Capstone Proposal for Machine Learning Engineer Nanodegree

Proposal

As a capstone project I would like to provide solution for Home Credit Group's "Home Credit Default Risk" challenge posted on Kaggle.com[1]. The main challenge is to predict how each candidate is capable of repaying a loan.

Domain Background

"Default risk is the chance that companies or individuals will be unable to make the required payments on their debt obligations. Lenders are exposed to default risk in virtually all forms of credit extensions. To mitigate the impact of default risk, lenders often charge rates of return that correspond the debtor's level of default risk. A higher level of risk leads to a higher required return."[2] To evaluate candidate's willingness and ability to repay loan creditors and lenders utilize a number of financial tools. Credit bureaus maintain records of consumers' experience with banks, retailers, doctors, hospitals, finance companies etc.. This solution only works when relevant, financial, data is recorded and available.

Problem Statement

It is more challenging for banking institutions to evaluate people that have insufficient or non-existent credit histories. Due to that some group of people is excluded from positive borrowing experience. The goal of this project is to use alternative data to predict candidate's repayment abilities. This will enable more people to get loan from banks and other financial institutions, so they won't be taken advantage of by untrustworthy lenders.

Dataset and Inputs[3]

Dataset and inputs are provided by Home Credit Group and it contains 8 files:

- application_{train|test}.csv – main table, broken into two files for Train and Test.
- bureau.csv – previous credits provided by other financial institutions that were reported to Credit Bureau.
- bureau_balance.csv - monthly balances of previous credits in Credit Bureau.
- POS_CASH_balance.csv – monthly balance snapshots of previous point of sales and cash loans that the applicant had with Home Credit.
- credit_card_balance.csv - monthly balance snapshots of previous credit cards that the applicant has with Home Credit.
- previous_application.csv - all previous applications for Home Credit loans of clients who have loans.
- installments_payments.csv – repayment history for the previously disbursed credits in Home Credit.
- HomeCredit_columns_description.csv - descriptions for the columns in the various data files.

---

[1] Challenge description: https://www.kaggle.com/c/home-credit-default-risk

[2] Taken from https://www.investopedia.com/terms/d/defaultrisk.asp

[3] All descriptions are taken from: https://www.kaggle.com/c/home-credit-default-risk/data

Solution Statement

Work will be started with exploratory data analysis to detect missing values, distribution and correlation of features. Then feature importance will be checked. After that data cleaning and normalization will be done. Engineering of the solution will be started with simplest model – Logistic Regression. After that it's planned to use Random Forest method. Grid Search will be used to tune hyperparameters of Random Forest algorithm. After that LightGBM will be used due to its sophistication, fast training time, high efficiency and good accuracy.

Benchmark Model

The simplest benchmark model can do random guessing with equal probability of 0.5 for both repaying loan and not.

Evaluation Metrics

Because this is Kaggle's competition it's evaluated on area under the ROC curve between the predicted probability and the observed target. "The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings."[4]. Values for this metric are between 0 and 1. Model with better scoring will have higher value. Model that does random guessing will have score of 0.5.

Project Design

Due to several input files work will be started with exploratory data analysis. It needs to be checked if training data is balanced, which features have missing records, which features are correlated with each other etc. First steps will focus on finding solution that is based only on one input – *application_train.csv*. Then more sophisticated solutions that use Random Forest or LightGBM will be engineered.

[4] Taken from https://en.wikipedia.org/wiki/Receiver_operating_characteristic