Forecasting India S&P BSE SENSEX and USA S&P-500 Benchmark Indices Using SARIMAX and Facebook Prophet Library

Authors: Suraj Prakash Sharma[1], Jeyanthi R[2], and K.Deepa[3]
Presenter: Suraj Prakash Sharma

[1]Department of Computer Science & Engineering, [2]Department of Electronics & Communication Engineering, and [3]Department of Electrical & Electronics Engineering
Amrita School of Engineering, Bangalore, Amrita Vishwa Vidyapeetham, India

ICICCS Conference Presentation, May 26, 2022

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.

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/organization wants to build an investment strategy/policies [1].

Literature Review

- Forecasting stock market returns plays a pivotal role whenever an investor/investment firm/organization wants to build an investment strategy/policies [1].
- Recent computational advancements have led to various econometric models which have been used consistently to anticipate market movements/irregularities/risks and thus forecast the future prices/returns [2] [3].

Literature Review

- Forecasting stock market returns plays a pivotal role whenever an investor/investment firm/organization wants to build an investment strategy/policies [1].
- Recent computational advancements have led to various econometric models which have been used consistently to anticipate market movements/irregularities/risks and thus forecast the future prices/returns [2] [3].
- ARIMA and its variants are a 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 used ARIMA or variants of ARIMA models to forecast econometric variables like GDP, CPI, WPI, or price of a listed financial asset etc.

Literature Review

- Forecasting stock market returns plays a pivotal role whenever an investor/investment firm/organization wants to build an investment strategy/policies [1].
- Recent computational advancements have led to various econometric models which have been used consistently to anticipate market movements/irregularities/risks and thus forecast the future prices/returns [2] [3].
- ARIMA and its variants are a 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 used ARIMA or variants of ARIMA models to forecast econometric variables like GDP, CPI, WPI, or price of a listed financial asset etc.
- The Jenkins ARIMA approach is more efficient than other econometric models, which are based on regression and exponential smoothing.

Different Types of Markets

Weak, Semi-Strong and Strong Markets

- There are various studies done by prominent economists like Paul Samuelson, Mandelbrot about the nature of the market, but it was Eugene Fama [4] 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 Hypothesis.

Different Types of Markets

Weak, Semi-Strong and Strong Markets

- There are various studies done by prominent economists like Paul Samuelson, Mandelbrot about the nature of the market, but it was Eugene Fama [4] 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 Hypothesis.
- Efficient Market Hypothesis (EMH): It states that the price of any financial asset/product at any time reflects all the public and private information which the market has processed. The direct conclusion of EMH is that it is impossible to beat the market, i.e. generating alpha (α) consistently is impossible.

- There are various studies done by prominent economists like Paul Samuelson, Mandelbrot about the nature of the market, but it was Eugene Fama [4] 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 Hypothesis.
- Efficient Market Hypothesis (EMH): It states that the price of any financial asset/product at any time reflects all the public and private information which the market has processed. The direct conclusion of EMH is that it is impossible to beat the market, i.e. generating alpha (α) consistently is impossible.
- According to EMH there are 3 types of tests proposed to categorize the markets: **Weak Form, Semi-Strong, and Strong Form**.

Datasets & Methodology

- The datasets are collected using either the python library or by Yahoo Finance website as per table 1.

Table: Source of Datasets

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

Datasets & Methodology

- The datasets are collected using either the python library or by Yahoo Finance website as per table 1.

Table: Source of Datasets

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

- India & CBOE VIX indexes dataset 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.

Descriptive Statistics of S&P BSE SENSEX

Table: Descriptive Statistics of S&P BSE SENSEX Daily Close and Daily Returns

Sr.No	Statistics	Daily Close	Daily Returns
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)

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

Descriptive Statistics of S&P-500 Index

Table: Descriptive Statistics of S&P-500 Daily Close and Daily Returns

Sr.No	Statistics	Daily Close	Daily Returns
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)

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

Daily Returns Plots for S&P BSE SENSEX and S&P-500

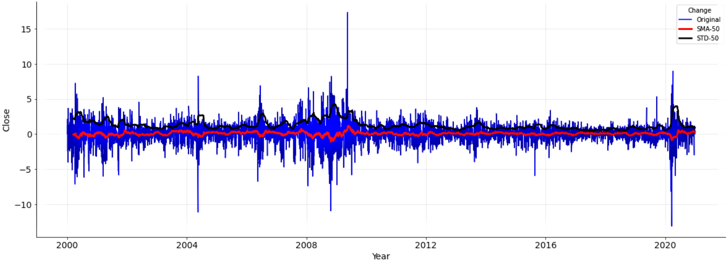

Figure: S&P BSE SENSEX Daily Returns Plot.

Daily Returns Plots for S&P BSE SENSEX and S&P-500

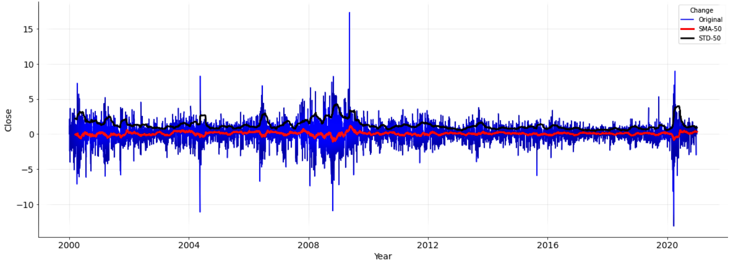

Figure: S&P BSE SENSEX Daily Returns Plot.

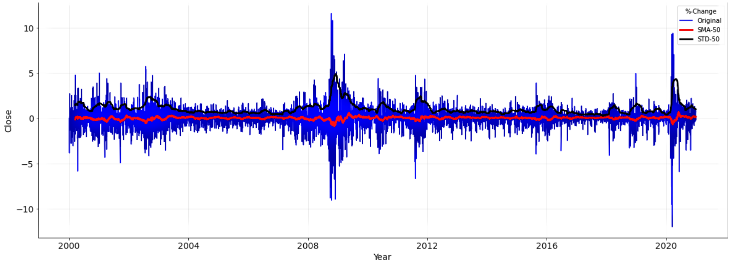

Figure: S&P-500 Daily Returns Plot.

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

- The daily returns 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 obtained is a **White Noise** with approximately 0 mean (μ) and constant standard deviation (σ) $\implies$ the given time series data of S&P BSE SENSEX and S&P-500 is a **Random Walk** process hence the given time series is **non-stationary**.

INDIA VIX Index

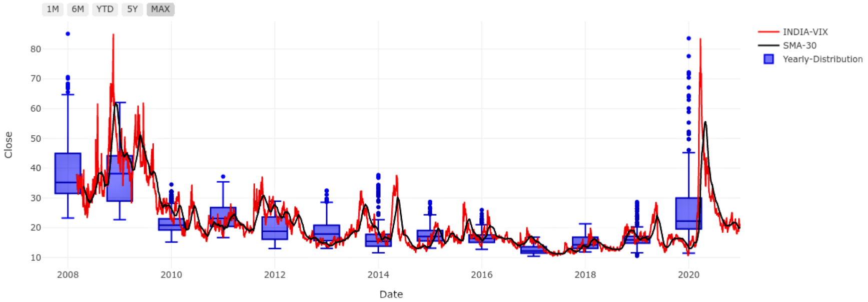

Figure: INDIA VIX (2008-2020)

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

CBOE VIX Index

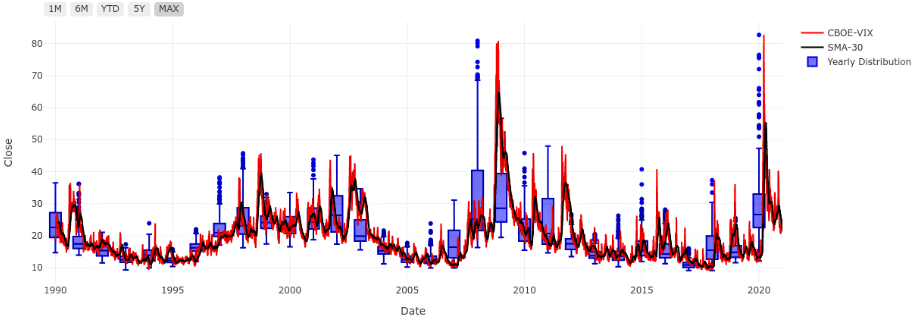

Figure: CBOE VIX Index

- 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 because in the FY-2009, Global Financial Crisis happened due to crash of Mortgage market in the United States and in FY-2021 COVID-19 crash happened due to great lockdown and the panic of recession.

ADF Test Results for S&P BSE SENSEX Index

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

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

- 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

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

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

- 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**.

Table: S&P BSE SENSEX Daily Returns ADF Test Results

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

- 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

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

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

- 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

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

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

- 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**.

Table: S&P-500 Daily Returns ADF Test Results

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

- 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 SARIMAX Model & Forecasting Using SARIMAX

```python
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

```python
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 [5]

Parameters of SARIMAX Model for S&P BSE SENSEX and S&P-500 Indexes

- 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 stands for exogenous features presence.

Parameters of SARIMAX Model for S&P BSE SENSEX and S&P-500 Indexes

- 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 stands for exogenous 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 $= 63798.806$ for S&P BSE SENSEX.

Parameters of SARIMAX Model for S&P BSE SENSEX and S&P-500 Indexes

- 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 stands for exogenous 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 $= 63798.806$ for S&P BSE SENSEX.
- 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 $= 35701.260$ for S&P-500.

SARIMAX and Prophet Predictions Combined for S&P BSE SENSEX

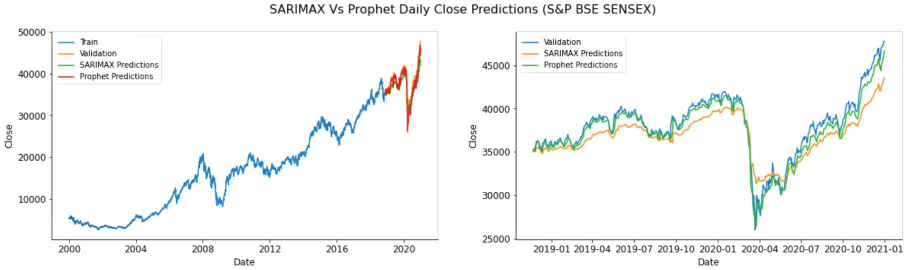

Figure: Prophet and SARIMAX Predictions on S&P BSE SENSEX Validation Set

SARIMAX and Prophet Predictions Combined for S&P BSE SENSEX

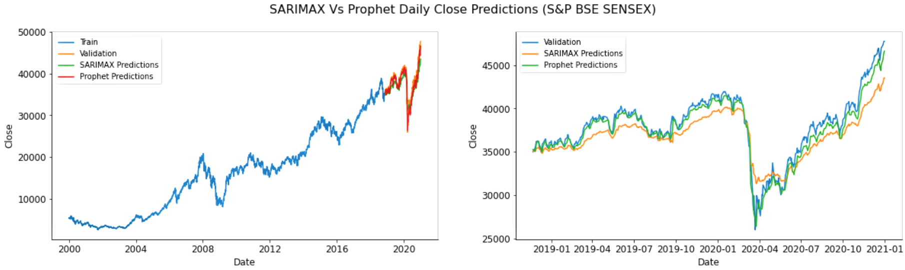

Figure: Prophet and SARIMAX Predictions on S&P BSE SENSEX Validation Set

Table: SARIMAX and Prophet Error Metric Values

Model	**MAE**	**MSE**	**RMSE**	**R^2-Value**	**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%

SARIMAX and Prophet Predictions Combined for S&P-500

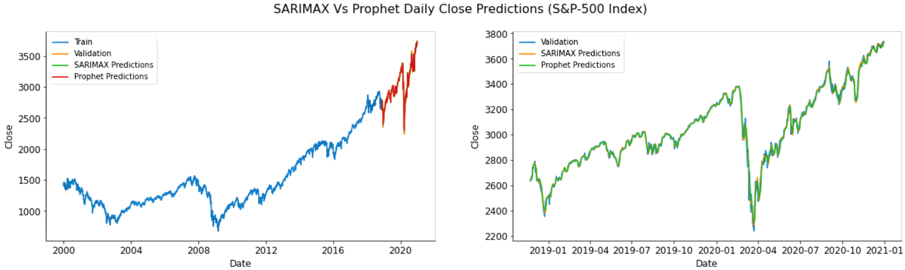

Figure: Prophet and SARIMAX Predictions on S&P-500 Validation Set

SARIMAX and Prophet Predictions Combined for S&P-500

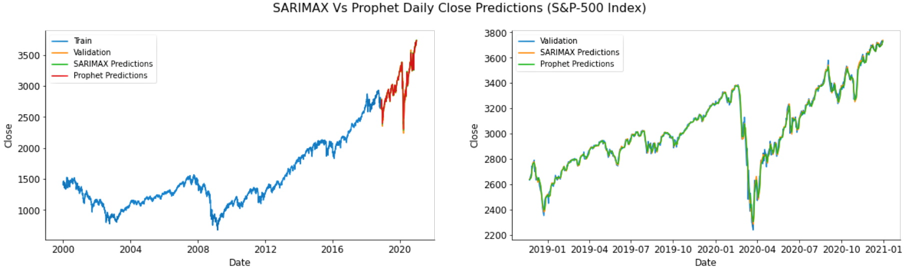

Figure: Prophet and SARIMAX Predictions on S&P-500 Validation Set

Table: SARIMAX and Prophet Error Metric Values

Model	**MAE**	**MSE**	**RMSE**	**R^2-Value**	**MAPE**
SARIMAX	18.189	755.416	27.484822	0.991685	0.62%
Prophet	18.452	770.787	27.764	0.991516	0.63%

Results & Conclusions

- SARIMAX $(2, 0, 0)(2, 0, 0, 7)$ for India benchmark index and SARIMAX $(5, 1, 1)(2, 0, 1, 7)$ for the USA benchmark index yielded reliable results with a R^2 values of 0.731307 and 0.991685 and mean average precision error (MAPE) of 4.05% and 0.61%.

Results & Conclusions

- SARIMAX $(2,0,0)(2,0,0,7)$ for India benchmark index and SARIMAX $(5,1,1)(2,0,1,7)$ for the USA benchmark index yielded reliable results with a R^2 values of 0.731307 and 0.991685 and mean average precision error (MAPE) of 4.05% and 0.61%.
- **Prophet has outperformed the corresponding SARIMAX models** when forecasting the underlying India and USA benchmark indexes, with R^2 values of 0.952397 and 0.991516, respectively, and mean average precision errors (MAPE) of 1.06% and 0.62%.

Results & Conclusions

- SARIMAX $(2, 0, 0)(2, 0, 0, 7)$ for India benchmark index and SARIMAX $(5, 1, 1)(2, 0, 1, 7)$ for the USA benchmark index yielded reliable results with a R^2 values of 0.731307 and 0.991685 and mean average precision error (MAPE) of 4.05% and 0.61%.
- **Prophet has outperformed the corresponding SARIMAX models** when forecasting the underlying India and USA benchmark indexes, with R^2 values of 0.952397 and 0.991516, respectively, and mean average precision errors (MAPE) of 1.06% and 0.62%.
- The models, **SARIMAX and Prophet, can capture the assorted trends and volatility present in both the indexes historical data** during the training phase, with **Prophet being better and more reliable than SARIMAX when tested on a validation dataset.**

Results & Conclusions

- SARIMAX $(2,0,0)(2,0,0,7)$ for India benchmark index and SARIMAX $(5,1,1)(2,0,1,7)$ for the USA benchmark index yielded reliable results with a R^2 values of 0.731307 and 0.991685 and mean average precision error (MAPE) of 4.05% and 0.61%.
- **Prophet has outperformed the corresponding SARIMAX models** when forecasting the underlying India and USA benchmark indexes, with R^2 values of 0.952397 and 0.991516, respectively, and mean average precision errors (MAPE) of 1.06% and 0.62%.
- The models, **SARIMAX and Prophet, can capture the assorted trends and volatility present in both the indexes historical data** during the training phase, with **Prophet being better and more reliable than SARIMAX when tested on a validation dataset.**
- These models can be used as a technical indicator of what values the indexes would take in short term according to forecasted trends in order to optimize the portfolios to maximize the profits in the market.

Findings from the Study

- The volatility which was observed 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 observed 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 investors in short-term but also gives us an idea about what are the variables which plays a major role like budget announcements, 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 observed 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 investors in short-term but also gives us an idea about what are the variables which plays a major role like budget announcements, 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 at least 2 years for the markets to recover to its early highs.

References

[1] V. Malepati, M. L. Challa, and S. N. R. Kolusu, "S&p bse sensex and s&p bse it return forecasting using arima," *Financial Innovation*, vol. 6, no. 1, 2020.

[2] M. L. Challa, V. Malepati, and S. N. R. Kolusu, "Forecasting risk using autoregressive integrated moving average approach: An evidence from s&p bse sensex," *Financial Innovation*, vol. 4, no. 3, 2018.

[3] G. Parambalath, M. E., B. P., *et al.*, "Big data analytics: A trading strategy of nse stocks using bollinger bands analysis," in *Data Management, Analytics and Innovation*, Singapore: Springer Singapore, 2019, pp. 143–154.

[4] E. F. Fama, "Efficient capital markets: A review of theory and empirical work," *The Journal of Finance*, vol. 25, no. 2, pp. 383–417, 1970.

[5] S. J. Taylor and B. Letham, "Forecasting at Scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, Jan. 2018.

Thank You

(The slides were created using LaTeX.)

For Interactive Plots, refer to Google Colab Notebook

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