TARTU ÜLIKOOL
MATEMAATIKA-INFORMAATIKATEADUSKOND
Matemaatilise statistika instituut

Paavo Binsol

Maksevõimetuse hindamine

Magistritöö
finants- ja kindlustusmatemaatika erialal (30 EAP)

Juhendaja: Raul Kangro (PhD)

Tartu 2015

Maksevõimetuse hindamine

Käesolevas magistritöös uuritakse maksevõimeliste ja maksevõimetute klientide õigesti klassifitseerimist, mis on kahe klassiga klassifitseerimisülesanne. Lisaks uuritakse erinevate kaofunktsioonide mõju klassifitseerimistäpsusele ja kahju suurusele. Leitud mudelid põhinevad logistilise regressiooni, otsustuspuude ja närvivõrkude metoodikal. Kahe andmestiku korral on võrreldud klassifitseerimise täpsust ja kahju suurust kasutades eelnevalt mainitud mudeleid koos kaofunktsioonidega.

Märksõnad: klassifitseerimine, riskianalüüs, logistiline regressioon, otsustuspuu, närvivõrk

Insolvency estimation

In this master's thesis there is evaluated the correct classification of solvent and insolvent customers, which is a two state classification problem. Additionally there is studied the effect of different loss functions to classification accuracy and loss size. Derived models are based on logistic regression, decision trees and neural networks methodology. On two datasets the classification accuracies and loss sizes are compared by using previously mentioned models with loss functions.

Keywords: classification, risk analysis, logistic regression, decision tree, neural network

Sisukord

Sissejuhatus

Krediidiriski efektiivne hindamine on üks olulisemaid ülesandeid tänapäeva panganduses pärast aastatel 2008 ja 2009 olnud majanduslangust. Sisereitingutepõhiste meetodite korral on pankadel kapitalinõude arvutamisel krediidiriski katteks kaks võimalust:

- põhivariant (*Foundation IRB*);
- täiustatud variant (*Advanced IRB*).

Kahe sisereitingutepõhise meetodi erisus seisneb selles, et kui põhivariandi puhul saab turuosaline ise määrata üksnes maksejõuetuse tõenäosuse (ülejäänud krediidiriski parameetrid on regulatiivselt sätestatud), siis täiustatud variandi puhul saab turuosaline määratleda ise kõik neli krediidiriski parameetrit (maksejõuetuse tõenäosus, maksejõetusest tingitud kahjususe määr, nõude prognoositav suurus maksejõuetuse hetkel, nõude lõpptähtaeg) vastavalt oma sisehinnangutele. (Basel II, s.a.)

Käesoleva töö uuritavaks parameetriks on mõlemas kapitalinõude meetodis turuosalise enda poolt hinnatav kliendi maksejõetuse tõenäosuse (*probability of default*) hindamine ja selle kasutamine klientide klassifitseerimisel. Töö eesmärgiks on anda ülevaade mitmest maksevõime hindamise statistilisest meetodist ning võrrelda nende meetodite poolt saadavaid tulemusi kahte näidisandmestikku kasutades.

Uuritavateks meetoditeks on:

- logistiline regressioon (*logistic regression*);
- otsustuspuu (*decision tree*);
- närvivõrk (*neural network*).

Töö on üles ehitatud järgmiselt. Esimeses peatükis defineeritakse vastavad terminid, millele hiljem toetutakse ning kirjeldatakse töös kasutatavaid meetodeid.
Teises peatükis on võrreldud kolme meetodi täpsust maksevõimetuse hindamisel kahe andmestiku korral. Riskimõõdikutena kasutatakse maksevõimeliste ja makseraskustega klientide õigesti klassifitseerimise osakaale ja kliendiportfellide kahju suurusi. Meetodite läbiviimiseks ja tulemuste illustreerimiseks kasutati vastavalt statistikapaketti R (Studio) ning Microsoft Excelit. Tööle on lisatud algsed andmestikud, kasutatud R-i programmikoodid ja failid (CD).

1. Klassifitseerimine

Paljudes olukordades tuleb teha valik mitme võimaluse vahel omamata täielikku informatsiooni. Pankade krediidiriski peamiseks osaks on otsustada, millistele klientidele laenu anda ja kelle laenutaotlus tagasi lükata. Otsuse tegemiseks leitakse hinnangud, mis saadakse subjekti/objekti iseloomustavaid abitunnuseid kasutades. Eeldatakse, et sarnaste näitajatega laenutaotlejad käituvad samamoodi nagu minevikus teadaolevad kliendid.

Klassifitseerija (*classification*) on teatud objekti liigitamine ühte etteantud klassidest.

Klassifitseeritava objekti kirjeldus esitatakse vektorina $\boldsymbol{x} \in \mathbb{R}^p$, mida nimetatakse tunnusvektoriks (*feature, pattern*).

Klassifitseerija seab igale tunnusvektorile $\boldsymbol{x}$ vastavusse ühe klassi võimalikest klassidest

$$Y := \{0, \ldots, K-1\},$$

kus praegusel juhul $K = 2$.

Seega matemaatiliselt on klassifitseerija funktsioon

$$g : \mathbb{R}^p \mapsto Y. \tag{1.1}$$

Iga sellist funktsiooni (1.1) saab esitada kujul

$$g(\boldsymbol{x}) = \sum_{i=0}^{K-1} i I_{C_i}(\boldsymbol{x}), \tag{1.2}$$

kus I_{C_i} on hulga C_i indikaatorfunktsioon ehk

$$I_{C_i}(\boldsymbol{x}) := \begin{cases} 0, & \text{kui } \boldsymbol{x} \notin C_i \\ 1, & \text{kui } \boldsymbol{x} \in C_i \end{cases}. \tag{1.3}$$

(Lember, 2013, lk 7)

Käesoleva töö uuritavateks klassideks on:

- 0 maksevõimega klient;
- 1 maksevõimetu klient.

Sellest tulenevalt lihtsustub klassifitseerija (1.2) kujule

$$g(\boldsymbol{x}) = I_{C_1}(\boldsymbol{x}).$$

1.2. Kaofunktsioon

Reaalses elus põhjustab klassifitseerimisviga konkreetset kahju. Näiteks kaotab pank raha, kui laenuvõtja ei suuda seda tagasi maksta ning samuti kaotab pank siis, kui jätab laenu andmata kliendile, kes oma kohustusi täidaks. Samuti võib erinevate klasside klassifitseerimisviga olla erineva tagajärjega. Pankade jaoks põhjustab makseraskustega kliendile laenu andmine enamasti suuremat kahju, kui õigeaegselt tagasimakseid tegevate klientide laenutaotluste tagasi lükkamine. Järgnevalt on defineeritud funktsioon, mille väärtusteks on kahjud vastavate valesti klassifitseerimise juhtude korral.

Definitsioon 1.1. Kaofunktsioon (*loss function*)

$$L : Y \times Y \to \mathbb{R}^{+}$$

seab klasside paarile (i, j) vastavusse kahju, mida toob endaga kaasa tegelikult klassi i kuuluva vaatluse lugemine klassi j kuuluvaks.

Tulenevalt kaofunktsiooni definitsioonist on näha, et on loomulik võtta õigesti klassifitseerimise korral kaofunktsiooni väärtuseks 0 ehk $L(i, j) = 0$. Klassi y tunnusvektoriga $\boldsymbol{x}$ klassifitseerimisel tekkiv kahju on $L(y, g(\boldsymbol{x}))$ ning mida väiksem on uute vaatluste klassifitseerimisel tekkiv keskmine kahju, seda parem on klassifitseerija. (Lember, 2013, lk 11)

Kui laenu pealt teenib pank kasu 5%, aga maksevõimetu kliendi korral on võimalik taastada ainult 50% algselt antud laenust (tagasimaksed kuni maksevõimetuse hetkeni, tagatise müümine jne), siis kahjude vahekord klassifitseerimisvigade korral on kümnekordne. Erinevatele klassifitseerimisvigadele vastavate kaofunktsioonide väärtuste erinevus sõltub suuresti pakutavate toodete (krediitkaart, kodulaen jne) iseärasusest.

Järgnevalt on defineeritud neli kaofunktsiooni, kus esimese korral on valesti klassifitseerimise kahju sama suur mõlema klassi korral ning ülejäänud kolmes on eeldatud, et maksevõimelise kliendi valesti klassifitseerimise kahju on vastavalt kaks, viis ja kümme korda väiksem kui maksevõimetu kliendi valesti prognoosimine.

Erinevad kaofunktsioonid, mille mõju hinnatakse on järgmised:

$$L_1(i,j)=\begin{cases}0, \text{ kui } i=j\\ 1, \text{ kui } i\neq j\end{cases} \tag{1.4}$$

$$L_2(i,j)=\begin{cases}0, \text{ kui } i=j\\ 1, \text{ kui } i=0, j=1\\ 2, \text{ kui } i=1, j=0\end{cases} \tag{1.5}$$

$$L_3(i,j)=\begin{cases}0, \text{ kui } i=j\\ 1, \text{ kui } i=0, j=1\\ 5, \text{ kui } i=1, j=0\end{cases} \tag{1.6}$$

$$L_4(i,j)=\begin{cases}0, \text{ kui } i=j\\ 1, \text{ kui } i=0, j=1\\ 10, \text{ kui } i=1, j=0\end{cases}. \tag{1.7}$$

1.3. Bayesi klassifikaator ja selle optimaalsus

Tulenevalt klassifitseerija g (1.1) ning kaofunktsiooni L definitsioonist (def. 1.1.) on loomulik uurida vaadeldava klassifitseerija keskmist kadu, mis on defineeritud järgnevalt.

Definitsioon 1.2. Klassifitseerija g risk on keskmine kahju üle tunnusvektori ja klasside ühisjaotuse $F(\boldsymbol{x}, y)$:

$$R(g) := \int_{\mathbb{R}^p \times Y} L\left(y, g(\boldsymbol{x})\right) dF(\boldsymbol{x}, y). \tag{1.8}$$

Käesoleva töö kontekstis lihtsustub valem (1.8) kujule

$$\begin{aligned} R(g) &= \int_{\mathbb{R}^p \times Y} L\left(y, g(\boldsymbol{x})\right) dF(\boldsymbol{x}, y) = \int_{\mathbb{R}^p} \int_{Y} L\left(y, g(\boldsymbol{x})\right) dF(y \mid \boldsymbol{x}) dF(\boldsymbol{x}) = \\ &= \int_{\mathbb{R}^p} \left(L\left(0, g(\boldsymbol{x})\right) p(0 \mid \boldsymbol{x}) + L\left(1, g(\boldsymbol{x})\right) p(1 \mid \boldsymbol{x}) \right) dF(\boldsymbol{x}), \end{aligned}$$

kus:

- $p(0 \mid \boldsymbol{x})$ $\left(p(1 \mid \boldsymbol{x})\right)$ on tõenäosus, et uuritava vaatluse klass on 0 (1) tingimusel, et tema abitunnuste vektor on $\boldsymbol{x}$.

On mõistetav, et eemärgiks on leida klassifitseerija üle kõikide võimalike klassifitseerijate hulga, mis minimiseeriks riski $R(g)$. Järgnevalt on defineeritud parim võimalik klassifitseerija ning minimaalne võimalik keskmine kahju.

Definitsioon 1.3. Klassifitseerijat

$$g^*(\boldsymbol{x}) := \arg\min_{i \in Y} \sum_{j=0}^{K-1} L(j, i) p(j \mid \boldsymbol{x}) \tag{1.9}$$

nimetatakse **Bayesi klassifitseerijaks** ning tema riski

$$\begin{aligned} R^* &:= R(g^*) = \int_{\mathbb{R}^p \times Y} L\left(y, g^*(\boldsymbol{x})\right) dF(\boldsymbol{x}, y) = \\ &= \int_{\mathbb{R}^p} \sum_{j=0}^{K-1} L\left(j, g^*(\boldsymbol{x})\right) p(j \mid \boldsymbol{x}) dF(\boldsymbol{x}) \end{aligned} \tag{1.10}$$

nimetatakse **Bayesi riskiks**.

Valemist (1.9) on näha, et Bayesi klassifitseerija on parim võimalik klassifitseerija, sest keskmine kadu $(\sum_{j=0}^{K-1} L\left(j, g^*\left(\boldsymbol{x}\right)\right) p\left(j \mid \boldsymbol{x}\right))$ üle tunnusvektori $\boldsymbol{x}$ on väikseim kõigi tunnusvektorite korral ning seega on ka kogukahju väikseim võimalik. (Lember, 2013, lk 12-13)

Käesolevas töös lihtsustuvad Bayesi klassifitseerija ja riski valemid vastavalt järgmisele kujule:

$$g^*\left(\boldsymbol{x}\right) := \arg\min_{i \in \{0,1\}} \left(L\left(0, i\right) p\left(0 \mid \boldsymbol{x}\right) + L\left(1, i\right) p\left(1 \mid \boldsymbol{x}\right)\right) \quad (1.11)$$

$$R^* := R\left(g^*\right) = \int_{\mathbb{R}^p} \left(L\left(0,1\right) p\left(0 \mid \boldsymbol{x}\right) + L\left(1,0\right) p\left(1 \mid \boldsymbol{x}\right)\right) dF\left(\boldsymbol{x}\right)$$

1.4. Logistilisel regressioonil põhinev klassifikaator

Logistilise regressiooni korral modelleeritakse etteantud argumenttunnuse korral klassidesse 1 ja 0 kuulumise tõenäosuste suhte logaritmi lineaarse funktsioonina abiinformatsiooni sisaldava vektori $\boldsymbol{X}$ väärtustest. Sellise lähenemise korral on tagatud, et klassidesse kuulumised tõenäosused on vahemikus $[0,1]$ ja summeeruvad üheks.

Mudeli kuju on järgmine

$$\ln \frac{P(Y=1|\boldsymbol{X}=\boldsymbol{x_i})}{P(Y=0|\boldsymbol{X}=\boldsymbol{x_i})} = \beta_{10} + \boldsymbol{\beta_1}^T \boldsymbol{x_i},$$

kus:

- $P(Y=1|\boldsymbol{X}=\boldsymbol{x_i})$ on tõenäosus i -s vaatlus kuulub klassi 1;
- $P(Y=0|\boldsymbol{X}=\boldsymbol{x_i})$ on tõenäosus i -s vaatlus kuulub klassi 0;
- $\boldsymbol{x_i}$ on i -nda vaatluse abitunnuste vektor;
- β_{10} mudeli vabaliige ja $\boldsymbol{\beta_1}^T$ tundmatute parameetrite vektor.

Tähistame huvipakkuva tõenäosuse kujul $P(Y=1|\boldsymbol{X}=\boldsymbol{x_i}) = p_1(\boldsymbol{x_i};\boldsymbol{\theta})$, kus $\boldsymbol{\theta}$ on tundmatute parameetrite hulk $\boldsymbol{\theta} = \{\beta_{10}, \boldsymbol{\beta_1}^T\}$.

Kuna hinnatakse ainult kahte klassi kuulumise tõenäosusi, siis $P(Y=0|\boldsymbol{X}=\boldsymbol{x_i}) = p_0(\boldsymbol{x_i};\boldsymbol{\theta}) = 1 - p_1(\boldsymbol{x_i};\boldsymbol{\theta})$ ning

$$\ln \frac{P(Y=1|\boldsymbol{X}=\boldsymbol{x_i})}{P(Y=0|\boldsymbol{X}=\boldsymbol{x_i})} = \ln \frac{p_1(\boldsymbol{x_i};\boldsymbol{\theta})}{1-p_1(\boldsymbol{x_i};\boldsymbol{\theta})} = \beta_{10} + \boldsymbol{\beta_1}^T \boldsymbol{x_i}.$$

Teisendamise tulemusena

$$p_1(\boldsymbol{x_i};\boldsymbol{\theta}) = \frac{\exp(\beta_{10} + \boldsymbol{\beta_1}^T \boldsymbol{x_i})}{1+\exp(\beta_{10} + \boldsymbol{\beta_1}^T \boldsymbol{x_i})}$$

ja

$$p_0(\boldsymbol{x_i};\boldsymbol{\theta}) = \frac{1}{1+\exp(\beta_{10} + \boldsymbol{\beta_1}^T \boldsymbol{x_i})}.$$

1.4.1. Tundmatute parameetrite hindamine

Üldjuhul hinnatakse logistilise regressiooni tundmatud parameetrid $\boldsymbol{\theta} = \{\beta_{10}, \boldsymbol{\beta_1}^T\}$ suurima tõepära meetodiga, kus eesmärgiks on maksimiseerida tõenäosust saada vastav $\boldsymbol{Y}$ väärtuste vektor $(y_1,\ldots,y_n)$ tingimusel, et $\boldsymbol{X_i} = \boldsymbol{x_i}$, $i = 1,\ldots,n$ ehk maksimeeritakse tõenäosusfunktsiooni

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p_{y_i}(\boldsymbol{x_i};\boldsymbol{\theta})$$

Maksimeerimisülesande lihtsustamiseks uuritakse logaritmitud tõenäosusfunktsiooni

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln p_{y_i}(\boldsymbol{x_i};\boldsymbol{\theta}). \tag{1.12}$$

Kuna uuritav tunnus saab kuuluda ainult kahte klassi, siis avaldub valem (1.12) kujul

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left\{ y_i \ln p_1(\boldsymbol{x_i};\boldsymbol{\theta}) + (1-y_i)\ln\left(1-p_1(\boldsymbol{x_i};\boldsymbol{\theta})\right)\right\}. \tag{1.13}$$

Kui $y_i = 1$, siis

$$\begin{aligned}
&y_i \ln p_1(\boldsymbol{x_i};\boldsymbol{\theta}) + (1-y_i)\ln\left(1-p_1(\boldsymbol{x_i};\boldsymbol{\theta})\right) = \\
&y_i\left(\ln\left(\exp\left(\beta_{10}+\boldsymbol{\beta_1}^T\boldsymbol{x_i}\right)\right) - \ln\left(1+\exp\left(\beta_{10}+\boldsymbol{\beta_1}^T\boldsymbol{x_i}\right)\right)\right) = \\
&y_i\left(\beta_{10}+\boldsymbol{\beta_1}^T\boldsymbol{x_i} - \ln\left(1+\exp\left(\beta_{10}+\boldsymbol{\beta_1}^T\boldsymbol{x_i}\right)\right)\right)
\end{aligned}$$

ja $y_i = 0$, siis

$$\begin{aligned}
&y_i \ln p_1(\boldsymbol{x_i};\theta) + (1-y_i)\ln\left(1-p_1(\boldsymbol{x_i};\theta)\right) = \\
&\ln\left(\frac{1}{1+\exp\left(\beta_{10}+\boldsymbol{\beta_1}^T\boldsymbol{x_i}\right)}\right) = \ln\left(\left(1+\exp\left(\beta_{10}+\boldsymbol{\beta_1}^T\boldsymbol{x_i}\right)\right)^{-1}\right) = \\
&-\ln\left(1+\exp\left(\beta_{10}+\boldsymbol{\beta_1}^T\boldsymbol{x_i}\right)\right).
\end{aligned}$$

Eelnevast tulenevalt avaldub valem (1.12) kujul

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left\{ y_i\left(\beta_{10}+\boldsymbol{\beta_1}^T\boldsymbol{x_i}\right) - \ln\left(1+\exp\left(\beta_{10}+\boldsymbol{\beta_1}^T\boldsymbol{x_i}\right)\right)\right\}.$$

(Hastie, Tibshirani, Friedman, 2009, lk 119-120)

1.4.2. Kaofunktsiooni kasutamine

Logistilise regressiooni korral me kaofunktsiooni ennast mudeli sobitamisel ei arvesta, kuid kaofunktsiooni kasutatakse selleks, et panna paika piir, millest alates lugeda vaatlust klassi 1 kuuluvaks.

Eesmärgiks on leida parim klassifitseerija g, mille korral risk R on väikseim. Käesoleva töö Bayesi klassifitseerija (1.11) on ekvivalentne valemiga:

$$g^*(\boldsymbol{x}_i) = \begin{cases} 0, \text{ kui } L(0,1)\,\hat{p}_0(\boldsymbol{x}_i;\boldsymbol{\theta}) > L(1,0)\,\hat{p}_1(\boldsymbol{x}_i;\boldsymbol{\theta}) \\ 1, \text{ kui } L(0,1)\,\hat{p}_0(\boldsymbol{x}_i;\boldsymbol{\theta}) \le L(1,0)\,\hat{p}_1(\boldsymbol{x}_i;\boldsymbol{\theta}) \end{cases} =$$

$$= \begin{cases} 0, \text{ kui } \hat{p}_1(\boldsymbol{x}_i;\boldsymbol{\theta}) < \dfrac{L(0,1)}{L(1,0)+L(0,1)} \\ 1, \text{ kui } \hat{p}_1(\boldsymbol{x}_i;\boldsymbol{\theta}) \ge \dfrac{L(0,1)}{L(1,0)+L(0,1)} \end{cases}. \tag{1.14}$$

Seega tuleks lugeda i-s vaatlus klassi 1 kuuluvaks juhul, kui vastava vaatluse tingliku tõenäosuse hinnang

$$\hat{p}_1(\boldsymbol{x}_i;\boldsymbol{\theta}) \ge \frac{L(0,1)}{L(1,0)+L(0,1)}.$$

Siit on näha, et mida suurem on klassi 1 vaatluste valesti klassifitseerimise kahju $L(1,0)$ võrreldes klassi 0 valesti klassifitseerimise kahjuga $L(0,1)$, seda väikseima tingliku tõenäosuse hinnangu $\hat{p}_1(\boldsymbol{x}_i;\boldsymbol{\theta})$ korral me loeme vaatluste klassi 1.

1.4.3. Parima mudeli valik

Erinevaid logistilise regressiooni mudeleid võib tulenevalt seletavate tunnuste arvust olla väga palju. Eesmärgiks on leida kõige lihtsam (väheste parameetritega) mudel, mis sisaldaks samas kõiki olulisi seletavaid tunnuseid. Mainitud eesmärgi 100% kindluse saavutamiseks tuleb läbi vaadata kõikvõimalike p regressori alamhulkadele vastavad mudelid, mis on vähegi suurema p korral väga aeganõudev tegevus. Selle asemel kasutatakse sageli **sammuviisilist regressiooni**, millel on kolm lähenemist:

- ettepoole liikuv (*forward selection*): alustatakse ainult vabaliikmega mudelist ning lisatakse juurde tunnus, mis parandab mudelit kõige rohkem ning eelnevat sammu korratakse nii kaua, kuni pole ühtegi olulist tunnust lisada;
- tagasi liikuv (*backward selection*): alustatakse mudelist, mis sisaldab kõiki seletavaid tunnuseid ning hakatakse eemaldama järjest kõige vähem olulisemaid tunnuseid, kuni pole eemaldada ühtegi tunnust, mis poleks oluline;
- eelnevalt kirjeldatud lähenemiste kombinatsioon: igal sammul uuritakse, kas on mõni oluline tunnus, mida tuleks lisada või ebaoluline tunnus, mida eemaldada.

(Hastie et al., 2009, lk 58-61)

Hindamaks, kuidas teatud tunnuse lisamine või eemaldamine mõjutab mudeli täpsust, defineerime Aikaike (AIC) informatsioonikriteeriumi:

$$AIC = -2\log(L) + 2p, \tag{1.15}$$

kus:

- L on uuritava mudeli tõepärafunktsiooni väärtus;
- p on parameetrite arv mudelis.

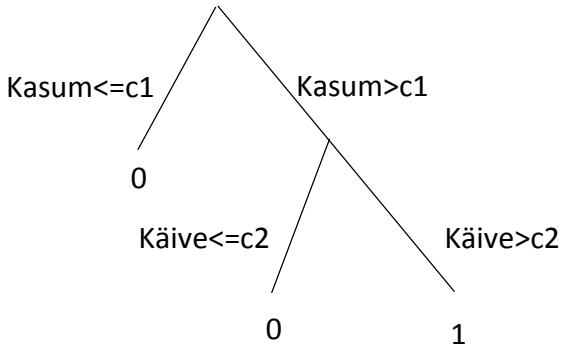
Aikaike informatsioonikordajas mõõdab mudeli täpsust logaritmitud tõepärafunktsiooni väärtus ning $2p$ trahvib liigselt keerulise mudeli kasutamise eest. Tulemusena on parim mudel väikseima AIC väärtusega.

Statistikatarkvaras R on sammuviisilise regressiooni kasutamiseks funktsioon *step*, kus on võimalik kõiki kolme sammregressiooni meetodit rakendada. Selleks tuleb funktsiooni parameetri *direction* väärtuseks omistada vastavalt kas *forward*, *backward* või *both*. (Ripley, s.a.)

Käesolevas töös kasutatakse kolmandat sammregressiooni meetodit (*direction=both*).

1.5. Otsustuspuu

Peatüki materjal on esitatud Hastie, Tibshirani, Friedmanni (2009) ning Lemberi (2013) põhjal. Otsustuspuu korral leitakse etteantud tunnusvektori klassihinnang lõpliku arvu sammudega, kus igal sammul võrreldakse ühte tunnusvektori komponenti (jagamistunnust) mingi arvuga (jagamiskohaga). Otsustuspuudel põhinevad meetodid jagavad abitunnuste ruumi ristkülikuteks ning sobitavad lihtsa mudeli (konstandi) igasse ristkülikusse. Klassifitseerimise probleemi lahendamiseks kasutatakse käesolevas töös otsustuspuu meetodit CART (*Classification And Regression Tree*). Joonisel 1 on esitatud kahe jagamiskohaga (c1, c2) otsustuspuu kahe klassiga klassifitseerimisülesande korral.



Joonis 1. Kahe jagamiskohaga otsustuspuu skeem kahe klassi korral

1.5.1. Puu kasvatamine

Olgu andmestik moodustatud p abitunnusest ja ühest uuritavast tunnusest iga n vaatluse jaoks. Algoritmi eesmärgiks on automaatselt valida parimad jagamistunnused ja jagamisarvud ning otsustuspuu suurus (jagamiste arv).

Olgu uuritavate tunnuste põhjal jaotus gruppidesse $R_1, R_2, \ldots, R_M$, siis sellele jaotusele vastava klassifikaatori võib kirja panna kujul

$$g(\boldsymbol{x}) = \sum_{m=1}^{M} c_m I(\boldsymbol{x} \in R_m),$$

kus:

- c_m on konstantne vastus gruppi R_m jaoks.

Selle töö klassifitseerimisülesande korral on $\hat{c}_m$ väärtuseks kaks võimalust, kas 0 (maksevõimeline klient) või 1 (makseraskustega klient).

Vaadeldakse suvalise ristküliku R jagamist kaheks osaks. Iga tunnuse j ja jagamispunkti s jaoks defineeritakse kaks alamregiooni

$$R_1(s, j) = \{\boldsymbol{X} \in R \mid X_j \le s\} \text{ ja } R_2(s, j) = \{\boldsymbol{X} \in R \mid X_j > s\}.$$

Järgnevalt otsitakse jagamistunnust j ja jagamispunkti s, mis oleks järgneva ülesande

$$\arg\min_{j,s}\left[n_1\varphi\left(\hat{p}_1\left(R_1(s, j)\right)\right) + n_2\varphi\left(\hat{p}_1\left(R_2(s, j)\right)\right)\right], \tag{1.16}$$

kus:

- $\hat{p}_{im} = \hat{p}_i(R_m) = \dfrac{n_{im}}{n_{0m} + n_{1m}}$, $m,\ i = 1,2$ ehk i-nda klassi osakaal m-ndas alapiirkonnas;
- $n_{im} = \#\{\boldsymbol{x} \in R_m \mid y = i\}$, $m,\ i = 1,2$ on i-nda klassi vaatluste arv m-ndas alapiirkonnas;
- $n_m = \#\{\boldsymbol{x} \in R_m\}$, $m = 1,2$ on m-ndasse regiooni kuuluvate vaatluste arv;

lahendiks.

Funktsiooni φ eesmärgiks on mõõta klassifitseerimisviga vastavas piirkonnas ning sellest tulenevalt on valemis (1.16) minimiseerimisülesanne üle tunnuste ja tunnuste väärtuste.

Enamasti on funktsiooniks φ üks järgnevatest:

- $\varphi(p) = \min(p, 1-p)$ klassifitseerimisviga;
- $\varphi(p) = 2p(1-p)$ Gini indeks;
- $\varphi(p) = -p\ln(p) - (1-p)\ln(1-p)$ entroopiafunktsioon (*cross-entropy*).

Olgu j ja s ülesande (1.16) lahendid, siis R jagamisest saadav võit defineeritakse kujul:

$$\frac{1}{n}\left(n_R \varphi\left(\frac{n_{11}+n_{12}}{n_R} \right) - n_1 \varphi\left(\hat{p}_1\left(R_1(s,j) \right) \right) - n_2 \varphi\left(\hat{p}_1\left(R_2(s,j) \right) \right) \right),$$

kus:

- $n_R = \#\{\boldsymbol{x} \in R\}$ ehk ristkülikusse R kuuluvate vaatluste arv;
- n on kõigi vaatluste arv.

Igal sammul valitakse jagamiseks selline ristkülik, mille jagamisel saadav võit on kõige suurem.

Järgnevaks probleemiks on, et kui suur puu tuleks „kasvatada“? Liiga suure puu korral on ülesobitamise oht, aga liiga väike puu ei taba olulisi seoseid. Samuti ei saa puud kasvatada nii kaua, kuni klassifitseerimisviga väheneb rohkem mingist etteantud suurusest, sest mõne väikese kasuteguriga paranemise eraldusele võib järgneda väga hea eraldus, mis jääks vastava strateegia korral leidmata.

Eelistatud strateegia on kasvatada võimalikult suur puu T_0 (määrates minimaalse sõlme suuruseks näiteks 5). Tekitatud suurt puud pügatakse *cost-complexity* pügamismeetodiga.

Tähistagu T suure puu T_0 sellist alampuud $(T \subset T_0)$, mida on võimalik saada T_0 pügamisel ehk sõlmede alampuude kustutamisel (kus sõlm jääb uueks puu leheks). Eesmärgiks on iga α korral leida selline alampuu $T_\alpha \subset T_0$, mis minimiseerib *cost-complexity* kriteeriumit:

$$C_\alpha(T) = \sum_{m=1}^{|T|} n_m Q_m(T) + \alpha |T|, \tag{1.17}$$

kus:

- $n_m = \#\{\boldsymbol{x} \in R_m\}$ ehk vaatluste arv regioonis R_m;
- $Q_m(T) = \varphi\left(\hat{p}_1\left(R_m \right) \right)$, kus φ ja $\hat{p}$ on valemis (1.16) kirjeldatud;
- $|T|$ on puu lehtede arv.

Eesmärgiks on leida iga $\alpha > 0$ jaoks alampuu T_α, mis minimeeriks funktsiooni $C_\alpha(T)$ (1.17). Parameeter (*tuning parameter*) α reguleerib puu suuruse ja puu täpsuse vahelist seost. Suure α korral on tulemuseks väiksemad puud T_α ning vastupidi väikese α korral, kui $\alpha = 0$, siis lahendiks on algselt leitud suurim puu T_0.

On võimalik näidata, et iga α jaoks on olemas väikseim alampuu T_α, mis minimiseerib funktsiooni $C_\alpha(T)$ ja on unikaalne. T_α leidmiseks kasutatakse nõrgima lüli pügamist (*weakest link pruning*). Esmalt ühendatakse sellise sõlme lehed, mille korral suuruse $\sum_{m=1}^{|T|} n_m Q_m(T)$ kasv ühe sõlme kohta on väikseim. Eelnevat kirjeldatud tegevust jätkatakse, kuni jõutakse väikseima, ühe sõlmega, puuni. Tulemuseks saadakse lõplik jada alampuudest ja on tõestatud, et jada sisaldab puud T_α (Breiman et al). Rist-valideerimist kasutatakse $\hat{\alpha}$ saamiseks. Lõplik puu on $T_{\hat{\alpha}}$.

Otsustuspuul põhinev klassifikaator on kujul, kus vaatlused m-ndas lehes klassifitseeritakse klassi, mille osakaal on kõige suurem vastavas sõlmes ehk käesolevas töös on m-nda lehe klassifitseerija:

$$g_m(\boldsymbol{x}_i) := \begin{cases} 0, \text{ kui } \hat{p}_{1m} < \frac{1}{2} \\ 1, \text{ kui } \hat{p}_{1m} \geq \frac{1}{2} \end{cases}.$$

1.5.2. Kaofunktsiooni kasutamine

Gini indeks iseloomustab jagamise kasulikust hästi, kui kaofunktsioon on sümmeetriline s.t. $L(0,1) = L(1,0)$. Järgnevalt näitame, kuidas ebasümmeetrilise kaofunktsiooni korral saab puu ehitamise taandada sümmeetrilise kaofunktsiooni juhule. Selleks paneme tähele, et me saame tinglikud tõenäosused $p_{im} = P(y = i \mid \boldsymbol{x} \in R_m)$ alati kirjutada kujul

$$p_{im} = P\left(y = i \mid \boldsymbol{x} \in R_m\right) = \frac{P\left(y = i \cap \boldsymbol{x} \in R_m\right)}{P\left(\boldsymbol{x} \in R_m\right)} =$$
$$= P\left(y = i\right) \frac{P\left(\boldsymbol{x} \in R_m \mid y = i\right)}{P\left(\boldsymbol{x} \in R_m\right)} \approx \pi_i \frac{n_{im}}{n_i} \frac{n}{n_m} \tag{1.18}$$

- π_0, π_1 on klasside eeltõenäosused $P\left(y = 0\right)$ ja $P\left(y = 1\right)$;
- n_{im} ja n_m on samad, mis valemis (1.16);
- $P\left(\boldsymbol{x} \in R_m \mid y = i\right)$ on tõenäosus, et vaatlus tunnusvektoriga $\boldsymbol{x}$ kuulub regiooni R_m tingimusel, et vaatluse õige klass on i;
- $P\left(\boldsymbol{x} \in R_m\right)$ on regiooni R_m sattumise tõenäosus.

Paneme tähele, et piirkondade R_m konstantse klassifitseerija risk avaldub kujul

$$R(g) = E\left(L\left(y, g\right)\right) = \sum_{m=1}^{M} P\left(\boldsymbol{x} \in R_m\right) E\left(L\left(y, g_m\right) \mid \boldsymbol{x} \in R_m\right) =$$
$$= \sum_{m=1}^{M} P\left(\boldsymbol{x} \in R_m\right) R(m) \tag{1.19}$$

kus:

- $R\left(m\right) = L\left(0, g_m\right) p_{0m} + L\left(1, g_m\right) p_{1m}$ ehk risk m-ndas sõlmes;
- g_m on mingi konstant igas piirkonnas.

Valemit (1.18) kasutades saame

$$R\left(m\right) \approx \pi_0 L\left(0, g_m\right) \frac{n_{0m}}{n_0} \frac{n}{n_m} + \pi_1 L\left(1, g_m\right) \frac{n_{1m}}{n_1} \frac{n}{n_m}, \tag{1.20}$$

kus:

- n on vaatluste arv valimis.

Eeldame, et leiduvad $\tilde{\pi}_i$ ja $\tilde{L}\left(i, j\right)$ nii, et

$$\tilde{\pi}_i \tilde{L}\left(i, j\right) = \pi_i L\left(i, j\right),$$

siis riski hinnang sõlmes m

$$R(m) \approx \tilde{\pi}_0 \tilde{L}(0, g_m) \frac{n_{0m}}{n_0} \frac{n}{n_m} + \tilde{\pi}_1 \tilde{L}(1, g_m) \frac{n_{1m}}{n_1} \frac{n}{n_m},$$

on sama iga ristkülikus R_m (ning seetõttu on ka kogurisk) uue kaofunktsiooni ja eeltõenäosuste korral. Järelikult võime optimaalse puu otsimisel kasutada sümmeetrilise kaofunktsiooni jaoks sobivat lähenemist, kuid tükeldamisel tuleb arvestada modifitseeritud eeljaotustele vastavaid tõenäosuseid.

Säärasel juhul

$$\tilde{\pi}_i = \frac{\pi_i L_i}{\pi_0 L_0 + \pi_1 L_1}, i = 0,1, \tag{1.21}$$

kus:

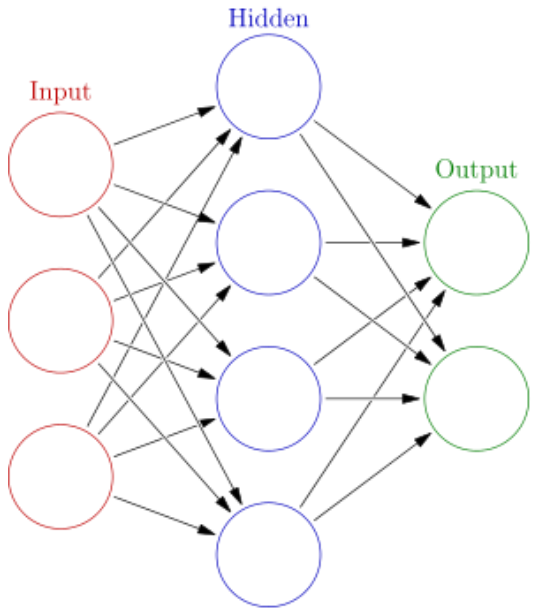
- $L_0 = L(0,1)$ ning $L_1 = L(1,0)$.

(Therneau, Atkinson, 2015, lk 1-8)

1.6. Närvivõrgud

Peatüki materjal on esitatud Hastie, Tibshirani ja Friedmanni (2009) põhjal. Närvivõrkude puhul moodustatakse klassifikaator, kombineerides omavahel paljusid mitme muutuja funktsioone, mis arvutavad oma argumentidest (sisendmuutujatest) lineaarkombinatsioone ning rakendavad saadud tulemusele mingeid mittelineaarseid ühe muutuja funktsioone.

Käesolevas töös vaatleme ühe peidetud kihiga (*hidden layer*) närvivõrku. Klassifitseerimisülesande korral on närvivõrgul vastavalt klasside arvule K väljundit (*output*'i), kus iga klass on kirjeldatud indikaatortunnusega. Joonisel 2 on esitatud kolme sisendtunnuse, ühe peidetud kihi, nelja peidetud kihi elemendiga närvivõrgu skeem kahe klassiga klassifitseerimisülesande korral.



Joonis 2. Ühe peidetud kihiga närvivõrgu skeem kahe klassi korral (Colored_neural_network.svg, s.a.)

Peidetud kihi elemendid leitakse funktsionaalse seosena sisendtunnuste lineaarsest kombinatsioonist:

$$Z_{im} = \sigma(a_{0m} + a_m{}^T \boldsymbol{x_i}),\ m = 0,\ldots,M-1,\ i = 1,\ldots n \qquad (1.22)$$

kus:

- Z_{im} on i-nda vaatluse peidetud kihi m-s element;
- a_{0m} ja $\boldsymbol{a_m}^T$ on tundmatud parameetrid;
- $\boldsymbol{x_i}^T = \left(x_{i1},\ldots,x_{ip}\right)$ $\boldsymbol{i}$-nda vaatluse seletavate tunnuste vektor;
- M ja n on vastavalt esimese peidetud kihi elementide ja vaatluste arv.

Valemis (1.22) kasutatud σ nimetatakse aktiveerimisfunktsiooniks (*activation function*), milleks sageli on *sigmoid* funktsioon:

$$\sigma(v) = \frac{1}{\left(1+\exp(-v)\right)}.$$

Tunnusvektorile $\boldsymbol{x_i}$ vastavad klassifitseerimishinnangud leitakse, analoogselt esimese sammuga, funktsionaalse seosena nüüd peidetud esimese kihi elementide lineaarsest kombinatsioonist ehk:

$$T_{ik} = \beta_{0k} + \boldsymbol{\beta_k}^T \boldsymbol{Z_i},\ k = 0,\ldots,K-1 \qquad (1.23)$$

ja

$$f_k(\boldsymbol{x_i}) = g_k\left(\boldsymbol{T_i}\right)\ k = 0,\ldots,K-1,$$

kus:

- $\boldsymbol{Z_i}^T = \left(Z_{i0},\ldots,Z_{iM-1}\right)$ ehk i-nda vaatluse peidetud kihi elementide vektor;
- β_{0k} ja $\boldsymbol{\beta_k}^T$ on tundmatud parameetrid;
- $\boldsymbol{T_i}^T = \left(T_{i0},\ldots,T_{iK-1}\right)$ ehk vektorist $\boldsymbol{Z_i}$ lineaarsete kombinatsioonidena saadud tulemusvektor;
- g_k viimase transformatsiooni funktsioon.

Klassifitseerimisülesande korral on g_k funktsiooniks valitud üldjuhul *softmax* funktsioon

$$g_k(\boldsymbol{T_i}) = \frac{\exp(T_{ik})}{\sum_{l=0}^{K-1}\exp(T_{il})},$$

sest sellel puhul rahuldavad hinnangud tingimusi $g_k(\boldsymbol{T_i}) \in [0,1]$ ning $\sum_{k=0}^{K-1} g_k(\boldsymbol{T_i}) = 1$, mis on sobilik klassifitseerimisülesande jaoks.

Tulenevalt käesoleva töö klassifitseerimisülesandest, kus on ainult kaks klassi ehk $K = 2$ (0 ja 1), saadakse närvivõrke kasutades igale kliendile mõlemasse klassi kuulumise hinnangud:

$$\hat{p}_{0i} = g_0(\hat{\boldsymbol{T}}_i),\ \hat{p}_{1i} = g_1(\hat{\boldsymbol{T}}_i)\ i = 1,\ldots,n.$$

1.6.1. Närvivõrgu sobitamine andmetele

Närvivõrkudel on tundmatud parameetrid, mida nimetatakse kaaludeks (*weights*) ning eesmärgiks on hinnata kaalud, mis sobitaksid mudeli võimalikult hästi meie andmestikule. Tulenevalt valemitest (1.22) ja (1.23) on ühe peidetud kihiga närvivõrgu tundmatute parameetrite hulk $M(p+1)+K(M+1)$, sest vajaminevad kaalud on:

$$\begin{aligned} &\{a_{0m}, \boldsymbol{a_m};\ m = 0,\ldots,M-1\} \\ &\{\beta_{0k}, \boldsymbol{\beta_k};\ k = 0,\ldots,K-1\} \end{aligned} \tag{1.24}$$

Parameetrite hindamiseks kasutatakse entroopiafunktsiooni :

$$R(\boldsymbol{\theta}) = -\sum_{i=1}^{n}\sum_{k=0}^{K-1} \boldsymbol{y_{ik}} \ln f_k(\boldsymbol{x_i}) \tag{1.25}$$

kus:

- $\boldsymbol{y_{ik}}$ on i-nda vaatluse klassi k iseloomustav indikaatorvektor;
- $\boldsymbol{\theta}$ on tundmatute kaalude (1.24) hulk.

Vastav klassifitseerija on kujul

$$G(\boldsymbol{x_i}) = \arg\max_k f_k(\boldsymbol{x_i}). \tag{1.26}$$

(Hastie et al., 2009, lk 395)

Valemist (1.25) on näha, et summa R on võimalikult väike, kui uuritavad klassid hinnatakse hästi funktsiooniga f_k.

Käesolevas töös (klasside arv $K = 2$) lihtsustub valem (1.25) kujule:

$$R(\theta) = -\sum_{i=1}^{n} \left(y_{i0} \ln f_0(x_i) + y_{i1} \ln f_1(x_i) \right).$$

ning klassifitseerija:

$$G(x_i) = \arg\max_{k \in \{0,1\}} f_k(x_i).$$

Funktsiooni $R(\theta)$ miinimumpunktide leidmiseks kasutatakse iteratiivseid numbrilisi meetodeid.

Erinevaid kaofunktsioone kasutatakse analoogselt logistilise regressiooni meetodiga (1.4.2), kus vaatluse klassifitseerimine muutub tulenevalt tingliku tõenäosuse lävepunktist.

1.6.2. Olulised kohad närvivõrgu sobitamisel

Minimiseerimisülesande funktsioon R (1.25) on sageli mittekumer ning kaalude hinnangud sõltuvad nende algväärtustest. Lahenduseks on soovitatud leida lõplikuks klassifitseerimishinnanguks keskmine üle mitme erinevate närvivõrkude hinnangute.

Sisendtunnuste jaotus võib mõjutada kaalude väärtusi. Sellest tulenevalt on soovitav standardiseerida kõik sisendtunnused keskmisega null ning standardhälbega üks. Tulemusena koheldakse kõiki sisendtunnuseid samaväärselt mudeli sobitamisel.

Peidetud kihi elementide arv (*hidden units*) on üldjuhul vahemikus $5-100$, suurenedes koos sisendtunnuste ja vaatluste arvuga. (Hastie et al., 2009, lk 397-401)

1.6.3. Ülesobitamine

Sageli on närvivõrkudel liiga palju kaalusid ning tulemuseks on ülesobitamise oht lahendades funktsiooni R (1.25) miinimumülesannet. Eelneva vältimiseks kasutatakse kaalude vähendamise (*weight decay*) meetodit, kus on lisatud trahv piiramaks liigselt suuri hinnanguid kaaludele. Uus minimiseeriv funktsioon (1.24) on järgmisel kujul:

$$R^*(\theta) = R(\theta) + \lambda J(\theta),$$

kus:

- $J(\theta) = \sum_{k=0}^{K-1}\sum_{m=0}^{M-1}\beta^2{}_{km} + \sum_{m=0}^{M-1}\sum_{l=1}^{p}a^2{}_{ml}$ ehk ruutude summa üle kõikide kaalude;
- $\lambda \geq 0$ (*tuning parameter*) määrab kaalude suuruse mõju närvivõrgu mudeli sobitamisel.

Väikse λ korral võetakse kaalude mõju sobitamisel vähem arvesse kui suure λ korral, millal kaalud lähenevad nullile. Sobiva λ leidmiseks kasutatakse ristvalideerimist (vt peatükk 2. Sobitamise protseduur). (Hastie et al., 2009, lk 398)

Tulenevalt kasutatud vea hindamise funktsioonist valemis (1.25) on soovitav λ 't otsida ligikaudu vahemikust $0.01-0.1$ (Ripley, 1993, 1994, viidatud Venables, Ripley, 2002).

Statistikatarkvaras R on võimalik närvivõrke genereerida funktsiooniga *nnet*.

2. Sobitamise protseduur

Antud peatüki eesmärgiks on võrrelda eelnevalt kirjeldatud meetodite täpsust ja tõhusust kasutades nii eraisikust kliendi kui ettevõtte kohta olevat informatsiooni. Töös esimesena vaadeldud andmestik sisaldab maksevõimeliste ja makseraskustesse sattunud klientide isikuomadusi (sugu, töökoha staaž jne) ning majanduslikku seisu (arvelduskonto summa, eelnevate võlgnevuste ajalugu jne) iseloomustavaid tunnuseid. Teises andmestikus on 2008. aasta majandusaasta aruande näitajad Eesti ettevõtete kohta ja eesmärgiks on võimalikult hästi klassifitseerida ettevõtteid vastavat sellele, kas järgmise 3 aasta jooksul (2009-2011) ollakse pankrotistunud või mitte.

Algandmetele lisaks on leitud juurde finantssuhtarve ja lisatunnuseid, uurimaks, kas genereeritud tunnused võivad anda klassifitseerimisülesandele täpsema hinnangu või väiksema kahju. Lisaks on uuritud, kuidas mõjutavad tulemusi erinevad kaofunktsioonid.

Kummagi andmestiku korral valiti kogu andmestikust 80% vaatlusi treeninguks (parameetrite hindamine, puu kasvatamine) ja 20% testimiseks (täpsuse ja kahju leidmine). Iga meetodi korral sobitati mudel nii algandmetele, lisatunnustele kui ka eelnevatest ühendatud andmestikule. Otsustuspuu korral leiti lisaks parimad puud iga kaofunktsiooni korral eraldi. Logistilise regressiooni ja närvivõrkude korral muutus kaofunktsioonist tulenevalt tõenäosuspiir, mille järgi klient klassi liigitati. Tabelis 1 on väljatoodud eelpool mainitud tõenäosuspiirid tulenevalt kaofunktsioonist. Kui $\hat{P}\left(Y=1 \mid \boldsymbol{X}=\boldsymbol{x}\right) \geq piir$, siis $g\left(\boldsymbol{x}\right)=1$.

Tabel 1. Tõenäosuspiirid

Kaofunktsioon	Piir
$L_1\ (1.4)$	0.50
$L_2\ (1.5)$	0.33
$L_3\ (1.6)$	0.17
$L_4\ (1.7)$	0.09

Parima logistilise regressiooni mudeli leidmisel kasutati *step* funktsiooni statistikapaketis R, mis teostab sammuviisilist regressiooni kasutades täpsusnäitajana Aikaike informatsioonikriteeriumit. Kõigi juhtude korral kasutati *step* funktsiooni kaks korda. Algselt anti parameetrina ette ainult vabaliikmega mudel ning sellest tulenevalt oli esimesel sammul võimalik ainult lisada mõni tunnus. Teisel juhul kasutati kõiki seletavaid tunnuseid sisaldavat mudelit ja esimese sammuna oli võimalik ainult eemaldada kõige ebaolulisem tunnus. Lõpptulemusena valiti parimaks väiksema AIC-ga mudel.

Otsustuspuu kasvatamiseks ja pügamiseks kasutati R-i *rpart* paketi funktsiooni *rpart*. Uute jagamiskohtade ja – punktide otsimine lõpetati, kui vaatluste arv oli sõlmes väiksem kui 20. Puu pügamisel kasutati ristvalideerimist leidmaks parim α (*tuning parameter* valemis (1.17)). Vastava kaofunktsiooni mõju avaldumiseks kasutati leituid eeltõenäosuseid funktsiooni *rpart* parameetris *priors*. Tabelis 2 on väljatoodud kasutatud eeltõenäosused mõlema andmestiku ja kolme kaofunktsiooni korral (juhul (1.4) on $\tilde{\pi}_0$ ja $\tilde{\pi}_1$ vastavalt klasside 0 ja 1 vaatluste osakaalud andmestikes ja neid kasutatakse vaikimisi funktsioonis *rpart*). (Atkinson, Ripley, Therneau, 2015)

Tabel 2. Muudetud eeltõenäosused (altered priors)

	Andmestik 1		Andmestik 2	
Kaofunktsioon*Altered priors*	$\tilde{\pi}_0$	$\tilde{\pi}_1$	$\tilde{\pi}_0$	$\tilde{\pi}_1$
$L_2\ (1.5)$	0.54	0.46	0.97	0.03
$L_3\ (1.6)$	0.32	0.68	0.94	0.06
$L_4\ (1.7)$	0.19	0.81	0.88	0.12

Närvivõrgu parameetrite hindamiseks kasutati R-i funktsiooni *nnet*. Olenevalt andmestiku suurusest valiti peidetud kihi elementide arvuks vastavalt 5 (andmestik 1) ja 10 (andmestik 2). Seletavad tunnused standardiseeriti, et kõiki tunnuseid koheldaks sobitamisel samaväärselt. Ülesobitamist ennetava parameeter λ leidmiseks kasutati ristvalideerimist, mille korral jaotati andmestik 10-ks (ligikaudselt) võrdse suurusega osaks ning igas osas leiti, milline on täpsus ja kahju suurus juhul, kui kasutada kaalude hindamiseks ülejäänud osasid. Keskmiselt väikseima kahju ja suurima täpsuse saavutanud $\hat{\lambda}$ valiti sobivaimaks. Eelneva protseduuri läbiviimiseks tehti vastav funktsioon (*cross_validation.nnet.Rdat*). Lõplikuks testandmete hinnanguks kujunes 10 närvivõrgu mudeli hinnangute keskmine, vältimaks kaalude juhuslike algväärtuste liigset mõju (*final_estimate.nnet.Rdat*).

2.1. Andmestik 1 (*German Credit data*)

Andmetabel on kättesaadav *UCI Machine Learning Repository* kodulehelt, mis sisaldab andmestikke erinevate meetodite rakendamiseks (Statlog (German Credit Data) Data Set, s.a.). Informatsiooni on 1000 kliendi kohta, kus vastavalt 700 on maksevõimelised ja 300 makseraskustega laenu taotlejad. Iga inimese kohta on teada 21 tunnust, kus üks on uuritav tunnus (maksevõimeline/maksevõimetu klient) ja 20 seletavat tunnust. Käesolevas töös on kasutatud andmetabelit, kus faktortunnuste korral on iga tase defineeritud indikaatortunnusega. Tunnuste nimekiri ja kirjeldus on esitatud Lisas 1.

Algandmete põhjal arvutati juurde kaheksa lisatunnust. Kuna andmestik sisaldas ainult faktortunnuseid, siis pidevate tunnuste (arvelduskrediidi ja hoiuse summa) arvutamiseks kasutati tulemusena faktori tasemete vahelist keskmist. Järgnevalt on väljatoodud kasutatud lisatunnused:

- Checking_account – arvelduskonto suurus kui positiivne väärtus, muidu 0;
- Depth – indikaatortunnus, kui arvelduskonto maht negatiivne või konto puudulik;
- Savings_account – hoiusekonto suurus, kui hoiusekontot pole, siis väärtus 0;
- Savings_account_unknown – indikaatortunnus, kas hoiusekonto eksisteerib või mitte;
- Employment – indikaatortunnus iseloomustamaks, kas kliendi kestev töökoha kestvus all 1 aasta või puudulik;
- Amount_Savings_account – laenu ja hoiusekonto summade suhe;
- Amount_Checking_account – laenu ja arvelduskonto summade suhe;
- Amount_per_month – laenu summa ja kestvuse suhe (kohustuste igakuine maht laenu kestvuse ajal).

Vastava andmestiku ametlikuks kaofunktsiooniks on märgitud funktsioon L_3. Kõikide meetodite, tunnusvektorite komplektide ja kaofunktsioonide korral on hinnangute ja tegelike klasside sagedustabelid ning kahju suurused esitatud Lisas 1.

2.1.1. Tulemused

Tabelis 3-5 on tulemused testandmestikus iga meetodi, kasutatud andmete ja kaofunktsiooni korral, kus:

- Klass 0 – näitab maksevõimeliste klientide õigesti klassifitseerimise osakaalu;
- Klass 1 – näitab makseraskustega klientide õigesti klassifitseerimise osakaalu;
- Kahju – ühikuline kahju, mis on tekkinud klasside valesti klassifitseerimisest vastavalt kaofunktsioonile.

Kui valesti klassifitseerimise viga oli sama mõlema klassi korral (L_1), siis väikseim kahju oli logistilise regressiooni mudeliga, mis sisaldas nii algandmeid kui lisatunnuseid (Lisa 1., Tabel 9.). Mudelis on 21 tunnust ja vabaliige, algandmetele lisaks on kasutusel üks lisatunnus (*Amount_saving_account*). Indikaatortunnustest avaldasid suurimat mõju kliendi maksevõime hinnangule tunnused, mis näitasid, kas klient on käendaja mõnes teises lepingus (-1.411) või puudub arvelduskonto (-1.333). Pidevate tunnuste parameetritest oli suurima väärtusega (0.038) laenu pikkust iseloomustava tunnuse *Duration* parameeter. Kõige ebatäpsemalt ja suurima kahjuga klassifitseeris otsustuspuu kõigi vaadeldud tunnusvektorite komplektide korral.

Närvivõrkude mudel sobitatuna algandmetel andis kõige väikseima kahju (83), kui kaofunktsioon on kujul L_2. Suurima kahju ja ebatäpsusega hindas ainult lisatunnuseid sisaldanud otsustuspuu. Kahju suurus oli võrreldes parima mudeliga ligikaudu 40% suurem.

Vastava andmestiku eelistatud kaofunktsiooni L_3 korral prognoosis väikseima kahjuga viie lisatunnuse ja ühe vabaliikmega logistilise regressioonimudel (Lisa 1., Tabel 10.). Klassi 0 kuuluvate klientide klassifitseerimistäpsus oli halvem kui teiste meetodite korral, aga klassi 1 paigutati õigesti ligikaudu 95% makseraskustega klientidest. Sellest tulenevalt oli suure kahjuga vigu väga vähe (3 tk). Mudeli indikaatortunnustele hinnatud parameetritest avaldas suurimat negatiivset mõju *Credit_history* (-0.779), mis kirjeldas, kas kliendil on olnud makseraskusi eelnevalt mõne krediidikohustuse täitmisega. Pidevate väärtustega tunnustest sisaldas mudel igakuist laenukohustuse suurust (*Amount_per_month*) ja laenu osakaalu hoiusekontost (*Amount_Savings_account*).

Kaofunktsioonile L_4 andis väikseima kahjuga hinnangu närvivõrkude meetodil põhinev mudel kasutades kõiki olemasolevaid tunnuseid. Tulemustest on näha, et vastava kaofunktsiooni korral andsid väikseima kahju mudelid, mis klassifitseerisid kõik kliendid klassi 1. Kõige rohkem prognoosis testandmestikus kliente klassi 0 logistilise regressiooni mudel, aga sellega kaasnenud üksikute makseraskustega klientide valesti klassifitseerimisel oli suur mõju kogu portfelli kahjule (Lisa 1. Kaofunktsioon L_4).

Tabel 3. Õigesti klassifitseerimise osakaalud ja portfelli kahju kasutades algandmeid erinevate kaofunktsioonide korral

	Algandmed								
	Logit			Otsustuspuu			Närvivõrk		
Kaofunktsioon	Klass 0	Klass 1	Kahju	Klass 0	Klass 1	Kahju	Klass 0	Klass 1	Kahju
L_1	81.2%	40.3%	63	80.4%	32.3%	69	80.3%	35.5%	68
L_2	71.0%	59.7%	90	70.3%	41.9%	113	70.3%	66.1%	83
L_3	50.7%	82.3%	123	47.8%	77.4%	142	47.8%	82.3%	127
L_4	31.9%	91.9%	144	26.8%	88.7%	171	1.4%	100.0%	136

Tabel 4. Õigesti klassifitseerimise osakaalud portfelli kahju kasutades lisatunnuseid erinevate kaofunktsioonide korral

	Lisatunnused (finantssuhtarvud, indikaatorid)								
	Logit			Otsustuspuu			Närvivõrk		
Kaofunktsioon	Klass 0	Klass 1	Kahju	Klass 0	Klass 1	Kahju	Klass 0	Klass 1	Kahju
L_1	91.3%	21.0%	61	79.7%	25.8%	74	89.9%	22.6%	62
L_2	71.0%	51.6%	100	59.4%	51.6%	116	68.8%	46.8%	109
L_3	25.4%	95.2%	118	19.6%	80.6%	171	10.9%	95.2%	138
L_4	2.2%	98.4%	145	0.0%	100.0%	138	0.0%	100.0%	138

Tabel 5. Õigesti klassifitseerimise osakaalud ja portfelli kahju kasutades algandmeid ja lisatunnuseid erinevate kaofunktsioonide korral

	Algandmed+Lisatunnused								
	Logit			Otsustuspuu			Närvivõrk		
Kaofunktsioon	Klass 0	Klass 1	Kahju	Klass 0	Klass 1	Kahju	Klass 0	Klass 1	Kahju
L_1	82.6%	43.5%	59	73.9%	43.5%	71	79.0%	33.9%	70
L_2	68.8%	56.5%	97	63.8%	62.9%	96	71.0%	61.3%	88
L_3	50.7%	79.0%	135	47.8%	77.4%	142	38.4%	85.5%	130
L_4	34.1%	91.9%	141	0.0%	100.0%	138	3.6%	100.0%	133

2.1.2. Järeldused

Lineaarsel regressiooni ja närvivõrkude meetodil põhinevate mudelite klassidesse õigesti klassifitseerimise osakaalud ja kahju suurused olid analoogsed, aga otsustuspuude kahjud olid võrreldes teiste meetoditega suuremad.

Tabelitest 3-5 on märgata, et klassi 1 kuuluva kliendi valesti klassifitseerimise kahju suurenemisega, muutub klasside 0 ja 1 õigesti klassifitseerimise osakaalud vastupidiseks. See on mõistetav, sest eemärgiks on vältida suuremat kahju põhjustava vea tegemist ning sellest tulenevalt ollakse klassi 0 paigutamisel ettevaatlikumad. Kui tegelik olukord vastaks kaofunktsioonile L_3 ja seda mitte arvestada, siis võrdsete kahjudega parim mudeliga (algandmetel põhinev logistiline regressioon mudel) oleks tegelik kahju ligikaudu 68% (199 vs 118) suurem võrreldes vastavat kaofunktsiooni arvesse võttes leitud parima mudeli korral (lisatunnustel põhinev logistiline regressiooni mudel). Seega tuleks vastava andmestiku korral arvestada kaofunktsiooni mõju modelleerimisel, sest kahju suuruse vähenemine on märgatav.

Märkimisväärne on, et kaofunktsiooni L_4 korral hindasid otsustuspuu ja närvivõrgu mudel peaaegu kõik kliendid klassi 1, vältimaks kümnekordse kahju põhjustamist valesti klassifitseerimisest.

2.2. Andmestik 2

Andmestik 2 sisaldab 56 497 Eesti ettevõtte 2008 aasta majandusaasta aruande andmeid, mis on kättesaadavad äriregistri kodulehelt. Uuritavaks tunnuseks on, kas ettevõte läheb pankroti järgmise kolme aasta jooksul ning selliseid vaatlusi oli andmestikus 722 tükki (1.28%). Seletavate tunnustena on majandusaasta aruannetest kasutusel võetud 22 pidevat tunnust (Lisa 2.1). Algandmete põhjal arvutati juurde 19 lisatunnust, mis sisaldasid erinevaid finantssuhtarve ja indikaatortunnuseid. Ettevõtte finantssuhtarv väärtustati 0-ga, kui nimetajas oleva tunnuse väärtus oli 0 ning vastava puuduva mõju mõõtmiseks defineeriti indikaatortunnus. Järgnevalt on väljatoodud kasutatud lisatunnused:

- current_ratio – käibevara ja lühiajaliste kohustuste suhe;
- quick_ratio – kassa/arvelduskonto, lühiajaliste finantsinvesteeringute ja nõuete summa osakaal lühiajalistest kohustustest ;
- missing_short_term_liabilities – indikaatortunnus näitamaks, kas lühiajalised kohustused andmetabelis 0 või mitte;
- missing_total_equity –indikaatortunnus näitamaks, kas omakapital andmetabelis 0 või mitte;
- return_on_sales – kasumi osakaal käibest;
- missing_sales – indikaatortunnus näitamaks, kas käive andmetabelis 0 või mitte;
- profit_total_assets – kasumi osakaal koguvarast;
- liabilities_assets – kohustuste osakaal koguvarast;
- sales_assets – käibe osakaal koguvarast;
- current_assets_total_assets – käibevarade osakaal kohustustest;
- equity_total_assets – omapakitali osakaal koguvaradest;
- missing_assets - indikaatortunnus näitamaks, kas varad andmetabelis 0 või mitte
- short_liabilities_total_liabilities – lühiajaliste kohustuste osakaal kogu kohustustest;
- missing_total_liabilities – indikaatortunnus näitamaks, kas kohustused andmetabelis 0 või mitte;
- cash_bank_current_assets – kassa/arvelduskonto osakaal käibevarast;
- missing_current_assets – indikaatortunnus näitamaks, kas käibevarad andmetabelis 0 või mitte;

- liabilities_larger_equity – indikaatortunnus näitamaks, kas kohustusi on rohkem kui omakapitali;
- profit – indikaatortunnus näitamaks, kas ettevõte lõpetas majandusaasta kasumi või kahjumiga;
- empl_nbr_missing – indikaatortunnus näitamaks, kas töötajate arv puudulik (0) või mitte.

Kõikide meetodite, tunnusvektorite komplektide ja kaofunktsioonide korral on klassifitseerimishinnangute ja tegelike klasside sagedustabelid ning kahju suurused esitatud Lisas 2.3.

2.2.1. Tulemused

Tabelis 6-8 on vastava testandmestiku tulemused iga meetodi, kasutatud andmete ja kaofunktsiooni korral, kus:

- Klass 0 – näitab järgmise kolme aasta jooksul mitte pankrotistuvate ettevõtete õigesti klassifitseerimise osakaalu;
- Klass 1 – näitab järgmise kolme aasta jooksul pankrotistuvate ettevõtete õigesti klassifitseerimise osakaalu;
- Kahju – ühikuline kahju, mis on tekkinud klasside valesti klassifitseerimisest vastavalt kaofunktsioonile.

Algandmeid sisaldanud närvivõrkudel põhinevates mudelites oli maksimaalne klassi 1 kuulumise tingliku tõenäosuse hinnanguks $\hat{P}(Y=1|\boldsymbol{X}=\boldsymbol{x})$ ligikaudu 0.087 ning sellest tulenevalt on kõik ettevõtted kasutatud kaofunktsioonide korral klassifitseeritud klassi 0. Ainult kaofunktsiooni L_3 ja L_4 korral, kui kasutati lisatunnuseid, hinnati mõningad ettevõtted pankrotistuvateks närvivõrkude mudelite poolt.

Järgnevalt on tehtud ülevaade tulemustest iga kaofunktsiooni juhu jaoks.

Kaofunktsiooni L_1 korral andis väikseima portfelli kahju närvivõrkude ja logistilise regressioonil meetoditel põhinevad mudelid, mille korral klassifitseeriti kõik ettevõtted järgmise kolme aasta jooksul mitte pankrotistuvateks ehk jätkusuutlikeks. Otsustuspuu

klassifitseeris võrreldes teiste meetoditega märgatavalt rohkem vaatlusi klassi 1, kuid halva täpsusega.

Kaofunktsiooni L_2 korral oli tähelepanuväärne see, et otsustuspuu ja närvivõrkude mudelid ei hinnanud ühtegi ettevõtet majandusaasta aruandest saadud informatsiooni ja arvutatud lisatunnuste põhjal kolme aasta jooksul pankrotistuvaks. Kahe mudeli korral tegi seda logistilise regressiooni mudel, aga tulenemata sellest oli portfelli kahju suurem, kui eelnevalt mainitud meetodite korral, sest enamus klassi 1 klassifitseeritud vaatlustest olid mitte pankrotistunud ettevõtted (Lisa 2., Tabel 26 ja 28). Seega vastava kaofunktsiooni korral ei saa ühtegi meetodit selgelt eelistada.

Tabelitest 6-8 on näha, et kaofunktsiooni L_3 korral oli otsustuspuu hinnanguid kasutades portfelli kahju väiksem. Võrreldes teiste meetoditega hinnati rohkem vaatlusi õigesti klassi 1. Klassi 0 kuuluva ettevõtte valesti klassifitseerimine ei avaldanud kogu kahjule nii suurt mõju, sest kahju oli viis korda väiksem võrreldes klassi 1 kuuluva vaatluste valesti klassifitseerimisega. Parima tulemuse andis ainult algandmeid kasutanud 13 jagamiskohaga otsustuspuu, mis on väljatoodud Lisas 2., Joonisel 3.

Kui erinevate klasside valesti klassifitseerimise kahju suuruste vahe oli kümnekordne (L_4), siis parima tulemuse andis ainult algandmeid kasutanud 23 jagamiskohaga otsustuspuu (Lisa 2., Joonis 4.). Võrreldes teiste meetodite mudelitega andis otsustuspuu peaaegu kõikide erinevate tunnusvektorite komplektide korral ligikaudu 7-14%-lise väiksema kahju. Ainult lisatunnuseid kasutanud närvivõrkude mudeli kogu kahju oli sarnane otsustuspuu omaga. Kahju suuruste vahe tuleb sellest, et ülejäänud mudelid klassifitseerisid üksikuid ettevõtteid klassi 1 ning sellest tulenevalt oli rohkem suurema kahjuga vigu.

Tabel 6. Õigesti klassifitseerimise osakaalud ja portfelli kahju kasutades algandmeid erinevate kaofunktsioonide korral

	Algandmed								
	Logit			Otsustuspuu			Närvivõrk		
Kaofunktsioon	Klass 0	Klass 1	Kahju	Klass 0	Klass 1	Kahju	Klass 0	Klass 1	Kahju
L_1	<100.0%	0.7%	145	99.5%	5.6%	189	100.0%	0.0%	142
L_2	99.9%	0.7%	288	100.0%	0.0%	284	100.0%	0.0%	284
L_3	99.9%	1.4%	710	99.2%	17.6%	670	100.0%	0.0%	710
L_4	99.9%	1.4%	1411	97.5%	33.8%	1222	100.0%	0.0%	1420

Tabel 7. Õigesti klassifitseerimise osakaalud ja portfelli kahju kasutades lisatunnuseid erinevate kaofunktsioonide korral

	Lisatunnused (finantssuhtarvud, indikaatorid)								
	Logit			Otsustuspuu			Närvivõrk		
Kaofunktsioon	Klass 0	Klass 1	Kahju	Klass 0	Klass 1	Kahju	Klass 0	Klass 1	Kahju
L_1	100.0%	0.0%	142	99.8%	2.1%	161	100.0%	0.0%	142
L_2	100.0%	0.0%	284	100.0%	0.0%	284	100.0%	0.0%	284
L_3	<100.0%	0.0%	711	100.0%	0.0%	710	<100.0%	0.0%	711
L_4	<100.0%	0.0%	1421	98.0%	22.5%	1323	99.1%	12.0%	1346

Tabel 8. Õigesti klassifitseerimise osakaalud ja portfelli kahju kasutades algandmeid ja lisatunnuseid erinevate kaofunktsioonide korral

	Algandmed+Lisatunnused								
	Logit			Otsustuspuu			Närvivõrk		
Kaofunktsioon	Klass 0	Klass 1	Kahju	Klass 0	Klass 1	Kahju	Klass 0	Klass 1	Kahju
L_1	<100.0%	0.7%	144	99.5%	4.2%	187	100.0%	0.0%	142
L_2	<100.0%	0.7%	287	100.0%	0.0%	284	100.0%	0.0%	284
L_3	99.9%	0.7%	711	99.3%	16.2%	676	100.0%	0.0%	710
L_4	99.9%	0.7%	1424	97.7%	30.3%	1243	100.0%	0.0%	1420

2.2.2. Järeldused

Erinevate kaofunktsioonide mõju võttis kõige paremini arvesse otsustuspuu, mille korral klassi 1 kuuluva vaatluste valesti klassifitseerimise kahju suurenemisel (võrreldes klassi 0 kahjuga) hinnati järjest rohkem ettevõtteid pankrotistuvateks, mis vähendas tekkinud kogu kahju suurust ning seda ei mõjutanud ka piisavalt otsustuspuu ebatäpsus klassi 0 suhtes. Logistilise regressiooni ja närvivõrkude meetodite korral olid ettevõtete tinglikud tõenäosused kuulumaks klassi 1 väga madalad. Seega hinnati järgmise kolme aasta jooksul enamus tegelikult pankrotistuvatest ettevõtetest jätkusuutlikeks.

Kokkuvõte

Käesolevas magistritöö eesmärgiks on võrrelda kahe andmestiku korral erinevaid meetodeid klientide klassifitseerimiseks maksevõimetuteks ja maksevõimelisteks, lähtuvalt valesti klassifitseerimise poolt põhjustatud kahjusid kirjeldavatest kaofunktsioonidest. Meetoditena kasutati logistilisel regressioonil, otsustuspuul ja ühe peidetud kihiga närvivõrkudel põhinevaid mudeleid. Lisaks uuriti, kuidas mõjutab tulemusi erinevate kaofunktsioonide ja algandmetest leitud lisatunnuste kasutamine.

Kaofunktsiooni mõju arvestamiseks oli otsustuspuu meetodil erinev lähenemine võrreldes ülejäänud kahe meetodiga. Otsustuspuu korral leiti vastavad eeltõenäosused (Tabel 2.) ning logistilise regressiooni ja närvivõrkudele põhinevate mudelite klassifitseerimise hinnangute jaoks leiti tõenäosuspiirid (Tabel 1.), mida kasutati klassifitseerimisel pärast tinglike tõenäosushinnangute leidmist. Kolme meetodi ja nelja kaofunktsiooni kahju suuruste ja õigesti klassifitseerimise tulemuste võrdlemiseks viidi läbi simulatsioon kahe andmestikuga. Parameetrite hindamiseks ja puu kasvatamiseks jäeti mõlema andmestiku korral 80% andmetest ning ülejäänud osal hinnati riskimõõdikud.

Lõpptulemusena ei olnud kindlat meetodit, mille kasutamisel oleks kahju väikseim mõlema andmestiku ja iga kaofunktsiooni korral. Erinevad kaofunktsioonid avaldasid selget mõju mudelite klassifitseerimise hinnangutele ja sellest tulenevalt ka (v.a Andmestik 2 Närvivõrkudel põhinevad mudelid) täpsusele. Klasside vaheliste valesti klassifitseerimise kahjude erinevuste suurenemisel vältisid mudelid järjest rohkem suuremat kahju põhjustavat klassi.

Kasutatud kirjandus

1. Atkinson, Beth., Ripley B. D., Therneau, T. (2015). *Package 'rpart'*. http://cran.r-project.org/web/packages/rpart/rpart.pdf. Külastatud 04.05.2015.

2. Basel II (s.a.). http://www.fi.ee/index.php?id=3226. Külastatud 03.05.2015.

3. Breiman, L. (1984). *Classification and regression trees*. Belmont, California.: Wadsworth International Group.

4. Colored_neural_network.svg. (s.a.). https://commons.wikimedia.org/wiki/File:Colored_neural_network.svg#file. Külastatud 12.05.2015.

5. Hastie, T., Tibshirani, R., Friedmann, J. (2009). *The Elements of Statistical Learning Second Edition*. Springer.

6. Lember, J. (2013). *Tehisõpe 1*. Loengukonspekt. http://www.ms.ut.ee/sites/default/files/ms/tehisope_2013.pdf. Külastatud 03.05.2015.

7. Ripley B. D. (s.a.) *Choose a model by AIC in a Stepwise Algorithm*. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/step.html. Külastatud 03.05.2015.

8. Statlog (German Credit Data) Data Set. (s.a.). https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29. Külastatud 04.05.2015.

9. Therneau, T. M., Atkinson, E. J. (2015). *An Introduction to Recursive Partitioning Using the RPART Routines*.

http://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf. Külastatud 03.05.2015.

10. Venables, W. N., Ripley, B. D. (2002). *Modern Applied Statistics With S Fourth edition*. New York. Springer.

Lisad

Lisa 1. Andmestik 1

1.1. Kasutatud tunnused

Feature	Explanation
Duration	Duration of credit in months
Amount	Credit amount
InstallmentRatePercentage	Installment rate as % of disposable income
ResidenceDuration	Applicants present residence since
Age	Applicants age
NumberExistingCredits	Applicants number of number of existing credits at this bank
NumberPeopleMaintenance	Number of people being liable to provide maintenance for
Telephone	No/Yes
ForeignWorker	No/Yes
Class	Bad Loan/Good Loan
	Applicants checking account status
CheckingAccountStatus.lt.0	No/Yes
CheckingAccountStatus.0.to.200	No/Yes
CheckingAccountStatus.gt.200	No/Yes
CheckingAccountStatus.none	No/Yes
	Applicants credit history
CreditHistory.NoCredit.AllPaid	No/Yes
CreditHistory.ThisBank.AllPaid	No/Yes
CreditHistory.PaidDuly	No/Yes
CreditHistory.Delay	No/Yes
CreditHistory.Critical	No/Yes
	Applicants purpose of credit
Purpose.NewCar	No/Yes
Purpose.UsedCar	No/Yes
Purpose.Furniture.Equipment	No/Yes
Purpose.Radio.Television	No/Yes
Purpose.DomesticAppliance	No/Yes
Purpose.Repairs	No/Yes
Purpose.Education	No/Yes
Purpose.Vacation	No/Yes
Purpose.Retraining	No/Yes
Purpose.Business	No/Yes
Purpose.Other	No/Yes
	Applicants average balance in savings account
SavingsAccountBonds.lt.100	No/Yes
SavingsAccountBonds.100.to.500	No/Yes
SavingsAccountBonds.500.to.1000	No/Yes
SavingsAccountBonds.gt.1000	No/Yes
SavingsAccountBonds.Unknown	No/Yes
	Applicants present employment since
EmploymentDuration.lt.1	No/Yes
EmploymentDuration.1.to.4	No/Yes
EmploymentDuration.4.to.7	No/Yes
EmploymentDuration.gt.7	No/Yes
EmploymentDuration.Unemployed	No/Yes
	Applicants personal status and sex
Personal.Male.Divorced.Seperated	No/Yes
Personal.Female.NotSingle	No/Yes
Personal.Male.Single	No/Yes
Personal.Male.Married.Widowed	No/Yes
Personal.Female.Single	No/Yes
	Other debtors / guarantors
OtherDebtorsGuarantors.None	No/Yes
OtherDebtorsGuarantors.CoApplicant	No/Yes
OtherDebtorsGuarantors.Guarantor	No/Yes
	Applicants property
Property.RealEstate	No/Yes
Property.Insurance	No/Yes
Property.CarOther	No/Yes
Property.Unknown	No/Yes
	Applicant has other installment plan credit
OtherInstallmentPlans.Bank	No/Yes
OtherInstallmentPlans.Stores	No/Yes
OtherInstallmentPlans.None	No/Yes
	Applicants housing
Housing.Rent	No/Yes
Housing.Own	No/Yes
Housing.ForFree	No/Yes
	Nature of job
Job.UnemployedUnskilled	No/Yes
Job.UnskilledResident	No/Yes
Job.SkilledEmployee	No/Yes
Job.Management.SelfEmp.HighlyQualified	No/Yes

1.2. Mudelid

Tabel 9. Logistilise regressiooni mudel kaofunktsiooni L_1 korral

```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -3.474736   0.816688  -4.255 2.09e-05 ***
CheckingAccountStatus.none     -1.332913   0.251193  -5.306 1.12e-07 ***
Duration                        0.038092   0.008969   4.247 2.17e-05 ***
SavingsAccountBonds.lt.100      0.614459   0.304524   2.018 0.043615 *
CreditHistory.Critical         -0.753172   0.241136  -3.123 0.001788 **
OtherDebtorsGuarantors.Guarantor -1.411344   0.491666  -2.871 0.004098 **
Purpose.NewCar                  0.771501   0.226899   3.400 0.000673 ***
Personal.Male.Single           -0.559318   0.200400  -2.791 0.005254 **
OtherInstallmentPlans.None     -0.556105   0.240894  -2.309 0.020971 *
InstallmentRatePercentage       0.262878   0.091912   2.860 0.004235 **
Property.RealEstate            -0.442542   0.237945  -1.860 0.062907 .
Purpose.UsedCar                -1.045586   0.402871  -2.595 0.009450 **
Amount_Savings_account          0.007310   0.002703   2.704 0.006845 **
Telephone                       0.488985   0.207207   2.360 0.018280 *
CreditHistory.NoCredit.AllPaid  0.852192   0.452992   1.881 0.059938 .
CreditHistory.ThisBank.AllPaid  0.749769   0.414235   1.810 0.070294 .
EmploymentDuration.4.to.7      -0.571106   0.272281  -2.097 0.035950 *
SavingsAccountBonds.100.to.500  0.703615   0.346876   2.028 0.042516 *
CheckingAccountStatus.lt.0      0.486372   0.223884   2.172 0.029823 *
Purpose.Education               0.801787   0.421860   1.901 0.057355 .
ForeignWorker                   1.038860   0.670364   1.550 0.121215
Housing.Rent                    0.424755   0.247768   1.714 0.086470 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 973.97  on 799  degrees of freedom
Residual deviance: 696.89  on 778  degrees of freedom
AIC: 740.89
```

Tabel 10. Logistilise regressiooni mudel kaofunktsiooni korral

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)            -0.0950012  0.2256871  -0.421  0.67380
Amount_Savings_account  0.0133999  0.0019321   6.935 4.05e-12 ***
Credit_history         -0.7788521  0.1810087  -4.303 1.69e-05 ***
Amount_per_month       -0.0026260  0.0008257  -3.180  0.00147 **
Employment             -0.5150441  0.1864581  -2.762  0.00574 **
Depth                  -0.3834663  0.1720086  -2.229  0.02579 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 973.97  on 799  degrees of freedom
Residual deviance: 878.87  on 794  degrees of freedom
AIC: 890.87
```

1.3. Täpsustabelid ja kahju suurused kaofunktsioonist tulenevalt

Kaofunktsioon L_1

Tabel 11. Algandmed

Algandmed	Logistiline regression		Otsustuspuu		Närvivõrk	
Tegelik\Hinnang	Maksevõimeline	Maksevõimetu	Maksevõimeline	Maksevõimetu	Maksevõimeline	Maksevõimetu
Maksevõimeline	112	26	111	27	114	28
Maksevõimetu	37	25	42	20	40	22
Kahju	63		69		68	

Tabel 12. Lisatunnused

Lisatunnused	Logistiline regression		Otsustuspuu		Närvivõrk	
Tegelik\Hinnang	Maksevõimeline	Maksevõimetu	Maksevõimeline	Maksevõimetu	Maksevõimeline	Maksevõimetu
Maksevõimeline	126	12	110	28	124	14
Maksevõimetu	49	13	46	16	48	14
Kahju	61		74		62	

Tabel 13. Algandmed ja lisatunnused

Algandmed+Lisatunnused	Logistiline regression		Otsustuspuu		Närvivõrk	
Tegelik\Hinnang	Maksevõimeline	Maksevõimetu	Maksevõimeline	Maksevõimetu	Maksevõimeline	Maksevõimetu
Maksevõimeline	114	24	102	36	109	29
Maksevõimetu	35	27	35	27	41	21
Kahju	59		71		70	

Kaofunktsioon L_2

Tabel 14. Algandmed

Algandmed	Logistiline regression		Otsustuspuu		Närvivõrk	
Tegelik\Hinnang	Maksevõimeline	Maksevõimetu	Maksevõimeline	Maksevõimetu	Maksevõimeline	Maksevõimetu
Maksevõimeline	98	40	97	41	97	41
Maksevõimetu	25	37	36	26	21	41
Kahju	90		113		83	

Tabel 15. Lisatunnused

Lisatunnused	Logistiline regression		Otsustuspuu		Närvivõrk	
Tegelik\Hinnang	Maksevõimeline	Maksevõimetu	Maksevõimeline	Maksevõimetu	Maksevõimeline	Maksevõimetu
Maksevõimeline	98	40	82	56	95	43
Maksevõimetu	30	32	30	32	33	29
Kahju	100		116		109	

Tabel 16. Algandmed ja lisatunnused

Algandmed+Lisatunnused	Logistiline regression		Otsustuspuu		Närvivõrk	
Tegelik\Hinnang	Maksevõimeline	Maksevõimetu	Maksevõimeline	Maksevõimetu	Maksevõimeline	Maksevõimetu
Maksevõimeline	95	43	88	50	98	40
Maksevõimetu	27	35	23	39	24	38
Kahju	97		96		88	

Kaofunktsioon L_3

Tabel 17. Algandmed

Algandmed	Logistiline regression		Otsustuspuu		Närvivõrk	
Tegelik\Hinnang	Maksevõimeline	Maksevõimetu	Maksevõimeline	Maksevõimetu	Maksevõimeline	Maksevõimetu
Maksevõimeline	70	68	66	72	66	72
Maksevõimetu	11	51	14	48	11	51
Kahju	123		142		127	

Tabel 18. Lisatunnused

Lisatunnused	Logistiline regression		Otsustuspuu		Närvivõrk	
Tegelik\Hinnang	Maksevõimeline	Maksevõimetu	Maksevõimeline	Maksevõimetu	Maksevõimeline	Maksevõimetu
Maksevõimeline	35	103	27	111	15	123
Maksevõimetu	3	59	12	50	3	59
Kahju	118		171		138	

Tabel 19. Algandmed ja lisatunnused

Algandmed+Lisatunnused	Logistiline regression		Otsustuspuu		Närvivõrk	
Tegelik\Hinnang	Maksevõimeline	Maksevõimetu	Maksevõimeline	Maksevõimetu	Maksevõimeline	Maksevõimetu
Maksevõimeline	68	70	66	72	53	85
Maksevõimetu	13	49	14	48	9	53
Kahju	135		142		130	

Kaofunktsioon L_4

Tabel 20. Algandmed

Algandmed	Logistiline regression		Otsustuspuu		Närvivõrk	
Tegelik\Hinnang	Maksevõimeline	Maksevõimetu	Maksevõimeline	Maksevõimetu	Maksevõimeline	Maksevõimetu
Maksevõimeline	44	94	37	101	2	136
Maksevõimetu	5	57	7	55	0	62
Kahju	144		171		136	

Tabel 21. Lisatunnused

Lisatunnused	Logistiline regression		Otsustuspuu		Närvivõrk	
Tegelik\Hinnang	Maksevõimeline	Maksevõimetu	Maksevõimeline	Maksevõimetu	Maksevõimeline	Maksevõimetu
Maksevõimeline	3	135	0	138	0	138
Maksevõimetu	1	61	0	62	0	62
Kahju	145		138		138	

Tabel 22. Algandmed ja lisatunnused

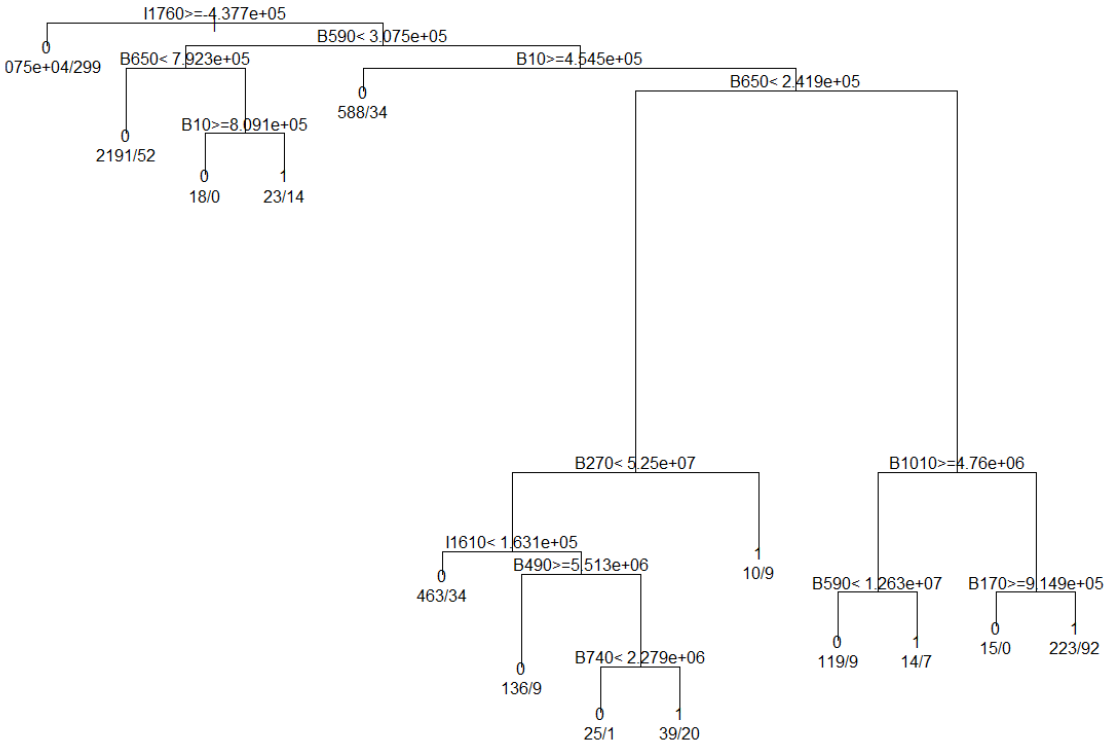
Algandmed+Lisatunnused	Logistiline regression		Otsustuspuu		Närvivõrk	
Tegelik\Hinnang	Maksevõimeline	Maksevõimetu	Maksevõimeline	Maksevõimetu	Maksevõimeline	Maksevõimetu
Maksevõimeline	47	91	0	138	5	133
Maksevõimetu	5	57	0	62	0	62
Kahju	141		138		133	

Lisa 2. Andmestik 2

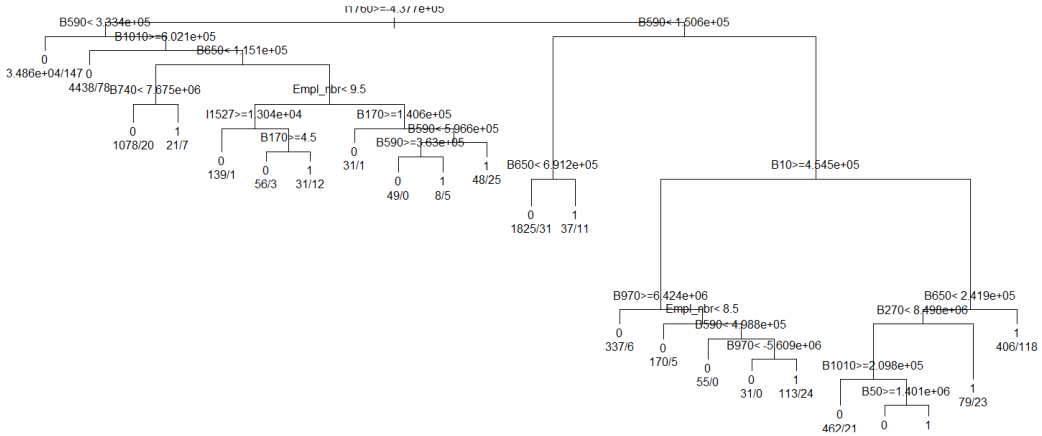
2.1. tunnuste kirjeldus

	Feature	Explanation
Balance sheet	B10	Cash And Bank
	B1010	Total Owners Equity
	B130	Accrued Income
	B170	Prepald Expenses
	B20	Marketable Securities
	B270	Total Current Assets
	B490	Total Fixed Assets
	B50	Accounts Receivable
	B590	Supplier Payables
	B650	Tax Liabilities
	B660	Accrued Expenses
	B740	Total Short Term Liabilities
	B870	Long Term Liabilities
	B950	Legal Reserves
	B970	Retained Earnings
Income sheet	I1040	Net Sales
	I1110	Other Revenues
	I1170	Personal Expenses
	I1527	Operating Profit (-loss)
	I1610	Interest Expense
	I1760	Net Result
	Empl_nbr	Number of employees

2.2. Otsustuspuud



Joonis 3. Kaofunktsiooni L_3 korral otsustuspuu kasutades ainult algandmeid



Joonis 4. Kaofunktsiooni L_4 korral otsustuspuu kasutades ainult algandmeid

2.3. Täpsustabelid ja kahju suurused kaofunktsioonist tulenevalt

Kaofunktsioon L_1

Tabel 23. Algandmed

Algandmed	Logistiline regression		Otsustuspuu		Närvivõrk	
Tegelik\Hinnang	Tegutsev	Pankrot	Tegutsev	Pankrot	Tegutsev	Pankrot
Tegutsev	11153	4	11102	55	11157	0
Pankrot	141	1	134	8	142	0
Kahju	145		189		142	

Tabel 24. Lisatunnused

Lisatunnused	Logistiline regression		Otsustuspuu		Närvivõrk	
Tegelik\Hinnang	Tegutsev	Pankrot	Tegutsev	Pankrot	Tegutsev	Pankrot
Tegutsev	11157	0	11135	22	11157	0
Pankrot	142	0	139	3	142	0
Kahju	142		161		142	

Tabel 25. Algandmed ja lisatunnused

Algandmed+Lisatunnused	Logistiline regression		Otsustuspuu		Närvivõrk	
Tegelik\Hinnang	Tegutsev	Pankrot	Tegutsev	Pankrot	Tegutsev	Pankrot
Tegutsev	11154	3	11106	51	11157	0
Pankrot	141	1	136	6	142	0
Kahju	144		187		142	

Kaofunktsioon L_2

Tabel 26. Algandmed

Algandmed	Logistiline regression		Otsustuspuu		Närvivõrk	
Tegelik\Hinnang	Tegutsev	Pankrot	Tegutsev	Pankrot	Tegutsev	Pankrot
Tegutsev	11151	6	11157	0	11157	0
Pankrot	141	1	142	0	142	0
Kahju	288		284		284	

Tabel 27. Lisatunnused

Lisatunnused	Logistiline regression		Otsustuspuu		Närvivõrk	
Tegelik\Hinnang	Tegutsev	Pankrot	Tegutsev	Pankrot	Tegutsev	Pankrot
Tegutsev	11157	0	11157	0	11157	0
Pankrot	142	0	142	0	142	0
Kahju	284		284		284	

Tabel 28. Algandmed ja lisatunnused

Algandmed+Lisatunnused	Logistiline regression		Otsustuspuu		Närvivõrk	
Tegelik\Hinnang	Tegutsev	Pankrot	Tegutsev	Pankrot	Tegutsev	Pankrot
Tegutsev	11152	5	11157	0	11157	0
Pankrot	141	1	142	0	142	0
Kahju	287		284		284	

Kaofunktsioon L_3

Tabel 29. Algandmed

Algandmed	Logistiline regression		Otsustuspuu		Närvivõrk	
Tegelik\Hinnang	Tegutsev	Pankrot	Tegutsev	Pankrot	Tegutsev	Pankrot
Tegutsev	11147	10	11072	85	11157	0
Pankrot	140	2	117	25	142	0
Kahju	710		670		710	

Tabel 30. Lisatunnused

Lisatunnused	Logistiline regression		Otsustuspuu		Närvivõrk	
Tegelik\Hinnang	Tegutsev	Pankrot	Tegutsev	Pankrot	Tegutsev	Pankrot
Tegutsev	11156	1	11157	0	11156	1
Pankrot	142	0	142	0	142	0
Kahju	711		710		711	

Tabel 31. Algandmed ja lisatunnused

Algandmed+Lisatunnused	Logistiline regression		Otsustuspuu		Närvivõrk	
Tegelik\Hinnang	Tegutsev	Pankrot	Tegutsev	Pankrot	Tegutsev	Pankrot
Tegutsev	11150	7	11076	81	11157	0
Pankrot	141	1	119	23	142	0
Kahju	712		676		710	

Kaofunktsioon L_4

Tabel 32. Algandmed

Algandmed	Logistiline regression		Otsustuspuu		Närvivõrk	
Tegelik\Hinnang	Tegutsev	Pankrot	Tegutsev	Pankrot	Tegutsev	Pankrot
Tegutsev	11146	11	10875	282	11157	0
Pankrot	140	2	94	48	142	0
Kahju	1411		1222		1420	

Tabel 33. Lisatunnused

Lisatunnused	Logistiline regression		Otsustuspuu		Närvivõrk	
Tegelik\Hinnang	Tegutsev	Pankrot	Tegutsev	Pankrot	Tegutsev	Pankrot
Tegutsev	11156	1	10934	223	11061	96
Pankrot	142	0	110	32	125	17
Kahju	1421		1323		1346	

Tabel 34. Algandmed ja lisatunnused

Algandmed+Lisatunnused	Logistiline regression		Otsustuspuu		Närvivõrk	
Tegelik\Hinnang	Tegutsev	Pankrot	Tegutsev	Pankrot	Tegutsev	Pankrot
Tegutsev	11143	14	10904	253	11157	0
Pankrot	141	1	99	43	142	0
Kahju	1424		1243		1420	

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina: Paavo Binsol ,

(*autori nimi*)

(sünnikuupäev: 01.08.1991)

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

(*lõputöö pealkiri*)

mille juhendaja on Raul Kangro ,
(*juhendaja nimi*)

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 **13.05.2015**