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. Machine learning algorithms have been already applied to such type of problems. Different classification algorithms were checked regarding their performance in application to credit risk assessment. Details can be found in paper: *"Multiple classifier architectures and their application to credit risk assessment"* by Steven Finlay[2]

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."[3] 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.

This problem is a standard supervised classification task. It's expected that for each candidate probability of loan repayment will be calculated. As an inputs relevant data such as income, education, occupation, age, previous credits, repayment history etc. is provided.

Dataset and Inputs[4]

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.

---

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

[2] Finlay, S M (2008) *Multiple classifier architectures and their application to credit risk assessment*. http://eprints.lancs.ac.uk/48931/

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

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

- 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.

In this project *application_train.csv* data will be mainly used. If there is enough time other inputs will be used. *Application_train.csv* contains 120 features and 307510 examples. There is 16 categorical features, 41 of type int and 65 of type float. Target for training data can be either 0 or 1 where '0' means that candidate won't be able to repay its loan. Data in test set is not well balanced, examples with target equal to '0' are dominating. There is 282686 examples of candidates who application was rejected and 24825 of applications which were accepted.

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 XGBoost will be used due to its sophistication, fast training time, high efficiency and good accuracy.

Benchmark Model

Because test dataset mostly contains, data of rejected applications simplest benchmark model will be always guessing that application is rejected – target value equal to 0.

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."[5]. 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 will be started with reading first few rows of the train data. It needs to be known what features there are and in what form data is presented. Then number of missing values will be calculated. After that some features will be normalized. For example following features: DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION have negative values. Those features need to be transformed into positive values and scaled to be represented in years. Next step is one-hot encoding of categorical features in training and testing data. When all of the data is prepared search for anomalies and correlations will be started.

First step in model engineering will be creating the benchmark model described in previous paragraph. Next model that will be implemented is Logistic Regression from Scikit-Learn. For hyperparameter optimization Grid Search method is going to be used. After that improved model - Random Forest will

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

be used. It’s expected to have better results since there is a lot of features in test data. At this step feature importance will be checked. As a last model XGBoost is going to be used. It’s planned to use it on data with missing values and then on same data but with missing values replaced through imputation.