TARTU ÜLIKOOL
LOODUS-JA TÄPPISTEADUSTE VALDKOND
MATEMAATIKA JA STATISTIKA INSTITUUT

Lagle Sammelsaar

Pankrotistumise tõenäosuse prognoosimine otselaenamisettevõtte Bondora andmetel

Magistritöö

Finants- ja kindlustusmatemaatika erialal (30 EAP)

Juhendaja: Raul Kangro, PhD

Tartu 2016

Pankrotistumise tõenäosuse prognoosimine otselaenamisettevõtte Bondora andmetel

Magistritöö

Lagle Sammelsaar

Lühikokkuvõte. Käesoleva magistritöö põhieesmärk on hinnata otselaenamisettevõtte Bondora klientide oodatava kahju ühte komponenti, milleks on pankrotistumise tõenäosus, et võimalikult täpselt määrata riskitase nende klientide korral, kellele ollakse valmis laenu andma. Analüüs teostatakse avaliku klientide kohta käiva andmestiku põhjal ([1]). Pankrotistumise tõenäosuse prognoosimisele lähenetakse interaktiivse ja automatiseeritud metoodikatega, toetudes logistilisele regressioonile ning regressioonipuule *CART*, et välja selgitada parim krediidiriski kirjeldav mudel. Lõplik mudeli valik tugineb prognoosihinnangute võrdlemisel keskmise ruutvea põhjal.

CERCS teaduseriala: P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika

Märksõnad: pankrotistumise tõenäosus, prognostika, regressioonanalüüs, tehisõpe

Estimating the 'Probability of Default' on a dataset from a lending platform Bondora

Master's thesis

Lagle Sammelsaar

Abstract. The main objective of this Master Thesis is to estimate Bondora clients' Probability Of Default, one of the components of Expected Loss, and predicting risk rate as accurately as possible for the applicants that Bondora is ready to lend to. Analysis is conducted on by publicly accessible client's dataset ([1]). In this Thesis the Probability of Default has been estimated using interactive and automated techniques, leaning on logistic regression and regression tree *CART* methodology to achieve the best credit risk model. The final model is validated by comparing it to the estimated prediction using mean squared error.

CERCS research specialisation: P160 Statistics, operation research, programming, actuarial mathematics

Keywords: probability of default, prediction, regression analysis, automatic learning

Sisukord

Sissejuhatus

Suurenenud konkurents ning surve tulu tekitamiseks on pannud laenuettevõtteid ja erinevaid finantsinstitutsioone otsima efektiivsemaid meetodeid krediidivõimeliste klientide eristamiseks. Lisaks on laenu väljastavatele ettevõtetele oluline netotootlikuse kontrolli all hoidmine. Sellest ajendatult on hakatud potentsiaalsete klientide riske põhjalikumalt hindama ning nende kohta käivat informatsiooni kiiremini ja tõhusamalt töötlema. Lisaks selgelt kahjulike klientide välistamisele on eesmärgiks laenutoodete paindlik hinnastamine vastavalt kliendi riskitasemele. Sellistes oludes krediidiriski mudelid pakuvad tõhusat, ettevõtete vajadustele vastavaid empiirilistele andmetele tuginevaid lahendusi ([2], lk 1-2). Krediidimudeli meetodika võimaldab objektiivselt hinnata erinevaid riskifaktoreid.

Käesoleva magistritöö eesmärgiks on hinnata võimalikult täpselt kliendi laenulepingu pankrotistumise tõenäosust, et määrata klientide riskitase, kellele ollakse valmis laenu andma. Huvipakkuva suuruse prognoosimiseks kasutatakse statistilisi mudeleid. Antud magistritöö on üles ehitatud viiele peatükile.

Esimeses peatükis tutvustatakse krediidiriski hindamise probleemi ning peatutakse laenulepingu oodatava kahju kujunemisel. Põhjalikumalt käsitletakse pankrotistumise tõenäosuse prognoosimist ja selle probleemile lähenemist konkreetses töös. Lisaks täpsustatakse prognoosimudelite võrdlemise printsiipe.

Teises peatükis antakse ülevaade andmete eeltöötlusest ja korrastamisest ning kirjeldatakse lisatud arvutatud tunnuseid. Lisaks käsitletakse töös kasutatud prognoosimudelite headuste mõõdikuid.

Kolmas ja neljas peatükk puudutab teoreetilisi ning praktilisi käsitlusi. Kirjeldatakse parima mudelite otsimise protseduure koos tulemustega logistilise regressiooni ning regressioonipuu meetodite juhtudel. Tulemusi on illustreeritud graafiliselt ning võrreldud prognoosimudeleid. Kolmas osa keskendub logistilisele regressioonile, neljas regressioonipuu $CART$ meetodile.

Viiendas osas koostatud koondtabeli põhjal ostutakse erinevate lähenemisega saadud parim mudel. Lisaks võrreldakse Bondora süsteemi poolt arvutatud krediidiskoore, et uurida pankrotistumise tõenäosuse hindamiseks väljatöötatud parima mudeli täpsust.

Töö on vormistatud tekstitöötlusprogrammis LaTeXning praktiline pool on teostatud statistika tarkvaras *R*. Magistritööle on kaasa pandud andmestik koos failide, koodide ja tulemustega.

Töö valmimisele on kaasa aidanud juhendaja Raul Kangro, kellele töö autor soovib tänu avaldada arvukate konsultatsioonide, suunamiste, paranduste ja täienduste eest, olles inspireeriv juhendaja kui ka õppejõud.

Bondorast

Sotsiaalpanganduse platvorm Bondora on välja kasvanud laenuettevõttest IsePankur OÜ 2009. aastal. Antud ettevõte pakub tagatiseta laenude vahendamise teenust eraisikutelt eraisikutele (*Peer-to-Peer lending*), selleks ühendades kahte osapoolt- investorid ja laenuvõtjad. Bondora administratiivteenusteks on laenutaotlejate krediidiinfo kontrollimine, laekuvate maksete kogumine, lepingute sõlmimine ja klienditugi. Pärast kliendipoolset laenutingimuste ja pakkumise aksepteerimist avatakse investoritele oksjon. Investoritel on võimalus investeerida täielikult või osaliselt nendele sobiva riskitasemega eraisiku laenu ja seeläbi teenida väljalaenatud summalt intressitulu. Bondorast pakutavate tarbimis- või väikelaenude krediidi intressimäär on väiksem võrreldes erinevate pankade ja kiirlaenufirmadega, mistõttu kliendid saavad soodsamatel tingimustel laenu. Lisaks võimaldatakse Bondoras ka laenude refinantseerimist ehk konsolideerimist lisatagatise olemasolul. Investeerimisvõimalus Bondora keskkonnas on antud kõikidele Euroopa Liidu liikmesriikidele ning Šveitsi residentidele alates 2012. aastast ning minimaalne võimalik investeering laenusummale algab viiest eurost. Hetkel pakub Bondora laenuteenuseid Eestis, Soomes ja Hispaanias. Laenatavad summad on 500-10000 eurot ning võimaldatav laenuperiood kolm kuni 60 kuud. Laenaja intressimäär sõltub konkreetselt kliendi krediidiriskigrupist, mille kinnitab Bondora iga laenaja andmete ja kreedidiajaloo põhjal, kasutades selleks Krediidiinfo AS andmebaasidest saadavat informatsiooni. Enamikul juhtudel raha on laenatud eelnevate laenude refinantseerimiseks, mis tähendab kõrgema intressimääraga laenu ennetähtaegset tagasimaksmist soodsama laenu abil ning planeeritud või planeerimata kulutuste katmiseks ([1] ja [3]).

1 Krediidiriski hindamise probleem

Antud kontekstis krediidiriski all on mõistetud laenu väljastamisega kaasnevat riski, mis seisneb võimaluses, et klient muutub enne laenulepingu lõpptähtaega maksejõuetuks, millele järgneb laenaja laenulepingutingimuste rikkumine ning võlausaldajale kahju tekitamine.
Käesolevas töös iseloomustame krediidiriski tõenäosusega, et klient muutub maksejõuetuks.

Krediidiriski põhjal on võimalik moodustada erineva riskitasemega laenulepingute portfelle, millesse investeerimine võib pakkuda huvi nii suurinvestoritele kui üksikisikutele, võimaldades vähendada investeerimisriske ilma oodatava tulu kahanemiseta. Laenuvõtjad jaotatakse rühmadesse vastavalt riskantsusastmele, kus eristatakse madala riskiga kliente kõrgema riskiga klientidest. Laenuettevõtetele on oluline sobiva intressi määramiseks võimalikult täpne ja korrektne rühmadesse jaotamine, et katta oodatav kahju vaadeldavas rühmas. Vaadeldavate suuruste leidmiseks kasutatakse statistilisi mudeleid ning toetutakse eeldusele, et minevikuandmed klientide võlgnevuste tasumise kohta võimaldavad hästi prognoosida sarnaste taustaandmetega klientide maksekäitumist tulevikus.

1.1 Oodatav kahju

Järgnevas alampeatükis käsitletakse laenulepingu oodatava kahju kujunemist.

Laenuettevõtete jaoks tähendab kahju väljalaenatud summa mitte tagasi maksmist kliendi pankrotistumise tagajärjel. Iga laenu väljastamisel leitakse kliendile vastav intressimäär, mis sõltub oodatava kahju (*EL*, *Expected Loss*) ja oodatava tulu (*ER*, *Expected Return*) lihtprotsendist ning ettevõtte poolt määratud plat-

vormtasudest.

Olgu i-nda laenulepingu kahju juhuslik suurus L_i, mida saab esitada kujul

$$L_i = \begin{cases} 0, \text{ kui pankrotistumine ei toimunud} \\ EAD_i \cdot LGD_i, \qquad \text{kui pankrotistumine toimus} \end{cases}$$

ning i-nda kliendi pankrotistumist kirjeldab juhuslik suurus D_i, omandades väärtuseid

$$D_i = \begin{cases} 1, \text{tõenäosusega } p(x_i) \\ 0, \text{tõenäosusega } 1 - p(x_i). \end{cases}$$

Suurus EAD_i (*Exposure at Default*) on konkreetse laenu korral oodatav protsent põhiosast, mis on tagasi maksmata pankrotistumise hetkel ehk vaadeldavast juhuslikust suurusest on võetud keskväärtus. Suurus LGD_i (*Loss given Default*) on oodatav osa (protsent), mis jääb pankrotimenetluse käigus tagasi saamata pankroti tekkimise hetkel maksmata põhiosast.

Eelnevaid tähistusi kasutades võime oodatava kahju leida valemiga

$$EL_i = P(D_i = 1) \cdot E(EAD_i \cdot LGD_i \mid D_i = 1). \tag{1.1}$$

Suurus $P(D_i = 1)$, lühemalt PD (*Probability of Default*), tähistab kliendi tõenäosust pankrotistuda laenulepingu perioodil. Olgu mainitud, et oodatavat kahju antud magistritöös on käsitletakse kui protsentuaalset kahju.

Bondora dokumentidele kohaselt ning täiendavate eelduste kehtimisel i-nda laenulepingu oodatav kahju on arvutatav komponentide kaupa ning leitav valemiga

$$EL_i = P(D_i = 1) \cdot E(EAD_i \mid D_i = 1) \cdot E(LGD_i \mid D_i = 1). \tag{1.2}$$

Valemis 1.2 on eeldatud maksmata summa ja pärast pankrotimenetluse lõppu allesjääva osade sõltumatust pankrotistunud lepingute korral ehk vaadeldavad suurused EAD_i ja LGD_i on sõltumatud tingimusel, et pankrotistumine toimus.

Klientide individuaalne intressimäär on tuletatud oodatava kahju, oodatava tulu ja platvormtasude protsentide summana. Ettevõtte platvormtasud on Bondora poolt fikseeritud, sisaldades lepingutasusid, mis on 5.95 % väljalaenatud summalt ja laenuhaldustasud 2.6 % aastas. Laenaja intressimäär on leitav järgnevalt

$$I = EL + ER + Platvormtasud. \tag{1.3}$$

1.2 Pankrotistumise tõenäosuse prognoosimine ja erinevate prognoosimudelite võrdlemise printsiibid

Oodatav kahju on leitav komponentide kaupa 1.2, millest üks seda kirjeldav suurus on laostumis- ehk pankrotistumistõenäous. Käesoleva magistritöö põhieesmärgiks on hinnata võimalikult täpselt kliendi laenulepingu pankrotistumise tõenäosust. Selliste mudelite loomisel võib olla erinevaid eesmärke:

1. Otsustada laenuandmine või mitteandmine ning sellisel juhul on tegemist klassifitseerimisülesandega. Eristavateks piirideks on pankrotistunud ja mittepankrotistunud laenuvõtjad, defineerides sellega head ja halvad kliendid.

2. Hinnata võimalikult täpselt kliendi pankrotistumise tõenäosust väljastatava laenu hinnastamise grupeerimiseks ja intressimäära arvutamiseks.

Siin tasub rõhutada, et eesmärgiks ei ole laenu andmise/mitteandmise otsustamine, mis vastaks klassifitseerimisülesandele, vaid võimalikult täpne riskitaseme määramine nende klientide korral, kellele ollakse valmis laenu andma.

Laenusoovijate hulgas võib leiduda ka pahatahtlikke kliente ehk pettureid, kuid selliste klientide identifitseerimist käesolevas töös ei vaadelda.

Eeldame, et leidub selline funktsioon $p : H \mapsto [0,1]$, kus H on parameetervektori võimalike väärtuste hulk ja iga kliendi pankrotistumise tõenäosus on antud selle funktsiooni poolt. Kui kliendi parameetrite vektor on x_i, siis selle tõenäosus on $p(x_i)$. Iga meetodi kasutamise korral teeme mingi eelduse p kuju kohta ning andmete põhjal valime parima (mingis mõttes) funktsiooni $\hat{p}$. On üsna selge, et $\hat{p}$ ei pruugi olla võrdne p-ga ning isegi, kui eeldatud kuju on õige, siis üldjuhul andmete põhjal ei saa me leida täpselt õigeid parameetreid. Loomulikult tahaks olla kindlad, et p ja $\hat{p}$ erinevad võimalikult vähe üksteisest ning sellest lähtuvalt on tekkinud küsimus, kuidas hinnata funktsioonide lähedust. Ideaalne oleks, kui saaks väita, et

$$\left|p(x) - \hat{p}(x)\right| \leq \varepsilon \quad \forall \quad x \tag{1.4}$$

mingi piisavalt väikese epsiloni korral, kuid see on võimatu eesmärk.

Käesolevas töös on oluline leida kriteerium, millega võrrelda hinnatava ja tegeliku funktsioonide lähedust ning üheks mooduseks funktsioonide erinevust mõõta on järgmine lähenemine.

Eeldame, et argumenttunnused on juhuslikud, mis tähendab, et laenulepingute tulemine vastab sõltumatu valimi moodustamisele mingi kindla jaotusega juhusliku vektori (X, D) väärtustest, kus X tähistab fikseeritud komplekt inimesi ning D on juhuslik suurus. Juhuslik suurus D on mingi jaotusega ning laenulepingud on sõltumatud sama jaotusega juhusliku suuruste valim, milleks on potentsiaalsete Bondora klientide hulk. See tähendab, et juhuslike suuruste X ja D korral on pankrotistumine teadaolevate argumentvektori väärtuste korral (juhuslik $D \mid X$)

Bernoulli jaotusega $B_e(p(X))$.

Sel juhul saame mingi meetodiga leitud funktsioonile $\hat{p}$ lähedust õigele funktsioonile p kirjeldada ruutvea keskväärtuse abil, see tähendab suuruse $E\left[p(X)-\hat{p}(X)\right]^2$ kaudu. Kui funktsioon p oleks teada, siis saaks sel juhul $\hat{p}$ headust hinnata vaadeldava valimi põhjal kujul

$$E\left[p(X)-\hat{p}(X)\right]^2 \approx \frac{1}{n}\sum_{i=1}^{n}\left[p(x_i)-\hat{p}(x_i)\right]^2. \tag{1.5}$$

Kuna tegelikule pankrotistumise tõenäosusele vastav funktsioon p ei ole teada, siis on meil vaja leida moodus erinevate p lähendite headuse omavaheliseks võrdlemiseks nii, et me ei pea kasutama p väärtuseid. Selleks paneme tähele, et

$$E\left[\hat{p}(X)-D\right]^2 = E\left[\big(\hat{p}(X)-p(X)\big)+\big(p(X)-D\big)\right]^2. \tag{1.6}$$

Lihtsustades ning summa ruudu valemit rakendades saame

$$E\left[\hat{p}(X)-D\right]^2 = E\left[\hat{p}(X)-p(X)\right]^2+2E\left[\big(\hat{p}(X)-p(X)\big)\big(p(X)-D\big)\right]+E\left[p(X)-D\right]^2.$$

Kasutades tingliku keskväärtuse omadusi ([4], lk 9), leiame

$$E\left[\big(\hat{p}(X)-p(X)\big)\big(p(X)-D\big)\right] = E\left[E\left((\hat{p}(X)-p(X))(p(X)-D)\mid X\right)\right] =$$

$$= E\left[\hat{p}(X)-p(X))(p(X)-E(D\mid X)\right] = 0.$$

$$E\left[\hat{p}(X)-p(X)\right]^2 = E\left[\hat{p}(X)-D\right]^2 - E\left[p(X)-D\right]^2, \tag{1.7}$$

millest parempoolseim suurus ei sõltu vaadeldavast funtksioonist $\hat{p}$, mistõttu võime öelda, et $\hat{p_1}$ on parem lähend p-le kui $\hat{p_2}$, kui

$$E\left[(\hat{p_1}(X)-D)^2\right] < E\left[(\hat{p_2}(X)-D)^2\right]. \tag{1.8}$$

Seega valitud prognoosimudelile vastava funktsiooni $\hat{p}$ korral hindame suurust

$$E\left[\hat{p}(X)-D\right]^2, \tag{1.9}$$

ehk rakendame keskmise ruutvea valemit.
Antud magistritöö põhieesmärgiks on leida võimalikult hea lähend funktsioonile

$$\hat{p}(x) = Pr(D = 1 \mid X = x), \tag{1.10}$$

et vastava tõenäosuse abil hinnata klientide riskitaset.

Pankrotistumise tõenäosust erinevates laenuettevõtetes prognoositakse vastavate krediidiriskimudelitega. Krediidiriskimudelid võimaldavad kogutavate andmete põhjal arvutada iga kliendi jaoks riskihinnangu, et kliente riskipõhiselt paremini võrrelda. Enamasti on tegemist statistiliste mudelitega, mis prognoosivad kliendi maksekäitumist kliendiandmestikus leiduvate argumenttunnuste kaudu ([2], lk 5).

1.3 Pankrotistumise tõenäosuse prognoosimismudeli koostamine

Käesolevas magistritöös lähenetakse pankrotistumise tõenäosuse prognoosimisele kahe meetodiga. Tulemuste omavaheliseks võrdlemiseks on andmestik jaotatud test- ja treeningandmestikuks, kus kõik mudeli koostamisel tehtud otsused langetatakse treeningandmestiku põhjal ning mudeleid võrreldakse testandmestikus olevate lepingute pankrotistumise tõenäosuste prognooside põhjal. Eesmärgiks on välja selgitada, kas logistiline regressioon annab eelise teiste lähenemiste ees ning leida sobviaim krediidiriski hindav mudel.
Kasutatavad lähenemiste metoodikad on alljärgnevad.

- Interaktiivne lähenemine
  - Interaktiivne lähenemine on teostatud logistilise regressiooni mudeli sobitamise näitel. Sel juhul arvutatakse uusi argumenttunnuseid vastavalt sobitaja arusaamisele sellest, mis võiks anda olulist informatsiooni

pankrotistumise tõenäosuse kohta. Sobitamisel kasutatakse vajadusel splaine mittelineaarsete sõltuvuste kirjeldamiseks argumenttunnustest ning argumenttunnuseid lisatakse sisulistel kaalutlustel. Otsuseid langetatakse järk-järgult, arvestades tunnuste statistilist olulisust ja vähima *Akaike* informatsioonikriteeriumi (AIC) mudeli leidmist.

- Automatiseeritud lähenemine
  - logistilise regressiooni sobitamisel valitakse mudel vastavalt sammuviisilisele regressioonimeetodile, kus alustatakse võimalikult suure arvu regressoritega mudelist ning igal sammul lisatakse/eemaldatakse üks tunnus vastavalt *Akaike* kriteeriumile. Antud juhul andmeid ei täiendata paremate krediidiriski kirjeldavate kordajatega ning arvutatavaid tunnuseid juurde ei lisata. Mudelis kasutatakse enamikku üldkogumis leiduvatest argumenttunnustest, mis ei inditseeri kliendile eelnevalt määratud riskitasemele. Lisaks kasutatakse regressioonipuu meetodit ehk $CART$ (vt peatükki 4.3) kahel juhul, täiendatud ja täiendamata andmetele.

2 Andmestiku kirjeldus

Antud magistritöös on kasutatud Bondora koduleheküljelt kättesaadavat avalikku andmestikku ([1]). Alla on laetud fail nimega *Loan dataset* kuupäeva seisuga 21/01/2016. Vaadeldav statistiline andmestik sisaldab erinevaid arvutatud ning mõõdetuid numbrilisi ja mittearvulisi tunnuseid Bondora klientide ehk laenu võtnute kirjeldamiseks. Üldkogum mahutab 50052 laenulepingu informatisooni koos 186 tunnusega.

2.1 Andmestiku eeltöötlus ja korrastamine

Kasutatav andmestik ei olnud täielik, sisaldades mitmeid puuduvaid väärtusi ja loogikavigu. Valimi moodustamisel on teostatud mitmeid kitsendusi, et andmestikku saaks kasutada uuritava suuruse pankrotistumistõenäosuse prognoosimiseks. Esimese sammuna võeti vaatlusesse ainult need laenud, mille korral on täidetud järgmised tingimused:

- tegemist on Eesti klientidega,
- leping on lõppenud või leidub pankrotistumise alguskuupäev,
- laenutaotlus rahastati.

Täpsemate tulemuste saamiseks ja usaldusväärsemate andmete kasutamiseks on filtreeritud lähteandmeid veel selle järgi, kelle sissetulek ja kulud olid Bondora poolt kontrollitud (*VerificationType=4*). Pärast kitsendusi jäi vaatlusesse kokku 5397 lepingut.

Magistritöö analüüsi eesmärgist lähtuvalt on moodustatud uus uuritav nimitunnus *Default*, mis tähistab kliendi pankrotistumist väärtusega üks ning vastasel juhul null, kui ei pankrotistunud.
Pärast teostatud filteerimisi ja uuritava tunnuse lisamist jaotati andmestik juhuslikult kaheks osaks, vastavate osakaaludega 1/4 ning 3/4 test- ja treeningandmestikuks. Treeningandmestik sisaldab 4047 ja testandmestik 1350 vaatlust. Üldkogumis oleva 186 tunnuse seast on välja valitud 43 klienti kirjeldavat argumenttunnust, mis on välja toodud Lisas 1. Andmefailis kõikidest olevatest tunnustest valiti ainult sellised kliendilt kogutud algandmed, mis ei vihjaks Bondora poolt määratud intressimäärale ning mille põhjal oleks meil võimalik ise tuletada ning hinnata pankrotistumise tõenäosust.

Andmestikus esineb puuduvaid väärtuseid, kuid neid ei ole ühegi vaadeldava tunnuse puhul palju, jäädes enamasti alla 1%. Puudumist iseseisva väärtusena käsitlemist seetõttu vaadeldava andmestiku põhjal ei ole otstarbekas. Samas, puuduvate väärtuste olemasolu tekitab osade prognoosimeetodite korral probleeme, mistõttu otsustati need ka valimimahu säilitamiseks imputeerida ehk asendada. Puudvaid väärtusi on käsitletud alljärgnevalt.

Pidevate ja diskreetsete arvtunnuste korral on puuduvad väärtused asendatud treeningandmete põhjal leitud vastava argumenttunnuse aritmeetilise keskmisega. Mõned üldkogumis esinevad arvudega kodeeritud, kuid sisuliselt mittearvulised tunnused, näiteks *UseOfLoan*, *occupation_area*, *education_id*, *marital_status_id* ja *employment_status_id*, , on asendatud argumenttunnuste moodiga. Mittearvulised järjestustunnused on esmalt kodeeritud arvuliste väärtustega klassideks ning puuduvad väärtused antud juhul on asendatud kõige sagedamini esineva väärtusega. Testandmestikus esinenud puuduvad väärtused on asendatud treeningandmestiku tunnuste keskmistega või enim esinenud väärtusega, vastavalt argumenttunnuste iseloomule. Leiduvate seletamatute tasemetega tegutseti analoogselt puuduvatele väärtustele.

Selleks, et vähendada mittearvulise tunnuse faktortunnusena kasutamisest tingitud parameetrite arvu, valiti osade tunnuste puhul sisulistest kaalutlustest lähtuvalt kokkukuuluvate väärtuste rühma ning mudelites kasutati esialgse tunnuse asemel valitud rühmade baasil defineeritud indikaatortunnuseid. Näiteks arvudega kodeeritud väärtustega juhuslike suuruste puhul kasutati väärtuste rühmitamist indikaatorfunktsioonidega järgnevatel tunnustel: *marital_status_id* on jagatud kaheks sisulisel kaalutlusel- koosolijad ja üksikud või *education_id* korral on kokku võetud alg- ja põhiharidusega ning kutse- ja keskkooliharidusega kliendid. Tunnus

UseOfLoan on selekteeritud laenu refinantseerijate, soetatud materialistlike objektide ning tarbitud teenuste alusel. Argumenttunnsut *occupation_area* on proovitud jaotada vastavalt ameti kvalifikatsioonile, lähtudes peamiselt tööstus- ja avaliku sektori töötajatest. Tunnuse *employment_status_id* korral on eristatud täistööajaga töötajad pensionäridest, osakoormusega ja füüsilisest isikust ettevõtjatest.

Selliste mittearvuliste tunnuste puhul, mis vastavad mingi arvulise väärtuse kuulumisele teatud vahemikku, defineeriti vastav arvuline tunnus, kus väärtuseks võeti vaatlusele vastava vahemiku keskpunkt. Näiteks tunnuse *Employment_Duration_Current_Employer* erinevad tasemed on asendatud klassi keskmistega. Näiteks antud juhul kehtiva töösuhte ajavahemik jaotati kaheks tasemeks: [katseaeg; kuni kaks aastat] ja (kaks aastat; rohkem kui viis aastat], kus esimesel juhul asendati väärtusega 0.68 (natuke üle ühe aasta) ning teisel juhul 2.5 (kuni viis aastat).

2.2 Arvutatavate tunnuste lisamine

Krediidiriski paremaks kirjeldamiseks on kliendilt kogutud andmete põhjal lisatud arvutatavaid tunnuseid. Kasutatavas andmestikus leidus eelnevalt arvutatud tunnuseid, kuid kuna käesoleva töö autorile ei ole nende arvutamise algoritmid teada, on alustatud kliendilt kogutud andmetega ning juurde leitud uusi tunnuseid ja suhtarve, mis on välja toodud Lisas 2. Arvutatavate tunnuste leidmisel ei ole kasutatud Bondora poolt fikseeritud intressimäära, sest käsitletud pankrotistumise tõenäosus on üks oodatava kahju kompnonentidest, mille alusel hinnastatakse laenusoovija leping. Selgitavalt leitud tunnuste kohta:

- *income_total1* - kuine netosissetulek, mis on leitud kliendi kõikide sissetulekute summana. Selleks summeeriti andmestikust leitavad põhisissetulek, pension, peretoetused, sotsiaaltoetused, koondamistasud, lastetoetused ning

muud igakuised tasud.

- *DTI1* - protsentuaalne suurus, mis kirjeldab kliendi võlgnevuste ehk kohustuste osakaalu sissetulekust (*Debt to Income*). Suhtarv on leitud kuiste kohustuste kogusumma ja sissetuleku jagatisena.

Tavapäraselt laenusoovija krediidiriski kirjeldavad suhtarvud ei peaks omandama ühest suuremaid väärtuseid. Küll aga kasutatava andmestiku põhjal saab märgata juhte, kui näiteks *DTI1* suhe on suurem kui üks. Sellise nähtuse jaoks on töö autor sisse toodud lisatunnuse, mis põhjendaks laenuandmist klientidele ka suuremate võlgnevuste või finantseeritud laenu osakaaluga sissetulekusse. Mudeliehitusel on käsitletud tunnust *laenutingimused1* nii binaarse tunnusena või lihtsalt kordajana vastavalt sellele, mis andis treeningandmestikul ehitatud mudelile väiksema AIC väärtuse.

- *laenutingimused1* - binaarne tunnus, mis kirjeldab kinnisvara omandistaatuse ja lisasissetulekute, milleks on pensioni, peretoetuste, lastetoetuste või sotsiaaltoetuste olemasolu. Mudelisse listatud tunnus *laenutingimused1* põhjendab klientidele laenuandmist suuremate suhtarvude korral.

- *uusDTI* - võlgnevuste osakaalu sissetulekusse kirjeldav suhtarv, mis sisaldab tunnust *laenutingimused1*. Uus võlgnevuste osakaalu sissetulekusse kirjeldav suurus on leitud arvude $DTI1$ ja *laenutingimused1* korrutamisega.

- *Cash1* - igakuine kliendile kätte jääv vaba raha pärast kohustuste katmist. Vaba raha on leitud kuise sissetuleku ja kuiste kohustuste kogusumma vahena ning sisaldab tunnust *laenutingimused1*.

- *FAT1* - protsentuaalne suurus, mis kirjeldab finantseeritud laenu osakaalu sissetulekust. Arv on leitud finantseeritud laenu igakuise makse ja sissetuleku jagatisena, sisaldades paranduskordajat *laenutingimused1*.

- *NLMP1* - igakuine finantseeritud laenu osamakse vastavalt laenuperioodi pikkusele. Arv on leitud finantseeritud laenu ja laenuperioodi jagatisena.
- *NPTI1* - protsentuaalne suurus, mis kirjeldab kuise finantseeritud laenu osamakse osakaalu sissetulekusse. Arv on leitud kuise finantseeritud laenu ja sissetuleku jagatisena, sisaldades paranduskordajat *laenutingimused1*.
- *ATI1* - protsentuaalne suurus, mis kirjeldab taotletud laenu osakaalu sissetulekusse. Arv on leitud taotletud laenu igakuise makse ja sissetuleku jagatisena.
- *PerPerson* - summa sissetulekust, mis kirjeldab sissetulekut inimese kohta ilma kuiseid kohustusi arvestamata. Suurus on leitud kuise sissetuleku ja ülalpeetavate ning kliendi koguarvu jagatisena.
- *PerPerson2* - protsentuaalne suurus, mis kirjeldab kuiste kohustuste ja ülalpeetavatele ning kliendile minimaalse kuluva summa (sajaga korrutatud) osakaalu sissetulekusse. Arvu leidmisel on summeeritud kuised kohustused ning arvestuslikud kulud laenaja ja tema ülalpeetavate kohta (arvestusega 100 eurot inimese kohta) ning saadud summa on jagatud sissetulekuga.
- *DTC* - protsentuaalne suurus, mis kirjeldab kuiste kohustuste osakaalu vabasse rahasse. Arv on leitud kuiste kohustuste ja vaba raha jagatisena.

2.3 Prognoosimudeli headuse mõõdikud

Prognoosimudeli headuse mõõdikuid on erinevaid, kuid antud töös prognoosimudelite omavaheliseks võrdlemiseks testandmestikul on enamasti kasutatud keskmist ruutviga (*mean square error*), mis avaldub kujul

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(d_i - \hat{p}(x_i))^2, \tag{2.1}$$

kus suurus $\hat{p}(x_i)$ on vaadeldava mudeli abil leitud hinnang i-nda kliendi pankrotistumise tõenäosusele ning d_i konkreetse lepingu puhul juhusliku suuruse D_i realiseerunud väärtus ([5], lk 19).

Varasemast teame, et erinevate meetodite korral valemis 2.1 suuruseid võrreldes saame teha kindlaks, millise meetodi abil on leitud tõenäosusfunktsioon lähedasem tegelikule. Samas ei anna leitud veahinnang otseselt aimu sellest, kui hästi on prognoositud tõenäosused kooskõlas tegelikkusega erinevate kliendirühmade korral. Hea kooskõla tegelikkusega mitmesuguste oluliste kliendirühmade korral on täiendavaks garantiiks, et mudel töötab jätkuvalt hästi ka siis, kui klientide struktuur muutub, näiteks mingile kliendigrupile suunatud reklaami tõttu.

Selleks, et analüüsida prognooside käitumist erinevates kliendirühmades, on moodustatud erinevaid klientide alamrühmasid vanuse, soo, hariduse või sissetuleku alusel. Klassidesse jaotamisel on leitud hinnangud juhuslike suuruste D_i keskmisele vaadeldavas klassis (lepingu argumente x_i vaadeldakse fikseerituna, mis tähendab mittejuhuslikena) koos 95% usalduspiiridega eeldusel, et mudeli poolt leitud pankrotistumise tõenäosused vastavad tegelikule D_i jaotusele ning võrreldud saadud tulemusi empiirilise pankrotistunud lepingute protsendiga vaadeldavas klassis.

Tehtud eeldustel avaldub suuruste D_i keskmine vaadeldavas klassis kujul

$$E\left(\frac{1}{n}\sum_{i=1}^{n} D_i\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(D_i\right) = \frac{1}{n}\sum_{i=1}^{n} \hat{p}(x_i),$$

kus lihtsuse mõttes on eeldatud, et vaadeldavasse klassi kuulub n laenulepingut indeksitega 1, 2, $\cdots$, n. Kasutades juhusliku suuruse D_i sõltumatust ning disper-

siooni omadusi, saame

$$Var\left(\frac{1}{n}\sum_{i=1}^{n} D_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(D_i) = \frac{1}{n^2}\sum_{i=1}^{n}\left(\hat{p}(x_i))(1-\hat{p}(x_i)\right).$$

Eeldades, et juhuslike suuruste D_i keskmine on ligikaudu normaaljaotusega, mis on motiveeritud näiteks Lindbergi tsentraalse piirteoreemiga ([6], ptk 16.2), saame kirjutada

$$\frac{1}{n}\sum_{i=1}^{n} D_i \sim \mathcal{N}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{p}(x_i), \frac{1}{n}\sqrt{\sum_{i=1}^{n}(\hat{p}(x_i))(1-\hat{p}(x_i))}\right)$$

ning normaaljaotuse lähendil põhinev vahemikhinnag olulisuse nivool 0.95 pankrotistunud laenuvõtjate osakaalule vaadeldavas valemis on kujul

$$\frac{1}{n}\sum_{i=1}^{n}\hat{p}(x_i) \pm 1.96\frac{1}{n}\sqrt{\sum_{i=1}^{n}(\hat{p}(x_i))(1-\hat{p}(x_i))}. \tag{2.2}$$

Eelnevast järeldub, kui mingi kliendirühma korral leiame selle rühma keskmise pankrotistumise arvu ning see jääb toodud piiridesse, siis vaadeldav mudel ei tee süstemaatilisi vigu selle kliendirühma korral ja vastava kliendirühma osakaalu muutumine klientuuris ei põhjusta prognoosimudeli omaduste olulist halvenemist.

3 Logistiline regressioon pankrotistumise tõenäosuse hindamiseks

Eeldame, et i-nda kliendi pankrotistumist kirjeldav juhuslik suurus D_i on binoomjaotusega ehk $D_i \sim B(1, p(x_i))$, kus x_i tähistab vastava lepinguga seotud argumenttunnuseid ning p on mingi funktsioon.

Andmetega sobiva p leidmiseks on üheks levinud võimaluseks teha täiendav eeldus, et p on mingi konkreetsel lõplikust arvust parameetritest sõltuval kujul olev funktsioon ning seejärel valitakse parameetrid nii, et kooskõla andmetega oleks võimalikut hea. Sellisel lähenemisel baseeruvad paljud klassikalised prognoosimeetodid. Laialt levinud meetod binaarse uuritava tunnuse prognoosimiseks on logistiline regressioon. Antud magistritöös kasutatakse logistilist regressiooni kliendi laenulepingu pankrotistumise tõenäosuse prognoosimieks.

3.1 Logistiline regressioon

Antud alampeatükis, mis põhineb allikatel ([7], lk 50-51), ([8], lk 135-136) ja ([9], lk 110-111), käsitletakse logistilist regressiooni hinnatava pankrotistumise tõenäosuse leidmiseks. Logistiline regerssioon on üks üldistatud lineaarse regressiooni meetoditest.

Olgu meil iga laenulepingu jaoks komplekt argumenttunnuseid kujul

$$\mathbf{x}_i = (1, x_{i1}, x_{i2}, ..., x_{ip}),$$

kus vektori $\mathbf{x}_i$ algusesse on lisatud arv 1 ning mitterasvane x_i on tunnuste vektor ilma lisatud arvuta.
Vektor $\mathbf{x}_i$ koosneb erinevatest arv- ja kategoriaalsetest tunnustest. Antud töö kontekstis on vektori $\mathbf{x}_i$ korral tegemist laenutaotleja ja laenulepingu karakteristikute vektoriga, milleks võivad olla näiteks kliendi vanus, kuine sissetulek, ülalpeetavate arv perekonnas, eelnevate laenude arv jne.

Kuna tõenäosuse väärtused peavad kuuluma lõiku [0,1], siis ei ole mõistlik eeldada, et otsitav tõenäosus oleks argumenttunnuste lineaarne funktsioon. Logistilise

regressiooni ideeks on hinnata šansi logaritmi tunnuste lineaarkombinatsioonina. Ehk pankrotistumise šanss on pankrotistumise tõenäosuse jagatis mittepankrotistumise tõenäosusega, mis tähendab, et otsitakse krediidimudeli kordajate vektorit $\beta = (\beta_0, \cdots, \beta_p)$ nii, et kehtib

$$\log\left(\frac{p(x_i)}{1-p(x_i)}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} = \beta {\mathbf{x}_i}^T, \tag{3.1}$$

kus $\beta_0, \cdots, \beta_p$ tähistavad regressioonikordajaid ning $p(x_i) = Pr(D_i = 1)$ on vaadeldava kliendi pankrotistumise tõenäosus.

Võttes võrrandi 3.1 mõlemad pooled e astmesse, saame kirjutada

$$\frac{p(x_i)}{1-p(x_i)} = e^{\beta {\mathbf{x}_i}^T}.$$

Seega logistilise regressiooni korral eeldatakse, et sündmuse esinemise tõenäosus ehk prognoos tõenäosusele on avaldatav kujul

$$p(x_i) = \frac{e^{\beta {\mathbf{x}_i}^T}}{1 + e^{\beta {\mathbf{x}_i}^T}}. \tag{3.2}$$

Samaväärselt saame kirjutada

$$p(x_i) = \frac{e^{\beta_o + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}}{1 + e^{\beta_o + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}}. \tag{3.3}$$

Parameetrite $\beta = (\beta_0, \beta_1, \beta_2, \cdots, \beta_p)$ leidmiseks kasutatakse tõepära maksimismeerimise meetodit. Valimi tõepärafunktsioon n vaatluse korral avaldub kujul

$$L(\beta) = \prod_{i:D_i=1}^{n} p(x_i) \prod_{i':D_{i'}=1}^{n} (1 - p(x_{i'})), \tag{3.4}$$

kus parameetri β suurima tõepära hinnang on selline β väärtus, mis maksimiseerib antud tõepärafunktsiooni 3.4 ning saame

$$\hat{\beta}_{ST} = \arg\max_{\beta} L(\beta), \tag{3.5}$$

samaväärselt kirjutades

$$\max_{\beta} L(\beta) = L(\hat{\beta}_{ST}).$$

Siinjuures on kasulik aru saada, et logistilise regressiooni leitud funktsioon p ei ole peaaegu kunagi täpselt võrdne tegeliku funktsiooniga:

1. tehtud eeldus $\log\left(\frac{p(x)}{1-p(x)}\right)$ esitumise kohta argumentide lineaarkombinatsioonina ei pruugi olla õige

2. isegi, kui eeldus kuju kohta on õige, ei saa me andmete põhjal parameetreid hinnates peaaegu kunagi täpselt õigeid parameetreid.

3.2 Potentsiaalselt oluliste argumenttunnuste valik

Potentsiaalselt oluliste argumenttunnuste selekteerimisel on lähtutud nende sisulisest interpreteerimisest, tuginedes autori tunnetuslikule valikule eesmärgiga eristada parimad pankrotistumise tõenäosust kirjeldavad tunnused. Kasutatud tunnused on välja toodud koondtabelis Lisas 3.

Esimese sammuna valiti käsitsi kõikide argumenttunnuste hulgast potentsiaalselt olulised tunnused ning hinnati vajadust lisatunnuste kasutamiseks antud mudelis. Uuriti võimalikke koosmõjusid, kui oli alust arvata, et mingi tunnuse mõju pankrotistumise tõenäousele sõltuvalt mingi teise argumenttunnuse väärtustest ning toodi sisse paranduskordajaid. Kui treeningandmete põhjal oli alust arvata, et tõenäosuse käitumine vaadeldava argumenttunnuse muutumisel ei vasta logit funktsiooni käitumisele, siis kasutati šansside suhte kirjeldamisel kuupsplaine vaadeldavast argumendist. Splainidega lähendamisel on erinevate sõlmkohtade valikuks kasutatud lokaalset kaalutud regressiooniga silumist hinnatava tõenäosuse jaoks. Saadud graafikutelt valiti visuaalselt sobivaim sõlmede arv. Üldiselt antud

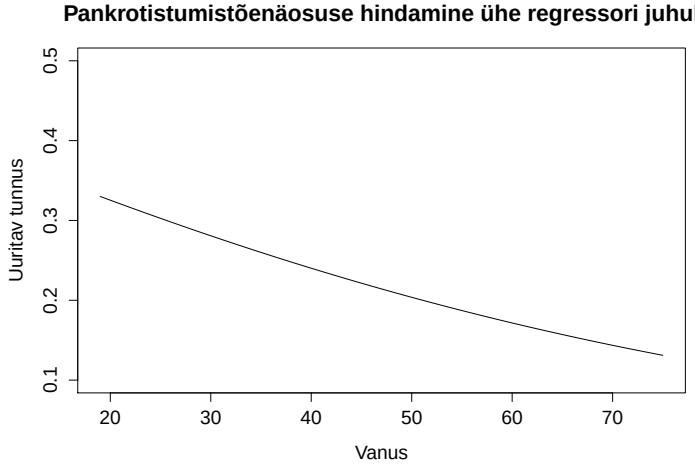
töös on eelistatud 1/3 ja 2/3 kvantiilide kasutamist, kuid mõnel juhul on sõlmkohad valitud käsitsi, vastavalt vähima AIC leidmiseni. Splainide kohta saab täpsemalt *Wikipediast* [10] ning baassplainidest [11].

Pankrotistumist potentsiaalselt mõjutavad tunnused on alljärgnevad:

- Vanus - laenutaotleja vanusel on oluline tähtsus laenulepingu pankrotistumise tõenäosuse hindamisel. On alust arvata, et nooremad laenutaotlejad võivad olla riskantsemad kliendid võrreldes vanemaealistega ning seetõttu pankrotistuvad ka suurema tõenäosusega.

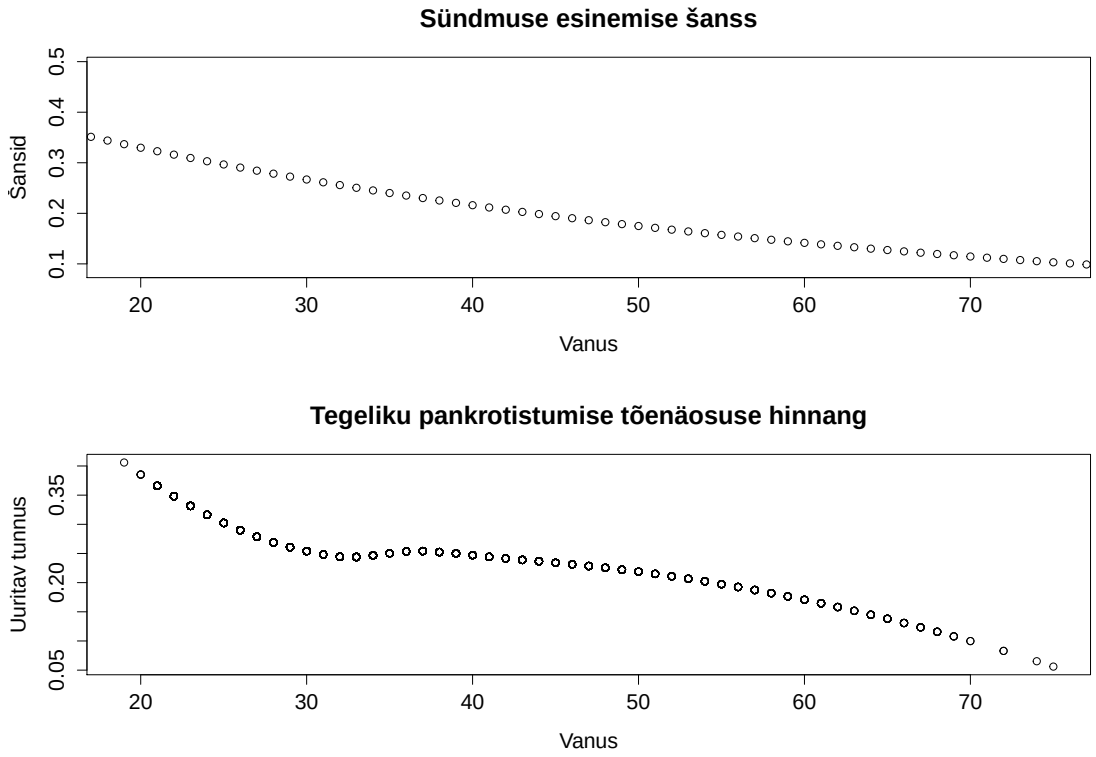
Interpreteerimise hõlbustamiseks on visandatud joonised 3.1 ja 3.2. Prognoosides i-nda laenulepingu pankrotistumise tõenäosust logistilise regressiooniga ühe regressori juhul, milleks olgu kliendi vanus, saab esitada kujul

$$\hat{p}(x_i) = \frac{e^{\beta_0+\beta_1 x_{i1}}}{1 + e^{\beta_0+\beta_1 x_{i1}}}. \tag{3.6}$$



Joonis 3.1: Seos uuritava tunnuse ja regressori vahel

Jooniselt 3.1 saame näha, et uuritava tunnuse ja regressori vaheline seos on monotoonselt kahanev ning tegemist on sujuva kõveraga. Tegeliku pankrotistumise tõenäosuse hinnanguks, sõltuvalt lepingu sõlmija vanusest, on kasutatud lokaalse kaalutud regressiooniga silumist, mille abil saadud tulemus on kujutatud joonisel 3.2, mis on kõrvutatud pankrotistumise šansside käitumisega.



Joonis 3.2: Sündmuse esinemise šansid ja tegeliku pankrotistumise tõenäosuse hinnang lokaalse regressiooniga

Jooniselt 3.2 on näha, et šansside käitumise korral ei ole tegemist lineaarse funktsiooniga ning tegeliku pankrotisumise tõenäosuse hindamisel lokaalse kaalutud regressiooniga märkame, et vanus mõjutab pankrotistumise tõenäosust erinevalt, eristades kiiremaid ja aeglasemaid kahanemisi. Sellest lähtuvalt šansside lähendamiseks kasutatakse keerulisemaid funktsioone ja antud töös tõenäosuse pa-

remaks kirjeldamiseks on kasutatud splaine. Põhiliselt keskendutakse naturaalsplainidele.
Eeldefineeritud regressori sõlmkohtade valikuks on kasutatud kvantiile, milleks antud juhul on vanused 28 ja 38. Naturaalsplainidega lähendatud funktsiooniga saadud tõenäosuse hinnangu näide vanuse korral on välja toodud peatükis 3.4.1.

Järgnevate potentsiaalselt oluliste argumenttunnuste uurimine on läbi viidud analoogselt eelnevale.

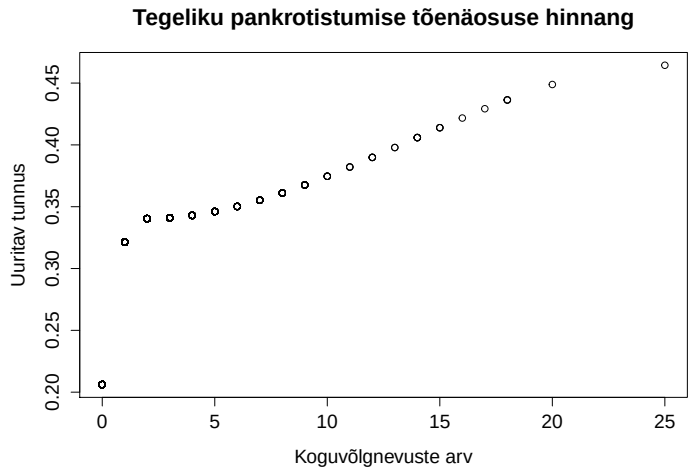
- Sissetulek - laenutaotleja kuine netosissetulek. On alust arvata, sissetulek ei mõjuta uuritavat tunnust ühesuguselt ning väiksema sissetulekuga kliendid võivad pankrotistuda suurema tõenäosusega kui suurema sissetulekuga kliendid. Sõlmkohtade valikuks on kasutatud kvantiile, milleks antud juhul on 600, 800 ja 1100 eurot.

- Võlgnevuste koguarv - võib arvata, et laenutaotlejad, kellel on rohkem võlgnevusi, võivad suurema tõenäosusega pankrotistuda.

Oluline on märkida, et mingist võlgnevuste arvust alates võib pankrotistumise tõenäosus ühesuguselt sõltuda võlgnevuste koguarvust arvust, mis tähendaks seda, et ei ole vahet kas kliendil on näiteks 15 ja enam võlgnevust. Alljärgneval joonisel 3.3 on visualiseeritud tegeliku pankrotistumise tõenäosuse hinnang lokaalse regressiooniga, mille põhjal on valitud käsitsi kaks sõlmkohta.

Sõlmpunktid antud juhul on valitud käsitsi, milleks on valitud 1 ja 3 võlgnevust.

- Perekonnaseis - kliendid on jaotatud kaheks rühmaks. Oluliseks on arvatud abielus ja kaaslasega kliendid, moodustades sellega esimese rühma; üksikud, lahutatud ning lesed kuuluvad teise rühma. Hüpoteetiliselt on oletatud, et kliendid, kes ei ela koos, võivad olla riskantsemad.



Joonis 3.3: Käsitsi hinnatud sõlmkohtade valik lokaalse regressiooniga

- Tegevusala - võib arvata, et tegevusvaldkond mõjutab mingil määral pankrotistumise tõenäosust. Vastavalt tegevusvaldkonnale on jaotuatud kliente gruppidesse oletatava teenitava sissetuleku põhjal.

- Haridus - on alust arvata, et kvalifitseeritud ja kõrgharitud inimesed pankrotistuvad väiksema tõenäosusega, mis on seotud ametikohaga ja teenitava sissetulekuga. Kliendid on eristatud kolmeks alamgrupiks, moodustades alg- ja põhiharidusega; kutse- ja keskkooli haridusega ning kõrgharidusega rühmad.

- Tööstaaž - loomuliku järjestuse esinemisel on otsustatud jätta tööstaaž ühe parameetrina mudelisse. Loomulik oleks mõelda, et inimesed, kellel on rohkem töökogemust, pankrotistuvad väiksema tõenäosusega ja vastupidisel juhul, väiksema töökogemusega laenuvõtjad pankrotistuvad suurema tõenäousega. Tööstaaži mõjutab kliendi vanus. Sõlmpunkt 0.5 on valitud käsitsi, mis tähendab kahe kuni viie aastast töökogemust ning võib arvata, et tööstaaži pikkus võib mõjutada oluliselt uuritavat tunnust seal, kus staaž on lühike. Mingist hetkest alatest tööstaaž ei pruugi enam nii suurt mõju tõenäosusele avaldada.

- Laenuperiood - pankrotistumise tõenäosuse prognoosimisel üks väga oluline näitaja on laenuperioodi pikkus, mida antud juhul on mõõdetud kuudes. Loomulik oleks arvata, et pikemaajalised laenud võivad suurema tõenäousega pankrotistuda kui lühilaenud. Sõlmpunktide valikul on kasutatud kvantiile, milleks on 24 ja 48 kuud.

- Teiste laenuvõlgnevuste kogusumma - sõltuva tunnuse tõenäosuse määramisel oluliseks faktoriks on teistest laenuettevõtetest saadud laenude kogusumma. Võib arvata, et suuremat laenukoormat kandev klient võib pankrotistuda suurema tõenäosusega võrreldes kliendiga, kellel ei ole eelnevaid laene. Sõlmpunkt on valitud käsitsi, milleks on 2000 eurot.

- Eelnevate laenutaotluste kogusumma - klientide poolt taodeldud krediitide kogusumma võib mõningal määral inditseerida teatavale käitumisele. Üldiselt võib oletada, et palju ja suuremaid laenutaotlusi esitanud kliendid võivad pankrotistuda suurema tõenäosusega. Sõlmpunktiks on antud juhul on valitud mediaan, milleks on 383.4 eurot.

- Võlgnevuste ja sissetuleku suhe - olulisemaks suhtarvuks pankrotistumise tõenäoususe kirjeldamisel on võlgnevuste ja sissetuleku suhe. Loomulik on eeldada, et mida suurem osa sissetulekust läheb võlgnevuste alla, seda suurem tõenäosus on kliendil pankrotistuda.

Vastavalt peatükis 2.2 käsitletule ei ole suhtarvu $DTI1$ eraldi vaadeldud. Koosmõjuna antud suhtarvule kasutatakse paranduskordajat *laenutingimused1*. Lisades mudelile suurusele $DTI1$ vastavale splainile ning kordajaga korrutatud teise $DTI1$ abil defineeritud splaini, saame mudeli, kus vaadeldav kordaja tagab ühtemoodi käitumise, kui *laenutingimused1* on 0 ning teist tüüpi käitumise, kui tunnus on 1. Üldiselt leitud paranduskordaja põhjendab laenu andmist ka väiksema sissetuleku ja suuremate võlgnevuste juhul. Selleks on valitud kinnisvara omamise staa-

tus või lisasissetulekute, näiteks pensioni, peretoetuste, lastetoetuste või sotsiaaltoetuste olemasolu. Mudelisse on valitud $DTI1$ ja *laenutingimused1* korrutis kujul *laenutingimuse1·ns(DTI1, df=3)* üle leitud suuruse $uusDTI$, andes mudelisse väiksema *Akaike* kriteeriumi väärtuse. Sõlmkohtade valikuks tunnuse $DTI1$ jaoks hinnati kaks sõlmkohta kvantiilidega väärtustega 0.487 ning 0.681.

- Taotletud laenu ja sissetuleku suhe - kliendi soovitud laenutaotluse osakaal kuisest sissetulekust võib anda vajalikku informatsiooni tõenäosuse hindamiseks. Loomulik oleks arvata, et suurema suhtarvu korral klient pankrotistub tõenäolisemalt. Sõlmkohad on valitud kvantiilidega, milleks on 0.058 ja 0.105.

- Ühele isikule kasutamiseks mõeldud summa kuisest sissetulekust, arvestamata kohustusi ja võlgnevusi - antud suurus kirjeldab kliendile ja ülalpeetavatele võimalikku kulutatavat summat isiku kohta. On alusta arvata, et väiksema kulutatava summa korral klient pankrotistub tõenäolisemalt. Funktsiooni paremaks lähendamiseks on sõlmkohad valitud käsitsi, milleks on 500 ja 1000 eurot.

- Laenuotstarve - laenu võtmise põhjus võib anda lisainformatsiooni kliendi kohta. Antud mudelis on laenuotstarve jaotatud materiaalsete väärtuste ja teenuste alusel kolmeks. On eristatud auto, kinnisvara soetanud või renoveerinud kliendid; laenu refinantseerijad ja muud ning reisimiseks, hariduseks, terviseks või äriks laenu võtnud kliendid.

- Kohustuste koguarv - pankrotistumise tõenäosuse hindamisel oluliseks suuruseks on kliendi kohustuste arv. Loomulik oleks eeldada, et suurema arvu kohustustega kliendid pankrotistuvad tõenäolisemalt kui väiksema kohustuste arvuga kliendid. Funktsiooni lähendamisel on hinnatud graafikult kaks sõlmkohta, milleks on vastavalt 5 ja 10 kohustust.

- Kiirlaenude koguarv - pankrotistumise tõenäosuse hindamisel väga oluliseks suuruseks võlgnevuste ja sissetuleku suhtarvu kõrval on kliendi eelnevate kiirlaenud arv. Rohkemate kiirlaenude arvuga kliendid pankrotistuvad tõenäolisemalt võrreldes klientidega, kellel on vähem kiirlaene. Sõlmpunkt on valitud käsitsi, milleks on 2 kiirlaenu.

3.3 Parima mudeli otsimise protseduur

Parima pankrotistumise tõenäosuse hindava mudeli leidmiseks on andmestik esmalt jaotatud test- ja treeningandmestikuks. Kõik argumenttunnuste sobivuseks langetatatud otsused põhinevad treeningandmestikul, seejuures testandmestikku parima mudeli leidmisel ei kaasata. Testandmestiku põhjal leitakse huvipakkuvad prognoosihinnangud. Lisatavate argumenttunnuste sobivuse mõõdikuks, mida laialdaselt kasutatakse krediidiriskimudelite ehitamisel, on *Akaike* informatsioonikriteerium (*Akaike Information Criterion*), mis on avaldatav kujul

$$AIC = 2 \cdot k - 2 \cdot lnL, \tag{3.7}$$

kus suurus lnL tähistab sobitatud mudeli logaritmilist tõepära ning k on parameetrite arv mudelis ([5], lk 45). Mudeli otsimise protseduur peamiselt lähtus vähima AIC väärtustega mudeli leidmisest ning arvestati parameetrite olulisust, võttes statistiliselt olulisuse kindlaks tegemisel p-väärtuseks 0.05.

3.4 Parim interaktiivselt leitud mudel

Parim pankrotistumise tõenäosust kirjeldav mudel on moodustatud treeningandmestikul 16 sisuliselt erineva argumenttunnuse põhjal. Vastavalt vajadusele on lähendatud funktsioone splainidega või sõltumatuid tunnuseid rühmitatud nende

karakterile. Mudeli väljavõte on Lisas 4.

Vähim saavutatud Aikaike informatsioonikriteerium antud mudeli korral on 3586.6. Peaaegu kõik sisuliselt erinevad lisatud tunnused on statistiliselt olulised. Saab märgata, et tunnus *UseOfLoan1* ei ole statistiliselt oluline, kuid selle eemaldamise tulemusena AIC väärtus lõppmudelis suureneb ning seetõttu on otsustatud antud tunnus sisse jätta.
Treeningandmestiku põhjal tõenäosuse hindamisel oluliseks kujunenud rühmadeks olid kliendid, kelle tegevusvaldkond oli seotud põllumajandusega, tootmisega või kinnisvaraga ning need, kes olid ajateenijad, meelelahutajad või transporttöölised. Lisaks statistiliselt oluliseks grupiks kujunes kutse- ja keskkooli haridusega kliendid, laenuotstarbe kohaselt materiaalse varaga ja abielustaatuse korral abielus ja partneritega kliendid.

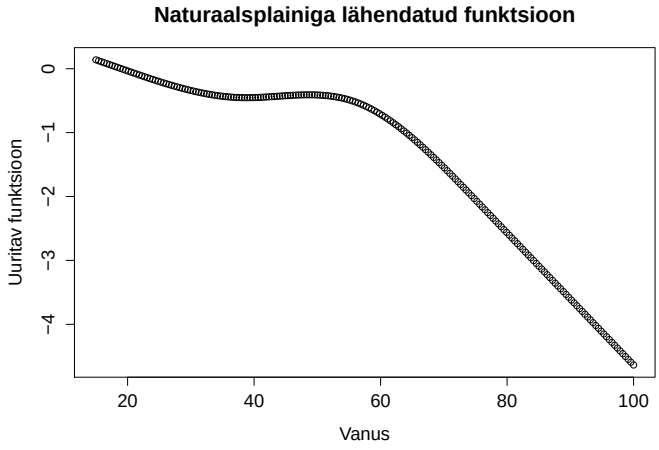
Antud mudelit on kontrollitud sammuviisilise regressiooniga mõlemas suunas, et kindlaks teha, kas leitud mudelit ei saa parandada ühe muutuja lisamise või eemaldamise teel parima pankrotistumise tõenäosust prognoosiva mudeli leidmiseks.

3.4.1 Leitud mudeli analüüs

Antud alampeatükis analüüsitakse mõningate näidete põhjal argumenttunnuste mõju laostumistõenäosuse prognoosile interaktiivselt leitud mudeli korral.
Regressioonikordaja ees olev postiivne märk näitab samapidist seost argumendi ja uuritava tunnuse vahel ning negatiivse kordaja korral on tegemist vastupidise seosega. Seejuures olgu mainitud, et negatiivne vabaliige, antud juhul $\beta_0 = -3.08415$, ei ole interpreteeritav ([9], lk 111). Kõiki parameetreid eraldi tõlgendada ei ole otstarbekas ja seetõttu vaadeldakse ainult huvipakkuvaid tunnuseid.

Lihtsama argumenttunnuste korral, näiteks *marital_status_id1*, saame negatiivse regressioonikordaja -0.27235 väärtuse põhjal väita, et inimesed, kes on abielus või elavad koos partneriga, pankrotistuvad vähem tõenäolisemalt kui need, kes elavad üksi. Tunnuse *UseOfLoan1* korral positiivne parameetriväärtuse 0.15583 põhjal saab väita, et kliendid, kes on on laenu võtnud korteri ostuks, renoveerimiseks või auto soetamiseks, pankrotistuvad tõenäolisemalt võrreldes teiste klientidega, kes on laenanud refinantseerimiseks või muudeks teenusteks.

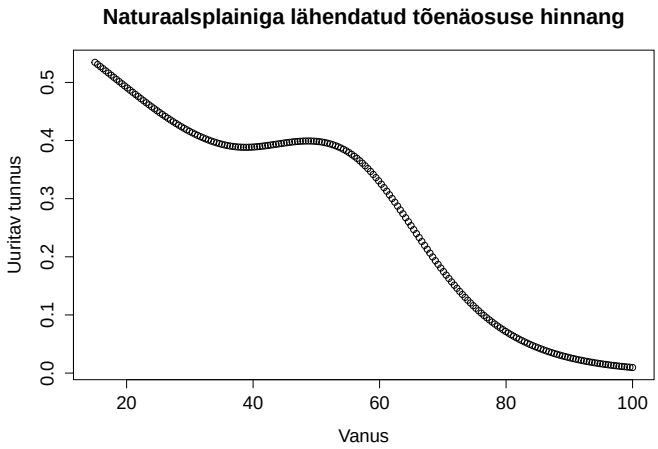
Teiste argumenttunnuste interpreteerimiseks splainidega lähendamise korral vaatleme üksikuid näited. Täpsemalt väljavõttest Lisas 5 saame näha, et argumenttunnus vanus on kvantiilidega 1/3 ja 2/3 jaotatud kahe sõlmkohaga kolmeks intervalliks: rajapunktist kuni esimese valitud sõlmkohani ehk lõiguks $[19, 28]$, $[28, 38]$ ning kliendid vanuses $[38, 75]$. Naturaalsplainiga lähendatud funktsioon asub alljärgneval joonisel 3.4, kus splaini baasmaatriksit on korrutatud interaktiivselt leitud mudeli regressioonikordajatega, milleks on 0.05732, -1.95445 ja -1.68853.



Joonis 3.4: Naturaalsplainiga lähendatud funktsioon

Saame märgata, et rajapunktides 19 ja 75 on splaini lähendatud lineaarfunkt-

siooniga ning siselõikudes kuupfunktsiooniga, mis defineeribki naturaalkuupsplainid. Saadud tõenäosuse hinnangud vastava näite kohaselt asuvad joonisel 3.5.



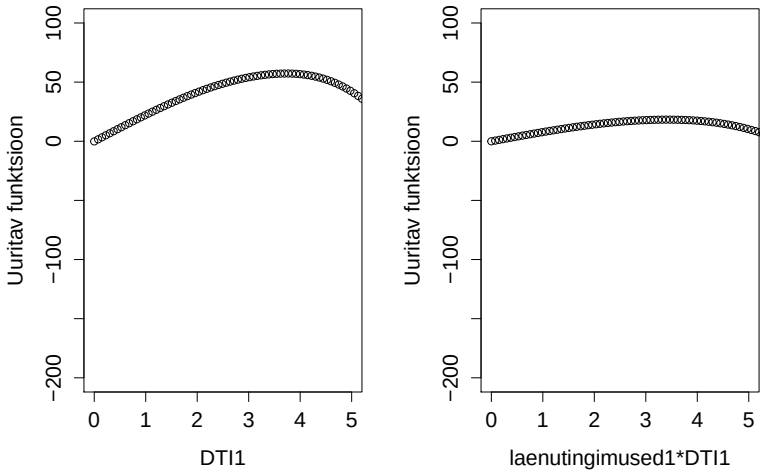
Joonis 3.5: Naturaalsplainiga lähendatud funktsioon

Jooniselt 3.5 saame näha, et üldiselt on tõenäosusel kahanev trend, kuid klientidel vanuses $[38, 50]$ tõenäosus pankrotistuda veidike suureneb vanuse kasvades.

Huvipakkuvamaks tunnuseks antud mudelis on suhtarv *DTI1*. Üldiselt võiks arvata, et suurema *Debt to Income* suhte korral kliendid pankrotistuvad tõenäolisemalt ja suhtarv võib omandada väärtuseid lõigust $[0, 1]$. Tegelikkuses antud andmestik ei kirjelda oodatavat käitumist 0.5 suuremate *DTI1* väärtuste korral. Laenu on antud ka klientidele, kelle *DTI1* on ühest suurem ning üllataval kombel suure *DTI1* suhtarvudega lepingute puhul on pankrotistumise protsent küllalt väike. Selle efekti võimalikuks seletuseks võib olla see, et suuremate *DTI1* väärtuste korral anti laenu enamasti oluliste lisagarantiide olemasolul. Kasutatava andmestiku põhjal võiks nendeks lisagarantiideks olla kinnisvara omamise staatus või lisasissetulekute olemasolu, mis on kokku võetud tunnusega *laenutingimused1* ning mudelil võimaldati kirjeldada erinevat *DTI1* mõju juhul, kui lisagarantii oli olemas ja ju-

hul, kui seda ei olnud.

Et näha, kuidas funktsioonide *DTI1* ja *laenutingimusted1·DTI1* mõjud vastavatel juhtudel avaldusid, vaatleme graafikuid 3.6. Kui teame splaini sõlmi ja vastavate baassplainide kordajaid, siis saame splaini välja arvutada igas punktis. Kui *laenutingimus1* on null, siis kirjeldab *DTI1* mõju splain kordajatega 2.47653, -333.10153 ja -658.27161, mille kuju on toodud vasakpoolsel joonisel. *Laenutingimused1* korral liituvad eelnevatele kordajatele koosmõjule vastavate baasfunktsioonide kordajad, seega tunnuse avaldatav mõju on sel juhul kirjeldatud parempoolsel joonisel. Olgu mainitud, et funktsioonkujude paremaks eristuseks on graafikut 3.6 suurendatud.



Joonis 3.6: *DTI1* ja *laenutingimusted1·DTI1*

Jooniselt 3.6 saame näha, et mõistlike (ühest väiksemate) *DTI1* väärtuste korral mõlemal juhul võlgnevuste osakaalu kasvamisel sissetulekusse suurendab pankrotistumise tõenäosust. Kuid, mõlemal juhul on näha ka efekti, et alates mingist *DTI1* väärtusest hakkab pankrotistumise tõenäosus kahanema, mis on vastuolus intuitsiooniga. Võib arvata, et selliste ekstremaalsete *DTI1* väärtuste korral on laenuandmise otsuse tegemisel kasutatud mingit täiendavat informatsiooni, mi-

da kättesaadavates andmetes ei ole. Samas aga eeldades, et ka tulevikus tehakse laenuandmise otsus samasugustel kaalutlustel, võib saadud mudel ikkagi sobida pankrotistumise tõenäosuse prognoosimiseks ka suurte *DTI1* suhtarvude korral. Ülejäänud tunnuste uurimiseks ja kirjeldamiseks kehtib analoogia.

3.5 Leitud sammuviisilise regressiooniga mudel

Krediidiriskimudelite koostamisel kasutatakse erinevaid automatiseeritud meetodid, millest üks on sammuviisiline regressioon. Antud alampeatükis kirjeldatakse automatiseeritud lähenemisega mudeli leidmise protseduuri ja proovitakse välja selgitada, kas sellel on eeliseid võrreldes intuitsioonil põhineval modelleeritud mudelil.

Sammuviisilise regressiooni meetodi korral on kasutatud pärast puuduvate väärtuste töötlemist saadud andmestikku, kuhu ei ole lisatud autoripoolseid krediidiriski kirjeldavaid tunnuseid. Tulemuste omavaheliseks paremaks võrdlemiseks on eelistatud kasutada täpse arvutusalgoritmi puudumise tõttu autori poolt arvutatud tunnuseid (*NLMP1*, *DTI1*, *ATI1*, *NPTI*, *Cash*, *FAT*), kuhu ei ole kaasatatud binaarset tunnust *laenutingimused1*. Ehk andmed on töötlemata kujul ning laenuandmist põhjendavat lisatunnust ei ole kasutatud.

Sammuviisilise regressiooni meetodiga on ehitatud kaks mudelit, mis on läbi viidud tarkvarapaketis *R*, funktsiooniga *step*. Esimeses mudelis, kuhu lisati kokku 49 töötlemata seletavat tunnust, kasutati suunda *both*, mis iga sammu korral otsutab *Akaike* kriteeriumi põhjal, kas lisada mõni tunnus mudelisse või eemaldada mõni olemasolev tunnus. Teise mudeli korral on kasutatud piirangut 16 olulisema tunnuse lõppmudelisse jätmiseks ning suunaga $forward$ ning analoogselt tunnuste lisamine/eemaldamine tehakse *Akaike* kriteeriumi kohaselt.

Esimeses mudelis sammuviisiline regressioon valis välja 29 statistiliselt olulist tunnust ja saadud Akaike kriteerium on 3738.2. Mudeli väljavõte on Lisas 5. Teise, 16 tunnusega saadud mudeli AIC on 3749.6 ning saadud mudeli väljavõte on Lisas 6.

3.6 Prognoosimudeli headuse otsustamine

Interaktiivse ja automatiseeritud logistilise regressiooni tulemused (AIC treeningandmetelt ning keskmine ruutviga testandmetelt) on esitatud koondtabelis 1.

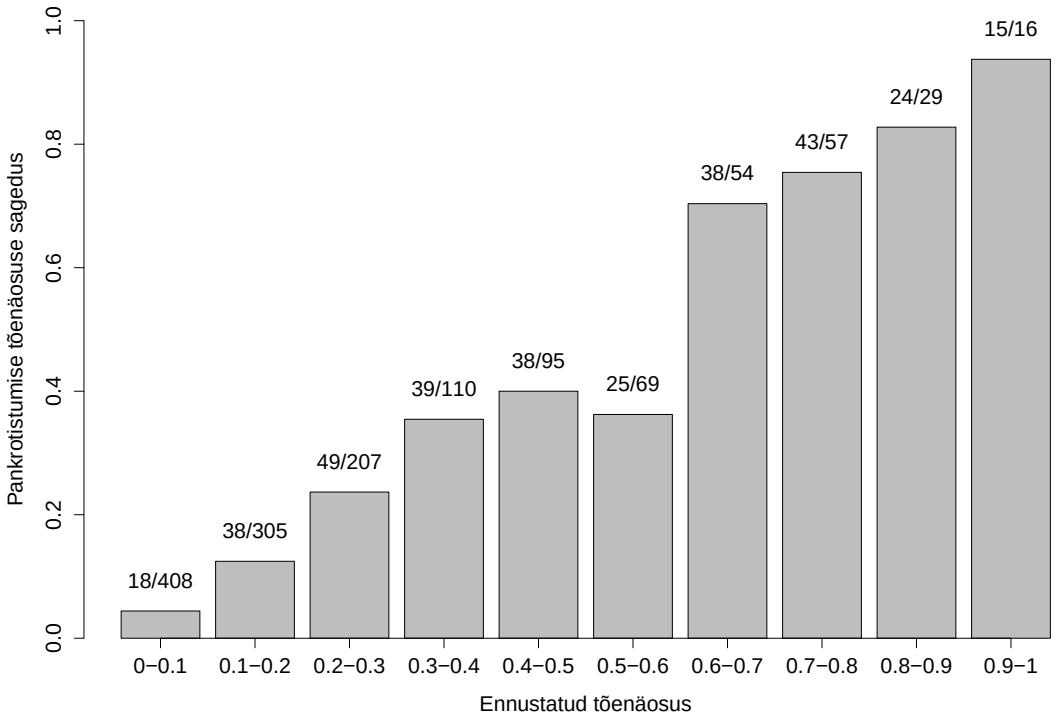
Mudel	AIC	MSE
Interaktiivne mudel	3586.6	0.1333
Sammuviisilise regressiooni mudel	3738.2	0.1383
Sammuviisilise regressiooni mudel (16 tunnust)	3749.6	0.1407

Tabel 1: Erinevate lähenemistega saadud mudelite võrdlus treening- ja testandmetel

AIC kriteeriumile ning leitud MSE väärtusele tuginedes võime väita, et interaktiivselt ehitatud mudel kaalub üle sammuviisilise regressiooniga leitud mudeli, saades vähimad väärtused mõlemal juhul. Antud tulemustest lähtuvalt prognoosimudeli hinnangud on leitud interaktiivselt ehitatud mudeli põhjal.

Interaktiivse lähenemisega leitud mudeli korral saadava prognoosimudeli headuse üle otsustamiseks on kasutatud alljärgnevaid meetodeid.

Esmalt illustreerime prognoositavad pankrotistumise tõenäosused ja tegeliku pankrotistumise tõenäosuse sagedused joonisel 3.7.



Joonis 3.7: Pankrotistumise tõenäosuse riskiklassid ja sagedus

Jooniselt 3.7 saame näha, et kliendid on jaotatud kümnesse riskiklassi. Näiteks, esimeses riskiklassis $0-0.1$ on ennustatud 408 kliendile, et nad pankrotistuvad tõenäosusega $0-0.1$. Tegelikult 408-st kliendist läks pankrotti 18, mis antud juhul teeb 4.4%. Võib öelda, et konkreetne hinnatud riskiklass kirjeldab hästi tegelikkust. Analoogselt eelnevale, riskigrupis $0.3-0.4$ on 110 kliendile ennustatud pankrotistumist tõenäosusega $0.3-0.4$, kuid tegelikult 110-st kliendist läks pankrotti 39 ehk 35.5%. Riskiklassis $0.5-0.6$ on 69 kliendile ennustatud pankrotistumist tõenäosusega $0.5-0.6$, neist tegelikult läks pankrotti 25 klienti ehk 36.2%. Antud riskirühmas on pankrotistumise tõenäosust ülehinnatud. Seevastu seitsmendas riskirühmas $0.6-0.7$ on tõenäosust natukene alahinnatud, saades tegelikuks pankrotistumise tõenäosuseks 70.1%. Järgnevate riskirühmade kirjeldamisel kehtib

analoogia ning üldiselt saame öelda, et prognoosimudel kirjeldab hästi tegelikkust väikeste erinevustega.

Võimalike kliendistruktuurimuutuste hindamiseks on kliente jaotatud valikuliselt erinevatesse klassidesse vanuse, sissetuleku, hariduse või soo järgi. Hinnatavad tõenäosused ja tegelikud sagedused koos vahemikhinnangutega on esitatud koondtabelis 2.

A	$\frac{\sum_{i\in A}\hat{p}(x_i)}{n}$	$\frac{\sum_{i\in A} d_i}{n}$	95%	95%	n
Kliendi klass	Sagedus		Usaldusintervall		Klassi suurus
	Prognoos	Tegelik	Alumine	Ülemine	
Sissetulek > 500	0.251	0.238	0.229	0.272	1162
Vanus ≤ 25, mehed	0.315	0.317	0.255	0.375	167
Vanus >25, mehed	0.219	0.218	0.189	0.249	544
Vanus $(30, 60)$, naised	0.274	0.246	0.235	0.314	350
Vanus ≤ 30	0.288	0.268	0.257	0.320	578
Sissetulek > 500, vanus < 50	0.256	0.243	0.233	0.278	1026
Sissetulek < 500, kutse- ja keskharidus	0.294	0.278	0.227	0.362	126
Sissetulek < 500, laenuperiood > 6	0.318	0.285	0.259	0.376	172
Sissetulek ≥ 500, laenuperiood > 6	0.264	0.252	0.242	0.287	1099
Sissetulek < 1200, kinnisvaraomand	0.283	0.250	0.255	0.311	688

Tabel 2: Prognoosimudeli hinnatud ja tegelik sagedus valitud kliendi klassidele usaldusintervallidega

Tabelist 2 saame näha, et kõikide rühmade korral tegelik pankrotistumise sagedus ei ole 95% piirides. Üldiselt saab mudeli kooskõla tegelikkusega vaadeldavates rühmades lugeda heaks ning mudeli prognoositud pankrotistumise protsendid rühmades on lähedased realiseerunud pankrotistumise protsentidele, jäädes kõigis

rühmades peale ühe usalduspiiridesse. Viimase kliendirühma korral on realiseerunud pankrotistumise arv ainult veidi usalduspiiridest väljas. Kuna vaadeldud rühmade arv on küllalt suur, siis sellise kõrvalkalde esinemine ühes rühmas oleks isegi mudeli põhjal genereeritud andmete põhjal üsna tõenäoline. Kui rühmad oleks ilma ühisosadeta, siis oleks kõikide rühmade korral piiridesse jäämise tõenäosus mudeli kehtivuse korral ainult $0.95^{10} \approx 0.6$. Kokkuvõtvalt saame öelda, et interaktiivselt ehitatud mudeli prognoosihinnangud testandmestikul on kooskõlas tegelikkusega ning mudel töötab hästi ka uute, mudeli sobitamisel mittekasutatud andmete korral.

4 *CART* meetod

Antud peatükis käsitletakse $CART$ meetodit uuritava tunnuse prognoosimiseks. Üheks võimaluseks uuritava ehk sõltuva tunnuse prognoosimiseks on otsustuspuu meetod, mis üldiselt koosneb argumenttunnuste abil moodustatud kriteeriumite, milleks on piirangud (nt tunnus on väiksem etteantud arvust) või faktrorite (tunnuse väärtus kuulub etteantud faktortasemete väärtuste hulka), hierarhilisest kogust. Otsuseid langetatakse kindlate kriteeriumite alusel, mis jagunevad faktoriteks ja piiranguteks. Vastavalt uuritava tunnuse karakterile saab jagada otsustuspuid klassifikatsiooni- ja regressioonipuudeks (*Classification and Regression Trees*), lühendatult *CART*. Klassifikatsioonipuude abil otsustatakse uuritava tunnuse klassikuuluvust ning regressioonipuude korral prognoositakse reaalarvulisi väärtuseid ([12], lk 118-119). Antud magistritöös keskendutakse ühemõõtmelistele regressioonipuudele, prognoosides kliendi laenulepingu pankrotistumist.

4.1 Prognoosimine rekursiivse binaarse tükeldamisega

Järgnevad alampeatükid põhinevad autorite Gareth James, Daniela Witten, Trevor Hastie ja Robert Tibshirani õpikul ([8], lk 303-311).
Olgu antud n vaatlust, millest igaüks sisaldab uuritava tunnuse väärtust d ning p argumenttunnustest koosnev vektor komponentidega $x_1, x_2, \cdots, x_p$. Uuritava tunnuse ehk pankrotistumise prognoosimine argumenttunnuste ruumi tükeldamisega on esitatud alljärgneva algoritmiga, mille eesmärgiks on leida parimad faktorid ning regressioonipuu suurus.

- Argumenttunnustest moodustatav ruum $(x_1, x_2, \cdots, x_p) \in H$, mille me jagame j-iks erinevaks mittekattuvaks piirkonnaks. Tähistame antud piirkonnad vastavalt $R_1, R_2 \cdots, R_J$.

- Iga vaatluse korral, mis satub piirkonda R_J, prognoositakse suuruse D jaoks sama väärtus. Selleks prognoosiks on uuritava tunnuse treeningandmete keskväärtus antud piirkonnas R_J.

Eelmainitud piirkondade $R_1, R_2 \cdots, R_J$ valikus võivad olla mistahes kujuga alad, kuid lihtsuse huvides argumenttunnuste ruum jaotatakse mitmemõõtmelisteks risttahukateks. Näiteks reaalarvulise argumenttunnuse korral jaotatakse ruum ruutudeks või ristkülikuteks.
Eesmärgiks on leida sellised ristkülikud $R_1, R_2 \cdots, R_J$, mis minimiseerivad jääkide ruutude summat RSS ($Residual\ sum\ of\ Squares$), mis on esitatav kujul

$$\sum_{j=1}^{J} \sum_{i \in R_j} (d_i - \hat{p}(x_i)_{Rj})^2, \tag{4.1}$$

kus $\hat{p}(x_i)_{Rj}$ tähistab uuritava tunnuse treeningandmestiku keskmist j-ndas ristkülikus.
Iga võimaliku seletavate tunnuste ruumi jaotust J-iks risttahukaks on arvutustlikult mahukas käsitleda, mistõttu lähenetakse antud probleemile ülalt-alla (*top-*

down) meetodiga ehk rekursiivse binaarse tükeldamisega (*recursive binary splitting*).

Otsustuspuu hargnemine algab juurest, kus vaadeldav ruum jagatakse kaheks tütarharuks. Igal tütarharul olev ruum jagatakse omakorda kaheks. Vaatleme täpsemalt olukorda, kus meil on juba olemas mingi komplekt ristkülikuid. Iga olemasolev ristkülik R_i jagatakse kaheks mingi argumenttunnuse järgi.
Kui vaatleme R_i jagamist arvulise või järjestustunnuse järgi, siis leiame sellise lõikekoha s, et

$$\{x \mid x_j < s\} \quad \text{ja} \quad \{x \mid x_j \geq s\}.$$

Selle tulemusena saame kaks uut ristkülikut $R_{ij1} = \{x \mid x_j < s\}$ ja $R_{ij2} = \{x \mid x_j \geq s\}$.

Kui vaatleme jagamist tunnuse x_j järgi, mis on faktortunnus, vaatleme hulga kõikvõimalikke jaotamisi kaheks alamhulgaks A ja B nii, et

$$R_{ij1} = \{x \mid x_j \in A\} \quad \text{ja} \quad R_{ij2} = \{x \mid x_j \in B\},$$

mille korral on

$$\sum_{i:x_i \in R_{ij1}} (d_i - \hat{p}(x_i))^2 + \sum_{i:x_i \in R_{ij2}} (d_i - \hat{p}(x_i))^2$$

minimaalne.
Kokkuvõttes valitakse R_i jagamiseks selline tunnus j, mille korral

$$\sum_{i:x_i \in R_{ij1}} (d_i - \hat{p}(x_i))^2 + \sum_{i:x_i \in R_{ij2}} (d_i - \hat{p}(x_i))^2 \tag{4.2}$$

on minimaalne.
Igal sammul valitakse jagamiseks välja selline ristkülik, mille parima j korral muutus

$$\sum_{i:x_i \in R_i} (d_i - \hat{p}(x_i))^2 - \sum_{i:x_i \in R_{ij1}} (d_i - \hat{p}(x_i))^2 - \sum_{i:x_i \in R_{ij2}} (d_i - \hat{p}(x_i))^2 \tag{4.3}$$

on kõige suurem (maksimaalne).
Antud protsessi korratakse kuni edasine jagamine ei ole enam võimalik. Näiteks, kui on saavutatud mingi fikseeritud ristkülikute arv, igas piirkonnas on ainult üks vaatlus või edasine jagamine tekitaks ristkülikuid, kus on vähem vaatlusi kui eelnevalt lubatud.

Piirkonna jagamist ristküliketeks võib esitada puuna, kus juureks on kogu piirkond, mis vastab kogu ruumile ehk ühele suurele ristkülikule. Seejärel liigutakse järjest allapoole ning iga jagamine tekitab ühe olemasoleva ristküliku asemele kaks uut ristkülikut. Seejuures jagamiskohta tähistatakse kahe uue haruga. Igast sõlmest läheb edasi kaks haru järgmiste sõlmede või lehtedeni. Sõlmedeks nimetatakse neid ristkülikud, mida on edasi tükeldatud. Regressioonipuu lehtedeks nimetatakse selliseid ristkülikud, mida enam edasi pole jaotatud. Lehed määravadki piirkonna tükeldamise.

Kui piirkonnad $R_1, \cdots, R_J$ on valitud, saame prognoosida funktsioontunnust antud testandmetel, kasutades treeningandmete keskväärtust selles piirkonnas, kuhu testandmestik kuulub.

4.2 Ristvalideerimise idee parameetritide valikuks

Rekursiivse binaarse tükeldamise meetod võib anda häid prognoose treeningandmetel, kuid regressioonipuu keerukusest treeningandmestikku ülesobitamine võib anda ebatäpseid tulemusi testandmetel. Selleks, et kasutada treeningandmestikku täielikult ning leida sobiv tükelduste arv, rakendatakse antud töös K-kordset ristavalideerimise metoodikat.

4.2.1 Regressioonipuu pügamine

Tähistagu T_0 regressioonipuud ning T selle alampuud, see tähendab, et kehtib $T \subset T_0$. Esmalt soovitakse treeningandmetel kasvatada võimalikult suur puu, millel kasutatakse rekursiivset binaarset tükeldamist ning peatutakse siis, kui kõikides lehtedes (jaotuse tükkides) on selline arv vaatlusi, et edasisel jagamisel jääks vähemalt ühes ristkülikus vaatluste arv väiksemaks kui etteantud minimaalne arv ühes tükis. Pügamise eesmärk on leida selline alampuu T, mis annab väikseima vea testandmetel. Veamäära leidmiseks kasutatakse ristvalideerimise metoodikat, mis kõikvõimalike alampuude moodustamise kriteeriume vaadates võib olla väga ajamahukas. Sellest lähtuvalt valitakse vaatlusse väiksem arv alampuid ning kasutatakse *Cost complexity pruning* kriteeriumit, et saada parim alampuude järjestus, mis on tähistatud mittenegatiivse reguleeriva parameetriga α. Iga α väärtusele vastab alampuu $T \subset T_0$ selliselt, et

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (d_i - \hat{p}(x_i)_{R_m})^2 + \alpha \mid T \mid \tag{4.4}$$

on minimaalne. Suurus $\mid T \mid$ tähistab alampuu lehtede arvu. R_m on m-s lõplik sõlm ehk leht argumenttunnuste alamruumis ja $\hat{p}(x_i)_{R_m}$ on prognoos, milleks on treeningandmete keskmine piirkonnas R_m.

Parameeter α (*tuning parameter*) reguleerib alampuude keerukust ja sobivust treeningandmetele. Kui $\alpha = 0$, siis $T = T_0$ ehk on saadud esialgne puu. Kui suurendame parameetrit α, siis puude harusid pügatakse ja saadakse väiksemad puud.

4.2.2 K-kordne ristvalideerimise idee

Valitud parameetriga mudeli täpsuse hindamiseks kasutatakse K-kordset ristvalideerimise metoodikat, mis tagab parima α leidmise. Selleks jaotatakse treeningandmestik juhuslikult K osasse. Iga $k = 1, \cdots, K$ korral

1. kasutatakse rekursiivset binaarset tükeldamist treeningandmestiku osasse k mittekuuluvatest vaatlustest moodustatud andmestiku põhjal ning peatutakse viimases sõlmes määratletud minimaalse vaatluste arvu korral

2. leitakse etteantud α korral parim puu

3. leitakse prognoosiviga k-ndasse osasse kuuluvate andmete korral (valideerimisviga).

Leitakse keskmine valideerimisviga ning regressioonipuu parameetriks valitakse selline α, mis minimiseerib keskmist ruutviga. Antud magistritöös on kasutatud 10-kordset ristvalideerimist.

4.3 *CART* rakendamine treeningandmetele

Järgnevas alampeatükis kirjeldatakse *CART* meetodi rakendamist treeningandmetele. Sobitamise protseduurides on kasutatud statistika tarkvaraprogrammi *R* pakette *rpart*, *rpart.plot* meetodiga *anova*.
CART meetodit rakendatakse kahel erineval andmestikul.
Esmalt on kasutatud pärast puuduvate väärtuste töötlemist saadud andmestikku, kuhu ei ole lisatud leitud paremaid krediidiriski kirjeldavaid kordajaid. Samuti on eelistatud kasutada autori kalkuleeritud tunnuseid, milleks olid *Cash*, *NPTI*, *NLMP1*, *ATI1* ja *FAT* ja *DTI1*, mis ei sisalda sisalda binaarset tunnust *laenutingimused1* (vt Lisast 2) ehk antud lähenemise korral *CART* meetodit rakendades ei ole andmestikku lisatud tunnused töödeldud kujul.
Esialgse T_0 regressioonipuu parameetri valikuks on kasutatud K-kordset ristvalideerimist, jaotades treeningandmestiku argumenttunnuse vaatlusruumi juhuslikult kümneks ligikaudselt võrdseks osaks, jättes igasse osasse antud juhul ligikaudu 405 vaatlust. Pärast korduvalt osadesse jaotamist ning parima keskmistatud valideerimishinnangu leidmist on valitud parameeter *cp*. Konkreetsel juhul parimaks

parameetriks on saadud $cp = 0.006$ ning antud parameetrit on kasutatud kogu treeningandmestikul parima regresioonipuu leidmiseks. Testandmestikult, mida ei kasutatud sobitamisel, hinnatakse leitud mudeli headust.

CART meetodit rakendades teise andmestiku korral on kasutatud lisaks töötlemata andmetele ka arvutatud tunnuseid (vt Lisa 2) ning kõiki sisulistel kaalutlustel rühmitatud faktortunnuseid analoogselt intuitiivsele logistilise regressiooni mudeli juhule. Veel on sisse jäetud paranduskordajat *laenutingimused1* sisaldavad suhtarvud, näiteks *uusDTI* või *FAT1*. Konkreetsel juhul parimaks pärast ristvalideerimishinnagu leidmist parameetriks on $cp = 0.006$.

Üllataval kombel on kahe erineva andmestikuga saadud samad regressioonipuud ning *CART* meetodit rakendades autoripoolsed arvutatud tunnused ei aita parema mudeli leidmisele kaasa. Kuna leitud puud on väikese arvu jagamistega, siis arvutatud parameetrite mõju ei tule mängu. Sellest lähtuvalt vaatleme edasi töötlemata andmestikuga saadud regressioonipuud, mis on välja toodud Lisas 7.

4.4 Parim *CART* mudel

Töötlemata andmestikuga saadud tulemused on esitatud koondtabelis 3, kuhu on lisatud võrdluseks valideerimishinnang.

Mudel	MSE	cp	*Valideerimishinnang*
CART mudel 1	0.154	0.006	0.1627

Tabel 3: Testandmestikul leitud MSE ja treeningandmestikul leitud parameeter ning valideerimishinnang

Lisast 7. saame näha, et olulisemateks tunnusteks on kujunenud *CountOfPaydayLoans*, *TotalMaxDebtMonths*, *AmountOfPreviousApplications*, *LoanDuration*, *DTI1* ja *TotalMonthlyLiabilities*.

4.4.1 Leitud mudeli analüüs

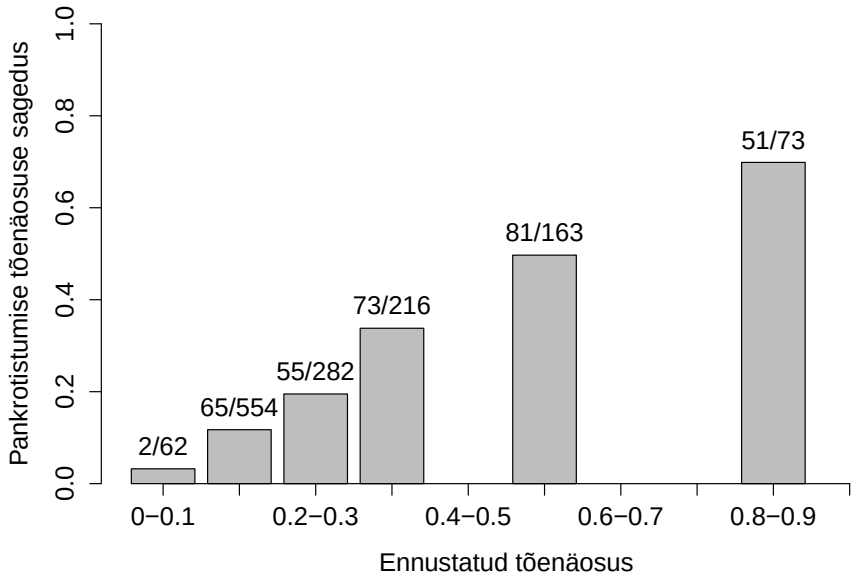
Antud alampeatükis käsitletakse lühidalt treeningandmestiku põhjal saadud regressioonipuud.

Puu alguses ehk juurel on kokku 4047 kliendi laenulepingut, millest tegelikult on pankrotistunud 26%. Konkreetse treeningandmestiku põhjal ning eeldusel, et klientide tuleviku käitumine on sama, saame väita, et regressioonipuu on eristanud vähemriskantsemateks klientideks tõenäosusvahemikus 0-0.1 need, kellel on vähem kui 1.5 kiirlaenu, maksimaalne võlgnevuses oldud aeg on suurem kui 1.5 kuud, laenuperiood on pikem kui 14 kuud ning võlgnevuste suhe sissetulekusse suurem kui 0.95 ning kliendid, kellel on rohkem kui 1.5 kiirlaenu ning $DTI1$ suhe on suurem kui 1. Vaatlustes olevatel lehekestel, antud tingimusi on täitunud mõlemal juhul kokku 191 klienti, mis on ligikaudu 4.7% kogu valimist. Analoogselt saame väita kasutatava andmestiku põhjal, et riskantsemad kliendid on need, kellel on rohkem kui 1.5 kiirlaenu, võlgnevuste suhe sissetulekusse on väiksem kui 0.74, laenuperiood suurem või võrdne 21 kuud ning kuine kohustuste summa väiksem kui 363 eurot. Antud tingimuste korral 192 kliendile on ennustatud pankrotistumist tõenäosusega 0.81.

Ülejäänud analüüs harudesse jaotamisel on analoogne eelnevale.

4.5 Prognoosimudeli headuse otsustamine

CART lähenemisega leitud mudeli prognoosimudeli headuse üle otsustamiseks on illustreeritud ennustatud klassi keskmiste ja tegeliku pankrotistumise tõenäosuse sagedused joonisel 4.1.



Joonis 4.1: *CART* meetodi pankrotistumise tõenäosuse riskiklassid ja sagedus

Sarnaselt paragrahvi 3.6 joonisele 3.7, kehtib *CART* meetodi riskirühmade kirjeldamiseks analoogia. Jooniselt 4.1 saame näha, et kümnest neljas riskiklassis 0.4-0.5, 0.6-0.7, 0.7-0.8 ning 0.9-1 ei ole pankrotistunud kliente. Esimeses riskiklassis $0-0.1$ on ennustatud 62 kliendile, et nad pankrotistuvad tõenäosusega $0-0.1$. Tegelikult 62-st kliendist läks pankrotti 2, mis on 3% ning see kirjeldab hästi hinnatud ja tegelikku tõenäosust. Riskigrupis $0.3-0.4$ on 216 kliendile ennustatud pankrotistumist tõenäosusega $0.3-0.4$, kuid tegelikult 216-st kliendist läks pankrotti 73, ehk 34%. Riskiklassis $0.8-0.9$ on 73 kliendile ennustatud pankrotistumist tõenäosusega $0.8-0.9$. Tegelikult pankrotistus 51 klienti ehk 69.9%, millest järeldub, et pankrotistumise tõenäosust on ülehinnatud. Järgnevate riskirühmade kirjeldamisel kehtib analoogia ning üldiselt saame öelda, rühma keskmine on paigas ning prognoosimudel kirjeldab tegelikkust.
Valikuliselt moodustatud kliendirühmad *CART* meetodi juhul on välja toodud koondtabelis 4.

A Kliendi klass	$\frac{\sum_{i\in A}\hat{p}(x_i)}{n}$ Sagedus	$\frac{\sum_{i\in A} d_i}{n}$	95% Usaldusintervall	95%	n Klassi suurus
	Prognoos	Tegelik	Alumine	Ülemine	
Sissetulek > 500	0.260	0.238	0.237	0.283	1162
Vanus ≤ 25, mehed	0.268	0.317	0.208	0.327	167
Vanus >25, mehed	0.251	0.218	0.218	0.284	544
Vanus $(30, 60)$, naised	0.270	0.246	0.229	0.311	350
Vanus ≤ 30	0.269	0.268	0.237	0.301	578
Sissetulek > 500, vanus < 50	0.262	0.243	0.237	0.286	1026
Sissetulek < 500, kutse- ja keskharidus	0.260	0.278	0.192	0.325	126
Sissetulek < 500, laenuperiood > 6	0.277	0.285	0.221	0.334	172
Sissetulek ≥ 500, laenuperiood > 6	0.268	0.252	0.245	0.291	1099
Sissetulek < 1200, kinnisvaraomand	0.269	0.250	0.240	0.298	688

Tabel 4: *CART* meetodi prognoosimudeli hinnatud ja tegelik sagedus valitud kliendi klassidele usaldusintervallidega

Tabelist 4 saame näha, et hinnatavad ja tegelikud tõenäosused on ligilähedased ning prognoositud tõenäosused jäävad lubatud usaldusvahemikku. Kokkuvõtvalt saame öelda, et leitud *CART* meetodi prognoosimudeli hinnangud on rahuldavad.

5 Parima mudeli valik

Eeldusel, et kliendirühmad on tulevikus samas proportsioonis, siis logistiline regressioon ja *CART* meetodid töötavad üldiselt hästi ning valikuliselt segmenteeritud kliendirühmad mõlema meetodi korral on stabiilsed muutuste suhtes. Koondtabel *CART* ja logistilise prognoosimudelite keskmise ruutvea (testandmestikult) jaoks

on alljärgnev.

Mudel	MSE
Interaktiivne mudel	0.1333
Sammuviisilise regressiooni mudel	0.1383
Sammuviisilise regressiooni mudel (16 tunnust)	0.1407
CART meetod 1	0.1545

Tabel 5: Erinevate lähenemistega saadud mudelite keskmise ruutvea võrdlus

Tabelist 5 saame näha, et prognoosimudeli keskmise ruutvea tulemused ei erine palju, kuid interaktiivselt ehitatud prognoosimudeli korral on saadud minimaalsem väärtus. Lisaks, leitud interaktiivne logistilise regressiooni mudel eristab riskiklasse täpsemalt ning seetõttu saame prognoosida pankrotistumise tõenäosust paremini. *CART* meetodi korral on prognoosimudeli tulemused kehvemad, eristades võimalikke riskiklasse vähem ning seetõttu prognoosib pankrotistumise tõenäosust robustsemalt.

Töös vaadeldud mudelite sobitamisel ei kasutatud Bondora andmestikus olevaid krediidiskoore. Küll aga võib arvata, et laenusoovijatele hinnatud krediidiskoor peaks olema tugevalt seotud pankrotistumise tõenäosusega. Seega pakub huvi, kas töös otseselt pankrotistumise tõenäosuse hindamiseks on väljatöötatud parim mudel, milleks on töö autor arvanud interaktiivselt tuletatud logistilise regressioonmudeli, võimaldab seda teha täpsemalt, kui vastava sündmuse prognoosimine krediidiskooride põhjal.

Selleks on võetud Bondora kodulehelküljelt kättesaadavast andmestikust ([1]) kaks krediidiskoori *Rating* ning *Rating_V2*, mis on erinevatel ettevõtte tegutsemise

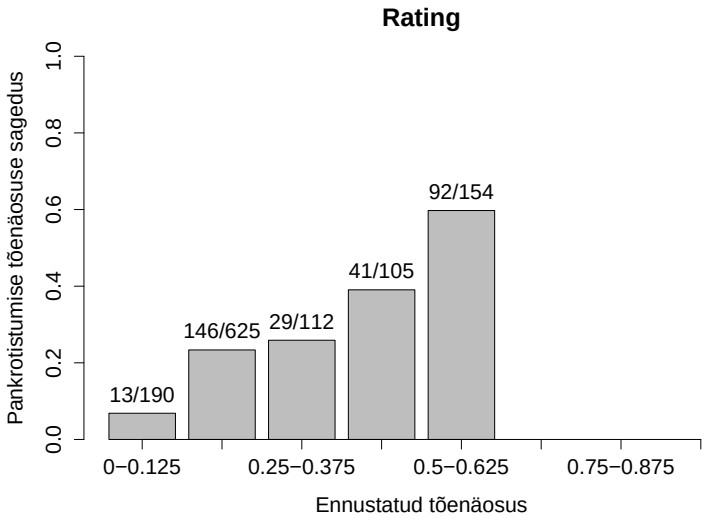
ajaperioodidel hinnatud. Andmeid on töödeldud ja filtreeritud analoogselt peatükis 2 käsitletule ning kasutati vaid neid vaatlusi, millele oli määratud Bondora poolne krediidiskoor. Üldkogum sisaldab kokku 4576 laenulepingut, mis on jaotati juhuslikult treening- ja testandmestikuks, sisaldades 3390 ja 1186 vaatlust.

Läbi viies analoogseid samme paragrahvis 3 kirjeldatule, kasutades mittelineaarse funktsiooni käitumise kirjeldamiseks splaine, on saadud kolm erinevat mudelit. Esimeses mudelis on kasutatud ühe regressorina hilisemat krediidiskoori *Rating*, teises mudelis uuemat *Rating_V2* ning kolmandas interaktiivselt ehitatud mudelit. Saadud tulemused *Akaike* kriteeriumite väärtuste ja keskmiste ruutvigade järg on välja toodud alljärgnevas tabelis 6.

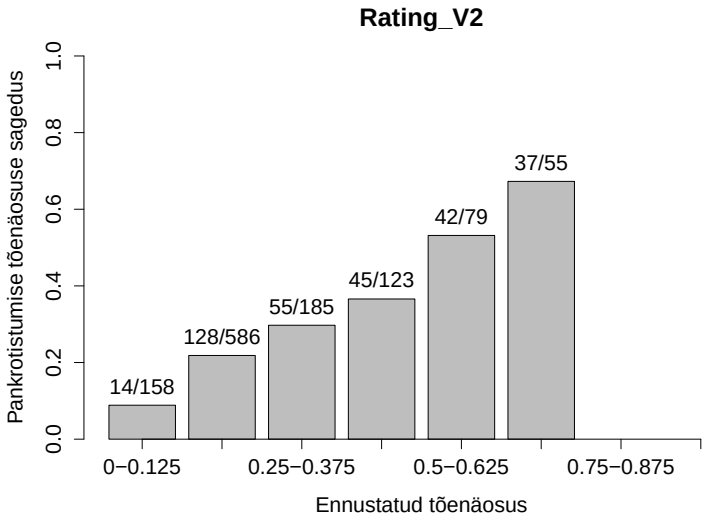
Mudel	AIC	MSE
Rating regressorina	3399.1	0.175
Rating_V2 regressorina	3408.4	0.178
Interaktiivne mudel	2778.4	0.151

Tabel 6: Bondora krediidiskooriga ning interaktiivse mudeliga saadud AIC (treeningandmestik) ning MSE (testandmestik) väärtused

Krediidiskooridele vastavad prognoosimudelite hinnangud asuvad alljärgnevatel joonistel 5.1 ja 5.2.

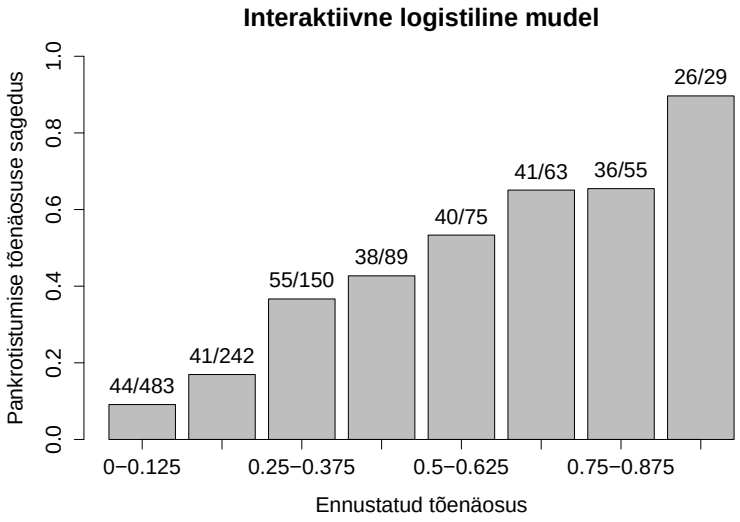


Joonis 5.1: Pankrotistumise tõenäosuste riskiklassid ja sagedus vanema krediidiskoori järgi



Joonis 5.2: Pankrotistumise tõenäosuste riskiklassid ja sagedus uuema krediidiskoori järgi

Interaktiivselt tuletatud prognoosimudeli graafik asub joonisel 5.3.



Joonis 5.3: Pankrotistumise tõenäosuste riskiklassid ja sagedus interaktiivse mudeliga

Tabelist 6 ning joonistelt 5.1, 5.2 ja 5.3 saadavate tulemuste põhjal võib väita, et antud magistritöös tuletatud mudeleid kasutades on lootust välja töötada krediidiskoori arvutamise süsteem, mis annab täpsema hinnangu kliendiga seotud riskidele, kui seni kasutusel olnud süsteemid.

Kokkuvõte

Käesolevas magistritöös on vaadeldud mitmeid aspekte- kasutatavate tunnuste väljatöötamist parema krediidiriski mudeli saamiseks ning mudeli ehitust erinevate lähenemismeetoditega, interaktiivselt kui ka automatiseeritult. Interaktiivsel lähenemisel kasutati logistilist regressiooni, mittelineaarse sõltuvuse kirjeldamiseks argumenttunnustest kasutati vajadusel splainidega lähendamist ning mudelisse valiti argumenttunnused vastavalt sobitaja arusaamisele. Automatiseeritud lähenemise korral kasutatati sammuviisilist regressioonmeetodit töötlemata andmetele ja regressioonipuud *CART* töödeldud ja töötlemata andmetele. Tihtipeale on levinud arvamus, et automatiseeritud fikseeritud algoritmiga mudelid töötavad paremini. Küll aga vähest eeltööd nõudvad automatiseeritud mudelid ei pruugi häid tulemusi anda, milleks osutus regressioonipuu meetod *CART*. Antud töös leiti kinnitust sellele, et interaktiivse lähenemisega on võimalik leida veidi parem mudel. Järelikult andmete eeltöötlus ja lisatunnuste leidmine võib anda olulist efekti.

Parima mudeli otsimise protseduuris argumenttunnuste sobivuseks langetatud otsused põhinevad treeningandmestikul ning headuse mõõdikuks kasutatakse *Akaike* informatsiooni kirteeriumit. Prognoosimudelite omavaheliseks võrdlemiseks testandmestikul on kasutatud keskmist ruutviga. Prognooside erinevate kliendirühmade käitumise analüüsimiseks on kliendistruktuurimuutuste jaoks moodustatud klientide alamrühmasid peamiselt vanuse, soo või sissetuleku alusel. Viimases peatükis on vaadeldud ka Bondora süsteemi poolt tuletatud krediidiskoore, et võrrelda, kas pankrotistumise tõenäosuse hindamiseks on väljatöötatud parim mudel. Saadud tulemuste põhjal võib väita, et antud magistritöös tuletatud mudeleid kasutades on võimalik välja töötada krediidiskoori arvutamise süsteem, mis annab täpsema hinnangu kliendiga seotud riskidele.

A Lisad

A.1 Tunnused ja seletus

Kasutatud tunnused	
Tunnus	Seletus
Age	Laenutaotleja vanus, D
TotalNumDebts	Võlgnevuste koguarv, D
TotalMaxDebtMonths	Maksimaalne võlgnevuses oldud periood, P
TotalLiabilitiesBeforeLoan	Kogu kohustuste arv, D
TotalMonthlyLiabilities	Kuine kohustuste summa, P
NumDebtsFinance	Tähtaja ületanud maksete arv, D
MaxDebtMonthsFinance	Maksimaalne tähtaja ületanud maksete periood, P
NumDebtsTelco	Telekommunikatsiooni teenuste võlgnevuste arv, D
MaxDebtMonthsTelco	Maksimaalne telekommunikatsiooni teenuste eest võlgnevuses oldud periood, P
NumDebtsOther	Teiste võlgnevuste arv, D
MaxDebtMonthsOther	Maksimaalne teistes võlgnevustes oldud periood, P
AppliedAmount	Taotletud summa, P
FundedAmount	Rahastatud summa, P
UseOfLoan	Laenu sihtotstarve, F, T=9
income_from_principal_employer	Põhitöötasu, P
IncomeFromPension	Pension, P
IncomeFromFamilyAllowance	Peretoetus, P
IncomeFromSocialWelfare	Sotsiaaltoetused, P
IncomeFromLeavePay	Koondamistasud, P
IncomeFromChildSupport	Lastetoetus, P
income_other	Muu igakuine tasu, P
DebtLiabilitiesBeforeLoan	Võlgnevuskohustuste arv enne laenu, D

OtherLiabilitiesBeforeLoan	Teiste kohustuste arv enne laenu, D
CountOfBankCredits	Pankade poolt väljastatud kohustuste arv, D
SumOfBankCredits	Pankade poolt väljastatatud kohustuste kogusuurus, P
CountOfPaydayLoan	Kiirlaenude koguarv, D
SumOfPaydayLoans	Kiirlaenude kogusuurus, P
CountOfOtherCredits	Teiste kohustuste arv, D
SumOfOtherCredits	Teiste kohustuste kogu suurus, P
marital_status_id	Perekonnaseis, F, T=5
Gender	Sugu, B
Occupation_area	Tegevusala, F, T=19
education_id	Haridus, F, T=5
employment_status_id	Ametinimetus, F, T=6
Employment_Duration_Current_Employer	Kehtiva töösuhte aeg, F, T=7
nr_of_dependants	Ülalpeetavate arv, D
work_ experience	Tööstaaž, F, T=6
home_ ownership_ type_ id	Omandisuhe elukohaga, F, T=10
LoanDuration	Laenuperiood, P
NoOfPreviousApplications	Eelnevate laenutaotluste arv, D
AmountOfPreviousApplications	Eelnevate laenutaotluste kogusumma, P
NoOfPreviousLoans	Teiste laenuvõlgnevuste arv, D
AmountOfPreviousLoans	Teiste laenuvõlgnevuste kogusumma, P

Tähistuste selgitused: P- pidev, D- diskreetne, F- faktortunnus, B-binaarne, T- tasemete arv

A.2 Arvutatud uued tunnused ja suhtarvud selgitustega

Arvutatud tunnused ja suhtarvud	
Tunnus	Seletus
income_total1	Neto sissetulek kokku (kuine)
DTI1	(*Debt-to-Income*) võlgnevuste ja sissetuleku suhe (kuine)
laenutingimused1	Leitud laenuandmist põhjendav paranduskordaja, koosneb koduomamise staatuse ja lisasissetulekute koosmõjust
Cash	Vabaraha pärast kohustuste katmist (kuine)
Cash1	Vabaraha pärast kohustuste katmist (kuine), sisaldab paranduskordajat
uusDTI	Paranduskordajat sisaldav võlgnevuste ja sissetuleku suhe (kuine)
FAT	(*Founded-amount-to-Income*) rahastatud laenu ja sissetuleku suhe (kuine) laenuintresse arvestamata
FAT1	(*Founded-amount-to-Income*) rahastatud laenu ja sissetuleku suhe (kuine) laenuintresse arvestamata, sisaldab paranduskordajat
ATI1	(*Applied-amount-to-Income*) taotletud laenu ja sissetuleku suhe (kuine)
NLMP1	(*New-Loan-Monthly-Payment*) taotletud laenu igakuine makse laenuintresse arvestamata
NPTI	(*New-Payment-to-Income*) taotletud laenu igakuise makse ja sissetuleku suhe laenuintresse arvestamata
NPTI1	(*New-Payment-to-Income*) taotletud laenu igakuise makse ja sissetuleku suhe laenuintresse arvestamata, sisaldab paranduskordajat

PerPerson	Ühele isikule kasutamiseks mõeldud summa kuisest sissetulekust ilma kohustusi ja võlgnevusi arvestamata
PerPerson2	Võlgnevuste ja kohustuste ning ülalpeetavatele kuluv keskmise rahasumma osakaal kuisest sissetulekust
DTC	(*Debt-to-Cash*) Võlgnevuste suhe vabarahasse

A.3 Potentsiaalselt olulised tunnused koos tähistustega

Valitud tunnus ja tähistus	
Tunnus	Nimi
Vanus	*ns(Age, df=3)*
Sissetulek	*ns(income_total1, df=4)*
Võlgnevuste koguarv	*ns(TotalNumDebts, knots=c(1,3))*
Perekonnaseis	*marital_status_id1*
Tegevusala	*occupation_area19123*, *occupation_area8151718*
Haridus	*education_id34*
Tööstaaž	*work_experience*
Laenuperiood	*ns(LoanDuration, df=3)*
Teiste laenuvõlgnevuste kogusumma	*ns(AmountOfPreviousLoans, knots=2000)*
Eelnevate laenutaotluste kogusumma	*ns(AmountOfPreviousApplications, df=2)*
Võlgnevuste osakaal sissetulekusse koos laenuandmist põhjendava kordajaga	*laenutingimused1*·*ns(DTI1, df=3)*

Ühele isikule kasutamiseks mõeldud summa kuisest sissetulekust ilma kohustusi ja võlgnevusi arvestamata	*ns(PerPerson, knots=c(500,1000))*
Kohustuste koguarv	*ns(TotalLiabilitiesBeforeLoan, knots=c(5,10))*
Kiirlaenude koguarv	*ns(CountOfPaydayLoans, knots=2)*
Laenusihtotstarve	*UseOfLoan1*

A.4 Parima mudeli väljavõte

```
Call:
glm(formula = Default ~ ns(Age, df = 3) + ns(income_total1, df = 4) +
    ns(TotalNumDebts, knots = c(1, 3)) + marital_status_id1 +
    occupation_area19123 + occupation_area8151718 + education_id34 +
    ns(work_experience, knots = 0.5) + ns(LoanDuration, df = 3) +
    ns(AmountOfPreviousLoans, knots = 2000) + laenutingimused1:ns(AmountOfPreviousApplications,
    df = 2) + ns(DTI1, df = 3) + laenutingimused1:ns(DTI1, df = 3) +
    ns(ATI1, df = 3) + ns(PerPerson, knots = c(500, 1000)) +
    ns(TotalLiabilitiesBeforeLoan, knots = c(5, 10)) + ns(CountOfPaydayLoans,
    knots = 2) + UseOfLoan1, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3083  -0.7042  -0.4094   0.5061   2.8199

Coefficients:
                                                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                                                      -3.08415    0.65578  -4.703 2.56e-06 ***
ns(Age, df = 3)1                                                  0.05732    0.27208   0.211 0.833154
ns(Age, df = 3)2                                                 -1.95445    0.51540  -3.792 0.000149 ***
ns(Age, df = 3)3                                                 -1.68853    0.48562  -3.477 0.000507 ***
ns(income_total1, df = 4)1                                       -0.42093    0.52543  -0.801 0.423068
ns(income_total1, df = 4)2                                       -2.07303    1.06254  -1.951 0.051054 .
ns(income_total1, df = 4)3                                        6.41565    3.40795   1.883 0.059761 .
ns(income_total1, df = 4)4                                       13.73126    6.80282   2.018 0.043543 *
ns(TotalNumDebts, knots = c(1, 3))1                               2.66789    0.51767   5.154 2.55e-07 ***
ns(TotalNumDebts, knots = c(1, 3))2                               2.60679    0.60728   4.293 1.77e-05 ***
ns(TotalNumDebts, knots = c(1, 3))3                               0.95804    1.27682   0.750 0.453053
marital_status_id1                                               -0.27235    0.09702  -2.807 0.004998 **
occupation_area19123                                              0.40528    0.10649   3.806 0.000141 ***
occupation_area8151718                                            0.31866    0.11092   2.873 0.004067 **
education_id34                                                   -0.18801    0.08686  -2.165 0.030416 *
ns(work_experience, knots = 0.5)1                                -0.80353    0.24794  -3.241 0.001192 **
ns(work_experience, knots = 0.5)2                                -0.13615    0.18054  -0.754 0.450786
ns(LoanDuration, df = 3)1                                         2.49005    0.23158  10.752  < 2e-16 ***
ns(LoanDuration, df = 3)2                                         5.08193    0.59361   8.561  < 2e-16 ***
ns(LoanDuration, df = 3)3                                         1.33902    0.12422  10.779  < 2e-16 ***
ns(AmountOfPreviousLoans, knots = 2000)1                          0.95033    0.56446   1.684 0.092258 .
ns(AmountOfPreviousLoans, knots = 2000)2                         -0.84744    1.35179  -0.627 0.530722
ns(DTI1, df = 3)1                                                 2.47653    3.74132   0.662 0.508010
ns(DTI1, df = 3)2                                              -333.10153  112.86577  -2.951 0.003164 **
ns(DTI1, df = 3)3                                              -658.27161  226.04974  -2.912 0.003590 **
ns(ATI1, df = 3)1                                                 0.80684    0.61960   1.302 0.192855
ns(ATI1, df = 3)2                                                 7.67441    1.98717   3.862 0.000112 ***
ns(ATI1, df = 3)3                                                10.45886    3.98765   2.623 0.008721 **
ns(PerPerson, knots = c(500, 1000))1                              2.35069    1.52263   1.544 0.122628
ns(PerPerson, knots = c(500, 1000))2                            -26.46344   13.26280  -1.995 0.046009 *
ns(PerPerson, knots = c(500, 1000))3                            -52.29745   26.80477  -1.951 0.051051 .
ns(TotalLiabilitiesBeforeLoan, knots = c(5, 10))1                 0.85987    0.35982   2.390 0.016862 *
ns(TotalLiabilitiesBeforeLoan, knots = c(5, 10))2                 3.52488    0.87485   4.029 5.60e-05 ***
ns(TotalLiabilitiesBeforeLoan, knots = c(5, 10))3                 3.03069    1.42477   2.127 0.033408 *
ns(CountOfPaydayLoans, knots = 2)1                                2.16116    0.40665   5.315 1.07e-07 ***
ns(CountOfPaydayLoans, knots = 2)2                                0.54251    0.79384   0.683 0.494359
UseOfLoan1                                                        0.15583    0.09572   1.628 0.103516
laenutingimused1:ns(AmountOfPreviousApplications, df = 2)1     -111.40003   34.91656  -3.190 0.001420 **
laenutingimused1:ns(AmountOfPreviousApplications, df = 2)2     -248.07524   76.02800  -3.263 0.001103 **
laenutingimused1:ns(DTI1, df = 3)1                              -10.45254    4.19750  -2.490 0.012768 *
laenutingimused1:ns(DTI1, df = 3)2                              186.27606  123.61858   1.507 0.131846
laenutingimused1:ns(DTI1, df = 3)3                              378.19515  247.78469   1.526 0.126934
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4650.4  on 4046  degrees of freedom
Residual deviance: 3502.6  on 4005  degrees of freedom
AIC: 3586.6

Number of Fisher Scoring iterations: 12
```

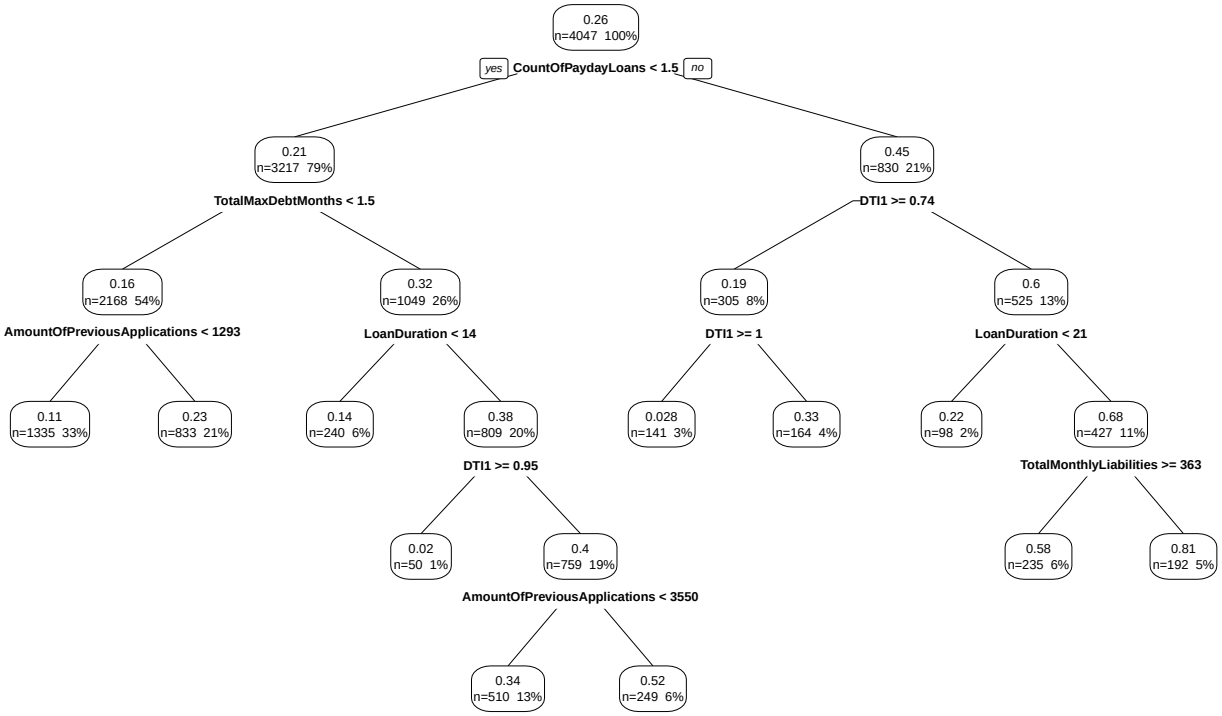
A.5 Sammuviisilise regressiooni mudeli 1 väljavõte

```
Call:
glm(formula = Default ~ Age + TotalNumDebts + TotalMaxDebtMonths +
    NumDebtsFinance + NumDebtsTelco + MaxDebtMonthsTelco + MaxDebtMonthsOther +
    AppliedAmount + UseOfLoan + income_from_principal_employer +
    IncomeFromPension + IncomeFromFamilyAllowance + IncomeFromChildSupport +
    income_other + TotalLiabilitiesBeforeLoan + TotalMonthlyLiabilities +
    CountOfPaydayLoans + marital_status_id + education_id + employment_status_id +
    Employment_Duration_Current_Employer + nr_of_dependants +
    work_experience + LoanDuration + NoOfPreviousApplications +
    AmountOfPreviousApplications + NoOfPreviousLoans + AmountOfPreviousLoans +
    DTI1, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2697  -0.7160  -0.4697   0.4782   2.8092

Coefficients:
                                     Estimate Std. Error z value Pr(>|z|)
(Intercept)                        -9.827e-01  3.588e-01  -2.739 0.006166 **
Age                                -8.660e-03  5.308e-03  -1.631 0.102809
TotalNumDebts                       2.702e-01  6.828e-02   3.957 7.59e-05 ***
TotalMaxDebtMonths                  1.084e-02  3.405e-03   3.183 0.001456 **
NumDebtsFinance                    -1.573e-01  7.681e-02  -2.048 0.040567 *
NumDebtsTelco                       3.289e-01  1.201e-01   2.739 0.006170 **
MaxDebtMonthsTelco                 -8.236e-03  4.508e-03  -1.827 0.067721 .
MaxDebtMonthsOther                 -9.492e-03  4.708e-03  -2.016 0.043802 *
AppliedAmount                       1.728e-04  2.649e-05   6.523 6.87e-11 ***
UseOfLoan                          -3.239e-02  1.560e-02  -2.076 0.037869 *
income_from_principal_employer     -3.711e-04  1.726e-04  -2.151 0.031506 *
IncomeFromPension                  -1.676e-03  6.667e-04  -2.514 0.011936 *
IncomeFromFamilyAllowance          -2.144e-03  8.796e-04  -2.438 0.014786 *
IncomeFromChildSupport             -2.131e-03  1.187e-03  -1.796 0.072539 .
income_other                       -9.471e-04  3.158e-04  -2.999 0.002707 **
TotalLiabilitiesBeforeLoan          1.299e-01  1.632e-02   7.961 1.71e-15 ***
TotalMonthlyLiabilities            -1.313e-03  3.465e-04  -3.789 0.000151 ***
CountOfPaydayLoans                  1.433e-01  3.303e-02   4.337 1.45e-05 ***
marital_status_id                   8.738e-02  4.705e-02   1.857 0.063294 .
education_id                       -7.123e-02  4.295e-02  -1.659 0.097196 .
employment_status_id                1.201e-01  7.229e-02   1.661 0.096761 .
Employment_Duration_Current_Employer -6.814e-02  4.527e-02  -1.505 0.132292
nr_of_dependants                    1.034e-01  4.881e-02   2.118 0.034149 *
work_experience                    -1.149e-01  6.916e-02  -1.661 0.096648 .
LoanDuration                        2.033e-02  2.738e-03   7.428 1.10e-13 ***
NoOfPreviousApplications            1.087e-01  3.118e-02   3.487 0.000488 ***
AmountOfPreviousApplications       -2.605e-05  7.715e-06  -3.377 0.000734 ***
NoOfPreviousLoans                  -1.375e-01  5.044e-02  -2.725 0.006424 **
AmountOfPreviousLoans               1.382e-04  3.712e-05   3.723 0.000197 ***
DTI1                               -1.882e+00  3.094e-01  -6.083 1.18e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4650.4  on 4046  degrees of freedom
Residual deviance: 3678.2  on 4017  degrees of freedom
AIC: 3738.2

Number of Fisher Scoring iterations: 6
```

A.6 Sammuviisilise regressiooni mudeli 2 väljavõte

```
Call:
glm(formula = Default ~ TotalMonthlyLiabilities + CountOfPaydayLoans +
    FundedAmount + TotalNumDebts + TotalLiabilitiesBeforeLoan +
    LoanDuration + DTI1 + NumDebtsTelco + Age + AmountOfPreviousLoans +
    TotalMaxDebtMonths + UseOfLoan + income_other + work_experience +
    AmountOfPreviousApplications + NoOfPreviousApplications,
    family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1825  -0.7204  -0.4880   0.4804   2.8239

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -8.580e-01  1.855e-01  -4.625 3.75e-06 ***
TotalMonthlyLiabilities      -1.894e-03  2.275e-04  -8.325  < 2e-16 ***
CountOfPaydayLoans            1.581e-01  3.243e-02   4.876 1.08e-06 ***
FundedAmount                  1.665e-04  2.621e-05   6.352 2.13e-10 ***
TotalNumDebts                 1.423e-01  2.306e-02   6.173 6.71e-10 ***
TotalLiabilitiesBeforeLoan    1.226e-01  1.575e-02   7.786 6.91e-15 ***
LoanDuration                  2.039e-02  2.635e-03   7.737 1.02e-14 ***
DTI1                         -1.398e+00  2.243e-01  -6.232 4.61e-10 ***
NumDebtsTelco                 3.960e-01  8.437e-02   4.693 2.69e-06 ***
Age                          -1.139e-02  4.748e-03  -2.399  0.01644 *
AmountOfPreviousLoans         8.635e-05  3.265e-05   2.645  0.00817 **
TotalMaxDebtMonths            5.275e-03  2.212e-03   2.384  0.01710 *
UseOfLoan                    -3.413e-02  1.547e-02  -2.206  0.02739 *
income_other                 -6.471e-04  2.896e-04  -2.235  0.02542 *
work_experience              -1.614e-01  6.657e-02  -2.425  0.01532 *
AmountOfPreviousApplications -1.299e-05  4.839e-06  -2.685  0.00725 **
NoOfPreviousApplications      4.325e-02  1.660e-02   2.605  0.00918 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4650.4  on 4046  degrees of freedom
Residual deviance: 3715.6  on 4030  degrees of freedom
AIC: 3749.6

Number of Fisher Scoring iterations: 6
```

A.7 Regressioonipuu meetod 1



Kasutatud kirjandus

[1] Bondora koduleheküIg. [Internet]. Saadaval: https://www.bondora.com/en/invest/statistics/data_export [Allalaetud 21. jaanuar 2016].

[2] Siddiqi, Naeem . 2006. *Credit Risk Scorecards Developing and Implementing Intelligent Credit Scoring*. Hoboken: John Wiley & Sons, Inc.

[3] Wikipedia. 2014. *Bondora* [Internet]. Saadaval: https://et.wikipedia.org/wiki/Bondora [Allalaetud 09.märts 2016].

[4] Pärna, Kalev. 2014. *Lecture notes: Martingales (MTMS.02.010).* Tartu Ülikool.

[5] Kangro, Raul. 2011. *Loengukonspekt: Aegridade analüüs (MTMS.01.023).* Tartu Ülikool.

[6] Lember, Jüri. 2013. *Loengukonspekt ja ülesanded: Tõenäosusteooria II (MTMS.02.004).* Tartu Ülikool.

[7] Lyn C. Thomas, David B. Edelman ja Jonathan N. Crook. 2002. *Credit Scoring and Its Applications*. Philadelphia: Society for Industrial and Applied Mathematics.

[8] Gareth James, Daniela Witten, Trevor Hastie ja Robert Tibshirani. 2013. *An Introduction to Statistical Learning with Applications in R*. New York: Springer Science+Business Media.

[9] Käärik, Ene. 2013. *Loengukonspekt: Andmeanalüüs II (MTMS.01.007).* Tartu Ülikool.

[10] Wikipedia. 2016 *Spline* [Internet]. Saadaval: https://en.wikipedia.org/wiki/Spline_(mathematics) [Kasutatud 12. mai 2016]

[11] Wikipedia. 2016 *B-spline* [Internet]. Saadaval: https://en.wikipedia.org/wiki/B-spline [Kasutatud 12. mai 2016]

[12] Kaasik, Ants; Remm, Kalle. 2012. *Ruumiliste loodusandmete statistiline analüüs.* Tartu: Tartu Ülikooli Ökoloogia ja Maateaduste Instituut.

[13] Kaasik, Ülo. 2002. *Matemaatikaleksikon.* Tartu: As Atlex.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Lagle Sammelsaar,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

"Pankrotistumise tõenäosuse prognoosimine otselaenamisettevõtte Bondora andmetel",

mille juhendaja on Raul Kangro,

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, 12.05.2016