Airline Stock Prediction Based on Alternate Data

Mohapatra, Jyotirmoy
New York University,
New York, USA
jm7432@nyu.edu

Panse, Shubhankar
New York University,
New York, USA
ssp573@nyu.edu

Shah, Sohil
New York University,
New York, USA
sss857@nyu.edu

***Abstract*—**

Stock prices have a lot of volatility and become very difficult to deterministically predict. But, this volatile nature could have some correlation with various factors. In this paper, we intend to take a use case of Airline industry stocks and find the correlation between the stock market price of top performing airlines in the USA, interest over time data of the terms searched on Google, and the crude oil US market prices. We use regression models like Linear Regression and Logistic Regression using MLLib on Spark and predict prices for these airline stocks. Based on results of our analysis, it is possible to reasonably predict if the stock prices will go up or down based on the orthogonal data used.

Keywords*— *airlines, analytics, crude oil, regression, stocks, trends

I. Introduction

A lot of research has been concentrated on predicting stock markets or bitcoin prices using public sentiments mined from Twitter and other sources. We believe stock prices could be dependent on various other factors. Through this project, we intend to showcase the relationship between daily stock prices and factors such as macroeconomic variables like oil prices, interest over time data of a certain curated list of terms on Google's search engine.

The application built, based on this correlation, will be able to predict the stock price of a company which can then be used to make decisions about whether to buy or sell the stocks. For demonstration purposes, we have focused on airline industry stocks. We have selected stocks of major 10 airlines in the United States of America. The application finds correlation numbers between the stock prices and crude oil index and Google interest over time data and helps in predicting the stock price in near future. We have used regression algorithms like Logistic regression, Isotonic regression, and Decision Tree regression to make the predictions for test data.

II. Motivation

The primary users of this application will be stock market investors, traders, the airline industry as well as airline customers. Stock buyers, brokers, financial investment firms can use this application to predict the future movements in the stock market. It will also help the airline companies in predicting their stock prices. Moreover, they can analyze what search terms are strongly correlated to their growth which would give an idea of how they could optimize their business strategies. These airlines can utilize this application further, for making pricing decisions based on topics that are trending and the jet oil prices based on US crude oil index.

III. Related Work

There has been significant research focused on predicting stock prices based on the sentiment of the people mined from various social media applications. Bollen et al. [13] study how the collective mood tracked using twitter feeds impacts Dow Jones Industrial Average (DJIA) over time. They achieve an accuracy of 87.6% in

predicting if the daily stock prices will go up or down. Attigeri et al. [1] use sentiments mined from Twitter and news articles along with historical data to predict stock market prices. These papers have dealt with sentiment mining of available data for prediction. We now look at how simple hit counts of certain words on Google's search engine are good indicators for stock market prediction.

Preis et al. [8] discuss the correlation between the interest over time value of 98 random terms and Stock Market prices. This work is in great alignment with our idea that stock market behavior is correlated with Google query volumes of certain words. The results in this paper show that when terms like "debt", "profit" and other such financial terms have a high volume of searches on Google's Search Engine, the stock markets tend to fall in the near future. One possible interpretation of this behavior explained in the paper is that selling on the financial market at lower prices is generally after a period growing concern during which people will go to the Internet to search for important information regarding the market.

Studies have shown that oil prices data is also a significant pointer specific to the Airline industry. The work of Ryota et al. [15] relates a lot to our idea of using interrelated data to predict stock markets. This paper brings in the exact motivation of what we intend to work on i.e. extracting inter-relationships of predicted stock price and various other time series data, such as other stocks time series, foreign exchanges and oil prices from real data automatically. The method they have used calculates the variation patterns which represent how referenced time series have changed using Evolution strategy of historical data, and then utilized extracted data (interrelated time series data) to predict the stock using correlations. They also used 'Likely Interrelated Data' (LIDs) technique to filter only the appropriate and interrelated time series. This is very similar to our current approach.

IV. Application Design

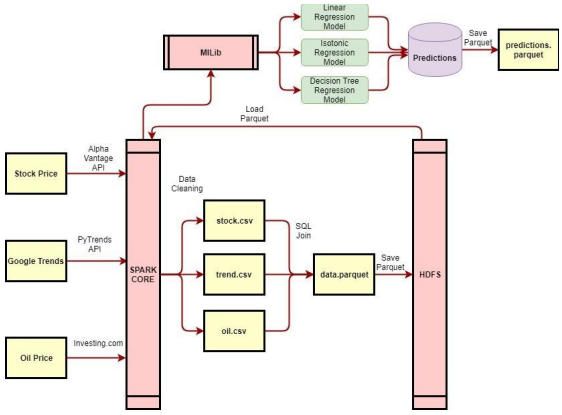

Fig. 1

Our approach is to combine the three datasets – major US airline stock prices data, Google interest over time daily data for a curated list of words and the US crude oil index data. As shown in Fig. 1, the three datasets are passed to Spark Core as csv files. They are then cleaned and transformed to the required format using SparkSQL Dataframes. This cleaned and transformed data is then stored in HDFS in the form of parquet files. We then use this parquet file to load the required Dataframe for the regression models. We use Spark's ML library for leveraging Linear regression, Isotonic Regression and Decision Tree Regression models to find correlations and predict the near future stock prices. At the front-end of the application, the results obtained are visualized as trend charts on Tableau.

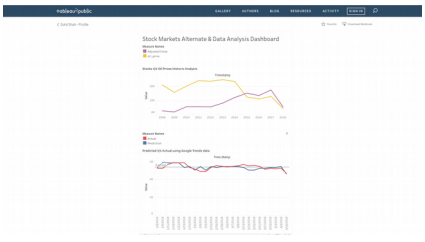

Fig. 2

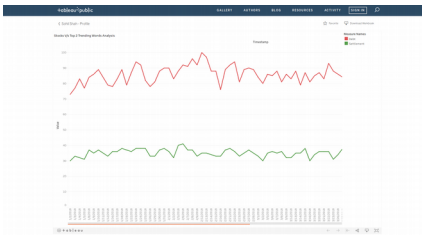

Fig. 3

The visualization shown in Fig. 2 and Fig. 3 has three graphs. The first one in Fig. 2 is the historical correlation between oil and stock prices. The second one is a comparison of the stock prices predicted by our model versus the actual stock prices. The third one in Fig. 3 has the time-series of two words which affect the regression model the most.

V. DATASETS

We are using three orthogonal datasets. The first is the stock prices data of United Airlines from AlphaVantage. It has the timestamp, opening, and closing prices, highest and lowest price for the day as well as volume traded for the day. The second dataset is the daily Google Trends interest over time data. This data has interest over time values of the curated list of words which are most correlated to stock prices. Interest over time is a measure of how relatively popular a certain term was on the Google Search Engine (the value ranges from 0 to 100). The list of words was carefully curated to include words that could positively as well as negatively affect the stock prices. Some of the positive words used were 'Christmas', 'New Year's Eve', etc. while the negative words were 'settlement', 'debt', 'crash', etc. The third dataset is of Crude Oil prices. The fields are the timestamp, average price, opening price, highest price, lowest price, volume traded for the day, change with respect to the previous day.

VI. REMEDIATION

The primary aim of this work is to be able to predict stock market prices. This is done using the orthogonal datasets. This is a powerful tool for stock brokers, buyers, and sellers. Being able to predict stock prices for the future to a reasonable level of accuracy will result in right decisions being made by all parties involved. This includes buying or selling stocks, constructing the strategies of your company based on stock market fluctuations, etc. These predictions can also be used by companies to compare the predictions with their actual stock prices. The prediction is a good indicator as to where the stock price should be considering all the values for the fields in the dataset. If the actual stock prices don't reflect the same intuition, the user can be sure that it is an anomaly which needs to be found and fixed. Another remediation is finding the terms that have the most impact on a company's stock prices. The linear regression model is fit on the training datasets by varying the coefficients for each term. The equation for linear regression is shown below.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots. + \beta_n X_n + \varepsilon$$

Y is the predicted value and ß1… ßn are the coefficients assigned to each independent variable X1… Xn. The ß values are varied over the training iterations to come up with the best hypothesis. At the end of the training period, we get a set of ß that are a measure how well correlated the corresponding independent variable is to the dependent predicted variable Y. A value close to 0 means there is little or no correlation while high negative and positive values (> |0.5|) means there is a high correlation.

VII. EXPERIMENTS

In this section, we discuss our experimental setup, obstacles we faced with our data and the limitations of the application.

We have used three orthogonal data sources. We prepared the experimental setup by joining all 3 datasets – major US airline company stocks, crude oil USA index and Google Trends interest over time data. After collecting the data we cleaned it and removed the features that were not useful for our experiments. From the stock price dataset, we used the timestamp and the adjusted closing values of the stock. From the oil price dataset, we used the timestamp and the opening price values.

We conducted three experiments in order to find out which factors affect the variation in stock prices the most. In our experiments, we used three regression models, Linear Regression, Isotonic Regression and Decision Tree Regression. We evaluated these models and extracted insights about the performance of each model and the correlation measure of each feature used.

Our first experiment was using all the data that we had. We used oil data, interest over time data and stock prices data. Our train/test data split was 70:30. We found that the regression model performed reasonably well with the best Root Mean Squared Error (RMSE) of 2.55 on Decision Tree Regression model.

We then tried to improve our model and find out which dataset was hindering the model. We ran the model on just the oil prices dataset. It performed very poorly with a best RMSE of 5.9 using Decision Tree Regression model. Our hypothesis is that this was because oil prices have been stable lately and hence have not had much impact on the stock prices.

Following up on the previous experiment, we then used only the interest over time data which produced improved results with an RMSE of 2.1 on the Decision Tree Regression model.

Our experiments have ruled out crude oil prices as an indicator for predicting stock prices. We also learned that the coefficients that were collected from Linear Regression suggest that “debt” and “settlement” are the terms which have a high negative correlation with stock prices. This is because as people start getting anxious, the number of Google searches for these terms goes high and the stock prices fall as a result of this anxiety. Thus our application is a good indicator of how the general sentiment of the people affects the stock prices. This result in alignment with the results of Preis et al. [8]

We would like to generalize our work better and not have it specific to the Airline industry. That entails using other orthogonal datasets, which we plan to do in the future.

Data Problems:

We faced several problems with cleaning and profiling our data. Firstly, the crude oil prices dataset had a different timestamp format than the other two datasets. We had to use spark tools to get them in the same format. Another problem we had was that the Pytrends API used to collect the interest over time data had a limitation that it could pull daily data for a timeframe of only four months. We had to set up a loop to iteratively pull the data for contiguous four-month periods to get our final dataset.

Limitations of Application:

One of the limitations of the application is that it can only predict the stock prices for the current day as we cannot get the interest over time data for a day that has not yet started. This means that this application cannot be used to forecast stock market prices for an upcoming quarter. This can be solved by using Time Series forecasting to first predict the interest over time data for

the future and then use that set for our model.

Although we have seen a reasonable accuracy in our regression models, we think this can be improved by using more orthogonal datasets. Predicting stock prices itself is a very difficult task and using just two datasets does not do justice to the complexity of the problem. We will be pursuing this aspect further in the future.

VIII. Conclusion

Through this work, we put forth the hypothesis that interest over time data can be used to reasonably solve the problem of predicting stock market prices. We also confirm that the words 'settlement' and 'debt' have the most impact on the stock market prices. This application will be helpful to stock market buyers, sellers and brokers to make data-driven decisions on when to sell or buy a stock. It will also help airline companies to understand which words are impacting their stock prices the most and can then leverage this information when deciding their business strategies. We have scope for further improvement in this work.

Acknowledgment

We would like to thank Prof. Suzanne McIntosh for the support she extended to the project. We would also like to thank the NYU HPC team for being generous with the resources provided.

References

[1] Girija V Attigeri , Manohara Pai M M, Radhika M Pai, Aparna Nayak, "Stock Market Prediction: A Big Data Approach", TENCON 2015 - 2015 IEEE Region 10 Conference.

[2] Todsanai Chumwatana, Ichayaporn Chuaychoo, "Using social media listening technique for monitoring people's mentions from social media: A case study of Thai airline industry", 2017 2nd International Conference on Knowledge Engineering and Applications (ICKEA)

[3] Yahya Eru Cakra, Bayu Distiawan Trisedya, "Stock Price Prediction using Linear Regression based on Sentiment Analysis"

[4] Md. Rafiul Hassan, Baikunth Nath, "Stock Market Forecasting Using Hidden Markov Model: A New Approach", 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS)

[5]K. Tziridis, Th. Kalampokas, G.A. Papakostas, "Airfare Prices Prediction Using Machine Learning Techniques", 2017 25th European Signal Processing Conference (EUSIPCO)

[6] Yuling Li1,2, Ping Zhu, Sujuan Qin, "Optimization of ticket purchasing strategy based on machine learning", 2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC)

[7] Yauheniya Shynkevichl, T.M. McGinnityl, Sonya Colemanl, Ammar Belatrechel, "Stock Price Prediction based on Stock-Specific and Sub-Industry-Specific News Articles"

[8] Tobias Preis, Helen Susannah Moat & H. Eugene Stanley, "Quantifying Trading Behavior in Financial Markets Using GoogleTrends", Scientific Reports volume 3, Article number: 1684 (2013)

[9] Helen Susannah Moat, Chester Curme, Adam Avakian, Dror Y. Kenett, H. Eugene Stanley & Tobias Preis, "Quantifying Wikipedia Usage Patterns Before Stock Market Moves" , Scientific Reports volume 3, Article number: 1801 (2013)

[10] Min-Hsua Fan, Mu-Yen Chen, En-Chih Liao, "A TAIEX Forecasting Model Based on Changes of Keyword Search Volume on Google Trends", 2014 IEEE International Symposium on Independent Computing (ISIC)

[11] Hyunyoung Choi and Hal Varian, "Predicting the Present with Google Trends", UC Berkeley

[12] Roberto Cervelló-Royo, Francisco Guijarro, Karolina Michniuk, "Stock market trading rule based on pattern recognition and technical analysis: Forecasting the DJIA with intraday data", Science Direct

[13] Johan Bollen, Huina Mao, Xiao-Jun Zeng, "Twitter mood predicts the stock market", Journal of Computational Science, 2(1), March 2011

[14] Chungyong Li, Zhongying Qi, Tan Li, Tan Jie, Xiaona Wang, " Notice of Retractiob Dynamic relationship between oil price and China stock market", 2011 International Conference on Business Management and Electronic Information. (May 2011)

[15] K. Ryota, N. Tomaharu, "Stock market prediction based on interrelated time series data", IEEE Symposium on Computers and Informatics (March 2012)