Machine Learning Engineer Nanodegree

Capstone Proposal

Rabin Aryal

May 31th, 2018

Proposal

Domain Background

Financial distress is the term used in corporate finance to indicate a condition when business, household, or individual runs into a tight cash situation and cannot pay the owed amount to creditor on a due date. Financial institutions use a credit scoring algorithm, which determines what is the probability of credit default.

Machine learning techniques can be an important tool for financial risk management. [1] Predictive machine learning models are able to predict the probability of credit default could reduce the risk for credit originators. Better credit score systems could lead to lower overall borrowing costs. This could benefit both lenders and borrowers.

The data that is being used to predict the somebody will experience financial distress will be used from Kaggle competition website Give Me Some Credit - data.

Problem Statement

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.[2] For markets and society to function, individuals and companies need access to credit. Credit scoring algorithms, which make a guess at the probability of default, are the method banks use to determine whether or not a loan should be granted. The credit scoring algorithm can be improved using a machine learning that will predict the probability that somebody will experience financial distress in the next two years. The proposed model will be supervised classification [3] problem as the output is credit score (probability) that the customer will experience financial distress next two years based based on the features in the dataset as input. Based on the probability score, creditor as well as a borrower can make a wise decision whether to give/take loan or not.

Datasets and Inputs

The data is available from the Kaggle competition website Give Me Some Credit - data. The data includes

1. cs-training.csv (7.21 mb)
2. cs-test.csv (4.75 mb)
3. Data Dictionary.xls(14.50 kb)
4. sampleEntry.csv (1.82 mb)

The training and test dataset contains eleven features that is used to determine the probability that the customer will default the loan. Some of those features are as follows:

- Serious delinquency in 2 years
- Revolving utilization of unsecured line
- Age
- Debt ratio
- Monthly Income
- Number of open credit lines and loans
- Number of dependent

The training dataset consist of 150,000 borrowers, and the test dataset contains 101,503 observations, for a sum of over 250,000 borrowers gathered from historical data. The outcome will be the probability that the customer will default the loan in the next two years. The 30% of the training dataset will be used as validation dataset to test the model.

At quickly glancing the data, I notice that some of the feature have a missing or incorrect data. To be specific, "Monthly Income" and "Number of dependent" have missing data. The two column seems to be very important feature on determining a probability score, so they can not be discarded. The option here will be to replaced the data with a median value for each feature. which may be discarded while implementing the model.

The feature "Number of times borrower has been 90 days or more past due" have some incorrect data of value 98, which seems not to be true. The median for this feature seems to be 0 and does not make replace the incorrect data with median. I am thinking to discard this feature on my model.

Finally, the feature "Serious delinquency in 2 years" will be target variable and will be discarded from the feature that determines the probability score. At the end, the probability score can be compared with the "Serious delinquency in 2 years". Out of 150,000 training dataset, "Serious delinquency in 2 years" consists 139,974 (93.3%) records that are predicted to not have any

financial distress in next 2 year and 10,026 (6.7%) records that are predicted to have any financial distress in next 2 year. The dataset is highly unbalanced and "Serious delinquency in 2 years" cannot be used to evaluate the performance. Thus, area under curve will be used to evaluate the performance of the model being built.

Solution Statement

To solve this problem, a prediction model would be developed using several machine learning algorithms on this dataset to see which will give the best performance and will be judged against a predefined evaluation metric. Binary classification [4] algorithms including random forests, bayesian networks, and logistic regression will be evaluated. XGBoost [5] (trusted) and LightGBM [6](fast) algorithm will be evaluated in order to compare with prediction model using binary classification.

Benchmark Model

As this is a Kaggle competition a benchmark model is evaluated for area under curve [7], which is a metric for binary classifier. I will run the random forest classifier to get a base AUC. Then, I can compare our next model with it to see of it can beat it and by what extent. I will take the best model and for satisfying the curiosity, run it on test set provided by Kaggle as test dataset. The submission file will be uploaded on kaggle website to check the score. Then, I can also compare my model with the benchmark model hosted by Kaggle. A personal goal would be to be in the top 30% of the Kaggle Private Leaderboard.

Evaluation Metrics

The model prediction for this problem can be evaluated in several ways.Since the official evaluation of this project is done by Kaggle using area under curve (higher it is, better the model), same will be used for evaluation of models.

Project Design

The first step in my project will be data collection. As discussed above, I will download the dataset from Kaggle website. Then I will process/clean the data if needed, also remove the feature from the model if some data is missing or is outlier. The second step will be to divide 30% of the training dataset into validation dataset to cross validate the model (algorithms) that will be used. Then the final step will be to build a model using random forest classifier (x out of y [x < y], features will be randomly selected to predict the output), followed by other binary

classification algorithms bayesian networks, and logistic regression to evaluate area under curve. Then I will use XGBoost and followed by LightGBM to evaluate area under curve.

Whichever model from above gives higher area under curve will be the final model from which the prediction that the customer will default loan in next two year will be submitted .

Tools and Libraries used:

Python, Jupyter Notebook, pandas, numpy, scikit learn, matplotlib, XGBoost, LightGBM. Other libraries will be added if necessary.

References

1. Accurate Prediction of Financial Distress of Companies with Machine Learning Algorithms (https://www.researchgate.net/publication/225150865_Accurate_Prediction_of_Financial_Distress_of_Companies_with_Machine_Learning_Algorithms)
2. Give Me Some Credit (https://www.kaggle.com/c/GiveMeSomeCredit)
3. Classification Analysis(https://en.wikipedia.org/wiki/Statistical_classification)
4. Binary Classification (https://en.wikipedia.org/wiki/Binary_classification)
5. XGBoost (https://en.wikipedia.org/wiki/Xgboost)
6. LightGBM (https://github.com/Microsoft/LightGBM)
7. Area under curve (http://fastml.com/what-you-wanted-to-know-about-auc/)