Machine Learning Engineer Nanodegree

Capstone Proposal

Hyungsuk Kang
September 27st, 2017

Proposal for Study

(approx. 2-3 pages)

Domain Background

(approx. 1-2 paragraphs)

Banks play a crucial role in market economies. They decide who can get finance and on what terms and can make or break investment decisions. For markets and society to function, individuals and companies need access to credit[^1]:wikipedia. To prove the credit, FICO score was introduced in 1989 by FICO, originally Fair, Isaac and Company, which is a is a data analytics company based in San Jose, California focused on credit scoring services[^2]:wikipedia. Nowadays, Credit Karma is produces free service to look up one's creditworthiness based on their scoring model called 'VantageScore'[3]:investopedia.

However, FICO score does not actually prove that one has no credit. It creates a score by looking at your file from the three major credit reporting bureaus – TransUnion, Equifax and Experian. What's worse, 'VantageScore' model shows rather boasted score and waits for their customers responding to their tailored ads after taking customers' personal information and financial condition. In conclusion, there is no suitable financial model which can determine a person will have financial difficulty in upcoming years and therefore cannot get the loan.

The good service have to provide accurate score for customers to support their financial decision, predicting their future finance. Knowing future is impossible, but it can be approximated with the help of machine learning generated from past cases. In this project, I will compare the performance between machine learning and deep learning to predict whether the borrower will face serious financial problem in 2 years, and the optimal model will help borrowers to make the best financial decision with iOS11 app with CoreML framework. The application uses "Give Me Some Credit" dataset to determine the condition.

Problem Statement

(approx. 1 paragraph)

The goal is to create an iOS11 app that will predict the borrower will face a financial difficulty or not; The tasks are involved are the following:

```
1. Download and preprocess the "Give Me Some Credit" data.
2. Train a classifier that can determine if the borrrower is in financial crisis.
3. Make the model run on iPhone.
```

Datasets and Inputs

(approx. 2-3 paragraphs)

The "Give Me Some Credit" dataset has the data of 250,000 borrowers with 10 features:

RevolvingUtilizationOfUnsecuredLines

[Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits]

age

[Age of borrower in years]

NumberOfTime30-59DaysPastDueNotWorse

[Number of times borrower has been 30-59 days past due but no worse in the last 2 years.]

DebtRatio

[Monthly debt payments, alimony,living costs divided by monthy gross income]

MonthlyIncome

[Monthly income]

NumberOfOpenCreditLinesAndLoans

[Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)]

NumberOfTimes90DaysLate

[Number of times borrower has been 90 days or more past due.]

NumberRealEstateLoansOrLines

[Number of mortgage and real estate loans including home equity lines of credit]

NumberOfTime60-89DaysPastDueNotWorse

[Number of times borrower has been 60-89 days past due but no worse in the last 2 years.]

NumberOfDependents

[Number of dependents in family excluding themselves (spouse, children etc.)]

And 1 label:

SeriousDlqin2yrs

[Serious financial distress in the next two years]

Solution Statement

(approx. 1 paragraph)

The solution for this problem would be to fit to machine learning model(LogisticRegression, DecisionTree, SVM from sklearn) or deep neural networks(Keras). The feature engineering, such as PCA or ICA will be used to get accurate results from the model.

Benchmark Model

(approximately 1-2 paragraphs)

The benchmark model for this project will be the random classifier. This is same as shooting an arrow blindfolded, which is why the model is such a great baseline model. The model's prediction from the test data will be compared to financial crisis classifier with the evaluation metrics below.

Evaluation Metrics

(approx. 1-2 paragraphs)

The model will be evaluated using both F1 score and area under the receiver operating characteristic curve(AUROC). Here are the explanation for each of them.

F1 score is the harmonic mean of precision and recall. Precision is the ratio of true positives to all positive predictions, and it is used to evaluate the model's suitability. Recall is the ratio of true positives to all positive examples, and it is used to evaluate the precision of the model.

AUROC curve is the trade-off for between the true positive rate and the false positive rate as the classifier's threshold is varied. The metric is only used for binary classification. The advantage of AUROC curve to F1 score is that it is not sensitive to imbalanced classes.

Project Design

(approx. 1 page) The workflow of the project would be:

Setup and Data Preprocessing

- Import all necessary packages and modules to jupyter notebook
- Load CSV file into Python
- Split features and labels
- Split the data with the ratio of 8 to 2 (8 for training set, 2 for testing set)

Data Exploration and Feature Engineering

- Analyze correlation for each feature and label with heatmap
- utilize sklearn.pca to combine features into 4 components
- Analyze correlation again

Training classifier

First of all, random classifier is trained. Training divides into 2 part; machine learning methods and deep learning methods.

1. Machine learning method

- Utilize sklearn.SVC.svm to train predictive model.
- Utilize sklearn.tree.DecstionTree to train predictive model.
- Utilize GridSearchCV to optimize training method and parameters.
- Utilize sklearn.linear_model.LogisticRegression

1. Deep learning method

- Use Keras and Tensorflow as backend to construct a neural net classifier.
- Use Randomclass from sklearn to create baseline which only guesses randomly.