Tartu Ülikool

Loodus- ja täppisteaduste valdkond

Matemaatika ja statistika instituut

Martin Aasmäe

Krediidiriski hindamisel kasutatavate mudelite võrdlus ühe Eesti laenuandmestiku näitel

Matemaatilise statistika eriala

Bakalaureusetöö (9 EAP)

Juhendajad: dots. Meelis Käärik, prof. Kalev Pärna

Tartu 2019

Krediidiriski hindamisel kasutatavate mudelite võrdlus ühe Eesti laenuandmestiku näitel

Lühikokkuvõte.

Krediidiasutustel on väga oluline tunda oma klienti ning konkreetselt laenutoodete puhul on tähtis olla teadlik kliendi maksejõulisust. Töö eesmärgiks on uurida, kas maksejõulisuse hindamiseks eelnevalt tuntud logistilise regressiooni uurimismeetodile lisaks leidub ka teistsuguseid alternatiive. Selle tarbeks valitakse välja 3 konkureerivat meetodit, kus uuritav tunnus on binaarsel kujul – *probit*, *c-log-log*, *cauchit*. Kõigi eeltoodud meetodite abil konstrueeritakse mudelid hindamaks laenusaaja maksejõulisust. Töös antakse ka teoreetiline ülevaade kõigi nelja meetodi kohta. Töö praktilises osa alguses viiakse läbi analüüsid kõigi nelja uurimise all oleva mudeliga. Praktilise osa lõpus võrreldakse erinevate meetodite kasutamisel saadud tulemusi ning valitakse välja parim mudelivariant olemasolevatest. Parima mudeliga tehakse ka süvaanalüüs ning esitatakse mudeli interpretatsioon. Lõpuks esitatakse kokkuvõte ja järeldused tehtud tööst.

Märksõnad: krediidirisk, logistiline regressioon, üldistatud lineaarsed mudelid

CERCS teaduseriala: Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika (P160)

Comparison of credit risk assessment models based on one Estonian loan dataset

Abstract. It is very important for credit institutions to know their clients and, in the case of loan products in particular, it is important to be aware of the customer's solvency. The purpose of this study is to investigate whether there are other alternatives to the previously known logistic regression research method for assessing solvency. For this purpose, 3 competing methods are selected in which the investigated character is binary - probit, c-log-log, cauchit. Using all of the forementioned methods, models are constructed to assess the borrower's solvency. The thesis also provides a theoretical overview of all four methods. At the beginning of the practical part of the work, analyzes are carried out on all four models under investigation. At the end of the practical part, the results obtained using the different methods are compared and the best model of the available models is selected. For the best model, in-depth analysis and interpretation are also provided. Finally, a summary and conclusions are drawn of the work done.

Keywords: credit risk, logistic regression, generalized linear models

CERCS research specialisation: Statistics, operations research, programming, actuarial mathematics (P160)

Sisukord

Sissejuhatus

Krediidiasutuste (pankade) üheks põhitegevusalaks on laenude väljastamine. Selle tegevuse puhul on tarvis esmalt analüüsida laenutaotlejaid ning selgitada välja nende maksejõulisus. See tegevus kannab nime krediidiriski hindamine, kus lõpptulemusena on, lihtsustatult öeldes, tarvis klassifitseerida kliendid kaheks – headeks ehk maksevõimelisteks ning halbadeks ehk maksejõuetuteks klientideks. Selle jaoks on mitmeid võimalusi. Näiteks logisitilise regressiooni mudeli abil teostatud klassifitseerimisprotsess on maksejõulisuse tõenäosuse hindamise osas kujunenud finantsinstitutsioonide seas seni üheks levinumaks praktikaks.

Käesoleva töö laiemaks eesmärgiks ongi välja selgitada, kas krediidiriski hindamisel leidub lisaks logistilisele regressioonile ka alternatiivseid mudelivõimalusi. Välja valitud mudelitega teostatakse põhjalikud analüüsid ning omavahelised võrdlused. Samuti antakse lisaks praktilisele kasutatavusele ka teoreetiline ülevaade valikus olevatest mudelitest.

Töö kitsamaks eesmärgiks on tegeleda süvitsi andmeanalüüsi erinevate tahkudega. Näiteks pakub huvi see, kas ja kui palju aitab argumenttunnuste teisendamine ja grupeerimine kaasa mudeli hindamisvõime paranemisele. Samuti tegeletakse andmestikku puudutavate iseärasustega, sealhulgas vigade sümmetriaga.

Alternatiivsete variantidena logistilise regressioonile on valikus sellised üldistatud lineaarsed mudelid, mis põhinevad *probit*, täiend-log-log (*c-log-log*) ja *cauchit* tüüpi seosefunktsioonidel. Tegevuse mõte on hoida uuritavate mudelite arv väiksena, aga see-eest uurida neid rohkem süvitsi. Kuigi *probit* mudel on teoreetilise sisu mõttes logitile üpris sarnane, proovitakse selle abil otsida teatud erinevusi mudeli tulemustes. Nendele lisaks on töösse kaasatud ka kaks ebasümmeetrilist seosefunktsiooni (*c-log-log* ja *cauchit*), mille abil püütakse anda käesolevasse uuringusse suuremat võrdlusmomenti.

Töö on kirjutatud kahes osas. Esimeses osas antakse teoreetiline ülevaade võrdluse all olevatest mudelitest ning kõik muu sellega seonduv. Teises osas ehk töö praktilises osas viiakse läbi analüüsid kõigi eeltoodud mudelitega ning uuritakse, missugune on vaatluse all oleva mudeli prognoosivõime. Suur rõhk on mudeli sobivuse hindamisel. Teise osa lõpus võrreldakse erinevate meetodite kasutamisel saadud tulemusi ning valitakse välja parim mudelivariant olemasolevatest. Viimasega tehakse ka süvaanalüüs ning esitatakse mudeli interpretatsioon. Lõpuks esitatakse kokkuvõte ja järeldused tehtud tööst.

Töö praktilise osa läbiviimisel kasutati peamiselt rakendustarkvara R ning tabelarvutussüsteemi MS Excel.

1 Valdkonna tutvustus

Käesoleva töö arusaadavuse ja mõistetavuse huvides on järgnevas peatükis lahti seletatud uuritava valdkonnaga seotud põhilised mõisted ja terminid.

Krediidirisk on tõenäosus kaotada raha vastaspoole suutmatuse, tahtmatuse või mitteõigeaegsuse tõttu rahalise kohustuse (antud juhul laenu) täitmisel/tagasi maksmisel. Millal iganes on võimalus, et vastaspool ei maksa võlgnetavat summat, ei täida rahalist kohustust või ei austa kokkulepitud nõuet, on olemas krediidirisk (Bouteille ja Coogan-Pushner, 2012).

Krediidiskooring on otsustusmudelite kogum, mis aitab krediidiasutustel (laenuandjatel) läbi viia laenuandmise protsessi. Neid meetodeid kasutatakse selleks, et otsustada, kes saab laenu, mis on saadava laenu suurus ning milline tegevuskava aitab suurendada laenuvõtjate kasumlikkust laenuandjate jaoks (Thomas, Edelman ja Crook, 2002).

Kõige laialdasemalt kasutatud krediidiskooringu süsteem on FICO skoor (FICO *Scores*), mis on loodud *Fair Isaac Corporation*'i poolt. FICO skoor on kasutusel enamike krediidiasutuste puhul ning see aitab asutustel iga aasta langetada miljardeid krediidialaseid otsuseid. Skoor arvutatakse üksnes krediidiasutustes olevate tarbijakrediidi aruannete põhjal

FICO skoor arvutatakse krediidiraportist saadud andmete põhjal, mis on jaotatud viite kategooriasse: maksete ajalugu (35%), võlgnetav summa (30%), krediidiajaloo pikkus (15%), uue krediidikohustise teke (10%) ja krediidikohustiste jaotus (10%) (*Fair Isaac Corporation*, kasutatud 07.05.2019).

Antud töö ülesanne on prognoosida kliendi maksejõulisust, võttes arvesse teda iseloomustavat tunnuste vektorit. Töö käigus kasutatakse erinevaid mudeleid ning lõpuks jõutakse selgusele, missugune nende hulgast annab parima tulemuse.

2 Kasutatav metoodika

Järgnev peatükk põhineb õpikutel Koenker, R., Yoon, J. (2009, lk 1-3) ja Tutz, G. (2012, lk 29-30), kui pole viidatud teisiti.

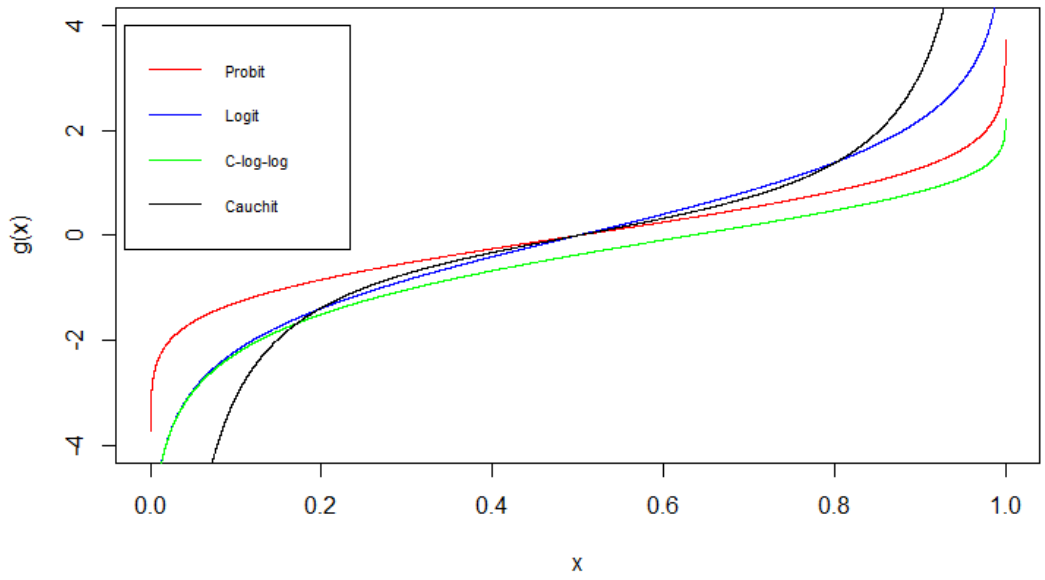
Käesolevas töös kasutatakse binaarse funktsioontunnusega (uuritava tunnusega) Y regressioonimudeleid, st mudeleid, mille funktsioontunnusel esineb kaks võimalikku väärtust, mis kodeeritakse järgnevalt: 1 tähistab juhtu, kui vaadeldav sündmus toimus, ja 0, kui vaadeldav sündmus ei toimunud. Tähistame sündmuse toimumise tõenäosuse $\pi_i = P(Y_i = 1)$ ja sündmuse mittetoimumise tõenäosuse $1 - \pi_i = P(Y_i = 0)$, kus Y_i on funktsioontunnuse Y väärtus i-ndal objektil.

Vaatleme järgmist üldistatud lineaarset mudelit:

$$g(\pi_i) = \boldsymbol{x_i^T \beta},$$

kus $\boldsymbol{x_i^T}$ on transponeeritud kujul kovariantide vektor ning $\boldsymbol{\beta}$ on mudeli argumenttunnuste parameetrite vektor ja g on nn. seosefunktsioon, mis seob regressortunnuste põhjal ehitatud lineaarse prognoosi ja meid huvitava tõenäosuse. Kuna prognoositakse tõenäosust, mis teatavasti on tõkestatud lõigul [0,1], siis on otstarbekas leida tõenäosuse teisendus (üksühene, pidev, diferentseeruv) kogu reaalteljele. Binaarse funktsioontunnusega mudelite jaoks on kasutusel mitmeid seosefunktsioone. Käesoleva töö jaoks on nendest välja valitud järgmised: *logit*, *probit*, *c-log-log*, *cauchit*. Järgnevates peatükkides antakse neist kõigist ülevaade.

Lisaks on joonisel 1 visuaalselt esitatud kõigi nelja eelnevalt mainitud seosefunktsiooni graafikud.



Joonis 1. Seosefunktsioonide võrdlusgraafik

2.1 *Logit* mudel

Järgnev peatükk põhineb õpikul Tutz, G. (2012, lk 30-42). Lisaks sellele on veel kasutatud raamatut Hosmer ja Lemeshow (2000:6-7, 1-3) ning Andmeanalüüs II loengukonspekti.

Viimastel aastakümnetel on logistilise regressiooni mudel kujunenud üheks oluliseks analüüsimeetodiks ja seda ka krediidiriski valdkonnas. Prognoositava tõenäosuse leidmiseks kasutatakse siin *logit* seosefunktsiooni, mis avaldub kujul:

$$g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right),$$

kus suhet $\frac{\pi_i}{1-\pi_i}$ nimetatakse sündmuse esinemise šansiks.

Logistilise regressioonimudeliga prognoositakse seega šansi logaritmi

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k,$$

kus $\pi_i = P(Y_i = 1)$ on sündmuse esinemise tõenäosus ja k on argumenttunnuste arv. Parameetrid $\beta_0, \beta_1, \beta_2, ..., \beta_k$ on on vastavate argumenttunnuste $x_1, x_2, ..., x_k$ kordajad mudelis.

Logit seosest saame avaldada sündmuse esinemise tõenäosuse järgnevalt:

$$\pi_i = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}.$$

2.2 *Probit* mudel

Järgnev peatükk põhineb õpikul Tutz, G. (2012, lk 123).

Probit mudeli näol on tegu laialt kasutatud mudeliga (eriti majandusvaldkonnas), mis baseerub standardse normaaljaotuse jaotusfunktsioonil $\Phi(\eta) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\eta} e^{-\frac{x^2}{2}}\, dx$.

Probit seosefunktsioon avaldub kujul

$$g(\pi_i) = \Phi^{-1}(\pi_i).$$

Seega *probit* mudel on järgmine:

$$\Phi^{-1}(\pi_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

Probit mudeli rakendamisel saadakse üldjuhul sarnaseid tulemusi logit mudeliga. Võrreldav on ka mudeli sobivus, samuti ei ole erinevusi statistiliselt oluliste tunnuste vahel. Ka olulisuse tõenäosused (ehk p-väärtused) on sisuliselt samad. Teisalt tuleb silmas pidada asjaolu, et mudeli parameetrite hinnanguid ei tohi üks-ühele võrrelda *logit* mudeli omadega.
Logit ja *probit* mudeli eristamiseks on vaja väga suurt valimit. Kohati võib osutuda takistuseks, et *probit* seosefunktsioonil pole ilmutatud kuju ja et mudeli parameetrid pole sama lihtsasti interpreteeritavad kui *logit* mudelis. Sellest hoolimata on tulemused sarnased.

2.3 Täiend-log-log mudel (*c-log-log* mudel)

Järgnev peatükk põhineb õpikul Tutz, G. (2012, lk 124-125).

C-log-log mudel on seotud Gompertzi jaotusega, mille jaotusfunktsioon avaldub kujul

$$F(\eta) = 1 - \exp(-\exp(\eta)).$$

Eeldades, et sündmuse tõenäosus π avaldub kujul $\pi = F(\eta)$, saame seose

$$\eta = \log(-\log(1-\pi)).$$

Sellest tulenebki täiend-log-log mudeli seosefunktsiooni kuju:

$$g(\pi_i) = \log(-\log(1-\pi_i)).$$

Seega täiend-log-log mudel näeb välja järgmine:

$$\log(-\log(1-\pi_i)) = \boldsymbol{x_i^T \beta}\,.$$

Nimetus „täiend-log-log“ tuleb mudeli kujust, kus on näha linkfunktsiooni log-log mõju täiendtõenäosusele $1 - \pi_i$.

Märgime, et kui *logit* ja *probit* funktsiooni puhul on tegemist sümmeetrilise jaotusega, siis Gompertzi jaotus on asümeetriline.

2.4 *Cauchit* mudel

Järgnev peatükk põhineb õpikul Tutz, G. (2012, lk 126).

Cauchit seosefunktsioon kasutab (standardset) Cauchy jaotusfunktsiooni

$$F(\eta) = \tan^{-1}\frac{\eta}{\pi} + \frac{1}{2},$$

kus $\tan^{-1} = arctan$ puhul on tegu tangensi pöördfunktsiooniga, $\pi = 3.14159$ ....

Cauchy jaotuse eripära seisneb selles, et tal ei ole ei keskväärtust ega dispersiooni. Teisalt on defineeritud mood ja mediaan ning need võrduvad mõlemad nulliga. Cauchy jaotus langeb Studenti t-jaotusega kokku vabadusastmega 1.

Cauchit seosefunktsioon avaldub kujul

$$g(u) = \tan\left(\pi\left(u - \frac{1}{2}\right)\right)$$

ning selle abil saadakse mudelid

$$\pi_i = \tan^{-1}\frac{\boldsymbol{x_i^T\beta}}{\pi} + \frac{1}{2} \qquad \text{ja} \qquad \tan\left(\pi\left(\pi_i - \tfrac{1}{2}\right)\right) = \boldsymbol{x_i^T\beta}\,,$$

kus π_i tähistab sündmuse esinemise tõenäosust.

Võrreldes normaaljaotusega on Cauchy jaotusel raskemad sabad, mis lubab esineda ekstreemsetel väärtustel sagedamini kui normaaljaotuse korral. See asjaolu muudab mudelit (võrreldes *logiti* ja *probiti* mudeliga) tolerantsemaks erindite suhtes.

2.5 Regressiooni parameetrite hindamine

Järgnev peatükk põhineb õpikul Tutz, G. (2012, lk 82-83).

Kui mudeli eesmärk on kirjeldada suhet sõltuva tunnuse (Y) ja mitme seletava tunnuse (X) vahel, siis üks esimesi samme analüüsi osas on mudeli sobitamine ehk tundmatute parameetrite hindamine.

Üks sagedane parameetrite hindamisemeetod kannab nime suurima tõepära meetod (*maximum likelihood method,* STP-hinnang). Selle põhiprintsiip on konstrueerida nn tõepärafunktsioon valimiandmete jaoks ning leida parameetritele sellised väärtused, mis maksimeerivad selle tõepärafunktsiooni. Tõepära esindab ühistõenäosust või vaadeldavate andmete tõenäosuse tihedust, mida käsitletakse tundmatute parameetrite funktsioonina.

Olgu meil n sõltumatut vaatluste paari $(\boldsymbol{x}_i, y_i)$, $i = 1,2, \dots, n$, kus $y_i \in \{0,1\}$ on prognoositav muutuja ja $\boldsymbol{x}_i$ tähistab sõltumatute muutujate komplekti i-ndal objektil.

Olgu $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)$ STP-meetodil hinnatud mudeli parameetrite vektor, kus k tähistab kirjeldavate tunnuste arvu. Näiteks logistilise regressiooni puhul on hinnatavate parameetrite arv $k + 1$.

Siis tõepärafunktsioon avaldub kujul:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\boldsymbol{x_i})^{y_i}(1 - \pi(\boldsymbol{x_i})^{1-y_i} \qquad (2.5)$$

Suurima tõepära hinnang β maksimiseerib funktsiooni (2.5). Samas on matemaatiliselt lihtsam kasutada saadud avaldise naturaallogaritmi, mis ei muuda ülesande sisu. See avaldub järgmiselt:

$$\mathrm{l}(\boldsymbol{\beta}) = \ln[L(\boldsymbol{\beta})].$$

2.6 Mudeli headuse näitaja

Järgnev peatükk põhineb õpikul Fox, J. (2016, lk 673-677).

Parima mudeli valikul sammregressiooni abil kasutame Akaike' informatsioonikriteeriumit (AIC), mis on üks levinumaid mudeli headuse näitajaid. See leitakse kujul:

$$AIC = -2 \cdot l(\boldsymbol{\beta}) + 2 \cdot p,$$

kus $l(\boldsymbol{\beta})$ on maksimeeritud log-tõepära , $\boldsymbol{\beta}$ on parameetervektori STP hinnang ning p on mudeli parameetrite arv. Sammregressiooni täpsemad skeemid on kirjeldatud allpool (vt. punkt 3.2.1).

3 Töö praktiline osa

Selles töö peatükis tutvustatakse täpsemalt töös kasutatud andmestikku ja selles sisalduvaid tunnuseid. Samuti antakse seletused tunnuste teisenduste kohta.

Edasine tegevus on jaotatud kahte etappi:

1. etapis uuritakse kõigi 4 seosefunktsiooniga mudeleid, kus tunnused on algsel, töötlemata kujul ning koosmõjusid pole lisatud. Selle etapi lõpuks valitakse võrdluse tulemusena välja kõige sobivam seosefunktsioon.
2. etapis analüüsitakse väljavalitut mudelit detailselt, teisendatakse/grupeeritakse tunnuseid ning katsetatakse potentsiaalsete koosmõjude olulisust. Samuti teostatakse mudeli diagnostika ning mudeli interpretatsioon.

Mudelite loomisel ja võrdlemisel on kasutatud statistikatarkvara R ning selle lisapakette.

3.1 Andmestiku kirjeldus, ülevaade tunnustest

Töö praktilises osas kasutatakse empiirilisi andmeid, kus on 3800 andmerida ning uuritavaid tunnuseid on 16. Vaatluse all on kiirlaenude andmestik, millele on iseloomulikud lühikesed laenuperioodid. Andmestik sisaldab laenusaajate kohta järgmisi tunnuseid:

- Staatus (*Status*) – 1 (hea) ja 0 (halb)
- Sugu (*Sex)* – M (mees) ja F (naine)
- Vanus aastates (*Age)*
- Maakonna nimetus (*Region)*
- Emakeel (*Language*)
- Laenusumma eurodes (*Sum*)
- Laenuperiood päevades (*Period*)
- Kuine sissetulek eurodes (*Income*)
- Kuine väljaminek eurodes (*Outcome*)
- Perekonnaseis (*Family*)
- Haridustase (*Education*)
- Töökogemus (*WorkExperience*)
- Laste arv (*Children*)
- Kinnisvaraobjektide arv (*Estate*)
- Maksehäirete arv kokku (*PaimentAlertsTotal*)
- Aktiivsete maksehäirete arv (*PaimentAlertsActive*)
- Lõpetatud maksehäirete arv (*PaimentAlertsClosed*)

Töö eesmärk on prognoosida tunnust „Staatus“ (väärtustega 1/0) teiste ülaltoodud tunnuste järgi, kasutades selleks erinevaid mudeleid ning lõpuks jõuda selgusele, missugune mudel on selliseks prognoosiks kõige parem. Kasutatavad mudelid on *logit*, *probit*, *c-log-log* ja *cauchit*, kusjuures modelleeritakse tõenäosust olla hea staatusega klient (staatus = 1).

3.2 Esimene etapp: mudelite hindamine

Selles etapis uuritakse läbi kõik konkureerivad mudelid, tuuakse välja nende tulemused ja erisused ning lõpuks valitakse välja parima prognoosivõimega mudel.

Enne mudelite juurde jõudmist tuuakse veel eraldi välja vajalik teave kasutatud metoodika ja muu teoreetilise informatsiooni kohta, mis puudutab kõiki võrdluse all olevaid mudeleid.

3.2.1 Kirjeldavate tunnuste valik (sammregressioon)

Järgnev peatükk põhineb Tutz, G. (2012, lk 359-360) õpikul, kui pole viidatud teisiti.

Käesolevas töös on regressioonimudeli kasutamisel kasutatud sammregressiooni meetodit (*stepwise regression)*, mille eesmärk on automatiseerida ning sealjuures lihtsustada argumentide valikut loodavasse mudelisse.

Käesoleva töö puhul kasutatakse kahte sammregressiooni strateegiat/põhimõtet:

1) ettepoole valik (*forward);*
2) tahapoole valik (*backward*).

Esimesel juhul lähtutakse mudelist, mis sisaldab üksnes vabaliiget ning seejärel lisatakse uus muutuja, mis annab üksikuna parima tulemuse mudeli hindamisvõimele ehk mille lisamisel üksikuna oleks mudeli AIC väikseim. Tegevust jätkatakse seni, kuni ühegi argumendi lisamine AIC väärtust enam ei vähenda. Kehtib põhimõte, et kord mudelisse valitud argumenti mudelist enam kõrvale ei jäeta.

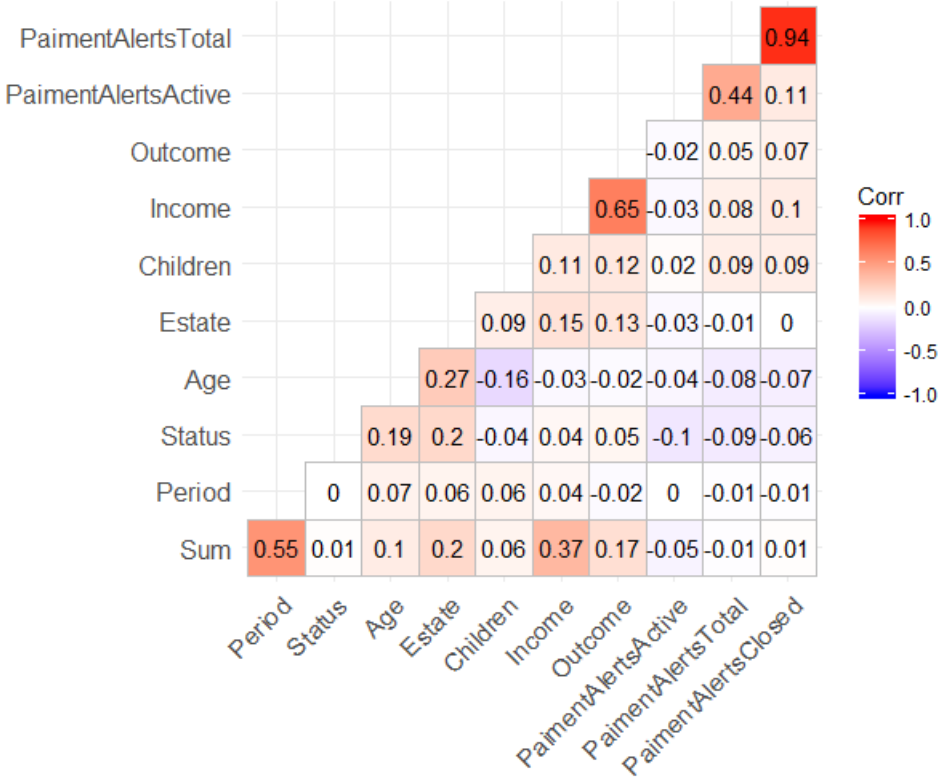
Teisel juhul on tegu esimesele vastupidise protsessiga. Tegevust alustatakse täismudeliga, mis sisaldab kõiki muutujaid ning igal sammul jäetakse kõrvale muutuja, mille üksikuna mudelist väljajätmine muudab mudelit täpsemaks ehk mille üksikuna välja jätmisel oleks mudeli AIC väikseim. Kehtib põhimõte, et kord mudelist väljajäetud argumenti mudelisse enam uuesti ei lisata.

3.2.2 Korrelatsioonide uurimine

Korrelatsioonide uurimine on vajalik selleks, et teha selgeks, missugused on seosed erinevate tunnuste vahel ning missugused on nende seoste tugevused. Selle uurimiseks on koostatud korrelatsioonimaatriks, et saada ülevaade kõikvõimalike tunnustepaaride korrelatsioonidest.
Käesoleva andmestiku puhul on enne sammregressiooni rakendamist eemaldatud maksimaalsest mudelist (*maximum model)* tunnus „Maksehäirete arv kokku“, kuna see on väga tugevalt korreleeritud tunnusega „Suletud maksehäirete arv kokku“ – vastava korrelatasioonikordaja väärtus on 0.94. See tähendab seda, et maksehäirete koguarvu muutus kirjeldab väga suurel mõjul ka suletud maksehäirete arvu muutust. Kuna see aga vähendab mudeli hindamisvõimet ning -täpsust, on mõistlik üks nendest tunnustest välja jätta.

Samuti võeti arvesse tunnuste „sissetulek“ ja „väljaminek“ kõrget korrelatsiooni, kuid katsetamise tulemusel selgus, et kummagi muutuja üksikult väljajätmine mudeli hindamisvõimet ei paranda. Seetõttu otsustati siinkohal muutusi mitte sisse viia.

Järgnevalt on ära toodud ka korrelatsioonimaatriks:



Joonis 2. Korrelatsioonimaatriks

3.2.3 Mudelite esmane võrdlus

Järgnevalt on vaatluse all kõik neli konkureerivat mudelit. Antud juhul rakendatakse kõigi mudelite puhul sammregressiooni meetodit vastava mudeliga, sealjuures teisendamata seletavaid tunnuseid, ning seejärel valitakse Akaike’ informatsioonikriteeriumi põhjal nende seast sobivaim.

Igast mudelist on multikollineaarsuse tõttu eelnevalt eemaldatud tunnus „maksehäirete arv kokku“.

Logit mudeli korral saadakse AIC väärtuseks 4233 (programmiväljund asub lisades, vt lisa nr 1). Mudeli kuju avaldub järgnevalt:

$$\begin{aligned}\log\left(\frac{\pi}{1-\pi}\right) = & -0.095 \\ & + 0.610 \cdot kinnisvaraobjektide\ arv \\ & + 0.024 \cdot vanus \\ & - 0.240 \cdot aktiivsete\ maksehäirete\ arv \\ & - 0.001 \cdot laenusumma \\ & - 0.360 \cdot I_{\{sugu=mees\}} \\ & - 0.038 \cdot suletud\ maksehäirete\ arv \\ & + 0.001 \cdot väljaminek \\ & - 0.130 \cdot laste\ arv \\ & + 0.199 \cdot I_{\{emakeel=vene\}},\end{aligned}$$

kus π tähistab sündmuse esinemise tõenäosust ning kus I on indikaatorfunktsioon, mille väärtus on 1, kui vastav tingimus on täidetud ning 0 vastasel juhul.

Probit mudeli korral saadakse AIC väärtuseks 4236,3 (programmiväljund asub lisades, vt lisa nr 2). Mudeli kuju avaldub järgnevalt:

$$\begin{aligned}\Phi^{-1}(\pi) = & -0.046 \\ & + 0.327 \cdot kinnisvaraobjektide\ arv \\ & + 0.014 \cdot vanus \\ & - 0.146 \cdot aktiivsete\ maksehäirete\ arv \\ & - 0.0004 \cdot laenusumma \\ & - 0.224 \cdot I_{\{sugu=mees\}} \\ & + 0.0005 \cdot väljaminek \\ & - 0.024 \cdot maksehäirete\ arv\ kokku \\ & + 0.124 \cdot I_{\{emakeel=vene\}} \\ & - 0.07 \cdot laste\ arv,\end{aligned}$$

kus π tähistab sündmuse esinemise tõenäosust.

Kuna *logiti* mudel on krediidiriski hindamisel väga levinud mudel, siis võrreldakse seda mudelit teiste „konkurentidega". Niisiis, kui võrrelda *logiti* ja *probiti* mudelit, siis ainuke erinevus seisneb selles, et *probiti* puhul on „suletud maksehäirete arvu" asemel mudelis „maksehäirete arv kokku". Veel võib välja tuua tunnuse „kinnisvaraobjektide arv" kordaja erinevuse - *logiti* puhul 0.610, *probiti* puhul 0.327.

C-log-log mudeli korral saadakse AIC väärtuseks 4242 (programmiväljund asub lisades, vt lisa nr 3). Mudeli kuju avaldub järgnevalt:

$$\begin{aligned}\log(-\log(1-\pi)) = \; & -0.365 \\ & +0.250 \cdot kinnisvaraobjektide\ arv \\ & +0.014 \cdot vanus \\ & -0.157 \cdot aktiivsete\ maksehäirete\ arv \\ & -0.0004 \cdot laenusumma \\ & -0.234 \cdot I_{\{sugu=mees\}} \\ & +0.0005 \cdot väljaminek \\ & -0.003 \cdot suletud\ maksehäirete\ arv \\ & +0.127 \cdot I_{\{emakeel=vene\}} \\ & -0.058 \cdot laste\ arv,\end{aligned}$$

kus π tähistab sündmuse esinemise tõenäosust.

Logiti ja *c-log-log* mudelil on statistiliselt oluliseks osutunud tunnuste nimekiri ühesugune. Kordajate erinevuse mõttes võib siinkohal välja tuua tunnused „laste arv“ (*logiti* puhul -0.130, *c-log-log* -0.058) ning taaskord „kinnisvaraobjektide arv“ (*logit* 0.610, *c-log-log* 0.250).

Cauchit mudeli korral saadakse AIC väärtuseks 4248 (programmiväljund asub lisades, vt lisa nr 4). Mudeli kuju avaldub järgnevalt:

$$\begin{aligned}\tan\left(\pi^*\left(\pi-\frac{1}{2}\right)\right) = \; & -0.191 \\ & +0.783 \cdot kinnisvaraobjektide\ arv \\ & +0.023 \cdot vanus \\ & -0.228 \cdot aktiivsete\ maksehäirete\ arv \\ & -0.00056 \cdot laenusumma \\ & -0.328 \cdot I_{\{sugu=mees\}} \\ & -0.154 \cdot laste\ arv \\ & +0.0007 \cdot väljaminek \\ & +0.181 \cdot I_{\{emakeel=vene\}},\end{aligned}$$

kus $\pi^* = 3{,}14159\ldots$ ja π tähistab sündmuse esinemise tõenäosust.

Kui kõigis kolmes eelnevas mudelis on maksehäiretega seonduvaid tunnuseid kaks, siis *cauchit* mudelis on neid kõigest üks. Selles seisneb ka *cauchit* ainuke erinevus teistest mudelitest (sh ka *logit* mudelist).

Eelnevate tulemuste põhjal on koostatud järgmine kokkuvõtlik tabel, kus on iga mudeli puhul välja toodud statistiliselt oluliseks osutunud kordajate väärtused. Viimases tabelireas on kirjas AIC väärtus vastava meetodi korral.

Mudel / Tunnus	*Logit*	*Probit*	*C-log-log*	*Cauchit*
Vabaliige	-0.095	-0.046	-0.36	-0.191
Kinnisvaraobjektide arv	0.610	0.327	0.250	0.783
Vanus	0.024	0.014	0.0143	0.023
Aktiivsete maksehäirete arv	-0.240	-0.146	-0.157	-0.228
Laenusumma	-0.001	-0.0004	-0.0004	-0.00056
Sugu (=mees)	-0.360	-0.224	-0.234	-0.328
Suletud maksehäirete arv	-0.038	-	-0.0029	-
Väljaminek	0.001	0.0005	0.00046	0.0007
Laste arv	-0.130	-0.07	-0.0583	-0.154
Emakeel (=vene)	0.199	0.124	0.127	0.181
Maksehäirete arv kokku	-	-0.024	-	-
AIC	**4233**	**4236,3**	**4242**	**4248**

Tabel 1. Kordajate hinnangud sobitatud mudelites

Ülaltoodud võrdlev analüüs võimaldas välja valida sobivaima mudeli, milleks osutus *logit* mudel. Selline otsus langetati Akaike' informatsioonikriteeriumi põhjal.

3.3 Teine etapp: parima mudeli süvaanalüüs

Selles etapis jätkatakse detailset analüüsi 1. etapi lõpus väljavalitud mudeliga, milleks osutus logistilise regressiooni mudel. Selles alapeatükis on ära toodud informatsioon tunnuste töötlemise ning mudeli analüüsi ja interpretatsiooni kohta.

3.3.1 Argumenttunnuste teisendamine

Lõppmudelis kasutatavate tunnustega tehti läbi mitmeid teisendusi eesmärgiga parandada mudeli prognoosivõimet. Tegevuse ajendiks oli idee, et nii mõnigi argumenttunnus võiks osutuda mudelis statistiliselt oluliseks mõnel muul kujul kui tema algne kuju ning aidata seetõttu kaasa lõpptulemuse paranemisele. Järgnevalt on ära toodud täpsemad seletused nende tunnuste kohta, mis ei esinenud lõppmudelis oma algsel kujul.

Teisenduse järjekord oli järgmine: tegevust alustati nendest mittearvulistest tunnustest, mis algsel kujul olid mudelis statistiliselt ebaolulised. Kui kõik mittearvulised tunnused olid läbi proovitud, siis tegeleti edasi arvulistega tunnustega.

Järnevalt on välja toodud logit-mudeli n-ö üldmudel, mille baasil hakati teostama edasisi teisendusi.

```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  0.0345254  0.4743439   0.073  0.94198
Estate                       0.5711481  0.0636027   8.980  < 2e-16 ***
Age                          0.0177701  0.0035259   5.040 4.66e-07 ***
Educationei ole             -0.1257244  0.5070395  -0.248  0.80417
Educationkeskharidus         0.1079916  0.3984797   0.271  0.78638
Educationkutseharidus        0.2501803  0.3995628   0.626  0.53123
Educationkõrgharidus         0.5245467  0.4091083   1.282  0.19978
Educationpõhiharidus        -0.2670078  0.4081668  -0.654  0.51301
PaimentAlertsActive         -0.1946043  0.0492375  -3.952 7.74e-05 ***
WorkExperienceKuni aasta    -0.3252617  0.2438281  -1.334  0.18221
WorkExperienceRohkem kui aasta  0.1444185  0.2351106   0.614  0.53904
WorkExperienceTöötu          0.1742961  0.2931136   0.595  0.55209
Sum                         -0.0008045  0.0001766  -4.555 5.24e-06 ***
SexM                        -0.3336082  0.0826638  -4.036 5.44e-05 ***
PaimentAlertsTotal          -0.0404358  0.0165187  -2.448  0.01437 *
Outcome                      0.0005777  0.0001848   3.126  0.00177 **
Children                    -0.1021868  0.0437502  -2.336  0.01951 *
Languagerus                  0.1642221  0.0783343   2.096  0.03604 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4549.2  on 3799  degrees of freedom
Residual deviance: 4156.6  on 3782  degrees of freedom
AIC: 4192.6

Number of Fisher Scoring iterations: 4
```

Tabel 2. Üldmudeli programmiväljund

Maakonnad olid algsel kujul andmestikus 15 tasemega faktorina. Kuna tunnus sellisel kuju statistiliselt oluliseks ei osutunud, siis otsustati maakonnad jaotada üldmudeli koefitsientide põhjal kahte gruppi. See muudatus tegi maakonnatunnuse statistiliselt oluliseks ning parandas ka mudeli AICi.

Hariduse tunnus oli algsel kujul 6-tasemeline ning esmases mudelis ei osutunud kuigivõrd oluliseks tunnuseks. Seetõttu moodusati 6-tasemelisest faktorist 3-tasemeline, kusjuures tasemeteks valiti "alg/põhi/ei ole", "Kesk/kutseharidus", "Kõrgharidus". Selline grupeering tehti lähtuvalt üldmudeli koefitsientidest ning see parandas mudeli hindamisvõimet.

Vanus oli algselt mudelis lineaarsel kujul, kuid katsetamise käigus selgus, et mudelis osutub AIC paremaks juhul, kui vanusetunnus on mudelisse kaasatud lisaks lineaarkujule ka ruutliikmena. Kuna see on ka üsna levinud praktika vanuse tunnuse puhul, sest vanusel

ei tavatse olla mittelineaarne suhe sõltumatu muutujaga, siis otsustatigi selline muudatus sisse viia.

Laenuperiood (päevades) oli algsel kujul üldmudelis statistiliselt ebaoluline. Pärast tunnuse jaotuse ja histogrammi uurimist selgus, et mõistlik on laenuperiood jaotada kaheks – pikkadeks (st üle 30 päeva) ja lühikesteks (30 ja vähem päeva). Sellisel moel tuli ka tunnus statistiliselt oluline ning paranes ka mudeli hindamisvõime.

Kinnisvara objektide arvu puhul prooviti mudelisse sobitamisel kahte varianti – pidev kuju ning 0/1 kuju. Analüüsi tulemusel selgus, et teine variant on üldmudeli seisukohast parem.

Kõigi maksehäirete tunnuste (kokku, aktiivsed, suletud) puhul prooviti mudelisse sobitamisel kahte varianti - pidev kuju ning 0/1 kuju. Analüüsi tulemusel selgus, et teine variant on üldmudeli seisukohast parem.

Järgnevalt on välja toodud teave kõigi tunnuste lõppkujude kohta:

- Staatus – 1 (hea) ja 0 (halb) – pidev tunnus (mitte faktor)
- Sugu – M (mees) ja F (naine) – 2-tasemeline faktor (algne kuju)
- Vanus (aastates) – lineaarliikme ja ruutliikmena pidev tunnus
- Maakonna nimetus – 15 maakonda on jaotatud kahe grupi vahel (koefitsientide põhjal). Sellisel moel on see tunnus mudelis oluline ja see parandab ka AIC väärtust. Grupid moodustusid järgnevalt:
    - Grupp 1: Ida-Virumaa, Järvamaa, Pärnumaa, Raplamaa, Tartumaa, Jõgevamaa, Lääne-Virumaa, Põlvamaa, Saaremaa, Viljandimaa, Võrumaa
    - Grupp 2: Valgamaa, Hiiumaa, Läänemaa, Harjumaa
- Emakeel – 2-tasemeline faktor (algne kuju)
- Laenusumma (eurodes) – pidev tunnus (algne kuju)
- Laenuperiood (päevades) – pidevast tunnusest on tehtud 2-tasemega kategooriline tunnus tasemetega "üle 30" ja "30 ja alla". Nii on see mudelis oluline ja mudeli hindamisvõime paranes oluliselt
- Kuine sissetulek (eurodes) – pidev tunnus (algne kuju)
- Kuine väljaminek (eurodes) – pidev tunnus (algne kuju)
- Perekonnaseis - Prooviti jaotada tasemed järgmiselt: "Abielus","Lahutatud/lesk", "Vabaabielu/vallaline". Aga kuna ka see ei osutunud oluliseks, siis perekonnaseisu arvesse ei võetud.
- Haridustase – Moodustati 6-tasemelisest faktorist 3-tasemelise. Tasemed: "alg/põhi/ei ole", "Kesk/kutseharidus", "Kõrgharidus". Nii on see mudelis oluline ja mudeli hindamisvõime paranes.
- Töökogemus – Moodusati 4-tasemelisest faktorist 2-tasemelise. Tasemed: "Kuni aasta", "Katseaeg/töötu/Rohkem kui aasta". Nii on see mudelis oluline ja mudeli hindamisvõime paranes.
- Laste arv – pidev tunnus (algne kuju)
- Kinnisvaraobjektide arv – pidev tunnus on muudetud 1/0 (jah/ei) tunnuseks (1 – on kinnisvara, 0 – ei ole)

- Maksehäirete arv kokku - on muudetud pidevast tunnusest 1/0 tunnuseks (1- esines maksehäireid, 0 – ei esinenud maksehäireid)
- Aktiivsete maksehäirete arv - on muudetud pidevast tunnusest 1/0 tunnuseks (1- esines aktiivseid maksehäireid, 0 – ei esinenud aktiivseid maksehäireid)
- Lõpetatud maksehäirete arv - on muudetud pidevast tunnusest 1/0 tunnuseks (1- esines lõpetatud maksehäireid, 0 – ei esinenud lõpetatud maksehäireid)

Peale tunnuste teisendamist ning multikollineaarsete tunnuste eemaldamist otsustati luua eraldi mudelid ka koosmõjude katsetamiseks, et uurida, kas sellel on mudeli headuse seisukohast lisandväärtust.

Koostati mudelid järgmiste koosmõjudega, mis tundusid töö autori jaoks olevat kõige loogilisemad ning tõenäolisemad:

- Sugu ja laste arv
- Sugu ja perekonnaseis
- Vanus ja regioon
- Vanus ja haridus
- Sugu ja haridus
- Sugu ja regioon

Uuringu tulemusena selgus, et kahjuks ükski neist koosmõjudest ei osutunud lõppmudelis statistiliselt oluliseks. Kõige lähedasem sellele oli esimesel koosmõjupaaril „sugu ja laste arv“, kus vastav hinnangukordaja oli 0.266, olulisuse tõenäosus 0.11 ning sealjuures paranes ka üldmudeli AIC. Kuna aga valitud olulisuse nivoo oli 0.05 ja olulisuse tõenäosus tuli 0.11, siis otsustati see siiski mudelist välja jätta.

3.3.2 Mudeli lõikepunkti määramine

Krediidi väljastamisel on ülimalt oluline veenduda kliendi maksejõulisuses. Antud töö puhul on võetud mudeli konstrueerimisel aluseks põhimõte, et kasutatav mudel peab olema väga täpne selgitamaks, kas inimene maksab tagasi või mitte.

Käesoleva andmestiku puhul esineb vigade asümmetria – see tähendab, et vead on erineva kaaluga. Näiteks olukord, kus pank väljastab laenu, aga klient tagasi ei maksa, võib olla pangale ca 5 korda kulukam kui olukord, kus pank laenu ei väljasta, aga tegelikkuses oleks klient olnud võimeline laenu saamise korral seda tagasi maksma.

Eelneva probleemi lahenduseks on rakendatud päriselus levinud praktikat, kus kõrgem viga 5 ühikut on määratud nendele klientidele, keda mudel ennustas maksujõulisteks ning kellele laen väljastati, aga kes tegelikkuses ei olnud võimelised laenu tagasi maksma. Vastupidisele juhule määrati vea suuruseks1 ühik. Selliselt toimides saadi tabel, kuhu oli koondatud vaadeldav lõikepunkt (vahemikust 0,1 – 0,9 sammuga 0,01) ning vaadeldava kirje (rea) veahinnang (0; 1 või 5). Veahinnang 0 määrati juhul, kes said laenu ning maksid ka selle tagasi; 1 määrati juhul, kui pank laenu ei andnud, aga tegelikkuses oleks

klient olnud võimeline laenu saamise korral seda tagasi maksma; 5 määrati juhul, kus pank väljastas laenu, aga klient ei suutnud seda tagasi maksta. Seejärel otsiti lõikepunkti, mille korral vigade summaarne hind oleks minimaalne – teisisõnu, otsiti, millise lõikepunkti korral eksib mudel kõige vähem. Selliselt toimides saadi lõikepunktiks 0,83.

3.3.3 Lõppmudeli interpretatsioon

Tuginedes Akaike' informatsioonikriteeriumile, osutus parimaks mudeliks järgmine logistilise regressiooni mudel. Sellel on 11 statistilist olulist tunnust, Akaike' informatsioonikriteeriumi väärtuseks on 4120 ning see avaldub järgmisel kujul:

$$
\begin{aligned}
Logit(\pi) = {} & 1.564 \\
& + 0.766 \cdot I_{\{kinnisvaraobjektide\ olemasolu\ =\ jah\}} \\
& - 0.512 \cdot I_{\{\text{üldine maksehäirete olemasolu} = jah\}} \\
& - 0.572 \cdot I_{\{periood=\text{üle } 30\}} \\
& - 0.056 \cdot vanus \\
& - 0.001 \cdot vanus^2 \\
& - 0.593 \cdot I_{\{haridus\ =\ põhi/ei\ ole\}} \\
& - 0.249 \cdot I_{\{haridus\ =kesk/kutseharidus\}} \\
& + 0.478 \cdot I_{\{töökogemus\ =katseaeg/töötu/rohkem\ kui\ aasta\}} \\
& + 0.298 \cdot I_{\{regioon\ =grupp\ 2\}} \\
& - 0.267 \cdot I_{\{sugu\ =\ mees\}} \\
& + 0.001 \cdot väljaminek.
\end{aligned}
$$

Uuritavas mudelis tähistab π tõenäosust olla hea staatusega klient (staatus = 1).

```
> print(summary(fwd2),digits=3)

Call:
glm(formula = Status ~ Estate + PaimentAlertsTotal + Period +
    Age + I(Age^2) + Education + WorkExperience + Region + Sex +
    Outcome, family = binomial(link = "logit"), data = algnedata)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.429  -1.073   0.579   0.836   1.693

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                               1.563754   0.476111    3.28  0.00102 **
EstateJah                                 0.765746   0.085978    8.91  < 2e-16 ***
PaimentAlertsTotalJah                    -0.512336   0.077173   -6.64  3.2e-11 ***
Periodüle 30                             -0.572577   0.079507   -7.20  6.0e-13 ***
Age                                      -0.055792   0.023806   -2.34  0.01910 *
I(Age^2)                                  0.000887   0.000283    3.13  0.00173 **
Educationalg/põhi/ei ole                 -0.593146   0.146930   -4.04  5.4e-05 ***
EducationKesk/kutseharidus               -0.248524   0.114111   -2.18  0.02941 *
WorkExperienceKatseaeg/töötu/Rohkem kui aasta  0.477817   0.098366    4.86  1.2e-06 ***
Regiongrupp2                              0.297747   0.078343    3.80  0.00014 ***
SexM                                     -0.266657   0.080266   -3.32  0.00089 ***
Outcome                                   0.000465   0.000181    2.57  0.01032 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4549.2  on 3799  degrees of freedom
Residual deviance: 4096.1  on 3788  degrees of freedom
AIC: 4120
```

Tabel 3. Lõppmudeli programmiväljund

Kõik järgnevad võrdlused on mõeldud selliselt, et ühe tunnuse väärtused kahel isikul on erinevad ja kõikide teiste tunnuste väärtused on samad.

Analüüsides kinnisvara omamise tähtsust selgub, et kinnisvara omanikel on $(e^{0.766} = 2.151; \frac{1}{2.151} = 0.465)$ 46,5 % suuremad šansid osutuda paremaks kliendiks kui kinnisvara mitteomavatel klientidel.

Kui võrrelda kahte klienti, kellest ühel esineb maksehäireid ja ja teisel mitte, siis maksehäiretega kliendil on ($e^{-0.512} = 0.599; \frac{1}{0.599} = 1{,}67$) 67% väiksemad šansid osutuda maksujõuliseks kliendiks võrreldes inimesega, kellel maksehäireid pole.

Kui võrrelda kahte klienti, kellest ühel on laenuperiood alla 30 päeva ning teisel üle 30 päeva (pikk laen), siis pika laenuga kliendil on 77% $\left(e^{-0.572} = 0.564; \frac{1}{0.564} = 1{,}77\right)$ väiksemad šansid osutuda hea maksekäitumisega kliendiks.

Kuna mudelisse kaasati vanusetunnus nii lineaarliikme kui ka ruutliikmena, otsustati sellest lähtuvalt interpreteerida seda tunnust natuke erinevalt. Nimelt leiti selline vanuse väärtus, kuhuni toimub kasv/langus ja kust alates toimub muutus (sõltuvalt eelnevast kas kasv või langus) kliendi maksejõulisuse osas.
Selleks leiti vanusetunnuse kordajate ekstreemum, st võrdsustati esimene tuletis (kus x tähistab vanust) nulliga ning leiti see punkt, mis hetkel toimub muutus laenu tõenäosuses. See avaldub valemite kujul järgnevalt:

$$(-0.056 \cdot x + 0.00089 \cdot x^2)' = 0$$
$$(-0.056 + 0.00178 \cdot x) = 0$$
$$x = 31{,}46$$

Selliseks vanuseks saadi 31 – kuni selle vanuseni tõenäosus laenu tagasi maksmiseks kahaneb, kuid sellest edasi suureneb.

Kui võrrelda kahte klienti, kellest üks kuulub haridusetasemelt gruppi „ei ole/alg/põhi“ ning teine on kõrgharidusega, siis esimesse gruppi kuuluval on ($e^{-0.593} = 0.552; \frac{1}{0.552} = 1.812$) 81,2% väiksemad šansid osutuda heaks kliendiks.

Kui võrrelda kahte klienti, kellest üks kuulub haridusetasemelt gruppi „kesk/kutseharidus“ ning teine on kõrgharidusega, siis esimesse gruppi kuuluval on ($e^{-0.248} = 0.780; \frac{1}{0.780} = 1{,}282$) 28,2% väiksemad šansid osutuda heaks kliendiks.

Kui võrrelda kahte klienti, kellest üks kuulub tööstaaži pikkuselt gruppi „katseaeg/töötu/rohkem kui aasta“ ning teine klient on töötanud alla ühe aasta, siis esimesse gruppi kuuluval on ($e^{0.477} = 1.611$) 61.1% suuremad šansid osutuda korralikuks laenu tagasimaksjaks.

Klient, kes kuulub elukoha piirkonnalt gruppi 2, omab ($e^{0.297} = 1.346$) 34,6% suuremaid šansse osutuda paremaks kliendiks võrreldes gruppi 1 kuuluva kliendiga.

Meessoost laenusaaja omab võrreldes naissoost kliendiga ($e^{-0.267} = 0.766; \frac{1}{0.766} = 1.305$) 30,5% väiksemaid šansse osutuda heaks laenu tagasimaksjaks.

Kliendil, kelle väljaminekud on teisega võrreldes 100 euro võrra suuremad, on ($e^{0.0465} = 1.0476$) 4,76 % suuremad šansid olla hea klient. Veel võib täpsustuseks lisada, et antud juhul on siin varjatult seos ka sissetulekuga, st kui kliendil on suurem sissetulek, siis ta saab rohkem kulutada ja jõuab ka laenu tagasi maksta.

Järgnevalt on välja toodud loetelu nendest tunnustest, mille esinemine suurendab tagasimaksmise šanssi: kinnisvaraobjektide arv, vanus, töökogemus (=katseaeg/töötu/rohkem kui aasta), regioon (= grupp 2), väljaminek.

Täpsuse huvides on ära toodud ka gruppi nr 2 kuuluvad maakonnad: Valgamaa, Hiiumaa, Läänemaa, Harjumaa.

3.3.4 Mudeli prognooside täpsus

Järgnev peatükk põhineb õpikutel Fawcett, T. (2005, lk 861-874), kui pole viidatud teisiti.

Prognoosimise jaoks kasutatavate testide ja meetodite puhul on alati tähtsal kohal nende abil saadavate prognooside täpsus. Binaarse funktsioontunnuse puhul, nagu ka antud

uurimuse puhul, on prognoosi korrektsuse hindamiseks vajaminevad suurused koondatud järgmises tabelis.

Prognoos	Tegelik olek		
	Y = 0 (negatiivne)	Y = 1 (positiivne)	Kokku
Y = 0 (negatiivne)	TN	FN	TN + FN
Y = 1 (positiivne)	FP	TP	FP + TP
Kokku	TN + FP	FN + TP	TN + FN + FP+ TP

Tabel 4. Eksimismaatriks

Alljärgnevalt on välja toodud seletused tabelis esinevatele tähistustele:

- TN - nende juhtude arvu, kui uuritavat sündmust ei oleks prognoosi kohaselt tohtinud toimuda ega toimunud ka tegelikkuses. Seega tegu on tõeselt negatiivse juhtudega (*true negative, TN).*
- FN – nende juhtude arv, kus sündmuse toimumist prognoositi negatiivseks, kuid sündmus tegelikkuses toimus. Seega on tegu valenegatiivsete juhtude arvuga (*false negative, FN).*
- TP – nende juhtude arv, kus sündmust prognoositi positiivseks ning see toimus ka tegelikkuses. Seega on tegu tõeselt positiivsete juhtudega (*true positive, TP).*
- FP – nende juhtude arv, kus sündmust ennustati toimuvaks, kuid tegelikkuses see aset ei leidnud. Seega on tegu valepositiivsete juhtudega (*false positive, FP).*

Kasutades eeltoodud suurusi (TN, FN, TP, FP) on võimalik leida mitmeid prognoosi korrektsust hindavaid karakteristikuid, millest enim on kasutusel tundlikkus ja spetsiifilisus.

Tundlikkus (sensitiivsus, *sensitivity*) näitab, kui mitu protsenti uuritava sündmuse toimumisest ennustab kasutusel olev mudel õigesti:

$$Tundlikkus = \frac{TP}{TP + FN}.$$

Mõnes valdkonnas nimetatakse seda valemit „tõeselt positiivsete määraks“ (*true positive rate, TPR).*

Spetsiifilisus (*specificity)* näitab, kui mitu protsenti uuritava sündmuse mittetoimumisest ennustab kasutusel olev mudel õigesti:

$$Spetsiifilisus = \frac{TN}{TN + FP}.$$

Kasutades spetsiifilisuse valemit on võimalik leida karakteristik nimega “valepositiivsete määr” (*false positive rate, FPR)*, mis avaldub

$$FPR = 1 - Spetsiifilisus.$$

Järgnevad lõigud ROC-kõvera teemal põhinevad õpikul Tutz, G. (2012, lk 448-451), kui pole viidatud teisiti.

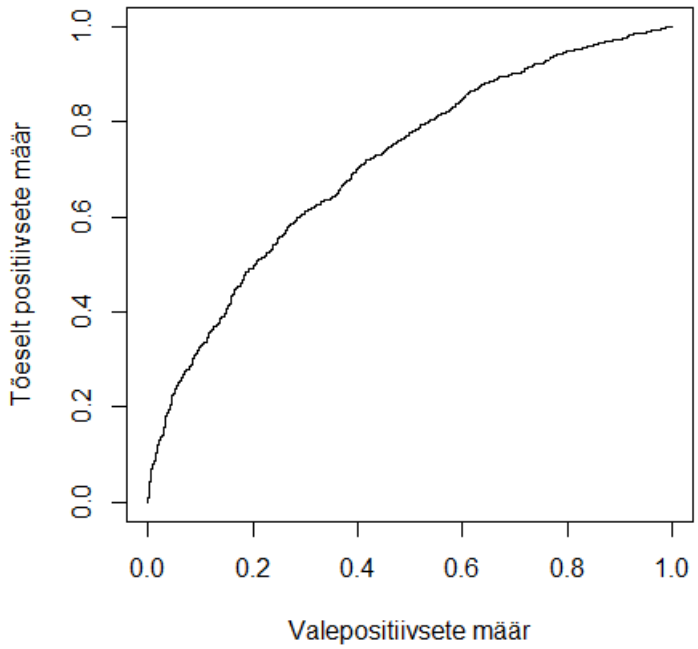
ROC-kõver on graafiline esitus, kus y-teljel on kujutatud tundlikkuse väärtused ning x-teljel väärtused valepositiivsete määr (FPR). Selle puhul on tegemist väga kasuliku vahendiga hindamaks ning võrdlemaks ennustusmudeleid. ROC kõvera pealt on hea näha, kuidas ennustusmudel suudab eristada tõeselt posiitivseid ja tõeselt negatiivseid juhte.

Seega, parim otsustusreegel on kõrge tundlikkusega ning madala FPRga. See reegel näeb ette, et tegelikult positiivsete hulgas enamus osutub ka mudeli järgi positiivseks ning tegelikult negatiivsete hulgas enamus osutub ka mudeli järgi negatiivseks.

Kui võrrelda näiteks kahte suvalist punkti ROC graafikul (vt all), siis punkt, mille TP on kõrgem ja FP madalam (y väärtus suurem ja x väiksem), on parem kui teine.

ROC-kõvera aluse pindala puhul on tegu ühe tuntud mudeli headuse näitajaga. See näitab, kui hästi on mudel suuteline ennustama klasside vahel – mida kõrgem AUC, seda paremini ennustab mudel nulle nullideks ja ühtesid ühtedeks. Lihtsustatult öeldes näitab AUC mudeli võimet õigesti ennustada sündmuse toimumise tõenäosust (Hosmer ja Lemeshow, 2000).

Joonisel nr 3 on toodud laenusaajate ROC-kõver. Selle kõvera alune pindala on 0,713, mistõttu võib seda pidada aktsepteeritavaks mudeliks. Otsus on tehtud lähtudes üldist otsustusreeglit AUC väärtuste kohta, kus AUC = 0,5 puhul on tegu prognoosivõimetu mudeliga ning AUC > 0,8 puhul on tegu juba väga hea või suurepärase mudeliga.



Joonis 3. Laenusaajate mudeli ROC-kõver

Kokkuvõte

Krediidiasutustel on väga oluline tunda oma klienti ning konkreetselt laenutoodete puhul on tähtis olla teadlik kliendi maksejõulisusest. Viimase uurimine hõlmab endas klientide jaotamist maksejõulisteks ja maksejõuetuteks klientideks. Selle protsessi läbiviimiseks on mitmeid võimalusi ning viise, kuid samas on kujunenud välja teatud mustrid.

Käesoleva uuringu eesmärgiks oli uurida, kas lisaks juba tuntud logistilise regressiooni uurimismeetoditele leidub ka teistsuguseid alternatiive. Selle tarbeks valiti välja 3 konkureerivat meetodit, kus uuritav tunnus on binaarsel kujul – *probit*, *c-log-log*, *cauchit*. Kõigi eeltoodud meetodite abil konstrueeriti mudel hindamaks laenusaaja maksejõulisust. Töös anti ka teoreetiline ülevaade kõigi nelja meetodi kohta.

Pärast esmaseid analüüse selgus, et parima hindamisvõimega mudeliks on siiski logistilise regressiooni mudel, mis on selles vallas juba pikalt kasutuselolev praktika. Kuigi alternatiivsete mudelite hindamisvõime ei küündinud samale tasemele, joonistus siiski kohati välja teatav erinevus statistiliselt oluliseks osutunud tunnuste vahel.

Seejärel jätkati täpsemat analüüsi *logit* mudeliga ning selles kasutatavate tunnustega tehti läbi mitmeid teisendusi eesmärgiga parandada mudeli prognoosivõimet. Selle käigus kasutati grupeerimist, pideva tunnuse lõikamist (*truncation)* ja tunnuse viimist 1/0 kujule. Samuti võeti mudeli konstrueerimisel arvesse multikollineaarsusi ja koosmõjusid. Mudeli testimise osas arvestati ka vigade ebasümmeetriaga. Sellise protseduuri tulemusena loodud mudelil oli 11 statistiliselt olulist tunnust ning selle Akaike' informatsioonikriteeriumi väärtuseks osutus 4120.

Seega on saadud tulemused üsna loogilised ja ootuspärased ning kuigi parimaks prognoosimeetodiks osutus ennast juba ammu tõestanud logistiline regressioon, siis sellegipoolest sai töö käigus katsetada sama protsessi ka teiste mudelite peal ning leida peamised seletused ja põhjused sellele, miks uuritavas valdkonnas siiski üks mudel niivõrd dominantne on.

Kasutatud kirjandus

1. Bouteille, S., Coogan-Pushner, D. (2012). *The Handbook of Credit Risk Management: Originating, Assessing, and Managing Credit Exposures*. John Wiley & Sons Inc.
2. *Fair Isaac Corporation*, kasutatud 07.05.2019. https://www.myfico.com/credit-education/credit-scores/
3. *Fair Isaac Corporation*, kasutatud 07.05.2019. https://www.myfico.com/credit-education/whats-in-your-credit-score/
4. Fawcett, T. (2005, lk 861-874). *An introduction to ROC analysis*. Pattern Recognition Letters, 27. doi:10.1016/j.patrec.2005.10.010
5. Fox, J. (2016). *Applied Regression Analysis and Generalized Linear Models.* SAGE Publications, Inc.
6. Hosmer, D. W., Lemeshow, S. (2000). *Applied Logistic Regression* (2nd Edition). New York: Wiley. http://dx.doi.org/10.1002/0471722146
7. Koenker, R., Yoon, J. (2009). *Parametric links for binary choice models: A Fisherian-Bayesian colloquy.* Journal of Econometrics.
8. Käärik, E. (2014). Andmeanalüüs II. Loengukonspekt. Tartu: Tartu Ülikool, matemaatika ja statistika instituut. http://dspace.ut.ee/bitstream/handle/10062/35401/AndmeanaluusII.pdf?sequence=1
9. Thomas, L. C., Edelman, D. B., Crook, J. N. (2002). *Credit Scoring and Its Applications.* Society for Industrial and Applied Mathematics.
10. Tutz, G. (2012). Alternative Binary Regression Models. *Regression for Categorical Data* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511842061

LISAD

Lisa 1. Logit mudeli programmiväljund.

```
> print(summary(fwd3),digits=3)

Call:
glm(formula = Status ~ Estate + Age + PaimentAlertsActive + Sum +
    Sex + PaimentAlertsClosed + Outcome + Children + Language,
    family = binomial(link = "logit"), data = algnedata)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-3.344  -1.224   0.617   0.872   1.757

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.094587   0.163255   -0.58  0.56233
Estate               0.609881   0.063371    9.62  < 2e-16 ***
Age                  0.023840   0.003395    7.02  2.2e-12 ***
PaimentAlertsActive -0.240193   0.044580   -5.39  7.1e-08 ***
Sum                 -0.000655   0.000173   -3.78  0.00016 ***
SexM                -0.359543   0.081040   -4.44  9.1e-06 ***
PaimentAlertsClosed -0.038116   0.016400   -2.32  0.02012 *
Outcome              0.000759   0.000182    4.18  2.9e-05 ***
Children            -0.129647   0.042875   -3.02  0.00250 **
Languagerus          0.198761   0.076018    2.61  0.00893 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4549.2  on 3799  degrees of freedom
Residual deviance: 4213.1  on 3790  degrees of freedom
AIC: 4233

Number of Fisher Scoring iterations: 4
```

Lisa 2. Probit mudeli programmiväljund.

```
> summary(fwd3)

Call:
glm(formula = Status ~ Estate + Age + PaimentAlertsActive + Sum +
    Sex + Outcome + PaimentAlertsClosed + Language + Children,
    family = binomial(link = "probit"), data = algnedata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6841  -1.2341   0.6225   0.8706   1.7636

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.0461369  0.0981209  -0.470 0.638209
Estate               0.3267691  0.0350205   9.331  < 2e-16 ***
Age                  0.0145310  0.0019989   7.270 3.60e-13 ***
PaimentAlertsActive -0.1458348  0.0268352  -5.434 5.50e-08 ***
Sum                 -0.0003854  0.0001031  -3.737 0.000186 ***
SexM                -0.2242221  0.0479884  -4.672 2.98e-06 ***
Outcome              0.0004656  0.0001056   4.409 1.04e-05 ***
PaimentAlertsClosed -0.0241963  0.0099374  -2.435 0.014897 *
Languagerus          0.1237714  0.0451571   2.741 0.006127 **
Children            -0.0728641  0.0256970  -2.836 0.004575 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4549.2  on 3799  degrees of freedom
Residual deviance: 4216.3  on 3790  degrees of freedom
AIC: 4236.3

Number of Fisher Scoring iterations: 5
```

Lisa 3. C-log-log mudeli programmiväljund.

```
> print(summary(fwd3),digits=3)

Call:
glm(formula = Status ~ Estate + Age + PaimentAlertsActive + Sum +
    Sex + Outcome + PaimentAlertsClosed + Language + Children,
    family = binomial(link = "cloglog"), data = algnedata)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-4.037  -1.249   0.633   0.875   1.693

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -3.65e-01   9.60e-02   -3.80  0.00014 ***
Estate               2.50e-01   2.96e-02    8.44  < 2e-16 ***
Age                  1.43e-02   1.87e-03    7.63  2.4e-14 ***
PaimentAlertsActive -1.57e-01   3.15e-02   -4.99  6.1e-07 ***
Sum                 -3.56e-04   9.90e-05   -3.59  0.00033 ***
SexM                -2.34e-01   4.52e-02   -5.17  2.4e-07 ***
Outcome              4.62e-04   9.54e-05    4.84  1.3e-06 ***
PaimentAlertsClosed -2.87e-02   1.05e-02   -2.74  0.00608 **
Languagerus          1.27e-01   4.27e-02    2.97  0.00302 **
Children            -5.83e-02   2.51e-02   -2.32  0.02010 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4549.2  on 3799  degrees of freedom
Residual deviance: 4222.1  on 3790  degrees of freedom
AIC: 4242

Number of Fisher Scoring iterations: 12
```

Lisa 4. Cauchit mudeli programmiväljund.

```
> print(summary(fwd3),digits=3)

Call:
glm(formula = Status ~ Estate + Age + PaimentAlertsActive + Sum +
    Sex + Children + Outcome + Language, family = binomial(link = "cauchit"),
    data = algnedata)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.482  -1.206   0.627   0.853   1.720

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.191347   0.157355   -1.22  0.22398
Estate               0.782628   0.091095    8.59  < 2e-16 ***
Age                  0.022581   0.003690    6.12  9.4e-10 ***
PaimentAlertsActive -0.228161   0.043552   -5.24  1.6e-07 ***
Sum                 -0.000561   0.000174   -3.23  0.00126 **
SexM                -0.328091   0.083630   -3.92  8.7e-05 ***
Children            -0.154153   0.042293   -3.64  0.00027 ***
Outcome              0.000684   0.000199    3.43  0.00061 ***
Languagerus          0.180882   0.076379    2.37  0.01787 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4549.2  on 3799  degrees of freedom
Residual deviance: 4230.3  on 3791  degrees of freedom
AIC: 4248

Number of Fisher Scoring iterations: 5

  '
```

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Martin Aasmäe,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Krediidiriski hindamisel kasutatavate mudelite võrdlus ühe Eesti laenuandmestiku näitel“, mille juhendajad on Meelis Käärik ja Kalev Pärna, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Martin Aasmäe
07.05.2019