RMDS 2021 - How News Sentiment Affects the Stock Market

Erik R. Endress, Marisa N. Nakanishi, Jaineel Patel, Ryan Fredric P. Peji, Matthew Reynaga

I. Introduction

Predicting the potential stock market has been an interest to researchers for generations. Understanding different trends and how external drivers lead to shifts in the market is valuable when developing business strategies. Creating a model to visualize the commodities exchange affects how businesses pursue their day-to-day operations. Analyzing all independent sources' aptitudes, such as the news or consumer behavior, is initially time-consuming and redundant. By creating a model, we can visualize the data and forecast the market's potential directions.

II. Background

When going about business operations, finding the profitability of the specific market one is reaching is crucial when making valuable decisions. Being aware of the different opportunities makes creating a model to predict potential trends significant. An emphasis on the digital market can lead to tremendous success in pursuing multiple departments' strategies, such as marketing [1]. Focusing on the crude oil industry may be of great interest to investors with a 73.5% increase since late 2020 [3]. Large corporations purify the oil to distribute for consumer use for fueling their vehicles [2].

III. Methodology

A. Initial Thoughts

To solicit imminent points in the stock market, employing a linear regression model is most appropriate. Using news sentiment, Apple's mobility index, and the anticipated crude oil prices, we gauged understanding of the oil industry's stock prices.

We determined that pursuing a simple linear regression would lead to a common problem within machine learning: data overfitting. To counter this imminent issue, we aimed to regularize the data using a regression model.

B. Cleaning and Mining

In the data cleaning phase, we determined that aligning data values using their dates as a key would be the most feasible option.

In the data mining phase, we took the average value within the Global Stock Exchange Price, Mobility Index, and the Crude Oil Price categories to condense the data. To prevent the loss of accuracy from the source data, we accounted for all the points within a specific date.

We implemented one extensive adoption to condense the data within the Apple Mobility Index data set. We took the data primarily from the New York Stock Exchange leading to our decision to restrict the data to only states and regions in North America. Furthermore, we focused our focus on data revolving around driving or transportation because they are the industry's most significant consumers.

The inclusion of crude oil's projected price as a trait also followed a similar train of thought; if the price of crude oil were to rise, this would cause gasoline/oil companies would expend more income to buy them. We assumed that this would cause the stock price to rise proportionally to the crude oil price.

The final model would attempt to align the target variable, the average stock prices, and the different characteristics using their respective dates to align the data and uniformly create it.

IV. Error

A. Data Cleansing

Our first challenge consisted of data cleansing and creating a consistent way to access and organize the entries. For example, we found instances of inconsistencies when analyzing the date's formatting, creating a challenge to align values. Furthermore, the Apple Mobility Index and the Anticipated Crude Oil prices exclusively included the

most recent year (2020). It made for complex calibration when creating our model to align any pertinent data with dates from older years. By formatting in Excel, we could mold the data to fit our needs and fill in any values deemed to be missing but necessary.

B. Extracting

We discovered a critical issue while extracting data, including inconsistencies within the Global Stock Exchange (GSE) dataset prices. We observed ten occurrences per value in most of the GSE database dates. Once identified, we created a formula that would calculate the average based on the key (date). After learning that the number of occurrences per date was inconsistent within a few variables, we decided to remove them to have more pertinent data.

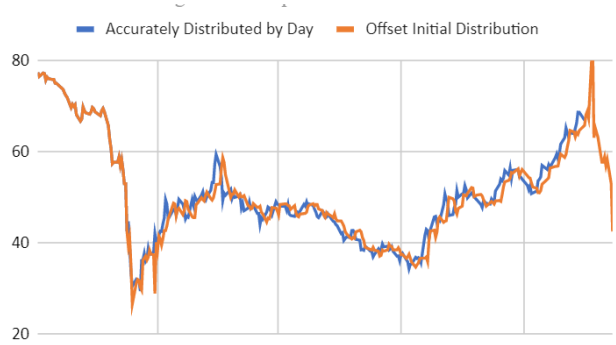

Fig. 1. Offset from inconsistent data we found during the data cleansing phase.

C. *Model Testing*

After creating and testing the Ridge regression model using Python's sci_kit_learn library, it became apparent there were glaring issues with the model. One main concern was the abysmally low test score, hovering around 10%. We also found it difficult to produce any future prediction data using the model as it also became apparent that it was incomplete. Reflecting upon our timeline within the competition, we decided to take our experience as a learning experience and submit our attempted work.

VI. Conclusion

Although we were unsuccessful in developing a model that would consistently predict the stock market, the experience through the challenge has provided a real-world application for our team to explore. Allowing the opportunity to explore industries' multitudes has broadened our perspective regarding data applications.

References

[1] Baccardax, Martin. “Global Oil Prices Extend Gains: Jim Cramer Says Crude 'In Driver's Seat' For Stock Market Direction.” TheStreet, *TheStreet*, 1 Mar. 2021, www.thestreet.com/investing/oil-prices-gain-jim-cramer-says-crude-in-drivers-seat.

[2] *Frequently Asked Questions (FAQs) - U.S. Energy Information Administration (EIA)*, www.eia.gov/tools/faqs/faq.php?id=727&t=6#:~:text=The top five source countries, Arabia, Russia, and Colombia.&text=Note: Ranking in the table, imports by country of origin.

[3] Wilcox, Kade. “What Is the Effect of Economic Change on Business?” *Primitive*, www.leadwithprimitive.com/blog/the-effect-of-economic-growth-on-business#:~:text=The effect of economic growth, for further growth and expansion.