TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Riivo Kiljak 232086TARM
Eva-Anna Klugman 231850IABM
Jolanta Rudus 231851IABM

Bondora laenuvõtmise platvormi laenuvõtjate maksevõimekuse analüüs

Uurimisprojekt

Juhendaja: Heili Hein

Tallinn 2023

Sisukord

1 Uurimisprobleem 3
2 Andmed 4
2.1 Kasutatud tunnused 4
3 Metoodika 6
4 Tulemused ja järeldused 7
Kasutatud kirjandus 11
Lisa 1 – Kvantitatiivsete tunnuste statistilised näitajad 12
Lisa 2 – Kvalitatiivsete tunnuste ülevaade 13

1 Uurimisprobleem

Laenuvõtmise platvormid seisavad silmitsi pideva väljakutsega hinnata laenutaotlejate maksevõimekust. Antud uurimistöö eesmärk on analüüsida peamisi tegureid, mis mõjutavad laenudele võlgnevuse tekkimist Bondora platvormil. Bondora on keskkond laenude vahendamiseks, mis loodi 2008. aastal. Platvorm võimaldab taotleda laenu iga päev ööpäevaringselt. Selle missiooniks on võimaldada inimestel elada oma unistuste elu raha pärast muretsemata. [1]

Laenuturgudel on laenuvõtjate maksevõimekuse hindamine oluline, et vähendada võimalikult efektiivselt laenuandjate riske ja tagada selliste ettevõtete jätkusuutlik finantsseis. Varasemad uuringud on näidanud, et mitmed tegurid – alates sissetulekust ja haridusest kuni laenude suuruse ning muude demograafiliste andmeteni – võivad oluliselt mõjutada laenuvõtjate tagasimaksevõimet. [2]

Käesoleva uurimistöö uurimisküsimused on järgmised:

1. Mis järgnevatest mudelitest osutub kõige efektiivsemaks maksevõimetuse ennustamiseks: logistiline regressioon, juhuslik mets (*random forest*) või XGBoost?

2. Millised on peamised tegurid, mis mõjutavad laenuvõtjate maksevõimekust Bondora laenuplatvormil?

Teises uurimisküsimuses nimetatud meetodid on valitud nende võimekuse tõttu hinnata keerukaid seoseid erinevate tegurite vahel ning anda täpseid prognoose laenuvõtjate maksevõimekuse kohta. Sellised järeldused tehti teadusartiklite tulemuste põhjal, milles kirjeldati nende eeliseid võrreldes teiste mudelitega. Teadusartikliteks võeti “Study comparing classification algorithms for loan approval predictability (Logistic Regression, XG boost, Random Forest, Decision Tree)” [3]“, Bankruptcy Prediction using the XGBoost Algorithm and Variable Importance Feature Engineering” [4], “A study on predicting loan default based on the random forest algorithm” [5] ja “Loan default prediction using decision trees and random forest: A comparative study” [6].

2 Andmed

Andmed on pärit Bondora avalike aruannete lehelt ning hõlmavad erinevaid tunnuseid, mis võimaldavad analüüsida laenuvõtjate maksevõimekust [7]. Andmefail oli CSV formaadis ning see laeti Bondora keskkonnast alla 17. novembril 2023 kell 14.20. Andmed uuenevad Bondora veebis igapäevaselt.

Esialgsetes andmetes oli 340 704 vaatlust ja 112 tunnust. Kuna töö eesmärgiks on maksevõimekuse (maksevõimetuse tõenäosuse) ennustamine, otsustati laenude müümise ja reitingutega seotud muutujad välja jätta. Samuti jäeti välja muutujad, mida Bondora kasutab laenude identifitseerimiseks. Peale valitud teemaga mitteseonduvate muutujate väljajätmist, eemaldati andmetest omapärased vaatlused. Nende eemaldamisel katsetati kahte meetodit: omapäraste vaatluste eemaldamine standardhälbe põhjal ning kvartillide põhjal. Kuigi mõlemad meetodid olid sobivad, otsustati eemaldada omapärased vaatlused kvartillide põhjal, kuna selle meetodi lõppandmed võimaldasid saavutada suuremat ennustusvõimet nendel treenitud mudelites.

Kvalitatiivsete tunnuste kategooriad, milles oli vähem kui 10 vaatlust, ühendati omavahel kategooriasse `Other`. Kõik kvantitatiivsed tunnused, mille kõik väärtused olid nullid, kustutati. Lisaks eelnevale eemaldati tunnused, milles oli väga tugev vaatluste ülekaal mingis kindlas kategoorias.

Maksevõimetuse hindamiseks loodi uus kvalitatiivne muutuja `Defaulted`, mille väärtuteks oli `TRUE` või `FALSE`. Väärtused loodi `DefaultDate` tunnuse põhjal – kui see polnud tühi, järelikult läks laen pankrotti ja sai väärtuseks `TRUE` ning vastupidi.

Peale andmete puhastamist jäi alles 244 278 vaatlust ja 17 muutujat.

2.1 Kasutatud tunnused

Kvantitatiivsed tunnused

1. “`Amount`” ehk esialgselt taotletud laenusumma;

2. “AmountOfPreviousLoansBeforeLoan” ehk varasemalt võetud laenude summa;
3. “AppliedAmount” ehk summa, mis Bondora välja laenas;
4. “Age” ehk laenuvõtja vanus;
5. “ExistingLiabilities” ehk laenuvõtja olemasolevad kohustused;
6. “IncomeTotal” ehk laenuvõtja kogusissetulek;
7. “LiabilitiesTotal” ehk kõik laenaja kohustused taotlemise hetkel;
8. “LoanDuration” ehk väljaantud laenu lepinguline kestus;
9. “MonthlyPayment” ehk väljaantud laenu kuumakse;
10. “NoOfPreviousLoansBeforeLoan” ehk eelnevalt võetud laenude arv.

Kvantitatiivsete muutujate andmete kirjeldavat statistikat on võimalik näha Lisast 1.

Kvalitatiivsed tunnused:

11. “Gender” ehk laenuvõtja sugu;
12. “Education” ehk laenuvõtja haridustase;
13. “LanguageCode” ehk portaali keel laenu võtmisel;
14. “Country” ehk laenuvõtja riik;
15. “EmploymentDurationCurrentEmployer” ehk laenuvõtja olemasolevas töökohas töötamise aeg;
16. “HomeOwnershipType” ehk laenuvõtja koduomandi tüüp;
17. “Defaulted” ehk kas välja antud laen tasuti täielikul või mitte.

Kvalitatiivsete muutujate ülevaadet on võimalik näha Lisast 2.

3 Metoodika

Maksevõimetuse prognoosimiseks kasutati kolme mudelit: logistiline regressioon, juhuslik mets (*random forest*) ja XGBoost. Mudeleid treeniti treeningandmetel, mis moodustasid ¾ andmete koguhulgast. Andmete jaotamiseks treening- ja testvalimiks kasutati funktsiooni `initial_split` teegist `tidymodels`. Antud funktsioon võimaldab jagada andmed kahte gruppi, säilitades samasugust maksevõimetute laenude vaatluste osakaalu mõlemas valimis. Mudelite võrdlemisel kasutati järgmiseid võtmemõõdikuid:

- Mudeli prognoosimistäpsus testandmetel. Selle mõõtmiseks kasutati funktsiooni `accuracy` teegist `yardstick`, mis oli `tidymodels`'ga kaasa antud.

- ROC-kõvera aluse pindala osakaal. Selle mõõtmiseks kasutati funktsiooni `roc_auc`, mis oli samuti `yardstick` teegist.

Logistilise regressiooni teostamiseks kasutati R keele funktsiooni `glm`, mille `family` parameetri väärtuseks on `binomial`. Esimese mudeli järgi osutusid kõik tunnused olulisteks peale laenuvõtjate vanuse. Mudeli koostamist korrati uuesti. Selleks eemaldati andmetest vanuse muutuja ning seejärel jagati andmed uuesti test-ja treeningandmeteks.

Juhusliku metsa mudeli loomiseks kasutati funktsiooni `rand_forest tidymodels` teegist. `Tidymodels` teek omab teiste R keele teekide jaoks liidese rolli. Juhusliku metsa modelleerimisel kasutati `ranger` teegi funktsionaalsust ning mudelis olevate otsustus-puude arvuks osutus 500.

XGBoosti mudeli koostamiseks kasutati funktsiooni `boost_tree`, mis oli samuti `tidymodels` teegist. Mudeli koostamisel kasutati samu parameetreid nagu juhusliku metsa mudeli loomise puhul, milleks olid puude arv 500 ja `mode=classification`.

4 Tulemused ja järeldused

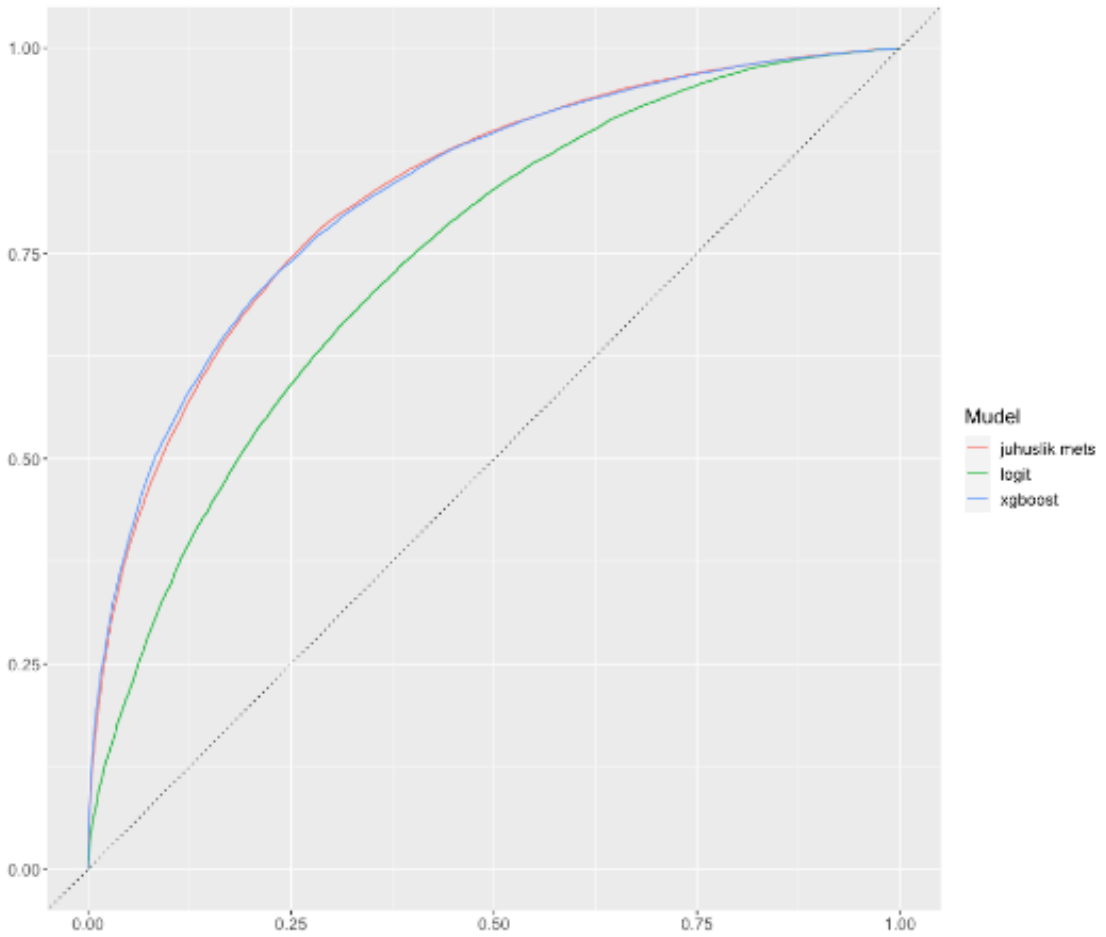
Esimese uurimisküsimuse vastamiseks loodi kolm mudelit ning võrreldi nende võtmemõõdikuid, et hinnata mudelite prognoosimisvõimet. Võtmemõõdikute väärtuseid on võimalik näha Tabelis 1.

Tabel 1. Mudelite võtmemõõdikute väärtused.

Mudel	**Prognoosimistäpsus testandmetel**	**ROC-kõvera aluse ala osakaal**
Logistiline regressioon	73,2%	74,3%
Juhuslik mets	77,5%	82,3%
XGBoost	77,5%	82,4%

Võtmemõõdikute järgi andis kõige halvema tulemuse logistilise regressiooni mudel. Vaatamata sellele, oli ka logit mudeli prognoosimisvõime suhteliselt täpne. Juhusliku metsa ja XGBoost mudelid on võtmemõõdikute järgi peaaegu võrdväärsed. Sellise järelduse saab teha ka Joonise 1 põhjal, kus on kujutatud kõikide mudelite ROC-kõverad.

Tabelid 2 ja 3 esitavad juhusliku metsa ja XGBoost mudelite segadusmaatrikseid. Juhuslik mets kipub maksevõimet ülehindama. See oli oodatav, kuna vaatluste hulgas oli ülekaal maksevõimelistel laenudel. XGBoost mudeli prognoosid on võrreldes juhusliku metsaga rohkem tasakaalus. Sellest tulenevalt võib eeldada, et suur andmemaht võimaldas vältida üle sobitamist, mis on antud meetodi puhul risk. Tuginedes ROC-kõvera aluse pindalale ja segadusmaatriksitele saab järeldada, et parimaks mudeliks maksevõimetuse prognoosimiseks on XGBoost, mis vastab antud töö esimesele uurimisküsimusele.



Joonis 1. Mudelite ROC-kõverad.

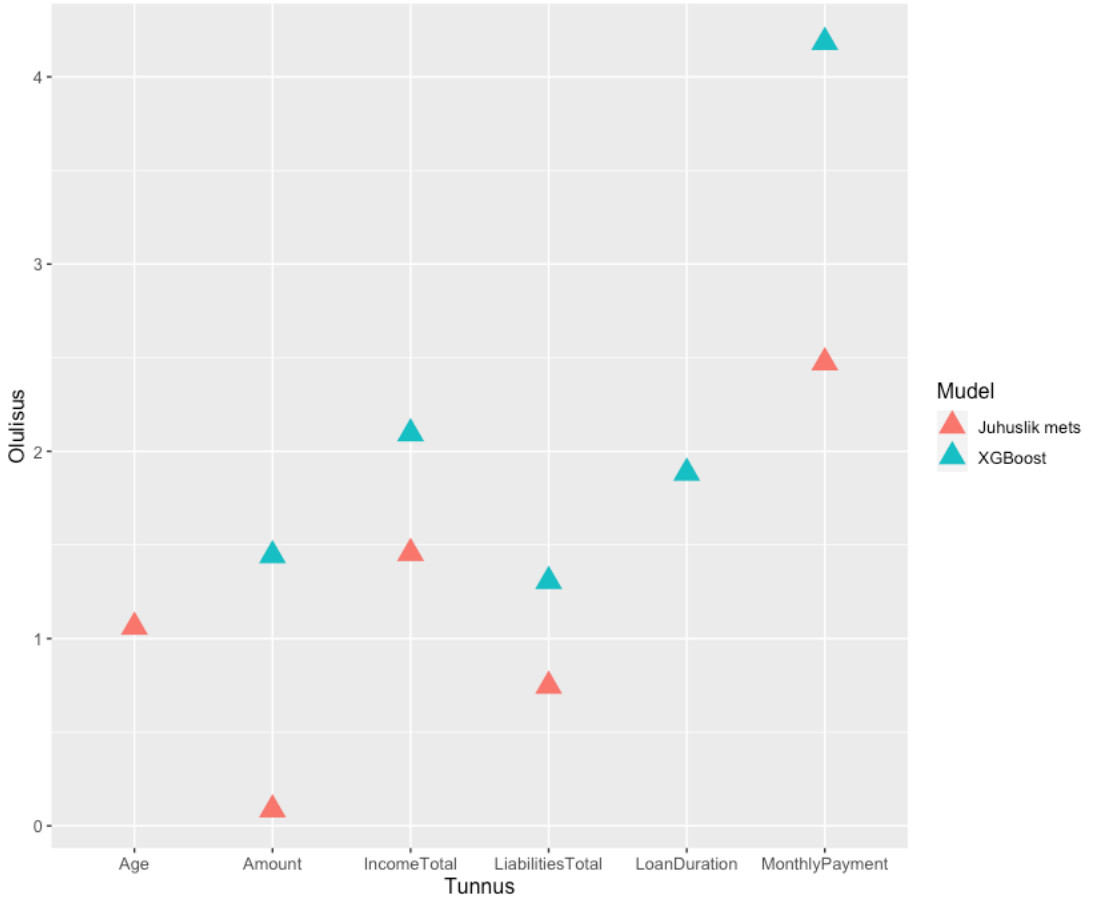
Tabel 2. Juhusliku metsa segadusmaatriks.

Ennustus	**FALSE**	**TRUE**
FALSE	35904	8400
TRUE	4900	10024

Tabel 3. XGBoost mudeli segadusmaatriks.

Ennustus	**FALSE**	**TRUE**
FALSE	36789	9327
TRUE	4015	9097

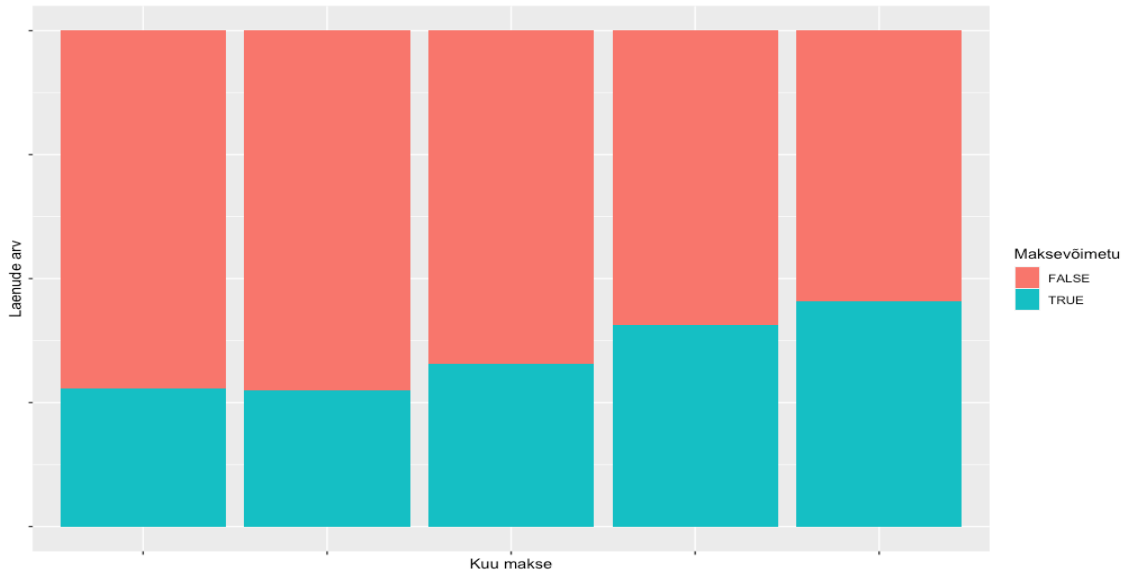
Teisele uurimisküsimusele vastamiseks võrreldi tunnuste olulisust juhusliku metsa ja XGBoost mudelites, kuna need osutusid tunduvalt täpsemateks kui logistilise regressiooni mudel. Viis kõige olulisemat tunnust on esitatud Joonisel 2.



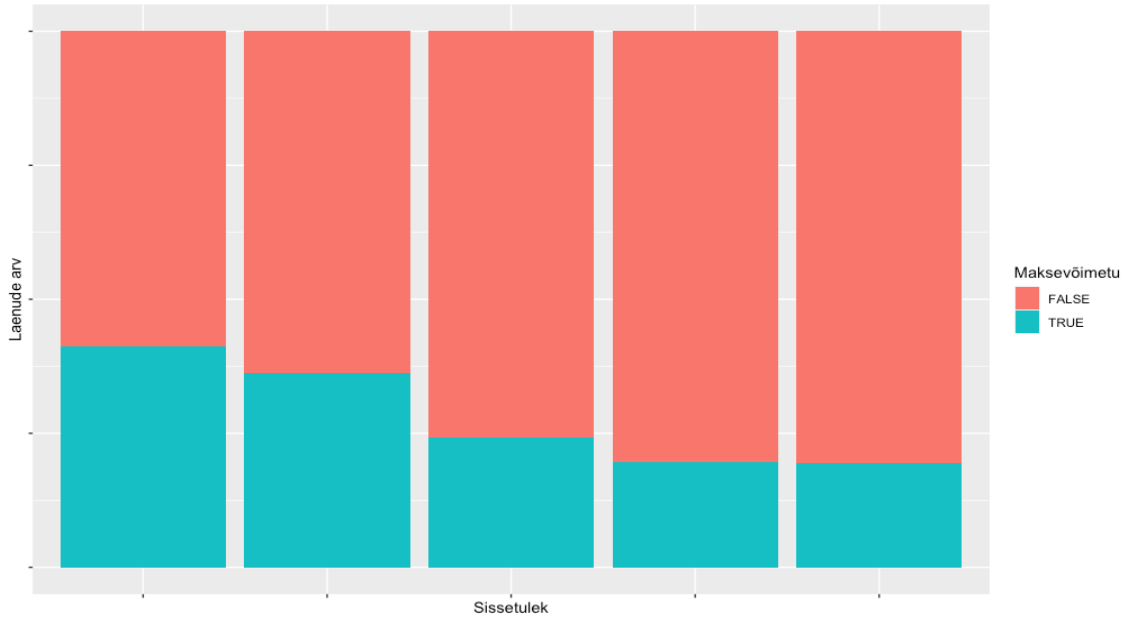
Joonis 2. Tunnuste olulisus juhusliku metsa ja XGBoost mudelites.

Mõlemas mudelis kaheks kõige olulisemaks tunnuseks osutusid `MonthlyPayment` ja `IncomeTotal`. Joonistel 3 ja 4 on näha maksevõimetute laenude osakaalude muutust sõltuvalt nende tunnuste väärtusest. Ülevaatlikkuse eesmärgil on vaatlused jaotatud viite võrdsesse valimisse.

Mõlema mudeli kolmandaks kõige olulisemaks tunnuseks osutus tunnus, mis teise mudeli puhul ei pääsenud viie olulisema tunnuse hulka. Juhusliku metsa mudelis selleks tunnuseks on `Age` ehk vanus ja XGBoost mudelis `LoanDuration` ehk laenu pikkus kuudes. Neljandaks ja viiendaks tunnuseks olulisuse järgi mõlemas mudelis olid `LiabilitiesTotal` ehk võlakohustuste summaarne arv isiku kohta ning `Amount` ehk laenu kogusumma.



Joonis 3. Maksevõimetute laenude osakaalu muutus kuise makse kasvamisel.



Joonis 4. Maksevõimetute laenude osakaalu muutus sissetuleku kasvamisel.

Suurimaks piiranguks antud tulemuste analüüsi puhul on tunnuste omavaheline korrelatsioon. On oodatav, et tunnuste vahel, nagu laenu suurus, kestvus ja igakuine makse võib esineda multikollineaarsus. Selline seos mõjub negatiivselt nii regressiooni tulemuste usaldusväärsusele kui ka otsustuspuu mudelite tulemuste tõlgendamisele. Kuna oluliste tunnuste pingerea tipus on mitu seotud tunnust, on töö autoritel põhjust kahtlustada, et mingisugune muu tunnus, mis on samuti oluline, jäi nende tunnuste tõttu märkamatuks. Veel peab arvesse võtma, et töös kasutatud andmetesse jõudsid ainult Bondora poolt välja antud laenud ehk puuduvad laenutaotlused, mida Bondora ei rahuldanud pidades maksevõimetuse riski liiga kõrgeks.

Kasutatud kirjandus

[1] Bondora, „Bondorast,“ bondora.ee, [Võrgumaterjal]. Available: https://www.bondora.ee/bondorast/. [Kasutatud 16 detsember 2023].

[2] C. Taimre, „Laenuvõtja maksejõuetuse modelleerimine,“ 2015. [Võrgumaterjal]. Available: https://dspace.ut.ee/server/api/core/bitstreams/4a4a288f-39a3-4bd7-9347-dc0cd96b4cec/content. [Kasutatud 16 detsember 2023].

[3] P. K. Reddy ja D. R, „Study comparing classification algorithms for loan approval predictability (Logistic Regression, XG boost, Random Forest, Decision Tree),“ 2023. [Võrgumaterjal]. Available: https://sifisheriessciences.com/journal/index.php/journal/article/view/475/458. [Kasutatud 22 detsember 2023].

[4] Bondora, „Public Reports,“ bondora.com, 17 detsember 2023. [Võrgumaterjal]. Available: https://www.bondora.com/et/public-reports. [Kasutatud 17 detsember 2023].

[5] S. B. Jabeur, N. Stef ja P. Carmona, „Bankruptcy Prediction using the XGBoost Algorithm and Variable Importance Feature Engineering,“ 23 jaanuar 2022. [Võrgumaterjal]. Available: https://link.springer.com/article/10.1007/s10614-021-10227-1. [Kasutatud 22 detsember 2023].

[6] L. Zhu, D. Qiu, D. Ergu, C. Ying ja K. Liu, „A study on predicting loan default based on the random forest algorithm,“ 2019. [Võrgumaterjal]. Available: https://www.sciencedirect.com/science/article/pii/S1877050919320277?ref=pdf_download&fr=RR-2&rr=83995be95fd25c28. [Kasutatud 22 detsember 2023].

[7] M. Madaan, A. Kumar, C. Keshri, R. Jain ja P. Nagrath, „Loan default prediction using decision trees and random forest: A comparative study,“ 2020. [Võrgumaterjal]. Available: https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012042/pdf. [Kasutatud 22 detsember 2023].

Lisa 1 – Kvantitatiivsete tunnuste statistilised näitajad

Tunnus	Keskmine	Standardhälve	Mediaan	Min	Max
Amount	2445,35	1684,97	2126	103	9037
AmountOfPreviousLoansBeforeLoan	2484,57	3479,79	518	0	13490
AppliedAmount	2467,34	1696,97	2126	103	9037
Age	40,21	12,42	39	18	70
ExistingLiabilities	1,72	1,64	1	0	8
IncomeTotal	1709,53	839,03	1550	0	4355
LiabilitiesTotal	234,54	258,15	152,86	0	1183,18
LoanDuration	49,95	15,87	60	3	96
MonthlyPayment	88,27	56,80	89,98	0	255,32
NoOfPreviousLoansBeforeLoan	1,07	1,35	1	0	5

Lisa 2 – Kvalitatiivsete tunnuste ülevaade

Gender	Kirjeldus	Vaatluste arv
0	Mees	134872
1	Naine	95644
2	Määramata	13762

Education	Kirjeldus	Vaatluste arv
-1	Määramata	7696
1	Algharidus	33938
2	Põhiharidus	1124
3	Kutseharidus	83424
4	Keskharidus	68296
5	Kõrgharidus	49800

LanguageCode	Kirjeldus	Vaatluste arv
1	Eesti keel	77670
2	Inglise keel	2734
3	Vene keel	26399
4	Soome keel	111113
6	Hispaania keel	21579
19	Hollandi keel	4777
Other	Keeled, mis ei kuulunud Bondora keelte loetelu hulka	6

Country	Kirjeldus	Vaatluste arv
EE	Eesti	105183
ES	Hispaania	21623
FI	Soome	112621
NL	Holland	4851

EmploymentDurationCurrentEmployer	Kirjeldus	Vaatluste arv
(Tühi)	Määramata	7218
TrialPeriod	Katseajal	173
UpTo1Year	< 1 aasta	46919
UpTo2Years	< 2 aastat	949
UpTo3Years	< 3 aastat	757
UpTo4Years	< 4 aastat	523
UpTo5Years	< 5 aastat	67609
MoreThan5Years	> 5 aastat	82441
Retiree	Pensionil	18810

HomeOwnershipType	Kirjeldus	Vaatluste arv
0	Kodutu	79
1	Omanik	86508
2	Elab koos vanematega	28095
3	Mööbliga üürnik	86403
4	Mööblita üürnik	609
5	Sotsiaaleluase	1757
6	Üürib kellegagi	220
7	Ühisomand	551
8	Hüpoteeklaenuga omanik	16920
9	Omaniku koorem	138
10	Muu	22998

Defaulted	Kirjeldus	Vaatluste arv
FALSE	Laen maksti tagasi või on graafikus	168995
TRUE	Tekkis maksevõimetus	75283