Forecasting S&P BSE SENSEX and S&P-500 Indices Using ARIMA and Prophet

Suraj Prakash Sharma
(BLENP2DSC20038)

M.Tech Data Science (IInd Semester)
Amrita Vishwa Vidyapeetham, School of Engineering, Bengaluru

Project Presentation, May 2021

Table of Contents

Objective & Problem Statement

Objective

Forecasting S&P BSE SENSEX and S&P-500 Using Autoregressive Integrated Moving Average (ARIMA) & Prophet Library.

Objective & Problem Statement

Objective

Forecasting S&P BSE SENSEX and S&P-500 Using Autoregressive Integrated Moving Average (ARIMA) & Prophet Library.

Problem Statement

- There are several studies in the research community which were carried out on the predictions of stock market returns using ARIMA and other powerful models especially for developed markets i.e. USA, European Markets. However very few have focused on emerging/developing and less developed markets.
- This study fills the gap by forecasting S&P BSE SENSEX (Emerging Market Index) and S&P-500 (Developed Market Index) in order to help the investors to make a more informed decision related to their investments regarding both the markets.

Literature Review

- Forecasting stock market returns plays a pivotal role whenever an investor/investment firm/organisation wants to build an investment strategy/policies [CMK20].

Literature Review

- Forecasting stock market returns plays a pivotal role whenever an investor/investment firm/organisation wants to build an investment strategy/policies [CMK20].
- Computational advancements have led to various econometric models which have been used consistently to anticipate market movements/irregularities and thus forecast the future prices/returns.

Literature Review

- Forecasting stock market returns plays a pivotal role whenever an investor/investment firm/organisation wants to build an investment strategy/policies [CMK20].
- Computational advancements have led to various econometric models which have been used consistently to anticipate market movements/irregularities and thus forecast the future prices/returns.
- ARIMA and its variants are one kind of time series models which can be used for short-term forecasting of financial time series data and various studies done in the research community uses ARIMA or variants of ARIMA models to forecast econometric variables like GDP, CPI, HPI, or price of an indexed financial assets etc.

Literature Review

- Forecasting stock market returns plays a pivotal role whenever an investor/investment firm/organisation wants to build an investment strategy/policies [CMK20].
- Computational advancements have led to various econometric models which have been used consistently to anticipate market movements/irregularities and thus forecast the future prices/returns.
- ARIMA and its variants are one kind of time series models which can be used for short-term forecasting of financial time series data and various studies done in the research community uses ARIMA or variants of ARIMA models to forecast econometric variables like GDP, CPI, HPI, or price of an indexed financial assets etc.
- The Jenkins ARIMA approach is more efficient then other econometric models which are based on regression and exponential smoothing.

Different Types of Markets [Fam70]

- There are various studies done by prominent economists like Paul Samuelson, Mandelbrot about the nature of the market but it was Eugene Fama [Fam70] who gave a framework to classify the nature of the market in his influential 1970 paper where he discussed his famous, controversial theory known as Efficient Market Hyphothesis.

Different Types of Markets [Fam70]

- There are various studies done by prominent economists like Paul Samuelson, Mandelbrot about the nature of the market but it was Eugene Fama [Fam70] who gave a framework to classify the nature of the market in his influential 1970 paper where he discussed his famous, controversial theory known as Efficient Market Hyphothesis.
- Efficient Market Hyphothesis (EMH): It states that the price of any financial asset or product at any time reflects all the public and private information which the market has processed.

Different Types of Markets [Fam70]

- There are various studies done by prominent economists like Paul Samuelson, Mandelbrot about the nature of the market but it was Eugene Fama [Fam70] who gave a framework to classify the nature of the market in his influential 1970 paper where he discussed his famous, controversial theory known as Efficient Market Hyphothesis.
- Efficient Market Hyphothesis (EMH): It states that the price of any financial asset or product at any time reflects all the public and private information which the market has processed.
- Direct conclusion of EMH is that it is impossible to beat the market consistently (i.e. generate alpha (α)).

Different Types of Markets [Fam70]

- There are various studies done by prominent economists like Paul Samuelson, Mandelbrot about the nature of the market but it was Eugene Fama [Fam70] who gave a framework to classify the nature of the market in his influential 1970 paper where he discussed his famous, controversial theory known as Efficient Market Hyphothesis.
- Efficient Market Hyphothesis (EMH): It states that the price of any financial asset or product at any time reflects all the public and private information which the market has processed.
- Direct conclusion of EMH is that it is impossible to beat the market consistently (i.e. generate alpha (α)).
- According to EMH there are 3 types of tests proposed to categorize the markets: 1. Weak Form (Dependent only on historical prices), 2. Semi-Strong (Depends on publicly available information i.e. annoucements of annual earnings, stock splits etc), 3. Strong Form (Depends on private information about the asset i.e. insider trading etc).

Datasets & Methodology

- The datasets are collected using either the library or by Yahoo Finance website.

Sr.No	Dataset Name	Time-Period	Source
1	S&P BSE SENSEX	2000-2020	quandl library
2	India VIX	2008-2020	investpy library
3	S&P-500	2000-2020	Yahoo Finance Website
4	CBOE VIX	1990-2020	investpy library

Table: Source of Datasets.

- India VIX & CBOE VIX Index datasets are used for explaining the market volatility which was high during the year 2007-2008 & same kind of volatility was also observed in the year 2020-2021.

Exploratory Data Analysis (EDA)

- There are various kinds of interactive plots which were created in order to understand the datasets clearly before developing time series based models.

Exploratory Data Analysis (EDA)

- There are various kinds of interactive plots which were created in order to understand the datasets clearly before developing time series based models.
- This section is divided into two parts:
  - S&P BSE SENSEX Index EDA.
  - S&P-500 Index EDA.

Descriptive Statistics of S&P BSE SENSEX

Sr.No	Stats	Close	%-Change
1	Mean	17930.2	0.0525427
2	Median	17222.6	0.0952336
3	Min	2600.12	-13.1526
4	Max	47751.3	17.3393
5	Std. Dev	11379.9	1.4641
6	Skewness	0.401981	-0.13972
7	Kurtosis	-0.875044	9.65161
8	Jarque Bera Test*	(307.514, 0.0)	(20257.592, 0.0)

Table: Descriptive Statistics of S&P BSE SENSEX Close and %-Change Values.

$*$ Jarque Bera Test is used to find out whether the values are normally distributed or not.

S&P BSE SENSEX EDA: Line Plot (2000-2020)

Figure: S&P BSE SENSEX Line Plot.

S&P BSE SENSEX EDA: Simple Moving Average & Yearly Distribution Plot

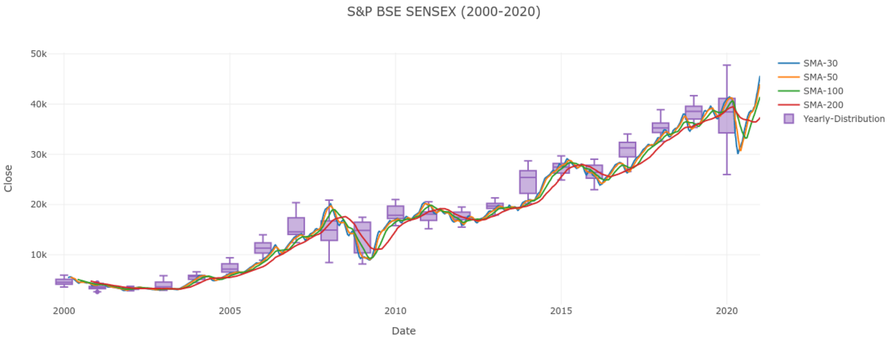


Figure: S&P BSE SENSEX Simple Moving Average (30, 50, 100, 200) and Yearly Distribution

S&P BSE SENSEX EDA: %-Change Plot

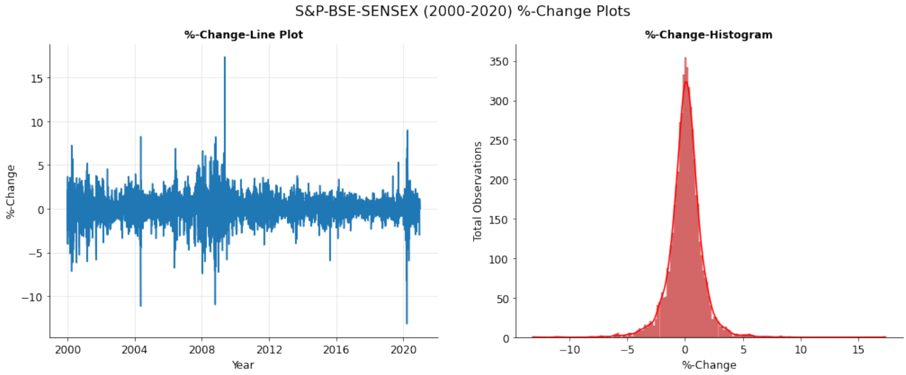


Figure: %-Change Plot in the Value of S&P BSE SENSEX.

Descriptive Statistics of S&P-500 Index

Sr.No	Stats	Close	%-Change
1	Mean	1653.27	0.025695
2	Median	1386.95	0.0593618
3	Min	676.53	-11.9841
4	Max	3735.36	11.58
5	Std. Dev	673.836	1.25313
6	Skewness	1.03217	-0.1538
7	Kurtosis	0.0629593	10.736
8	Jarque Bera Test*	(938.363, 0.0)	(25339.385, 0.0)

Table: Descriptive Statistics of S&P-500 Close and %-Change Values.

∗ Jarque Bera Test is used to find out whether the values are normally distributed or not.

S&P-500 EDA: Line Plot

Figure: S&P-500 Line Plot

S&P-500 EDA: Simple Moving Average & Yearly Distribution

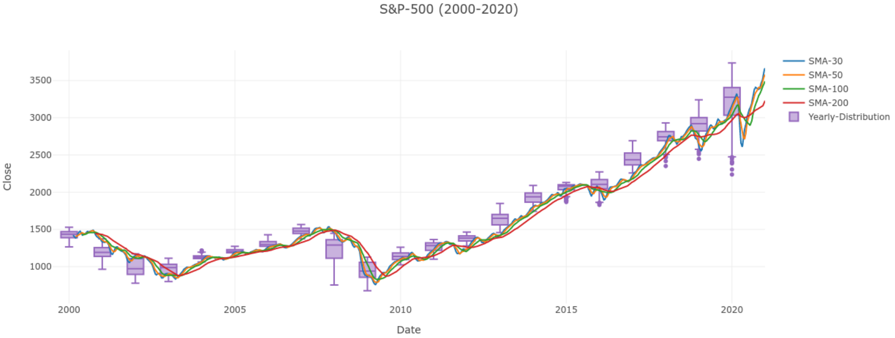


Figure: S&P-500 SMA (30, 50, 100, 200) and Yearly Distribution.

S&P-500 EDA: % Change Plots

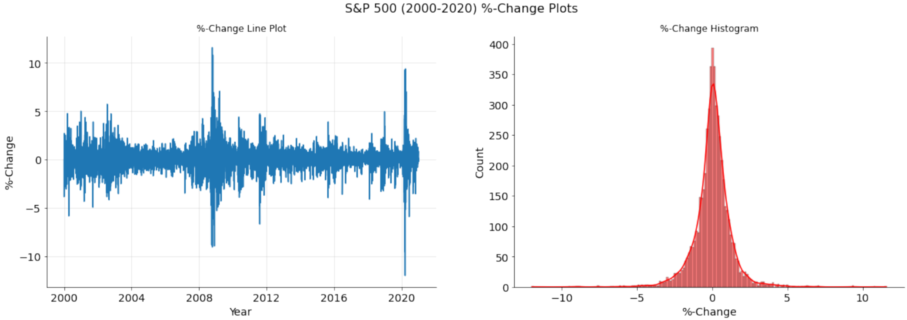


Figure: S&P-500 %-Change Plot.

Insights from EDA of S&P BSE SENSEX and S&P-500

- The %-Change plots shows the change in the value of indexes from previous day x_{t-1} to current day x_t and it turns out that the sequence which we got is a **White Noise** with approximately 0 mean (μ) and constant standard deviation (σ) which implies that the given time series data of indexes i.e. S&P BSE SENSEX and S&P-500 is a **Random Walk** process hence the given time series is **non-stationary**.

Insights from EDA of S&P BSE SENSEX and S&P-500

- The %-Change plots shows the change in the value of indexes from previous day x_{t-1} to current day x_t and it turns out that the sequence which we got is a **White Noise** with approximately 0 mean (μ) and constant standard deviation (σ) which implies that the given time series data of indexes i.e. S&P BSE SENSEX and S&P-500 is a **Random Walk** process hence the given time series is **non-stationary**.

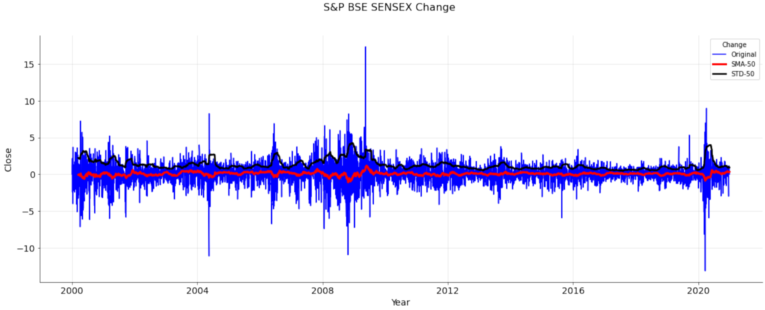


Figure: S&P BSE SENSEX %-Change Plot.

Random Walk Process Eqn's

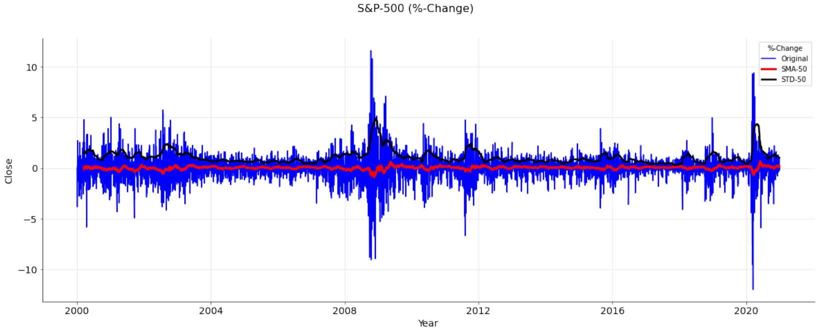


Figure: S&P-500 %-Change Plot.

Random Walk Process Eqn's

- Random Walk Process Eqn's:

$$(i)\ P_t = P_{t-1} + \epsilon_t$$

$$(ii)\ P_t = d + P_{t-1} + \epsilon_t$$

$$(iii)\ P_t = P_0 + dt + \sum_{t=1}^{n} \epsilon_t$$

where P_t = Value of underlying series at time t.
P_{t-1} = Value of underlying series at time $t - 1$.
d = Drift (which is just a trend like property for a random walk process i.e.
$d > 0 \implies$ Upward trend and $d < 0 \implies$ downward trend).
ϵ_t = White Noise or Gaussian White Noise.

INDIA VIX Index

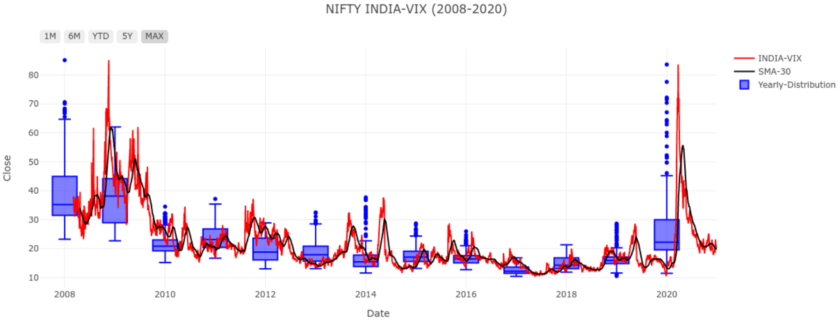


Figure: INDIA VIX (2008-2020)

- The above plot tells us about the total volatility present in the Indian Markets from the perpective of NIFTY-50.

CBOE VIX Index

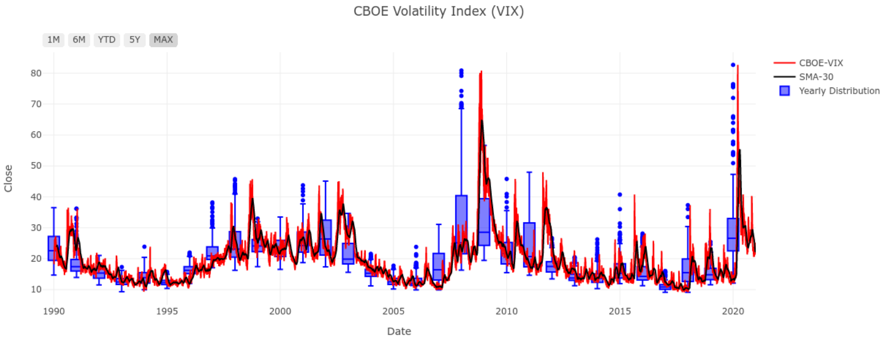


Figure: CBOE VIX Index

- The above figure gives us an idea about the volatility in the United States Markets from the perspective of S&P-500.

Insights from EDA of Volatility Indexes (India VIX & CBOE VIX)

- The interesting insights (labelled in the figure) to note is that in the FY-2009 and FY-2021, the volatility in the market (both Indian & United States) are pretty high beacause in the FY-2009, Global Financial Crisis happened due to crash of Mortgage market in United States and in FY-2021 COVID-19 crash happened due to great lockdown and the panic of recession.

S&P BSE SENSEX ACF and PACF Plots

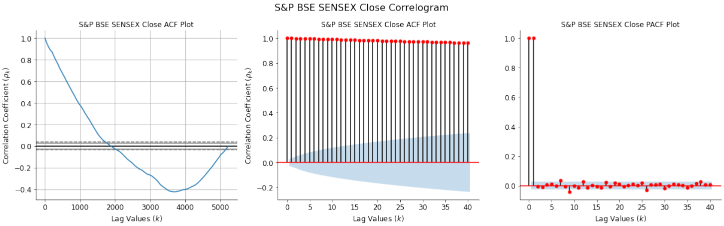


Figure: S&P BSE SENSEX ACF and PACF Plot.

S&P BSE SENSEX ACF and PACF Plots

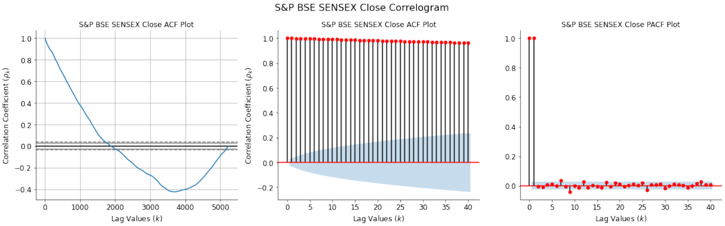


Figure: S&P BSE SENSEX ACF and PACF Plot.

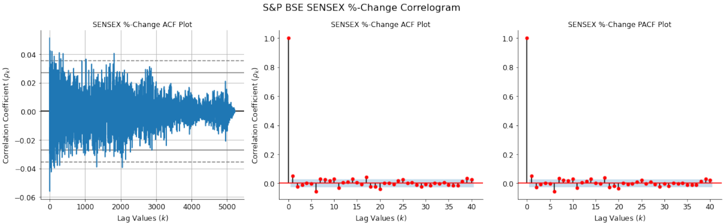


Figure: S&P BSE SENSEX %-Change ACF, PACF Plot.

S&P BSE SENSEX Time Series Components

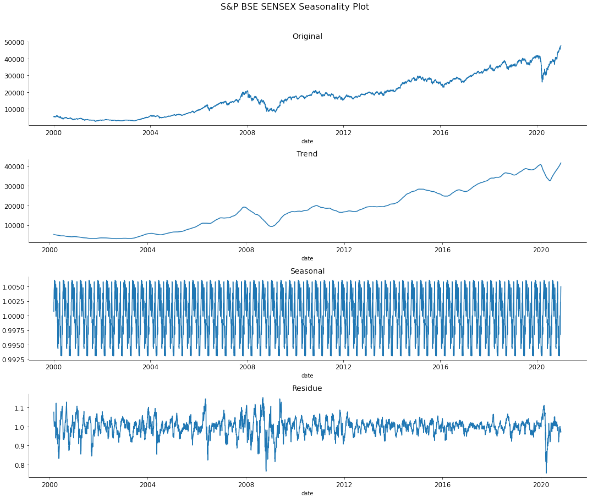


Figure: S&P BSE SENSEX Trend, Seasonal, Residual Plot (Quaterly).

S&P BSE SENSEX Time Series Cyclicity Component

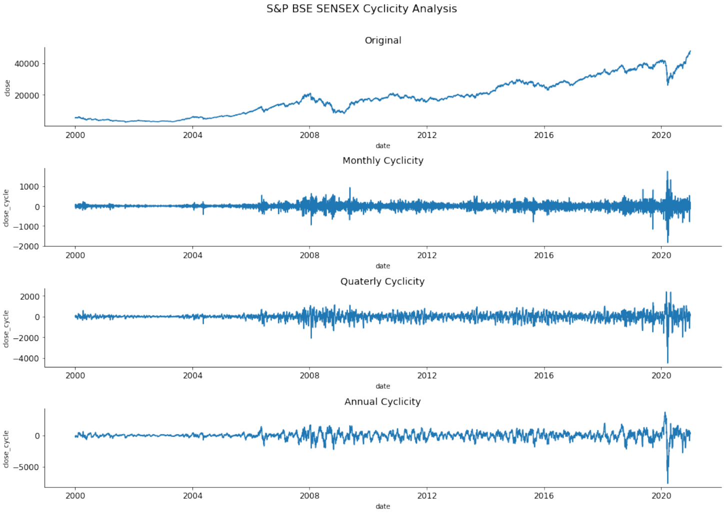


Figure: S&P BSE SENSEX Montly, Quaterly and Annualy Cyclicity.

S&P-500 ACF, PACF Plots

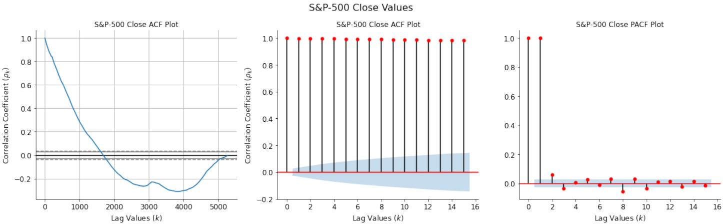


Figure: S&P-500 ACF, PACF Plots.

S&P-500 ACF, PACF Plots

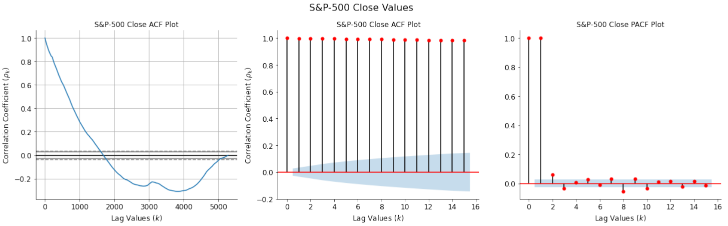


Figure: S&P-500 ACF, PACF Plots.

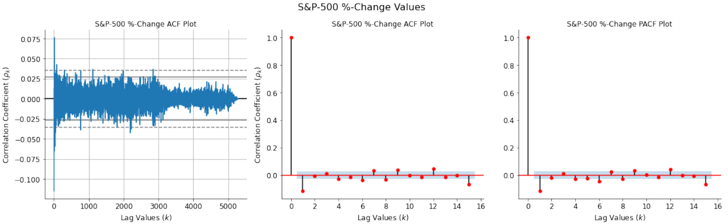


Figure: S&P-500 %-Change ACF, PACF Plots.

S&P-500 Time Series Components

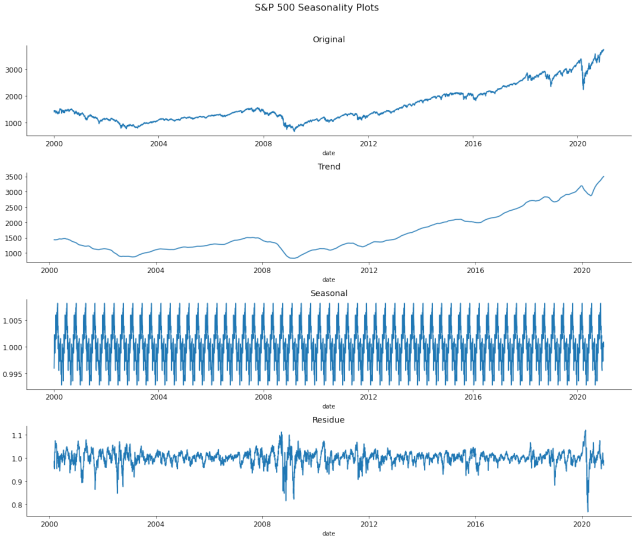


Figure: S&P-500 Trend, Seasonal, Residual Plot (Quaterly).

S&P-500 Time Series Cyclicity Component

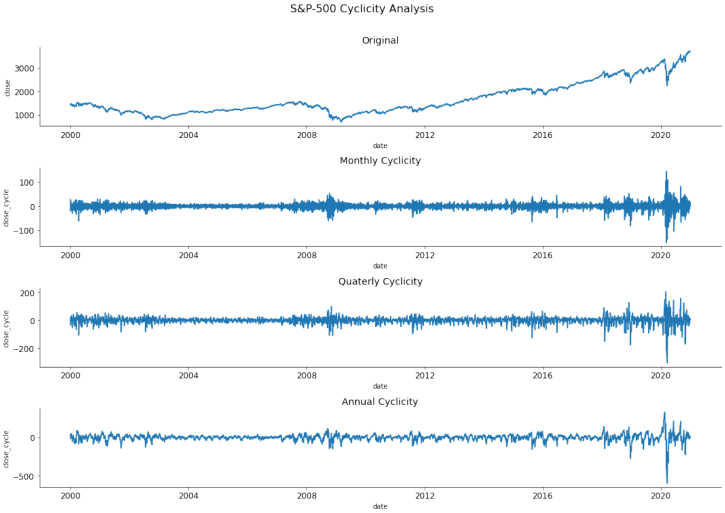


Figure: S&P-500 Montly, Quaterly and Annualy Cyclicity.

Statistical Tests for Stationarity: Augmented Dickey Fuller Test

- H_0: Given time series data is non-stationarity $\implies$ Time series data has time dependent statistical properties i.e. mean, variance, auto-correlation etc present in it.
  H_A : Given time series data is stationary $\implies$ Time series data doesn't have any time dependent statisitcal properties present in it.
- $p - value \leq 0.05$ then we reject the H_0 otherwise we don't have enough evidence to reject the $H_0 \implies$ failed to reject H_0.
- The results of ADF Test performed on both the indexes is shown on the next slide.

ADF Test Results for S&P BSE SENSEX Index

Sr.No	Data	t-statistic	p-value	Verdict
1	Close Values	0.688133	0.688133	Failed to Reject H_0.

Table: S&P BSE SENSEX Close Value ADF Test Results.

- As p-value obtained i.e. $0.998199 > 0.05(\alpha) \implies$ Failed to Reject $H_0 \implies$ S&P BSE SENSEX Close values is **non-stationary**.

ADF Test Results for S&P BSE SENSEX Index

Sr.No	Data	t-statistic	p-value	Verdict
1	Close Values	0.688133	0.688133	Failed to Reject H_0.

Table: S&P BSE SENSEX Close Value ADF Test Results.

- As p-value obtained i.e. $0.998199 > 0.05(\alpha) \implies$ Failed to Reject $H_0 \implies$ S&P BSE SENSEX Close values is **non-stationary**.

Sr.No	Data	t-statistic	p-value	Verdict
1	%-Change Values	-16.080128	5.389647×10^{-29}	Reject H_0.

Table: S&P BSE SENSEX %-Change Values ADF Test Results.

- As p-value obtained i.e. $5.389647 \times 10^{-27} << 0.05(\alpha) \implies$ Reject the $H_0 \implies$ S&P BSE SENSEX %-Change values is **stationary**.

ADF Test Results for S&P-500 Index

Sr.No	Data	t-statistic	p-value	Verdict
1	Close Values	1.631761	0.99795	Failed to Reject H_0.

Table: S&P-500 Close Value ADF Test Results.

- As p-value obtained i.e. $0.99795 > 0.05(\alpha) \implies$ Failed to Reject $H_0 \implies$ S&P-500 Close values is **non-stationary**.

ADF Test Results for S&P-500 Index

Sr.No	Data	t-statistic	p-value	Verdict
1	Close Values	1.631761	0.99795	Failed to Reject H_0.

Table: S&P-500 Close Value ADF Test Results.

- As p-value obtained i.e. $0.99795 > 0.05(\alpha) \implies$ Failed to Reject $H_0 \implies$ S&P-500 Close values is **non-stationary**.

Sr.No	Data	t-statistic	p-value	Verdict
1	%-Change Values	-13.74109	1.094051×10^{-25}	Reject H_0.

Table: S&P-500 %-Change Values ADF Test Results.

- As p-value obtained i.e. $1.094051 \times 10^{-25} << 0.05(\alpha) \implies$ Reject the $H_0 \implies$ S&P-500 %-Change values is **stationary**.

Estimating Parameters of ARIMA Model i.e. SARIMAX & Forecasting Using SARIMAX

```
def timeseries_forecast_using_arima(timeseries_data: pd.DataFrame, forecast_col_name: str,
exog_features: list = None, train_data_size: float = 0.90):
    if not isinstance(timeseries_data, pd.DataFrame):
        raise Exception("Given timeseries data is not an instance of Data-Frame class.")
    train_data, validation_data = tts(timeseries_data, train_size = train_data_size)
    auto_arima_model = auto_arima(train_data[forecast_col_name],
                                  X = train_data[exog_features] if exog_features else None,
                                  m = 7, # For Daily Forecasts
                                  stepwise = True,
                                  trace = True,
                                  error_action = "ignore",
                                  supress_warnigs = True)
    model_predictions = pd.Series(auto_arima_model.predict(validation_data.shape[0],
                                                          validation_data[exog_features] if
                                                          exog_features else None),
                                  index = validation_data.index)
    return train_data, validation_data, model_predictions, auto_arima_model
```

Figure: Source Code for Estimating Parameters and Forecasting Using SARIMAX.

Prophet Based Forecasting

```
def timeseries_forecast_using_prophet(timeseries_data: pd.DataFrame, exogenous_features,
train_data_size: float = 0.90):
    if not isinstance(timeseries_data, pd.DataFrame):
        raise Exception("Given timeseries data is not an instance of Data-Frame class.")
    columns_of_interest = ['date', 'close'] + exogenous_features
    timeseries_data = timeseries_data[columns_of_interest].rename(columns = dict(date = 'ds',
                                                                                  close = 'y'))
    train_data, validation_data = tts(timeseries_data, train_size = train_data_size)
    prophet_model = Prophet()
    for efeature in exogenous_features:
        prophet_model.add_regressor(efeature)
    prophet_model.fit(train_data)
    prophet_forecast = prophet_model.predict(validation_data)
    return prophet_model, prophet_forecast
```

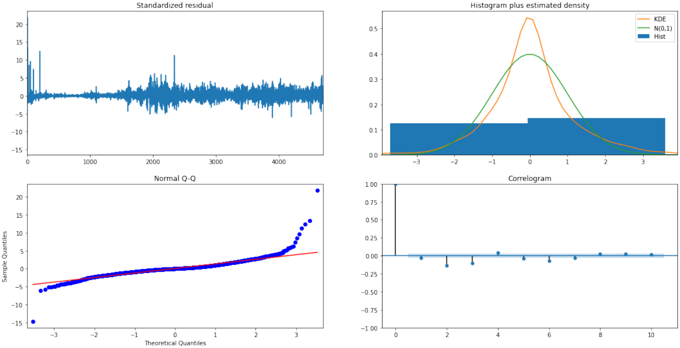
Figure: Source Code for Forecasting Using Prophet.

Parameters of SARIMAX Model for S&P BSE SENSEX

- The model used is SARIMA but as we are also exposing the SARIMA model to exogenous features i.e. those features which are not used to fit the model but have an influence on the model forecast. Hence the model used is SARIMAX where X is for exogeneous features presence.

Parameters of SARIMAX Model for S&P BSE SENSEX

- The model used is SARIMA but as we are also exposing the SARIMA model to exogenous features i.e. those features which are not used to fit the model but have an influence on the model forecast. Hence the model used is SARIMAX where X is for exogeneous features presence.
- Estimated hyperparameters of SARIMAX $(p, d, q) \times (P, D, Q, M)$ are: **SARIMAX** $(2, 0, 1) \times (2, 0, 0, 7)$ with an AIC $= 63798.806$.



SARIMAX Performance on Validation Set (S&P BSE SENSEX)

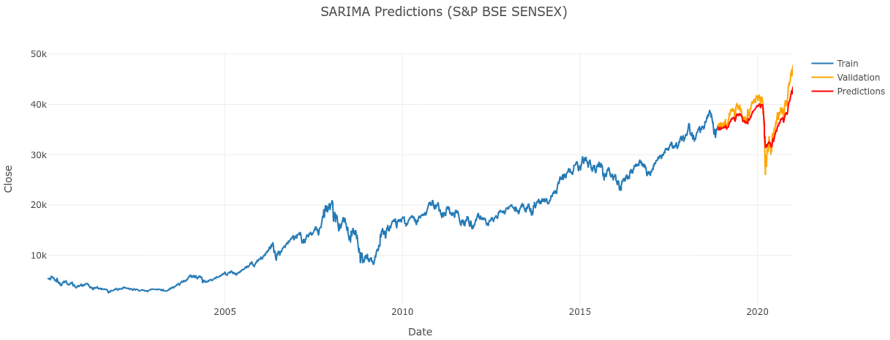


Figure: SARIMA Model Predictions for Validation Data (S&P BSE SENSEX).

Prophet Performance on Validation Set (S&P BSE SENSEX)

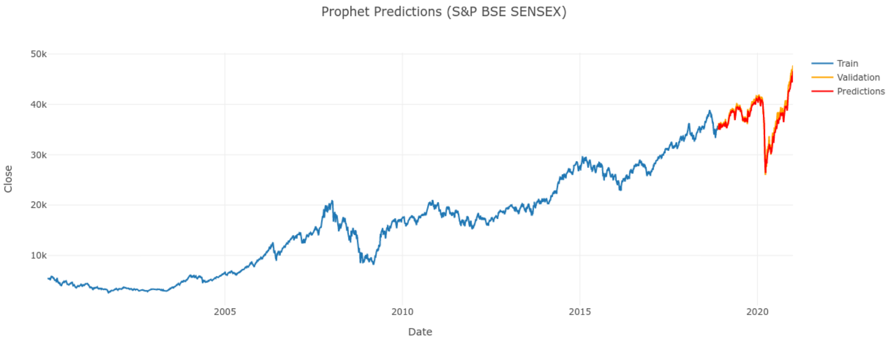


Figure: Prophet Model Predictions for Validation Data (S&P BSE SENSEX).

SARIMAX and Prophet Predictions Combined (S&P BSE SENSEX)

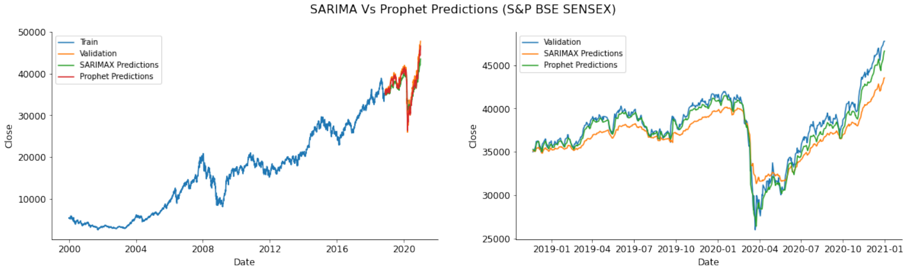


Figure: Prophet and SARIMAX Predictions on S&P BSE SENSEX Validation Set.

SARIMAX and Prophet Predictions Combined (S&P BSE SENSEX)

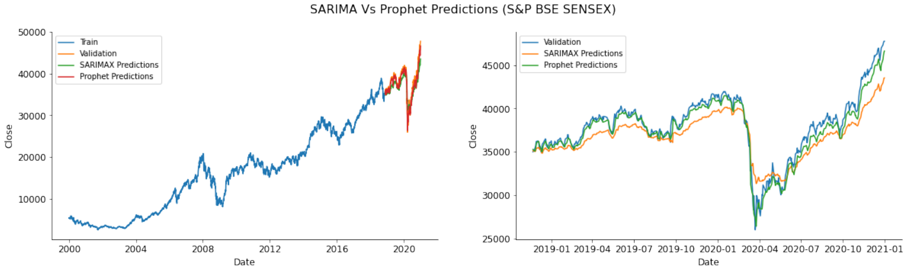


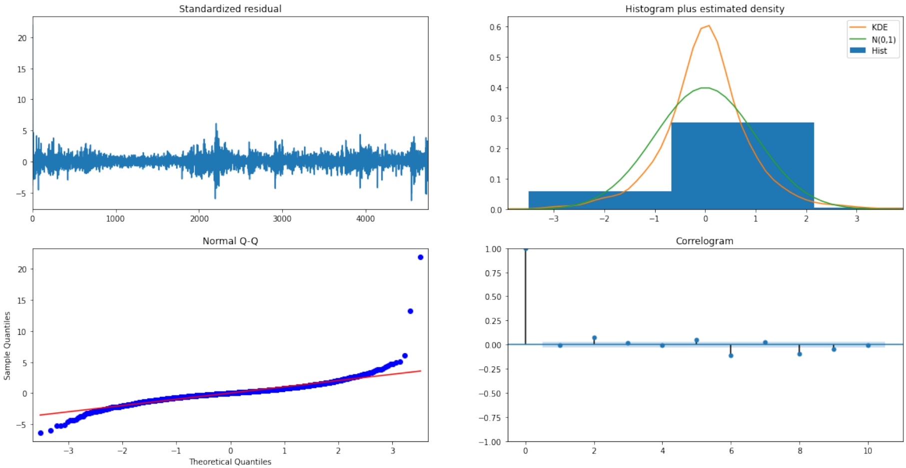
Figure: Prophet and SARIMAX Predictions on S&P BSE SENSEX Validation Set.

Model	MAE	MSE	RMSE	R2-Score	MAPE
SARIMAX	1553.135	3.386×10^6	1839.904	0.731307	4.05%
Prophet	611.323	5.998×10^5	774.435	0.953	1.60%

Table: SARIMAX and Prophet Error Metric Values.

Notes on Parameters of SARIMAX Model for S&P-500

- Estimated hyperparameters of SARIMAX$(p, d, q) \times (P, D, Q, M)$ are: **SARIMAX** $(5, 1, 1) \times (2, 0, 1, 7)$ with an AIC=35701.260.



SARIMAX Performance on Validation Set (S&P-500)

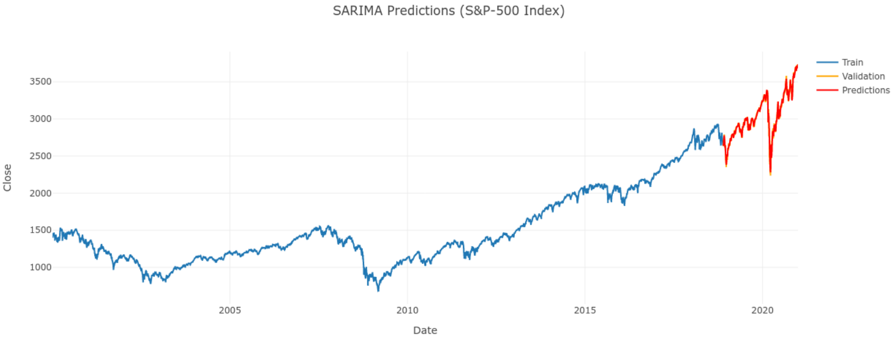


Figure: SARIMA Model Predictions for Validation Data (S&P-500).

Prophet Performance on Validation Set (S&P-500)

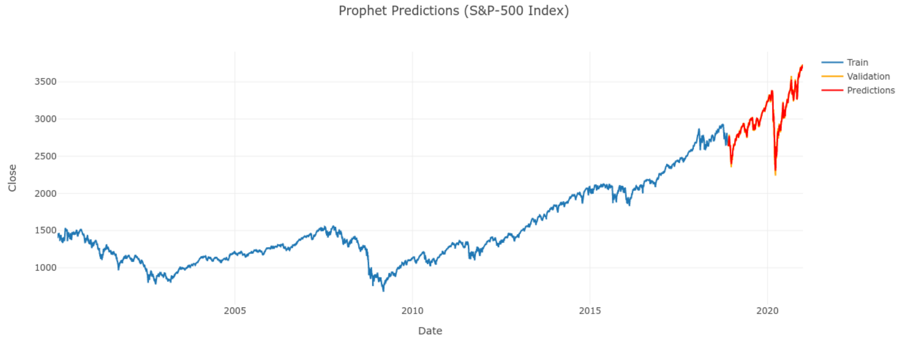


Figure: Prophet Model Predictions for Validation Data (S&P-500).

SARIMAX and Prophet Predictions Combined (S&P-500)

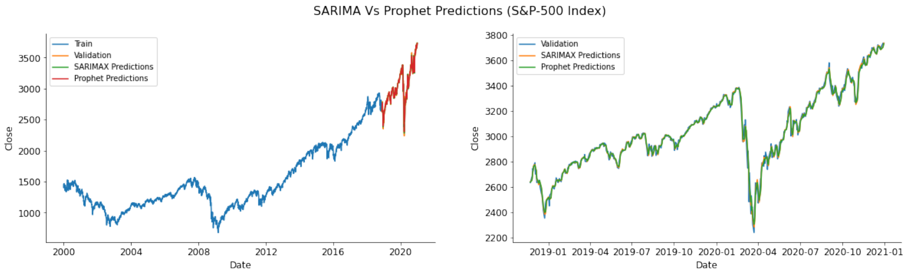


Figure: Prophet and SARIMAX Predictions on S&P-500 Validation Set.

SARIMAX and Prophet Predictions Combined (S&P-500)

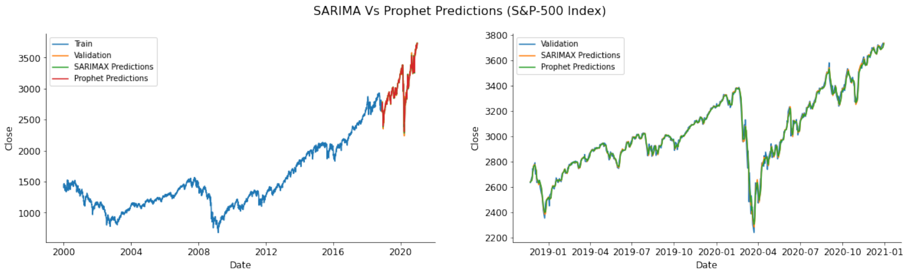


Figure: Prophet and SARIMAX Predictions on S&P-500 Validation Set.

Model	MAE	MSE	RMSE	R2-Score	MAPE
SARIMAX	18.189	755.416	27.484822	0.991685	0.62%
Prophet	18.452	770.787	27.764	0.991516	0.63%

Table: SARIMAX and Prophet Error Metric Values.

Findings from the Study

- The volatility which was obeserved during the Global Financial Crisis of 2008, the same kind of volatility was observed in the COVID Crash of 2020 in both the markets i.e. USA and India.

Findings from the Study

- The volatility which was obeserved during the Global Financial Crisis of 2008, the same kind of volatility was observed in the COVID Crash of 2020 in both the markets i.e. USA and India.
- The markets independent of geography not only tells us the sentiment of the investor in short-term but also gives us an idea about what are the variables which plays a major role like budget annoucements, stimulus plans, FDIs etc to move the market in either directions, but in long term only those stocks/equities/securities performed well which represents high quality businesses.

Findings from the Study

- The volatility which was obeserved during the Global Financial Crisis of 2008, the same kind of volatility was observed in the COVID Crash of 2020 in both the markets i.e. USA and India.
- The markets independent of geography not only tells us the sentiment of the investor in short-term but also gives us an idea about what are the variables which plays a major role like budget annoucements, stimulus plans, FDIs etc to move the market in either directions, but in long term only those stocks/equities/securities performed well which represents high quality businesses.
- Both the markets i.e. Developed & Emerging markets have rebounded quickly i.e. in approximately 7 months after the COVID-19 crash which itself is a thing to discuss because normally it takes atleast 2 years for the markets to recover to its early highs.

Conclusions

- In this analysis, **SARIMAX**$(2,0,0) \times (2,0,0,7)$ for S&P BSE SENSEX and **SARIMAX**$(5,1,1) \times (2,0,1,7)$ for S&P-500 yielded a highly accurate results with a MAPE of 4.05% and 0.61%.

Conclusions

- In this analysis, **SARIMAX**$(2,0,0) \times (2,0,0,7)$ for S&P BSE SENSEX and **SARIMAX**$(5,1,1) \times (2,0,1,7)$ for S&P-500 yielded a highly accurate results with a MAPE of 4.05% and 0.61%.
- Prophet has performed better than both the ARIMA models when forecasting S&P BSE SENSEX and S&P-500 with MAPE of 1.06% and 0.62%.

Conclusions

- In this analysis, **SARIMAX**$(2, 0, 0) \times (2, 0, 0, 7)$ for S&P BSE SENSEX and **SARIMAX**$(5, 1, 1) \times (2, 0, 1, 7)$ for S&P-500 yielded a highly accurate results with a MAPE of 4.05% and 0.61%.
- Prophet has performed better than both the ARIMA models when forecasting S&P BSE SENSEX and S&P-500 with MAPE of 1.06% and 0.62%.
- This model can be used as a techinical indicator of what values the indexes would take in short term in order manage the portfolios to maximize the profits in the market.

Future Work

- In addition to forecasting the closing price, it will also be more strategic if we can also forecast the β value i.e. measure of risk with respect to benchmark indices or broader market indices [CMK18].

Future Work

- In addition to forecasting the closing price, it will also be more strategic if we can also forecast the β value i.e. measure of risk with respect to benchmark indices or broader market indices [CMK18].
- Adding more exogenous variables like P/E ratio, P/B ratio, Market Capitalisation etc as external features in the dataset can help in better forecasting.

References

Eugene F. Fama. "Efficient Capital Markets: A Review of Theory and Empirical Work". In: *The Journal of Finance* 25.2 (1970), pp. 383–417. ISSN: 00221082, 15406261.

Madhavi Latha Challa, Venkataramanaiah Malepati, and Siva Nageswara Rao Kolusu. "Forecasting risk using auto regressive integrated moving average approach: an evidence from SP BSE Sensex". In: *Financial Innovation* 4.3 (2018). DOI: 10.1186/s40854-018-0107-z.

Madhavi Latha Challa, Venkataramanaiah Malepati, and Siva Nageswara Rao Kolusu. "SP BSE Sensex and SP BSE IT return forecasting using ARIMA". In: *Financial Innovation* 6.1 (2020).

Thank You

(The sildes were created using LaTeX.)

For Interactive Plots refer Google Colab Notebook

Notebook Link: https://tinyurl.com/uvs7drpa