**Predicting Stock Growth based on historical data 2015 to 2025**

**ADTA - 5410**

**By**

**Sai Satish Pallapolu**

**Introduction:**

Stock Market analysis has always been the vital area for financial analysis to help investors to make data-driven decisions. Forecasting the investment return allows the investors to analyze risks and maximize their profits. The growth of Artificial Intelligence and Machine learning models has greatly contributed to predictive analysis and now it has become an important tool in understanding stock market fluctuations.

**Problem Statement:**

*Based on historical stock price trends from 2015 to 2025, which industries are expected to demonstrate the most sustained growth beyond 2025?*

This study has significant effects on investors’ investment decisions by predicting the stock market. This project analyzes the 10 years historical Close price of 9 industries such as Technology, Finance, Retail, Healthcare, Automotive, Semiconductors, Ecommerce, Entertainment and Telecommunications. This research provides insights on what industries have shown growth until 2025 and are subject to show growth in future impacting portfolio diversification, industry growth strategies and corporate investments.

**Literature Review:**

Stock Market predictions using Machine Learning models have been extensively studied using various models. The Auto Regressive Integrated Moving Average (ARIMA) model is one of the most widely used models for stock market analysis (Ma, 2020). “The Autoregressive Integrated Moving Average model can be considered as generalization of an Autoregressive Moving Average (ARMA)model”, both the models combine to form ARIMA model (Dhyani et al., 2020). ARIMA models do not assume any underlying relationships unlike other models and the results are purely based on the input variables, t s model has the best capacity to forecast the short-term prediction values (Adebiyi et al., 2014).

**Gaps in Existing Literature:**

The Existing research mainly focus on forecasting the individual companies' stock performance. So, this project aims to fill the gap by analyzing the industry trends for the past 10 years and to forecast the sector’s trends for the next six years. As the ARIMA model has been only used for individual stock performances, there is a lack of studies for implementing ARIMA for sector analysis. This research focuses on addressing those gaps by analyzing the industry trends over time and predicting growth beyond 2025.

**Exploratory Data Analysis Report:**

Exploratory Data Analysis (EDA) is an important step to be followed to understand the structure and characteristics of a dataset before applying the predictive models. In our study, we have analysed stock market data spanning from 2015 to 2025 to uncover the trends, patterns, and useful insights that may contribute to the stock growth prediction. By deriving the statistical summaries of the data, visualizations, and correlation analysis, we can identify the key attributes and prepare the dataset for the forecasting models.

**Data Collection:**

The dataset, "Stocks.csv," was extracted within Excel using stocks under the data tab. 25 organizations were randomly selected for the dataset. The data spans from 2015 to 2025.

**Data Preprocessing:**

Data preprocessing ensures that the collected dataset is clean, structured, and suitable for analysis. The process involves handling the missing values, converting the data types, and extracting relevant features to improve the predictive modelling.

To prepare the data for the predictive model, the 'Date' column is changed into a datetime format and by extracting the 'Year,' 'Month,' and 'Day' components for time-based analysis. The 'Organization' column was split into 'Company' and 'Ticker,' to enhance clarity. A dictionary mapping the companies to their respective industries was created to enable the industry-level analysis. To maintain the data integrity, the 'High' and 'Low' columns were converted to numeric values. Additional attributes were introduced to improve the accuracy, which includes 'Daily Range' (difference between 'High' and 'Low'), 'Net Change' (difference between 'Close' and 'Open'), and 'Percentage Change' (relative daily price change).

To handle the missing values in dataset, the 'High' and 'Daily Range' columns were filled using the median value to ensure consistency without introducing the biases. The 'Volume' column was converted to an integer data type for efficiency. Finally, the duplicate rows were identified and removed to prevent redundancy in analysis. Outlier detection and elimination is a crucial process to ensure the data quality, an IQR-based outlier detection method was applied to key features. Stocks with values falling beyond 1.5 times the interquartile range (IQR) were removed to reduce the impact of extreme fluctuations. This process helps in refining the dataset for more reliable forecasting.

These preprocessing steps refine the dataset, making it more reliable for exploratory data analysis and predictive modelling.

**Data Overview:**

The stocks dataset contains 3001 observations and 13 attributes, the columns are Organization, Close, Open, High, Low, year, month, day, Daily\_range, Net\_change, Percentage\_change, Company and, Ticker. The attributes Daily\_range, Net\_change, and Percentage\_change are included to conduct predictive analysis. The Daily\_range variable shows us the daily high and low margins, Net\_change is calculated by taking the difference between the closing price and the last day’s closing price, and Percentage\_change shows the change from the previous day’s closing price.

**Independent Variables:** Organization, Open, High, Low

**Dependent Variables:** Close**,** Daily\_range, Net\_change, Percentage\_change, Year, Month

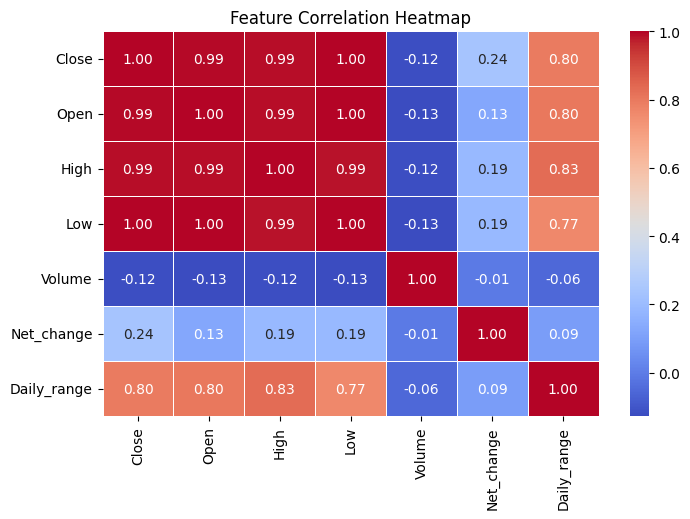
The **“Industry”** field is used to classify the companies based on their sectors. These were grouped by based on their sectors in the dataset are **Automotive, E-Commerce, Entertainment, Finance, Healthcare, Retail, Semiconductors, Technology, and Telecommunications for which** for which stock prices were analysed and used to forecast for future growth.

**Descriptive Statistics:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Statistic | Close | Open | High | Low | Volume | Daily\_range | Net\_change | Percentage\_change | Log\_Close | Log\_Daily\_range | Winsorized\_Close | Winsorized\_Daily\_range |
| count | 2484.00 | 2484.00 | 2484.00 | 2484.00 | 2484.00 | 2484.00 | 2484.00 | 2484.00 | 2484.00 | 2484.00 | 2484.00 | 2484.00 |
| mean | 54.85 | 54.46 | 57.31 | 51.96 | 643105400.00 | 5.35 | 0.39 | 0.77 | 3.63 | 1.61 | 55.06 | 5.38 |
| std | 47.67 | 47.31 | 49.38 | 45.97 | 1846168000.00 | 4.61 | 2.80 | 5.94 | 1.03 | 0.69 | 47.43 | 4.58 |
| min | 0.07 | 0.07 | 0.07 | 0.00 | 9810.00 | 0.00 | -7.00 | -15.06 | 0.07 | 0.01 | 5.18 | 0.63 |
| 25% (Q1) | 23.62 | 23.62 | 24.93 | 22.16 | 99760670.00 | 2.19 | -1.11 | -2.99 | 3.20 | 1.16 | 23.62 | 2.19 |
| 50% (Median) | 42.39 | 42.23 | 44.84 | 38.91 | 192904100.00 | 4.02 | 0.20 | 0.78 | 3.77 | 1.61 | 42.39 | 4.02 |
| 75% (Q3) | 67.96 | 67.71 | 70.81 | 63.91 | 508393300.00 | 7.04 | 1.89 | 4.55 | 4.23 | 2.08 | 67.96 | 7.04 |
| max | 279.08 | 284.03 | 293.07 | 276.16 | 23445520000.00 | 25.58 | 8.09 | 16.35 | 5.64 | 3.28 | 279.08 | 25.58 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |

**Key Insights from EDA:**

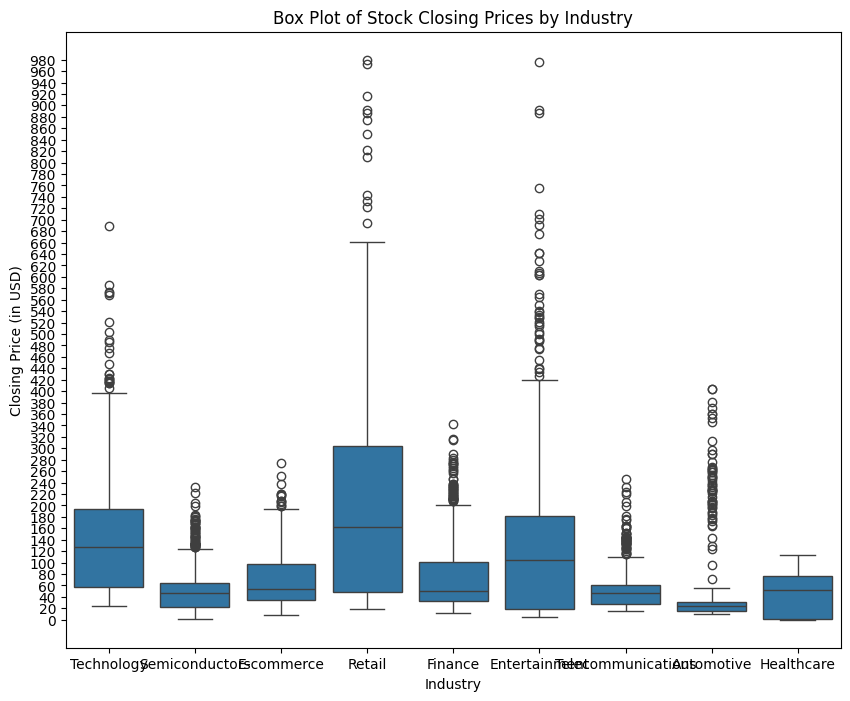
1. **Correlation Analysis**



*Fig1: Correlation Analysis*

The correlation heatmap was generated to understand the relationships between key variables present in the dataset. The heatmap shows a strong positive correlation between 'Close,' 'Open,' 'High,' and 'Low' prices it explains that these stock price metrics move together.

1. **Box Plot of stock closing prices by industry**

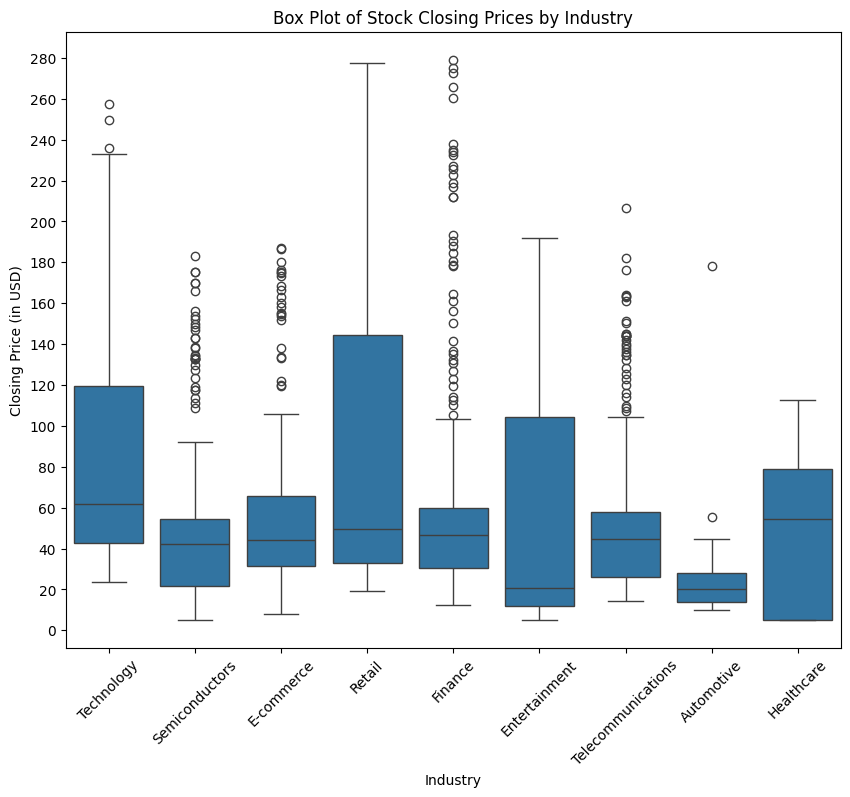


*Fig2: Box Plot before Outliers*

The above box plot analysis of closing prices by industries shows the significant price variations. The Retail and Entertainment industries show higher volatility, with a wider interquartile range and more outliers. Semiconductors and Telecommunications exhibit relatively stable stock prices, making them potential choices for the investors.

1. **Box Plot with Winsorization**

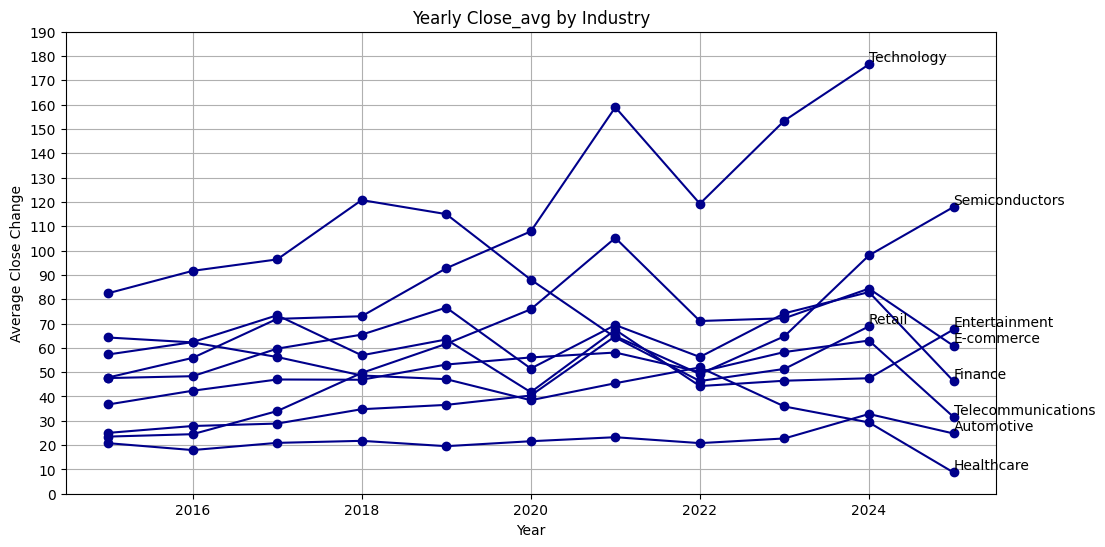
Outlier analysis using IQR revealed a significant number of extreme values in stock price-related features such as **Close, Daily Range, and Net Change**. The application of **Winsorization** effectively capped extreme values at the 5th and 95th percentiles, preserving data integrity while reducing distortion caused by outliers.



*Fig3: Box Plot After Outliers*

The box plot after removing the outliers and refining the data, we can see that the Retail and Entertainment industries shows the highest variability in stock closing prices, with Retail having the highest median. The Automotive stocks have the lowest median and least variability, making them more stable but potentially offering lower short-term gains.

1. **Yearly close\_avg by industry**



*Fig4: Line graph of yearly closing of industries.*

The line chart shows that the Technology and Semiconductors industries have shown significant growth in average closing prices over the years, particularly post-2020. Healthcare and Automotive industries have remained relatively stagnant or declined, suggesting lower volatility but possibly weaker returns.

**Observations and Limitations**

* Needed to evaluate the missing data present in the dataset and evaluate where necessary.
* The potential outliers present in the stock prices require further investigation and needed to be eliminated to maintain data integrity.
* The Retail and Entertainment industries show high volatility, while Semiconductors, Automotive, and Telecommunications display more stability.
* Lack of research paper using ARIMA for sector analysis.

**Methodology:**

This research study makes use of the time series forecasting methodology using **ARIMA (1,1,1)** (Auto **R**egressive **I**ntegrated **M**oving **Average)** for analysing the historic stock price trends from 2015 to 2025 among various industries. The primitive goal of our objective is to identify the industries that has shown stable growth and to forecast their growth for the coming years.

**Justification of Methodology:**

This methodology selected is appropriate because of the following reasons

Using historical data to forecast stock price patterns requires the use of time series forecasting, Linear trends, seasonality, and autoregressive patterns in stock prices are effectively captured in ARIMA models. Also, ARIMA model uses an approach to estimate its parameters and it minimizes the Sum of Squared Errors (SSE).

**Dep/Input Variable:**

**Winsorized\_Close:** The predicted stock price for future dates (selected as the primary dependent variable).

**Analytical Approach**

ARIMA(p, d,q)Model is used here for forecasting the stock prices. The model parameters are adjusted to (1,1,1) after observing the Autocorrelation function and Partial Auto Correlation Function (PACF). The forecast table generated will print all the predicted values for each industry for the upcoming year by year.

The ARIMA model used in this study follows the standard formulation:

**Yt' = c + Σ (i=1 to p) ϕiYt-i' + Σ (j=1 to q) θjϵt-j + ϵt**

Where **Yt':** The time series value at time t.

**Σ (i=1 to p) ϕiYt-i' is** theautoregressive (AR) component

Where **p =** The order of the AR component.

**Φi =** AR coefficients

**Σ (j=1 to q) θjϵt-j** is the MA moving average component

Where **q=** The order of the MA component

**Θj** = MA coefficients

**ϵt:** The current error term.

**Model Evaluation**

The models were assessed by using the following metrics Mean Squared Error (MSE), Mean Absolute Error (MSE) and Mean Absolute Percentage Error (MAPE).

**forecast = model\_fit.forecast(steps=6)**

This model generates predictions for six future points which means for the years like 20205 to 20230. 

**Results:**

Here are the evaluation metrics for each industry:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Industry** | **MAE** | **MSE** | **RMSE** | **MAPE (%)** |
| **Automotive** | 1.88 | 3.53 | 1.88 | 7.58 |
| **E-commerce** | 2.76 | 7.61 | 2.76 | 4.54 |
| **Entertainment** | 17.68 | 312.57 | 17.68 | 26.18 |
| **Finance** | 19.13 | 365.99 | 19.13 | 41.32 |
| **Healthcare** | 4.18 | 17.50 | 4.18 | 47.81 |
| **Retail** | 0.12 | 0.015 | 0.12 | 0.16 |
| **Semiconductors** | 10.59 | 112.23 | 10.59 | 8.98 |
| **Technology** | 9.47 | 89.76 | 9.47 | 5.37 |
| **Telecommunications** | 8.41 | 70.71 | 8.41 | 26.64 |

From the above results we observe that Retail forecast accuracy is the best with the lowest MAE that is 0.12 % and with MAPE is 0.16 %. Also, the Finance and Entertainment show the highest errors with MAPE resulting in poor accuracy. On the other hand, E Commerce and Technology performed well with MAPE (Mean Absolute Percentage Error) having under 6 %.The evaluations conclude that the model is a moderate fit.The below tables show the overall forecast for the future six years starting 2025 to 2030 forecast.

|  |  |
| --- | --- |
| Forecast for Automotive Industry: | |
| Year | **Forecast\_Close** |
| 2025 | 26.658034 |
| 2026 | 26.381952 |
| 2027 | 26.422538 |
| 2028 | 26.416571 |
| 2029 | 26.417449 |
| 2030 | 26.417320 |

|  |  |
| --- | --- |
| Forecast for E-commerce Industry: | |
| Year | **Forecast\_Close** |
| 2025 | 63.5695 |
| 2026 | 63.42893 |
| 2027 | 63.43609 |
| 2028 | 63.43573 |
| 2029 | 63.43574 |
| 2030 | 63.43574 |

|  |  |
| --- | --- |
| Forecast for Entertainment Industry | |
| Year | **Forecast\_Close** |
| 2025 | 49.862869 |
| 2026 | 56.758074 |
| 2027 | 54.068886 |
| 2028 | 55.117692 |
| 2029 | 54.708649 |
| 2030 | 54.868179 |

|  |  |
| --- | --- |
| Forecast for Finance Industry | |
| Year | **Forecast\_Close** |
| 2025 | 65.43077 |
| 2026 | 63.3518 |
| 2027 | 63.57772 |
| 2028 | 63.55317 |
| 2029 | 63.55584 |
| 2030 | 63.55555 |

|  |  |
| --- | --- |
| Forecast for Healthcare Industry | |
| Year | **Forecast\_Close** |
| 2025 | 4.566295 |
| 2026 | 0.382723 |
| 2027 | -3.80071 |
| 2028 | -7.98402 |
| 2029 | -12.1672 |
| 2030 | -16.3502 |

|  |  |
| --- | --- |
| Forecast for Retail Industry | |
| Year | **Forecast\_Close** |
| 2025 | 77.10196 |
| 2026 | 77.22325 |
| 2027 | 77.22503 |
| 2028 | 77.22505 |
| 2029 | 77.22505 |
| 2030 | 77.22505 |

|  |  |
| --- | --- |
| Forecast for Semiconductors Industry | |
| Year | **Forecast\_Close** |
| 2025 | 128.504 |
| 2026 | 138.4554 |
| 2027 | 147.8031 |
| 2028 | 156.5839 |
| 2029 | 164.832 |
| 2030 | 172.5799 |

|  |  |
| --- | --- |
| Forecast for Technology Industry | |
| Year | **Forecast\_Close** |
| 2025 | 186.0416 |
| 2026 | 195.5152 |
| 2027 | 204.988 |
| 2028 | 214.4599 |
| 2029 | 223.9311 |
| 2030 | 233.4014 |

|  |  |
| --- | --- |
| Forecast for Telecommunications Industry | |
| Year | **Forecast\_Close** |
| 2025 | 23.15111 |
| 2026 | 27.45722 |
| 2027 | 25.2521 |
| 2028 | 26.38132 |
| 2029 | 25.80306 |
| 2030 | 26.09918 |

**Division of work:**

|  |  |
| --- | --- |
| **Team Member** | **Role (Contribution Plan)** |
| **Venkata Rajyalakshmi Gudala** | **Descriptive Statistics, Visualizing, Documentation regarding the above** |
| **Sai Satish Pallapolu** | **Methodology, Predictive Modelling and Model Analysis, Documentation regarding the above** |
| **Indusree Terala** | **Data Collection, Research question, EDA, Data cleaning, Insight generation, Documentation of Literature Review.** |

**References:**

Adebiyi, A. A., Adewumi, A. O., & Ayo, C. K. (2014). Comparison of ARIMA and artificial

neural networks models for stock price prediction. *Journal of Applied Mathematics*, *2014*, 1–7.

<https://doi.org/10.1155/2014/614342>

Dhyani, B., Kumar, M., Verma, P., & Jain, A. (2020). Stock Market Forecasting Technique using

Arima Model. *International Journal of Recent Technology and Engineering (IJRTE)*, *8*(6), 2694

2697. <https://doi.org/10.35940/ijrte.f8405.038620>

Ma, Q. (2020). Comparison of ARIMA, ANN and LSTM for stock price prediction. *E3S Web of*

*Conferences*, *218*, 01026. <https://doi.org/10.1051/e3sconf/202021801026>

Noor Zaman, Mustansar Ali Ghazanfar, Madiha Anwar, et al. Stock market prediction based on

machine learning and social sentiment analysis. TechRxiv. March 27, 2023.

<https://www.techrxiv.org/doi/full/10.36227/techrxiv.22315069>