S&P 500 INDEX FORECASTING

Team Project

[DATE]

[COMPANY NAME]
[Company address]

Table of Contents

1. Introduction---------- 2

2. Literature Review---------- 2

3. Data Description---------- 3

4. Data Pre-Processing---------- 5

5. Data Exploration---------- 7

6. Modeling---------- 10

7. Findings and Recommendations---------- 21

8. References---------- 22

Introduction:

The S&P 500 index is one of the most widely used measures of the health of the United States stock market. It tracks the performance of the 500 largest publicly traded companies in the country and is used by investors and analysts as an indicator of overall market trends. Forecasting the movement of this index is of great importance to many stakeholders, including traders, investors, and financial institutions.

This project aims to provide an accurate prediction of the future movement of the S&P 500 index by analyzing the trends in crude oil and gold prices, as well as the daily flow of global news. Given the close relationship between the S&P 500 index and various market factors, it is essential to understand how these factors may impact the market.

The project will analyze historical data on crude oil and gold prices to identify relevant trends and patterns that could impact the S&P 500 index. Additionally, the project will monitor global news sources to assess the potential impact of any significant global events on the market.

The primary focus of this project is to explore the role of these factors in predicting the movement of the S&P 500 index. By utilizing machine learning algorithms and statistical models, the project aims to provide a forecast with a higher level of accuracy than traditional forecasting methods.

The findings of this project will be of great relevance to several stakeholders, including investors, traders, regulators, and policymakers. The predictions can be used to make informed investment decisions, manage risk, and develop effective market strategies. Overall, this project aims to contribute to a better understanding of the S&P 500 index and its predictors, helping stakeholders to make better financial decisions.

Literature Review:

Hyndman and Khandakar (2008) used an exponential smoothing model to forecast the S&P 500 index. The authors found that the model performed well in forecasting the index, particularly during periods of high volatility. However, the model did not perform well during the financial crisis of 2008-09.

Topic modeling has been used to identify the news which causes changes in the S&P 500 index. For instance, Li et al. (2017) used a probabilistic topic model to identify the news that impacts the S&P 500 index. They found that geopolitical risk, economic growth, and corporate earnings were the most important topics that impacted the index.

In our study, we used top news headlines to predict if the S&P 500 index will go up or down based on VaderSentiment and TextBlob sentiment analysis. VaderSentiment is a lexicon and rule-based sentiment

analysis tool that assigns positive, negative, and neutral scores to text. TextBlob is a Python library used for performing sentiment analysis. We also used other major indicators like crude oil prices and gold prices for the forecast.

Data Description:

Data Source:

Data	Source
S&P 500 Index	Historical data downloaded from WSJ
Gold Price	Historical data downloaded from Trading Economics
Crude Price	Historical data downloaded from Macro Trends
News Headlines	Historical data downloaded from Reddit.com

The dataset contains information on the S&P 500 Index forecast based on three predictors - crude oil prices, gold prices, and daily global news inflow. The data is collected from various sources, including the Wall Street Journal, Trading Economics, MacroTrends, and Reddit. The data covers the period from 8th August 2008 to 1st July 2016.

Time Series Data:

Variable Name	Data Type	Description
Date	Date	Dates from 08/0808 to 07/01/16
Gold Price	Numerical	The closing price of Gold on that day
Crude Price	Numerical	The closing price of Crude on that day
S&P Index Price	Numerical	Closing stock price of the S&P 500 on that day

Top News Data:

Variable Name	Description
Date	Date 08/0808 to 07/01/16
Top1-Top25	Top 25 global news headlines

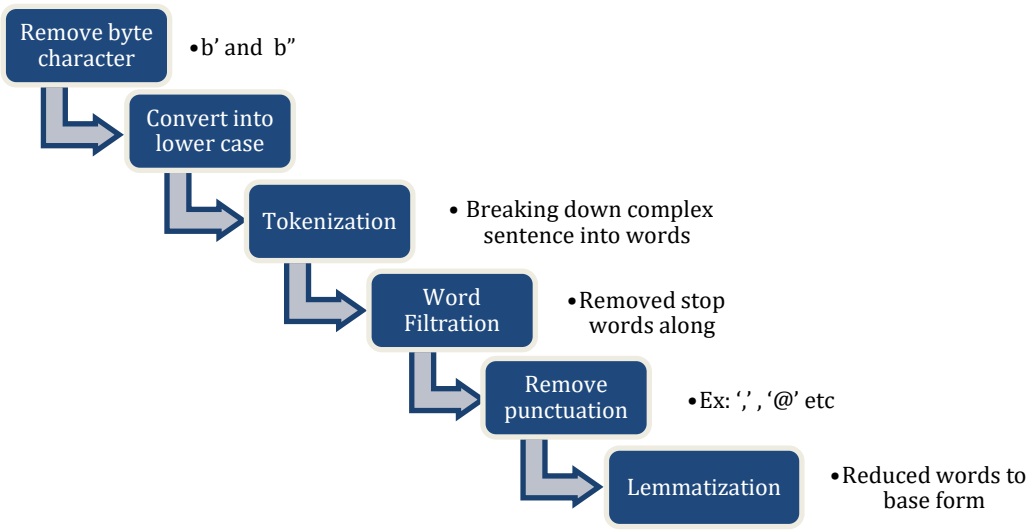
	Close S&P	Close Gold	Close Crude
count	1989.000000	1973.000000	1989.000000
mean	1492.818909	1292.232083	77.447929
std	405.984114	258.107480	22.280555
min	676.530000	704.900024	26.050000
25%	1164.970000	1133.099976	58.630000
50%	1394.230000	1257.900024	83.330000
75%	1913.850000	1482.300049	95.750000
max	2130.820000	1888.699951	115.640000

Figure-1: Basic Stats of all the numerical variables

The dataset helps us acquire insights and make predictions on the S&P 500 stock price through its variables. We can observe correlations between crude oil and gold prices, assess the global news' influence on markets, and forecast possible trends of the S&P 500 index. Our research further allows us to gain a deeper understanding of how these factors interact with each other.

Data Pre-Processing:

Text Mining:



1. Converting byte string to regular string: We converted the byte string text to a regular string using the decode() method, which allowed us to further process the text for NLP tasks. For example, we were able to tokenize the text and filter out stop words.
2. Converting to lowercase: We converted all words to lowercase using the lower() method. This step standardized the text and made it easier to compare and analyze.
3. Tokenization: We tokenized the text into individual words using the NLTK library's word_tokenize() function. This step created a list of tokens that we could further process.
4. Word filtration: We filtered out stop words such as "the", "a", and "an" using NLTK's stopwords module. This step helped to reduce the noise in the data and improve the accuracy of our NLP models.
5. Remove punctuation characters: We removed all punctuation characters from the text using Python's string module. This step helped to ensure that our models weren't influenced by punctuation marks, which are not meaningful in NLP tasks.
6. Lemmatization: We used the WordNetLemmatizer from NLTK to lemmatize the text. This step involved reducing words to their base form, which helped to reduce the complexity of the data and improve model accuracy. For example, the words "running", "ran", and "runs" would all be reduced to their base form, "run".

Data Transformation:

To make the S&P 500 index time series continuous, we inserted missing dates for the whole range. Since the dataset included only trading days, we had missing dates on weekends and a few holidays. To address this issue, we first created a continuous date range that covered the entire period of interest, including weekends and holidays. Next, we merged this date range with our original time series data, filling in any missing dates with NaN or null values. This step ensured that we had a complete set of data for the entire time range, which is necessary for many time series analyses.

By making the time series continuous, we were able to avoid issues with missing data and ensure that our time series data was suitable for analysis using various statistical and machine learning models. This step is a crucial part of the preprocessing pipeline for time series data and helps to ensure that we have high-quality, reliable data for further analysis.

To further preprocess the S&P 500 index data, we introduced a binary variable called 'Label', which was based on the close price of the index.

- 0 - if the index decreased from the previous day's close price
- 1 - if the index remained the same or increased concerning the previous day's close price

This new variable allowed us to categorize the changes in the index over time, which could be useful for various analyses. Additionally, we introduced a new variable called 'percent_change' which represented the percentage change in the index from the previous day. This variable provided us with a more granular view of the changes in the index over time and allowed us to detect trends and patterns that might have been missed using other metrics.

Together, the Label and percent_change variables provided us with a more comprehensive picture of the S&P 500 index data, allowing us to perform more nuanced analyses and make more informed decisions. These variables were a crucial part of our preprocessing pipeline and were used extensively in subsequent analyses.

Sentiment Analysis:

To gain insights into the sentiment of the text data derived from the top 25 news headlines, we performed sentiment analysis using two different libraries - VaderSentiment and TextBlob. From the VaderSentiment library, we extracted the polarity of the sentiment as positive, negative, or neutral. This polarity metric ranged from -1 to +1, where a score of -1 indicated a highly negative sentiment and a score of +1 indicated a highly positive sentiment. The neutral score was set at 0.

Next, using the TextBlob library, we extracted the polarity of the sentiment as a float value between -1 and +1, providing a more granular view of the sentiment polarity. Additionally, we also used TextBlob to calculate the subjectivity of the text data, which measures the degree to which the language used in the document is subjective or opinion-based. The subjectivity score ranged from 0 to 1.

- 0 - An objective document
- 1 - Highly subjective or opinion-based document.

By performing sentiment analysis on the text data derived from the top 25 news headlines, we were able to gain a deeper understanding of the underlying sentiment of the language used in the S&P 500 index dataset. This information could be useful for various analyses, such as understanding how market sentiment affects stock prices, and could help inform investment decisions. The sentiment analysis pipeline we used, incorporating both VaderSentiment and TextBlob libraries, was an essential step in our preprocessing pipeline for the text data.

Data Exploration:

The time series data of the S&P 500 from 2008 to 2016 reveals several ups and downs in the index. The most significant dip occurred in 2008 due to the Global Financial Crisis, with the index reaching a low of 666 in March 2009. However, the index gradually recovered over the years, buoyed by government policies and a stabilizing economy. The European debt crisis and the US government shutdown of 2013 also caused temporary dips in the index. The decline in oil prices from 2014 to 2016 had a significant impact on the index as well. Finally, the US presidential election in 2016 caused a brief drop in the index but rebounded with expectations of business-friendly policies from the incoming administration. Overall, the time series data of the S&P 500 from 2008 to 2016 demonstrates the inherent volatility of the stock market and the impact of global events and government policies on the index.

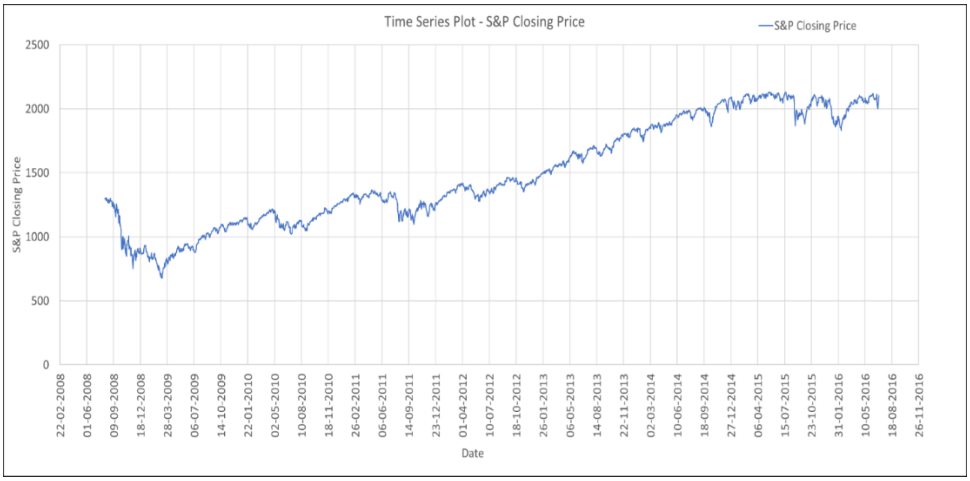


Figure 2: Time Series Plot for S&P 500 Index from 2008 to 2016

Gold and crude oil prices are key economic indicators that experienced significant volatility from 2008 to 2016, and their fluctuations can impact the S&P 500 index. Gold is often seen as a hedge against economic uncertainty and market volatility, and its price reflects the level of investor confidence in the economy. During the financial crisis of 2008, the price of gold reached an all-time high of around $1,000 per ounce as investors flocked to the safe-haven asset. Similarly, the price of crude oil peaked in mid-2008, reaching over $140 per barrel, before falling sharply due to weak demand and a global supply glut. By early 2016, crude oil prices had fallen to around $30 per barrel, a significant decline from its peak. These fluctuations in gold and crude oil prices can impact the S&P 500 index as they are key components of the global economy. High crude oil prices can increase the cost of production and transportation, leading to reduced corporate profits and negatively impacting the stock market. Similarly, a rise in gold prices may indicate inflationary pressures or investor uncertainty, causing investors to shift their investments away from equities and towards safe-haven assets like gold, which can cause the stock market to decline. Therefore, changes in gold and crude oil prices can have a significant impact on investor sentiment and the S&P 500 index.

Figure 3: Time Series Plot for Gold from 2008 to 2016

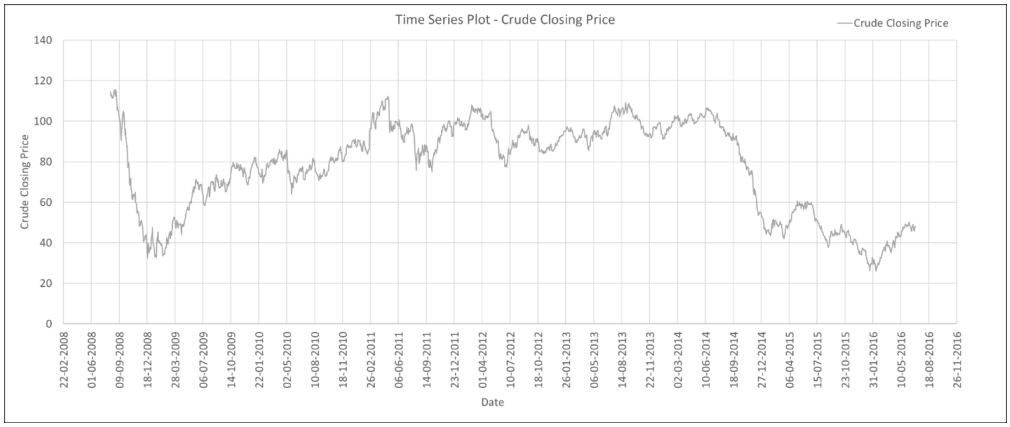


Figure 4: Time Series Plot for Crude from 2008 to 2016

Finding Correlations:

We analyzed the S&P 500 index and explored the correlations between various numerical variables and the target variable. Our analysis revealed several significant observations, with the most notable being the high negative correlation of -0.667 between the volume of S&P 500 stocks traded and the index. This negative correlation means that as the volume of stocks traded increases, the index tends to decline, and vice versa.

	Open	High	Low	Volume	Adj Close	Close S&P
Open	1.000000	0.999592	0.999436	-0.691621	0.998991	0.994099
High	0.999592	1.000000	0.999373	-0.686997	0.999546	0.994921
Low	0.999436	0.999373	1.000000	-0.699572	0.999595	0.994177
Volume	-0.691621	-0.686997	-0.699572	1.000000	-0.694281	-0.667542
Adj Close	0.998991	0.999546	0.999595	-0.694281	1.000000	0.994925
Close S&P	0.994099	0.994921	0.994177	-0.667542	0.994925	1.000000

Figure 5: Correlation Matrix

There are several possible reasons for this relationship. Firstly, higher trading volumes may indicate increased investor uncertainty or anxiety, which can cause investors to sell off their holdings and drive down stock prices. Similarly, low trading volumes may indicate investor confidence, causing prices to rise. Additionally, higher trading volumes may indicate greater market volatility, which can lead to increased price swings and larger declines in the index. Moreover, high trading volumes may also be indicative of large institutional investors, such as pension funds and hedge funds, who may have a greater impact on the market due to their larger trades. Finally, the relationship between trading volume and the S&P 500 index may also be influenced by technical factors, such as algorithmic trading and program trading, which can exacerbate price movements and volatility. Overall, the high negative correlation between trading volume and the S&P 500 index highlights the importance of market activity and investor sentiment in driving stock prices and underscores the need for careful monitoring and analysis of market trends.

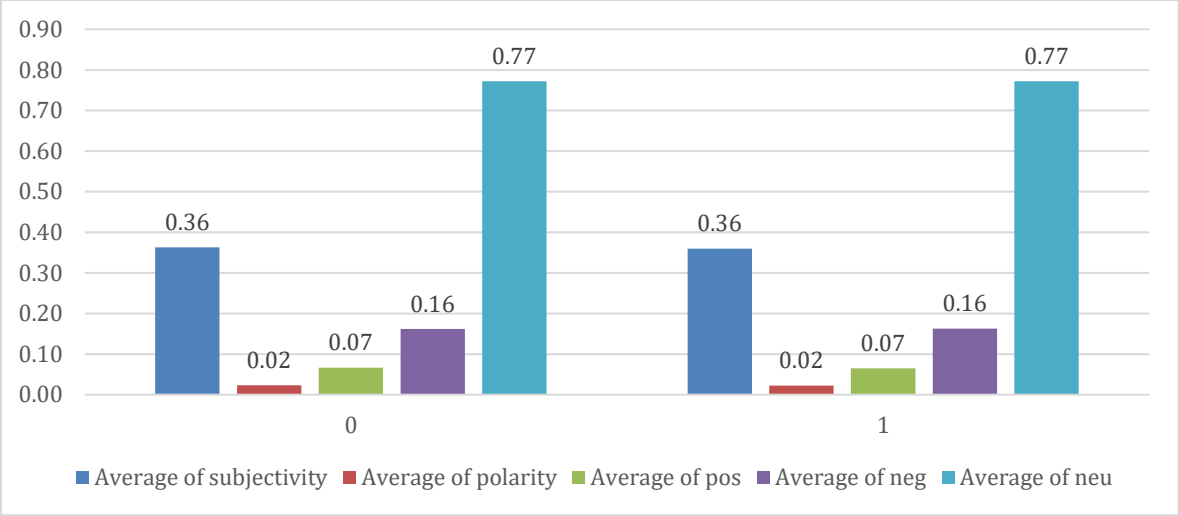


Figure 6: Relationship between target variables and results of Sentiment Analysis

We examined the distribution of the binary target variable, which represents the daily rise or fall of the S&P 500 index, concerning features obtained from sentiment analysis. However, our analysis from Figure 6 revealed that there were no significant differences in distribution, meaning that no insights could be drawn from this analysis. While sentiment analysis can be a useful tool for understanding investor sentiment and predicting stock prices, our findings suggest that sentiment alone may not be a strong predictor of the daily movements of the S&P 500 index. Other factors, such as market fundamentals, economic indicators, and geopolitical events, may play a more significant role in driving stock prices. Therefore, while sentiment analysis can be a valuable tool for understanding investor sentiment and market trends, it should be used in conjunction with other data sources and analytical techniques to gain a comprehensive understanding of market dynamics.

Modeling:

Classifier Models:

To predict the behavior of the S&P 500 index, we employed various classification models by utilizing the sentiment analysis results such as 'Subjectivity', 'Polarity', 'neg' & 'pos'. Specifically, we employed models like Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting to predict the movement of the S&P 500 index relative to the previous day's price, i.e., whether it would fall or rise.

We split the data into two parts - a training set and a test set - with the training set accounting for 80% of the total dataset. After splitting the data, we used the predictors obtained from sensitivity analysis and

developed models to predict the target variable ‘Change’ and understand if the sensitivity of the news helps in predicting whether the index will move up or down compared to the previous day.

After building and validating several models, we found that all the models had very low accuracy and the AUC of the models was just around 50%. Though the Gradient Boosting model was the top-performing model, the accuracy and AUC were not good. As the AUC and the Accuracy of the model were not satisfactory, we cannot rely on the classification models to predict the behavior of the S&P 500 index. It was really hard to predict the direction of the movement of the S&P index only based on the news inflow. We even tried to predict the amount of percentage change when compared to the previous day’s price and the results were not encouraging at all. The performance metrics of the Gradient Boosting model were as below:

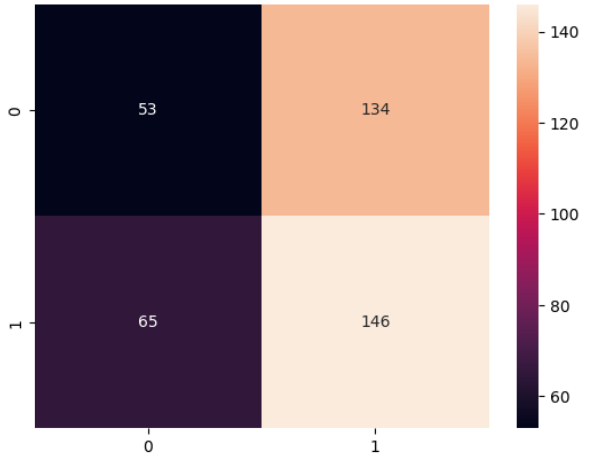


Figure 7: Confusion Matrix of Gradient Boosting Model

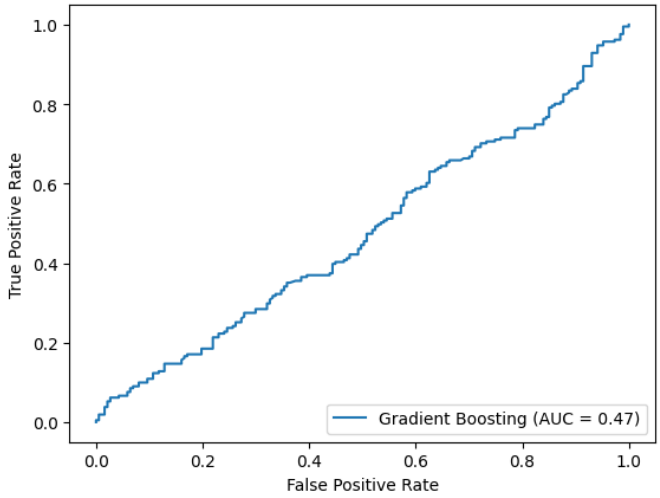


Figure 8: ROC curve of Gradient Boosting Model

The above classification models were built in Python. As the accuracy and ROC were low, we tried to build these models using the same predictors and target variables in SAS Enterprise Miner. However, the results didn't improve as the accuracy and the AUC of the models were just around 50% in SAS Enterprise Miner as well. The diagram & the results of the models developed in SAS Enterprise Miner are as below.

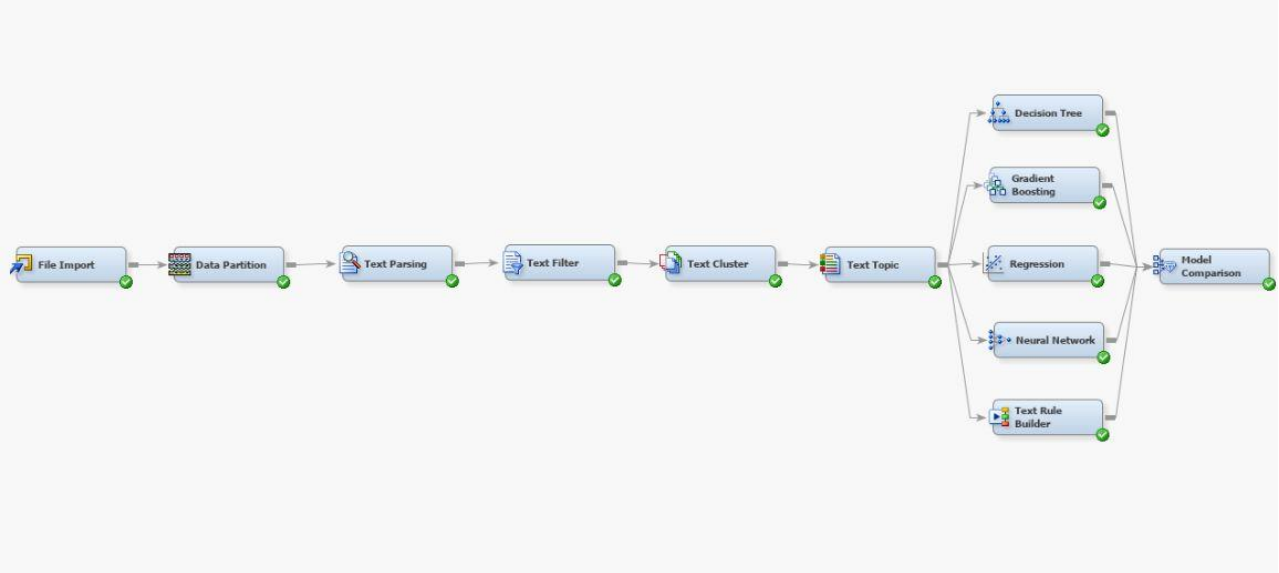


Figure 9: Diagram from SAS Enterprise Miner

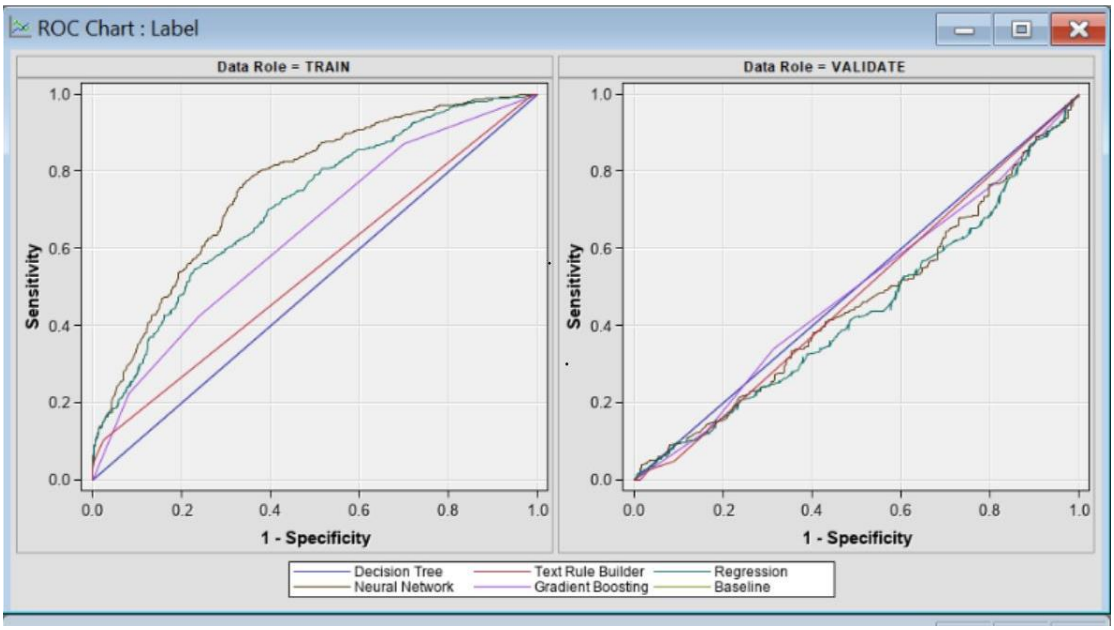


Figure 10: ROC curves from SAS Enterprise Miner

Time Series Analysis:

To predict the S&P 500 index, we tried to develop various time series models using the past data of the S&P 500 and we also used various other regressors such as Gold Price, Crude Price, results from sensitivity analysis, etc. We used the following procedure to develop the time series forecasting models in SAS.

1. Set a Holdout sample:

Historical data in our dataset was available from 8th Aug 2008 to 1st Jul 2016 with the period of each entry as 1 day. We have set aside a holdout sample of 200 data points to validate the time series model which we had developed. This holdout sample was used to obtain the model's accuracy and evaluate its performance.

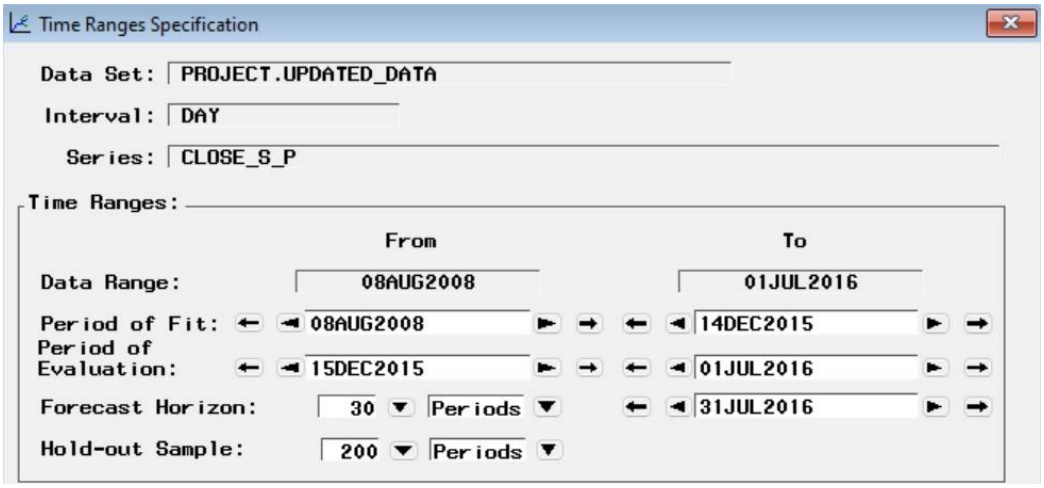


Figure 11: Hold-out sample allocation for the Time Series

2. Check the stationarity of the time series:

Stationarity is a crucial assumption in time series analysis because it ensures that the statistical properties of the data remain constant over time. As we can develop time series models only after ensuring that the time series is stationary, we checked whether the S&P 500 index time series data was stationary by examining the white noise tests, unit root tests, and seasonality tests.

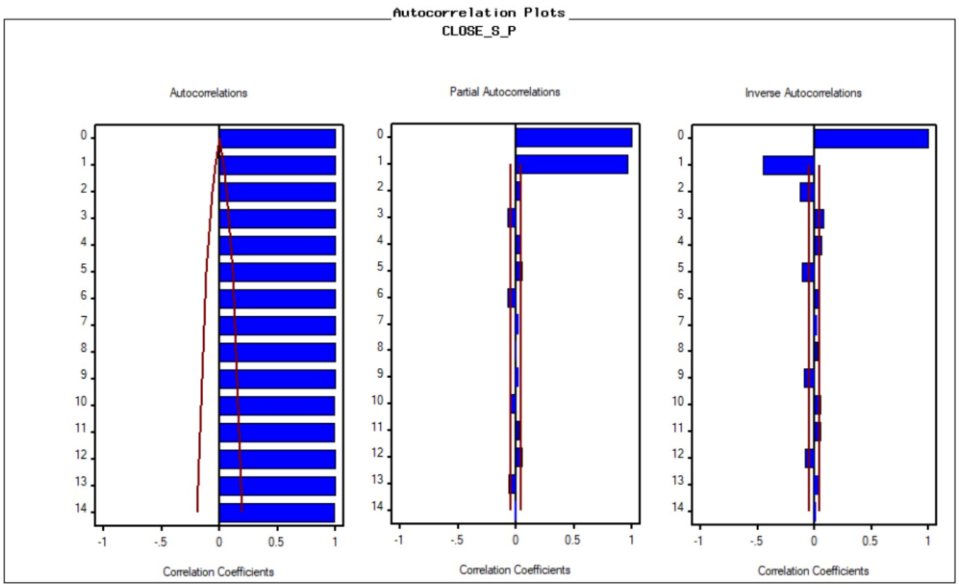


Figure 12: Auto-Correlation Plots for S&P 500 index

From the above plots, it was observed that the autocorrelation was decaying exponentially at a very slow rate across the lags, which indicated high autocorrelation in the dataset.

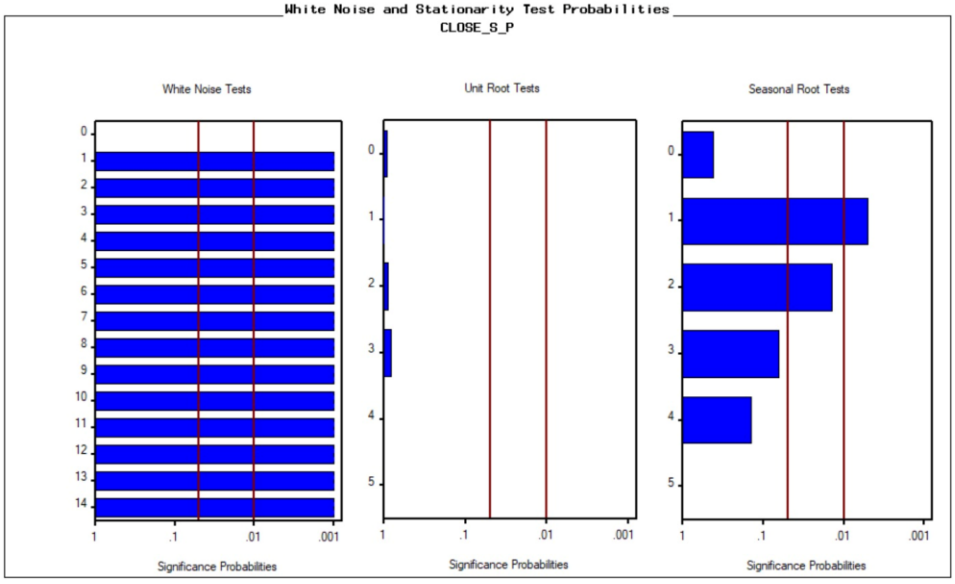


Figure 12: Stationary Tests for the S&P 500 index

Based on the above plots, we have concluded the following:

- From the White Noise Test, we observed that the p-values for all the instances were around 0.01, i.e., significant. Hence, we rejected the null hypothesis and concluded that the time series was not white noise.

- From the Unit Root Test, we observed that the p-values at all the instances were closer to 1, i.e., non-significant. Hence, we cannot reject the null hypothesis which indicated that the time series had a clear trend.
- From the Seasonal Root Test, we observed that p-values for a few instances were significant (p-value less than 0.05) and p-values for other instances were non-significant (p-value greater than 0.05). From this seasonal root test, we were not able to conclude whether this time series had seasonality or not.

From the above tests, we can conclude that the time series is not stationary, and we cannot directly build an ARIMA time series model with this time series data.

3. Made the time series stationary:

As the time series had a clear trend, we applied the first simple difference to the time series to make it stationary. The stationary tests after applying the first difference were as below.

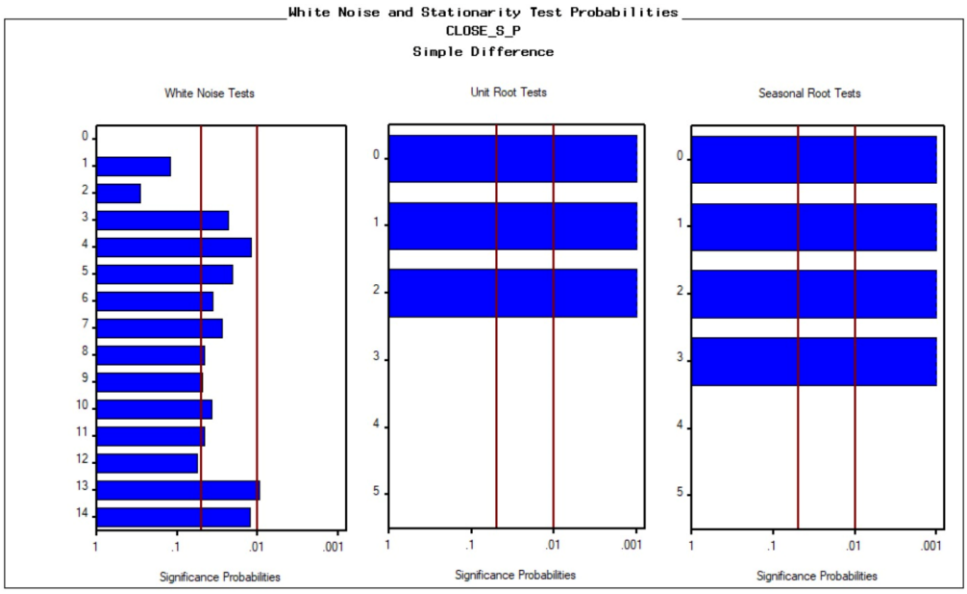


Figure 13: Stationary Tests for the S&P 500 index after the first difference

From the above stationarity tests after we carried out the first simple difference, the p-values were less than 0.05, i.e., they were significant in both the Unit Root Test and Seasonal Root Test due to which we rejected the null hypothesis in both the cases, and we concluded that there was no trend and seasonality after first simple difference.

4. Identified p & q values from the correlation plots of the first difference applied time series:

After we made the time series stationary, we examined the correlation plots of the first differenced time series to identify the p and q values for the ARIMA model. The p-value represents the number of lag observations included in the model, while the q-value represents the size of the moving average window. The autocorrelation plots of the time series after the first difference were as below.

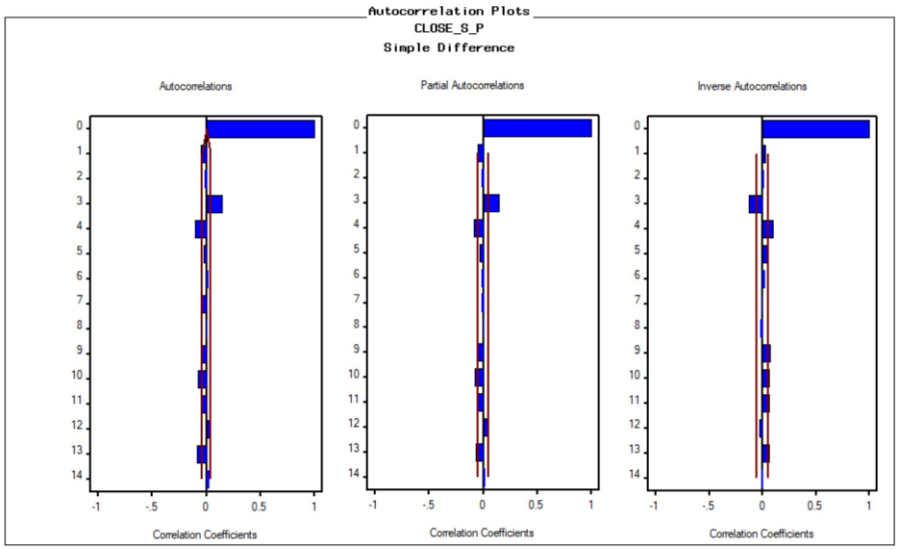


Figure 13: Autocorrelation plots of the time series after applying the first difference.

From the ACF plot, we observed that the highest lag with a significant spike was lag 4. Based on this information, we concluded that the values for p in our ARIMA models can range from 0 to 4. From the PACF and IACF plots, we observed that the highest lag with a significant spike was also lag 4. Based on this, we concluded that the values for q in our ARIMA models can range from 0 to 4.

5. Built multiple models:

Based on the p, q, and d values obtained, we built various ARIMA models using different combinations of p and q to identify the best model that fits the data. We also built other time series models such as exponential smoothing models, multiplicative and additive models, etc., and observed the performance of various time series models. We evaluated the models based on the performance parameters such as RMSE value, AIC, and BIC.

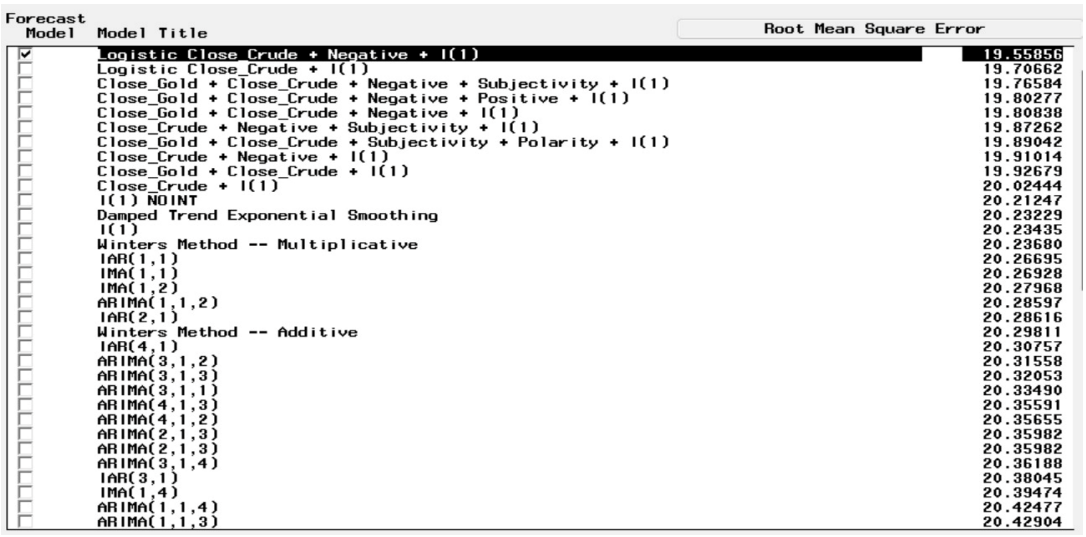
6. Induced other regressors:

To improve the accuracy of the model, we used other regressors like gold price, crude oil price, and results from sensitivity analysis of the news along with the best-performing model ARIMA models at that point. These additional regressors helped in capturing the effects of external factors that influenced the behavior

of the S&P 500 index. We even used a few transfer functions with these regressors and checked for possible improvement in the performance of the times series prediction.

Among all the models which we built, the best performing time series model was 'Logistic Crude + Negative + I(1)' with an RMSE value of 19.558 which was the lowest among all the other models that were fitted. This model also had very low AIC and BIC when compared with the other models. The parameters of the best time series model were as follows:

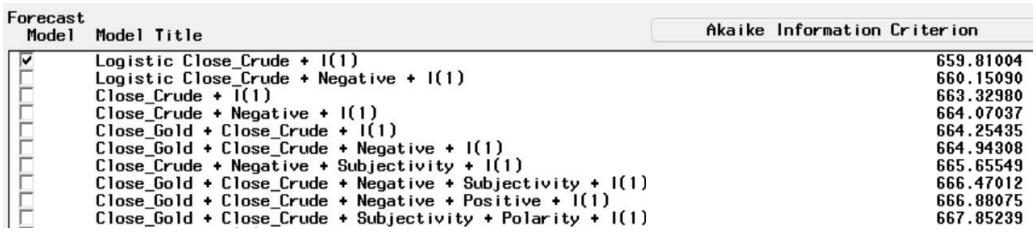
- Logistic Crude – Logistic transfer function applied to the crude price.
- Negative – Negative Sentiment obtained from the sensitivity analysis of the news.
- I(1) – ARIMA model with simple difference applied (ARIMA(0,1,0)).



Forecast Model	Model Title	Root Mean Square Error
☑	Logistic Close_Crude + Negative + I(1)	19.55856
☐	Logistic Close_Crude + I(1)	19.70662
☐	Close_Gold + Close_Crude + Negative + Subjectivity + I(1)	19.76584
☐	Close_Gold + Close_Crude + Negative + Positive + I(1)	19.80277
☐	Close_Gold + Close_Crude + Negative + I(1)	19.80838
☐	Close_Crude + Negative + Subjectivity + I(1)	19.87262
☐	Close_Gold + Close_Crude + Subjectivity + Polarity + I(1)	19.89042
☐	Close_Crude + Negative + I(1)	19.91014
☐	Close_Gold + Close_Crude + I(1)	19.92679
☐	Close_Crude + I(1)	20.02444
☐	I(1) NOINT	20.21247
☐	Damped Trend Exponential Smoothing	20.23229
☐	I(1)	20.23435
☐	Winters Method -- Multiplicative	20.23680
☐	IAR(1,1)	20.26695
☐	IMA(1,1)	20.26928
☐	IMA(1,2)	20.27968
☐	ARIMA(1,1,2)	20.28597
☐	IAR(2,1)	20.28616
☐	Winters Method -- Additive	20.29811
☐	IAR(4,1)	20.30757
☐	ARIMA(3,1,2)	20.31558
☐	ARIMA(3,1,3)	20.32053
☐	ARIMA(3,1,1)	20.33490
☐	ARIMA(4,1,3)	20.35591
☐	ARIMA(4,1,2)	20.35655
☐	ARIMA(2,1,3)	20.35982
☐	ARIMA(2,1,3)	20.35982
☐	ARIMA(3,1,4)	20.36188
☐	IAR(3,1)	20.38045
☐	IMA(1,4)	20.39474
☐	ARIMA(1,1,4)	20.42477
☐	ARIMA(1,1,3)	20.42904

Figure 14: Develop Models window

In all the performance metrics, the model 'Logistic Crude + Negative + I(1)' had really good performance when compared to all other models that were fitted. The top 10 best models based on various performance metrics like MAPE, AIC, and BIC are mentioned below.



Forecast Model	Model Title	Akaike Information Criterion
☑	Logistic Close_Crude + I(1)	659.81004
☐	Logistic Close_Crude + Negative + I(1)	660.15090
☐	Close_Crude + I(1)	663.32980
☐	Close_Crude + Negative + I(1)	664.07037
☐	Close_Gold + Close_Crude + I(1)	664.25435
☐	Close_Gold + Close_Crude + Negative + I(1)	664.94308
☐	Close_Crude + Negative + Subjectivity + I(1)	665.65549
☐	Close_Gold + Close_Crude + Negative + Subjectivity + I(1)	666.47012
☐	Close_Gold + Close_Crude + Negative + Positive + I(1)	666.88075
☐	Close_Gold + Close_Crude + Subjectivity + Polarity + I(1)	667.85239

Figure 15: Performance on AIC

Forecast Model	Model Title	Schwarz Bayesian Information Criterion
☑	Logistic Close_Crude + I(1)	665.21100
☐	Logistic Close_Crude + Negative + I(1)	668.25235
☐	Close_Crude + I(1)	668.73076
☐	Close_Crude + Negative + I(1)	672.17181
☐	Close_Gold + Close_Crude + I(1)	672.35579
☐	Close_Gold + Close_Crude + Negative + I(1)	675.74501
☐	Close_Crude + Negative + Subjectivity + I(1)	676.45741
☐	Close_Gold + Close_Crude + Negative + Subjectivity + I(1)	679.97252
☐	Close_Gold + Close_Crude + Negative + Positive + I(1)	680.38315
☐	Close_Gold + I(1)	681.16232

Figure 16: Performance on BIC

Forecast Model	Model Title	Mean Absolute Percent Error
☐	Logistic Close_Crude + Negative + I(1)	0.75585
☐	Logistic Close_Crude + I(1)	0.76403
☐	Close_Gold + Close_Crude + Negative + Subjectivity + I(1)	0.76112
☐	Close_Gold + Close_Crude + Negative + Positive + I(1)	0.76160
☐	Close_Gold + Close_Crude + Negative + I(1)	0.76178
☐	Close_Crude + Negative + Subjectivity + I(1)	0.76316
☐	Close_Gold + Close_Crude + Subjectivity + Polarity + I(1)	0.76650
☐	Close_Crude + Negative + I(1)	0.76387
☐	Close_Gold + Close_Crude + I(1)	0.76616
☐	Close_Crude + I(1)	0.76815

Figure 17: Performance on MAPE

After finalizing our best-performing model, we analyzed the residual plot for the model. From the prediction error plot, we observed that the prediction errors were very high (both negative and positive) between the end of 2008 and the beginning of 2009. The model was not accurate during this period because of the high volatility in the S&P 500 index due to the Global financial crisis. We also observed that whenever there was a big spike on the either upside or downside, there was a significant big spike on the opposite side as well which indicates high volatility in those periods. Apart from these observations, we found that prediction errors were mostly white noise and there was no significant pattern observed.

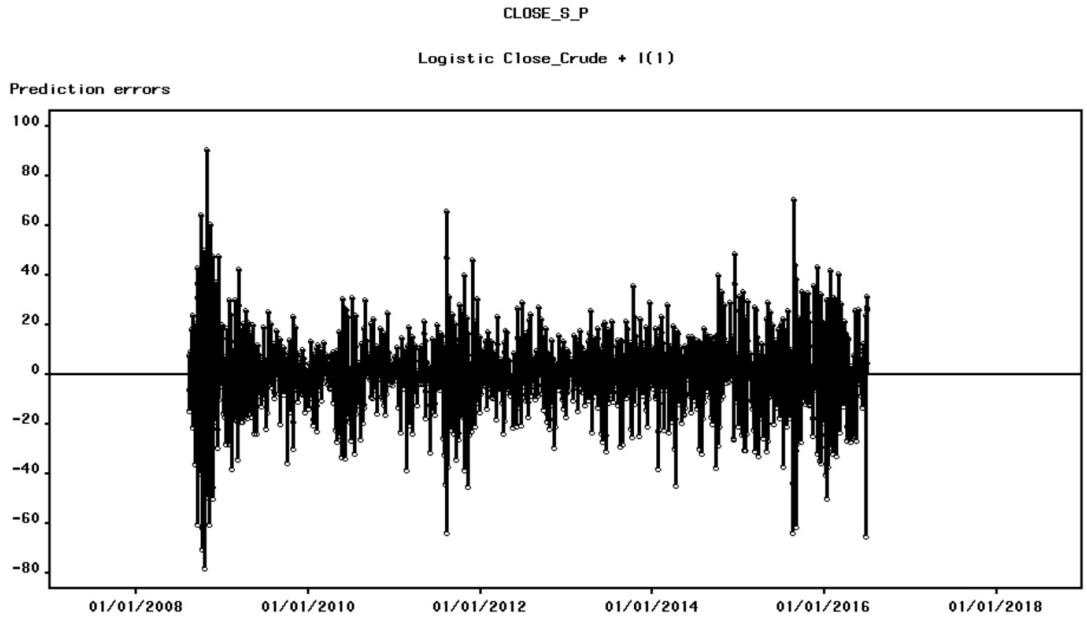


Figure 18: Residuals of the best-performing model

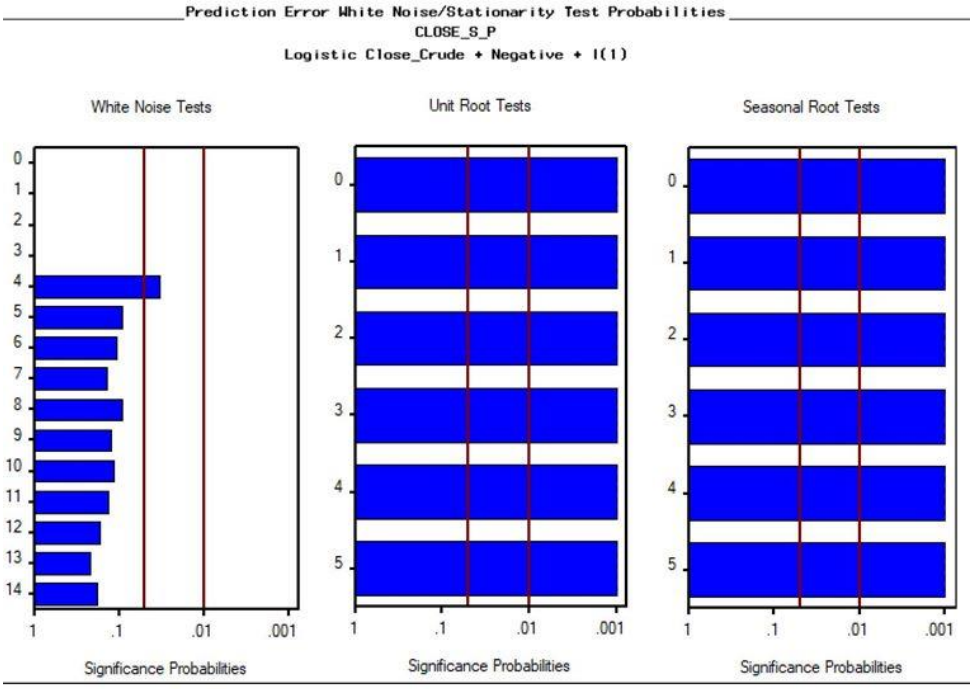


Figure 19: Stationary Tests for Residuals

We also verified the white noise and stationarity test plots for the prediction errors, and we concluded that the prediction errors do not have any trend and seasonality as the p-values were significant. Also, the White Noise Test indicated that the p-value was insignificant, and the prediction errors were white noise.

We refitted our best model on the whole historical data available by removing the holdout sample, and we compared the parameter estimates of this model without the holdout sample with the earlier model with the holdout sample. We observed that parameter estimates were almost the same in both cases, which further increased our confidence in the best model. The parameter estimates with and without hold-out samples were as below.

Parameter Estimates

CLOSE_S_P

Logistic Close_Crude + Negative + I(1)

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	0.0005504	0.000382	1.4398	0.1528
Close_Crude	0.00291	0.000269	10.8335	<.0001
Negative	0.01228	0.0066	1.8695	0.0643
Model Variance (sigma squared)	0.0002114	.	.	.

Figure 20: Parameter Estimates with Hold-out Sample.

Parameter Estimates

CLOSE_S_P

Logistic Close_Crude + Negative + I(1)

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	0.0005227	0.000365	1.4335	0.1519
Close_Crude	0.00298	0.000260	11.4410	<.0001
Negative	0.01369	0.0062	2.1998	0.0280
Model Variance (sigma squared)	0.0002071	.	.	.

Figure 21: Parameter Estimates without Hold-out Sample.

Time Series Forecasting:

We used our best-performing model to forecast the S&P 500 index for 200 periods and the predictions are as below.

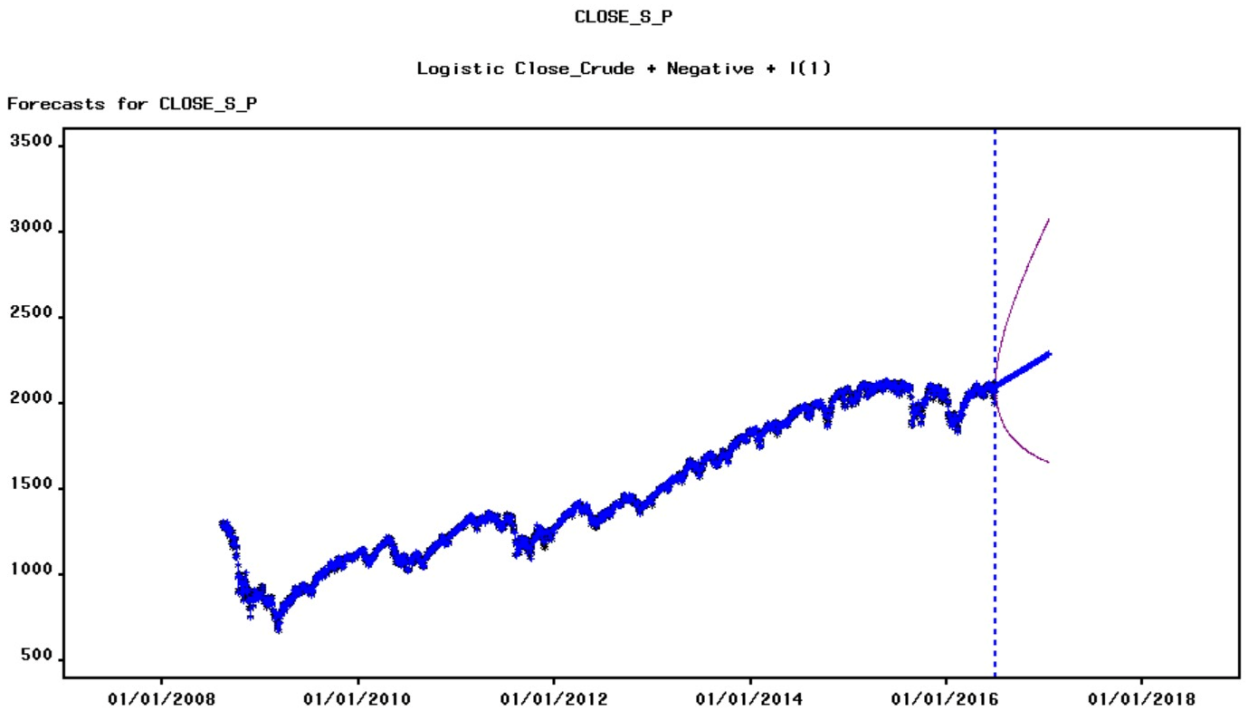


Figure 22: Time Series Forecasting for 200 periods.

Findings and Recommendations:

After analyzing various data mining and time series models and exploring the data, we have arrived at the following findings along with recommendations:

1) The S&P 500 index is highly dependent on crude price: Our best time series model shows that the price of the S&P 500 index can be predicted using the crude price. This can be because crude prices have a significant impact on the global economy and the performance of the companies that make up the S&P 500 index. The index includes many energy companies, which are highly sensitive to fluctuations in crude prices. We recommend that investors who are interested in investing in the S&P 500 index should closely monitor the price of crude. Changes in crude oil prices can have a significant impact on the performance of the index, and investors should consider these factors while making investment decisions.

2) Negative news inflow has a significant impact on S&P 500 index: Our best time series model also indicates that negative news affects the S&P 500 index. Negative news can cause investors to become anxious and sell off their stocks, leading to a decrease in the price of the S&P 500 index. We recommend that investors should keep themselves up to date with the latest economic news and events and be aware of the impact of negative news on the S&P 500 index.

3) The volume of trade has a strong negative impact on S&P 500 index: Our analysis shows that there is a negative correlation of -0.67 between the S&P 500 index and trading volume. This finding suggests that high trading volumes can lead to a decrease in the price of the S&P 500 index. This may be because high trading volumes often occur during periods of market instability or uncertainty, which can lead to a decrease in investor confidence and a decrease in stock prices. We recommend that traders exercise caution and be risk-averse during high trading volumes to minimize risk.

4) Gold price doesn't seem to affect S&P 500 index: Our analysis suggests that the price of gold does not have a significant impact on the performance of the S&P 500 index. Based on this finding, we recommend that investors should invest in gold to diversify their portfolio, as any drastic fall in the S&P 500 index will be balanced out by the investment in gold.

5) The strong upward trend in S&P 500 index, but seasonality doesn't seem to exist: The S&P 500 index has been on an upward trend over the long term, but there is no significant seasonal pattern to the index. This trend may be due to good economic growth over the years, and the non-existence of seasonality may be because the index is made up of a diverse group of companies that are affected by a variety of factors, rather than a single industry that may be more susceptible to seasonal trends. We recommend that investors should be aware of the long-term upward trend of the S&P 500 index and invest in the index when the price reaches closer to the long-term trend line. We also recommend that investors should not rely on seasonal patterns to invest in the S&P 500 index.

References:

1. S&P 500 Index: SPX | S&P 500 Index Historical Prices - WSJ

2. Gold Prices: Gold - 2023 Data - 1968-2022 Historical - 2024 Forecast - Price - Quote - Chart (tradingeconomics.com)

3. Crude Oil prices: Crude Oil Prices - 70 Year Historical Chart | MacroTrends

4. News Headlines: Stock Market News (reddit.com)

5. Kaggle

6. Stack Over Flow: https://stackoverflow.com/questions/tagged/time-series

7. Chat Gpt

8. Data Mining & Business Intelligence Lectures

Responsibilities of Each Team Member:

Team Member	Responsibilities
Abhinaya Kumar Singampalli	Data collection, Modeling & White paper Editing
Srinath Reddy Pamula	Time series forecasting modeling & White paper draft
Sumanth Kumar Goud Golla	Sentiment Analysis from News & White Paper Editing
Sai Dheeraj Reddy Gopannagiri	Data preprocessing, Exploration & Classification Modeling(SAS)
Naveen Abboju	Classification modeling(Python) & White paper draft
Venkata Vinayaka Durga Prasanth Lakamsani	Data Exploration & Time series Modeling