TALLINNA TEHNIKAÜLIKOOL
SCHOOL OF ENGINEERING
Department of Electrical Power Engineering and Mechatronics

AN XGBOOST-BASED FRAUD DETECTION MODEL FOR MICRO-FINANCE USING ADVANCED FEATURE SELECTION METHODS AND BAYESIAN HYPERPARAMETER OPTIMIZATION

XGBOOSTI-PÕHINE PETTUSTE TUVASTAMISE MUDEL KASUTADES KAUGELE ARENENUD MUUTUJATE VALIKU MEETODEID JA BAYESI HÜPERPARAMEETRITE OPTIMISATSIOONI

MASTER THESIS

Student:	Sten Sinimets
Student Code:	163381MAHM
Supervisor:	Kristina Vassiljeva , Associative Professor
Supervisor:	Robert Hudjakov , Engineer, Research Scientist
Supervisor:	Mykola Herasymovych , Head of Data Science Department at Creditstar Estonia AS

Tallinn 2020

SUMMARY

The aim of the thesis was to deleveop a fraud detection model for international micro-finance company using machine learning algorithm XGBoost with advanced feature selection methods and Bayesian Hyperparameter Optimization.

As the company did not have clear flag of fraud in the database, fraud was defined as a case where a person did not pay anything within 90 days after the first installment due date of the loan. The final dataset was split into train, test and validation set with the ratio of 60-20-20. The initial feature set included more than 800 features, while after the initial data pre-processing, the final feature set included 605 features.

The Chi Square and Information Value tests were performed in order to reduce the number of features. In total 60 tests were performed on the top-performing features according to the feature selection tests, while also on the full feature set using Bayesian Hyperparameter optimization in order to find the best set of parameters. The highest result of the train set AUC of 0.766 and validation set AUC of 0.765 came when using all of the features. The number of features was trimmed down to 68, while using XGBoost native feature importance method with the final model having a performance of 0.765 on the training set, 0.764 on the validation set and 0.742 on the test set. The results exceeded the expected objectives of the thesis of achieving at least 0.7 AUC base performance of the model and a maximum of 0.03 overfitting.

Analysis was done while comparing the existing scoring model to the final model and it was found that the final model performs significantly better compared to the existing scoring model. Based on the analysis, it was decided by the company management, that the model was fit to go into production, which happened on 17.09.2019.

At the moment of writing this summary, none of the loans issued have matured enough to bring conclusions based on the target variable. Nevertheless, the initial conclusion can be brought , based on a client failing to make payment within 30 days after the due date of the first installment. For online customers with bank statement the improvement in the FPD_30 rate compared to the previous model has been 22%, while for offline clients the improvement compared to the previous model has been over 48% given similar acceptance levels.

To conclude, the objectives of the thesis were achieved and the production-ready model created. The model has been in production for over 3 months and hundreds of loans have been issued based on the suggestion made by the model.

KOKKUVÕTE

Lõputöö eesmärgiks oli luua pettuste tuvastamise mudel rahvusvahelisele mikrolaenudega tegelevale ettevõttele, kasutades masinõppealgoritmi XGBoost koos kaugele-arenenud muutujate valikute meetodite ja Bayesi hüperparameetrite optimisatsiooniga.

Kuna ettevõttel ei olnud andmebaasis selget märki, kas kasutaja on petis või mitte, , defineeriti pettus olukorrana, kus klient ei maksnud midagi 90 päeva jooksul pärast laenu esimese tagasimakse kuupäeva. Lõplik andmestik jaotati treenimis-, testimis ja valideerimisandmestikeks suhtega 60-20-20. Esialgne tunnuste kogum koosnes enam kui 800 tunnusest, kuid pärast esialgset andmete töötlust jäi lõplikkusse tunnuste kogumisse 605 tunnust.

Hii-Ruudu ja Informatsiooni Väärtuse testid viidi läbi, et vähendada tunnuste hulka. Enam kui 60 testi viidi läbi kõige paremini esinevate tunnustega nendes testides, kuid ka kasutades kõiki tunnuseid andmestikus. Testid viidi läbi kasutades Bayesi hüperparameetrite optimatsiooni, et leidi kõige sobivamad parameetrid. Kõige parem AUC tulemus treeningandmestikul oli 0.766 ja validatsiooniandmestikul 0.765 ning tuli kasutades kõiki tunnuseid. Tunnuste arv vähendati 68-le kasutades XGBoosti funktsiooni, mis näitas tunnuste olulisust mudeli. Lõpliku mudeli tulemus treeningandmestikul oli 0.765, validatsiooniandmestikul 0.764 ja testimisandmestikul 0.742. Tulemused ületasid esialgseid eesmärke saavutada vähemalt 0.7 AUC baastulemus mudelil maksimaalselt 0.03 ülesobitusega.

Analüüs viidi läbi, et võrrelda olemasolevat mudelit lõpliku mudeliga ja leiti, et lõplik mudel on silmnähtavalt parem võrreldes mudeliga. Analüüsi põhjal otsustas ettevõtte juhtkond antud mudeli reaalelus käiku panema ning mudel läks tootmisesse 17.09.2019.

Kokkuvõtte kirjutamise hetkel pole ükski laenudest, mis mudeli alusel väljastati jõudnud järku, et teha järeldusi mudeli eesmärgi muutuja põhjal. Samas, esialgsed järeldused saab teha, võttes arvesse olukorda, kus klient ei suuda midagi maksta 30 päeva jooksul pärast esimest maksetähtaega. Läbi veebi avalduse teinud panga väljavõttega klientidel langes FPD_30 22% võrreldes eelmise mudeliga ning väljaspool veebi avalduse teinud klientidel 48% võttes arvesse, et akptsepteerimismäär oli sarnane.

Kokkuvõttes võib öelda, et lõputöö eesmärgi täideti ja tootmiseks valmis mudel loodi. Mudel on olnud tootmises üle 3 kuu ja sadu laene on väljastatud mudeli soovituse alusel.