Analyzing & Predicting Tesla stock market behavior from Elon Musk Tweets

Team X: Final Project for Data Mining
ISM6316.901F21.91252

Team Members:
Veera Mukesh Aripaka
Tejaswini Tippanaboina
Sravani Chowdary Dondapati
Siva Sankari Ravipati

Problem Statement:

This project aims to analyze and predict the Tesla stock behavior based on Elon Musk's tweets.

Introduction:

Stock markets are very important in an open economy. Volatility is the major challenge when it comes to stock markets. Most of the time, to analyze the performance of an economy, we often look at the stock markets of that economy. Also, predicting the future trends in the stock market can decrease investment risks, discover the future value of companies, and financial assets and increase strategic investments, etc. In our project, we analyzed the correlation between Elon Musk's tweets and Tesla stock prices. We propose to build a model that predicts the future trends in Tesla stocks based on Musk's tweets.

Motivation:

Twitter is one of the social media applications that people actively follow to get the latest news and interesting information. It allows us to see what is happening in the world within seconds. Many influential people and celebrities post regular content daily. Elon Musk is among them and is well known for his controversial tweets. So, we attempted to analyze the relationship between his tweets and his company's stock price.

Goal:

From this project, we would like to check for the presence of any dependency between Tesla's stock performance and Elon Musk's tweets. If such a relationship is established to be true, we would analyze the impact of that particular tweet on the following day's stock.

Data source:

For this project, we required the historical stock quotes of Tesla for the year 2021 which we gathered from NASDAQ's official site, and Elon musk's tweets for the same time which were taken from the Kaggle website

https://www.kaggle.com/datasets/ayhmrba/elon-musk-tweets-2010-2021
https://www.nasdaq.com/market-activity/stocks/tsla/historical

Prior work:

The paper which inspired us to build our model, performs sentimental analysis on financial news to predict the future stock market trends with little error ratio.

In our project, we have considered the twitter data which has become the widely used source of information exchange instead of just the financial news for sentimental analysis and used this data to forecast the behaviour of a single stock

Reference paper:

https://www.researchgate.net/publication/318298991_Predicting_Stock_Market_Behavior_using_Data_Mining_Technique_and_News_Sentiment_Analysis

Method:

This model attempts to predict Tesla's stock behavior with respect to Elon Musk's tweets, to help investors make wiser investment decisions.
The suggested model's purpose is to anticipate not only the rise and fall of the stock price but also the depth of the movement and categorize based on Tesla handlers' assumptions. The suggested architecture combines the evaluation of Musk's tweets and stock indices to improve stock market behavior classification accuracy.
The two main steps involved in the method:

1) Sentimental analysis: After pre-processing the tweet data we use Naive Bayes algorithm to analyze tweet sentiment and categorize it as positive, negative, and neutral.
2) Numeric analysis: We classify the numerical outcome as rise, high rise, fall and steep fall based on the change in the stock price

These two different datasets are joined by the date of the following day and we apply algorithms like K-NN and Neural networks to predict the stock behavior.

Data Description:

As mentioned, we initially collected Elon musk's tweets from January 2021 to December 2021 and the respective stock prices over the last year. The collected tweets are categorized under text data. Tesla Stock price values of the same dates have been considered as numeric data. In numeric data, we have considered the opening, high, low and closing prices since these have a direct effect on future stock prediction.

Attributes	Description
Date	Date at which tweet made
Tweet	Tweets related to Tesla
Open	Opening stock value
Close	Closing stock value
Sentiment Analysis	Sentiment drawn on tweet
Numeric Analysis	positive/negative stock
Rise/Fall	Rise, High rise, Fall and Steep fall of stock price
Difference	Difference between closing values of consecutive days

Data Challenges:

- There was no single data set suitable for our needs, so we had to tailor a new one based on the historical stock prices and tweet data we already had, we used the attribute date to amalgamate the two datasets.
- The effect of a tweet can be perceived on the following day's stock price, to accommodate this, we introduced a lag into the dataset aligning the closing price to the next day's tweet using SQL.
- There were times when there were multiple tweets concerning the stock on the same day, but we had to consolidate their effect into that of a single tweet.

Model:

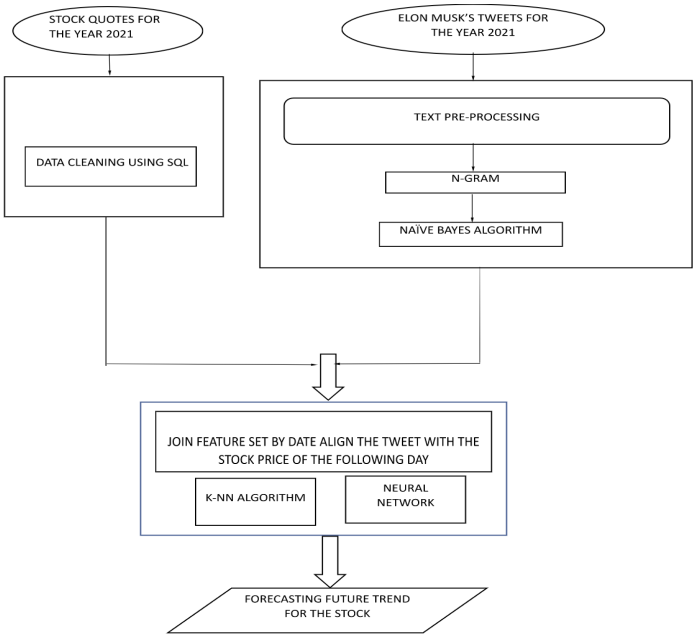


Fig: Proposed model

Model Components:

In our proposed model we have the following major components

1) *Sentiment Analysis Component:*

➢ Text-Processing:

•Tokenization: Each tweet is split into meaningful words called tokens.

•Data standardization: Transforming all words in the tweet about Tesla in a document into lowercase.

•Stop-word-removal: Words that do not have a significant meaning in the documents such as the, a, of...etc. are removed.

•Stemming: Removing suffixes such as (-Ed, -ing, -ion...etc.) to reduce the complexity.

•Filter tokens: Words that consist of two or fewer characters are filtered.

➢ N-Gram:

N-grams are continuous sequences of words or tokens in the tweets, also referred to as the neighboring sequences of items. N-gram is one of the important tools when we are dealing with text data in NLP(Natural Language Processing) tasks.

➢ Naïve Bayesian classifiers:

We have used the Naive Bayes classifier to categorize whether the tweet is positive or negative with reasonable accuracy of the model.

2)*Numeric Analysis Component:*

For numeric analysis, first, we converted all the numeric data into positive and negative by considering the closing values of stock prices. If the stock value of a particular day is higher than its previous day's stock value, then it is considered positive otherwise it is negative. Then we categorized the rise or fall based on the average fluctuations of the stock price. This is an assumption we made for categorizing. We also added the difference to the data set to check if we can predict the different ranges of the stock price.

SQL coding:

```
SELECT A.DAT,
CASE WHEN A.CLOSE - B.CLOSE = 0 THEN 'na'
WHEN A.CLOSE - B.CLOSE > 0.02*(B.CLOSE) THEN 'HIGH-RISE'
WHEN A.CLOSE - B.CLOSE > 0 AND  A.CLOSE - B.CLOSE <= 0.02*(B.CLOSE) THEN 'RISE'
WHEN A.CLOSE - B.CLOSE < 0 AND  A.CLOSE - B.CLOSE >= -0.02*(B.CLOSE) THEN 'FALL'
WHEN A.CLOSE - B.CLOSE < -0.02*(B.CLOSE) THEN 'STEEP FALL' ELSE 'EMO' END
FROM ELON A
INNER JOIN ELON B ON A.SEQ1 = B.SEQ2
ORDER BY 1 DESC
```

	A	B	C	D	E	F	G	H
1	DATE	TWEET	OPEN	CLOSE	SENTIMENTAL ANA	NUMERIC ANALYSIS	RISE/FALL	DIFFERENCE
5	2021-12-29	@CSmithson80 @he	366.213318	362.063324	neutral	negative	FALL	90.049988
10	2021-12-28	@BLKMDL3 @mims	369.829987	362.823334	negative	positive	FALL	90.809998
15	2021-12-27	@mims If history is a	357.890015	364.646667	neutral	neutral	HIGH-RISE	94.25
17	2021-12-23	@waitbutwhy üiØ	335.600006	355.666656	positive	neutral	HIGH-RISE	85.269989
18	2021-12-22	@T_Ball5 Probably n	321.886658	336.290009	neutral	neutral	HIGH-RISE	-19.376647
27	2021-12-21	@heydave7 üşî	305.623322	312.843323	neutral	positive	HIGH-RISE	31.176667
29	2021-12-20	https://t.co/OCUqr6	303.566681	299.980011	neutral	negative	STEEP FALL	18.313355
34	2021-12-17	https://t.co/mVhCpx	304.92334	310.856659	neutral	neutral	RISE	35.470001
35	2021-12-16	@PPathole I do	331.5	308.973328	neutral	neutral	STEEP FALL	-1.883331
37	2021-12-15	@GailAlfarATX @Saw	317.736664	325.329987	neutral	neutral	RISE	31.729981
43	2021-12-14	@SawyerMerritt Ash	315	319.503326	positive	neutral	FALL	25.90332
45	2021-12-13	@esprit_tesla @chaz	333.696655	322.136658	positive	neutral	STEEP FALL	27.773316
53	2021-12-10	@risermaker @engir	336.25	339.01001	neutral	neutral	RISE	60.533356
61	2021-12-09	@SirineAti @enginee	353.546661	334.600006	positive	negative	STEEP FALL	70.089996
66	2021-12-08	Lex asks great questi	350.90332	356.320007	positive	neutral	RISE	91.809997
67	2021-12-07	@cleantechnica Mos	348.066681	350.583344	positive	neutral	HIGH-RISE	-5.736663
71	2021-12-06	@SpaceXMR ü§£	333.83667	336.33667	positive	neutral	FALL	53.006683
76	2021-12-03	@engineers_feed Th	361.59668	338.323334	neutral	neutral	STEEP FALL	54.246674
80	2021-12-02	@karpathy All of real	366.353333	361.533325	neutral	neutral	FALL	73.726654
89	2021-12-01	@lexfridman Yeah, th	386.899994	365	positive	neutral	STEEP FALL	81.84668
96	2021-11-30	@BillyM2k Now that	381.456665	381.58667	neutral	neutral	RISE	113.313324
100	2021-11-29	@BillyM2k It is simul	366.996674	378.996674	neutral	positive	HIGH-RISE	108.443329
103	2021-11-26	@PPathole People ar	366.48999	360.640015	neutral	neutral	STEEP FALL	88.600006
105	2021-11-24	Physics formulas are	360.130005	372	positive	neutral	RISE	106.593323
108	2021-11-23	So much of AI is abou	389.170013	369.676666	neutral	positive	STEEP FALL	104.269989
111	2021-11-22	@stats_feed @engin	387.443329	385.623322	positive	neutral	RISE	125.190003
113	2021-11-19	@nichegamer ü§£	366.290009	379.019989	neutral	positive	HIGH-RISE	140.853317

Fig: Final dataset

Solution methodology:

The logic behind the algorithm is to predict the rise or fall of the stock price on a particular day, based on the tweets made by Elon Musk on the previous day. The category of the rise/fall can be based on the assumptions above mentioned and for the difference, we subtracted the previous day's stock price.

In our project, we are considering the following algorithms:

- KNN
- Neural networks
- Regression model

Here we have 3 output variables and 2 target variables, which are rise/fall and difference. We allotted 40 % of the data to train the model, 30% of the data for validation, and the other 30% of the data for testing. In the end, we compared models to know which model performs better on our dataset. The Model Comparison outputs from SAS EMiner are discussed below.

Output from tools:

Model in SAS miner:

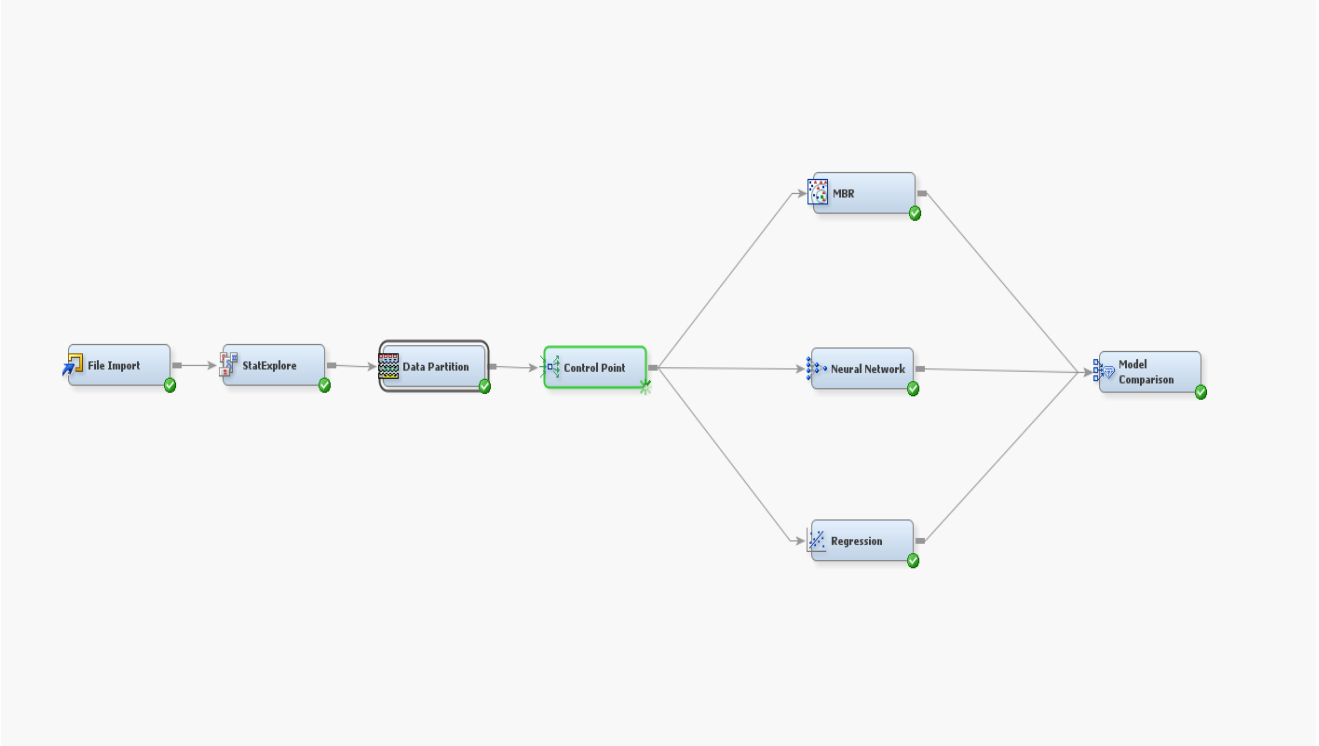
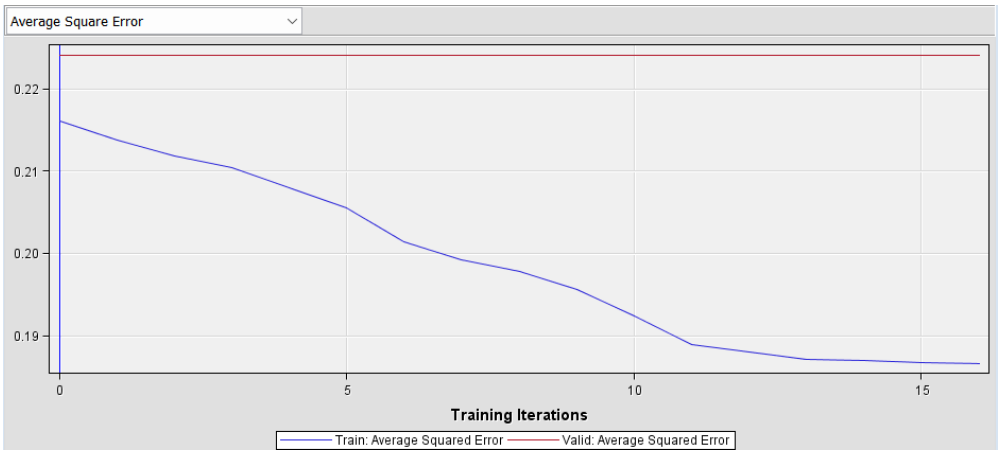


Fig : Model in SAS Miner

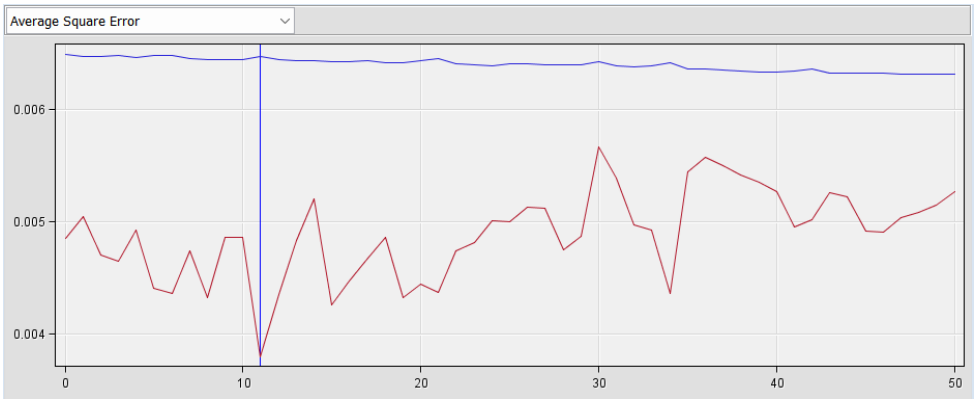
For the target variable rise/fall category here are the results:



Evaluation metrics:

Model	Misclassification rate	Sum of squared error	Average squared error	Final prediction error
MBR	**0.02631**	**156.722**	**0.00910**	**0.00082**
Neural Network	**0.91732**	**187.463**	**0.02287**	**0.02160**
Regression	**0.91732**	**174.328**	**0.02243**	**0.01866**

For the target variable difference here are the results:



Evaluation metrics:

Model	Misclassification rate	Sum of squared error	Average squared error	Final prediction error
MBR	**0.0012**	**159.11**	**0.003801**	**0.0004**
Neural Network	**0.21739**	**165.295**	**0.04867**	**0.00643**
Regression	**0.9000**	**155.30**	**0.07155**	**0.00618**

Conclusion:

From the above tables, MBR i.e., KNN model has the least average squared error and also has the least misclassification rate when compared to Neural network and Regression. So we can conclude that Knn is the best model for our project.

Future work:

Since the beginning, the stock market has always been a volatile entity, the inner workings of which, if decoded precisely, could be a lucrative source of income. Through this project, we attempted to tackle one of the factors that affect stock performance, the opinions of influential people with a direct interest in the stock. By deciphering the sentiment behind the opinion with this model, the stock behavior is forecasted to help a layman make better investment choices

- With more time, we would work on broadening our model to incorporate the social network activity of influential people, such as world leaders from various sectors and the effect of their opinions on a particular stock price that holds a direct interest of the individual and the stock market as a whole.
- we would also like to incorporate technical indicators into the model to get a more precise price prediction