Dr. D. Y. PATIL VIDYAPEETH, PUNE
(Deemed to be University)

DR. D. Y. PATIL SCHOOL OF SCIENCE AND TECHNOLOGY

TATHAWADE, PUNE

A Foundation of Data Science-Report on

S&P-500 Index Price Prediction.

SUBMITTED BY:

NAME OF STUDENT	ROLL NUMBER
1.TEJAS BHAVSAR	BTAI-204
2. ROHAN MASHERE	BTAI-228
3.PRATHMESH NANGARE	BTAI-230

GUIDED BY:

Dr. Mily Lal

ARTIFICIAL INTELLIGENCE & DATA SCIENCE

ACADEMIC YEAR 2024-2025

Dr. D. Y. PATIL VIDYAPEETH, PUNE
(Deemed to be University)

DR. D. Y. PATIL SCHOOL OF SCIENCE AND TECHNOLOGY

TATHAWADE, PUNE

CERTIFICATE

This is to certify that the Foundation of Data Science-Report entitled.

S&P-500 Index Price Prediction.

is a Bonafide work carried out by **Mr. Tejas Bhavsar** under the supervision of **Dr. Mily Lal** and it is submitted towards the partial fulfillment of the requirement Foundation of Data Science.

Dr. Mily Lal
Project Guide

Prof. Manisha Bhende
Director I/C

ARTIFICIAL INTELLIGENCE & DATA SCIENCE

ACADEMIC YEAR 2024-2025

Dr. D. Y. PATIL VIDYAPEETH, PUNE
(Deemed to be University)

DR. D. Y. PATIL SCHOOL OF SCIENCE AND TECHNOLOGY

TATHAWADE, PUNE

CERTIFICATE

This is to certify that the Foundation of Data Science-Report entitled.

S&P-500 Index Price Prediction.

is a Bonafide work carried out by **Mr. Rohan Mashere** under the supervision of **Dr. Mily Lal** and it is submitted towards the partial fulfillment of the requirement Foundation of Data Science.

Dr. Mily Lal
Project Guide

Prof. Manisha Bhende
Director I/C

ARTIFICIAL INTELLIGENCE & DATA SCIENCE

ACADEMIC YEAR 2024-2025

Dr. D. Y. PATIL VIDYAPEETH, PUNE
(Deemed to be University)

DR. D. Y. PATIL SCHOOL OF SCIENCE AND TECHNOLOGY

TATHAWADE, PUNE

CERTIFICATE

This is to certify that the Foundation of Data Science-Report entitled.

S&P-500 Index Price Prediction.

is a Bonafide work carried out by **Mr. Prathmesh Nangare** under the supervision of **Dr. Mily Lal** and it is submitted towards the partial fulfillment of the requirement Foundation of Data Science.

Dr. Mily Lal
Project Guide

Prof. Manisha Bhende
Director I/C

ARTIFICIAL INTELLIGENCE & DATA SCIENCE

ACADEMIC YEAR 2024-2025

ABSTRACT

Predicting stock prices is a significant challenge in finance and a popular application of machine learning. This project aims to forecast the S&P 500 index using historical data and various machine learning models, focusing on technical and fundamental analysis. Stock prices are influenced by multiple factors, including market conditions, industry trends, leadership changes, and acquisitions. While sentiment analysis from news and social media can improve predictions, this study relies solely on historical price data and key technical indicators. Data from the last 20 years was extracted using the Yahoo Finance API, along with five major tech stocks (Apple, Amazon, Microsoft, Netflix, and Google) for comparison. Technical indicators such as future price change, simple moving averages, relative strength index, and exponential moving averages were incorporated as features. Several Machine learning models, including Decision Trees, Random Forests, Gradient Boosting, Neural Networks, and Facebook's Prophet, were tested. Traditional models struggled with predictive accuracy, while Facebook's Prophet showed promising results, forecasting an upward trend in the S&P 500 over the next 2–3 years. This study highlights the potential of machine learning in stock price prediction while acknowledging its limitations, suggesting that time-series forecasting models like Prophet may be more effective. Future work could enhance predictions by integrating sentiment analysis and macroeconomic indicators.

Keywords: Stock Price Prediction, S&P-500, Machine Learning, Decision Trees, Random Forests, Gradient Boosting, Neural Networks, Facebook's Prophet, Historical Data, Market Trends.

INDEX

List of Figures

List of Tables

Chapter 1

INTRODUCTION

Stock market prediction is a crucial area of research in financial analytics and machine learning. The S&P-500 index, which tracks the performance of all large-cap 500 U.S. companies, serves as a benchmark for market trends and investor sentiment. Predicting the future prices of this index presents a challenging problem due to the complex and dynamic nature of financial markets. In this study, historical stock market data was obtained using the Yahoo Finance API, covering 20 years of S&P 500 prices to analyze long-term trends. From 5 years of stock data for the Big-5 technology companies like Apple, Amazon, Microsoft, Netflix, and Google were collected to investigate their relationship with the S&P-500 index. Initial exploratory data analysis revealed a strong correlation of over 90% between these technology stocks and the index, indicating that their price movements are largely in sync. Further analysis of daily returns showed that the S&P-500 generally fluctuates within the range of -4% to +4%, with extreme losses observed during financial crises, such as the -11.98% drop in March 2020 due to the COVID-19 Crises. The average daily percentage change over the last 20 years was found to be 0.033%, reinforcing the notion that the S&P-500 index exhibits positive returns in the long run.

To improve predictive performance, several technical indicators were introduced as features, including Future Price Change, Simple Moving Averages (SMA), Relative Strength Index (RSI), and Exponential Moving Averages (EMA). SMA captures long-term trends by averaging past prices, RSI measures price momentum and identifies overbought/oversold conditions, while EMA assigns greater importance to recent price changes. Using these features, five machine learning models were implemented to forecast the S&P-500 price: Decision Trees, Random Forests, Gradient Boosting, Neural Networks, and Facebook's Prophet. Decision Trees suffered from overfitting, performing well on training data but poorly on test data. Random Forests showed a slight improvement but still lacked sufficient predictive power. Gradient Boosting improved test accuracy but was not sufficient for reliable S&P-500 price predictions. Neural Networks, though generally powerful, did not perform well in this study due to basic model tuning and optimization. Finally, Facebook's Prophet emerged as the best-performing model, effectively capturing seasonality and long-term trends, and its forecast indicated that the S&P-500 index is likely to rise in the next 2-3 years. This study highlights both the potential and challenges of using machine learning for stock market forecasting. While historical prices and technical indicators provide valuable insights, stock prices are also influenced by macroeconomic conditions, industry-specific trends, corporate leadership changes, and global events. Future research could enhance prediction accuracy by incorporating sentiment analysis through news articles and social media data. Despite some limitations, the results highlight the S&P-500's long-term growth trend and show that machine learning models, especially Facebook's Prophet, can offer useful forecasting insights.

1.1 PROBLEM STATEMENT

Stock prices are influenced by various factors, making their prediction challenging. This project aims to forecast S&P-500 stock prices using historical data and machine learning models, to identify potential trends and improve predictive accuracy.

1.2 OBJECTIVE

This project aims to develop a predictive model using historical S&P 500 data to forecast stock price trends. The primary goal is to leverage past market behavior and identify patterns that can inform future movements. A key part of the analysis involves examining the impact of technical indicators such as moving averages, Relative Strength Index (RSI), and exponential moving averages. These indicators serve as crucial features in training the predictive models. Various machine learning algorithms will be explored, including Decision Trees, Random Forests, Gradient Boosting, Neural Networks, and Facebook's Prophet. The models will be trained and validated on structured historical data without incorporating market sentiment or unstructured inputs. Their performance will be compared using metrics like accuracy, confusion matrix, and ROC curves. The objective is to determine which model performs best under given conditions. Insights from the analysis will guide improvements in model selection and feature engineering. Ultimately, the project seeks to enhance forecasting accuracy for better financial decision-making.

1.3 SCOPE

In this project, we focus on evaluating various machine learning methods to predict the S&P 500 index based on historical stock data. The goal is to identify which models perform best using past trends and technical indicators. However, I will not consider market sentiment in this analysis. Sentiment analysis typically involves examining news articles, political events, market emotions, and social trends. These factors require natural language processing and access to unstructured data. To keep the project manageable and objective, I am limiting the scope to structured historical data only. This ensures a focused, data-driven approach to predicting market performance. The dataset includes features like RSI, EMA, SMA, and trading volumes. I evaluate models such as decision trees, random forests, gradient boosting, and Facebook's Prophet using metrics like ROC curves, accuracy, and confusion matrices. The aim is to determine which approach offers the most reliable insight into future market movements.

1.4 SYSTEM ARCHITECTURE

The stock price prediction process involves collecting historical data (prices, volumes) and engineering features (e.g., moving averages). Data is normalized and split into training/testing sets. Models like decision trees, random forests, gradient boosting, neural networks, or Prophet are trained and evaluated using metrics (MAE, RMSE). The best model generates future price forecasts and is deployed for real-time predictions. While helpful, stock markets are volatile, so predictions should be used cautiously with risk management strategies. Continuous model updates improve accuracy over time.

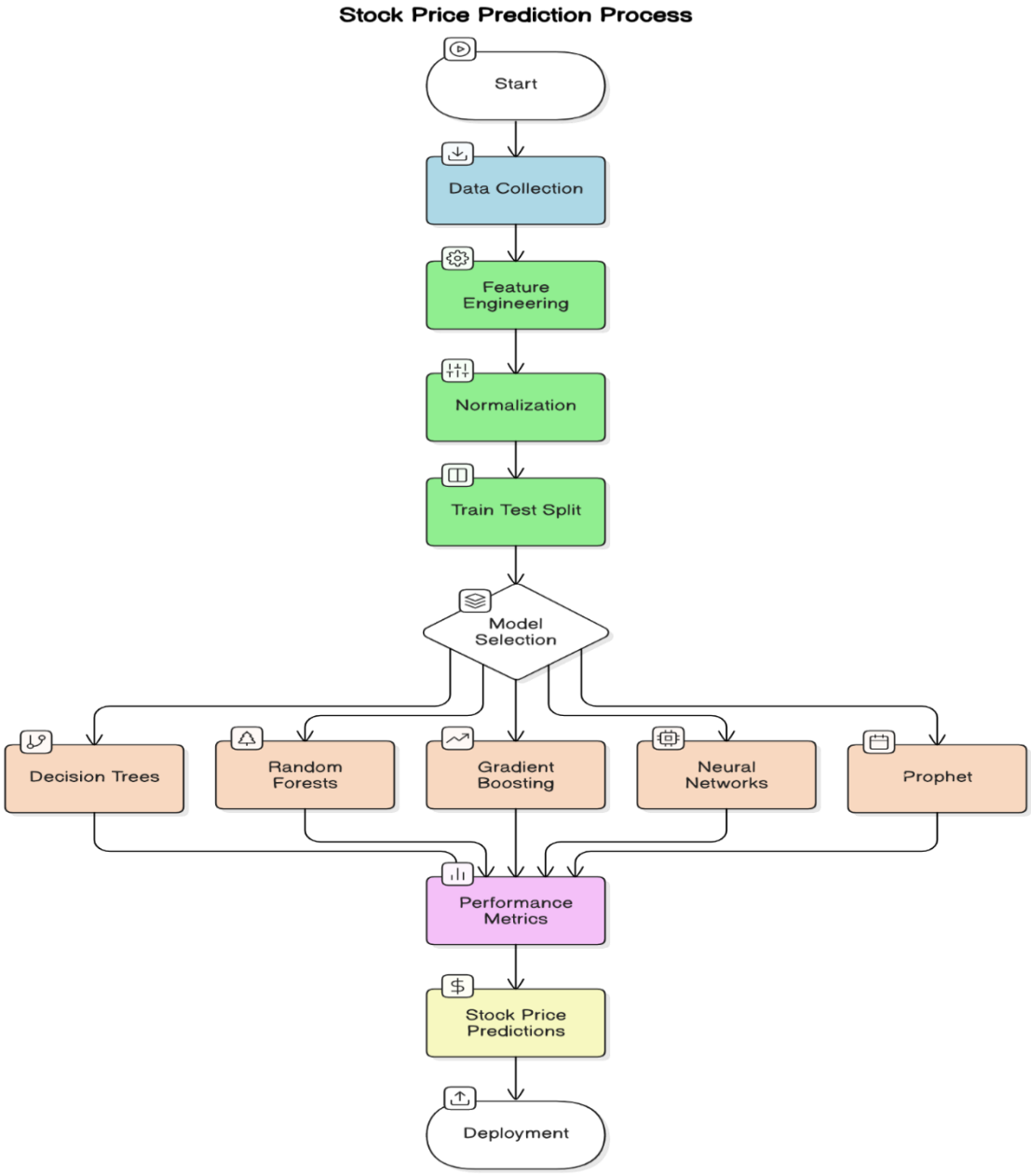

Fig.1 System Architecture

Chapter 2

DATA COLLECTION & PREPROCESSING

2.1 DATASET

The dataset for this project was obtained from the Yahoo Finance API, which provides historical stock market data dating back to the 1990s. Using the Yahoo Finance API, covering 20 years of S&P-500 stock price data were collected. Additionally, stock prices of five major technology companies (Apple, Amazon, Microsoft, Netflix, and Google) from the last five years were extracted for comparison. This dataset includes key stock market features such as Open, Close, High, Low, Volume, and Adjusted Close, which are essential for analyzing price trends and market behavior. This data is highly relevant for stock price prediction as it captures long-term market trends and daily fluctuations. The inclusion of major tech stocks allows for correlation analysis with the S&P-500, helping to identify patterns and dependencies. By leveraging historical data, technical indicators, and machine learning models, this dataset serves as a strong foundation for predicting future stock prices.

Features available in the dataset:

Table 1: Features of Dataset

Variable name	Description
Date	Date of the stock trade
High	Highest price of the stock on the day
Low	Lowest price of the stock on the day
Open	Opening price of the stock on the day
Close	Closing price of the stock on the day
Volume	Volume (number of shares exchanged) on the day
Adj Close	Adjusted closing price of the stock on the day. Adjusted price accounts into dividends and other price adjustments happened on the day. So that's why this variable is used as the final price of the stock.

2.2 DATA PREPROCESSING

Data preprocessing is a critical step before applying machine learning models, as it ensures data quality, consistency, and suitability for analysis. The preprocessing steps involve data cleaning, handling missing values, checking data types, computing summary statistics, feature scaling, and exploratory data analysis (EDA).

2.2.1 Handling Missing Values and Data Types:

Before feeding the dataset into the models, it is essential to check for missing values and incorrect data types. The dataset was checked using data_500.isnull().sum(), confirming that no missing values were present in key columns such as Date, Open, High, Low, Close, Volume, and Adjusted Close. Ensuring the correct data types is also crucial since numerical values should be in a format suitable for mathematical computations.

2.2.2 Data Cleaning and Transformation:

Since raw financial data often contains unnecessary features or outliers, irrelevant columns were removed, and inconsistencies were addressed. Adjustments were made for stock splits and dividend payouts by using the Adjusted Close price instead of the raw closing price.

2.2.3 Feature Scaling and Normalization:

Financial data varies significantly in scale (e.g., stock prices in thousands while volume in millions). To ensure fair weight distribution, Min-Max Scaling and Standardization techniques were applied where necessary. This helps machine learning models converge faster and perform better.

2.2.4 Exploratory Data Analysis (EDA):

EDA was performed to understand the dataset's characteristics and detect potential patterns. Key steps included:

In this project, summary statistics such as mean, median, variance, and standard deviation were computed to measure variations in stock prices and gain a foundational understanding of market behaviour. Various visualization techniques were employed to explore the distribution of daily stock returns, including plotting daily simple returns for clearer insights. Correlation analysis was conducted to examine the relationship between the S&P-500 index and the Big 5 tech stocks Apple, Amazon, Microsoft, Netflix, and Google revealing a strong positive correlation among them. Additionally, annualized returns were visualized through bar charts to compare the total returns of these major tech stocks with the S&P 500. To assess market volatility, rolling standard deviation was calculated, offering a dynamic view of price fluctuations over time. Furthermore, technical analysis tools such as the 14-day and 200-day simple moving averages, along with the Exponential Moving Average (EMA), were utilized to identify and interpret long-term trends in stock prices.

- Daily Returns of S&P 500 (Simple Returns):

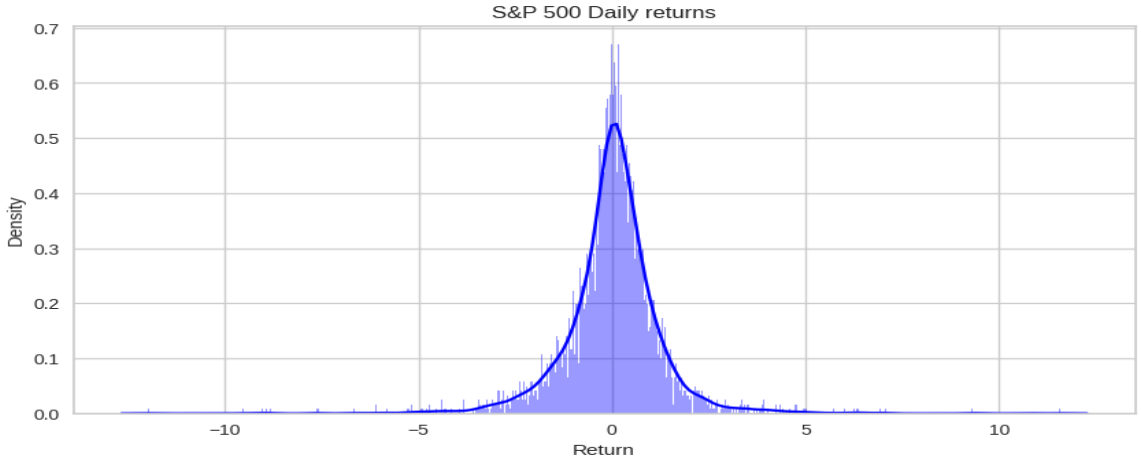

Fig.2 Daily Returns

- Daily simple returns to visualize it better:

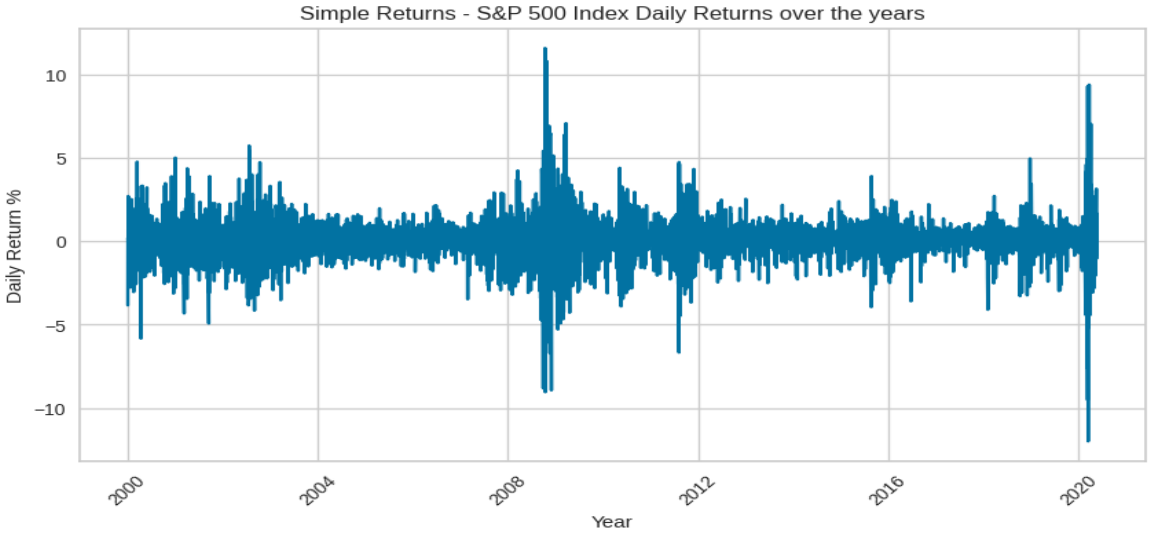

Fig.3 Simple Returns

- Correlation Analysis:

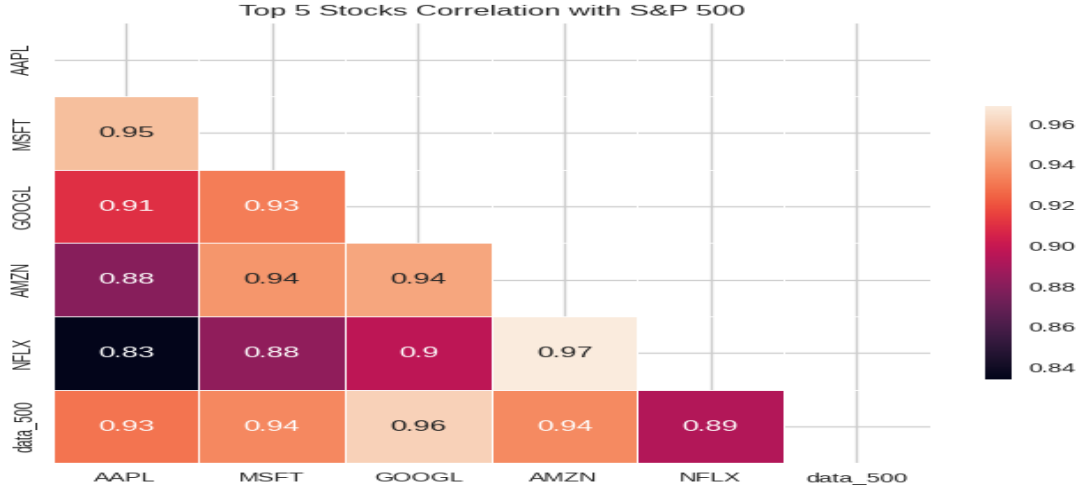

Fig.4 Correlation of Top-5 and S&P-500

- Annualized Returns:

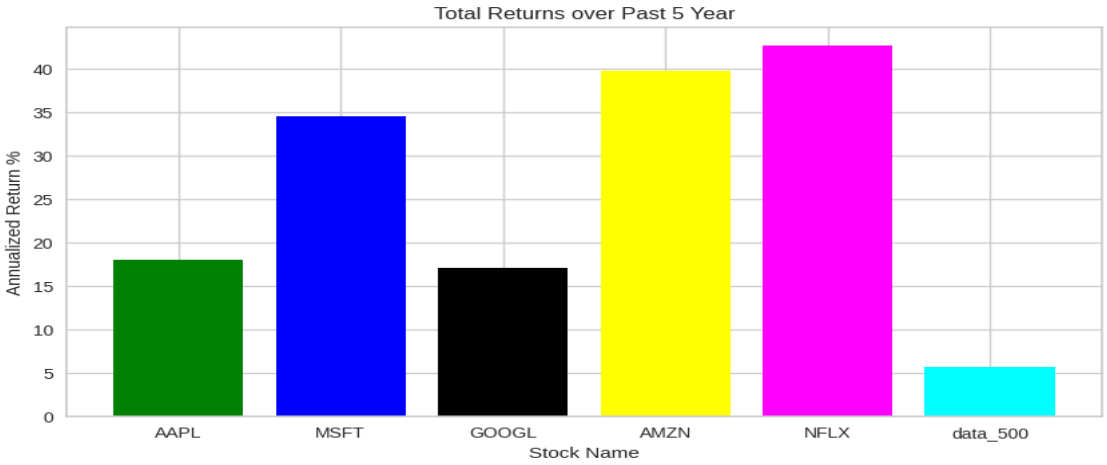

Fig.5 Annualized Returns

- Volatility Analysis:

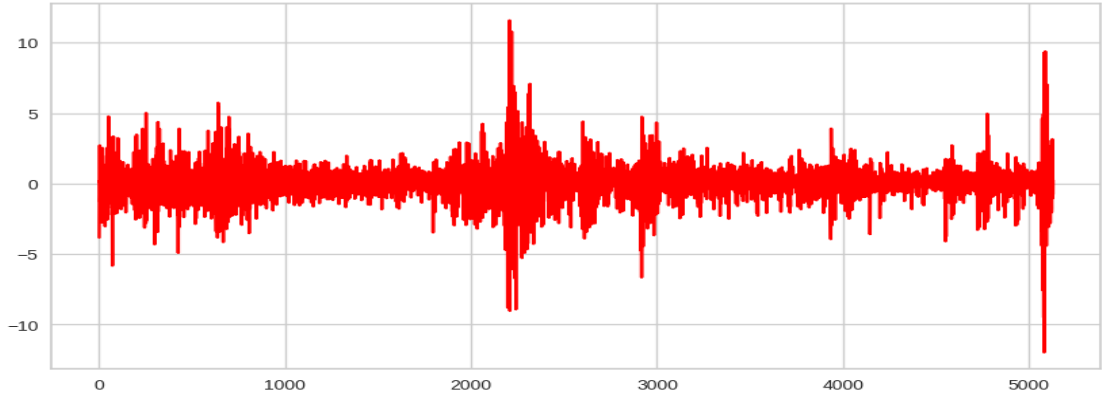

Fig.6 Volatility Analysis

- Moving Averages:

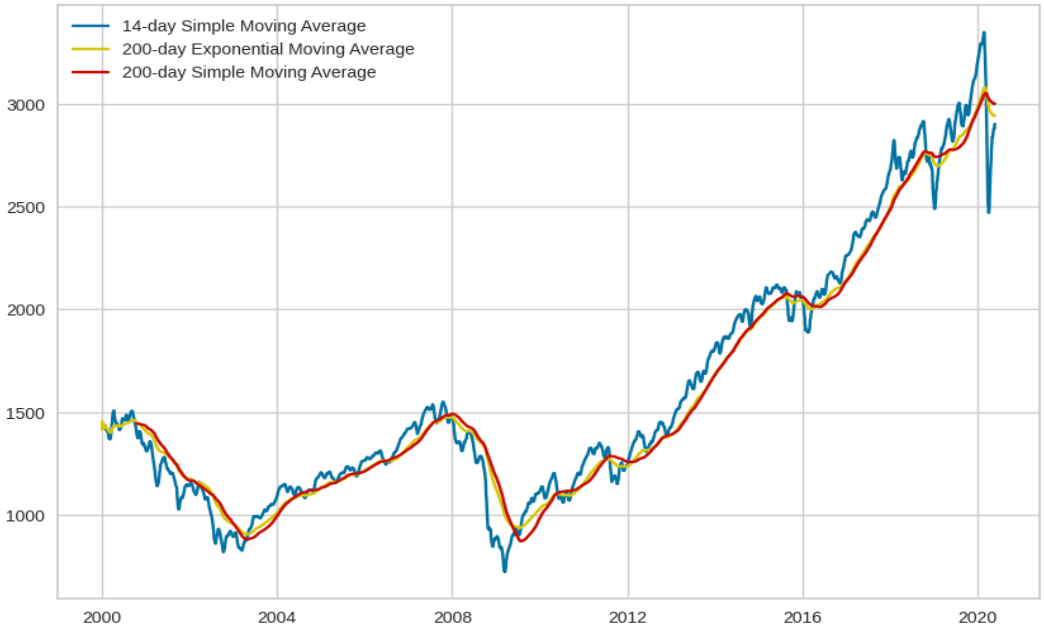

Fig.7 Moving Averages

Chapter 3

MODEL SELECTION & TRAINING

3.1 FEATURE ENGINEERING

Feature engineering is a crucial step in predictive modeling, as it involves transforming raw data into meaningful features that enhance the performance of machine learning models. In this project, several technical indicators and statistical features were engineered to improve the predictive accuracy of S&P 500 stock prices.

3.1.1 Technical Indicators: Indicators are essential for capturing stock price trends, momentum, and volatility, making them valuable for predictive modeling. In this project, key indicators like the Simple Moving Average (SMA) over 14-day and 200-day periods were used to smooth price fluctuations and identify trends. The Exponential Moving Average (EMA) was included to give more weight to recent prices, helping track momentum. The Relative Strength Index (RSI) measured the speed and magnitude of recent price changes, highlighting overbought or oversold conditions. Together, these indicators enhanced the feature set and improved the model's forecasting accuracy.

3.1.2 Statistical Feature: To better capture stock price behavior, several statistical metrics were incorporated into the analysis. Daily returns were calculated as the percentage change in closing prices, providing insights into short-term price fluctuations and helping to identify patterns in market movement. Volatility was measured using rolling standard deviation over various time windows, allowing for an assessment of market risk and the consistency of price changes over time. Additionally, annualized returns were computed to facilitate the comparison of performance trends across longer periods, offering a clearer perspective on long-term investment potential. These metrics enriched the dataset and supported more accurate and robust predictive modeling.

3.1.3 Correlation Features: A strong correlation was observed between S&P 500 and the Big 5 tech stocks (Apple, Amazon, Microsoft, Netflix, and Google), with over 90% correlation. This feature was used to predict stock price movements based on broader market trends.

3.1.4 Target Variable: Future Price Change: A key feature indicating whether the stock price is expected to rise or fall over a certain period.

3.1.5 Train-Test-Split: Time series data cannot be sliced randomly. So that I sliced first 85% data to train the model and remaining 15% (latest data) to test the models.

- 14-day future closing price, 14-day and 200-day moving average and 14-day and 200-day EMA:

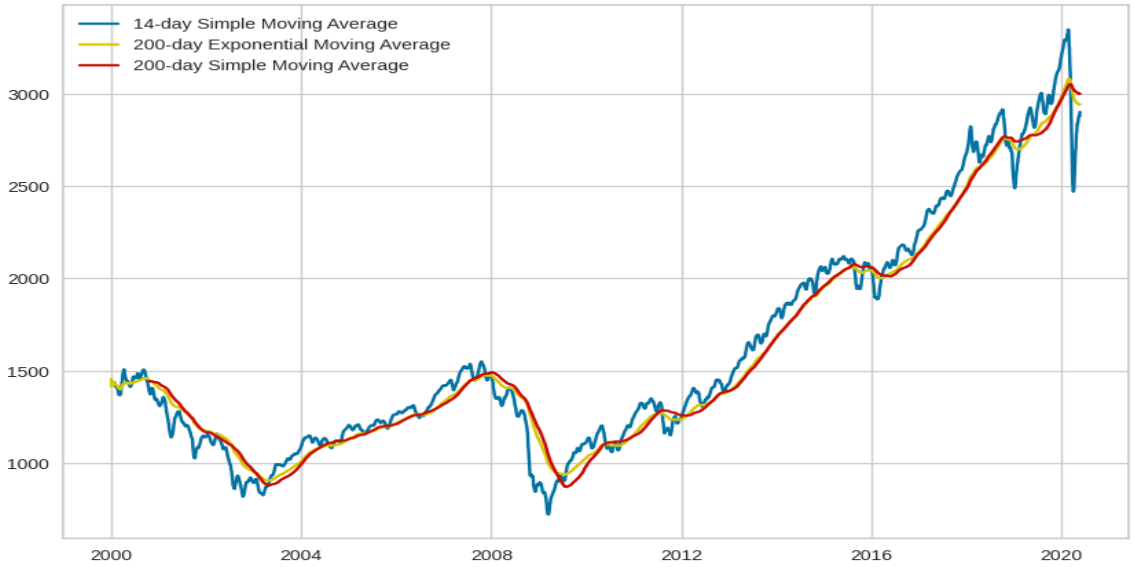

Fig.8 14-Day Future Closing Price

- Correlation:

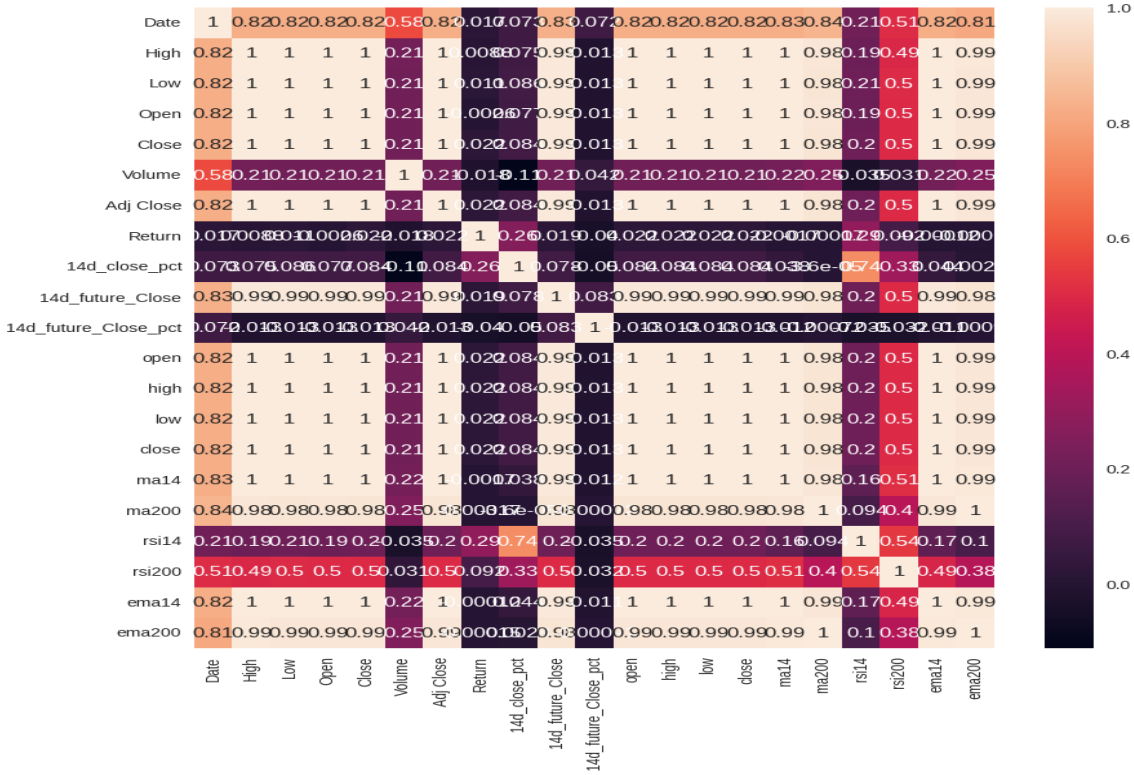

Fig.9 Correlation

- Returns - S&P 500 and Big 5:

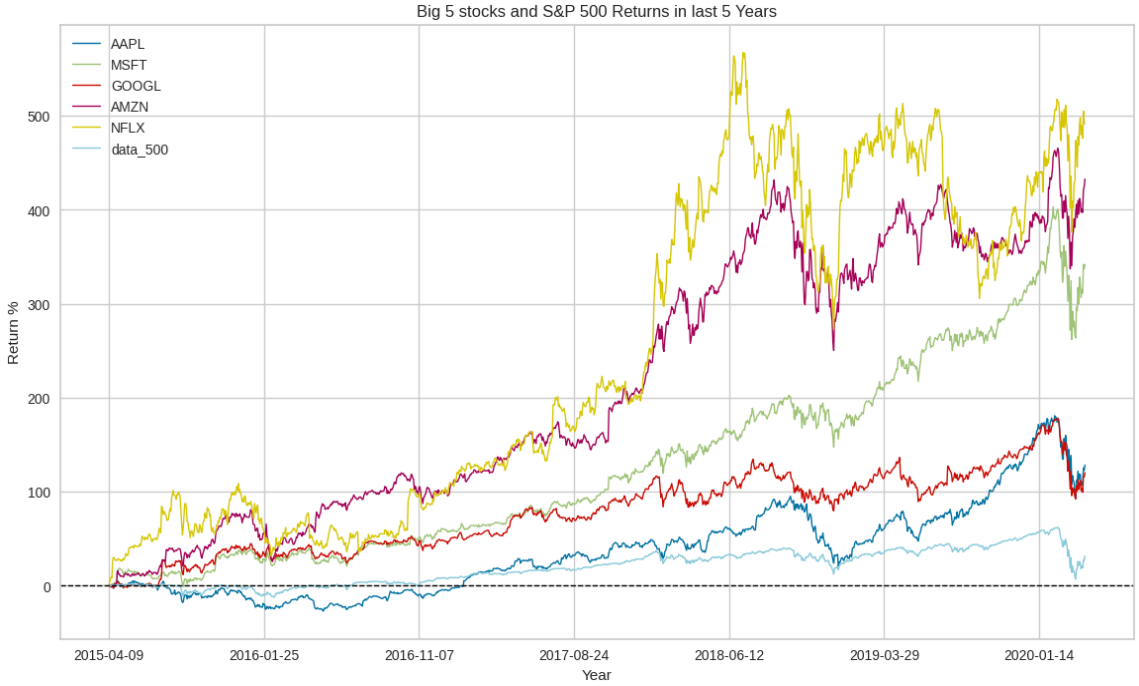

Fig.10 Returns-S&P-500 and Big-5.

3.2 MACHINE LEARNING MODEL

To predict the S&P 500 stock prices, various machine learning models were implemented and evaluated for their predictive performance. Each model was tested using historical stock price data and technical indicators to identify potential trends.

3.2.1 Decision Tree:

The Decision Tree model was trained on the dataset to predict S&P 500 stock movements. During training, the model achieved an accuracy of 77% (0.77), indicating strong performance on the training data. However, when evaluated on the test dataset, the accuracy dropped significantly to 18% (0.18).
This drastic drop suggests that the Decision Tree has overfitted to the training data. Overfitting occurs when the model learns patterns specific to the training set, including noise, but fails to generalize well to new, unseen data. As a result, its performance on the test set is poor.

It looks like your model is performing well on the training dataset but not generalizing well to the test dataset.

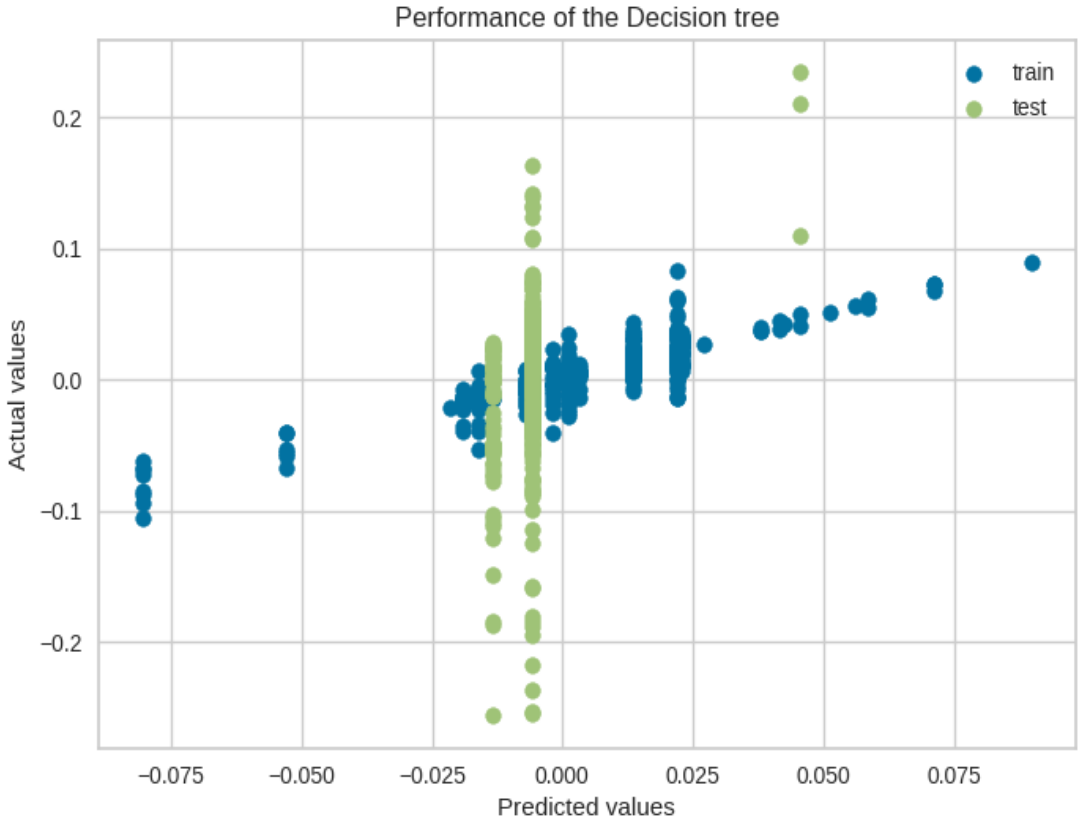

Fig.11 Performance of Decision Tree

3.2.2 Random Forests:

I initially used a Decision Tree model, but it did not perform well on the test dataset. To improve performance, I applied Random Forests, which help reduce variance by combining multiple trees. The model achieved 85.08% training accuracy but only 5.5% test accuracy, indicating severe overfitting. Further tuning and optimization are needed to improve generalization.

3.2.3 Gradient Boosting:

Boosted models are a class of machine learning algorithms that improve predictions by iteratively fitting models, such as Decision Trees, to the data. Each new model learns from the residual errors of the previous one. Gradient Boosting performed slightly better, achieving 0.024 accuracy, but it is still not sufficient for accurate stock price prediction. Further improvements are needed for better performance.

3.2.2 Neural networks:

I tried Neural Networks next, but they performed poorly on the dataset, even giving negative accuracy. This happened because the model wasn't properly tuned. Neural Networks need a lot of customization, like using different activation functions or custom loss functions. There's good scope for improvement, but due to limited time, I decided to try one more forecasting model.

3.2.3 Facebook's Prophet:

Prophet is a procedure for forecasting time series data based on an additive model where nonlinear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well.

When I ran Prophet model with target variable as "Adjusted close", R-square value came out to be 0.98 for the fitted model which means the model is a great fit. To find out if its an overfit using Mean squared error and Mean absolute error.

Mean Squared Error: 6214

Mean square error is decent so the model is a good fit.

Mean Absolute Error: 51.49

It means the predicted value can be 51 basis points away from the actual value either side of the curve at the maximum.

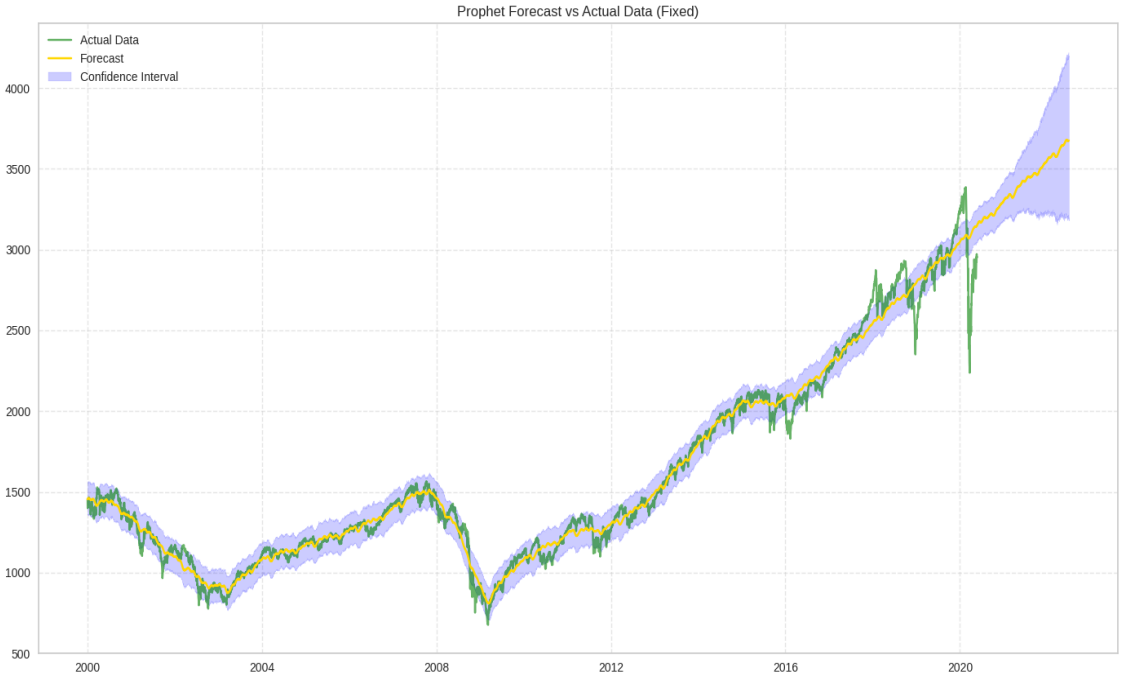

Fig.12 Prophet Forecast VS Actual Data

Chapter 4

MODEL EVALUTION & VALIDATION

4.1 PERFORMANCE MATRICES

Table 2: Performance Matrices

Model	Train Accuracy	Test Accuracy	R^2 Score	MSE	MAE
Decision Tree	77.08%	5.5%	-	-	-
Random Forest	85.08%	2.3%	-	-	-
Gradient Boosting	74.37%	2.4%	-	-	-
Neural Network	-	Negative	-	-	-
Prophet	-	-	0.98	6244.43	51.49

Chapter 5

CONCLUSION AND FUTURE SCOPE

5.1 CONCLUSION

In conclusion, this project demonstrated the application of various machine learning models and time series forecasting techniques to predict the S&P 500 index price. Through extensive data collection and analysis, it became evident that the Big 5 tech stocks have a strong influence on the overall index, as indicated by their high correlation. The S&P-500 has shown a long-term trend of positive returns, with notable fluctuations during major economic events such as the 2008 financial crisis and the COVID-19 pandemic. Despite experimenting with multiple Machine learning models, including Decision Trees, Random Forests, Gradient Boosting, and Neural Networks, the results revealed limitations in their predictive capabilities, especially in capturing the complex patterns of financial time series data. However, the Facebook Prophet model stood out by delivering reliable forecasts, suggesting an upward trend in the index over the next two to three years. This project highlights the importance of model selection, feature engineering, and domain knowledge in financial forecasting, and sets a foundation for more advanced modeling and optimization in the future.

5.2 FUTURE SCOPE

This project focuses exclusively on evaluating different machine learning models for predicting the S&P-500 index based solely on historical stock data. While market sentiment derived from news articles, political developments, investor behavior, and public mood plays a significant role in influencing stock prices, it is intentionally excluded from the scope of this project. Sentiment analysis typically requires advanced natural language processing techniques to extract insights from unstructured data sources such as social media, financial news, and public forums. To maintain a clear and manageable focus, this study limits itself to quantitative data and technical indicators derived from historical price movements. By doing so, the analysis remains data-driven and avoids the complexity and subjectivity involved in assessing human sentiment.

Chapter 6

REFERENCE

1. Machine Learning Algorithms: https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/
2. https://campus.datacamp.com/courses/importing-and-managing-financial-data-in-python/importing-stock-listing-data-from-excel?ex=1
3. https://finance.yahoo.com/quote/API/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAELGmvfYdKo2hByDb2ET2XdTIMmXobkuvOk99Qzeq4qMvfi1PfJst4O1G1LoHLBRee8LsHgCUUm5LrbVHASmnG-D-xs5-q6QhzBw873KSX5fr1RVuF_BVihIjcAkx5vyy0zR6reH2RJWewDdwIxQqKFWdb7iVSvAp0RM1Z_fjDF6
4. Using Machine Learning Models to Predict S&P500 Price Level and Spread Direction Alex Fuster (akfuster@stanford.edu), Zhichao Zou (zzou@stanford.edu)
5. https://medium.com/@randerson112358/stock-price-prediction-using-pythonmachine-learning-e82a039ac2bb
6. https://www.analyticsvidhya.com/blog/2018/05/generate-accurate-forecasts-facebook-prophet-python-r/
7. APPLICATION OF FACEBOOK'S PROPHET ALGORITHM FOR SUCCESSFUL SALES FORECASTING BASED ON REAL-WORLD DATA Emir Žunić1,2, Kemal Korjenić1 , Kerim Hodžić2,1 and Dženana Đonko2 1 Info Studio d.o.o. Sarajevo, Bosnia and Herzegovina 2Faculty of Electrical Engineering, University of Sarajevo, Bosnia and Herzegovina
8. https://in.investing.com/indices/us-spx-500
9. https://stockanalysis.com/list/sp-500-stocks/
10. Nair, Anjana & Narayanan, Jayasree. (2022). Indian Stock Market Forecasting using Prophet Model. 1-7. 10.1109/CSI54720.2022.9924117.
11. https://www.advancinganalytics.co.uk/blog/2021/7/26/facebook-prophet-and-the-stock-market-part-2
12. Machine learning approaches in stock market prediction: A systematic literature review
13. Predicting stock market index using fusion of machine learning techniques
14. Stock Price Prediction and Forecasting using Stacked LSTM https://www.analyticsvidhya.com/blog/2021/05/stock-price-prediction-and-forecasting-using-stacked-lstm/
15. A performance comparison of machine learning models for stock market prediction with novel investment strategy https://pmc.ncbi.nlm.nih.gov/articles/PMC10513304/