Tartu Ülikool

Loodus- ja täppisteaduste valdkond

Matemaatika ja statistika instituut

Oskar Kärmas

Krediidiskooringu süsteemi loomine laenu mittesaajate kaasamisega

Matemaatilise statistika eriala

Bakalaureusetöö (9 EAP)

Juhendaja prof. Kalev Pärna

Tartu 2016

Krediidiskooringu süsteemi loomine laenu mittesaajate kaasamisega

Üksnes laenusaajate andmestiku põhjal leitud krediidiskooringu süsteemi võimaliku kallutatuse vältimiseks peab süsteemi loomisel kaasama ka laenu mittesaajad. Käesoleva bakalaureusetöö eesmärgiks on välja selgitada, millise kaasamismeetodi rakendamise tulemusena välja töötatud krediidiskooringu süsteem prognoosib laenutaotleja staatust kõige täpsemalt ning rakendada seda meetodit reaalsel andmestikul. Töös uuritakse nelja kaasamismeetodit. Krediidiskooringu süsteemi loomiseks kasutatakse logistilist regressioonanalüüsi. Töö tulemusena töötatakse välja aktsepteeritava kvaliteediga krediidiskooringu süsteem, mis on loodetavasti vähem kallutatud ja mida saab rakendada kõikidel laenutaotlejatel.

Märksõnad: *krediidireiting, regressioonanalüüs, mudelid, kaasamismeetod*
P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika

Developing a credit scoring system with reject inference

Rejected applicants must be involved in the process of developing a credit scoring system in order to avoid possible bias. Reject inference methods are used to develop credit scoring systems and the aim of this thesis is to find which of these systems is the most precise to predict the status of the applicant and therefore which is the best reject inference method to use. The best method is used on a real data set. Logistic regression is used for creating the credit scoring system. As a result of this thesis, an acceptable and hopefully less biased credit scoring system is developed which can be applied to all loan applicants.

Keywords: *credit rating, regression analysis, models, reject inference method*
P160 Statistics, operation research, programming, actuarial mathematics

Sisukord

Sissejuhatus

Krediidiasutuste äritegevuse üks osa on laenude väljastamine. Laenuandmise otsustusprotsessis kasutatakse ühe meetodina krediidiskooringu süsteemi, mis hindab tõenäosust, et klient osutub ettevõtte seisukohast heaks kliendiks. Hea või halva kliendi definitsioon on iga ettevõtte enda määrata – hea klient võib olla näiteks ilma makseviivitusteta klient. Kliendi reaalne staatus on ettevõttele teada vaid nende klientide puhul, kellele on laenu antud. Seega on esialgne krediidiskooringu süsteem võimalik välja töötada laenusaajate andmete pealt, kuid tulemuseks on kallutatud mudel, mida ei saa rakendada kõikide laenutaotlejate peal. Kõikide potentsiaalsete klientide peal rakendatava krediidiskooringu süsteemi loomisel tuleb kaasata ka laenu mittesaajate andmed.

Laenu mittesaajate kaasamismeetodeid on mitmeid, millest käesolevas bakalaureusetöös uuritakse nelja. Töö eesmärgiks on välja selgitada, millise kaasamismeetodi rakendamise tulemusena välja töötatud krediidiskooringu süsteem prognoosib laenutaotleja staatust kõige täpsemalt, ning rakendada seda meetodit reaalsel andmestikul. Krediidiskooringu süsteemi loomisel kasutatakse logistilist regressioonanalüüsi.

Töö esimeses peatükis antakse ülevaade krediidiskooringu süsteemi olemusest ning laenu mittesaajate kaasamisest. Teises peatükis kirjeldatakse logistilise regressioonanalüüsi meetodit. Kolmandas peatükis kirjeldatakse kasutatud reaalseid andmestikke, viiakse läbi tunnuste esmaanalüüs ning luuakse laenusaajate andmestiku pealt logistiline regressioonimudel. Neljandas peatükis antakse ülevaade laenu mittesaajate kaasamismeetoditest ning teostatakse eksperiment parima kaasamismeetodi välja selgitamiseks. Viiendas peatükis rakendatakse parimat kaasamismeetodit reaalsel andmestikul ning luuakse krediidiskooringu süsteem.

Töö on kirjutatud tekstitöötlusprogrammiga Microsoft Word 2010 ning analüüsid on läbi viidud statistikatarkvaraga SAS (versioon 9.2).

Autor tänab juhendajat professor Kalev Pärnat rohkete nõuannete ja soovituste eest.

1 Ülevaade krediidiskooringust

1.1 Krediidiskooringu süsteem

Laenu ehk krediidi pakkumisega kaasneb ettevõttele risk, et klient ei suuda oma kohustusi ettevõtte ees täita ning osutub halvaks kliendiks. Seda riski arvestatakse toote või teenuse pakkumise otsustusprotsessis. Traditsiooniline meetod on hinnata riski vastavalt intuitsioonile, mis on kujunenud eelnevate kliendikäitumiste põhjal. Majandusliku surve ning tehnoloogia arengu tulemusena on välja töötatud statistilised mudelid, mis on abivahenditeks otsustusprotsessil. Neid statistilisi mudeleid nimetatakse krediidiskooringu süsteemideks. Krediidiskooringu süsteemid prognoosivad kliendi kohta olemas olevate andmete toel tõenäosuse, et klient osutub heaks kliendiks. Klient on madala riskiga, kui tal on suur tõenäosus olla hea klient, ning kõrge riskiga, kui tal on väike tõenäosus olla hea klient. [1]

Krediidiasutuse kui ettevõtte eesmärk on kasumi maksimeerimine. Kasumi maksimeerimise üks strateegiaid on ettevõtte tegevuste ja protsesside efektiivistamine ning nendega seotud kulude vähendamine. Krediiditoodete lai levik ja kliendibaasi suurenemine on seadnud krediidiasutustele eesmärgi muuta laenutaotluste otsustusprotsessi kiiremaks, stabiilsemaks ja vähem kulukaks. Samuti on oluline, et otsustusprotsess minimeeriks tagasilükatud heade klientide ning laenusaanud halbade klientide arvu. [2, lk 1-2]

Varasemalt ostsid krediidiasutused klientidel rakendatava krediidiskooringu süsteemi vastavat teenust pakkuvatelt asutustelt. Tänu tehnoloogia arengule on tänapäeval laialt levinud praktika, et ettevõtted töötavad ise välja oma krediisiskooringu süsteemid. Ettevõttesiseselt luuakse süsteemid kiiremini, odavamalt ja paindlikumalt kui varem. Statistika rakendustarkvarade kättesaadavuse suurenemise ja andmete salvestamise süsteemide lihtsustumise tõttu ei pea ettevõtted tegema väga suuri investeeringuid infrastruktuuri ja programmeerijatesse. Seetõttu on ise krediidiskooringu süsteemide loomine odavam kui nende sisseostmine. Samuti võimaldab ettevõtte põhjalikum arusaam oma äriprotsessidest välja töötada paremini töötavaid süsteeme. [2, lk 2-3]

Krediidiskooringu süsteemide põhjal tehtud analüüside toel saab krediidiasutus parema ülevaate klientide riskikäitumisest, mis võimaldab välja töötada efektiivsemad äristrateegiad. Näiteks klientide puhul, kellel on suur prognoositud tõenäosus olla halb klient, on ühed võimalikud strateegiad:

- laenu/krediiditaotlus tagasi lükata;
- määrata madalam krediidilimiit (nt krediitkaardil või arvelduskrediidil);
- määrata kõrgem intressimäär;
- anda laenu, kuid võtta klient rangema jälgimise alla. [2, lk 7-8]

Seega klientide puhul, kellel on suur prognoositud tõenäosus olla hea klient, võib määrata kõrgema krediidilimiidi, madalama intressimäära ning pakkuda eksklusiivsemaid tooteid (nt kuld- ja plaatinumkrediitkaarte). [2, lk 8]

Kuigi krediidiskooringu fookuses on tõenäosus, et klient osutub heaks või halvaks kliendiks, ei ole see tõenäosus alati monotoonselt seotud ettevõtte kasumlikkusega. Näiteks madala riskiga kliendid, kes tasuvad krediitkaardi kasutatud krediidi enne kui ettevõtte hakkab intressi koguma, ei ole selle toote seisukohast kasumlikud. Samas kõrge riskiga kliendid saavad olla kasumlikud, sest kõrge riski tõttu rakendatakse ka märgatavalt kõrgemat intressimäära. [1] Seetõttu on krediidiskooringu süsteem krediiditaotluste otsustusprotsessis vaid üks komponent. Praktikas kasutatakse lisaks krediidiskoorile ka ekspertarvamust ning vastavate lähenemiste osakaal on iga ettevõtte enda määrata.

Krediidiskooringu süsteemi loomisel kasutatakse laialdaselt logistilist regressioonanalüüsi, kuna uuritav tunnus, kas klient on hea või halb, on binaarne. Samuti kasutatakse lineaarset regressioonanalüüsi, diskriminantanalüüsi ning tehisõppe meetodeid. Kuigi loetletud meetodid on laialt kasutusel, on krediidiskooringu süsteemide alane teadusliku kirjanduse hulk kesine. Selle põhjuseks on andmete konfidentsiaalsuse nõue ja asjaolu, et krediidiskooringu süsteem on osa krediidiasutuse ärisaladusest ning selle välja töötamise meetodite avaldamine võib kahjustada ettevõtte huve. [1]

1.2 Laenu mittesaajate kaasamine

Krediidiskooringu süsteemi eesmärk on prognoosida kõikide laenutaotlejate puhul tõenäosus, et klient osutub heaks. Kliendi reaalne staatus (hea või halb) selgub aga ainult nende klientide puhul, kellele on laenu antud. Seega ainult laenu saanud klientide andmete põhjal välja töötatud krediidiskooringu süsteem võib olla kallutatud ning seda ei saa rakendada kõikide laenu taotlevate klientide peal [3].

Kui klient taotleb laenu, edastab ta krediidiasutusele enda andmed, mis on ettevõtte jaoks vajalikud, et teha otsus, kas laenu anda või mitte. Seega on ettevõttel olemas vastavad andmed

nii laenusaajate kui ka laenu mittesaajate kohta, kuid kliendi staatus vaid laenusaajate kohta. Sellest tulenevalt on tekkinud idee, mis seisneb staatuste tuletamises laenu mittesaajatele. Laenusaajate andmestiku pealt töötatakse välja mudel, millega prognoositakse laenu mittesaajatele hea staatuse tõenäosused. Seejärel rakendatakse mõnda meetodit mittesaajate kaasamiseks liitandmestikku. Liitandmestiku, kus on nii laenusaajad kui ka laenu mittesaajad, pealt töötatakse välja krediidiskooringu süsteem, mis peaks olema vähem kallutatud kui ainult laenusaajate pealt loodud süsteem ning mida saab kasutada kõikide laenutaotlejate puhul. [3]

Laenu mittesaajate kaasamisel on lisaks kallutatuse vähendamise eesmärgile ka ärilised põhjused. Mittesaajate kohta info kaasamine võimaldab ettevõttel majandustegevust täpsemalt ja realistlikumalt prognoosida. Näiteks, kui ettevõtte soovib laenuandmise tingimusi lõdvendada, siis antaks laenu osale klientidest, kes siiani on olnud tagasilükatud. Nende klientide prognoositud staatuste tõenäosuste pealt on ettevõttel võimalik tingimuste lõdvendamisega seotud riske hinnata. Samuti võimaldab laenu mittesaajate kaasamine vähendada klientide arvu, kelle laenutaotlust ei rahuldatud, kuid kes oleksid osutunud headeks klientideks. [2, lk 100]

Laenu mittesaajatele staatuse prognoosimise alternatiiviks on reaalsete andmete kasutamine. See eeldab, et ettevõte annab info kogumise eesmärgil laenu ka nendele klientidele, kellele senise krediidiskooringu süsteemi põhjal laenu ei antaks. Ärilises mõttes on selline praktika kasumlik siis, kui halbadest laenudest lisanduv kulu on väiksem kui täiendava info abil välja töötatud täpsema krediidiskooringu süsteemi rakendamisest saadav tulu. Praktikas kasutatakse sellist lähenemist harva. Levinum on laenu mittesaajate kohta info hankimine krediidiasutustelt, kes on vastavatele klientidele juba laenu andnud. [1]

Äritegevuse käigus koguneb ettevõttele andmeid klientide kohta pidevalt juurde. Samuti toimub vastavalt muutlikule majandusseisule muutused laenu taotlevate isikute rahvastikuprofiilis [1]. Seetõttu kalibreerivad krediidiasutused vastavalt vajadusele oma krediidiskooringu süsteeme. Näiteks, Swedbank AS, Eesti hindab vähemalt kord aastas, kas krediidiskooringu süsteemid töötavad korrektselt [4, lk 21].

Laenu mittesaajate kaasamismeetodeid on mitmeid. Käesolevas bakalaureusetöös võrreldakse nelja meetodit. Hand ja Henley [3] on aga märkinud, et usaldusväärset ning igas olukorras rakendatavat universaalset laenu mittesaajate kaasamismeetodit, millega krediidiskooringu süsteem muutub paremaks, ei eksisteeri. Paremate tulemustega süsteemid on loodud kas tänu

juhusele, kasutades lisainformatsiooni või muutes kaasamismeetodi tingimusi paremale tulemusele orienteeritud suunas.

2 Ülevaade logistilisest regressioonanalüüsist

2.1 Logistilise regressioonimudeli kuju

Järgnev alapeatükk põhineb Kleinbaumi, Kupperi, Mulleri ja Nizami raamatul „Applied Regression Analysis and Other Multivariable Methods“ [5, lk 656-659] ja E. Kääriku loengukonspektil „Andmeanalüüs II“ [6, lk 110-111].

Uuritava tunnuse ja ühe või mitme seletava tunnuse omavahelise seose kirjeldamise lahutamatuks osaks on regressioonanalüüsi meetodite rakendamine. Tihti pakub analüüsijale huvi diskreetne uuritav tunnus, millel on ainult kaks võimalikku väärtust: on/ei ole, jah/ei, esineb/ ei esine. Seega on uuritava tunnuse puhul tegemist binaarse tunnusega Y, mille kodeerimisel kasutatakse tavaliselt väärtusi 1 ja 0. Logistiline regressioonimudel on binaarse uuritava tunnuse puhul kõige sagedamini kasutatav analüüsimeetod.

Seletavate tunnuste $X_1, X_2, \dots, X_k$ korral on logistiline regressioonimudel uuritava tunnuse keskväärtusele kujul

$$E(Y) = \frac{1}{1 + e^{-\left(\beta_0 + \sum_{j=1}^{k} \beta_j X_j\right)}}.$$

Tähistame binaarse uuritava tunnuse Y sündmuse esinemise tõenäosuse $P(Y = 1) = \pi$ ning sündmuse mitte esinemise tõenäosuse $P(Y = 0) = 1 - \pi$. Uuritav tunnus on Bernoulli jaotusega $Y \sim B(1, \pi)$, mille puhul keskväärtus $E(Y) = \pi$ ja dispersioon $D(Y) = \pi(1 - \pi)$. Seega on uuritava tunnuse keskväärtus võrdne sündmuse esinemise tõenäosusega ning logistiline regressioonimudel hindab sündmuse toimumise tõenäosust

$$\pi = \frac{1}{1 + e^{-\left(\beta_0 + \sum_{j=1}^{k} \beta_j X_j\right)}}. \tag{2.1}$$

Eelneva valemi (2.1) parem pool on üldistatuna kujul

$$f(z) = \frac{1}{1 + e^{-z}},$$

kus $z = \beta_0 + \sum_{j=1}^{k} \beta_j X_j$. Funktsiooni $f(z)$ nimetatakse logistiliseks funktsiooniks. Logistiline funktsioon sobib hästi tõenäosuste prognoosimiseks, sest kuigi z varieerub vahemikul $(-\infty; \infty)$, siis $f(z)$ väärtused asuvad lõigul $[0; 1]$. Seega on logistilise regressioonimudeli poolt prognoositud sündmuse esinemise tõenäosused alati 0 ja 1 vahel.

Logistilise regressioonimudeli kirjeldamise puhul kasutatakse lähtekuju (2.1) asemel tihti ka *logit* kuju. *Logit* on seosefunktsioon uuritava tunnuse keskväärtusest ehk huvipakkuva sündmuse toimumise tõenäosusest kujul

$$logit(\pi) = ln\frac{\pi}{1-\pi}, \tag{2.2}$$

kus $\frac{\pi}{1-\pi}$ on sündmuse esinemise šanss. *Logit* on seega šansi logaritm.

Asendades valemis (2.2) sündmuse toimumise tõenäosuse π logistilise regressioonimudeli valemiga (2.1), saame logistilise regressioonimudeli *logit* kuju

$$logit(\pi) = ln\frac{\pi}{1-\pi} = \beta_0 + \sum_{j=1}^{k} \beta_j X_j. \tag{2.3}$$

2.2 Parameetrite hindamine

Järgnev alapeatükk põhineb Kleinbaumi, Kupperi, Mulleri ja Nizami raamatul „Applied Regression Analysis and Other Multivariable Methods“ [5, lk 671-673], v.a seal, kus on märgitud teisiti.

Logistilises regressioonimudelis (2.3) olevate tundmatute β_i leidmiseks kasutatakse suurima tõepära meetodit. Järgnevalt anname ülevaate tõepärafunktsioonist, mida kasutab statistikatarkvara *SAS* protseduuri *LOGISTIC* korral.

Logistilise regressioonimudeli korral on uuritavaks tunnuseks Bernoulli jaotusega binaarne tunnus Y, tõenäosustega $P(Y = 1) = \pi$ ja $P(Y = 0) = 1 - \pi$. Bernoulli valem on sellisel juhul

$$P(Y; \pi) = \pi^Y (1-\pi)^{1-Y}, \quad Y = 0, 1.$$

Bernoulli valem üldistatuna n liikmelise valimi jaoks, kus Y_i on vaatluse i uuritava tunnuse väärtus ning $i = 1, 2, \dots, n$, on kujul

$$P(Y_i; \pi_i) = {\pi_i}^{Y_i}(1 - \pi_i)^{1-Y_i}, \quad Y_i = 0, 1,$$

kus π_i on i-nda vaatluse sündmuse esinemise tõenäosus.

Kuna vaatluste uuritavad tunnused $Y_1, Y_2, \dots, Y_n$ on omavahel sõltumatud, siis on tõepärafunktsioon kujul

$$L(\boldsymbol{Y}; \boldsymbol{\pi}) = \prod_{i=1}^{n} P(Y_i; \pi_i) = \prod_{i=1}^{n} [{\pi_i}^{Y_i}(1 - \pi_i)^{1-Y_i}], \tag{2.4}$$

kus $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_n)$ ja $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_n)$.

Olgu esimese n_1 vaatluse uuritava tunnuse väärtuseks 1 ning ülejäänud $n - n_1$ vaatluse uuritava tunnuse väärtuseks 0. Sellisel juhul on tõepärafunktsioon (2.4) kujul

$$L(\boldsymbol{Y}; \boldsymbol{\pi}) = \left(\prod_{i=1}^{n_1} \pi_i \right) \left[\prod_{i=n_1+1}^{n} (1 - \pi_i) \right]. \tag{2.5}$$

Tähistagu $\boldsymbol{X_i} = X_{i1}, X_{i2}, \dots, X_{ik}$ vaatluse i seletavate tunnuste komplekti $X_1, X_2, \dots, X_k$. Logistilise regressioonimudeli (2.1) kohaselt on π_i ja X_{ij} vaheline seos

$$\pi_i = \frac{1}{1 + e^{-\left(\beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}\right)}}, \tag{2.6}$$

kus $i = 1, 2, \dots, n$ ja $\beta_j, j = 0, 1, \dots, k$ on parameetrid, millele peab leidma väärtused.

Asendades tõepärafunktsiooni valemis (2.5) π_i logistilise regressioonimudeli valemiga (2.6) ning lihtsustades, saame tõepärafunktsiooniks

$$L(\boldsymbol{Y};\boldsymbol{\beta}) = \frac{\prod_{i=1}^{n_1} e^{\left(\beta_0+\sum_{j=1}^{k}\beta_j X_{ij}\right)}}{\prod_{i=1}^{n}\left[1+e^{\left(\beta_0+\sum_{j=1}^{k}\beta_j X_{ij}\right)}\right]}, \tag{2.7}$$

kus $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)$

Suurima tõepära meetod seisneb sellise $\boldsymbol{\beta}$ väärtuse leidmises, mille korral tõepärafunktsioon (2.7) saavutab maksimumi. Statistikatarkvara *SAS* kasutab selle leidmiseks kas Newton – Raphsoni meetodit või Fisheri skoorimeetodit. [6, lk 107]

2.3 Mudeli ja parameetrite olulisus

Järgnev alapeatükk põhineb Hosmeri, Lemeshow ja Sturdivanti raamatul „Applied Logistic Regression“ [7, lk 12-14, 39-42] ja E. Kääriku loengukonspektil „Andmeanalüüs II“ [6, lk 107].

Logistilise regressioonimudeli olulisuse testimiseks kasutatakse tõepärasuhte statistikut, Waldi statistikut ja skooristatistikut. Statistikud testivad hüpoteesipaari

- H_0: $\beta_j = 0,\ j = 0, 1, \dots, k$;
- H_1: $\exists j', \beta_{j'} \neq 0$.

Nullhüpoteesi kehtides on kõik eelnevad statistikud hii-ruut jaotusega. Praktikas soovitatakse kasutada tõepärasuhte statistikut selle paremate omaduste tõttu.

Tõepärasuhte statistik (*likelihood ratio*) on kujul

$$G = -2\,ln\left[\frac{(\text{ainult konstanti sisaldava mudeli tõepära})}{(\text{uuritava mudeli tõepära})}\right].$$

Nullhüpoteesi kehtides on tõepärasuhte statistik hii-ruut jaotusega vabadusastmete arvuga *k*. Kui tõepärasuhte statistiku olulisuse tõenäosus on väiksem kui etteantud olulisuse nivoo, võime kummutada nullhüpoteesi. Sel juhul leidub üks või mitu parameetrit, mis on statistiliselt olulised ning sellest järelduvalt on ka uuritav mudel statistiliselt oluline.

Mudeli üksikute parameetrite olulisuse kontrollimiseks kasutab statistikaprogramm *SAS* Waldi hii-ruut statistikut

$$W^2 = \left(\frac{\hat{\beta}_j}{\widehat{SE}(\hat{\beta}_j)}\right)^2,$$

kus $j = 0, 1, \ldots, k$, $\hat{\beta}_j$ on *j*-nda parameetri hinnang ning $\widehat{SE}(\hat{\beta}_j)$ on *j*-nda parameetri hinnangu standardvea hinnang.

Waldi hii-ruut statistik on hii-ruut jaotusega, vabadusastmete arvuga üks. Kui mudelisse kaasatud tunnuse Waldi hii-ruut statistiku olulisuse tõenäosus on väiksem kui olulisuse nivoo, siis see tunnus on statistiliselt oluline.

2.4 *ROC* kõver ja Hosmer – Lemeshow test

Järgnev alapeatükk põhineb Hosmeri, Lemeshow ja Sturdivanti raamatul „Applied Logistic Regression“ [7, lk 157-158, 173-178].

Logistilise regressionimudeli headuse analüüsimiseks kasutatakse enamasti *ROC* kõverat. Mudeliga prognoositakse igale vaatlusele sündmuse esinemise tõenäosus. Binaarse sündmuse toimumise prognoosimiseks tuleb aga valida lävend, mille alusel määratakse prognoos, kas antud vaatlusel sündmus toimub või mitte. Näiteks võib valida lävendiks 0,5, mille korral määratakse sellest suurema tõenäosuse prognoosiga vaatlustel sündmuse toimumise väärtuseks 1 ja väiksema tõenäosuse prognoosiga vaatlustel 0.

Klassifitseerimise tulemusena saab vaadelda klassifitseerimise tundlikkust (*sensitivity*) ja spetsiifilisust (*specificity*). Tundlikkus on tõeselt positiivsete vaatluste arvu jagatis vaatluste arvuga, millel huvipakkuv sündmus reaalselt toimus. Spetsiifilisus on tõeselt negatiivsete vaatluste arvu jagatis vaatluste arvuga, millel huvipakkuv sündmus reaalselt ei toimunud. Seega näitab tundlikkus tõeselt positiivsete vaatluste määra ning (*1 – spetsiifilisus*) valepositiivsete vaatluste määra.

Tundlikkus ja spetsiifilisus sõltuvad klassifitseerimise lävendi valikust. *ROC* kõver konstrueeritakse kuvades horisontaalteljel valepositiivsete vaatluste määra ja vertikaalteljel tõeselt positiivsete määra üle lävendi kõikide võimalike väärtuste.

ROC kõvera aluse pindala (*AUC*) väärtus iseloomustab mudeli võimet korrektselt prognoosida sündmuse esinemise tõenäosusi. Vastavalt *ROC* kõvera aluse pindala väärtusele jagatakse mudelid järgnevatesse klassidesse:

- $AUC = 0{,}5$ korrektne prognoosimisvõime puudub,
- $0{,}5 < AUC < 0{,}7$ halb mudel,
- $0{,}7 \leq AUC < 0{,}8$ aktsepteeritav mudel,
- $0{,}8 \leq AUC < 0{,}9$ hea mudel,
- $AUC \geq 0{,}9$ väga hea mudel.

Hosmer – Lemeshow testi kasutatakse mudeli sobitusastme uurimiseks. Vaatlused järjestatakse prognoositud sündmuse esinemise tõenäosuste järgi ning jagatakse vaatluste arvu järgi kümneks võrdseks grupiks. Hosmer – Lemeshow teststatistik on kujul

$$C = \sum_{k=1}^{10} \frac{(O_k - N_k\overline{\pi_k})^2}{N_k\overline{\pi_k}(1-\overline{\pi_k})},$$

kus O_k on huvipakkuva sündmuse esinemise arv grupis k, N_k on grupi k vaatluste arv ja $\overline{\pi_k}$ on prognoositud tõenäosuste keskmine grupis k. Teststatistik on hii-ruut jaotusega vabadusastmete arvuga kaheksa. Kui Hosmer – Lemeshow statistiku olulisuse tõenäosus on suurem kui olulisuse nivoo, ei saa ümber lükata nullhüpoteesi, et mudel sobib andmestikuga.

2.5 Mudeli interpretatsioon

Järgnev alapeatükk põhineb E. Kääriku loengukonspektil „Andmeanalüüs II“ [6, lk 111].

Logistilise regressioonimudeli parameetrite β_j interpretatsioon seisneb šansside suhte muutuse kirjeldamises. Šansside suhe on i-nda ja i'-nda vaatluse šansside suhe kujul

$$OR = \frac{\frac{\pi_i}{1-\pi_i}}{\frac{\pi_{i'}}{1-\pi_{i'}}}.$$

Tingimusel, et teiste argumentide väärtused ei muutu, kehtib reegel, et j-nda argumendi muutusega c ühiku võrra kaasneb šansside suhte muutus $e^{c\beta_j}$ korda. Positiivne parameeter β_j argumendi X_j ees näitab samapidist seost uuritava sündmuse tõenäosuse ja argumendi X_j vahel ning negatiivne parameeter vastupidist seost.

3 Laenusaajate logistiline regressioonimudel

3.1 Andmestike kirjeldus

Käesolevas bakalaureusetöös on kasutatud kahte andmestikku – laenusaajate andmestikku ja laenu mittesaajate andmestikku. Mõlema puhul on tegemist fragmendiga reaalsest andmestikust. Laenusaajate andmestikus on 1800 vaatlust ning laenu mittesaajate andmestikus 1599 vaatlust. Laenusaajate andmestikus on 17 tunnust:

- laenusaaja staatus, kus 1 – hea, 0 – halb (*staatus*)
- laenusaaja sugu, kus M – mees, F – naine (*sugu*)
- laenusaaja vanus aastates (*vanus*)
- laenusaaja elukoha maakonna nimetus (*maakond*)
- laenusaaja emakeel, kus est – eesti keel, rus – vene keel (*keel*)
- laenusumma eurodes (*summa*)
- laenuperiood päevades (*periood*)
- laenusaaja kuine sissetulek eurodes (*sissetulek*)
- laenusaaja kuine väljaminek eurodes (*valjaminek*)
- laenusaaja pereseis (*pereseis*)
- laenusaaja haridustase (*haridus*)
- laenusaaja töökogemus (*tookogemus*)
- laenusaaja laste arv (*lapsed*)
- laenusaaja omanduses olevate kinnisvaraobjektide arv (*kinnisvara*)
- laenusaaja aktiivsete maksehäirete arv (*mh_akt*)
- laenusaaja lõpetatud maksehäirete arv (*mh_lop*)
- laenusaaja maksehäirete arv kokku (*mh_koik*)

Laenu mittesaajate andmestikus on kõik ülaltoodud tunnused laenu mittesaajate kohta v.a *staatus*, mida pole võimalik laenu mittesaajatel määrata. Seega on laenu mittesaajate andmestikus 16 tunnust.

Laenusaajate andmestikus on hea staatusega kliente 1247 (69,3%) ning halva staatusega kliente 553 (30,7%).

3.2 Laenusaajate andmestiku tunnuste esmaanalüüs ja teisendamine

Sugu

Laenusaajate hulgas on mehi ja naisi peaaegu võrdselt – 905 meest ja 895 naist. Tabelist 1 näeme, et 62,9% meestest (569 vaatlust) on head kliendid ning naiste puhul vastavalt 75,8% (678 vaatlust). Tunnuse *sugu* ja tunnuse *staatus* sõltuvust testiti hii-ruut testiga. Hii-ruut statistiku väärtus on 35,08 ja p-väärtus $< 0{,}0001$. Seega on *staatus* ja *sugu* sõltuvad ning võime öelda, et naiskliendid on parema maksekäitumisega kui mehed.

Tabel 1. Tunnuse *sugu* sagedustabel

staatus/*sugu*	M	F	**Kokku**
1	569	678	**1247**
0	336	217	**553**
Kokku	**905**	**895**	**1800**

Vanus

Tunnuse *vanus* minimaalne väärtus on 18 ning maksimaalne 67. Tunnuse keskväärtuseks on 39,4.

Vanuse ja staatuse seose uurimiseks sorteeriti andmestik tunnuse järgi kasvavas järjekorras ning jaotati detsiilrühmadeks. Iga rühma puhul leiti tunnuse keskväärtus ning šansi logaritm kujul

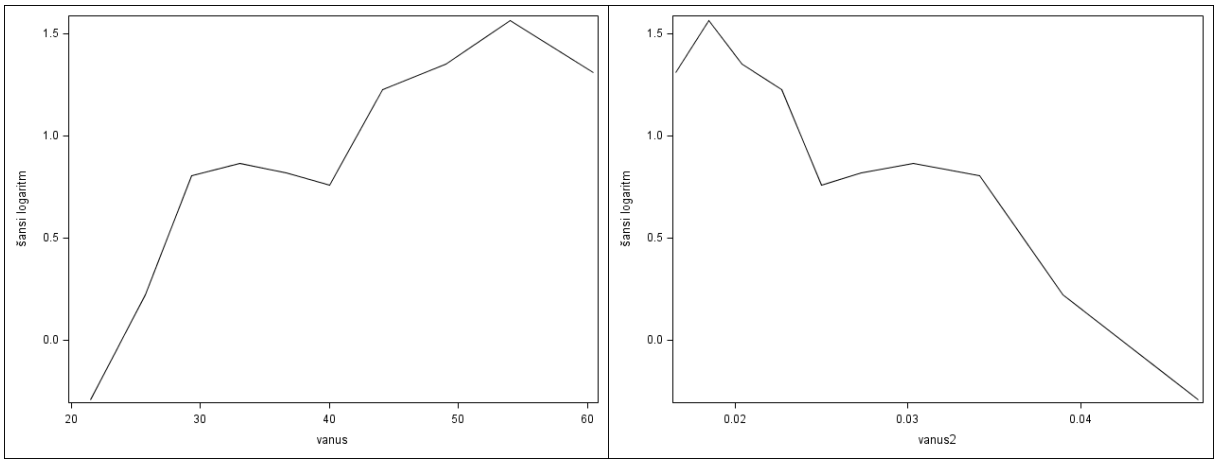
$$\text{šansi } logaritm = \ln\left(\frac{\overline{staatus}}{1-\overline{staatus}}\right),$$

kus $\overline{staatus}$ on tunnuse *staatus* keskväärtus detsiilrühmas. Joonisel 1 vasakul on kujutatud šansside logaritmid tunnuse *vanus* erinevates detsiilrühmades.

Proovimise tulemusena selgus, et parema tulemuse annab tunnuse *vanus* pöördväärtuse

$$vanus2 = \frac{1}{vanus}$$

kasutamine, mille seos staatusega on toodud joonisel 1 paremal. Jooniselt on näha, et esineb ligikaudu lineaarne seos tunnuse *vanus2* ja šansside logaritmide vahel.



Joonis 1. Šansside logaritmid tunnuste *vanus* (vasakul) ja *vanus2* (paremal) detsiilrühmades

Maakond

Tunnuse *maakond* väärtused on välja toodud tabelis 2 koos heade klientide osakaalude ja kõikide klientide arvudega.

Tabel 2. Heade klientide osakaalud tunnuse *maakond* väärtuste lõikes

maakond	heade klientide osakaal	klientide arv kokku
Hiiumaa	100,0%	3
Valgamaa	80,0%	25
Harjumaa	73,4%	1019
Läänemaa	70,0%	20
Ida-Virumaa	67,4%	242
Tartumaa	65,9%	173
Raplamaa	64,1%	39
Jõgevamaa	62,5%	24
Pärnumaa	62,0%	71
Lääne-Virumaa	61,8%	55
Võrumaa	58,8%	17
Saaremaa	57,9%	19
Põlvamaa	57,7%	26
Järvamaa	54,1%	37
Viljandimaa	36,7%	30

Tunnuse *maakond* väärtused grupeeriti kolme rühma põhimõttel, et rühmas oleksid homogeensete heade klientide osakaaludega maakonnad. Grupeerimise tulemusena loodi tunnus *maakond2*, kus rühma MK1 kuulub Harjumaa, rühma MK2 kuuluvad Valgamaa, Läänemaa, Ida-Virumaa, Tartumaa, Raplamaa ja rühma MK3 kuuluvad Jõgevamaa, Pärnumaa, Lääne-Virumaa, Võrumaa, Saaremaa, Hiiumaa, Põlvamaa, Järvamaa ning Viljandimaa. Harjumaa kaasati eraldi rühmana rohke vaatluste arvu tõttu (1019 vaatlust).

Tabelist 3 näeme, et maakondade gruppi MK1 kuuluvate laenusaajate seas on häid kliente 73,4% (748 vaatlust). Grupi MK2 puhul on häid kliente 67,3% (336 vaatlust) ning grupi MK3 puhul 57,8% (163 vaatlust). Tunnuse *maakond2* ja tunnuse *staatus* sõltuvust testiti hii-ruut testiga. Hii-ruut teststatistiku väärtus on 26,49 ning p-väärtus $< 0{,}0001$. Seega on tunnused *staatus* ja *maakond2* sõltuvad ning võime öelda, et Harjumaal elavad kliendid on parema maksekäitumisega, kui teistes maakondade gruppides elavad kliendid.

Tabel 3. Tunnuse *maakond2* sagedustabel

staatus/*maakond2*	MK1	MK2	MK3	**Kokku**
1	748	336	163	**1247**
0	271	163	119	**553**
Kokku	**1019**	**499**	**282**	**1800**

Keel

Laenusaajate hulgas on eesti keelt kõnelevaid kliente 1000 ning vene keelt kõnelevaid 800. Tabelist 4 näeme, et eesti keelt kõnelevate laenusaajate seas on häid kliente 66,8% (668 vaatlust). Vene keelt kõnelevate laenusaajate seas on häid kliente 72,4% (579 vaatlust). Tunnuse *keel* ja tunnuse *staatus* sõltuvust testiva hii-ruut teststatistiku väärtuseks on 6,49 ning p-väärtus on 0,01. Seega on tunnused *staatus* ja *keel* sõltuvad ning võime öelda, et vene keelt kõnelevad kliendid on parema maksekäitumisega.

Tabel 4. Tunnuse *keel* sagedustabel

staatus/*keel*	est	rus	**Kokku**
1	668	579	**1247**
0	332	221	**553**
Kokku	**1000**	**800**	**1800**

Summa

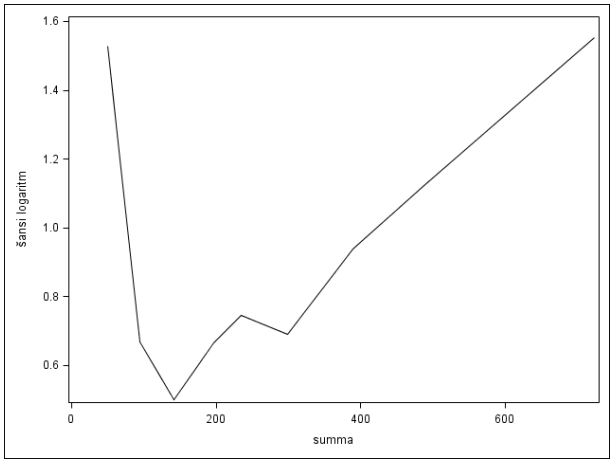
Laenusaajate poolt laenatud summa minimaalne väärtus on 50 EUR ning maksimaalne 2000 EUR. Keskmine laenusumma on 284,8 EUR.

Tunnus *summa* ei osutunud oluliseks tunnuseks logistilistes regressioonimudelites. Laenusumma ja kliendi staatuse seose uurimiseks jaotati laenusaajad vastavalt tunnuse *summa* väärtustele üheksaks grupiks ning leiti iga grupi šansi logaritm samal põhimõttel nagu tunnuse *vanus2* korral. Joonisel 2 on näha, et tunnuse *summa* ja šansside logaritmide vahel arvatavasti ei esine lineaarset seost, küll aga on märgata tükati lineaarset seost. Selleks tuleb tunnus *summa* poolitada *summa* = 150 pealt ja moodustada tunnused *summa1* ja *summa2* järgmiselt:

$$summa1 = \begin{cases} summa, & kui\ summa \leq 150 \\ 0, & kui\ summa\ > 150 \end{cases}$$

ja

$$summa2 = \begin{cases} 0, & kui\ summa \leq 150 \\ summa, & kui\ summa\ > 150. \end{cases}$$



Joonis 2. šansside logaritmid tunnuse *summa* erinevates suurusklassides

Olenemata eeldatavast tükati lineaarsest seosest, ei osutunud ka tunnused *summa1* ja *summa2* mudelites olulisteks tunnusteks.

Periood

Laenusaajate laenuperioodi tunnuse *periood* minimaalne väärtus on 1 päev ning maksimaalne 720 päeva. Keskmine laenuperioodi pikkus on 92,9 päeva.

Tunnuse *periood* väärtused grupeeriti viide rühma ning moodustati tunnus *periood2* järgmiselt:

$$periood2 = \begin{cases} "\leq 30", & kui\ periood \leq 30 \\ "31\text{-}60", & kui\ 31 \leq periood \leq 60 \\ "61\text{-}120", & kui\ 61 \leq periood \leq 120 \\ "121\text{-}180", & kui\ 121 \leq periood \leq 180 \\ ">180", & kui\ periood > 180. \end{cases}$$

Tabelist 5 näeme, et vähem kui 30-päevase laenuperioodiga laenusaajate seas on häid kliente 77,3% (678 vaatlust), 31-60-päevase laenuperioodi korral 58,2% (170 vaatlust), 61-120-päevase laenuperioodi korral 50,2% (117 vaatlust), 121-180-päevase laenuperioodi korral 63,1% (128 vaatlust) ja rohkem kui 180-päevase laenuperioodi korral 79,0% (154 vaatlust).

Tunnuste *periood2* ja *staatus* sõltuvust testiva hii-ruut teststatistiku väärtus on 95,45 ning p-väärtus $< 0{,}0001$. Seega on tunnused sõltuvad.

Tabel 5. Tunnuse *periood2* sagedustabel

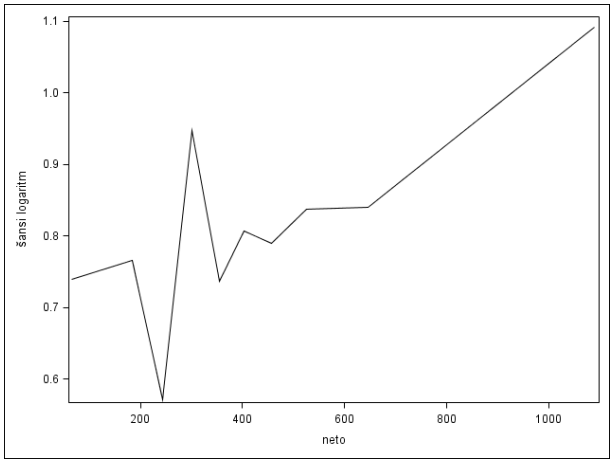
staatus/*periood2*	≤ 30	31-60	61-120	121-180	> 180	**Kokku**
1	678	170	117	128	154	**1247**
0	199	122	116	75	41	**553**
Kokku	**877**	**292**	**233**	**203**	**195**	**1800**

Sissetulek ja väljaminek

Tunnused *sissetulek* ja *valjaminek* ei osutunud logistilistes regressioonimudelites olulisteks tunnusteks. Seetõttu katsetati tunnuste kombinatsiooni. Tunnustest *sissetulek* ja *valjaminek* moodustati uus tunnus *neto*, mis väljendab laenusaaja netorahavoogu. Tunnus neto on kujul

$$neto = \text{sissetulek} - \text{valjaminek}.$$

Laenusaajad jaotati tunnuse *neto* detsiilide alusel kümneks grupiks ja leiti iga grupi šansi logaritm. Jooniselt 3 näeme, et tunnuse *neto* ja šansside logaritmide vahel arvatavasti esineb ligikaudu lineaarne seos. Siiski ei osutunud ka tunnus *neto* mudelites oluliseks tunnuseks.



Joonis 3. Šansside logaritmid tunnuse *neto* detsiilrühmades

Pereseis

Tunnusel *pereseis* on viis väärtust, mis on välja toodud tabelis 6 koos heade klientide osakaaludega ning kõikide klientide arvudega.

Tabel 6. Heade klientide osakaalud tunnuse *pereseis* väärtuste lõikes

pereseis	Lahutatud	Abielus	Lesk	Vabaabielus	Vallaline
heade klientide osakaal	79,1%	77,4%	77,2%	65,1%	61,0%
klientide arv kokku	187	530	57	478	548

Tunnuse *pereseis* väärtused grupeeriti kolme rühma põhimõttel, et gruppides oleksid sarnaste heade klientide osakaaludega pereseisud. Grupeerimise tulemusena loodi tunnus *pereseis2*, kus gruppi PS1 kuuluvad lahutatud, abielus ja lesestunud laenusaajad, gruppi PS2 kuuluvad vabaabielus laenusaajad ja gruppi PS3 kuuluvad vallalised laenusaajad.

Tabelist 7 näeme, et pereseisu gruppi PS1 kuuluvate laenusaajate seas on häid kliente 77,8% (602 vaatlust). Grupi PS2 puhul on häid kliente 65,1% (311 vaatlust) ning grupi PS3 puhul 61,0% (334 vaatlust). Tunnuste *pereseis2* ja *staatus* sõltuvust testiti hii-ruut testiga. Hii-ruut teststatistiku väärtus on 48,13 ja p-väärtus $< 0{,}0001$. Seega on tunnused sõltuvad.

Tabel 7. Tunnuse *pereseis2* sagedustabel

staatus/*pereseis2*	PS1	PS2	PS3	**Kokku**
1	602	311	334	**1247**
0	172	167	214	**553**
Kokku	**774**	**478**	**548**	**1800**

Haridus

Tunnusel *haridus* on kuus väärtust, mis on välja toodud tabelis 8 koos heade klientide osakaaludega ning kõikide klientide arvudega.

Tabel 8. Heade klientide osakaalud tunnuse *haridus* väärtuste lõikes

haridus	Kõrgharidus	Kutseharidus	Keskharidus	Algharidus	Ei ole	Põhiharidus
heade klientide osakaal	84,1%	69,2%	68,1%	60,0%	59,0%	50,5%
klientide arv kokku	345	535	670	15	39	196

Tunnus *haridus* grupeeriti kolme rühma samuti selliselt, et rühmad oleksid heade klientide osakaalude lõikes homogeensed. Grupeerimise tulemusena loodi tunnus *haridus2*, kus gruppi HAR1 kuuluvad kõrgharidusega laenusaajad, gruppi HAR2 kuuluvad kesk- ja kutseharidusega laenusaajad ning gruppi HAR3 kuuluvad põhi- ja algharidusega ning ilma hariduseta laenusaajad.

Tabelist 9 näeme, et gruppi HAR1 kuuluvate laenusaajate seas on häid kliente 84,1% (290 vaatlust). Gruppi HAR2 kuuluvate laenusaajate seas on häid kliente 68,6% (826 vaatlust) ning

grupi HAR3 puhul 52,4% (131 vaatlust). Tunnuse *haridus2* ja *staatus* sõltuvust testiva hii-ruut statistiku väärtus on 69,17 ning p-väärtus $< 0{,}0001$. Seega on tunnused sõltuvad.

Tabel 9. Tunnuse *haridus2* sagedustabel

staatus/*haridus2*	HAR1	HAR2	HAR3	**Kokku**
1	290	826	131	**1247**
0	55	379	119	**553**
Kokku	**345**	**1205**	**250**	**1800**

Töökogemus

Tunnusel *tookogemus* on neli väärtust, mis on välja toodud tabelis 10 koos heade klientide osakaaludega ning kõikide klientide arvudega.

Tabel 10. Heade klientide osakaalud tunnuse *tookogemus* väärtuste lõikes

tookogemus	Rohkem kui aasta	Töötu	Katseaeg	Kuni aasta
heade klientide osakaal	73,5%	62,3%	62,0%	54,4%
klientide arv kokku	1356	61	50	333

Tunnuse *tookogemus* grupeerimine tehti samal põhimõttel nagu eelnevad grupeerimised. Grupeerimise tulemusena loodi tunnus *tookogemus2*, kus gruppi TK1 kuuluvad laenusaajad, kellel on töökogemust rohkem kui aasta, gruppi TK2 kuuluvad laenusaajad, kes on töötud või on katseajal ning gruppi TK3 kuuluvad laenusaajad, kellel on töökogemust kuni aasta.

Tabelist 11 näeme, et gruppi TK1 kuuluvate laenusaajate seas on häid kliente 73,5% (997 vaatlust). Gruppi TK2 kuuluvate laenusaajate seas on häid kliente 62,2% (69 vaatlust) ning gruppi TK3 kuuluvate laenusaajate seas on 54,4% (181 vaatlust). Tunnuste *tookogemus2* ja *staatus* sõltuvust testiva hii-ruut teststatistiku väärtus on 48,98 ning p-väärtus $< 0{,}0001$. Seega on tunnused sõltuvad.

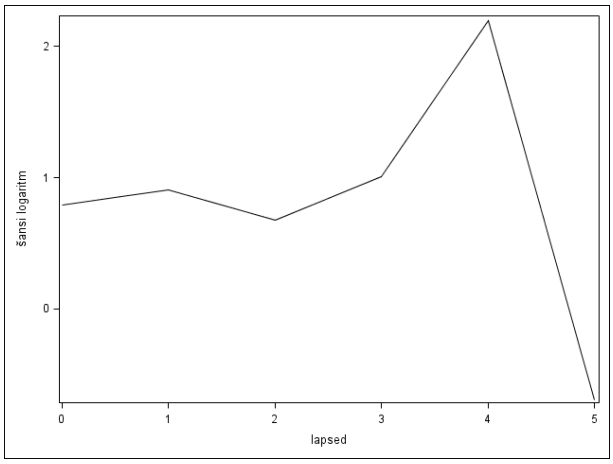
Tabel 11. Tunnuse *tookogemus2* sagedustabel

staatus/*tookogemus2*	TK1	TK2	TK3	**Kokku**
1	997	69	181	**1247**
0	359	42	152	**553**
Kokku	**1356**	**111**	**333**	**1800**

Lapsed

Laenusaajate minimaalne laste arv on 0 ning maksimaalne 5. Laste arvu keskväärtus on 0,57.

Tunnust *lapsed* ei grupeeritud ega teisendatud. Antud tunnuse puhul vaadeldi šansside logaritme tunnuse *lapsed* võimalike väärtuste lõikes. Jooniselt 4 näeme, et tunnuse *lapsed* ja šansside logaritmide vahel arvatavasti ei esine lineaarset seost. Vaatlusi, kellel on laste arvuks viis, on vaid kolm. Seega tasub nende vaatluste grupi šansi logaritmi väärtust ignoreerida.



Joonis 4. Šansside logaritmid tunnuse *lapsed* väärtuste lõikes

Kinnisvara

Laenusaajate omanduses olevate kinnisvaraobjektide arvu väljendava tunnuse *kinnisvara* minimaalne väärtus on 0 ning maksimaalne 8. Keskväärtus on 0,61.

Tunnus *kinnisvara* teisendati binaarseks tunnuseks *kinnisvara2*, kus kinnisvara omavale laenusaajale (*kinnisvara* $>$ 0) omistati väärtus „On“ ja kinnisvara mitte omavale laenusaajale (*kinnisvara* = 0) omistati väärtus „Ei ole“.

Tabelist 12 näeme, et kinnisvara omavate laenusaajate seas on häid kliente 81,7% (653 vaatlust) ja kinnisvara mitte omavate laenusaajate seas 59,3% (594 vaatlust). Tunnuste *kinnisvara2* ja *staatus* sõltuvust testiva hii-ruut teststatistiku väärtus on 104,62 ning p-väärtus $<$ 0,0001. Seega on tunnused sõltuvad ning võime öelda, et kinnisvara omavad kliendid on parema maksekäitumisega kui kinnisvara mitte omavad kliendid.

Tabel 12. Tunnuse *kinnisvara2* sagedustabel

staatus/*kinnisvara2*	on	ei ole	**Kokku**
1	653	594	**1247**
0	146	407	**553**
Kokku	**799**	**1001**	**1800**

Aktiivsed ja lõpetatud maksehäired

Lõpetatud maksehäirete tunnust *mh_lop* ei kaasatud analüüsi, sest antud tunnus on ära kirjeldatud aktiivsete maksehäirete ja kõikide maksehäirete tunnuste kaudu.

Aktiivsete maksehäirete tunnuse *mh_akt* minimaalne väärtus on 0 ning maksimaalne 12. Keskväärtus on 0,32.

Tunnus *mh_akt* teisendati binaarseks tunnuseks *mh_akt2*, kus aktiivseid maksehäireid omavale laenusaajale (*mh_akt* > 0) omistati väärtus „On“ ja aktiivseid maksehäireid mitte omavale laenusaajale (*mh_akt* = 0) omistati väärtus „Ei ole“.

Tabelist 13 näeme, et aktiivseid maksehäireid omavate laenusaajate seas on häid kliente 46,8% (123 vaatlust) ja aktiivseid maksehäireid mitte omavate laenusaajate seas 73,1% (1124 vaatlust). Tunnuste *mh_akt2* ja *staatus* sõltuvust testiva hii-ruut teststatistiku väärtus on 73,32 ning p-väärtus < 0,0001. Seega on tunnused sõltuvad ning võime öelda, et aktiivseid maksehäireid mitte omavad kliendid on parema maksekäitumisega kui aktiivseid maksehäireid omavad kliendid.

Tabel 13. Tunnuse *mh_akt2* sagedustabel

staatus/*mh_akt2*	on	ei ole	**Kokku**
1	123	1124	**1247**
0	140	413	**553**
Kokku	**263**	**1537**	**1800**

Maksehäired kokku

Kõikide maksehäirete tunnuse *mh_koik* minimaalne väärtus on 0 ning maksimaalne 27. Keskväärtus on 1,39.

Tunnus *mh_koik* teisendati binaarseks tunnuseks *mh_koik2*, kus maksehäireid omavale laenusaajale (*mh_koik* > 0) omistati väärtus „On“ ja maksehäireid mitte omavale laenusaajale (*mh_koik* = 0) omistati väärtus „Ei ole“.

Tabelist 14 näeme, et maksehäireid omavate laenusaajate seas on häid kliente 58,6% (475 vaatlust) ning maksehäireid mitte omavate laenusaajate seas 78,1% (772 vaatlust). Tunnuste *mh_koik2* ja *staatus* sõltuvust testiva hii-ruut teststatistiku väärtus on 79,52 ning p-väärtus <

0,0001. Seega on tunnused sõltuvad ning võime öelda, et maksehäireid mitte omavad kliendid on parema maksekäitumisega kui maksehäireid omavad kliendid.

Tabel 14. Tunnuse *mh_koik2* sagedustabel

staatus/*mh_koik2*	on	ei ole	**Kokku**
1	475	772	**1247**
0	336	217	**553**
Kokku	**811**	**989**	**1800**

3.3 Laenusaajate logistilise regressioonimudeli loomine

Logistilise regressioonimudeli loomisel kasutati statistikaprogrammi *SAS* protseduuri *logistic* ning mudelisse kaasatud tunnuste valikul kasutati automatiseeritud protseduuri *STEPWISE*.

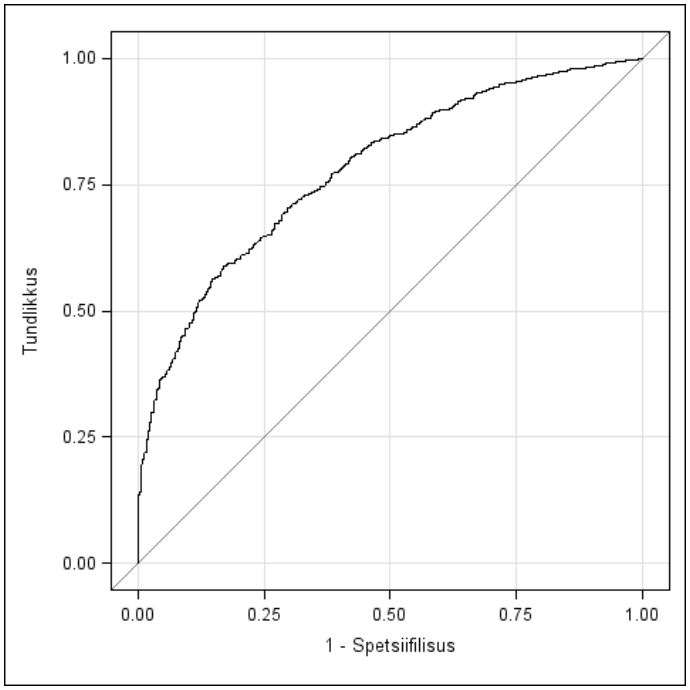
Laenusaajate mudelisse kaasatud tunnused koos Waldi hii-ruut statistiku W^2 väärtuste ja neile vastava *p*-väärtustega on välja toodud tabelis 15. Näeme, et statistiliselt oluliseks osutusid ka viis koosmõju, kuid kõik koosmõjudes esindatud tunnused (nt *sugu*) ei ole statistiliselt olulised. Hosmer ja Lemeshow [7, lk 73] on öelnud, et koosmõjude korral peavad mudelis olema esindatud ka mõlemad koosmõjus olevad tunnused isegi siis, kui need ei ole statistiliselt olulised. Seetõttu on mudelisse kaasatud statistiliselt mitteolulised tunnused *sugu*, *haridus2*, *tookogemus2* ja *kinnisvara2*.

Tabel 15. Laenusaajate mudeli tunnused

Tunnus	**Vabadusastmete arv**	W^2	***p* - väärtus**
vanus2	1	21,05	< 0,0001
sugu	1	0,07	0,7927
maakond2	2	8,58	0,0137
haridus2	2	3,19	0,2034
tookogemus2	2	3,25	0,1966
periood2	4	57,56	< 0,0001
kinnisvara2	1	1,87	0,1710
mh_koik2	1	25,60	< 0,0001
mh_akt2	1	19,24	< 0,0001
*vanus2*haridus2*	2	6,51	0,0386
*vanus2*kinnisvara2*	1	6,73	0,0095
*sugu*tookogemus2*	2	6,77	0,0340
*kinnisvara2*mh_koik2*	1	10,42	0,0012
*kinnisvara2*mh_akt2*	1	6,60	0,0102

Mudeli tõepärasuhte statistiku väärtus on 408,59 ning olulisuse tõenäosus $< 0{,}0001$. Seega on mudel statistiliselt oluline.

Joonisel 5 on kuvatud mudeli *ROC* kõver. *ROC* kõvera aluse pindala *AUC* väärtus on 0,7813. Seega on tegemist aktsepteeritava mudeliga. Hosmer – Lemeshow teststatistiku väärtus on 8,38 ning olulisuse tõenäosus 0,3974. Seega ei saa ümber lükata nullhüpoteesi, et mudel ei sobi andmestikuga.



Joonis 5. Laenusaajate mudeli *ROC* kõver

Laenusaajate mudel on kujul

$$logit(\pi) = 1{,}83 - 93{,}52vanus2 - 0{,}06(sugu = \text{F}) + 0{,}40(maakond2 = \text{MK1}) + 0{,}10(maakond2 = \text{MK2}) - 0{,}66(haridus2 = \text{HAR1}) - 1{,}00(haridus2 = \text{HAR2}) + 0{,}13(tookogemus2 = \text{TK1}) + 0{,}62(tookogemus2 = \text{TK2}) + 0{,}02(periood2 = \text{„}\leq 30\text{“}) - 0{,}80(periood2 = \text{„31-60“}) - 1{,}07(periood2 = \text{„61-120“}) - 0{,}60(periood2 = \text{„121-180“}) - 0{,}69(kinnisvara2 = \text{Ei ole}) + 1{,}22(mh\_koik2 = \text{Ei ole}) + 1{,}21(mh\_akt2 = \text{Ei ole}) + 50{,}79vanus2*(haridus2 = \text{HAR1}) + 40{,}20vanus2*(haridus2 = \text{HAR2}) + 41{,}79vanus2*(kinnisvara2 = \text{Ei ole}) + 0{,}71(sugu = \text{F}, tookogemus2 = \text{TK1}) + 0{,}21(sugu = \text{F}, tookogemus2 = \text{TK2}) - 0{,}93(kinnisvara2 = \text{Ei ole}, mh\_koik2 = \text{Ei ole}) - 0{,}88(kinnisvara2 = \text{Ei ole}, mh\_akt2 = \text{Ei ole}).$$

3.4 Laenusaajate mudeli analüüs

Tabelis 16 on välja toodud laenusaajate mudelisse tunnuste lisamise järjekord protseduuri *STEPWISE* käigus. Igal sammul on mudelisse lisatud kõige väiksema olulisuse tõenäosusega tunnus.

Tabel 16. Laenusaajate mudelisse tunnuste lisamise järjekord

Samm	Tunnus
1.	*kinnisvara2*
2.	*mh_akt2*
3.	*periood2*
4.	*vanus2*
5.	*kinnisvara2*mh_akt2*
6.	*mh_koik2*
7.	*haridus2*
8.	*sugu*
9.	*kinnisvara2*mh_koik2*
10.	*tookogemus2*
11.	*vanus2*kinnisvara2*
12.	*maakond2*
13.	*sugu*tookogemus2*
14.	*vanus2*haridus2*

Järgnevad šansside võrdlused on tehtud eeldusel, et teiste tunnuste, peale käsitletava tunnuse, väärtused ei muutu.

Tunnuse *vanus2* ees oleva parameetri märk on negatiivne, mis näitab vastupidist seost tunnuse *vanus2* ja hea staatuse tõenäosuste vahel. Antud parameetri väärtus (-93,52) kehtib isikute kohta, kellel on haridustase HAR3 ning kes omavad kinnisvara. Kui parameetrit korrigeerida selliste isikute puhul, kellel on teine haridustase või kellel ei ole kinnisvara, siis jääb parameeter siiski negatiivseks. Seega arvestades, et *vanus2* on vanuse pöördväärtus, võime öelda, et mida suurem on kliendi vanus, seda suurem on hea staatuse šanss.

Tunnuse (*sugu* = F) ees oleva parameetri märk on negatiivne. Seega on naistel töökogemusega TK3 väiksem hea staatuse šanss kui meestel, kes kuuluvad töökogemuse gruppi TK3. Teiste töökogemuste gruppide puhul parameetrit korrigeerides on naistel suurem hea staatuse šanss kui meestel.

Maakonna gruppi MK1 kuuluvatel klientidel on suurem hea staatuse šanss kui maakondade gruppi MK2 või MK3 kuuluvatel klientidel. Gruppi MK3 kuuluvatel klientidel on võrreldes teiste gruppidega kõige madalam hea staatuse šanss.

Tunnuste (*haridus2* = HAR1) ja (*haridus2* = HAR2) ees olevaid parameetreid korrigeeritakse vastavalt kliendi vanusele. Seega sõltub kliendi haridustaseme mõju hea staatuse šanssidele kliendi vanusest.

Meessoost klientide puhul on töökogemuse gruppi TK2 kuuluvatel klientidel suurem hea staatuse šanss kui teiste töökogemuse gruppide puhul. Naissoost klientidel on kõige suurem hea staatuse šanss töökogemuse gruppi TK1 kuuluvatel klientidel.

Kõige suurem hea staatuse šanss on klientidel, kellel laenuperiood on väiksem kui 30 päeva. Kõige väiksem hea staatuse šanss on 61-120-päevase laenuperioodiga klientide puhul.

Tunnuse (*kinnisvara2* = Ei ole) ees olev parameeter on negatiivne. Korrigeerides parameetrit vastavalt tunnuste *mh_koik2* ja *mh_akt2* väärtustele, jääb parameetri märk siiski negatiivseks. Seega võime öelda, et kinnisvara omavatel klientidel on suurem hea staatuse šanss kui kinnisvara mitte omavatel klientidel.

Tunnuse (*mh_koik2* = Ei ole) ja (*mh_akt2* = Ei ole) ees olevad parameetrid on positiivse märgiga. Korrigeerides parameetrite väärtust klientide puhul, kes ei oma kinnisvara, jäävad parameetrid siiski positiivseteks. Seega on maksehäireid mitte omava kliendi hea staatuse šanss suurem kui maksehäireid omava kliendi hea staatuse šanss ning sama kehtib ka aktiivsete maksehäirete korral.

Laenusaajate logistilise regressioonimudeli näol on tegemist kallutatud krediidiskooringu süsteemiga, mida saab rakendada vaid laenusaajatel ning mitte kõikidel laenutaotlejatel.

4 Laenu mittesaajate kaasamine

4.1 Kaasamismeetodid

Käesolevas bakalaureusetöös käsitletakse nelja kaasamismeetodit – randomiseeritud, kahestamise, lävendi ja nullide meetodit. Meetodite eesmärgiks on täita laenu mittesaajate staatuste tühikud kasutades laenusaajate logistilise regressioonimudeliga laenu mittesaajatele prognoositud hea staatuse tõenäosusi $\hat{\pi}_i$. Järgmisena tuuakse välja meetodite kirjeldused.

Randomiseeritud meetod

Igale vaatlusele genereeritakse juhuslik suurus X_i ühtlasest jaotusest $X_i \sim U(0,1)$. Juhuslikke suurusi X_i võrreldakse prognoositud hea staatuse tõenäosustega $\hat{\pi}_i$. Lühidalt on eeskiri järgmine:

$$\hat{Y}_i = \begin{cases} 1, & kui\ X_i \leq \hat{\pi}_i \\ 0, & vastasel\ juhul \end{cases},$$

kus $\hat{Y}_i$ on i-nda vaatluse staatuse prognoos. Pärast staatuse prognooside leidmist liidetakse liitandmestiku saamiseks laenusaajate ja laenu mittesaajate andmestikud. [8]

Kahestamise meetod

Laenu mittesaajad kaasatakse liitandmestikku kahekordselt – osaliselt hea kliendina ja osaliselt halva kliendina. Hea staatusega (*staatus* = 1) kliendi kaaluks määratakse hea staatuse tõenäosuse prognoos $\hat{\pi}_i$ ja halva staatusega (*staatus* = 0) kliendi kaaluks määratakse $1 - \hat{\pi}_i$. Liitandmestiku pealt logistilise regressioonimudeli loomisel kasutatakse vaatluste kaalusid, kusjuures kõigile laenusaajatele määratakse liitandmestikus kaal väärtusega 1. [8]

Lävendi meetod

Laenu mittesaajate klassifitseerimisel headeks või halbadeks klientideks määratakse lävend π_c. Lävendit π_c võrreldakse hea staatuse tõenäosuste prognoosidega $\hat{\pi}_i$. Lühidalt on eeskiri järgmine:

$$\hat{Y}_i = \begin{cases} 1, & kui\ \hat{\pi}_i \geq \pi_c \\ 0, & vastasel\ juhul \end{cases}.$$

Lävendi valikul lähtutakse sellest, et heade klientide osakaal oleks laenu mittesaajate andmestikus võrdne heade klientide osakaaluga laenusaajate andmestikus. Liitandmestiku saamiseks liidetakse laenusaajate ja laenu mittesaajate andmestikud.

Nullide meetod

Kõikidele laenu mittesaajatele omistatakse staatus 0 ehk halb. Liitandmestiku saamiseks liidetakse laenusaajate ja laenu mittesaajate andmestikud.

4.2 Ülevaade praktikas kasutatavatest kaasamismeetoditest

Randomiseeritud ja kahestamise meetod on praktikas laialt kasutusel olevad meetodid. Lävendimeetod ja nullide meetod on autori poolt välja pakutud meetodid, mille kohta võib eeldada, et meetodid prognoosivad vaatluste staatuste väärtusi ebatäpsemalt kui randomiseeritud ja kahestamise meetodid. Lävendi meetodi puhul esineb arvatav probleem eelduses, et laenusaajate seas on heade klientide osakaal võrdne heade klientide osakaaluga laenu mittesaajate seas. Nullide meetodi puuduseks on see, et kõik mittesaajad klassifitseeritakse halbadeks klientideks, kuigi nende seas võib olla ka häid kliente.

Laenu mittesaajate kaasamiseks kasutatakse praktikas ka meetodit, mille esimeseks sammuks on laenusaajate ja laenu mittesaajate andmestike pealt välja töötada mudel, mis prognoosib igale vaatlusele laenusaamise tõenäosuse. Igale laenusaajale omistatakse kaal, mis on laenusaamise tõenäosuse hinnangu pöördväärtus. Seega on väikse laenuaamise tõenäosusega vaatlustele määratud suuremad kaalud kui suure laenusaamise tõenäosusega vaatlustele ning laenusaajate valim kirjeldab täpsemalt laenutaotlejate populatsiooni. Seejärel luuakse kaalusid kasutades laenusaajate andmestikult mudel, mis hindab hea staatuse tõenäosust. [9]

Kaasamismeetodina kasutatakse ka iteratiivset meetodit, mille puhul esialgu töötatakse välja laenusaajate mudel. Mudelit rakendatakse laenu mittesaajatel ning mittesaajad klassifitseeritakse mingi lävendi pealt headeks ja halbadeks klientideks. Laenusaajate ja laenu mittesaajate andmestikud liidetakse ning luuakse uus mudel. Saadud mudelit rakendatakse jälle laenu mittesaajatel ning toimub uus vaatluste klassifitseerimine lävendi järgi. Seejärel liidetakse jälle andmestikud ning luuakse uus mudel. Protsessi korratakse, kuni uuesti klassifitseerides laenu mittesaajatele omistatud staatused ei muutu. [9]

Eelnevalt kirjeldatud meetodite ning ka randomiseeritud, kahestamise ja lävendi meetodite võimalik probleem seisneb eelduses, et laenusaajate abil saab prognoosida laenu mittesaajate staatusi. Laenusaajate andmestikus ei pruugi alati olla piisavalt halva staatusega kliente, et korrektselt prognoosida staatust laenu mittesaajatele, kelle hulgas on halva staatusega kliente arvatavasti rohkem. [9]

On ilmne, et laenu mittesaajate kaasamisega välja töötatud mudeli ja ainult laenusaajate mudeli vahelise seose uurimisel on võimalik parim tulemus saavutada valimiga, kus on nii laenusaajate kui ka laenu mittesaajate staatused teada. Leitud seost saab rakendada tulevastel laenusaajate mudelitel, et mudeleid korrigeerida. Hand ja Henley [3] on välja toonud kolm meetodit, mis kõik kasutavad ühte või mitut taolist valimit. Praktikas on aga selliste valimite kättesaadavus kesine.

Kaasamismeetodite rakendamisel on oluline, et esialgses mudelis, mille alusel toimub laenutaotlejate jagamine laenusaajateks ja laenu mittesaajateks, ning laenusaajate mudelis oleksid kaasatud samad tunnused. Olgu X tunnuste hulk, mis on teada laenutaotlejate kohta. Oletame, et mingile osale laenutaotlejatest, kelle tunnustekomplekti vektor on x, antakse laenu. Siis peab heade klientide osakaal tunnustekomplektiga x laenusaajate seas olema võrdne heade klientide osakaaluga laenu mittesaajate seas, kelle tunnustekomplekti väärtus on x. Kui aga laenuandmise otsuse juures kasutatakse lisainformatsiooni, siis ei ole tavaliselt ühegi tunnustekomplekti X väärtuse puhul heade klientide osakaalud laenusaajate ja laenu mittesaajate seas võrdsed. [3]

Kui esialgne mudel, mida kasutatakse laenuandmise otsustusprotsessis, kasutab tunnuste hulka X ning laenusaajate pealt töötatakse välja mudel, mis kasutab tunnuste hulka Y nii, et Y on hulga X alamhulk, siis laenusaajate mudeli kasutamisel kaasamismeetodis saadakse kallutatud tulemus. Seega tuleks kaasamismeetodi kallutatuse vältimiseks kasutada laenusaajate mudelis samu tunnuseid, mis on kasutusel esialgses mudelis, mille alusel toimub taotlejate jagunemine laenusaajateks ja laenu mittesaajateks. [3]

4.3 Kaasamismeetodite võrdluseksperimendi kirjeldus

Kliendi reaalne staatus Y_i on teada vaid laenusaajate andmestikus olevate klientide kohta. Selleks, et võrrelda, millise kaasamismeetodi rakendamise tulemusena välja töötatud krediidiskooringu süsteem prognoosib kliendi staatust kõige täpsemalt, tuleks võrrelda krediidiskooringu süsteemi poolt prognoositud hea kliendi tõenäosust kliendi reaalse

staatusega. Sellest tulenevalt tekkis idee jaotada laenusaajate andmestik kaheks – pseudo-laenusaajateks ja pseudo-laenumittesaajateks. Pseudo-laenumittesaajate reaalsed staatused kustutatakse ning kaasamismeetodeid rakendades töötatakse välja krediidiskooringu süsteem. Krediidiskooringu süsteemiga prognoositud hea kliendi tõenäosusi saab võrrelda klientide reaalsete staatustega ning leida, millise kaasamismeetodi rakendamisega prognoositakse kliendi staatust kõige täpsemalt. Järgnevalt on toodud eksperimendi täpsem kirjeldus.

1) Laenusaajate andmestikul rakendatakse välja töötatud laenusaajate logistilist regressioonimudelit ning vaatlused klassifitseeritakse mingi lävendi pealt pseudo-laenusaajateks ja pseudo-laenumittesaajateks. Pseudo-laenumittesaajate tunnuse *staatus* väärtused kustutatakse.
2) Pseudo-laenusaajate andmestikul töötatakse välja pseudo-laenusaajate logistiline regressioonimudel, kusjuures pseudo-laenusaajate mudelisse kaasatakse samad tunnused, mis on laenusaajate mudelis, et vältida kallutatust kaasamismeetodite rakendamisel.
3) Pseudo-laenusaajate mudeliga prognoositakse pseudo-laenumittesaajatele hea kliendi tõenäosused ning rakendatakse kaasamismeetodeid. Iga kaasamismeetodi rakendamise tulemusena luuakse liitandmestik, kus on nii pseudo-laenusaajate kui ka pseudo-laenumittesaajate andmed.
4) Iga liitandmestiku põhjal luuakse logistiline regressioonimudel, kusjuures kõikidesse mudelitesse kaasatud tunnustekomplektid on samad.
5) Iga liitandmestiku puhul prognoositakse liitandmestiku pealt välja töötatud mudeliga igale vaatlusele hea kliendi tõenäosus. Prognooside täpsust kirjeldatakse keskmise absoluutveaga *MAE* ning keskmise ruutveaga *MSE*, mis on kujul

$$MAE = \frac{\sum_{i=1}^{n} |\hat{\pi}_i - Y_i|}{n},$$

$$MSE = \frac{\sum_{i=1}^{n} (\hat{\pi}_i - Y_i)^2}{n},$$

kus n on vaatluste arv, $\hat{\pi}_i$ on i-nda vaatluse hea staatuse tõenäosuse prognoos ja Y_i on i-nda vaatluse reaalne staatus (1 – hea, 0 – halb).

6) Liitandmestike pealt arvutatud statistikute *MAE* ja *MSE* põhjal tehakse otsus, millist kaasamismeetodit rakendades prognoositakse kliendi staatus kõige täpsemalt.

4.4 Võrdluseksperimendi tulemused

Laenusaajate andmestik jaotati kaheks, kasutades peatükis 3.3 välja toodud logistilist regregressioonimudelit. Lävendi väärtuseks valiti 0,7, kuna sellisel juhul on pseudo-laenusaajaid ja pseudo-laenumittesaajaid peaaegu võrdselt (vastavalt 957 ja 843 vaatlust). Pseudo-laenusaajate andmestikul loodi logistiline regressioonimudel, milles kasutati samu tunnuseid kui laenusaajate mudelis. Kaasamismeetodite rakendamise tulemusena loodi neli liitandmestikku. Liitandmestikel loodi logistilised regressioonimudelid, kuhu kõikide mudelite puhul kaasati tunnused *sugu*, *haridus2*, *periood2*, *kinnisvara2*, *mh_akt2* ja *mh_koik2*.

Tabelis 17 on välja toodud rakendatud meetodite täpsust hindavate statistikute (keskmine absoluutviga *MAE* ja keskmine ruutviga *MSE*) väärtused. Näeme, et kõige väiksem keskmine absoluutviga esineb lävendi meetodi rakendamisel. Samas on kõige väiksema keskmise ruutveaga randomiseeritud meetod, millega praktiliselt võrdne on kahestamise meetod.

Tabel 17. Kaasamismeetodite täpsust hindavate statistikute väärtused lävendi 0,7 korral

meetod/statistik	***MAE***	***MSE***
Randomiseeritud	0,352	0,194
Kahestamise	0,353	0,195
Lävendi	0,333	0,222
Nullide	0,375	0,274

Eelnev tulemus saadi, kui laenusaajate andmestik jagati kaheks lävendi 0,7 pealt. Selleks, et veenduda parima kaasamismeetodi valikus, viidi eksperiment läbi ka lävendite 0,5, 0,6, ja 0,8 pealt. Tabelis 18 on välja toodud statistikute väärtused meetodite ja lävendite lõikes. Näeme, et lävendite 0,6, 0,7 ja 0,8 korral on keskmine absoluutviga kõige väiksem lävendi meetodit rakendades. Lävendi 0,5 korral on väikseim keskmine absoluutviga nullide meetodi puhul. Paneme tähele, et esialgse lävendi suurenedes lävendi meetodi keskmine absoluutviga väheneb. Keskmine ruutviga on kõige väiksem randomiseeritud ja kahestamise meetoditel kõikide lävendite lõikes, kusjuures lävendi suurenedes kasvavad vahed lävendi meetodi ja nullide meetodi keskmiste ruutvigadega.

Tabel 18. Kaasamismeetodite täpsust hindavate statistikute väärtused lävendite 0,5, 0,6, 0,7 ja 0,8 korral

	MAE				***MSE***			
meetod/lävend	**0,5**	**0,6**	**0,7**	**0,8**	**0,5**	**0,6**	**0,7**	**0,8**
Randomiseeritud	0,353	0,352	0,352	0,374	0,180	0,182	0,194	0,200
Kahestamise	0,353	0,351	0,353	0,374	0,180	0,182	0,195	0,200
Lävendi	0,358	0,346	0,333	0,327	0,196	0,208	0,222	0,258
Nullide	0,347	0,352	0,375	0,415	0,196	0,221	0,274	0,339

Autori seisukohast on krediidiskooringu süsteemi puhul oluline vältida hea staatuse prognooside puhul suuri vigu. Keskmine ruutviga *MSE* on tundlik suurte vigade suhtes, mistõttu lähtume parima meetodi valikul keskmisest ruutveast. Seega on parimateks kaasamismeetoditeks randomiseeritud ja kahestamise meetodid, mille keskmised ruutvead on samaväärsed ning kõikide lävendite lõikes kõige väiksemad. Rakendatud meetoditest oodatult halvim on nullide meetod.

5 Parima kaasamismeetodi rakendamine

5.1 Logistiline regressioonimudel liitandmestikule

Laenu mittesaajate kaasamiseks kasutati randomiseeritud meetodit, kuna selle rakendamine on tehniliselt põhjustel lihtsam kui kahestamise meetodi rakendamine. Laenu mittesaajate andmestikus on 1599 vaatlust. Randomiseeritud meetodi rakendamise tulemusena klassifitseeriti 994 laenu mittesaajat heaks kliendiks ja 605 halvaks kliendiks. Laenu mittesaajate ja laenusaajate andmestikud liideti. Saadud liitandmestikus on 3399 vaatlust, millest 2241 on hea staatusega ning 1158 halva staatusega.

Liitandmestikult loodi logistiline regressioonimudel, kus tunnuste valikul kasutati automatiseeritud protseduuri *STEPWISE*. Tabelis 19 on välja toodud mudelisse kaasatud tunnused koos Waldi hii-ruut statistiku W^2 väärtuste ja neile vastava *p*-väärtustega.

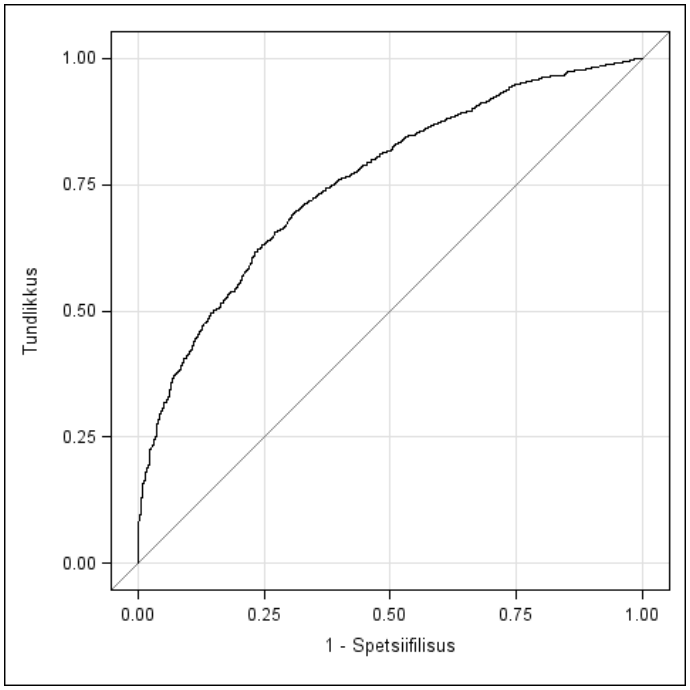
Tabel 19. Liitandmestiku mudeli tunnused

Tunnus	Vabadusastmete arv	W^2	*p* - väärtus
vanus2	1	24,61	< 0,0001
sugu	1	5,56	0,0183
maakond2	2	16,30	0,0003
haridus2	2	4,14	0,1261
tookogemus2	2	7,80	0,0203
periood2	4	79,62	< 0,0001
kinnisvara2	1	2,25	0,1332
mh_koik2	1	24,82	< 0,0001
mh_akt2	1	31,57	< 0,0001
*vanus2*haridus2*	2	13,44	0,0012
*vanus2*kinnisvara2*	1	7,98	0,0047
*vanus2*mh_koik2*	1	4,95	0,0260
*sugu*tookogemus2*	2	18,71	< 0,0001
*sugu*haridus2*	2	6,08	0,0478
*kinnisvara2*mh_koik2*	1	6,97	0,0083
*kinnisvara2*mh_akt2*	1	13,85	0,0002

Mudeli tõepärasuhte statistiku väärtus on 672,98 ning olulisuse tõenäosus $< 0{,}0001$. Seega on mudel statistiliselt oluline.

Joonisel 6 on kuvatud mudeli *ROC* kõver. *ROC* kõvera aluse pindala *AUC* väärtus on 0,7573. Seega on tegemist aktsepteeritava mudeliga. Hosmer – Lemeshow teststatistiku väärtus on

9,24 ning olulisuse tõenäosus 0,3222. Seega ei saa ümber lükata nullhüpoteesi, et mudel ei sobi andmestikuga.



Joonis 6. Liitandmestiku mudeli *ROC* kõver

Liitandmestiku mudel on kujul

$$logit(\pi) = 1{,}27 - 67{,}71vanus2 - 0{,}54(sugu = \text{F}) + 0{,}41(maakond2 = \text{MK1}) + 0{,}20(maakond2 = \text{MK2}) - 0{,}58(haridus2 = \text{HAR1}) - 0{,}73(haridus2 = \text{HAR2}) - 0{,}06(tookogemus2 = \text{TK1}) + 0{,}42(tookogemus2 = \text{TK2}) + 0{,}07(periood2 = \text{„}\leq 30\text{“}) - 0{,}68(periood2 = \text{„31-60“}) - 0{,}83(periood2 = \text{„61-120“}) - 0{,}52(periood2 = \text{„121-180“}) - 0{,}55(kinnisvara2 = \text{Ei ole}) + 1{,}46(mh\_koik2 = \text{Ei ole}) + 1{,}23(mh\_akt2 = \text{Ei ole}) + 44{,}55vanus2*(haridus2 = \text{HAR1}) + 32{,}51vanus2*(haridus2 = \text{HAR2}) + 32{,}19vanus2*(kinnisvara2 = \text{Ei ole}) - 18{,}07vanus2*(mh\_koik2 = \text{Ei ole}) + 0{,}83(sugu = \text{F}, tookogemus2 = \text{TK1}) + 0{,}29(sugu = \text{F}, tookogemus2 = \text{TK2}) + 0{,}79(sugu = \text{F}, haridus2 = \text{HAR1}) + 0{,}30(sugu = \text{F}, haridus2 = \text{HAR2}) - 0{,}60(kinnisvara2 = \text{Ei ole}, mh\_koik2 = \text{Ei ole}) - 0{,}94(kinnisvara2 = \text{Ei ole}, mh\_akt2 = \text{Ei ole}).$$

5.2 Liitandmestiku mudeli ja laenusaajate mudeli võrdlus

Liitandmestiku ja laenusaajate mudelid on oma kujult sarnased. Laenu mittesaajate kaasamise tulemusena osutusid liitandmestiku mudelis võrreldes laenusaajate mudeliga oluliseks veel kaks koosmõju – *vanus2*mh_koik2* ja *sugu*haridus2*. Teised kaasatud tunnused on samad, mis laenusaajate mudelis.

Liitandmestiku mudeli *ROC* kõvera alune pindala (0,7573) on väiksem kui laenusaajate mudeli *ROC* kõvera alune pindala (0,7813). See tähendab, et liitandmestiku mudel ei ole nii hea kirjeldamisvõimega kui laenusaajate mudel. See on oodatav tulemus, kuna laenu mittesaajate kaasamise teel saadud liitandmestik on heterogeensem kui ainult laenusaajate andmestik.

Nii liitandmestiku mudeli kui ka laenusaajate mudeli puhul on tegemist kirjeldamisvõime ja sobitusastme poolest aktsepteeritavate mudelitega. Liitandmestiku mudeli välja töötamisel on aga kaasatud ka laenu mittesaajad, mistõttu on mudel loodetavasti vähem kallutatud ja seega saab mudelit kui krediidiskooringu süsteemi rakendada kõikide uute laenutaotlejate korral.

Kokkuvõte

Bakalaureusetöös selgitati krediidiskooringu süsteemi sisu ning laenu mittesaajate kaasamise olulisust nii statistilisest kui ka ärilisest seisukohast. Laenusaajate põhjal töötati välja aktsepteeritava kvaliteediga logistiline regressioonimudel, mille puhul on tegemist kallutatud krediidiskooringu süsteemiga ning mida ei ole õige otseselt rakendada uute laenutaotlejate korral.

Bakalaureusetöö eesmärgiks oli välja selgitada, millise laenu mittesaajate kaasamismeetodi rakendamise tulemusena välja töötatud krediidiskooringu süsteem prognoosib laenutaotleja staatust kõige täpsemalt, ning rakendada seda meetodit reaalsel andmestikul. Käesolevas töös võrreldi nelja kaasamismeetodit – randomiseeritud, kahestamise, lävendi ja nullide meetodit.

Parima kaasamismeetodi välja selgitamiseks teostati eksperiment, mille käigus jaotati laenusaajad pseudo-laenusaajateks ja pseudo-laenumittesaajateks. Pseudo-laenumittesaajate reaalsed staatused kustutati ning prognoositi staatuste hinnangud kasutades kaasamismeetodeid. Liitandmestikelt loodud logistiliste regressioonimudelite hea staatuse tõenäosuste prognooside põhjal selgitati välja, milline kaasamismeetod on kõige parem. Selgus, et parimad kaasamismeetodid on randomiseeritud ja kahestamise meetod, mille prognooside keskmised ruutvead olid kõige väiksemad. Oodatult halvim kaasamismeetod on nullide meetod.

Parimaks osutunud meetoditest rakendati laenu mittesaajate andmestikul randomiseeritud meetodit, prognoosides laenu mittesaajatele staatused. Seejärel liideti laenusaajate ja laenu mittesaajate andmestikud ning liitandmestikult töötati välja aktsepteeritava kvaliteediga ühine logistiline regressioonimudel. Loodud mudeli näol on tegemist krediidiskooringu süsteemiga, mis on loodetavasti vähem kallutatud ja mida saab seega rakendada kõigi uute laenutaotlejate korral.

Kasutatud kirjandus

[1] Hand, D.J., Henley, W.E. (1997). Statistical Classification Methods in Consumer Credit Scoring: a Review. *Journal of the Royal Statistical Society: Series A (Statistics in Society), Vol. 160, No. 3,* 523-541

[2] Siddiqi, N. (2005). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring.* New Jersey: John Wiley & Sons, Inc.

[3] Hand, D.J., Henley, W.E. (1993). Can reject inference ever work? *IMA Journal of Mathematics Applied in Business & Industry, 5,* 45-55

[4] *Swedbank AS Aastaaruanne 2014* (2015).
https://www.swedbank.ee/static/pdf/about/finance/reports/info_annual-report-2014_est.pdf
[06.04.2016]

[5] Kleinbaum, D.G., Kupper, L.L., Muller, K.E., Nizam, A. (1998). *Applied Regression Analysis and Other Multivariable Methods*. Duxbury Press.

[6] Käärik, E. (2014). *Andmeanalüüs II. Loengukonspekt.* Tartu: Tartu Ülikool, matemaatika ja statistika instituut.
Saadaval: http://dspace.ut.ee/bitstream/handle/10062/35401/AndmeanaluusII.pdf

[7] Hosmer, D.W., Lemeshow, S., Sturdivant, R.X. (2013). *Applied Logistic Regression*. New Jersey: John Wiley & Sons, Inc.

[8] Montrichard, D. (2008). *Reject Inference Methodologies in Credit Risk Modeling*. SESUG, Inc.
Saadaval: http://analytics.ncsu.edu/sesug/2008/ST-160.pdf

[9] Joanes, D.N. (1993). Reject inference applied to logistic regression for credit scoring. *IMA Journal of Mathematics Applied in Business & Industry, 5,* 35-43

Lisad

Lisa 1. Laenusaajate logistilise regressioonimudeli kood

```
ods graphics on;
proc logistic data = ok.laenusaajad descending plots=roc;
class sugu maakond2 keel pereseis2 haridus2 tookogemus2
kinnisvara2 mh_koik2 mh_akt2 periood2/param=ref ref=last;
model staatus = vanus2 sugu maakond2 keel pereseis2 haridus2
tookogemus2 periood2 summa1 summa2 lapsed neto kinnisvara2
mh_koik2 mh_akt2 vanus2*haridus2 vanus2*kinnisvara2
vanus2*mh_koik2 sugu*tookogemus2 kinnisvara2*mh_koik2
kinnisvara2*mh_akt2/ selection=stepwise lackfit;
run;
ods graphics off;
```

Lisa 2. Kaasamismeetodite koodid

Randomiseeritud meetod

```
data ok.koguandmed_rand;
set ok.laenusaajad (in=a) ok.tagasilukatud_rand (in=b);
if b then do;
     rand = ranuni(10);
     if rand <= P_1 then staatus_uus = 1;
     else staatus_uus = 0;
end;
else staatus_uus = staatus;
run;
```

Kahestamise meetod

```
data ok.tagasilukatud_kah;
set ok.tagasilukatud_kah;
staatus_uus = 1;
kaal = P_1;
output;
staatus_uus = 0;
kaal = P_0;
output;
run;

data ok.koguandmestik_kah;
set ok.laenusaajad (in=a) ok.tagasilukatud_kah (in=b);
if a then do;
     staatus_uus = staatus;
     kaal = 1;
end;
```

```
run;
```

Lävendi meetod

```
data ok.tagasilukatud_cutoff;
set ok.tagasilukatud_cutoff;
if P_1 < 0.109 then staatus_uus=0;
else staatus_uus=1;
run;

data ok.koguandmestik_cutoff;
set ok.laenusaajad (in=a) ok.tagasilukatud_cutoff (in=b);
if a then do;
     staatus_uus = staatus;
end;
run;
```

Nullide meetod

```
data ok.koguandmestik_nullid;
set ok.laen_skooridega;
if staatus07=0 then staatus_uus=0; /*staatus07 on pseudo-
laenusaajate ja pseudo-laenumittesaajate staatuse tunnus*/
else staatus_uus=staatus;
run;
```

Lisa 3. Liitandmestiku logistilise regressioonimudeli kood

```
ods graphics on;
proc logistic data = ok.koguandmestik descending plots=roc;
class sugu maakond2 keel pereseis2 haridus2 tookogemus2
kinnisvara2 mh_koik2 mh_akt2 periood2/param=ref ref=last;
model staatus_uus = vanus2 sugu maakond2 keel pereseis2
haridus2 tookogemus2 periood2 summa1 summa2 lapsed neto
kinnisvara2 mh_koik2 mh_akt2 sugu*tookogemus2 sugu*haridus2
vanus2*haridus2 vanus2*kinnisvara2 kinnisvara2*mh_koik2
kinnisvara2*mh_akt2 vanus2*mh_koik2/ selection=stepwise
lackfit;
run;
ods graphics off;
```

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Oskar Kärmas,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Krediidiskooringu süsteemi loomine laenu mittesaajate kaasamisega“, mille juhendaja on prof. Kalev Pärna,

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, **28.04.2016**