STOCK INDEX PREDICTION

Sentiment & Technical Analysis to predict stock movement based on daily news & technical data

Thank you, Springboard Mentor - Dipanjan Sarkar, Data Science Lead - Google, Author

HOW?

Data Source
Historical Technical Data
Daily News Headlines

Data Wrangling
Exploratory Data Analysis
Feature Engineering

Modeling
Natural Language Processing(NLP)
Deep Learning – Neural Network
Time Series Analysis

DATA SOURCE

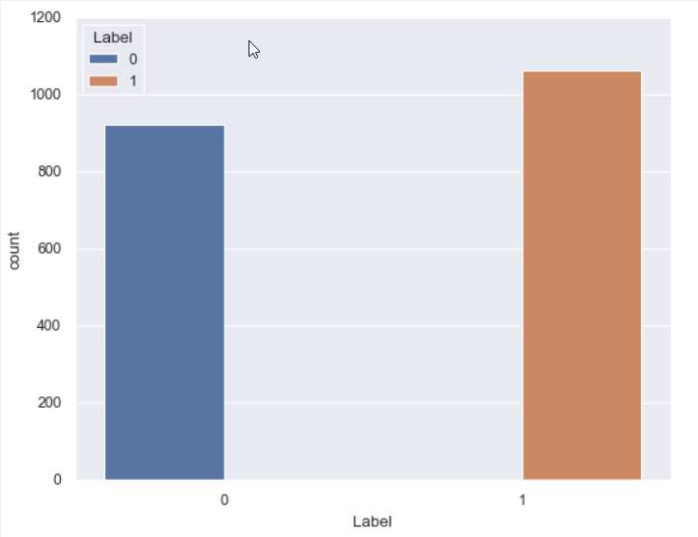
- Kaggle - https://www.kaggle.com/aaron7sun/stocknews
- Technical data and top25 daily news headline
- DJIA_table.csv:
  Downloaded from Yahoo Finance
- Stock data: Dow Jones Industrial Average (DJIA) - Range: 2008-08-08 to 2016-07-01

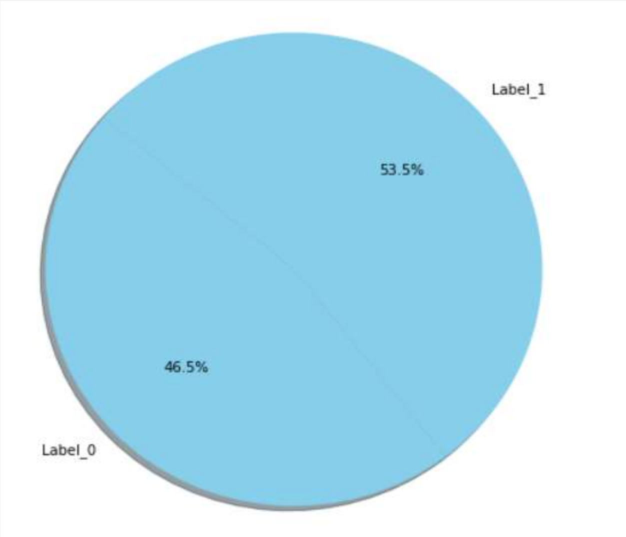
DATA WRANGLING

- Tokenized Stemming and Lemmatizing sentence
- Converted text into lowercase as required
- Removed all the punctuations and commas
- Replace emojis by using a pre-defined dictionary containing emojis along with their meaning
- Replacing characters except Digits and Alphabets with space
- Ignored Stopwords (e.g.: “the”, “he”, “have”)
- Lemmatization is the process of converting a word to its base form. (e.g., “Great” to “Good”)
- Removed NaN, Null values
- Combined relevant features and ignored irrelevant

EXPLORATORY DATA ANALYSIS

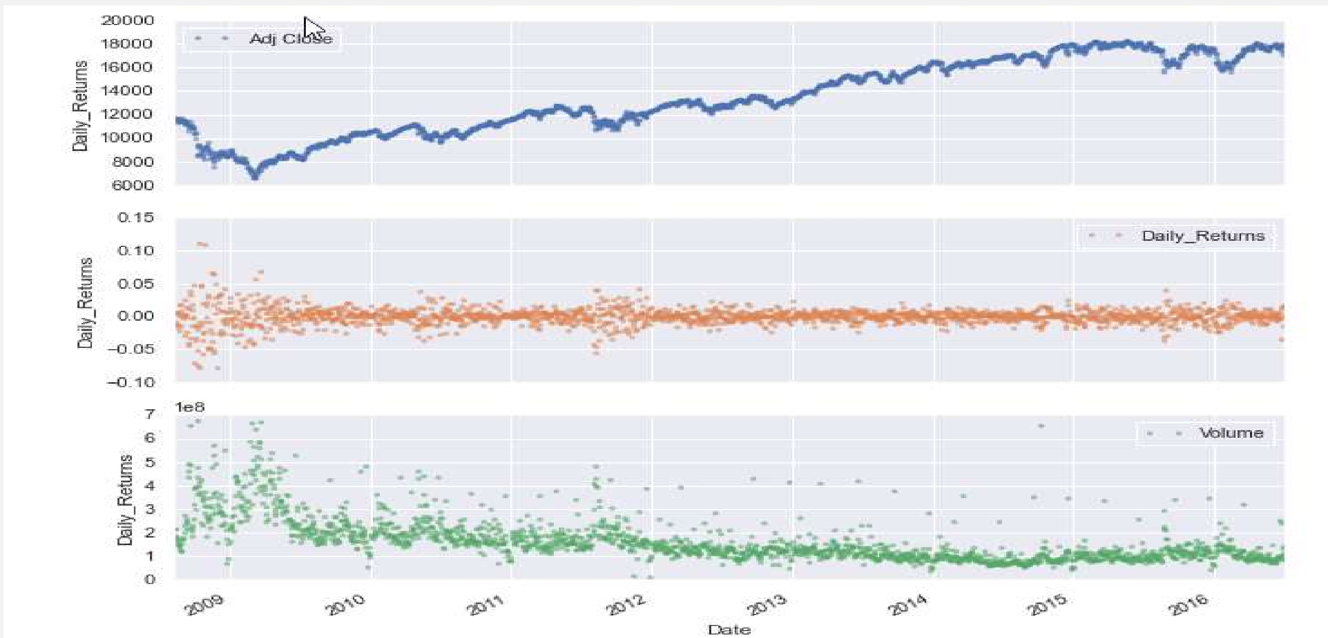
Balanced Data. News with Positive Sentiment- Label "1" is when DJIA Adj Close value has been risen or stayed as the same. News with Negative Sentiment- Label "0" is when DJIA Adj Close value decreased





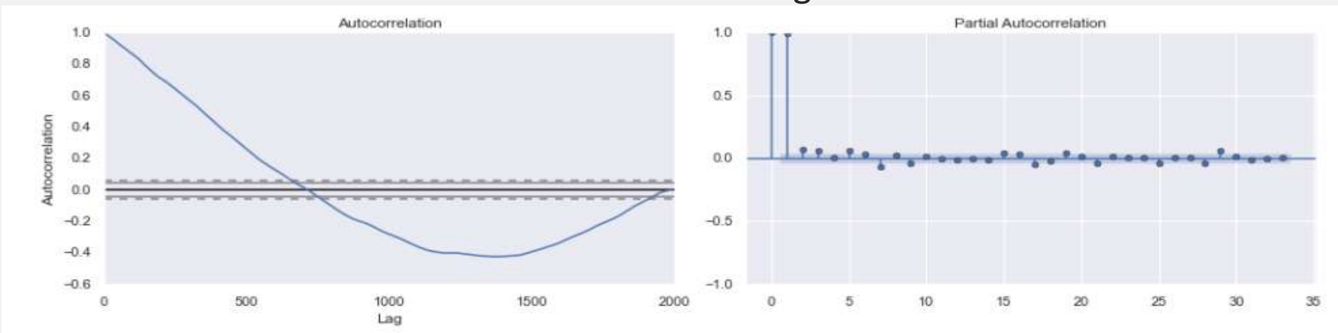
EXPLORATORY DATA ANALYSIS

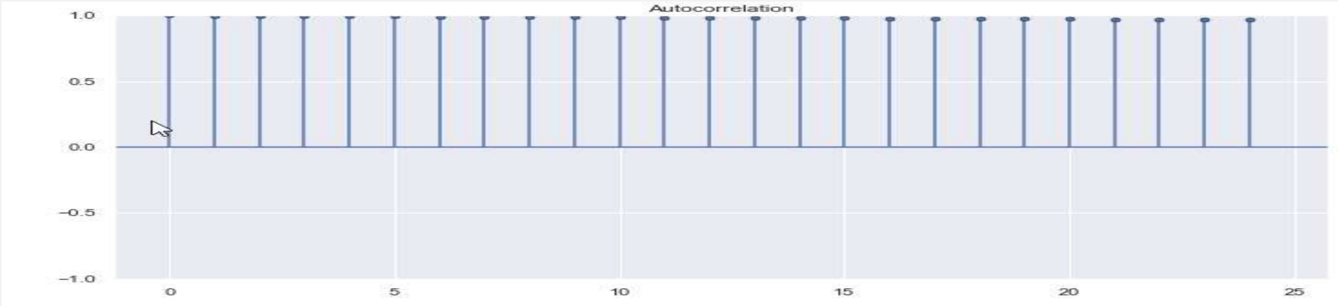
Daily return with respect to the volume, Adjusted closing price, and the daily return



EXPLORATORY DATA ANALYSIS

Correlation with the prices and their lagged values. The ACF shows values outside the confidence bands around 0 indicating variable autocorrelation

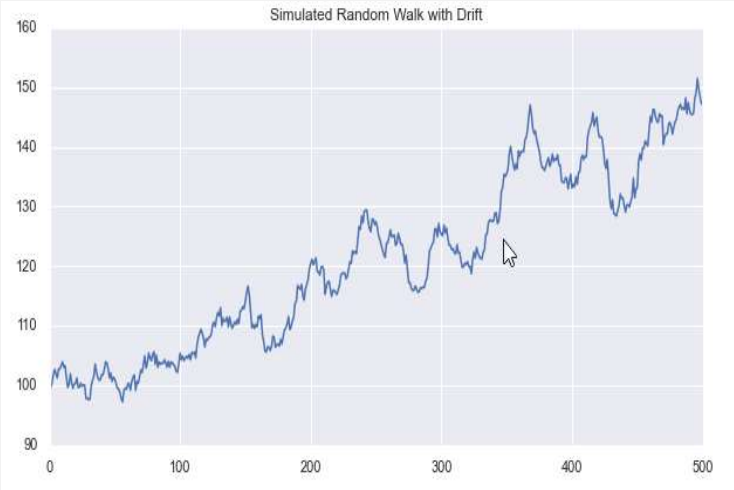




EXPLORATORY DATA ANALYSIS

Smoothing and Random Walk
Adj Close price with the moving averages this time Calculating Moving Average with lag of 100, 200 and 365 days

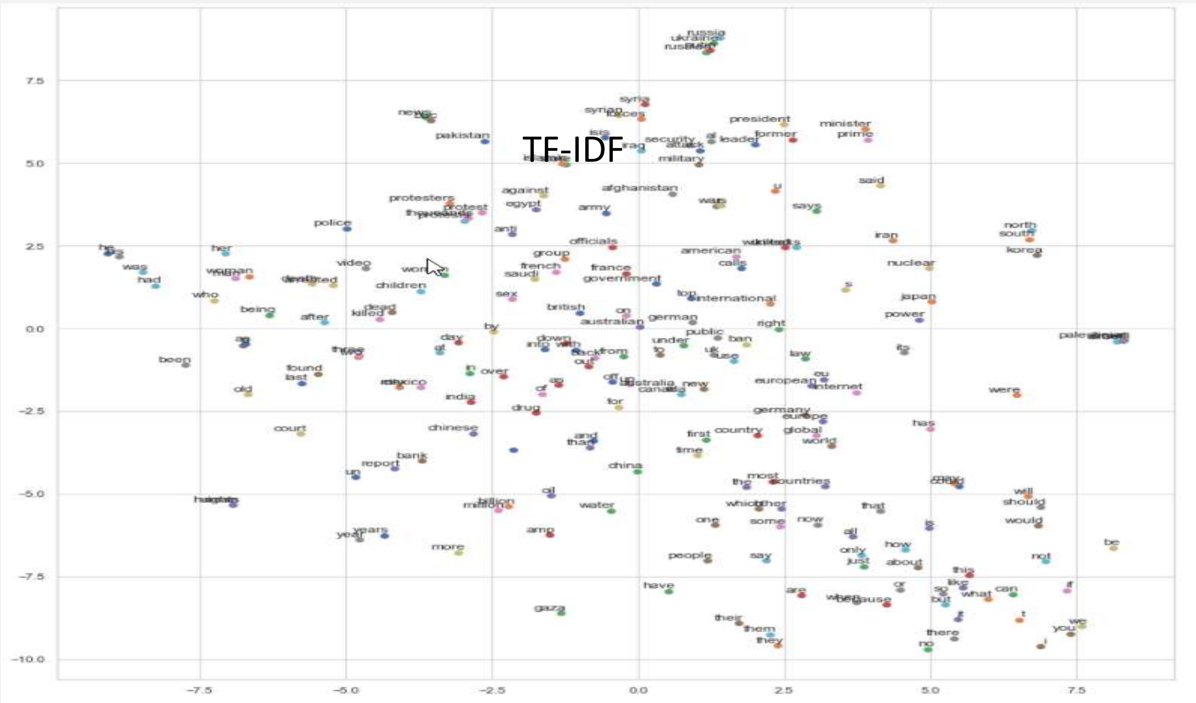
Random walk with a drift
•stock prices, are random walks but tend to drift up over time
•When adding noise, we may theoretically get negative prices



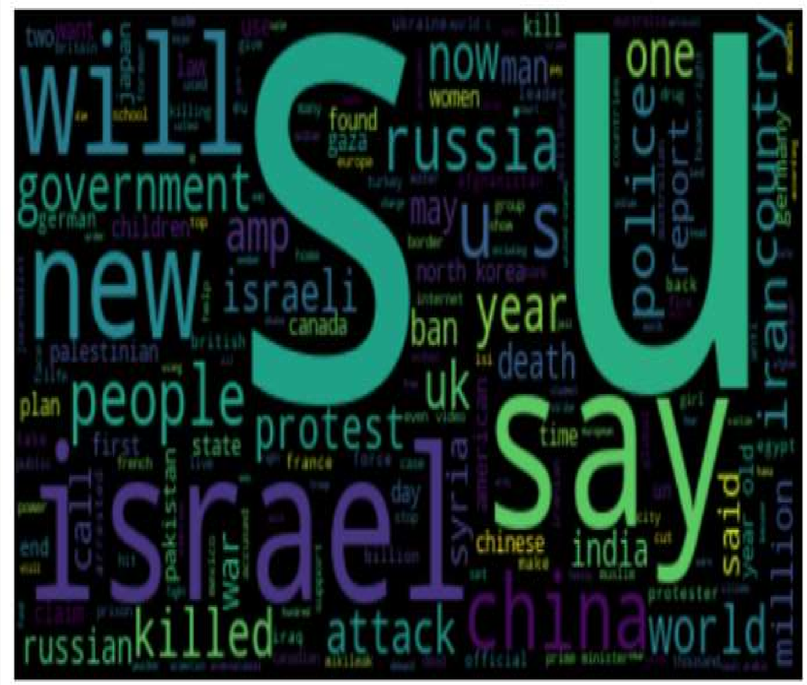
MODELING-NATURAL LANGUAGE PROCESSING(NLP)

Word Embedding and Word2Vec: is capable of capturing context of a word in a document, semantic and syntactic similarity, relation with other words, etc. Word2Vec is a method to construct such an embedding. It can be obtained using two methods (both involving Neural Networks): Skip Gram and Common Bag Of Words (CBOW)

Word2Vec Keyword Visualization



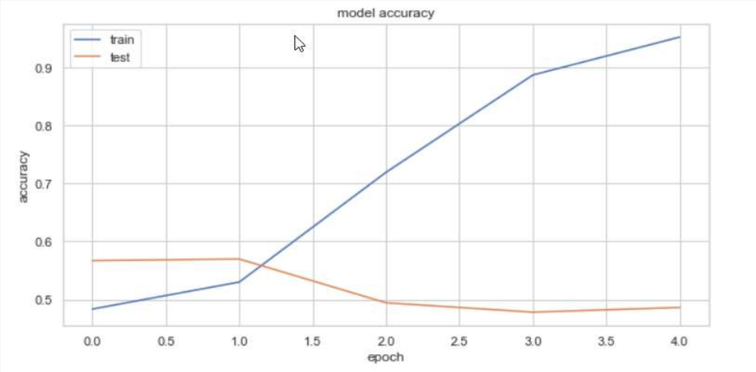
Tokenized Keyword Visualization



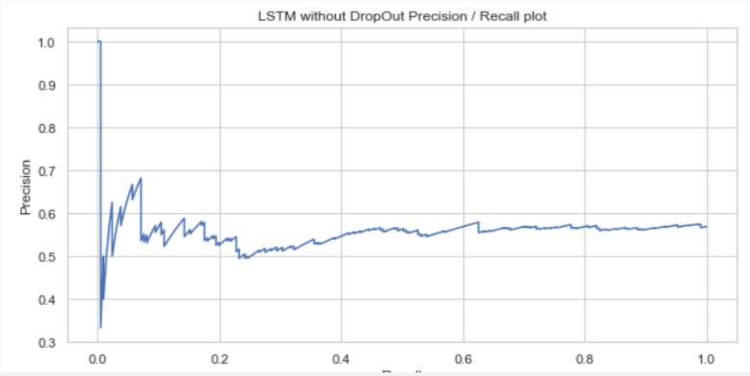
MODELING-NATURAL LANGUAGE PROCESSING(NLP)

Dropout can be applied between layers using the Dropout Kera's layer. We can do this easily by adding new Dropout layers between the Embedding and LSTM layers and the LSTM and Dense output layers.

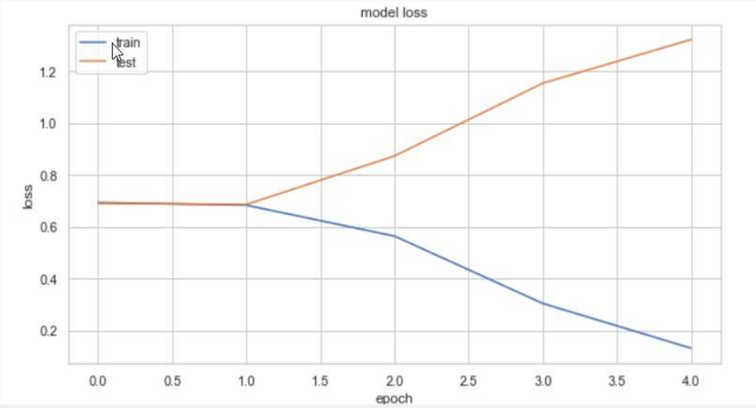
Accuracy LSTM without Dropout



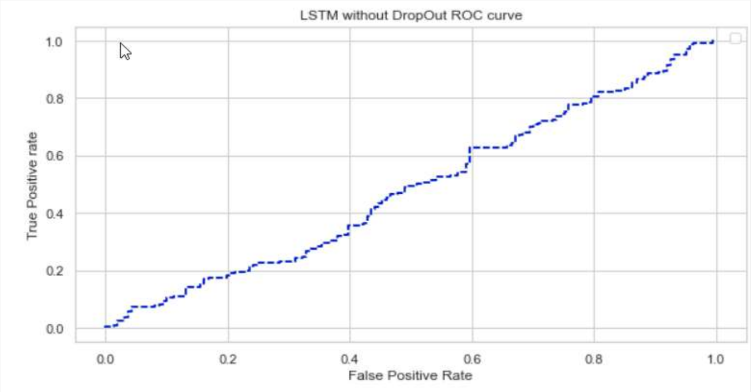
Precision/Recall LSTM without Dropout



Loss LSTM without Dropout

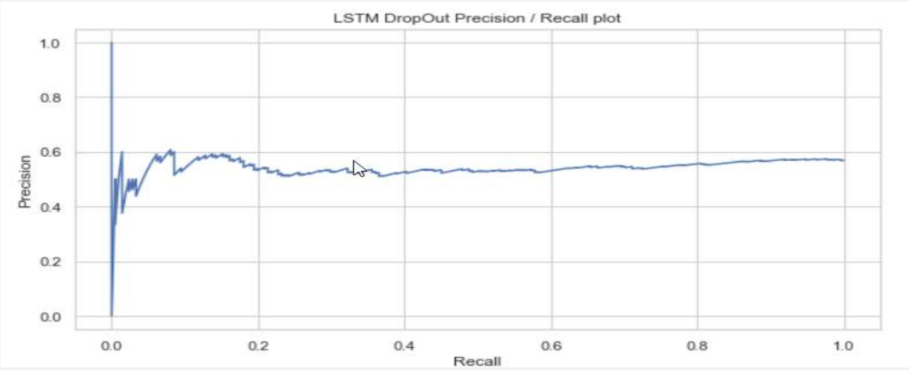


ROC-AUC Curve LSTM without Dropout

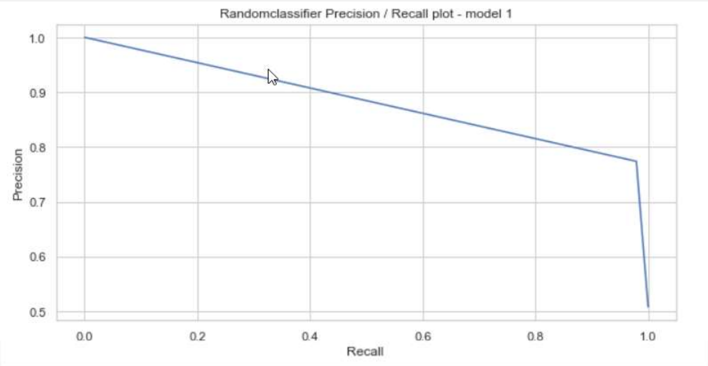


MODELING-NATURAL LANGUAGE PROCESSING(NLP)

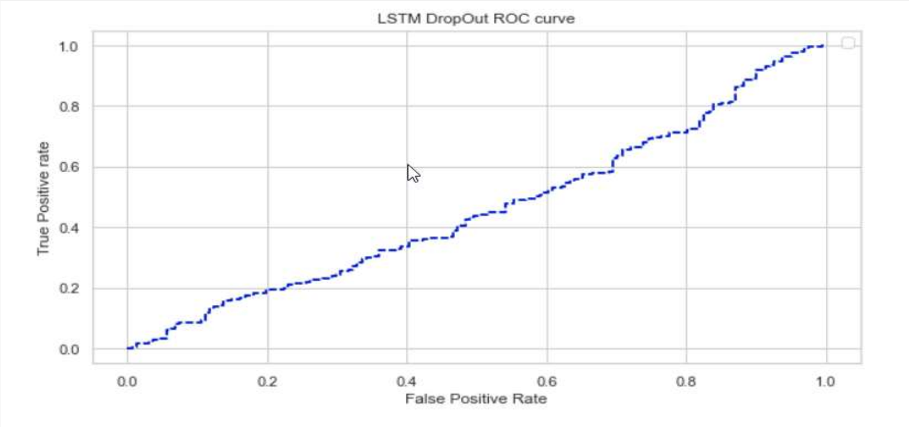
Precision/Recall LSTM Dropout



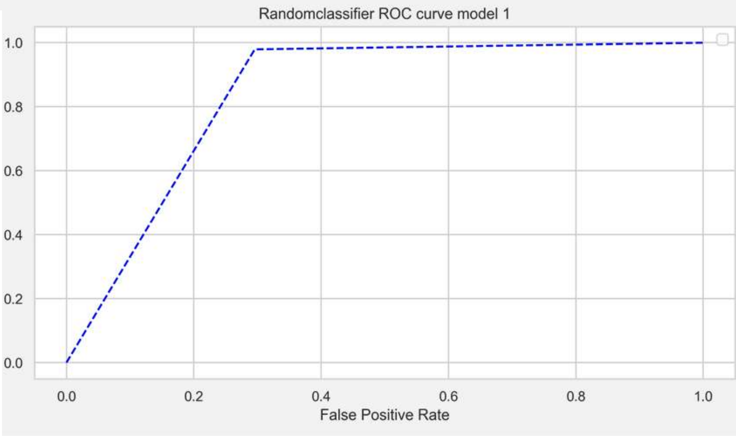
Precision/Recall Random Forest



ROC-AUC LSTM Dropout



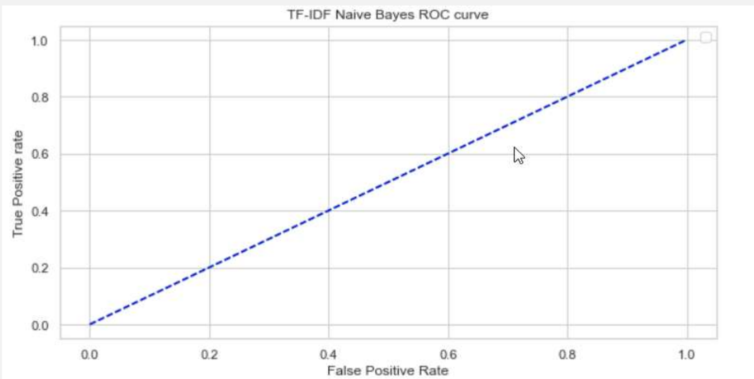
ROC-AUC Random Forest

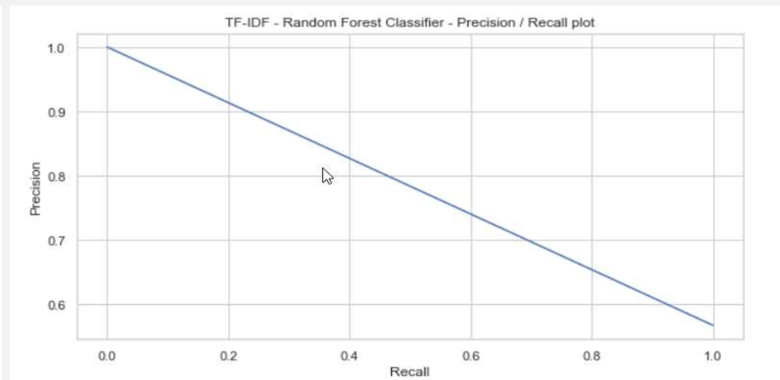


MODELING-NATURAL LANGUAGE PROCESSING(NLP)

TF-IDF stands for "Term Frequency — Inverse Document Frequency". By vectorizing the documents, we can further perform multiple tasks such as finding the relevant documents, ranking, clustering, etc. It is a text vectorizer that transforms the text into a usable vector.Term frequency indicates how important a specific term in a document. Inverse document frequency (IDF) is the weight of a term, it aims to reduce the weight of a term if the term's occurrences are scattered throughout the documents.

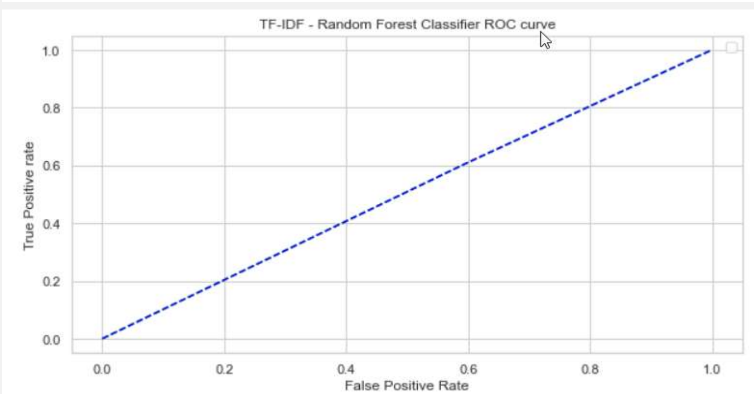
TF-IDF Naïve Bayes ROC-AUC Curve

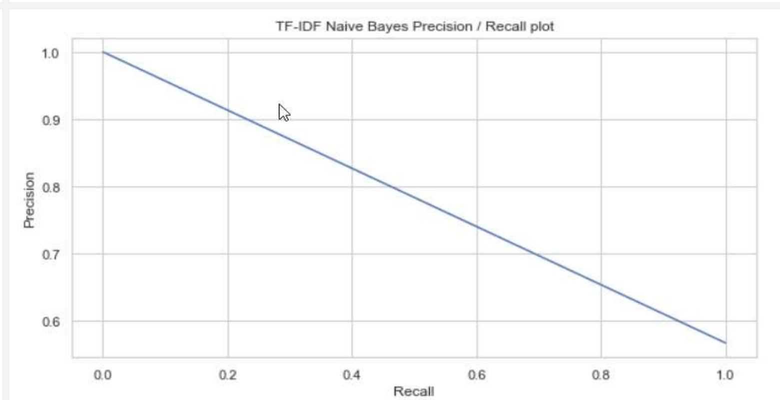




Precision/Recall TF-IDF Random Forest

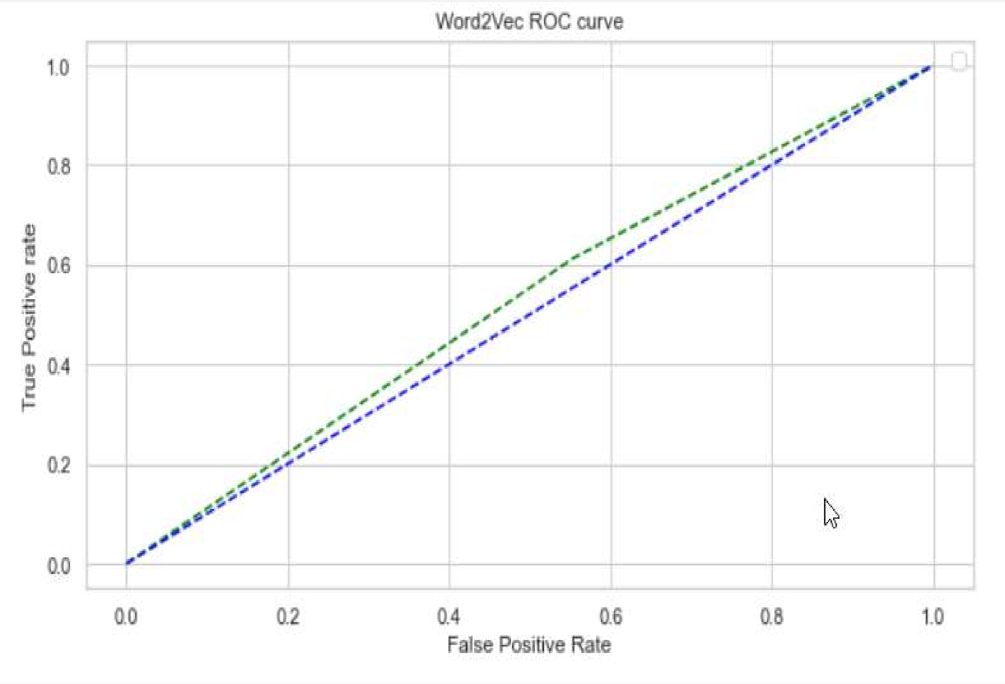
TF-IDF Random Forest ROC-AUC Curve

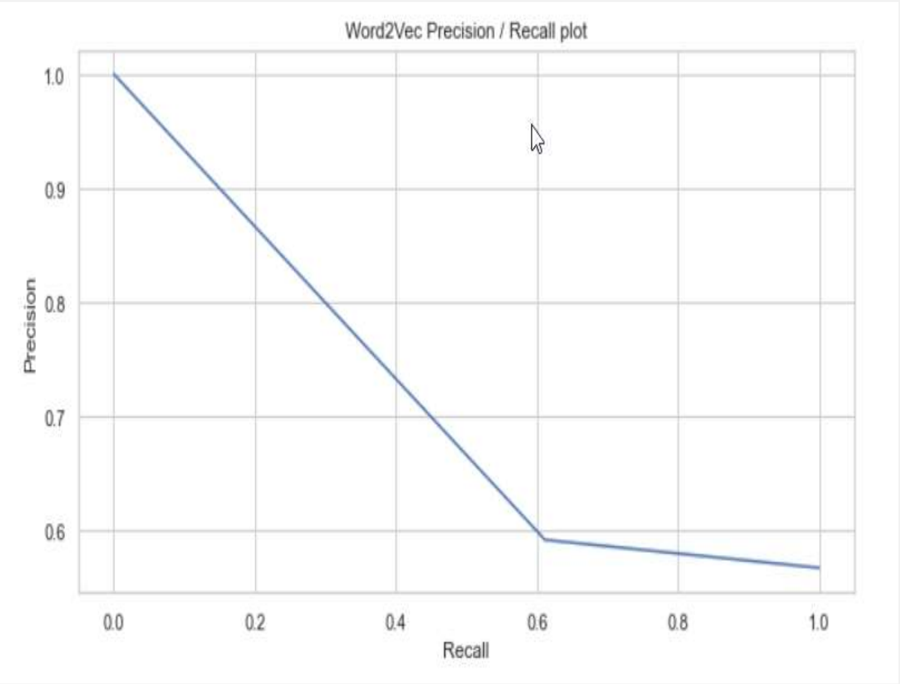


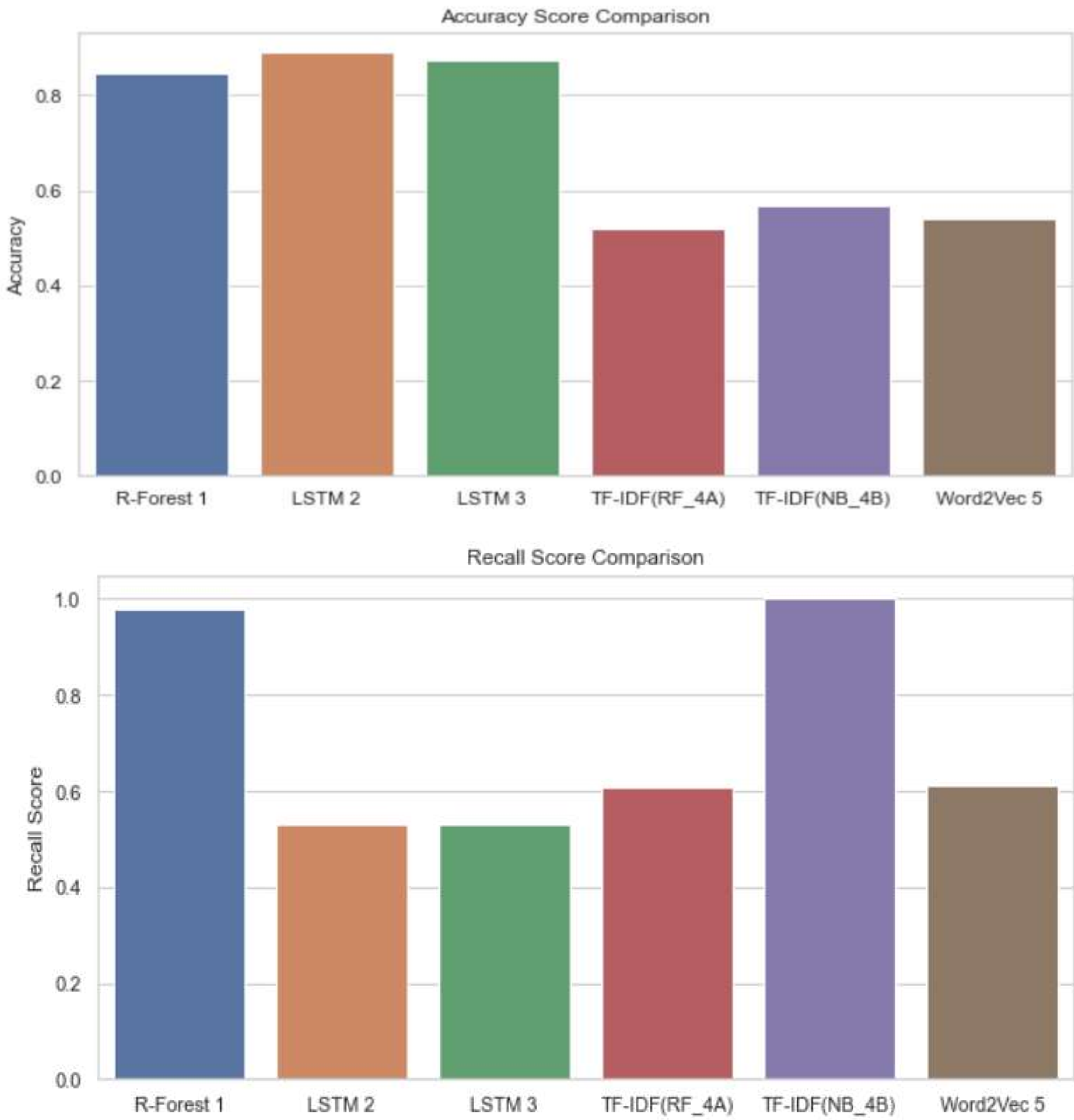


Precision/Recall TF-IDF Naïve Bayes

MODELING-NATURAL LANGUAGE PROCESSING(NLP)







MODEL I : RANDOM FOREST CLASSIFIER

ACCURACY VALUE WITH LSTM WITHOUT DROPOUT MODEL WHICH IS 84.39%

MODEL 2 : RECURRENT NEURAL NETWORKS(LSTM)

ACCURACY VALUE FOR LSTM
- WITHOUT DROPOUT MODEL IS 87.2%
- DROPOUT MODEL IS 87.96%

MODEL 3 : WORD2VEC-NEURAL NETWORK MODEL

ACCURACY SCORE IS 51.07%

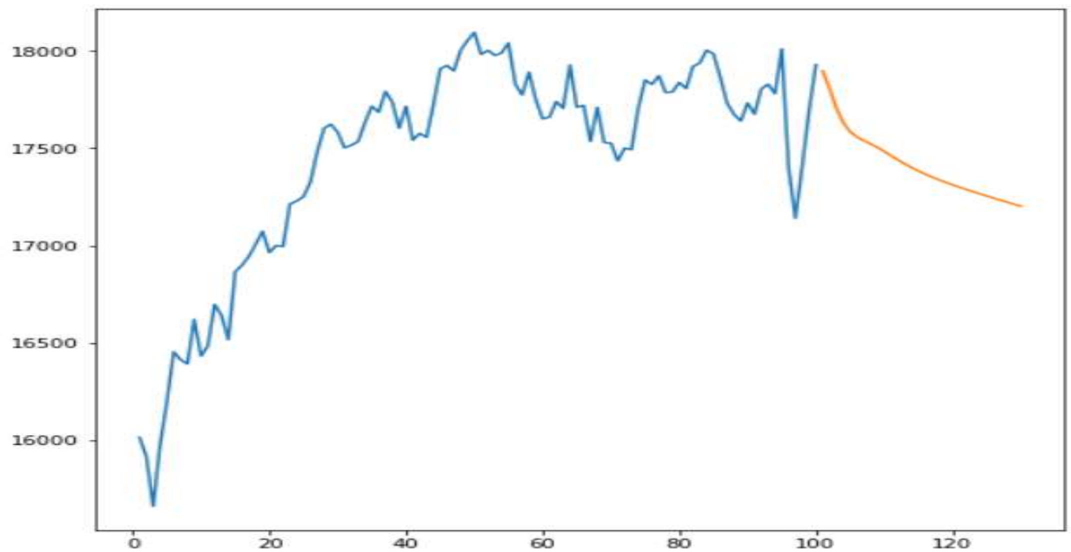
MODEL 4 : TF-IDF - TERM FREQUENCY–INVERSE DOCUMENT FREQUENCY

ACCURACY SCORE

- WITH RANDOM FOREST CLASSIFIER - 51.88%
- WITH NAIVE BAYES - 56.72%

SUMMARY:

MODEL 2 LSTM WITH AND WITH DROPOUT ARE GOOD MODELS



DEEP LEARNING - RECURRENT NEURAL NETWORKS(LSTM)

MODEL 1: LSTM NETWORK FOR REGRESSION:
TRAIN RMSE SCORE: 180.79
TEST RMSE SCORE: 340.61

MODEL 2: STACKED LSTM WITH MEMORY BETWEEN BATCHES:

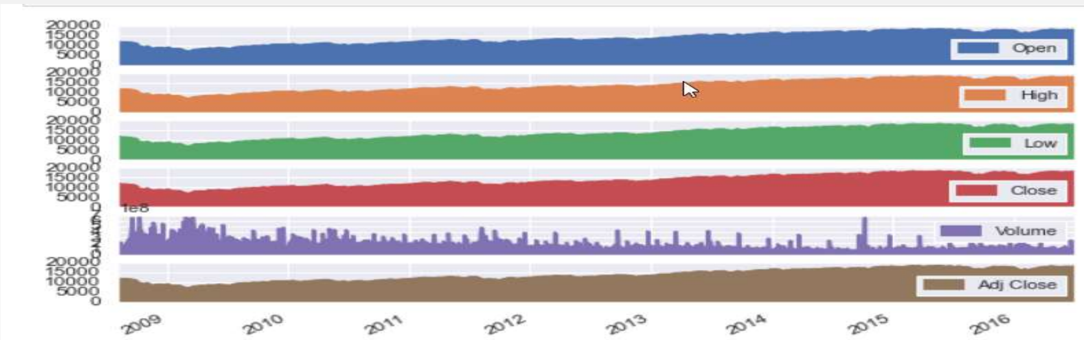
TRAIN RMSE SCORE: 126.95
TEST RMSE SCORE: 158.12

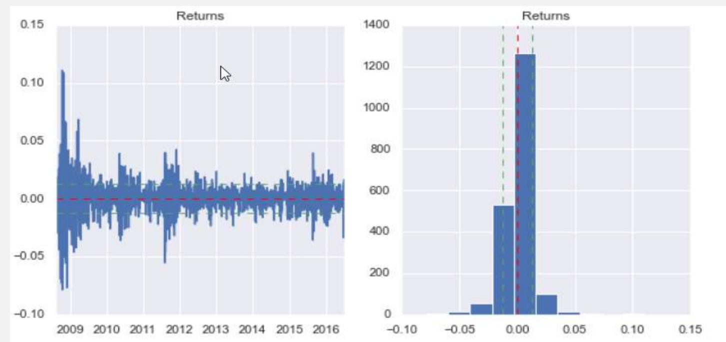
MODEL 2 IS THE BEST MODEL WITH SMALL ROOT MEAN SQUARED VALUE.

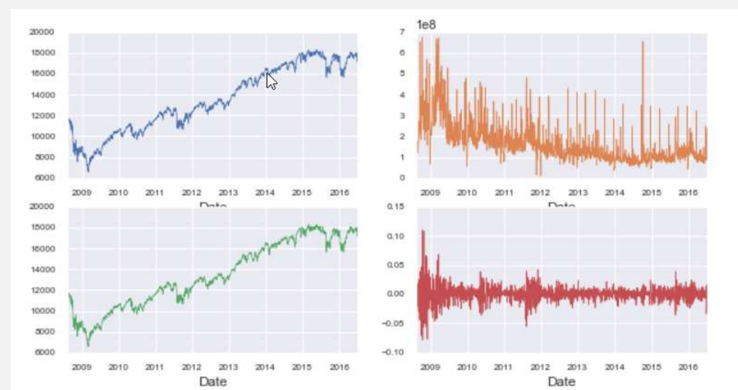
FIG1 - THIS IS THE PREDICTED DATASET AND IT IS PLOTTED, SHOWING THE ORIGINAL DATASET IN BLUE, THE PREDICTIONS FOR THE TEST DATASET IN GREEN, AND THE PREDICTIONS ON THE TRAINED DATASET IN ORANGE. RMSE TRAIN SCORE: 126.95 RMSE TEST SCORE: 158.12 ARE LOOKING GOOD.

FIG2 - FUTURE 30 DAYS FORECAST:
THIS IS THE FITURE 30 DAYS FORECAST AND PLOTTED THE VALUE. THE ORANGE COLORED VALUE IS FOR THE NEXT 30 DAYS.

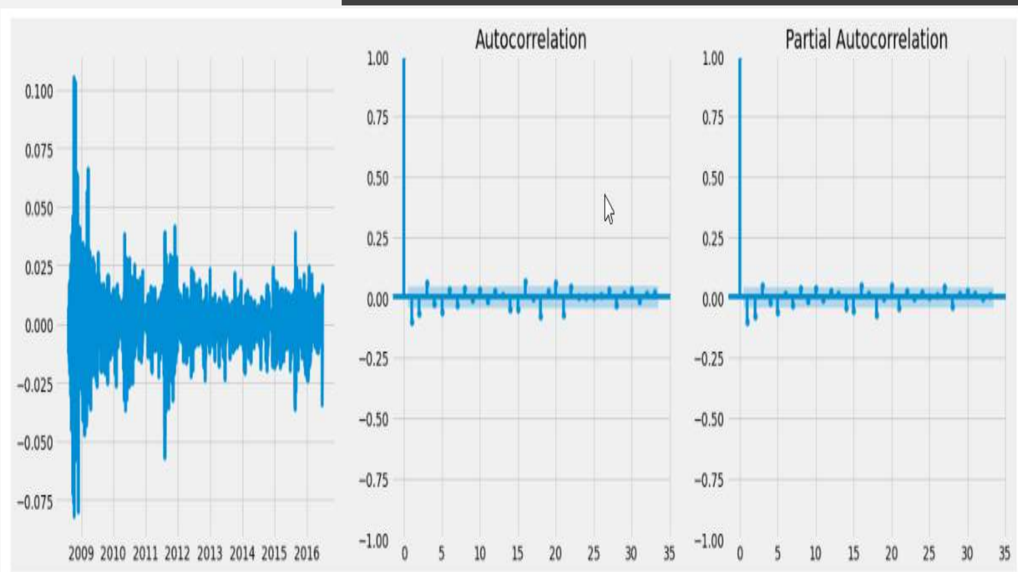
MODELING-TIME SERIES ANALYSIS-ARIMA

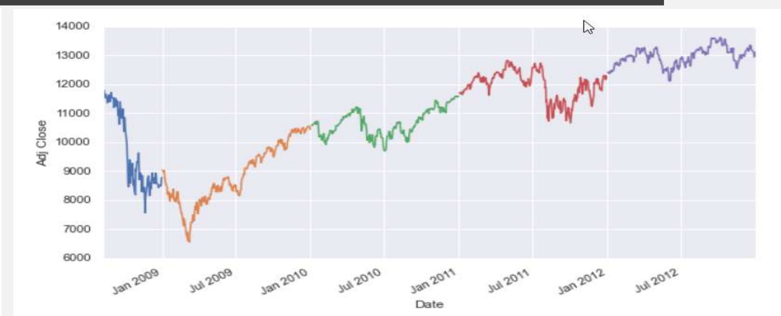


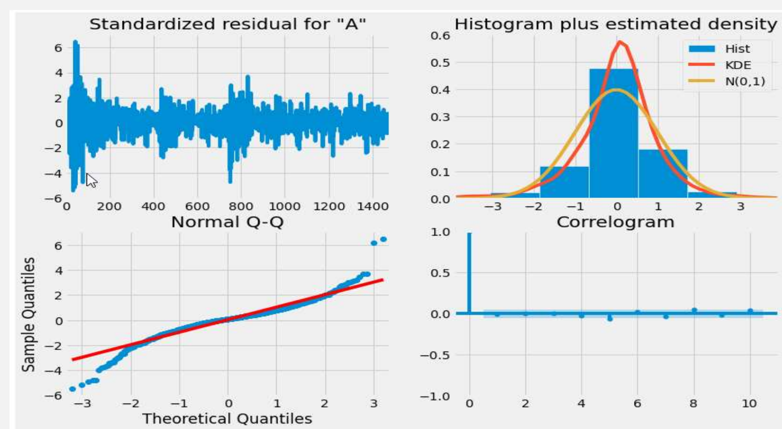


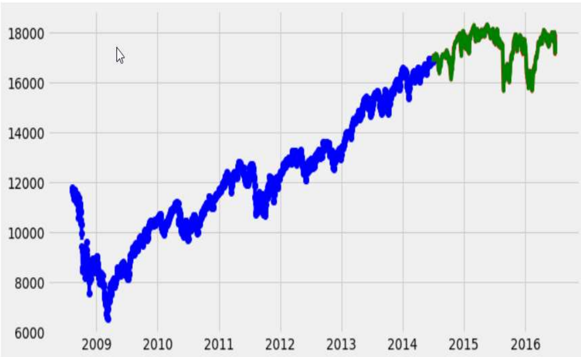


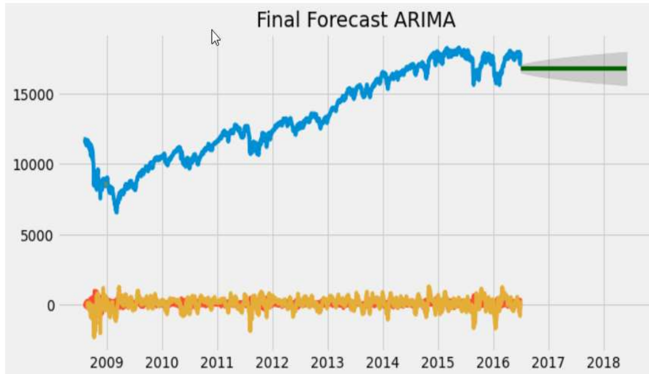
MODELING-TIME SERIES ANALYSIS-ARIMA












```
# Build ARIMA with recommended order
model_arima = ARIMA(X_train, order=(0,1,3))
model_fit = model_arima.fit()
forecast = model_fit.predict()

print(model_fit.summary())
```

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Adj Close   No. Observations:                 1480
Model:                 ARIMA(0, 1, 3)   Log Likelihood               -9384.697
Date:                Fri, 11 Feb 2022   AIC                          18777.393
Time:                        20:30:58   BIC                          18798.590
Sample:                             0   HQIC                         18785.295
                               - 1480
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.0978      0.018     -5.363      0.000      -0.134      -0.062
ma.L2         -0.0460      0.013     -3.439      0.001      -0.072      -0.020
ma.L3          0.0374      0.018      2.099      0.036       0.002       0.072
sigma2      1.903e+04    383.910     49.578      0.000    1.83e+04    1.98e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):              1754.48
Prob(Q):                              0.93   Prob(JB):                         0.00
Heteroskedasticity (H):               0.31   Skew:                            -0.36
Prob(H) (two-sided):                  0.00   Kurtosis:                         8.29
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

1) THE AVERAGE VALUE IN THE SERIES IS CALLED THE LEVEL.

2) THE INCREASING OR FALLING VALUE IN THE SERIES IS REFERRED TO AS THE TREND.

3) SEASONALITY IS THE SERIES' RECURRING SHORT-TERM CYCLE.

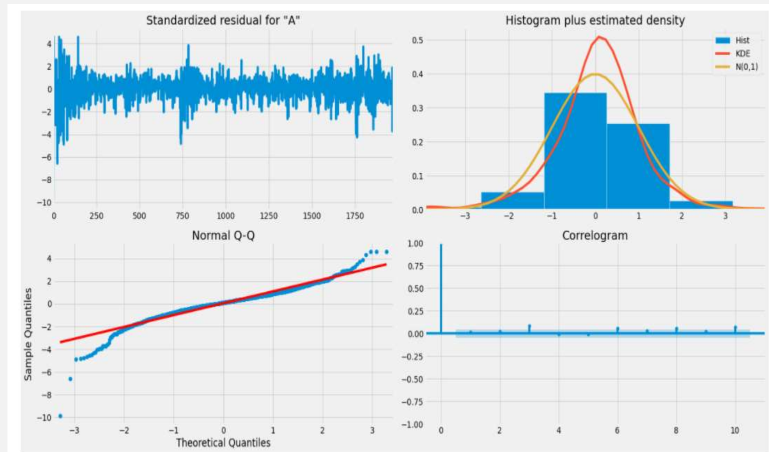
4) THE RANDOM VARIANCE IN THE SERIES IS REFERRED TO AS NOISE.

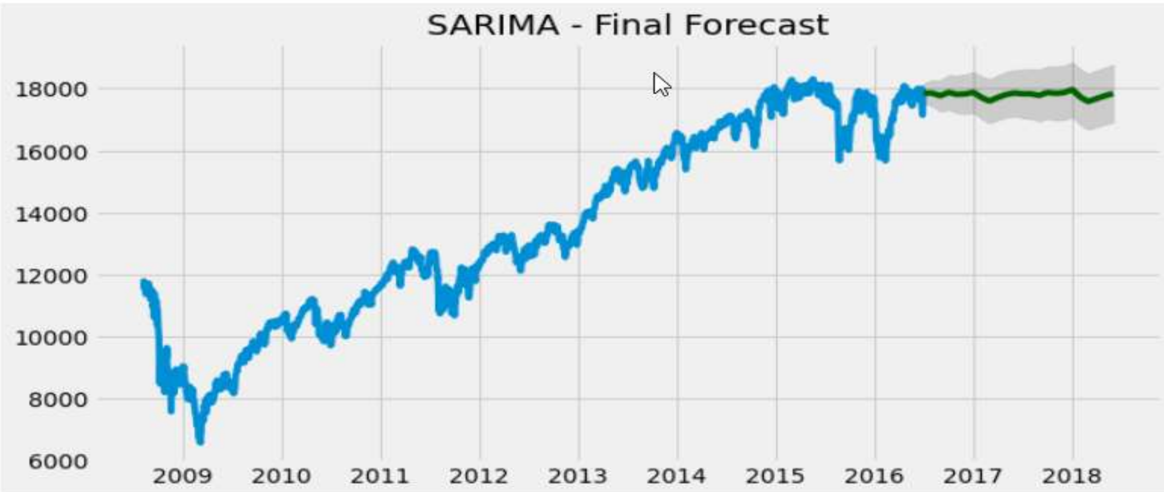
MAKING TIME SERIES STATIONARY
RUN THE AUGMENTED DICKY-FULLER TEST ON THE TIME SERIES TO TEST FOR STATIONARITY.
THE P-VALUE IS OBTAINED IS GREATER THAN SIGNIFICANCE LEVEL OF 0.05 AND THE ADF STATISTIC IS HIGHER THAN ANY OF THE CRITICAL VALUES.
SEASONALITY DIFFERENCING

MANY TIME SERIES EXHIBIT STRONG SEASONAL BEHAVIOR. FOR SEASONAL ADJUSTMENTS, INSTEAD OF TAKING FIRST DIFFERENCES, WE WILL TAKE DIFFERENCES WITH A LAG CORRESPONDING TO THE PERIODICITY.

THE P-VALUE IS EXTREMELY SMALL(P-VALUE 0.1), SO WE CAN EASILY REJECT THE HYPOTHESIS THAT PRICES ARE A RANDOM WALK AT ALL LEVELS OF SIGNIFICANCE.
ROOT MEAN SQUARED ERROR IS 157.23

MODELING-TIME SERIES ANALYSIS-SARIMA





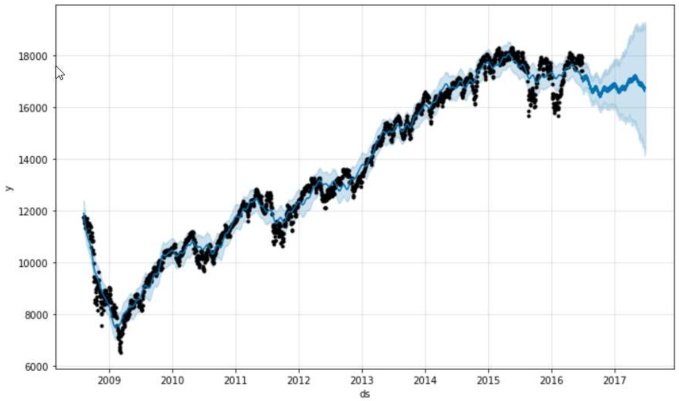

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                          Adj Close   No. Observations:                 1985
Model:             SARIMAX(3, 1, 0)x(2, 1, 0, 12)   Log Likelihood              -12834.944
Date:                            Fri, 11 Feb 2022   AIC                          25681.887
Time:                                    21:58:42   BIC                          25715.408
Sample:                                         0   HQIC                         25694.204
                                           - 1985
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.0659      0.016     -4.154      0.000      -0.097      -0.035
ar.L2         -0.0384      0.014     -2.744      0.006      -0.066      -0.011
ar.L3          0.0336      0.015      2.318      0.020       0.005       0.062
ar.S.L12      -0.6763      0.015    -45.108      0.000      -0.706      -0.647
ar.S.L24      -0.3515      0.016    -21.641      0.000      -0.383      -0.320
sigma2      2.626e+04    517.587     50.745      0.000    2.53e+04    2.73e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):               896.50
Prob(Q):                              0.93   Prob(JB):                         0.00
Heteroskedasticity (H):               0.83   Skew:                            -0.01
Prob(H) (two-sided):                  0.02   Kurtosis:                         6.30
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

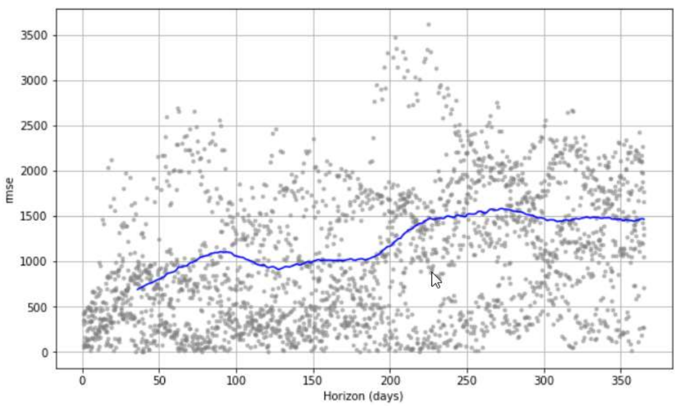
HIGH-LEVEL UNDERSTANDING OF TIME SERIES, STATIONARITY, SEASONALITY, FORECASTING, AND MODELING WITH SARIMAX

“SARIMA” MODEL IS A CONVENTIONAL MODEL BASED ON STATISTICS THAT ARE OFTEN USED TO PREDICT THE STOCK MARKET. THIS IS BECAUSE STOCK MARKET PRICES ARE NOT STATIC AND WOULD OFTEN VARY OVER TIME WHICH “SARIMA” IS ABLE TO PREDICT.

SEASONAL COMPONENTS CAN BE EXTRACTED OUT FROM THE ORIGINAL SERIES AND MODELED DIFFERENTLY

MEAN SQUARED ERROR : 169.99

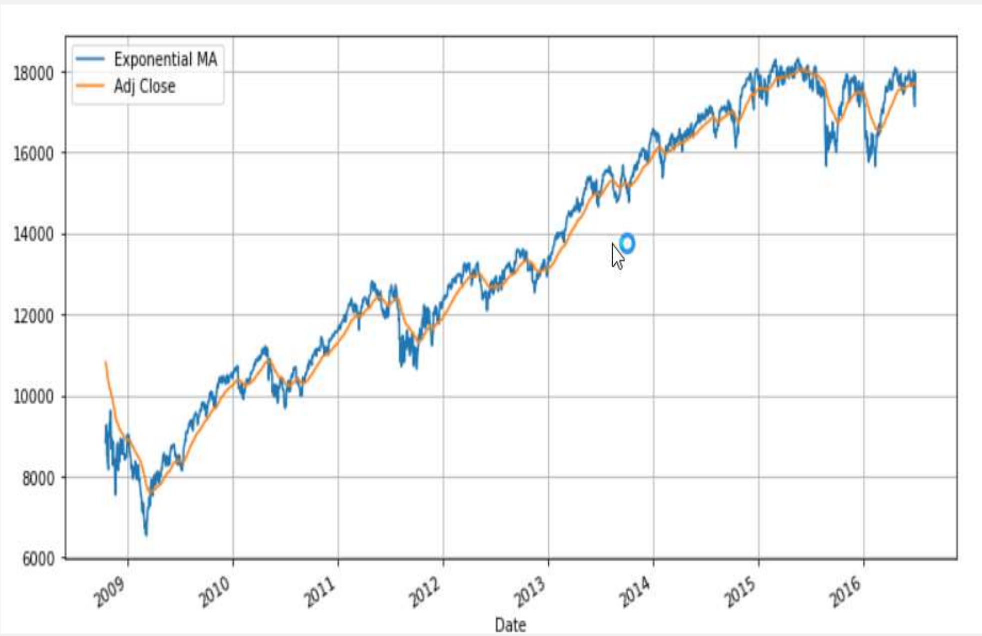


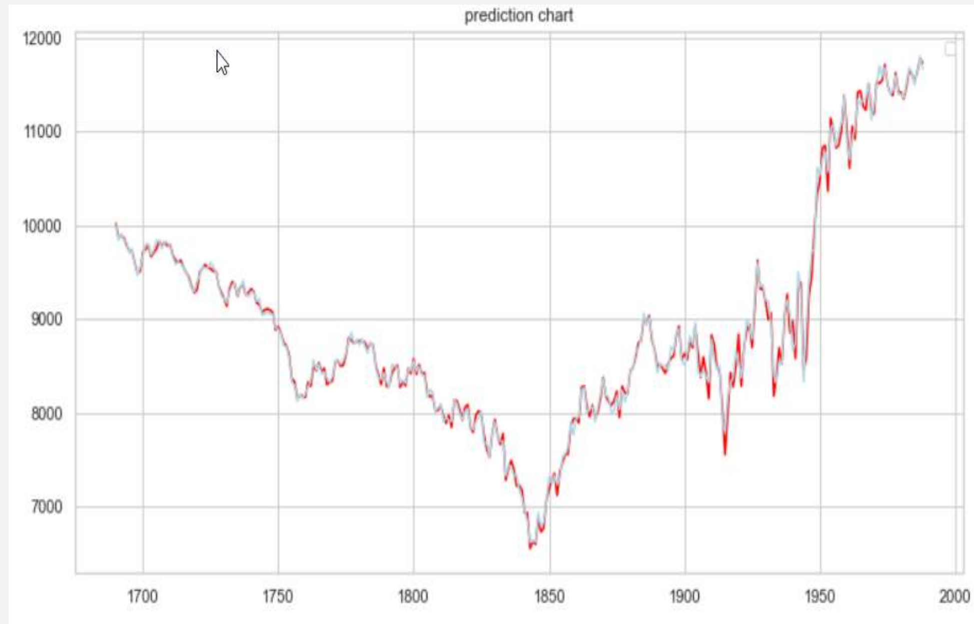


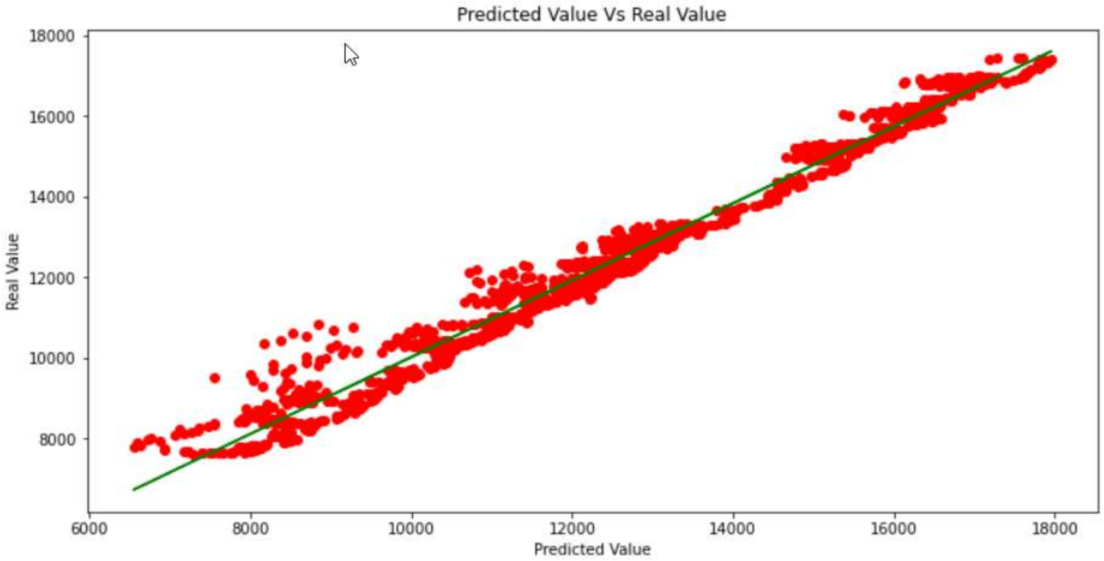
	horizon	mse	rmse	mae	mape	mdape	coverage
0	36 days	471794.024430	686.872641	549.432790	0.040527	0.029344	0.370518
1	37 days	484532.244901	696.083504	561.076271	0.041343	0.030327	0.357427
2	38 days	497090.656950	705.046564	572.081826	0.042192	0.031690	0.346614
3	39 days	509215.366675	713.593278	580.611555	0.042936	0.032093	0.334661
4	40 days	521542.844957	722.179233	589.208581	0.043647	0.032616	0.332669

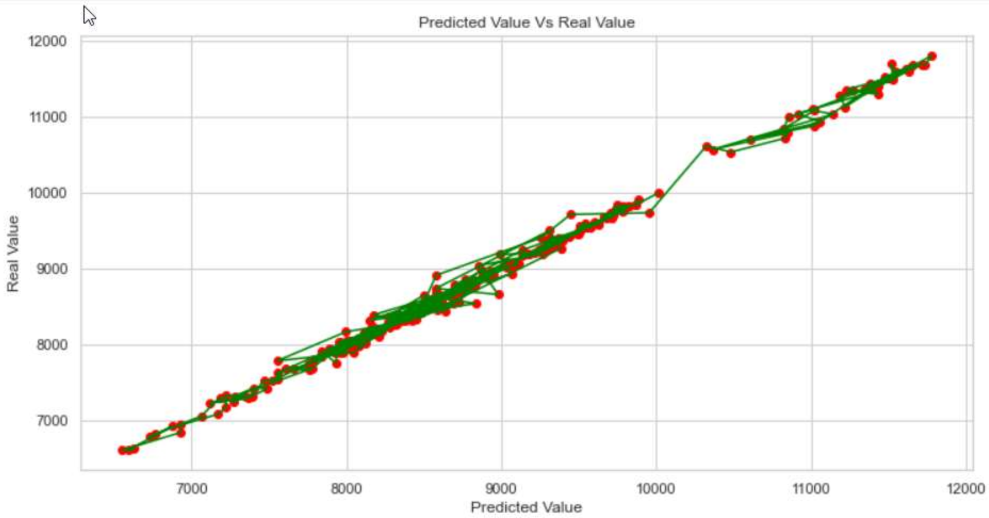
MODELING-TIME SERIES - PROPHET

MODELING-TIME SERIES – LINEAR REGRESSION (EMA)









PORTFOLIO GROWTH

THANK YOU

Archana Robin

archuskrishna@gmail.com