Data Scientist Assignment

Home Credit Vietnam

Thank you for your interest in Data Scientist position at Home Credit Vietnam. To evaluate your data science skills, we have prepared an assignment which is inspired by one of our typical task. Each client applying for a loan needs to be assesed on his/her ability to repay the loan back. To make automated assesments, we develop a model estimating probability of repayment of each loan application and then use it for decision whether the application is to be approved or rejected. The same task is the goal of the assignment - to develop and evaluate a binary classification model for probability of loan repayment based on the provided dataset.

Dataset can be found here and description of the columns is following:

- `ID_APPLICATION`: ID of the application

- `TIME`: Timestamp of the application

- `NUMERICAL_0-41`: Numerical features that can be used in the model

- `CATEGORICAL_0-10`: Categorical features that can be used in the model

- `TARGET` - Binary flag whether loan was repaid or not (dependent variable)

Features in the provided dataset can suffer from various issues concerning stability in time. These problems are quite usual in real data problems. Please choose 2 binary classification algorithms according to your preference and compare their results. Preferred programming language is Python. The expected result is all source codes and a short report including:

- Explanatory analysis of features and their relation to target

- Brief description of the used algorithms and their pros and cons

- Performance evaluation of the model on the last month (May 2015), which would be used as holdout testing data. Please include AUC-ROC among your evaluation metrics

- Short conclusion and recommendation

Do not spend too much time maximizing model performance, we are rather interested in how you approach the problem and level of data analysis, machine learning and programming skills.