Machine Learning Engineer Nanodegree

Capstone Proposal

Rahul Sharma January 3rd, 2018

Proposal

Domain Background

In an attempt to explore a problem most closely resembling the work we do at Capital One, I propose to work on a problem related to credit approval using the data from Lending Club. This is a well known challenge pertaining to the marketplace which matches borrowers seeking loans to investors offering funds. As a typical credit risk problem, I am primarily interested in identifying characteristics that can help identify factors leading to loan defaults. We try to predict such behavior using various attributes describing credit worthiness of a borrower. These characteristics include (but are not limited to) prior lending history of the boorower. Based on the eligibility of a borrower, the loans are either approved or declined. If approved, the loans may have variable interest rates based on the risk profiling. I find the challenge of predicting human behavior quite interesting and because this deals with the type of data we usually handle at Capital One, I thought this may be a fruitful exercise to undertake for the capstone project.

Details of how Lending Club works can be found here. We will build a model and calculate the dollar amount of money saved by rejecting these loan requests with the model, and then I plan to explore if we can combine this with the profits lost in rejecting good loans. We will use Lending Club's Open Data to obtain the probability of a loan request defaulting or being charged off.

- References:

  HBS Summary of Lending Club

Problem Statement

The primary motivation for this exercise is to identify how much money could have been saved if we have a model that can identify the loan defaults appropriately. Similarly, we can also cross check how close our model resemebles that of Lending Club because we are also provided with the data for *Decline Loans*.

For the purpose of this exercise, we will focus on two questions:

1. How well our model predicts loan defaults? To predict the accuracy, we will take into consideration Area under the ROC along with F1 ratio. We use F1 ratio because based on personal experience, I know that such datasets are highly imbalanced where the default rates are usually less than *20%*
2. If our model were to be used, how much money would Lending Club be able to save by rejecting the defaulted loans?

Datasets and Inputs

The datasets required to conduct the analysis are available at Lending Club where the data has been broken down by both Accepted and Rejected as well as origination year/quarter of the loan originations.

- **Time Period** : The loan level data is available starting from 2007 until 2017 broken by years and quarters.
- **Attributes** : As a quick overview, it appears that there are more than 100 attributes available including credit history for the borrower, funding amount, origination dates and interest rate.

We will take these dataset into consideration by breaking them into training set, validation set for hyper-parameter tuning and both a test set as well as out of time test set to see the effectiveness of our model over long term.

Solution Statement

The applicable solution to the problem will be a classification model which most accurately identify the loans which have the highest potential to default. As a typical *Supervised Learning* problem, we will try to explore the data to build a model which uses the credit history, data on loans for users to predict whether the loan will default or not.

In addition to that, I will also make use of these predictions to cross-check against the existing actual data to assess the total dollars Lending Club could have potentially saved.

However, to measure our model's effectiveness, we will conduct out of time testing as well to see if our model holds it's strength across time periods.

Benchmark Model

As we have the actual data for all the defaulted loans, we can take this data into consideration to compare how our model performs. As Lending Club offers the data for the loans it rejects, I am hoping to use this dataset as part of the benchmark exercise. While we will not have the properietary model of the company, we will use the output of the model (i.e. Declined loans) to conduct a benchmarking exercise. If our model is as good as that of Lending Club, majority of the loans in the declined dataset should also be declined by our model.

Evaluation Metrics

The applicable solution to the problem will be a classification model which most accurately identify the loans which have the highest potential to default. To measure the effectiveness of the model, we will take into consider F1 score:

$$F1 = \frac{2 * precision * recall}{precision + recall}$$

where:

$$precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

$$recall = \frac{TruePositive}{TruePositive + TrueNegative}$$

As previously discussed, credit risk problems tend to have imbalanced targets which requires us to look at how well we identify the imbalanced classes. A typical meausre like *Accuracy* can be misleading in such scenario. Hence, I decided to use F1 ratio for the purpose of this exercise. The main goal is to identify as many defaulted loans accuratly as possible while not being overly restrictive to classify good loans as the ones which will default. Latter is important as it can lead to rejecting loans which can contribute towards the profit of the company. Our solution should minimize risk and not avoid risk by being overly strict around the decision criteria. To convey this point, it's important to keep in mind that easiest way to not have any risk in our portfolio is by not offering any loans. However, our intention is to offer loans while assessing the risk associated with the loans apporpriately.

Project Design

As a typical workflow for any data science exercise, I will go through the following steps to conduct the analysis:

1. **Data Cleaning**: Often raw data tends to have data issues which stem from data handling at the source and can lead to issues at a later stage, e.g. unrecognizable characters in text fields, unexpected header rows of data. In addition, our data is split by different time periods, which will be required to combine (row-wise) before we can conduct further analysis.
2. **Data Exploration**: As part of this exercise, we will conduct an analysis of data types, check data distributions, conduct data visualizations to get ourselves familarized with the data fields.
3. **Feature Expansion**: Based on the existing data, we may need to further devise techniques to derive new fields from existing raw data fields e.g. splines, normalization, text mining etc.
4. **Feature Selection**: It is generally true that not all of the data is relevant for the purpose of building

models, so we will try to limit the features by conducting feature selection/elimination using various techniques including *variable imporatance ranking*, *PCA* etc..

5. **Model Development**: In this step, we will try to train models on our data to identify a good fitting model for our needs.
6. **Hyper-parameter Tuning**: Using the validation dataset, we will try to identify the appropriate hyperparameters for our given model.
7. **Model Testing**: To measure the effectiveness, of our models, we will test our models on a test dataset which our model would have not seen previously during the training exercise.
8. **Model Selection**: Instead of just one type of model, we will try a variety of models to assess which one meets our needs the best. Specifically, I plan to levearge the following algorithms:

    - Random Forest
    - Gradient Boosting
    - Generalized Linear Models
    - Stacked Ensembles
    - Deep Learning