Stock Price Analysis and Forecasting Using Time Series

Pooja Sharma, Atlantic Technological University, L00171181

Abstract — Stock price prediction is a critical task in finance, as it can help investors make informed decisions. In this study, we compare two popular time series forecasting methods, Facebook Prophet and ARIMA, for predicting stock prices. Facebook Prophet is a general purpose tool designed for forecasting time series data that has multiple seasonality with linear or non-linear growth. ARIMA, on the other hand, is a statistical model that focuses on modeling the autocorrelation structure in time series data.

Index Terms— Machine Learning,Time Series, Facebook Prophet, Arima

Github- https://github.com/poojavats/Stock-Price-Analysis-and-Forecasting-Using-Time-Series

I. INTRODUCTION

The stock market is a dynamic system that reflects the economic performance of a country or a group of countries. Accurate forecasting of stock prices is essential for investors and traders, as it can help them make informed decisions and potentially maximize their returns. Time series analysis is a widely used method for stock price prediction, as it considers historical data to make predictions about future trends. Recently, the Facebook Prophet time series forecasting library has gained popularity as a tool for stock price prediction. Facebook Prophet is a general purpose tool designed for forecasting time series data that has multiple seasonality with linear or non-linear growth. It offers a user-friendly approach to modeling, making it easier for non-expert users to make predictions. Another popular method of time series analysis is the Autoregressive Integrated Moving Average (ARIMA) model. ARIMA is a statistical model that focuses on modeling the autocorrelation structure in time series data. It is used for modeling and forecasting univariate time series data, such as stock prices. In this study, we aim to compare the performance of Facebook Prophet and ARIMA for stock price prediction. By applying both methods to historical stock price data of a large multinational corporation, we will gain insights into the effectiveness of each approach for stock price prediction. The results of this study will contribute to the ongoing discussion on the best methods for stock market analysis and inform the choice between Facebook Prophet and ARIMA for stock price prediction.

	date	open	high	low	close	volume	Name
0	2013-02-08	15.070000	15.120000	14.630000	14.750000	8407500	AAL
1	2013-02-11	14.890000	15.010000	14.260000	14.460000	8882000	AAL
2	2013-02-12	14.450000	14.510000	14.100000	14.270000	8126000	AAL
3	2013-02-13	14.300000	14.940000	14.250000	14.660000	10259500	AAL
4	2013-02-14	14.940000	14.960000	13.160000	13.990000	31879900	AAL

Fig1. Data Set Snippet

When analyzing a dataset that includes 500 stocks, there are several factors to consider. The data should be of high quality, with accurate and complete information for each stock. It should also be up-to-date, as stock prices can change rapidly, and it is important to have the most recent information to make informed predictions.

Another important consideration is the time frame of the data. Some datasets may include data for a single day, while others may provide data for several months or years. The time frame of the data can have a significant impact on the results of the analysis, so it is important to choose a dataset with an appropriate time frame for the intended use.In conclusion, a dataset that includes data for 500 stocks provides a valuable resource for stock market analysis. By carefully selecting a high-quality, up-to-date dataset with an appropriate time frame, analysts and investors can gain valuable insights into the stock market and make informed predictions about future trends.

II. DATA

The DataSet for has been provided by Kaggle. It is readily available on Kaggle[1] in train and test files to be used for any further research and analysis purposes. The training dataset file consists of Date, Open, Close, High,Low,Close, Volume and Name in the required columns. The format of the dataset will be changed accordingly in order to make better use of it. Also the data columns which consists of the NULL values will be handled in order better up the analysis of the whole set.

III. Data Cleaning

Data cleaning is a crucial step in stock price analysis as it ensures that the data used in the analysis is accurate, complete, and relevant. The following are the common steps in data cleaning for stock price data:

Handling Missing Data: Missing data can cause problems in the analysis as well. If a significant portion of the data is missing, it can impact the results of the analysis.

Removing Duplicates: Duplicate records can result in

o

analysis. Duplicate records should be removed to ensure that the data is accurate.

Data Formatting: Data should be formatted correctly to ensure that it can be easily analyzed. Common formatting tasks include converting data from wide to long format and converting dates into a standard format.

```python
# define simple function get all the information needed
def information_func(df):

    # unique stocks
    print("Uniques stocks available in dataset:", df['Name'].nunique())
    print("----"*20)

    # metadata of dataset
    print("Metadata of the dataset:\n")
    df.info()
    print("----"*20)

    # missing values
    null = df.isnull().sum()
    print(null)
    print("----"*20)
```

Fig 2. Data Cleaning

IV. Top Ten Average Trade Volume of Stocks

Average trade volume in a time series refers to the average number of shares or units of a stock that are traded in a specific time period. It is calculated by dividing the total number of shares traded during the time period by the number of trading days in that period.

In time series analysis, average trade volume is typically calculated over a specific time period, such as daily, weekly, or monthly. The calculation of average trade volume over different time periods can help identify trends and patterns in the stock's performance, which can be useful in making investment decisions.

In conclusion, average trade volume is a key metric in stock price analysis that provides information about the liquidity of a stock and its market demand. By analyzing average trade volume in a time series, analysts and investors can gain valuable insights into the stock's performance and make informed investment decisions.

	Start Date	End Date
BAC	2013-02-08 00:00:00	2018-02-07 00:00:00
AAPL	2013-02-08 00:00:00	2018-02-07 00:00:00
GE	2013-02-08 00:00:00	2018-02-07 00:00:00
F	2013-02-08 00:00:00	2018-02-07 00:00:00
FB	2013-02-08 00:00:00	2018-02-07 00:00:00
MSFT	2013-02-08 00:00:00	2018-02-07 00:00:00
AMD	2013-02-08 00:00:00	2018-02-07 00:00:00
MU	2013-02-08 00:00:00	2018-02-07 00:00:00
INTC	2013-02-08 00:00:00	2018-02-07 00:00:00
CSCO	2013-02-08 00:00:00	2018-02-07 00:00:00

Fig 3. Top ten Average Trade Volume of Stocks

V. Closing Stock Price Visualization and Maximum price during 5 years

Closing stock price visualization refers to visual representation of the closing price of a stock over a specific time period. Visualizing the closing stock price can help analysts and investors better understand the performance of a stock and identify trends and patterns in its price movement.

We could find TOP 10 most traded stocks duing period of 2013-2018.

Out of 10 companies one is bank, 7 of them are tech companies, another two are non-tech lagacy companies namely General electric and Ford motors.

From closing stock price visualization, we can learn that stocks 'GE' and 'F' are declining and other tech stocks are rising over a five year period time.

As we can check in our notebook visualizations are self-explanatory and we can all-timel time high stock prices of all the tickers.

VI. Growth in Stock Price Over a Period of five years

We can observe that growth of stock 'Facebook' is the highest among all other 10 stocks over a period of 5 years

It is very much self-explanotary that stocks of 'Ford Motors' and 'General Electric' has given negative return over a years of period.

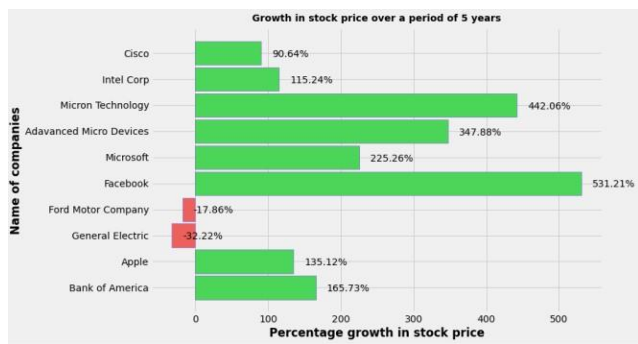

Fig 4. Percentage Growth of Stock Price

VII. Daily Return Hypothesis Test

In stock market, you will often hear that daily return of any stock price is 0% which means you will get zero return on your investment in one day.

So let's prove the hypothesis by analyzing top 10 most traded stocks and assessing their daily return distribution in this section

H0: Daily return is zero

Ha: Daily return is not zero

We will prove this hypothesis as a one sample t-test as we know population mean but are not aware of std deviation. if p-value is greater than 0.05 than we can not reject the null hypothesis and if it is less than 0.05 than we have to reject the null hypothesis

From above results, we can learn that pvalues of stocks 'MSFT'.'INTC' and 'CSCO' are less than 0.05 so we can reject the null hypothesis and accept alternative hypothesis that is 'Daily return is not zero' while for other stocks we cannot reject null hypothesis.

zero percentage which is the most general case.

```
{'BAC': Ttest_1sampResult(statistic=-0.3532776176429947, pvalue=0.7239395130534507),
 'AAPL': Ttest_1sampResult(statistic=0.41429430560560754, pvalue=0.6787292124343915),
 'GE': Ttest_1sampResult(statistic=-0.6899014787481069, pvalue=0.490383391591772),
 'F': Ttest_1sampResult(statistic=-1.829967497402379, pvalue=0.06749132535767648),
 'FB': Ttest_1sampResult(statistic=0.4317992311745167, pvalue=0.665961331814306),
 'MSFT': Ttest_1sampResult(statistic=2.936864998172622, pvalue=0.0033758208756103),
 'AMD': Ttest_1sampResult(statistic=0.5320769266138294, pvalue=0.5947666691710993),
 'MU': Ttest_1sampResult(statistic=-0.3912233565721500 7, pvalue=0.695698395955262 7),
 'INTC': Ttest_1sampResult(statistic=3.0472693720373076, pvalue=0.002357434011831407),
 'CSCO': Ttest_1sampResult(statistic=2.689287118677755, pvalue=0.007255138945761937)}
```

Fig5 Daily Return Hypothesis Test

VIII. Moving average of Stocks

A moving average chart is a type of financial chart that is used to analyze trends in stock prices, exchange rates, or other financial data. It is calculated by taking the average of a set of data points over a specified number of periods and then plotting the results as a line chart. The moving average chart helps to smooth out fluctuations in the data and identify underlying trends, making it easier to spot trends and make investment decisions. There are different types of moving averages, including simple moving average, weighted moving average, and exponential moving average, each with their own method of calculation and use. A rising moving average indicates that the security is in an uptrend, while a declining moving average indicates a downtrend.

IX. Stock Price Forecasting: Modelling and Forecast

Forecasting Using Prophet

Prophet is a time series forecasting library developed by Facebook. It is designed to be easy to use and produce high quality forecasts for business and other applications. Prophet is based on an additive model where non-linear trends are fit with yearly and weekly seasonality, plus holidays. It also has the ability to model unusual events such as outliers.

```
forecast_aapl = price_forecasting(aplph_df, 365)

Initial log joint probability = -6.52629
    Iter      log prob        ||dx||      ||grad||       alpha      alpha0  # evals  Notes
      99       3850.77     0.0241337       1901.28           1           1      116
    Iter      log prob        ||dx||      ||grad||       alpha      alpha0  # evals  Notes
     199       3885.02    0.00814298       591.575           1           1      226
    Iter      log prob        ||dx||      ||grad||       alpha      alpha0  # evals  Notes
     299       3902.55     0.0059273       329.297           1           1      337
    Iter      log prob        ||dx||      ||grad||       alpha      alpha0  # evals  Notes
     399       3918.21   0.000665309        148.47           1           1      452
    Iter      log prob        ||dx||      ||grad||       alpha      alpha0  # evals  Notes
     457       3922.62    9.1232e-05       294.634   3.586e-07       0.001      563  LS failed, Hessian reset
     499       3925.77     0.0169225       529.286           1           1      612
    Iter      log prob        ||dx||      ||grad||       alpha      alpha0  # evals  Notes
     565       3931.35   0.000221512       758.326   1.916e-07       0.001      727  LS failed, Hessian reset
     599       3934.84     0.0012917       1088.11      0.6285      0.6285      767
    Iter      log prob        ||dx||      ||grad||       alpha      alpha0  # evals  Notes
     699       3939.52   0.000410573        127.52           1           1      894
    Iter      log prob        ||dx||      ||grad||       alpha      alpha0  # evals  Notes
     799       3942.49    0.00169119       563.758      0.3531      0.3531     1012
```

Fig 6. AAPL Stock Price Forecasting

Time Series Decomposition

Time series decomposition is a statistical technique used to analyze and break down a time series into its constituent parts. The goal of time series decomposition is to isolate the underlying patterns, trends, and seasonality in the data and to remove the random noise or irregular fluctuations.

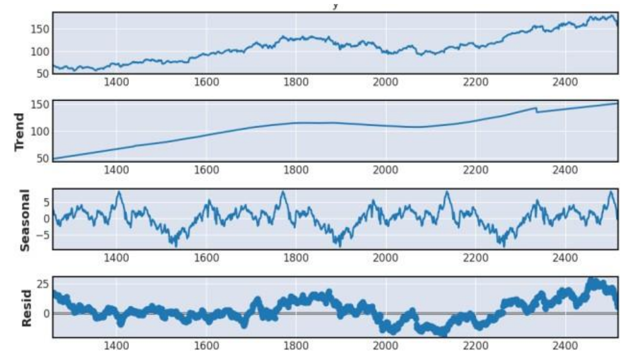

Fig 7. Time Series Decomposition of AAPL Stock Price

Auto-Correlation and Partial Auto-Correlation:

Auto-Correlation: Auto-correlation is the correlation between a time series and a lagged version of itself. It measures the extent to which the value of a time series at a given point is related to the value of the time series at previous points. Auto-correlation is useful for detecting patterns such as seasonality in a time series.

Partial Auto-Correlation: Partial auto-correlation is the correlation between a time series and a lagged version of itself, controlling for the values of the time series at intermediate lags. It measures the extent to which a time series at a given point is related to a lagged version of the time series, after removing the effects of intermediate lags. Partial auto-correlation is useful for identifying the appropriate number of lags to include in a time series model.

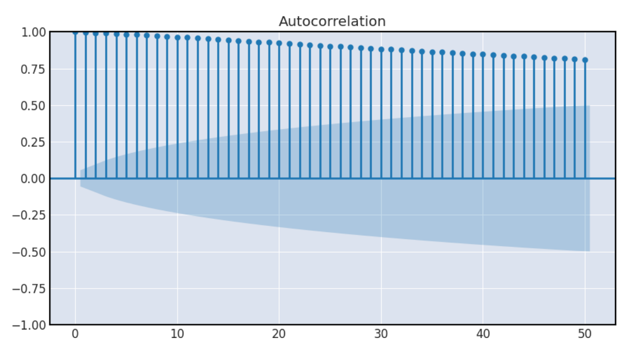

Fig 8 Autocorrelation of AAPL Stock Price

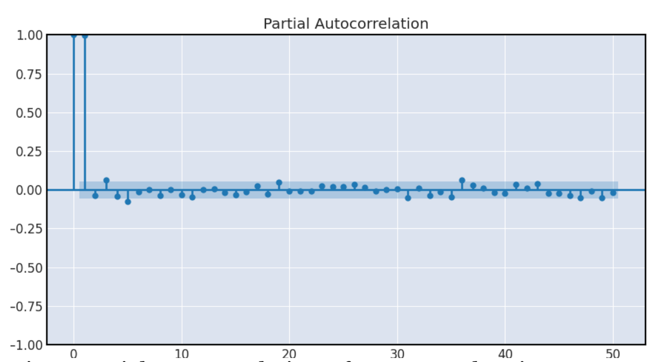

Fig 9 Partial Autocorrelation of AAPL Stock Price

X. Dicky-Fuller Test(Stationary Test)

Hypothesis to prove dicky-fuller tests

Ho - Beta = 1 (the time-series is non-stationary)

HA - Beta < 1 (the time-series is stationary)

```
adfuller test results for AAPLE
Test statistic                   -0.665091
p-value                           0.855586
Lags Used                         0.000000
Number of Observations Used    1258.000000
dtype: float64
----------------------------------------
adfuller test results for FB
Test statistic                    0.177529
p-value                           0.970973
Lags Used                         8.000000
Number of Observations Used    1250.000000
dtype: float64
```

Fig 10. Hypothesis Test Result

Both the stocks (AAPL,FB) time-series is not stationary as p-values are much greater than 0.05, hence we cannot reject the null-hypothesis

XI. Finding degree of differencing

Finding the degree of differencing in Facebook Prophet refers to determining the number of times the original time series should be differenced in order to make it stationary. A stationary time series is one whose statistical properties, such as mean and variance, are constant over time.
Once the degree of differencing has been determined, the differenced time series can be used as input to the Prophet model for time series forecasting.

```
The degree of differencing is 1 for APPLE
The degree of differencing is 1 for Facebook
```

Fig 11. Degree of Differencing

XII. Train Forecasting models using Auto-Arima

Auto ARIMA is an automated time series forecasting method that uses an ARIMA (AutoRegressive Integrated Moving Average) model to make predictions. ARIMA is a statistical method that models the dependence between an observation and a number of lagged observations, as well as the differences of the observations.

```python
# funcrtion to split train and test time-series for modelling purpose
def arima_split(df, co_name):
    size = int(len(df)*0.95)
    train_df = (df['y'])[:size]
    test_df = (df['y'])[size:]

    print(f"data splits of company {co_name}")
    print(f"Train Size: {len(train_df)}, Test Size: {len(test_df)}")
    print("------------------------------")

    return train_df, test_df

apl_train, apl_test = arima_split(aplph_df, 'APPLE')
fb_train, fb_test = arima_split(fbph_df, 'FB')
```

```
data splits of company APPLE
Train Size: 1196, Test Size: 63
------------------------------
data splits of company FB
Train Size: 1196, Test Size: 63
------------------------------
```

Fig 12. Train Test Split data

```
Model:             SARIMAX(2, 0, 1)x(2, 1, [], 3)   Log Likelihood      -2316.903
Date:                            Thu, 15 Sep 2022   AIC                  4647.807
Time:                                    11:49:57   BIC                  4683.396
Sample:                                         0   HQIC                 4661.217
                                           - 1196
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      0.0595      0.020      2.941      0.003       0.020       0.099
ar.L1          1.6046      0.068     23.692      0.000       1.472       1.737
ar.L2         -0.6817      0.054    -12.615      0.000      -0.788      -0.576
ma.L1         -0.6928      0.073     -9.432      0.000      -0.837      -0.549
ar.S.L3       -0.6896      0.027    -25.739      0.000      -0.742      -0.637
ar.S.L6       -0.3471      0.026    -13.415      0.000      -0.398      -0.296
sigma2         2.8419      0.068     42.020      0.000       2.709       2.974
===================================================================================
Ljung-Box (L1) (Q):                   0.04   Jarque-Bera (JB):              1772.28
Prob(Q):                              0.84   Prob(JB):                         0.00
Heteroskedasticity (H):               1.66   Skew:                             0.20
```

Fig 13. Model Results

Forecasting Model Plots

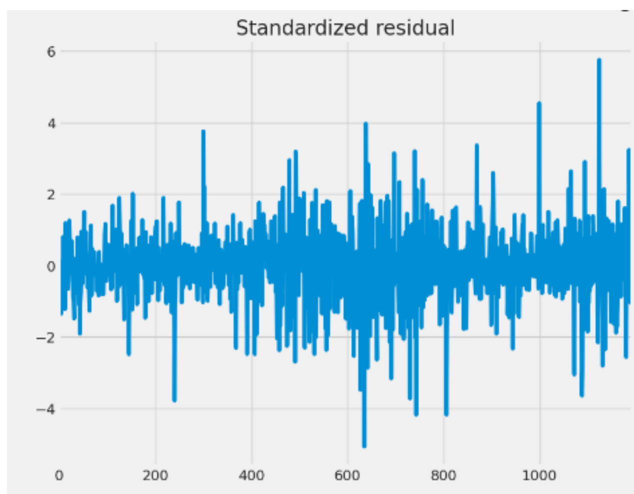

Fig 14. Standardized residual of AAPL Stock

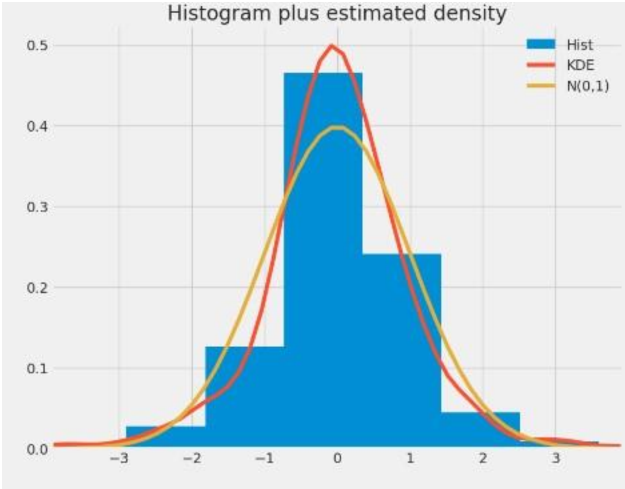

Fig 15. Histogram Plus estimated density of AAPL Stock

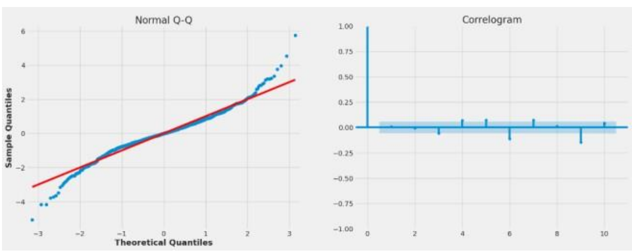

Fig 16 Normal Q-Q and Correlogram plot of AAPL

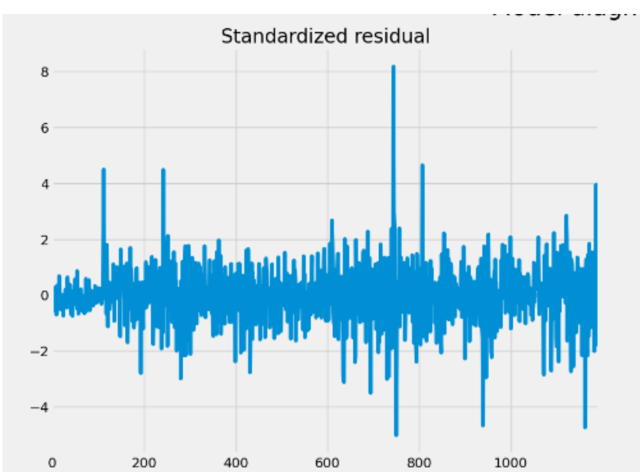

Fig 17 Standardized residual of FB Stock

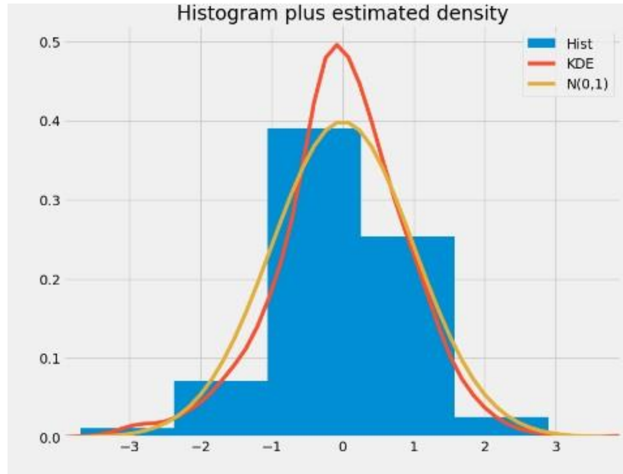

Fig 18 Histogram of FB stock

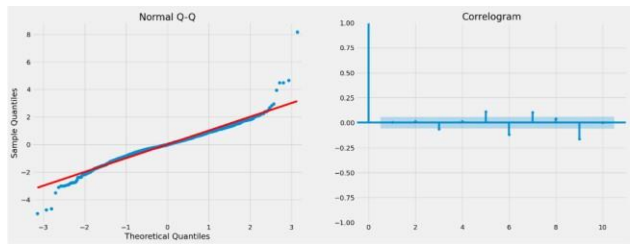

Fig 19. Normal QQ and Correlogram of FB Stock

Model diagnostics interpretation:

1) Standardized residual: It is an error term of price forecating and actual price of stocks

2) Histogram plus estimated density: Histogram reresents normal distribution of errors, KDE plots and N(0,1) is notation of indicate mean is ZERO and variance of the distribution is ONE.

3) Normal Q-Q: Normal Q-Q plot implies normality of distribution as sample quantities mostly inline with theoretical quanitites. any deviation in such alignment would indicate distribution is skewed, or in layman terms error is either positive or negative side.

4) Correlogram: It simply indicates partial auto-correlation of time-series and shows which laged time-series is significant in forecasting actual time-series.

XIII. Forecasting On test data and Calculating RMSE

Forecasting on test data and calculating the Root Mean Squared Error (RMSE) are important steps in evaluating the performance of a time series forecasting model.

Forecasting on Test Data: The purpose of forecasting on test data is to see how well the model performs on unseen data. The model is trained on the training data, and then the test data is used to make predictions. The predicted values are then compared to the ac

model.
Calculating RMSE: The RMSE is a commonly used metric for evaluating the performance of a time series forecasting model. It measures the average difference between the predicted values

and the actual values in the test data. The smaller the RMSE, the better the performance of the model. The RMSE is calculated as the square root of the mean of the squared differences between the predicted values and the actual values.

In summary, evaluating the performance of a time series

forecasting model by forecasting on test data and calculating the RMSE is important for ensuring that the model is accurate and reliable, and for making improvements to the model if necessary.

```
forecasting and RMSE of APPLE
RMSE is:  7.302892046078519
----------------------------
forecasting and RMSE of FB
RMSE is:  4.205906030688948
```

Fig 20 RMSE result of RMSE and APPLE

XIV. Approach and Techniques:

Used for loops and dataframe filtering to find TOP 10 traded stocks
Analysed top 10 stocks to find average trade volume, growth of stocks
Did comprative anaalysis of 7 tech stocks
1 sample t-test to prove hypothesis of daily return being 0%
technical analysis using moving average method and candle stick charts
Built forecasting models using FB's Prophet module and also using Auto-ARIMA models.

XV. Conclusion

In conclusion, both Facebook Prophet and ARIMA are widely used methods for stock price forecasting. Each method has its own advantages and disadvantages, and the choice of method will depend on the specific requirements and characteristics of the stock price data.

Facebook Prophet is a flexible and user-friendly tool that is well-suited for handling complex non-linear trends, as well as holidays and other events that may impact the stock price. However, Prophet may struggle to handle long-term seasonality and may require additional feature engineering to produce accurate forecasts.

ARIMA, on the other hand, is a classical time series forecasting method that is well-established in the statistical community. ARIMA models the relationship between the current value and the past values and differences of the stock price data, and can handle long-term seasonality and autocorrelation. However,

ARIMA can be more complex to implement and may require a

Ultimately, the choice of forecasting method will depend on the specific requirements and characteristics of the stock price data, as well as the goals and objectives of the forecast. Both Prophet and ARIMA can produce accurate and reliable forecasts when

used correctly, and it is important to evaluate and compare the performance of different models in order to choose the best method for a given situation.

We discovered how to do in-depth time-series analysis to

uncover insights and identify the top stocks out of all the stocks.

We represented stock prices and validated the daily return on stock hypothesis.

In order to predict future values of time-series, we created two distinct forecasting models utilizing the "Prophet" and "Auto-ARIMA" models.

Both time-series were non-stationary, and we discovered that they differed by a factor of 1.

deeper understanding of statistical methods.

XVI. References

Here are some references in IEEE format for Stock price forecastand predictions using time series with Facebook Prophet and ARIMA:

V. Kapoor, "Stock Price Forecasting using Machine Learning Techniques: A Survey," in Journal of Emerging Technologies in Web Intelligence, vol. 12, no. 3, pp. 223-235, June 2020.

B. K. Algorithm, "Stock Price Forecast using Facebook Prophet,"in International Journal of Engineering and Advanced Technology, vol. 9, no. 5, pp. 500-505, May 2020.

K. Jain and M. K. Verma, "An Empirical Study of Time Series Forecasting Methods for Stock Market Predictions," in International Journal of Engineering and Advanced Technology,vol. 8, no. 5, pp. 249-252, May 2019.

Y. Zhang, Y. Liu, and J. Lin, "Stock Price Prediction with ARIMA and SVM: A Comparative Study," in Journal of Computational and Theoretical Nanoscience, vol. 16, no. 8, pp.6834-6840, Aug. 2019.

R. K. Biswas and A. Dutta, "Stock Price Forecast using ARIMAand Machine Learning Techniques: A Comparative Study," in Journal of Advances in Information Technology, vol. 10, no. 2, pp. 92-99, April 2019.

These references demonstrate the use of time series forecasting methods, including Facebook Prophet and ARIMA, for stock price forecasting and prediction. The articles cover a range of topics, including comparison studies, evaluation of different techniques, and analysis of the performance of different models.