SUMMARY

The purpose of this work was to discover and apply different feature engineering techniques to create well-performing variables, use them in machine learning models and improve their performance.

Most part of development was spent on getting new features from bank transactional data. Reason comes from the new regulations regarding data in EU, GDPR and open banking. As bank data becomes more accessible for clients, the demand and usage of this data will increase significantly. Accordingly, this work is relevant in the current regulatory and business environment and can be considered as a contribution to future applications.

Experiments were undertaken using Python 3 and its open source data science libraries. One of the most significant part of techniques was about natural language processing, for which NLTK library was researched and implemented.

As a result, best performing 7 different techniques have been applied for the real-world application usage; 1) Ratcliff/Obershelp sequence matcher. 2) Levenshtein distance. 3) lemmatization. 4) stemming. 5) tokenization. 6) tf-idf feature selection. 7) count-based featurization.

At the beginning of thesis development, there have been approximately 300 variables used for ML modeling. Applying abovementioned techniques resulted in the creation of more than 400 new features. Accordingly, at the end of the experiment and advanced feature engineering process, 700+ features were deployed into production. This improved performance of the model by 0.016 unit.

Further research

Feature engineering and data science overall is based on experiments. The more approaches are tried out, the better performance is achieved. Accordingly, this work can be improved in the future by discovering new techniques and improving the existing ones. Besides that, the performance of features depends on ML models on which they are tested. Applying exactly the same approaches and testing the same features, but in a different production environment in order to compare final performances, could be beneficial for future development.