Capstone Project

Machine Learning Engineer Nanodegree August 4th, 2018

Domain Background

In the field of credit lending, people hope to solve some short-term funding gaps by using loans. The lender wants to make a profit with a small amount of interest, but there is currently no 100% way to determine/predict the credit repayment ability. This may result in people who are unable to repay being used by bad lenders, therefore leads to bankruptcy of credit, which indirectly causes certain public security/economic problems [1]. Various federal agencies also required lenders to assess a consumer's ability to repay a home loan, and the creditor must consider and evaluate at least eight factors, including “Current employment status”, “Monthly payments on a simultaneous loan”, “Monthly debt-to-income ratio or residual income”, and so on [1]. Several improvements have been developed in the credit scoring domain. In [2], an empirical comparison of various resampling techniques has been discussed in order to deal with imbalanced data. [3] shows that ensembles of classifiers achieve better results for credit risk assessment, and [4] demonstrates that Bayesian hyper-parameter optimization performs better than random search, grid search, and manual search. “Home-Credit” has launched a machine learning competition “Home Credit Default Risk” on Kaggle [5].

By collecting enough data to accurately predict/analyze each borrowers' repayment ability(but not too much to prevent privacy disclosure), both sides and even the social security level can benefited from it. And it could be considered as some kind of “changing the world”. Problem Statement

Problem Statement

The goal is to build a prediction model to classify the clients’ repayment abilities. By using a variety of alternative data--including telco and transactional information, if we can accurately predict the clients who have ability to repay, but with insufficient or non-existent credit histories, then we can expand the business for the company and avoid untrustworthy lending to this population. This problem is a machine learning **classification** task, TARGET=0 means the client will repay on time and TARGET=1 means the client have some difficulty repaying loan. And because the targets for training are already known and labeled, **supervised machine learning algorithms including logistic regression and random forest are potential solutions.**

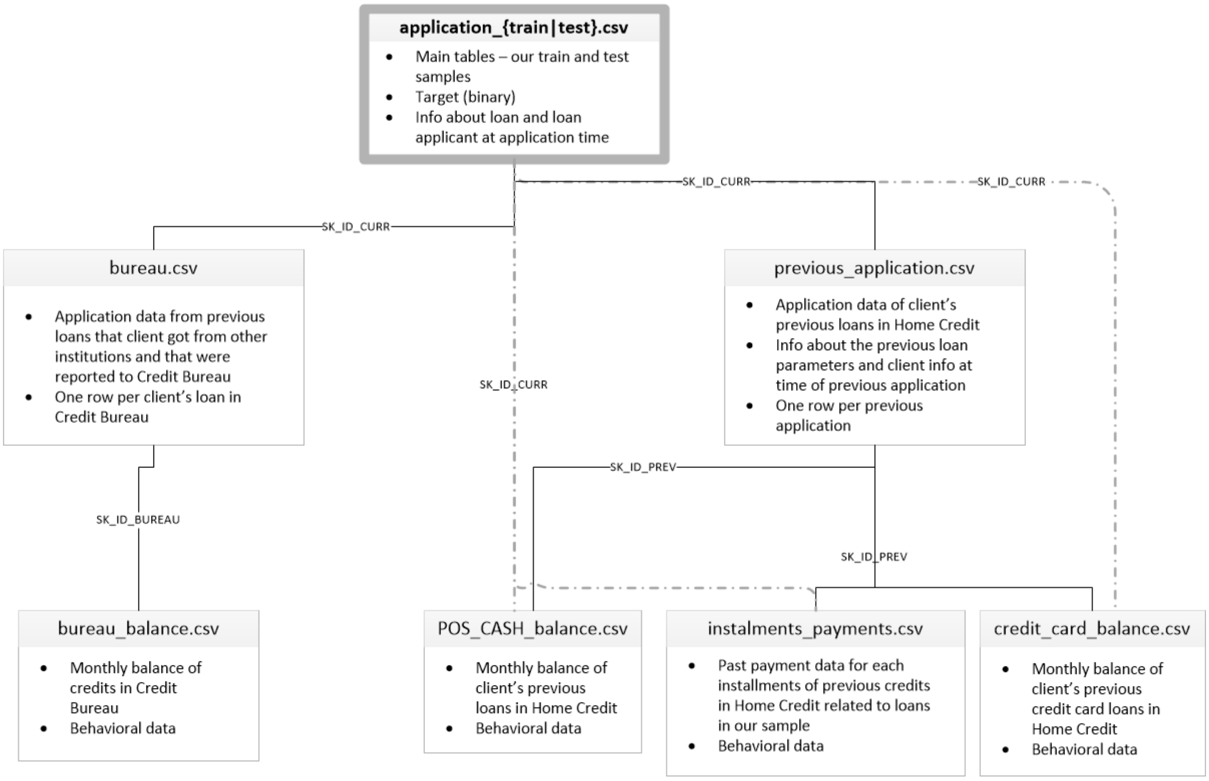
Datasets and Inputs

The datasets are provided by Home-Credit-Group for use in the competition, and are available for download on the Kaggle competition data page [6]. There are mainly 9 CSV files where “HomeCredit_columns_description.csv” contains all descriptions and “application_{train|test}.csv” contains all applications’ static data for the use of training and testing. There are **totally 307511 examples and 122 features** in the training dataset. The datatype of 65 features are float number(continuous), 41 features are integer(continuous), and 16 features are String(categorical). And 282686 examples’ TARGET equal 0, 24825 examples’ TARGET equal 1, Which means the percentage of “TARGET==1” is only **8.1%**. In order to deal with this **imbalanced** data, we would resampling from the dataset or assign misclassification costs to force the classifier to concentrate on the minority classes. We will split the training data into training and validation sets to evaluate/optimize our model, then apply the optimized model on the testing data provided to get the final prediction.

- The other 6 CSV files contain additional informations briefly described below:

bureau.csv	All client's 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 POS (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 in our sample.
installments_payments.csv	Repayment history for the previously disbursed credits in Home Credit related to the loans

- The diagram below shows the relationships between all of the dataset:


application_{train|test}.csv
• Main tables – our train and test samples
• Target (binary)
• Info about loan and loan applicant at application time
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
bureau.csv
• Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
• One row per client's loan in Credit Bureau
previous_application.csv
• Application data of client's previous loans in Home Credit
• Info about the previous loan parameters and client info at time of previous application
• One row per previous application
SK_ID_PREV
SK_ID_BUREAU
SK_ID_PREV
bureau_balance.csv
• Monthly balance of credits in Credit Bureau
• Behavioral data
POS_CASH_balance.csv
• Monthly balance of client's previous loans in Home Credit
• Behavioral data
instalments_payments.csv
• Past payment data for each installments of previous credits in Home Credit related to loans in our sample
• Behavioral data
credit_card_balance.csv
• Monthly balance of client's previous credit card loans in Home Credit
• Behavioral data

Solution Statement

The solution to this problem would utilize supervised machine learning algorithms. We will mainly use Gradient-boosting algorithms [7], such as **XGBoost**(eXtreme Gradient Boosting) [8], **LightGBM**(Light Gradient Boosting Machine) [9] and **CatBoost** [10]. By using model-tuning methods, such as Random Search and **Bayesian optimization**, we could further improve the prediction model. Some preprocessing techniques will also be used firstly, including data-cleaning, EDA(exploratory data analysis), and some **feature-engineering**/feature-selection.

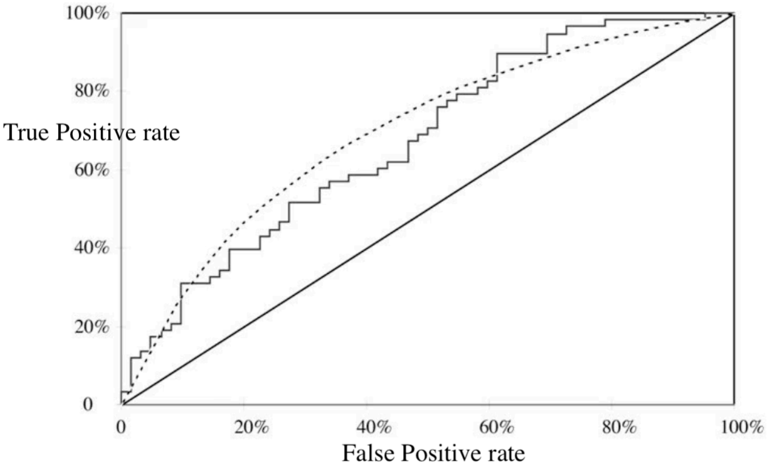
Benchmark Model

We use a **logistic regression classifier with default hyperparameters** from scikit-learn library as the benchmark model [11]. The benchmark model has an **AUC score of 0.677** on the testing set. In the benchmark model, we only use one file for training, which is "application_train.csv". Necessary but simplest "Data cleaning and formatting" methods, such as one-hot encoding, missing values handling, and scaling, are applied in order to run the logistic regression model.

Evaluation Metrics

The evaluation metric used is “**Area under the ROC curve**”(also called AUC) [12], where the ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. AUC is between 0 and 1(the higher AUC, the better model). The AUC can be calculated by the equation below, where T is the threshold parameter and X1(X0) is the score for a positive(negative) instance.

$$\mathrm{AUC} = \int^{-\infty} \mathrm{TPR}(T)\mathrm{FPR}'(T)\,dT = P(X_1 > X_0)$$



ProjectDesign

In this project, we will follow the typical machine learning workflow below:

EDA(Exploratory Data Analysis)	• Data Exploration and Visualization
Data Preprocessing	• Convert categorical features(One-Hot Encoding for binary variables, and Label Encoder for multi-class variables) • Transforming Skewed Continuous Features • Normalizing/Scaling Numerical Features • Missing values handling(Drop or Fill-in with mean/ median/most frequent value) • Outliers detection and handling
Model Performance Evaluation	• Implement a pipeline to choose the best model(s) • Evaluate testing data on Kaggle by AUC score • Feature Engineering and/or Feature Selection • Using model-tuning methods, such as Random Search and Bayesian optimization, to further improve the prediction model(s) • (Optional) OOF(out-of-fold prediction, i.e. stacking features)

Reference Links

1. An Overview of the Consumer Financial Protection Bureau's Ability-to-Repay and Qualified Mortgage Rule
   https://www.americanbar.org/publications/blt/2013/04/02_shatz.html
2. Anahita Namvar, Mohammad Siami, Fethi Rabhi, Mohsen Naderpour: "Credit risk prediction in an imbalanced social lending environment"
   https://arxiv.org/abs/1805.00801
3. Joaquín Abellán,Javier G.Castellano: "A comparative study on base classifiers in ensemble methods for credit scoring", Expert Systems with Applications Volume 73, 1 May 2017, Pages 1-10
   https://www.sciencedirect.com/science/article/pii/S0957417416306947?via%3Dihub
4. Xia Yufei, Liu Chuanzhe, Li YuYing, Liu Nana: "A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring", Expert Systems with Applications Volume 78, 15 July 2017, Pages 225-241
   https://www.sciencedirect.com/science/article/pii/S0957417417301008
5. “Home Credit Default Risk”, kaggle.com, 2018. [Online]
   https://www.kaggle.com/c/home-credit-default-risk/
6. “Home Credit Default Risk Data”, kaggle.com, 2018. [Online]
   https://www.kaggle.com/c/home-credit-default-risk/data
7. A Gentle Introduction to Gradient Boosting
   http://www.ccs.neu.edu/home/vip/teach/MLcourse/4_boosting/slides/gradient_boosting.pdf
8. XGBoost, eXtreme Gradient Boosting (Github)
   https://github.com/dmlc/xgboost
9. LightGBM, Light Gradient Boosting Machine (Github)
   https://github.com/Microsoft/LightGBM
10. CatBoost: based on gradient boosting over decision trees (Github)
    https://github.com/catboost/catboost
11. Capstone project benchmark model code (Gist)
    https://gist.github.com/jo4x962k7JL/06af77c0d82da5dfbc2d82788d42659b
12. wikipedia: Receiver_operating_characteristic
    https://en.wikipedia.org/wiki/Receiver_operating_characteristic