UNIVERSITY OF TARTU
Institute of Computer Science
Software Engineering Curriculum

Eldar Hasanov

Automated Balance Depletion Prediction in Retail Banking

Master's Thesis (30 ECTS)

Supervisor: Marlon Dumas

Tartu 2019

Automated Balance Depletion Prediction in Retail Banking

Abstract: Retail banks employ various solutions and techniques to analyze data of customers with the business goal of delivering better service. In general, customer transactions and cash flow may provide useful information or pattern about customer's behavior. One of the machine learning techniques that is employed on the cash flow and transactions of a customer is balance depletion prediction which estimates whether or not a customer will reach a balance of zero, or close to zero, within a given time interval. The balance depletion prediction may provide a better economic strategy for customers and help retail banks to offer more competent risk management services to the bank's customers. These models have also been exploited by several other companies to identify potential problems in their business and to mitigate the adverse outcomes during project development. Although there have been few studies to analyze the cash flow of companies, a limited number of research studies has addressed the problem of cash flow and balance depletion prediction in retail banking.

Here, we present a case study where we employ machine learning solution to build balance depletion model. Our task is estimating the depletion of balance after the given prediction window. Our partner financial institution provided datasets that contain a time series of balance records for six months and data related to the customer and bank account. Initially, we propose a baseline approach where we train LightGBM classifier on the input data. To reduce computational complexity, we integrate two feature selection techniques into the pipeline (Boruta and BoostaRoota). Next, to improve model performance, we incorporate three feature engineering techniques: manual, Featuretools and TSFRESH. Each model is evaluated on a real anonymized dataset extracted by the financial institution.

Boruta and BoostaRoota don't provide expected improvement due to input dataset size and computation time of the algorithm. Besides, the feature engineering techniques don't also provide significant improvement over the baseline approach. Feature extraction with TSFRESH is computationally expensive while other two feature engineering techniques perform in short time.

Keywords: machine learning, balance depletion, feature engineering, feature selection

CERCS: P170, Computer science, numerical analysis, systems, control

Automatiseeritud konto tühjenemise ennustamine jaepanganduses

Lühikokkuvõte: Jaepangad kasutavad mitmeid eri lahendusi ja metoodikaid selleks, et töödelda klientide andmeid eesmärgiga pakkuda paremat teenindust. Üldiselt võivad kliendi tehingud ja rahavood anda kasulikku infot kliendi käitumise või selle mustrite kohta. Üks kliendi rahavoogude ja tehingute puhul kasutatavatest tehisintellektil põhinevatest tehnoloogiatest on konto tühjenemise ennustamine. Teame, et laekumiste ja väljamaksete vaheline tasakaal määrab kliendi majandusliku seisu ning selle tasakaalu

ebaefektiivne haldamine võib viia kliendi pankrotti. Konto tühjenemise ennustamise abil on võimalik klientidele pakkuda paremat majandusstrateegiat ja toetada jaepanku, et need saaksid oma klientidele kompetentsemaid riskihaldusteenuseid pakkuda. Neid mudeleid on kasutanud ka paljud teised ettevõtted, et tuvastada potentsiaalseid probleeme ja hallata projekti arenduse käigus tekkivaid ebasoodsaid tagajärgi. Kuigi mõnedes uurimustes on ettevõtete rahavooge analüüsitud, pühenduvad vähesed uurimused rahavoogude ja tühjenemise ennustamise probleemidele jaepanganduses.

Selles töös näitame juhtumianalüüsi, kus kasutame tühjenemise ennustamise mudeli loomiseks masinõppelahendust. Meie töö on hinnata konto tühjenemist pärast määratud ennustusvahemikku. Meie finantsasutusest partneri pakutud andmekogum sisaldab aegrida kontojäägi andmetest kuue kuu jooksul ning kliendi ja pangakontoga seotud tunnuseid. Esmalt pakume välja algse lähenemisviisi, kus treenime sisendandmete abil LightGBM-i klassifitseerija. Arvutuskeerukuse vähendamiseks integreerime konveieriga Boruta ja BoostARoota tunnuste valiku metoodikad. Seejärel lisame mudeli jõudluse parandamiseks kolm tunnuste loomise metoodikat: manuaalne, FeatureTools ja TSFRESH. Iga mudelit hinnatakse finantsasutuse anonümiseeritud andmekogumi väljavõtte põhjal.

Boruta ja BoostARoota ei näita oodatud paranemist sisendandmekogumi suuruse ja algoritmi arvutusaja tõttu. Lisaks ei näita tunnuste loomise metoodikad algse lähenemisviisiga võrreldes olulist paranemist. TSFRESHi jaoks on arvutuskeerukus probleem, teised metoodikad aga töötavad kiiremini.

Võtmesõnad: masinõpe, konto tühjenemine, tunnuste konstrueerimine, tunnuste valimine

CERCS: P170, Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine

Appendix

I. Licence

Non-exclusive licence to reproduce thesis

I, **Eldar Hasanov**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for the purpose of preservation in the DSpace digital archives until the expiry of the term of copyright,

   Automated Balance Depletion Prediction in Retail Banking,

   supervised by Marlon Dumas.

   Publication of the thesis is not allowed.

2. I am aware of the fact that the author retains the right specified in p. 1

3. This is to certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Eldar Hasanov
Tartu, 14.05.2019