TARTU ÜLIKOOL
Majandusteaduskond

Ekke Sakkov

PANKROTISTUNUD LAENUDESSE INVESTEERIMINE BONDORA.EE LAENUKESKKONNAS

Magistritöö sotsiaalteaduse magistrikraadi taotlemiseks majandusteaduses

Juhendajad: Kurmet Kivipõld ja Hendrik Luuk

Tartu 2018

Soovitan suunata kaitsmisele ……..
(juhendaja allkiri)

Kaitsmisele lubatud “ “..... 2018. 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

Investeerimine on Eestis viimastel aastatel järsult populaarsust kogunud. Enam ei ole see vaid rikkale eliidile reserveeritud tegevus, vaid ühisrahastuse tulek on võimaldanud investeerimist alustada ka vaid viieeurosest sissemaksest. Investeerimiskeskkondi ja nende poolt pakutavaid võimalusi tekib pidevalt juurde. Kõik pakuvad uutele potentsiaalsetele investoritele teed finantsvabaduseni ja selleks pakutavaks teeks on investeerimine just nende pakutavatesse investeerimistoodetesse.

Käesoleva töö peamiseks fookuseks on ennustavate statistilise meetodite rakendamine investeerimiskontekstis. Selleks, et teoreetilist tagapõhja mõne Eestis pakutava investeerimisvõimaluse raames proovile panna, viib autor läbi juhtumiuuringu pankrotistunud laenudesse investeerimise strateegia kohta laenukeskkonnas Bondora. Juhtumiuuring sisaldab endas teoorias käsitletud statistiliste ennustusmeetodite kasutamist eraisikutele antavate laenude taastumise etteennustamiseks, et saavutada olukord, kus üle 90 päeva viivises olnud ehk pankrotistunud laenudesse investeerimise strateegiat rakendav investor võib pankrotistunud laenusid kokku ostes teenida kasumit. See võib suurte allahindluste korral tähendada kasvõi 80% aastatootlust ja seega mäekõrguselt tulususe poolest ületada kõik turul pakutavad investeerimistooted.

Käesolev töö annab muuhulgas ülevaate ka Eesti investeerimisturu hetkeseisust ja sellest, kuidas statistilisi ennustusmeetodeid Eesti investeerimisturu kontekstis autori arvates rakendada saaks. Kindlasti omab väärtust autori poolt loodav tabel, mis võimaldab võrrelda erinevaid ennustusmeetodeid muuhulgas ka nende täpsuse alusel. Autor ei pretendeeri siinjuures absoluutsele tõele, sest olukord igas käesoleva töö problemaatikat puudutavas aspektis on pidevalt muutuv ja seega saab investeerimislähenemisi võrreldes rääkida vaid autori arvamusest, mis tugineb igal võimalusel faktilistele tõenditele.

Käesoleva magistritöö eesmärgiks on läbiviidava juhtumiuuringu käigus uurida, kas bondora.ee laenukeskkonnas on võimalik pikaajaliselt viivises olnud ehk pankrotistunud laenudesse investeerimise strateegiat edukalt rakendada. Eesmärgi saavutamiseks kasutab autor investeerimisotsuste põhistamisel statistilisi ennustusmeetodeid. Juhtumiuuring hõlmab endas krediidiandmete korrastamist, töötlemist ja statistilist analüüsi.

Magistritöö uurimisülesanded tulenevad läbiviidavast juhtumiuuringust:

- kirjeldada Eesti investeerimismaastiku hetkeolukorda ja võimalusi;
- süstematiseerida enam levinud ennustusmeetodeid, mida saab mõne olulise muutuja etteennustamiseks kasutada;
- anda ülevaade eelnevalt kirjutatud töödest investeeringute tulemuse ennustamise kontekstis;
- kirjeldada tööks kasutatavaid andmeid ja nende analüüsiks ettevalmistamist;
- analüüsida pankrotist taastumise etteennustamise võimalikkust;
- teha järeldused juhtumiuuringu tulemustest.

Magistritöö teoreetiline osa hõlmab seega Eesti investeerimismaastiku tutvustust, erinevate statistiliste ennustusmeetodite kirjeldamist ja ülevaadet eelnevatest töödest, mis käesoleva töö kontekstiga oluliselt kattuvad. Näiteks nii Kisand (2015) kui ka Haltuf (2014) on mõlemad lähenenud inimeselt-inimesele laenukeskkondade problemaatikale ja tootluse optimeerimisprobleemile analüütiliselt. Üldisemat tausta annavad näiteks Zhou (2012) õpik ansambelmeetodite kohta ja Williams (2011) õpikuga R-is tehtava masinõppe kohta. Autori panuseks teoreetilises osas on luua eelnevate autorite töödest sünteesida üldistatum arusaam laiast meetodivalikust kitsa probleemi lahendamisel, pakkuda ajakohastatud ülevaade Eesti ühisrahastusturust ja pakkuda võimalik laiendus portfelli optimeerimist käsitlevale metoodikale. Teoreetiline osa on otseselt empiirilise osa eeltingimuseks, sest juhtumiuuring viiakse läbi teoreetilises osas väljatoodud meetodeid kasutades.

Magistritöö empiirilise osa moodustab juhtumiuuring pankrotistunud laenudesse investeerimise strateegia rakendatavuse ja edu kohta bondora.ee laenukeskkonnas. Esmalt annab autor ülevaate kasutavatest andmetest. Andmestikuks on bondora.ee laenukeskkonna poolt väljastatav LoanData andmefail, mis käsitleb kõiki nende keskkonnas tehtud laenutehinguid ja nende tulemusi andmete väljavõtmise kuupäevaga. Andmestikku tuleb seega enne kohandada, et see ennustamiseks sobiks – näiteks ei saa

arvesse võtta laene, mis pole kunagi pankrotti läinud, sest mitte ükski neist ei ole ka pankrotist taastunud. Pankrotist taastumise ennustamiseks rakendab autor teoreetilises osas toodud meetodeid ja leiab võimalused meetodite sooritusvõime hindamiseks ja omavaheliseks võrdlemiseks. Juhtumiuuringu tulemustest järelduste tegemiseks võrdleb autor leitud tulemusi eelnevate autorite poolt leituga ja annab hinnangu läbiviidud uuringu edukusele.

Selle magistritööga seonduvad märksõnad: ühisrahastamine, investeerimine, ennustamine, masinõpe, närvivõrk, tugivektormasin, otsustuspuu, võimendamine, logistiline regressioon, inimeselt-inimesele, mudel, tootlus, risk, krediidiskooring, laenud, klassifitseerimine.

1. INVESTEERIMINE JA SELLE TULEMUSLIKKUSE ENNUSTAMINE

1.1. Investeerimisvõimalused ja ühisrahastus Eestis

Käesoleva töö mõistmise jaoks on tähtis teha vahet investeerimisel ja spekuleerimisel ning saada aru, kuidas need suhestuvad investori tegevusega inimeselt-inimesele laenuturgudel. Investeerimise all mõeldakse üldiselt kasu saamise eesmärgil tehtud pikaajalist kapitalimahutust (Mereste 2003). Spekuleerimine seevastu on mingisuguse vara ülesostmine ja kasusaamiseks kõrgema hinnaga mahamüümine (*ibid*). Mõlemad mõisted seostuvad investori tegevusega inimeselt-inimesele laenuturgudel küllaltki tugevasti, sest laenuturg pakub nii võimalusi investeerimiseks kui ka spekuleerimiseks. Kui vaadelda käesoleva töö fookuseseks olevat investeerimistegevust, pankrotistunud (üle 90 päeva viivises olnud) laenude ülesostmist, siis selles tegevuses on elemente nii investeerimise kui ka spekuleerimise mõistetest.

Ostes kelleltki üles juba tehtud, kuid siis ebaõnnestunuks osutunud investeeringu, nagu näiteks osaluse pankrotistunud laenus, on laenu taastumise korral võimalik teenida kasu nii hinnamuutusest (spekuleerimine) kui ka uuest seatavast laenu tagasimaksegraafikust (investeerimine). Kuigi kindlat eralduspiiri nende kahe mõiste (investeerimine ja spekuleerimine) vahel ei ole, siis vaadeldavas näites eraldab spekulatsiooni ja investeerimist väga selgelt ajahorisont. Laenu tagasimaksegraafik on enamasti viie aasta pikkune, ehk võimaldab investorile igakuiseid rahavooge viieaastase perioodi vältel. Pankrotist taastumine seevastu suurendab tehtud investeeringu väärtust allahindlusprotsendi võrra, sest enamasti müüakse osalusi pankrotistunud laenutehingutes väga suure allahindlusprotsendiga, lootes midagigi tehtud investeeringust tagasi saada. See tähendab, et spekulatiivne väärtuse kasv võib juhtuda põhimõtteliselt hetkega. Spekulandi huvides on seega võimalikult hästi ette ennustada, millised laenud kõige suurema tõenäosusega taastuvad. Käesoleva töö kontekstis

käsitleb autor investeerimist ja spekuleerimist kui raskesti eristatavaid mõisteid ja andes ülevaadet investeerimisvõimalustest ja nende arengust Eestis ei pea autor oluliseks nendel igal sammul vahet teha.

1.1.1. Ühisrahastamine kui investeeringute kategooria

Üldiselt võib investeerimisvõimalused Eestis jaotada kolme suurde kategooriasse. Nendeks on klassikalised varaklassid nagu aktsiad ja kinnisvara ning kõige uuem investeerimismoodus, viimasel ajal järsult populaarsust kogunud ühisrahastus. Klassikaliste investeerimisobjektide loetlemisel poleks käesoleva magistritöö kontekstis väärtust, seega keskendub autor just ühisrahastusvõimaluste võrdlemisele ja analüüsile. Kuigi küllaltki raske on välja selgitada ühe või teise investeerimismooduse oodatavat tulusust, siis enamasti on ühisrahastusest saadav tootlus kõrgem kui kinnisvarasse investeerides hinnamuutusest ja renditulust saadav tootlus ning kõrgem on see ka aktsiaturgude keskmisest tootlusest. See fakt põhjendab veelgi enam fokusseerimist just ühisrahastuse raames pakutavatele investeerimisvõimalustele.

Ühisrahastamine on viimasel kümnendil jõudsalt arenenud uus finantseerimisviis. Raha, mida on äriühingutel või ka füüsilistel isikutel vaja projekti elluviimiseks, kogutakse väikeste summade kaupa paljudelt inimestelt. Ühisrahastamisega seotud ettevõtlus areneb kiiresti ka Eestis. Ühisrahastamine toimub tavapäraselt avalike internetiplatvormide kaudu. (Finantsinpektsioon 2018)

Eestis tähendab ühisrahastus enamasti ühel või teisel viisil laenudesse investeerimist, kuid on ka erandeid. Ühisrahastuse põhimõte on äärmiselt lihtne – viiakse kokku inimesed, kellel on raha, mida välja laenata ja inimesed, kellel on raha vaja (Mild *et al* 2013: 1291). Ühisrahastuse parimaks omaduseks on sisenemisbarjääride puudumine. Viieeurose sissemaksega alustamine ei ole vaid loosung, vaid investeerimismaailma uus reaalsus, mis pakub kõigile võimaluse oma rahaga arvestatavat tootlust teenida. Oma lihtsusele vaatamata peidab ühisrahastus endas esmapilgul nähtamatuid riske ja erisusi, millest ülevaadet omada on kindlasti kasulik.

Esimeseks ühisrahastusportaaliks Eestis oli 2009. aastal tegevust alustanud isePankur, nüüdseks Bondora nime all tegutsev ettevõte. Populaarsed on ka Omaraha, MoneyZen

ja Läti portaalid nagu Twino ja Mintos. Üldiselt on mainitud portaalide kaudu investeerimine tehtud küllaltki lihtsaks. Investori jaoks lihtsuse ja vahendaja jaoks rahastusvõimekuse tagamise huvides on portaalid loonud automaatpakkujaid, mis reinvesteerivad teenitud tulu automaatselt investori poolt etteantud parameetrite järgi. Populaarsust on kogumas ka end kinnisvarasse investeerimise võimalusena turundavad CrowdEstate, EstateGuru ja BitOfProperty.

Ühisrahastuse kaudu väljastatavaid laene on kolme põhilist tüüpi (Madalvee 2016):

- Laenud eraisikutele – enamasti tagatiseta paarikuisest perioodist ja mõnesajast eurost 5+ aasta ja 10000+ euroni pakutavad tarbimislaenud.
- Laenud ettevõtetele – laenajaks on ettevõte, enamasti tagatisega.
- Faktooring – arvete ost ehk lühiajaline laen.

Nagu kõigi investeerimisvõimaluste puhul, tuleb ka ühisrahastusest rääkides välja tuua selle plussid ja miinused. Ühisrahastuse eelised ja puudused on toodud tabelis 1.

Tabel 1. Ühisrahastuse eelised ja puudused.

Eelised	**Puudused**
Väikesed või olematud teenustasud	Tulumaksu ei saa edasi lükata
Kättesaadavad andmestikud analüüsi läbiviimiseks	Riskantne varaklass
Sisenemisbarjääri puudumine	Tegeliku tootluse arvutamine on keerukas
Suhteliselt kõrge oodatav tootlus	Suhteliselt madal likviidsus – investeeringute müümine nõuab aega
Investeeringuid on võimalik maha müüa	Küllaltki reguleerimata valdkond
Automaatpakkuja kasutamise võimalus	Pole selge, milline oleks võimaliku majandussurutise mõju
Portaalid tegelevad laenuvõtjate taustakontrolliga ja pankrotistunud laenudega	Lühike ajalugu ja palju ebaselgust osapoolte õiguste kaitse osas

Allikas: (Madalvee 2016: 39)

Autori arvates on ühisrahastuse suurimateks tugevusteks väikesed või olematud teenustasud, sisenemisbarjääri puudumine ja see, et analüüsiks vajalikud andmestikud on kättesaadavad. Olematud teenustasud võimaldavad investoril suurema osa teenitud tulust endale jätta ja seeläbi suurendavad investeeringu tootlust. Sisenemisbarjääride puudumine annab võimaluse kõigil huvilistel investeerimisega alustada ja ühtlasi tähendab see võimalikult suurt potentsiaalset laenukapitali mahtu. Kättesaadavad andmestikud tähendavad andmeteaduse valdkonnas oskuslike investorite jaoks

võimalust teha ennustavaid mudeleid oma portfelli tootluse optimeerimiseks (Kisand 2015) või ka näiteks riski minimeerimiseks.

Samas on tegemist aga üsna riskantse varaklassiga (Mild *et al* 2013:1291) ja see on autori arvates ka ühisrahastuse suurimaks nõrkuseks. Ühisrahastuse riskantsus väljendub mitmes erinevas riskis (Finantsinspektsioon 2018):

- Raha kaotamise risk – investeerimisel on võimalik kaotada sissemakstud raha, eriti kui pakutakse laenu tegevusajaloota ettevõttele. Kinnisvara tagatisel antava laenu ebaõnnestumisel tagatiseks antud kinnisvara väärtus osutuda oodatust palju väiksemaks.
- Investeeringu pikaajalisus – ettevõtetele laenu andmise õnnestumine sõltub suuresti ettevõtte käekäigust ja tuleks arvestada pika tasuvusperioodiga.
- Finantsjärelevalve puudumine – ühisrahastusettevõte ei pruugi alati olla riikliku järelevalve all.
- Usalduse kuritarvitamine – investoritel tuleks enne investeerimist kontrollida platvormi omanike ja juhtkonna usaldusväärsust.
- Majanduse tsüklilisus – ühisrahastuse kaudu vahendatud varasemate projektide õnnestumine ei garanteeri sarnaste projektide õnnestumist tulevikus.

Lisaks sellele, mida Finantsinspektsioon mainib majanduse tsüklilisuse all on autori arvates selle punkti alla sobiv ka tõsiasi, et ühisrahastamine kui investeerimisliik ei ole veel üle elanud suuremat majanduslanguse perioodi ja seega ei ole üldse selge, kuidas mõjutab majanduslangus näiteks Bondorast võetavate laenude kvaliteeti, inimeste kalduvust minna petturluse teed või ka üldist tagasimaksmiste või pankrotist taastumiste protsenti. Pole ka näiteks selge, mis täpsemalt juhtuks siis, kui kinnisvara ühisrahastust pakkuvatel ettevõtetel poleks enam järku võtta ühtegi klienti, kellele raha pakkuda, sest keegi lihtsalt ei taha enam kinnisvara arendada.

1.1.2. Ühisrahastusportaalid Eestis

Järgnevalt annab autor tabelis 2 lühiülevaate valitud ühisrahastusportaalidest, kuhu on võimalik Eestis investeerida. Valik on tehtud silmas pidades eelkõige portaalide populaarsust arvestades, kuid ka autori subjektiivse arvamuse põhjal ja maine põhjal, mida portaalid on suutnud autori jaoks Eestis kujundada. Vaatlusaluseid portaale on kokku seitse, aga ühisrahastust kui investeerimisvõimalust pakkuvaid portaale on veelgi rohkem ja kuna tegemist on ka piirideülese äriga, ei oleks käesolevas töös praktiline loetleda kõiki võimalusi, kuhu investor ühisrahastuse raames oma raha paigutada saab.

Tabel 2. Valitud ühisrahastusportaalide lühiülevaade.

Portaal	**Lühitutvustus**
Mintos	Portaalil on 47763 investorit, nende poolt pakutud investeeringute keskmine tootlus on olnud 11,94% ja kokku on väljastatud laene 492 mln euro väärtuses.
Twino	Portaal on mõnevõrra eriline, kuna pakub suhteliselt kõrge intressiga tagatud laene. Kokku on portaal väljastanud laene üle 500 mln euro väärtuses ja keskmine tootlus investoritele on olnud üle 11% aastas.
Omaraha	Portaalil on 133144 investorit, kokku on väljastatud laene 41 mln euro väärtuses, pakutakse nii kõrge intressimääraga tagamata laene kui ka madala intressimääraga tagatud laene. Keskmine tootlus investoritele on olnud 21% aastas.
Estateguru	Portaal pakub võimalust investeerida mahukatesse kinnisvaraprojektidesse. Laenud on tagatud hüpoteegiga, kusjuures laenu suhe tagatisvarasse on olnud 58%. Kokku on laene väljastatud 41 mln euro väärtuses ja keskmine tootlus on olnud 11,09%. Portaalil on 10115 investorit.
Crowdestate	Erinevalt Estategurust pakub see kinnisvaralaenudega tegelev portaal laene ilma hüpoteegita. Sellest tulenevalt on ka intressimäärad kõrgemad ja investoritele on suudetud pakkuda 25,5% tootlust. Portaalil on 17350 investorit.
Bondora	Suurima Eesti portaalina on Bondora väljastanud laene 116 mln euro väärtuses ja ajalooliselt investoritele pakkunud väga laia võimalike tootluste spektrit, kusjuures keskmiseks tootluseks on olnud 12,6%. Portaal pakub investoritele väga detailset andmestikku paremate investeerimisotsuste tegemiseks ja analüüsi läbiviimiseks.
Investly	Portaal pakub ettevõtetele faktooringut, hankides vajaminev raha ühisrahastusest. Faktooringuga vahendatud arvete summaks on 19 mln eurot ja investoritele on suudetud pakkuda 11-13% aastatootlust.

Allikas: Autori koostatud ühisrahastusportaalide kodulehtedelt saadaval oleva info põhjal.

Tabelist 2 on näha, et ühisrahastusportaalide pakutav tootlus investori jaoks on küllaltki sarnane, kuid siin on mõned erandid. On näha, et peamiseks tootluse suurust määravaks teguriks on investeeringu tagatise olemasolu. Kui investeering on tagatud mingisuguse alusvaraga (auto, kinnisvara), siis on investori risk rahast ilma jääda oluliselt väiksem ja seetõttu pole suur ka riskipreemia. Kõrgeima riskiga on tabeli 2 loetelus toodud investeeringutest autori arvamuse kohaselt Omaraha ja Bondora tagamata tarbimislaenud. Bondora ajalugu on näidanud, et kergekäeliselt laenatakse välja investorite raha Hispaaniasse ja Sloveeniasse, mis on riikide lõikes ka seni näidanud kõrgeimat pankrotistumise ja madalaimat taastumise taset. Kõige vähem „mängimisruumi“ pakuvad kinnisvarainvesteeringute platvormid Estateguru ja Crowdestate, kus investori eeltöö enne investeerimist piirdub tavaliselt laenu võtva

ettevõtte majandusaasta aruande lugemisega. Kõige rohkem on aga investoril võimalik oma analüütilisi oskuseid rakendada seal, kus on olemas selleks vajalikud andmed.

Käesoleva magistritöö fookuses oleva juhtumiuuringu läbiviimise tarbeks on autor valinud just Bondora, kuna nende detailne ja pika ajalooga andmestik võimaldab teha kõige täpsemat analüüsi ja ühtlasi on autoril kogemus Bondora kaudu investeerimisel. Autor näeb ka võimalusena Bondora poolt pakutavat väga laia võimalike tootluste spektrit ja hästi toimivat järelturgu. Järelturu olemasolu on Bondoras analüütilisi meetodeid kasutades äärmiselt oluline, sest võimaldab arvesse võtta ka laenu võtnud inimese maksekäitumist alates laenu võtmise hetkest kuni investeeringu müükipanekuni.

Järgmises alapeatükis kirjeldab autor võimalikke statistilisi ennustusmeetodeid, mida võiks Bondora poolt väljastatava andmestiku peal kasutada, et ette ennustada laenude pankrotist taastumist ja seeläbi investorina võita kordades suuremat tootlust, kui tabelis 2 välja toodud ühisrahastusportaalide keskmised.

1.2. Levinumad meetodid investeeringu tulemuse ennustamisel

Kasutades statistilisi ennustusmeetodeid on võimalik andmetes leiduvad trende ja mustreid iseloomustada ja selle põhjal teha ennustusi tuleviku suhtes. Tööriistakast, mida ennustamisanalüütika valdkond pakub, sisaldab suurel hulgal erinevaid meetodeid. Piltlikult öeldes annab see analüütikule võimaluse panna vastavusse andmed ja mingisugune oluline muutuja, mida on vaja ette ennustada, et sel viisil tehtavaid ennustusi ära kasutada tuleviku etteennustamiseks. Mõnikord nimetatakse ka eelkirjeldatud protsessi masinõppeks.

Masinõppe peamiseks eesmärgiks on ehitada andmestikest häid mudeleid. Andmestikud koosnevad enamasti vaatlustest ja muutujatest. Mudeli all mõeldakse enamasti ennustavat mudelit, mille eesmärgiks on mingi muutuja tulemust teisi muutujaid arvestades ette ennustada. Mudeli ehitamise protsessi nimetatakse nii treenimiseks kui ka õppimiseks, sest tegemist on mudeli headuse kasvatamisega õppimisalgoritmi rakendamise läbi. Mudeleid nimetatakse mõnikord ka õppijateks ja kui neil on ennustamisvõime, siis nimetatakse neid ka ennustajateks. Kui ennustatavat muutujat

võib jagada kategooriatesse, ehk tegemist on kategoorilise muutujaga, siis saab mudeli ennustustööd nimetada klassifitseerimiseks. (Zhou 2012: 2)

Ennustusmeetodite arv on ajas pidevalt kasvanud ja seetõttu on analüütiku käsutuses palju erinevaid meetodeid. Isegi pärast otsuseid, mis puudutavad andmete korrastamist ja analüüsistrateegiat, võib ikkagi alles jääda rohkem kui üks sobilik meetod vaatlusaluse probleemi lahendamiseks. Seetõttu on mõistlik proovida oma andmetel kasutada erinevaid meetodeid, et aru saada nende erinevustest ja välja selgitada, milline meetod sobib kõige paremini. On ebatõenäoline, et üks meetod tõuseb teistest mäekõrguselt paremaks, aga mõningad erinevused meetodite vahel on siiski oodatud. (McCue et al 2007: 126)

Andmaks ülevaade käesoleva magistritöö fookuses oleva juhtumiuuringu raames kasutatavatest ennustusmeetoditest, on autor meetodid koos lühikirjeldusega koondanud tabelisse 3. Meetodeid on kokku 6 ja nende kasutamist ennustamisel on kirjeldatud peatükis 2.2.

Tabel 3. Lühiülevaade töös kasutatavatest ennustamismeetoditest.

Ennustamismeetod	**Lühikirjeldus**
Närvivõrk	Sobilik meetod siis, kui sisendandmed ja väljundmuutuja on lihtsasti tõlgendatavad, aga nendevahelise seose protsess on keerukas. Võimalik suur täpsus.
Tugivektormasin	Meetod ei nõua kuigi suurt arvutusvõimsust, suudab toime tulla ka kategooriliste muutujatega, võib teoreetiliselt pakkuda väga head täpsust, kuid tulemused on raskesti tõlgendatavad.
Otsustuspuu	Võimalik luua lihtsasti arusaadav joonis, sarnaselt tugivektormasinale ei nõua kuigi suurt arvutusvõimsust, aga ei paku teiste meetoditega võrreldes kuigi head täpsust ja omab ohtu treeningandmestikule ülesobituda.
Otsustuspuude mets	Robustne sisendandmete muutumise ja müra suhtes. Nõuab suurt arvutusvõimsust, aga omab paindlikkust suure muutujate arvuga toimetulemusel. Ei ülesobita.
Võimendamine	Suudab suurte andmemahtude juures pakkuda väga head täpsust. Õpib vigadest.
Logistiline regressioon	Mainitud meetoditest vanim ja lihtsaim. Võimaldab uurida marginaalseid efekte. Laialdaselt kasutuses panganduses krediidiskooringu teostamisel.

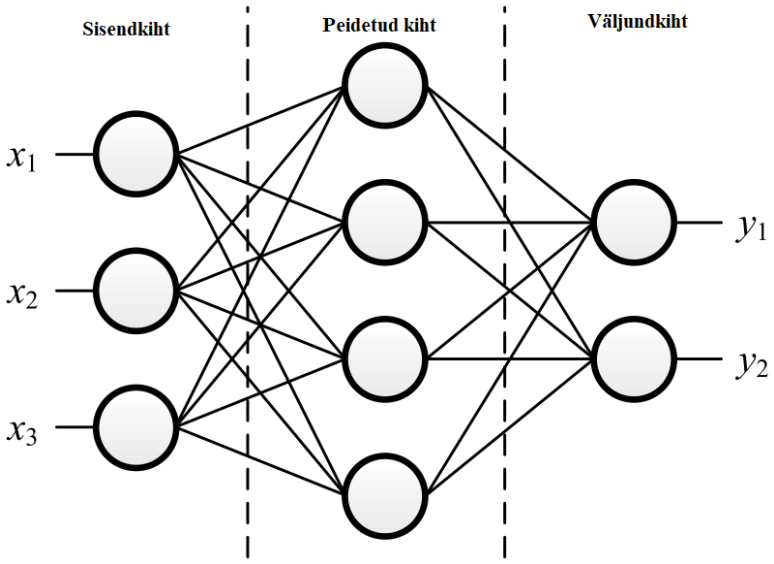
Allikas: autori koostatud peatükis 1.2 kasutatud allikate põhjal.

Üldiselt saab investeeringu tulemuse etteennustamist pidada klassifitseerimiseks, sest ennustatakse mingi olulise tunnuse väärtust või teisisõnu seda, millisesse kategooriasse see suurima tõenäosusega langeb. Kindlasti saab klassifitseerimiseks pidada ka käesoleva töö fookuses tehtavat tegevust, milleks on pankrotti jäävate ja pankrotist taastuvate laenude eristamine. Klassifitseerimine on ka investeeringu tulemuse ennustamise jaoks äärmiselt tähtis, kuna parem informatsioon võimaldab teha paremaid investeerimisotsuseid. Järgnevalt annab autor ülevaate valitud meetoditest, mida käesolevas magistritöös klassifitseerimiseks kasutatakse.

1.2.1. Närvivõrk

Närvivõrgu (neural network) tööpõhimõtteks on luua seos sisendiks olevate andmete ja tulemuste vahele kombineerides lihtsaid mittelineaarseid mooduleid, mis igaüks transformeerivad klasside esindatust ühel tasemel, jätkates klasside esindatuse transformeerimist kõrgematel ja abstraktsematel tasemetel (LeCun et al 2015). Üheks aspektiks, miks autori arvates närvivõrk võib käesoleva töö raames sobivaks meetodiks olla, on see, et närvivõrk toimib hästi siis, kui sisendiks olevaid andmeid ja nendest saadavat tulemust on lihtne mõista, aga kogu protsess, mis tulemuseni viib, on keeruline.

Närvivõrkude edukaim ja autori hinnangul käesoleva magistritöö spetsiifikat arvesse võttes sobivaim algoritm, BP (Back-Propagation), toimib nii, et kõigepealt söödetakse sisendmuutujad sisendite kihist peidetud kihi kaudu väljundkihini, kus arvutatakse veamäär, võrreldes võrgu väljundit tegeliku tõega. Seetõttu töötavad närvivõrgud klassifitseerimismeetoditena väga hästi (Elizondo 2006: 330). Seejärel saadetakse tehtud vead tagasi peidetud kihti ja sisendite kihti, mille tulemusena seadistatakse neuronitevaheliste ühenduste kaalusid, et veamäära vähendada. Kihid on kujutatud alljärgneval joonisel 1. Seda protsessi korratakse väga palju kordi, kuni veamäär on minimeeritud. (Zhou 2012: 8)



Joonis 1. Närvivõrgu kihid Back-Propagation algoritmi korral. Allikas: Zhou (2012: 8)

Samas ei saa muidugi mainimata jätta, et tehislikud närvivõrgud on looduslike närvivõrkude väga ligikaudsed lihtsustused. Ajus on neuronite arv väga palju suurem kui tehisnärvivõrkudes (u 100 miljardit vs mõni tuhat). Tehisneuronid on äärmiselt lihtsad. Neil on vähem sisendeid ja informatsiooni töötlus toimib rohkematel tasanditel. Sellele vaatamata on siiski idee ja struktuur tehislike närvivõrkude tööks võetud just bioloogilistelt närvivõrkudelt. (Lints 2004)

Järgnevalt kirjeldab autor veel ühte äärmiselt huvitava tööpõhimõttega ennustusmeetodit, mida krediidiandmete peal edukalt rakendatakse.

1.2.2. Tugivektormasin

Tugivektormasin (ingl. SVM – support vector machine) on väga tõhus klassifitseerimismeetod. Selle toimimispõhimõte on järgnev: lineaarselt mitteeralduvad klassid kujutatakse kõrgema dimensiooniga ruumi ja lineaarne eraldamine teostatakse seal (Cortes et al 1995: 273). SVMide parimaks küljeks on see, et püütakse leida optimaalseim klasse eraldav hüpertasand ja seetõttu on SVMi ennustusvõime küllaltki kõrge. SVMi töö käib kahes etapis (Tint 2003: 1):

1. Mittelineaarne sisendvektori kujutamine kõrgdimensionaalsesse varjatud ruumi.
2. Varjatud ruumis optimaalse tasandi leidmine klasside eraldamiseks.

Tugivektormasinate kaks tugevaimat omadust on suvalise sobiva optimaalse tasandi leidmise asemel optimaalseima leidmine, mis võib tähendada omakorda kõrget ennustusvõimet või ennustuste täpsust, ja võimalus kõrgdimensionaalsesse ruumi üleminekul hoiduda arvutusraskuse liiga järsust kasvust, mis kiirendab oluliselt analüüsi läbiviimise protsessi (Tint 2003:12). Arvutusraskuse järsk kasv on käesoleva magistritöö fookuses oleva juhtumiuuringu kontekstis üsnagi oluline tegur, sest krediidiriski puudutavad andmed, nagu ka näiteks Bondora poolt väljastatav „LoanData“ andmekogum on tavaliselt küllaltki mahukad, sisaldades kümneid tuhandeid vaatlusi ja sadu muutujaid. Samas on tugivektormasinal ka küllaltki suur nõrkus. Kui negatiivsete instantside arv on positiivsetest väga palju suurem, siis selle täpsus kannatab märkimisväärselt ja Akbani et al (2004) hinnangul ei ole tugivektormasin krediidiandmetele rakendamiseks kuigi sobiv juhul, kui negatiivsete vaatluste osakaal on liialt väike. Autori hinnangul võib tugivektormasin osutuda ka käesoleva töö jaoks sobimatuks, sest põhimõtteliselt on tegemist krediidiandmetega, kuid kui negatiivsete vaatluste osakaal ei ole liiga väike, võib tugivektormasin siiski heaks ennustusmeetodiks osutuda. Huang et al (2007) peab tugivektormasinat siiski krediidiandmetele sobivaks ennustusmeetodiks.

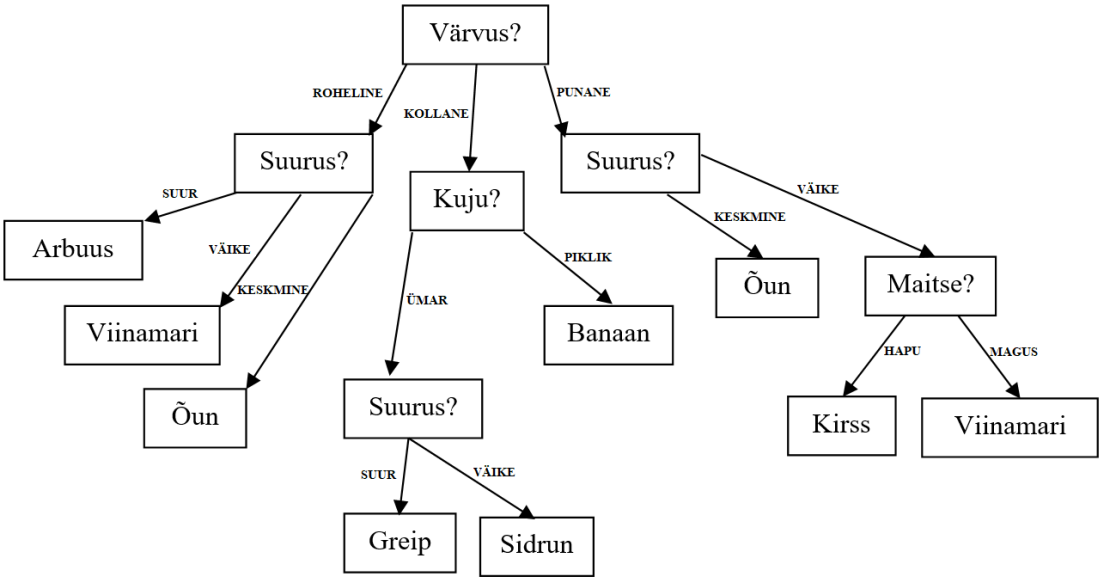
Käärmanni (2003) järgi võib väita, et klassifitseerimisprotsessis üldiselt treenitakse välja teatud adapteeruv algoritm (nt. neurovõrk, otsustuspuu, Bayes-i võrk, SVM) teadaolevate objektidega, milleks käesoleva magistritöö fookuses oleva juhtumiuuringu kontekstis on vaatlused pankrotistunud, kuid seejärel taastunud laenude kohta ning rakendatakse sedasama väljaõpetatud algoritmi hiljem tundmatute objektide klassifitseerimiseks. Tundmatuteks objektideks võib siinjuures pidada siis näiteks pankrotis olevaid laene, mille puhul pole veel kindel, kas need taastuvad või jäävadki pankrotistunuks. Hsu et al (2016) pakub välja praktilise juhendi tugivektormasina ülesseadmiseks ja käesoleva töö autor võtab seda töö koostamisel arvesse.

1.2.3. Otsustuspuu

Enamasti kasutatakse klassifitseerimismeetodeid reaalmaailma nähtuste äratundmiseks. Laialdast kasutust on klassifitseerimismeetodid leidnud just krediidiriski hindamisel, aga nendel on ka hulgaliselt muid kasutusalasid. Näiteks võib kardiogrammi andmetel ära tunda südamehaige, pildilt tuvastada kujutist või serveri logi põhjal avastada

pahatahtliku ründe. Üheks huvitavaks meetodiks, mis võib osutuda kasulikuks pankrotist taastuvate laenude äratundmisel on otsustuspuu. Hierarhilise otsustuspuu rekursiivne konstrueerimine on end tõestanud mitmete reaalelu probleemide lahendamisel.

Üheks otsustuspuude populaarsuse põhjuseks ning eeliseks teiste klassifitseerimismeetodite ees on nende intuitiivne arusaadavus (Brownlee 2013), sest otsustuspuud on võimalik visuaalselt kujutada kõigile väga lihtsasti arusaadavalt (vt. joonis 2) ning kasutuslihtsus. Sellest hoolimata ei suuda iga puu igas kontekstis mõistlikku tulemust anda ning ka otsustuspuuga klassifitseerimine nõuab piisavaid teadmisi ülesande põhiolemusest ja otsustuspuude kombineerimise võimalikkusest. (Käärmann 2003: 16)



Joonis 2. Lihtne näide otsustuspuust puuvilja äratundmise kohta. Allikas: (Käärmann 2003: 18)

Joonisel 2 kujutatud otsustuspuu on lihtne illustratsioon selle kohta, kuidas on võimalik otsustuspuu meetodil saadud analüüsi tulemusi graafiliselt kujutada. Kui proovida mõttes iga joonisel olev küsimus asendada mõne krediidiriski või maksekäitumist kirjeldava muutujaga, hakkab kujunema pilt selle kohta, milline võiks otsustuspuu meetodi väärtus olla käesoleva magistritöö fookuses oleva juhtumiuuringu jaoks. Käärmann (2003) kirjeldab otsustuspuud kui mängu, mida lapsed mõnikord igavuse peletamiseks mängivad. Tema sõnul võib ette kujutada tundmatu objekti

(andmevektori) äratundmist, küsides järjepanu küsimusi objekti tunnuste kohta ning olles kogunud piisavalt infot, otsustada millise objektiga on tegemist (objekti klassikuuluvus).

1.2.4. Otsustuspuude mets

Ühe otsustuspuu ehitamine annab meile ennustatavast nähtusest lihtsustatud ülevaate, kuid tihtipeale võib see vaade osutuda liiga lihtsaks või liiga spetsiifiliseks. On selge, et mitu ennustavat mudelit koos annavad kokku üksikust mudelist parema tulemuse (Williams 2011: 245). Seetõttu on mõistlik koondada otsustuspuud kokku otsustuspuude metsaks, et meil oleks mudelite ansambel. Seda tegevust võib võrrelda näiteks tõsiasjaga, et tihtipeale koondatakse eksperdid kokku paneelidesse, et siis konsensusliku otsuseni jõuda. Sarnaselt toimivad ka valitsused, ettevõtete juhatused ja ülikoolid üle kogu maailma. Tihtipeale annab just mitmete otsustajate kokkukoondamine parema tulemuse kui lootmine üksiku eksperdi arvamusele.

Otsustuspuude metsa täpsust omistatakse tõsiasjale, et see vähendab üksikutele otsustuspuudele omast ebastabiilsust, mida saab ilmestada näiteks siis, kui andmestikust ära võtta kasvõi väike arv vaatlusi ning seejärel vaadata otsustuspuud, mis selle väike muudatuse korral ka ise üsna suuri muutusi läbi teeb. Ansamblina on otsustuspuude mets andmemuudatuste vastu palju robustsem ja sellest tulenevalt ka üsna vähe mõjutatud „mürast“ või teisisõnu, muutujatest, millel on ennustatava muutujaga nõrk seos (Breiman 2001: 5, Williams 2001: 246). Erinevalt eelmainitud tugivektormasinatest saavad otsustuspuude metsad väga hästi hakkama andmestikega, kus ennustatava muutuja huvialune klass on esindatud väga vähe – näiteks 5% vaatlustest või vähemgi. (Williams 2011: 246)

Suurte arvude seaduse tõttu ei teki otsustuspuude metsa kasutades mudeli ülesobitamist ja seetõttu peetakse seda efektiivseks ennustusmeetodiks. Nende sisseehitatud juhuslikkus üksikute puude ehitamisel tehtavate vaatlusvalikute asjus teeb neist täpse meetodi klassifitseerimiseks (Breiman 2001: 29) ja seega on see kahtlemata sobiv meetod käesoleva töö fookuses oleva juhtumiuuringu läbiviimiseks.

1.2.5. Võimendamine

Veel üheks tuntud ansambelmeetodiks klassifitseerimise valdkonnas on võimendamine (boosting) mida peetakse efektiivseks (Brownlee 2013) ja samas ka lihtsasti kasutatavaks ennustamismeetodiks. Selle tööpõhimõte seisneb mitmete mudelite ehitamises, mis iseenesest ei pea olema kuigi head õppijad (Zhou 2012: 23, Bastos 2008: 4, Witten 2011: 321). Võimendamise käigus annab algoritm suuremad kaalud nendele vaatlustele, mida on keeruline õigesti ennustada. Iga uue mudeli loomise järel muudetakse kaalusid, et raskesti ennustatavatele vaatlustele veelgi kaalu juurde lisada. Võimendamine vajab aga õigesti toimimiseks väga suurt andmemahtu ja on ka küllaltki vastuvõtlik andmetes leiduva „müra“ suhtes. Kokkuvõttes võib siiski öelda, et võimendamine on otsustuspuude metsale üsnagi sarnane, sest mõlema meetodi käigus kombineeritakse kokku meetodite „ansambel“, mis annab parema tulemuse kui üksik mudel. (Williams 2011: 269)

Sarnaselt otsustuspuude metsale võib ka võimendamise aluseks olev „nõrk õppija“ olla otsustuspuu, kuid siiski säilib üks väga oluline erinevus. Kui otsustuspuude mets luuakse muutujatest suvalisi alamvalikuid tehes, siis võimendamise käigus luuakse mudelid üks teise järel tuginedes sellele, mida eelmine mudel teha ei suutnud. Põhimõtteliselt muudetakse valesti klassifitseeritud vaatlused prominentsemaks, et need andmestikus paremini esile tuleks ja mudeli kujunemisel suuremat kaalu omaks.

Võimendamisalgoritm ei ole väga sobilik andmestike jaoks, kus on palju müra, sest see loodi algselt kasutatama nii-öelda puhastel andmestikel. Suuresti saab müratundlikkuse panna võimendamisalgoritmi suurima tugevuse süüks, sest mudel annab valesti ennustatud vaatlustele (seal hulgas ka lihtsalt mürale) väga suure kaalu uute mudelite loomisel ja ennustuste kujunemisel. See võib viia võimendamisalgoritmi ennustusvõime languseni. (Zhou 2012: 41)

Järgnevalt tutvustab autor ka logistilist regressiooni.

1.2.6. Logistiline regressioon

Regressioonmudelid on ennast tõestanud andmeanalüüsi lahutamatu osana, kui on olnud vajadus selgitada mingisuguse sõltuva muutuja suhet ühe või enama sõltumatu

muutujaga. Käesoleva magistritöö kontekstis tasub keskenduda logistilistele regressioonmudelitele, sest need võimaldavad erinevalt lineaarsest regressioonist klassifitseerimist binaarse tunnuse alusel. Kui seda erisust arvesse võtta, siis logistilise regressiooni läbiviimine on väga sarnane vähimruutude meetodil saadud lineaarse regressiooni läbiviimisega ja on seetõttu lihtne ja küllaltki efektiivne klassifitseerimismeetod. (Hosmer *et al* 2013: 1-2)

Logistilise regressiooni sobitamiseks olgu meil mingi arv vaatlusi, millel on muutujate paar (X_i, Y_i), i =1,2,…,n, kus Y_i väljendab i-nda vaatluse binaarset muutujat, mida me ennustada püüame ja X_i väljendab i-nda vaatluse sõltumatut muutujat. Lisaks sellele eeldame, et ennustatav muutuja on kodeeritud nullide ja ühtedena, esindades näiteks mingi nähtuse toimumist või mittetoimumist. Sobitamine tähendab sellisel juhul leitava mudeli parameetrite β_1 ja β_2 leidmist. Lineaarses regressioonis leitakse parameetrid vähimruutude meetodil, mille põhimõtteks on ennustavate väärtuste ja tegelike väärtuste vahelise ruutvigade summa minimeerine. Logistilises regressioonis on sobitamiseks kasutusel aga suurima tõepära meetod, mis annab parameetritele väärtused, mis maksimeerivad vaadeldud andmete saamise tõenäosust. Selleks loodav tõepärafunktsioon ilmestab vaadeldavate andmete saamist tõenäosust tundmatute parameetrite funktsioonina. (Hosmer *et al* 2013: 6-7)

Peatükis 1.2 toodud meetodite lühikirjeldused annavad küll meetoditest hea ülevaate, kuid vaja on ka teada, kuidas meetodite lühikirjelduse all mainitavat „täpsust“ kui mudeli kvaliteedi mõõdikut mõõdetakse ja ilmestatakse. Selleks toob autor järgnevalt välja valitud mõõdikud ja võimalused, mida käesoleva magistritöö empiirilises osas kasutatakse.

1.2.7. Mudeli ennustusvõime kirjeldamine

Andmeteaduses öeldakse üldiselt, et kui mudeli ennustusjõud on liiga hea, et tõsi olla, siis see tõenäoliselt ongi. Mudel võib liiga heaks osutuda näiteks siis, kui kaasata mudelis mõni muutuja, mis on ennustatava muutujaga otseselt seotud. Käesolevas töös võiks pankrotist taastumise ennustamise korral selliseks muutujaks olla näiteks maksete kogusumma pärast taastumist, sest kõik vaatlused, kus on toimunud makseid pärast

taastumist, ongi pankrotist taastunud laenud ja mudel saaks vaid ühe muutuja põhjal täiuslikult ette ennustada kõik tulevased vaatlused.

Mudeli ennustusvõime hindamine annab meile ettekujutuse sellest, et mida oodata, kui me rakendame mudelit uutele andmetele. Näiteks 80% täpsusega mudel suudaks uutest vaatlustest korrektselt klassifitseerida 80 vaatlust sajast. Mudeli ennustusvõime parandamine võib krediidiandmete korral olla erakordselt tähtis, sest kui eesmärgiks võtta pankrotistuvate laenude vältimine, siis väikseimgi täpsuseparandus võib kaasa tuua rahalise võidu selle tõttu, et pankrotistuvaid laene antakse mudeli parema täpsuse tõttu vähem välja. Toetudes aga arutluses rohkem käesolevas töös läbiviidavale juhtumiuuringu spetsiifikale, on iga vaatlus, mida mudel ennustab taastuvaks, kuid tegelikkuses ei taastu, väga suure kaaluga investeeringu kogutootluse vähendamise suunas, sest iga selline investeering läheb 100-protsendiliselt kirja kuluna.

Williamsi (2011) järgi peaks mudelite ennustusvõime kirjeldamise protsess nägema välja selline, et kõigepealt luuakse mudel kasutades treeningandmestikku. Mudeli parameetrid pannakse paika kasutades valideerimisandmestikku ja kõik muudatused katsetataksegi kasutades vaid valideerimisandmestikku. See tähendab, et testandmestik on kuni lõpuni puutumata. Testandmestikku kasutatakse vaid lõplike hinnangute andmiseks mudelite töövõimele.

Ristvalideerimine kui mudeli töövõime hindamise lähenemisviis on oma tööpõhimõttelt küllaltki lihtne. Võtame näiteks mingi andmestiku ja jagame selle kümneks suvaliseks, või veelgi parem, ennustava klassi järgi stratifitseeritud osaks. Seejärel ehitame mudeli, kasutades nendest kümnest suvalisest osast üheksat, mis kokku moodustavad meie treeningandmestiku üheksa osa. Nüüd on võimalik saadud treeningandmestiku põhjal loodud mudelit testida kümnenda andmestiku peal. Protsessi on võimalik korrata kümme korda nii, et iga kord jääb testandmestikuks erinev osa koguandmestikust, et anda meile mõõde mudeli arvatavast ennustusvõimest. (Williams 2011)

Veel enne, kui saame rääkida mudeli ennustusvõime mõõdikutest, tuleb rääkida, sellest, mida mõistetakse õigepositiivsete, õigenegatiivsete, valenegatiivsete ja valepositiivsete ennustuste all. Võttes näidiseks käesoleva magistritöö käigus läbiviidava juhtumiuuringu oodatavaid tulemusi, siis võib neid mõisteid selgitada nii, et kui mudel

ennustab pankrotist taastumist ja tegelikkuses laen taastubki, siis on tegu õigepositiivse ennustusega, teistpidi, kui mudel ennustab, et laen ei taastu ning see ei taastugi, siis on tegu õigenegatiivse ennustusega. Riskikartliku investori jaoks ehk olulisemgi on aga valenegatiivsete ja valepositiivsete ennustuste esinemine. Kui mudel ennustab, et laen pankrotist ei taastu, aga tegelikkuses taastub, siis seda nimetatakse valenegatiivseks ennustuseks. Investori seisukohalt on tegemist lihtsalt tegemata investeeringuga ja ei tähenda otsest kahju, kui olemas on mõni teine samaväärne investeerimisvõimalus. Olulisimateks ennustusteks on aga siinjuures valepositiivsed ennustused, kus mudel ennustab pankrotist taastumist, aga tegelikkuses taastumist ei toimu. See on väga kahjulik investori saadavale tootlusele, sest mittetaastuv pankrot on põhimõtteliselt maha visatud raha, kui seda investeeringut ei õnnestu kellelegi teisele edasi müüa.

Kuna käesolevas töö on ennustatava muutuja näol tegemist binaarse tunnusega, siis on lihtne ja otstarbekas mudelite ennustusvõimet ilmestada ka maatriksis, mille näide koos õigepositiivsete, õigenegatiivsete, valepositiivsete ja valenegatiivsete ennustuste paiknemisega maatriksis on toodud alljärgnevas tabelis 4.

Tabel 4. Ennustustulemuste paigutumine maatriksisse.

		Ennustatud	
		Ei taastu pankrotist	**Taastub pankrotist**
Tegelik	**Ei taastu pankrotist**	Õigenegatiivsed	Valepositiivsed
	Taastub pankrotist	Valenegatiivsed	Õigepositiivsed

Allikas: autori koostatud Williams (2011: 312) põhjal.

Üldiselt tähendab mudeli ennustusvõime hindamine seda, et rakendatakse mudelit mingisugusele andmestikule, kus tegelik tulemus on teada ja seejärel võrreldakse, kas mudel ennustas õigesti või mitte. Kõige lihtsamaks mõõdikuks seejuures ongi veamäär (error rate), mis on arvutatud kui valesti ennustatud vaatluste arvu suhe vaatluste tegelikku arvu. (Williams 2011: 312)

Järgnevalt annab autor ülevaate kolmest olulisemast mõõdikust, mida mudelite võrdlemisel ja headuse hindamisel kasutada võib (Williams 2011: 312):

1. **Täpsus** (precision) on õigepositiivsete ennustuste arvu suhe õigete ennustuste koguarvu. Täpsus näitab seda, kuivõrd õigesti mudeli suudab ennustada, või teisisõnu, kui täpne mudel ennustamises on.

2. **Tundlikkus** (sensitivity või ka recall) on õigepositiivsete ennustuste määr, mis näitab, kui suure osa positiivsetest ennustustest suudab mudel ära teha.
3. **Spetsiifilisus** (specificity) on õigenegatiivsete ennustuste määr, mis näitab, kui suure osa negatiivsetest ennustustest suudab mudel ära teha.

Lisaks lihtsatele suhtarvudele nagu eelpool kirjeldatud, saab mudeli headuse hindamisele läheneda ka graafiliselt. Üheks heaks näiteks on siinjuures riskigraafikud (risk charts).

Riskigraafik võrdleb huupi arvavat mudelit vaatlusaluse mudeliga, pannes vastavusse protsendi vaatlustest, mida kokku on hinnatud (x-teljel) ja õigesti hinnatud vaatluste protsendi (y-teljel). See on koostatud nii, et vaatlused on järjestatud mudeli poolt hinnatud esinemistõenäosuse järgi ja esmalt kontrollitaksegi just kõrgema tõenäosusega vaatlusi. Niimoodi tekib vahe huupi pakkuva triviaalse mudeli ja vaatlusaluse mudeli vahel, sest vaatlusalune mudel ei hinda vaatlusi huupi, vaid teeb otsuseid andmetest leitud seoste põhjal. (Williams 2011: 317)

Heaks võrdluskohaks riskigraafikutel triviaalse mudeli ja vaatlusaluse mudeli vahel on näiteks koht, kus hinnatud on pooled vaatlused (Caseload = 50%). Triviaalne mudel on selleks ajaks leidnud pooled etteennustavad tulemused, kuid ennustusjõudu omava mudeli puhul võib sel juhul vaadata riskikõverat, mille y-telje väärtus 50% vaatluste läbivaatamise järel näitab, kui suur osa ennustatavatest väärtustest on selleks ajaks leitud. Mida suurem on erinevus triviaalse mudeli sirge ja saadud riskikõvera vahel, või teisisõnu, mida lähemal asub riskikõver graafiku ülemisele paremale nurgale, seda parem on vaatlusalune klassifitseerimismudel. (Williams 2011: 317)

Veel saab mudeli headust graafiliselt kujutada ROC-graafikute, tundlikkuse/spetsiifilisuse graafikute, lift-graafikute ja täpsuse/tundlikkuse graafikutega. Kõige enam kasutatakse tõenäoliselt ROC-graafikuid (Engelmann et al 2003: 8), kuid olenevalt uurimisprobleemi eripärast võib mõne klassi etteennustamine olla teistest oluliselt tähtsam ja seetõttu võib parima graafilise ettekujutuse mudeli ennustusvõimest anda mõni teine graafiline meetod. ROC-graafikud panevad vastavusse õigepositiivsete ennustuste määra valepositiivsete ennustuste määraga (Hand 2010: 1502). ROC graafikutelt võib leida näitaja AUC (area under the curve), mis on Huang et al (2003: 2)

sõnul täpsust väljendavast suhtarvust parem mudeli kvaliteedi näitaja. Lift-graafikud kujutavad y-teljel mudeli õigesti tehtavate ennustuste võimendusmäära triviaalse mudeliga võrreldes, pannes see vastavusse õigete ennustuste määraga x-teljel. Lift-kordaja näitab mitu korda rohkem huvialuseid vaatluseid mudel võrreldes triviaalse mudeliga ära tunneb. Täpsuse/tundlikkuse graafikud panevad vastavusse õigepositiivsete ennustuse suhte õigetesse ennustustesse ja õigepositiivsete ennustuste määra. Esmapilgul tunduvad erinevused graafikute x- ja y-teljel esitatavates suhtarvudes küllaltki väikesed, kuid need omavad suurt mõju nii graafiku kujule kui ka selle kujutatavale tulemusele. (Williams 2011: 317)

Järgnevalt toob autor välja eelnevate autorite tööd Bondora andmestikku kasutaval temaatikal, krediidiriski arvutusi puudutaval temaatikal ja ka investeeringu tulemuse ennustamisel üldisemalt.

1.3. Eelnevate autorite tööd investeeringu tulemuse ennustamisel

Esimeseks ühisrahastusplatvormiks, kes oma andmestiku avaldas ja sellega teaduslike panuste tulva vallandas oli 2007. aastal Prosper (Bachmann 2011: 3). Bondora on oma väljastatava andmestiku poolest ühisrahastusmaastikul pigem erandiks, sest paljud oma andmeid ei avalda. Seetõttu on sealt kättesaadavat andmestikku kasutatud erinevate andmetöötlus- ja masinõppemeetodite proovilepanekuks. Käesoleva töö autor seega ei ole esimene, kes Bondora poolt väljastatavat andmestikku masinõppe kontekstis kasutab. Näiteks Haltuf (2014) uuris küllaltki põhjalikult tugivektormasinate kasutusvõimalusi krediidiskooringu kontekstis. Tema sõnul on kvantitatiivsetel masinõppemeetoditel kindel koht krediidiriski hindamisel ja laenuäri läbipaistvamaks muutmisel. Ta tõi ka välja, et väikseimgi parandus krediidiriski hindamisel toob kaasa märgatava rahalise võidu. Haltuf tõdes, et pangad ei kipu logistilist regressiooni krediidiskooringu meetodina veel välja vahetama, kuid uurimine selles suunas käib. Samuti on Haltufi tulemustest selge, et vähemasti Bondora andmestikul laenu pankrotti minemise tõenäosuse hindamisel oli logistiline regressioon nõrgem ennustaja kui tugivektormasin. Samas tunnistab aga Haltuf tugivektormasinate puudujääke müratundlikkuse ja treenimisaja suhtes.

Kui Haltuf (2014) uuris tugivektormasinaid, siis Kisand (2015) tugines hoopis otsustuspuude metsadele, püüdes elu viia küllaltki ambitsioonikat eesmärki, milleks oli optimaalse portfelli loomine. Kisand võrdles otsustuspuude meetodil tehtavaid otsuseid krediidireitingu järgi investeerimisega ja tõdes, et otsustuspuude meetodil tehtavad otsused on paremad. Tema loodud otsustuspuude mets suutis välja pakkuda 94% kõikidest headest investeeringutest ilma eksimata. Käesoleva töö autori arvates tuleks siin ära märkida, et Kisand rakendas oma metoodikas kulude maatriksit tegemaks valepositiivsed ennustused väga kulukaks – see viis positiivsete ennustuste täpsuse küll üles, aga üldine täpsus läks selle tõttu alla. Samuti tuleks märkida, et käesolev magistritöö on otsekui juurdeehitus Kisandi tööle, sest lisab optimeerimismetoodikale veel ühe mõõtme – pankrottidesse investeerimise. Kisand tõdes, et otsustuspuude mets omab üksikust otsustuspuust suuremat täpsust. Ajaloolise täpsuse huvides oleks ehk ka huvitav märkida, et Kisand ennustas Bondora platvormile investorisõbralikuma kasutajaliidese teket, mis võimaldaks mudeli poolt tähtsaks peetud parameetrite järgi laene otse veebilehel filtreerida, kuid käesoleva töö kirjutamise hetkel (3 aastat hiljem) pole seda veel juhtunud.

Kangas (2014) lähenes Bondora andmestikust tulenevale problemaatikale teistest küllaltki erineval viisil. Ta uuris, kuidas on Bondoras laenavaid inimesi võimalik klastritesse jaotada nii, et riskantsemad inimesed oleks teistest eraldi. Tema eesmärgiks oli leida mingeid muutujakombinatsioone, mis aitavad ennustada isiku pankrotti minemise tõenäosust. Kangase kasutatavateks meetoditeks olid põhikomponentide analüüs, iseorganiseeruvad kaardid ja klasterdamine. Kahuks ei olnud tema tulemused väga põhjapanevad – leiti vaid, et laenu kasutamine reisiks, äriks ja tervisekulutuste katmiseks on pankrotti minemise ennustamisel tähtsad. Samuti leiti mõned grupid, mis on pigem sobivad investeerimistunnused.

Sammelsaar (2016) kasutas logistilist regressiooni kui panganduses kasutatavat standardset meetodit pankrotistumise tõenäosuse prognoosimiseks Bondora andmetel. Kuigi andmestikus on muutujaid tunduvalt enam, võttis Sammelsaar kasutusele vaid 16 muutujat. Sellele vaatamata lähenes ta ennustamisele interaktiivselt ning koostöös mudeliga suutis ta teha häid ennustusi. Ta kasutas ka ennustamiseks ka otsustuspuu meetodit, kuid tõdes, et logistiline regressioon on parem.

Taastumistemaatikat käsitles Lopes et al (2016), kes ennustas raskustesse sattunud laenude taastumist ühe Brasiilia panga andmetel. Parimaks ennustusmeetodiks sellise uurimisprobleemi lahendamisel osutus võimendamine, näidates seejuures väga head täpsust. Töös kasutati väga erinevaid meetodeid. Heterogeensetest ansambelmeetoditest kasutati meetodeid nagu HCES-Bag, AVG-W, GASEN. Homogeenseteks ansambelmeetoditeks, mida Lopes kasutas, olid BagNN, võimendamine ja otsustuspuude mets. Individuaalseteks klassifikaatoriteks olid logistiline regressioon, lineaarne diskriminantanalüüs ja tugivektormasin. Lopes hindas meetodite ennustusjõudu ROC-kõverate, AUC näitaja ja klassifitseerimistabelitega.

Suhteliselt palju on eelnevad autorid uurinud krediidiskooringu metoodikat üldisemalt, põhjendades uurimisprobleemi aktuaalsust inimeselt-inimesele krediidisektori plahvatusliku kasvuga (Berger 2009: 39) ja vajadusega teha kiireid (reaalajas) krediidiotsuseid veebiplatvormidel pakutavate laenude kontekstis. Võimendamist käsitleb ühe potentsiaalse kandidaadina seda probleemi lahendama Bastos (2008), kelle väitel sobib mitmete nõrkade õppijate agregeerimine väga hästi krediidiskooringu problemaatikat lahendama. Kõige populaarsemaks võimendamisalgoritmiks on seejuures AdaBoost (Bastos 2008: 4). Bastos kasutas võimendamist, närvivõrku ja tugivektormasinat ühel Saksa krediidiandmestikul, mis oli küll kahjuks küllaltki väike (1000 vaatlust) ja ühel austraalia andmestikul, mis oli veelgi väiksem (690 vaatlust), kuid siiski õnnestus tal saada huvitavaid tulemusi. Kõige paremaks meetodiks krediidiennustuste tegemisel osutus võimendamine, millele järgnes tugivektormasin ja närvivõrk, kusjuures tugivektormasin ja närvivõrk olid põhimõtteliselt võrdse ennustusjõuga ja võimendamine oli märgatavalt parem nii suhtarvudes kui ka graafiliselt analüüsides ja seda mõlemal andmestikul. Bastose loodud võimendamine suutis näidata 94-protsendilist täpsust krediidiotsuste tegemisel.

Krediidiskooringu ja andmeteaduse suhet uuris lähemalt ka Yap (2011), kes tõi välja, et krediidiskooringu mudelid tõepoolest aitavad pankadel ja teistel finantsasutustel paremini hinnata inimeste krediidivõimet ja klassifitseerida inimesi „headeks“ ja „halbadeks“. Yap toob ka välja, et ennustusmeetoditel ei laialdaselt kasutusvõimalusi ka teistes valdkondades, nagu näiteks kindlustuses, kinnisvaras, telekommunikatsioonis või mujal.

Sarnaselt Kisandile (2015) käsitleb optimaalse laenuportfelli koostamise temaatikat ka Polak (2017), küll aga mitte Bondora, aga suurema ühisrahastusplatvormi, Lending Clubi andmestikul, mis on Bondora andmestikust vaatluste arvult üle 16 korra suurem. Polak kasutab krediidiotsuste tegemisel logistilist regressiooni ja tõdeb sarnaselt Kisandile (2015), et krediidireitingu järgi investeerimine ei aita nii hästi pankrotte vältida, kuid seda teeb ise mudeli loomine ja selle tehtavate ennustuste põhjal investeerimine. Polak kasutas ka simulatsioone tõestamaks, et paljudesse odavatesse laenudesse investeerimine on parem kui vähestesse kallitesse, sest see aitab riske hajutada.

Lending Clubi andmestikku kasutas ka Polena (2017), kes võrdles kümmet erinevat klassifitseerimisalgoritmi krediidiskooringu tarbeks. Tema tulemuste järgi osutusid parimateks klassifitseerijateks logistiline regressioon, närvivõrk ja lineaarne diskriminantanalüüs. Töös kasutatud meetoditeks olid logistiline regressioon, lineaarne diskriminantanalüüs, tugivektormasin, närvivõrk, k-nearest neighbours, Bayesi võrk, otsustuspuu ja otsustuspuude mets.

Arandi (2017) kasutas masinõppe metoodikat pangakaardipettuste tuvastamiseks. Tema eesmärgiks oli luua masinõppel põhinevate meetoditega mudelid, mis suudaksid reaalajas vaadelda pangakaartidega tehtavaid tehinguid, et vältida käsitsi ülekontrollitavate tehingute mahtu. Arandi saavutas parimad tulemused otsustuspuude metsa, võimendamise ja närvivõrkudega. Kuna aga pettuste osakaal kõikidest tehingutest on kaduvväike, oli täpsus kõige parem otsustuspuude metsa puhul, mis on just selliste väga väiksesse vähemusse kuuluvate vaatluste leidmisel hea.

Üldisemalt läheneb krediidiennustuste problemaatikale Baesens et al (2003), kes toob välja, et logistilise regressioonile, diskriminantanalüüsile, närvivõrkudele ja otsustuspuudele tasub krediidiennustuste tegemisel kasutada ka tugivektormasinaid. Sarnane on ka Bellotti (2009) lähenemine. Ta tõdeb, et logistiline regressioon töötab krediidiandmeid analüüsides väga hästi, kuid leiab ka, et teatud tingimustel võivad tugivektormasinad paremad olla.

Baesens et al (2003) lähenemisele lisab 10 aastat hiljem Lessmann et al (2013) infot uudsete ennustusmeetodite, täpsusmõõdikute ja lähenemisviiside kohta, mida

krediidiskooringut käsitlev metoodika veel ei olnud käsitlenud. Lessmann toob välja 48 eelnevat uuringut krediidiennustuste teemal, 16 individuaalset klassifikaatorit, 8 homogeenset ansambelmeetodit ja 17 heterogeenset ansambelmeetodit, mida on krediidiandmetel ennustuste tegemisel eelnevalt kasutatud. Lessmanni läbiviidud analüüs näitas, et parimaks individuaalseks klassifikaatoriks on närvivõrk ja parimaks ansambelmeetodiks on otsustuspuude mets, kuid see oleneb täpsusmõõdikust, mida vaadelda. Väga selgelt toodi välja, et igal meetodil on omad tugevused ja nõrkused. Lessmann kinnitas ka juba ka eelnevate autorite poolt korduvalt tõestatut, et keerukamad meetodid suudavad tegevusala standardiks olevat logistilist regressiooni edestada.

Järgnevalt esitab autor käesoleva magistritöö tarbeks läbiviidud juhtumiuuringu käigu, kirjeldades andmeid, analüüsi ja seejärel tehes tulemustest järeldusi.

2. PANKROTISTUNUD LAENUDESSE INVESTEERIMINE

2.1. Ülevaade andmetest

2.1.1. LoanData krediidiandmestik

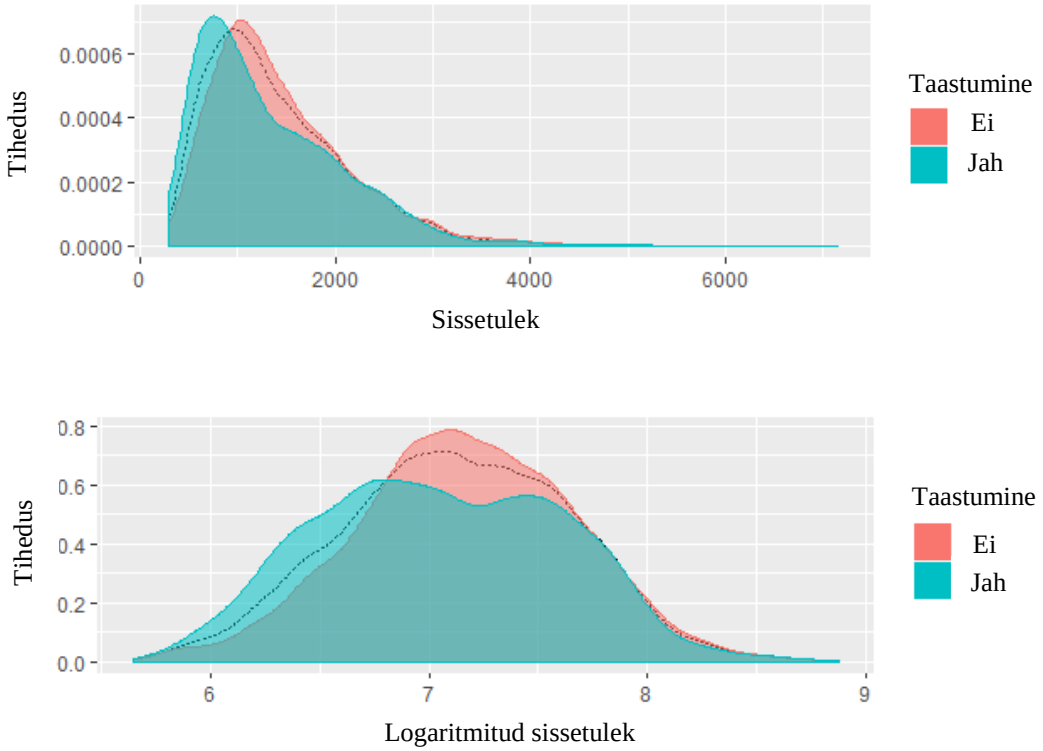
Andmestik, mida juhtumiuuringu läbiviimiseks kasutatakse, on pärit laenukeskkonnast Bondora. Analüüs viiakse läbi statistikapaketis R. Bondora poolt väljastatav investoritele suunatud LoanData nimeline andmestik on Bondora kodulehelt avalikult kättesaadav ja seda uuendatakse igapäevaselt uute laenulepingute lisandudes. Käesoleva töö tarbeks kasutatav andmestik on alla laetud 7. veebruaril 2018, seega sellest uuemaid vaatlusi andmestikus ei ole. Töötlemata andmestik sisaldab 51340 vaatlust, millest analüüsikõlbulik ei ole sugugi terve andmestik, vaid võrdlemisi väike osa sellest – analüüs on viidud läbi 14028 vaatlusele tuginedes. Suurima vähenemise põhjustas see, et välja tuli võtta kõik vaatlused, mis ei ole olnud pankrotis ja mõned vaatlused tuli ka kustutada ebausaldusväärsete andmete tõttu. Autor eemaldas ka erindit sisaldavad vaatlused. Vaatluste eemaldamine on tehtud „müra“ vähendamise kaalutlustel.

Andmestikus on kokku 112 muutujat. Muutujaid on nii selle kohta, mis on laenu võtnud inimese kohta teada laenu võtmise hetkel kui ka selle kohta, milline on tema käitumine pärast laenu võtmist. Käesoleva töö jaoks ei kasutata kõiki muutujaid ja seda väga erinevatel põhjustel. Lisaks sellele on väljundmuutuja autori poolt tuletatud taastumisjärgsete maksete alusel ja indikaatormuutujate loomisel on muutujaid juurde tekkinud. Kokku osaleb analüüsis 123 muutujat.

Andmestikus on võrdlemisi palju puuduvaid ja valesti (või puuduva kodeeringuga) kodeeritud väärtusi, mis vajavad eraldi tähelepanu. Kuna autor kavatseb analüüsis kasutada ka meetodeid, mille puhul on vajalik nii muutujate jaotuse parandamine kui ka kategoorilise muutujate indikaatormuutujateks teisendamine, siis tuleb lisaks puuduvate

ja valesti kodeeritud väärtustega tegelemisele muundada ka suur osa kõikidest kasutatavatest muutujatest.

Kuna sissetulek on enamasti muutuja, mis kipub nii päriselus kui ka sellest tulenevalt andmestikes olema küllaltki ebaühtlase ja normaaljaotusest erineva jaotusega, siis on see käesoleva juhtumiuuringu jaoks koos teiste sarnaste, parandust vajavate muutujatega, logaritmitud. Logaritmimise põhjustatud jaotuse ja varieeruvuse paranduse näide on illustreeritud joonisel 3.



Joonis 3. Sissetulekumuutuja logaritmimine ja sellest tulenev muutus jaotuses ja varieeruvuses taastumisklasside lõikes. Allikas: autori arvutused.

Nagu on eelolevalt joonisel näha, on logaritmimine drastiliselt parandanud sissetulekumuutuja jaotust ning see tuleb ennustavaid mudeleid luues kasuks, aidates mudelitel hõlpsamini leida andmete ja väljundmuutuja vahelisi seoseid. Tasub ka ära märkida, et jaotused on lahku löödud väljundmuutujaks olevate klasside lõikes, ehk pankrotist taastumist näitava muutuja järgi. Pankrotist taastumist näitav muutuja on saadud selle järgi, et kas on toimunud laenu tagasimakseid pärast pankroti

väljakuulutamist ja seejärel taastumist. Kui on, on muutuja väärtuseks 1 ja kui ei ole, siis on muutuja väärtuseks 0.

Andmete töötlemise, korrastamise, ülevaatamise, erindite eemaldamise, valekodeeringute eemaldamise ja puuduvate väärtustega tegelemise tulemusena jäi andmestikku alles 14028 vaatlust ja 123 muutujat, moodustades töös kasutatava andmestiku. Järgnevalt kirjeldab autor töös kasutatavat andmestikku lähemalt.

2.1.2. Analüüsiks kasutatava andmestiku kirjeldus

Selleks, et anda analüüsis kasutatavast andmestikust hea ülevaade, tuleks uurida erinevaid tähtsamaid muutujaid ja hea oleks seda teha ka väljundmuutujaks oleva pankrotist taastumist näitava muutuja lõikes. Lisaks sellele on vajalik ka arutleda andmestikus leiduvate muutujate üle, mis ei näita otseselt laenu võtva inimese mingit omadust, vaid on Bondora poolt arvutatud muutujad, mis on lisatud andmestikku abistamaks investoreid investeerimisotsuste tegemisel.

Bondora poolt arvutatud muutujateks on:

1. oodatav tootlus,
2. oodatav kahju
3. oodatav kahju pankroti korral,
4. pankroti tõenäosus (üheaastase perioodi vältel),
5. oodatav laenu põhiosa pankroti korral,
6. planeeritud saadav intress,
7. oodatav intress pankroti korral.

Arvestades käesoleva töö spetsiifikat ja tõsiasja, et juhtumiuuringu eesmärgiks on panna investor parimasse võimalikku olukorda pankrottidesse investeerimisel, võtab autor analüüsis neid muutujaid arvesse. Kui aga eesmärgiks oleks olnud pankrotist taastumise protsessi kirjeldamine (mitte etteennustamine), siis oleks ehk olnud õigem jätta välja Bondora poolt tehtud arvutused ja lähtuda vaid tooretest andmetest.

Suur osa analüüsis kasutatavatest muutujatest on logaritmitud, et parandada nende jaotust ja suurt varieeruvust. Üheks võimaluseks, kuidas logaritmimise asemel oleks võinud muutujaid normaliseerida, oleks olnud Z-skooride arvutamine.

Käesolevas töös on logaritmitud muutujateks:

1. taodeldud laenusumma,
2. saadud laenusumma,
3. intressimäär,
4. sissetulek,
5. võlgade kogusumma,
6. planeeritud saadav intress,
7. oodatav kahju.

Analüüsis kasutatavad kategoorilised muutujad on muudetud indikaatormuutujateks, sest vastasel korral poleks võimalik analüüsi korrektselt läbi viia. Näiteks hariduse muutuja on muundatud viieks indikaatormuutujaks, mis iga üks väljendavad ühte haridustaset andes seejuures väljendatavale haridustasemele väärtuseks 1 ja teistele 0. Sedasi väljendub kategooriline haridusmuutuja viie indikaatormuutujana.

Indikaatormuutujateks on käesolevas töös:

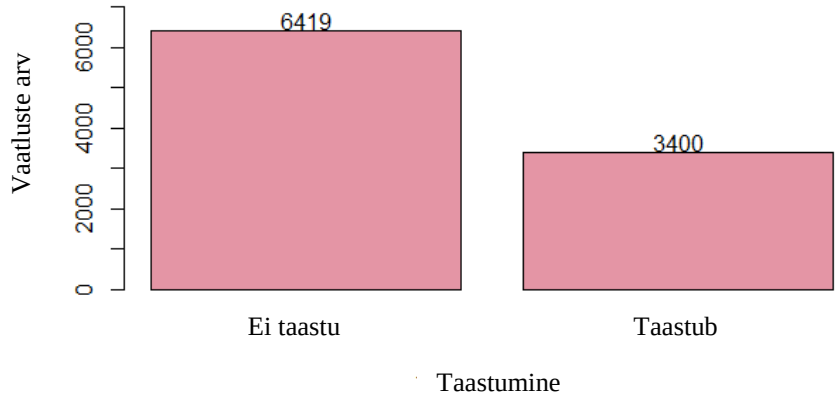
1. riik,
2. tööstaaži klassid praeguse tööandja juures,
3. tööstaaži klassid,
4. krediidireiting,
5. uus klient,
6. sissetulekute ja väljaminekute kinnitamise viis,
7. keel,
8. sugu,
9. laenu otstarve,
10. haridus,
11. abielustaatus,
12. tööhõivestaatus,
13. koduomandlus.

Järgnevalt annab autor graafilise ülevaate valitud muutujatest, et ilmestada mõningaid kasutatavale andmestikule iseloomulikke jooni. Autori arvates on siinkohal kõige otstarbekam iseloomustada just seda andmestikku, mille pealt luuakse peatükis 2.2 ennustavad mudelid. Ennustavate mudelite loomiseks ja hindamiseks on autor juhuslikult jaotanud andmestiku kolmeks osaks, mille suurused on 70%, 15% ja 15%. 70% suurusega andmestikku nimetatakse treeningandmestikuks ja see on ka andmestik, mille põhjal tasub siinkohal teha kirjeldavat statistikat. Treeningandmestiku suuruseks on seega 70% koguandmestikust (14028) ehk siis $14028 \times 0{,}7 = 9819$ vaatlust. 15%

suurustega andmestikke nimetatakse vastavalt valideerimisandmestikuks ja testandmestikuks. Valideerimisandmestikku (2104 vaatlust) kasutab autor mudelite ennustusvõime võrdlemiseks ja testandmestik (2105 vaatlust) jääb esialgu puutumata, et hiljem hinnata parimaks osutunud mudeli töövõimet ennenägemata andmetel.

2.1.3. Kirjeldav statistika

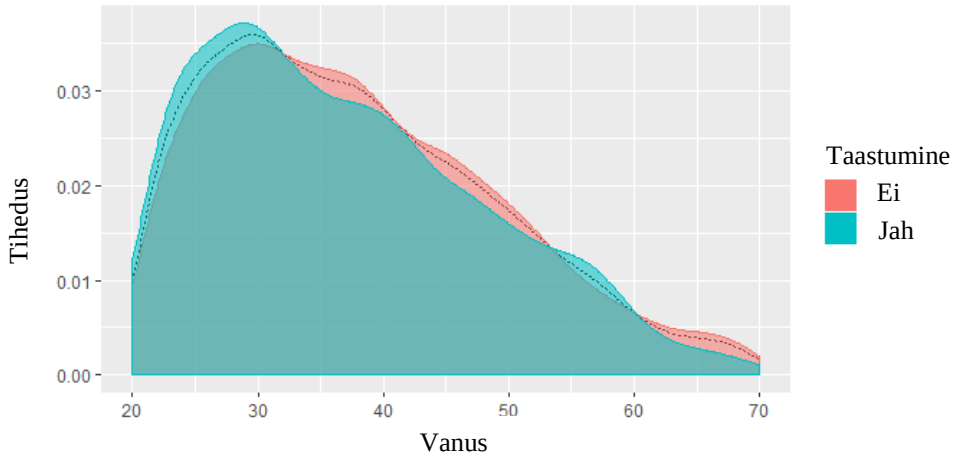
Käesoleva peatüki tarbeks on autor teinud töös kasutatavatest muutujatest valiku, sest kõiki muutujaid kirjeldada ei oleks kuigi mõttekas, sest neid on väga palju (123 sisendmuutujat + 1 väljundmuutuja). Valiku tegemisel on lähtutud sellest, et milliste muutujate ilmestamine annab andmestikust kõige parema ülevaate. Kõigepealt kirjeldatakse väljundmuutujaks olevat pankrotist taastumise muutujat.



Joonis 4. Taastumise jaotus treeningandmestikus. Allikas: autori koostatud.

Nagu on jooniselt 4 näha, siis taastumismuutuja jaotub treeningandmestikus küllaltki ebaühtlaselt. On selge, et taastuvaid laene (3400 vaatlust) on vähem kui mittetaastuvaid (6419 vaatlust). Taastumismääraks on seega 34,6%, mis tähendab, et ennustaval mudelil tuleb leida vähemusse kuuluvaid vaatlusi.

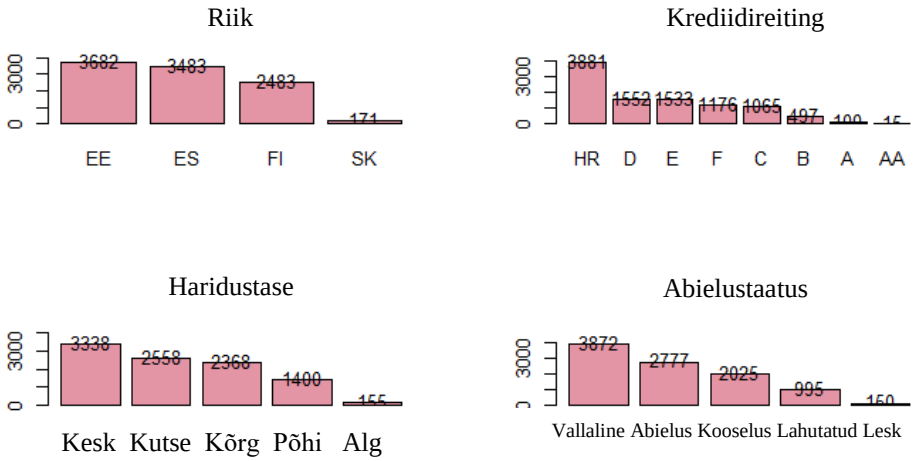
Kindlasti tasub ka uurida, milline on tavaline laenuvõtja, kelle võetud laen on pankrotti läinud. Kuna analüüsis kasutatavas andmestikus on vaid pankrotistunud laenud, siis selleks mingeid vaatlusi välja võtma ei pea ja tasub muutujate jaotusi ilmestada joonisel. Laenuvõtjate vanuse ilmestamiseks on autor koostanud alljärgneva joonise 5.



Joonis 5. Vanuse jaotus andmestikus. Allikas: autori koostatud.

Noorimateks laenuvõtjateks andmestikus on 20-, 21- ja 22-aastane inimene ja vanimateks 68-, 69- ja 70-aastane. Nagu on näha joonisel 4, langeb suur osa laenuvõtjatest vanusesse 25-40. Laenuvõtjate keskmine vanus on 37,78 aastat. Suurt erinevust taastumise lõikes vanuse järgi märgata ei ole. Graafiliselt hinnates tundub vanus taastumise kirjeldamisel küllaltki väikest rolli mängivat. Lisaks vanusele on autor veel välja valinud 4 laenu võtvaid isikuid hästi kirjeldavat muutujat, et veelgi paremini näidata, milline on tavaline laenuvõtja analüüsi aluseks olevas andmestikus.

Alljärgneval joonisel 6 on toodud muutuja laenu võtnud isiku riigi kohta. Andmestikus on andmeid laenude kohta, mis on antud nelja erinevasse riiki, milleks on Eesti, Soome, Hispaania ja Slovakkia. Lisaks riigile on toodud ka muutuja isiku krediidireitingu kohta, mis varieerub kõrge riski kategooriast kuni AA-reitinguga laenudeni. Ära on toodud ka muutuja haridustaseme kohta. Haridustasemeid on andmestikus ära toodud viis erinevat. Lisaks eeltoodule on joonisel ka muutuja abielustaatuse kohta, mis on samuti jaotatud viide kategooriasse.



Joonis 6. Riigi, krediidireitingu, haridustaseme ja abielustaatuse jaotus treeningandmestikus. Allikas: autori koostatud.

Nagu on jooniselt 6 näha, siis muutujate jaotus ei ole ühegi vaadeldava muutuja puhul ühtlane, vaid mõned grupid on võrreldes teistega rohkem esindatud ja sisaldavad seega rohkem vaatlusi. Näiteks kui võrrelda laenuvõtjaid riigiti, siis on näha, et kõige enam on laenuvõtjaid Eestist (3682) ja Hispaaniast (3483). Veidi vähem on andmestikus laenuvõtjaid Soomest (2483) ja väga vähe on laenuvõtjaid Slovakkiast (171). Krediidireitingu alusel on suurim grupp just kõrge riskiga grupp, mis on ka oodatud, sest andmestikus on vaid pankrotistunud laenud. Heale reitingule vaatamata on ka andmestikku sattunud 212 vaatlust, kus laenuvõtja reiting on B või kõrgem, kuid on läinud pankrotti. Hariduse järgi on kõige rohkem inimesi (3338) keskhariduse grupis, millele järgnevad kahanevas järjekorras kutseharidus, kõrgharidus, põhiharidus ja algharidus. Abielustaatuselt on suurimaks grupiks vallalised (3872 vaatlust), kellele järgnevad abielus olevad inimesed (2777 vaatlust) ja seejärel kooselavad, lahutatud ja lesestunud inimesed.

Omades ettekujutust pidevatest, kategoorilistest ja arvulistest muutujatest, mida töös kasutatakse on võimalik edasi liikuda andmete ja väljundmuutuja vaheliste seoste otsimisele, ehk pankrotistunud laenude taastumise etteennustamisele. Ennustamiseks vajalikku analüüsiprotsessi kirjeldab järgnev peatükk 2.2.

2.2. Pankrotist taastumise ennustamine

Käesolevas peatükis läbiviidav analüüs on tehtud teoreetilise osa peatükis 1.2 kirjeldatud metoodika alusel ja võtab arvesse kõiki kuut ennustusmeetodit, mida teoreetilises osas kirjeldatud sai. Tegemist on juhtumiuuringuga, et uurida pankrottidesse investeerimise strateegia võimalikkust laenukeskkonnas Bondora. Selleks, et juhtumiuuringu tulemusi saaks lugeda positiivseteks, peab küllaltki hea täpsusega olema võimalik ette ennustada, millised laenud pankrotist taastuvad ja millised mitte.

Määratlemaks, et mida võiks mõista „küllaltki hea täpsuse“ all, tuleb see termin lahti seletada. Võime eeldada, et pankrotistunud laenusid müüakse Bondora järelturul allahindlusega. Selle eelduse tõestamiseks tasub vaid internetis lahti teha Bondora järelturg ja veenduda oma silmaga, et pankrotistunud laenusid tõepoolest müüakse väga suure allahindlusega. Pole sugugi ebatavaline leida laenusid, mille allahindlusprotsendiks on üle 50%. Ennustaval mudelil on alati mingisugune eksimistõenäosus ja seda peame me arvestama kuluna. Seega võime eeldada, et nulli jäämise punkt on investori jaoks pankrottidesse investeerimise strateegiat rakendades oodatav tulu, mis katab ära oodatava kulu.

Investori oodatavat tulu allahindlusega müüdava laenuosaku ostmisel võib väljendada järgnevalt:

(1) $$ER = P * T + P * (1 + I) * T$$

kus ER – investeeringult oodatav tulu,
P – laenuosaku hind,
T – mudeli täpsusprotsent,
I – laenu intressimäär.

Valemile 1 vastavalt võib investori oodatavat kulu samas situatsioonis väljendada nii:

(2) $$EC = P * R * F$$

kus EC – investeeringu oodatav kulu,

P – laenuosaku hind,

R – allahindlusprotsent,

F – mudeli eksimismäär.

Nii lihtsuse kui ka riskikartlikkuse huvides võime käesoleva töö tarbeks „küllaltki heaks“ mudeliks pidada mudelit, mille eksimismäär ei ole suurem kui 10%. Sellise täpsusega mudelit kasutades investeerimine tooks edu korral kaasa vägagi kõrge potentsiaalse tootluse, mis oleks konkurentsivõimeline teiste turul pakutavate investeerimistoodetega, seda küll arvestamata taastumisega seonduvat ajalist juhuslikkust.

2.2.1. Närvivõrk

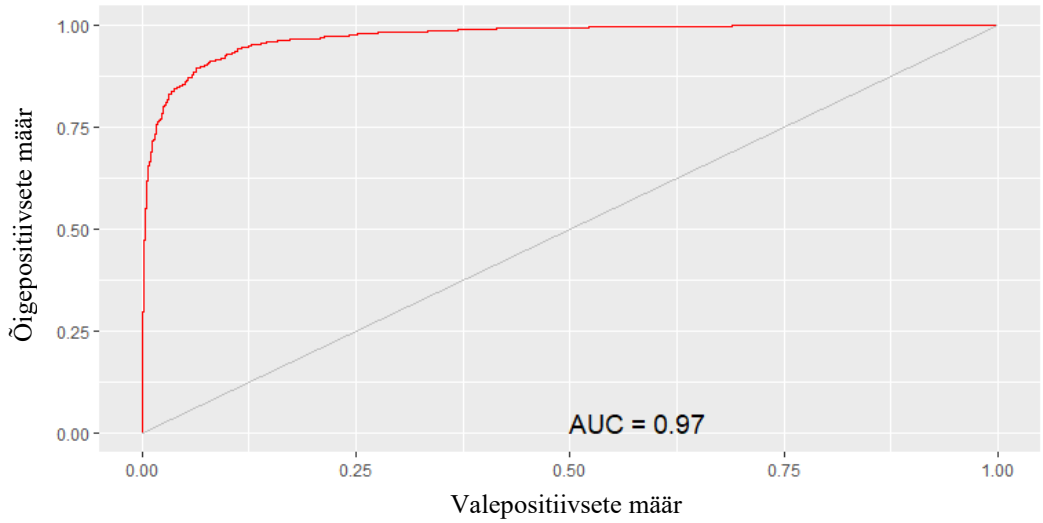
Kõigepealt uuris autor väljundmuutuja Recovery (pankrotist taastumine) ennustamisel *Back-Propagation* metoodikal (vt peatükk 1.2.1) toimivat närvivõrku. Nagu ka kõik teised mudelid, mida peatükis 2.2 on käsitletud, on ka närvivõrku rakendatud testandmestikul, tagades andmestiku puhtuse, sest tegemist on andmestikuga, mis ei ole osalenud ei treeningul ega valideerimisel. Närvivõrgu struktuuriks on 123-2-1 (sisendmuutujad-peidetud muutujad-väljundmuutujad). Muutujad on normaliseeritud Z-skooridega. Ennustustulemused on autor koondanud alljärgnevasse tabelisse 5.

Tabel 5. Närvivõrgu ennustustulemuste paigutumine maatriksisse.

		Ennustatud		
		Ei taastu pankrotist	**Taastub pankrotist**	**Viga klassi ennustamisel**
Tegelik	**Ei taastu pankrotist**	1278	55	4,1%
	Taastub pankrotist	117	652	15,2%
Mudeli veamäär		8,2%		
Keskmine klassiviga		9,65%		
Viga investeerimisel		7,8%		

Allikas: autori arvutused statistikapaketis R.

Nagu on näha tabelist 5, oli mudel ennustamisel küllaltki täpne. Pankrotist taastumist suudeti õigesti ette ennustada 1278 korral ja mittetaastumist 652 korral. Riskikartlikku investorit huvitab ehk aga kõige enam see, kui suure osa mittetaastuvatest pankrottidest suudab mudel üles leida. Siinkohal tuleb tunnistada, et mudel ei ole väga täpne. 15,2% mittetaastuvatest pankrottidest jäi mudelil leidmata. Üldine mudeli veamäär on aga sellest oluliselt madalam, vaid 8,2%. Kui oletada, et investor investeerib müügil olevasse laenuosakusse vaid siis, kui mudel selle taastumist ennustab, siis mudeli tegelik viga investori poolt kaalutavate otsuste tegemisel on 55 vaatlust 707 positiivsest ennustusest, ehk 7,8%. Kuigi autor peab sedasi arvutatud investeerimisviga tähtsaimaks täpsusnäitajaks, mida käesoleva juhtumiuuringu tarbes vaja läheb, siis tuleks siiski täpsushinnangu kontrollimiseks ja sügavamaks analüüsiks uurida mudeli ennustusvõimet graafiliselt. Selleks on autor koostanud alljärgneva joonise 7.

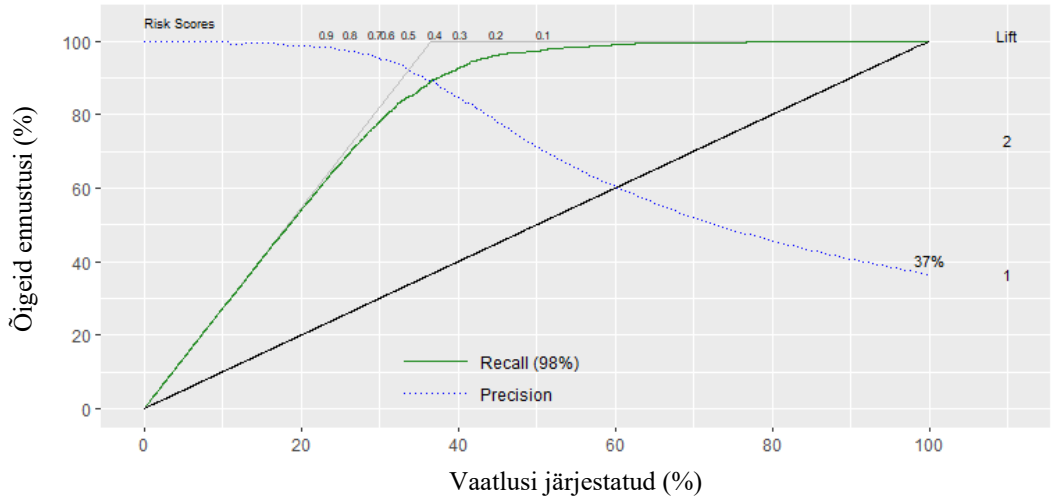


Joonis 7. Närvivõrgu ROC-kõver. Allikas: autori arvutused statistikapaketis R.

Ideaalset ennustajat kirjeldaks joonisel 7 ülemise vasaku nurgani ulatuv ROC-kõver. Autori loodud närvivõrk ei ole küll ideaalne ennustaja, aga sellegi poolest on sellel väga tugev ennustusjõud. Ala graafikul oleva sirge ja ROC-kõvera vahel (AUC) näitab mudeli diskrimineerimisvõimet või teisisõnu võimet teha vahet pankrotist taastuvatel ja mittetaastuvatel laenudel. Närvivõrgu poolt saavutatud AUC = 0,97 näitab väga head diskrimineerimisvõimet. Diskrimineerimisvõime tähendab siinjuures tõenäosust, et närvivõrk hindab suvaliselt valitud taastuvat laenu taastumisvõime poolest kõrgemalt kui suvaliselt valitud mittetaastuvat laenu. Seega saab öelda, et 97% tõenäosusega

hindab mudel taastuva laenu taastumisvõimet kõrgemalt kui mittetaastuva laenu taastumisvõimet.

ROC-kõver näitab küll hästi mudeli üldist täpsust, kuid mudeli poolt antava taastumistõenäosuse järgi saab vaatlusi ka järjekorda panna. Oluline oleks seejuures vaadata, kui hästi suudab mudel klassifitseerida siis, kui alustada selle tehtud tugevamatest hinnangutest. Selleks on autor koostanud alljärgneva joonise 8.



Joonis 8. Närvivõrgu riskigraafik. Allikas: autori arvutused statistikapaketis R.

Peatükis 1.2.7 kirjeldatud riskigraafikuid puudutava metoodika järgi on heaks punktiks joonisel 8 kujutatud riskigraafiku analüüsimisel mõni kindel protsent, mis väljendab järjest ärahinnatud vaatluste osakaalu kõigist vaatlustest. Autori hinnangul võiks joonisel 8 kujutatud riskigraafikut analüüsida kohas, kus on hinnangu tugevuse järgi ära reastatud 40% vaatlustest. Selleks hetkeks on mudel veel väga täpne ja vigu on tehtud väga vähe. Mudel on selleks hetkeks üles leidnud rohkem kui 90% kõigist mittetaastuvatest laenudest ja lift-kordaja näitab sellel hetkel, et mudel on huupi arvavast mudelist üle kahe korra parem. See näitab, et mudel suudab küllaltki kiiresti üles leida väga suure osa mittetaastuvatest laenudest.

Järgmisena uurib autor peatükis 1.2.2 käsitletud tugivektormasina ennustusvõimet testandmestikul.

2.2.2. Tugivektormasin

Sarnaselt Haltufile (2014) kasutab autor ka käesolevas töös tugivektormasinat Bondora andmestikul. Haltuf kasutas analüüsiks kahte erinevat tugivektormasina funktsiooni (kernel), milleks olid lineaarne ja Gaussi funktsioon. Ühtlasi tõestas Haltuf tugivektormasinate kasutamise võimalikkust ja mõttekust P2P laenuandmetel. Seega on tugivektormasinaid kasutatud nii tavalistel krediidiandmetel kui ka andmetel, mis on sarnased käesolevas magistritöös kasutatavale andmestikule. Tuginedes Hsu (2016) metoodikale funktsiooni valikul proovis käesoleva töö autor valideerimisandmestikul läbi järgnevad funktsioonid:

1. Gauss – veamäär 10%
2. Polünomiaal – veamäär 6,9%
3. Lineaarne – veamäär 6,9%
4. Laplace – veamäär 23%

Kuigi nii polünomiaalse kui ka lineaarse funktsiooni puhul oli veamääraks 6,9%, oli lineaarse funktsiooni puhul keskmine klassiviga väiksem ja seetõttu kasutab autor analüüsi läbiviimiseks just lineaarset funktsiooni. Järgnevalt rakendatakse lineaarse funktsiooniga treenitud mudelit testandmestikul, et luua võrdlusalus teiste mudelitega ja ühtlasi eemaldada võimalik valideerimisest tulenev liigne täpsus. Seega on tabelis 6 näidatud tulemused saadud testandmestikule rakendades ja simuleerivad mudeli täpsust uutel, ennenägematutel andmetel.

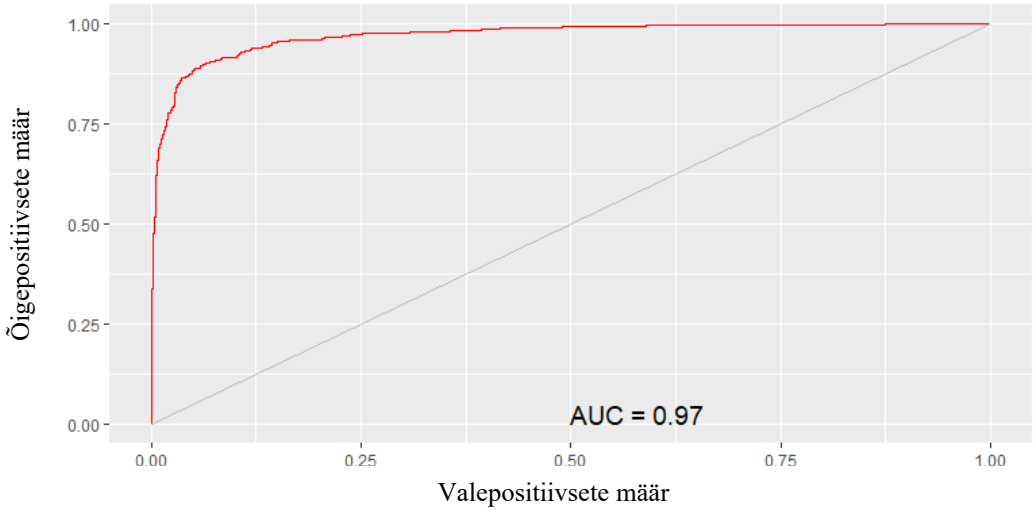
Tabel 6. Tugivektormasina ennustustulemuste paigutumine maatriksisse.

		Ennustatud		
		Taastub pankrotist	**Ei taastu pankrotist**	**Viga klassi ennustamisel**
Tegelik	**Taastub pankrotist**	1286	47	3,5%
	Ei taastu pankrotist	109	660	14,2%
Mudeli veamäär		7,4%		
Keskmine klassiviga		8,85%		
Viga investeerimisel		6,6%		

Allikas: autori arvutused.

Kui võrrelda tabelis 6 näidatud tugivektormasina tulemusi tabelis 5 näidatud närvivõrgu tulemustega, on selge, et tugivektormasin on igas aspektis parem. Nii mudeli üldine

veamäär (7,4%) kui ka käesoleva töö jaoks olulisim näitaja, viga investeerimisel (6,6%), on mõlemad küllaltki madalad ja näitavad seega mudeli tugevat ennustusvõimet. Autorile tundub, et tugivektormasin teeb väga vähe valepositiivseid ennustusi, mida riskikartlik investor proovibki vältida. Mudeli ennustusvõimet hindab autor ka graafiliselt alljärgnevatel joonistel 9 ja 10.

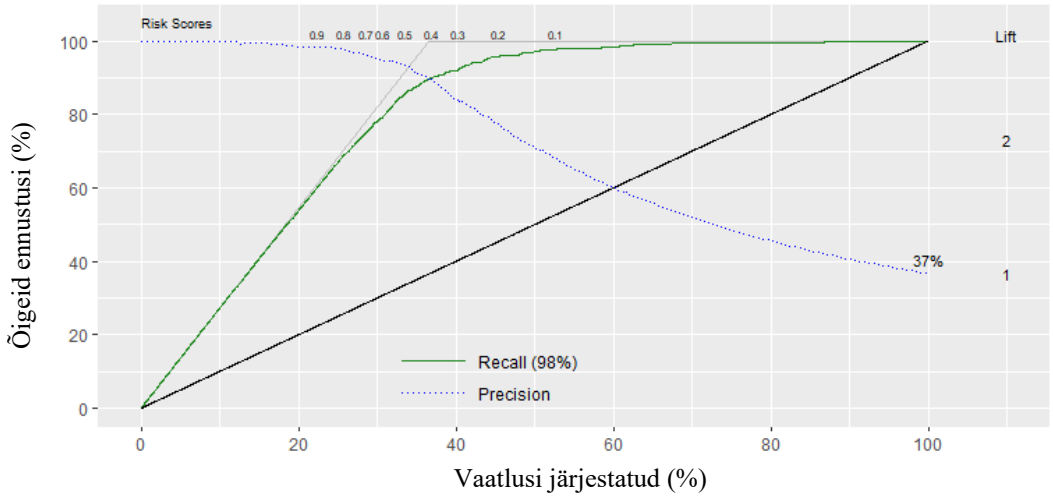


Joonis 9. Tugivektormasina ROC-kõver. Allikas: autori arvutused statistikapaketis R.

Joonisel 9 kujutatud tugivektormasina ROC-kõver näitab, sarnaselt joonisel 7 kujutatud närvivõrgu ROC-kõverale, et nii närvivõrgu kui ka tugivektormasina puhul on tegemist tugeva ennustajaga, mis sobib hästi käesolevas töös kasutatud andmestikust leidma seoseid sisendmuutujate ja väljundmuutuja vahel ja seega ette ennustama pankrotist taastumist. Joonisel kujutatud AUC (0,97) väljendab tõenäosust, et tugivektormasin hindab suvaliselt valitud taastuvat laenu taastumisvõime poolest kõrgemalt kui suvaliselt valitud mittetaastuvat laenu. Seega saab öelda, et 97% tõenäosusega hindab mudel taastuva laenu taastumisvõimet kõrgemalt kui mittetaastuva laenu taastumisvõimet.

ROC-kõvera poolt näidatavat üldist täpsushinnangut peab autor küll oluliseks, aga mõeldes sellele, kuidas ennustav mudel päriselus kasutust leiaks, on olulisem vaadata mitte suvaliselt valitud vaatlusi, vaid taastumistõenäosuse järgi reastatud vaatlusi alates tugevamast. See tuleneb tõsiasjast, et riskikartlik investor ei alustaks investeerimist mitte suurima allahindlusprotsendiga laenudest vaid laenudest, mille puhul ennustav mudel annab suurima tõenäosuse taastumiseks. Seetõttu ongi oluline vaadata, kui hästi

suudab mudel klassifitseerida siis, kui alustada selle tehtud tugevamatest hinnangutest. Selleks on autor koostanud alljärgneva joonise 10.



Joonis 10. Tugivektormasina riskigraafik. Allikas: autori arvutused statistikapaketis R.

Tugivektormasina riskigraafik on sarnane närvivõrgu riskigraafikuga. See võib tuleneda sellest, et mudelite täpsusnäitajad olid küllaltki sarnased. Autori hinnangul võiks joonisel 10 kujutatud tugivektormasina riskigraafikut analüüsida kohas, kus on hinnangu tugevuse järgi ära reastatud 30% vaatlustest. Selleks hetkeks pole veel võimalik, et ära on tuntud kõik taastuvad laenud, sest neid oli andmestikus umbes 34%. 30% vaatluste reastamise järel on mudel veel väga täpne ja vigu on tehtud väga vähe. Mudel on selleks hetkeks üles leidnud rohkem kui 75% kõigist mittetaastuvatest laenudest ja lift-kordaja näitab sellel hetkel, et mudel on huupi arvavast mudelist üle kahe korra parem. Veamäär on selles punktis umbes 5%. See näitab, et mudel suudab küllaltki kiiresti üles leida väga suure osa mittetaastuvatest laenudest.

Kui närvivõrgu ja tugivektormasina puhul on otsustamise protsessi keeruline ilmestada, siis otsustuspuu puhul on see lihtne järgnevalt analüüsib autor otsustuspuu meetodit pankrotist taastumise etteennustamisel.

2.2.3. Otsustuspuu

Selles peatükis käsitleb autor teoreetilise osa peatükis 1.2.3 väljatoodud meetodit otsustuspuu abil ennustuste tegemisel. Otsustuspuu kui üksik ennustaja ei pruugi olla kuigi tugeva ennustusvõimega, sest võtab arvesse suhteliselt väikest osa kõikidest

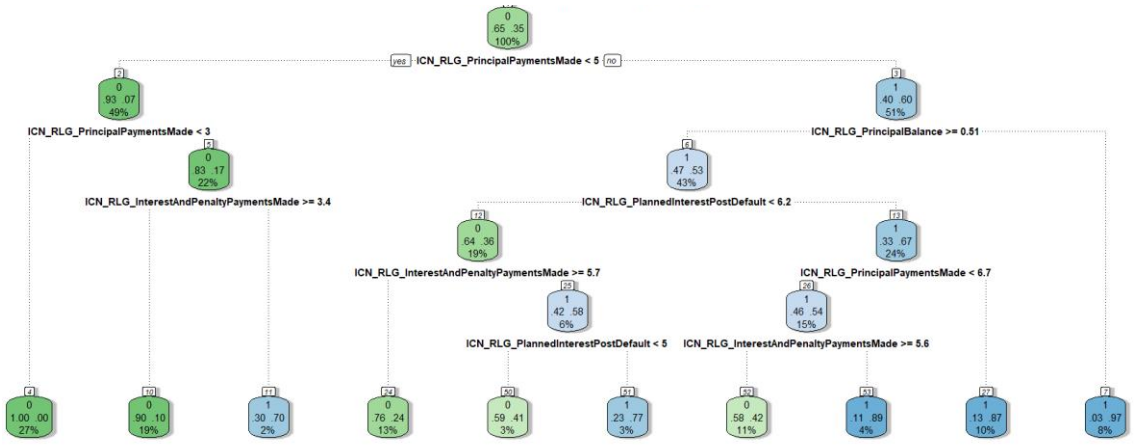
andmestikus leiduvatest muutujatest. Otsustuspuu meetodil saadud ennustustulemused on koondatud alljärgnevasse tabelisse 7.

Tabel 7. Otsustuspuu ennustustulemuste paigutumine maatriksisse.

		Ennustatud		
		Ei taastu pankrotist	**Taastub pankrotist**	**Viga klassi ennustamisel**
Tegelik	**Ei taastu pankrotist**	1249	86	6,4%
	Taastub pankrotist	262	508	34%
Mudeli veamäär		16,6%		
Keskmine klassiviga		20,2%		
Viga investeerimisel		14,5%		

Allikas: autori arvutused statistikapaketis R.

Nagu on tabelist 7 näha, ei ole otsustuspuu tulemused oma täpsuselt närvivõrgu ja tugivektormasinaga võrreldavad, jäädes kahele eelpool analüüsitud meetodile tugevalt alla. On näha, et mudel on liiga optimistlik. Kõik veahinnangud on umbes kaks korda suuremad kui närvivõrgu või tugivektormasina korral. Autori hinnangul annaks seda mudelit üsna palju parandada, andes mudelite ette teistsugune eeljaotus (priors) ja lisades juurde mudeli lubatavat sügavust. Mudeli sügavust on autor piiranud viieni, et seda paremini joonisel ilmestada. Otsustuspuu on graafiliselt kujutatud alljärgneval joonisel 11.

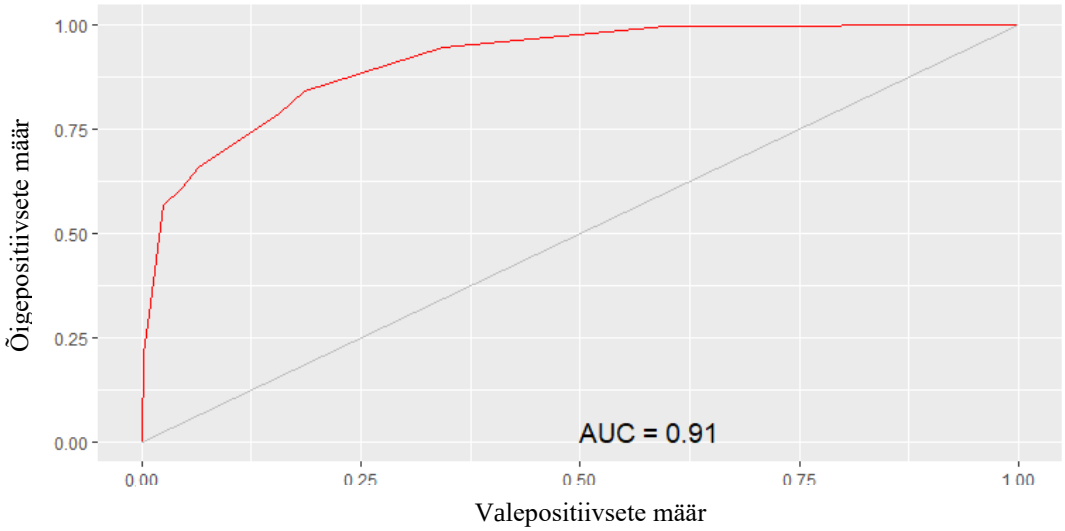


Joonis 11. Otsustuspuu. Allikas: autori arvutused statistikapaketis R.

Autor viitab peatükis 1.2.3 Käärmannile (2003), kes kirjeldab otsustuspuud kui mängu, mida lapsed mõnikord igavuse peletamiseks mängivad. Siinkohal tasuks proovida

loodud otsustuspuud lahti kirjutada nii, et suhteliselt keerukast joonisest saaks lihtsasti mõistetav tekst. Mudel on reeglitena välja kirjutatud lisas 1.

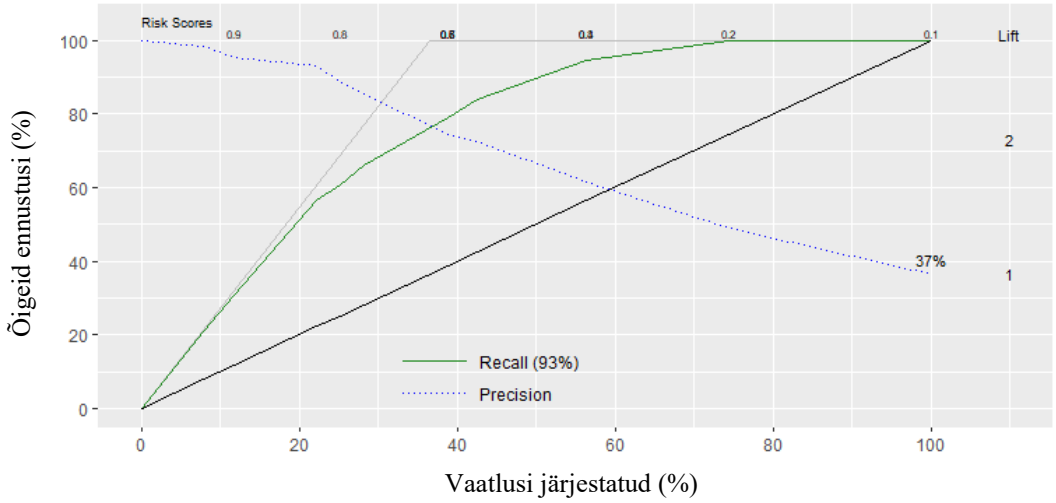
Hakates joonisel 11 kujutatud otsustuspuud mööda ülevalt alla tulema üritab mudel jaotada vaatlused kahte gruppi – laenud, mis taastuvad pankrotist (kodeeritud kui 1) ja laenud, mis pankrotist ei taastu (kodeeritud kui 0). Kuna kõik otsustuspuule koondatud muutujad on logaritmitud, saame rääkida jagunemisel mingist muutuja tasemest, millel toimub lahknevus. Otsustuspuu vasakpoolset haru vaadates on selge, et mida vähem on tehtud laenu tagasimakseid, seda suurema tõenäosusega laen pankrotist ei taastu. Samas on aga positiivne mõju laenu taastumisele sellel, kui tehtud on palju eelnevaid intressimakseid. Otsustuspuu parempoolset haru vaadates näeme, et eelnevate tagasimaksete summal on tugev positiivne mõju pankrotist taastumise tõenäosusele. Samuti on positiivne mõju sellel, kui laenu pole enam väga palju jäänud tagasi maksta. Mudel on pidanud oluliseks ka Bondora poolt arvutatud muutujat planeeritud pankrotijärgsete intressimaksete kohta ja ka muutujat eelnevate intressimaksete kohta. See kõik tundub autori arvates olevat kooskõlas loogika ja terve mõistusega.



Joonis 12. Otsustuspuu ROC-kõver. Allikas: autori arvutused statistikapaketis R.

Sarnaselt närvivõrgu ja tugivektormasina jaoks läbi viidud analüüsiga koostas autor ka otsustuspuu jaoks ROC-kõvera, mis seekord on aga teistest küllaltki erinev. Nimelt ei ulatu see kõver nii kaugele vasakule ülemisse nurka kui seda tegid närvivõrgu ja tugivektormasina ROC-kõverad. Põhjus peitud otsustuspuu väiksemas ennustusvõimes

ja suuremas ebatäpsuses. Selle valepositiivsete ennustuste määr kasvab palju kiiremini kui teiste meetodite puhul. Samuti ei ole ka AUC (0,91) nii kõrge, nagu eelnevate mudelite puhul.



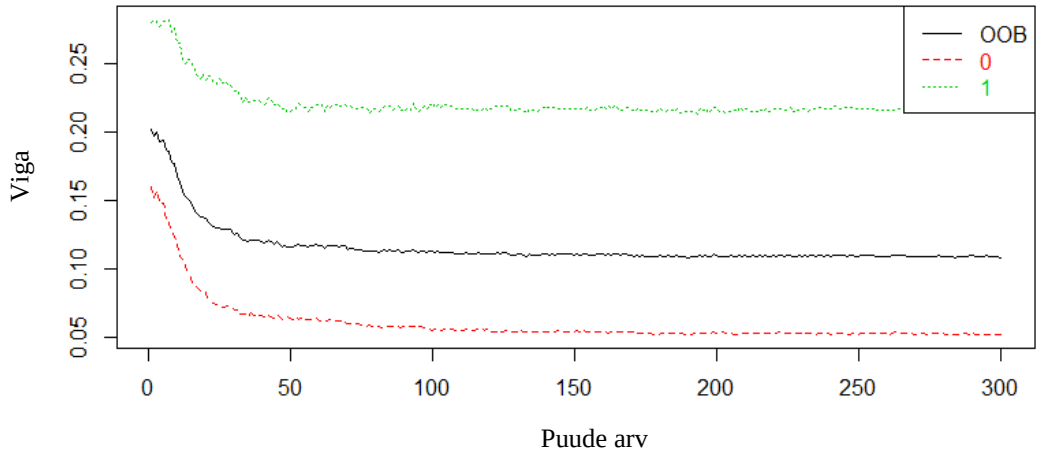
Joonis 13. Otsustuspuu riskigraafik. Allikas: autori arvutused statistikapaketis R.

Otsustuspuu riskigraafik ei sarnane närvivõrgu ega tugivektormasina riskigraafikutega. Tundlikkuse (recall, pidev roheline) joon on laugem ja täpsuse (precision, katkendlik sinine) joon langeb kiiremini. See tuleneb mudeli oluliselt halvematest täpsusnäitajatest. Autori hinnangul võiks joonisel 13 kujutatud otsustuspuu riskigraafikut analüüsida kohas, kus on hinnangu tugevuse järgi ära reastatud 30% vaatlustest. Nagu sai ka tugivektormasina analüüsis märgitud, pole selleks hetkeks veel võimalik, et ära on tuntud kõik taastuvad laenud, sest neid oli andmestikus umbes 34%. 30% vaatluste reastamise järel ei ole otsustuspuu mudel kuigi täpne ja tehtud on küllaltki palju vigu. Mudel on selleks hetkeks üles leidnud vaid veidi alla 70% kõigist mittetaastuvatest laenudest, kuid lift-kordaja näitab sellel hetkel, et mudel on huupi arvavast mudelist üle kahe korra parem. Veamäär on selles punktis peaaegu 20%. See näitab, et mudel ei ole pankrotist taastumise etteennustamisel kuigi hea ennustusjõuga.

Eelnevale tuginedes saab seega väita, et üksik otsustuspuu ei ole kuigi hea ennustaja. Otsustuspuid on aga võimalik teha mitu ja need kokku koondada ansambliks, ehk otsustuspuude metsaks.

2.2.4. Otsustuspuude mets

Otsustuspuid metsaks koondades pole seda vaja teha lõputult, vaid kuskil on piir, millest edasi puude lisandumine mudeli ennustusvõimet enam ei paranda. Selle piiri otsimiseks tegi autor 300 otsustuspuud, millest iga puu sügavuseks oli 11 taset ja uuris mudeli veamäära muutumist. Tulemused on koondatud alljärgnevale joonisele 14.



Joonis 14. Otsustuspuude metsa veamäär puude lisandudes. Allikas: autori arvutused statistikapaketis R.

Joonisel 14 on kolm joont. Roheline katkendlik joon näitab taastuvate laenude leidmise veamäära, punane katkendlik joon näitab taastumatute laenude leidmise veamäära ja must pidev joon näitab mudeli üldist veamäära. Kuigi kõigi kolme joone käitumine on esimeste puude lisandudes küllaltki erinev, saabub siiski hetk (umbes 100 puu juures), kus edasine ennustusvõime võit iga puu lisandudes on väga väike. Vaatamata sellele on käesolevas peatükis läbiviidud analüüs tehtud 300 puu pealt treenitud mudeliga.

Otsustuspuude metsa eripäraks on see, et väga edukalt on võimalik välja tuua muutujate tähtsushierarhia, ehk teisisõnu, on võimalik näidata, millised muutujad on kõige olulisemad ja omavad mudelis kõige suuremat kaalu. Järgnevalt toob autor välja kümme olulisemat muutujat (vt lisa 2) nende tähtsuse järjekorras:

1. Eelnevate tagasimaksete summa
2. Laenu jääk
3. Planeeritud pankrotijärgne intress (Bondora arvutatud)
4. Planeeritud pankrotijärgne põhiosa (Bondora arvutatud)
5. Intressi jääk

6. Intressimaksete summa
7. Oodatav kahju pankroti korral (Bondora arvutatud)
8. Graafikujärgne oodatav põhiosa tagasimaksete summa hetkel
9. Oodatav tootlus
10. Graafikujärgne oodatav intressimaksete summa hetkel

Muutujate tähtsushierarhia ei sisalda üllatusi. Autori arvates on siin näha pöördumatute kulude efekti (sunk cost effect), ehk väga suur mõju laenu pankrotist taastumisele on just sellel, et kui palju on inimene sinna juba sisse maksnud. Olulised olin ka mõned Bondora poolt arvutatud muutujad, mis ilmselt näitavad inimese riskitaset. Oluline oli ka see, et kui kaugel on laenu tagasimaksmise lõpp – kui maksta pole enam palju, siis see mõjub taastumistõenäosusele positiivselt. Siit leitud muutujate tähtsushierarhia on seega kooskõlas ka peatükis 2.2.3 leitud otsustuspuu meetodil leituga.

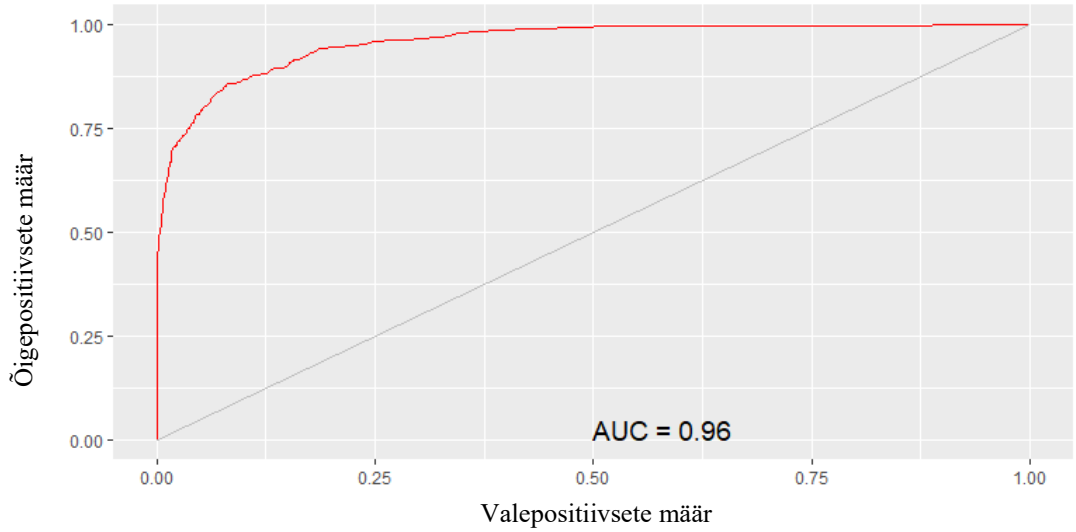
Järgnevalt tasub uurida, kas otsustuspuude metsa loomine annab paremad ennustustulemused kui üksiku puu tegemine, nagu viitab joonis 14, kus mudeli veamäär puude lisandudes väheneb. Selleks on autor koostanud alljärgneva tabeli 8.

Tabel 8. Otsustuspuude metsa ennustustulemuste paigutumine maatriksisse.

		Ennustatud		
		Ei taastu pankrotist	**Taastub pankrotist**	**Viga klassi ennustamisel**
Tegelik	**Ei taastu pankrotist**	1276	57	4,3%
	Taastub pankrotist	177	592	23%
Mudeli veamäär		11,1%		
Keskmine klassiviga		13,65%		
Viga investeerimisel		8,8%		

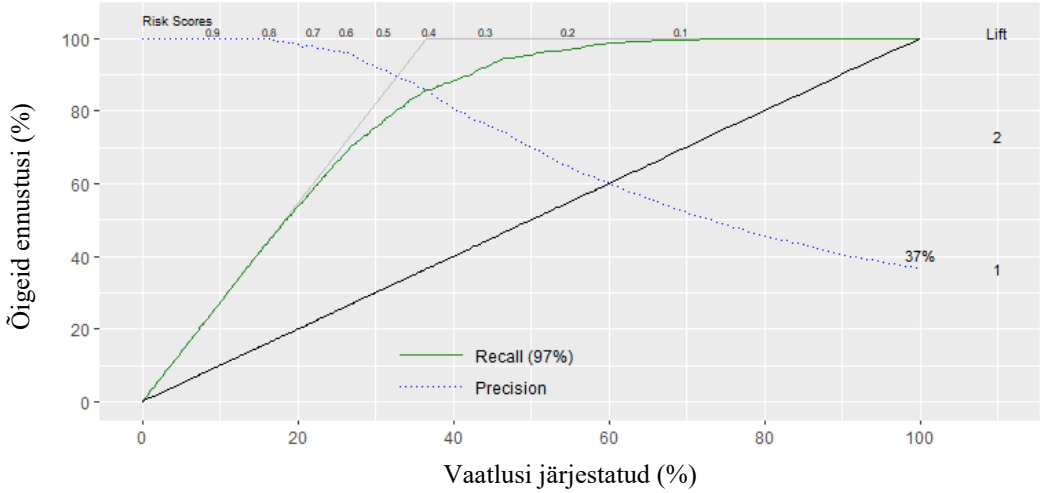
Allikas: autori arvutused statistikapaketis R.

Tabelist 8 on näha, et mudeli tulemused on üksiku otsustuspuuga võrreldes oluliselt paranenud, kuid siiski eksib mudel rohkem kui närvivõrk või tugivektormasin. See tulemus ei ole kooskõlas Kisandiga (2015), kes kasutab portfelli optimeerimismeetodina just otsustuspuude metsa. Kindlasti tuleb siinkohal ka ära märkida, et tegemist ei ole sama ennustusülesandega ja ka erinevusi on ka andmestikus. Kui võtta arvesse peatükis 2.2 alguses seatud eesmärki mahtuda investeerimisveaga 10% alla, siis saab mudelit siiski küllalt täpseks lugeda. Mudeli täpsust uuris autor graafiliselt alljärgneval joonisel 15.



Joonis 15. Otsustuspuude metsa ROC-kõver. Allikas: autori arvutused statistikapaketis R.

Nagu on näha jooniselt 15, siis on mudel graafiliselt uurides küllaltki hea. Samas ei ole aga näha nii suurt täpsust kui närvivõrgu või tugivektormasina korral. Mudel tunneb oskab 96% tõenäosusega hinnata suvaliselt valitud pankrotist taastuv laen suvaliselt valitud taastumata laenust suurema tõenäosusega taastuvaks või teisisõnu, mudel oskab neil 96 korral sajast vahet teha.



Joonis 16. Otsustuspuude metsa riskigraafik. Allikas: autori arvutused statistikapaketis R.

Ka joonisel 16 kujutatud otsustuspuude metsa riskigraafikult on näha, et mudel on üksikust otsustuspuust parem. Kui analüüsida joonisel sama punkti nagu üksiku

otsustuspuu korral, siis võrdluses otsustuspuuga on mets selleks hetkeks leidnud rohkem kui 5% enam pankrotist taastuvaid vaatlusi ja ka lift-kordaja on kõrgem

30% vaatluste reastamise järel ei ole otsustuspuu mudel kuigi täpne ja tehtud on küllaltki palju vigu. Mudel on selleks hetkeks üles leidnud vaid veidi alla 70% kõigist mittetaastuvatest laenudest, kuid lift-kordaja näitab sellel hetkel, et mudel on huupi arvavast mudelist üle kahe korra parem. Veamäär on selles punktis peaaegu üle 10 protsendi madala. See näitab, et mudel ei ole pankrotist taastumise etteennustamisel üksikust otsustuspuust oluliselt parema ennustusjõuga.

Järgnevalt uurib autor vigadest õppivat ansambelmeetodit, ehk võimendamist.

2.2.5. Võimendamine

Mõeldes peatükis 1.2.5 kirjeldatud võimendamise nõrkusele, milleks on müra andmestikus ja tõsiasjale, et autor on vähemasti proovinud peatükis 2.1.1 müra vähendada, on autori arvates võimendamise meetodil suur potentsiaal õppida oma vigadest ja tõusta käesoleva juhtumiuuringu parimaks ennustusmeetodiks. Esmaseks analüüsiks suhtarvude osas on autor tulemused kandnud alljärgnevasse tabelisse 9.

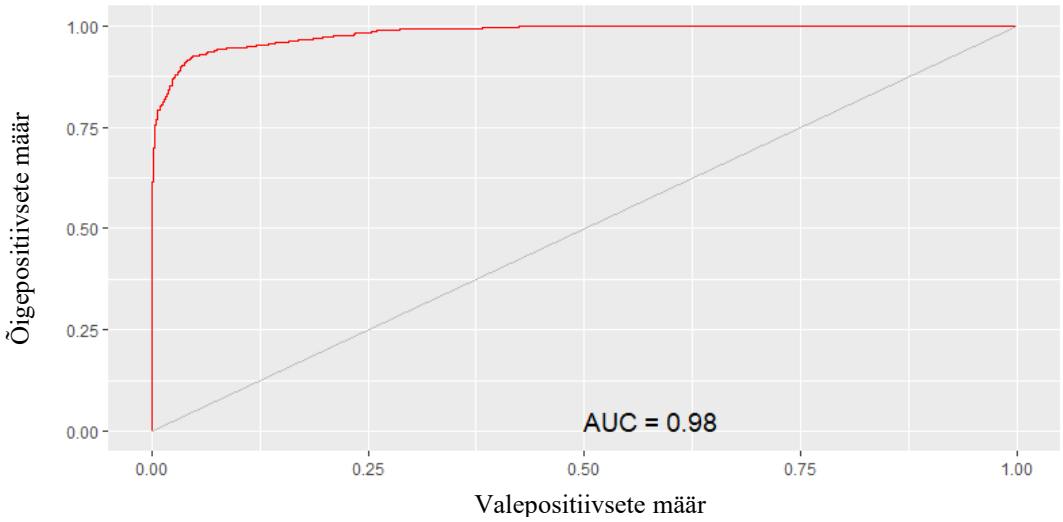
Tabel 9. Võimendamise ennustustulemuste paigutumine maatriksisse.

		Ennustatud		
		Ei taastu pankrotist	**Taastub pankrotist**	**Viga klassi ennustamisel**
Tegelik	**Ei taastu pankrotist**	1295	38	2,9%
	Taastub pankrotist	93	676	12,1%
Mudeli veamäär		6,2%		
Keskmine klassiviga		7,5%		
Viga investeerimisel		5,3%		

Allikas: autori arvutused statistikapaketis R.

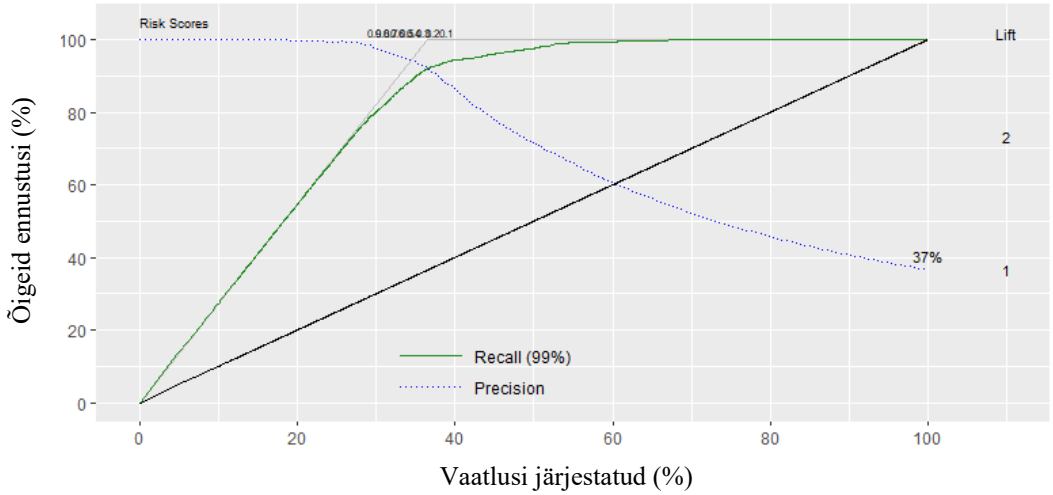
Kui investor seda mudelit investeerimisel kasutaks ja investeeriks vaid siis, kui mudel ennustab pankrotist taastumist, on mudeli eksimismääraks vaid 5,3%. See on väga madal eksimismäär. Mudel on väga täpne ka mittetaastuvate laenude etteennustamisel, eksides vaid 2,9% juhtudest. Nii mudeli veamäär kui ka keskmine klassiviga on seni vaadeldud mudelitest kõige madalamad. Selleks, et uurida, kas tabelis 9 näidatud heade

suhtarvude poolt vihjatav mudeli kvaliteet ka graafiliselt paika peab, et on autor koostanud alljärgneva joonise 17.



Joonis 17. Võimendamise ROC-kõver. Allikas: autori arvutused statistikapaketis R.

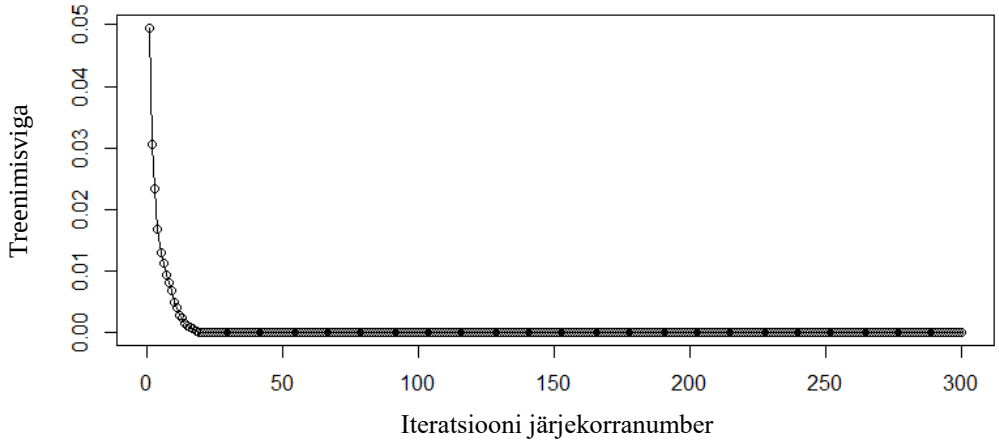
Vaadates joonist 17 on näha, et mudel on ka graafiliselt seni analüüsitutest parim, sest ulatub kõige kaugemale vasakusse ülemisse nurka. Ka mudeli AUC (0,98) on väga kõrge, mis näitab, et mudelil on väga hea diskrimineerimisvõime. Järgnevalt analüüsib autor võimendamise teel saadud mudelit riskigraafikul, mis on kujutatud alljärgneval joonisel 18.



Joonis 18. Võimendamise riskigraafik. Allikas: autori arvutused statistikapaketis R.

Uurides joonisel 18 kujutatud riskigraafikut sarnaselt eelnevalt analüüsitud meetoditega kohas, kus reastatud on 30% vaatlustest ennustuse tugevuse alusel, on võimendamine

leidnud üles 80% kõigist taastuvatest laenudest ning mudeli veamäär on selles punktis vaid 3%. Ka lift-graafiku järgi on mudel selles kohas eelnevalt hinnatud mudelitest parem, olles üle 2,5 korra suurema ennustusjõuga kui triviaalne mudel. Kuna tegemist on vigadest õppiva meetodiga, siis on siin oluline uurida, kui mitu korda peab mudel oma vigadest õppima, et saavutada optimaalne täpsus. Selle uurimiseks on autor koostanud alljärgneva joonise 19.



Joonis 19. Võimendamise veamäära vähenemine iteratsioonide lisandudes. Allikas: autori arvutused statistikapaketis R.

Jooniselt 19 on näha, et mudel õpib oma vigadest küllaltki kiiresti. Autor kordas õppimisprotsessi 300 korda, kuid samaväärse tulemuse saavutamiseks oleks piisanud ka umbes 25 korrast. Nähes, et mudeli veamäär enam ei vähene, võib antud mudeliga rahul olla. Sarnaselt peatükis 2.2.4 otsustuspuude metsa puhul koostatud tähtsaimate muutujate hierarhiale saab selle hierarhia koostada ka võimendamise puhul (vt lisa 3). Mudel andis kümneks tähtsaimaks muutujaks:

1. Eelnevate tagasimaksete summa
2. Intressimaksete summa
3. Planeeritud pankrotijärgne intress (Bondora arvutatud)
4. Laenu põhiosa jääk
5. Intressimaksete jääk
6. Oodatav kahju pankroti korral (Bondora arvutatud)
7. Graafikujärgne oodatav põhiosa tagasimaksete summa hetkel
8. Laenu pikkus
9. Laenuvõtja riik = Hispaania
10. Planeeritud pankrotijärgne põhiosa (Bondora arvutatud)

Muutujate hierarhia on seega väga sarnane otsustuspuude metsa poolt antud hierarhiaga, kuid siiski on siin mõned erisused. Võimendamine toob välja, et oluline on kõige muu kõrval ka näiteks see, kui laenuvõtja riik on Hispaania, tähtis on ka laenu pikkus, võlgade ja sissetuleku suhe ja eelnevate laenude arv. Autori arvates on võimendamise poolt antud muutujate hierarhia loogiline ja on kooskõlas Bondora poolt välja antud statistikaga Hispaania laenude halvemate taastumismäärade kohta ja teiste muutujate osas on ka võimalik lihtsasti leida põhjendusi, miks need muutujad mudelisse sobivad.

Viimasena analüüsib autor logistilist regressiooni laenude pankrotist taastumise ennustamisel.

2.2.6. Logistiline regressioon

Võrreldavuse huvides on kindlasti vajalik näidata ka logistilise regressiooni ennustusvõimet, et oleks tugev võrdlusbaas teistele mudelitele. Võttes arvesse, et logistiline regressioon ongi loodud binaarse tunnuse klassifitseerimiseks, siis on täiesti võimalik, et see meetod sobib pankrotistunud laenude taastumise ennustamiseks väga hästi.

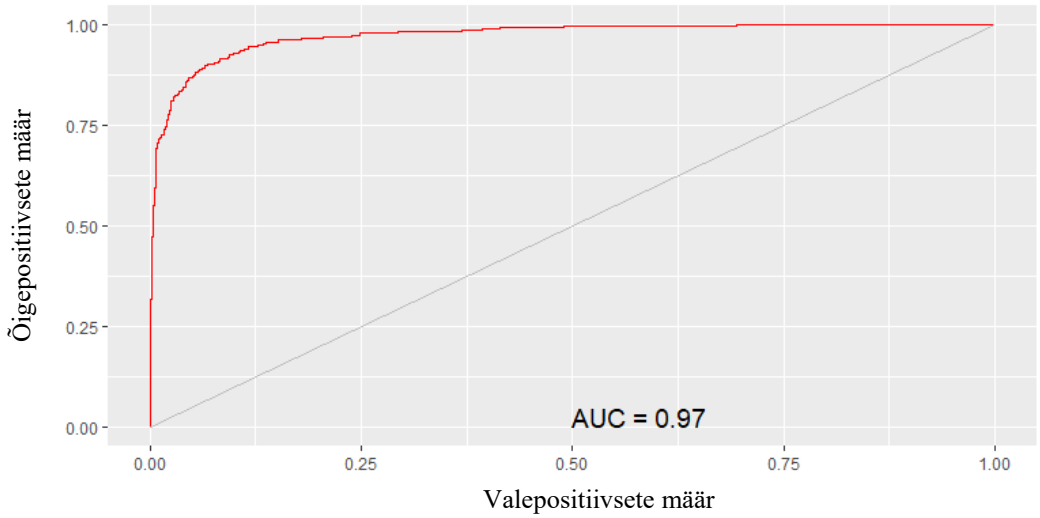
Tabel 10. Logistilise regressiooni ennustustulemuste paigutumine maatriksisse.

		Ennustatud		
		Ei taastu pankrotist	**Taastub pankrotist**	**Viga klassi ennustamisel**
Tegelik	**Ei taastu pankrotist**	1274	59	4,4%
	Taastub pankrotist	107	662	13,9%
Mudeli veamäär		7,9%		
Keskmine klassiviga		9,15%		
Viga investeerimisel		8,2%		

Allikas: autori arvutused statistikapaketis R.

Nagu on tabelist 10 näha, siis paigutuvad logistilise regressiooni ennustustulemused sarnaselt paremate ennustusmeetoditega (närvivõrk, tugivektormasin ja võimendamine). Viga, millega investor peaks arvestama, on 8,2%, ehk siis, kui mudel ennustab pankrotistunud laenu taastumist, võib see 8,2% tõenäosusega valeks osutuda. Üldiselt on mudeli veanäitajad küllaltki head – üldine veamäär jääb alla 8 protsendi ja ka keskmine klassiviga on küllaltki madal. Nii, nagu ka teiste meetodite puhul, uurib autor

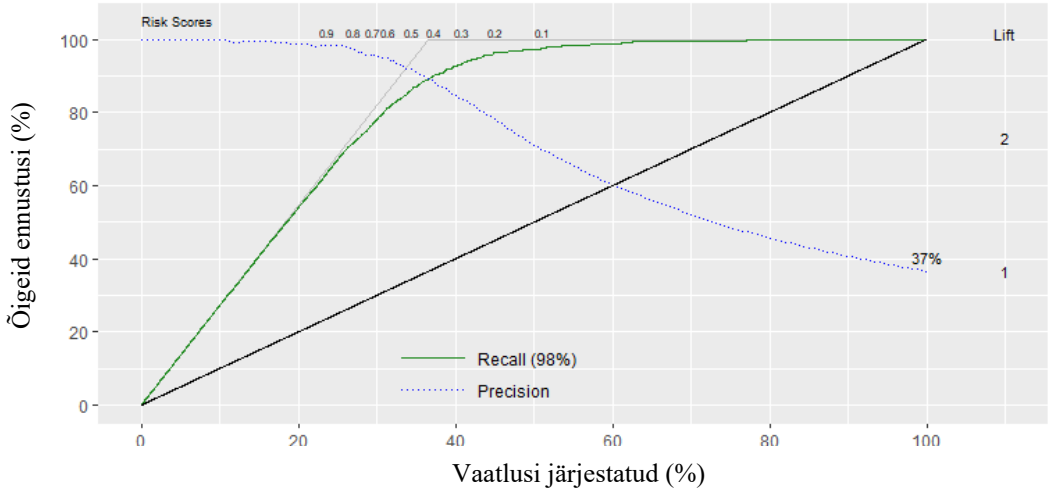
ka logistilise regressiooni ennustusvõimet graafiliselt. Logistilise regressiooni ROC-kõver on esitatud alljärgneval joonisel 20.



Joonis 20. Logistilise regressiooni ROC-kõver. Allikas: autori arvutused statistikapaketis R.

Ka joonisel 20 näeme juba eelpool analüüsitud paremate ennustusmeetoditega sarnast pilti, kus ROC-kõver ulatub küllaltki kaugele vasakusse ülemisse nurka ja mudeli diskrimineerimisvõime on küllaltki kõrge. AUC (0,97) näitab, et 97% tõenäosusega suudab mudel tegelikult taastuva laenu taastumistõenäosust kõrgemalt hinnata, kui mittetaastuva laenu oma.

Kuna riskikartlik investor ei investeeriks suvalistesse pankrotistunud laenudesse, mille taastumist mudel ette ennustab, vaid laenudesse, mille puhul ennustav mudel annab suurima tõenäosuse taastumiseks, on oluline vaadata, kui hästi suudab mudel klassifitseerida siis, kui alustada selle tehtud tugevamatest hinnangutest. Selleks on autor koostanud alljärgneva joonise 21.



Joonis 21. Logistilise regressiooni riskigraafik. Allikas: autori arvutused statistikapaketis R.

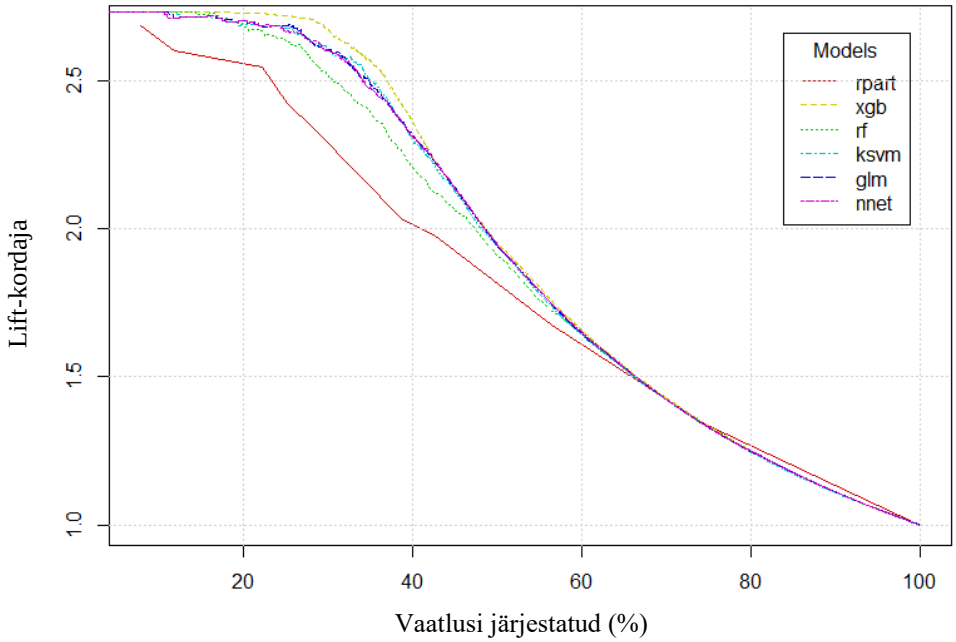
Joonisel 21 on x-teljel järjestatud logistilise regressiooni poolt tehtud ennustused ennustuse tugevuse järgi nii, et suurema taastumistõenäosuse saanud laenud on joonisel vasakul pool ja väiksema ennustustõenäosuse saanud laenud on paremal pool. Uurides punkti, kus järjestatud on 30% kõigist vaatlustest (caseload = 30%), saame uurida mudeli täpsusnäitajaid selles punktis. Kui 30% vaatlustest on järjestatud, on mudel üles leidnud 78% taastuvatest laenudest. Mudeli veamäär on seejuures vaid umbes 5% ja lift-kordaja on samuti küllaltki kõrge. Tuleb välja, et kas logistiline regressioon sobib väga hästi pankrotist taastumise ennustamiseks.

Peatükis 2.2 tehtud analüüsi võtab autor kokku peatükis 2.3, mis annab kokkuvõtlikuma ülevaate mudelite erinevustest ja uuritavatest täpsusnäitajatest ning teeb saadud informatsioonist järeldused.

2.3. Järeldused ja tulemuste analüüs

Käesolev peatükk koondab kokku kuue ennustusmeetodi tulemused ja täpsusnäitajad pankrotist taastumise ennustamisel. Graafiliselt esitatakse tulemused kahel joonisel. Esmalt vaadeldakse kõiki mudeleid ühisel lift-graafikul ja seejärel kõiki mudeleid täpsusgraafikul. Analüüsi lõpetab kokkuvõttev tabel ja järeldused leitud tulemuste kohta.

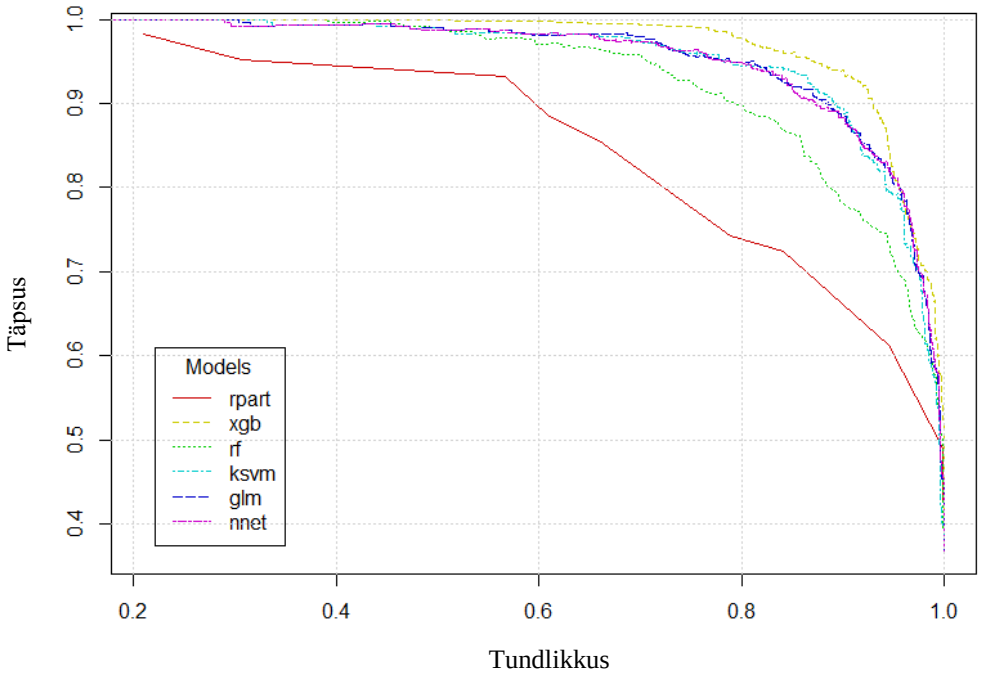
Selleks, et näidata kõiki mudeleid ühisel lift-graafikul, on autor koostanud alljärgneva joonise 22.



Joonis 22. Kõik vaatlusalused mudelid lift-graafikul. Allikas: autori arvutused statistikapaketis R.

Joonisel 22 kujutatud lift-graafik ilmestab seda, kui kaua vaatlusi suudab ennustuse tugevuse järgi vaatlusi reastades iga mudel säilitada oma eelist (lift) triviaalse või teisisõnu huupi arvava mudeli ees. Väga lihtsalt öeldes on mudel parem, kui selle joon ulatub teistest rohkem üles ja paremale. Jooniselt on näha, et kollane katkendlik joon (xgb), mis tähistab võimendamist, püsib kõige kauem kõrgel ja hakkab langema alles siis, kui reastatud on natuke alla 30% vaatlustest. Võrdluseks lift-graafiku järgi halvima mudeli, pideva punase joonega (rpart) tähistatud otsustuspuu joon hakkab langema koheselt ja langeb teistest oluliselt kiiremini ja ei ulatu nii kaugele vasakule ega üles kui teiste mudelite jooned. Nagu oli ka näha peatükis 2.2 läbiviidud analüüsist, on närvivõrgu, tugivektormasina ja logistilise regressiooni ennustusvõime enam-vähem võrdselt hea ning otsustuspuude mets näitab ka siin joonisel teistest mudelitest pisut halvemat tulemust. Oluline on ka märkida, et kui reastatud on umbes 70% vaatlustest, on kõik mudelid enam-vähem sama lift-väärtusega ja pole enam triviaalsest mudelist kuigi palju paremad.

Järgnevalt uurib autor täpsusgraafikut, mis paneb omavahel vastavusse õigepositiivsete ennustuste arvu suhe õigete ennustuste koguarvu (precision, y-teljel) ja õigepositiivsete ennustuste määra (recall, x-teljel). Graafik on kujutatud joonisel 23.



Joonis 23. Kõik vaatlusalused mudelid täpsusgraafikul. Allikas: autori arvutused statistikapaketis R.

Joonis 23 näitab käesoleva magistritöö fookuses oleva juhtumiuuringu jaoks väga olulist omadust. See näitab mudeli võimet teha õigepositiivseid ennustusi, säilitades seejuures kõrge täpsus ja vältida valepositiivseid ennustusi. Siin on oluline, et mudeli joon püsiks võimalikult kaua kõrgel ja jõuaks paremale serva võimalikult vara. Sarnaselt joonisele 22 tuleb ka siit välja, et võimendamist väljendab kollane katkendlik joon (xgb) ilmestab parimat mudelit. Eriti tugevalt tuleb siin joonisel välja otsustuspuu meetodi nõrkus valepositiivsete ennustuste vältimisel. Ka siin joonisel koonduvad üksteise lähedale kokku närvivõrk, tugivektormasin ja logistiline regressioon, millest jääb natukene maha otsustuspuude mets.

Mida selle graafiku tulemus tähendab investori jaoks on see, et võimendamise abil saadud mudelit saab kasutada suurema koguse raha investeerimiseks, sest selle ennustusvõime püsib kõige kauem kõrge ja see väldib kõige edukamalt valepositiivseid ennustusi, mis investorile kahju tooks.

Järgnevalt koondab autor juhtumiuuringu tulemused alljärgnevasse koondtabelisse 11.

Tabel 11. Pankrotist taastumise ennustamiseks kasutatavate meetodite võrdlus.

Ennustusmeetod	Viga investeerimisel	Üldine veamäär	Lift kordaja säilimine	Võime vältida valepositiivseid ennustusi
Närvivõrk	7,8%	8,2%	Hea	Hea
Tugivektormasin	6,6%	7,4%	Hea	Hea
Otsustuspuu	14,5%	16,6%	Halb	Väga halb
Otsustuspuude mets	8,8%	11,1%	Keskmine	Keskmine
Võimendamine	5,3%	6,2%	Väga hea	Väga hea
Logistiline regressioon	8,2%	7,9%	Hea	Hea

Allikas: autori koostatud.

Tuginedes peatüki 2.2 alguses toodud „küllaltki hea täpsuse“ määratlusele, tuleb märkida, et viis mudelit kuuest osutusid küllaltki heaks. Kui investor lepiks investeeringute tegemisel 10-protsendilise veamääraga, sobiks kasutusele võtta seega nii närvivõrk, tugivektormasin, otsustuspuude mets, võimendamine kui ka logistiline regressioon (vt tabel 11). Kui aga investorit huvitaks pigem lähtuda sellest, et mudeli üldine veamäär ei tohi olla rohkem kui 10%, siis otsustuspuude mets enam kasutamiseks ei sobiks. Ükskõik, kumba pidi ka vaadata, on parimaks ennustusmeetodiks kasutatud andmestiku abil pankrottidest taastumise etteennustamisel võimendamine, mis suutis investeerimisel saavutada 94,7% täpsuse ja üldise veamäära alusel 93,8% täpsuse. Samal ajal kui otsustuspuu meetod säilitas nii oma lift-kordajat kui ka võimet vältida valepositiivseid ennustusi üsna halvasti, suutsid nii närvivõrk, tugivektormasin ja logistiline regressioon ülesannetega hästi hakkama saada. Otsustuspuude mets oli nendest kolmest meetodist pisut halvem igas vaadeldavas aspektis. Kindlaks võitjaks osutus siiski suurimat täpsust näidanud, kõige kauem lift-kordajat säilitanud ja kõige paremini valepositiivseid ennustusi vältinud võimendamine.

Tabelis 11 näidatud tulemused on kooskõlas Haltufi (2014) leituga, sest vähemasti pankrotist taastumise ennustamisel (lisaks Haltufi leitule) on tugivektormasin logistilisest regressioonist märgatavalt tugevam. Kuna käesolevas töö on suudetud tõestada tugivektormasinate efektiivsust krediidiotsuste tegemisel, ei ole tulemus kooskõlas Akbani et al (2004) väitega, et tugivektormasinad on krediidiotsuste tegemisel pigem ebasobivad. Autori arvates on siin põhjuseks see, et analüüsis

kasutatud andmestik oli küllaltki tasakaalus ja see ei põhjustanud probleeme. Kooskõla on näha ka Kisandiga (2015), sest otsustuspuude mets on ka käesoleva töö käigus läbiviidud juhtumiuuringu kontekstis oluliselt parem ennustaja kui üksik puu. Autori arvates võib ka öelda, et tulemused on samuti kooskõlas Sammelsaare (2016) leituga, sest logistiline regressioon edestab ennustusvõimelt otsustuspuud. Kõikidele kooskõladele vaatamata on aga käesoleva töö autori tulemused küllaltki erinevad eelnevalt leitust, sest parimaks mudeliks osutus võimendamine, mida pole autorile teadaolevalt veel keegi teaduslikus kontekstis Bondora andmetele rakendanud. Samuti on autor leidnud, et närvivõrkude kasutamine krediidiskooringu kontekstis on teiste meetoditega konkurentsivõimeline lahendus – eriti selle pärast, et see edestas investeerimisvea järgi isegi standardiks peetavat logistilist regressiooni.

Kõige paremini on tabelis 11 näidatud tulemused kooskõlas Bastosega (2008), kes suutis austraalia krediidiandmestikul sarnaselt käesoleva töö autoriga näidata nii seda, et võimendamine on parim meetod krediidiostuste tegemisel kui ka seda, et närvivõrk ja tugivektormasin jäävad sellele täpsuselt natukene alla ja on omavahel üpris sarnase ennustusjõuga. Kooskõlas on ka Bastose leitud võimendamise täpsusprotsent (94,03%) ja käesoleva töö autori leitud võimendamise täpsusprotsent (93,8%), mis näitab, et kahel erineval andmestikul on võimalik krediidiotsuste tegemisel saavutada sarnane täpsuse tase. Kooskõla on ka Lopesega (2016), sest ka käesolevas töös osutus taastumisproblemaatikat kõige paremini lahendama just võimendamine. Lessmanniga (2013) ei ole käesoleva töö tulemused kooskõlas, sest käesolevas töös osutus võimendamine parimaks ansambelmeetodiks ja tugivektormasin parimaks üksikuks klassifitseerijaks, kuid see on vägagi andmetest sõltuv määratlus.

Käesoleva töö autori arvates võib juhtumiuuringu tulemusi positiivseteks lugeda, sest on näidatud, et pankrotist taastumist on võimalik ette ennustada inimesele omaste näitajate ja eelneva maksekäitumise põhjal ja seda on võimalik teha küllaltki täpselt. Töö tulemusi kinnitavad ka eelnevate autorite tööd, kes on tunnistanud keerukate ennustusmeetodite kasulikkust krediidiotsuste tegemisel.

KOKKUVÕTE

Käesolev magistritöö käsitles andmeteaduses tuntud statistilisi ennustusmeetodeid ühisrahastusvõimalusse investeerimise kontekstis. Autor viis läbi juhtumiuuringu, mille käigus analüüsiti inimeselt-inimesele laenukeskkonna Bondora poolt väljastatavat andmestikku, et testida pankrottidesse investeerimise strateegia võimalikkust. Juhtumiuuring sisaldas endas andmete töötlemist ja närvivõrgu, tugivektormasina, otsustuspuu, otsustuspuude metsa, võimendamise ja logistilise regressiooni kasutamist laenude pankrotist taastumise etteennustamisel.

Autor andis ülevaate Eesti investeerimismaastiku hetkeolukorrast, keskendudes eriti just ühisrahastusportaalidele kui uusima investeeringute liigi pakkujatele. Tuli välja, et väga suuresti erinevad nii portaalide pakutavad investeerimistooted kui ka nende poolt pakutavad aastatootlused investoritele. Samas sai aga selgeks, et statistilisi ennustusmeetodeid ei ole võimalik kõigil Eestis turul olevatel ühisrahastusportaalidel investeeringute tegemiseks kasutada, sest sugugi kõik portaalid ei paku andmebaasi mineviku tehingute kohta. Analüüsi jaoks sobivaks portaaliks oligi just Bondora, sest nende pakutav LoanData andmestik sisaldab küllaltki palju vaatlusi ja on muutujate poolest sobilik analüüsi läbiviimiseks.

Kuna autor sai kasutada inimese maksekäitumist väljendavaid muutujaid, osutus pankrotist taastumise etteennustamine vägagi võimalikuks. Tähtsaks ei osutunud mitte ainult see, kui palju on laenult esinenud põhiosa- ja intressimakseid, vaid ka näiteks laenu jääk, osutades autori arvates psühholoogilistele faktoritele nagu pöördumatute kulude efekt ja „peaaegu finišis olemise“ tunne. Veel pidasid autori poolt loodud mudelid olulisteks selliseid laenutegureid nagu intressimäär, laenu pikkus ja laenuvõtja päritolumaa.

Esmalt viis autor analüüsi läbi kasutades meetodina närvivõrku, mis on üldiselt sobiv meetod keerukate protsesside etteennustamiseks. Kasutades Back-Propagation algoritmi

mudeli treenimiseks, suutis närvivõrk näidata 8,2% veamäära ja suhteliselt head võimet vältida valepositiivseid ennustusi. Seejärel näitas autor ära, et kuna LoanData andmestikus oli negatiivsete vaatluste osakaal piisavalt suur, suutis väga head täpsust näidata ka tugivektormasin, mille veamääraks oli 7,4% ja mille võime vältida valepositiivseid ennustusi oli samuti kõrge. 16,6% veamäära näitas autori poolt loodud üksik otsustuspuu, mida oli lihtsama tõlgendamise huvides piiratud. Luues aga otsustuspuudest 300-puune mets, näitas säärane meetod 11,1% veamäära. Võttes järjest otsustuspuid ja pannes neid oma vigadest õppima, suutis autor võimendamise läbi näidata vaid 6,2% veamäära ja kõige madalamat valepositiivsete ennustuste määra. Krediidiskooringul standardina kasutatav logistiline regressioon näitas aga seejuures 7,9% veamäära. Kokkuvõttes suutis vaid võimendamine näidata logistilisest regressioonist oluliselt paremat ennustustäpsust.

Kõige täpsemaks mudeliks laenude taastumise ennustamisel osutus seega võimendamine, millele järgnesid tihedasti koos logistiline regressioon, närvivõrk ja tugivektormasin ning nendele järgnes otsustuspuude mets, jättes kõige viimasele kohale üksiku otsustuspuu. Selline mudelite täpsuse hierarhia oli oodatud ja tulemused olid kooskõlas mitmete eelnevate uurijatega. Kuna võimendamine suutis näidata suurt täpsust ja allahindluste korral pole isegi nii suurt täpsust pankrottidesse investeerides kasumi teenimiseks vaja, võib läbiviidud juhtumiuuringu tulemusi positiivseteks lugeda.

Käesolev magistritöö tähendab riskialdi investori jaoks võimalust investeerida potentsiaalselt väga suure tootlusega, kuid samas ka suure riskiga. Peamiselt seisneb riski ajahorisondis, sest käesolevas töös ei ole tehnilistel põhjustel arvesse võetud taastumiseks kuluvat aega. Seega saab väita, et pankrottidest taastumise etteennustamine ja sedasi investeerimine on võimalik, aga kahjuks ei ole võimalik arvutada sedasi investeerides oodatavat tootlust.

Tööd oleks võimalik edasi arendada proovides veelgi ennustamisel kasutada veelgi keerukamaid ennustusmeetodeid, et investeeringu tulemust paremini ennustada. Kuna ansambelmeetodid osutusid ennustuste tegemisel väga headeks ennustajateks, tasuks korrata ka Lessmann et al (2013) metoodikat ülevaatliku meetodite võrdluse koostamisel. Tööd on võimalik ka edasi arendada koostades parimatest meetoditest veel omakorda ansambel.

VIIDATUD ALLIKAD

1. **Akbani, R., Kwek, S., Japkowicz, N.** Applying Support Vector Machines to Imbalanced Datasets. 2004. [https://link.springer.com/content/pdf/10.1007%2F978-3-540-30115-8_7.pdf] 23.05.2018

2. **Arandi, I.** Masinõppe meetoditel põhinevate tehingute pettuste tuvastamise algoritmide välja töötamine ja testimine. Tallinna Tehnikaülikool. 2017. [https://digi.lib.ttu.ee/i/file.php?DLID=8881&t=1] 23.05.2018

3. **Bachmann, A., Becker, A., Buerckner, D., Hilker, M., Kock, F., Lehmann, M., Tiburtius, P., Funk, B.** Online peer-to-peer lending - A literature review. - Journal of Internet Banking and Commerce. Vol 16. 2011. [https://www.researchgate.net/publication/236735575_Online_Peer-to-Peer_Lending--A_Literature] 25.10.2017

4. **Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., Vanthienen, J.** Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society. 2003. [http://booksc.org/dl/18405730/d604a9] 25.10.2017

5. **Bastos, J.** Credit scoring with boosted decision trees. Lissaboni Tehnikaülikool. 2008. [https://raider.io/characters/eu/twisting-nether/Shanoa] 23.05.2018

6. **Bellotti, T., and Crook, J.** Support vector machines for credit scoring and discovery of significant features. Expert Systems with Applications. 3302–3308. 2009. [http://www.research.ed.ac.uk/portal/en/publications/support-vector-machines-for-credit-scoring-and-discovery-of-significant-features(c407098f-9bf3-4077-b546-9af603485c12).html] 25.10.2017

7. **Berger, S., Gleisner, F.** Emergence of financial intermediaries in electronic markets: The case of online p2p lending. BuR Business Research Journal. 2009. [https://link.springer.com/article/10.1007/BF03343528] 25.10.2017

8. Bondora. Statistika. 2018. [https://www.bondora.com/et/public-statistics] 23.05.2018

9. **Breiman, L.** Random forests. - Machine learning. Vol 45. 2001. [https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf] 25.10.2017

10. **Brownlee, J.** A Tour of Machine Learning Algorithms. 2013. [https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/] 23.05.2018

11. **Cortes, C., Vapnik, V.** Support-vector networks. Machine learning. 1995. [https://link.springer.com/article/10.1023/A:1022627411411] 25.10.2017

12. Crowdestate. Koduleht. 2018. [https://crowdestate.eu/et/home] 23.05.2018

13. **Elizondo, D.** The linear separability problem: Some testing methods. Neural Networks, IEEE Transactions. 2006. [http://booksc.org/dl/34421400/554bff] 25.10.2017

14. **Engelmann, B., Hayden, E., Tasche, D.** Measuring the discriminative power of rating systems. Tech. rep., Discussion paper, Series 2: Banking and Financial Supervision, 2003. [https://www.bundesbank.de/Redaktion/EN/Downloads/Publications/Discussion_Paper_2/2003/2003_10_01_dkp_01.pdf?__blob=publicationFile] 25.10.2017

15. Estateguru. Statistika. 2018. [https://estateguru.co/home/statistics] 23.05.2018

16. Finantsinspektsioon. Ühisrahastamine. 2018. [https://digi.lib.ttu.ee/i/file.php?DLID=7350&t=1] 04.02.2018

17. **Haltuf, M.** Support Vector Machines for Credit Scoring. University of Economics in Prague. 2014. [http://www.quantitative.cz/data/files/support-vector-machines-for-credit-scoring-michal-haltuf-2014.pdf] 25.10.2017

18. **Hand, D. J.** Evaluating diagnostic tests: the area under the roc curve and the balance of errors. Statistics in Medicine. 2010. [http://onlinelibrary.wiley.com/doi/10.1002/sim.3859/full] 25.10.2017

19. **Hosmer, D. W., Lemeshow, S., Sturdivant, R. X.** Introduction to the Logistic Regression Model. Applied Logistic Regression, Third Edition. 2013. [http://booksc.org/dl/22192871/0bd42e] 23.05.2018

20. **Hsu, C.-W., Chang, C.-C., Lin, C.-J.** A practical guide to support vector classification, 2016. [http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf] 25.10.2017

21. **Huang, C.-L., Chen, M.-C., Wang, C.-J.** Credit scoring with a data mining approach based on support vector machines. Expert Systems with Applications. 2007. [http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.330.1474&rep=rep1&type=pdf] 25.10.2017

22. **Huang, J., Ling, C. X.** Using AUC and Accuracy in Evaluating Learning Algorithms. The University of Western Ontario. Department of Computer Science. 2003. [http://home.cse.ust.hk/~qyang/Teaching/537/Papers/AUC-evaluation.pdf] 25.10.2017

23. Investly. Investeeri. 2018. [https://www.investly.co/invest] 23.05.2018

24. **Kangas, R.** Online peer to peer lending: clustering borrowers using self-organizing maps. Lappeenranta Tehnikaülikool. 2014. [https://www.doria.fi/bitstream/handle/10024/99399/Online%20P2P%20Lending.pdf?sequence=2] 23.05.2018

25. **Kisand, M-L.** Optimaalse tootlusega laenuportfelli koostamine eraisikust investorile - otsustuspuude meetodite rakendamine inimeselt-inimesele laenukeskkonnas Bondora. Tallinna Tehnikaülikool. 2015. [https://digi.lib.ttu.ee/i/file.php?DLID=3314&t=1] 04.02.2018

26. **Käärman, K.** Otsustuspuudega klassifitseerimine. Arvutiteaduse instituut, Tartu Ülikool. [http://www.quretec.com/u/vilo/edu/2003-04/DM_seminar_2003_II/Raport/P02/main.pdf] 23.05.2018

27. **LeCun, Y., Bengio, Y., Hinton, G.** Deep learning. Nature 521. 2015. [https://www.nature.com/articles/nature14539] 23.05.2018

28. **Lessmann, S., Seowb, H.-V., Baesens, B., Thomas, L. C.** Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update. In Credit Research Centre, Conference Archive. 2013. [https://lirias.kuleuven.be/bitstream/123456789/497988/1/Credit+scoring+benchmark+-+R1+-+2015-02-23.pdf] 25.10.2017

29. **Lints, T.** Tehis- vs. bioloogilised närvivõrgud. Tallinna Tehnikaülikool. 2004. [http://taivo.net/downloads/TL_TehisVsBiolANN.pdf] 23.05.2018

30. **Lopes, R. G., Carvalho, R. N., Ladeira, M., Carvalho, R. S.** Predicting Recovery of Credit Operations on a Brazilian Bank. 2016. [https://www.h2o.ai/wp-content/uploads/2017/01/Predicting-Recovery-of-Credit-Operations-on-a-Brazilian-Bank.pdf] 23.05.2018

31. **Madalvee, H.** Investeerimisvõimaluste analüüs erinevate varaklasside lõikes kogumisel pensioniks. 2016. [https://digi.lib.ttu.ee/i/file.php?DLID=7350&t=1] 23.05.2018

32. **McCue, T., Keller, C.** Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis, 2nd Edition. 2007. [https://doc.lagout.org/Others/Data%20Mining/Data%20Mining%20and%20Predictive%20Analysis_%20Intelligence%20Gathering%20and%20Crime%20Analysis%20%5BMcCue%202007-05-01%5D.pdf] 23.05.2018

33. **Mereste, U.** Majandusleksikon. Tallinn, Eesti Entsüklopeediakirjastus. 2003.

34. **Mild, A., Waitz, M., Wöckl, J.** How low can you go? — Overcoming the inability of lenders to set proper interest rates on unsecured peer-to-peer lending markets. 2013. [http://booksc.org/dl/39514869/a4409b] 23.05.2018

35. Mintos. Statistika. 2018. [https://www.mintos.com/en/statistics/] 23.05.2018

36. Omaraha. Statistika. 2018. [https://omaraha.ee/et/invest/stats/] 23.05.2018

37. **Polak, P.** Portfolio diversification on P2P loan markets. Charles University. 2017. [https://www.doria.fi/bitstream/handle/10024/99399/Online%20P2P%20Lending.pdf?sequence=2] 23.05.2018

38. **Polena, M.** Performance Analysis of Credit Scoring Models on lending Club Data. Charles University. 2017. [https://dspace.cuni.cz/bitstream/handle/20.500.11956/86490/DPTX_2016_1_11230_0_519394_0_185577.pdf?sequence=1] 23.05.2018

39. **Sammelsaar, L.** Pankrotistumise tõenäosuse prognoosimine otselaenamisettevõtte Bondora andmetel. Tartu Ülikool. 2016. [http://dspace.ut.ee/bitstream/handle/10062/52464/sammelsaar_lagle_msc_2016.pdf?sequence=1&isAllowed=y] 23.05.2018

40. **Zhou, Z-H.** Ensemble Methods – Foundations and Algorithms. CRC Press. 2012. [http://www2.islab.ntua.gr/attachments/article/86/Ensemble%20methods%20-%20Zhou.pdf] 04.02.2018

41. **Tint, H.** Sissejuhatus tugivektor-masinatesse. Arvutiteaduse instituut, Tartu Ülikool. 2003. [http://www.quretec.com/u/vilo/edu/2003-04/DM_seminar_2003_II/ver1/P09/main.pdf] 05.02.2018

42. Twino. Statistika. 2018. [https://www.twino.eu/en/statistics] 23.05.2018

43. **Williams, G.** Data mining with Rattle and R. The Art of Excavating data for Knowledge Discovery. 2011. [https://mineriaddatos.wikispaces.com/file/view/Data+Mining+With+Rattle+and+R_+The+Art+of+Excavating+Data+for+Knowledge+Discovery+-+Graham+Williams.pdf] 23.05.2018

44. **Witten, I., Frank, E., Hall, M.** Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science. 2011. [ftp://ftp.ingv.it/pub/manuela.sbarra/Data%20Mining%20Practical%20Machine%20Learning%20Tools%20and%20Techniques%20-%20WEKA.pdf] 25.10.2017

45. **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. Vol. 38. 2011. [http://isiarticles.com/bundles/Article/pre/pdf/22235.pdf] 25.10.2017

SUMMARY

VIABILITY OF INVESTING IN BANKRUP LOANS ON A PEER-TO-PEER LOAN MARKET SUCH AS BONDORA.EE

Ekke Sakkov

This master's thesis provides a look into predictive analytics in the context of investing into peer-to-peer consumer loans. A case study was conducted in order to analyze the dataset provided by Bondora, a leading peer-to-peer loan provider in Estonia, and to test the viability of investing in bankruptcies. The case study entailed data preprocessing and the creation of a neural network, a support vector machine, a decision tree, a random forest and the utilization of boosting and logistic regression as predictive methods for loan recovery.

An overview of the current investing climate in Estonia was provided, with particular regard towards different peer-to-peer solutions. It quickly became apparent that there are major differences between investment products and also between the expected return rates. It also became clear that predictive analytics do not apply to all of the peer-to-peer investment opportunities that are available to investors because of the prerequisite of a suitable dataset. Bondora was chosen as the suitable candidate due to it's comprehensive and large dataset that's also readily available on their home page.

Since the author had the opportunity to use variables that express a person's credit behavior, predicting loans that do actually recover did not turn out to be very difficult. The most important variables expressed what the author believes to be the "sunk cost effect", namely amount of previous repayments in that loan. It would seem that when someone has already made several repayments, they are very unlikely to walk away. Current principal balance also turned out to be a strong predictor because of what the author believes to be the "feeling of almost being at the finish line" and therefore not quitting. Other important variables included the interest rate, loan duration and country.

A neural network trained on the Back-Propagation algorithm turned out to be a good predictor with an error rate of 8,2%. Having a sufficiently high proportion of negative observations meant that the linear support vector machine was able to perform very well, showing an error rate of 7,4%. The author also trained a decision tree, which was limited in order to better allow explanation and provided an error rate of 16,6%. Training a random forest gave an error rate of 11,1%. Boosting trees to be better learners yielded an error rate of only 6,2%. The standard of the industry, logistic regression, managed an error rate of 7,9%. All in all, it became clear, that boosting provided more accurate predictions than logistic regression.

What these results mean for a risk-seeking investor is simply yet another possibility to invest money at a high risk and high possible return. The risk in this sense is mainly constituted by the temporal dimension of recoveries – this means that when using the methods that the author has described, one can predict reasonably well if, but not when loans will recover. Therefore, it must be concluded that it's possible to predict recoveries, but it's not possible to calculate the expected return.

The author suggests that in order to further investigate this way of making investments, one should look into more complex ensemble methods, because they seem to be able to predict credit-related information very well. One should look towards Lessmann *et al* and repeat their way of comparing predictive methods.

Lisa 1. Otsustuspuu reeglitena

Rule number: 7 [Recovery=1 cover=800 (8%) prob=0.97]
  ICN_RLG_PrincipalPaymentsMade>=5.02
  ICN_RLG_PrincipalBalance< 0.5148
 Rule number: 53 [Recovery=1 cover=375 (4%) prob=0.89]
  ICN_RLG_PrincipalPaymentsMade>=5.02
  ICN_RLG_PrincipalBalance>=0.5148
  ICN_RLG_PlannedInterestPostDefault>=6.232
  ICN_RLG_PrincipalPaymentsMade< 6.676
  ICN_RLG_InterestAndPenaltyPaymentsMade< 5.592
 Rule number: 27 [Recovery=1 cover=934 (10%) prob=0.87]
  ICN_RLG_PrincipalPaymentsMade>=5.02
  ICN_RLG_PrincipalBalance>=0.5148
  ICN_RLG_PlannedInterestPostDefault>=6.232
  ICN_RLG_PrincipalPaymentsMade>=6.676
 Rule number: 51 [Recovery=1 cover=293 (3%) prob=0.77]
  ICN_RLG_PrincipalPaymentsMade>=5.02
  ICN_RLG_PrincipalBalance>=0.5148
  ICN_RLG_PlannedInterestPostDefault< 6.232
  ICN_RLG_InterestAndPenaltyPaymentsMade< 5.689
  ICN_RLG_PlannedInterestPostDefault>=5.032
 Rule number: 11 [Recovery=1 cover=223 (2%) prob=0.70]
  ICN_RLG_PrincipalPaymentsMade< 5.02
  ICN_RLG_PrincipalPaymentsMade>=3.024
  ICN_RLG_InterestAndPenaltyPaymentsMade< 3.416
 Rule number: 52 [Recovery=0 cover=1096 (11%) prob=0.42]
  ICN_RLG_PrincipalPaymentsMade>=5.02
  ICN_RLG_PrincipalBalance>=0.5148
  ICN_RLG_PlannedInterestPostDefault>=6.232
  ICN_RLG_PrincipalPaymentsMade< 6.676
  ICN_RLG_InterestAndPenaltyPaymentsMade>=5.592
 Rule number: 50 [Recovery=0 cover=323 (3%) prob=0.41]
  ICN_RLG_PrincipalPaymentsMade>=5.02
  ICN_RLG_PrincipalBalance>=0.5148
  ICN_RLG_PlannedInterestPostDefault< 6.232
  ICN_RLG_InterestAndPenaltyPaymentsMade< 5.689
  ICN_RLG_PlannedInterestPostDefault< 5.032
 Rule number: 24 [Recovery=0 cover=1233 (13%) prob=0.24]
  ICN_RLG_PrincipalPaymentsMade>=5.02
  ICN_RLG_PrincipalBalance>=0.5148
  ICN_RLG_PlannedInterestPostDefault< 6.232
  ICN_RLG_InterestAndPenaltyPaymentsMade>=5.689
 Rule number: 10 [Recovery=0 cover=1909 (19%) prob=0.10]
  ICN_RLG_PrincipalPaymentsMade< 5.02
  ICN_RLG_PrincipalPaymentsMade>=3.024
  ICN_RLG_InterestAndPenaltyPaymentsMade>=3.416
 Rule number: 4 [Recovery=0 cover=2633 (27%) prob=0.00]
  ICN_RLG_PrincipalPaymentsMade< 5.02
  ICN_RLG_PrincipalPaymentsMade< 3.024

Lisa 2. Otsustuspuude metsa muutujate tähtsushierarhia

	0	1	MeanDecreaseAccuracy
ICN_RLG_PrincipalPaymentsMade	42.31	40.31	46.88
ICN_RLG_PrincipalBalance	34.73	17.89	36.40
ICN_RLG_PlannedInterestPostDefault	24.93	20.12	31.84
ICN_RLG_PlannedPrincipalPostDefault	22.06	25.87	31.02
ICN_RLG_InterestAndPenaltyBalance	28.89	14.39	29.70
ICN_RLG_InterestAndPenaltyPaymentsMade	24.96	-0.93	24.82
R01_EAD2	18.20	12.00	24.32
R01_EAD1	16.80	11.14	20.75
ICN_RLG_PlannedPrincipalTillDate	18.56	7.17	20.56
ExpectedReturn	18.98	3.72	19.88
RLG_PlannedInterestTillDate	13.42	11.05	19.79
RLG_AppliedAmount	16.24	8.90	18.96
RLG_Interest	15.73	7.74	17.90
RLG_IncomeTotal	15.19	6.71	17.09
RLG_Amount	15.47	7.18	16.52
LossGivenDefault	12.85	13.04	15.82
ProbabilityOfDefault	13.13	8.09	15.27
LoanDuration	11.42	8.87	14.57
RRC_ExpectedLoss	13.34	3.14	14.23
TIN_Country_ES	9.97	12.60	13.79
ICN_RLG_IncomeFromPrincipalEmployer	10.03	8.40	13.73
DebtToIncome	11.91	5.05	13.61
ICN_RLG_FreeCash	8.94	9.56	13.45
ExistingLiabilities	10.64	6.02	12.94
TIN_Country_EE	9.28	11.98	12.59
RLG_LiabilitiesTotal	11.50	4.69	12.50
TIN_TFC_LanguageCode_.5.6.	9.09	10.97	11.79
ICN_RLG_InterestAndPenaltyWriteOffs	10.05	-0.20	10.32
TIN_TFC_LanguageCode_.1.1.	8.28	9.03	9.64
RefinanceLiabilities	9.33	2.27	9.20
TIN_Country_FI	6.70	6.31	8.74
ICN_RLG_PreviousRepaymentsBeforeLoan	5.89	5.91	8.61
TIN_TFC_LanguageCode_.3.4.	5.24	6.15	8.60
TIN_TFC_Restructured_.FALSE.TRUE.	7.38	3.86	8.36
ICN_RLG_AmountOfPreviousLoansBeforeLoan	6.99	2.29	7.78
Age	10.23	-1.20	7.28
TIN_TFC_Restructured_.FALSE.FALSE.	6.63	2.69	7.16
TIN_Rating_HR	7.07	2.24	6.99
TIN_TFC_Gender_.1.2.	4.85	4.32	6.96
TIN_TFC_VerificationType_.1.1.	5.60	3.59	6.59
ICN_RLG_IncomeFromPension	4.37	4.24	5.97
ICN_RLG_IncomeFromSocialWelfare	4.32	3.96	5.69
TIN_TFC_EmploymentStatus_.5.6.	3.30	4.58	5.69
TIN_TFC_VerificationType_.3.4.	4.56	3.59	5.55
TIN_TFC_NewCreditCustomer_.FALSE.FALSE.	3.78	2.60	5.30
NrOfDependants	5.19	1.97	5.25
ICN_RLG_PreviousEarlyRepaymentsBefoleLoan	5.26	0.70	5.17
TIN_TFC_UseOfLoan_.0.0.	4.80	1.87	4.78
TIN_WorkExperience_MoreThan25Years	4.14	2.07	4.75
NoOfPreviousLoansBeforeLoan	3.53	2.20	4.29
TIN_TFC_VerificationType_.2.3.	2.69	2.81	4.29
TIN_TFC_Gender_.0.1.	3.07	2.90	4.23
ICN_RLG_IncomeFromFamilyAllowance	4.31	0.61	4.12
TIN_TFC_MaritalStatus_.2.3.	2.49	3.25	4.00
TIN_TFC_NewCreditCustomer_.FALSE.TRUE.	3.60	1.08	3.96
PreviousEarlyRepaymentsCountBeforeLoan	3.23	1.88	3.88
TIN_TFC_UseOfLoan_.1.2.	3.32	1.41	3.53
ICN_RLG_IncomeOther	2.78	1.73	3.47

Lisa 3. Võimendamise muutujate tähtsushierarhia

	Feature	Gain
1:	ICN_RLG_PrincipalPaymentsMade	0.41800507261
2:	ICN_RLG_InterestAndPenaltyPaymentsMade	0.11786634118
3:	ICN_RLG_PlannedInterestPostDefault	0.07362140913
4:	ICN_RLG_PrincipalBalance	0.07259265933
5:	ICN_RLG_InterestAndPenaltyBalance	0.05531675710
6:	ICN_RLG_PlannedPrincipalTillDate	0.03215253589
7:	R01_EAD1	0.02871662585
8:	ICN_RLG_PlannedPrincipalPostDefault	0.01717146804
9:	LoanDuration	0.01694322906
10:	R01_EAD2	0.01430003243
11:	TIN_Country_ES	0.01381798559
12:	RLG_Interest	0.01277000803
13:	ExpectedReturn	0.01053103722
14:	ProbabilityOfDefault	0.00799368589
15:	ICN_RLG_IncomeFromPrincipalEmployer	0.00752769511
16:	RLG_PlannedInterestTillDate	0.00707005371
17:	Age	0.00701125678
18:	RLG_IncomeTotal	0.00698420790
19:	DebtToIncome	0.00578483193
20:	RLG_LiabilitiesTotal	0.00571847473
21:	ICN_RLG_FreeCash	0.00538865123
22:	LossGivenDefault	0.00506596775
23:	RRC_ExpectedLoss	0.00420080825
24:	ICN_RLG_PreviousRepaymentsBeforeLoan	0.00389157825
25:	RLG_AppliedAmount	0.00377644126
26:	ICN_RLG_AmountOfPreviousLoansBeforeLoan	0.00340680315
27:	TIN_TFC_LanguageCode_.1.1.	0.00325349119
28:	TIN_TFC_Restructured_.FALSE.FALSE.	0.00264599060
29:	NrOfDependants	0.00219746019
30:	ExistingLiabilities	0.00213733922
31:	RefinanceLiabilities	0.00206480882
32:	TIN_Country_EE	0.00202876258
33:	TIN_TFC_Education_.3.4.	0.00176552856
34:	ICN_RLG_PreviousEarlyRepaymentsBefoleLoan	0.00164950556
35:	RLG_Amount	0.00147739431
36:	ICN_RLG_IncomeFromFamilyAllowance	0.00138815643
37:	ICN_RLG_IncomeOther	0.00114427491
38:	ICN_RLG_IncomeFromSocialWelfare	0.00096484368
39:	ICN_RLG_IncomeFromPension	0.00095692064
40:	TIN_TFC_HomeOwnershipType_.0.1.	0.00095656836
41:	TIN_TFC_UseOfLoan_.6.7.	0.00092942180
42:	TIN_TFC_UseOfLoan_.1.2.	0.00074106762
43:	TIN_TFC_Education_.2.3.	0.00073253951
44:	NoOfPreviousLoansBeforeLoan	0.00072528727
45:	TIN_EmploymentDurationCurrentEmployer_UpTo5Years	0.00071804966
46:	TIN_TFC_MaritalStatus_.1.2.	0.00067013346
47:	ICN_RLG_IncomeFromChildSupport	0.00062424284
48:	TIN_EmploymentDurationCurrentEmployer_UpTo1Year	0.00057416467
49:	TIN_EmploymentDurationCurrentEmployer_MoreThan5Years	0.00053264959
50:	TIN_TFC_Gender_.0.1.	0.00053192012
51:	TIN_WorkExperience_10To15Years	0.00050099016
52:	TIN_TFC_MaritalStatus_.3.4.	0.00046132210
53:	TIN_TFC_VerificationType_.1.1.	0.00041599120
54:	TIN_Country_FI	0.00041283182
55:	TIN_TFC_VerificationType_.2.3.	0.00040048348
56:	TIN_TFC_Gender_.0.0.	0.00039854522
57:	TIN_TFC_MaritalStatus_.1.1.	0.00038426279
58:	TIN_TFC_Education_.4.5.	0.00038019848
59:	TIN_TFC_HomeOwnershipType_.2.3.	0.00036065085
60:	TIN_TFC_EmploymentStatus_.2.3.	0.00035420194

Lisa 4. Logistilise regressiooni mudel

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	3.267e+13	2.881e+13	1.134	0.256816	
Age	-5.245e-03	7.482e-03	-0.701	0.483304	
LoanDuration	6.564e-02	9.486e-03	6.919	4.54e-12	***
NrOfDependants	-8.752e-03	5.655e-02	-0.155	0.876995	
ExistingLiabilities	2.165e-02	2.501e-02	0.866	0.386610	
RefinanceLiabilities	2.028e-02	4.000e-02	0.507	0.612173	
DebtToIncome	7.215e-03	4.555e-03	1.584	0.113202	
LossGivenDefault	3.231e+00	1.256e+00	2.572	0.010123	*
ExpectedReturn	2.933e+00	1.748e+00	1.678	0.093403	.
ProbabilityOfDefault	-2.325e+00	1.431e+00	-1.624	0.104326	
NoOfPreviousLoansBeforeLoan	2.983e-02	7.739e-02	0.385	0.699881	
PreviousEarlyRepaymentsCountBeforeLoan	-7.683e-02	2.043e-01	-0.376	0.706827	
RLG_AppliedAmount	8.225e-02	1.731e-01	0.475	0.634567	
RLG_Amount	-1.380e+00	3.305e-01	-4.174	2.99e-05	***
RLG_Interest	1.206e+00	6.908e-01	1.745	0.080933	.
RLG_IncomeTotal	6.861e-02	2.233e-01	0.307	0.758696	
RLG_LiabilitiesTotal	-7.627e-02	1.331e-01	-0.573	0.566546	
RLG_PlannedInterestTillDate	-6.534e-01	2.934e-01	-2.227	0.025974	*
RRC_ExpectedLoss	7.635e-01	4.344e-01	1.758	0.078797	.
R01_EAD1	-3.525e+00	9.570e-01	-3.684	0.000230	***
R01_EAD2	1.133e+01	1.554e+00	7.288	3.15e-13	***
ICN_RLG_IncomeFromPrincipalEmployer	8.981e-02	5.703e-02	1.575	0.115295	
ICN_RLG_IncomeFromPension	1.662e-03	3.487e-02	0.048	0.961978	
ICN_RLG_IncomeFromFamilyAllowance	-2.523e-02	3.181e-02	-0.793	0.427750	
ICN_RLG_IncomeFromSocialWelfare	-2.741e-02	4.512e-02	-0.608	0.543452	
ICN_RLG_IncomeFromLeavePay	7.752e-02	6.062e-02	1.279	0.200989	
ICN_RLG_IncomeFromChildSupport	1.386e-02	4.492e-02	0.308	0.757762	
ICN_RLG_IncomeOther	-1.961e-03	2.753e-02	-0.071	0.943213	
ICN_RLG_PreviousEarlyRepaymentsBefoleLoan	1.161e-03	3.969e-02	0.029	0.976672	
ICN_RLG_PreviousRepaymentsBeforeLoan	6.192e-02	7.356e-02	0.842	0.399894	
ICN_RLG_AmountOfPreviousLoansBeforeLoan	-3.435e-02	5.717e-02	-0.601	0.547929	
ICN_RLG_InterestAndPenaltyBalance	-2.633e-01	5.782e-02	-4.553	5.29e-06	***
ICN_RLG_PrincipalBalance	-6.547e-01	8.733e-02	-7.498	6.50e-14	***
ICN_RLG_InterestAndPenaltyWriteOffs	1.487e-01	2.959e-01	0.502	0.615363	
ICN_RLG_InterestAndPenaltyPaymentsMade	-1.282e+00	5.701e-02	-22.485	< 2e-16	***
ICN_RLG_PrincipalPaymentsMade	3.252e+00	1.014e-01	32.088	< 2e-16	***
ICN_RLG_PlannedInterestPostDefault	1.385e+00	8.758e-02	15.815	< 2e-16	***
ICN_RLG_PlannedPrincipalPostDefault	2.339e-01	6.250e-02	3.742	0.000182	***
ICN_RLG_PlannedPrincipalTillDate	-5.612e-01	1.472e-01	-3.813	0.000137	***
ICN_RLG_FreeCash	-3.517e-02	5.083e-02	-0.692	0.489071	
TIN_Country_EE	2.605e+13	1.792e+13	1.454	0.145921	
TIN_Country_ES	2.605e+13	1.792e+13	1.454	0.145921	
TIN_Country_FI	2.605e+13	1.792e+13	1.454	0.145921	
TIN_Country_SK	NA	NA	NA	NA	
TIN_EmploymentDurationCurrentEmployer_MoreThan5Years	-2.442e-01	1.681e-01	-1.452	0.146407	
TIN_EmploymentDurationCurrentEmployer_TrialPeriod	-2.300e-01	4.103e-01	-0.560	0.575199	
TIN_EmploymentDurationCurrentEmployer_UpTo1Year	-2.504e-01	1.872e-01	-1.338	0.180948	
TIN_EmploymentDurationCurrentEmployer_UpTo2Years	-2.045e-01	1.907e-01	-1.073	0.283485	
TIN_EmploymentDurationCurrentEmployer_UpTo3Years	-3.825e-01	2.031e-01	-1.884	0.059632	.
TIN_EmploymentDurationCurrentEmployer_UpTo4Years	-2.960e-01	2.162e-01	-1.369	0.170984	
TIN_EmploymentDurationCurrentEmployer_UpTo5Years	NA	NA	NA	NA	
TIN_WorkExperience_10To15Years	-1.312e-01	1.829e-01	-0.717	0.473156	
TIN_WorkExperience_15To25Years	1.350e-02	1.522e-01	0.089	0.929304	
TIN_WorkExperience_2To5Years	7.525e-02	2.368e-01	0.318	0.750644	
TIN_WorkExperience_5To10Years	-7.570e-02	2.015e-01	-0.376	0.707213	
TIN_WorkExperience_LessThan2Years	-7.462e-02	2.702e-01	-0.276	0.782371	
TIN_WorkExperience_MoreThan25Years	NA	NA	NA	NA	

Lisa 4 jätk

```
TIN_Rating_A                                    -5.059e-01  4.993e-01  -1.013 0.310962
TIN_Rating_AA                                   -1.368e+00  9.850e-01  -1.389 0.164755
TIN_Rating_B                                    -7.300e-01  3.517e-01  -2.076 0.037896 *
TIN_Rating_C                                    -6.646e-01  2.986e-01  -2.226 0.026045 *
TIN_Rating_D                                    -5.413e-01  2.639e-01  -2.051 0.040273 *
TIN_Rating_E                                    -4.927e-01  2.392e-01  -2.060 0.039410 *
TIN_Rating_F                                    -2.311e-01  2.104e-01  -1.098 0.272077
TIN_Rating_HR                                           NA         NA      NA       NA
TIN_TFC_NewCreditCustomer_.FALSE.FALSE.         -3.601e-02  2.692e-01  -0.134 0.893592
TIN_TFC_NewCreditCustomer_.FALSE.TRUE.                  NA         NA      NA       NA
TIN_TFC_VerificationType_.1.1.                   1.567e-01  1.383e-01   1.133 0.257112
TIN_TFC_VerificationType_.1.2.                  -4.616e-01  5.020e-01  -0.920 0.357756
TIN_TFC_VerificationType_.2.3.                   2.163e-01  1.335e-01   1.621 0.105059
TIN_TFC_VerificationType_.3.4.                          NA         NA      NA       NA
TIN_TFC_LanguageCode_.1.1.                      -5.872e+13  2.591e+13  -2.266 0.023454 *
TIN_TFC_LanguageCode_.1.2.                      -5.872e+13  2.591e+13  -2.266 0.023454 *
TIN_TFC_LanguageCode_.2.3.                      -5.872e+13  2.591e+13  -2.266 0.023454 *
TIN_TFC_LanguageCode_.3.4.                      -5.872e+13  2.591e+13  -2.266 0.023454 *
TIN_TFC_LanguageCode_.4.5.                      -5.872e+13  2.591e+13  -2.266 0.023454 *
TIN_TFC_LanguageCode_.5.6.                      -5.872e+13  2.591e+13  -2.266 0.023454 *
TIN_TFC_LanguageCode_.6.7.                              NA         NA      NA       NA
TIN_TFC_LanguageCode_.7.9.                      -3.267e+13  2.881e+13  -1.134 0.256816
TIN_TFC_LanguageCode_.9.10.                     -3.267e+13  2.881e+13  -1.134 0.256816
TIN_TFC_LanguageCode_.10.13.                    -3.267e+13  2.881e+13  -1.134 0.256816
TIN_TFC_LanguageCode_.13.15.                    -5.872e+13  2.591e+13  -2.266 0.023454 *
TIN_TFC_LanguageCode_.15.21.                    -5.872e+13  2.591e+13  -2.266 0.023454 *
TIN_TFC_LanguageCode_.21.22.                    -5.872e+13  2.591e+13  -2.266 0.023454 *
TIN_TFC_Gender_.0.0.                            -1.883e-01  3.363e-01  -0.560 0.575519
TIN_TFC_Gender_.0.1.                             1.213e-02  3.386e-01   0.036 0.971433
TIN_TFC_Gender_.1.2.                                    NA         NA      NA       NA
TIN_TFC_UseOfLoan_.0.0.                         -2.448e-01  2.584e-01  -0.947 0.343539
TIN_TFC_UseOfLoan_.0.1.                         -3.018e-01  3.814e-01  -0.791 0.428725
TIN_TFC_UseOfLoan_.1.2.                         -1.543e-01  2.477e-01  -0.623 0.533471
TIN_TFC_UseOfLoan_.2.3.                         -4.949e-01  3.172e-01  -1.560 0.118748
TIN_TFC_UseOfLoan_.3.4.                         -4.835e-01  3.604e-01  -1.341 0.179777
TIN_TFC_UseOfLoan_.4.5.                         -4.452e-01  3.138e-01  -1.419 0.155908
TIN_TFC_UseOfLoan_.5.6.                         -6.705e-02  2.828e-01  -0.237 0.812559
TIN_TFC_UseOfLoan_.6.7.                         -5.176e-02  2.464e-01  -0.210 0.833631
TIN_TFC_UseOfLoan_.7.8.                                 NA         NA      NA       NA
TIN_TFC_Education_.1.1.                          2.024e-01  4.286e-01   0.472 0.636722
TIN_TFC_Education_.1.2.                          5.221e-02  1.576e-01   0.331 0.740493
TIN_TFC_Education_.2.3.                         -2.062e-02  1.366e-01  -0.151 0.879997
TIN_TFC_Education_.3.4.                         -1.186e-01  1.229e-01  -0.965 0.334487
TIN_TFC_Education_.4.5.                                 NA         NA      NA       NA
TIN_TFC_MaritalStatus_.1.1.                     -1.600e-02  3.490e-01  -0.046 0.963441
TIN_TFC_MaritalStatus_.1.2.                      1.643e-01  3.545e-01   0.463 0.643078
TIN_TFC_MaritalStatus_.2.3.                     -1.931e-01  3.575e-01  -0.540 0.589109
TIN_TFC_MaritalStatus_.3.4.                     -3.864e-03  3.639e-01  -0.011 0.991528
TIN_TFC_MaritalStatus_.4.5.                             NA         NA      NA       NA
TIN_TFC_EmploymentStatus_.0.0.                  -9.369e-01  1.198e+00  -0.782 0.434264
TIN_TFC_EmploymentStatus_.0.2.                  -4.635e-02  4.593e-01  -0.101 0.919617
TIN_TFC_EmploymentStatus_.2.3.                   2.118e-01  4.055e-01   0.522 0.601488
TIN_TFC_EmploymentStatus_.3.4.                   4.394e-01  4.724e-01   0.930 0.352281
TIN_TFC_EmploymentStatus_.4.5.                  -1.599e-01  4.522e-01  -0.354 0.723691
TIN_TFC_EmploymentStatus_.5.6.                          NA         NA      NA       NA
TIN_TFC_HomeOwnershipType_.0.0.                 -1.378e+00  1.487e+00  -0.927 0.354071
TIN_TFC_HomeOwnershipType_.0.1.                 -6.967e-01  3.478e-01  -2.003 0.045192 *
TIN_TFC_HomeOwnershipType_.1.2.                 -6.770e-01  3.631e-01  -1.865 0.062240 .
TIN_TFC_HomeOwnershipType_.2.3.                 -8.937e-01  3.590e-01  -2.490 0.012785 *
TIN_TFC_HomeOwnershipType_.3.4.                 -9.524e-01  3.688e-01  -2.583 0.009804 **
TIN_TFC_HomeOwnershipType_.4.5.                 -1.490e+00  4.500e-01  -3.311 0.000930 ***
TIN_TFC_HomeOwnershipType_.5.6.                 -9.501e-01  4.029e-01  -2.358 0.018355 *
TIN_TFC_HomeOwnershipType_.6.7.                 -6.201e-01  3.794e-01  -1.635 0.102146
TIN_TFC_HomeOwnershipType_.7.8.                 -4.543e-01  3.624e-01  -1.254 0.209946
TIN_TFC_HomeOwnershipType_.8.9.                         NA         NA      NA       NA
TIN_TFC_Restructured_.FALSE.FALSE.              -4.469e-01  1.102e-01  -4.054 5.03e-05 ***
TIN_TFC_Restructured_.FALSE.TRUE.                       NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 12657.5  on 9811  degrees of freedom
Residual deviance:  3619.2  on 9702  degrees of freedom
  (7 observations deleted due to missingness)
AIC: 3839.2

Number of Fisher Scoring iterations: 25

Log likelihood: -1809.589 (110 df)
Null/Residual deviance difference: 9038.365 (109 df)
Chi-square p-value: 0.00000000
Pseudo R-Square (optimistic): 0.87388542
```

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina Ekke Sakkov
(sünnikuupäev: 11.11.1992)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose
„Pankrotistunud laenudesse investeerimine bondora.ee laenukeskkonnas“,
mille juhendajad on Kurmet Kivipõld ja Hendrik Luuk,

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 kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

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

Tartus 24.05.2018