Machine Learning Engineer Nanodegree

Capstone Proposal

Huiyan Xu

February 16th, 2018

Proposal

Domain Background

One of the principal tasks of bank is to provide loans and generate profits. Individual loans are often of large size, but relatively smaller amount comparing to company loans. Due to this characteristic, banks tend to use automated way to censor the individuals, including passing credit checks. Referring to *Machine Learning Application in Online Leading Credit Risk Prediction (Xiaojiao Yu)*, bank's traditional way of assessing credit risk is by credit score models fall short in applying big data technology in building risk model. Yu also examined predicting loan default rate using ensemble machine learning models, random forest and XGBoost model. Machine Learning is getting more and more important in finance roles and it could make use of larger amount of customer data and make better decisions.

Problem Statement

Our problem is cited from *Kaggle Competition - Loan Default Prediction - Imperial College London*. Our goal is to train a model that takes 778 inputs and determine whether a loan will default, as well as the severity of the losses by measuring the amount of loss given if it defaults. The target is to achieve a low Mean Absolute Error:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}|$$

Datasets and Inputs

The data is a set of financial transactions associated with individuals sourced from Imperial College London. We have 105471 individuals and 778 features (x) in total and a loss value (0 for non-default, 0<y<=100 for default and relative loss). To protect user's information, the features we have received are anonymized. We have categorical features as well as features with NA values. We would consider all the features available and use some feature selection technique to filter the most valuable ones.

Solution Statement

I would like to divide the project into two parts:

(1) Predicting default/non-default. This would be a classification problem. We could do feature selections and train models such as Decision Trees, GBM classifier to predict if individual would default or not (y>0, default or y=0, non-default).

(2) For those predicted to be default (y>0), predict the amount of loss using Machine Learning technique such as DecisionTreeRegressor, GBM Regressor.

Benchmark Model

I have built a random number distribution on loss generated by training dataset, and generated a set of random numbers of same size as test dataset. This method only takes loss data characteristics into consideration and does not make use of any features. Thus this would be a good benchmark model for later comparison.

The benchmark model's default/non-default accuracy is 83.05%. And Mean Average Error is 1.51. For detailed computation, please check capstone_proposal_report.html.

Evaluation Metrics

(1) Regarding default/non-default prediction, we could use accuracy rate as evaluation metric.

(2) The evaluation metric used in the Kaggle competition is MAE, which calculate the error between predicted loss and actual loss. When non-default is predicted, y_i is zero. MAE is a suitable evaluation metric as it shows more on the actual loss. For example, when an individual loss is small, but prediction result is non-default; default/non-default accuracy would take this as a 1 error out of test size, but MAE would consider this as a smaller loss on model training.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}|$$

Project Design

Step1. Data Exploration

Given this project's feature size of 778 inputs, I plan to go through the features and check NA values and look for categorical variables. For NA values, I consider fill in mean values. For categorical variables, one-hot encode method will be used. With this large feature size, I believe some should be removed, if high correlation exists; or making diff between two features as a new feature if they are similar.

Step2. Classification Problem: Default/Non-Default

Techniques to be implemented: Boosted Decision Trees, GBM Classifier and XGBoost.

Cross-validation on Default/Non-Default accuracy rate.

Step3. Regression Problem: Loss on Default

Techniques to be implemented: DecisionTreeRegressor, GBM Regressor.

Cross-validation on MAE.

Step4. Summary

Compare the Accuracy Rate and MAE on trained models and benchmark model. Analyze the best model trained.