Loan Approval Prediction

Madis Puu (Rühm 4)

Rando Roosik (Rühm 6)

Ülesanne 2

Oma ärieesmärkide tuvastamine

Taust

Laenude väljastamine eraisikutele on finantssektori oluline ja keeruline protsess, mis mõjutab nii pankade kui ka laenu taotlejate huve. Pankade eesmärgiks on alati tagada tasuvus, minimeerida finantsriske ja pakkuda kõrgetasemelist teenindust, samal ajal püüdes rahuldada klientide vajadusi ja soove. Siiski, laenutaotluste hindamine on tihti manuaalne ja ajamahukas protsess, mis võib tugineda vananenud või ebaefektiivsetele meetoditele. Laenutaotluste sobivuse hindamine eeldab põhjalikku analüüsi laenutaotleja finantsseisundi, makseajalugu ja usaldusväärsuse kohta. See kõik võib olla üsna töömahukas ja ebatäpne, kui seda tehakse käsitsi või tuginedes ainult traditsioonilistele hindamismeetoditele. Seetõttu on vajalik täiustada otsustusprotsessi täpsust ja kiirust, et luua efektiivsemad lahendused laenuhindamiseks.

Ärieesmärgid

Projektis püstitatud eesmärk on automatiseerida laenutaotluste hindamine, et ennustada väga täpselt need isikud, kes ei ole kvalifitseeritud laenu saama. Selline lähenemine võimaldab vähendada vajadust käsitsi hindamise järele ning muudab laenuväljastamise protsessi sujuvamaks ja kiiremaks. Lisaks sellele on oluline parandada kliendikogemust, pakkudes kiiremaid ja täpsemaid laenuotsuseid. Pank saab ennustada laenutaotlejate sobivuse ning pakkuda neile paremaid lahendusi.

Samuti tuleb keskenduda finantsriskide minimeerimisele, et tuvastada tõhusamalt need taotlejad, kellel on suurem tõenäosus maksejõuetuseni jõuda. Kõrgema riskiga laenuvõtjate varajane tuvastamine aitab pankadel vältida suuremaid kahjusid, võimaldades neil vältida laenude väljastamist isikutele, kellel on madal tagasimaksmise usaldusväärsus.

Äri edu kriteeriumid

Projekt on edukas, kui suudame ennustada laenu sobivuse väga kõrge täpsusega. Eriti oluline on see, et vale-negatiivsete juhtumite arv oleks väga madal. Vale-negatiivne tähendab seda, et klient, kellel tegelikult on õigus laenu saada, ei saa seda süsteemi vea tõttu. Kui vale-negatiivsete juhtumite protsent on liiga suur, võib see põhjustada tulu kaotuse, kuna pank jääb ilma võimalikest laenutuludest. Samas on oluline, et algoritm ennustaks piisavalt palju päris-negatiivseid vastuseid, sest seda suurem on siis ka säästetud aeg.

Oma olukorra hindamine

Ressursside inventuur

Meie projekti tiim koosneb kahest inimesest, kellel on baasteadmised andmeteadusest ja masinõppest. Projektitööd arendame koostöös, kasutades Kaggle'i andmestikke (test.csv,

train.csv), kus on võimalik õppida teiste probleemidest. Meil on olemas vajalikud arvutivahendid – kooli sülearvutid ja isiklikud lauaarvutid – et viia läbi vajalikud analüüsid ja töötada läbi masinõppe algoritme.

Nõuded, eeldused ja piirangud

Projekti tähtaeg on 13. detsember, mis seab ajapiirangu. Seetõttu peame tagama, et meie mudel on piisavalt täpne ja toimiv enne tähtaega. Lisaks on oluline, et projekt sisaldaks ka postrit, mis tutvustab meie töö tulemusi ja protsessi teistele huvilistele.

Riskid ja ettenägematud asjaolud

Projektiga kaasnevad mõned riskid, mis on seotud andmete kvaliteedi ja modelleerimise täpsusega. Masinõppe mudelid ei pruugi olla täiuslikud ning võib tekkida vajadus täiendavate andmete või algoritmide järele. Täiendavad riski faktorid võivad hõlmata ka ajapiiranguid ja tööjaotuse efektiivsust, kuna tegeleme projektiga kahekesi.

Terminoloogia

Mõisted, mida kasutame oma projektis:

• Vale-negatiivne: Laenu taotleja, kes oleks pidanud laenu saama, kuid ei saanud süsteemi vea tõttu.

• Vale-positiivne: Laenu taotleja, kellele ei oleks pidanud laenu andma, kuid süsteem annab positiivse otsuse.

• Masinõpe: Meetod, kus arvuti õpib andmete põhjal tegemas ennustusi või otsuseid, ilma et oleks vajadust iga samm ette näha.

Kulud ja tulud

Projekti elluviimine ei too endaga kaasa suuri kulutusi. Me kasutame olemasolevaid tööriistu, näiteks Kaggle’i platvormi tasuta versioone ja kooli sülearvuteid. See tähendab, et projekti põhikulud on seotud ainult aja ja ressursi panusega.

Andmekaevandamise eesmärkide määratlemine

Andmekaevandamise eesmärgid

Data-mining eesmärkideks on kasutada masinõppe ja statistiliste meetodite abil suurte andmekogumite analüüsimist, et tuvastada mustrid ja suundumused, mis aitavad meil automaatselt hinnata laenutaotlejate sobivust laenu saamiseks. Peamine eesmärk on arendada täpsed ennustusmudelid, mis suudavad ennustada laenutaotlejate maksejõuetuse riski ja sellega aidata määrata, milliseid taotlusi tuleks käsitleda ja milliseid mitte

Andmekaevandamise edukriteeriumid

Täpsus (Accuracy): Mudel peab olema piisavalt täpne, et vähendada vale-negatiivide ja vale-positiivide arvu. Vale-negatiivide arv peab olema minimaalne, kuna see tähendab, et laenu taotleja, kellel on õigus laenu saada, jääb selleks ilma. Vale-positiivide arv peab olema samuti kontrollitud, et mitte põhjustada liigset töökoormust ja manuaalset hindamist, mis on seotud valede otsustega.

Ülesanne 3

Andmete kogumine

Andmete nõuete määramine

Analüüsi jaoks vajalikud andmed peavad kajastama laenutaotlejate isikuandmeid, finantsolukorda ja laenuomadusi. Vajalikud väljad peavad andma ulatusliku ülevaate iga taotleja taustast ja laenuga seotud andmetest. Eesmärk on analüüsida mustreid, mis on seotud laenude heakskiidu ja maksehäiretega, lähtudes taotlejate profiilidest.

Kontrollige andmete saadavust

Antud andmestik sisaldab mitmeid üksikasju, mis seonduvad laenutaotlejate isikuomaduste ja laenutingimustega. Andmestik on kinnitatud kergesti kergendamiseks ja sisaldab piisavalt andmeid, et analüüsida taotlejate omaduste ja laenu tulemuslikkuse seoseid.

Valikukriteeriumide määramine

Analüüsi valikukriteeriumid hõlmavad selliste andmete hindamist, mis pakuvad piisavat andmekogust tähenduslike järelduste tegemiseks. Eesmärk on hinnata laenutaotlejaid, kes on esitanud laenutaotlusi erinevate summade, intressimäärade ja staatustega, samuti hinnata erinevatest sissetulekutasemetest ja krediidiajaloo pikkusest lähtuvaid mustreid.

Veergude kirjeldused

id: Unikaalne identifikaator iga kirje jaoks.

person_age: Isiku vanus, kategooriline jaotamine vanusevahemikeks.

person_income: Isiku sissetulek, kategooriline jaotamine sissetulekuvahemikeks.

person_home_ownership: Kodu omamise seisund, mis hõlmab kategooriaid nagu 'RENT', 'MORTGAGE' jne.

person_emp_length: Isiku tööhõive pikkus, kategooriline jaotamine aastate põhjal.

loan_intent: Laenu eesmärk, kategooriad nagu 'HARIDUS', 'MEDITSIIN' jne.

loan_grade: Laenu krediitklass, näiteks 'A', 'B' jne.

loan_amnt: Laenusumma, kategooriline jaotamine.

loan_int_rate: Laenu intressimäär, kategooriline jaotamine protsendivahemikeks.

loan_percent_income: Laenu osa, mis on isiku sissetulekust, kategooriline jaotamine vahemikeks.

cb_person_default_on_file: Kas isikul on olnud laenu maksehäire ajalugu, väärtused 'true' või 'false'.

cb_person_cred_hist_length: Isiku krediidiajaloo pikkus, kategooriline jaotamine.

loan_status: Laenu staatus, väärtused, mis näitavad, kas laen on heaks kiidetud (binaarsed väärtused).

Andmete uurimine

Andmete uurimise etapis (EDA) tuleks analüüsida andmete jaotust, korrelatsioone ja mustreid. Näiteks:

Vanus ja laenustaatus: Kontrollida, kas teatud vanuserühmad on rohkem kalduvad maksehäiretele.

Sissetulek ja laenusumma: Uurida, kas kõrgemad sissetulekud on seotud suuremate laenusummade taotlemisega.

Kodu omamine ja laenude maksehäired: Hinnata, kas koduomanikud on vähem tõenäoliselt maksehäiretega võrreldes rentnikega.

Krediidiajalugu ja laenustaatus: Kontrollida, kuidas krediidiajaloo pikkus mõjutab laenu heakskiitmist ja maksehäirete tõenäosust.

Andmete kvaliteedi kontrollimine

Kõik vajalikud väljad, nagu loan_status ja person_income, on täidetud, mis tagab andmete täielikkuse. Samuti on kinnitatud, et kõik veergude väärtused vastavad eeldatavale andmetüübile, näiteks on loan_int_rate õigesti määratletud kui numbriline väärtus, tagades andmete kehtivuse. Lisaks on andmed kontrollitud, et need vastaksid reaalsusele ja et kõik vead või anomaaliad oleksid kõrvaldatud, kinnitades andmete täpsuse.

Ülesanne 4

1) Andmete analüüsimine: Saada andmetest ülevaade, uurida iga veergu ning määrata kindlaks kõige mõjukamad omadused. - [Madis ja Rando (~2h)]

2) Decision tree: - Andmete analüüsimiseks ja peamiste otsustusreeglite tuvastamiseks loome ja treenime Decision tree mudeli. - [Madis (~4h)]

3) Random Forest: - Tugevama ja täpsema klassifitseerimise jaoks rakendame Random Forest mudelit. - [Rando (~4h)]

4) Neural Network: - Ehitame ja treenime Neural Network'i, et testida andmestiku süvaõppe toimivust. - [Madis ja Rando (5-6h)]

5) Võrdleme mudeleid – Võrdleme Decision tree, Random Forest ja Neural Network mudelite jõudlust sobivate mõõdikute abil. [Madis ja Rando (~2h)]

6) ROC graafik (FN ja TN): Luua ROC graafik, kus on kujutatud vale negatiivsed (FN) ja tõelised negatiivsed (TN), et hinnata mudelid. - [Madis and Rando (~2h)]

7) Plakati tegemine: Kokkuvõtte tegemine projekti protsessist ja tulemustest. - [Madis ja Rando (~3h)]

Meetodid:

Decision tree: Kasutame selle lihtsuse ja tõlgendatavuse tõttu.

Random Forest: Kasutame stabiilsema ja täpsema prognoosi jaoks.

Neural Network: Süvaõppe lähenemisviis mittelineaarsete suhete uurimiseks.

ROC graafik (FN ja TN): Vale negatiivsete ja tõeliste negatiivsete analüüsiks.

Tööriistad:

Jupyter Notebook: Koodi ja andmete uurimiseks.

Python: Programmeerimis keel, mida kasutades oma tööd teeme.

Github: Töökoht, kus hoiame andmeid projekti jaoks, et saaks nendega koostööd teha.

Discord ja Messenger: Suhtlus keskkonnad, mida kasutame, et infot jagada või arutada.

Kaggle: Lisandmete ja ressursside leidmiseks.