Optimizing Investment Portfolios: Comprehensive Analysis of Stock Price Prediction Models, Sectoral Analysis and Correlation

Sudhanshu Agarwal | SI 671: Data Mining: Methods and Applications | Instructor: Prof. Paramveer Dhillon

INTRODUCTION/OBJECTIVE

The aim of this research project is to meticulously assess the forecasting prowess of three renowned time series models — ARIMA, LSTM, and Prophet — by applying them to the stock prices of a curated selection of firms featured on prominent indices such as NYSE, NASDAQ, S&P 500, and Forbes 500. Our endeavor is to pinpoint the most accurate and reliable model for stock price prediction through comprehensive quantitative analysis and stringent performance benchmarking. The project extends beyond mere predictive analysis to include a sectoral evaluation, focusing on key industry segments like technology, finance, and manufacturing, to offer a granular view of the models' performance across different market sectors. By correlating the models' outputs with real-world events, we seek to provide a nuanced understanding of the factors driving stock price movements, thereby furnishing investors with data-driven insights for strategic portfolio management.

KEY RESEARCH/BUSINESS QUESTIONS

- How well do ARIMA, LSTM, and Prophet models perform in predicting future stock prices for a diverse set of companies from different stock exchanges?
- What are the differences in predictive accuracy and reliability among the three models?
- Which model exhibits the highest predictive power and consistency across various benchmarks?
- How do the models compare in terms of their ability to handle different types of stock price data (e.g., highly volatile, seasonal, or trend-driven)?
- Can the findings from this analysis provide insights into selecting the most suitable model for stock price prediction in practical investment scenarios?"
- Implementing relevant financial benchmarks and evaluation metrics, such as the Efficient Market Hypothesis (EMH),Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE) to compare the models against market efficiency standards.
- Furthermore, our project entails conducting a sectoral analysis of stocks by categorizing them into distinct industry sectors, including finance, technology, and manufacturing etc. Through a comparative evaluation of returns from each sector over a specified time frame, our goal is to pinpoint the sector that generated the highest returns. Additionally, we will delve into the interplay between these returns and pertinent news events. This holistic analysis promises to offer invaluable insights, allowing investors to make data-driven decisions regarding portfolio restructuring and diversification based on sector-specific performance and its alignment with market news and trends.

DATASET(S) USED

Our dataset is a comprehensive collection retrieved from Kaggle, featuring weekly updates of key financial metrics from an array of companies listed on premier stock exchanges such as NASDAQ, S&P 500, and NYSE. This rich dataset encompasses vital statistics like Date, Volume, High, Low, and Closing Price, presenting an extensive chronological record up until the year 2022 for a majority of the stocks. With an estimated size of 9.5 GB post-extraction from the zip file, the dataset stands as a robust repository for historical stock market analysis.

Link to the data source - https://www.kaggle.com/datasets/paultimothymooney/stock-market-data/data

Alternative – Use Yahoo finance library

MODEL SELECTION/WORKFLOW

Model Selection and Development:

- Chose three time series forecasting models: ARIMA for its statistical rigor in handling stationary data, LSTM for its deep learning capabilities in capturing complex patterns, and Prophet for its ease of use and handling of seasonal trends.
- Configured each model with appropriate hyperparameters, utilizing cross-validation to tune them for optimal performance.

Model Training and Validation:

- Trained each model on historical stock price data, dividing the dataset into training and testing subsets to evaluate model performance.
- Employed validation techniques such as walk-forward validation for ARIMA and time-based splits for LSTM and Prophet to ensure the models were not overfitting and to gauge their predictive accuracy on unseen data.

Sectoral Analysis:

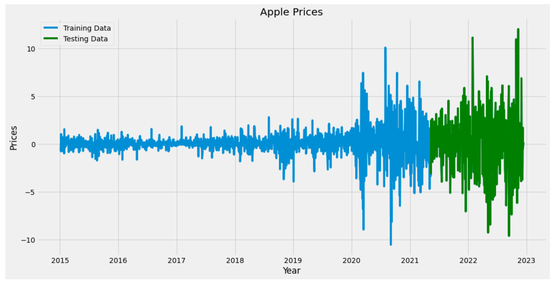
- Categorized stocks into their respective industry sectors such as technology, finance, and manufacturing and predicted which stock had the highest closing price in that particular sector.

Results Interpretation and Comparative Analysis:

- Used performance metrics such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) to quantitatively compare the models' predictive accuracies.
- Visualized the results using Plotly to create interactive graphs, facilitating a clear comparison between actual and predicted stock prices, as well as among the different sectors.
- Examined the correlation between the models' predictions and actual market events to interpret the real-world applicability of the findings.

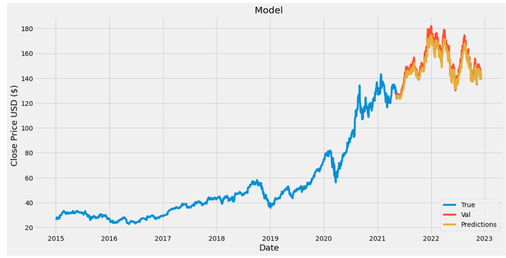
DATA PREPROCESSING FOR TIME SERIES ANALYSIS

- Perform the Augmented Dickey-Fuller (ADF) test to check the stationarity of the time series.
- Apply differencing to the time series to remove trends and seasonality, which can help in achieving stationarity.
- Determine the order of differencing (d) required to make the series stationary.
- Use the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots to identify the order of the ARIMA model (p and q parameters).
- Identify any strong seasonal patterns that need to be modeled separately.
- Prophet can handle seasonality internally, but any known seasonal effects should be specified in the model.
- For LSTM, tune hyperparameters such as the number of layers, number of neurons, learning rate, and epochs.
- For Prophet, adjust trend flexibility and seasonality parameters if necessary.



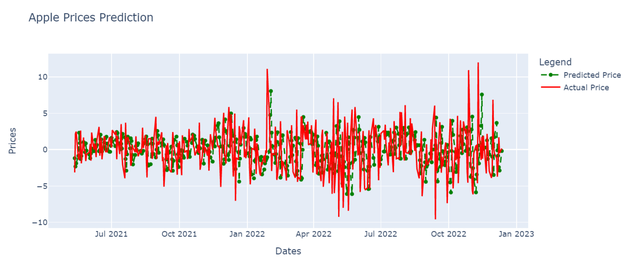
RESULTS FOR DIFFERENT MODELS

LSTM –



Forecast Horizon: The model appears to be making short-term forecasts as the predictions follow the actual prices closely. Long-term forecasts often diverge more from the actual prices due to accumulating prediction errors and the inherent uncertainty in the market.

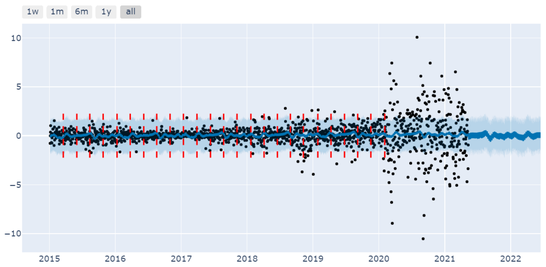
ARIMA -



Close Tracking: The predicted prices (green) closely track the actual stock prices (red), indicating the model is responsive to changes in the stock's price on a day-to-day basis.

Volatility Representation: Both lines exhibit volatility, which the model seems to capture well, reflecting the inherent variability in the stock market.

Prophet -

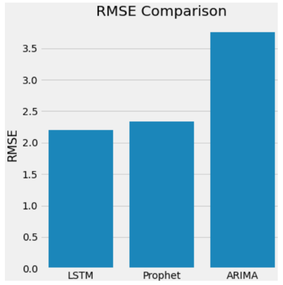


Extreme Changes: The occurrence of data points that reach the extremes of the y-axis, especially those above +5 or below -5, suggests days with significant price movements, which could be due to market shocks, earnings reports, product announcements, or global economic events.

Consistency with Historical Events: The increased volatility in the later years could correlate with real-world events affecting the financial markets, such as economic uncertainty, changes in technology, or shifts in consumer behavior that could impact Apple's stock specifically.

COMPARATIVE ANALYSIS

The bar chart compares the Root Mean Squared Error (RMSE) of three different forecasting models: LSTM, Prophet, and ARIMA. The ARIMA model has the highest RMSE, indicating less accuracy in forecasting compared to the other two models. LSTM shows the lowest RMSE, suggesting it was the most accurate in predicting stock prices in this analysis. Prophet's performance is moderate, with its RMSE falling between that of LSTM and ARIMA.

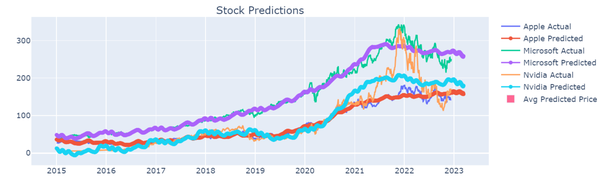


SECTORAL ANALYSIS

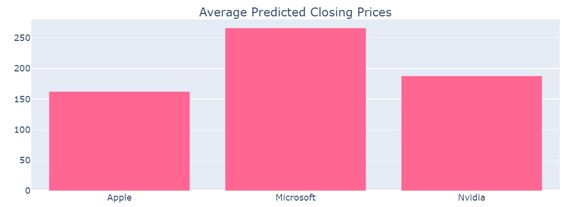
Sector Growth: The overall uptrend in the actual prices of Apple, Microsoft, and Nvidia reflects the robust growth within the Technology sector, indicative of innovation, market expansion, and increasing consumer and enterprise demand for technology solutions.

Technological Advancements: Breakthroughs in cloud computing, AI, and other emerging technologies, often spearheaded by companies like Microsoft and Nvidia, can lead to increased investment and stock valuation.

Technology Sector Stock Analysis



The bar chart represents the average predicted closing prices for stocks from three key players in the Technology sector: Apple, Microsoft, and Nvidia. The bar heights indicate that Microsoft's average predicted closing price is the highest among the three, suggesting that, according to the model's predictions,.



CONCLUSION/FUTURE WORK

This study has demonstrated that among the ARIMA, LSTM, and Prophet models, LSTM provided the most accurate predictions for stock prices, as indicated by its lowest RMSE value. The analysis highlighted the varying capabilities of each model to capture the underlying patterns in stock market data.

FUTURE WORK

Further research could explore the integration of external variables such as economic indicators and sentiment analysis to enhance prediction accuracy. Additionally, investigating the application of ensemble methods that combine the strengths of individual models may yield improvements in forecasting