TARTU ÜLIKOOL

Majandusteaduskond

Keit Adamson

ERAISIKU KREDIIDIRISKI MODELLEERIMINE ETTEVÕTTE KAUPMEHE JÄRELMAKS OÜ NÄITEL

Magistritöö sotsiaalteaduse magistrikraadi taotlemiseks majandusteaduses

Juhendaja: teadur Oliver Lukason

Tartu 2016

Soovitan suunata kaitsmisele .....

(juhendaja allkiri)

Kaitsmisele lubatud „ “.....2016. a.

Olen koostanud töö iseseisvalt. Kõik töö koostamisel kasutatud teiste autorite tööd, põhimõttelised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on viidatud.

.....

(töö autori allkiri)

SISUKORD

SISSEJUHATUS

Käesoleva magistritöö teemaks on eraisiku krediidiriski modelleerimine ettevõtte Kaupmehe Järelmaks OÜ näitel. Teema aktuaalsust tõstab asjaolu, et viimastel aastatel on Eestis eraisikutele väljastatavate tarbimislaenude maht olnud tõusvas trendis. Selle väite tõestuseks võib vaadelda Eesti Panga statistikat, mille kohaselt on Eestis aastatel 2013, 2014 ja 2015 kodumajapidamistele antud tarbimislaenude jääk, mis peegeldab ühtlasi reguleeritud tarbimislaenuturu mahtu, olnud 591.1, 602.6 ja 632.5 miljonit eurot (Kodumajapidamistele antud ... 2016). Nimetatud perioodil on kõrgema risikitasemega tagatiseta laenude osakaal tarbimislaenude jäägist kasvanud, moodustades vastavalt 55.24%, 58.25% ja 67.40% (*Ibid*.).

Mahtude kasv on kaasa toonud konkurentsi tihenemise, mille tulemusena on eraisiku maksevõime võimalikult täpne prognoosimine muutunud kreeditoride jaoks üha olulisemaks, kuna turul valitseva hinnasurve ja valitsusepoolsete regulatsioonide tõttu on laenuandmisega tegelevate ettevõtete eksimisruum muutunud väiksemaks. Mida efektiivsemalt hinnatakse krediidiandmisega seonduvat krediidiriski, seda täpsemini on võimalik seada provisjone, mis alandavad laenuandja jaoks kasutatava kapitali hinda. Täiendavalt võimaldab kõrgem klassifitseerimistäpsus hinnastada laenulepingut konkreetse taotleja riskitasemest lähtuvalt, võimaldades seeläbi pakkuda väiksemate kuludega laenu madalama krediidiriskiga klientidele. Samuti võib suurtemate laenumahtude korral väike klassifitseerimistäpsuse paranemine kreeditori jaoks kaasa tuua olulise kulude kokkuhoiu. Ühiskondlikult kasulik efekt seisneb asjaolus, et efektiivsema selektsiooni korral laenatakse vähem isikutele, kes tegelikkuses ei ole võimelised võetud kohustusi teenindama ja mille tulemusena halveneb pikemas perspektiivis selliste deebitoride majanduslik seisukord veelgi.

Magistritöö eesmärgiks on koostada eraisiku krediidiriski hindamise mudel otsustuspuu meetodil ettevõtte Kaupmehe Järelmaks OÜ näitel. Uurimustöö on piiritletud otsustuspuu meetodi kasutamisega, kuna meetod on akadeemilises kirjanduses hinnatud

interpreteeritavuse ja hea klassifitseerimistäpsuse pärast. Ka võib üheks eesmärgi valiku põhjuseks pidada Eesti akadeemilise kirjanduse vähesust eraisiku krediidiriski modelleerimisel antud meetodiga. Uurimustöö tulemused võivad leida kasutust ja edasiarendamist uuritava ettevõtte krediidiriski poliitika ja mudelite täiendamisel. Eesmärgi saavutamiseks on püstitatud järgmised uurimisülesanded:

- anda erialakirjanduse põhjal ülevaade eraisiku krediidiriskist, selle hindamisel kasutatavast metodoloogiast ja erinevate meetodite klassifitseerimistäpsusest;
- käsitleda krediidiriski hindamise kontekstis enimkasutatud selgitavaid muutujaid ja nende mõju;
- luua krediidiriski hindamise mudelid C4.5 meetodil;
- analüüsida mudelitepõhiseid muutujate mõju suundasid ja kõrvutada neid erialakirjanduses saadud tulemustega;
- hinnata ja analüüsida saadud mudelite klassifitseerimistäpsust.

Krediidiriski hindamist (*„credit scoring“*) kui tegevust on defineeritud erialases kirjanduses mitmete autorite poolt, sealhulgas on märgatavalt panustanud sellealasesse teadustöösse näiteks D.J. Hand, R. Anderson, L.C. Thomas, B. Baesens. Näiteks kirjeldavad D.J. Hand ja W.E. Henley krediidiriski hindamist, kui formaalset protsessi määramaks tõenäosust, millega taotleja tagasimaksete osas maksejõetuks osutub (Hand, Henley 1997: 524). L.C. Thomas ja kaasautorid defineerivad krediidiriski hindamise läbi otsustusmudelite ja nendes kasutatavate meetodite, mis abistavad kreeditore tarbimiskrediidi väljaandmisel ja mille tulemusena otsustatakse, kellele ja kui palju krediiti peaks väljastama ning milliseid operatsioonilisi strateegiaid peaks parendama laenuandja kasumlikkuse suurendamiseks (Thomas *et al*. 2002: 1).

Töös kasutatavad andmed pärinevad ettevõtte OÜ Kaupmehe Järelmaks infosüsteemi andmebaasist. Valimi suuruseks on 3901 vaatlust, mis on moodustatud juhuvalimina 2011. aasta järelmaksulepingutest. Iga andmestikus oleva lepingu kohta on teada kliendi sugu, vanus taotlemise hetkel, perekonnaseis, haridustase, ülalpeetavate arv, elukoha tüüp, postiaadressi maakond, postiaadressi linn, tegevusala, taotlemise hetkel praegusel ametikohal töötatud aeg kuudes, igakuine sissetulek eurodes, maksehäirete arv

taotlemise hetkel, laenusumma eurodes, laenuperiood kuudes ja maksehäire esinemine või mitteesinemine lepingus.

Eraisiku krediidiriski hindamiseks kasutatakse J.S. Quinlani poolt välja töötatud otsustuspuu algoritmi C4.5, mille raames jagatakse valim kaheks. Esimese saadud alamvalimi peal töötatakse välja mudel ja teist kasutatakse mudeli prognoosivõime hindamiseks kasutades selleks PCC („*percentage correctly classified*“) ja ROC („receiver operating characteristics“) kõvera aluse pindala mõõtusid. Saadud tulemusi põhjendatakse ja kõrvutatakse erialases kirjanduses tehtud järeldustega.

Eraisiku krediidiriski hindamise valdkonnas on erialase kirjanduse põhjal alust arvata, et klassifitseerimismeetodite kasutamine on hetkel enimlevinud lähenemine krediidiriski hindamise mudelite loomisel (Lessmann *et al*. 2013: 2). Erinevate autorite poolt on nimetatud teemal avaldatud mitmeid meetodite võrdlusele keskenduvaid artikleid nagu näiteks „Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring.“ (Baesens *et al*. 2003: 627-635), „Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update.“ (Lessmann *et al*. 2013: 1-60), „Neural Network ensemble strategies for financial decision applications.“ (West *et al*. 2005: 2543-2559), „An experimental comparison of classification algorithms for imbalanced credit scoring data sets.“ (Brown, Mues 2012: 3446-3453) ja „Building credit scoring models using genetic programming.“ (Ong *et al*. 2005: 41–47).

Otsustuspuu algoritmi C4.5 klassifitseerimistäpsus kõigub erinevate teadusartiklite ja valimite lõikes oluliselt. Nimelt on klassifitseerija osutunud väga täpseks Baesens *et al*. (2003) poolt avaldatud teadustöös diskreetsete väärtuste ja puu kärpimise korral näitaja PCC järgi, kuid kümme aastat hiljem Lessmann *et al*. (2013) poolt avaldatud võrdlevas uuringus osutus meetodi klassifitseerimistäpsus pigem madalaks või keskmiseks. Ka on Ong *et al*. (2005) poolt tehtud töös otsutuspuu algoritm C4.5 võrreldes teiste klassifitseerijatega saavutanud häid tulemusi. Enamasti on nimetatud teadustöödes peetud kõrge prognoosivõimega klassifitseerijateks nii juhumetsa kui ka närvivõrkude meetodeid.

Käesoleva magistritöö sisuline osa on jaotatud kaheks peatükiks, millest esimese alapeatükkides antakse ülevaade erialases kirjanduses kasutusel olevatest

definitsioonidest mõistele krediidirisk, avatakse krediidiriski hindamise tausta ja kirjeldatakse teemakohases kirjanduses enamlevinud statistilisi meetodeid. Täiendavalt käsitletakse peatükis uuritavas valdkonnas kasutust leidnud selgitavaid muutujaid ja nende mõju suunda. Autor koostab töö raames kasutatud kirjanduse põhjal kokkuvõtte statistiliselt oluliseks osutunud muutujatest kasutussageduse järgi. Ka kirjeldatakse varasemaid uurimusi ja nende raames saadud tulemusi erinevate meetodite lõikes – kordajad, klassifitseerimistäpsused ja mudelitega seonduvad probleemid.

Teises peatükis ehk töö empiirilises osas tutvustatakse lühidalt uuritavat ettevõtet Kaupmehe Järelmaks OÜ ja selle tegevusvaldkonda ning kehtinud laenuandmise põhimõtteid. Järgmisena kirjeldatakse töös kasutatavaid algandmeid, millele järgnevad rakendatava otsustuspuu meetodi ülevaade, koostatud mudelid, nende klassifitseerimistäpsused ja seonduvad probleemid. Viimasena arutletakse empiirilise uurimuse käigus saadud tulemuste üle ja kõrvutatakse neid erialases kirjanduses saadud resultaatidega.

Märksõnad: eraisiku krediidirisk, otsustuspuu, klassifitseerimine, krediidiriski modelleerimine

1. ERAISIKU KREDIIDIRISK JA SELLE MODELLEERIMISE TEOREETILISED ALUSED

1.1. Eraisiku krediidiriski hindamine

Krediit ehk laen omab olulist rolli kaubandustehingutes ja on mõjukas tegur riikide majanduse funktsioneerimisel, millega on rahvamajanduslikul tasandil võimalik siluda majanduslanguseid riigi majandust elavdades. Ka võimaldab krediit nii avaliku sektori kui ka eraettevõtetel teha tulevikku suunatud investeeringuid, milleks äriüksusel sisemised vabad vahendid puuduvad või on muul põhjusel väliskapitali kaasamine rentaablim. Investeeringute kõrval kasutatakse laene ka teistel otstarvetel - näiteks kaetakse käibelaenuga operatiivkulusid ja tagatakse äriüksuse likviidsust.

Nii nagu ettevõtted, kasutavad ka eraisikud investeeringu või tarbimise tulevikku lükkamise asemel krediiti. Samuti leiab laenuraha rakendust ootamatute kulutuste katmisel, mille tegemist ei ole erinevatel põhjustel mõistlik edasi lükata. Eraisikud on ühe enam hakanud eluaseme ja investeeringute finantseerimise kõrval kasutama krediiti ka kaupade ja teenuste tarbimise finantseerimiseks. Sellise muutuse üheks põhjuseks võib pidada hedonistlikke väärtuste tähtsuse ja hetkele orienteeritud mõtteviisi suurenevat levikut ühiskonnas.

Seda seisukohta toetab osaliselt 1991 kuni 2001. aasta Saksamaa eraisiku tarbimislaenu turu kohta tehtud uuring, milles järeldatakse, et kompulsiivse ostukäitumise tõus on mõjutanud eraisiku tarbimislaenu turu kasvu. Ühe põhjusena tuuakse välja kaubaartikli esteetika osatähtsuse suurenemist, mis hõlmab endas riski tarbimist kannustava praktilise vajaduse ja tegeliku tarbimise eraldamiseks. Ka nimetatakse varasemast efektiivsemat reklaamindust mõjurina, mis survestab täiendavalt eraisikute ostukäitumist. (Neuner *et al.* 2005: 509–522)

Guardia (2002: 2) jaotab kodumajapidamistele antavad laenud kaheks – eluasemelaen ja tarbijakrediit. Nimetatud autor defineerib tarbijakrediidi ehk tarbimislaenu läbi kahe

erineva krediidikategooria, millest esimest iseloomustab eraisikul lasuv lepinguline kohustus kasutada saadud laenu lepingus määratud teenuse või toote soetamiseks. Teise krediidikategooria korral eelnimetatud lepingust tulenev nõue puudub ja tarbija võib laenust saadud raha kasutada vabalt valitud toodete ja teenuste finantseerimiseks. Täiendavalt peab autor oluliseks tuua välja tarbijakrediidi ja eluasemelaenu erinevustena asjaolu, et valdavalt ei ole tarbimislaen tagatisega tagatud, aga eluasemelaenu tagatiseks on ostetav kinnistu. (Guardia 2002: 2)

Euroopa Keskpanga definitsiooni kohaselt on tarbijakrediit laen, mida väljastakse kodumajapidamistele isiklikuks kaupade või teenuste tarbimiseks (Statistics glossary 2015). Euroopa Komisjoni poolt tellitud uuringus määratletakse tarbijakrediit kui eraisikule antav laen, mis ei ole tagatud kinnistuga, mille eesmärk ei ole kinnistu omandamine, mille pakkujaks on pank või muu kreeditor ja mille eesmärk ei ole seotud isiku majandus- või kutsetegevusega (Study on... 2013:11).

Eelnimetatud definitsioonide ühisosaks on eraisiku poolt ostetavade teenuste ja kaupade finantseerimine tarbimislaenuga. Guardia ja Euroopa Komisjoni poolt tellitud uuringus peetakse tarbijakrediidi määratluses oluliseks selget eristust eluasemelaenust. Täpsustusena on oluline eluaseme- ja tarbimislaenu seisukohalt välja tuua asjaolu, et kui eluasemelaenu korral peab krediit olema tagatud soetatava kinnistuga, siis tarbijakrediit võib olla tagatud elamispinnana kasutatava kinnistuga. Võrreldes Euroopa Keskpanga ja Euroopa Komisjoni poolt tellitud uuringu tarbimislaenu definitsiooniga ei sea Guardia määratlus otseselt piiranguid krediidi kasutamiseks isiku majandus- või kutsetegevuses. Laenuraha kasutamise eesmärgi piiramine tarbijakrediidi definitsioonis on töö autori seisukohalt oluline, kuna ettevõtlusesse suunatud finantsvahendeid ei saa oma olemuselt pidada eraisiku tarbimiseks. Tegelikkuses puudub kreeditoril tihti kontroll laenuraha kasutamise üle, välja arvatud juhtudel, kus lepinguliselt on sätestatud krediidi seos konkreetse teenuse või kauba ostmisega nagu näiteks järelmaksutoote puhul, kus deebitorile rahalist väljamakset ei tehta. Nimetatud asjaolu võib teatud valimite korral täiendavalt moonutada tulemusi.

Kui kreeditor astub eraisikuga krediidisuhtesse, kaasnevad sellega laenuandja jaoks erinevad riskid, kaasaarvatud krediidirisk. Üheks sellise riski põhjuseks on osapoolte vahel valitsev tugev informatsiooni asümmeetria. Kuna tarbimislaenuturul on üheks

oluliseks teguriks teenuse osutamise kiirus, siis tehakse laenuotsus tihti põhjalikumaid kontrolle teostamata ja lähtutakse suuresti informatsioonist, mille klient taotlusel esitas ja mida on võimalik pärida erinevatest registritest. Kuigi mõlemad pooled astuvad laenu väljastamisel lepingulisse suhtesse, kus on õiguspäraselt ära määratud krediidiandja poolsed nõuded krediidisaaja vastu, ei ole kreeditoril kindlust, et deebitor kavatseb ja suudab laenulepingust tulenevaid kreeditoripoolseid nõudeid lepingujärgselt täita. Eelnimetatud olukorda iseloomustab krediidirisk, mida Baseli Pangajärelevalve Komitee defineerib kui tõenäosust, mille puhul laenaja või vastaspool ei täida kreeditori ees nõuetekohaselt kokkuleppejärgseid kohustusi (Principles for ... 2000: 1). Brown ja Moles defineerivad krediidiriski läbi kolme karakteristiku (Brown, Moles 2008: 2)

- avatus lepingulisele osapoolele, kellel võib esineda maksehäire („*default*“) või kelle maksekäitumine võib oluliselt halveneda;
- tõenäosus, et lepingulisel osapoolel esineb kohustuste osas maksehäire;
- sissenõudmismäär ehk kui suure osa nõudest suudab kreeditor maksehäire esinemisel sisse nõuda.

Anderson määratleb krediidiriski, kui mistahes riski, mis on põhjustatud tegelikust või tunnetuslikust muutusest vastaspoole võimes täita krediidikohustusi. See ei kata ainult nõude potentsiaalse mittelaekumisega seotud riski, vaid ka edasimüüdava nõude turuväärtuse vähenemisega seotud riski. Täiendavalt peab Anderson krediidiriski üheks komponendiks võla sissenõudmiskulude esinemisega seotud riski. (Anderson 2007: 98)

Baseli Pangajärelevalve Komitee definitsioon krediidiriskist on ülejäänud kahest fundamentaalselt erinev, kuna esimese määratlus piirdub deebitori ja kreeditori vahelise kokkuleppe rikkumisega, kuid teistel juhtudel tuuakse sisse täiendavalt hinnang kuludele, mis võivad esineda, kui risk peaks realiseeruma. Krediidiriski määratluses viidatakse lepingujärgsete kohustuste mittetäitmisele, kuid täpsustavat informatsiooni krediidiriski realiseerumise kohta ülalnimetatud definitsioonid endas ei hõlma. Kuna tarbijakrediidi tooted võivad oma tingimustelt erinevad olla, siis oleks ühtse definitsiooni andmine krediidiriski realiseerumisele erinevate finantstoodete lõikes keeruline. Üldlevinud praktikas peetakse nimetatud riski realiseerumisega seotud sündmuseks maksehäiret, mille täpne sisemiselt kasutatav määratlus on krediidiandja poolt defineerida, kuid enamasti lähtutakse rahvusvahelisest praktikast. Olemuselt

eelneb maksehäirele lühiajaline viivitus ühe lepingujärgse maksega, kuid sellisel juhul ei ole deebitoril enamasti põhjust pidada tekkinud olukorda püsivaks, sest kliendipoolse makse mittetegemise põhjuseks võib olla näiteks maksekuupäeva unustamine. Olles viivituses mitme makse ulatuses, on deebitori hinnang tõenäosusele, et võlas olevad ja lepingujärgsed tulevikus olevad maksed laekuvad, vähenenud. Sellises olukorras on lepingul esinenud maksehäire. Kui maksehäire ei ole ajutist laadi, võib see eskaleeruda püsivaks maksejõuetuseks.

Traditsiooniliselt kasutati krediidiriski realiseerumise hindamise seisukohalt spetsiifilist maksehäire esinemisega seotud riski (Thomas 2009: 6). Enimlevinud oli lähenemine, kus hinnati hetkeseisust lähtuvalt tõenäosust, et taotleja satub ühe või mitme maksega üle 90 päeva viivitusse järgmise 12 kuu jooksul (*Ibid.*: 6). Mis juhtub peale seda perioodi kliendi maksekäitumises ja kas krediidileping osutub laenaja jaoks kasumlikuks, olid aspektid, mida selline mudel ei käsitlenud (*Ibid.*: 6). Hiljem leidsid kasutust ka mudelid, kus kasutati maksehäire määratluses mõnda muud fikseeritud ajalist raamistikku (*Ibid*.: 6). Tongi ja teiste kirjutatud uurimuses ei kasutatud võrreldes eelmise definitsiooniga ajalist piirangut, vaid määratleti maksehäirena olukord, kus vähemalt ühe maksega oldi viivituses 90 või enam päeva (Tong *et al.* 2012: 136). Baseli Pangajärelevalve Komitee definitsioon hõlmab endas teatud täpsustusi defineerides maksehäirena situatsiooni, kus laenulepinguga on aset leidnud vähemalt üks järgmistest sündmustest (International Convergence ... 2006: 100):

- pank leiab, et võlgniku poolt krediidikohustuste tagasimaksmine täies ulatuses on vähetõenäoline ilma väärtpaberite realiseerimiseta juhul, kui neid omatakse;
- võlgnik on mistahes krediidikohustusega panga ees viivituses enam kui 90 päeva kohustuse tekkimise kuupäevast. Arvelduslaenu peetakse viivituses olevaks, kui klient on ületanud soovitatud limiidi või talle on soovitatud madalam limiiti, kui sellel hetkel kasutusel olev laenusumma.

Krediidiriski tähtsus, täpsemalt kui suure osa väljalaenatud rahast laenuandja maksehäirete tõttu kaotab, on tarbijakrediidi laenuportfelli seisukohalt Baseli kapitali adekvaatsuse raamistikus toimunud muudatuste tõttu veelgi tõusnud (Thomas 2009: 8). Basel II raamistik, mis jõustus 2007. aastal, võimaldab pankadel kasutada sisemisi krediidiriski hindamise mudeleid, et ära määrata, millises mahus peavad pangad

provisjone looma, et katta maksehäiretest tingitud potentsiaalseid laenuportfelliga seotud kahjusid (*Ibid*.: 8). Mida täpsemini suudetakse hinnata krediidiandmisega seonduvat krediidiriski, seda täpsemini on võimalik provisjone seada, mis omakorda alandab laenuandja jaoks kasutatava kapitali hinda. Järelikult on kreeditorid motiveeritud töötama välja ettevõttesisese definitsiooni maksehäirele ja arendama eraisiku krediidiriski mudeleid, mis tagaksid võimalikult täpsed prognooside tasemed.

Enne statistiliste meetodite kasutuselevõttu eraisiku krediidiriski hindamisel tegid laenuandmise otsuse selleks spetsialiseerunud töötajad („*underwriter"*). Subjektiivse laenuandmise otsuse tegemisel kasutati kliendi varasemat maksekäitumist, kui tegemist oli laenuandja olemasoleva kliendiga, ja kliendi poolt avaldatud täiendavat informatsiooni. Selleks, et laenutaotlejast ja tema finantsolukorrast saaks täielikuma pildi, kasutati taotlemisprotsessis paralleelselt intervjuu meetodit.

Sellise krediidiriski hindamise metoodika miinusteks on tööjõumahukus, taotlemisprotsessi pikkus ja inimfaktorist tingitud subjektiivsed vead otsuste tegemisel. Viimane on tingitud konkreetse töötaja eelarvamusest selle osas, millised omadused on heal või halval deebitoril. Samas võib nimetatud puudust pidada ka manuaalse laenuotsuse tegemise eeliseks, kui selleks spetsialiseerunud töötajad omavad kõrget kompetentsi taotleja hindamisel ja nende otsused on põhjendatud kasutades sarnaste laenusaajatega seotud faktilisi hindamisaluseid.

Majanduslik surve ettevõtetele krediidi nõudluse suurenemisest, tehnoloogiline areng ja tihenev konkurents on toonud kaasa keerulisemate statistiliste meetodite kasutamise laenuandmise otsustamisel. Statistiliste meetodite kasutamisel lähtutakse eeldusest, et taotleja maksevõimelisust on võimalik automaatselt hinnata taotleja kohta saadava informatsiooni põhjal kasutades otsuse tegemiseks eelmiste taotlejate ja nende maksekäitumisega seotud andmeid.

Krediidiriski hindamiseks („*credit scoring*") nimetatakse statistilistele meetoditele põhinevat krediidiriski hindamissüsteemi, mille eesmärgiks on grupeerida krediidi taotlejad krediidiriski järgi „hea riski" gruppi, kes tõenäoliselt täidavad oma finantskohustusi nõuetekohaselt, ja „halva riski" gruppi, kelle puhul on suur maksehäiresse sattumise tõenäosus (Yap *et al.* 2011: 13274-13283). Kasutades

ajaloolisi andmeid maksekäitumise, demograafiliste, finantseisu kajastavate ja muude käitumuslike tunnuste kohta, aitab krediidiriski hindamise mudel identifitseerida krediidiriski hindamise seisukohalt olulised tunnused ja anda neist lähtuvalt igale kliendile krediidiskoor (*Ibid.*: 2). Traditsiooniliselt lõppes krediidiriski hindamine skoorikaardi loomisega, kuid tänapäeval laialt kasutatavate klassifitseerimispuu algoritmide ja ekspertsüsteemide korral on lõpptulemuseks reeglite kogum, mille alusel on võimalik uue taotleja klassifitseerimine (Thomas 2000: 158; Thomas 2009: 97).

Hand ja Henley kirjeldavad krediidiriski hindamist, kui formaalset protsessi määramaks tõenäosust, millega taotleja tagasimaksete osas maksejõetuks osutub (Hand, Henley 1997: 524). Mõnikord kasutatakse krediidiriski hindamise asemel terminit taotluse põhine krediidiriski hindamine („*application scoring*“) eristamaks seda käitumuslikust krediidiriski hindamisest („behavioral scoring“), mis hõlmab endas jooksvaid monitoorimis- ja prognoosimistegevusi laenu saanud klientide maksekäitumise hindamiseks (*Ibid.*: 524). Krediidiriski hindamisel kasutatakse statistilisi mudeleid, nagu näiteks skoorikaardid või klassifikaatorid, taotluse vormidelt ja muudest allikatest kogutud sõltumatute muutujate abil maksehäire esinemise tõenäosuse hindamiseks (*Ibid.*: 524). Thomas *et al*. (2002: 1) täiendavad definitsiooni kasumlikkuse aspektiga. Nad defineerivad krediidiriski hindamise läbi otsustusmudelite ja nendes kasutatavate meetodite, mis abistavad kreeditore tarbimiskrediidi väljaandmisel (*Ibid.*: 1). Nende meetodite tulemusena otsustatakse, kellele ja kui palju krediiti peaks andma ning milliseid operatsioonilisi strateegiaid peaks parendama laenuandja kasumlikkuse suurendamiseks (*Ibid.*: 1).

Andersoni hinnangul peab krediidiriski hindamise defineerimiseks mõiste lahutama kaheks erinevaks komponendiks- krediit ja hindamine („*scoring*“). Terminit „krediit“ võib mõista, kui konspetsiooni „osta kohe, maksa hiljem“. Krediit pärineb ladinakeelsest sõnast „credo“, mis tähendab uskumist ja usaldamist. Teiseks, termin „hindamine“ viitab numbrilisele töövahendile, mida kasutatakse objektide või nähtuste järjestamiseks eristades neid faktipõhiste kvaliteeditunnuste alusel eesmärgiga tagada objektiivsed ja järjepidevad otsused. Seega võib skoori pidada numbriliseks väärtuseks, mis iseloomustab ühte konkreetset omadust ja mida kasutatakse reastamiseks, hinne seevastu iseloomustab ühte või mitut omadust. Lihtsustatult öeldes on krediidiriski

hindamine statistiliste mudelite kasutamine asjakohaste andmete transformeerimiseks numbrilisteks mõõtmeteks, mida kasutatakse laenuotsuste tegemiseks. (Anderson 2007: 3-6)

Eelpool mainitud krediidiriski hindamise definitsioonidest nähtub, et kõik nimetatud autorid peavad oluliseks statistiliste meetodite rakendamist maksehäire esinemise tõenäosuse hindamisel. Nimetatud autoritest nimetavad Thomas ja tema kaasautorid krediidiriski hindamise ühe täiendava olulise eesmärgina laenutaotleja krediidisumma üle otsustamist, mis võimaldab laenupakkujal suurendada kasumlikkust pakkudes kliendi poolt taotletud laenusummast madalamat krediidisummat eraisikutele, kes taotletud laenusummale ei kvalifitseeru, kuid suudaksid laenuandja hinnangul väiksemaid kohustusi teenindada. Ka võimaldab selline lähenemine pakkuda kõrgemat laenusummat eraisikutele, kelle laenu teenindamise võimet hinnatakse taotletud laenusummast kõrgemaks.

Krediidiriski hindamine statistilistel meetoditel on olemuselt empiiriline, mille üheks eeliseks subjektiivse otsusprotsessi ees on inimlikust subjektiivsusest tingitud nihke minimeerimine. Objektiivsus laenuotsuste tegemisel aitab kreeditoridel vältida diskrimineerimisest tingitud süüdistusi, mis võivad oluliselt kahjustada organisatsiooni mainet või kaasa tuua kohtukaasusi. Objektiivsetel alustel otsustamine on kindlasti ootus inimestele, kes vastutavad laenuotsuste tegemise eest, kuid inimene on oma loomuselt avatud üldistamisele ja stereotüüpide kujundamisele. Indiviidid loovad eeldusi, mis suunavad nende elu- eelkõige kui kogetakse samade sündmuste ja tulemuste, situatsioonide ja tagajärgede kordumist. Selliseid eeldusi ei looda vaid isiklike kogemuste põhjal, vaid ka isiklikust kommunikatsioonist teistega ja meedia põhjal. Kui sellistel eeldustel puudub faktuaalne taust, on tegemist pigem eelarvamusega. (Anderson 2007: 17-18)

Autorid Abdou ja Pointon toovad krediidiriski hindamise objektiivsuse põhjusena välja täiendavalt asjaolu, et mudeli loomisel kasutatakse märksa suuremat valimit, kui laenuandmisele spetsialiseerunud töötaja suudaks meelde jätta. Krediidiriski hindamise mudelid võtavad arvesse nii hea kui halva maksekäitumisega deebitore, kuid subjektiivse meetod on pigem nihkes kehva maksekäitumisega laenutaotlejate poole. (Abdou, Pointon 2011: 4-5)

Krediidiriski hindamise kasutamine on võrreldes subjektiivse laenuotsustuse meetodiga kiirem, mis on tarbimislaenuturu seisukohalt oluline konkurentsivõime tagamiseks, kuna täiendavat finantseerimisvajadust ei ole pigem pikemalt ette planeeritud ja laenuotsused peavad olema kiired. Osa laenuandjatest rakendavad kahetasandilist skoori lävendit, kus esimese korral välistatakse need kliendid, kes ei kvalifitseeru laenu saamiseks. Teine skoori lävend on mõeldud klassifitseerimaks kliente, kelle maksehäire esinemise tõenäosust hindab laenuandja laenutaotlusel esitatud andmete põhjal krediidiriski hindamise mudeli automaatotsus piisavalt madalaks ja sellisel juhul ei peeta vajalikuks laenuotsuse andmisele spetsialiseerunud töötaja sekkumist. Järelikult on krediidiriski hindamist rakendades võimalik teatud osale taotlejatest anda laenuotsus praktiliselt hetkega võimaldades kreeditori tähelepanu suunata enam taotustele, kus on küsitavusi. Selliselt saab laenuandmise protsessi kuni laenusumma väljamakseni teatud kanalite ja klientide jaoks täielikult automatiseerida.

Statistilistel meetoditel põhinevatele krediidiriski mudelitele on ette heidetud mudeli keerukust ja selliste muutujate kasutamist, millel ei ole tõlgendaja jaoks selgelt põhjendatavat seost maksehäiresse sattumise tõenäosusega. Samuti nähakse probleemina minevikuliste andmete kasutamist ja mudeli tundlikkust püstitatud kujule, mis võib eksimuse korral suurtemate mahtude korral kaasa tuua märkimisväärseid kahjusid. (Abdou, Pointon 2011: 5)

Ajalooliselt on eraisiku krediidiriski hindamiseks kasutatud enamlevinud statistilisteks meetoditeks olnud diskriminantanalüüs ja lineaarne regressioonanalüüs (Hand, Henley 1997: 531-532). Tänapäeval on eraisiku krediidiriski hindamise mudelite loomisel kasutust leidnud lai valik statistilisi meetodeid, millest on krediidianalüütikute, teadurite, kreeditoride ja teemakohase arvutitarkvara tootjate poolt enim kasutatavateks regressioonanalüüs, lineaarne programmeerimine („*linear programming*“), Coxi proportsionaalsete (võrdeliste) riskide mudel („*Cox proportional hazards model*“), tugivektor-masinate mudel (SVM - *„support vector machines*“), tehisnärvivõrgud („*artificial neural networks*“), otsustuspuud („*decision trees*“), lähima naabri meetod („*k-nearest neighbour*“), geneetilised algoritmid („*genetic algorithms*“), juhumetsa („*random forest*“) meetod ja geneetiline programmeerimine („*genetic programming*“) (Abdou, Pointon 2011: 13; Anderson 2007, 163; Brown, Mues 2012: 3446; Thomas

2009: 98). Järgmisena antakse detailsem ülevaade enamlevinud kredidiidiriski hindamiseks kasutatavatest statistilistest meetoditest.

Diskriminantanalüüs on parameetriline statistiline tehnika, mis võimaldab vaatlusi klassifitseerida sõltuva muutuja gruppidesse, milleks krediidiriski hindamise seisukohalt on laenude jaotamine maksehäirega ja maksehäireta laenudeks (Abdou, Pointon 2011: 69). Diskriminantfunktsioon, millega antakse igale objektile diskriminantskoor, avaldub järgmiselt (Lee *et al.* 2002: 245-254):

(1) $D=\beta_0+\beta_1X_1+\beta_2X_2+...\beta_nX_n$

kus D - vaatluse diskriminantskoor,

β_0 - vabaliige,

β_i - sõltumatule muutujale X_i antud kaal (i=1,…, n),

X_i - sõltumatu muutuja *(i=1,...,n).*

Kuigi diskriminantanalüüs oli esimene laiemalt kasutust leidnud statistiline meetod krediidiriski hindamise mudelites, on seda kritiseeritud kehva klassifitseerimistäpsuse pärast, kuna see on eelkõige loodud avastama lineaarseid sõltuvusi muutujate vahel (Lee, Chen 2005: 743-752). Võrreldes logistilise regressiooniga peetakse meetodi puuduseks suuremat arvu eeldusi, millest üheks olulisemaks on selgitavate muutujate normaaljaotus (Anderson, 2007: 170).

Logistiline regressioon on laialt kasutatust leidnud regressioonanalüüsi vorm, kus binaarse väljundi tõenäosus on seotud potentsiaalsete selgitavate muutujatega järgmisel kujul (Cox, Snell 1989: 19; Lee *et al.* 2002: 245-254):

(3) $$log\left[\frac{p}{(1-p)}\right]=\beta_0+\beta_1X_1+\beta_2X_2+\ldots\beta_nX_n$$

kus p - modelleeritava sündmuse esinemise tõenäosus,

β_0 - vabaliige,

β_i - sõltumatu muutuja X_i kordaja *(i*=1,…,*n),*

X_i - sõltumatu muutuja *(i*=1,…,*n).*

Logit mudeli korral on sõltuvaks muutujaks logaritmiline šansside suhe, kus šansside suhe on tõenäosus, et sündmus toimub, jagatud tõenäosusega, et sündmust ei toimu (Lee

et al. 2002: 245-254). Üheks meetodi eelduseks on lineaarne seos sõltumatute muutujate ja logaritmilise šansside suhte vahel (Anderson 2007: 170).

O. L. Mangasarian järeldas oma teadustöös, et lineaarset programmeerimist („*linear programming*“) on võimalik kasutada kahe grupiga klassifitseerimisprobleemi lahendamiseks eraldades need hüpertasandiga (Mangasarian 1965: 451). Eeldame, et valimisse kuulub n arv laenutaotlejaid, mille korral n_G on maksehäireta ja n_B maksehäirega taotlused. Taotleja i kohta on teada m selgitava muutuja väärtust x_{i1}, x_{i2},..., x_{im} (Thomas *et al.* 2002: 64). Sellisel juhul klassifitseeritakse taotlused krediidiriski järgi minimeerides selleks valesti klassifitseerimist täiendava muutuja („*slack variable*“) absoluutväärtuste summa minimeerimise kaudu, mis on esitatav järgmise lineaarse programmina (*Ibid.*: 64):

$$(4) \qquad min\left(\alpha_1 + \alpha_2 + \cdots + \alpha_{n_G+n_B}\right)$$

$$\text{eeldusel, et } \begin{cases} w_1x_{i1} + w_2x_{i2} + \cdots + w_mx_{im} \geq c - \alpha_i, & 1 \leq i \leq n_G, \\ w_1x_{i1} + w_2x_{i2} + \cdots + w_mx_{im} \leq c + \alpha_i, & n_G + 1 \leq i \leq n_G + n_B, \\ \alpha_i \geq 0, & 1 \leq i \leq n_G + n_B. \end{cases}$$

kus α_i - *täiendav muutuja,*

w_m - selgitava muutuja kaal,

x_{im} - selgitav muutuja,

c - lõikeväärtus.

Holland (1992: 1-211) oli esimene, kes tutvustas geneetiliste algoritmide („*genetic algorithms*“) meetodit, mis on oma olemuselt bioloogilise evolutsiooni abstraktsioon. Geneetiline algoritm kasutab geneetikast inspireeritud operaatoreid arendamaks esialgsest populatsioonist uut populatsiooni. Iga populatsiooni liige koosneb kromosoomidest, mis kujutavad endast geneetiliselt kodeeritud lahendust konkreetsele probleemile. Igale populatsiooni liikmele omistatatakse väärtusfunktsiooni väärtus („*fitness score*“), mis iseloomustab selle võimekust lahenduse seisukohalt. Uus populatsioon areneb välja kasutades ristamise operaatoreid, mutatsioone ja selektsiooni. (Kozeny 2015: 2998-3004)

Tugivektor-masinad on andmete klassifitseerimismeetod, mis klassifitseerib binaarandmed kasutades hüpertasandit selliselt, et klassi puntki kaugus tasandist oleks

maksimaalne (Bellotti, Crook 2009: 3302-3308). Kui klassid ei ole hüpertasandiga eraldatavad, tuuakse õpiandmete kontekstis sisse täiendav muutuja („*slack variable*“), mis võimaldab vaatlusel esineda valel pool hüpertasandit (*Ibid*.: 3302-3308). Sellisel juhul rakendatakse vaatlusele vea hinda („*penalty*“), mis sõltub sellest, kui kaugel valel pool vaatlus asub (*Ibid*.: 3302-3308). Nii minimeeritakse klassifitseerimisprobleemi lahendamisel vea hindade summat ja maksimeeritakse kaugust tasandist (*Ibid*). Kui $y_i \in \{-1, +1\}$, $i = 1,..,n$, siis on tugivektor-masinate meetodi optimeerimisprobleem kujutatav järgmiselt (*Ibid*.: 3302-3308):

(5) $$\max_{\alpha} \left(\sum_{i=1}^{n} \alpha_i + \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j k(x_i, x_j) \right)$$

eeldusel, et $0 \leq \alpha_i \leq C$, $\sum_{i=1}^{n} y_i \alpha_i = 0$.

kus α_i - Lagrange kordaja iga vaatluse i kohta,

x_i - tunnuse vektor,

y_i - klass, kuhu x_i kuulub,

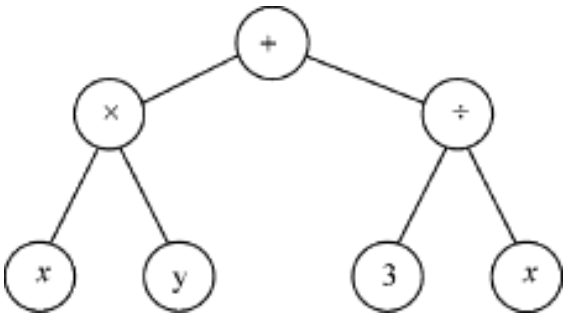
$k(x_i, x_j)$ - tuuma funktsioon,

C - konstant.

Üheks tugivektor-masinate piiranguks peetakse pikka treenimisaega ja ebatäpse hüpertasandi loomist, kui mudelisse on kaasatud mitteolulisi muutujaid ja andmemahud on suured. Kuigi meetod on robustne ja tagab üldjuhul hea klassifitseerimistäpsuse, ei võimalda see interpreteerida saadud tulemusi, kuna seost sõltuva ja sõltumatute muutujate vahel ei ole võimalik otseselt selgitada. Järelikult on meetodi praktilisel rakendamisel olulisi piiranguid, kuna ei võimalda saadud otsuseid lihtsasti põhjendada. (Han *et al.* 2013: 848-862).

Geneetilise programmeerimise („*genetical programming*“) meetodit võib esitada, kui puulaadset struktuuri, mis koosneb funktsioonide ja terminalide kogumitest (Ong *et al.* 2005: 41-47). Funktsioonide kogumi alla kuuluvad operaatorid, funktsioonid ja avaldised ning terminali kogumi alla kuuluvad sisendparameetrid, konstandid ja muud

null väärtust omavad argumendid (*Ibid*.: 41-47). Joonisel 2 on esitatud avaldise xy+3/x geneetilise programmeerimise puu näide (*Ibid*.: 41-47).

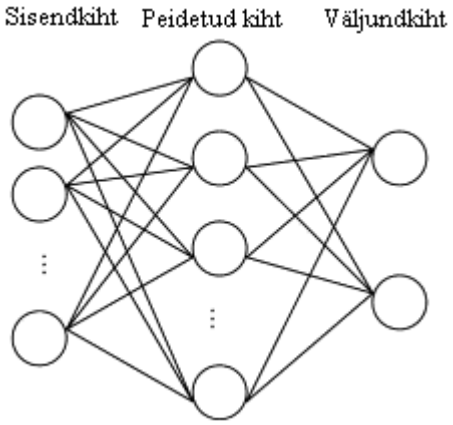


Joonis 2. Geneetilise programmeerimise puu (Ong *et al.* 2005: 41-47).

Kui geneetilise programmeerimise puu genereerimine käivitatakse, sarnaneb protseduur geneetiliste algoritmidega, kasutades väärtusfunktsiooni, ristamist, mutatsiooni ja reprodutseerimist. Geneetilise programmeerimise korral kasutatakse ristamise operaatorit erinevate puude alampuude väljavahetamiseks eesmärgiga luua uus puu struktuur rakendades selektiivsusreegleid, mitte nagu geneetiliste algoritmide puhul, kus vahetatakse bittide jadasid („*bit strings*“). (Ong *et al.* 2005: 41-47)

Tehisnärvivõrkude (ANN – „*artificial neural networks*“) meetod arendati matkima inimaju neuropsühholoogiat ja hõlmab endas mittelineaarseid regressioon-, diskriminant- ja klastermudeleid (Ong *et al.* 2005: 41-47). Tehisnärvivõrkude arhitektuuri võib tavaliselt kujutada kolme kihilise süsteemina, mis koosneb sisendi, peidetud ja väljundi kihtidest (*Ibid*.: 41-47). Sisendkihis töödeldakse sisendandmeid ja antakse need ette peidetud kihile, kus arvutatakse enne väljundkihile edastamist aktiveerimisfunktsiooni kasutades välja vastavad kaalukoefitsiendid (*Ibid*.: 41-47). Aktiveerimisfunktsiooniks võib olla näiteks hüperboolne tangens või logistiline funktsioon (*Ibid*.: 41-47). Selliselt neuroneid seotud süsteemiks ühendades, on andmestikus võimalik tuvastada keerulisi mittelineaarseid seoseid (*Ibid*.: 41-47). Joonisel 1 on kujutatud lihtne, kolmekihiline pertseptron („*perceptron*“), mis on enim kasutatud krediidiriski hindamisel (*Ibid*.: 41-47). Tehisnärvivõrkude meetodit on kritiseeritud kehva klassifitseerimistäpsuse pärast, kui mudelisse on kaasatud ebaolulisi muutujaid või andmestik on väike (*Ibid*.: 41-47). Meetodi piiranguks peetakse ka

läbipaistmatust, kuna klassifitseerimisotsuse teeb justkui „must kast“ (Hand, Henley 1997: 536).



Joonis 1. Kolmekihilise pertseptroniga tehisnärvivõrk (Ong *et al.* 2005: 41-47).

K-lähima naabri (kNN – „*k-nearest neighbour*“) meetod võimaldab määrata gruppikuulumist leides õpiandmete vaatlused, mis on grupeerimata vaatlusele kõige lähedasemad (Anderson, 2007: 177). Täht „k“ viitab naabrite arvule, mida vaatluse grupeerimisel arvesse võetakse (*Ibid*.: 177). Üheks kasutatavamaks sarnasuse mõõduks k-lähima naabri meetodi raames on Euclideani distants, mis avaldub järgmiselt (*Ibid*.: 177):

$$(6) \qquad d(x_i, x_j) = \sqrt{\sum_{i=1}^{n} (x_i - x_j)^2}$$

kus x_i - objekti i sisendvektor,

x_j - objekti j sisendvektor.

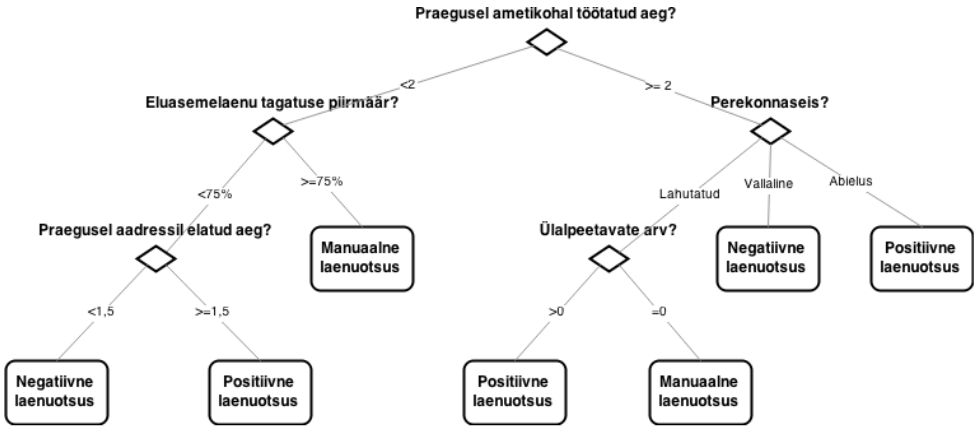
Krediidiriski hindamise seisukohalt võib k-lähima naabri meetodi piiranguteks pidada mudeli mitteloomise asjaolu, mille tõttu ei ole antud krediidiotsus läbipaistev. Ka on uute vaatluste grupeerimine suurte andmemahtude korral ajakulukas, mis võib praktikas automaatsete krediidiotsuste süsteemi korral problemaatiliseks osutuda. (*Ibid*.: 177)

Statistikas, andmekaevanduses ja masinõppimises mõistetakse otsustuspuu all ennustavat mudelit, mida on võimalik esitada klassifitseerimis- või

regressioonmudelina, millest viimast kasutatakse pideva prognoositava muutuja modelleerimiseks (Rokach, Maimon 2007: 5-6). Kui otsutuspuud kasutatakse klassifitseerimisülesandeks, viidatakse sellele täpsustavalt kui klassifitseerimispuule (*Ibid.*: 5-6). Kuna krediidiriski hindamise seisukohalt on oluline jaotada taotlejad maksehäire esinemise tõenäosuse alusel kahte gruppi, siis käsitletakse töös erinevaid klassifitseerimispuude algoritme. Tuntumad klassifitseerimispuu algoritmid, mida krediidiriski hindamisel kasutatakse, on ID3, C4.5, CART, CHAID ja MARS (Baesens et al. 2003: 631; Chuang, Lin 2009: 1685-1694; Brown, Mues 2012: 3446-3453; Ince, Aktan 2009: 236).

Otsustuspuu koosneb sisemistest sõlmpunktidest („*internal node*“), mis tähistavad individuaalse muutuja või atribuudi väärtuse kontrollimist. Järgmisena jaotatakse kontrollimise tulemust kirjeldavatest harudest lähtuvalt andmestik väiksemateks alamosadeks, mis lõppevad klasse või klasside jagunemist tähistavate lehtedega. (Han *et al.* 2012: 291)

Joonisel 3 kujutatakse tüüpilisel klassifitseerimispuul panga eluasemelaenu taotluse laenuotsustuse protsessi.



Joonis 3. Laenuotsuse otsustuspuu näide eluasemelaenu taotluse kohta (Rokach, Maimon 2007: 7).

Taotlemise protsessi ühe osana saadakse taotluse pealt järgmised andmed: ülalpeetavate arv, eluasemelaenu tagatuse piirmäär, perekonnaseis, osamakse suhe sissetulekusse, intressimäär, praegusel aadressil elatud aastate arv ja praegusel ametikohal töötatud aastate arv. Teatud osa nimetatud muutujatest kasutatakse sisemiste sõlmpunktide

loomiseks, millest hargnevad harud. Otsustuspuu meetodit rakendades klassifitseeritakse esitatud taotlused kolme erinevasse klassi, milleks on „Positiivne laenuotsus“, „Negatiivne laenuotsus“ ja „Manuaalne laenuotsus“. Näiteks peetakse nimetatud otsustuspuu järgi maksevõimeliseks kliente, kes on abielus ja praegusel ametikohal töötanud kaks või enam aastat. (Rokach, Maimon 2007: 6)

Üldjuhul on otsustuspuu meetoditega tehtud otsused läbipaistvad ja nendega saadud tulemusi lihtne implementeerida, kuid teatud juhtudel võib puu keerukusastme tõustes interpreteeritavus kannatada. Ka on seotud meetodid avatud ülesobitumisele, mille tulemusena ei ole saadud tulemused usaldusväärsed. Nimetatud probleemi lahendamiseks on enamasti vajalikud suured valimid. (Anderson, 2007: 174)

Otsustuspuu algoritm ID3 on Quinlani poolt loodud otsustuspuu algoritm, mis kasutab sõlmpunkti hargnemiskriteeriumina infohulga suurenemise (*„information gain“*) mõõtu (Quinlan 1986: 81-106). ID3 algoritm põhineb Shannoni informatsiooniteoorial (Hssina *et al.* 2014: 13). Kui tõenäosuse jaotus on $P = (p_1, p_2, ..., p_n)$ ja $S = (s_1, s_2, ..., s_n)$ on vaatluste hulk, siis selles jaotuses sisalduv informatsiooni hulk ehk entroopia on kujutatav järgmiselt (*Ibid.*: 13):

$$(7) \qquad Entroopia(P) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

kus P - tõenäosusjaotus,

p_i - tõenäosus, et S võtab väärtuse s_i.

Täpsem ülevaade Quinlani poolt välja töötatud algoritmide teoreetilise tausta kohta antakse töö empiirilises osas. ID3 algoritmi korral lõpetatakse otsustuspuu ehitamine, kui kõik vaatlused kuuluvad mõne lehe all või kui parim infohulga suurenemise kriteeriumi väärtus ei ole nullist suurem (Rokach, Maimon, 2007: 71). Nimetatud algoritm ei suuda hästi toime tulla pidevate muutujatega, kuna peab parimaks tipu hargnemise tuvastamiseks konstrueerima suure arvu otsustuspuid (*Ibid.*: 71).

Algoritmi ID3 tipu hargnemiskriteeriumil on tõsine puudus. Nimelt on sellel tugev kalduvus eelistada muutujaid, millel on palju erinevaid väärtusi. Selle iseloomustamiseks vaadeldakse hüpoteetilist olukorda, mille ülesandeks on patsiendile

anda meditsiiniline diagnoos, kus üheks muutujaks on patsienti identifitseeriv tunnus. Kuna see tunnus on unikaalne, siis toob õpiandmete jaotamine kaasa väga palju alamhulkasid, mis sisaldavad endas vaid ühe patsiendi juhtumit. Kuigi sellises olukorras on infohulga suurenemine maksimaalne, ei ole tegelikkuses hinnangu ja projektsioonide loomise koha pealt saadud tulemusel praktilist väärtust. Nimetatud ja ka mõndade teiste probleemide adresseerimiseks arendas Quinlan välja algoritmi C4.5. (Quinlan 1993: 23)

Otsustuspuu algoritm C4.5 on algoritmi ID3 edasiarendus sama autori poolt (Quinlan 1993: 23), mis kasutab sõlmpunkti hargnemiskriteeriumina („*splitting criteria*“) infohulga suurenemise määra („*gain ratio*“), millega lahendatakse varem mainitud ID3 algoritmiga eelistusega seotud probleemi. Meetodi korral konstrueeritakse otsustuspuu rekursiivset eeskirja rakendades. C4.5 algoritmiga loodud puude puhul esineb tihtipeale ülesobitumist, kuna loodud puud on liiga keerulised. Selle probleemi lahendamiseks kasutatakse retrospektiivselt tagasilõikamise protseduuri, mis kujutab endast puu kärpimist sõlmpunktide ühendamise teel. (Baesens *et al.* 2003: 631)

Lisaks ID3 algoritmiga seotud ülesobitumise ja paljude väärtustega muutujate eelistuse probleemi lahendamisele, suudab C4.5 algoritm toime tulla ka vaatlustega, mille korral teatud muutujate väärtused puuduvad või on oma olemuselt pidevad. Ka võimaldab algoritm anda muutujatele erinevaid kaalusid. (Hssina *et al.* 2014: 15- 17)

Täiendades C4.5 meetodit, arendas Quinlan välja kommertsliku otsustuspuu algoritmi C5.0, mis on autori sõnul teatud juhtudel täpsem, kiirem ja väiksema mälukasutusega kui C4.5 algoritm (Is See5/C5.0 ... 2015). Ka võimaldab C5.0 algoritm eraldi määrata muutujatega seotud vea hinna, mille kasutamisel minimeeritakse oodatavat vea hinda (*Ibid.*). Ruggieri hinnangul sisaldab C5.0 algoritm täiendavat funktsionaalsust, mida C4.5 puhul ei eksisteeri, muutes esimese sellelt seisukohalt aeglasemaks, aga samas genereerib C5.0 väiksemaid otsustuspuid, mis aitab teisalt kogu protsessis aega säästa (Ruggieri 2002: 443).

Klassifitseerimis- ja regressioonipuud (CART - „*Classification and Regression Trees*“) meetod on Breiman *et al.* (1984) poolt loodud statistiline protseduur, mille eesmärgiks on klassifitseerida vaatlusobjekt ühte või mitmesse kategooriasse (Chuang, Lin 2009:

1685-1694). CART analüüs koosneb tavaliselt kolmest eri sammust, millest esimeseks on binaarse hargnemise protseduuri kasutades küllaltki täpselt treeningandmeid kirjeldava ülekasvanud puu konstrueerimine (*Ibid*.: 1685-1694). Järgmise sammuna toimub olemasoleva ülesobitunud puu kärpimine, mille käigus tuletatakse mitmeid vähemkeerulisi puid (*Ibid*.: 1685-1694). Lõpuks valitakse optimaalse suurusega puu kasutades ristkontrolli („*cross-validation*“) protseduuri (*Ibid*.: 1685-1694). CART meetodi testid on erinevalt algoritmide C4.5 ja C5.0 testidest alati binaarsed (Hssina *et al*. 2014: 18). Eelnimetatud põhjusel kannatab CART meetodiga loodud suurtemate puude interpreteeritavus.

CHAID („*Chi-square Automatic Interaction Detector*“) meetod on John A. Hartigan poolt 1975. aastal esmakordselt avaldatud klassifitseerimispuu algoritm, mis kasutab hii-ruut testi prognoositava muutujale mitteolulist mõju omavate sõltumatute muujute kategooriate liitmiseks, hargnemis- ja peatumiskriteeriumina (Linoff, Berry 2011: 182-183). Kui hii-ruut testi kasutatakse puu hargnemise puhtuse hindamiseks, indikeerib kõrgem hii-ruudu väärtus hargnemise suuremat statistilist olulisust (*Ibid*.: 181). Kuna hii-ruut-test on mõeldud kategooriliste muutujate jaoks, siis saavad klassikalise CHAID algoritmi sisendmuutujud olla vaid kategoorilised (*Ibid*.: 183). Teatud CHAID algoritmi edasiarendus kasutavad hargnemiskriteeriumina F test, mis võimaldab pidevate muutujate kaasamist (*Ibid*.: 183). CHAID algoritmi eeliseks on kiirus ja võimalus luua laiemaid puid, kuna ei ole piiratud binaarse hargnemisega, kuid piiranguks on suurte andmemahtude nõue usaldusväärsete tulemuste saamiseks (Miner *et al*. 2009: 147).

Juhumets on defineeritud kui kärpimata klassifitseerimis- ja regressioonipuude grupp (Brown, Mues 2012: 3449). Meetodi korral luuakse esimeses iteratsioonis palju erinevaid otsutuspuid, kus iga puu genereerimiseks valitakse esialgsest valimist juhuslikult potentsiaalsed hargnemiseks kasutatavad tunnused ja osavalim (Thomas 2009: 98). Järgmisena arvutatakse iga puu järgi hinnang, mille alusel kujuneb lõplik tulemus (Brown, Mues 2012: 3449). Juhumetsa klassifitseerimismeetodi korral vajavad häälestamist kaks parameetrit, milleks on puude arv ja iga puu kasvatamiseks kasutavate tunnuste arv (*Ibid*.: 3449). Nagu otsutuspuude korral, sõltub ka juhumetsa meetodi loodavate puude lõplik kuju eelkõige valitud hargnemiskriteeriumist (Thomas 2009: 98). Nagu teiste kogumike klassifitseerimismeetodite (*„ensemble methods“*)

korral, on ka juhumetsa meetodi tulemusena loodud mudeli tõlgendamine keeruline (Gislason *et al*. 2006: 295).

Splain-regressiooni mitmetunnuseline variant MARS („*multivariate adaptive regression splines*“) on Friedmani (1991) poolt välja pakutud mitteparameetriline ja mittelineaarne meetod, mis võimaldab modelleerida seoseid sobitades mitmemõõtmelisi seosejooni (Chuang, Lin 2009: 1685-1694). Optimaalne MARS mudel rakendatakse kaheetapilise protsessina, mille esimeses etapis luuakse väga palju erinevaid andmetele ülesobituvaid alusfunktsioone, mille sisendiks võivad olla pidevad, kategoorilised või järjestikulised muutujad (*Ibid*.: 1685-1694). Alustades funktsioonist, mis panustab kõige vähem, kasutades selleks üldistatud ristkontrolli („*GCV*“) kriteeriumit („*generalized cross-validation criterion*“), kustutatakse teises etapis alusfunktsioone (*Ibid*.: 1685-1694). Jälgides üldistatud ristkontrolli väärtuse vähenemist konkreetse muutuja eemaldamisel, on võimalik hinnata muutuja olulisust (*Ibid*.: 1685-1694). Tegevust jätkatakse, kuni kõik allesjäänud alusfunktsioonid on mudelile eelmääratud nõuetega kooskõlas (*Ibid*.: 1685-1694). Üldistatud ristkontrolli funktsioon on järgmine (Friedman 1991: 20):

$$(8) \qquad GCV(M) = \frac{1}{N}\sum_{i=1}^{n}[y_i - f_M(x_i)]^2 \Big/ \left[1 - \frac{C(M)}{N}\right]^2$$

kus N - vaatluste arv,

y_i - sõltumatu muutuja väärtus ("*data response value*"),

$C(M)$ - keerukuse kulufunktsioon,

$f_M(x_i)$ - sõltumatu muutuja hinnatud väärtus.

Coxi võrdeliste riskide mudel on eraisiku kredidiiriski hindamisega seotud kirjanduses enimkasutatud elulemusanalüüsi („*survival analysis*“) meetod, mille semiparameetriline lähenemine riskimäärale avaldub järgmiselt (Tong *et al.* 2012: 132-139):

$$(9) \qquad h(t|X) = h_0(t)e^{X\beta}$$

kus $h(t|X)$ - riskimäär ajahetkel t sõltuvalt selgitava muutuja X vektorist,

$h_0(t)$ - baasrisk, mille kuju on määratlemata.

Elulemusanalüüsi abil on peale maksehäire esinemise tõenäosuse hindamise võimalik täiendavalt hinnata, millal maksehäire kõige tõenäolisemalt esineb. Teisisõnu

võimaldavad elukestusmudelid hinnata maksehäire esinemise tõenäosust mistahes ajahetkel vaatlusperioodi lõikes. (*Ibid*.: 132-139)

1.2. Eraisiku krediidiriski hindavate teadustööde tulemused

Baesensi ja tema kaasautorite hinnangul võib erialase kirjanduse põhjal järeldada, et klassifitseerimismeetodite kasutamine on hetkel enimlevinud lähenemine krediidiriski hindamise mudelite loomisel (Lessmann *et al.* 2013: 2). Mitmed eri autorid on uurinud erinevate klassifitseerimisalgoritmide efektiivsust krediidiriski hindamise seisukohalt, kus kasutatatakse erinevate mudelite prognoosivõime hindamiseks ja omavahel võrdlemiseks õigesti klassifitseeritud vaatluste osakaalu kõikidest vaatlustest (PCC – *„percentage correctly classified*"), suhtelise toimimise karakteristikute (ROC – „*receiver operating characteristics*") kõvera alust pindala (AUC – „*area under curve*"), H-näitajat („*H-measure*") ja Brieri skoori („*Brier Score*") (Lessmann *et al.* 2013: 29; Baesens *et al*. 2003: 631- 632; Paleologo *et al.* 2010: 490-499; West *et al.* 2005: 2543-2559). Üldistatult jaotuvad nimetatud näitajad kolme gruppi (Lessmann *et al.* 2013: 9):

- diskrimineerimisvõimet mõõtvad näitajad (AUC, H-näitaja),
- tõenäosuse hinnangute täpsust mõõtvad näitajad (Brieri skoor),
- kategooriliste hinnangute täpsust mõõtvad näitajad (PCC ja klassifitseerimisviga („*classification error*", „*error rate*")).

PCC, mis põhineb vigade maatriksil („*confusion matrix*"), on defineeritav kui korrektselt klassifitseeritud vaatluste arv jagatud kõikide vaatluste arvuga (PCC = (TP + TN) / (TP + TN + FP + FN)) (vt tabel 1) (Lessmann *et al.* 2013: 29).

Tabel 1. Vigade maatriks

		Hinnatud klass		TN - õige-negatiivne
		-1	+1	FN – vale-negatiivne
Tegelik klass	-1	TN	FP	FP – vale-positiivne
	+1	FN	TP	TP – õige-positiivne

Allikas: Lessmann *et al*. 2013: 29.

Nimetatud näitaja ei pruugi teatud juhtudel olla parim prognoosivõime hindamiseks, kuna eeldab võrdset valesti klassifitseerimise kulu nii vale-positiivsetele (FP – *false positive*) kui ka vale-negatiivsetele (FN – *false negative*) prognoosidele. Sellise eelduse tegemine on problemaatiline, kuna enamikel juhtudel on ühte tüüpi klassifitseerimisvea hind kõrgem kui teisel tüübil. (Baesens *et al.* 2003: 631)

ROC kõver on kahedimensiooniline graafiline esitus, mille Y-teljel on mudeli tundlikkus („*sensitivity*“, „*recall*“) ehk õige-postiivsete prognooside määr (TPR – „*true positive rate*“) ja X-teljel vastavalt õige-negatiivsete prognooside määr (TNR – „*true negative rate*“), mis saadakse spetsiifilisuse („*specificity*“) lahutamisel ühest, klassifitseerimise aluseks olevate erinevate piirväärtuste korral. Tundlikkus mõõdab õige-positiivsete prognooside osakaalu, mis avaldub kui õige-positiivsete (TP – „*true positive*“) prognooside suhe õige-positiivsetesse ja vale-negatiivsetesse prognoosidesse (TP/(TP+FN)). Spetsiifilisus on õige-negatiivsete (TN – „*true negative*“) prognooside suhe vale-positiivsetesse ja õige-negatiivsetesse prognoosidesse (TN/(FP+TN). (Baesens *et al.* 2003: 631)

ROC kõvera alune pindala on laialt kasutust leidnud prognoosivõime hindamise mõõt (Flach *et al*. 2015: 1). Näitaja hindab olukorra, kus juhuslikult valitud positiivne vaatlus klassifitseeritakse korrektselt kõrgemaks kui juhuslikult valitud negatiivne vaatlus, esinemise tõenäosust (*Ibid*.: 1). Näitaja AUC väärtused 1 ja 0.5 tähistavad vastavalt täiuslikku ja täiesti juhuslikku klassifitseerimist (Lessmann *et al.* 2013: 30). Handi hinnangul on ROC kõvera aluse pindala kasutamisel erinevate mudelite prognoosivõime hindamisel tõsine puudus (Hand 2009: 103). Nimelt on nimetatud autori hinnangul AUC fundamentaalselt seostamatu valesti klassifitseerimise kulude lõikes, kuna kasutab erinevaid valesti klassifitseerimise kulude jaotusi erinevate klassifitseerijate korral (*Ibid*.: 103). See on võrreldav erinevate mõõtühikute rakendamisega eri klassifitseerimismeetodite korral (*Ibid*.: 103). Mainitud puudus esineb olukorras, kus klassifitseerijat rakendades esinevad kahte eri tüüpi valesti klassifitseerimisel erinevad kulud (*Ibid*.: 103). Eelduseks on veel asjaolu, et mudeli kasutaja ei tea, milline on tegelik veaga seonduv kulu, vaid ta aimab, millised on tõenäolised valesti klassifitseerimisega seotud kulude määra väärtused („*values of the ratio of the misclassification costs*“) (*Ibid*.: 107).

Hand pakub ROC kõvera aluse pindala näitajaga seotud võtmeprobleemi, mis seisneb valesti klassifitseerimise kulude ja optimaalse klassifitseerimise piirväärtuse („*classification threshold*“) valiku vahel, lahendamiseks välja H-näitaja (Hand 2009: 105). H-näitaja on oodataval minimaalsel valesti klassifitseerimise kaol põhinev normaliseeritud klassifitseerija hinnang, mille väärtuse korral 0 on tegemist juhusliku klassifitseerijaga ja väärtuse 1 korral täiusliku klassifitseerijaga (Lessmann *et al.* 2013: 30). Brieri skoor on sündmuse toimumise tõenäosuste prognooside keskmise ruutvea mõõt, mida kasutatakse binaarse sündmuse korral prognoosivõime hindamiseks (Hamill, Juras 2006: 2906).

Baesens ja tema kaasautorid võrdlesid 41. erinevat klassifitseerimismeetodit prognoosimisvõime seisukohalt näitajate AUC, PCC, Bieri skoor ja H-näitaja lõikes, kasutades selleks andmestikku Baesensi ja tema kaasautorite poolt 2003. aastal avaldatud empiriilisest teadustööst, millele kaasati täiendavalt kaks uut märkimisväärse suurusega andmestikku. Mitmeid kasutatud meetodeid ei oldud krediidiriski hindamise kontekstis selleks ajaks kasutatud. (Lessmann *et al.* 2013: 4)

Mues ja Brown uurisid oma teadustöös kümne erineva klassifitseerija prognoosimisvõimet ROC kõvera aluse pindala alusel, kasutades selleks viite erinevat valimit. Uuritavateks klassifitseerimismeetoditeks olid logistiline regressioon, C4.5, tehisnärvivõrgud, gradientvõimendus („*gradient boosting*“), juhumetsad, lineaarne LS-SVM, k-lähima naabri meetod („k“ väärtuse kohal 10 ja 100), lineaarne ja mittelineaarne diskriminantanalüüs. (Brown, Mues 2012: 3446-3453)

2003. aastal B. Baesensi ja tema kaasautorite poolt avaldatud uurimuses võrreldi 17. klassifitseerimisalgoritmi prognoosimisvõimet kasutades selleks üle kaheksa erineva valimi näitajaid PCC ja AUC. Erinevatele klassifitseerimisalgoritmidele arvutati PCC alusel valimite ülene keskmine järk, arvestades lõikeväärtuse määramisel heade ja halbade lepingute suhet valimis. (Baesens *et al*. 2003: 632 - 633)

Uurimusest järeldati, et PCC seisukohalt on kõige kõrgema prognoosimisvõimega lineaarne tugivektor-masinate meetod. Leiti, et võrdluseks kasutatud näitaja järgi on lineaarse programmeerimisega, standardse tugivektor-masinate algoritmist tuletatud RBF LS-SVM-ga („*Least squares support vector machine with radial basis function*“),

lineaarse LS-SVM-ga, tehisnärvivõrkude meetodil ja diskreetsete muutujatega otsutuspuu algoritmiga C4.5 saadud tulemused praktiliselt samad. Tulemustest nähtub, et nii algoritm C4.5 kui ka algoritmi C4.5 tagasilõikamata puust loodud reeglite kaudu konstrueeritud mudeli korral on mõlemad meetodid enamasti parema prognoosimisvõimega disktreetsete muutujate korral. Siiski osutub valimi „Bene2" korral, kus heade laenude osakaal on 70%, algoritmi C4.5 tagasilõikamata puust loodud reeglite kaudu konstrueeritud mudel kõige kõrgema prognoosivõimega meetodiks, mille PCC on 69,7%, kusjuures sama meetodi vastav näitaja diskreetsete muutujate korral on 50,5%. Diskreetsete muutujatega C4.5 algoritm osutub meetoditest täpseimaks kõige enam tasakaalustamata klassidega valimi korral, kus on halbade laenude osakaal 10%, saades PCC väärtuseks 89,5%. (Baesens *et al.* 2003: 632-633)

Osaliselt toetab saadud tulemusi Baesensi ja kaasautorite poolt 2013. aastal avaldatud uurimus, kus osutusid PCC võrdluses kõrgeima prognoosivõimega individuaalseteks klassifitseerimismeetoditeks tehisnärvivõrkude meetod, lineaarne tugivektor-masinate meetod, ELM-K („*Kernalized ELM*"), reguleeritud logistiline regressioon („*regularized logistic regression*") ja RBF LS-SVM. Eelnevalt mainitud uurimustööga võrreldes ei saavutanud otsustuspuu meetod PCC näol kõrget prognoosimisvõimet. Nimelt oli algoritmi J4.8 PCC nelja valimi puhul pigem keskmine, kuid kolme korral madalaim, saades valimi „GMC" korral PCC väärtuseks 50% – täpseima meetodi vastavaks väärtuseks oli 86% (Lessmann *et al.* 2013: 32). Uurimustööst ei selgu, millised võiksid olla valimite ülese küllaltki suure suhtelise klassifitseerimistäpsuse kõikumise põhjused. J4.8 PCC väärtus on kõrgeim ehk 91,5% valimi „AC" korral (Lessmann *et al*. 2013: 32), kui 2003. aastal avaldatud uurimuses osutub otsutuspuu meetoditest sama valimi korral kõrgeima PCC väärtusega algoritmiks diskreetsete muutujatega algoritmi C4.5 tagasilõikamata puust loodud reeglite kaudu konstrueeritud mudel väärtusega 91,7% (Baesens *et al.* 2003: 633).

Kaasates PCC võrdlusesse ka homogeensete kogumike klassifitseerimismeetodid („*homogeneous ensemble classifiers*"), osutub nende sooritus enamasti individuaalsetest klassifitseerimismeetodiest paremaks (Lessmann *et al.* 2013: 35). Teostatud statistilisest analüüsist järeldub, et ühe erandiga on individuaalsete klassifitseerijate prognoosivõime märgatavalt madalam, kui juhumetsa meetodil, mis oli uuritud meetoditest täpseim

(Lessmann *et al.* 2013: 36-37). Nimelt ei olnud logistilise regressiooni korral piisavalt alust null hüpoteesi ümberlükkamiseks, mis oli püstitatud järgmiselt: klassifitseerija on statistiliselt võrdväärne juhumetsa meetodiga (Lessmann *et al.* 2013: 37). Uuringusse kaasati ka heterogeensed kogumike klassifitseerimismeetodid („*heterogeneous ensemble classifiers*"), millest täpseimaks osutus algoritm HCES-Bag („*hill-climbing ensemble selection with bootstrap sampling*"), kuid PCC võrdluses osutus juhumetsa meetod erinevate valimite üleselt keskmiselt paremaks (Lessmann *et al.* 2013: 37, 55, 56). Erinevalt eelmainitud teadustöö tulemustest, oli C. L. Devasena poolt koostatud võrdlusuuringus juhumetsa meetodi ja C4.5 algoritmi klassifitseerimistäpsus sarnaselt kõrge (Devasena 2015: 35).

2003. aastal avaldatud uurimustöös osutusid kehvema prognoosivõimega algoritmideks PCC järgi algoritmi C4.5 tagasilõikamata puust loodud reegli kaudu konstrueeritud mudel, mittelineaarne diskriminantanalüüs (QDA – „*quadratic discriminant analysis*"), naiivne Bayesi klassifitseerija („*naive Bayes classifier*") ja k-lähima naabri meetodil (k = 10 ja k = 100) (Baesens *et al*. 2003: 632 - 633). Kümme aastat hiljem järjeuuringuna avaldatud uurimustöös osutusid PCC osas madalama prognoosimisvõimega meetoditeks näiteks CART, J4.8, naiivne Bayesi klassifitseerija, k-lähima naabri meetod, mittelineaarne disktriminantanalüüs (Lessmann *et al.* 2013: 36).

Muesi ja Browni poolt tehtud võrdlusuuringu järgi oli ROC kõvera aluse pindala võrdluses pigem tasakaalus klassidega valimite (halbade laenude osakaal vastavalt 30%, 15% ja 10%) korral statistiliselt ($\alpha = 0{,}05$) kõrgema prognoosivõimega meetoditeks lineaarne LS-SVM, gradientvõimendus ja juhumetsad (Brown, Mues 2012: 3446-3453). Kui halbade laenude osakaal oli 2,5% ja 1%, saavutati parimaid tulemusi gradientvõimendusega, juhumetsadega ja k-lähima naabri meetodiga (k = 100), millest viimane oli tasakaalustatud klassidega valimite korral keskmise prognoosimisvõimega (Brown, Mues 2012: 3446-3453). Uurimusest nähtub, et tehisnärvivõrkude prognoosivõime on kõikide valimite korral keskmine ja teiste algoritmidega võrreldes paraneb C4.5 algoritmi suhteline prognoosimisvõime koos halbade laenude osakaalu langemisega valimis (Brown, Mues 2012: 3446-3453). Kui Muesi ja Browni võrdlusuuringus jäi tehisnärvivõrkude prognoosimisvõime pigem keskmiseks, siis Baesensi *et al.* 2003. aastal avaldatud uurimuses osutuvad ROC kõvera aluse pindala

võrdluses parima prognoosivõimega meetoditeks RBF LS-SVM ja tehisnärvivõrkude meetod, madalama prognoosivõimega meetoditeks aga mittelineaarne diskriminantanalüüs, lineaarne programmeerimine, otsustuspuu meetodid ja k-lähima naabri meetod (k = 10) (Baesens *et al.* 2003: 632-633). Baesens ja kaasautorid järeldasid saadud tulemustest, et enamik krediidiriski hindamisega seotud andmestikest on nõrgalt mittelineaarsed (Baesens *et al*. 2003: 632-633).

Sarnased tulemused saadi kümme aastat hiljem avaldatud krediidiriski hindamise klassifitseerimismeetodite võrdlusuuringus, kus individuaalsete klassifitseerijate seas osutusid AUC näitaja võrdluses parimateks meetoditeks RBF LS-SVM, logistiline regressioon, tehisnärvivõrgud ja Bayesi võrgustik („*Bayes network*“) (Lessmann *et al.* 2013: 31). Mõlemas teadustöös oli individuaalsete klassifikaatorite võrdluses kõrgeim AUC näitaja tehisnärvivõrkude meetodil (Lessmann *et al.* 2013: 31; Baesens *et al.* 2003: 634). Nii nagu PCC korral, osutusid ka ROC kõvera aluse pindala võrdluses homogeensete kogumike klassifitseerimismeetodite prognoosimisvõime individuaalsete klassifitseerijatega kõrvutades paremaks kuue valimi korral seitmest (Lessmann *et al.* 2013: 31). Ainsaks erandiks on logistiline regressioon, mille ROC kõvera alune pindala on võrreldes juhumetsa meetodiga, mis osutus parimaks homogeensete klassifitseerijate seast, 0,0005 võrra suurem, vastavalt 0,9315 ja 0,9310 (Lessmann *et al*. 2013: 31). Uurimusest järeldub, heterogeensete kogumike klassifitseerijate grupist on kõikide valimite lõikes täpseim HCES-Bag (Lessmann *et al.* 2013: 37). Tähelepanuväärne on asjaolu, et väga hea tulemuse saavutas ka HCES, mis on lihtsustatud versioon HCES-Bag-st (Lessmann *et al.* 2013: 37). Sarnaselt Baesensi ja tema kaasautorite 2003. aasta võrldusuuringule (Baesens *et al.* 2003: 627-635), osutusid ROC kõvera aluse pindala võrdluses madala prognoosimisvõimega klassifitseerimismeetoditeks mittelineaarne diskriminantanalüüs ja otsustuspuu meetod, kuid täiendavalt ka CART ja naiivne Bayesi klassifitseerija (Lessmann *et al.* 2013: 36). Nimetatutest viimane oli varasemas teadustöös AUC võrdluses pigem keskmise prognoosimisvõimega (Baesens *et al*. 2003: 634). Ka Muesi ja Browni koostatud uurimuses oli mittelineaarse diskriminantanalüüsi klassifitseerimistäpsus madal kõikide valimite lõikes, kuid C4.5 algoritmi prognoosimisvõime hinnati madalaks valimite korral, kus halbade laenuda osakaal oli 5% või kõrgem (Brown, Mues 2012: 3446-3453). Tähelepanuväärne on, et valimi, kus halbade laenude osakaal moodustas 1% kogu valimi mahust, korral osutus üheks

madalaima klassifitseerimistäpsusega meetodiks logistiline regressioon (*Ibid.*: 3446-3453), mis aga Baesensi ja tema kaasautorite uuringus oli ROC kõvera aluse pindala järgi hea klassifitseerimistäpsusega (Lessmann *et al.* 2013: 36). Uuringust nähtub, et mida väiksem on halbade laenude osakaal valimi mahust, seda madalam on võrreldes teiste meetoditega logistilise regressiooni prognoosivõime näitaja AUC seisukohalt (Brown, Mues 2012: 3446-3453).

H-näitaja võrdluses jääb meetodite järjestus prognoosimisvõime järgi hindamisel PCC ja AUC järjestusega võrreldes üldjoontes samaks- individuaalsete ja homogeensete kogumike klassifitseerijate grupis osutub parimaks juhumetsa meetod ning heterogeensete kogumike klassifitseerijate grupis vastavalt HCES-Bag. Uurimusest nähtub, et varem mainitud ROC kõvera aluse pindala kontseptuaalsed puudused ei kajastu suurel määral klassifitseerimismeetodite võrdluses. Klassifitseerijate järjestamine prognoosimisvõime järgi annab nii H-näitajat kui ka AUC-i kasutades peaaegu samad tulemused. Ühelt poolt on see indikatsiooniks, et praktikas on ROC kõvera aluse pindala kasutamine piisav, kuid teisalt ei ole selleks mõjuvat põhjust, kuna kontseptuaalselt sobivam näitaja eksisteerib. (Lessmann *et al.* 2013: 36-39)

Kui tehisnärvivõrkude prognoosimisvõime oli nii PCC, H-näitaja kui ka ROC kõvera aluse pindala võrdluses individuaalsete klassifitseerijate grupis parim, siis Brieri skoori järgi võib pidada klassifitseerija prognoosivõimet pigem madalaks. Tulemuse järgi grupi esimesteks on järjestatud meetodid nagu logistiline regressioon, lineaarne diskriminantanalüüs ja Bayesi võrgustik, kuid kaasates võrdlusesse ka homogeensed kogumiku klassifitseerijad, on kokkuvõttes parima prognoosimisvõimega juhumetsa meetod. Heterogeensete kogumiku klassifitseerijate grupis osutub sarnaselt teistele prognoosivõime hindamise näitajatele täpseimaks HCES-Bag. (Lessmann *et al.* 2013: 36-38)

Lessmann *et al.* (2013: 40-41) kõrvutasid tehtud võrdlusuuringus prognoosivõime seisukohalt iga kategooria parimat klassifitseerijat (individuaalsed, homogeensed ja heterogeensed kogumike klassifitseerijad) ja populaarsuse tõttu täiendavalt ka logistilist regressiooni. Selleks arvutati välja keskmine järjenumber iga meetodi jaoks kõikide valimite ja prognoosivõime näitajate üleselt (*Ibid.*: 40-41). Saadud tulemustest järeldati, et kõige täpsemad prognoosid saavutati HCES-Bag-ga, millele järgnevad juhumetsa

meetod, tehisnärvivõrkude meetod ja logistiline regressioon, millest viimane oli nimetatutest madalaima prognoosimisvõimega (*Ibid*.: 40-41). 2003. aastal B. Baesens *et al*. (2003: 634) poolt avaldatud teadustöös järeldati, et PCC ja AUC lõikes andsid parima tulemuse RBF LS-SVM ja tehisvärvivõrgud ning enamasti ei olnud logistilise regressiooni ja lineaarse diskriminantanalüüsi vastavad tulemused statistiliselt eelnimetatud meetodite tulemustest erinevad. Ong *et al.* (2005: 45) poolt tehtud uuringus võrreldi kahe valimi lõikes geneetilise programmeerimise, logistilise regressiooni, tehisnärvivõrkude, CART, C4.5 ja ebatasaste hulkade („*Rough sets*“) meetodite klassifitseerimistäpsust klassifitseerimisvea alusel. Tulemustest järeldati, et parima prognoosimisvõimega on geneetiline programmeerimine, millele järgnesid tehisnärvivõrgud, algoritm C4.5 ja logistiline regressioon (*Ibid*.: 45). Kuigi Baesens *et al*. (2003: 634) järeldab, et krediidiriskiga seotud andmestikud on vaid nõrgalt mittelineaarsed, siis nähtub tehtud ülevaatest, et parima prognoosivõimega meetoditeks on osutunud mittelineaarset seost modelleeritavad meetodid.

Autorite D. J. Hand ja W.E. Henley (1997) poolt avaldatud artikli, mis käsitles näiteks logistilist regressiooni, otsustuspuu ja närvivõrkude meetodeid, järgi ei leidu üldist parimat meetodit krediidiriski hindamiseks, vaid see sõltub lahendatavast probleemist. Meetodi valikut ja saadud tulemusi mõjutavad andmestruktuur, kasutatavad selgitavad muutujad ning klasside erinevus kasutatavatest muutujatest ja klassifitseerimiseesmärgist lähtuvalt. Näiteks võib eesmärgiks olla üldise või kuludega kaalutud valesti klassifitseerimise määra minimeerimine, teatud kasumlikkuse näitaja maksimeerimine. Kui klassid ei ole kergesti eristatavad ja pind, mis eristab klasse, ei ole täpselt hinnatav, on paindlike meetodite nagu tehisnärvivõrgute ja k-lähima naabri meetodi korral ülesobitamise oht, mida on võimalik lahendada silumise teel (nt. muutujale „k“ väga suure väärtuse andmine). Lisaks klassifitseerimistäpsusele, on täiendavalt mudeli juures oluline klassifitseerimiskiirus, minevikulise otsuse revideerimise kiirus ja klassifitseerimisotsuste mõistetavus. Lihtsasti interpreteeritavad klassifitseerimismeetodid nagu otstustuspuu meetodid on kasutajate jaoks tihti atraktiivsemad, kui „musta kasti“ meenutavad meetodid nagu näiteks tehisnärvivõrgud, kuna võimaldavad lihtsamini ratsionaalselt argumenteerida saadud otsuse üle. Tehisnärvivõrgud sobivad kasutamiseks hästi siis, kui puudub arusaam andmestruktuurist. (Hand, Henley 1997: 535-536)

Tabelis 2 kajastuvad erinevate autorite poolsed poolt- ja vastuargumendid erinevate krediidiriski hindamiseks enim kasutatavate meetodite kohta, mida arutleti põhjalikumalt krediidiriski hindamise metodoloogia peatükis ja mida käsitleti töös esitatud võrdlusuuringutes.

Tabel 2. Kokkuvõte krediidiriski hindamises kasutatavate meetoditest

Meetod	Tugevused/ nõrkused
Diskriminantanalüüs	• madal klassifitseerimistäpsus PCC ja AUC osas; • statistiliste eelduste, eriti selgitavate muutujate normaaljaotuse, täidetuse probleem.
Logistiline regressioon	• lineaarsus sõltumatute muutujate ja logaritmilise šansside suhte vahel, • pigem kõrge klassifitseerimistäpsus PCC ja AUC osas.
Lineaarne programmeerimine	• kõrge klassifitseerimistäpsus PCC alusel, kuid madal näitaja AUC järgi.
Tugivektor-masinad	• kõrge klassifitseerimistäpsus PCC ja AUC lõikes; • pikk treenimisaeg; • ebatäpse hüpertasandi loomise oht mitteoluliste muutujate ja suure andmemahu korral; • ei võimalda interpreteerida saadud tulemusi, kuna seost sõltuva ja sõltumatute muutujate vahel ei ole võimalik otseselt selgitada.
Tehisnärvivõrgud	• keskmine kuni kõrge klassifitseerimistäpsus PCC ja AUC järgi, kuid pigem madal Brieri skoori korral; • kritiseeritud kehva klassifitseerimistäpsuse pärast, kui mudelisse on kaasatud ebaolulisi muutujaid või valim on väike; • krediidiotsuse läbipaistmatus, kuna tegemist on tõlgendaja jaoks sisuliselt „musta kastiga“.
Otsustuspuu (C4.5)	• madal kuni kõrge klassifitseerimistäpsus näitajate PCC ja AUC järgi, • krediidiotsuse läbipaistvus ja tulemuste interpreteeritavus, kuna luuakse mudel.
K-lähima naabri meetod	• pigem madal klassifitseerimistäpsus PCC ja AUC osas; • meetodi lihtsus: • krediidiotsuse läbipaistmatus, kuna mudelit ei looda; • uute vaatluste grupeerimise ajakulukus suurte andmemahtude korral.
Juhumets	• pigem kõrge klassifitseerimistäpsus PCC ja AUC osas, • saadud mudeli tõlgendamise keerukus.

Allikas: autori koostatud.

Enamikes kasutatud artiklites on logistilise regressiooni klassifitseerimistäpsust peetud pigem kõrgeks nii PCC kui ka AUC osas, kuid Browni ja Muesi teadustöös osutus see

teiste kasutatud meetoditega võrreldes näitaja AUC järgi madalaks. Nii logistilise regressiooni, juhumetsa kui ka tugivektor-masinate klassifitseerimistäpsus on eraisiku krediidiriski hindamisel olnud kõrge. Tehisnärvivõrkude klassifitseerimistäpsus on kirjanduse ülevaate põhjal PCC järgi enamasti kõrge, kuid keskmine kuni kõrge AUC korral ja pigem madal Brieri skoori kasutades. K-lähima naabri meetodi korral on saadud häid tulemusi, kui halbade laenude osakaal valimis on olnud 1% või 2,5%, kuid enamasti on meetodi täpsus olnud pigem madal. Otsustuspuu algoritmi C4.5 klassifitseerimistäpsus kõigub erinevate teadusartiklite ja valimite lõikes oluliselt. Nimelt on klassifitseerija osutunud väga täpseks Baesens *et al.* (2003) poolt avaldatud teadustöös diskreetsete väärtuste korral näitaja PCC järgi, kuid kümme aastat hiljem Baesens *et al.* (2013) poolt avaldatud võrdlevas uuringus osutus klassifitseerimistäpsus pigem madalaks või keskmiseks. Ka on Ong *et al*. (2005) poolt tehtud töös otsutuspuu algoritm C4.5 võrreldes teiste klassifitseerijatega saavutanud häid tulemusi. Näitaja AUC alusel on meetodi klassifitseerimistäpsust hinnatud pigem madalaks, kuid keskmiseks, kui klasside tasakaalustamatus on suur.

Nii nagu autorid D. J. Hand ja W.E. Henley (1997) järeldasid, nähtub ka tehtud kokkuvõttest, et eraisiku krediidiriski hindamiseks ei leidu ühte parimat meetodit, vaid see sõltub mitmetest aspektidest peale klassifitseerimistäpsuse. Praktikas on tihti oluline saadud tulemuste interpreteeritavus, mille korral on eelistatavam kasutada näiteks otsustuspuude põhist lähenemist, kuigi klassifitseerimistäpsuse järgi on mitmetes uurimustöödes tehisnärvivõrkude, juhumetsa ja tugivektor-masinate kasutamine osutunud valimite üleselt keskmiselt täpsemaks.

1.3. Eraisiku krediidiriski hindamiseks kasutatavad muutujad

Mitmete klassifitseerimisprobleemide korral on üheks olulisemaks küsimuseks asjaolu, kui mitu sõltumatut muutujat mudelisse kaasama peaks (Hand, Henley 1997: 528-529). Kuna üldjuhul on andmemahud suured, siis on oht ülesobitumise (*„overfitting“*) esinemiseks väiksem, mille tõttu võidakse püüda kaasata nii palju muutujaid kui võimalik, mis tegelikkuses on piiratud praktiliste kaalutlustega (*Ibid.*: 528-529). Nimelt ei ole liiga paljude küsimustega taotluse või liialt pika kontrollprotseduuride korral klient valmis laenutaotlust esitama, mille tulemusena pöördub potentsiaalne deebitor

konkureeriva kreeditori poole (*Ibid.*: 528-529). Krediidiriski hindamisega seotud muutujate valimiseks on kolm enamkasutatavat lähenemist (*Ibid.*: 528-529).

- Ekspertide teadmiste ja kogemuste kasutamine sõltumatute muutujate valimise protsessis on mõeldud formaalsete statistiliste meetodite täiendusena. Viimane neist aitab välistada statistiliselt ebaoluliste üleliigsete muutujate jätmist mudelisse ajaloolistel ja/või subjektiivsetel põhjustel. Ekspertide tagasiside on oluline andmaks põhjendusi mudeli muutujate seisukohalt tehtud valikute kohta.
- Teiseks võimaluseks on sammuviisiliste statistiliste valikuprotseduuride (*„stepwise statistical procedure"*) rakendamine. Näiteks võib alustada sõltumatute muutujate lisamisega, kus igas sammus hinnatakse, millise sõltumatu muutuja lisamine parendas enim mudeli hindamistäpsust.
- Kolmandaks lähenemiseks on muutujate valimine kasutades mõõtu, mis iseloomustab maksehäires olevate ja mitteolevate lepingute jaotuse erinevust. Üheks selliseks mõõduks on näiteks informatsiooniline väärtus („*information value*").

Üldlevinud on arvamus, et ei leidu optimaalset muutujate arvu, mis sobib kasutamiseks mistahes krediidiriski hindamise mudeli loomisel (Abdou, Pointon 2011: 13). Sõltumatute muutujate valik varieerub uurimustes sõltuvalt andmete iseloomust ja sellest, millised kultuurilised ja majanduslikud muutujad võivad mõjutada mudeli kvaliteeti (*Ibid.*: 13). Järelikult on sõltumatute muutujate valik erinev ka riigipõhiselt (*Ibid.*: 13). Mitmetes uurimustöödes on kasutatud krediidiriski hindamise mudelite kontrueerimisel esialgu ligi 20. sõltumatut muutujat ja on läbi muutujate eemaldamise jõutud ligi 5 kuni 10 statistiliselt olulise muutujani (Hand *et al.* 2005: 684–690; Banasik, Crook 2010: 473-485; Avery *et al.* 2004: 835–856).

Autorite Jacobson ja Roszbach (2003: 615–633) uurimsutöö valimis oli kokku 57 erinevat muutujat. Valim koosnes krediidiandjatelt saadud, avalikult kättesaadavatest ja valitsuse poolt antud muutujatest nagu näiteks isiku sugu, kodakondsus, perekonnaseis, elukohaaadressi postikood, maksustatava tulu suurus, informatsioon kinnistu omandi kohta, tagatiseta laenude arv (*Ibid.*: 615–633). Lõplikust krediidiriski hindamise mudelist jäeti 41 muutujat välja, kuna puudus seos sõltuva muutujaga või esines väga

kõrge korrelatsioon kasutatava sõltumatu muutujaga, mis kirjeldas sama asja ja oli parema kirjeldusvõimega (*Ibid*.: 615–633). Šušteršič *et al*. (2009: 4736–4744) poolt kasutatavas valimis oli kokku 67 sõltumatut muutujat, millest statistiliselt oluliseks osutusid 21.

Mitmetel juhtudel ei ole uurimustesse kaasatud muutujate valikut selgelt argumenteeritud, vaid enamasti on andmestik antud uurijatele erinevate institutsioonide poolt. Järelikult sõltub sõltumatute muutujate valik krediidiriski hindamise mudelite konstrueerimisel eelkõige andmepakkujast ja nende andmete olemasolust. Sellise situatsiooni ohuks on selliste muutujate vaikimisi mõjusaks pidamine. (Abdou, Pointon 2011: 11)

Kirjanduse ülevaatest nähtub, et sõltumatute muutujate nagu vanus, perekonnaseis, praeguses elukohas elatud aeg, praeguse tööandja juures töötatud aeg, kliendisuhte kestus kreeditoriga, igakuine sissetulek, ülalpeetavate või laste arv, laenu tüüp ja elukoha tüüp on krediidiriski hindamise mudelites laialdaselt kasutusel olnud ja hindamise seisukohalt statistiliselt oluliseks osutunud (Abdou *et al*. 2008: 1275–1292; Avery *et al.* 2004: 835–856; Banasik, Crook 2010: 473-485; Bellotti, Crook 2009: 3302-3308; Hand *et al.* 2005: 684-690; Jacobson, Roszbach 2003: 615–633; Lee, Chen 2005: 743-752; Lee *et al.* 2002: 245–254; Marshall *et al.* 2010: 501–512; Šušteršič *et al*. 2009: 4736–4744; Tong *et al.* 2012: 132–139; Yap *et al*. 2011: 13274–13283).

Mida suurem on taotleja leibkonnas olevate laste arv, seda kõrgem on taotleja krediidirisk (Marshall *et al.* 2010: 506). Üllatavalt jõudis avaldatud teadusartiklis Yap *et al*. (2011: 13280) koostatud logistilise mudeliga vastupidisele tulemusele, mille järgi väheneb deebitori maksehäire esinemise tõenäosus ülalpeetavate arvu suurenedes. T. Jacobson ja K. Roszbach (2003: 624) järeldavad koostatud teadustöös, et kuigi tavaliselt seostatakse kõrgemat sissetulekut madalama riskitasemega, siis nende töös saadud empiirilised tulemused viitavad vastassuunalisele mõjule. Viidates asjaolule, et valimist on eelselektsiooni tõttu välja jäänud negatiivse otsuse saanud taotlejad, seavad nad teatud muutujate mõju suuna kahtluse alla (*Ibid*.: 624).

Kui laenu tüübiks on kaastaotlejaga laen, siis väheneb sellest tulenevalt koostatud elulemusanalüüsi järgi maksehäire esinemise risk 43% (Tong *et al*. 2012: 132–139).

Saadud tulemusi kaastaotleja olemasolu mõju suuna osas toetab ka T. Jacobsoni ja K. Roszbachi (2003: 624) loodud mudel. Koostatud MARS tüüpi krediidiriski hindamise mudelist järeldub, et tagamata laenuga kaasneb kõrgem maksehäire esinemise tõenäosus (Lee, Chen 2005: 743-752).

Empiirilised tulemused näitavad, et klientidel, kelle elukoha tüübiks ei ole isiklikus omandis olev maja või kes ei teeni igakuist palka on kõrgem tõenäosus maksehäire esinemiseks (Marshall *et al.* 2010: 506). Ka Bellotti ja Crooki (2009: 3302–3308) tehtud uurimuse tulemused toetavad eelnimetatud väidet, et kodu omanikul on võrreldes üürnikuks olemisega madalam krediidirisk. Mida pikemalt on taotleja elanud praeguses elukohas või töötanud praeguse tööandja juures, seda madalam on sellise kliendiga seonduv risk (Marshall *et al.* 2010: 506).

Koostatud mudeli järgi oli madalaim maksehäire esinemise tõenäosus vanusegrupis 40-64 eluaastat ja kõrgeim vastavalt üle 64 eluaasta (Avery *et al.* 2004: 835–856). Banasik ja Crook (2010: 473-485) poolt avaldatud elulemusanalüüsil põhineval uurimuse vanuse muutuja koefitsiendi funktsioonist nähtub, et alates 18. eluaastast kuni 24. eluaastani langeb risk jätkates laugemat langust kuni 49. eluaastani. Alates 49. eluaastast oli riski langus taaskord järsem ja alates 55. eluaastast hakkas tõusma (*Ibid.*: 473-485). B. W. Yap *et al*. (2011: 13280) poolt avaldatud teadusartiklis koostatud logistilise regressiooni mudeli järgi väheneb kliendi vananedes ka tõenäosus maksehäire esinemiseks.

Avery *et al.* (2004: 835–856) uurimusest tuleneb, et pikemaajaliselt abielus olnud taotlejatel oli 1,2 protsendipunkti väiksem tõenäosus maksehäire esinemiseks võrreldes klientidega, kes ei olnud kunagi abielus olnud. See viitab asjaolule, et pikemaajalises abielus olevad indiviidid on vähem avatud leibkonna sissetuleku häiretele, kuna leibkonnas on kaks sissetulekuallikat (*Ibid.*: 835–856). Hiljuti lahutanud taotlejate maksehäire esinemise tõenäosus on aga 2,2 protsendipunkti kõrgem, kui mitte kunagi abielus olnud isikutel (*Ibid.*: 835–856). Eelmainitud tulemustega vastuoluliselt järeldasid Yap *et al*. (2011: 13274–13283) logistilise regressiooni mudeli tulemustest, et abielus olevatel isikutel on kõrgem tõenäosus maksehäire esinemiseks, kui mitte abielus olevatel isikutel.

Teatud osa muutujatest nagu praeguses elukohas elatud aeg ja praeguse tööandja juures töötatud aeg peegeldavad taotleja stabiilsust, teisalt krediitkaardi omamine, hoiusekonto olemasolu ja kliendisuhte pikkus kreeditoriga iseloomustavad pigem finantsteadlikkust. Elukoha tüüp, tegevusala ja elukaaslase tegevusala võiksid anda ülevaate lepinguosapoolega seotud ressurssidest ning kaudselt viitavad kuludele muutujad nagu laste või ülalpeetavate arv. (Thomas *et al.* 2002: 5)

Ka on mitmes krediidiriski hindamise mudelis maksehäire tõenäosuse esinemise hindamise seisukohalt statistiliselt oluliseks osutunud taotleja sugu, tegevusala, haridustase, igakuised kohustused, koduse lauatelefoni numbri esitamine taotlusel, laenusumma, laenu pikkus, laenu eesmärk ja kliendi krediidiajaloo kohta registritesse tehtud päringute arv (Abdou *et al*. 2008: 1275–1292; Banasik, Crook 2010: 473-485; Bellotti, Crook 2009: 3302-3308; Jacobson, Roszbach 2003: 615–633; Lee *et al.* 2002: 245–254; Marshall *et al.* 2010: 501–512; Šušteršič *et al*. 2009: 4736–4744; Tong *et al.* 2012: 132–139; Yap *et al*. 2011: 13274–13283). Ilmselt on üheks levinumaks stiliseeritud faktiks naiste madalam krediidiriski tase võrreldes meestega, kuid kui laenuandjad suudavad mudelisse kaasata muutujaid, mis soost tingitud riskiallikaid kajastavad, siis soo muutuja statistiline kirjeldusvõime langeb oluliselt (Schreiner 2004: 11). Levinud arusaama toetab Yap *et al*. (2011: 13280) poolt genereeritud logistiline mudel, mille kohaselt on naiste maksehäire esinemise tõenäosus madalam kui meestel.

Kõrgema haridustasemega isikud on oluliselt madalama krediidiriskiga (Kočenda, Vojtek 2009: 15-16). Nii on näiteks keskharidusega isikute puhul suurem tõenäosus maksehäire esinemiseks, kui keskerihariduse omajatel (*Ibid*.: 15-16). Kõrg- ja keskerihariduseta isikutel on tööjõuturul keerulisem saada hästi tasustatud töökohta ja sellele lisaks on neil suurem tõenäosus jääda töötuks, kui ettevõtte, piirkonna või riigi majanduslik olukord halveneb (*Ibid*.: 15-16). Üllatavalt leidsid A. Marshall *et al*. (2010: 506) avaldatud teadustöös, et tudengite riskitase on madalam kui mittetudengitel. Koostatud MARS tüüpi krediidiriski hindamise mudelist lähtub, et mida suurem on laenusumma või lepingu kuumakse ja sissetuleku suhe, seda kõrgem on maksehäire esinemise tõenäosus (Lee, Chen 2005: 743-752). Elulemusanalüüsi tulemusena järeldati, et koduse lauatelefoni numbri esitamine taotlusel vähendab maksehäire esinemise riski 47% (Tong *et al*. 2012: 132–139).

Erialases kirjanduses vähemkasutatud sõltumatuteks muujateks on näiteks kliendi krediidiajalugu, vähemusrahvuse samas piirkonnas elamine, taotlejale väljastatud kreeditoriga seotud pangakaartide arv, ettevõtte omamise asjaolu, laenuga soetatava toote hind, omaosalus laenuga soetatava toote hinnast, teenindava konto debiteerivate ja krediteerivate transaktsioonide summade suhe (Avery *et al*. 2004: 835–856; Marshall *et al*. 2010: 501–512). Tulemused kliendi ja panga suhet iseloomustavate sõltumatute muutujate kohta viitavad asjaolule, et pika nõudmiseni hoiuse konto ajalooga ja mitut pangakaarti omavatel taotlejatel on madalam maksehäire esinemise tõenäosus (Marshall *et al*. 2010: 506).

Kliendi eelnevate maksehäirete ajalugu on tugevalt seotud laenu maksekäitumisega. Nii kõikide maksehäirete arv kui ka maksehäire esinemine viimase kuue kuu jooksul suurendavad oluliselt tõenäosust, et kliendil esineb ka tulevikus maksehäire. Ka leiti, et kreeditoriga seotud pika nõudmiseni hoiuse konto ajalooga ja mitme pangakaardiga klientidel on madalam krediidirisk. (Marshall *et al*. 2010: 501–512)

Koostatud mudelist järeldati, et maksehäire esinemise tõenäosus on tugevas statistilises seoses taotleja elukohas elavate vähemusrahvustesse kuuluvate elanike osakaaluga piirkonna elanikkonda (Avery *et al*. 2004: 835–856). Seda põhjendati asjaoluga, et vähemusrahvusest leibkonnad on haavatavamad erinevatele maksevõimet mõjutavatele sündmustele (*Ibid*.: 835–856). Klientidel, kes ostavad laenuga odavamaid tooteid ja kelle omapoolne sissemakse moodustab suurema osa toote hinnast, on madalam tõenäosus maksehäire esinemiseks (Marshall *et al*. 2010: 506).

Kokkuvõtteks võib öelda, et krediidiriski hindamise mudelite seisukohalt ei leidu optimaalset sõltumatute muutujate arvu, kuid enamasti jääb praktikas lõplikusse mudelisse kaasatud statistiliselt oluliste muutujate arv vahemikku 5-10. Samuti erinevad selgitavad muutujad sisuliselt erinevate mudelite raames. Tehtud erialase kirjanduse ülevaatest nähtub, et krediidiriski hindamise seisukohalt on kõrge kasutussagedusega sõltumatuteks muutujateks vanus, perekonnaseis, praeguses elukohas elatud aeg, praeguse tööandja juures töötatud aeg, kliendisuhte kestus kreeditoriga, sissetulek, ülalpeetavate või laste arv, laenu tüüp ja elukoha tüüp (vt tabel 3). Kõik nimetatud tunnused on vähemalt nelja kasutatud artikli korral lõplikes mudelites statistiliselt oluliseks osutunud. Keskmise ja madala kasutussagedusega sõltumatuteks muutujateks

on näiteks taotleja sugu, tegevusala, haridustase, laenusumma ja kliendi krediidiajalugu (vt tabel 4).

Tabel 3. Erialases kirjanduses eraisikute krediidiriski hindamisel kõrge kasutussagedusega statistiliselt oluliseks osutunud sõltumatud muutujad

Sõltumatu muutuja	Allikad
vanus	Abdou et al. 2008; Avery et al. 2004; Banasik, Crook 2010; Bellotti, Crook 2009; Hand et al. 2005; Jacobson, Roszbach 2003; Lee et al. 2002; Tong et al. 2012; Yap et al. 2011;
perekonnaseis	Abdou et al. 2008; Avery et al. 2004; Banasik, Crook 2010; Jacobson, Roszbach 2003; Lee, Chen 2005; Yap et al. 2011;
praeguses elukohas elatud aeg	Banasik, Crook 2010; Hand et al. 2005; Marshall et al. 2010; Tong et al. 2012;
praeguse tööandja juures töötatud aeg	Banasik, Crook 2010; Hand et al. 2005; Marshall et al. 2010; Tong et al. 2012;
kliendisuhte kestus kreeditoriga	Avery et al. 2004; Bellotti, Crook 2009; Hand et al. 2005; Marshall et al. 2010;
igakuine sissetulek	Abdou et al. 2008; Jacobson, Roszbach 2003; Lee et al. 2002; Šušteršič et al. 2009;
ülalpeetavate või laste arv	Banasik, Crook 2010; Marshall et al. 2010; Tong et al. 2012; Yap et al. 2011;
laenu tüüp	Banasik, Crook 2010; Jacobson, Roszbach 2003; Lee, Chen 2005; Tong et al. 2012;
elukoha tüüp	Abdou et al. 2008; Banasik, Crook 2010; Bellotti, Crook 2009; Jacobson, Roszbach 2003; Marshall et al. 2010; Tong et al. 2012;

Allikas: autori koostatud.

Ootuspäraselt on teadustöödes lõplikesse mudelitesse kaasatud sõltumatute muutujate seas mitmeid demograafilisi muutujaid ning kreeditori ja kliendi vahelise suhte pikkust iseloomustav tunnus. Mõneti üllatavalt on madala sagedusega kasutust leidnud taotleja krediidiajalugu, kuid see võib olla tingitud asjaolust, et esialgsed kreeditoride poolt rakendatud laenuandmise põhimõtted on muutujat arvestanud ja halva krediidiajalooga taotlejad välistanud.

Tabel 4. Erialases kirjanduses eraisikute krediidiriski hindamisel keskmise ja madala kasutussagedusega statistiliselt oluliseks osutunud sõltumatud muutujad

Sõltumatu muutuja	Kasutus-sagedus*	Allikad
sugu	keskmine	Banasik, Crook 2010; Lee et al. 2002; Yap et al. 2011;
tegevusala	keskmine	Lee et al. 2002; Yap et al. 2011;
haridustase	keskmine	Abdou et al. 2008; Marshall et al. 2010;
igakuised kohustused	keskmine	Abdou et al. 2008; Marshall et al. 2010;
koduse lauatelefoni numbri esitamine	keskmine	Abdou et al. 2008; Banasik, Crook 2010; Tong et al. 2012;
laenusumma	keskmine	Abdou et al. 2008; Banasik, Crook 2010; Marshall et al. 2010;
laenuperiood	keskmine	Banasik, Crook 2010; Šušteršič et al. 2009;
laenu eesmärk	keskmine	Banasik, Crook 2010; Tong et al. 2012;
kliendi krediidiajaloo päringute arv registritesse	keskmine	Bellotti, Crook 2009; Jacobson, Roszbach 2003;
kliendi krediidiajalugu	madal	Marshall et al. 2010;
vähemusrahvuste osakaal taotleja elukoha piirkonnas	madal	Avery et al. 2004;
väljastatud kreeditoriga seotud pangakaartide arv	madal	Marshall et al. 2010;
ettevõtte omamine	madal	Jacobson, Roszbach 2003;
teenindava konto deebit ja kreedit kannete suhe	madal	Marshall et al. 2010;
laenuga soetatava toote hind	madal	Marshall et al. 2010;
omaosalus laenuga soetatava toote hinnast	madal	Marshall et al. 2010;

Märkused: Kasutussagedus tähistab artiklite arvu, kus muutuja lõplikus mudelis oluliseks osutus (keskmine – kaks või kolm korda; madal – üks kord).
Allikas: autori koostatud.

Soo keskmine kasutussagedus mainitud artiklites on samuti mõneti üllatav arvestades asjaolu, et naiste madalam krediidiriski tase võrreldes meestega on M. Schreineri järgi üks stiliseeritumaid fakte krediidiriski hindamisel. Vähem kasutust leidnud muutujate hulgas on näiteks taotleja elukohas elavate vähemusrahvustesse kuuluva elanike osakaal piirkonna elanikkonda. Üheks vähese kasutatavuse põhjuseks võib pidada sedalaadi informatsiooni kättesaamatust kreeditori jaoks.

2. ERAISIKU KREDIIDIRISKI HINDAMISE EMPIIRILINE UURIMUS ETTEVÕTTES KAUPMEHE JÄRELMAKS OÜ

2.1. Ettevõtte Kaupmehe Järelmaks OÜ tutvustus ja ülevaade töös kasutavast andmestikust

Kaupmehe Järelmaks OÜ on 2010. aastal asutatud ettevõte, mille majandusaasta aruandes näidatud põhitegevusalaks on Äriregistri teabesüsteemi järgi "Muu laenuandmine, v.a pandimajad" (Äriregistri teabesüsteem 2016). UNO Järelmaks on BIGBANKi gruppi kuuluva ettevõtte Kaupmehe Järelmaks OÜ teenus, mida reklaamitakse kodulehel kui alternatiivset võimalust eraisikute sisseostude rahastamiseks (UNO Järelmaks 2015). Lisaks Eestile pakutakse UNO Järelmaksu teenust ka Lätis ja Leedus (UNO Järelmaks 2015). UNO Järelmaksu taotlemine toimub eraisiku jaoks läbi koostööpartnerite, kes on kliendi poolt soetatava kauba või teenuse müüjaks.

2011. aastal kehtisid ettevõttes Kaupmehe Järelmaks OÜ järgmised laenuandmise põhimõtted (UNO järelmaksu... 2011):

- Laenusaajal on maksehäireregistri andmetel lõpetamata või vähem kui kuus kuud tagasi lõpetatud pangandus- või muu finantseerimisvõlg, sealhulgas SMS või kiirlaenu võlg.
- Käendajal on maksehäireregistri andmetel lõpetamata või vähem kui kuus kuud tagasi lõpetatud pangandus- või muu finantseerimisvõlg, sealhulgas SMS või kiirlaenu võlg.
- Käendajal on muu lõpetamata võlg peale pangandus- või finantseerimisvõla.
- Laenusaaja või käendaja ainsaks või põhiliseks sissetulekuallikaks on vanemahüvitis, mis on alampalgast väiksem või sellega võrdne.

- Laenusaaja või käendaja töötab katseajaga, välja arvatud juhul kui klient on enne katseajaga töötamist töötanud teisel töökohal ning eelnevalt ja olemasolevalt töökohalt saadud palgalaekumiste vahe ei ole rohkem kui kaks kuud.
- Laenusaaja laenukalkulaatori skoor on väiksem kui 275. Vastavalt pädevusele võib antud piirangus teha erandeid maa krediidikomitee eraisikulaenude alamkomitee või maa krediidikomitee.
- Isikul on ettevõttega kehtiv leping ja vähemalt üks viimasest neljast maksegraafikujärgsest maksest on tasutud rohkem kui viie päevase hilinemisega. Kui kliendi leping on kehtinud lühema tähtaja jooksul, peavad korrektselt tasutud olema kõik möödunud tähtpäevaga maksed.
- Isikul on ettevõttega kehtiv leping, millest tulenevalt on võlg üle viie euro.
- Isikul on ettevõttega kehtiv leping, mida on viimase nelja kuu jooksul kaetud, välja arvatud maksepäevade muutmiseks või tegeliku laenu väljastamiseks tehtud katmised.

Kõik eespool nimetatud asjaolud takistasid isikul laenu saamist, välja arvatud laenukalkulaatori skoori piirang, mille korral võidi rakendada erandi tegemise õigust. Siinkohal on oluline mainida, et ettevõtte konfidentsiaalsuspoliitika tõttu ei ole võimalik käesolevas töös esitada krediidiskoori arvutuskäiku. Krediidireeglite tulemusena on teatud osa valimist kunstlikult välja jäänud, mille tulemusena oleks valimis nihe, kui käsitleda seda representatiivse valimina kõigi taotlejate kohta.

Töös kasutatav andmestik pärineb ettevõtte Kaupmehe Järelmaks OÜ pangatarkvara andmebaasist. Valimi mahuks on 3901 vaatlust, mis on moodustatud juhuvalimina 2011. aasta jooksul Eestis nimetatud ettevõtte ja kliendi poolt sõlmitud järelmaksulepingutest. Ettevõttelt saadud juhuvalimis ei esine ühtegi klienti, kellel oleks rohkem kui üks leping. 2011. aastal esitatud kõikidest järelmaksutaotlustest moodustasid negatiivse otsusega taotlused ligi 40%. Eraisiku krediidiriski hindamise mudeli seisukohalt on ettevõtte jaoks oluline uute taotlejate võimalikult täpne klassifitseerimine, kuid arvestades varemtehtud eelselektsiooni taotlejate osas, on uusi mudeleid luues võimalik kasutada informatsiooni taotlejate kohta, kes said eelnevalt positiivse otsuse ja otsustasid laenu võtmise kasuks. Kuna järgneva analüüsi käigus hakatakse eristama kliente, kes esialgse ettevõttes kasutusel olnud laenuandmise

põhimõtete järgi klassifitseeriti ekslikult mitteprobleemseteks, siis on loodavate mudelite klassifitseerimistäpsus maksehäirega klientide osas tõenäoliselt madalam, kui kasutusel olnud mudelil.

Valimis on iga lepingu kohta kokku 14 sotsiaal-demograafilist, käitumuslikku, finantsilist ja lepinguga seotud tunnust (vt tabel 5), millest kategooriliste tunnuste võimalikud väärtused on leitavad töö lisadest (vt lisa 1).

Tabel 5. Valimi muutujate kirjeldus

Tunnus	Nimetus mudelis	Tüüp
sugu	gender	kategooriline
vanus taotlemisel	ageWhenApplying	diskreetne
perekonnaseis	maritalStatus	kategooriline
haridustase	education	kategooriline
ülalpeetavate arv	numberOfDependants	kategooriline
elukoha tüüp	residenceType	kategooriline
postiaadressi maakond	postalAddressCounty	kategooriline
postiaadress linn	postalAddressTownArea	kategooriline
tegevusala	activity	kategooriline
taotluse hetkel töötatud aeg kuudes (praegusel ametikohal)	workedAtMonths	diskreetne
kuine sissetulek eurodes	monthlyIncome	pidev
maksehäirete arv taotlemise hetkel	paymentDisturbances	kategooriline
laenusumma eurodes	loanAmount	pidev
laenuperiood kuudes	loanPeriodMonths	diskreetne

Allikas: autori koostatud.

Valimisse kaasatud muutujate valikul lähtuti tunnustest, mis olid varasemates eraisiku krediidiriski hindamisega seotud teadustöödes kasutust leidnud. Täiendavalt kaasati andmestikku postiaadressi maakond ja linn, kuna need muutujad võiksid sarnaselt muutujaga „taotleja elukohas elavate vähemusrahvustesse kuuluvate elanike osakaal piirkonna elanikkonda“ peegeldada deebitori haavatavust erinevatele maksevõimet mõjutavatele sündmustele. Täiendavaks kitsenduseks oli muutujate selekteerimisel andmekvaliteet ja ettevõtte poolt andmebaasi salvestatud tunnuste valik.

Valimis esinevaks uuritavaks tunnuseks on maksehäire esinemine või mittesinemine, mille nimetus mudelis on „default“. Koostöös ettevõtte Kaupmehe Järelmaks OÜ

krediidiriski hindamisega tegelevate isikutega, defineeriti töö raames maksehäire, kui olukord, kus klient on lepingu esimese kaheteistkümne kuu jooksul ühe või mitme maksega üle 90. päeva viivituses olnud. Sõltuv muutuja on tüübilt binaarne ehk fiktiivne muutuja, mille väärtus saab olla 0 või 1. Laenulepingutele, millel on nimetatud definitsiooni järgi esinenud maksehäire, on omistatud muutuja väärtuseks 1 ja vastasel juhul on muutuja väärtuseks 0. Ühe või mitme maksega esimese kaheteistkümne kuu jooksul üle 90. päeva viivituses olnud lepinguid on valimis kokku 246, mis moodustab valimist 6,3% (vt tabel 6).

Tabel 6. Muutujate kirjeldav statistika

Tunnus	Keskväärtus	Standard-hälve	Mediaan	Min.	Maks.
default	0,063		0,00	0,00	1,00
loanAmount	511,575		468,00	125,00	2818,00
loanPeriodMonths	20,810		18,00	3,00	60,00
monthlyIncome	644,730		580,00	130,00	2500,00
ageWhenApplying	37,914		37,00	18,00	72,00
workedAtMonths	67,101		47,00	1,00	498,00

Allikas: autori koostatud.

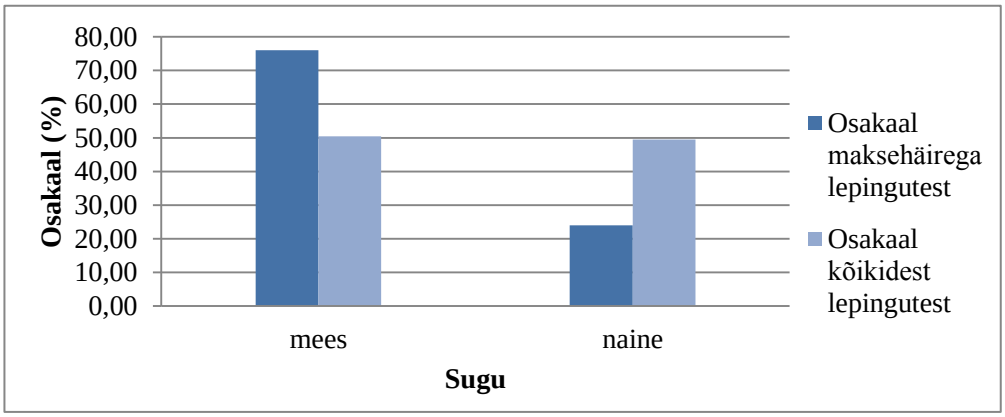
Maksehäirega kliendi vanus taotlemise hetkel oli keskmiselt 31,29 aastat, sissetulek 602,63 eurot, praegusel ametikohal töötatud aeg 35,34 kuud, laenusumma 641,46 eurot ja laenuperiood 27,2 kuud (vt lisa 3). Kliendid, kellel ei esinenud maksehäiret, olid keskmiselt 38,36 aastat vanad ja töötanud praegusel ametkohal pea poole pikemalt ehk 69,24 kuud (vt lisa 2). Ka oli selliste deebitoride keskmine laenusumma ja -periood vastavalt 502,83 eurot ja 20,38 kuud.

Järgmisena antakse ülevaade valimis olevate kateroogiliste tunnuste väärtuste jaotusest maksehäirega ja kõikide lepingute lõikes. Täiendavalt seotakse empiirilises osas kasutatavate tunnuste oodatav mõju suund kirjanduse ülevaatega. Mitmete töös kasutatud artiklite raames oli kirjeldatud ühesuunalist mõju uuritavale tunnusele sõltumatu muutuja erinevate väärtuste korral. Selle põhjuseks on lineaarset seost eeldavad meetodid nagu näiteks logistiline regressioon, mille eelduseks on lineaarsus sõltumatute muutujate ja logaritmilise šansside suhte vahel. Käesoleva töö empiirilises osas kasutatava otsustuspuu meetodi korral võib sõltumatu muutuja mõju suund olla ühes

sõlmpunktis positiivne, teises aga negatiivne. Sellegipoolest püstitatakse erialase kirjanduse põhjal valimis kasutatavate sõltumatute muutujate mõju suunale ootused.

MARS tüüpi krediidiriski hindamise mudeli tulemustest (Lee, Chen 2005: 743-752) lähtudes on ootuspärane, et töö empiirilises osas koostatud mudelites toob laenusumma suurenemine kaasa kõrgema eraisiku krediidiriski. Sarnaselt B. W. Yap *et al*. (2011: 13280) ja Avery *et al.* (2004: 835–856) teadusartiklites saadud tulemustele võiks taotleja vanuse tõustes maksehäiresse sattumise tõenäosus pigem langeda. Autori poolt oodatav taotleja praeguse tööandja juures töötatud aja mõju suund ühtib Marshall *et al.* (2010: 506) tehtud järeldustega, mille järgi on kliendiga seonduv risk madalam, kui ta on pikemalt töötanud sama tööandja juures.

Töö raames kasutatavas valimis on laenulepingu klientideks soo järgi 1932 naist ja 1969 meest (vt lisa 4). Jooniselt 4 nähtub, et lepingud, mille kliendiks on naine, moodustavad maksehäirega lepingutest 23,98% ja meeste korral on sama näitaja väärtus 76,02%.

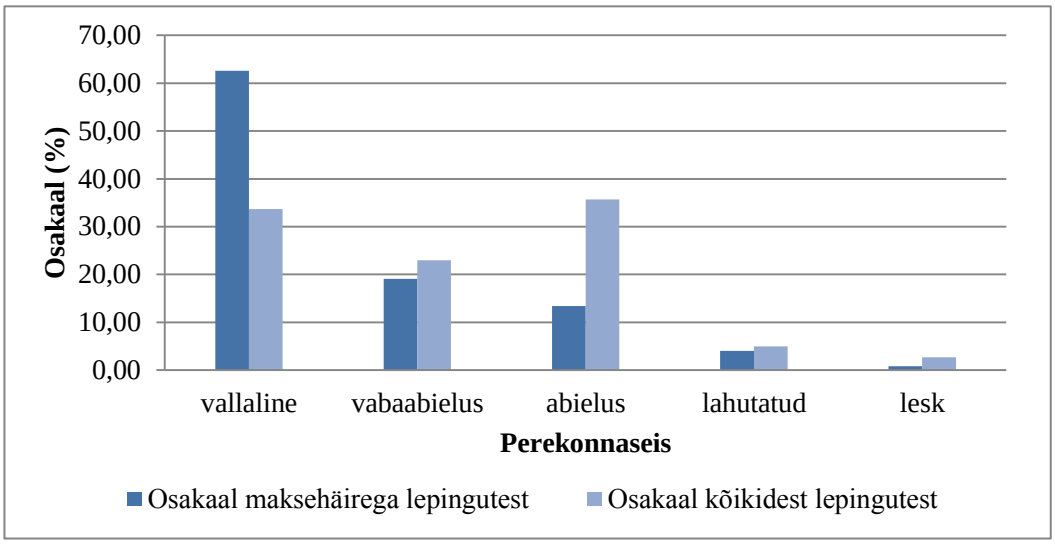


Joonis 4. Erinevast soost lepinguliste klientide osakaal maksehäirega ja kõikidest lepingutest (autori koostatud).

Autori ootus ühtib sarnaselt Schreineri (2004: 11) ja Yap *et al*. (2011: 13280) poolt avaldatud tulemustele, mille järgi on naise krediidiriski tase võrreldes meestega madalam.

Tunnus „maritalStatus“ indikeerib kliendi perekonnaseisu, mille taotleja valis laenutaotluse peal olemasolevatest valikutest. Vallaliste klientidega lepingute, kus on

esinenud maksehäire, osakaal kõikidest maksehäirega lepingutest on 62,60% ja kõikide vallaliste kliendidega lepingute osakaal valimist on 33,66% (vt joonis 5).

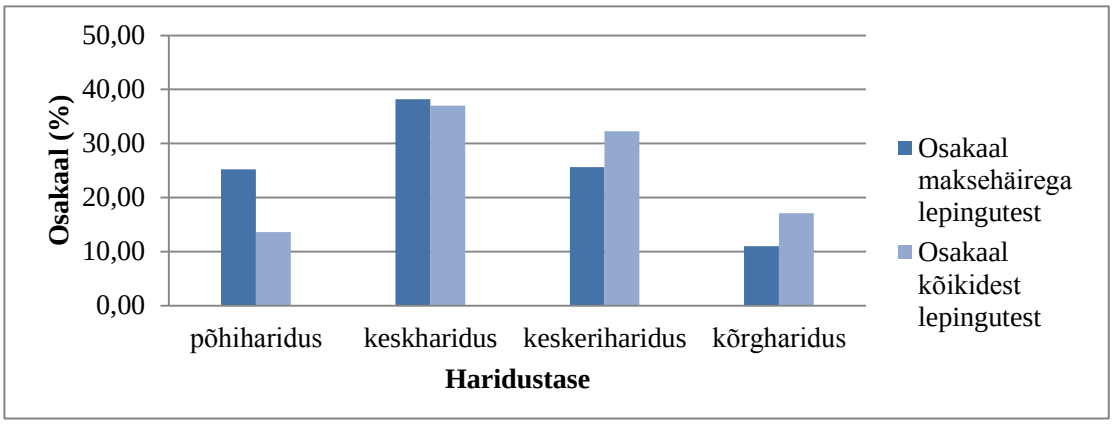


Joonis 5. Maksehäire esinemise osakaal valimist lepingulise kliendi perekonnaseisu järgi (autori koostatud).

Abielus klientidega lepingute osakaal moodustab lepingutest 35,68%, kusjuures maksehäirega lepinguga abielus klientide osakaal kõikidest maksehäirega lepingutest on 13,41%. Vabaabielus, lahutatud ja lesestunud klientidega lepingute, kus on esinenud maksehäire, osakaal maksehäirega lepingutest on vastavalt 19,11%, 4,07% ja 0,81%. Kirjanduse ülevaatest nähtus, et abielus olemine omas erinevates artiklites vastassuunalist mõju. Autori hinnangul võiks koostatavatest mudelitest tuleneda, et abielus olevad taotlejad on võrreldes mitte kunagi abielus olnud deebitoridega parema maksekäitumisega nagu järeldas ka Avery *et al.* (2004: 835–856). Põhjendusena võib tuua välja väiksema avatuse leibkonna sissetulekute häiretele, kuna töö kaotuse korral püsib üks sissetulekuallikas. Samasugune põhjendus on teatud olukordades relevantne ka vabaabielus olevate taotlejate korral. Sellest tulenevalt võib arvata, et perekonnaseisu muutuja kirjeldusvõime langeb, kui valimisse oleksid kaasatud tunnused, mida muutuja kaudselt peegeldab.

Kõrgharidusega klientide lepingud moodustavad maksehäirega lepingutest 10,98% ja kõikidest lepingutest 17,10% (vt joonis 6). Keskharidusega klientide lepingute osakaal maksehäirega lepingutest ja kõikidest lepingutest on peaaegu sama, nimelt 38,21% ja

36,99%. Seevastu põhiharidusega klientide korral on sama näitaja vastavalt 25,20% ja 13,64% ning keskeriharidusega klientidel 25,61% ja 32,27%.

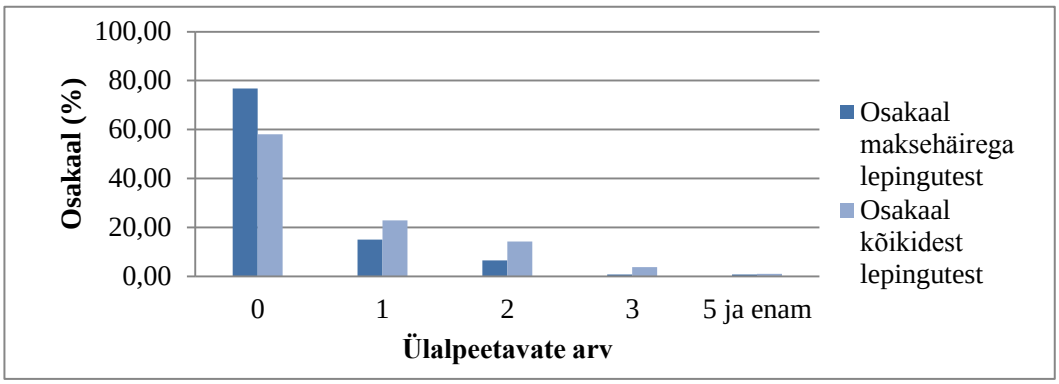


Joonis 6. Maksehäire esinemise osakaal valimist lepingulise kliendi haridustaseme järgi (autori koostatud).

Töö autor nõustub autorite E. Kočenda ja M. Vojteki (2009: 15-16) poolse põhjendusega kõrgema haridustasemega isikute madalama krediidiriski kohta, mille kohaselt on kõrgema haridustasemega deebitoridel lihtsam saada kõrgemalt tasustatud töökohta ja väiksem tõenäosus jääda töötuks ettevõtte, piirkonna või riigi majandusliku seisundi halvenedes. Käesoleva magistritöö autori jaoks üllatavalt järeldasid A. Marshall *et al*. (2010: 506), et tudengite riskitase on madalam kui mittetudengitel, kuid argumenteeritud arutluskäiku saadud tulemustele ei pakutud.

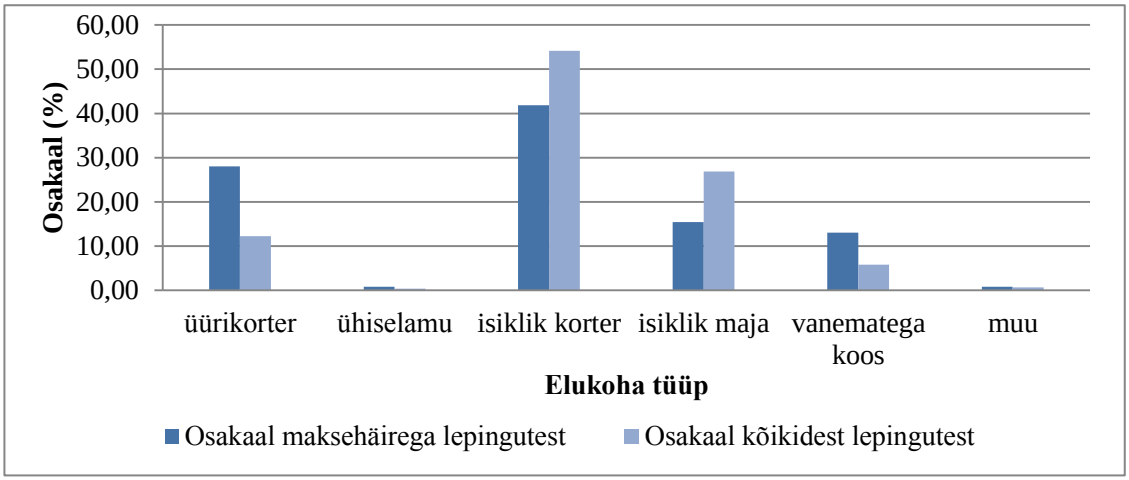
Ülalpeetavate arv on tunnus, mis on defineeritud kui kliendi poolt ülalpeetavate laste ja täiskasvanute arv. Kõrgeima osakaalu valimist ehk 58,01% moodustavad lepingud, mille klientidel ei olnud taotlemise hetkel ülalpeetavaid, ja madalaima osakaalu ehk 1,08% lepingud, mille klientidel oli 5 või rohkem ülalpeetavat (vt joonis 7). Mitte ühegi ülalpeetavaga, ühe, kahe, kolme, viie ja enama ülalpeetavaga klientide lepingute osatähtsus maksehäirega lepingutest on vastavalt 76,83%, 15,04%, 6,50%, 0,81% ja 081%. Valimis ei ole ühtegi lepingut, mille klient oleks taotlusel valinud ülalpeetavate arvuks neli. Erinevalt Yap *et al*. (2011: 13280) poolt saadud tulemustele, mille järgi maksevõime paraneb ülalpeetavate arvu suurenedes, ühtib autori ootus ülalpeetavate arvu mõju suuna osas Marshall *et al.* (2010: 506) poolt järeldatuga, mille järgi tõuseb eraisiku krediidirisk leibkonnas olevate laste arvu suurenedes. Kuna erinevalt Marshall

et al. (2010: 506) poolt kasutatud tunnusele kuuluvad käesoleva töö valimis ülalpeetavate alla lisaks täiskasvanud isikud, keda taotleja majanduslikult üleval peab, siis ei ole tunnused üheselt võrreldavad. Sellegipoolest viitavad nii ülalpeetavate laste kui ka täiskasvanute arv kaudselt taotlejaga seotud kuludele.



Joonis 7. Maksehäire esinemise osakaal valimist lepingulise kliendi ülalpeetavate arvu järgi (autori koostatud).

Lepingute, mille kliendi elukoha tüübiks on isiklik korter, osatähtsus kõikidest lepingutest on kõrgeim ehk 54,17% ja osakaal maksehäirega lepingutest on 41,87% (vt joonis 8).

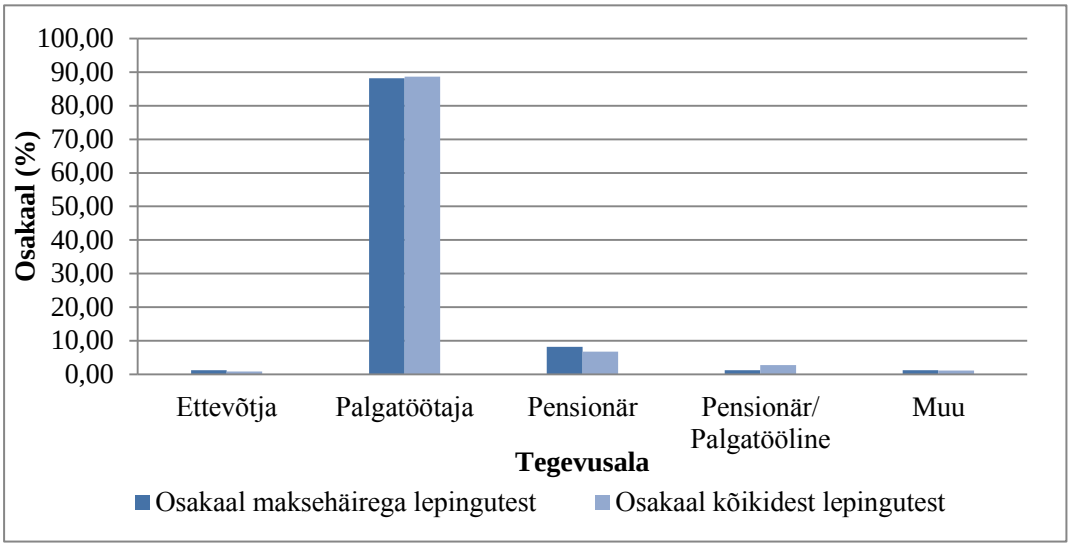


Joonis 8. Maksehäire esinemise osakaal valimist lepingulise kliendi elukoha tüübi järgi (autori koostatud).

Isiklikus majas, üürikorteris või vanematega majas elavate klientide lepingute osakaal on maksehäirega ja kõikide lepingute lõikes vastavalt 15,45% ja 26,86%, 28,05% ja

12,23% ning 13,01% ja 5,77%. Lepingute, mille klientide elukoha tüübiks on märgitud muu või ühiselamu, osatähtus jääb alla 1% mõlema nimetatud grupi arvestuses. Kirjanduse ülevaatele tuginedes on ootuspärane, et töö empiirilises osas kujuneb isiklikus omandis oleva korteri või majaga taotlejate krediidirisk võrreldes üürnikuks olejatega madalamaks nagu järeldasid teadustöös ka Bellotti ja Crook (2009: 3302–3308). Sama seisukohta toetab osaliselt ka Marshall et al. (2010: 506) poolt tehtud järeldus, mille järgi hinnati maja omavate deebitoride maksehäiresse sattumise tõenäosust madalamaks.

Kliendi tegevusala järgi on andmestikus enim lepinguid, mille kliendi tegevusalaks oli taotlusel märgitud palgatöötaja. Osatähtus kõikidest ja maksehäirega lepingutest on vastavalt 88,67% ja 88,21% (vt joonis 9).



Joonis 9. Maksehäire esinemise osakaal valimist lepingulise kliendi tegevusala järgi (autori koostatud).

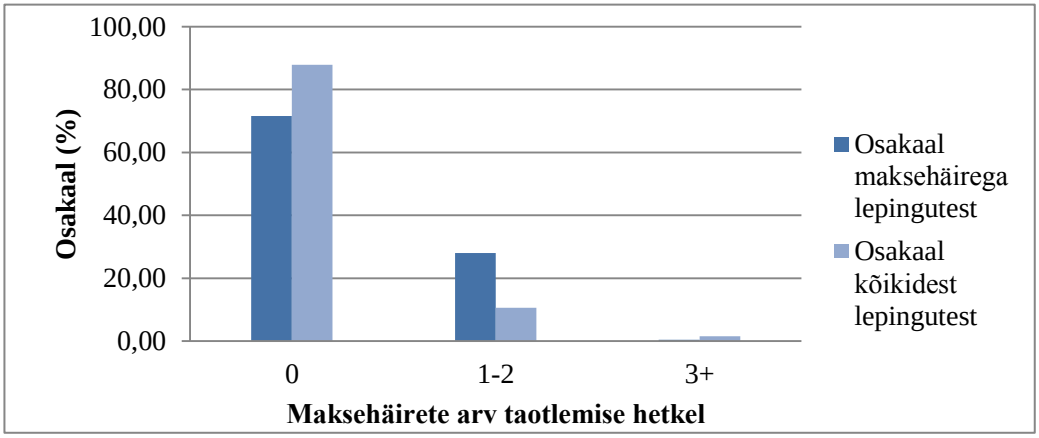
Valikute muu tegevusala, pensionär ja ettevõtja osakaalud maksehäirega lepingute ja kõikide lepingute võrdluses on küllaltki võrdsed. Kui lepingud, mille tegevusala valik on pensionär/palgatööline, moodustavad kõikidest lepingutest 2,69%, siis osatähtsus maksehäirega lepingutest on 1,22%.

Tunnus postiaadress linn tähistab kliendi poolt esitatud postiaadressijärgset linna, kus on eraldi eristatud Eesti suurimad linnad nagu Tallinn, Pärnu, Tartu, Narva, Kohtla-Järve ja Viljandi. Ülejäänud asulad/ piirkonnad on koondatud väärtuse "Muu" alla, kuna

ühegi väljajäänud haldusüksuse osakaal valimist ei ületanud eraldiseisvana ühte protsenti ja mittekodeerimine omaks analüüsile tõenäoliselt negatiivset mõju. Nii maksehäirega kui ka kõikide lepingute lõikes on suurima osakaaluga, nimelt 36,59% ja 49,53%, lepingud, mille postiaadressi linn ei ole üks selle töö raames eristatud Eesti suurtematest linnades. Tallinna ja Tartu korral on nimetatud osakaalud vastavalt 43,90% ja 31,43% ning 12,20% ja 7,77%. (vt lisa 6)

Postiaadressi maakond on tuletatud kliendi poolt taotlusel esitatud postiaadressist. Harjumaa postiaadressiks märkinud klientide lepingud moodustavad maksehäirega lepingutest 50,82% ja kõikidest lepingutest 43,89%. Tartumaa korral on samade näitajate väärtused vastavalt 15,04% ja 11,77%. Kui lepingute, mille kliendi postiaadressijärgseks maakonnaks on Pärnumaa, osakaal kõikidest lepingutest on 9,25%, siis osatähtsus maksehäirega lepingutest on 3,25%. (vt lisa 5)

Tunnus maksehäirete arv taotlemise hetkel näitab, kui mitu aktiivset maksehäiret, mille tüübiks ei olnud pangandus- või muu finantseerimisvõlg, oli kliendil ettevõtte Krediidiinfo AS eraisikute maksehäirete registris (Krediidiinfo 2016) taotluse esitamise hetkel. Kõige rohkem ehk 87,90% on valimis vaatlusi, mille lepingute klientidel ei olnud taotlemisel ühtegi maksehäiret (vt joonis 10).



Joonis 10. Maksehäire esinemise osakaal valimist lepingulise kliendi taotlemise hetkel olevate maksehäirete arvu järgi (autori koostatud).

Lepingud, mille klientidel ei olnud välise registri andmetel ühtegi maksehäiret, moodustavad maksehäirega lepingutest 71,54%. Ühe või kahe registreeritud

maksehäirega klientide lepingute osatähtsus maksehäirega ja kõikidest lepingutest on 28,05% ja 10,59%, kusjuures samad näitajad kolme või enama maksehäirega klientide korral on vastavalt 0,41% ja 1,51%. Töö raames ei ole deebitori maksehäirete arvu taotlemise hetkel kodeeritud, vaid ettevõtte salvestas tunnuse selliste väärtustega infosüsteemi. Sarnaselt Marshall *et al*. (2010: 501–512) poolt avaldatud teadustöös järeldatule on ootuspärane, et deebitoride, kellel ei esinenud taotlemise hetkel registri järgi ühtegi eelmainitid definitsiooni järgset maksehäiret, krediidirisk osutub empiirilises osas koostatavates mudelites võrreldes maksehäires olevate klientidega madalamaks.

2.2. Eraisiku krediidiriski modelleerimine otsustuspuu meetodil

Käesolevas töös kasutatakse ettevõtte Kaupmehe Järelmaks OÜ poolt antud valimi alusel eraisiku krediidiriski hindamiseks tarkvarapaketis R algoritmi J48 (Hornik *et al.* 2015: 18), mis on Java programmeerimiskeele põhine implementatsioon J.R. Quinlani poolt väljatöötatud C4.5 otsustuspuu meetodist (Quinlan 1993: 23).

Algoritm C4.5 põhineb informatsiooniteooria konseptsioonidel (Baesens *et al*. 2003: 631), mille kohaselt on Claude E. Shannoni poolt välja arendatud entroopia andmestikus olevate juhuslike muutujate vaheline määramatuse mõõt, mis viitab ühtlasi informatsiooniallika tulemusest saadavale keskmisele informatsiooni hulgale (Mazid *et al*. 2010: 298). Olgu p_1 klassi 1 ja p_o klassi 0 kuuluvate vaatluste osakaal valimis *S*, mille entroopia arvutatakse välja järgmiselt (Baesens *et al*. 2003: 631):

$$(10) \qquad Entroopia(S) = -p_1 \log_2(p_1) - p_0 \log_2(p_0)$$

kus S *- valim,*

p_i *- vaatluse i osakaal valimis.*

Muutuja p_0 väärtus saadakse p_1 lahutamisel ühest (p_0 = 1- p_1). Entroopia on maksimaalne ehk võrdne ühega, kui $p_1 = p_0 = 0{,}5$, minimaalne ehk võrdne nulliga, kui p_1 või p_0 on võrdne nulliga.

Infohulga suurenemine on defineeritud, kui oodatava entroopia vähenemine sõlmpunkti hargnemisel tunnuse x_i alusel: (Baesens *et al*. 2003: 631)

$$(11) \quad \textit{Infohulga suurenemine}(S, x_i) = \textit{Entroopia}(S) - \sum_{v \in \textit{väärtused}(x_i)} \frac{|S_v|}{|S|} \textit{Entroopia}(S_v)$$

kus S *- valim,*

S_v *- S-st moodustatud alamvalim konkreete* x_i *väärtuse kohal,*

x_i *- hargnemistunnus.*

Algoritm eelistab paljude erinevate väärtustega tunnuste hargnemisi, kui sõlmpunktide hargnemise üle otsustamiseks kasutatakse infohulga suurenemise kriteeriumit. Kui valimis on näiteks üheks tunnuseks vaatluse unikaalne identifitseeriv väärtus, siis valitakse see tunnus algoritmi poolt parimaks hargnemise otsuseks. Nimetatud probleemi lahendamiseks rakendatakse algoritmi C4.5 korral normaliseerimist kasutades infohulga suurenemise määra hargnemiskriteeriumina, mis avaldub järgmisel kujul: (Baesens *et al*. 2003: 631)

$$(12) \quad \textit{Infohulga suurenemise määr}(S, x_i) = \frac{\textit{Infohulga suurenemine}(S, x_i)}{-\sum_{k \in \textit{väärtused}(x_i)} \frac{|S_k|}{|S|} \log_2 \frac{|S_k|}{|S|}}$$

kus S *- valim,*

S_k *- S-st moodustatud alamvalim konkreete* x_i *väärtuse kohal,*

x_i *- hargnemistunnus.*

Valemist nähtub, et C4.5 algoritm eelistab suurima infohulga suurenemisega hargnemisi täiendava piiranguga, milleks on tingimus, et infohulga suurenemine peab olema vähemalt sama suur kui keskmine infohulga suurenemine üle kõikide jälgitud hargnemiste (Baesens *et al*. 2003: 631). Kärpimata puu luuakse iga sõlmpunkti kohta järgmist pseudokoodi rakendades (Ruggieri 2002: 438-439):

1. Arvutatakse välja iga klassi kaalutud sagedus valimis *S*, kuhu kuuluvad selle sõlmpunktiga seotud vaatlused.
2. Kui alamvalimis esineb üks klass või on teatud väärtusest vähem vaatlusi, tagastatakse leht vastava klassiga ja algoritm peatub, vastasel juhul luuakse sõlmpunkt *N*.

3. Iga tunnuse *A* kohta arvutatakse välja infohulga suurenemise määr. Pidevate muutujate korral jagatakse osavalim tunnuse *A* väärtuse järgi kaheks lokaalset lõikeväärtust kasutades.
4. Suurima infohulga suurenemise määraga tunnus valitakse sõlmpunkti hargnemiseks.
5. Pideva tunnuse korral leitakse lõikeväärtus kõikide õpiandmete seast. Nimelt määratakse väärtuseks suurim tunnuse väärtus, mis on väiksem kui kohalik lõikeväärtus.
6. Tulemusena on sõlmpunktil *s* arv hargnemisi ja $S_1,...,S_s$ on alamvalimid peale hargnemist. Kui tegemist on pideva tunnuse, on hargnemisi kaks, kategoorilise tunnuse korral on hargnemisi nii palju, kui on valimis *S* tunnuse *A* korral erinevaid väärtuseid.
7. Järgmise sammuna hinnatakse iga alamvalimi S_i (i=[1, s]) korral, kas see on tühi. Kui alamvalimisse ei kuulu ühtegi vaatlust, siis määratakse alanev sõlmpunkt leheks andes klassi väärtuseks sõlmpunktiga *N* seotud enim esinenud klass ja klassifitseerimisveale omistatakse väärtuseks 0. Kui S_i ei ole tühi, rakendatakse valimil S_i, millele kaasatakse ka *A* tunnuse puuduvate väärtustega vaatlused, rekursiivselt sama pseudokoodi.

Kirjeldatud otsustuspuu meetodi poolt rakendatav algoritm „jaga ja valluta“ (*„divide and conquer“*) jaotab tavaliselt valimid väiksemateks alamvalimiteks kuni igale lehele vastab üks klass või täiendav vaatluste jaotamine ei ole võimalik, kuna kahel või enamal vaatlusel on iga tunnuse lõikes samad väärtused, kuid erinev klass. Kui ühegi alamvalimi korral ei esine viimasena nimetatud klassifitseerimisprobleemi, siis klassifitseerib otsustuspuu korrektselt kõik õpiandmetega seotud vaatlused, mille korral on tegemist ülesobitumisega. Ülesobitumist saab vältida peatumiskriteeriumit rakendades, mis põhineb enamasti mõnel statistilise olulisuse testil. Sellisel juhul ei jaotata teatud alamvalimeid enam väiksemaks. Alternatiivne lahendus ülesobitumisega tegelemiseks on teatud osa otsustuspuu struktuuri eemaldamine peale puu loomist. Enamik autoreid peavad viimast lähenemist eelistatavamaks, kuna see võimaldab erinevate tunnuste vahelist potentsiaalset koosmõju uurida ja alles peale seda otsustada tulemuste jätmise või eemaldamise kasuks. C4.5 algoritm rakendab neist viimast. (Kohavi, Quinlan 1999: 6)

Peale esialgne puu loomist „jaga ja valluta“ algoritmi poolt, kärbitakse seda ülesobitumise probleemi lahendamiseks liikudes mööda puud alt üles. Puu kärpimisega seotud algoritmi selgitamiseks defineeritakse esmalt ära otsustuspuu *T*, mis on õpiandmetest loodud valimile *S* vastav mitte ühegi otsese lehega otstustuspuu. Nimetatud puu hargneb tunnuse *A* alusel alanevateks puudeks, millest iga puud T_i on eelnevalt kärbitud. Valimis *S* kõige rohkem esinenud tunnuse *A* väärtusega seotud alanev puu on tähistatud kui T_f ja *L* on leht, millele on omistatud valimis *S* enim esinenud klass. Meetodi C4.5 puu kärpimise algoritmis kasutatakse hinnatud tõelisi veamäärasid („*estimated true error rate*“) $U_{CF}(E_T,|S|)$, $U_{CF}(E_{Tf},|S|)$ ja $U_{CF}(E_L,|S|)$, kus *CF* viitab kasutatavale usaldusnivoole ning E_T, E_{Tf} ja E_L on vastavalt *T*, T_f ja *L* korral valesti klassifitseeritud vaatluste arv. Puu *T* jäetakse muutmata kujule, kui $U_{CF}(E_T,|S|)$ on nimetatud määradest madalaim, kuid asendatakse lehega *L*, kui E_L osutub madalaimaks. Sarnaselt asendatakse *T* alaneva puuga T_f juhul, kui hinnatud vigade määr $U_{CF}(E_{Tf},|S|)$ on kolmest madalaim. (Kohavi, Quinlan 1999: 6-8)

Otsustuspuu mudelite genereerimiseks ja hindamiseks C4.5 meetodil jaotatakse esialgne andmestik kaheks, nimelt treening- ja testvalimiks. Selle pinnalt tõstatub koheselt dilemma, kuna võimalikult hea klassifitseerija loomiseks on vaja kasutada treenimise etapis nii palju andmeid kui võimalik. Sama kehtib ka testvalimi kohta, sest suurem andmemaht võimaldab saada parema hinnangu loodud mudeli klassifitseerimistäpsusele. Juhuslike alamvalimite koostamisel rakendatakse kahte erinevate strateegiat, millest esimese korral muudetakse test- ja treeningvalimite suhet üldvalimisse. Täiendavalt rakendatakse uuritava tunnuse klasside suhte piirangut, mille käigus jaotatakse juhuslikult andmekirjeid alamvalimitesse piiranguga, et säiliks algandmetega võrreldes võimalikult täpne maksehäirega lepingute suhe valimisse, milleks on 6,3%. Tulemusena saadi järgmised valimid:

- treening- ja testvalim moodustavad esialgsest valimist vastavalt 50% ja 50% (strateegia tunnus on S1),
- treening- ja testvalim moodustavad esialgsest valimist vastavalt 60% ja 40% (strateegia tunnus on S2),
- treening- ja testvalim moodustavad esialgsest valimist vastavalt 70% ja 30% (strateegia tunnus on S3),

- treening- ja testvalim moodustavad esialgsest valimist vastavalt 50% ja 50% ning uuritavate klasside suhe säilib (strateegia tunnus on S4),
- treening- ja testvalim moodustavad esialgsest valimist vastavalt 60% ja 40% ning uuritavate klasside suhe säilib (strateegia tunnus on S5),
- treening- ja testvalim moodustavad esialgsest valimist vastavalt 70% ja 30% ning uuritavate klasside suhe säilib (strateegia tunnus on S6).

Kõikide treeningvalimite alusel genereeritakse J48 algoritmiga otsustuspuu mudel. Täiendavalt koostatakse täieliku esialgse valimi pealt sama algoritmi kasutades mudel M7, mille prognoosivõimet hinnatakse jagades esialgne valim kümneks eraldiseisvaks testvalimiks, kus on üritatud säilitada klasside esialgset suhet. Ülesobitumise lahendamiseks kärbitakse esialgset puud, kasutades usaldusnivoo väärtusena 0.25, mis on ühtlasi sisendparameetri vaikimisi väärtus. B. Baesensi ja kaasautorite poolt avaldatud teadustöös kasutati meetodi C4.5 rakendamisel nimetatud parameetri jaoks sama väärtust (Baesens *et al*. 2003: 632). Töö teoreetilises osas käsitleti mitmeid eraisiku krediidiriski prognoosimisega seotud mudelite klassifitseerimistäpsuse hindamiseks mõeldud indikaatoreid, kuid käesolevas töös kasutatakse selleks PCC ja ROC kõvera alust pindala. Täiendava informatsiooni andmiseks tuuakse mudeli kohta välja lisaks näitajad TPR ja TNR, kuna klasside suure tasakaalustamatuse korral ei peegelda PCC adekvaatselt klassifikaatori prognoosivõimet klasside eristamise seisukohalt. Nimetatud meetrikud valiti, kuna need peegeldavad mudeli prognoosivõime erinevaid külgi ning H-näitaja kasutamine ei ole põhjendatud, sest mudelite loomisel ei arvestata kulumomenti. Kuna ROC kõvera loomisel eeldatakse üldjuhul hinnatud tõenäosuse olemasolu iga vaatluse kohta, siis tekib otsustuspuu korral tavapärase mitmete punktidega kõvera asemel ühe punktiga kõver.

Tabelis 7 kajastuvad töö raames loodud seitse mudelit ning nendega strateegia järgi seotud treening- ja testvalimid. Kuna mudeli M7 treenimisel kasutatakse algvalimit treeningvalimina, siis ei kajastu tabelis nimetatud mudeli korral uuritava klasside osakaale. Valimites, mis on loodud esialgset klasside suhet säilitava strateegia kaudu, moodustavad maksehäirega lepingud nii treening – kui ka testvalimis üldvalimiga sarnaselt 6,3% (vt tabel 7). Mudeli M2 korral osutub uuritava klassi osakaal võrreldes

algandmete sama näitajaga kõige suuremaks, nimelt erineb testvalimis meetrik 0,5 protsendipunkti võrra.

Tabel 7. Koostatud mudelid, seotud valimi strateegiad ja uuritava klassi jaotus mudeli alamvalimites

Mudeli nimi	Valimi strateegia	Uuritava klassi osakaal treeningvalimis	Uuritava klassi osakaal testvalimis
M1	S1	6,1%	6,5%
M2	S2	6,0%	6,8%
M3	S3	6,4%	6,1%
M4	S4	6,3%	6,3%
M5	S5	6,3%	6,3%
M6	S6	6,3%	6,3%
M7	algvalim	-	-

Allikas: autori koostatud.

Kõikides töö raames genereeritud mudelites osutus oluliseks muutuja ageWhenApplying (vt tabel 8). Kordagi ei leidnud kasutust muutujad maritalStatus, numberOfDependants, residenceType, postalAddressCounty, activity ja paymentDisturbances. Mõned muutujad nagu näiteks perekonnaseis, ülalpeetavate arv, elukoha tüüp ja tegevusala, mille oluliseks osutumist töö teoreetilise osa alusel oleks põhjust oodata, võisid mudelitest välja jääda seetõttu, et nende tunnuste teatud väärtustega vaatluste sattumine töös kasutatavasse valimisse välistati laenuotsuse hetkel kehtinud krediidipoliitika tõttu. Ühe sellise näitena võib välja tuua maksehäirete arvu taotlemisel. Nimelt ei väljastatud krediidireeglite järgi laenu kliendile, kellel oli maksehäireregistri andmetel lõpetamata või vähem kui kuus kuud tagasi lõpetatud pangandus- või muu finantseerimisvõlg. Ootuspäraselt leiab mitmete mudelite raames kasutust ka taotleja sugu ja haridustase. Kõikides mudelites, peale M1 ja M7, on oluliseks muutujaks osutunud loanAmount, mis teoreetilise osa alusel oli lõplikesse mudelitesse kaasatud keskmise sagedusega. Tegur postiaadressi linn esines vaid mudelis M3.

Tähelepanuväärne on, et teemakohases kirjanduses leidis laenuperiood keskmise sagedusega kasutust, kuid käesolevas töös on nimetatud tunnus peale kärpimist otsustuspuu sisemise sõlmpunktina alles jäänud viie mudeli korral seitsmest. Samas olid tunnused praeguse tööandja juures töötatud aeg ja igakuine sissetulek artiklites tihedalt

kasutuses, kuid loodud mudelites osutusid need oluliseks ühel kuni kahel korral. Mõningane kasutust leidnud muutujate kõikumine mudelite lõikes võib viidata asjaolule, et mõni hea kirjeldamisvõimega karakteristik on valimist välja jäänud. Eelneva põhjal võib järeldada, et nihkega valimi korral sõltub konkreetsete muutujate oluliseks osutumine osaliselt sellest, milliseid selektsioone on tehtud laenuandmisel.

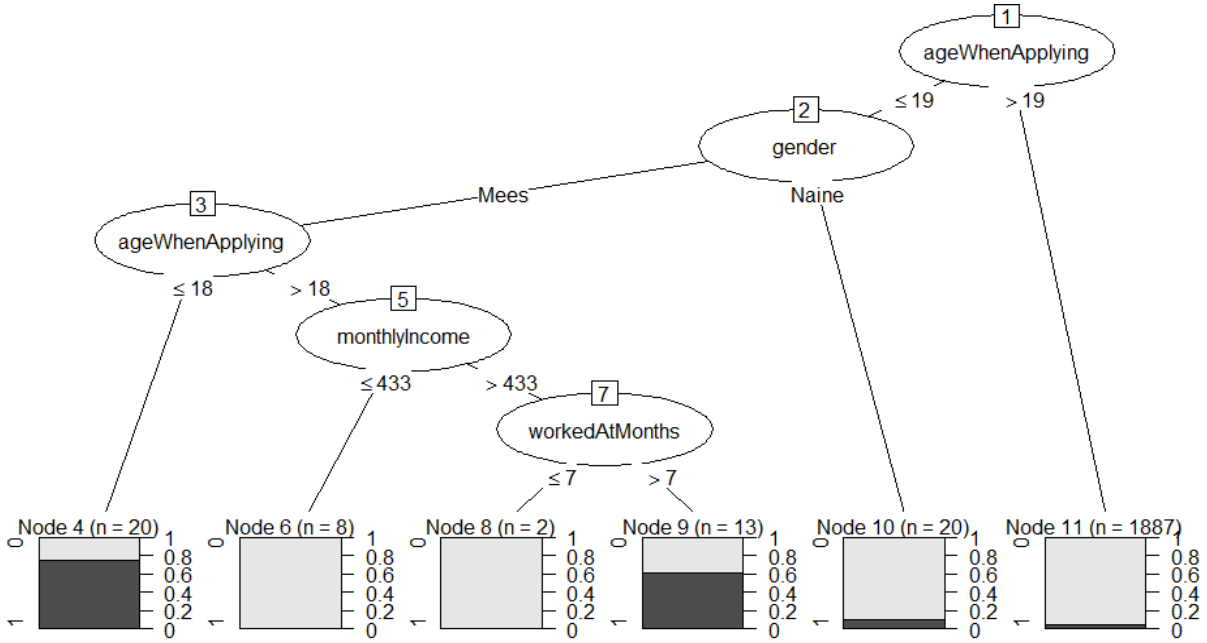
Tabel 8. Mudelites oluliseks osutunud muutujad kõikide muutujate lõikes

Mudel/ Muutuja	M1	M2	M3	M4	M5	M6	M7
ageWhenApplying	X	X	X	X	X	X	X
gender	X		X	X	X	X	X
loanAmount		X	X	X	X	X	
loanPeriodMonths		X	X	X	X		X
education		X	X		X		X
monthlyIncome	X		X		X		
workedAtMonths	X		X				
postalAddressTownArea			X				
maritalStatus							
numberOfDependants							
residenceType							
postalAddressCounty							
activity							
paymentDisturbances							

Allikas: autori koostatud.

Kuna otsustuspuude korral on oluline erinevate muutujate väärtuste kombinatsioon, siis vaadeldakse järgmisena detailsemalt eraldi igat mudelit. Mudeli M1 esimeseks hargnemiskriteeriumiks on osutunud kliendi vanus taotlemisel (vt joonis 11). Klassifitseerimispuu järgi ei esine maksehäiret kõikidel klientidel, kellel on vanust vähemalt 20 eluaastat. Siiski toob selline lihtsustus kaasa ebapuhta lehe, kus treeningvalimis kuulub lehe alla 1887 maksehäireta ja 94 maksehäirega klienti, mis moodustab 79% treeningvalimis olevatest probleemsetest lepingutest. Tulemus on osaliselt tingitud asjaolust, et klasse vaadeldakse võrdsetena ja samal ajal on klasside tasakaalustamatus kõrge. Võib väita, et saadud tulemus ei kattu suuresti erialases kirjanduses leiduvaga, kuna näiteks Banasik ja Crook (2010: 473-485) poolt avaldatud

elulemusanalüüsist järeldus, et risk langeb kuni 54. eluaastani, kuid koostatud mudel hindas ka noorte korral krediidiriski madalaks.



Joonis 11. Mudeli M1 otsustuspuu (autori koostatud).
Märkused: Helehalliga on tähistatud maksehäireta ja tumehalliga maksehäirega vaatluste (n) osakaal lehel („node“).

Mudeli järgi ei teki naissoost klientidel, kellel on vanust alla 20. eluaasta, probleeme uuritavas maksekäitumises, kuid otsustuspuu järgi ei saa sama väita 18. aastaste meeste kohta, kellele omistati lehena klass 1. Erialase kirjanduse põhjal on naiste madalam krediidiriski tase üks

stiliseeritumaid fakte. Siiski osutub mudeli põhjal 19. aastaste meessoost deebitori maksekäitumine heaks, kui taotleja sissetulek on alla 434 euro või sissetulek üle selle ja praegusel töökohal töötatud aeg jääb alla kaheksa kuu. Antud tulemus ei kattu Marshall *et al.* (2010: 501–512) poolt tehtud järeldustega, mille järgi pikemalt praegusel töökohal töötanud klientidel on madalam risk sattuda makseraskustesse.

Ka mudeli M2 esimeseks hargnemiskriteeriumiks on vanus samade väärtustega nagu mudelil M1 (vt lisa 13). Üle 19. aastastele omistatakse klass 0 lehega, kuhu kuulub 78,6% maksehäirega treeningvaatlustest. Alla 20. aastaste taotlejate korral on järgmiseks hargnemiskriteeriumiks haridustase, mille järgi ei teki lepingu esimese 12. kuu jooksul pikemaajalist viivitust põhi- või kõrgharidusega klientidel. Keskeriharidusega klientide, kelle laenuperiood on pikem kui 18 kuud, ja keskharidusega 18. aastaste klientide laenud klassifitseeritakse otsustuspuu järgi problemaatilisteks. Kočenda ja Vojtek (2009: 15-16) leidsid, et kõrgema haridustasemega kliendid on oluliselt madalama krediidiriskiga, mis on osaliselt vastuolus konkreetse mudeli tulemustega, kus põhiharidusega kliendi riski on hinnatud madalamaks kui keskeri- ja keskharidusega taotlejatel. Alla 20. aastaste põhiharidusega klientide madalam riskitase võib olla põhjendatud asjaoluga, et keskmine praeguses töökohas töötatud aeg kuudes on sellel grupil kõige kõrgem, mis peegeldab sissetuleku stabiilsust. Samuti võib töö kaotades pikem töökogemus tulla kasuks uue töö leidmisel. Täiendavalt on mudeli järgi riskantsem laen, kui kliendiks on 19. aastane keskharidusega klient, kelle laenusumma ületab 497,5. eurot.

Mudeli M3 esimene hargnemiskriteerium ja hargnemisväärtused on samad, mis mudelitel M1 ja M2 (vt lisa 14). Sarnaselt mudeliga M2, peetakse M3 korral riskantseks 18. aastast keskharidusega meessoost klienti, kuid täiendavalt samas vanuses meessoost keskeriharidusega deebitori. M3 on ainus mudel, kus osutus oluliseks postiaadressi linn ja seda 19. aastaste meeste grupis, kelle sissetulek ületab 433. eurot. Nimelt ei satu otsustuspuu järgi selliste omadustega kliendid maksehäiresse, kui postiaadressi linnaks on Tartu või Viljandi. Sama kehtib ka Tallinna kohta, kui kliendi praegusel töökohal töötatud aeg ei ületa 7,2. kuud. Kuigi samal tasemel omistatakse Kohtla-Järve, Narva ja Pärnu korral lehele klassiks üks, siis ei saa sellest lõplikke järeldusi teha, kuna antud lehtede alla ei kuulu ühtegi vaatlust. Kasutatav

otsustuspuu meetod omistab puuduvate vaatluste korral konkreetsete muutuja väärtuste lehtedele domineeriva klassi. Tulemuste usaldusväärsuse tõstmiseks peaksid valimisse kuuluma nimetatud tunnustega vaatlused. Kui mitte tõlgendada vaatlusteta lehtede tulemusi, siis on suuremates linnades elavate deebitoride krediidirisk ootuspäraselt madalam võrreldes väiksemate asulatega, mis on koondatud valiku „Muu“ alla.

Mudelite M4, M5, ja M6 korral, mis kasutavad klassi tasakaalu säilitavaid valimeid, on esimeseks hargnemiskriteeriumiks sarnaselt eelmiste mudelitega vanus taotlemise hetkel, kuid hargnemisväärtuseks on 19. eluaasta asemel 18 eluaastat (vt lisa 15, lisa 16, lisa 17). M4 klassifitseerib riskantseteks laenudeks lepingud, mis on väljastatud 18. aastastele meessoost klientidele, kelle laenuperiood on pikem kui 15 kuud. Ka osutuvad problemaatiliseks laenud, mis on antud 18. aastastele naissoost taotlejatele, kelle laenuperiood on pikem kui 15 kuud ja laenusumma on 407,5 eurot või vähem. Meessoost klientide klassifitseerimine riskantsemaks on kirjanduses tooduga kooskõlas, kuid suurem laenusumma peaks tõstma maksehäire esinemise tõenäosust (Lee, Chen 2005: 743-752).

18. aastased meesoost ja 18. aastased keskharidusega naissoost kliendid satuvad mudeli M5 järgi laenu esimese 12. kuu jooksul maksehäiresse. Samasugust tulemust prognoositakse klientidele, keda iseloomustavad samaaegaselt järgmised tunnused:

- lepingu laenusumma ületab 827 eurot,
- periood on pikem kui 13 kuud,
- sissetulek ületab 475. eurot,
- taotleja vanuseks on 19 või 20 eluaastat.

Taaskord ei ole tulemus sissetuleku ja laenusumma seisukohalt erialakirjanduse ülevaates kirjeldatud mõju suundadega kooskõlas, kuid otsustuspuu korral peab arvestama erinevate puu stuktuurist tulenevate muutujate koosmõjuga, mida kirjeldatud tööd ei kajastanud.

Mudelis M6 osutus maksehäirega ja maksehäireta lepingute klassifitseerimisel oluliseks kolm muutujat. Otsustuspuu järgi osutuvad „halbadeks“ laenudeks lepingud, mille kliendiks on 18. aastane meesterahvas või 19- 21. aastane isik, kelle poolt taotletav

laenusumma ületab 980 eurot. Mudelis kasutatavate tunnuste mõju suunad on vastavuses töö teoreetilises osas tooduga, mille kohaselt tõuseb maksehäire esinemise tõenäosus laenusumma suurenedes (Lee, Chen 2005: 743-752). Ka mudeli M7 lõplik puu struktuur on lihtne, koosnedes kokku 13. lehest või sõlmpunktist (vt lisa 18). Hinnatud klassifikaatori järgi satuvad maksehäiresse 18. aastased meessoost kliendid, kelle poolt taotletava laenu periood on pikem kui 21 kuud. Mudeli järgi esineb makseraskusi ka meessoost keskharidusega klientidel, kui laenuperiood on 21 kuud või lühem.

Kõikide loodud mudelite üldine klassifitseerimistäpsus PCC järgi on sarnane, jäädes 94,05% ja 94,96% vahele (vt tabel 9).

Tabel 9. Mudelite klassifitseerimistäpsused näitajate PCC, TPR, TNR ja AUC järgi

Mudel/ Näitaja	M1	M2	M3	M4	M5	M6	M7*
PCC (%)	94,53	93,47	94,96	94,05	94,29	94,44	94,54
TPR (klass = 1) (%)	22,05	14,15	18,31	9,76	16,33	21,92	14,62
TNR (klass = 1) (%)	99,56	99,24	99,91	99,73	99,52	99,27	99,92
AUC	0,61	0,57	0,59	0,55	0,58	0,61	0,57

Märkused: Mudeli M7 korral kuvatakse muutuja keskmist väärtus üle kümne testvalimi.
Allikas: autori koostatud.

Nimetatud meetriku järgi osutusid madalaima ja kõrgeima täpsusega mudeliteks vastavalt M4 ja M3, kuid teiste mudelite vahelised erinevused näitaja lõikes on väikesed. Tabelis üheksa toodud PCC, õige-positiivsete ja õige-negatiivsete määrade arvutused põhinevad lisadel 7, 8, 9, 10, 11 ja 12.

Klassi üks, mis tähistab maksehäire esinemist lepingus, jaoks välja arvutatud õige-positiivsete ja õige-negatiivsete prognooside määrade järgi tulevad mudelite vahelised erinevused selgemini välja. Kui mudeli M4 korral suudeti testvalimis korrektselt klassifitseerida 9,76% maksehäirega lepingutest kõikidesse maksehäirega vaatlustesse, siis parima tulemuse saavutasid mudelid M1 ja M6, kus sama näitaja väärtus oli üle kahe korra kõrgem ehk ligikaudu 22%. Sellest järeldub, et kuigi mudeli M4 klassifitseerimistäpsus on kõrge, siis klassidevahelise tasakaalustamatuse tõttu ei

peegeldu näitajas konkreetse mudeli väga madal võime uuritavaid klasse eristada. Kui mudelite, mille korral säilitati valimi loomisel klasside suhe, on muutuja TPR järgi näha klassifitseerimistäpsuse kasvu treeningvalimi suurendamisel, siis tasakaalustamata juhuvalimitega mudelite M1, M2 ja M3 korral sellist trendi ei avaldu ning täpseimaks osutub klassifitseerija, mille treeningvalim hõlmas endas 50% algandmetest. Kuna uuritava klassiga vaatlusi on algandmestikus 6,3% ja need ei ole homogeensed, siis sõltuvad saadavad tulemused suuresti kasutatud juhuslikest alamvalimitest. Kui krediidiriski mudeli üheks eesmärgiks on välistada laenu andmist taotlejatele, kes hiljem maksehäiresse satuvad, siis teiselt poolt soovitakse minimeerida taotlejate arvu, kes liigitatakse ekslikult halvaks kliendiks. Kõige täpsemini eristas maksehäireta vaatlusi mudel M7, mis kategoriseeris 99,92% juhtudest maksehäireta kliendi korrektselt. Saadud tulemus on mõneti ootuspärane, kuna M7 kasutab valideerimiseks samu andmeid, mida kasutati õpetamiseks. Nimetatud näitaja järgi oli täpsuselt teine M3 väärtusega 99,92%. Kõige madalama õige-negatiivsete määraga mudeliks osutus M2 väärtusega 99,24%.

Kuna ROC kõver konstrueeritakse meetrikute TPR ja TNR alusel teatud piirväärtuste korral, on AUC järgi ootuspäraselt kõrgeima klassifitseerimisvõimega mudeliteks M1 ja M6, mille korral arvutati AUC väärtuseks 0,61. Kõige madalama AUC oli mudelil M4, nimelt 0,55. Mudeli M7 AUC väärtus osutus teiste mudelitega võrreldes keskmiseks. Loodud mudelite klassifitseerimistäpsust võib näitaja AUC järgi pidada pigem madalaks, kuna väärtused 1 ja 0.5 tähistavad vastavalt täiuslikku ja täiesti juhuslikku klassifitseerimist. Siinkohal on oluline rõhutada asjaolu, et töös on kasutusel nihkega valim, millest on võrreldes üldkogumiga ligi 40% vaatlustest välja jäänud. Ettevõttes kehtinud laenuandmise põhimõtteid parimate mudelite reeglitega täiendades oleks saanud täiendavalt korrektselt klassifitseerida ligi iga viienda maksehäirega kliendi, mis oleks võinud suurtemate mahtude korral kreeditori jaoks kaasa tuua olulise kulude kokkuhoiu. Tegelik finantsiline mõju ettevõttele sõltub probleemsete laenude mahust, sissenõudetegevuse efektiivsusest ja sellega seotud kuludest. Täiendavalt peaks kreeditorid arvestama negatiivse mõjuga brändile, mis võib tekkida, kui maksehäirega lepingute osakaal moodustab liiga suure osa laenuportfellist.

Järgmisena kõrvutatakse lühidalt kõige kõrgema klassifitseerimistäpsusega mudeleid oluliseks osutunud karakteristikute seisukohalt. Mudelite M1 ja M6 korral osutusid olulisteks muutujateks kliendi vanus taotlemise hetkel ja sugu. Täiendavalt leidis mudelis M1 kasutust praeguse tööandja juures töötatud aeg kuudes ja igakuine sissetulek ning mudelis M6 laenusumma. Kõik nimetatud lõplikesse mudelitesse kaasatud tunnused on töös koostatud kirjanduse ülevaate alusel kõrge või keskmise kasutussagedusega. Mõlemas mudelis on naiste krediidirisk võrreldes meestega teatud alagruppide lõikes madalam, mis on kooskõlas erialases kirjanduses saaduga. Mudelitest nähtub, et kõrgema vanusega taotlejatel esineb otsustuspuude järgi harvemini makseraskuseid. Mudel M1 ei suuda eristada maksehäirega ja maksehäireta kliente, kui kliendi vanus ületab 19 eluaastat. Mudeli M6 puhul esineb samasugune probleem alates 21. eluaastast. Taotleja vanuse kohta saadud tulemused kattuvad osaliselt teadustöödes tehtud järeldustega, kuna enamasti langes maksehäiresse sattumise tõenäosus ka kõrgemas vanuses klientidel. Põhjustena võib välja tuua „heade“ ja „halbade“ laenude võrdse käsitlemise mudelites ning kasutatud meetodi parametriseeringu, mis mõjutab lõpliku otsustuspuu sügavust. Mudelis M6 ühtib laenusumma mõjusuund töö teoreetilises osas mainituga, mille järgi suureneb maksehäire esinemise tõenäosus laenusumma tõustes. Mudeli M1 järgi toob 19. aastaste meessoost deebitoride grupis suurem igakuine sissetulek ja pikemalt praeguses töökohas töötatud aeg kaasa kõrgema krediidiriski, mis ei ole kooskõlas erialase kirjanduse ülevaates toodud artiklite tulemustega. Siinkohal on oluline rõhutada, et kasutatud teadustöödes tehtud järeldused põhinesid enamasti lineaarset seost modelleerivatel mudelitel, millega ei ole erinevalt otsustuspuust võimalik modelleerida eri muutujate väärtuste koosmõju.

Magistritöö tulemuste põhjal saab järeldada, et kui ettevõttes Kaupmehe Järelmaks OÜ oleks lisaks kehtinud laenuandmise põhimõtetele rakendatud kahe kõige täpsema klassifikaatoriga seotud reeglistikku, oleks täiendavalt korrektselt klassifitseeritud ligi iga viies maksehäirega klient. Samal ajal oleks maksehäireta klientide valesti klassifitseerimise määr olnud väga madal. Käesolevas töös kasutatud muutujate mõju suund ei ole mitmete karakteristikute osas kooskõlas varasemate uurimustega, mis võib viidata töös kasutatava andmestiku mittelineaarsusele krediidiriski seisukohalt, kuna enamasti olid mõju suundasid kirjeldavates teadustöödes kasutusel lineaarset seost

modelleerivad meetodid. Püstitatud seisukoha addresseerimiseks on vajalik lineaarset seost modelleerivate meetodite kasutamine samal valimil ja prognoosivõime võrdlemine praeguse töö raames saadud mudelitega. See võiks olla üheks töö edasiarendamise suunaks.

Antud töö üheks piiranguks võib pidada valimi mittepiisavat mahtu, kuna kasutatud meetod C4.5 vajab usaldusväärsete tulemuste saamiseks küllaltki suurt vaatluste arvu. Seda eelkõige valimite korral, kus klassidevaheline tasakaalustamatus on suur. Koostatud mudelitest nähtus, et mitmetel juhtudel ei vastanud teatud lehtedele ühtegi vaatlust ja tulemusena rakendati domineerivat klassi. Nagu eraisiku krediidiriski hindamise töödes üldiselt, on üheks fundamentaalseks piiranguks negatiivse otsuse saanud kliendid, kelle tegeliku maksekäitumise kohta informatsioon puudub. Sellest tulenevalt võivad saadud tulemused olla moonutatud, kuna suure tõenäosusega on selliste taotlejate seas neid, kellel ei oleks laenu saamise korral esinenud maksehäiret. Järelikult on oht, et saadud mudel üle- või alahindab eraldiseisvana teatud muutujate mõju uuritavale tunnusele või on mõju suund sootuks vale. Täiendava piiranguna võib vaadelda uuritavade klassidega seotud kulude võrdsena käsitlemist mudelis, kuna praktikas kaasnevad maksehäires lepinguga enamasti kõrgemad kulud. Algoritm C4.5 võimaldab kaasata kulumaatriksi, kuid käesolevas töös seda ei tehtud, kuna ettevõte Kaupmehe Järelmaks OÜ ei avaldanud nimetatud informatsiooni uurimustöö koostajale.

Käesoleva töö üheks edasiarendamise võimaluseks on sama valimi alusel teiste teoreetilises osas mainitud eraisiku krediidiriski hindamisel enimkasutatud meetoditega mudelite koostamine. Ühe meetodina võiks täiendavalt uurimusse kaasata algoritmi C4.5 edasiarenduse C5.0, mille täpsus ületab teatud juhtudel algoritmi autori sõnul eelkäija oma. Saadud mudeleid saab võrrelda nii muutujate mõjusuundade kui ka klassifitseerimistäpsuse seisukohalt. Sellest tulenevalt oleks võimalik teha põhjalikemaid järeldusi töös kasutatavate selgitavate muutujate ja uuritava tunnuse vahelise seose mittelineaarsuse osas. Kuna koostatud mudelid ei suutnud vanemate taotlejate korral eristada maksehäirega ja maksehäireta kliente, siis võiks võimalusel sisse tuua täiendavaid karakteristikuid. Ettevõte Kaupmehe Järelmaks OÜ võiks kaaluda töös kasutatud metoodika rakendamist hilisemate lepingutega valimi peal tuvastamaks, kas antud meetod suudab võrreldes hetkel kehtiva krediidimudeliga täiendavalt

korrektselt klassifitseerida maksehäirega vaatlusi. Positiivse tulemuse korral on võimalik koostatud otsustuspuust saadud täiendavad reeglid implementeerida ettevõtte infosüsteemides. Kui kehtiva mudeli täiendamine ei ole mingil põhjusel otstarbekas, on üheks võimaluseks negatiivse otsuse saanud vaatluste kaasamine valimisse ja selle pealt uue mudeli genereerimine. Kuna ettevõtte eesmärgiks on kasumi maksimeerimine, näeb autor töö edasiarendamise võimalusena veel erinevate klassidega seotud kulumomendi sissetoomist. Selle saavutamiseks on tarvis välja selgitada, milline on valesti klassifitseeritud maksehäirega ja maksehäireta laenudega seonduv kulu. Ühelt poolt jääb ettevõttel saamata laenuga seotud tulu, kui laenu ei väljastata kliendile, kellel tegelikkuses ei esineks maksehäiret. Teiselt poolt kannab deebitor maksehäirega lepingu korral teatud kulusid, mis sõltuvad lõplikult sissenõutud summast ja selle tegevusega seotud kuludest.

KOKKUVÕTE

Eraisiku krediidiriski hindamisega seotud temaatika muudab Eesti kontekstis aktuaalseks asjaolu, et viimastel aastatel on eraisikutele väljastatavate tarbimislaenude maht olnud tõusva trendiga. Tihenenud konkurentsi tingimustes on eraisiku maksevõime võimalikult täpne prognoosomine muutunud kreeditoride jaoks üha olulisemaks, kuna turul valitseva hinnasurve ja valitsusepoolsete regulatsioonide tõttu on laenuandmisega tegelevate ettevõtete eksimisruum muutunud väiksemaks. Samuti võimaldab efektiivsem krediidiandmisega seonduva riski hindamine täpsemini seada provisjone, mis alandavad laenuandja jaoks kasutatava kapitali hinda. Kõrgem kasutatavate meetodite klassifitseerimistäpsus võimaldab hinnastada laenulepingut konkreetse taotleja riskitasemest lähtuvalt, andes seeläbi võimaluse pakkuda väiksemate kuludega laenu madalama krediidiriskiga klientidele. Suurtemate laenumahtude korral võib väike klassifitseerimistäpsuse parenemine tuua kreeditori jaoks kaasa märkimisväärse kulude kokkuhoiu. Ühiskondlikult kasulik efekt seisneb asjaolus, et efektiivsema selektsiooni korral laenatakse vähem isikutele, kes tegelikkuses ei ole võimelised võetud kohustusi teenindama ja mille tulemusena halveneb pikemas perspektiivis deebitori majanduslik seisukord veelgi.

Magistritöö eesmärgiks oli koostada krediidiriski hindamise mudel otsustuspuu meetodil ettevõtte Kaupmehe Järelmaks OÜ näitel. Uurimustöö on piiritletud otsustuspuu meetodi kasutamisega, kuna meetod on akadeemilises kirjanduses hinnatud interpreteeritavuse ja hea klassifitseerimistäpsuse pärast. Ka võib üheks eesmärgi valiku põhjuseks pidada Eesti akadeemilise kirjanduse vähesust eraisiku krediidiriski modelleerimisel antud meetodiga. Püstitatud eesmärgi saavutamiseks anti esmalt ülevaade erialakirjanduses eraisiku krediidiriski kohta kasutatavatest definitsioonidest, mis erinesid sisuliselt kasutatud allikate lõikes. Teatud juhtudel piirdutakse määratluses deebitori ja kreeditori vahelise kokkuleppe rikkumisega, kuid samas tuuakse mõnikord ka sisse hinnang kuludele, mis võivad esineda riski realiseerumisel. Mitmete definitsioonide korral viidatakse lepingujärgsete kohustuste mittetäitmisele, kuid

täpsustavat informatsiooni krediidiriski realiseerumise kohta ei pakutud. Põhjuseks võib olla tarbijakrediidi toodete tingimuste suur erinevus, mis muudab ühtse definitsiooni andmise krediidiriski realiseerumisele erinevate finantstoodete lõikes keeruliseks. Üldlevinud praktikas peetakse nimetatud riski realiseerumisega seotud sündmuseks maksehäiret, mille täpne sisemiselt kasutatav määratlus on krediidiandja poolt defineerida, kuid enamasti lähtutakse rahvusvahelisest praktikast. Olemuselt eelneb maksehäirele lühiajaline viivitus ühe lepingujärgse maksega, kuid sellisel juhul ei ole deebitoril enamasti põhjust pidada tekkinud olukorda püsivaks, sest kliendipoolse makse mittetegemise põhjuseks võib olla näiteks maksekuupäeva unustamine. Olles viivituses mitme makse ulatuses, on deebitori hinnang tõenäosusele, et võlas olevad ja lepingujärgsed tuleviku maksed laekuvad, vähenenud. Sellises olukorras on lepingul esinenud maksehäire. Kui maksehäire ei ole ajutist laadi, võib see eskaleeruda püsivaks maksejõuetuseks.

Järgmisena käsitleti eraisiku krediidiriski hindamisega tegelevate krediidianalüütikute, teadurite, kreeditoride ja teemakohase arvutitarkvara tootjate poolt enim kasutust leidnud statistiliste meetoditega, milleks on regressioonanalüüs, lineaarne programmeerimine, Coxi proportsionaalsete (võrdeliste) riskide mudel, tugivektor-masinate meetod, tehisnärvivõrgud, otsustuspuud, lähima naabri meetod, geneetilised algoritmid, juhumetsa meetod ja geneetiline programmeerimine. Erinevate meetodite võrdlusest järeldati, et kuigi klassifitseerimistäpsuse osas saavutati erialakirjanduses häid tulemusi logistilise regressiooni, tugivektor-masinate meetodi, tehisnärvivõrkude ja juhumetsa meetoditega, siis sõltub mudeli lõplik headus kreeditori jaoks ka mitmetest teistest aspektidest peale klassifitseerimistäpsuse. Praktikas on tihti oluline saadud tulemuste interpreteeritavus, mille korral on eelistatavam kasutada näiteks otsustuspuude põhist lähenemist, mille prognoosivõime kõikus artiklite lõikes madalast kõrgeni.

Selleks, et töö empiirilises osas oleks võimalik kõrvutada saadud tunnuste mõju suundasid varasemate teadustööde raames saadud tulemustega, anti järgmisena ülevaade eraisiku krediidiriski alases kirjanduses statistiliselt oluliseks osutunud karakteristikutest. Järeldati, et enamasti jääb praktikas lõplikusse mudelisse kaasatud statistiliselt oluliste muutujate arv vahemikku 5-10. Muutujate kasutussageduse põhjal

eristas töö autor kõrge-, keskmise ja madala kasutussagedusega tunnuseid. Kõrgeks peeti kasutussagedust, kui nimetatud muutuja jõudis lõplikkusse mudelisse vähemalt neljal korral, keskmise kasutussagedusega kahel või kolmel korral ning madala korral ühel juhul. Kirjanduse ülevaatest nähtus, et eraisiku krediidiriski hindamise seisukohalt olid kõrge kasutussagedusega sõltumatuteks muutujateks vanus, perekonnaseis, praeguses elukohas elatud aeg, praeguse tööandja juures töötatud aeg, kliendisuhte kestus kreeditoriga, sissetulek, ülalpeetavate või laste arv, laenu tüüp ja elukoha tüüp. Keskmise ja madala kasutussagedusega sõltumatuteks muutujateks olid näiteks taotleja sugu, tegevusala, haridustase, laenusumma ja kliendi krediidiajalugu.

Analüüsi läbiviimiseks kasutati ettevõtte OÜ Kaupmehe Järelmaks infosüsteemi andmebaasist pärit andmestikku. Valimi suuruseks oli 3901 vaatlust, mis oli moodustatud juhuvalimina 2011. aasta järelmaksulepingutest. Iga andmestikus oleva lepingu kohta oli teada kliendi sugu, vanus taotlemise hetkel, perekonnaseis, haridustase, ülalpeetavate arv, elukoha tüüp, postiaadressi maakond, postiaadressi linn, tegevusala, taotlemise hetkel töötatud aeg kuudes (praegusel ametikohal), igakuine sissetulek eurodes, maksehäirete arv taotlemise hetkel, laenusumma eurodes, laenuperiood kuudes ja maksehäire esinemine/ mitteesinemine lepingus.

Klassifitseerimismeetodina kasutati töös J.S.Quinlani poolt välja töötatud otsustuspuu algoritmi C4.5 implementatsiooni tarkvarapaketis R nimega J48. Juhuslike alamvalimite koostamisel rakendati kahte erinevate strateegiat, millest esimese korral muudeti test- ja treeningvalimite suhet üldvalimisse. Täiendavalt rakendati uuritava tunnuse klasside suhte piirangut, mille käigus jaotati juhuslikult andmekirjeid alamvalimitesse selliselt, et säiliks algandmetega võrreldes võimalikult täpne maksehäirega ja maksehäireta lepingute suhe. Tulemusena saadi järgmised valimid:

- treening- ja testvalim moodustasid esialgsest valimist vastavalt 50% ja 50%,
- treening- ja testvalim moodustasid esialgsest valimist vastavalt 60% ja 40%,
- treening- ja testvalim moodustasid esialgsest valimist vastavalt 70% ja 30%,
- treening- ja testvalim moodustasid esialgsest valimist vastavalt 50% ja 50% ning uuritavate klasside suhe säilis,

- treening- ja testvalim moodustasid esialgsest valimist vastavalt 60% ja 40% ning uuritavate klasside suhe säilis,
- treening- ja testvalim moodustasid esialgsest valimist vastavalt 70% ja 30% ning uuritavate klasside suhe säilis.

Iga treeningvalimi põhjal genereeriti J48 algoritmiga otsustuspuu mudel ehk kokku kuus mudelit. Täiendavalt koostati terve valimi pealt sama algoritmi kasutades mudel, mille prognoosivõimet hinnati jagades esialgne valim kümneks eraldiseisvaks testvalimiks, kus püüti säilitada klasside esialgset suhet. Otsustuspuude korral on levinud probleemiks ülesobitumine, mida lahendati esialgse puu kärpimise teel kasutades usaldusnivoona vaikimisi väärtust 0.25.

Kõikides töö raames genereeritud mudelites osutus oluliseks vanus taotlemise hetkel, kuid kordagi ei leidnud kasutust muutujad perekonnaseis, ülalpeetavate arv, elukoha tüüp, tegevusala, postiaadressi maakond ja maksehäirete arv taotlemise hetkel. Mõned muutujad nagu näiteks perekonnaseis, ülalpeetavate arv, elukoha tüüp ja tegevusala, mille oluliseks osutumist töö teoreetilise osa alusel oleks põhjust oodata, võisid mudelitest välja jääda seetõttu, et nende tunnuste teatud väärtustega vaatluste sattumine töös kasutatavasse valimisse välistati laenuotsuse hetkel kehtinud krediidipoliitika tõttu. Ühe sellise näitena võib välja tuua maksehäirete arvu taotlemisel. Nimelt ei väljastatud krediidireeglite järgi laenu kliendile, kellel oli maksehäireregistri andmetel lõpetamata või vähem kui kuus kuud tagasi lõpetatud pangandus- või muu finantseerimisvõlg. Kõikides mudelites, välja arvatud M1 ja M7 korral, osutus oluliseks lepingu laenusumma, mis teoreetilises osas käsitletud artiklites leidis kasutust keskmise sagedusega. Karakteristik postiaadressi linn leidis kasutust ainult mudelis M3, kuid laenuperiood viie mudeli korral. Tunnused praeguse tööandja juures töötatud aeg ja igakuine sissetulek olid käsitletud artiklites tihedalt kasutuses, kuid loodud mudelites osutusid need oluliseks ühel kuni kahel korral.

Kõige kõrgema klassifitseerimistäpsusega mudeliteks osutusid mudelid M1 ja M6, mille mõõdiku AUC väärtuseks oli 0,61 ning PCC väärtusteks vastavalt 94,53% ja 94,44%. Mudel M7, mis koostati terve algvalimi pealt, oli PCC lõikes nimetatud mudelitest täpsem, kuid maksehäirega lepingute eristamise osas jäi märkimisväärselt alla. Mudeli

M1 treening- ja testvalim moodustasid esialgsest valimist 50% ja 50%, aga M6 korral vastavalt 70% ja 30%. Mudelite M1 ja M6 korral osutusid olulisteks muutujateks kliendi vanus taotlemise hetkel ja sugu. Täiendavalt leidis mudelis M1 kasutust praeguse tööandja juures töötatud aeg kuudes ja igakuine sissetulek ning mudelis M6 laenusumma. Kõik nimetatud lõplikesse mudelitesse kaasatud tunnused on töös koostatud kirjanduse ülevaate alusel kõrge või keskmise kasutussagedusega. Mõlemas mudelis on naiste krediidirisk võrreldes meestega teatud alagruppide lõikes madalam, mis on kooskõlas erialases kirjanduses saaduga. Mudelitest nähtub, et kõrgema vanusega taotlejatel esineb otsustuspuude järgi harvemini makseraskuseid. Mudel M1 ei suuda eristada maksehäirega ja maksehäireta kliente, kui kliendi vanus ületab 19 eluaastat. Mudeli M6 puhul esineb samasugune probleem alates 21. eluaastast. Taotleja vanuse kohta saadud tulemused kattuvad osaliselt teadustöödes tehtud järeldustega, kuna enamasti langes maksehäiresse sattumise tõenäosus ka kõrgemas vanuses klientidel. Põhjustena võib välja tuua „heade“ ja „halbade“ laenude võrdse käsitlemise mudelites ning kasutatud meetodi parametriseering, mis mõjutab lõpliku otsustuspuu sügavust. Mudelis M6 ühtib laenusumma mõjusuund töö teoreetilises osas mainituga, mille järgi suureneb maksehäire esinemise tõenäosus laenusumma tõustes. Mudeli M1 järgi toob 19. aastaste meessoost deebitoride grupis suurem igakuine sissetulek ja pikemalt praeguses töökohas töötatud aeg kaasa kõrgema krediidiriski, mis ei ole kooskõlas erialase kirjanduse ülevaates saadud tulemustega. Siinkohal on oluline rõhutada, et kasutatud teadustöödes tehtud järeldused põhinesid enamasti lineaarset seost modelleerivatel mudelitel, millega ei ole erinevalt otsustuspuust võimalik modelleerida eri muutujate väärtuste koosmõju.

Magistritöö tulemuste põhjal saab järeldada, et kui ettevõttes Kaupmehe Järelmaks OÜ oleks lisaks kehtinud laenuandmise põhimõtetele rakendatud kahe kõige täpsema klassifikaatoriga seotud reeglistikku, oleks täiendavalt korrektselt klassifitseeritud ligi iga viies maksehäirega klient. Samal ajal oleks maksehäireta klientide valesti klassifitseerimine olnud väga madal. Käesolevas töös kasutatud muutujate mõju suund ei ole mitmete karakteristikute osas kooskõlas varasemate uurimustega, mis võib viidata töös kasutatava andmestiku mittelineaarsusele krediidiriski seisukohalt, kuna enamasti olid mõju suundasid kirjeldavates teadustöödes kasutusel lineaarset seost modelleerivad meetodid. Püstitatud seisukoha addresseerimiseks on vajalik lineaarset seost

modelleeritavate meetodite kasutamine samal valimil ja prognoosivõime võrdlemine praeguse töö raames saadud mudelitega. See võiks olla üheks töö edasiarendamise suunaks.

Antud töö üheks piiranguks võib pidada valimi mittepiisavat mahtu, kuna kasutatud meetod C4.5 vajab usaldusväärsete tulemuste saamiseks küllaltki suurt vaatluste arv. Seda eelkõige valimite korral, kus klassidevaheline tasakaalustamatus on suur. Koostatud mudelitest nähtus, et mitmetel juhtudel ei vastanud teatud lehtedele ühtegi vaatlust ja tulemusena rakendati domineerivat klassi. Nagu eraisiku krediidiriski hindamise töödes üldiselt, on üheks fundamentaalseks piiranguks negatiivse otsuse saanud kliendid, kelle tegeliku maksekäitumise kohta informatsioon puudub. Sellest tulenevalt võivad saadud tulemused olla moonutatud, kuna suure tõenäosusega on selliste taotlejate seas neid, kellel ei oleks laenu saamise korral esinenud maksehäiret. Järelikult on oht, et saadud mudel üle- või alahindab teatud muutujate mõju uuritavale tunnusele või on mõju suund sootuks vale. Täiendava piiranguna võib vaadelda uuritavade klassidega seotud kulude võrdsena käsitlemist mudelis, kuna praktikas kaasnevad maksehäires lepinguga enamasti kõrgemad kulud. Algoritm C4.5 võimaldab kaasata kulumaatriksi, kuid käesolevas töös seda ei tehtud, kuna ettevõte Kaupmehe Järelmaks OÜ ei avaldanud nimetatud informatsiooni uurimustöö koostajale.

Käesoleva töö üheks edasiarendamise võimaluseks on sama valimi alusel teiste teoreetilises osas mainitud eraisiku krediidiriski hindamisel enimkasutatud meetoditega mudelite koostamine. Ühe meetodina võiks täiendavalt uurimusse kaasata algoritmi C4.5 edasiarenduse C5.0, mille täpsus ületab teatud juhtudel algoritmi autori sõnul eelkäija oma. Saadud mudeleid saab võrrelda nii muutujate mõjusuundade kui ka klassifitseerimistäpsuse seisukohalt. Sellest tulenevalt oleks võimalik teha põhjalikemaid järeldusi töös kasutatavate selgitavate muutujate ja uuritava tunnuse vahelise seose mittelineaarsuse osas. Kuna koostatud mudelid ei suutnud vanemate taotlejate korral eristada maksehäirega ja maksehäireta kliente, siis võiks võimalusel sisse tuua täiendavaid karakteristikuid. Ettevõte Kaupmehe Järelmaks OÜ võiks kaaluda töös kasutatud metoodika rakendamist hilisemate lepingutega valimi peal tuvastamaks, kas antud meetod suudab võrreldes hetkel kehtiva krediidimudeliga täiendavalt korrektselt klassifitseerida maksehäirega vaatlusi. Positiivse tulemuse korral on

võimalik koostatud otsustuspuust saadud täiendavad reeglid implementeerida ettevõtte infosüsteemides. Kui kehtiva mudeli täiendamine ei ole mingil põhjusel otstarbekas, on üheks võimaluseks negatiivse otsuse saanud vaatluste kaasamine valimisse ja selle pealt uue mudeli genereerimine. Kuna ettevõtte eesmärgiks on kasumi maksimeerimine, näeb autor töö edasiarendamise võimalusena veel erinevate klassidega seotud kulumomendi sissetoomist. Selle saavutamiseks on tarvis välja selgitada, milline on valesti klassifitseeritud maksehäirega ja maksehäireta laenu kulu. Ühelt poolt jääb ettevõttel saamata laenuga seotud tulu, kui laenu ei väljastata kliendile, kellel tegelikkuses ei oleks maksehäiret esinenud. Teiselt poolt kannab deebitor maksehäirega lepingu korral teatud kulusid, mis sõltuvad lõplikult sissenõutud summast ja selle tegevusega seotud kuludest.

VIIDATUD ALLIKAD

1. **Anderson, R.** The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation. New York: Oxford University Press, 2007, 731 p.
2. **Abdou, H. A., Pointon, J.** Credit scoring, statistical techniques and evaluation criteria: A review of the literature. – Intelligent Systems in Accounting, Finance & Management, 2011, Vol 18, No. 2-3, pp. 59–88. URL: http://onlinelibrary.wiley.com/doi/10.1002/isaf.325/abstract
3. **Abdou, H., Pointon, J., El-Masry, A.** Neural nets versus conventional techniques in credit scoring in Egyptian Nanking. – Expert Systems with Applications, 2008, Vol. 35, No. 3, pp. 1275–1292. URL: http://www.sciencedirect.com/science/article/pii/S0957417407003326
4. **Avery, R.B., Calem, P.S, Canner, G.B.** Consumer credit scoring: Do situational circumstances matter? – Journal of Banking & Finance, 2004, Vol. 28, No. 4, pp. 835–856. URL: http://www.bis.org/publ/work146.htm
5. **Baesens, B., Gestel, T.V., Viaene, S., Stepanova, M., Suykens, J., Vanthienen, J.** Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring. – The Journal of the Operational Research Society, 2003, Vol. 54, No. 6, pp. 627-635. URL: http://www.jstor.org/stable/pdf/4101754.pdf?acceptTC=true&seq=1#page_scan_tab_contents
6. **Banasik, J., Crook, J.** Reject inference in survival analysis by augmentation. – Journal of the Operational Research Society, 2010, Vol. 61, No.3, pp. 473-485. URL: http://www.jstor.org/stable/pdf/40540274.pdf?acceptTC=true
7. **Bellotti, T., Crook, J.** Support vector machines for credit scoring and discovery of significant features. - Expert Systems with Applications, 2009, Vol. 36, No. 2, pp. 3302–3308. URL: http://www.sciencedirect.com/science/article/pii/S0957417408000857

8. **Benediktsson, J.A., Gislason, P.O., Sveinsson, J.R.** Random Forests for land cover classification. - Pattern Recognition Letters, 2006, Vol. 27, No. 4, pp. 294–300. URL: http://www.sciencedirect.com/science/article/pii/S0167865505002242
9. **Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.** Classification and Regression Trees (Wadsworth statistics/probability series). Boca Raton: CRC Press, 1984, 368 p.
10. **Brown, I., Mues, C.** An experimental comparison of classification algorithms for imbalanced credit scoring data sets. – Expert Systems with Applications, 2012, Vol. 39, No. 3, pp. 3446-3453. URL: http://www.sciencedirect.com/science/article/pii/S095741741101342X
11. **Brown, K., Moles, P**. Credit Risk Management. Edinburgh: Edinburgh Business School, 2008, 54 p.
12. **Chuang, C-L., Lin, R-H.** Constructing a reassigning credit scoring model. – Expert Systems with Applications, 2009, Vol. 36, No. 2, 1, pp. 1685-1694. URL: http://www.sciencedirect.com/science/article/pii/S0957417407005854
13. **Cox, D.R., Snell, E. J.** Analysis of Binary Data, Second Edition. Boca Raton: CRC Press, 1989, 19 p.
14. **Devasena, C.L.** Comparative Analysis of Random Forest, REP Tree and J48 Classifiers for Credit Risk Prediction. – International Journal of Computer Applications, 2015, pp. 30-36. URL: http://research.ijcaonline.org/icccmit2014/number3/icccmit7033.pdf
15. **Flach, P., Hernandez-Orallo, J., Ferri, C.** A Coherent Interpretation of AUC as a Measure of Aggregated Classification Performance. 8 p. [http://www.icml-2011.org/papers/385_icmlpaper.pdf]. 10.10.2015.
16. **Friedman, J.H.** Multivariate Adaptive Regression Splines. – The Annals of Statistics, 1991, Vol. 19, No. 1, pp. 1-67. URL: http://www.jstor.org/stable/2241837?seq=1#page_scan_tab_contents
17. **Guardia, N.D.** Consumer credit in the European Union. - European Credit Research Institute, 2002, No. 1, 2 p. URL: http://ec.europa.eu/consumers/citizen/my_rights/consumer-credit/index_en.htm

18. **Hamill, T.M., Juras, J.** Measuring forecast skill: is it real skill or is it the varying climatology?. - Quarterly Journal of the Royal Meteorological Society, 2006, Vol. 132, pp. 2905–2923. URL: http://www.esrl.noaa.gov/psd/people/tom.hamill/skill_overforecast_QJ_v2.pdf
19. **Han, L., Han, L., Zhao, H.** Engineering Applications of Artificial Intelligence. – Engineering Applications of Artificial Intelligence, 2013, Vol. 26, No. 2, pp. 848-852. URL: http://www.sciencedirect.com/science/article/pii/S0952197612002667
20. **Han, J., Kamber, M., Pei, J.** Data Mining Concepts and Techniques. Waltham: Morgan Kaufmann, 2012, 703 p.
21. **Hand, D.J.** Measuring classifier performance: a coherent alternative to the area munder the ROC curve. 2009, 21 p. [http://web.cs.iastate.edu/~cs573x/Notes/hand-article.pdf]. 07.11.2015.
22. **Hand, D.J., Henley, W.E.** Statistical Classification Methods in Consumer Credit Scoring: a Review. – Journal of the Royal Statistical Society: Series A (Statistics in Society), 1997, Vol. 160, No. 3, pp. 523–541. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-985X.1997.00078.x/pdf
23. **Hand, D.J., Sohn, S.Y., Kim, Y.** Optimal bipartite scorecards. – Expert Systems with Applications, 2005, Vol. 29, No.3, pp. 684–690. URL: http://www.sciencedirect.com/science/article/pii/S0957417405000850
24. **Hornik, K., Buchta, C., Hothorn, T., Karatzoglou, A., Meyer, D., Zeileis, A.** Package 'RWeka'. 2015, 34 p. [https://cran.r-project.org/web/packages/RWeka/RWeka.pdf]. 12.11.2015.
25. **Holland, J.H.** Adaptation in Natural and Artificial Systems: an Introductory Analysis with Applications to Biology, Control and Artificial Intelligence, 1992, 211 p.
26. **Hssina, B., Merbouha, A., Ezzikouri, H., Erritali, M.** A comparative study of decision tree ID3 and C4.5. – International Journal of Advanced Computer Science and Applications, 2014, 13 p. URL: http://thesai.org/Downloads/SpecialIssueNo10/Paper_3-A_comparative_study_of_decision_tree_ID3_and_C4.5.pdf

27. **Ince, H., Aktan, B.** A comparison of data mining techniques for credit scoring in Nanking: A managerial Perspective. – Journal of Business Economics and Management, 2009, Vol. 10, No. 3, 236 p. URL: http://www.tandfonline.com/doi/pdf/10.3846/1611-1699.2009.10.233-240
28. International Convergence of Capital Measurement and Capital Standards. Basel Committee on Banking Supervision, 2006, 333 p. [http://www.bis.org/publ/bcbs128.pdf]. 02.10.2015.
29. Is See5/C5.0 Better Than C4.5?. [http://rulequest.com/see5-comparison.html]. 10.11.2015.
30. **Jacobson, T., Roszbach, K.** Bank lending policy, credit scoring and value-at-risk. – Journal of Banking & Finance, 2003, Vol. 27, pp. 615–633. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.139.6221&rep=rep1&type=pdf
31. **Kočenda, E., Vojtek, M.** Default Predictors and Credit Scoring Models for Retail Banking. – Center for Economic Studies, 2009, No. 2862, pp 15-16. URL: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1519792
32. Kodumajapidamistele antud laenude jääk ja arv laenuliigi, valuuta ja tagatise lõikes. Eesti Pank. [http://www.statistika.eestipank.ee/#listMenu/1172/treeMenu/FINANTSSEKTOR/147/650]. 11.02.2016.
33. **Kohavi R., Quinlan R.** Decision Tree Discovery. 1999, 16 p. [http://ai.stanford.edu/~ronnyk/treesHB.pdf]. 17.11.2015.
34. **Kozeny, V.** Genetic algorithms for credit scoring: Alternative fitness function performance comparison. – Expert Systems with Applications, 2015, Vol. 42, 6, pp. 2998-3004. URL: http://www.sciencedirect.com/science/article/pii/S0957417414007143
35. Krediidiinfo. [http://www.krediidiinfo.ee]. 02.01.2016.
36. **Lee, T-S., Chen, I-F.** A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regressioon splines. – Expert Systems with Applications, 2005, Vol. 28, No. 4, pp. 743-752. URL: http://www.sciencedirect.com/science/article/pii/S0957417404001782

37. **Lee, T-S., Chiu, C-C., Lu, C-J., Chen, I-F.** Credit scoring using the hybrid neural discriminant technique. – Expert Systems with Applications, 2002, Vol. 23, No. 3, pp. 245–254. URL: http://www.sciencedirect.com/science/article/pii/S0957417402000441
38. **Lessmann, S., Seow, H-V., Baesens, B., Thomas, L.C.** Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update. 2013, 60 p. [http://www.business-school.ed.ac.uk/waf/crc_archive/2013/42.pdf] 17.11.2015.
39. **Linoff, G.S., Berry, M.J.A.** Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. Indianapolis: Wiley, 2011, 888 p.
40. **Mangasarian, O. L.** Linear and Nonlinear Separation of Patterns by Linear Programming. – Operations Research, 1965, Vol. 13, No. 3, pp. 444-452. URL: http://www.cs.iastate.edu/~honavar/mangasarian-1965.pdf
41. **Marshall, A., Tang, L., Milne, A.** Variable reduction, sample selection bias and bank retail credit scoring. – Journal of Empirical Finance, 2010, Vol. 17, No. 3, pp. 501–512. URL: http://www.sciencedirect.com/science/article/pii/S0927539809001042
42. **Mazid, M.M., Ali, A.B.M.S., Tickle, K.S.** Improved C4.5 Algorithm for Rule Based Classification. – Proceedings of the 9th WSEAS international conference on Artificial intelligence, knowledge engineering and data bases, 2010, pp. 296–301. URL: http://www.wseas.us/e-library/conferences/2010/Cambridge/AIKED/AIKED-47.pdf
43. **Miner, G., Nisbet, R., Elder, J.** Handbook of Statistical Analysis and Data Mining Applications. London: Elsevier, 2009, 147 p.
44. **Neuner, M., Raab, G., Reisch, L.A.** Compulsive buying in maturing consumer societies: An empirical re-inquiry. – Journal of Economic Psychology, 2005, Vol. 26, No. 4, pp. 509–522. URL: http://www.sciencedirect.com/science/article/pii/S0167487004000741
45. **Ong, C-S., Huang, J-J., Tzeng, G-H.** Building credit scoring models using genetic programming. – Expert Systems with Applications, 2005, Vol. 29, No. 1, pp. 41–47. URL: http://www.sciencedirect.com/science/article/pii/S0957417405000059

46. **Paleologo, G., Elisseeff, A., Antonini, G.** Subagging for credit scoring models. – European Journal of Operation Research, 2010, Vol. 201, No. 1, pp. 490-499. URL: http://www.sciencedirect.com/science/article/pii/S0377221709001532
47. Principles for the Management of Credit Risk. Basel Committee on Banking Supervision, 2000, 26 p. [http://www.bis.org/publ/bcbs75.pdf]. 02.10.2015.
48. **Quinlan, J.R.** Induction of Decision Trees. – Machine Learning, 1986, Vol. 1, pp. 81-106. URL: http://hunch.net/~coms-4771/quinlan.pdf
49. **Quinlan, J.R.** C4.5: Programs for Machine Learning. San Mateo, California: Morgan Kaufmann, 1993, 23 p.
50. **Rokach, L., Maimon, O.** Data Mining with Decision Trees. Series in Machine Perception and Artificial Intelligence: Vol. 69. Singapore: World Scientific Publishing Co, 2007, 264 p.
51. **Ruggieri. S.** Efficient C4.5. IEEE Transactions on Knowledge and Data Engineering, 2002, Vol. 14, No. 2, pp. 438–444. URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=991727
52. **Schreiner, M**. Scoring Arrears at a Microlender in Bolivia. – Journal of Microfinance, 2004, Vol. 6, No. 2, 11 p. URL: http://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=1078&context=esr
53. Statistics glossary. European Central Bank. [https://www.ecb.europa.eu/home/glossary/html/act2c.en.html]. 12.10.2015.
54. Study on the functioning of the consumer credit market in Europe. European Commission, 2013, 511 p. [http://ec.europa.eu/consumers/archive/rights/docs/consumer_credit_market_study_en.pdf]. 16.11.2015.
55. **Šušteršič, M., Mramor, D., Zupan, J.** Consumer credit scoring models with limited data. – Expert Systems with Applications, 2009, Vol. 36, No.3, pp. 4736–4744. URL: http://www.sciencedirect.com/science/article/pii/S0957417408002996

56. **Thomas, L.C.** A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. – International Journal of Forecasting, 2000, Vol. 16, pp. 149-172. URL: http://pages.ucsd.edu/~aronatas/project/academic/A%20survey%20of%20credit%20and%20behavioural%20scoring%20Forecasting%20fina.pdf
57. **Thomas, L.C.** Consumer Credit Models: Pricing, Profit and Portfolios. Oxford: Oxford University Press, 2009, 385 p.
58. **Thomas, L.C., Edelman, D.B., Crook, J.N.** Credit Scoring and Its Applications. – United States of America: Society for Industrial and Applied Mathematics, 2002, 248 p. URL: https://books.google.ee/books?id=qalLfKNve6sC&redir_esc=y
59. **Tong, E.N.C., Mues, C., Thomas, L.C.** Mixture cure models in credit scoring: If and when borrowers default. – European Journal of Operational Research, 2012, Vol. 218, No. 1, pp. 132-139. URL: http://www.sciencedirect.com/science/article/pii/S0377221711009064
60. UNO Järelmaks. [https://www.unojarelmaks.ee/et/meist]. 10.11.2015.
61. UNO järelmaksu olulised tingimused. OÜ Kaupmehe Järelmaks. 16.02.2011. (käskkiri)
62. **West, D., Dellana, S., Qian, J.** Neural Network ensemble strategies for financial decision applications. – Computers & Operations Research, 2005, Vol. 32, No. 10, pp. 2543-2559. URL: http://www.sciencedirect.com/science/article/pii/S0305054804000693
63. Äriregistri teabesüsteem. [https://ariregister.rik.ee]. 02.01.2016.
64. **Yap, B.W., Ong, S.H., Husain, N.H.M.** Using data mining to improve assessment of credit worthiness via credit scoring models. – Expert Systems with Applications, 2011, Vol. 38, No. 10, pp. 13274–13283. URL: http://www.sciencedirect.com/science/article/pii/S0957417411006749

LISAD

Lisa 1. Kategooriliste muutujate võimalikud väärtused

Muutuja	Võimalikud väärtused
sugu	• mees • naine
perekonnaseis	• vallaline • vabaabielus • abielus • lahutatud • lesk
haridustase	• põhiharidus • keskharidus • keskeriharidus • kõrgharidus
ülalpeetavate arv	• 0 • 1 • 2 • 3 • 5 ja enam
elukoha tüüp	• üürikorter • ühiselamu • isiklik korter • isiklik maja • vanematega koos • muu
postiaadressi maakond	• Viljandimaa • Harjumaa • Pärnumaa • Võrumaa • Tartumaa • Läänemaa • Valgamaa • Saaremaa • Ida-Virumaa • Lääne-Virumaa • Raplamaa • Järvamaa • Jõgevamaa • Põlvamaa • Hiiumaa

Allikas: autori koostatud.

Lisa 1 järg. Kategooriliste muutujate võimalikud väärtused

Muutuja	Võimalikud väärtused
postiaadress linn	• Tallinn • Pärnu • Tartu • Narva • Kohtla-Järve • Viljandi • Muu
tegevusala	• Ettevõtja • Palgatöötaja • Pensionär • Pensionär/Palgatööline • Muu
maksehäirete arv taotlemise hetkel	• 0 • 1-2 • 3+
maksehäire esinemine	• 0 • 1

Allikas: autori koostatud.

Lisa 2. Kirjeldav statistika maksehäireta lepingute korral (default = 0)

Tunnus	Keskväärtus	Standard-hälve	Mediaan	Min.	Maks.
loanAmount	502,83	247,58	464,00	125,00	2818,00
loanPeriodMonths	20,38	12,70	18,00	3,00	60,00
monthlyIncome	647,56	353,62	590,00	130,00	2500,00
ageWhenApplying	38,36	13,13	37,00	18,00	72,00
workedAtMonths	69,24	69,17	50,00	1,00	498,00

Allikas: autori koostatud.

Lisa 3. Kirjeldav statistika maksehäirega lepingute korral (default = 1)

Tunnus	Keskväärtus	Standard-hälve	Mediaan	Min.	Maks.
loanAmount	641,46	311,18	577,00	154,00	2324,00
loanPeriodMonths	27,20	13,44	24,00	6,00	60,00
monthlyIncome	602,63	313,64	542,00	130,00	2000,00
ageWhenApplying	31,29	13,02	27,00	18,00	64,00
workedAtMonths	35,34	38,92	20,00	1,00	278,00

Allikas: autori koostatud.

Lisa 4. Tunnuste väärtuste esinemise sagedus maksehäire esinemise järgi

Tunnus	Väärtus	Maksehäirega	Maksehäireta	Kokku
sugu				
	mees	187	1782	1969
	naine	59	1873	1932
perekonnaseis				
	vallaline	154	1159	1313
	vabaabielus	47	850	897
	abielus	33	1359	1392
	lahutatud	10	183	193
	lesk	2	104	106
haridustase				
	põhiharidus	62	470	532
	keskharidus	94	1349	1443
	keskeriharidus	63	1196	1259
	kõrgharidus	27	640	667
ülalpeetavate arv				
	0	189	2074	2263
	1	37	855	892
	2	16	539	555
	3	2	147	149
	5 ja enam	2	40	42
elukoha tüüp				
	üürikorter	69	408	477
	ühiselamu	2	11	13
	isiklik korter	103	2010	2113
	isiklik maja	38	1010	1048
	vanematega koos	32	193	225
	muu	2	23	25

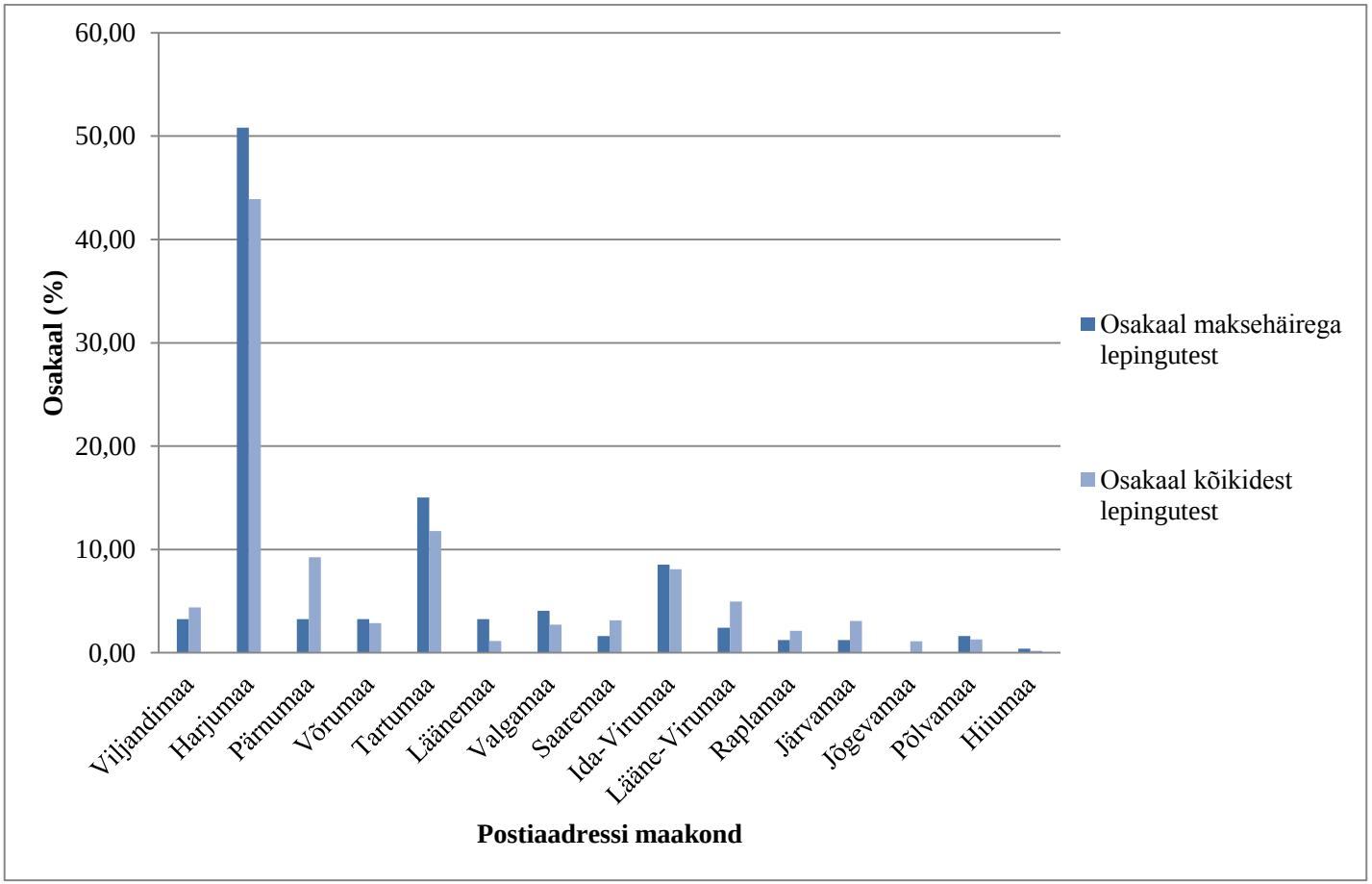
Allikas: autori koostatud.

Lisa 4 järg. Tunnuste väärtuste esinemise sagedus maksehäire esinemise järgi

Tunnus	Väärtus	Maksehäirega	Maksehäireta	Kokku
postiaadressi maakond				
	Viljandimaa	8	163	171
	Harjumaa	125	1587	1712
	Pärnumaa	8	353	361
	Võrumaa	8	104	112
	Tartumaa	37	422	459
	Läänemaa	8	37	45
	Valgamaa	10	96	106
	Saaremaa	4	119	123
	Ida-Virumaa	21	294	315
	Lääne-Virumaa	6	187	193
	Raplamaa	3	80	83
	Järvamaa	3	117	120
	Jõgevamaa	0	43	43
	Põlvamaa	4	46	50
	Hiiumaa	1	7	8
postiaadress linn/ piirkond				
	Tallinn	108	1118	1226
	Pärnu	4	188	192
	Tartu	30	273	303
	Narva	9	72	81
	Kohtla-Järve	4	95	99
	Viljandi	1	67	68
	Muu	90	1842	1932
tegevusala				
	Ettevõtja	3	29	32
	Palgatöötaja	217	3242	3459
	Pensionär	20	241	261
	Pensionär/Palgatööline	3	102	105
	Muu	3	41	44
maksehäirete arv taotlemise hetkel				
	0	176	3253	3429
	1-2	69	344	413
	3+	1	58	59

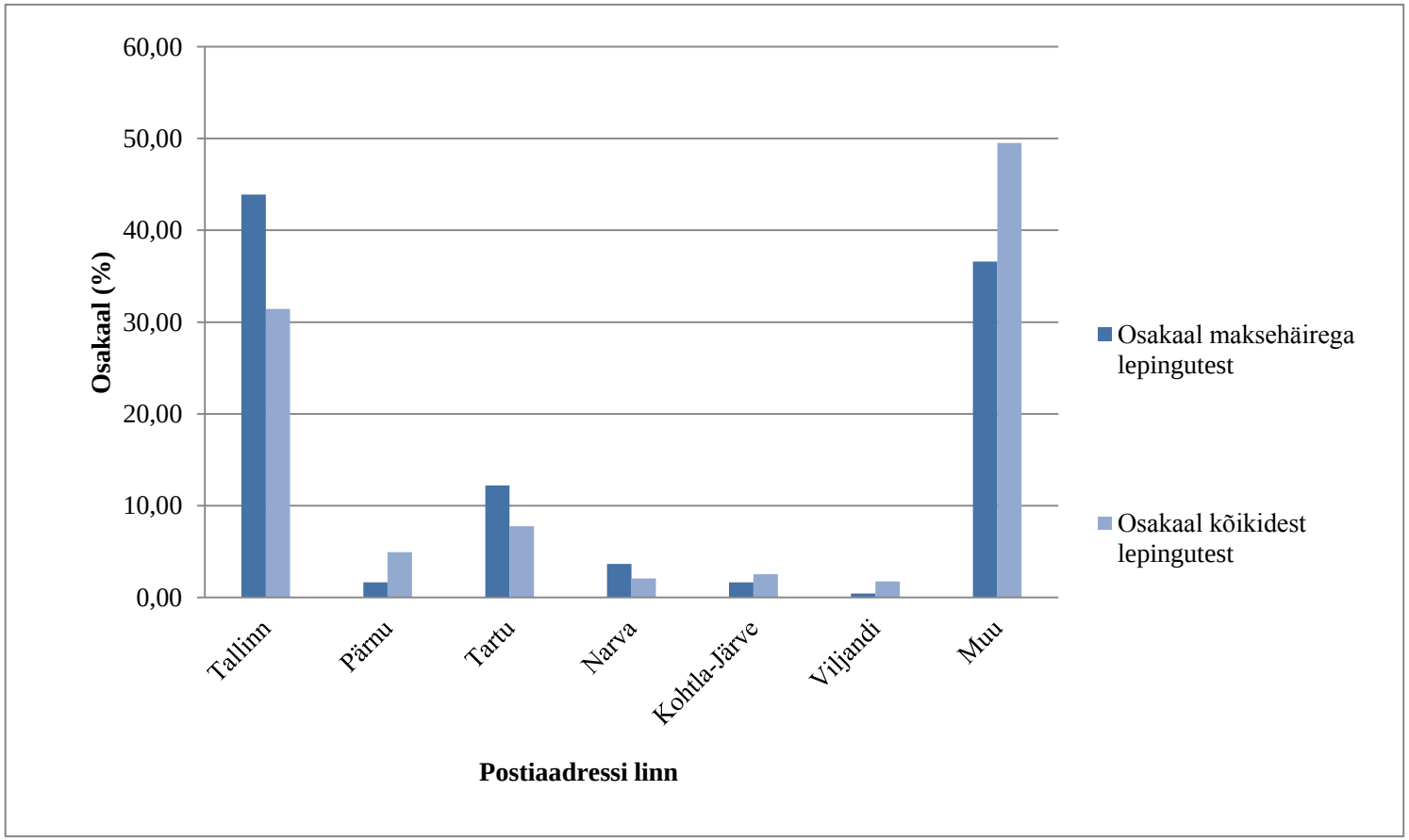
Allikas: autori koostatud.

Lisa 5. Maksehäire esinemise osakaal valimist lepingulise kliendi postiaadressi maakonna järgi



Allikas: autori koostatud.

Lisa 6. Maksehäire esinemise osakaal valimist lepingulise kliendi postiaadressi linna järgi



Allikas: autori koostatud.

Lisa 7. Mudeli M1 vigade maatriks

Hinnatud klass Tegelik klass	0 (maksehäireta)	1 (maksehäire)
0 **(maksehäireta)**	1821 (99,56%)	8 (0,44%)
1 **(maksehäire)**	99 (77,95%)	28 (22,05%)

Allikas: autori koostatud.

Lisa 8. Mudeli M2 vigade maatriks

Hinnatud klass Tegelik klass	0 (maksehäireta)	1 (maksehäire)
0 **(maksehäireta)**	1444 (99,24%)	11 (0,76%)
1 **(maksehäire)**	91 (85,85%)	15 (14,15%)

Allikas: autori koostatud.

Lisa 9. Mudeli M3 vigade maatriks

Hinnatud klass Tegelik klass	0 (maksehäireta)	1 (maksehäire)
0 **(maksehäireta)**	1098 (99,91%)	1 (0,09%)
1 **(maksehäire)**	58 (81,69%)	13 (18,31%)

Allikas: autori koostatud.

Lisa 10. Mudeli M4 vigade maatriks

Hinnatud klass Tegelik klass	0 (maksehäireta)	1 (maksehäire)
0 **(maksehäireta)**	1822 (99,73%)	5 (0,27%)
1 **(maksehäire)**	111 (90,24%)	12 (9,76%)

Allikas: autori koostatud.

Lisa 11. Mudeli M5 vigade maatriks

Hinnatud klass Tegelik klass	0 (maksehäireta)	1 (maksehäire)
0 **(maksehäireta)**	1455 (99,52%)	7 (0,48%)
1 **(maksehäire)**	82 (83,67%)	16 (16,33%)

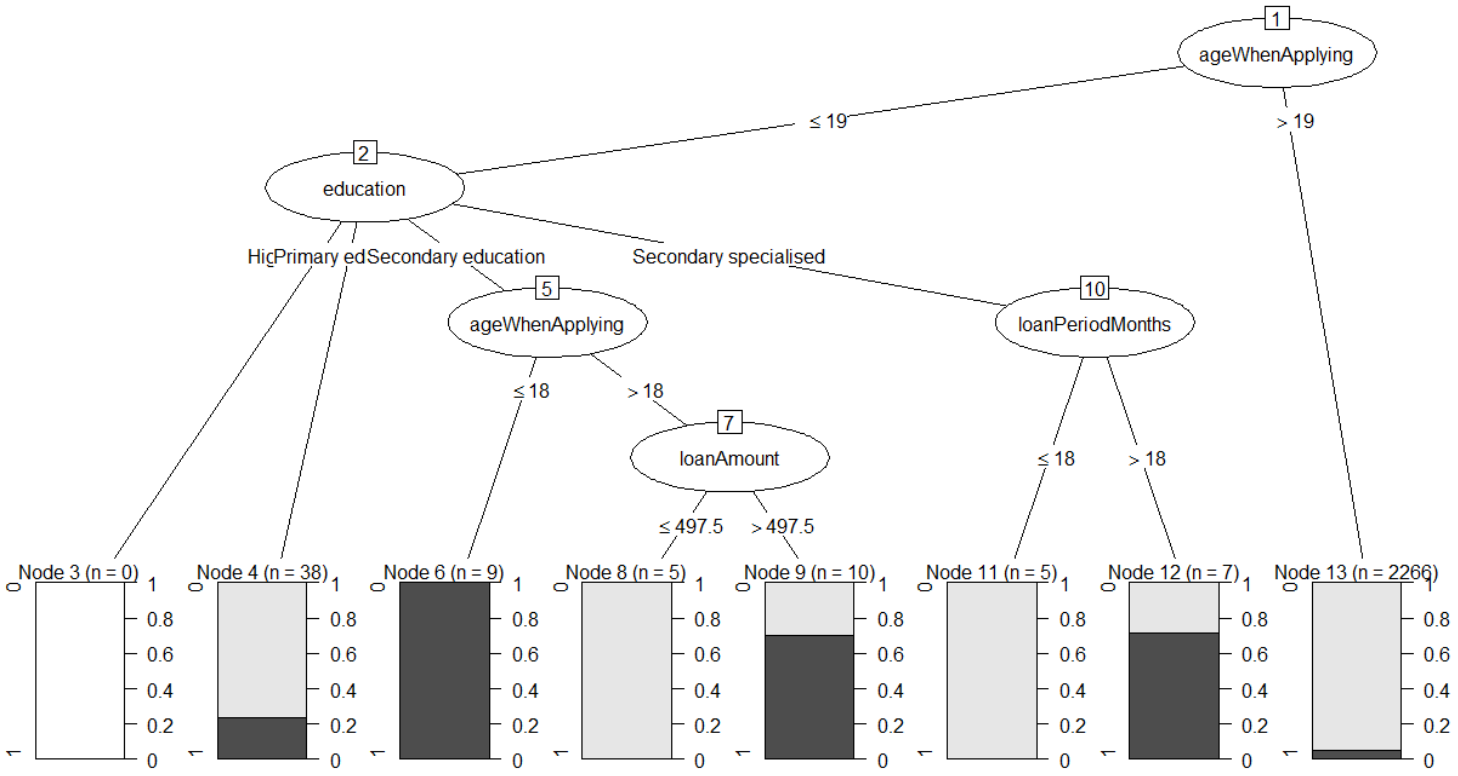
Allikas: autori koostatud.

Lisa 12. Mudeli M6 vigade maatriks

Hinnatud klass Tegelik klass	0 (maksehäireta)	1 (maksehäire)
0 **(maksehäireta)**	1088 (99,27%)	8 (0,73%)
1 **(maksehäire)**	57 (78,08%)	16 (21,92%)

Allikas: autori koostatud.

Lisa 13. Mudeli M2 otsustuspuu



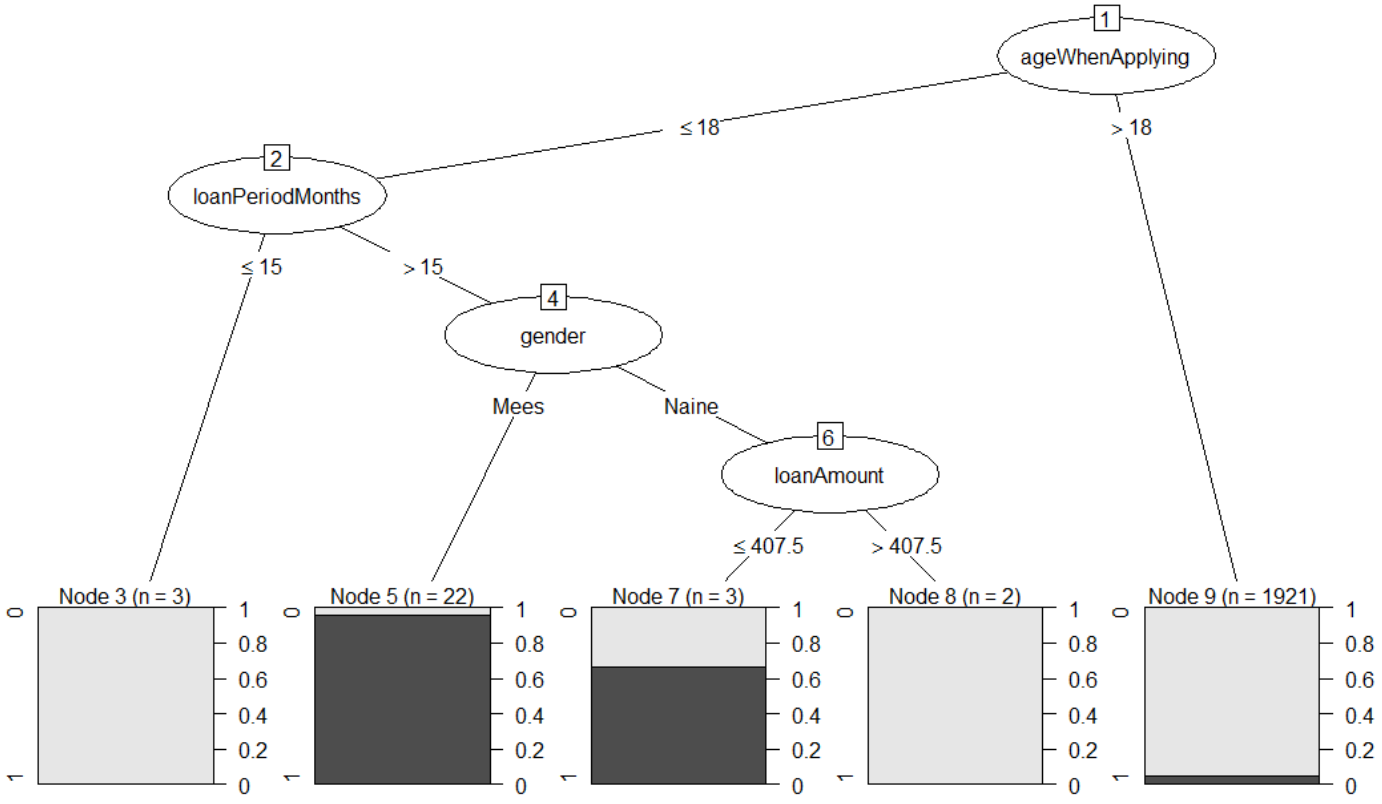
Märkused: Helehalliga on tähistatud maksehäireta ja tumehalliga maksehäirega vaatluste (n) osakaal lehel („*node*“). Valgega on tähistatud lehed, kuhu alla ei kuulu ühtegi vaatlust.
Allikas: autori koostatud.

Lisa 14. Mudeli M3 otsustuspuu

```
ageWhenApplying <= 19
|   gender = Mees
|   |   ageWhenApplying <= 18
|   |   |   education = Higher education: 1 (0.0)
|   |   |   education = Primary education
|   |   |   |   loanPeriodMonths <= 21: 0 (4.0)
|   |   |   |   loanPeriodMonths > 21
|   |   |   |   |   loanAmount <= 434: 0 (2.0)
|   |   |   |   |   loanAmount > 434: 1 (9.0/1.0)
|   |   |   education = Secondary education: 1 (10.0)
|   |   |   education = Secondary specialised: 1 (7.0)
|   |   ageWhenApplying > 18
|   |   |   monthlyIncome <= 433: 0 (14.0/1.0)
|   |   |   monthlyIncome > 433
|   |   |   |   postalAddressTownArea = Kohtla-Jarve: 1 (0.0)
|   |   |   |   postalAddressTownArea = Muu: 1 (7.0/2.0)
|   |   |   |   postalAddressTownArea = Narva: 1 (0.0)
|   |   |   |   postalAddressTownArea = Parnu: 1 (0.0)
|   |   |   |   postalAddressTownArea = Tallinn
|   |   |   |   |   workedAtMonths <= 7.2: 0 (2.0)
|   |   |   |   |   workedAtMonths > 7.2: 1 (8.0/1.0)
|   |   |   |   postalAddressTownArea = Tartu: 0 (2.0/1.0)
|   |   |   |   postalAddressTownArea = Viljandi: 0 (3.0)
|   gender = Naine: 0 (24.0/2.0)
ageWhenApplying > 19: 0 (2638.0/134.0)
```

Märkused: Lehega seotud klassi järel sulgudes on esimesena märgitud sama klassi omavate vaatluste arv, teisena teist klassi.
Allikas: autori koostatud.

Lisa 15. Mudeli M4 otsustuspuu



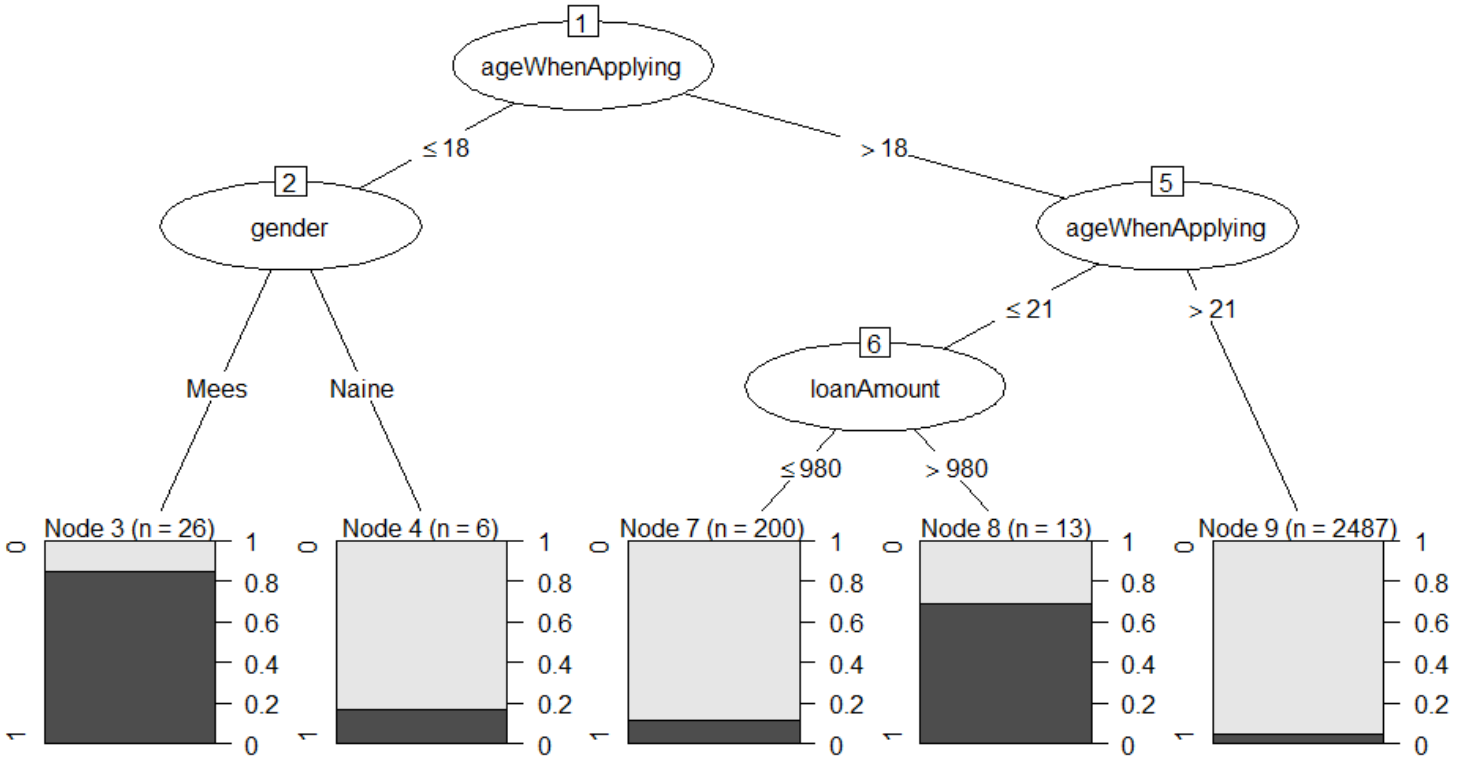
Märkused: Helehalliga on tähistatud maksehäireta ja tumehalliga maksehäirega vaatluste (n) osakaal lehel („*node*“).
Allikas: autori koostatud.

Lisa 16. Mudeli M5 otsustuspuu

```
ageWhenApplying <= 18
|   gender = Mees: 1 (25.0/3.0)
|   gender = Naine
|   |   education = Higher education: 0 (0.0)
|   |   education = Primary education: 0 (4.0)
|   |   education = Secondary education: 1 (3.0/1.0)
|   |   education = Secondary specialised: 0 (0.0)
ageWhenApplying > 18
|   loanAmount <= 827: 0 (2092.0/92.0)
|   loanAmount > 827
|   |   loanPeriodMonths <= 13: 0 (79.0)
|   |   loanPeriodMonths > 13
|   |   |   ageWhenApplying <= 20
|   |   |   |   monthlyIncome <= 475: 0 (4.0/1.0)
|   |   |   |   monthlyIncome > 475: 1 (7.0)
|   |   |   ageWhenApplying > 20: 0 (127.0/24.0)
```

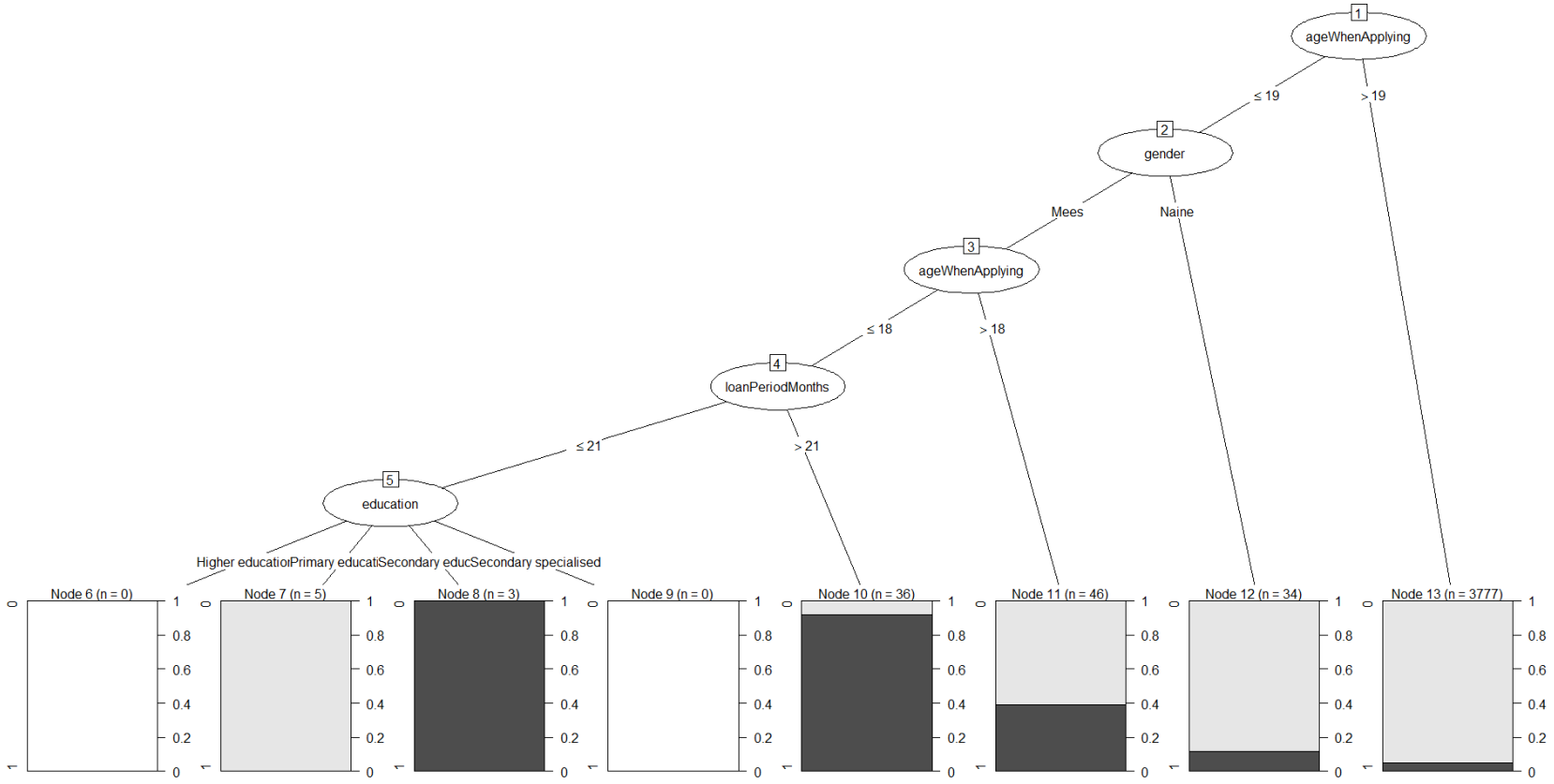
Märkused: Lehega seotud klassi järel sulgudes on esimesena märgitud sama klassi omavate vaatluste arv, teisena teist klassi.
Allikas: autori koostatud.

Lisa 17. Mudeli M6 otsustuspuu



Märkused: Helehalliga on tähistatud maksehäireta ja tumehalliga maksehäirega vaatluste (n) osakaal lehel („*node*“).
Allikas: autori koostatud.

Lisa 18. Mudeli M7 otsustuspuu



Märkused: Helehalliga on tähistatud maksehäireta ja tumehalliga maksehäirega vaatluste (n) osakaal lehel („*node*“). Valgega on tähistatud lehed, kuhu alla ei kuulu ühtegi vaatlust.
Allikas: autori koostatud.

SUMMARY

CONSUMER CREDIT RISK MODELLING IN THE EXAMPLE OF KAUPMEHE JÄRELMAKS LTD.

Keit Adamson

The actuality of consumer credit risk modelling is highlighted by the fact that in recent years the volume of consumer loans has been on a rising trend in Estonia. The increase in loans has increased competition to a level on which anticipating the solvency of private citizens has become progressively important to creditors, especially since the margin of error for loan companies has reduced due to prevailing price pressure in the market and government regulations. More efficient credit risk evaluation translates to more accurate provisioning, which reduces the cost of capital for the creditors. Furthermore, a higher classification accuracy enables the creditor to assess a specific credit case in accordance with its credit risk making it possible to provide a more favorable loan to clients with lower risk. Also, when dealing with larger volumes of loans even a slight improvement in the classification accuracy may result in a significant reduction of costs for the creditor. Credit risk modelling has a beneficial aspect to our society as a whole as well. Along with a more effective selection process fewer people with actual lack of means for servicing their loans will be met with their needs which prevents them from impairing their financial situation even more.

The purpose of the Master's thesis is to formulate a model for credit risk assessment using the decision tree method in the example of Kaupmehe Järelmaks Ltd. Research conducted is limited to the decision tree method since it is regarded in the academic literature as highly interpretable and with good classification accuracy. Also, the lack of Estonian academic literature on the subject may be regarded as one of the reason for the choice of methods. The results of the research may be useful and find further advancement in the credit risk policy and improvement of models within the examined company. Following research tasks are formulated to achieve set goals:

- Give a preview of consumer credit risk based on academic literature, methodology used in assessing consumer credit risks and classification accuracy of the methods.
- Examine consumer credit risk assessment in the context of most used explanatory variables and their effects.
- Create a credit risk assessment model on the basis of C4.5 method.
- Analyze the direction of effect of the model-based variables and compare them with the results acquired in academic literature.
- Assess and analyze the classification accuracy of the models.

Data used in the thesis have been acquired from the database of Kaupmehe Järelmaks Ltd. Sample consists of 3901 observations, which constitute a random sample from hire purchase contracts, that were concluded in 2011. For every contract in the database, applicant's gender, age at the time of application, marital status, level of education, number of dependants, type of residence, mailing address, county, postal address city, occupation, time of employment at the point of application (current occupation), monthly income in euros, payment defaults at the time of application, loan value in euros, loan period in months and payment default occurence/non-occurence in the contract are known.

To assess consumer credit risk a decision tree algorithm J48, which is an implementation of J.S. Quinlan's algoritm C4.5 in Java programming language, is used. Initial sample is divided in two. On the first sub-sample the model is developed and the other is used to assess classification accuracy using PCC („percentage correctly classified") and ROC („receiver operating characteristics") area under the curve. For further information about the model additional indicators TPR („true positive rate") and TNR („true negative rate") are used since in the case of greater imbalance between the classes PCC does not adequately reflect the classificator's capacity to predict when it comes to differentiating classes. Two different strategies are used to formulate random sub-samples. Firstly, test and training samples are altered in relation to the general sample. Additionally a strategy with class ratio restriction is used, where random data records are divided into sub-samples with an aim to retain the initial class balance of

defaulted and non- defaulted loan contracts. As a result the following samples were generated:

- Training and test sample make up 50% and 50% accordingly from the initial sample.
- Training and test sample make up 60% and 40% accordingly from the initial sample.
- Training and test sample make up 70% and 30% accordingly from the initial sample.
- Training and test sample make up 50% and 50% accordingly from the initial sample and the class ratio is retained.
- Training and test sample make up 60% and 40% accordingly from the initial sample and the class ratio is retained.
- Training and test sample make up 70% and 30% accordingly from the initial sample and the class ratio is retained.

Based on all the training samples a decision tree model is generated with the J48 algorithm. Additionally, a model using the same algorithm is generated from the whole sample where classification accuracy is assessed by dividing the original sample into ten independent test samples where the original class ratio is attempted to be retained. To resolve an issue of overfitting, the original tree is pruned using confidence level of 0.25, which is also the default value for it. In all of the models generated in the research age at the time of application turned out to be a relevant variable, inversely, marital status, number of dependants, type of residence, occupation, mailing address, county and the number of payment defaults were not used. Some variables like marital status, number of dependants, type of residence and occupation the importance of which based on the theoretical part of the research one would expect, could have been left out of the models because the occurence of the variables of observations with certain values were excluded from the thesis as a result of the credit policy at the time of the loan decision. To be precise, in accordance with credit regulations loans were not granted to clients with payment disturbances in banking sector or any other finance sector, that was unfinished or had ended less than six months ago.

Models M1 and M6 had the highest classification accuracy with the AUC value of 0.61 and PCC values 94.53% and 94.44% accordingly. Pursuant to models M1 and M6 the important variables were clients age at the time of application and applicant's gender. Additionally, in the model M1 the time spent under current employer in months and montly income was considered significant, as was the loan amount in model M6. All afore-mentioned variables included in the models have a high or average frequency of use based on the academic literature used in the research. In both models credit risk of women is lower compared to men in certain sub-groups which coincides with the results in previous findings. It is apparent from the models that applicants with a higher age have less chance to experience payment disturbances. Model M1 can't differentiate between clients with payment difficulty and those without if the clients age exceeds 19 years. A similar problem occurs with model M6 but in excess of age 21. The results from earlier research only partially coincide with the applicants age since predominantly the probability of experiencing payment difficulty decreases at a higher age group. Some of the reasons for this may be equal treatment of both „good" and „bad" loans in the models as well as the method's parametrization which influences the final depth of generated decision tree. In the model M6 the directional effect of the loan amount coincides with the theoretical part of the thesis according to which the probability of payment disturbance increases with the loan amount. According to model M1 higher monthly income and longer period of time spent under current employer entail higher credit risk for male creditors over the age of 19 which does not correspond to results from academic literature. At this point it is important to stress that in the scientific articles used in the thesis most of the conclusion made were based on models, that are modelling linear relationship and unlike the decison tree method do not allow interaction between different values of independent variables.

Based on the results of Master's thesis it can be concluded that if Kaupmehe Järelmaks Ltd. had complemented existing credit policy with rules from two classificators with the highest accuracy, every fifth customer with a payment disturbance would have been classified correctly in addition. At the same time the number of non-default clients classified as defaulting customers would have been very low.

In many cases the direction of effect of the variables does not coincide with the findings in academic literature, which may refer to the non-linear nature of the data used from the perspective of credit risk since most of the research describing the direction of effect used methods with an assumption of linearity of certain kind. To adress this position it is necessary to use such methods on the same sample and compare classification accuracy with current results. That would be one way to advance the subject matter.

A limitation of the thesis could be that the size of the sample is not large enough since the method C4.5 needs a substantial amount of observations for reliable results, especially in case of high class imbalance. It appeared from the models that in some cases no observations corresponded to leaves, thus a dominating class was applied. Reject inherence, which is caused by not knowing the payment behaviour of applicants who were rejected in loan granted process, is a fundamental restriction in consumer credit risk modelling. Consequently, the results may be distorted because there is a high probability of there being a number of applicants who would had no payment difficulties if a loan had been granted. As a result developed models can over- or underestimate variable's effect on the dependant variable or the direction of effect is false. Additionally, another restriction could be the equal handling of the costs related the researched classes in the model since in practice there are usually higher costs related to contracts with payment issues. The C4.5 algorithm enables the use of costs but it is not used in this thesis because necessary information was not disclosed to the author of the thesis by Kaupmehe Järelmaks Ltd.

Further development of the current paper is possible for example by using the same sample but other methods described in the theoretical part of the thesis thus creating new models for consumer credit risk assessment. One method could be the next version of the algorithm C4.5 which is C5.0 which according to the method's author surpasses it's predecessor in some cases with accuracy. Created models could be compared from the standpoint of directional effects of the variables or classification accuracy. Thus it would make further conclusions possible between the non-linear relation of explanatory variables and subject characteristics. Since the created model could not differentiate elderly applicants with or without payment difficulties, additional characteristics may be introduced. Kaupmehe Järelmaks Ltd. could consider implementing methodology from

the thesis on latter contract samples ascertaining whether the methodology is able to correctly classify along side the current credit model observations with payment difficulty. In the event of a positive result, it is possible to implement the newly acquired rules in to the company's banking system. If the supplementation of the current model is for some reason not efficient the observation with a negative decision could be incorporated to generate a new model. Since the goal of the company is maximizing profit it is possible to further develop the thesis by involving cost of different classes. To achieve this the costs of incorrectly classified non-defaulting and defaulting loans have to be determined. On the one hand the company loses profit if a loan is not granted to a client who in reality would not default. On the other hand the debitor with payment difficulty shall bear the costs which depend on the final amount recovered and costs related to collection.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Keit Adamson,

(sünnikuupäev: 13.02.1987)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Eraisiku krediidiriski modelleerimine ettevõtte Kaupmehe Järelmaks OÜ näitel“, mille juhendaja on Oliver Lukason,

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu alates 11.06.2021 kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, **25.05.2016**