MLND Capstone Proposal

Lancelot

16 October, 2018

Domain Background

As a MLND capstone project I am proposing to build a model to predict credit default risk for home loans. This is based on the Kaggle competition Home Credit Default Risk.

Home Credit has the vision of providing a safe environment to provide loans to the unbanked population. This group of people has been unfairly treated by unscrupulous money lenders and Home Credit wants to change that. However due to the lack of credit scores of these unbanked population, Home Credit needs to rely on alternative data to help them predict on their clients' repayment abilities. Traditional approach of using credit scores are out of scope and the project aims to develop a machine learning based model to solve the challenge.

Problem Statement

The problem we are solving is a binary classification problem. Given a set of data that is linked to a potential client, (examples include his income, age, education, his previous repayment history, his monthly snapshots of credit card loans, cash etc) we want to predict if he is able to repay the loans or not. Data provided is tabular and consists of a mixture of continuous and categorical data types.

The model we build will output the probability that the loan will not be repaid, given a set of features associated with the individual client. To measure the goodness of the model, we shall be using the metric area under ROC (Receiver Operating Curve)

Datasets and Inputs

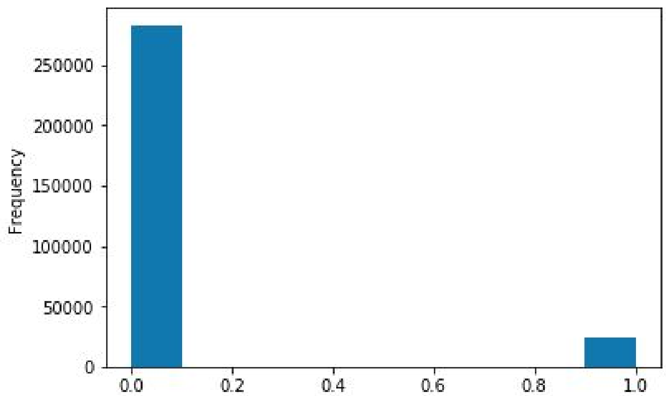
There are several datasets provided by Home Credit. The main training table has 307,511 rows and 121 columns. The columns are the key information that are required in an application for loan. Some additional tables related to monthly snapshots of previous credit balances, POS (Point of Sales) and cash loans are also provided. However, for this project we shall only focus on the main table called the application.

The table is of small to mid-size which means that we can potentially try out a range of models. Below we give a snapshot of some column fields in the main table. These are information required for a loan application and hence should provide a fair indicator of repayment default

Columns

- # TARGET
- A NAME_CONTRACT_TYPE
- A CODE_GENDER
- A FLAG_OWN_CAR
- A FLAG_OWN_REALTY
- # CNT_CHILDREN
- # AMT_INCOME_TOTAL
- # AMT_CREDIT
- # AMT_ANNUITY
- # AMT_GOODS_PRICE
- A NAME_TYPE_SUITE
- A NAME_INCOME_TYPE
- A NAME_EDUCATION_TYPE
- A NAME_FAMILY_STATUS
- A NAME_HOUSING_TYPE
- # REGION_POPULATION_RELATIVE
- # DAYS_BIRTH
- # DAYS_EMPLOYED
- # DAYS_REGISTRATION
- # DAYS_ID_PUBLISH
- A OWN_CAR_AGE

We did a quick check on the distribution of the target variable, noticed that it has an imbalanced distribution. This suggests that we should not use “accuracy” as our metric and will have to look into ways to create a more balanced training set and select the appropriate models that could address the issue of imbalance



Solution Statement

Since this is a classification problem, our approach will be to fit it with random forest or more advanced ensemble methods like XGBoost or LightGBM. With regards to the issue that there might be too many features, we will explore feature selection or dimensionality reduction methods.

Benchmark Model

To establish a benchmark model, I will simply use one which predicts “non-default” for all clients

Evaluation Metrics

We use area under ROC (AUC) as the metric to evaluate the model. AUC is a common metric used to evaluate binary classifier by integrating both the true positive rate and true negative rate into a single number

Project Design

The overall approach we will take is as follow. It is a rather common approach similar to [3]

1. Data cleansing. We will start by reviewing the data, looking at missing values, and one-hot encode categorical variables, and do sanity checks
2. Exploratory data analysis. Next we investigate correlation between the features and target, also just explore the different features searching for trends

3. Feature selection / engineering. Here potentially we could combine features or perform dimension reduction, or simply just select the more relevant features
4. Model training and selection.   We then split the data for training and cross validation, possibly scale the data for training, and tune hyper parameters for each of the models.
5. Test set evaluation. Finally we evaluate the final model on the test set

References

1. https://www.kaggle.com/c/home-credit-default-risk
2. https://www.kaggle.com/willkoehrsen/start-here-a-gentle-introduction
3. https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/
4. https://towardsdatascience.com/predicting-loan-repayment-5df4e0023e92
5. https://medium.com/henry-jia/bank-loan-default-prediction-with-machine-learning-e9336d19dffa
6. https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/
7.