NYU | TANDON SCHOOL OF ENGINEERING

Course: CS-GY 6053 Fall 2017 **Instructor**: Rumi Chunara, PhD

Project Members: Praphulla Bhawsar (pmb418), Paresh Patil (pap408)

PROJECT REPORT

- **Problem Motivation:**

  Banks and such other financial lending institutions often need to look at a loan applicant's credit history, economic status and such other factors to determine eligibility for loan, but the relationship between these factors is generally not well-defined and is most likely heuristic in nature. It is very often the case also that the company's current standing, such as its rise or fall in the immediate past is taken into consideration to determine its financial stability. This might lead to erroneous judgement with regards to the company's likelihood of defaulting on a loan. Using effective classification and time series analyses, we can generate a model that would not only be more precise but also cost-effective in solving this problem. With this objective, we shall analyze the data, augment it with data from other sets and try to understand the static factors that correlate most highly with a company's financial standing by way of creating a classification strategy and analyzing the steps taken by the model.

- **Target and Predictor Variables:**

  A total of 42 features were used in order to determine the target in the final classifier, which itself was a compounded variable obtained from the output of the SVM classifier and the ARIMA Time Series Analysis.

  The most important predictor variables used were: Accounts Payable, Capital Expenditures, Additional Income Expense Items, Accounts Receivable and After Tax Return on Equity.

  The Target Variable was a sum of the percentage change of the organization's stock price over a period of 2 years and the likelihood of it going bankrupt in the next 4 years.

- **Problem Statement:**

  To Identify various static features which are responsible for determining a company's growth trend and consequently its eligibility for a loan.

- **Type of Model:**

  The datasets taken up include data from companies that have filed for bankruptcy, 6-year stock trends of organizations from the New York Stock Exchange and financial data of these companies. We tried out multiple models on the bankrupt companies dataset, including Decision Trees, Linear Models and Logistic Regression, and concluded that a Support Vector Machine would fit the dataset best based upon the AUC values. An ARIMA Time Series analysis was made on the stock trends. Using these values, a compounded label was added to the financial dataset, and we used a Random Forest classifier on this dataset to solve the problem statement. Due to the presence of a lot of features and large variance resulting from the data being of different companies, the Random Forest model fit the data the best and provided the best overall accuracy.

- **Evaluation Approach:**

  The task of evaluating the data was completed in 4 steps:

  1. Data Cleansing: Since the data we operated upon came from multiple datasets from distinct sources, data cleaning was a major operation that was necessitated to ensure that the data representation was consistent across these data sets.
  2. Bankruptcy Prediction: The dataset of the companies that filed for bankruptcy did not contain the financial attributes of non-bankrupt organizations, and thus we did not have a dataset that could be used to train a model directly. To solve this problem, we used data for previous years from the larger financial dataset of companies listed on the NYSE and added it to this dataset to ascertain a ground rule. Then we trained an SVM on this dataset, getting an AUC of ~0.75.
  3. Time Series Analysis: The stock prices data contained daily closing values for about 500 organizations listed on the NYSE. This data was scaled down to contain averaged weekly stock prices. Since the time series would behave differently for different companies, it was necessary to model each of them separately. Quite expectedly, the data showed heavy trend and, in some cases, seasonality, which had to be removed via differencing to perform ARIMA on this dataset. The p and q values selected were 2 and 1 respectively based on ACF and PACF plots analysis and manual trials. Upon modelling, a Mean Absolute Error of ~0.05 was obtained, thus suggesting that the Time Series analysis was fairly accurate.
  4. Augmenting initial dataset with the predicted data: The dataset containing the financial information of the NYSE-listed organizations was augmented with a compounded label of the predicted bankruptcy values and a percentage change of their respective stock prices over a time period of 2 years. The label being continuous, we binned it by rounding it to the nearest tenth decimal place and multiplying it by 10 to get an integer. This label was used to train a Random Forest Classifier in order to ascertain the features it deemed to be most important to predicting the growth trends of the company. The features were identified and respective correlations were found between the features and the label. The analysis of the Random Forest classifier itself was evaluated by observing the confusion matrix generated and by it.

     An ROC curve could not be generated for this since the label was multi-class and not binary, though a good measure of confidence was found by the ~0.5 value of the Matthews correlation co-efficient.

- **Assumptions/Limitations:**

  1. Bankruptcy prediction is a major topic of Machine Learning research. There are several research papers on the topic, and several of them employ Neural Networks and advanced Machine Learning techniques to reliably predict the likelihood of

bankruptcy. In our own research, we found that some attributes were found to be commonly used across these models, and we made the assumption that these must correlate most highly to the probability of a company going bankrupt. Unavailability of public datasets containing more financial data of bankrupt companies was also a big factor in this choice. Therefore, only these most highly correlating features were used to train the predictor, and thus it is an extremely simplistic model of bankruptcy prediction.

2. The ACF and PACF plots were found to be quite ambiguous and were not conclusive enough to help decide the AR and MA parameter values with definitiveness. We therefore tried out a few values, and we have assumed that the (2,1) combination predicts the data the best.
3. The augmentation step included a merging of the binary predicted bankruptcy value and the continuous averaged Time Series prediction. We have assumed that this is a good indicator of whether the company is rising or declining, and consequently, whether it is safe to provide the company a loan.

- **Problem in Scope of Class:**

  Having used multiple Classification strategies as well as modelling the time series, this analysis can be taken further by increasing the number of features and data points to enable better bankruptcy prediction, as well as tuning the AR and MA parameters for the time series model. Given the current analysis, we found multiple features that understandably correlate with the target variable, and this analysis can help to supplement conventional, heuristic knowledge of the organization. The information store of banks and other financial lending institutions can use this analysis to focus more on these features, something that might not be possible by way of representative or inferential analysis.

- **Changes from original proposal and its reason:**

  The original proposal by us contained a strategy to use just the bankruptcy predictor to provide the label for the financial dataset. However, on careful analysis, we found that it could not be a comprehensive enough indicator of the company's overall standing. We therefore decided to augment it with a Time Series analysis of the company's stock trends, which together would be a much better indicator of organizational growth/decline.

TEAM EVALUATIONS:

Praphulla Bhawsar (pmb418) – 5 pts
Paresh Patil (pap408) – 5 pts