FINANCIAL MODELING - THE METEORIC RISE OF TESLA (TSLA)

Prepared for: MPCS - Applied Data Analysis
Prepared by: Dhruv Srikanth

APPLIED DATA ANALYSIS PROJECT PROPOSAL:

OBJECTIVE - WHAT DO I WANT TO DO?

The objective of this project is to understand Tesla's meteoric rise in stock price between 2010 and 2020. Additionally, I would like to investigate Elon Musk's influence over the price, specifically, his activity on Twitter, which may have led to the rise of TSLA over the span of 10 years. Finally, I aim to create a model to predict Tesla's stock price.

MOTIVATION - WHY DO I WANT TO DO THIS?

Tesla's stock has experienced a meteoric rise between the years 2010 and 2020. As most companies recovered slowly from the 2008 financial crisis, Tesla, along with other technology giants experienced tremendous growth and did not seem as adversely affected by the crisis as other companies did. Furthermore, Tesla, despite the dramatic effect the financial crisis had on the automobile industry, was able to rise above the rest to experience tremendous growth. I would like to investigate whether Elon Musk's presence on Twitter had a role to play in this. Specifically, does the sentiment of Elon Musk's often controversial tweets play on role in Tesla's stock price. The reasoning behind this is, companies of this size are often made up of thousands of employees, whose performance would logically dictate the trajectory of the company, thereby, influencing the stock price. It seems impossible that a single individual could hold such an influence over the company's stock. However, the value of the cryptocurrency DogeCoin appears to be highly influenced by Elon Musk's activity on Twitter. Therefore, I would like to investigate whether the same could be said for the Tesla stock price.

Finally, I would like to perform statistical analysis on the TSLA data to gain an understanding of what may influence the stock's price. Using this knowledge, I would like to build a prediction model to

estimate the stock's price. I want to do this to gauge the ability of machine learning algorithms and techniques to predict a stock's price. This could be useful in determining optimal trading strategies, understanding what factors, previously not included in these strategies could now be incorporated (for example, a single person's tweets), and even in building automated trading bots.

DATA - WHAT WILL I USE TO DO THIS?

I will be using a Kaggle dataset that collates the stock price of Tesla (TSLA) between 2010 and 2020 and can be found here. Additionally, I will be amalgamating Elon Musk's tweets into the dataset. The required twitter data can be found here.

SOFTWARE - WHAT WILL I USE TO FACILITATE THIS?

1. Language - I will be using python as the language to perform all of my analysis.
2. Sentiment Analysis - I will be using vaderSentiment to perform sentiment analysis in order to ascertain the nature of each tweet.
3. Data Cleaning - I will be using numpy, scipy and pandas to clean the data.
4. Statistical Analysis - I will be using sklearn and statsmodels to perform the statistical analysis on my data.
5. Data Visualization - I will be using matplotlib, seaborn and plotly to visualize the data I have gathered.
6. Model Creation - In addition to modules mentioned above, I will be using TensorFlow (including Keras) to build the model.
7. Running Environment - I will be using jupyter notebooks for the purpose of running my experiments, analysis and investigation.
8. Model Training - I will be performing the model training on Google Colab to take advantage of their GPUs.

I will add and use more libraries to the list as I go through the project.

RELATED WORK

I began researching the different kinds of machine learning models employed for the purpose of stock prediction and time series analysis. This led me to the following two papers -

1. Stock Prediction using Sentiment analysis and Long Short Term Memory
2. Harvesting social media sentiment analysis to enhance stock market prediction using deep learning

The first paper obtains information on stock data for Apple, Google and Microsoft through Yahoo finance and Twitter data by scraping tweets of the Twitter website, using Tweepy (Twitter API). The stock data is normalized and then split in a train-test 70-30 manner. The Twitter data is cleaned by removing all links and special characters. Finally, they are divided based on the polarity of the tweet (positive/negative). An LSTM architecture is employed for the prediction model due to its use of feedback connections. Root Mean Squared Error (RMSE) is used for evaluating the performance of the model. The authors of the paper achieved an RMSE of 1.2583.

The second paper approaches the problem of stock market prediction augmented with sentiment analysis on a more broad spectrum. News data was collected from Moneycontrol, IIFL, Economic Times, and Twitter. Stock data information was gathered from the National Stock Exchange of India (NSE). A similar approach to the previous paper was employed for cleaning the news data. For performing sentiment analysis, Naive Bayes, Maximum Entropy, Linear Support Vector Classifier, and Decision Tree classifier were used. The news data was then assigned polarities of -1 for negative sentiment, 0 for neutral sentiment and 1 for positive sentiment. For the stock data, a trading condition was determined based on the close price of the stock for the trading day which was added as a predictor to the dataset along with the sentiment information. An LSTM network was trained and employed to perform predictions on the data. The sentiment classifiers achieved the following results for accuracy - Naive Bayes - 86.72%, Maximum Entropy - 88.93%, Linear SVC - 89.46, LSTM - 92.45%. For the sentiment classifier, the performance matrix detailing precision, recall and F1-score were utilized in determining the performance. Unfortunately, I was unable to determine the evaluation metric used for the LSTM model from the information provided in the paper.

Based on this, I learned that recurrent neural networks (RNN) are often used due to their ability to accurately model sequential data by understanding the temporal context (retaining memory). Upon diving deeper into the kinds of models and architectures used, I came across a type of RNN, the long-short term memory unit (LSTM), that helps to avoid the vanishing gradient problem and has also shown promise in the performance it is able to achieve when applied to time series problems like a stock market prediction. I believe similar approaches to the network architecture of my model will be key in accurately predicting Tesla's stock price.

APPROACH - HOW WILL I DO THIS?

I will perform the following steps over the course of my project -

1. Data Cleaning - Clean the data to be suitable for analysis.
2. Data Analysis - This would involve performing some statistical analysis (investigating the correlation between different attributes of our data), data visualization (looking at the distribution of our data) and deriving insights based on this.
3. Model Selection - Trying out different models and measuring the performance of each one to make an informed decision on which model to choose.
4. Model Tuning - Tuning the model I have chosen in the previous step.
5. Model Training - Training the model on the cleaned data.
6. Model Testing - Testing the model's performance.

I plan on discussing the nature of the problem of stock market prediction and the potential to augment models with data through sentiment analysis, during the mid-quarter presentation. During the final presentation, I plan on displaying the methodology and architecture of the model and its performance. Additionally, I will discuss any challenges I encountered and insights I obtain from the data analysis.

EVALUATION

I will be using RMSE to evaluate my models performance quantitatively. Furthermore, I will visualize the predicted data against the ground truth in order to qualitatively determine how close my predictions are to the trend.

REFERENCES

Given below is a list of references that I have used in previous sections -

1. Tesla Stock Dataset - https://www.kaggle.com/timoboz/tesla-stock-data-from-2010-to-2020

2. Elon Musk Tweets Dataset - https://www.kaggle.com/ayhmrba/elon-musk-tweets-2010-2021

3. Stock Prediction using Sentiment analysis and Long Short Term Memory - https://ejmcm.com/pdf_3126_16b444b632c88db6fed0c6558dd6930a.html

4. Harvesting social media sentiment analysis to enhance stock market prediction using deep learning - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8053016/