

TARTU ÜLIKOOL
LOODUS- JA TÄPPISTEADUSTE VALDKOND
MATEMAATIKA JA STATISTIKA INSTITUUT

Annela Pajumets
**Krediidiriski hindamine valdavalt
kategoriaalsete andmete põhjal**
Kindlustus- ja finantsmatemaatika
Magistritöö (30 EAP)

Juhendaja: prof. Kalev Pärna

TARTU 2022

KREDIIDIRISKI HINDAMINE VALDAVALT KATEGORIAALSETE ANDMETE PÕHJAL

Magistritöö

Annela Pajumets

Lühikokkuvõte

Käesoleva magistritöö eesmärgiks on välja selgitada otstarbekad meetodid krediidiriski hindamiseks, kui argumenttunnused on valdavalt kategoriaalsed. Töös võrreldakse nelja erinevat prognoosimudelit – logistilist regressiooni, LASSO regressiooni, klassifitseerimispuud ning *gradient boosting* algoritmi. Töös kasutatav andmestik sisaldab infot väikelaenu saanud isikute kohta ning uuritavaks tunnuseks on laenu *staatus*, mis kirjeldab, kas laen on krediidiasutusele tagastatud või mitte.

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

Märksõnad: Krediidirisk, kategoriaalsed tunnused, logistiline regressioon, LASSO regressioon, klassifitseerimispuu, *gradient boosting*.

CREDIT SCORING PRIMARILY BASED ON CATEGORICAL DATA

Master thesis

Annela Pajumets

Abstract

The aim of this Master thesis is to find out methods which can be used for credit scoring when information about client is primarily categorical. Four different modelling methods are used and compared – logistic regression, LASSO regression, classification tree and *gradient boosting* algorithm. Data

set used in the thesis contains information about people who have received small loan. The target variable is loan *status*, which describes whether the loan is in default or not.

CERCS research specialisation: P160 Statistics, operations research, programming, financial and actuarial mathematics.

Key Words: Credit risk, categorical variables, logistic regression, LASSO regression, classification tree, *gradient boosting*.

Sisukord

Sissejuhatus	5
1 Ülevaade krediidiriskist ja -skooringu	6
2 Kasutatav metoodika	8
2.1 Logistiline regressioon	8
2.1.1 Mudeli parameetrite hindamine	10
2.2 LASSO regressioon	11
2.3 CART klassifitseerimispuu	12
2.4 <i>Gradient boosting</i>	14
2.5 <i>k</i> -kordne ristvalideerimine	17
2.6 Mudeli headuse näitajad	19
2.7 Hii-ruut test	20
3 Ülevaade andmestikust	22
3.1 Sõltumatud tunnused	23
3.1.1 Arvtunnused	24
3.1.2 Kategoriaalsed tunnused	28
3.2 Tunnuste informatiivsus	34
4 Mudelid	36
4.1 Mudelite ehitamine	36
4.2 Mudelite võrdlus	40
4.3 Tulemuste analüüs ja järeldused	44

Kokkuvõte	47
Kasutatud allikad	49
Lisa 1. Kordajad regressioonmudelites	50
Lisa 2. R-i kood	55

Sissejuhatus

Krediidiasutuste üheks peamiseks tegevusalaks on laenude väljastamine, mille eesmärgiks on intressi pealt kasumi teenimine. Kasumi maksimeerimisel mängib olulist rolli suurte kahjude vältimine, milleks laenu väljastavate krediidiasutuste jaoks on halvaks minevad laenud ehk laenud, mida krediidiasutusele täielikult tagasi ei maksta. Seetõttu on oluline enne laenu väljastamist klassifitseerida laenutaotlejad „headeks“ ja „halbadeks“ klientideks. Klassifitseerimine toimub läbi krediidiriski hindamise, mille käigus üldjuhul prognoositakse maksehäire tekkimise tõenäosust.

Krediidiriski hindamisel kasutatakse infot, mida laenutaotleja krediidiasutusele laenu taotlemise protsessi ajal annab. Saadav info võib olla nii arvuline (näiteks kuine sissetulek, kuine väljaminek) kui ka kategooriaalne (laenutaotleja haridustase, elukoht). Kategooriaalsete tunnuste modelleerimine ei ole aga nii lihtne. Esiteks, k -tasemelise ($k > 2$) kategooriaalse tunnuse kaasamine regressioonimudelisse tekitab mudelisse juurde $k - 1$ parameetrit, muutes mudeli keerulisemaks. Samuti tuleks mudeli prognoositäpsuse suurendamiseks kaasata kategooriaalsete tunnuste vahelised koosmõjud. Töö eesmärgiks on välja selgitada otstarbekad meetodid kategooriaalsete argumenttunnuste kasutamiseks krediidiriski hindamisel.

Töö koosneb neljast osast. Esmalt tutvustatakse krediidiriski ja -skooringu valdkonda. Seejärel tehakse ülevaade töös kasutatud meetoditest — logistilisest regressioonist, LASSO regressioonist, klassifitseerimispuust, *gradient boosting* algoritmist, k -kordsest ristvalideerimisest, hii-ruut testist ning mudeli headuse näitajatest, mida hiljem kasutatakse erinevate mudelite võrdlemiseks. Töö kolmandas osas on toodud andmestiku ülevaade ja kirjeldav statistika, samuti ülevaade argumenttunnuste informatiivsusest. Neljandas peatükis on esitatud erinevate klassifitseerimismeetodite võrdlus ning tulemuste analüüs ja järeldused.

Käesoleva magistritöö vormistamisel on kasutatud tekstitöötlusprogrammi $\text{\LaTeX}$. Andmete töötamiseks, kirjeldava analüüsi koostamiseks (sh. joonised) ning modelleerimiseks on kasutatud rakendustarkvara R.

1 Ülevaade krediidiriskist ja -skooringu

Krediidiriskiks nimetatakse võimalust saada kahju, mis tuleneb laenuvõtja suutmatusest laenu tagasi maksta või lepingulisi kohustusi täita. Üldjuhul viitab see riskile, kus krediidasutus või muu laenuandja ei pruugi tagasi saada laenu põhiosa ja intressi, mille tulemusena toimub rahavoogude katkemine. Lisaks suurenevad sellega ka laenu tagasisaamise kulud, näiteks inkassokulud. (The Investopedia Team, 2022)

Laenu väljastades ei ole krediidasutustel täpselt teada, kes laenuvõtjatest oma tagasimakse kohustused täidab ning kes makseraskustesse satub. Seetõttu võib korralik krediidiriski hindamine ja juhtimine vähendada kahju ulatust. Kõrgendatud krediidiriski saab maandada kõrgema intressimääraga, millest tulenevalt võib laenuvõtja intressimakseid käsitleda kui laenuandja tasu krediidiriski võtmise eest. (The Investopedia Team, 2022)

Krediidiskooringu kasutatakse krediidiriski hindamiseks. Täpsemalt kasutatakse krediidiskooringu selleks, et hinnata, kui tõenäoline on, et laenuvõtjal tekivad maksehäired (inglise keeles *default*) ehk teisisõnu hinnatakse maksehäire tekkimise tõenäosust (inglise keeles *probability of default*). Tõenäosust, et laenusaajal tekivad maksehäired, hinnatakse laenu taotlemise protsessi ajal ning selleks kasutatakse infot, mida laenutaotleja krediidasutusele annab. Erinevad krediidasutused on huvitatud erinevatest krediidiskooringu näitajatest. Näiteks ollakse huvitatud nii sellest, kui suur on tõenäosus, et laenusaaaja ei ole laenuperioodi lõpuks laenu tagastanud, kui ka sellest, kui suur on tõenäosus, et laenusaaaja on esimese aasta lõpuks kahe-kolme tagasimaksega maas. (Hand ja Henley, 1997)

Krediidasutused kasutavad krediidiskooringu, et otsustada, kes saab laenu ning kui riskantse kliendiga on tegemist. Laias laastus klassifitseeritakse krediidiskooringu abil laenutaotlejad kahte gruppi – „head“ ja „halvad“ kliendid. Selleks määratakse krediidiskooringu käigus hinnatud maksehäire tekkimise tõenäosusele lävend, kus

ühele poole lävendit jäävad „head“ laenutaotlejad, kellele antakse laenu ning teisele poole läve jäävad seega „halvad“ kliendid, kellele laenu ei väljastata. Krediidiskooringu saavad kasu nii laenuandja kasumi suurendamise ja kahjude vähendamise näol kui ka laenutaotleja finantsilise ülekoormuse vältimise näol. (Hand ja Henley, 1997)

Enne, kui statistilised krediidiskooringu meetodid kasutusele võeti, põhines krediidi andmise otsus subjektiivsel inimhinnangul. Tänapäeval on siiski laialdaselt levinud statistilised meetodid, mille eeliseks on võime käsitleda suurt tehingute arvu ning lisaks annavad need ka täpsema klassifikatsiooni võrreldes ekspertide subjektiivse hinnanguga. Enimlevinud statistilised meetodid krediidiriski hindamisel on järgmised: diskriminantanalüüs, lineaarne regressioon, logistiline regressioon, *probit* analüüs, lähima naabri meetod, erinevad otsustuspuud, Markovi mudelid (inglise keeles *Markov chain models*), närvivõrgud ning matemaatilise programmeerimise meetod. (Hand ja Henley, 1997)

Üldiselt ei eksisteeri ühte ja ainsat parimat meetodit krediidiriski hindamisel. Meetodi headus sõltub mitmest erinevast aspektist – andmete struktuurist, mudelis kasutatud tunnustest ning sellest, mil määral on võimalik klasse nende tunnuste abil eristada. Siiski peetakse atraktiivsemaks meetodeid, mis on nii laenuandjatele kui ka laenutaotlejatele kergesti mõistetavad (näiteks regressioonmudelid või otsustuspuud) ning vähem atraktiivsemaks meetodeid, mis on sisulised mustad kastid (näiteks närvivõrgud). (Hand ja Henley, 1997)

2 Kasutatav metoodika

Käesolevas peatükis antakse ülevaade töös kasutatud metoodikatest. Esmalt tutvustatakse logistilist ja LASSO regressiooni, klassifitseerimispuud ning *gradient boosting* meetodit. Seejärel kirjeldatakse ristvalideerimise algoritmi optimaalsete parameetrite leidmiseks mudelites. Viimaks kirjeldatakse mudelite headuse näitajaid, mida kasutatakse hiljem mudelite võrdlemiseks.

2.1 Logistiline regressioon

Binaarse sõltuva tunnuse modelleerimisel on laialdaselt kasutatud logistilist regressioonmudelit, mis on tuntud ka logit mudeli nime all. Logistilises regressioonis omab sõltuv tunnus kahte väärtust, mis on tavaliselt kodeeritud vastavalt 0-1. Nii on võimalik kodeerida mis tahes kaht väärtust ning sellisel kodeerimise viisil on ka oma eelis. Nimelt, sõltuva binaarse tunnuse keskmine on võrdne väärtuse 1 osakaaluga ning seda saab interpreteerida kui edu tõenäosust. (Pampel, 2000)

Binaarse uuritava tunnuse y_i , $i = 1, \dots, n$ korral modelleeritakse sündmuse toimumise tõenäosust. Tõenäosuse definitsioonist on teada, et tõenäosus ei saa ületada väärtust 1 ja olla väiksem kui 0, mistõttu ei ole võimalik seda modelleerida kasutades lineaarset regressiooni. Logistilises regressioonis kasutatakse prognoositava tõenäosuse leidmiseks logit seosefunktsiooni, mis avaldub järgneval kujul:

$$\eta_i = \ln \frac{\pi_i}{1 - \pi_i}, \quad (1)$$

kus $\pi_i = \pi(\mathbf{x}_i) = P(y_i = 1 | \mathbf{x}_i)$ on i -nda vaatluse edu tõenäosus. Valemist (1) selgub, et logit seosefunktsioon hõlmab endas kahte sammu. Esiteks leitakse sündmuse esinemise šanss, leides π_i ja $1 - \pi_i$ suhe. Seejärel võetakse leitud šansist naturaallogaritm. (Pampel, 2000)

Logistilise regressioonmudeli üldkuju on võimalik avaldada järgmiste omavahel ek-

vivalentsete valemitega:

$$\ln \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} = \mathbf{x}^T \boldsymbol{\beta}, \quad (2)$$

$$\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} = \exp(\mathbf{x}^T \boldsymbol{\beta}), \quad (3)$$

$$\pi(\mathbf{x}) = \frac{\exp(\mathbf{x}^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}^T \boldsymbol{\beta})}, \quad (4)$$

kus

$$\eta(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta} = \beta_0 + x_1 \beta_1 + \dots + x_p \beta_p.$$

Parameetrid $x_1, \dots, x_p$ on mudelis seletavad ehk sõltumatud tunnused ning parameetreid $\beta_0, \beta_1, \dots, \beta_p$ nimetatakse kordajateks. (Tutz, 2012)

Logistilises regressioonis võivad seletavad tunnused olla nii arvulised kui ka kategoriaalsed. Nendest viimaste puhul saab mudeli üldkuju kirja panna järgmiselt. Olgu x_A ja x_B kategoriaalsed tunnused $A \in \{1, \dots, I\}$ ja $B \in \{1, \dots, J\}$ tasemetega, siis

$$\begin{aligned} \ln \frac{\pi(A=i, B=j)}{1 - \pi(A=i, B=j)} &= \beta_0 + x_{A(1)} \beta_{A(1)} + \dots + x_{A(I-1)} \beta_{A(I-1)} + \\ &+ x_{B(1)} \beta_{B(1)} + \dots + x_{B(J-1)} \beta_{B(J-1)}, \end{aligned}$$

kus $x_{A(i)}$, $x_{B(j)}$ on 0-1 binaarsed tunnused ning $\beta_{A(i)}$ on muutus I -nda taseme suhtes ja $\beta_{B(j)}$ on muutus J -nda taseme suhtes. Seega vabaliige β_0 vastab taemetele $A = I$, $B = J$ ehk teisisõnu

$$\ln \frac{\pi(A=I, B=J)}{1 - \pi(A=I, B=J)} = \beta_0.$$

Tasemeid $A = I$, $B = J$ nimetatakse ka võrdluskategooriateks (inglise keeles *reference category*). (Tutz, 2012)

Logistilisse regressioonmudelisse saab kaasata ka kahe või enam kategoriaalse tunnuste vahelisi koosmõjusid. Näiteks tunnuste x_A , $A \in \{1, \dots, I\}$ ja x_B , $B \in \{1, \dots, J\}$ koosmõju modelleerides kaasatakse mudelisse kõik binaarsete tunnus-

te korrutised $x_{A(i)} \cdot x_{B(j)}$. (Tutz, 2012)

Nagu eespool näidatud, siis logistilist regressiooni saab kirja panna mitmel erineval kujul. Valem (4) näitab sündmuse esinemise tõenäosuse ja sõltumatute tunnuste vahelist seost. Valemis (3) on toodud sündmuse esinemise šansi seos sõltumatute tunnustega. Valemist (2) on aga näha, et logit seose ja sõltumatute tunnuste vahel eksisteerib lineaarne seos. (Tutz, 2012)

2.1.1 Mudeli parameetrite hindamine

Mudeli parameetrite $\beta^T = (\beta_0, \beta_1, \dots, \beta_p)$ hinnagud leitakse enamasti kasutades suurima tõepära meetodit. Olgu $(y_i, \mathbf{x}_i)$, $i = 1, \dots, n$, $y_i \in \{0, 1\}$ sõltumatud vaatlused ning $\pi_i = \pi(\mathbf{x}_i) = P(y_i = 1 | \mathbf{x}_i)$ edu tõenäosus i -ndal vaatlusel. Siis tõepärafunktsioon $y_i | \mathbf{x}_i$, $i = 1, \dots, n$ jaoks avaldub järgmiselt:

$$L(\beta) = \prod_{i=1}^n \pi(\mathbf{x}_i)^{y_i} (1 - \pi(\mathbf{x}_i))^{1-y_i}.$$

Parameetrite β suurima tõepära hinnanguks on väärtused $\hat{\beta}$, mis maksimeerivad tõepärafunktsiooni β suhtes. (Tutz, 2012)

Kuna logistilises regressioonimudelil avaldub π_i ka kujul (4), siis tihtipeale parameetrite β hinnangute saamiseks maksimeeritakse tõepärafunktsiooni asemel hoopis log-tõepärafunktsioon:

$$l(\beta) = \ln(L(\beta)) = \sum_{i=1}^n l_i(\beta) = \sum_{i=1}^n y_i \ln(\pi(\mathbf{x}_i)) + (1 - y_i) \ln(1 - \pi(\mathbf{x}_i)) \quad (5)$$

Log-tõepärafunktsiooni maksimeerimise tulemusena leitavad väärtused $\hat{\beta}$ on samad, mis tõepärafunktsiooni maksimeerimisest saadavad väärtused. (Tutz, 2012)

2.2 LASSO regressioon

Vahel võib seletavate tunnuste arv mudelis olla väga suur ning sellistel juhtudel on tunnuste arvu vähendamine oluline, et vältida mudeli ületreenitust. Ajalooliselt on kasutatud selleks sammregressiooni, kuid sammregressioon diskreetse protsessina on tundlik. Sammregressiooni tulemuse võib määrata juba väike muutus andmetes. Sammregressioonile alternatiiv seletavate tunnuste valimisel on reguleerimise meetod (inglise keeles *regularization method*), mis lisab parameetrite hindamisel log-tõepärafunktsioonile karistusliikme (inglise keeles *penalty term*). (Tutz, 2012)

Üheks selliseks reguleerimise meetodiks on LASSO (*Least Absolute Shrinkage and Selection Operator*) regressioon, mis kasutab L_1 regulatsiooni. Log-tõepärafunktsiooni karistusliikme lisamise tulemusena võtab valemis (5) esitatud log-tõepärafunktsioon järgmise kuju:

$$\begin{aligned} l_{\lambda}^L(\boldsymbol{\beta}) &= \sum_{i=1}^n l_i(\boldsymbol{\beta}) - \lambda \sum_{j=1}^p |\beta_j| \\ &= \sum_{i=1}^n y_i \ln(\pi(\mathbf{x}_i)) + (1 - y_i) \ln(1 - \pi(\mathbf{x}_i)) - \lambda \sum_{j=1}^p |\beta_j|, \end{aligned}$$

mille maksimeerimise tulemusena saadaksegi parameetrite $\beta_0, \beta_1, \dots, \beta_p$ hinnangud LASSO regressioonis. (Pereira, Basto ja Silva, 2015)

Piisavalt suure λ korral võrdsustuvad osade parameetrite hinnangud $\hat{\beta}_1, \dots, \hat{\beta}_p$ nulliga, mille tulemusena jääb lõplikusse mudelisse vähem kordajaid ning mudel muutub lihtsamini interpreteeritavaks. Seega saab LASSO regressiooni kasutada ka tunnuste valimisel. (Pereira, Basto ja Silva, 2015)

Kuna LASSO regressioon tuleb toime suure arvu seletavate tunnustega mudelis, siis käesolevas töös kaasati LASSO regressioonmudelisse lisaks seletavate tunnuste peamõjudele ka kõikide kategoriaalsete seletavate tunnuste paarikaupa koosmõjud (korrutistunnused). Koosmõjud, mille parameetrid LASSO mudelis osutusid nullist

erinevaks, kaasati hiljem logistilisse regressiooni. Seega kasutati LASSO regressiooni antud töös ka oluliste kategooriaalsete tunnuste vaheliste koosmõjude valimisel.

2.3 CART klassifitseerimispuu

CART (*Classification And Regression Tree*) on mitteparameetriline otsustuspuul põhinev meetod, mis loob olenevalt uuritavast tunnusest regressiooni- või klassifitseerimispuu. Käesolevas töös ollakse huvitatud klassifitseerimispuust binaarse (0/1) uuritava tunnuse korral. Olgu $\mathbf{x} = (x_1, \dots, x_p)$ seletavate tunnuste vektor, mille võimalike väärtuste ruumi nimetatakse prognoosiruumiks (inglise keeles *predictor space*). Klassifitseerimispuu konstrueerimine koosneb kahest peamisest sammust (James *et al.*, 2013):

1. Prognoosiruum jagatakse J -ks eristuvaks ja mittelõikuvaks piirkonnaks $R_1, \dots, R_J$;
2. Vaatlused, mis langevad piirkonda R_j , saavad sama prognoosi $\hat{y}_{R_j}$, milleks on kõige sagedamini esinev uuritava tunnuse klass (0/1) selles piirkonnas.

Piirkondadeks jagamiseks (punktis 1) kasutatakse ülalt alla lähenemist, mida tuntakse ka kui rekursiivne binaarne tükeldamine (inglise keeles *recursive binary splitting*). Lähenemine on ülalt alla ja binaarne, kuna see algab puu tipust, kus kõik vaatlused kuuluvad ühte piirkonda ning seejärel iga tükeldusega tekib kaks uut oksa allapoole. Puu ehitamise igal etapil tehakse tükeldus nii, et see oleks optimaalne just selles etapis, kuid ei vaadata ettepoole ning ei valita tükeldust, mis võiks viia parema puuni mõnes järgnevas etapis. Tükeldamise tagajärjel saadud lõplike piirkondasid $R_1, \dots, R_J$ nimetatakse puu lõplikeks sõlmedeks või lehtedeks ning punkte, kus toimub kriteeriumi järgi puu hargnemine, nimetatakse sisemisteks sõlmedeks. (James *et al.*, 2013)

Rekursiivne binaarne tükeldamine toimub teatud kriteeriumite alusel. Tükelduse tegemiseks valitakse kriteeriumi abil tunnus x_l ja määratakse lävend t nii, et prog-

noosiruum jagatakse kaheks lõikumatuks piirkonnaks R_1 ja R_2 . Klassifitseerimispuu korral on selliseks kriteeriumiks funktsioon $\phi : [0, 1] \rightarrow \mathbb{R}^+$, mille argument iga R_j korral on:

$$\hat{p}(R_j) := \hat{p}_j,$$

mis tähistab klassist 1 pärit vaatluste osakaalu piirkonnas R_j . Seega on üldiselt vaja leida x_l ja länd t , mille korral:

$$\min_{t,l} \left[n_1 \phi \left(\hat{p}(R_1(t,l)) \right) + n_2 \phi \left(\hat{p}(R_2(t,l)) \right) \right],$$

kus n_1 ja n_2 on vastavalt vaatluste arvud piirkondades R_1 ja R_2 . Väärtused n_1 ja n_2 sõltuvad argumentidest x_l ja t . Funktsiooni ϕ nimetatakse *impurity function*, milleks käesolevas töös on Gini indeks, mis avaldub järgmiselt (Lember, 2021):

$$\phi(\hat{p}_j) = 2\hat{p}_j(1 - \hat{p}_j).$$

Gini indeks mõõdab koguvareeruvust üle uuritava tunnuse klasside. Näitaja omandab väikese väärtuse, kui suurus $\hat{p}_j$ on lähedane 0-le või 1-le. Sel põhjusel nimetatakse Gini indeksit ka lehe „puhtuse“ mõõdikuks. Väike Gini indeksi väärtus näitab, et tegemist on „puhta“ lehega, kus üks uuritava tunnuse klass domineerib. (James *et al.*, 2013)

Pärast parima x_l ja t leidmist jagatakse prognoosiruum kaheks. Seejärel sama protsessi korratakse, tükeldades mõni eelnevas sammus saadud alampiirkond kaheks. Protsess jätkub kuni peatumiskriteeriumi saavutamiseni, milleks võib olla näiteks minimaalne vaatluste arv ühes piirkonnas. (James *et al.*, 2013)

Eelpool kirjeldatud protsessi tulemusena võidakse luua liiga keeruline klassifitseerimispuu, mille täpsus treeningandmete korral on hea, kuid mis tõenäoliselt sobitab andmed üle ning seega viib kehva täpsuseni testandmete korral. Parema tulemus testandmete jaoks saadakse, kui kasvatatakse väga suur puu T_0 , mida seejärel pügatakse väiksemaks, saades alampuu T . Üldjuhul on eesmärgiks leida selline alampuu

T , mis prognoosib testandmeid kõige täpsemini. (James *et al.*, 2013)

Klassifitseerimispuudel on nii häid omadusi kui ka mitmeid puudusi. Ühelt poolt on tulemusi võimalik graafiliselt esitada ning need on väga lihtsasti interpreteeritavad. Öeldakse isegi, et otsustuspuud peegeldavad inimeste mõtlemist otsustamisel. Lisaks saab meetod väga hästi hakkama kategoriaalsete seletavate tunnustega. Samas ei ole puude prognoositäpsus nii hea võrreldes mõne regressioonmudeliga ning nad ei pruugi olla kuigi stabiilsed ehk teisisõnu lõplikud tulemused võivad olla mõjutatud juba väikesest muutusest andmestikus. (James *et al.*, 2013)

2.4 *Gradient boosting*

Järgnev alapeatükk põhineb raamatul „The Elements of Statistical Learning: Data Mining, Inference, and Prediction“ (Hastie, Tibshirani ja Friedman, 2009).

Gradient boosting on otsustuspuudel põhinev meetod, mille idee seisneb nõrkade klassifikaatorite kombineerimisel. Nõrgaks klassifikaatoriks nimetatakse klassifikaatorit, mille veamäär on vaid veidi parem kui juhuslikul arvamisel. *Gradient boosting* kuulub võimendusmeetodite (inglise keeles *boosting methods*) kategooriasse, mis õpivad iteratiivselt nõrgalt klassifikaatorilt, et luua tugev mudel.

Nagu eelnevas alapeatükis kirjeldatud (alapeatükis 2.3), siis CART otsustuspuu jagab ennustajate ruumi mittelõikuvateks piirkondadeks $R_1, \dots, R_J$. Seejärel igale sellisele piirkonnale määratakse konstant, olgu selleks γ_j , $j = 1, \dots, J$. Seega saab prognoosireegli välja kirjutada järgmiselt:

$$\mathbf{x} \in R_j \Rightarrow f(\mathbf{x}) = \gamma_j.$$

Gradient boosting algoritmi üheks osaks on kaofunktsioon $l(y, \gamma)$ (inglise keeles *loss function*), mis binaarse ($k \in \{0, 1\}$) klassifitseerimisülesande korral on järgmine:

$$l(y, \gamma) = -[y \ln(\gamma) + (1 - y) \ln(1 - \gamma)], \quad (6)$$

kus y on uuritava tunnuse realiseerunud väärtus ning π on $y = 1$ esinemise tõenäosus. On teada, et tõenäosus π avaldub log-šansside $\gamma = \ln(\frac{\pi}{1-\pi})$ kaudu järgmiselt:

$$\pi = \frac{\exp(\gamma)}{1 + \exp(\gamma)}.$$

Seega on võimalik valemis (6) toodud kaofunktsioon avaldada järgmiselt:

$$\begin{aligned} l(y, \gamma) &= -[y \ln(\pi) + (1 - y) \ln(1 - \pi)] \\ &= -y \ln(\pi) - \ln(1 - \pi) + y \ln(1 - \pi) \\ &= -y[\ln(\pi) - \ln(1 - \pi)] - \ln(1 - \pi) \\ &= -y \ln\left(\frac{\pi}{1 - \pi}\right) - \ln(1 - \pi) \\ &= -y\gamma - \ln\left(1 - \frac{\exp(\gamma)}{1 + \exp(\gamma)}\right) \\ &= -y\gamma - \ln\left(\frac{1 + \exp(\gamma) - \exp(\gamma)}{1 + \exp(\gamma)}\right) \\ &= -y\gamma - \ln(1) + \ln(1 + \exp(\gamma)) \\ &= -y\gamma + \ln(1 + \exp(\gamma)). \end{aligned} \tag{7}$$

Olgu andmestikus n arv vaatlusi ning iga vaatluse kohta p seletavat tunnust: $(\mathbf{x}_i, y_i)$, kus $\mathbf{x}_i = (x_{i1}, \dots, x_{ip}) \forall i = 1, \dots, n$. Järgnevalt antakse ülevaade *gradient boosting* algoritmist, kui uuritav tunnus on binaarne ning selle võimalikud väärtused on $y_i \in \{0, 1\}$.

1. Esmalt leitakse optimaalne konstantne mudel, milleks tavaliselt on ühe lehega puu:

$$f_0(\mathbf{x}) = \arg \min_{\gamma} \sum_{i=1}^n l(y_i, \gamma).$$

kus $l(y_i, \gamma)$ on valemis (7) toodud kaofunktsioon. Seega on vaja leida selline γ , mis minimeeriks summa.

2. Iga $m = 1, \dots, M$ korral:

(a) Leitakse $y_i = 1$ esinemise tõenäosus:

$$\pi_i = \pi(\mathbf{x}_i) = \frac{\exp(f(\mathbf{x}_i))}{1 + \exp(f(\mathbf{x}_i))},$$

kus $f = f_{m-1}$.

(b) Iga $i = 1, \dots, n$ jaoks leitakse negatiivse gradiendi komponent:

$$r_{im} = - \left[\frac{\partial l(y_i, f(\mathbf{x}_i))}{\partial f(\mathbf{x}_i)} \right] = y_i - \pi(\mathbf{x}_i),$$

kus $f = f_{m-1}$.

(c) Eelmises punktis leitud väärtustele r_{im} sobitatakse regressioonipuu, mis jagab prognoosiruumi lõikumateks piirkondadeks R_{jm} , $j = 1, \dots, J_m$.

(d) Iga piirkonna R_{jm} , $j = 1, \dots, J_m$ jaoks leitakse:

$$\gamma_{jm} = \arg \min_{\gamma} \sum_{\mathbf{x}_i \in R_{jm}} l(y_i, f_{m-1}(\mathbf{x}_i) + \gamma).$$

(e) Seejärel leitakse väärtus $f_m(\mathbf{x})$:

$$f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \nu \sum_{j=1}^{J_m} \gamma_{jm} I(\mathbf{x} \in R_{jm}),$$

kus ν on õppimiskiiruse parameeter (inglise keeles *shrinkage*).

3. Lõpuks väljastab algoritm:

$$\hat{f}(\mathbf{x}) = f_M(\mathbf{x}).$$

Lõpliku prognoosi $\hat{y}_i$, $i = 1, \dots, n$ saamiseks leitakse tõenäosused

$$\hat{\pi}_1(\mathbf{x}_i) = \frac{\exp(\hat{f}(\mathbf{x}_i))}{1 + \exp(\hat{f}(\mathbf{x}_i))},$$

$$\hat{\pi}_0(\mathbf{x}_i) = 1 - \hat{\pi}_1(\mathbf{x}_i),$$

kus $\hat{\pi}_1(\mathbf{x}_i)$ ja $\hat{\pi}_0(\mathbf{x}_i)$ on vastavalt klassi 1 ja klassi 0 prognoositud esinemise tõenäosused i -ndal vaatlusel. Seejärel leitakse $\hat{y}_i$ kasutades reeglit:

$$\hat{y}_i = \arg \max_k \hat{\pi}_k(\mathbf{x}_i),$$

kus $k = 0, 1$.

Gradient boosting algoritmis on kolm parameetrit – regressioonipuu suurus J , iteratsioonide ehk puude arv M ning õppimiskiirus ν . Üldjuhul on algoritmis kasutatavate puude suurus pigem väike ning levinud on, et kõikidest iteratsioonides sobitatakse ühesuguse suurusega puud. Iga iteratsiooniga on võimalik mudeli täpsust treeningandmetel parandada, seega piisavalt suure M korral on mudeli viga treeningandmetel suvaliselt väike. Samas mudeli liiga hea sobitamine treeningandmetega võib viia mudeli ületreenimiseni, mistõttu on oluline leida optimaalne iteratsioonide arv algoritmis. Õppimiskiirus ν , $0 < \nu < 1$ määrab iga iteratsiooni panuse, kusjuures väiksem ν väärtus toob kaasa suurema iteratsioonide arvu algoritmis. Käesolevas töös leitakse optimaalne parameetrite komplekt k -kordse ristvalideerimise abil.

2.5 k -kordne ristvalideerimine

Töös rakendatavate mudelite algoritmid kasutavad mitmeid parameetreid, mis mõjutavad mudeli lõplikku kuju ning prognoositäpsust. Sellisteks parameetriteks on näiteks klassifitseerimispuu suurus või karistusliikme parameeter λ LASSO regressioonis. Optimaalsete parameetrite leidmiseks kasutatakse sageli k -kordset ristvalideerimist.

k -kordse ristvalideerimise idee seisneb selles, et andmestik jagatakse juhuslikult ligikaudu k võrdse suurusega gruppiks. Seejärel treenitakse mudel $k-1$ grupi andmetel ning k -ndat gruppi andmeid kasutatakse mudeli valideerimiseks. Valideerimisel lei-

takse mudeli headuse näitajad, milleks LASSO regressiooni korral on ROC kõvera alune pindala (AUC) ning klassifitseerimispuu ja *gradient boosting* korral mudeli täpsus (ACC). Seda protsessi korratakse k korda, mil iga kord kasutatakse erinevat gruppi valideerimisandmestikuna. Protsessi tulemusena leitakse k testviga, mille keskmistamisel leitaksegi k -kordne ristvalideerimisviga. (James *et al.*, 2013)

LASSO regressiooni puhul saab ristvalideerimisvea arvutusevalemi kirja panna järgmiselt:

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^k AUC_i$$

ning klassifitseerimispuu ja *gradient boosting* korral järgmiselt:

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^n ACC_i,$$

kus $ACC_i = I(y_i = \hat{y}_i)$.

Praktikas kasutatakse tavaliselt 5- või 10-kordset ristvalideerimist ($k = 5$ või $k = 10$) (James *et al.*, 2013). Seetõttu on käesolevas töös LASSO regressiooni ja klassifitseerimispuu optimaalsete parameetrite leidmiseks kasutatud just 10-kordset ristvalideerimist ning *gradient boosting* meetodi optimaalse parameetrite komplekti leidmiseks 5-kordset ristvalideerimist.

LASSO regressioonmudeli parameetri λ valimiseks hinnatakse mudeli ristvalideerimisviga kasutades erinevaid λ väärtusi. Seejärel valitakse λ , mille korral ristvalideerimisviga on väikseim. Lõpuks hinnatakse mudel uuesti kasutades kõiki vaatlusi andmestikus ning ristvalideerimise abil leitud λ väärtust. (James *et al.*, 2013)

Analoogiliselt leitakse optimaalsed mudelite parameetrid ja lõplikud mudelid ka klassifitseerimispuu ja *gradient boosting* korral.

2.6 Mudeli headuse näitajad

Käesolevas töös on kõigi mudelite puhul headuse näitajatenä kasutatud mudeli täpsust, tundlikkust ja spetsiifilisust. Lisaks on logistilisele ja LASSO regressioonimudelile leitud ka ROC kõver alune pindala (AUC). Käesolevas peatükis tehakse ülevaade eelpool mainitud mudeli headuse näitajatest.

Üheks võimaluseks, et hinnata mudelite prognoositäpsust, on kasutada klassifitseerimistabelit. Klassifitseerimistabel on risttabel, kus on ristatud binaarse uuritava tunnuse tegelikud väärtused ning mudeli hinnangud. Logistilise regressiooni ja LASSO regressiooni puhul tuleb määrata eralduspiir (lävend) c , mille vastu mudelist hinnatud tõenäosust võrreldaks. Kui hinnatud tõenäosus ületab lävendi c , siis võrdsustatakse mudeli prognoos väärtusega 1 ning vastasel juhul saab väärtuseks 0. (Lemeshow, Sturdivant ja Hosmer Jr., 2013)

Tabel 1: Klassifitseerimistabel.

Ennustus/Tegelik	0	1	Kokku
0	TN	FN	N*
1	FP	TP	P*
Kokku	N	P	n

Tabelis 1 on toodud näide klassifitseerimistabelist. Klassifitseerimistabeli põhjal on võimalik leida nii mudeli täpsus, tundlikkus, spetsiifilisus kui ka valepositiivsus ja valenegatiivsus. Nimetatud näitajate arvutusvalemid on järgmised (James *et al.*, 2013):

- Täpsus = $\frac{TN+TP}{n}$;
- Tundlikkus = $\frac{TP}{P}$;
- Spetsiifilisus = $\frac{TN}{N}$;
- Valepositiivsus = $\frac{FP}{N} = 1 - \text{Spetsiifilisus}$;
- Valenegatiivsus = $\frac{FN}{P} = 1 - \text{Tundlikkus}$.

Teine laialdaselt kasutatud logistilise regressioonimudeli (ja ka LASSO regressiooni) headuse näitajaid on ROC (*Receiver Operating Characteristic*) kõvera alune pindala (AUC). ROC kõvera konstrueerimisel leitakse kõikidele eralduspiiridele c tundlikkus ja valepositiivsus, millest esimene kuvatakse vertikaalteljele ja teine horisontaalteljel. ROC kõvera alune pindala näitab mudeli võimet eristada huvi-pakkuvad sündmused sündmustest, mis huvi ei paku (käesolevas töös häid laene halbadest laenudest), kusjuures näitaja väärtus jääb vahemikku 0,5-1. ROC kõvera aluse pindala põhjal on võimalik mudeli prognoositäpsus jagada järgnevatesse klassidesse (Lemeshow, Sturdivant ja Hosmer Jr., 2013):

- $AUC = 0,5$, mudeli prognoositäpsus on puudulik;
- $0,5 < AUC < 0,7$, mudeli prognoositäpsus on halb;
- $0,7 \leq AUC < 0,8$, mudeli prognoositäpsus on aktsepteeritav;
- $0,8 \leq AUC < 0,9$, mudeli prognoositäpsus on väga hea;
- $AUC \geq 0,9$, mudeli prognoositäpsus on suurepärase.

2.7 Hii-ruut test

Järgnev alapeatükk põhineb raamatu „Categorical Data Analysis“ lehekülgedel 18 ja 75-76 (Agresti, 2013).

Hii-ruut testiga on võimalik iseloomustada kahe tunnuse vahelist seost. Testi idee seisneb realiseerunud ja oodatavate sageduste võrdlemisel. Hii-ruut testi korral väidab nullhüpotees, et kaks tunnust on omavahel sõltumatud, kusjuures hüpoteesi saab kirja panna järgmiselt:

$$H_0 : \mu_{ij} = E(n_{ij}) = n\pi_{ij} = n\pi_{i+}\pi_{+j}, i = 1, \dots, I, j = 1, \dots, J, \quad (8)$$

kus μ_{ij} tähistab oodatavaid sagedusi ning suurused π_{i+} ja π_{+j} vastavalt i -nda rea ja j -nda veeru marginaalseid tõenäosusi. Sageli ei ole aga marginaalsed tõenäosused

teada ning seetõttu hinnatakse need valimist:

$$\hat{\pi}_{i+} = \frac{n_{i+}}{n}, \quad (9)$$

$$\hat{\pi}_{+j} = \frac{n_{+j}}{n}. \quad (10)$$

Valemeid (8), (9) ja (10) omavahel kombineerides saab valimist hinnatud oodatava sageduse kirja panna järgmiselt:

$$\hat{\mu}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = \frac{n_{i+}n_{+j}}{n}.$$

Hii-ruut test kasutab Pearsoni hii-ruut statistikut, mis leitakse järgmiselt:

$$X^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}.$$

Statistiku X^2 asümptootiline jaotus nullhüpoteesi eeldusel on hii-ruut jaotusega, mille vabadusastmete arv on järgnev:

$$df = (IJ - 1) - (I - 1) - (J - 1) = (I - 1)(J - 1).$$

Nullhüpoteesi kehtimise korral on realiseerunud sagedused n_{ij} lähedased oodatavate sagedustega $\hat{\mu}_{ij}$, millest annab aimu väike statistiku väärtus ning suur p-väärtus.

3 Ülevaade andmestikust

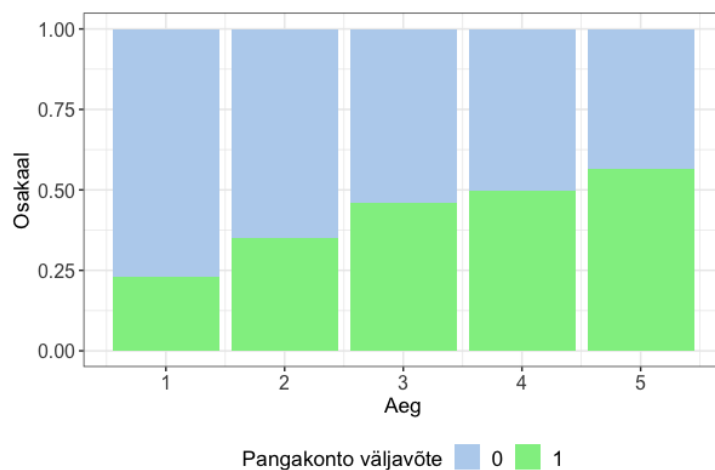
Käesoleva magistritöö praktilises osas on kasutatud andmestikku, mis sisaldab infot väikelaenu saanud isikute kohta. Täpsemalt on andmestikus toodud laenu taotlemise hetkel laenusaaaja kohta kogutud info ning samuti see, kas klient maksis saadud laenu tagasi või ei maksnud. Andmestikus toodud laenud pärinevad umbes kümne aasta tagusest ajast.

Algne andmestik sisaldas 6000 vaatlust ning 33 erinevat tunnust, kuid peale andmete korrastamist koosnes andmestik 5975 vaatlusest ning 17 tunnusest. Andmete korrastamine koosnes peamiselt kahest sammust. Esmalt kaotati andmestikust need tunnused, mis põhinesid laenusaaaja pangakonto väljavõttel, sest nendel tunnustel oli üle pooltel vaatlustel (täpsemalt 3475 real 6000-st) puuduvad väärtused. Nende tunnuste põhjal loodi uus binaarne tunnus *pangakontoinfo*, kus 1 tähistas, et laenusaaaja on krediitiasutusele edastanud väljavõtte oma pangakontost ning 0 tähistas vastupidist. Teisena eemaldati andmestikust puuduvaid väärtusi sisaldavad read ning seejärel saadi andmestik, mida kasutati mudelite ehitamisel. Saadud andmestikus olid järgmised tunnused: laenu *staatus*, laenusaaaja *sugu*, laenusaaaja *vanus* aastates, *maakond*, kus laenusaaaja elab, laenusaaaja peamine *emakeel*, *laenusumma* eurodes, *laenuperiood* päevades, laenusaaaja igakuine *sissetulek* eurodes, laenusaaaja igakuine *väljaminek* eurodes, laenusaaaja perekonnaseis, laenusaaaja *harridustase*, laenusaaaja *töökogemus*, laenusaaaja *laste arv*, *kinnisvara* omamine, *aktiivne maksehäire*, *suletud maksehäire* ning *pangakontoinfo*.

Töös on uuritavaks ehk sõltuvaks tunnuseks laenu *staatus*, kus vastusevariant 1 tähistab, et laen on krediitiasutusele tagastatud ning vastusevariant 0 tähistab, et laen ei ole tagasi makstud ehk teisisõnu laen on halvaks läinud. Andmestikust moodustavad „hea“ staatusega laenud 71,7% (4283) ning „halva“ staatusega laenud 28,3% (1692). Seega ei ole andmestik ideaalses tasakaalus.

Töö autorile oli teada, et andmestikus asuvad vaatlused ajalises järjekorras ning

seetõttu oli võimalik uurida, kas ja kuidas muutuvad pangakonto väljavõtte põhinevate tunnuste puuduvad väärtused ajas. Selgus, et andmestiku esimeses pooles on puuduvaid väärtusi märgatavalt rohkem, kui andmestiku teises pooles. Kirjeldatud seose uurimiseks jagati andmestik ajalisel järjekorras viieks võrdseks osaks ning vaadeldi igas alamandmestikus pangakonto väljavõtte olemasolu osakaalu. Kui ajaliselt kõige varajasemas alamandmestikus olid pangakonto väljavõtte esitanud vaid 22,9% laenusaaajatest, siis ajaliselt kõige hilisemas alamandmestikus oli sama näitaja juba 56,5% (joonis 1). Sellest saab järeldada, et krediidiasutus, kust andmed pärinevad, on aja jooksul oma äriprotsessi arendanud ning hakanud uutelt laenutaotlejatelt üha enam nõudma nende pangakonto väljavõtet.



Joonis 1: Pangakonto väljavõtte olemasolu ja aja seos.

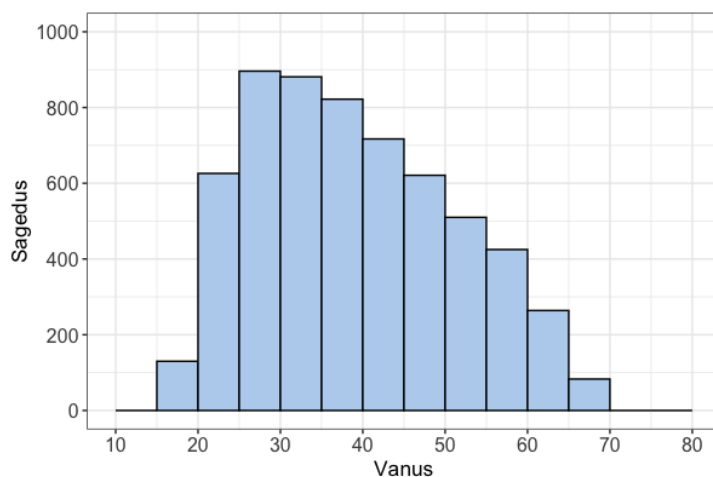
3.1 Sõltumatud tunnused

Käesolevas alapeatükis antakse ülevaade mudelites kasutatud sõltumatutest tunnustest. Arvtunnuse korral tehakse ülevaade tunnuse jaotusest ning selle peamistest arvkarakteristikutest. Kategoriaalsete tunnuste puhul tuuakse välja sõltumatu ja uuritava tunnuse ühisjaotus.

3.1.1 Arvtunnused

Laenusaaaja vanus

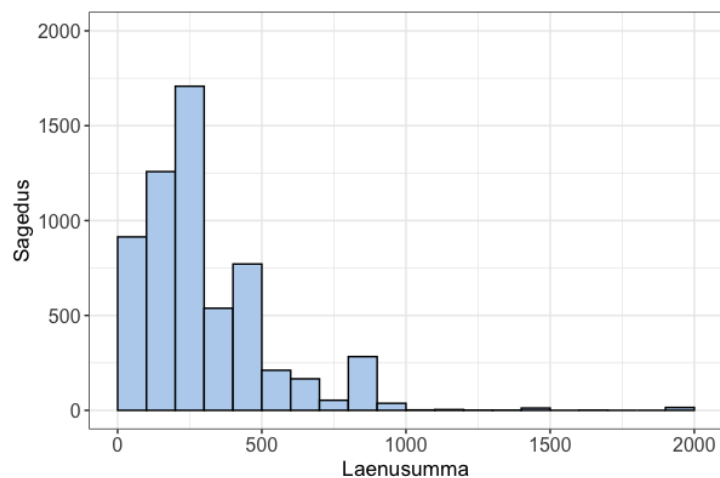
Laenusaajate keskmine vanus andmestikus on 39,6 aastat ning mediaaniks on 38 aastat, mis viitab sellele, et laenusaajate seas on rohkem keskmiselt nooremaid kliente. Noorim laenusaaaja on 18 aastat vana ning vanim laenusaaaja on 68-aastane. Joonisel 2 on toodud tunnuse *vanus* jaotus, kust selgub, et enim laenusaaajaid asub vanuserühmades 25-30 ja 30-35 eluaastat, mis omakorda kinnitab juba eelnevalt mainitut, et laenusaajate seas on rohkem keskmisest nooremaid inimesi.



Joonis 2: Laenusaajate vanuseline jaotus.

Laenusumma

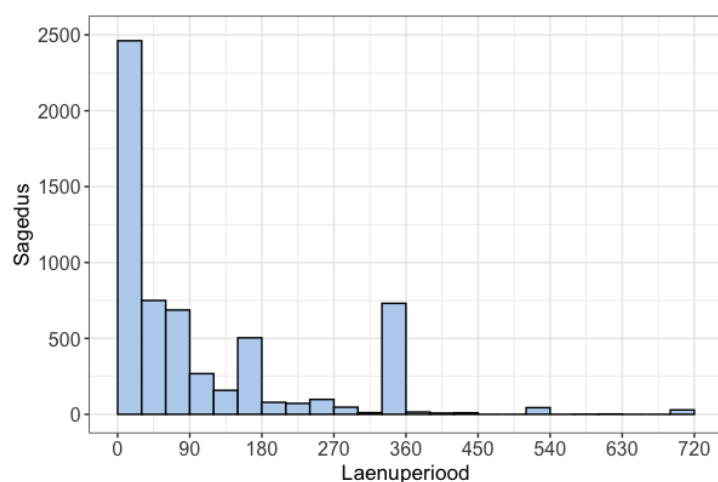
Keskmiseks laenatavaks summaks on 336,80 eurot, kusjuures suurim laenusumma on 2000 eurot ning väikseim 50 eurot. Joonisel 3 on näha tunnuse *laenusumma* jaotus, mille põhjal selgub, et enamik laenusummadest jääb vahemikku 0-300 eurot, millele viitab ka laenusumma mediaanväärtus 300. Lisaks on *laenusumma* jaotuse jooniselt näha, et tunnus on asümmeetriline sabaga paremale.



Joonis 3: Laenusumma jaotus.

Laenuperiood

Antud andmestiku puhul on tegemist väikelaenudega, millest tulenevalt on ka laenuperioodid lühikesed. Keskmine laenuperiood on 120,82 päeva ehk umbes 4 kuud, kõige lühem periood on 1 päev ning pikim periood 720 päeva ehk ligikaudu 2 aastata. Tunnust *laenuperiood* lähemalt uurides selgub, et enim laene on võetud perioodiks 30 päeva, kuid lisaks sellele on populaarsed laenuperioodid veel 60 päeva, 90 päeva ning 360 päeva. Sama tuleb välja ka jooniselt 4, millel on toodud tunnuse *laenuperiood* jaotus.



Joonis 4: Laenuperioodi jaotus.

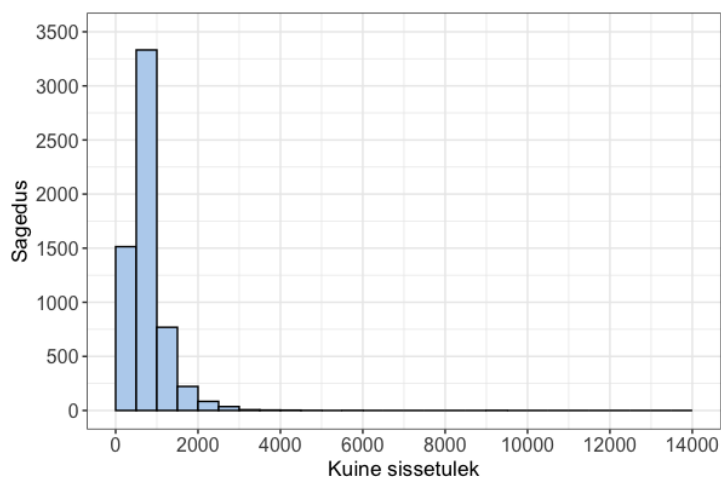
Kuine sissetulek

Laenusaaajate keskmine sissetulek on 811,96 eurot kuus ning kuise sissetuleku mediaanväärtus on 700 eurot, mis tähendab, et enam kui poolte laenusaaajate kuine sissetulek on alla keskmise. Kuise sissetuleku miinimumväärtus on 13 eurot ning maksimumväärtus 14004 eurot. Nendest viimase puhul on tegemist erindiga, mis võib tuleneda sellest, et laenutaotleja on teinud andmete sisestamisel näpuvea. Kui see vaatlus andmestikust välja jätta, siis on kuine sissetulek keskmiselt 809,75 eurot ning suurimaks sissetulekuks on 9500 eurot. Samas ei saa kindel olla, et tegemist on veaga, mistõttu ei saa seda vaatlust mudelite ehitamisel välja jätta.

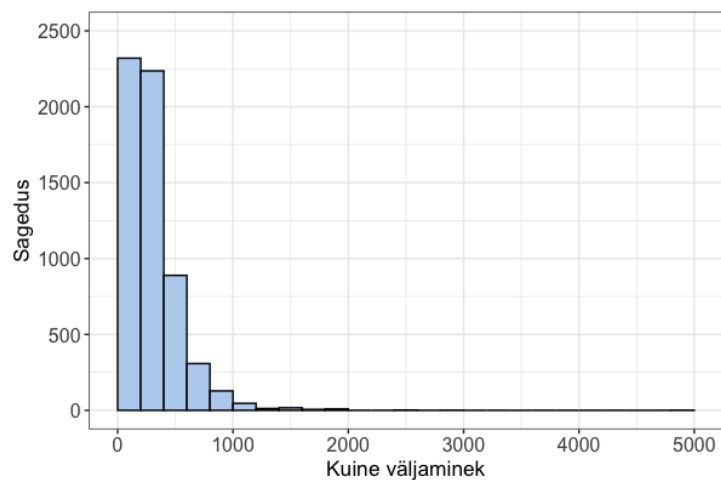
Jooniselt 5 on näha, et tunnuse *sissetulek* jaotus on asümmeetriline sabaga paremale ehk suurema osa laenusaaajate kuine sissetulek jääb vahemikku 0-800 eurot, kuid leidub ka inimesi, kelle kuine sissetulek on märgatavalt suurem.

Kuine väljaminek

Laenusaaajate keskmine väljaminek ühes kuus on 331,54 eurot, tunnuse miinimumväärtus on 4 eurot ning maksimumväärtus on 5000 eurot. Joonisel 6 on toodud tunnuse jaotus, millelt on näha, et enamike laenusaaajate kuine väljaminek jääb vahemikku 0-400 eurot, mida kinnitab ka väljamineku mediaanväärtus 300 eurot.



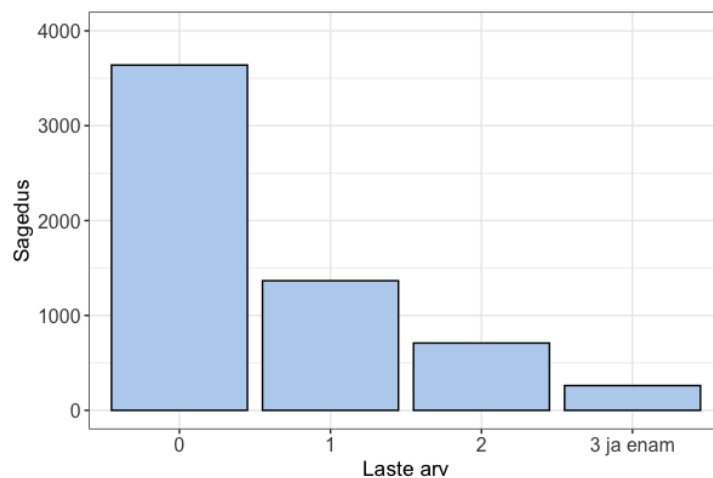
Joonis 5: Kuise sissetuleku jaotus.



Joonis 7: Laste arvu jaotus enne ülemise piiri rakendamist.

Laste arv

Laenusaaajate laste arv varieerub vahemikus 0-10 last, kuid siiski on enamik laenusaaajaid andmestikus lastetud (60,9%). Ainult 0,94% laenusaaajatel on rohkem kui 3 last, millest tulenevalt otsustati tunnusele *laste arv* rakendada ülemist piiri 3 last. Tunnuse jaotus enne ja pärast ülemise piiri rakendamist on toodud vastavalt joonistel 7 ja 8.



Joonis 8: Laste arvu jaotus pärast ülemise piiri rakendamist.

3.1.2 Kategoriaalsed tunnused

Sugu

Tabelist 2 on näha, et naised ja mehed on andmestikus enam-vähem sama palju – 2955 naist ja 3020 meest. Kui aga võrrelda naiste ja meeste laenu tagastamist, siis on näha, et naiste seas on „häid“ kliente rohkem. Soo ja laenu staatuse seost kontrolliti ka hii-ruut testiga, kus teststatistiku väärtuseks saadi 34,52 (p-väärtus $4,28 \cdot 10^{-9}$), mille põhjal saab järeldada, et laenu staatuse ja soo vahel tõepoolest leidub statistiliselt oluline seos.

Tabel 2: Tunnuste *sugu* ja *staatus* ühisjaotus ning tunnuse *staatus* tinglik jaotus.

Sugu/Staatus	0	1	Kokku
Naine	734 (24,8%)	2221 (75,2%)	2955
Mees	958 (31,7%)	2062 (68,3%)	3020
Kokku	1692	4283	5975

Maakond

Tunnus *maakond* kirjeldab laenusaaaja elukohta. Tabelist 3 selgub, et 3162 laenusaaajat ehk teisisõnu enamik (52,9%) laenusaaajatest on pärit Harjumaalt. Lisaks

on rohkelt laenusaaajaid pärit veel Ida-Virumaalt ja Tartumaalt, kus laenusaaajate arvud on vastavalt 908 ja 615. Vähim laenusaaajaid on pärit Hiiumaalt (14 laenusaaajat) ning Saaremaalt (42 laenusaaajat), kuid see on loomulik, kui võtta arvesse maakondade suurus.

Uurides laenu tagastanud ja tagastamata jätnud klientide osakaalu maakonniti, siis selgub, et kõige paremini maksavad laenu tagasi Harjumaal ning Tartumaal elavad laenusaaajad, kelle seast vastavalt 74,3% ja 72,2% moodustavad „head“ kliendid. Seevastu halvemini maksavad laenu tagasi Võrumaal ja Põlvamaal elavad laenusaaajad, kelle seas samad näitajad on vastavalt 58,3% ja 59,4%. Maakonna ja laenu staatuse vahelist seost uuriti ka hii-ruut testiga, mille statistiku väärtuseks saadi 41,087 (p-väärtus 0,0005). Test kinnitab, et tunnuste *maakond* ja laenu *staat* vahel leidub statistiliselt oluline seos.

Tabel 3: Tunnuste *maakond* ja *staat* ühisjaotus ning tunnuse *staat* tinglik jaotus.

Maakond/Staat	0	1	Kokku
Harjumaal	813 (25,7%)	2349 (74,3%)	3162
Hiiumaal	4 (28,6%)	10 (71,4%)	14
Ida-Virumaal	277 (30,5%)	631 (69,5%)	908
Jõgevamaal	29 (35,8%)	52 (64,2%)	81
Järvamaal	30 (28,0%)	77 (71,0%)	107
Lääne-Virumaal	78 (34,5%)	148 (65,5%)	226
Läänemaal	23 (30,7%)	52 (69,3%)	75
Põlvamaal	28 (40,6%)	41 (59,4%)	69
Pärnumaal	74 (28,7%)	184 (71,3%)	258
Raplamaal	37 (36,3%)	65 (63,7%)	102
Saaremaal	13 (31,0%)	29 (69,0%)	42
Tartumaal	171 (27,8%)	444 (72,2%)	615
Valgamaal	30 (30,3%)	69 (69,7%)	99
Viljandimaal	55 (37,9%)	90 (62,1%)	145
Võrumaal	30 (41,7%)	42 (58,3%)	72
Kokku	1692	4283	5975

Emakeel

Eesti keelt emakeelena kõnelevaid inimesi on laenusaaajate seas 3359 ning vene keelt kõnelevaid inimesi 2616, mis tähendab seda, et eesti keelt kõnelevaid inimesi on andmestikus rohkem. Tabelist 4 selgub, et vene keelt kõnelevate klientide hulgas on laenu tagastanud kliente rohkem (72,9%) kui eesti keelt kõnelevate laenusaaajate seas (70,7%). Antud erinevus ei ole siiski suur ning seda kinnitab ka hii-ruut test, andes teststatistiku väärtuseks 3,28 (p-väärtus 0,07), mille põhjal ei saa väita, et tunnuste *emakeel* ja *staatus* vahel esineks seos.

Tabel 4: Tunnuste *emakeel* ja *staatus* ühisjaotus ning tunnuse *staatus* tinglik jaotus.

Emakeel/Staatus	0	1	Kokku
Eesti keel	983 (29,3%)	2376 (70,7%)	3359
Vene keel	709 (27,1%)	1907 (72,9%)	2616
Kokku	1692	4283	5975

Perekonnaseis

Tunnusel *perekonnaseis* on andmestikus 5 võimalikku väärtust: vallaline, vabaabielus, abielus, lahutatud ja lesk. Tabelist 5 selgub, et rohkem laenusaaajaid on vallaliseid, abielus ja vabaabielus ning vähem on leski ja lahutatud inimesi. Kui võrrelda erinevate perekonnaseisust klientide maksekäitumist, siis on näha, et lahutanute seas on laenu tagastanud inimesi 80,2%, olles parima maksekäitumisega grupp. Samas vallaliste seas on laenu tagastanute osakaal 66,0%, olles sellega halvima maksekäitumisega grupp. Hii-ruut teststatistiku väärtuseks saadi 89,42 (p-väärtus $< 2,2 \cdot 10^{-16}$), mille põhjal saab kinnitada, et tunnuste *perekonnaseis* ja laenu *staatus* vahel eksisteerib statistiliselt oluline seos.

Haridustase

Andmestikus oli algselt tunnusel *haridustase* 6 võimalikku väärtust (haridus puudub, algharidus, põhiharidus, keskharidus, kutseharidus ja kõrgharidus), kuid erinevate haridustasemete maksekäitumist lähemalt uurides otsustati tunnus ümber

kodeerida järgmiselt: kuni põhiharidus, kesk- või kutseharidus ning kõrgharidus. Ümberkodeeritud tunnuse ühisjaotus tunnusega laenu *staatus* on toodud tabelis 6, kust on näha, et kõrgharidusega klientide seas on laenu tagastanud märgatavalt rohkem (82,5%) kui kesk- või kutseharidusega (71,5%) või kuni põhiharidusega (57,9%) laenusaaajate seas. Nende kahe tunnuse omavaheline seos on ka statistiliselt oluline, sest hii-ruut teststatistiku väärtuseks saadi 139,83 (p -väärtus $< 2,2 \cdot 10^{-16}$).

Tabel 5: Tunnuste *perekonnaseis* ja *staatus* ühisjaotus ning tunnuse *staatus* tinglik jaotus.

Perekonnaseis/Staatus	0	1	Kokku
Vallaline	629 (33,0%)	1222 (66,0%)	1851
Vabaabielus	511 (31,7%)	1103 (68,3%)	1614
Abielus	394 (22,8%)	1338 (77,2%)	1732
Lahutatud	119 (19,8%)	481 (80,2%)	600
Lesk	39 (21,9%)	139 (78,1%)	178
Kokku	1692	4283	5975

Tabel 6: Tunnuste *haridustase* ja *staatus* ühisjaotus ning tunnuse *staatus* tinglik jaotus.

Haridustase/Staatus	0	1	Kokku
Kuni põhiharidus	345 (42,1%)	474 (57,9%)	819
Kesk-, kutseharidus	1154 (28,5%)	2902 (71,5%)	4056
Kõrgharidus	193 (17,5%)	907 (82,5%)	1100
Kokku	1692	4283	5975

Töökogemus

Tabelist 7 selgub, et enamik laenusaaajaid omab rohkem kui aasta pikkust töökogemust – 4437 laenusaaajat (74,3%). Hii-ruut testiga uuriti ka töökogemuse ja laenu tagastamise vahelist seost. Teststatistiku väärtuseks saadi 69,11 (p -väärtus $6,63 \cdot 10^{-15}$) ehk nende kahe tunnuse vahel leidub statistiliselt oluline seos.

Kinnisvara omamine

Tabelist 8 selgub, et veidi üle poole laenusaaajatest ei oma kinnisvara. Kui võrrelda laenu tagastamist, siis kinnisvara omavad laenusaaajad on paremad kliendid kui

kinnisvara mitte omavad laenusaajad. Nimelt kinnisvara omavate laenusaaajate seas on laenu tagastamise osakaal 82,4%, kuid kinnisvara mitte omavate klientide seas on sama näitaja 63,2%. Kinnisvara ja laenu staatuse vahelist seost kinnitab ka hii-ruut teststatistik väärtusega 267,24 (p -väärtus $< 2,2 \cdot 10^{-16}$).

Tabel 7: Tunnuste *töökogemus* ja *staat*us ühisjaotus ning tunnuse *staat*us tinglik jaotus.

Töökogemus/Staat	0	1	Kokku
Tööt	73 (27,3%)	194 (72,7%)	267
Katseaeg	59 (33,9%)	115 (66,1%)	174
Kuni aasta	418 (38,1%)	679 (61,9%)	1097
Rohkem kui aasta	1142 (25,7%)	3295 (74,3%)	4437
Kokku	1692	4283	5975

Tabel 8: Tunnuste *kinnisvara* ja *staat*us ühisjaotus ning tunnuse *staat*us tinglik jaotus.

Kinnisvara/Staat	0	1	Kokku
0	1226 (36,8%)	2103 (63,2%)	3329
1	466 (17,6%)	2180 (82,4%)	2646
Kokku	1692	4283	5975

Aktiivne maksehäire

Tunnus *aktiivne maksehäire* kirjeldab, kas laenusaajal esineb aktiivseid maksehäireid või mitte. Tabelist 9 on näha, et pea enamikel laenusaaajatel andmestikus ei esine aktiivseid maksehäireid (5342, mis moodustab 89,4%). Samuti selgub, et klientid, kellel puuduvad aktiivsed maksehäired, maksavad laenu paremini tagasi kui kliendil, kellel esineb mõni aktiivne maksehäire. Vastavad laenu tagastanute osakaalud on 73,2% ja 59,2%. Kahe tunnuse vahelist seost testiti ka hii-ruut testiga, mille statistiku väärtuseks saadi 53,30 (p -väärtus $2,70 \cdot 10^{-13}$), mis kinnitab, et tunnuste vahel on statistiliselt oluline seos.

Tabel 9: Tunnuste *aktiivne maksehäire* ja *staatus* ühisjaotus ning tunnuse *staatus* tinglik jaotus.

Aktiivne maksehäire/Staatus	0	1	Kokku
0	1434 (26,8%)	3908 (73,2%)	5342
1	258 (40,8%)	375 (59,2%)	633
Kokku	1692	4283	5975

Suletud maksehäire

Tunnus *suletud maksehäire* näitab, kas laenusaaajal on enne laenu taotlemist esinenud maksehäireid. Tabeli 10 põhjal selgub, et kliente, kelle ei ole minevikus esinenud maksehäireid on andmestikus 3640 ja neid, kelle on esinenud mõni maksehäire on 2335. Sarnaselt tunnusele *aktiivne maksehäire* on ka tunnuse *suletud maksehäire* puhul näha, et laenusaaajad, kellel on minevikus esinenud maksehäireid, on kehvemad laenu tagastajad võrreldes klientidega, kellel ei ole minevikus maksehäireid esinenud. Laenu tagastamise osakaalud on vastavalt 64,2% ja 76,5%. Hii-ruut test-statistiku väärtuseks saadi 105,18 (p -väärtus $< 2,2 \cdot 10^{-16}$), mis samuti kinnitab, et suletud maksehäire ja laenu staatuse vahel esineb statistiliselt oluline seos.

Tabel 10: Tunnuste *suletud maksehäire* ja *staatus* ühisjaotus ning tunnuse *staatus* tinglik jaotus.

Suletud maksehäire/Staatus	0	1	Kokku
0	856 (23,5%)	2784 (76,5%)	3640
1	836 (35,8%)	1499 (64,2%)	2335
Kokku	1692	4283	5975

Pangakontoinfo

Tabeli 11 põhjal selgub, et andmestikus on kliente, kes on krediidasutusele esitanud oma pangakonto väljavõtte, vähem kui kliente, kes seda teinud ei ole (vastavalt 2513 ja 3462). Samas on heade klientide osakaal mõlemas grupis sarnane. Nimelt pangakonto väljavõtte esitanud klientide seas on häid kliente 70,6% ning pangakonto väljavõtte esitamata jätanud klientide seas on sama näitaja 72,5%. Hii-ruut

teststatistiku väärtuseks saadi 2,63 (p-väärtus 0,11), mille põhjal ei saa väita, et tunnuste vahel esineks seos.

Tabel 11: Tunnuste *pangakontoinfo* ja *staatus* ühisjaotus ning tunnuse *staatus* tinglik jaotus.

Pangakontoinfo/Staatus	0	1	Kokku
0	952 (27,5%)	2510 (72,5%)	3462
1	740 (29,4%)	1773 (70,6%)	2513
Kokku	1692	4283	5975

3.2 Tunnuste informatiivsus

Kõikidele sõltumatutele tunnustele leiti tunnuse informatiivsus (inglise keeles *information value*). Tunnuse informatiivsus on väärtus, mis näitab tunnuse ennustusvõimet – mida suurema väärtusega on näitaja, seda suurem on muutuja ennustusvõime kategoriseerimisel. Tunnuse informatiivsuse näitaja leitakse alltoodud valemi abil:

$$IV = \sum_{i=1}^k \ln \left(\frac{Halb_i}{Halb} \cdot \frac{Hea}{Hea_i} \right) \cdot \left(\frac{Halb_i}{Halb} - \frac{Hea_i}{Hea} \right),$$

kus k tähistab uuritava tunnuse tasemete arvu, $Halb$ ja Hea tähistavad vastavalt heade ja halbade laenude koguarvu andmestikus ning $Halb_i$ ja Hea_i on vastavalt heade ja halbade laenude arv uuritava tunnuse i -ndal tasemel. Panganduses peetakse tugeva ennustusvõimega tunnuseks sellist tunnust, mille informatiivsuse näitaja on suurem kui 0,2. (Kočenda ja Vojtek, 2009)

Pidevate ja täisarvuliste arvtunnuste informatiivsuse leidmiseks tunnused kategoriseeriti. Tunnuste kategoriseerimisel jälgiti, et ei tekiks väga väikese vaatluste arvuga kategooriaid.

Tabelist 12 selgub, et kõige informatiivsemad seletavad tunnused andmestikus on *kinnisvara*, *vanus* ning *haridustase*, mille informatiivsused on vastavalt 0,2343, 0,1721 ning 0,1174. Siiski tugeva ennustusvõimega tunnuseid on andmestikus vaid

üks – *kinnisvara*. Informatiivsuse näitaja põhjal on halvima ennustusvõimega tunnuseks *laste arv* (informatiivsus 0,0022).

Tabel 12: Tunnuste informatiivsus.

Tunnus	Informatiivsus
Kinnisvara	0,2343
Vanus	0,1721
Haridustase	0,1174
Laenuperiood	0,1082
Suletud maksehäire	0,0858
Perekonnaseis	0,0762
Töökogemus	0,0538
Aktiivne maksehäire	0,0408
Maakond	0,0323
Sugu	0,0289
Kuine väljaminek	0,0194
Laenusumma	0,0172
Kuine sissetulek	0,0136
Emakeel	0,0028
Pangakontoinfo	0,0022
Laste arv	0,0022

4 Mudelid

Järgnevalt antakse ülevaade eelnevas peatükis kirjeldatud andmetel põhinevate krediidiriski ennustusmodelite loomisest. Esmalt kirjeldatakse logistilisel regressioonil, LASSO regressioonil, klassifitseerimispuul ja *gradient boosting* algoritmil põhinevate mudelite ehitusprotsessi ja lõplikkeprodukte. Seejärel võrreldakse saadud mudelite headust.

Andmete analüüsimisel ja mudelite ehitamisel on kasutatud rakendustarkvara R (versioon 3.6.3). Peamisteks kasutatud pakettideks on dplyr, stats, glmnet, rpart, caret ning pROC.

4.1 Mudelite ehitamine

Mudelite ehitamisel jagati kogu andmestik kaheks – treening- ja testandmestikuks. Treeningandmestik moodustas 80% ja testandmestik 20% kogu andmestikust, kusjuures saadud andmestike suurused olid vastavalt 4780 ja 1195 vaatlust. Treening- ja testandmestiku moodustamisel jagati esmalt andmestik ajalises järjekorras viieks võrdse suurusega alamandmestikuks ning seejärel valiti juhuslikult igast osast 80% vaatlustest treeningandmestikku ja ülejäänud 20% vaatlustest läksid testandmestikku. Treeningandmestikku kasutati mudelite ehitamiseks ja ristvalideerimisel sobivaimate parameetrite leidmiseks ning testandmestikku hiljem mudelite headuse hindamiseks.

Algsetesse mudelitesse kaasati kõik peatükis 3.1 toodud tunnused. Tunnus *töökogemus* kodeeriti järgmiselt:

- 0 – töötu;
- 1 – katseaeg;
- 2 – kuni aastane töökogemus;

- 3 – töökogemust enam kui aasta.

Ülejäänutel tunnustel kasutati nende algset või logaritmitud kuju. Logaritmimist kasutati tunnuse hajuvuse vähendamiseks ja jaotuse sümmeetrilisemaks muutmiseks. Tunnuste logaritmitud kuju kasutati ainult regressioonimudelites.

Logistiline regressioon peamõjudega

Logistilise regressioonimudeli ehitamisel kasutati kahanevat sammregressiooni, mille lähtepunktiks oli mudel, mis sisaldas kõikide sõltumatute tunnuste peamõjusid. Kahaneva sammregressiooni tulemusena saadi järgnev logistiline mudel:

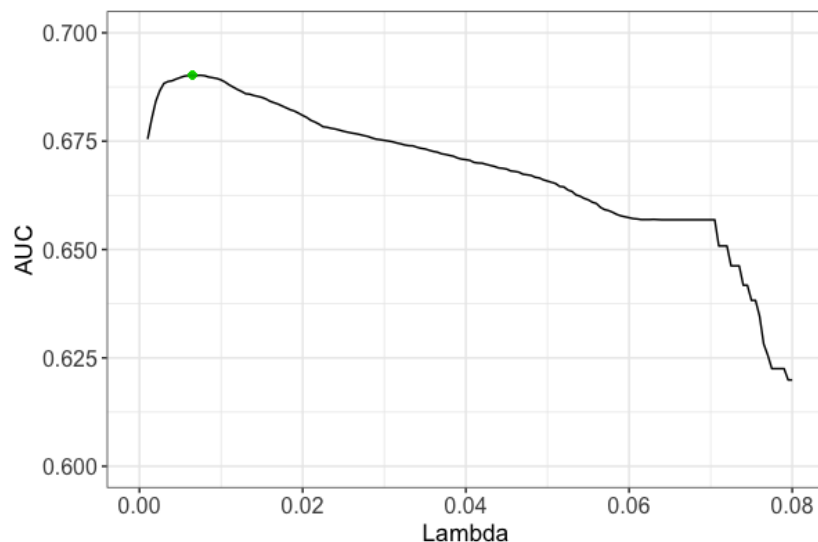
$$\begin{aligned} \text{logit}(\pi) = & 1,6328 - 0,1975 \cdot (\text{sugu} = \text{M}) + 0,0254 \cdot \text{vanus} + 0,3597 \cdot (\text{maakond} = \\ & = \text{Hiiumaa}) - 0,4369 \cdot (\text{maakond} = \text{Ida-Virumaa}) + 0,0511 \cdot (\text{maakond} = \\ & = \text{Järvamaa}) - 0,4206 \cdot (\text{maakond} = \text{Jõgevamaa}) - 0,3147 \cdot (\text{maakond} = \\ & = \text{Lääne-Virumaa}) - 0,0150 \cdot (\text{maakond} = \text{Läänemaa}) - 0,1155 \cdot (\text{maakond} = \\ & = \text{Pärnumaa}) - 0,7349 \cdot (\text{maakond} = \text{Põlvamaa}) - 0,4578 \cdot (\text{maakond} = \\ & = \text{Raplamaa}) - 0,3540 \cdot (\text{maakond} = \text{Saaremaa}) - 0,1089 \cdot (\text{maakond} = \\ & = \text{Tartumaa}) - 0,1288 \cdot (\text{maakond} = \text{Valgamaa}) - 0,2328 \cdot (\text{maakond} = \\ & = \text{Viljandima}) - 0,2463 \cdot (\text{maakond} = \text{Võrumaa}) - 0,3511 \cdot \ln(\text{laenusumma}) + \\ & + 0,0006 \cdot \text{väljaminek} + 0,5716 \cdot (\text{pangakontoinfo} = 1) - 0,4423 \cdot \\ & \cdot (\text{aktiivne maksehäire} = 1) - 0,5325 \cdot (\text{suletud maksehäire} = 1) + 0,9507 \cdot \\ & \cdot (\text{kinnisvara} = 1) + 0,3386 \cdot (\text{haridus} = \text{kõrgharidus}) - 0,3643 \cdot (\text{haridus} = \\ & = \text{kuni põhiharidus}), \end{aligned}$$

kus π tähistab laenu tagasimaksmise tõenäosust. Mudelis osutusid tunnuse *maakond* tasemetest vaid Ida-Virumaa ja Põlvamaa statistiliselt oluliselt erinevaks tasemest Harjumaa, mistõttu võrreldi tõepära suhte testi abil saadud mudelit mudeliga, kust tunnus *maakond* oli tervikuna välja jäetud. Tõepära suhte testi teststatistiku väärtuseks saadi 29,1220 (p-väärtus 0,0101), mis tähendab, et kasu keerulisemast

modelist on statistiliselt oluline. Seetõttu tuleks eelistada mudelit, kus tunnus *maakond* on kaasatud. Lõpliku mudelis olevate tunnuste kordajad ja nende p-väärtused on toodud tabelis 21 (lisa 1).

LASSO regressioon

LASSO regressioonimudelisse kaasati lisaks sõltumatute tunnuste peamõjudele ka kõikide kategooriaalsete tunnuste paarikaupa koosmõjud. Mudeli ehitamisel leiti esmalt ristvalideerimisega parameetri λ optimaalne väärtus, milleks saadi $\lambda = 0,0065$ (joonis 9). Seejärel optimaalset λ väärtust kasutades treeniti mudel tervel treeningandmestikul. Lõpliku LASSO regressioonimudeli nullist erinevad kordajad on toodud tabelis 22 (lisa 1). Tabelist on näha, et lisaks tunnuste peamõjudele mõjutavad laenu tagasimaksmise tõenäosusi ka mõnede kategooriaalsete tunnuste vahelised koosmõjud.



Joonis 9: LASSO regressiooni optimaalse parameetri leidmine ristvalideerimise abil.

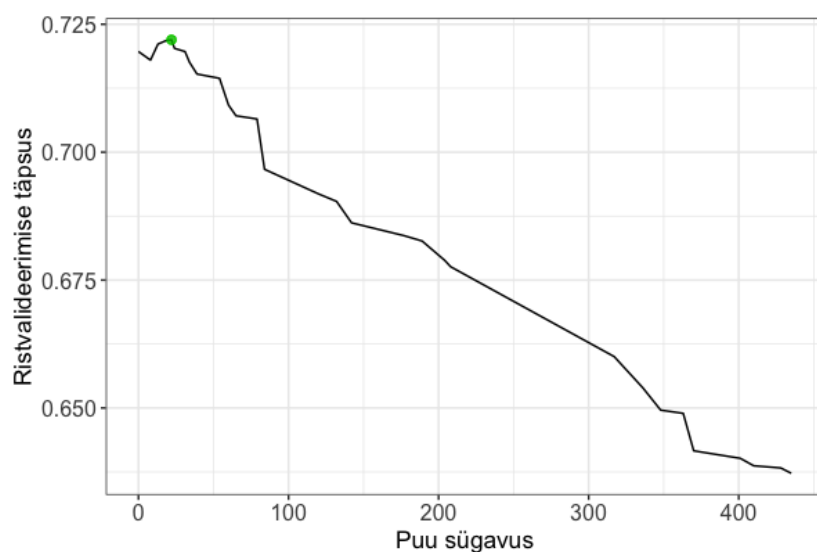
Logistiline regressioon pea- ja koosmõjudega

LASSO mudelis oluliseks osutunud tunnuste vahelised koosmõjud kaasati ka logistilisse regressiooni. Taaskord kasutati kahanevat sammregressiooni, mille lähtepunktiks oli seekord mudel, mis sisaldas kõikide sõltumatute tunnuste peamõjusid ning

LASSO mudeli põhjal valitud koosmõjud. Sammregressiooni tulemusena saadi mudel, kus statistiliselt olulised koosmõjud esinesid järgmiste tunnuste vahel: *perrekonnaseis* ja *suletud maksehäire*, *perrekonnaseis* ja *kinnisvara*, *pangakontoinfo* ja *kinnisvara* ning *aktiivne maksehäire* ja *kinnisvara*. Mudelis oluliseks osutunud tunnuste kordajad ja nende p-väärtused on toodud tabelis 23 (lisa 1).

Klassifitseerimispuu

Klassifitseerimispuu (CART) konstrueerimisel loodi esmalt ülesobitatud puu (sügavusega 435), mida seejärel pügama hakati. Puu optimaalse suuruse leidmiseks kasutati 10-kordset ristvalideerimist, mille tulemused on toodud joonisel 10. Jooniselt selgub, et klassifitseerimispuu optimaalseks sügavuseks on 22 ehk saadud otsustuspuu on märgatavalt väiksem kui algne puu.

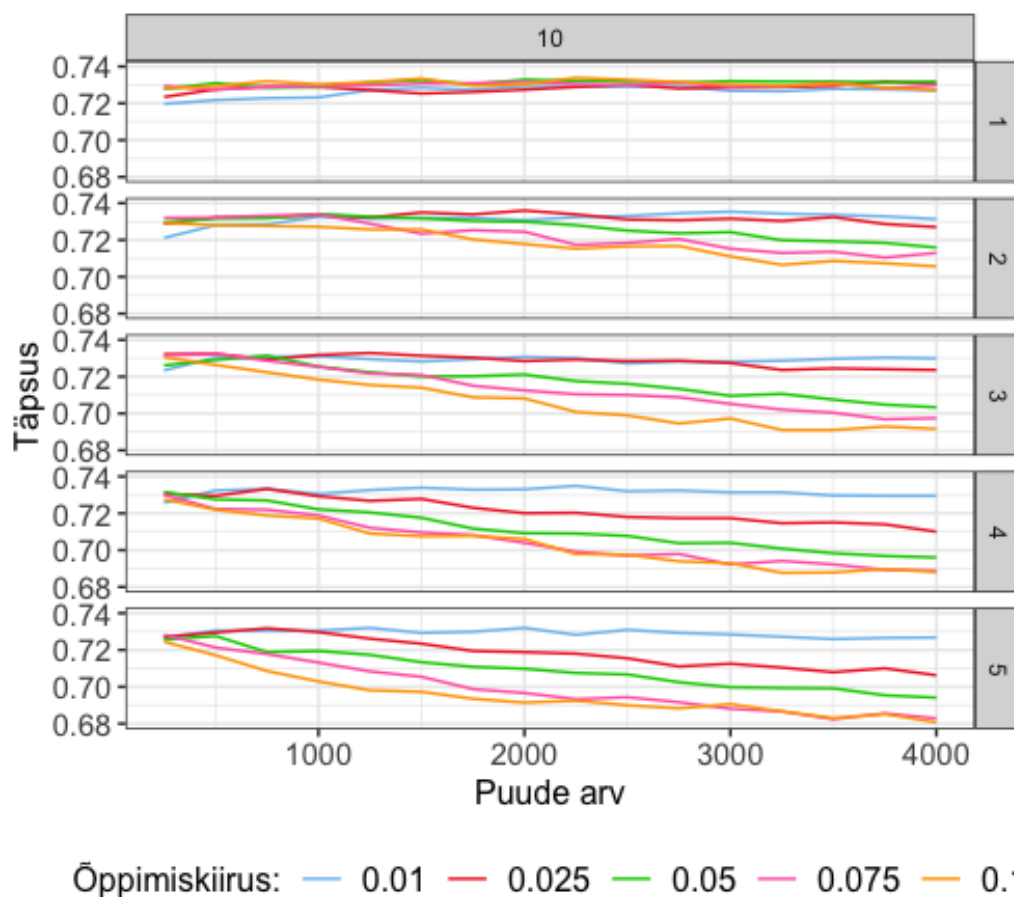


Joonis 10: Klassifitseerimispuu optimaalse suuruse leidmine ristvalideerimise abil.

Gradient Boosting

Viimaseks töös kasutatud meetodiks on *gradient boosting*, mis põhineb samuti klassifitseerimispuudel. Nii nagu eelnevate meetodite korral, leiti ka siin optimaalsed parameetrite väärtused ristvalideerimise abil. Joonisel 11 on näha, et sobivaimaks õppimiskiiruseks on 0,025, optimaalne klassifitseerimispuu sügavus on 2, kusjuures

optimaalne väikseim vaatluste arv lehes on 10 ning optimaalne puude ehk iteratsioonide arv algoritmis on 2000. Kasutades mainitud parameetrite väärtusi treeniti algoritm tervel treeningandmestikul.



Joonis 11: *Gradient boosting* meetodi optimaalsete parameetrite leidmine ristvalideerimise abil.

4.2 Mudelite võrdlus

Mudelite ennustustäpsuse hindamiseks rakendati treenitud mudeleid testandmetel. Kõikidele mudelitele leiti mudeli täpsus, tundlikkus ja spetsiifilisus. Regressioonil põhinevatele mudelitele leiti lisaks eelmainitud headuse näitajatele ka ROC kõvera alune pindala (AUC).

Klassifitseerimispuu ja *gradient boosting* meetodi võrdlemiseks loodi klassifitseerimistabelid, mis on toodud vastavalt tabelites 13 ja 14. Tabelites on näha, et mõlemad meetodit prognoosivad hea staatusega laene paremini kui halva staatusega laene, kuid klassifitseerimispuu üldine täpsus on veidi parem. Kõikidest hea staatusega klientidest suutis klassifitseerimispuu õigesti prognoosida 94,78% ning *gradient boosting* meetod 92,17%. Samas halva staatusega laene klassifitseerib paremini *gradient boosting* meetod. Nimelt, kõikidest halva staatusega laenudest suutis *gradient boosting* õigesti prognoosida 67 laenu, mis on 19,03% tegelikest halva staatusega laenudest, samas kui klassifitseerimispuu puhul oli sama näitaja 46 laenu, mis moodustab 13,07% kõikidest halva staatusega laenudest testandmestikus. Klassifitseerimispuu ja *gradient boosting* meetodi täpsus, tundlikkus ja spetsiifilisus on toodud tabelis 15.

Tabel 13: Klassifitseerimispuu klassifitseerimistabel.

Ennustus/Tegelik	0	1	Kokku
0	46	44	90
1	306	799	1105
Kokku	352	843	1195

Tabel 14: *Gradient boosting* klassifitseerimistabel.

Ennustus/Tegelik	0	1	Kokku
0	67	66	133
1	285	777	1062
Kokku	352	843	1195

Tabel 15: Otsustuspuudel põhinevate meetodite headuse näitajad.

Mudel	Täpsus	Tundlikkus	Spetsiifilisus
Klassifitseerimispuu	70,71%	94,78%	13,07%
<i>Gradient boosting</i>	70,63%	92,17%	19,03%

Omavahel sooviti võrrelda ka kolme regressioonmudeli ning *gradient boosting* meetodi prognoosivõimet. Selleks määrati regressioonmudelite „hea“ ja „halva“ kliendi

eralduspiir (lävend) nii, et kõik neli mudelit prognoosiks 1062 hea staatusega laenu ja 133 halva staatusega laenu, kusjuures antud arvud põhinesid *gradient boosting* meetodi marginaaljaotusel (tabel 14). Peamõjusid sisaldava logistilise regressiooni lävendiks sai 0,53338, LASSO regressioonil 0,55593 ning pea- ja koosmõjusid sisaldaval logistilisel regressioonimudelil 0,52913. Prognoosid (mudelist hinnatud tõenäosused), mis ületasid lävendi, said väärtuseks 1 ehk mudel ennustas, et laen makstakse tagasi ning kõik prognoosid, mis jäid alla otsustuspiiri, said väärtuseks 0 ehk teisisõnu mudel ennustas, et laenu ei maksta tagasi. Tabelites 16, 17 ja 18 on kujutatud selle tulemusena leitud regressioonimodelite klassifitseerimistabelid.

Tabel 16: Peamõjudega logistilise regressiooni klassifitseerimistabel.

Ennustus/Tegelik	0	1	Kokku
0	68	65	133
1	284	778	1062
Kokku	352	843	1195

Tabel 17: LASSO regressiooni klassifitseerimistabel.

Ennustus/Tegelik	0	1	Kokku
0	75	58	133
1	277	785	1062
Kokku	352	843	1195

Tabel 18: Koosmõjudega logistilise regressiooni klassifitseerimistabel.

Ennustus/Tegelik	0	1	Kokku
0	67	66	133
1	285	777	1061
Kokku	352	843	1195

Tabelist 19 on näha, et LASSO regressiooni täpsus on natukene kõrgem kui logistilistel regressioonidel või *gradient boosting* meetodil, mil näitajad on vastavalt 71,97%, 70,79%, 70,63% ja 70,63%. Kui võrrelda mudelite tundlikkust ja spetsiifi-

lisust, siis ka nende näitajate puhul toimib LASSO regressioon paremini võrreldes kolme teise mudeliga.

Tabel 19: Logit mudeli, LASSO regressiooni ja *gradient boosting* võrdlus.

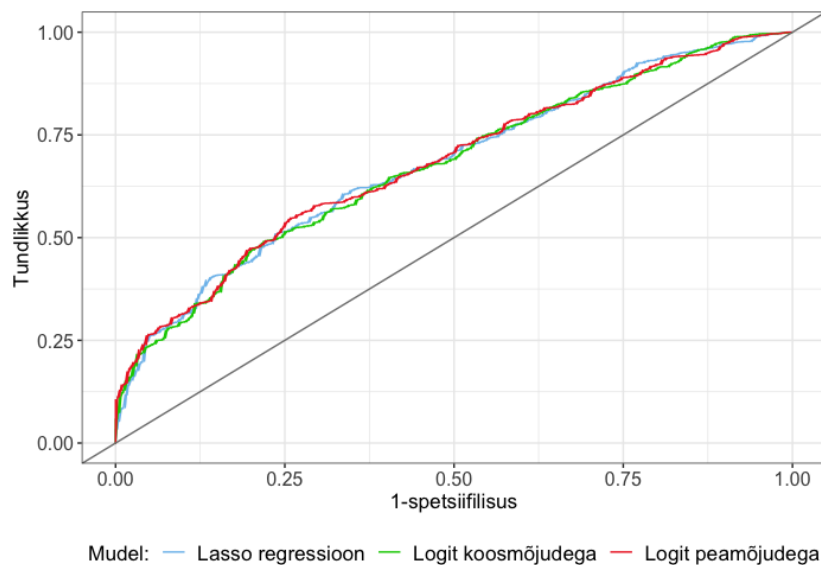
Mudel	Eralduspiir	Täpsus	Tundlikkus	Spetsiifilisus
Logit mudel peamõjudega	0,53338	70,79%	92,29%	19,32%
LASSO regressioon	0,55593	71,97%	93,12%	21,31%
Logit mudel koosmõjudega	0,52913	70,63%	92,17%	19,03%
<i>Gradient boosting</i>	–	70,63%	92,17%	19,03%

Kui võrrelda omavahel kahe logistilise mudeli prognoosivõimet, siis nii täpsus, tundlikkus kui ka spetsiifilisus on ainult peamõjusid sisaldaval mudelil paremad kui peaja koosmõjusid sisaldaval mudelil. Sellest saab järeldada, et koosmõjude kaasamine antud juhul ei anna mudeli ennustusvõimele midagi juurde ning tekitab pigem mudeli ületreenitust. Sellele viitab ka järgnevalt toodud ROC kõvera alune pindala.

Mõlemale logistilisele regressioonmudelile ja LASSO regressioonmudelile leiti ROC kõver (joonis 12) ning selle alune pindala (AUC). Kõigi kolme mudeli puhul jääb ROC kõvera alune pindala alla 0,7 ehk selle näitaja põhjal ei ole tegemist eriti heade mudelitega. Mudelite ROC kõvera alused pindalad on toodud tabelis 20, kust näeb, et suurim ROC kõvera alune pindala on ainult peamõjusid sisaldaval logistilisel regressioonil ($AUC = 0,6801$).

Tabel 20: Regressioonmudelite ROC kõvera alune pindala (AUC).

Mudel	AUC
Logit mudel peamõjudega	0,6801
LASSO regressioon	0,6781
Logit mudel koosmõjudega	0,6762



Joonis 12: Regressioonimudelite ROC kõverad.

Üldiselt võib öelda, et nii regressioonimudelid, klassifitseerimispuu kui ka *gradient boosting* prognoosivad hästi hea staatusega laene ja halvasti kehva staatusega laene. Selliste näitajatega mudelite kasutusele võtmine krediitdiasutuses ei ole otstarbekas, sest halbade laenude valesti klassifitseerimine võib krediitdiasutusele tekitada suuri kahjusid.

4.3 Tulemuste analüüs ja järeldused

Antud alapeatükis võetakse kokku erinevate prognoosimudelite rakendamisest saadud kogemused ning tehakse üldisemat laadi ettepanekud krediitdiskooringu sobiva mudeli leidmiseks olukorras, kus suur osa kliendiinfot on kategooriaalne.

- Kategooriaalsete sõltumatute tunnuste korral sobib krediitdiskooringu modelleerimiseks nii logistiline regressioon, LASSO regressioon, klassifitseerimispuu kui ka *gradient boosting* meetod. Kõik neli mudelit tulevad hästi toime kategooriaalsete tunnustega. Kasutades rakendustarkvara R, on logistilise regressiooni, klassifitseerimispuu ja *gradient boosting* meetodi korral võimalik

modelisse kaasata kategoriaalsed sõltumatud tunnused nii, et neid ei ole vaja ümber kodeerida binaarseteks 0-1 tunnusteks. Samas LASSO regressiooni korral tuleb kategoriaalsed sõltumatud tunnused kodeerida ümber binaarseteks tunnusteks.

- Mudeli prognoositäpsuses mängivad rolli mudelitesse kaasatud argumenttunnused ja nende informatiivsus. Tähtis on, et argumenttunnus suudaks eristada uuritava tunnuse erinevaid tasemeid. Suure tasemete arvuga kategoriaalse argumenttunnuse korral tasub kaaluda tunnuse ümberkodeerimist – sarnaste tasemete omavahel ühendamist. Esiteks aitab see vältida liiga väikeste vaatluste arvuga tasemeid ning teiseks vähendab see parameetrite arvu regressioonmudelites, muutes mudelid lihtsamini interpreteeritavamaks. Näiteks käesolevas töös oli tunnusel *haridustase* algselt 6 võimalikku väärtust (haridus puudub, algharidus, põhiharidus, keskharidus, kutseharidus ja kõrgharidus), kuid erinevate haridustasemete maksekäitumist uurides kodeeriti tunnus ümber nii, et tunnuse võimalikud tasemed olid kuni põhiharidus, kesk- ja kutseharidus ning kõrgharidus. Samas leiab töö autor, et tunnuste ümberkodeerimist oleks saanud töös veel paremini teha, näiteks tunnuse *maakond* näol.
- Kategoriaalsete tunnuste vahelised koosmõjud (korrutistunnused) võivad kanda endas olulist informatsiooni uuritava tunnuse kohta. Suure hulga kategoriaalsete tunnuste vaheliste koosmõjudega tuleb hästi toime LASSO regressioonimudel. LASSO regressioonil toimub tunnuste valik mudelisse „automaatselt“, mille käigus vähem oluliste tunnuste parameetrite hinnangud ($\hat{\beta}_1, \hat{\beta}_2, \dots$) võrdsustatakse nulliga. Seega on võimalik LASSO regressiooni kasutada ka tunnuste ja nende vaheliste koosmõjude valimisel.
- Käesolevas töös kasutatud andmete põhjal selgus, et nii logistiline regressioon, LASSO regressioon kui ka *gradient boosting* meetod eristavad häid laene halbadest laenudest enam-vähem ühesuguse täpsusega. Regressioon-

mudelite puhul on mudeli kasutajal võimalik näha, millised tunnused ja kuidas mõjutavad tulemust, samas kui *gradient boosting* on sisuliselt nagu „must kast“. Töö autori arvates tuleks sellistel juhtudel interpreteeritavust ja kasutajasõbralikkust silmas pidades eelistada regressioonmudeleid.

Kokkuvõte

Käesoleva magistritöö eesmärgiks oli leida otstarbekaid meetodeid kategooriaalse te seletavate tunnuste kasutamiseks krediidiriski hindamisel. Töös kasutati nelja erinevat prognoosimudelit — logistilist regressiooni nii peamõjudega kui ka pea- ja koosmõjudega, LASSO regressiooni, klassifitseerimispuud (CART) ning *gradient boosting* meetodit. Lisaks anti töös ülevaade krediidiriski olemusest, kasutatud meetoditest ning tehti esmane analüüs ja ülevaade mudelite ehitamiseks kasutatud andmetest.

Töös kasutatud andmestik sisaldas infot väikelaenu saanud isikute kohta. Algne andmestik sisaldas 6000 vaatlust ning 33 erinevat tunnust, kuid peale andmete korrastamist koosnes andmestik 5975 vaatlusest ning 17 tunnusest. Töös oli uuritavaks ehk sõltuvaks tunnuseks binaarne tunnus laenu *staatus*, mis kirjeldas, kas laen on krediidasutusele tagastatud või on halvaks läinud. Modelleerimisel kasutatud argumenttunnuste seas oli 6 arvulist ja 10 kategooriaalset tunnust. Kõikidele argumenttunnustele leiti ka tunnuse informatiivsuse näitaja, mille põhjal selgus, et kõige informatiivsemaks seletavaks tunnuseks oli *kinnisvara*, mis oli ka ainuke tugeva ennustusvõimega argumenttunnus käesolevas töös. Argumenttunnuste kehvast informatiivsusest annavad aimu ka töös koostatud prognoosimudelite headuse näitajad.

Töö praktilises osas rakendati nelja erinevat tüüpi prognoosimudelit krediidiriski hindamiseks. Mudelite omavaheliseks võrdlemiseks leiti kõikidele mudelitele testandmetel mudeli täpsus, tundlikkus ning spetsiifilisus. Lisaks leiti regressioonmudelitele ka ROC kõvera alune pindala (AUC). Mudelite headuse näitajate põhjal selgus, et kõik mudelid prognoosivad hea staatusega laene paremini kui halva staatusega laene. Klassifitseerimispuud ja *gradient boosting* meetodit omavahel võrreldes selgus, et klassifitseerimispuu täpsus ja tundlikkus on veidi paremad kui *gradient boosting* meetodil, kuid *gradient boosting* meetodil on jällegi spetsiifilisuse näitaja märgatavalt parem. Logistiliste regressioonmudelite, LASSO regressiooni

ja *gradient boosting* meetodi võrdlemisel saadi, et mudelite headuse näitajad on kõigi mudelite korral enam-vähem sarnased, vaid LASSO regressioonil on headuse näitajad veidi kõrgemad. Samas regressioonmudelite korral on ROC kõvera alune pindala suurim ainult peamõjusid sisaldaval logistilisel regressioonil. Seetõttu ei ole võimalik eristada ühte kõige paremat meetodit.

Üleüldiselt võib öelda, et kõik eelpool nimetatud prognoosimudelid sobivad krediitdiskooringu modelleerimiseks kategooriaalsete argumenttunnuste korral. Mudeli prognoositäpsuses mängib enim rolli argumenttunnuste valik mudelisse – tunnused peavad olema informatiivsed ehk kirjeldama hästi uuritava tunnuse erinevaid tasemeid.

Kasutatud allikad

- Agresti, A. (2013). *Categorical Data Analysis*. John Wiley Sons, Inc.
- Hand, D. J. ja W. E. Henley (1997). “Statistical Classification Methods in Consumer Credit Scoring: a Review”. *Journal of the Royal Statistical Society* 160.3, lk. 523–541.
- Hastie, T., R. Tibshirani ja J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
- James, G., D. Witten, T. Hastie ja R. Tibshirani (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.
- Kočenda, E. ja M. Vojtek (2009). “Default Predictors and Credit Scoring Models for Retail Banking”. *CESifo Working Paper* 2862.
- Lember, J. (2021). *Tehisõpe I. Loengukonspekt*. Tartu Ülikool, lk. 205–209.
- Lemeshow, S., R. X. Sturdivant ja D. W. Hosmer Jr. (2013). *Applied Logistic Regression*. John Wiley Sons, Inc.
- Pampel, F. C. (2000). “Logistic Regression: A Primer”. *Sage University Papers Series on Quantitative Applications in the Social Sciences* 07-132.
- Pereira, J. M., M. Basto ja A. Ferrera da Silva (2015). “The logistic lasso and ridge regression in predicting corporate failure”. *Procedia Economics and Finance* 39 (2016), lk. 634–641.
- The Investopedia Team (2022). *Credit Risk*. URL: <https://www.investopedia.com/terms/c/creditrisk.asp> (vaadatud 27.04.2022).
- Tutz, G. (2012). *Regression for Categorical Data*. Cambridge.

Lisa 1. Kordajad regressioonmudelites

Tabel 21: Peamõjudega logistilises regressioonis oluliseks osutunud tunnuste kordajad.

Tunnus	Kordaja (β)	p-väärtus
Vabaliige	1,6328	$1,75 \cdot 10^{-7}$
Sugu = Mees	-0,1975	0,0059
Vanus	0,0254	$8,62 \cdot 10^{-15}$
Maakond = Hiiumaa	0,3597	0,6624
Maakond = Ida-Virumaa	-0,4369	$1,27 \cdot 10^{-5}$
Maakond = Järvamaa	0,0511	0,8444
Maakond = Jõgevamaa	-0,4206	0,1486
Maakond = Lääne-Virumaa	-0,3147	0,0715
Maakond = Läänemaa	-0,0150	0,9618
Maakond = Pärnumaa	-0,1155	0,4983
Maakond = Põlvamaa	-0,7349	0,0087
Maakond = Raplamaa	-0,4578	0,0701
Maakond = Saaremaa	-0,3540	0,3895
Maakond = Tartumaa	-0,1089	0,3553
Maakond = Valgamaa	-0,1288	0,6328
Maakond = Viljandimaa	-0,2328	0,2979
Maakond = Võrumaa	-0,2463	0,4552
ln(laenusumma)	-0,3511	$4,96 \cdot 10^{-11}$
Väljaminek	0,0006	$9,99 \cdot 10^{-5}$
Pangakontoinfo = 1	0,5716	$6,75 \cdot 10^{-14}$
Aktiivne maksehäire = 1	-0,4423	$2,87 \cdot 10^{-5}$
Suletud maksehäire = 1	-0,5325	$4,29 \cdot 10^{-14}$
Kinnisvara = 1	0,9507	$< 2 \cdot 10^{-16}$
Haridus = kõrgharidus	0,3386	0,0010
Haridus = kuni põhiharidus	-0,3643	0,0002

Tabel 22: Lasso regressioonis oluliseks osutunud tunnuste kordajad.

Tunnus	Kordaja (β)
Vabaliige	0,0132
Sugu = Mees	-0,1354
Vanus	0,0204
Maakond = Ida-Virumaa	-0,0384
Laenusumma	-0,0004
Väljaminek	0,0004
Perekonnaseis = Vallaline	-0,0689
Töökogemus	0,0129
Pangakontoinfo = 1	0,3466
Aktiivne maksehäire = 1	-0,0474
Suletud maksehäire = 1	-0,3885
Kinnisvara = 1	0,8479
Haridus = kõrgharidus	0,0019
Haridus = kuni põhiharidus	-0,1602
Sugu = Mees; Maakond = Hiiumaa	0,1600
Sugu = Mees; Maakond = Ida-Virumaa	-0,0035
Sugu = Mees; Maakond = Raplamaa	-0,0888
Sugu = Mees; Maakond = Viljandimaa	0,0204
Sugu = Mees; Perekonnaseis = Lesk	-0,1513
Sugu = Mees; Pangakontoinfo = 1	0,0639
Maakond = Põlvamaa; Emakeel = est	-0,0929
Maakond = Tartumaa; Emakeel = est	-0,0412
Maakond = Ida-Virumaa; Perekonnaseis = Lahutatud	-0,1528
Maakond = Lääne-Virumaa; Perekonnaseis = Lahutatud	0,5774
Maakond = Läänemaa; Perekonnaseis = Lahutatud	-0,0418

Tabel 22 jätkub

Tunnus	Kordaja (β)
Maakond = Põlvamaa; Perekonnaseis = Lesk	-2,0931
Maakond = Ida-Virumaa; Perekonnaseis = Vabaabielus	-0,3919
Maakond = Jõgevamaa; Perekonnaseis = Vabaabielus	-0,2323
Maakond = Läänemaa; Perekonnaseis = Vabaabielus	0,1484
Maakond = Viljandimaa; Perekonnaseis = Vabaabielus	-0,2211
Maakond = Ida-Virumaa; Perekonnaseis = Vallaline	-0,0981
Maakond = Lääne-Virumaa; Perekonnaseis = Vallaline	-0,4067
Maakond = Põlvamaa; Perekonnaseis = Vallaline	-0,4457
Maakond = Järvamaa; Pangakontoinfo = 1	0,0958
Maakond = Ida-Virumaa; Aktiivne maksehäire = 1	-0,2700
Maakond = Jõgevamaa; Aktiivne maksehäire = 1	0,1806
Maakond = Pärnumaa; Aktiivne maksehäire = 1	-0,4892
Maakond = Põlvamaa; Aktiivne maksehäire = 1	-0,1475
Maakond = Viljandimaa; Aktiivne maksehäire = 1	-0,5335
Maakond = Võrumaa; Aktiivne maksehäire = 1	-1,6671
Maakond = Põlvamaa; Kinnisvara = 1	-0,2215
Maakond = Tartumaa; Kinnisvara = 1	0,0116
Maakond = Valgamaa; Kinnisvara = 1	-0,2808
Maakond = Jõgevamaa; Haridus = Kõrgharidus	0,2088
Maakond = Pärnumaa; Haridus = Kõrgharidus	0,4820
Maakond = Põlvamaa; Haridus = Kõrgharidus	0,1499
Maakond = Tartumaa; Haridus = Kõrgharidus	0,1579
Maakond = Valgamaa; Haridus = Kõrgharidus	0,1499
Maakond = Ida-Virumaa; Haridus = Kuni põhiharidus	-0,3621
Maakond = Järvamaa; Haridus = Kuni põhiharidus	0,5062
Maakond = Lääne-Virumaa; Haridus = Kuni põhiharidus	-0,2184
Maakond = Saaremaa; Haridus = Kuni põhiharidus	-0,5898

Tabel 22 jätkub

Tunnus	Kordaja (β)
Maakond = Jõgevamaa; Töökogemus	-0,0633
Maakond = Raplamaa; Töökogemus	-0,0496
Emakeel = est; Aktiivne maksehäire = 1	-0,0465
Emakeel = est; Suletud maksehäire = 1	-0,1032
Perekonnaseis = Lahutatud; Pangakontoinfo = 1	0,0247
Perekonnaseis = Vabaabielus; Aktiivne maksehäire = 1	-0,0441
Perekonnaseis = Lesk; Suletud maksehäire = 1	-0,3824
Perekonnaseis = Lesk; Kinnisvara = 1	0,2209
Perekonnaseis = Vallaline; Haridus = Kõrgharidus	0,1551
Perekonnaseis = Vabaabielus; Haridus = Kuni põhiharidus	-0,1937
Pangakontoinfo = 1; Kinnisvara = 1	-0,0917
Aktiivne maksehäire = 1; Kinnisvara = 1	-0,3000
Kinnisvara = 1; Haridus = Kõrgharidus	0,0723
Haridus = Kõrgharidus; Töökogemus	0,0590

Tabel 23: Pea- ja koosmõjudega logistilises regressioonis oluliseks osutunud tunnuste kordajad.

Tunnus	Kordaja (β)	p-väärtus
Vabaliige	1,9175	$4,96 \cdot 10^{-8}$
Sugu = Mees	-0,1892	0,0113
Vanus	0,0237	$6,23 \cdot 10^{-11}$
Maakond = Hiiumaa	0,3537	0,6685
Maakond = Ida-Virumaa	-0,4353	$1,71 \cdot 10^{-5}$
Maakond = Järvamaa	0,0557	0,8308
Maakond = Jõgevamaa	-0,4348	0,1361
Maakond = Lääne-Virumaa	-0,3130	0,0739
Maakond = Läänemaa	0,0137	0,9651
Maakond = Pärnumaa	-0,1076	0,5308
Maakond = Põlvamaa	-0,7297	0,0093
Maakond = Raplamaa	-0,4760	0,0606
Maakond = Saaremaa	-0,3283	0,4236
Maakond = Tartumaa	-0,1019	0,3914
Maakond = Valgamaa	-0,1178	0,6622
Maakond = Viljandimaa	-0,2556	0,2529
Maakond = Võrumaa	-0,2364	0,4749
ln(laenusumma)	-0,3670	$1,06 \cdot 10^{-11}$
Väljaminek	0,0006	0,0001
Perekonnaseis = Lahutatud	-0,0911	0,6864
Perekonnaseis = Lesk	-0,5656	0,1510
Perekonnaseis = Vabaabielus	-0,3123	0,0425
Perekonnaseis = Vallaline	-0,3934	0,0076
Pangakontoinfo = 1	0,7391	$< 2 \cdot 10^{-16}$
Aktiivne maksehäire = 1	-0,2654	0,0402
Suletud maksehäire = 1	-0,7132	$2,40 \cdot 10^{-7}$
Kinnisvara = 1	1,1393	$3,99 \cdot 10^{-14}$
Haridus = kõrgharidus	0,3384	0,0010
Haridus = kuni põhiharidus	-0,3643	0,0002
Perekonnaseis = Lahutatud; Suletud maksehäire = 1	0,0049	0,9860
Perekonnaseis = Lesk; Suletud maksehäire = 1	-0,7320	0,1574
Perekonnaseis = Vabaabielus; Suletud maksehäire = 1	0,2395	0,1984
Perekonnaseis = Vallaline; Suletud maksehäire = 1	0,4086	0,0233
Perekonnaseis = Lahutatud; Kinnisvara = 1	0,1531	0,5828
Perekonnaseis = Lesk; Kinnisvara = 1	1,3315	0,0122
Perekonnaseis = Vabaabielus; Kinnisvara = 1	-0,0258	0,8934
Perekonnaseis = Vallaline; Kinnisvara = 1	0,0188	0,9248
Pangakontoinfo = 1; Kinnisvara = 1	-0,5501	0,0004
Aktiivne maksehäire = 1; Kinnisvara = 1	-0,4699	0,0281

Lisa 2. R-i kood

Peamõjusid sisaldav logistiline regressioon:

```
# Logistiline regressioon peamõjudega
library(stats)
library(lmtest) #lrtest-i jaoks

# Andmed
load("train")
load("test")

m.logit1 = glm(Status ~ Sex+Age+Region+Language+log(Sum)+Period+Income+
               Outcome+ Family+work+Children+PangakontoInfo+AktiivneMaksehaire+
               SuletudMaksehaire+Kinnisvara+Education_new,
               family="binomial"(link = "logit"), data=train)
m.logit1 = update(m.logit1, .~.-Period)
m.logit1 = update(m.logit1, .~.-Language)
m.logit1 = update(m.logit1, .~.-Children)
m.logit1 = update(m.logit1, .~.-Income)
m.logit1 = update(m.logit1, .~.-work)

# Kas tunnus perekonnaseis (Family) peaks sisse jääma?
lrtest(m.logit1, update(m.logit1, .~.-Family))
#p-väärtus > 0.05 --> eelistada lihtsamat mudelit

m.logit1 = update(m.logit1, .~.-Family)
summary(m.logit1)
```

```
# Kas tunnus maakond peaks olema mudelisse kaasatud?
lrtest(m.logit1, update(m.logit1, ~.-Region))

#p-väärtus < 0.05 --> eelistada keerulisemat mudelit

# Lõplik mudel:
summary(m.logit1)
```

LASSO regressioon:

```
# LASSO regressioon
library(glmnet)
library(dplyr)

# Andmed
train_lasso <- train %>%
  mutate(Sex = ifelse(Sex == 'M', 1, 0)) %>%
  mutate(Language = ifelse(Language == 'est', 1, 0))

test_lasso <- test %>%
  mutate(Sex = ifelse(Sex == 'M', 1, 0)) %>%
  mutate(Language = ifelse(Language == 'est', 1, 0))

# Valem
initial_formula = Status ~ Sex+Age+Region+Language+Sum+Period+Income+
  Outcome+Family+work+Children+PangakontoInfo+AktiivneMaksehaire+
  SuletudMaksehaire+Kinnisvara+Education_new +
  Sex:Region+Sex:Language+Sex:Family+Sex:PangakontoInfo+
  Sex:AktiivneMaksehaire+Sex:SuletudMaksehaire+Sex:Kinnisvara+
  Sex:Education_new + Sex:work +Region:Language+Region:Family+
```

```

Region:PangakontoInfo+Region:AktiivneMaksehaire+Region:SuletudMaksehaire+
Region:Kinnisvara+Region:Education_new + Region:work + Language:Family+
Language:PangakontoInfo+Language:AktiivneMaksehaire+Language:SuletudMaksehaire+
Language:Kinnisvara+Language:Education_new + Language:work +
Family:PangakontoInfo+Family:AktiivneMaksehaire+Family:SuletudMaksehaire+
Family:Kinnisvara+Family:Education_new + Family:work +
PangakontoInfo:AktiivneMaksehaire+PangakontoInfo:SuletudMaksehaire+
PangakontoInfo:Kinnisvara+PangakontoInfo:Education_new + PangakontoInfo:work +
AktiivneMaksehaire:SuletudMaksehaire+AktiivneMaksehaire:Kinnisvara+
AktiivneMaksehaire:Education_new + AktiivneMaksehaire:work +
SuletudMaksehaire:Kinnisvara+SuletudMaksehaire:Education_new +
SuletudMaksehaire:work +Kinnisvara:Education_new + Kinnisvara:work +
Education_new:work

x_vol1 <- model.matrix(initial_formula, train_lasso)[,-1]
y_vol1 <- train_lasso$Status
test_x_vol1 <- model.matrix(initial_formula, test_lasso)[,-1]
test_y_vol1 <- test_lasso$Status

# Ristvalideerimisega sobiva lambda väärtuse leidmine
lambdas <- seq(from = 0.001, to = 0.5, by = 0.0005)

set.seed(6878)
cv.lasso <- cv.glmnet(x_vol1, y_vol1, alpha = 1, family = "binomial",
                      type.measure="auc", lambda = lambdas,
                      nfolds = 10)

plot(cv.lasso, type = "l")
cv.lasso$lambda.min #0.0065

```

```
# Lõplik mudel, treenitud kogu treeningandmestikul
m.lasso <- glmnet(x_voll, y_voll, alpha = 1, family = "binomial",
                  lambda = 0.0065)

coef(m.lasso)
```

Pea- ja koosmõjusid sisaldav logistiline regressioon:

```
# Logistiline regressioon pea- ja koosmõjudega
library(stats)

# valem

formula.log.interaction = Status ~ Sex+Age+Region+Language+log(Sum)+Period+
  Income+Outcome+Family+work+Children+PangakontoInfo+AktiivneMaksehaire+
  SuletudMaksehaire+Kinnisvara+Education_new+
  Sex:Region + Sex:Family + Sex:PangakontoInfo +
  Region:Language + Region:Family + Region:PangakontoInfo +
  Region:AktiivneMaksehaire + Region:SuletudMaksehaire + Region:Kinnisvara +
  Region:Education_new + Region:work +
  Language:AktiivneMaksehaire + Language:SuletudMaksehaire +
  Family:PangakontoInfo + Family:AktiivneMaksehaire + Family:SuletudMaksehaire +
  Family:Kinnisvara + Family:Education_new +
  PangakontoInfo:SuletudMaksehaire + PangakontoInfo:Kinnisvara +
  AktiivneMaksehaire:Kinnisvara + Kinnisvara:Education_new + work:Education_new

# Kahaneev sammregressioon

logit.interaction <- glm(formula.log.interaction,
                          family="binomial"(link = "logit"),
                          data=train)

logit.interaction = update(logit.interaction, ~.-Region:AktiivneMaksehaire)
```

```

logit.interaction = update(logit.interaction, .~.-Family:AktiivneMaksehaire)
logit.interaction = update(logit.interaction, .~.-work:Education_new)
logit.interaction = update(logit.interaction, .~.-Region:SuletudMaksehaire)
logit.interaction = update(logit.interaction, .~.-Region:PangakontoInfo)
logit.interaction = update(logit.interaction, .~.-Region:Family)
logit.interaction = update(logit.interaction, .~.-Region:Language)
logit.interaction = update(logit.interaction, .~.-Sex:Family)
logit.interaction = update(logit.interaction, .~.-Sex:Region)
logit.interaction = update(logit.interaction, .~.-PangakontoInfo:SuletudMaksehaire)
logit.interaction = update(logit.interaction, .~.-Language:AktiivneMaksehaire)
logit.interaction = update(logit.interaction, .~.-Region:PangakontoInfo)
logit.interaction = update(logit.interaction, .~.-Family:PangakontoInfo)
logit.interaction = update(logit.interaction, .~.-Language)
logit.interaction = update(logit.interaction, .~.-Region:PangakontoInfo)
logit.interaction = update(logit.interaction, .~.-Children)
logit.interaction = update(logit.interaction, .~.-Sex:PangakontoInfo)
logit.interaction = update(logit.interaction, .~.-Period)
logit.interaction = update(logit.interaction, .~.-Region:work)
logit.interaction = update(logit.interaction, .~.-Family:Education_new)
logit.interaction = update(logit.interaction, .~.-Region:Education_new)
logit.interaction = update(logit.interaction, .~.-Kinnisvara:Education_new)
logit.interaction = update(logit.interaction, .~.-Income)
logit.interaction = update(logit.interaction, .~.-work)
logit.interaction = update(logit.interaction, .~.-SuletudMaksehaire:Language)
logit.interaction = update(logit.interaction, .~.-Region:Kinnisvara)
summary(logit.interaction)

```

Klassifitseerimispuu (CART):

```
# Klassifitseerimispuu (CART)

library(tidymodels)
library(rpart)
library(rpart.plot)
library(dplyr)

# Andmed
train_tree <- train %>%
  mutate(Status = as.factor(Status)) %>%
  mutate(Sex = as.factor(Sex)) %>%
  mutate(Region = as.factor(Region)) %>%
  mutate(Language = as.factor(Language)) %>%
  mutate(Family = as.factor(Family)) %>%
  mutate(Education_new = as.factor(Education_new))

test_tree <- test %>%
  mutate(Status = as.factor(Status)) %>%
  mutate(Sex = as.factor(Sex)) %>%
  mutate(Region = as.factor(Region)) %>%
  mutate(Language = as.factor(Language)) %>%
  mutate(Family = as.factor(Family)) %>%
  mutate(Education_new = as.factor(Education_new))

# 10-kordne ristvalideerimine sobiva puu sügavuse leidmiseks
set.seed(6879)

m.tree <- rpart::rpart(Status~ Sex + Region + Language + Family + work +
  PangakontoInfo + AktiivneMaksehaire +
```

```

        SuletudMaksehaire + Kinnisvara + Education_new +
        Sum + Income + Outcome + Period + Age + Children,
        data=train_tree,cp=0,minsplrit=10,method="class", xval=10)

# cp = 0 - complexity parameter
# minsplrit = 10 - väikseim arv vaatlusi lehes
# xval - 10-kordne ristvalideerimine
printcp(m.tree)
as_tibble(m.tree$cptable) %>% filter(xerror==min(xerror))
m.tree$cptable # optimaalsel puul: cp = 0.0037313433

# Puu pügamine
m.tree_pruned <- rpart::prune(m.tree,cp=0.0037313433)

```

Gradient Boosting:

```

# Gradient Boosting
library(caret)
library(dplyr)

# Andmestikud
train_gbm = train %>%
  mutate(Sex = as.factor(Sex)) %>%
  mutate(Region = as.factor(Region)) %>%
  mutate(Language = as.factor(Language)) %>%
  mutate(Family = as.factor(Family)) %>%
  mutate(Education_new = as.factor(Education_new))
test_gbm = test %>%
  mutate(Sex = as.factor(Sex)) %>%
  mutate(Region = as.factor(Region)) %>%

```

```

mutate(Language = as.factor(Language)) %>%
mutate(Family = as.factor(Family)) %>%
mutate(Education_new = as.factor(Education_new))

# Ristvalideerimine
fitControl <- trainControl(method = "cv", number = 5)
gbmGrid <- expand.grid(interaction.depth = c(1, 2, 3, 4, 5),
                      n.trees = c((1:20)*250),
                      shrinkage = c(0.01, 0.025, 0.05, 0.075, 0.1),
                      n.minobsinnode = 10)

nrow(gbmGrid) #500

set.seed(6879)
cv.gbm <- train(as.factor(Status) ~ Sex + Region + Language + Family + work +
               PangakontoInfo + AktiivneMaksehaire + SuletudMaksehaire +
               Kinnisvara + Education_new + Sum + Income + Outcome +
               Period + Age + Children,
               data = train_gbm,
               method = "gbm",
               trControl = fitControl,
               verbose = FALSE,
               tuneGrid = gbmGrid)

cv.gbm
results.cv.gbm <- cv.gbm$results
results.cv.gbm <- results.cv.gbm %>%
  arrange(desc(Accuracy))
View(results.cv.gbm)
# optimaalsed parameetrid:
# shrinkage = 0.025

```

```

# interaction.depth = 2
# n.trees = 2000

# Lõplik mudel
set.seed(6879)
m.gbm <- train(as.factor(Status) ~ Sex + Region + Language + Family + work +
               PangakontoInfo + AktiivneMaksehaire + SuletudMaksehaire +
               Kinnisvara + Education_new + Sum + Income + Outcome +
               Period + Age + Children,
               data = train_gbm,
               method = "gbm",
               trControl = trainControl(method = "none"),
               verbose = FALSE,
               tuneGrid = data.frame(interaction.depth = 2,
                                     n.trees = 2000,
                                     shrinkage = 0.025,
                                     n.minobsinnode = 10),
               metric = "Accuracy")

```

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Annela Pajumets,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Krediidiriski hindamine valdavalt kategooriaalsete andmete põhjal“, mille juhendaja on 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 Commons'i 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.

Annela Pajumets

17.05.2022