DS5220 Project Iteration 05
Market Analysis of S&P 500

Sai Nikhil Kunapareddy & Dilep Shetty Ittanguru Venkatesh

July 2024

1. Introduction to S&P 500 and Time Series Analysis
The S&P 500 index a highly recognized and most sought after equity index in the United States. S&P 500 stands for Standard & Poor's 500, which is collection of top 500 publicly traded organizations in the country. It serves as a very important indicator of the overall performance of the country's economy. We can also gauge the performance of different industries with the help of this metric.

Time series forecasting involves predicting future values with the help of analysing a series of previous values. It is widely used in fields like finance, economics and weather forecasting. This forecasting technique depends on identifying patterns, trends and cycles from previous data and try to replicate it.

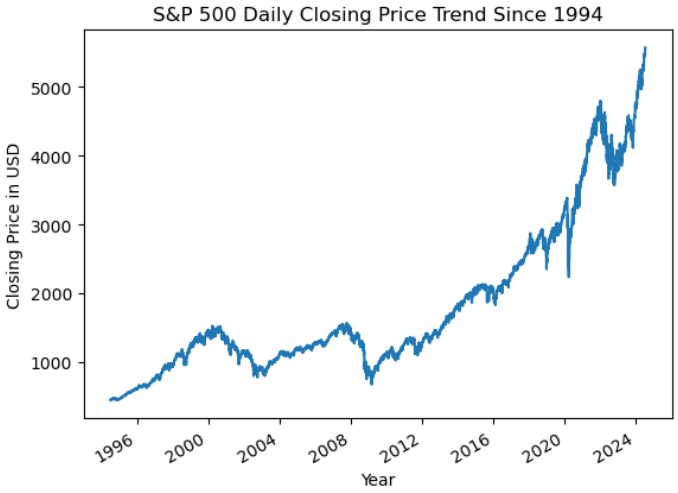
2. Data Sourcing and Handling
The data source for this project has been taken from yahoo finance api. Which provided us with daily stock information (daily low price, high price, closing price) of the requested stock index. The collection method involved calling the api (providing required stock index and time period of interest) in the notebook and storing the information in a pandas dataframe.

Duplicate values and 'NaN' values does not exist in the dataset as it is extracted from a reputed api. All the necessary corrections and precautions are taken care by the developers. Few preliminary checks were done to validate the same.

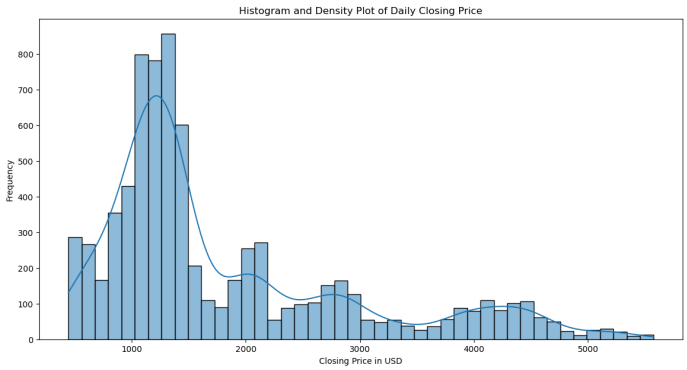
New columns were added into the dataset that contain the calculations of rolling averages and exponential averages for different time windows. In addition to this first order differencing and log transformation of the closing price are also done and added as separate columns for analysis.

3. Exploratory Data Analysis
The closing price of the index since 1994 is plotted to observe different trends over the years. We can correlate this plot with various global events to understand the influence of external factors in our analysis. We can also observe that

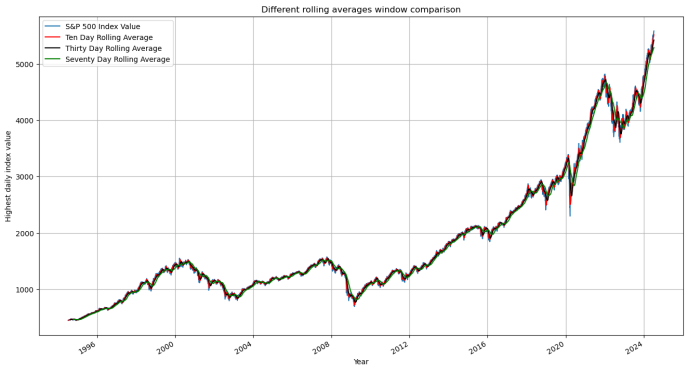
we don't really see constant seasonality or cyclic behaviors in the data. It is to be understood that this index is not solely a function of time.



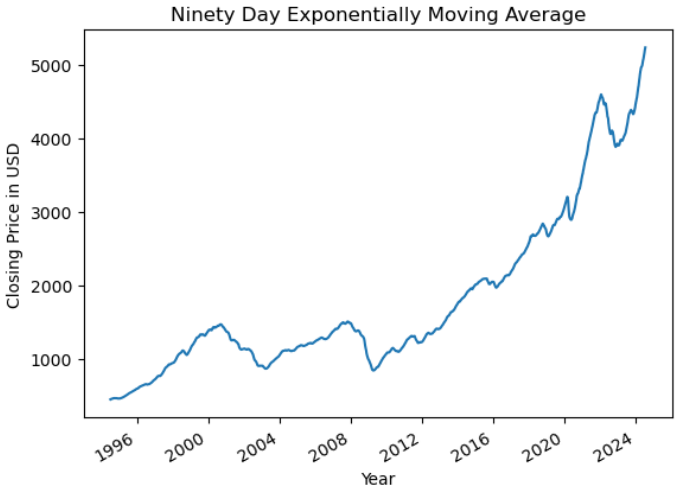
We are interested in looking at the distribution of the closing price of the index. Initially the expectation is to observe a Gaussian distribution, from the generated plot we can see multiple bell curves combined with around four peaks.



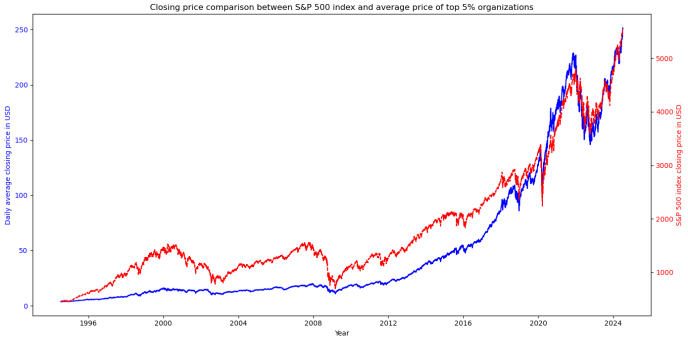
The rolling average of the index closing price for ten, thirty and seventy days is calculated to check the trend capturing ability. As expected the smoothest curve is at seventy days average but a shift is observed because of the window selection.



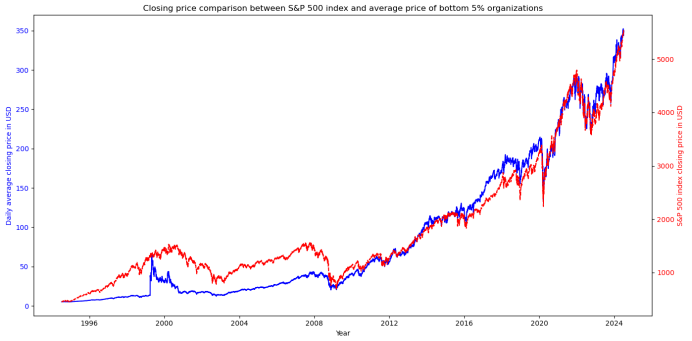
Exponential average is nice alternative to find the trends of the index without the shift observed due to rolling averages.



The influence of the top five percent organizations on the index is observed by calculating the average performance of the top twenty five organizations.



A similar plot is observed for the least five performing companies.



4. Analysis Methodology

ARIMA stands for AutoRegressive Integrated Moving Average. It is widely used for time series forecasting, capturing three aspects: autoregression(AR), differencing(I) and moving averages(MA). Autoregression is the ability of the model to predict with the help of past values. Differencing is a technique used to achieve stationarity (constant mean and standard deviation over time) for the dataset. The dependency between with observed values and residuals for lagged observations is the moving average part of the model.

SARIMA stands for Seasonal AutoRegressive Integrated Moving Average. This is an extension of previously discussed ARIMA model that focuses on seasonality. This model assumes that the data considered has periodic fluctuations.

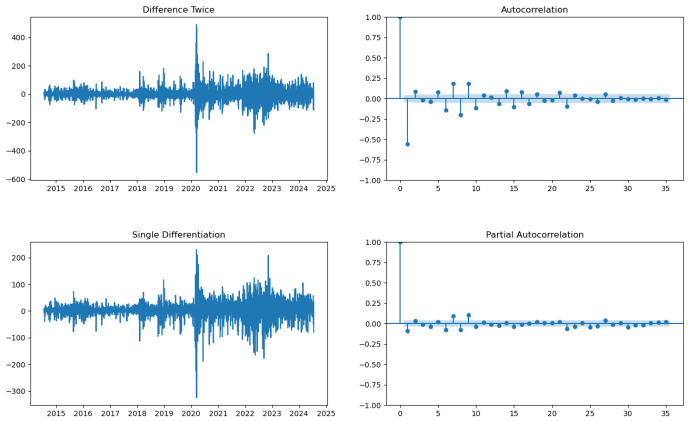
Meta's PROPHET is a time series forecasting model that relies on advanced statistical techniques. It is an open-source tool designed for forecasting time series data, particularly for datasets with daily observations that display patterns of seasonality, holidays, and trends. An API call is used to train and generate predictions.

5. Model Building

Initially an ARIMA model was built by checking for data stationarity. Augmented Dickey-Fuller test is used to judge if the data has stationarity. Following are the results for the data with and without differencing.

Model	ADF Statistic	p-value
Base Model	0.409	0.981
First Differencing	-15	0.000

After the first differencing stationarity is achieved in the model and a model is fit using 7, 1 and 7 as the p, d and q respectively. Autocorrelation and partial autocorrelation plots are included below to support the numerical selection of the parameters.



Automatic grid search is used to find the parameters for SARIMA model. Initially a seasonality window of one year is used to compute the model but it is turned out to be computationally expensive and hence it is dropped down to 90 days.

6. Model Evaluation metrics

Mean absolute error and root mean square error have been the constant model evaluation metrics for all the predictions considered. Mean Absolute Error is a measure of prediction accuracy in a regression model, calculated as the average of the absolute differences between the predicted and actual values. Root Mean Square Error is calculated as the square root of the average of the squared differences between predictions and actual values. These specific metrics are the most popular evaluation criteria for regression models and forecasting, hence they were considered as the first choice. Ideally the values of these metrics are supposed to be as close to zero as possible.
The formula for Mean Squared Error (MSE) is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

The formula for Root Mean Squared Error (RMSE) is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where y_i represents the actual observed value at the i-th instance. $\hat{y}_i$ represents the predicted value at the i-th instance.

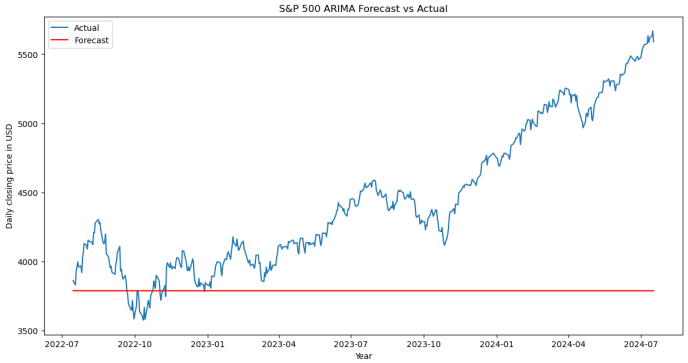
The sole values of these metrics doesn't mean anything due to lack of interpretability. They have to be considered comparatively to evaluate the performance of different models enhancing fair and robust evaluation. But these

Model	MAE	RMSE
ARIMA	657	825
SARIMA	1280	1597
PROPHET	784	955

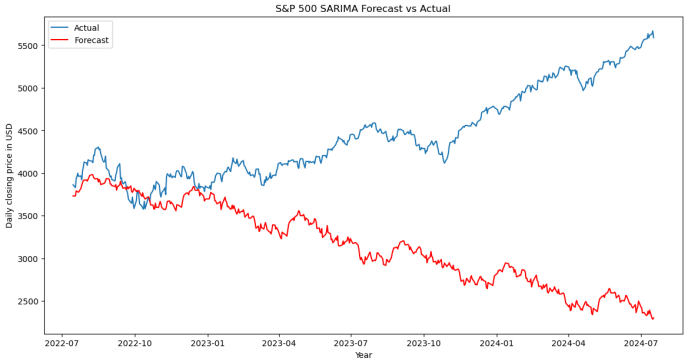
metrics have the advantage of penalizing large error values. Interestingly the least sophisticated model ARIMA has the best numbers, but when plotted PROPHET model seemed to be performing better. The difference in the error terms might be a compounding effect of many small errors due to trend changes.

7. Results and Discussion

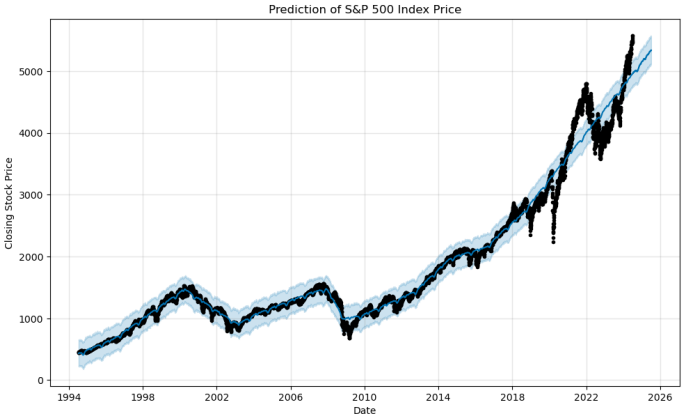
A constant prediction similar to a linear regression model is observed in ARIMA model. This model failed to capture the essence of the upward trend observed in the index curve. Changing the parameter values might improve the model performance but it involves enhanced grid search automation with high computational costs.



SARIMA model was able to capture the seasonality of the data but failed in following the general upward trend of the data. In contrary, it started to diverge and go down resulting in very high error values.



PROPHET model developed by Meta provided the best results by capturing. In contrary to earlier models this model is used to predict for the total dataset range.



Cross validation techniques are not used for any of the considered models as of now.

8. Reflections and Future work

An in-depth understanding of time series forecasting and widely used statistical techniques to achieve the same were attained. Important components and considerations within time series forecasting is looked up while building models. One glaring issue with the considered models is the need of seasonality in data which might not always be the case in stock market.

Model performance of ARIMA and SARIMA models can be further improved by adding hyper parameters. Including deep learining as a problem solution can also provide better results.