Machine Learning Engineer Nanodegree

Capstone Proposal

Satish Gautam Irrinki

Domain Background

Corporate bankruptcy prediction is a very beneficial tool to have for wide range of practitioners in the financial services industry. Investors and investment portfolio mangers can better manage their funds and portfolio allocations when armed with the information about potential bankruptcy of the companies they invent in. Career corporate finance professionals can better plan for financial liquidity and take possible mitigating risk and cost reduction activities in response to early detection of potential bankruptcy. Other sectors in the industry that can benefit from the tool include firms in Private Equity, Investment Banking, Mergers and Acquisitions advisors, and Credit rating agencies.

The concept of bankruptcy is quite old with early bankruptcy laws dating back to 1841. These laws govern corporations to ensure business continuity that result in well being of the society through supply of products and services. Because bankruptcy ensure orderly control of operations or liquidation of firms in extreme competitive situations, predicting bankruptcy early will cause less social and economic damage. There is considerable research both in academia and in industry to address the prediction of possible bankruptcy. The academic paper Bankruptcy Prediction with Financial Ratios, http://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=3918017&fileOId=3918043 provides an illustration of the concept of using financial ratios for prediction. I will use financial ratios as well to predict bankruptcies in this capstone project.

I am personally motivated to implement this project because I have a very deep interest in the discipline of Finance. I pursued MBA degree with focus in Finance. I also successfully pursed Chartered Financial Analyst program (CFA program) and passed all three levels of the program.

Problem Statement

All pubic companies have their financial information publicly available from their filings with the regulators in their respective countries. The data available can be leveraged to evaluate its current as well as future financial performance. I will use Machine Learning algorithms (Support Vector Machine and Decision Tree Classifier) to train the bankruptcy prediction algorithms to classify the data and then test the algorithms to predict bankruptcy prediction accuracy.

Datasets and Inputs

A representative data set from UCI Machine Learning Repository, Polish Companies Bankruptcies Data Dataset,

https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data is used to train and test the bankruptcy prediction model. The dataset has one-year financial rates and corresponding classification label indicating bankruptcy status after five years. The file has financial data for 7027 companies. Among these companies, 6756 companies did not file for bankruptcy after five years and 271 companies were bankrupt. The raw data file has been processed to be completely compliant to be in tabular form and populated the missing values as NaN. Although the data set is for companies in Poland, the underlying data can be changed to represent any set of companies with appropriate feature columns, primarily because the models can be agnostic to the names of the features as well as leverage dimensionality reduction using Principal Component Analysis (PCA).

There are 64 features in the data set. Below is a set of sample features. The full listing can be found at: https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data#

X1 net profit / total assets, X9 sales / total assets, X10 equity / total assets, X23 net profit / sales, X27 profit on operating activities / financial expenses, X28 working capital / fixed assets, X40 (current assets - inventory - receivables) / short-term liabilities, X48 EBITDA (profit on operating activities - depreciation) / total assets, X64 sales / fixed assets

For example, X1, net profit/total assets, is an indicating of how well a company is using its assets; higher the value for X1 less likely it is to get bankrupt. Similarly other features influence the prediction.

Solution Statement

The solution to solving the bankruptcy prediction problem can be treated as a classification problem. Because the data set has a label representing the real state of the company after five years, we can use appropriate techniques from supervised learning. After reducing the number dimensions using PCA, I will use SVM and Decision Tree Classification algorithms to train, optimize, and test the models and compare the results such as accuracy_score and fbeta_score with those of a baseline benchmark algorithm that uses few representative raw features from the dataset. The motivation for using SVM and Decision Tree algorithms can be found in the thesis report referenced in the link:

https://dspace.library.uu.nl/bitstream/handle/1874/351884/Thesis-Frank_Wagenmans-3870154.pdf?sequence=1&isAllowed=y

Benchmark Model

The thesis referenced in the links above form the basis for the baseline benchmark model. I will use six features from the dataset and leverage Decision Tree classification algorithm with default hyperparameters to train, test, and measure results using accuracy_score and fbeta_score. Then, I will compare the results with the tuned SVM and DecsionTreeClassfication algorithms.

Because Decision Tree classification does not measure distances between data points but uses conditions on features as the underlying logic, there is no need to normalize the features

to make them consistent for measurement. So, only raw data will be used to define and measure the benchmark model.

Evaluation Metrics

I will use accuracy_score and fbeta_score to evaluate the performance of the models I build as well as for the benchmark model. I will then compare the results of each of the models relative to the benchmark and note observations of the relative performance of the tuned bankruptcy prediction models. The metric accuracy_score represents the rate of accuracy of predicted values relative to true values, and fbeta_score represents the weighted average of precision and recall. We need to ensure that we capture all bankruptcies, so we need a high recall model.

Project Design

The project design follows the following steps:

- Pre process the input features to fill missing values with average of the feature values
- Normalize the data using MinMaxScalar so that the transformed data can be used for PCA
- Iteratively select suitable number of components for PCA for dimensionality reduction
- Use the principal components as inputs to classification algorithms SVM and DecisionTreeClassifer
- Optimize the algorithm to using GridSearchCV method
- Capture model performance metrics defined by accuracy_score and fbeta_score for both SVM and DecisionTreeClassifer
- Capture the baseline model metrics, accuracy_score and fbeta_score and compare to that of the model
- Analyze and report the observations

Scaling the input data is necessary to have an accurate dimensionality reduction using PCA. PCA uses distance and different scales for different features can skew the components. An iterative approach will be used to select the number of PCA components so that 94 to 95% of the explained variance is captured in the components.

These dimensions will be provided as inputs to the classification algorithms, SVM and DecisonTreeClassifer, identified earlier. The build models will then be optimized for best performance using combinations of hyperparameters. Then the models will be trained and tested for accuracy using the metrics accuracy_score and fbeta_score from sklearn. These results will then be compared with those of the baseline benchmark model. Conclusions will then be drawn from the observed results.