US Stock Performance Exploration and Prediction (Group P22)

Yash Mishra ymishra@ncsu.edu
Anirudh Pande apande@ncsu.edu
Mithil Ghinaiya mghinai@ncsu.edu
Vraj Chokshi vschoksh@ncsu.edu

1 Background

1.1 Problem:

There aren't many fields as lucrative and dynamic as the stock market. Since many stocks are traded on stock exchange, numerous factors influence the decision making process. Building an accurate stock prediction model is still a challenging problem in this age of machine learning and deep learning. In addition to historical prices, the current stock market is affected by the mood of the society and many other unpredictable factors. While predicting stock values flawlessly cannot yet be done even by a maestro, an increase in computational power and storage capacity have enabled us to try our hand at algorithmic predictions. In this paper, we aim to utilize the available financial data to classify companies where it is profitable to buy their stocks at the start of the year and sell it at the end of the year for a profitable transaction. Our refinement was to go generalize and handle data of past five years from more than 4000 unique companies and not focus on a few companies and their stocks for prediction. We look beyond the time series data which is commonly used by other researchers and focus on 10K filings which is reported by companies at the end of each year.

1.2 Literature Survey:

Stock market forecasting is one of the most popular topics in academic and real-world business. Many researchers have attempted to answer the question of stock market prediction using various Machine Learning and Deep Learning techniques. Mehtabhorn Obthong et al [11] have used Bayesian networks, time series method such as Auto Regressive Model and so on to discover the pattern in the data, their work does not consider the sentiments on the social media in this paper or other non-financial factors which may have a big influence on the stock prices. Their survey shows other important papers which have considered such factors.

Lahmiri S et al [8] have focused on ensemble methods like various bagging and boosting models for financial data prediction. While most of the work concerned with stock market prediction focuses on time-series data to predict the stock values, there is a prominent use of Recurrent Neural Network for this task. However, works by Attigeri et al[4] show 70% accuracy with Logistic Regression and Dang Duong [5] achieved an even higher accuracy of 73% with SVM in stock market value prediction by using news articles instead of Twitter Sentiments. With few works also incorporating twitter sentiments for the classification. For eg: Nguyen et al. [10] show a 54% accuracy with SVM model by incorporating twitter sentiments. Twitter has been shown to be an important source for public opinion extraction using Machine Learning models by many different papers [3]. With only few companies being covered by the above papers, the conclusion might be insufficient and unreliable. To the best of our knowledge, there is no research paper showing a good prediction result on a data consisting of many stocks in a long time period with use of 10k filings of companies.

2 Method

2.1 Approach

We will be utilizing 10k filing financial data of over four thousand companies from 2014 to 2018 for our classification task. We will start by cleaning the data and performing Exploratory Data Analysis to understand how our features are related to one another. We aim to use multiple Machine Learning models to classify companies where the transaction of purchasing their stocks at the start of the year and selling it at the end of the year is a profitable transaction. The classification task is done using seven different machine learning models: Decision Tree, Random Forest, Logistic Regression, Gradient Boosting, XG-Boosting, SVM with different kernels, Naive Bayes. For the classification task, we aim to perform it twice, once using the standardised version of the data and another using the normalised version of the data to understand the best scaling technique for our data. Also, principal component analysis is applied after scaling to reduce the number of features given to the classification models. Our work, independent of classification, uses Ridge and Lasso Regression to identify features which are important for the classification task. Finally, we augment our classification work by introducing Twitter Sentiments to understand if it can help us classify profitable companies any better.

Since, we found no work which uses 10K fillings to identify companies where it is profitable to buy their stocks at the start of the year and sell it at the end of the year, we believe that our work will introduce a new perspective of looking at stock market prediction. The use of seven different machine learning models and using twitter sentiments to bolster our model will likely lead to an exhaustive work in the field of machine learning for stock prediction using 10k filling reports.

We start with Data Cleaning of the input data so that the dataset contains records that can be worked on by Machine Learning Models without erratic behavior.

2.1.1 Data Preparation:

We combine all five data sets into one dataset. This combined dataset now contains all the data from the year 2014 to 2018.

Upon further data inspection, we found that the companies present in the dataset belonged to different countries. By naming, its is implied that columns such as revenue and costs contains monetary value. From cross comparison with openly available secondary sources, we found out that the numerical values did not follow a common denomination and before processing the data we would need to convert the currency columns into a singular currency.

As discussed above, the data contained companies have their financial information in currencies other than American Dollar. This lead to huge discrepancies. This was solved by using the Nasdaq ticker label data [2] which allowed us to perform inner join on our dataset with the Nasdaq data and filter out countries which were outside of United States. We finally deleted columns which were not relevant to predicting the final class.

2.1.2 Data Cleaning:

Initial Data Frame Shape (after merging) was 22077 rows, 226 columns. Few columns are renamed, and unnecessary columns are then removed. Null(NAN) values were cleaned by removing columns having more than 20% null values from the dataset. Similarly, rows with any null value is removed from the dataset and final shape after operation was 12804 rows and 171 columns. We observed that a few columns have more than 50% values 0. We hence deleted all such attributes having more that 50% values as 0 hence contributing to sparse Dataset. One Hot Encoding is applied on Sector columns as it had string value, and machine learning models need numerical data.If the column contained just one value, it was deleted as their variance was zero. Duplicate rows are checked and deleted. All the data cleaning task gave us the final Data Frame of shape 9358 rows and 182 columns

Next, we do Exploratory Data Analysis (EDA) on the data to find out patterns and relations between the attributes. We aim to get more insights and understand the data further using the analysis results.

2.1.3 Exploratory Data Analysis:

We check if the output class (Attribute to be predicted) is balanced or not. As the output class contains 59% profitable and 41% lossy companies, we conclude that the data is balanced enough to not perform imbalanced data handling. We then find correlation between different attributes and arrange and plot them according to their correlation value to the "Class" variable. We then plot Box plots to visually depict data variance of columns which have the highest average and standard deviation and lowest average and standard deviation. This helped us see that features with high average and high standard deviation are linearly dependent, hence, not much useful to keep all of them for modelling. Where as, columns with the lowest average and standard deviation showed very peculiar trend with data of one being explained almost entirely by the other.

We then proceed to outlier treatment. And employ multiple methods for outlier detection and below are the observations:

2.1.4 Outlier Detection and Treatment:

Standard deviation method: Number of identified outliers are 3335. IQR Method: Just 9 records were identified as non-outliers, leading to a heavy loss of data if we used this technique for outlier treatment. 99-1 percentile method: non-outlier observations: 7870. We finally select this method for outlier detection as large amount of data is otherwise being deleted from all the other methods. Our final data shape is: 7870 x 181. Finally, we get acceptable number of non-outliers by discarding data points above and below 2.5 standard deviation.

Now that our data have been cleaned to acceptable level, we are proceeding for Data Modelling. Our first step for the same is to find a appropriate scaling algorithm for our use-case.

2.1.5 Important Feature Identification:

We use Ridge and Lasso Regression to identify columns which are important for the identification of profitable and lossy companies. The level of importance is defined as the absolute value of coefficients which is assigned to the feature after performing ridge and lasso regression. One important thing to observe would be to see if there are features which are unanimously called important by both ridge and lasso regression.

2.1.6 Scaling:

We perform both normalisation and standardisation to understand which scaling technique is better for our machine learning models for classification. We then compare the results and decide the final scaler to be used.

Normalisation: We are using MinMax scaler to normalize the data. Post that with the help to PCA, we are trying to identify the number of dimensions needed to explain 95% variance. In such a case, we find that 47 such components are needed.

StandardScaler: StandardScaler scales each feature by the standard deviation value of the entire dataset. Post that with the help to PCA, we are trying to identify the number of dimensions needed to explain 95% variance. In such a case, we find that 74 such components are needed.

PCA is performed on StandardScaler and MinMaxScaler and the most significant 74 components are then used for Data Modeling.

2.1.7 Classification:

Random Forest: It takes the average of many decision trees where each tree is a weak classifier and weaker than a full decision tree, but the combination of all the trees delivers a superior overall

classifier. They are relatively fast to train but not easy to understand predictions and can become a black box. Few works [7] have shown good results with random forests, especially in the metric of precision.

Decision Tree: These are graph like structures which use branching methods to match possible outcome of a decision. They are easy to understand and implement but not often used for their simplicity is their undoing and they are unable to explain the variance of complex data sets. Margaret et al [9], in their paper in 2010 use decision tree and artificial intelligence for stock market prediction with good result.

Boosting Algorithms: It uses even weaker decision trees that increasingly focus on harder and more difficult examples. They are often high performing but even a small change in the future set or training set can create radical changes in the model. We have discussed about the history of its efficacy in stock market prediction in our literature survey [8].

Logistic Regression: It adopts the linear regression mathematical equation to classify problems. Though easy to understand, sometimes too simple to be able to capture the complex relationship present between variables and has a tendency to over-fit. Also, there are strong assumptions made to fit in their mathematical model and they are sensitive to outliers [13] [4]

Support Vector Machine: It divides the data points by a hyper plane. If the data points are not lineally separable, then "kernel trick" is used. [10]

Naive Bayes: It is a probabilistic model. It is not very popular in stock market prediction but we include it as a proof of concept and to understand how it performs anyway. There are few works which have used Naive Bayes Classifier for stock market prediction with the classifier augmented by Auto-Regressive Moving Average [12].

2.1.8 Twitter Sentimental Analysis:

Dataset [6] used was part of the paper published in the 2020 IEEE International Conference on Big Data under the 6^{th} Special Session on Intelligent Data Mining track, it contained tweets related to five different companies namely Google, Microsoft, Tesla, Amazon, and Apple. In preprocessing we segregated data year-wise and further divided company-wise. We have considered 5000 tweets for each year per company. Sentiments classified into positive, neutral, and negative for each tweet are noted and extra weights are added for likes and retweets on that tweet.

2.2 Rationale

We have tried to be exhaustive with our machine learning models for the classification task. We have studied the results achieved by previous researchers for different machine learning models and tried understood their limitations. We try to overcome it by adding twitter sentiments to help predict profitable and lossy companies better. Since there was no clear answer on whether normalisation is better than standardisation (or vice-versa), we have applied machine learning models on both the scalers and then compared the result.

The metric chosen was Precision because we agreed that it is okay if a person does not earn money from using our classifier in the real world but it is not acceptable if loses money. Therefore, we had to minimise false positives and that is why, precision was chosen.

The hyperparameters for all were machine learning models were tuned using 10-fold cross validation technique on our training dataset. We also performed several train-test split of our data and found that a train-test split of 80-20 gave the best precision. Since a lot many data points were lost because of outlier treatment, validation data was not used for hyper-parameter tuning.

3 Plan and Experiment:

3.1 Datasets:

200+ Financial Indicators of US stocks (2014-2018) | Kaggle [1] – Size: 22007 rows, 225 columns. The dataset has five sets, ranging from stock data from 2014 to 2018. Each dataset contains 225 financial indicators, that are commonly found in the 10-K filings each publicly traded company releases yearly (on average, 4k stocks are listed in each dataset). There are 22,007 total rows and 225 attributes. The last column of the dataset represents the class of each stock, where: if $\text{stock}_{\text{year start}} > \text{stock}_{\text{year end}}$, class = 1 else class = 0

The dataset is available in the following files containing company data from different years.*{ 2014_Financial_Data.csv, 2015_Financial_Data.csv, , 2016_Financial_Data.csv, 2017_Financial_Data.csv ,2018_Financial_Data.csv}*

Overall, when combined the dataset has around 20K rows and 225 columns. Some columns contain NaN cells (missing values). Additionally, the dataset contains outliers that are recognized and treated accordingly in our data cleaning phase.

3.2 Hypotheses:

Stock market is a dynamic and ever changing field in the knowledge of outcome is invaluable. The team wanted to explore data analytics techniques, especially machine learning to put into use in field of stock market. Broadly, we hope to address the following questions:

- **Model:** Which machine learning model is most suitable to predict if a company stock will rise by the end of the year compared to the stock price at start?
- **Parameter:** Which parameter of the Machine Learning model should be used in stock market use-case?
- **Important Feature Identification:** Weather Ridge and Lasso would anonymously agree on any important.
- **Scaling Techniques:** Which scaling technique *{ Standardization, Normalization }* is better for our use case.
- **Classification:** Weather classification on sector wise segregated data-set better than classification on the overall population ?
- **Satisfactory:** Are Machine Learning models enough to produce satisfactory classification results?
- **Improvement:** Can results obtained from ML models be improved with further analysis from twitter? Team will aim to employ sentiment analysis for this measure

3.3 Experimental Design:

We plan to answer the below questions related of ML Modeling utilizing the below setup.

- **Model:** We created seven different Models and saw which one bests predicts the final classification output. The best model was the one with the highest Precision.
- **Parameter:** We performed 10-fold cross validation on our training data to choose the best hyper-parameters for our machine learning models.
- **Important Feature Identification:** We performed ridge and lasso regression and had to use GridCV for cross-validation here as well, to come up with the best alpha value which acted as the penalty parameter. This was done independent of our classification task.

- **Scaling Techniques:** We used both *{ Standardization and Normalization }* to understand which scaling factor was best for us. We found that standardisation gave us the highest precision. Using two different scalers for classification, increased our work by one hundred percent but it was important to make no strong assumption.
- **Classification:** We performed classification of the normalised and the standardised version of the data and then filtered companies based on their sector to identify profitable and lossy companies. Though we did not see any significant difference in sector-wise analysis but it proved that sector-wise analysis gives no added advantage.
- **Satisfactory:** We used seven different classification model on two differently scaled data on sector-wise data as well as the whole data to identify profitable and lossy companies.
- **Improvement:** We plan to utilize sentiment analysis to see the general trend of public sentiment throughout the year regarding the company's stock we are tying to predict. This will reveal the exciting result that if the public opinion is a good enough indicator to aid our ML Model's finding.

4 Results:

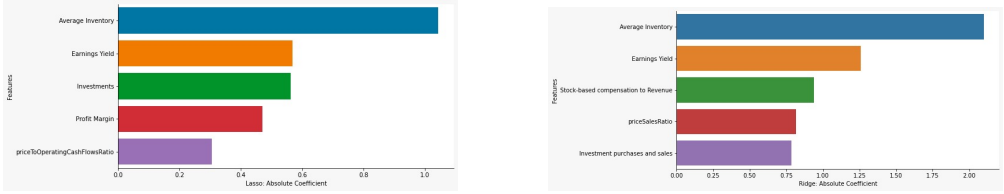
4.1 Results:

There are four important results of our experiments and they are as follows:

- **Classification Model:** Decision Tree turned out to be the best model for our classification task as it gave the highest precision. The table below shows the precision of all the machine learning models used in our experiment:

Machine Learning Model	Precision
Decision Tree	0.786
Random Forest	0.750
Logistic Regression	0.720
Gradient Boosting	0.665
XG-Boost	0.645
SVM (RBF-Kernel)	0.626
Naive Bayes	0.601

- **Important Feature Identification:** After performing ridge and lasso regression, we found that both of them unanimously agreed that "Average Inventory" and "Earning Yields" are the top two most important feature in identifying profitable and lossy companies.



(a) (b)

Figure 1: (a)Lasso Regression Coefficients (b) Ridge Regression Coefficients

- **Linear Regression Residual Analysis:** There was feature named "Next_year_variance" in our dataset which was removed for classification task but it was used regression. It was the percent change in stock value from the previous year. On performing linear regression assumption analysis, we found that not all assumptions of linear regression are preserved, indicating that the linear regression's result cannot be trues ted.

  There are four assumptions associated with a linear regression model:

  - **Linearity**: The relationship between X and the mean of Y is linear.

- **Homoscedasticity**: The variance of residual is the same for any value of X.
- **Independence**: Observations are independent of each other.
- **Normality**: For any fixed value of X, Y is normally distributed.

We found that the second assumption is violated and the variance of residual is not the same for any value of X.

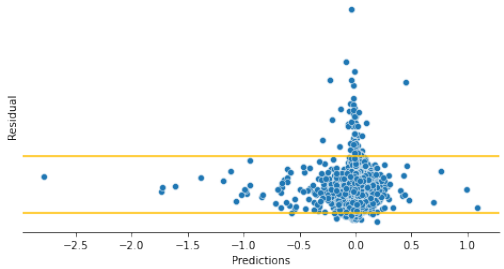


Figure 2: Homoscedasticity Violated

- **Twitter Sentiment Analysis:** For each company, graphs are plotted showing the percentage of positive, negative, and neutral tweets from 2015 to 2018. Trends show a rise in positive tweets and a dip in negative tweets is the sign of increase in stock prices for that year. After plotting we get 80% of companies shows this trend.

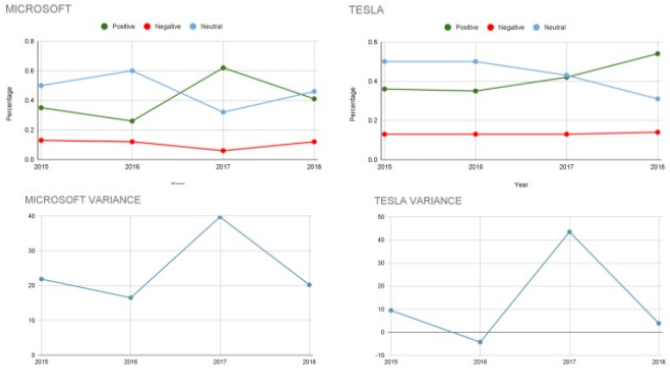


Figure 3: Twitter trends

4.2 Critical Evaluation:

Though the precision might seem to be on the lower side, it is better than most of the experimental strategies that we referred during this project. To enhance the results even further, we need to target specific companies and apply time series data approach along with 10K filing and twitter sentimental analyse on them. The trade off here is that we will not be able to cover all the companies and our model would then target certain companies only.

5 Conclusion:

In this project we used several machine learning models to classify profitable companies from lossy ones using 10k filing reports from companies. We found that standardisation is the best scaling technique and decision tree gave the highest precision out of the seven models used. We also saw how regression does not produce trustworthy results as one of the assumptions do not stand. And explored the use of Twitter Sentiment Analysis to bolster our classification task.

There is a reason why there is only one Warren Buffet in the world as no Machine learning model can guarantee you profit all the time in the dicey world of stock market. One needs to be wise and use his understanding of the market for his investment in this unpredictable chaos.

References

[1] 200+ financial indicators of us stocks (2014-2018). `https://www.kaggle.com/cnic92/200-financial-indicators-of-us-stocks-20142018?select=2018_Financial_Data.csv`. Accessed: 2021-11-25.

[2] Nasdaq ticker label data. `https://www.nasdaq.com/market-activity/stocks/screener`. Accessed: 2021-11-25.

[3] K. Arun and A. Srinagesh. Multi-lingual twitter sentiment analysis using machine learning. *International Journal of Electrical and Computer Engineering*, 10:5992–6000, 12 2020.

[4] Girija V Attigeri, Manohara Pai MM, Radhika M Pai, and Aparna Nayak. Stock market prediction: A big data approach. In *TENCON 2015-2015 IEEE Region 10 Conference*, pages 1–5. IEEE, 2015.

[5] Minh Dang and Duc Duong. Improvement methods for stock market prediction using financial news articles. In *2016 3rd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS)*, pages 125–129. IEEE, 2016.

[6] Mustafa Doğan, Ömer Metin, Elif Tek, Semih Yumuşak, and Kasım Öztoprak. Speculator and influencer evaluation in stock market by using social media. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 4559–4566, 2020.

[7] Luckyson Khaidem, Snehanshu Saha, and Sudeepa Roy Dey. Predicting the direction of stock market prices using random forest. *arXiv preprint arXiv:1605.00003*, 2016.

[8] Salim Lahmiri, Stelios Bekiros, Anastasia Giakoumelou, and Frank Bezzina. Performance assessment of ensemble learning systems in financial data classification. *Intelligent Systems in Accounting, Finance and Management*, 27:3–9, 1 2020.

[9] Margaret Miró-Julià, Gabriel Fiol-Roig, Andreu Pere, and Isern Deyà. Lnai 6096 - decision trees in stock market analysis: Construction and validation. pages 185–194, 2010.

[10] Thien Hai Nguyen, Kiyoaki Shirai, and Julien Velcin. Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 42(24):9603–9611, 2015.

[11] Mehtabhorn Obthong, Nongnuch Tantisantiwong, Watthanasak Jeamwatthanachai, and Gary Wills. A survey on machine learning for stock price prediction: algorithms and techniques. 2020.

[12] Mahajan Shubhrata, Deshmukh Kaveri, and Samel Bhavana. International journal on recent and innovation trends in computing and communication stock market prediction and analysis using naïve bayes.

[13] Lang Wu and Menggang Li. Applying the cg-logistic regression method to predict the customer churn problem. In *2018 5th International Conference on Industrial Economics System and Industrial Security Engineering (IEIS)*, pages 1–5. IEEE, 2018.

Appendix

GitHub Link: https://github.ncsu.edu/ymishra/P22-ALDA-fall2021