hour_of_day: int 8 7 12 days_since_prior_order: num NA 15 21
Užsakyti produktai	1 384 617	4	order_id: int 1 1 1 product_id: int 49302 11109 10246 add_to_cart_order: int 1 2 3 reordered: int 1 1 0
Anksčiau užsakyti produktai	32 434 489	4	order_id: int 2 2 2 product_id: int 33120 28985 9327 add_to_cart_order: int 1 2 3 reordered: int 1 1 0

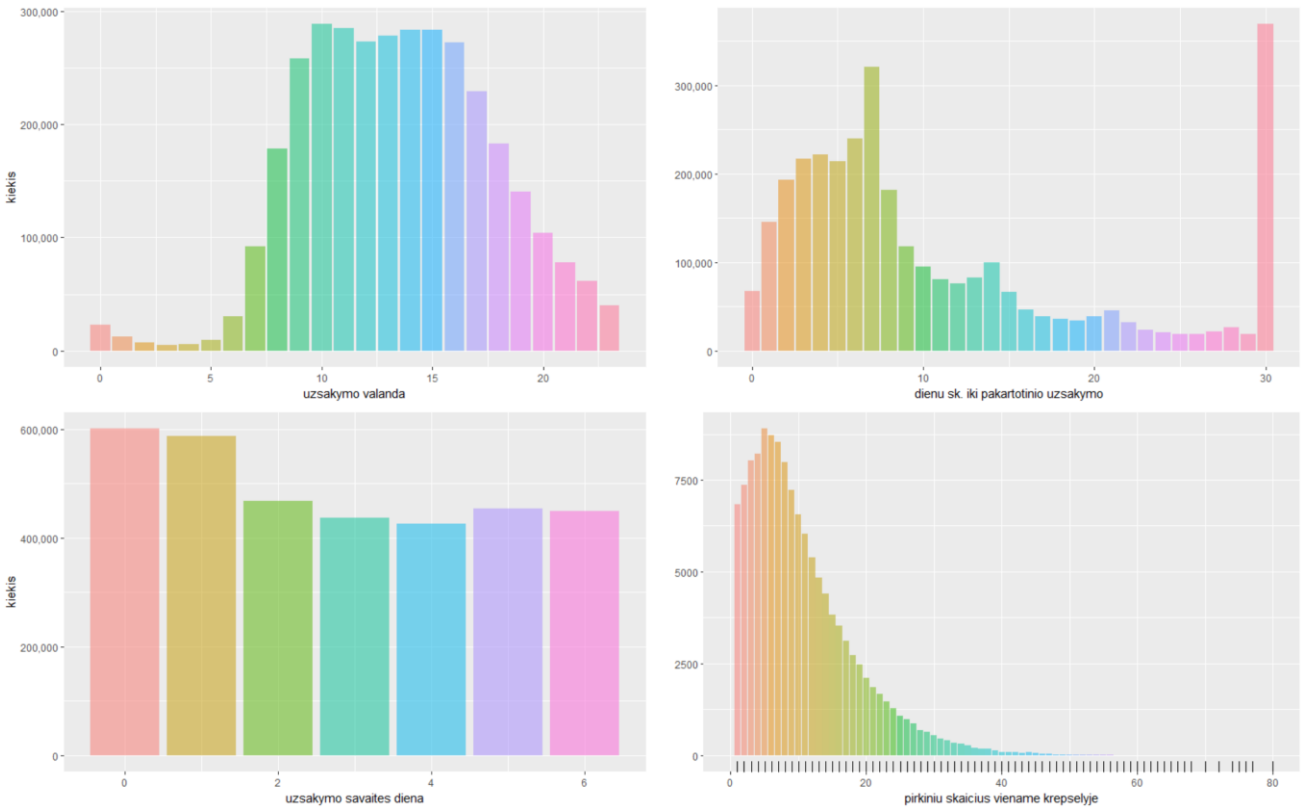
Analizuojant užsakymo valandas, nesunku pastebėti, kad aktyviausios užsakymų valandos yra dieną: nuo 10 iki 16 valandos. Kadangi užsakymai pristatomi labai greitai, todėl vartotojams yra patogu suplanuoti pietus ar vakarienę.

Atlikę užsakymą, klientai įprastai sugrįžta atlikti naujo užsakymo praėjus 2-8 dienoms. Kad klientas atliks dar vieną užsakymą tą pačią dieną yra mažiau tikėtina. Dažniausiai klientai atlieka naują užsakymą praėjus vienai savaitei. Prie to galimai prisideda periodiški atvejai, pavyzdžiui kai įmonė kas savaitę atlieka maisto produktų užsakymą savo darbuotojams. Pakartotinių užsakymų pikas grafike ties 30-ta diena yra dirbtinis, jo nereikia interpretuoti, kad dažniausiai žmonės vėl užsisako kas 30 dienų. Taip yra dėl to, kad visi pakartotiniai užsakymai 30-ies ir daugiau dienų yra laikomi kaip 30 dienų.

Pirkinių skaičius viename krepšelyje įprastai būna nedidelis, dažniausiai 5 produktai. Kuo užsakymas yra didesnis, tuo jis retesnis. Tokių atvejų duomenyse, kai buvo perkama 10 prekių, yra panašiai kaip pirktas tik 1 produktas.



13 pav. Prekių grupių ir kategorijų išsidėstymas pagal pardavimų dažnumą



14 pav. Klientų apsipirkimo Instacart internetinėje parduotuvėje įpročiai

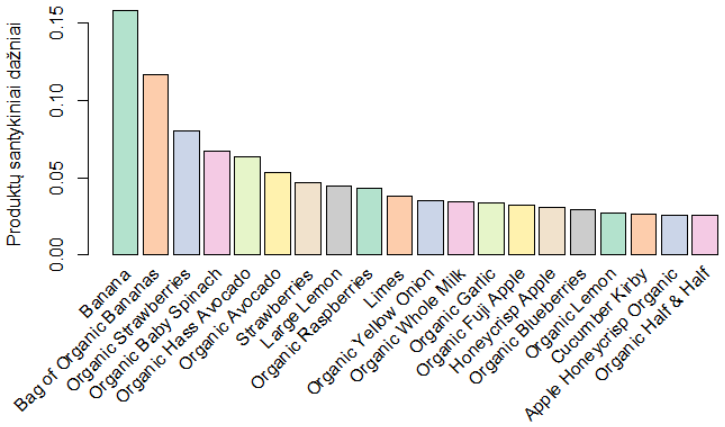
3.1.2. Susietumo taisyklių analizė

Šiame darbe bus sudaromos susietumo taisyklės prekių kategorijoms ir produktams siekiant atrasti įdomius sąryšius, kuriuos galima būtų panaudoti tikslingesniems klientų pasiūlymams.

Prieš sudarant susietumo taisykles yra aktualu ištirti kokios prekės dažniausiai pasitaiko pirkinių krepšelyje. 15 pav. yra pateiktos dažniausiai pirkinių krepšelyje pasitaikančios prekės pagal santykinį dažnumą. Dažniausiai perkama prekė yra bananai: daugiau nei 15% užsakymų yra bananai, o 12% užsakymų perkami ekologiški bananai. Tipiniame Instacart pirkinių krepšelyje yra vaisių ir daržovių, o tarp jų didžioji dalis yra ekologiški produktai. Iš to galima daryti išvadą, kad Instacart tikslinė auditorija nėra tipinė JAV, nes šioje internetinėje parduotuvėje populiariausias sveikas maistas.

Norint sudaryti stiprias susietumo taisyklės, reikia maksimizuoti įdomumo matų reikšmes:

- Dažnumas (angl. *support*): turėtų galioti dideliam atvejų kiekiui;
- Patikimumas (angl. *confidence*): taisyklė turėtų būti dažnai teisinga;
- Svarba (angl. *lift*): taisyklė nėra tik atsitiktinumas;
- Įsitikinimas (angl. *conviction*): taisyklė yra įdomi atsižvelgiant į jos kryptį.



15 pav. Dažniausiai užsakymuose pasitaikančios prekės

Tačiau, norint išvengti neįdomių (akivaizdžių) taisyklių, pavyzdžiui {švieži vaisiai, šviežios daržovės} => {supakuoti vaisiai ir daržovės}, praktikoje gali tekti naudoti žemesnę dažnumo reikšmę. Pirmiausiai sudaromos prekių kategorijų susietumo taisyklės, kurios turi didesnį dažnumą nei 0,001 ir didesnį patikimumą nei 0,3 (3 lentelė). Iš šios aibės atrinktos įdomesnės taisyklės, kurios turi arba dideles svarbos, arba įsitikinimo reikšmes. Aukščiausią svarbos reikšmę (44) turi taisyklė: jei pirks raudoną vyną, tai pirks ir baltą vyną (tokie atvejai pasikartoja 2 kartus iš 1000 su 30% patikimumu). Aukščiausią įsitikinimo reikšmę (9,5) turi taisyklė: jei pirks šviežias žoleles ir supakuotas jūros gėrybes tai užtikrintai pirks ir šviežias daržoves (su 0,94 patikimumu).

Verta atkreipti dėmesį, kad svarbos matas aukštas vertes įgija, kai patikimumas mažas, o įdomumo matas – kai patikimumas aukštas.

3 lentelė. Atrinktos įdomios prekių kategorijų susietumo taisyklės

Taisyklės	***support***	***confidence***	***lift***	***conviction***
{red wines} => {white wines}	0.002	0.30	44.0	1.4
{frozen vegan vegetarian,pasta sauce} => {tofu meat alternatives}	0.001	0.32	10.1	1.4
{laundry,paper goods} => {cleaning products}	0.003	0.31	9.8	1.4
{dish detergents,laundry} => {cleaning products}	0.001	0.31	9.6	1.4
{body lotions soap,laundry} => {paper goods}	0.001	0.52	6.7	1.9
{dry pasta,salad dressing toppings} => {pasta sauce}	0.002	0.44	6.6	1.7
{fresh herbs,packaged seafood} => {fresh vegetables}	0.001	0.94	2.1	9.5
{lunch meat,seafood counter} => {fresh vegetables}	0.001	0.91	2.0	5.8
{seafood counter,soy lactosefree} => {fresh vegetables}	0.002	0.90	2.0	5.5
{canned jarred vegetables,poultry counter} => {fresh vegetables}	0.005	0.90	2.0	5.5
{fresh herbs,packaged cheese} => {fresh vegetables}	0.030	0.90	2.0	5.2
{baby food formula,packaged produce} => {fresh fruits}	0.002	0.92	1.7	5.4

Vienas iš susietumo taisyklių atvaizdavimo būdų yra grafų diagrama (16 pav.). Šioje diagramoje yra atvaizduotos 16 aukščiausią svarbą turinčios prekių kategorijų taisyklės, kurių svarba yra bent 6 (su dažnumo 0,001 ir patikimumo 0,3 slenksčiais). Kiekvienas apskritimas žymi susietumo taisyklę, kuri sudaroma iš į apskritimą rodyklėmis nukreiptų prekių. Taisyklės santykinį dažnumą nusako apskritimo dydis, o taisyklės svarbą – spalvos ryškumas.

Nesunku pastebėti, kad diagramoje yra trys atskiri susietumo taisyklių klasteriai. Viršuje kairėje yra švaros prekių susietumo taisyklės. Keleto prekių taisyklės yra nukreiptos į popierinius reikmenis. Vadinasi, klientams perkant tokias prekes kaip muilas ar burnos higiena, galima pasiūlyti ką nors iš popierinių reikmenų kategorijos. Dešinėje viršuje yra aukščiausią svarbą turinti taisyklė: jei pirks raudoną vyną, tai pirks ir baltą vyną. Tokia taisyklė galioja 30% atvejų, kai perkamas raudonas vynas. Apačioje yra su maistu susijusios taisyklės, kurios daugumoje atvejų yra nukreiptos į makaronų padažus. Klientams, perkantiems šaldytus užkandžius ar picas, galima pasiūlyti makaronų padažus. Taip pat matomas makaronų išeinantis ryšys į įvairius maisto produktus. Šiuo atveju, perkant makaronus galima pasiūlyti nusipirkti produktų iš kelių skirtingų kategorijų.

16 pav. Prekių kategorijų tinklas iš 16 susietumo taisyklių

Susietumo taisyklės pagal konkrečius produktus gali būti naudingesnės siekiant vartotojams pateikti specialius pasiūlymus. Produktams susietumo taisyklės buvo atrinktos su dažnumu 0,003 ir patikimumu 0,25 (4 lentelė). Pagal gautas susietumo taisykles galima padaryti prielaidą, kad kiekviena taisyklė yra susijusi su tam tikru receptu. Svarbiausia taisyklė yra ta, kad saldžioji bulvė perkama kartu su raudonuoju svogūnu. O meksikietiškiems receptams galioja stipri taisyklė: jei pirks Jalapeno pipirus, tai pirks kartu ir žaliąją citriną. Lentelėje taip pat yra tokia mažiau įdomi taisyklė, kad avietės perkamos kartu su braškėmis.

Susietumo taisyklės produktams taip pat atvaizduotos grafų diagrama (17 pav.). Atrinktos taisyklės apribotos šiais minimaliais slenksčiais: svarba ≥ 4, dažnumas ≥ 0,003 ir patikimumas ≥ 0,25. Lyginant su prieš tai apžvelgta prekių kategorijų diagrama, čia taisyklių visuma labiau primena tinklą, o ne atskirus klasterius. Šiame tinkle praeinama pro ekologišką avokadą, saldžiąją bulvę, ekologiškus špinatus, bananus ir ekologiškas braškes. Šios prekės yra populiarios, dėl to į jas

veda įvairūs kiti produktai. Tačiau iš verslo pusės svarbu, kad klientui, kuris perka populiariausią prekę (bananą) yra sunkiau ką nors užtikrintai pasiūlyti nei klientui, kuris perka saldžiąją bulvę.

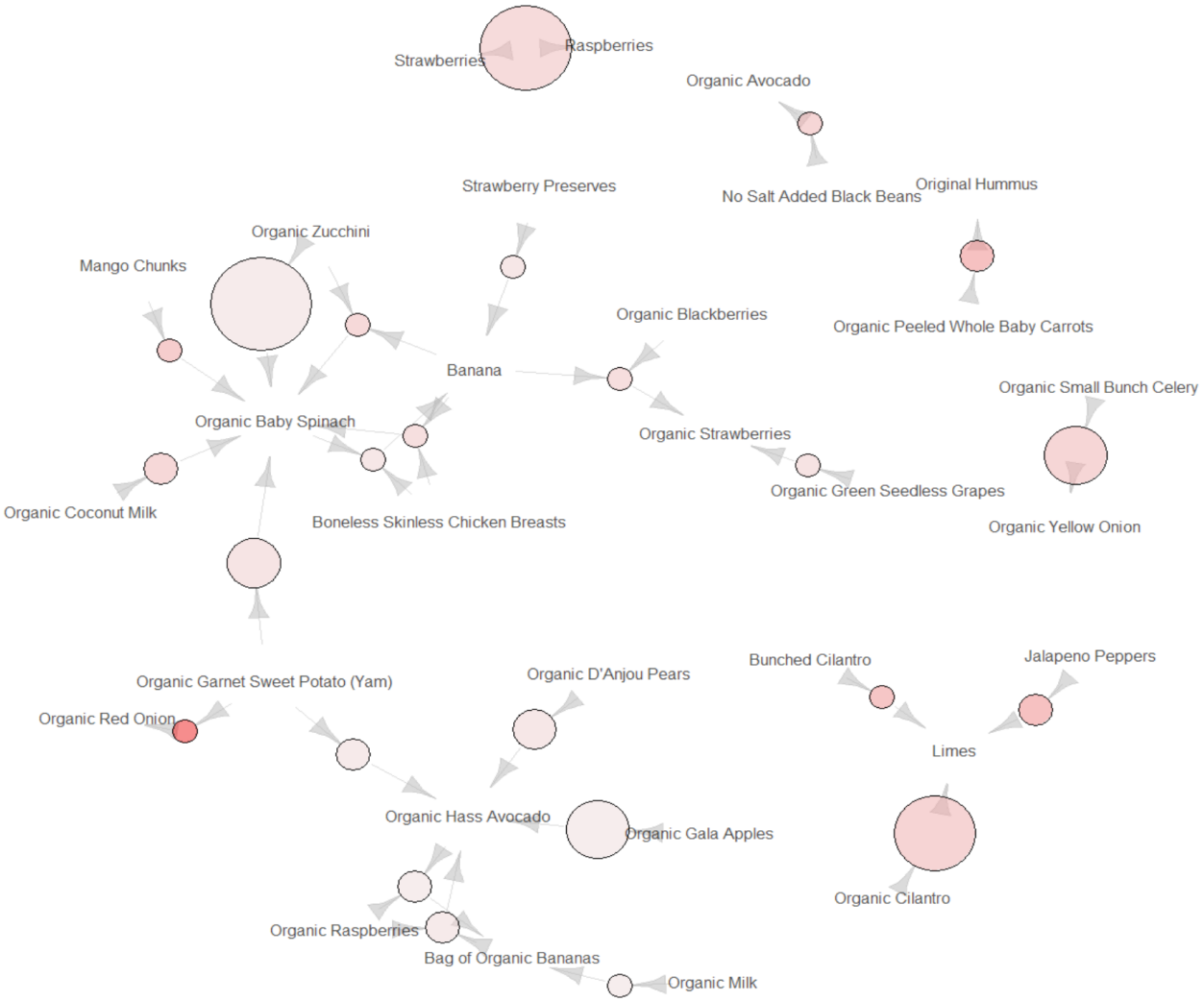
4 lentelė. Atrinktos įdomios produktų susietumo taisyklės

Taisyklės	***support***	***confidence***	***lift***	***conviction***
{Organic Garnet Sweet Potato (Yam)} => {Organic Red Onion}	0.003	0.25	13.7	1.3
{Jalapeno Peppers} => {Limes}	0.004	0.36	9.6	1.5
{Organic Peeled Whole Baby Carrots} => {Original Hummus}	0.004	0.25	9.6	1.3
{Bunched Cilantro} => {Limes}	0.003	0.35	9.2	1.5
{Mango Chunks} => {Organic Baby Spinach}	0.003	0.58	8.6	2.2
{No Salt Added Black Beans} => {Organic Avocado}	0.003	0.39	7.3	1.5
{Raspberries} => {Strawberries}	0.007	0.31	6.7	1.4
{Organic Garnet Sweet Potato (Yam)} => {Organic Baby Spinach}	0.005	0.36	5.3	1.5
{Strawberry Preserves} => {Banana}	0.003	0.78	4.9	3.8
{Boneless Skinless Chicken Breasts,Organic Baby Spinach} => {Banana}	0.003	0.78	4.9	3.8
{Unsweetened Original Almond Breeze Almond Milk} => {Banana}	0.005	0.44	2.8	1.5
{Green Beans} => {Banana}	0.004	0.41	2.6	1.4

Taip pat yra stiprių pavienių taisyklių, kurios padeda atpažinti tokius ryšius kaip morkos kartu su humusu, juodosios pupos kartu su avokadu ar salieras kartu su svogūnu. Arba kitas atvejis: trys taisyklės, kurios veda į žaliąją citriną. Taigi, šios mažiau akivaizdžios taisyklės gali būti identifikuotos ir panaudojamos pasiūlymams.

Apibendrinant pirkinių krepšelio analizę, sudaryti susietumo taisykles prekių kategorijoms gali būti patogu organizuojant prekių išdėstymą, pavyzdžiui šviežių vaisių ir jogurto skyriai netoliese (fizinės parduotuvės atveju). Žinant stiprius ryšius tarp prekių kategorijų, galima teikti klientui pasiūlymus nusipirkti ką nors iš konkrečios kategorijos. Internetinės parduotuvės atveju: „klientai, pirkę panašias prekes taip pat pirko daržovių“ arba „peržiūrėti X kategoriją“. Toks pasiūlymas yra ne toks konkretus, tačiau mažesnė tikimybė, kad klientui bus parodytas jam neaktualus produktas, nes jis galės pats pasirinkti kategorijoje.

Analizuojant susietumo taisykles produktų lygyje labiau tiktų tokie pasiūlymai kaip „kartu rekomenduojame“ arba „susijusios prekės“. Stiprų ryšį turinčioms prekėms taip pat gali būti pateikiami grupiniai (angl. *bundle*) pasiūlymai, pavyzdžiui braškės-avietės ar raudonas-baltas vynai.



17 pav. Produktų tinklas iš 24 susietumo taisyklių

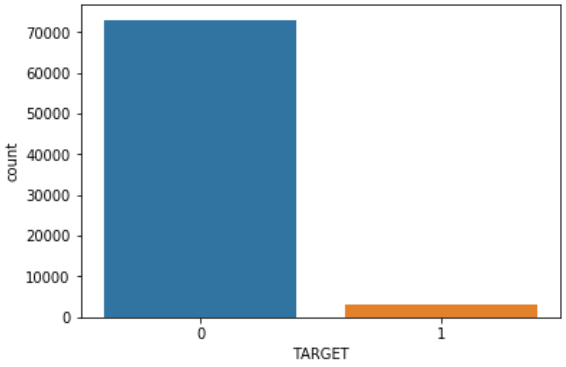
3.2. Vartotojų pasitenkinimo tyrimas

3.2.1. Žvalgomoji analizė

Vartotojų pasitenkinimas daugumai kompanijų yra sėkmės rodiklis. Nepatenkinti klientai gali būti greitai prarandami apie tai net nepranešę. Amerikos bankas Santander surengė Kaggle varžybas su tikslu identifikuoti nepatenkintus klientus. Tai leistų bankui geriau pasirūpinti klientais, kurie nėra patenkinti paslaugomis. Šios varžybos vyko prieš 2 metus, pritraukė 5123 komandas, o prizinis fondas pirmoms trims vietoms buvo 60000 JAV dol.

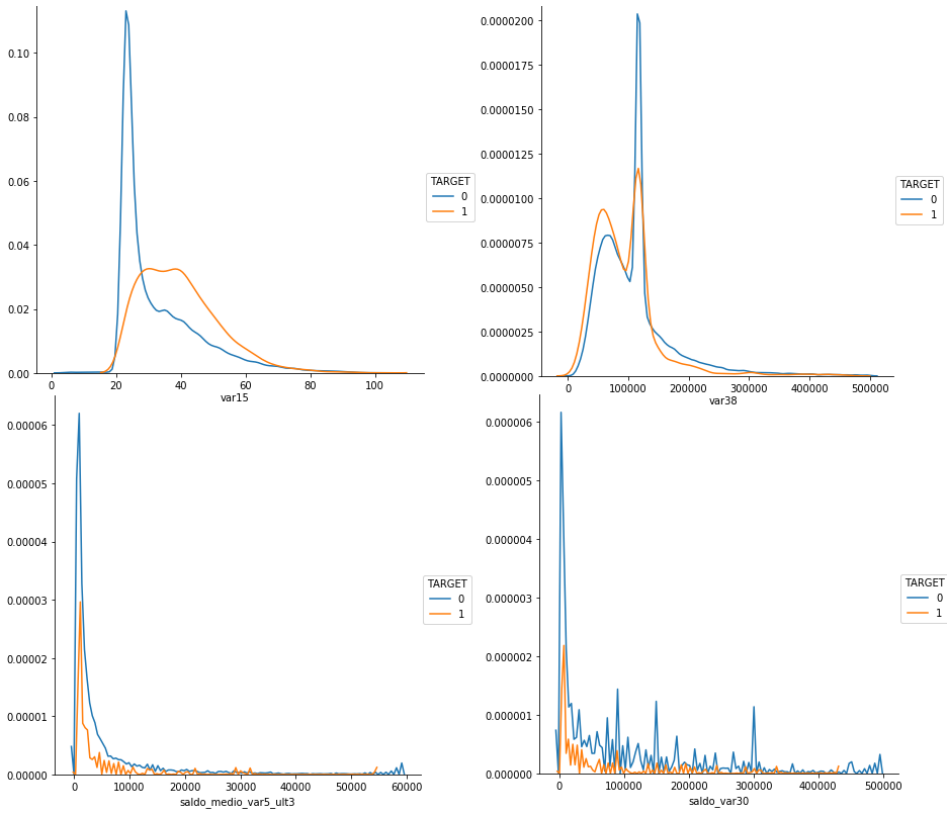
Dėl konfidencialumo, pateikti klientų duomenys yra anoniminiai. Nors kintamųjų vardai yra pervadinti, tačiau Kaggle diskusijose kai kuriuos kintamuosius varžybų dalyviai identifikavo. Pateikti du duomenų failai: apmokymo imtis (kurioje žinomas vartotojų pasitenkinimas) ir testavimo imtis (kurios vartotojų pasitenkinimas nėra viešai prieinamas). Norint dalyvauti varžybose reikia apmokyti modelį su apmokymo imtimi ir prognozuoti vartotojų pasitenkinimo tikimybes testavimo imčiai. Tada prognozės yra pateikiamos Kaggle, o Kaggle jas įvertina su *AUC* klasifikavimo matu (pagal viešai neprieinamas tikrąsias vartotojų pasitenkinimo reikšmes). Pagal tai sudaroma dalyvių rezultatų lentelė, o varžyboms pasibaigus laimi trys geriausios komandos. Kadangi šio darbo tikslas yra ne dalyvauti pačiose varžybose, o pritaikyti šiuolaikinius matematinius metodus vartotojų pasitenkinimo prognozavimui, todėl **visi skaičiavimai bus atliekami tik su apmokymo imtimi**.

Duomenyse iš viso yra 370 kintamųjų ir 76020 stebinių. Tai yra pakankamai didelis duomenų rinkinys mašininio mokymosi algoritmų taikymui. Santander duomenyse visi kintamieji yra skaitiniai ir nėra tuščių reikšmių. Tai palengvina sprendžiamą uždavinį, nes kai kurie metodai nepriima kategorinių kintamųjų ar tuščių reikšmių. Duomenyse yra 369 kintamieji (X) ir vartotojų pasitenkinimo kintamasis (Y), kuris įgyja reikšmes 0 (patenkinti vartotojai) ir 1 (nepatenkinti). Atsako kintamojo klasės yra nesubalansuotos (18 pav.): 0 priklauso daugumos klasei (96,04% atvejų), o 1 priklauso mažumos klasei (3,96% atvejų). Klasių disbalansas yra gana didelis, nes vienam nepatenkinto vartotojo atvejui tenka apytiksliai 24 patenkintų vartotojų atvejai.



18 pav. Vartotojų pasitenkinimo klasių pasikartojimo skaičius

Nors kintamieji yra anoniminiai ir jų prasmė nėra žinoma, tačiau yra aktualu paanalizuoti kuo skiriasi stebiniai pagal vartotojų pasitenkinimą. Palyginimui grafike atvaizduojami tankiai (angl. *kernel density estimation*) pagal kintamųjų klases (19 pav.). Šiame grafike atvaizduoti vieni svarbiausių kintamųjų prognozuojant vartotojų pasitenkinimą, kuriuos įvardino Kaggle dalyviai. Nesunku pastebėti, kad kintamojo *var15* tankis reikšmingai skiriasi klasėse. Likusių trijų kintamųjų tankiai taip pat turi skirtumų, tačiau mažiau akivaizdžių.



19 pav. Kintamųjų tankiai pagal vartotojų pasitenkinimą

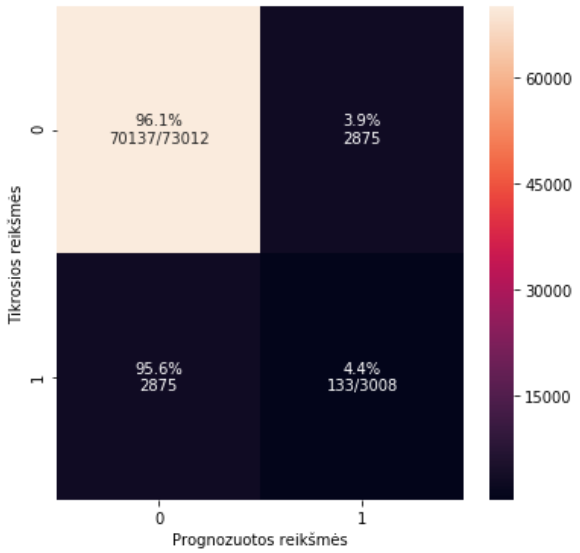
Toliau aktualu panagrinėti šių kintamųjų aprašomąją statistiką (5 lentelė). Kaggle diskusijose *var15* kintamasis buvo identifikuotas kaip kliento amžius. Nesunku pastebėti, kad nepatenkintų klientų amžius prasideda nuo 23 metų, o tarp patenkintų vartotojų šis amžius yra lygus 25% kvartiliui. Galima daryti išvadą, kad patys jauniausi klientai (jaunesni nei 23 metų) yra patenkinti banko paslaugomis. To priežastis gali būti, kad jauni klientai naudojasi banko paslaugomis trumpesnį laiką nei vyresni klientai. Kita tikėtina priežastis, kad jie yra nauji vartotojai ir naudoja mažiau banko paslaugų.

Taip pat Kaggle diskusijose buvo spėliota *var38* kintamojo prasmė. Manoma, kad šis kintamasis yra arba paskolos dydis (angl. *mortgage*), arba kliento grynoji vertė (angl. *net worth*). Pagal aprašomąją kintamojo *var38* statistiką galima tik pastebėti, kad patenkintų vartotojų reikšmės yra kiek plačiau išsisklaidžiusios (mažesnis minimumas ir didesnis maksimumas nei nepatenkintų klientų). Tačiau taip gali būti ir atsitiktinai, nes patenkintų vartotojų yra ženkliai daugiau. Tokia pati išvada galioja ir *saldo_medio_var5_ult3* ir *saldo_var30* kintamiesiems.

5 lentelė. Kintamųjų aprašomoji statistika pagal klientų pasitenkinimą

	var15		***var38***		***saldo_medio_var5_ult3***		***saldo_var30***	
Pasitenkinimas	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**
Minimumas	5	23	$5{,}2 \cdot 10^3$	$1{,}1 \cdot 10^4$	-476	-173	$-4{,}9 \cdot 10^3$	$-2{,}9 \cdot 10^3$
25% kvartilis	23	30	$6{,}8 \cdot 10^4$	$5{,}7 \cdot 10^4$	0	0	0	0
Vidurkis	33	39	$1{,}2 \cdot 10^5$	$9{,}9 \cdot 10^4$	1 080	282	$1{,}4 \cdot 10^4$	$2{,}2 \cdot 10^3$
Mediana	27	38	$1{,}1 \cdot 10^5$	$8{,}6 \cdot 10^4$	3	0	3	0
75% kvartilis	39	47	$1{,}2 \cdot 10^5$	$1{,}2 \cdot 10^5$	85	3	$2{,}8 \cdot 10^2$	6
Maksimumas	105	102	$2{,}2 \cdot 10^7$	$4 \cdot 10^6$	544 365	108 250	$3{,}5 \cdot 10^6$	$5{,}1 \cdot 10^5$

Toliau bus tiriama ar nepatenkinti klientai turi ką nors bendro su išskirtimis. Tam naudojamas izoliavimo miško algoritmas, kuris nesunkiai aptinka išskirtis net ir didelių matavimų erdvėse. Pasinaudojus Kaggle diskusijų informacija buvo pastebėta, kad *var15* ir *var38* yra du svarbiausi kintamieji vartotojų pasitenkinimo prognozavime. Pirmiausiai izoliavimo miškas buvo pritaikytas tik šiems kintamiesiems. Apmokymui naudota visa duomenų imtis (76020 stebinių) algoritmui nurodant tikslią išskirčių dalį, kurią reikia aptikti (3,96% nepatenkintų vartotojų). Gautas išskirtis galima palyginti su tikrosiomis reikšmėmis naudojant sumaišymų matricą (20 pav**.)**.



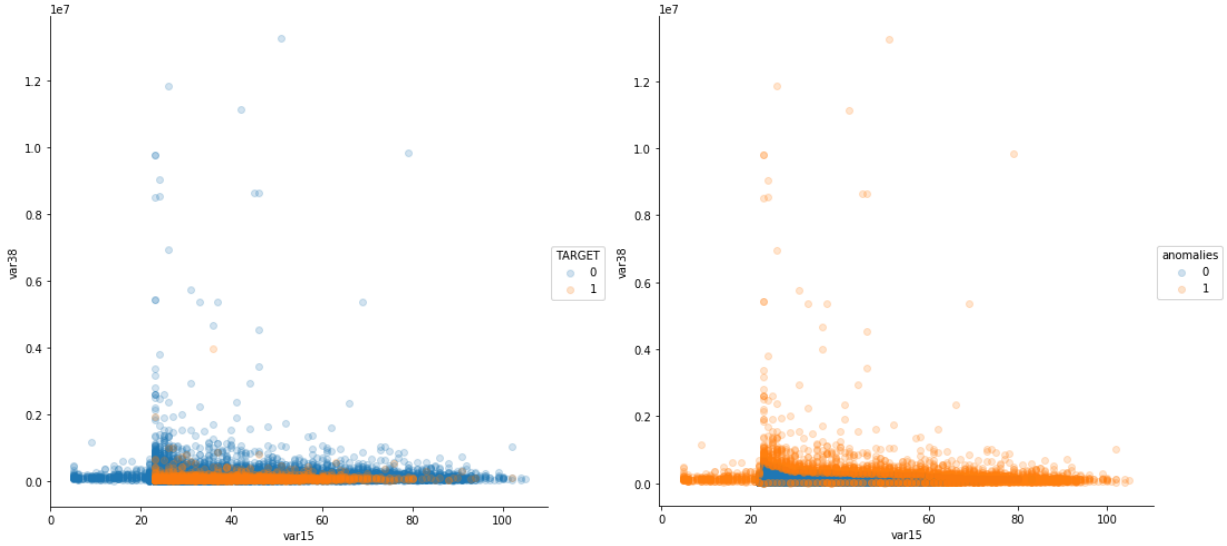
20 pav. Sumaišymų matrica: tikrosios reikšmės ir izoliavimo miško aptiktos išskirtys (kintamiesiems *var15* ir *var38*)

Nors pavyko teisingai aptikti 96,1% patenkintų klientų, tačiau dominančios klasės buvo aptikti vos 4,4% atvejų. Vadinasi, nepatenkintų klientų sutapimas tėra atsitiktinis. Išskirčių aptikimas su izoliavimo mišku taip pat buvo tirtas ir visam duomenų rinkiniui su 369 kintamųjų. Siekiant užtikrinti algoritmo tikslumą, buvo naudotas didelis 1000 medžių skaičius. Tačiau sumaišymų matricoje (pateiktoje 1 Priede) nepatenkintų vartotojų teisingai suklasifikuota tik 4,7%.

Toliau nubraižomos *var15* ir *var38* sklaidos diagramos (21 pav.) su pažymėtomis vartotojų pasitenkinimo grupėmis (kairėje diagrama su tikrosiomis, o dešinėje su prognozuotomis

reikšmėmis). Galima pastebėti, kad dvimačiu atveju izoliavimo miškas teisingai identifikavo išskirtis, tačiau tikrosios nepatenkintų vartotojų reikšmės net nėra tarp išskirčių.

Atlikus išskirčių ir nepatenkintų klientų palyginimą, galima apibendrinti, kad ši klientų grupė nepriklauso išskirčių aibei, nes tarp jų nebuvo aptikta reikšmingo sutapimo.



21 pav. Tikrosios vartotojų pasitenkinimo reikšmės (kairėje) ir prognozuotos su išskirčių metodu (dešinėje)

3.2.2. Duomenų paruošimas modeliavimui

Prieš atliekant vartotojų pasitenkinimo prognozavimą reikia tinkamai paruošti duomenis. Tam bus atliekama kintamųjų atranka ir klasių subalansavimas imties metodais.

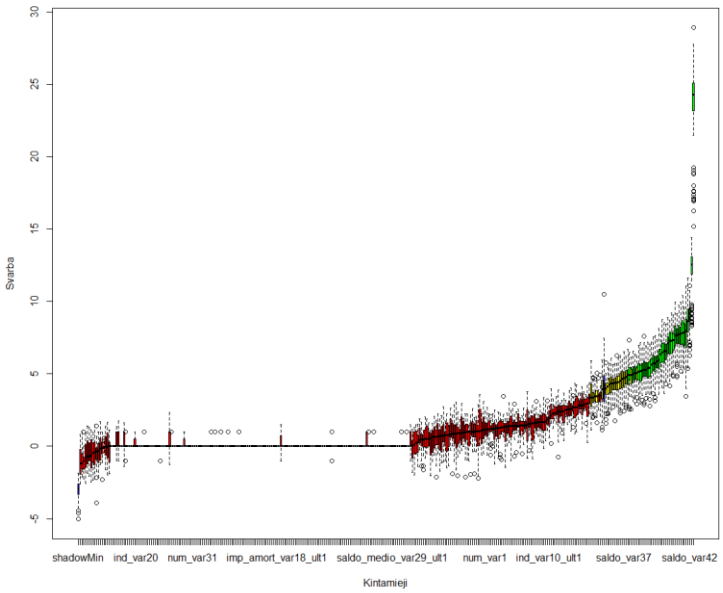
Kintamųjų atranka yra vienas iš žingsnių kurį atlieka komandos Kaggle varžybose. Dažnu atveju duomenyse yra per daug kintamųjų adekvačių modelių sudarymui. Nereikšmingi kintamieji modeliui neatneša jokios naudos, bet įneša triukšmo. Tokių kintamųjų atsisakius algoritmą galima apmokyti greičiau, o modelis tampa paprastesniu. Apmokymo greitis ypatingai aktualus, kai modeliuojami dideli duomenų rinkiniai arba apmokoma keletas algoritmų su dideliu parametrų kombinacijų skaičiumi. Taigi, dėl praktinių sumetimų verta surasti optimalų kintamųjų skaičių Santander duomenyse.

Kintamųjų atrankai atlikti duomenų rinkinys padalinamas į apmokymo (80%) ir testavimo (20%) imtis. Tuomet pritaikomas Boruta algoritmas. Duomenų rinkinys yra padvigubinamas, nes kiekvienam kintamajam yra sukuriama šešėlinė (sumaišyta) jo versija. Tuomet su *Random Forest* atliekamas klasifikavimas ir gaunami visų kintamųjų svarbos įverčiai. Šešėlinio kintamojo svarba didesnė už 0 gali būti tik dėl atsitiktinių svyravimų. Todėl šešėlinių kintamųjų svarba padeda nuspręsti, kurie kintamieji iš tikrųjų yra svarbūs.

Kintamųjų svarbos matas svyruoja dėl tokių veiksnių kaip *Random Forest* stochastinės savybės bei nereikšmingų ir šešėlinių kintamųjų įtraukimo į duomenų rinkinį. Dėl to yra būtina kartoti iš naujo kintamųjų sumaišymo procedūrą pakankamą iteracijų skaičių, kad būtų pasiekti

statistiškai patikimi rezultatai. Šiame darbe siekiant taupyti skaičiavimo laiką buvo pasirinkta 100 iteracijų.

Gauti tokie Borutos algoritmo rezultatai po 100 iteracijų: 35 kintamieji priskirti svarbiems, 19 kintamųjų nepriskirti (pritrūko iteracijų), o likę 315 – atmesti kaip nereikšmingi (22 pav.). Gauti du svarbiausi kintamieji yra *var15* ir *saldo_var30.* Toks drastiškas kintamųjų atmetimas (80-90%) sumažintų modelio tikslumą, todėl reikia alternatyvaus būdo kintamųjų skaičiui pasirinkti.



22 pav. Kintamųjų svarba pagal Boruta algoritmą

Borutos algoritmas R programoje ne tik pateikia atliktus sprendimus kintamųjų atžvilgiu, bet ir kiekvieno kintamojo svarbos aprašomąją statistiką (iš 100 iteracijų): vidurkį, medianą, minimumą ir maksimumą. Kintamųjų atranką galima atlikti ribojant kintamuosius pagal prieš tai minėtą aprašomąją statistiką.

Kintamųjų atrankos tikslas yra rasti mažesnę kintamųjų imtį, su kuria atliekama prognozė būtų neprastesnė nei su visu duomenų rinkiniu. Atliekant Santander banko kintamųjų atranką ištirti penki skirtingi atvejai (6 lentelė). Kintamųjų atrankos gerumo kriterijumi buvo panaudotas paprastas sprendimų medžio algoritmas prognozuojant testavimo imtį. Gauti prognozavimo rezultatai yra vertinami lyginant klasifikavimo tikslumo matus.

6 lentelė. Kintamųjų atrankos rezultatai prognozuojant testavimo imtį

Kintamųjų atrankos filtras	**Kintamųjų sk.**	***Accuracy***	***Specificity***	***Precision***	***Recall***	***F1***	***Kappa***
Be filtro	369	0.767	0.708	0.985	0.770	0.864	0.135
Svarbos vidurkis >= 0	318	0.768	0.708	0.985	0.770	0.864	0.135
Svarbos min reikšmė >= 0	253	0.779	0.723	0.986	0.782	0.872	0.148
Svarbos max reikšmė > 0	193	0.726	0.684	0.983	0.728	0.836	0.102
Svarbos mediana > 0	156	0.751	0.611	0.979	0.756	0.853	0.100

Sumažinus kintamųjų skaičių iki 318 tikslumas išlieka apytiksliai toks pat kaip ir pradinio duomenų rinkinio. Apribojant kintamųjų skaičių iki 253 gautas didžiausias tikslumas iš visų analizuotų atvejų. Tai yra net ~31% kintamųjų skaičiaus sumažinimas, o prognozavimo tikslumas nuo to tik išaugo. Tačiau taip pat pastebėta, kad klasifikavimo tikslumas yra prastesnis nei su pradiniu duomenų rinkiniu, kai kintamųjų skaičius yra mažesnis nei 200.

Remiantis kintamųjų atrankos analizės rezultatais, tolimesnėje darbo eigoje bus naudojami tik atrinkti 253 kintamieji.

Kitas svarbus duomenų paruošimo žingsnis yra vartotojų pasitenkinimo klasių subalansavimas. Tam bus naudojami trijų tipų imčių metodai: *downsampling, upsampling* ir hibdriniai. Įprastai imčių metodais nesubalansuotų klasių stebinių skaičius yra suvienodinamas po maždaug 50%. Tačiau toks dirbtinis klasių dažnumo suvienodinimas greičiausiai bus nenaudingas. *Downsampling* atveju, klasių suvienodinimas reiškia, kad duomenų rinkinys yra smarkiai sumažinamas daugumos klasės sąskaita. Todėl daugumos klasė yra prasčiau apmokoma nei prieš tai. Naudojant *upsampling* metodą klasių dažnumui suvienodinti yra atkartojami mažumos klasės stebiniai. Šiuo atveju didėja persimokymo problema, nes yra daug pasikartojančių mažumos klasės stebinių. Hibridiniai metodai, tokie kaip *SMOTE* ir *ROSE* naudoja *downsampling* metodą daugumos klasės imties sumažinimui ir tuo pačiu sintetina naujus mažumos klasės stebinius, kad šie nebūtų identiški jau esantiems.

Kadangi pradinėje apmokymo imtyje mažumos klasė sudaro 4% visų stebinių, tai imčių ėmimo metodais buvo sudaryti 8 atvejai, kai ši dalis padidinama tam tikrais atvejais iki 50% (7 lentelė). Gautų duomenų imčių tikslumas įvertintas apmokant sprendimų medžius su sluoksniuotu 10 dalių kryžminiu patikrinimu ir prognozuojant testavimo imtį. Lyginant prognozes, labiausiai atsižvelgiama į *F1* ir *kappa* matus.

Santander duomenų rinkinys yra pakankamai didelis, todėl *downsampling* metodas tinka labiau nei *upsampling*. Šiame darbe buvo ištirti keturi *downsampling* atvejai, kai mažumos klasės dalis yra 8%, 14%, 24% ir 50%. Tikslumas labiausiai sumažėja, kai mažumos klasės dalis pasiekia 50%, tokiu atveju duomenų rinkinio dydis susitraukia ~92%. Didžiausias tikslumas su *downsampling* pasiekiamas kai mažumos klasės dalis yra lygi 14%. Tokiu atveju taip pat yra atsisakoma nemažos duomenų dalies: ~72% eilučių.

Upsampling metodu buvo ištirti du atvejai: kai mažumos klasės dalis padidinama iki 8% ir 50%. Klasių dažnumo suvienodinimas šiuo metodu reiškia, kad duomenų rinkinio dydis išaugo du kartus, o tuo pačiu išaugo ir modelių apmokymo trukmė. Tačiau klasifikavimas su didesniu duomenų kiekiu pasižymėjo prastesniais rezultatais nei su pradine apmokymo imtimi. Antra vertus,

upsampling su 8% mažumos klasės dalimi pagerino pradinės apmokymo imties rezultatą. Kadangi šis metodas padidina ir taip nemažą duomenų rinkinį, tai į tolimesnę analizę įtrauktas nebus.

7 lentelė. Imčių metodų palyginimas

Metodas	**Mažumos kl. dalis**	**Eilučių sk.**	***Accuracy***	***Specificity***	***Precision***	***Recall***	***F1***	***Kappa***
Apmokymo imtis	**4%**	**60 816**	**0.767**	**0.708**	**0.985**	**0.770**	**0.864**	**0.135**
Downsampling	8%	31 612	0.763	0.778	0.988	0.763	0.861	0.147
Downsampling	14%	17 010	0.788	0.751	0.987	0.790	0.878	0.162
Downsampling	24%	10 196	0.748	0.733	0.986	0.749	0.851	0.127
Downsampling	50%	4 818	0.676	0.805	0.988	0.671	0.799	0.100
Upsampling	8%	63 225	0.779	0.731	0.986	0.780	0.871	0.148
Upsampling	50%	116 814	0.755	0.726	0.985	0.756	0.856	0.130
SMOTE	25%	17 842	0.862	0.546	0.979	0.875	0.924	0.188
ROSE	50%	60 816	0.755	0.726	0.985	0.756	0.855	0.129

Taip pat ištirti du hibridinių metodų atvejai: *SMOTE* su 25% mažumos klasės dalimi ir *ROSE* su 50%. Pastarasis metodas turėdamas vienodą eilučių skaičių kaip ir pradinė apmokymo imtis pasižymėjo prastesniais rezultatais. Su *SMOTE* metodu buvo pasiektas didžiausias tikslumas iš visų nagrinėtų atvejų (vertinant pagal *F1* ir *kappa* matus).

Apibendrinant, buvo ištirti 8 klasių disbalanso metodai, kurių naudingumas buvo įvertintas su paprastu sprendimų medžių modeliu prognozuojant testavimo imtį. Vertinant pagal tikslumo matus ir optimalų duomenų rinkinio dydį buvo pastebėta, kad *SMOTE* ir *downsampling* (su 14% mažumos dalimi) pasirodė itin gerai. Dėl to, šios dvi duomenų imtys bus išbandytos vartotojų pasitenkinimo prognozavimo žingsnyje.

3.2.3. Vartotojų pasitenkinimo prognozavimas

Atlikus duomenų paruošimą liko įvertinti kurią imtį tikslingiausia naudoti modeliavimui. Po to bus apmokomi algoritmai atliekant modelių parametrų derinimą. Taip pat svarbu įvertinti modelių galimybes prognozuoti vartotojų pasitenkinimą bei nustatyti, kurie kintamieji turi didžiausią įtaką.

Pirmame žingsnyje tiriama kaip algoritmai veikia su skirtingomis imtimis. Algoritmai apmokyti su apmokymo, *downsampling* ir *SMOTE* imtimis. Paprastumo dėlei, kryžminis patikrinimas ir parametrų derinimas nebuvo atliekami. *Random Forest, LightGBM ir XGBoost* algoritmai turi klasių svorio parametrą su kuriuo buvo padidintas mažumos klasės svoris (pagal konkrečios imties klasių santykį). *RGF* algoritmas tokio parametro neturi. Su šiais algoritmais buvo prognozuota testavimo imtis. Modelių rezultatai pateikti 8 lentelėje.

8 lentelė. Atrinktų geriausių imčių metodų palyginimas

Metodai		Priskirta dalis į mažumos klasę	*Accuracy*	*AUC*	*Precision**	*Recall**	*F1**	*Kappa*	Apmokymo trukmė (min)
Apmokymo imtis	*Random Forest*	28.3%	0.735	0.715	0.097	0.694	0.171	0.109	0.79
	LightGBM	16.9%	0.944	0.575	0.229	0.174	0.197	0.169	0.88
	RGF	0.0%	0.961	0.5	0	0	0	0	3.47
	XGBoost	22.4%	0.799	0.768	0.132	0.735	0.224	0.168	4.85
	Vidurkis	**16.9%**	**0.860**	**0.640**	**0.115**	**0.401**	**0.148**	**0.112**	**2.50**
Downsampling	*Random Forest*	33.6%	0.727	0.723	0.097	0.718	0.172	0.11	0.31
	LightGBM	23.4%	0.834	0.757	0.148	0.674	0.242	0.19	0.17
	RGF	4.8%	0.947	0.583	0.257	0.189	0.218	0.191	0.86
	XGBoost	28.6%	0.789	0.774	0.129	0.758	0.22	0.164	1.2
	Vidurkis	**22.6%**	**0.824**	**0.709**	**0.158**	**0.585**	**0.213**	**0.164**	**0.64**
SMOTE	*Random Forest*	35.2%	0.777	0.725	0.111	0.669	0.191	0.132	0.32
	LightGBM	28.4%	0.876	0.735	0.175	0.583	0.27	0.223	0.22
	RGF	20.3%	0.92	0.665	0.216	0.389	0.278	0.239	1.02
	XGBoost	30.6%	0.854	0.762	0.164	0.661	0.263	0.214	1.28
	Vidurkis	**28.6%**	**0.857**	**0.722**	**0.167**	**0.576**	**0.251**	**0.202**	**0.71**

** Pažymėti tikslumo matai skaičiuoti mažumos klasei.*

Rezultatų lentelėje galima pastebėti, kad net ir naudojant klasių subalansavimo parametrus, algoritmai yra linkę priskirti gerokai daugiau stebinių į mažumos klasę (testavimo imtyje mažumos klasę sudaro tik 3,9% stebinių). *RGF* algoritmas su apmokymo imtimi yra puiki iliustracija atvejo, kai algoritmas pasirinko lengviausią prognozavimo kelią – viską klasifikuoti į daugumos klasę. Tokiu atveju, *accuracy* rodo aukštą 0,961 tikslumą, nes tokia dalis teisingai priskirta prie patenkintų vartotojų. Kadangi *accuracy* neatsižvelgia į mažumos klasę, tai yra nenaudingas tikslumo matas nesubalansuotų klasių uždavinyje. *AUC* reikšmė lygi 0,5, o tai reiškia atsitiktinį spėliojimą. Kiti matai šiuo atveju grąžina 0 reikšmes.

Lyginant klasifikavimo rezultatus tarp trijų skirtingų imčių metodų, su *SMOTE* imtimi vidutinis tikslumas yra didžiausias. Tačiau su šia imtimi į mažumos klasę buvo priskirta didžiausia dalis stebinių. Šiame palyginime apmokymo imtis pasižymėjo mažiausiu tikslumu ir ilgiausia apmokymo trukme.

Apibendrinant, palyginus skirtingus algoritmus su trejomis imtimis, galima pastebėti, kad vartotojų prognozavimas yra nelengvas iššūkis. Šio tyrimo metu maksimali pasiekta *AUC* reikšmė 0,77 reiškia tik vidutinį klasifikatoriaus gebėjimą atskirti klases. Tokie tikslumo matai kaip *F1* (mažumos klasės) ir *kappa* rezultatuose svyruoja apie 0,2, o tai reiškia silpną klasifikavimo tikslumą. Taip pat verta pastebėti, kad *RGF* algoritmo rezultatai yra labai jautrūs nesubalansuotoms imtims, o taip pat *RGF* neturi parametro klasių svoriui balansuoti. Dėl to į kitus skaičiavimus *RGF* nėra įtraukiamas. Kadangi nebuvo atliekama kryžminio patikrinimo ir parametrų derinimo,

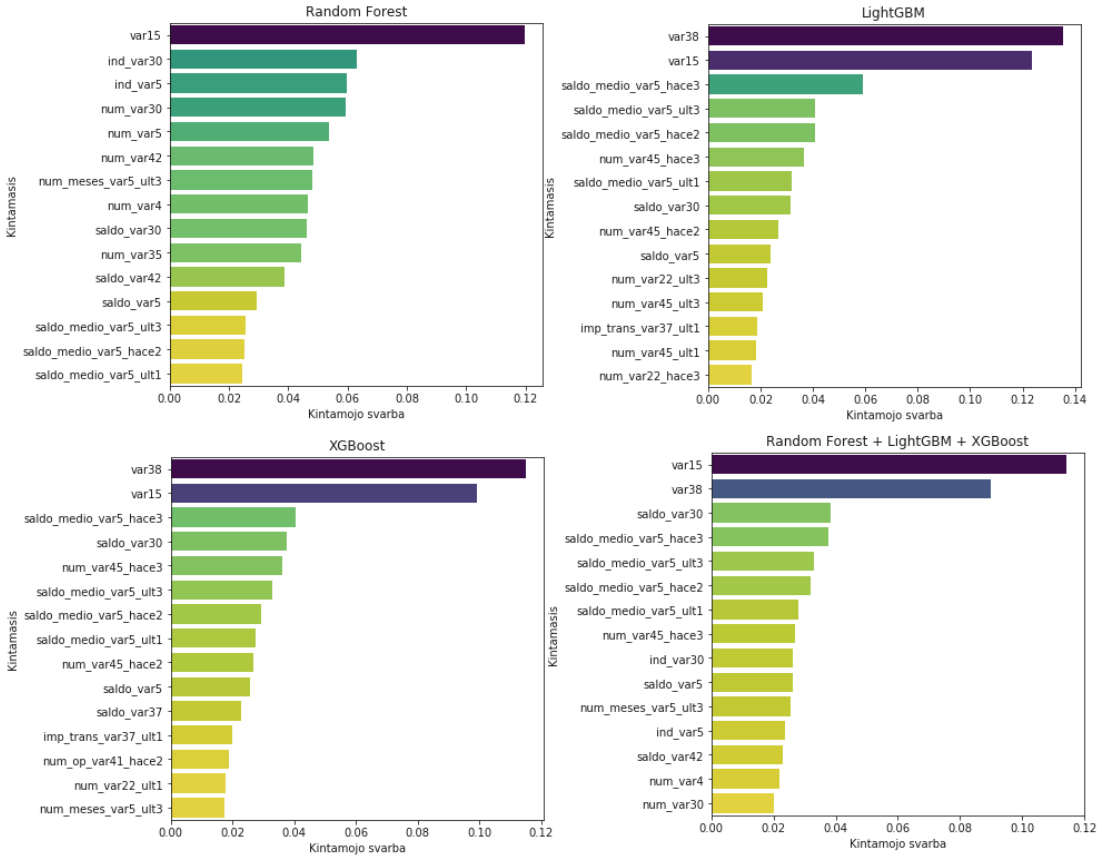
algoritmai buvo apmokomi pakankamai greitai – iki 5 min. Vartotojų pasitenkinimo prognozavimui toliau bus naudojama *SMOTE* imtis, nes ši imtis buvo geriausiai įvertinta ir yra mažesnė nei apmokymo imtis.

Toliau algoritmai apmokomi tokia tvarka:

1) Kiekvienam algoritmui yra pateikiamas parametrų sąrašas derinimui. Parametrai parinkti tokie, kurie didina modelio tikslumą ir tuo pačiu apsaugo nuo persimokymo:
    - *Random Forest* parinktas medžių skaičius (300, 500, 800), maksimalus požymių skaičius (8, 15, 30), maksimalus gylis (8, 10);
    - *LightGBM* parinktas medžių skaičius (500, 800, 1000), mokymosi tempas (0,01), maksimalus gylis (6, 8), kintamųjų dalis (0,7), stebinių dalis (0,6);
    - *XGBoost* parinktas medžių skaičius (600, 800), mokymosi tempas (0,01), maksimalus gylis (6, 8, 10), minimalus atžalos svoris (4, 7), kintamųjų dalis (0,9), stebinių dalis (0,9);
2) Algoritmams atliekama tinklelio paieška su 3 dalių sluoksniuotu kryžminiu patikrinimu. Nustatytas tinklelio paieškos tikslumo vertinimo matas – svorinis *F1*;
3) Visi trys algoritmai apmokomi su tinklelio paieška atitinkamai nustatytais geriausiais parametrais;
4) Su kiekvienu modeliu klasifikuojamas vartotojų pasitenkinimas testavimo imtyje. Gauta prognozė palyginama su žinomomis testavimo imties vartotojų pasitenkinimo reikšmėmis.

Su parametrų derinimu ir kryžminiu patikrinimu algoritmų apmokymo trukmė gerokai išaugo (iki valandos). Taip yra dėl to, nes tinklelio paieška tikrina kiekvieną parametrų kombinaciją ir šis procesas kartojamas 3 kartus (nes naudojamas 3 dalių kryžminis patikrinimas). Apmokymas trukmė galėjo būti dar kelis kartus ilgesnė, jeigu nebūtų atlikti kintamųjų atrankos ir duomenų imties ėmimo žingsniai.

Apmokius modelius, sprendimų medžių algoritmai įvertina kintamųjų svarbą atsako kintamojo prognozavimui. 23 pav. atvaizduotos modelių svarbos: *Random Forest*, *LightGBM*, *XGBoost* bei jų vidurkis. *LightGBM* ir *XGBoost* kintamųjų svarba yra labai panaši. Tai galima paaiškinti tuo, kad patys algoritmai yra labai gimininги, abu naudoja gradientinį *boosting* metodą, bet skirtingai augina medžius. Šie algoritmai išskyrė *var38* ir *var15* kaip svarbiausius kintamuosius, kurių jungtinė svarba vartotojų pasitenkinimo prognozavimui viršija 20%. *Bagging* tipo algoritmas *Random Forest* svarbiausią kintamąjį nustatė klientų amžių (*var15*), o *var38* net nepateko tarp svarbiausių 15. Patikimesnius rezultatus galima gauti tiesiog suskaičiuojant visų trijų modelių kintamųjų svarbos vidurkius. Apibendrintuose rezultatuose klientų amžius yra svarbiausias, *var38* liko antroje vietoje.



23 pav. Galutinių modelių kintamųjų svarba

Iš verslo perspektyvos žiūrint, buvo nustatyti du kintamieji kurie daro reikšmingą įtaką vartotojų pasitenkinimo prognozavimui. Tai gali būti panaudojama kaip papildoma informacija gerinant paslaugų kokybę. Pavyzdžiui, galbūt reikėtų kreipti daugiau dėmesio į tam tikras amžiaus arba kintamojo *var38* reikšmių grupes. Verslui taip pat aktualu kokiu tikslumu prognozuoja sudaryti modeliai. Galutiniai modelių prognozavimo rezultatai pateikti 9 lentelėje.

9 lentelė. Galutinių modelių rezultatai

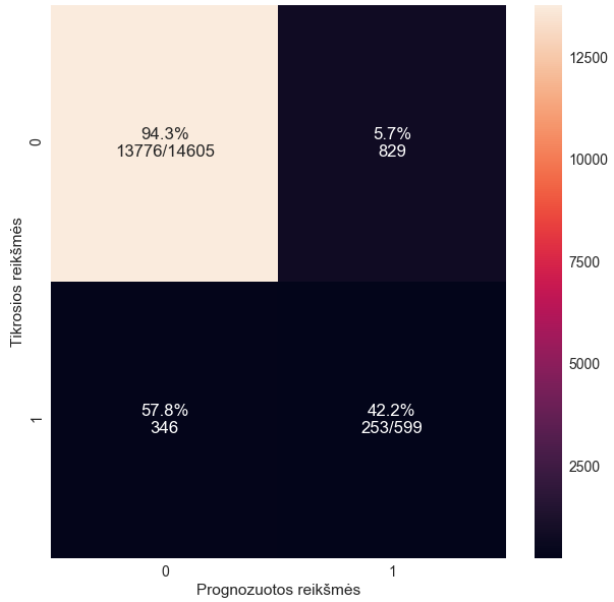
Modelis	**Priskirta dalis į mažumos klasę**	***Accuracy***	***AUC***	***Precision****	***Recall****	***F1****	***Kappa***
Random Forest	8.07%	0.914	0.686	0.214	0.437	0.287	0.247
LightGBM	6.92%	0.924	0.68	0.237	0.416	0.302	0.265
XGBoost	7.17%	0.922	0.682	0.231	0.421	0.298	0.261
Balsavimas	7.04%	0.924	0.686	0.239	0.427	0.307	0.27
Stacking	6.69%	0.926	0.676	0.239	0.406	0.301	0.264

** Pažymėti tikslumo matai skaičiuoti mažumos klasei.*

Vartotojų pasitenkinimui prognozuoti buvo sudaryti *Random Forest*, *LightGBM* ir *XGBoost* modeliai. Šie algoritmai po atlikto parametrų derinimo tiksliau atpažino mažumos klasės dalį testavimo imtyje. Lyginant su ankstesniais bandymais, *F1* ir *kappa* tikslumo matai išaugo, tačiau jie vis dar nesiekia net ir vidutinio klasifikavimo standartų. *AUC* reikšmės yra mažesnės nei 0,7, tai reiškia ganėtinai silpną klasių atskyrimą.

Praktikoje, ypatingai Kaggle varžybose, tikslumo maksimizavimui iš kelių modelių yra sudaromas „ensemble“ modelis. Dėl to šiame darbe papildomai sudaryti balsavimo ir *stacking* modeliai. Balsavimo modelis remiasi daugumos balsų principu – trijų modelių atveju jei du modeliai klientui priskiria 1 reikšmę, o likęs modelis priskiria 0, tai modeliai nubalsuoja, kad klientas yra nepatenkintas. Balsavimo modelis pranoko visus tris modelius pagal *F1* ir *kappa* tikslumo matus. Tiesa, tikslumo pagerinimas yra tik tūkstantosiomis dalimis, tačiau remiantis kelių skirtingų modelių prognozėmis gaunami patikimesni rezultatai. *Stacking* metodas buvo taikomas apmokant naują modelį su *Random Forest, LightGBM* ir *XGBoost* prognozuotomis reikšmėmis. Prognozės buvo apmokytos su *XGBoost* modeliu, tačiau *stacking* modelis nepagerino trijų modelių rezultatų ir nusileido *LigthGBM* modeliui. Praktikoje *stacking* modeliuose yra naudojamos dešimtys modelių, ir tai gali būti viena iš priežasčių kodėl modelis nebuvo naudingas šiuo atveju.

Daugumos balsų modelis buvo naudingiausias vartotojų pasitenkinimo prognozavime, todėl liko apžvelgti jo stiprybes ir silpnybes naudojant sumaišymų matricą (24 pav.)



24 pav. Sumaišymų matrica: tikrosios testavimo imties reikšmės ir daugumos balsų reikšmės

Iš sumaišymų matricos galima pamatyti, kad daugumos klasė (patenkinti vartotojai) yra gerai prognozuojama – 94,3% atvejų teisingai suklasifikuoti. Tačiau nepatenkintų vartotojų teisingai suklasifikuoti nepavyko nei pusės – 42,2%. Taigi, jeigu bankas imtųsi veiksmų, tuomet turėtų galimybę išsaugoti maždaug tokią dalį norinčių atsisakyti paslaugas vartotojų. Tačiau taip pat verta atkreipti dėmesį, kad 829 klientai buvo klaidingai modelio identifikuoti kaip nepatenkinti. Šiuo atveju, banko iniciatyva būtų tik papildomi kaštai, nes klientas ir taip patenkintas paslaugomis.

3.2.4. Rezultatų aptarimas

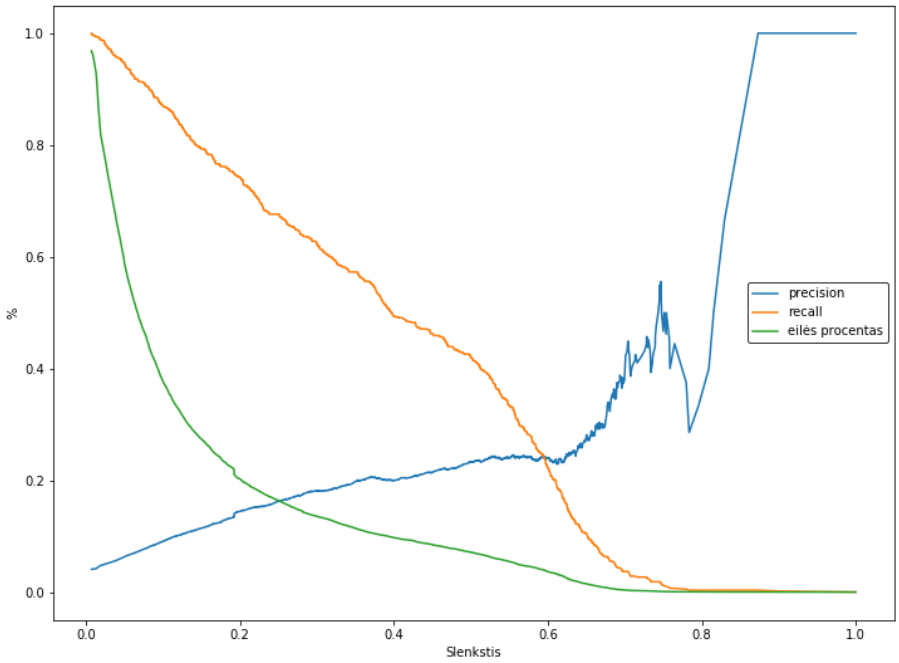
Ankstesniuose žingsniuose buvo atliekami duomenų paruošimo žingsniai bei pritaikomi mašininio mokymosi algoritmai siekiant sukurti praktinę vertę. Ši vertė įgalina atlikti sprendimus geriau nei tai buvo daroma turint primityvesnį modelį ar išvis nenaudojant modelio. Siekiant maksimizuoti verslui kuriamą vertę, klasifikavimo atveju geriausio modelio rezultatus galima iliustruoti pavaizduojant modelio *precision* ir *recall* kreives bei pasirenkant atitinkamą slenkstį. Slenksčio metodas yra populiarus dėl savo paprastumo ir lengvo pritaikymo. Sudėtingiausias klausimas yra kokį slenkstį pasirinkti, tačiau tai yra verslo sprendimas.

Renkantis slenkstį įprastai remiamasi trimis faktoriais:

1) Eilės procentas: į kokią dalį atvejų verta sureaguoti? Tai priklausys tiek nuo individualaus atvejo kaštų, tiek ir nuo bendrų pajėgumų. Pavyzdžiui, nepatenkintų vartotojų dalies vertinimas yra apribotas prie to dirbančios komandos ištekliais. Tačiau, jei yra galimybė visiems atvejams paprasčiausiai išsiųsti elektroninius laiškus, tuomet šie kaštai yra artimi nuliui, ir galima į eilę įtraukti daug atvejų;
2) *Precision* atsako į klausimą: koks klaidingai priskirto atvejo nuostolis? Tiriant vartotojų pasitenkinimą tai būtų iššvaistytas vertinančio žmogaus laikas. Kitu atveju, tai gali būti piniginis nuostolis. Pavyzdžiui, jeigu per klaida yra pasiūloma 10% nuolaida vartotojui, kuris yra klaidingai priskirtas prie nepatenkintųjų paslaugomis, tokiu atveju atitinkamai bus gauta mažiau pajamų (jei šis klientas būtų ir taip pirkęs paslaugas);
3) *Recall*: kokio dydžio nuostolis yra nepriskyrimas į eilę atvejo, kuris turėjo būti priskirtas? Perteikiant tai pagal vartotojų pasitenkinimo prognozavimą: prarasta galimybė atlikti veiksmus kliento išsaugojimui.

Siekiant rasti kompromisą tarp šių trijų faktorių yra nubraižomas grafikas (25 pav.). Šiame grafike x ašyje yra pavaizduotas klientų pasitenkinimo prognozuotos tikimybės slenkstis: artimi 0 – patenkinti klientai, o artimi 1 – prarasti klientai.

Pasirinkus 0,37 slenkstį, bus tikrinamas apytiksliai 12% klientų pasitenkinimas. Tiesa, šiuo atveju, eilės procentas kinta minimaliai lyginant jį su skirtingomis slenksčio reikšmėmis. Prie pasirinkto slenksčio *precision* reikšmė siekia apie 56%. Tai reiškia, kad tinkamu laiku nesiėmus veiksmų, iš tikrinamų 12% klientų tokia dalis klientų tikrai pasitrauks. *Recall* reikšmė yra apie 20%, vadinasi modeliui pavyks atpažinti tik penktadalį pasitraukti planuojančių klientų.



25 pav. *precision* ir *recall* kreivės

Žinant klientų pasitraukimo tikimybes galima pagal jas naudoti skirtingus slenksčius. Šie skirtingi slenksčiai leistų sudaryti skirtingas rizikos grupes, o pagal jas būtų galima efektyviau paskirstyti ribotus finansinius bei žmogiškuosius išteklius klientų išsaugojimui. Aukštesnės rizikos klientams būtų skiriamas didesnis dėmesys bei teikiami patrauklesni pasiūlymai, o žemesnės rizikos klientams teikiami pasiūlymai būtų nuosaikesni taupant išteklius.

Išvados

Šiame darbe literatūros apžvalgoje buvo aptarta, kad pirkinių krepšelio analizė atspindi tikrą vartotojų elgseną bei atskleidžia svarbius paslėptus pirkėjų įpročius. Yra pavyzdžių, kai susietumo taisyklės buvo sėkmingai pritaikytos rekomendavimo sistemoms, tačiau jos nėra itin populiarios. Mokslinėje literatūroje neretai yra išreiškiama kritika susietumo taisyklių naudojimui. Nors susietumo taisyklės yra įprastinis pirkinių krepšelio analizės įrankis, tačiau tipiškai reikia sugeneruoti didelį susietumo taisyklių skaičių, kurį išanalizuoti sudėtinga. Taip pat teigiama, kad įprastai naudojami taisyklių įdomumo matai *confidence* ir *lift* neatsižvelgia į tikimybines savybes, dėl to literatūroje galima rasti dešimtis alternatyvių matų. Siekiant kovoti su susietumo taisyklių trūkumais, mokslinėje literatūroje pateikiami išradingi pasiūlymai, pavyzdžiui išplėsti jas minimalios aprėpties medžiais. Šiuo metodu sudarytas tinklas sumažina susietumo taisyklių paieškos erdvę bei leidžia jas lengviau interpretuoti.

Vartotojų pasitenkinimo prognozavimas literatūroje dažniausiai sutinkamas tiriant telekomunikacijų srities duomenis. Apžvelgtoje mokslinėje literatūroje keletą kartų buvo akcentuota, kad naujo vartotojo pritraukimas yra brangesnis nei esamo išsaugojimas ir vienas pagrindinių santykių su klientais valdymo uždavinių yra išvengti klientų praradimo. Tipiškai, klientų praradimas yra retas atvejis lyginant su tais kurie naudojasi paslaugomis. Todėl klientų pasitenkinimo prognozavimas yra nesubalansuotų klasių uždavinys. Jam spręsti per pastarąjį dešimtmetį buvo pasiūlyti du pagrindiniai būdai: duomenų arba algoritmų lygio sprendimai. Duomenų lygio sprendimuose yra naudojami imčių metodai klasių disbalanso sumažinimui. Algoritmų lygio metodai yra populiaresni, šiuo atveju algoritmai adaptuojami uždaviniui su nesubalansuotomis klasėmis. Praktikoje populiariausi algoritmai yra sprendimų medžiai, atraminių vektorių metodas, dirbtiniai neuroniniai tinklai ir tokie paprastesni metodai kaip logistinė regresija ir naivus Bajeso algoritmas. Taip pat literatūroje yra pasiūlyti tokie sprendimai kaip jonvabalio algoritmas ar izoliavimo miškas, kurie siūlo netradicinius metodus retų atvejų prognozavimui.

Analizuojant internetinės maisto prekių parduotuvės Instacart duomenis, buvo pastebėta, kad šios kompanijos asortimente dominuoja su sveika mityba susiję produktai, dažniausiai vaisiai ir daržovės. Šiam duomenų rinkiniui buvo sudarytos susietumo taisyklės tiek prekių kategorijų, tiek produktų lygyje. Atrenkant įdomias susietumo taisykles buvo pastebėta, kad norint išvengti akivaizdžių taisyklių, reikia naudoti žemesnes dažnumo reikšmes ir aukštas svarbos ir įsitikinimo matų reikšmes. Susietumo taisyklių radimui buvo naudotas populiarus apriori algoritmas, tačiau taisyklių atrinkimas užtrunka, nes nėra iš anksto aišku kokias dažnumo ir patikimumo reikšmes reikia parinkti įdomių taisyklių radimui. Atrinkus įdomesnes taisykles buvo padaryta išvada, kad

prekių kategorijų susietumo taisyklės gali būti pritaikytos prekių išdėstymo tikslais arba abstraktesniems pasiūlymams. Tuo tarpu stiprios produktų susietumo taisyklės gali būti naudojamos klientui pasiūlant konkrečią prekę.

Tiriant banko Santander klientų pasitenkinimo duomenis pirmiausiai buvo pastebėta, kad jauniausi klientai (jaunesni nei 23 metų) yra patenkinti banko paslaugomis. Tam turi įtakos tai, kad jie paslaugomis naudojasi trumpesnį laiką ir tikėtina, kad naudojamų paslaugų skaičius yra mažesnis. Tačiau kituose kintamuosiuose yra sunku įžvelgti daugumos ir mažumos klasės skirtumus. Pagrindinės išvados:

- Su izoliavimo miško algoritmu buvo mėginta aptikti netipinius atvejus duomenyse ir padaryta išvada, kad nepatenkinti vartotojai nėra išskirtiniai. Dėl to vartotojų pasitenkinimo prognozavimas yra pakankamai sudėtingas uždavinys;
- Norint sumažinti statistiškai nereikšmingų kintamųjų triukšmą duomenyse buvo naudotas kintamųjų atrankos algoritmas Boruta. Pastebėta, kad šio metodo naudojimas reikalauja daug skaičiavimo resursų, todėl pati kintamųjų atranka dirbant su didžiaisiais duomenimis gali tapti iššūkiu. Banko Santander duomenims buvo atsisakyta 31% stulpelių nesumažinant prognozavimo tikslumo;
- Klasių disbalanso sumažinimui buvo ištirti trijų tipų imčių metodai: *downsampling*, *upsampling* ir hibdriniai. Šių imčių naudingumas buvo palygintas klasifikuojant testavimo duomenų imtį. Gauti rezultatai, kad naudingiausia duomenų transformacija yra sumažinti klasių disbalansą (bet ne suvienodinti). Su *downsampling* ir *SMOTE* metodais pavyko pasiekti didesnį tikslumą nei su pradine apmokymo imtimi;
- Vartotojų pasitenkinimo prognozavimas buvo atliktas su populiariais mašininio mokymosi algoritmais: *Random Forest, XGBoost, LightGBM* ir *Regularized Greedy Forest*. Didžiausias tikslumas pasiektas iš trijų algoritmų prognozių sudarius paprastą balsavimo modelį. Tačiau net ir tiksliausias modelis teisingai atpažino tik ~42% nepatenkintų vartotojų atvejų. Kadangi banko duomenys anoniminiai, tai nėra galimybės nustatyti, kurie kintamieji daro didžiausią įtaką vartotojų pasitenkinimui.

Literatūros sąrašas

1. Videla-Cavieres I.F., Ríos S.A. Extending market basket analysis with graph mining techniques: A real case. Expert Systems with Applications 2014; vol: 41 (4): 1928–1936.
2. Ahmed A.A.Q., Maheswari D. Churn prediction on huge telecom data using hybrid firefly based classification. Egyptian Informatics Journal 2017;18(3):215–220.
3. Kazemi A., Babaei M.E., Javad M.O.M. A data mining approach for turning potential customers into real ones in basket purchase analysis. International Journal of Business Information Systems. 2015;19(2):139.
4. Ascarza E., Neslin S.A., Netzer O. et al. In Pursuit of Enhanced Customer Retention Management: Review, Key Issues, and Future Directions. Customer Needs and Solutions. 2017;5(1–2):65–81.
5. Dyché J. The CRM Handbook: A Business Guide to Customer Relationship Management. 2001; 3-8.
6. Hahsler M., Hornik K. New probabilistic interest measures for association rules. Intelligent Data Analysis 2007; 11(5): 437–455.
7. Amatriain X., Jaimes A., Oliver N., et al. Data Mining Methods for Recommender Systems. Recommender Systems Handbook 2011; 39-71.
8. Valle M.A., Ruz G.A., Morrás R. Market basket analysis: Complementing association rules with minimum spanning trees. Expert Systems with Applications 2017; 146–62.
9. Mostafa M.M. Knowledge discovery of hidden consumer purchase behaviour: a market basket analysis. International Journal of Data Analysis Techniques and Strategies 2015; 7(4):384.
10. Vafeiadis T., Diamantaras K.I., Sarigiannidis G., et al. A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory 2015;55:1–9.
11. Huang Y., Zhu F, Yuan M., Deng K., et al. Telco Churn Prediction with Big Data. ACM Press 2015.
12. Zhu B., Baesens B., Backiel A. et al. Benchmarking sampling techniques for imbalance learning in churn prediction. Journal of the Operational Research Society 2017;69(1):49–65.
13. Haldankar A.N., Bhowmick K. A cost sensitive classifier for Big Data. IEEE International Conference on Advances in Electronics, Communication and Computer Technology (ICAECCT) 2016.
14. Fridrich M. Experimental Parameter Tuning of Artificial Neural Network in Customer Churn Prediction. Trends Economics and Management 2017;11(28):9.

15. Liu F.T., Ting K.M., Zhou Z.-H. Isolation Forest. 2008 Eighth IEEE International Conference on Data Mining 2008.
16. Johnson R., Zhang T. Learning Nonlinear Functions Using Regularized Greedy Forest. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2014;36(5):942–954.

Priedai

1 priedas

Sumaišymų matrica: tikrosios reikšmės ir izoliavimo miško aptiktos išskirtys (naudojant visus 369 kintamuosius):