Machine Learning Engineer Nanodegree

Capstone Proposal

Huiyan Xu

February 11th, 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. Machine Learning is getting more and more important in finance roles. Comparing to traditional ways of loss controlling which takes a smaller number of credit related variables for loan default prediction, machine learning techniques 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 k-NN, Decision Tree 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 Regression.

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.

Evaluation Metrics

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

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.

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

Project Design

(approx. 1 page)

In this final section, summarize a theoretical workflow for approaching a solution given the problem. Provide thorough discussion for what strategies you may consider employing, what analysis of the data might be required before being used, or which algorithms will be considered for your implementation. The workflow and discussion that you provide should align with the qualities of the previous sections. Additionally, you are encouraged to include small visualizations, pseudocode, or diagrams to aid in describing the project design, but it is not required. The discussion should clearly outline your intended workflow of the capstone project.