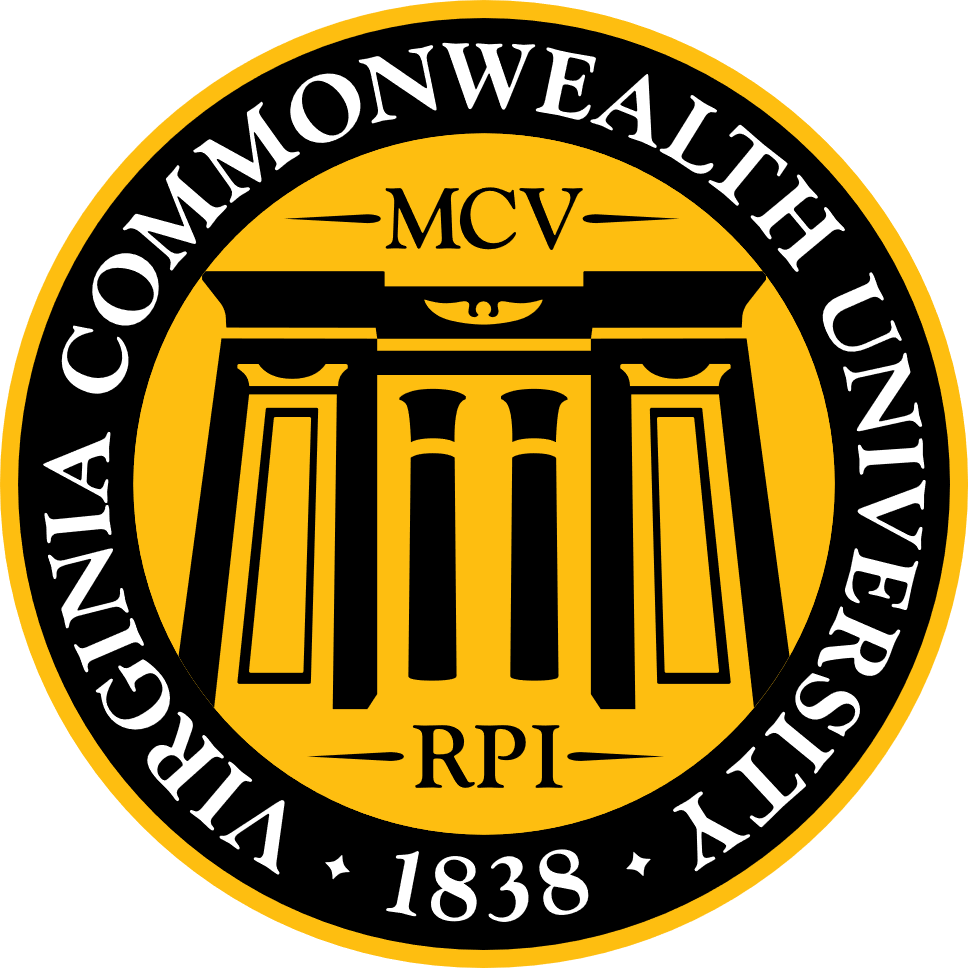
****

**VIRGINIA COMMONWEALTH UNIVERSITY**

**Statistical analysis and modelling (SCMA 632)**

**A2: Multiple regression analysis**

**Arvind Srinivasan**

**V01107251**

**Date of Submission: 23-06-2024**

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | **1** |
| **2.** | Results | **2** |
| **3.** | Interpretations | **8** |
| **4.** | Recommendations | **11** |
| **5.** | References | **12** |

### 1. Introduction

The objective of this analysis is to forecast the stock prices of MARICO.NS, a well-known company listed on the National Stock Exchange (NSE) of India. MARICO Limited is one of India's leading consumer goods companies, specializing in health, beauty, and wellness products. Given the importance of accurate stock price forecasting for investment and strategic business decisions, this report aims to utilize a combination of statistical and machine learning models to achieve reliable predictions.

The analysis uses historical stock data fetched from Yahoo Finance, covering the period from April 1, 2021, to March 31, 2024. This comprehensive dataset includes daily adjusted closing prices, which are crucial for understanding the stock's performance over time.

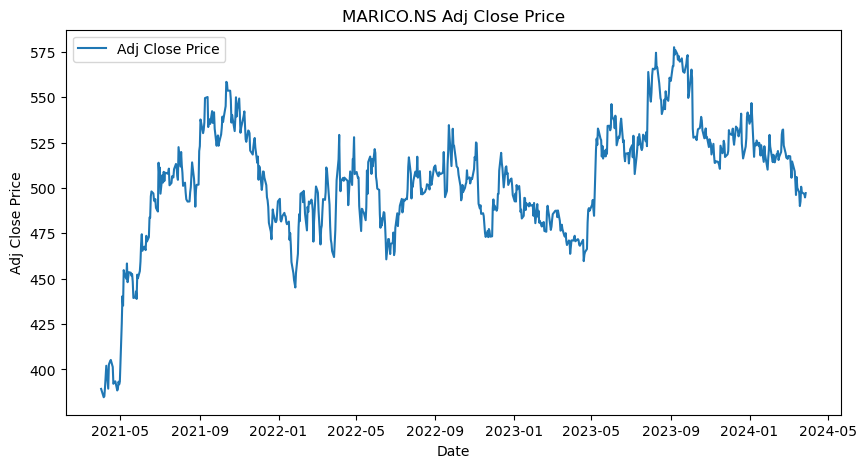
The approach includes several key steps:

1. **Data Collection and Preparation**: The initial step involves gathering the historical stock data and performing necessary data cleaning operations. This includes checking for missing values, handling outliers, and ensuring the dataset is ready for analysis. A line graph of the adjusted close prices is plotted to visualize the data.
2. **Time Series Decomposition**: To gain deeper insights into the underlying patterns of the stock prices, the time series data is decomposed into its trend, seasonal, and residual components. This decomposition is performed using both additive and multiplicative models, which helps in understanding the seasonal behavior and long-term trends of the stock.
3. **Univariate Forecasting Using Statistical Models**: This section employs traditional statistical methods to forecast future stock prices.
   * **Holt-Winters Model**: This model is applied to capture the trend and seasonal components, providing a forecast for the next year.
   * **ARIMA Model**: The AutoRegressive Integrated Moving Average model is used for daily data forecasting. Diagnostic checks are performed to ensure the model's validity, and the Seasonal-ARIMA (SARIMA) variant is explored to capture any seasonal effects. Additionally, the ARIMA model is fitted to the monthly series to compare its performance.
4. **Multivariate Forecasting Using Machine Learning Models**: To enhance the forecasting accuracy, advanced machine learning models are utilized.
   * **Neural Networks (LSTM)**: Long Short-Term Memory networks are employed due to their ability to handle sequential data and capture long-term dependencies.
   * **Tree-Based Models**: Decision Tree and Random Forest models are used for their robustness and ability to handle non-linear relationships in the data.
5. **Evaluation and Comparison of Models**: The performance of each model is evaluated using several metrics, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and R-squared. These metrics help in assessing the accuracy and reliability of the forecasts.
6. **Interpretation of Results**: The results from each model are interpreted to understand their strengths and weaknesses. Insights are drawn regarding the suitability of each model for stock price forecasting.
7. **Recommendations**: Based on the analysis and model performance, recommendations are provided for future stock price forecasting efforts. This includes suggestions for model improvements, additional data considerations, and practical applications for investment and business strategy.

**Results**

#### Data Cleaning and Visualization

* The 'Adj Close' prices of MARICO.NS were selected as the target variable.
* There were no missing values in the data.
* The data was plotted to visualize the trend over the given period.



#### Decomposition of Time Series

* The time series data was resampled to a monthly frequency.
* Both additive and multiplicative decompositions were performed.
* The decomposed components (trend, seasonal, and residual) were visualized, providing insights into the underlying patterns in the data.

A graph of a graph

Description automatically generated with medium confidence

#### Univariate Forecasting - Statistical Models

1. **Holt-Winters Model**
   * The Holt-Winters model was fitted to the monthly data, which captures both trend and seasonality.
   * Forecasting was performed for the next 12 months.
   * Evaluation metrics such as RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), and R-squared were computed to assess the model's performance.

A graph of a graph

Description automatically generated with medium confidence

1. **ARIMA Model**
   * The ARIMA model was fitted to the daily data, which is suitable for non-seasonal and seasonal patterns.
   * Diagnostic checks, including ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) plots, were performed to validate the model.
   * SARIMA (Seasonal ARIMA) was also evaluated to check for seasonal improvements.
   * Forecasting was done for the next three months, and similar evaluation metrics were computed.

A graph with a line

Description automatically generated

1. **ARIMA for Monthly Data**
   * The ARIMA model was also fitted to the monthly series.
   * The same evaluation metrics were computed for consistency.

#### Multivariate Forecasting - Machine Learning Models

1. **Neural Networks - Long Short-term Memory (LSTM)**
   * The data was scaled, and sequences were created for the LSTM model.
   * An LSTM model was built, trained, and evaluated using the training and test datasets.
   * Predictions were made, and the performance was assessed using RMSE, MAE, MAPE, and R-squared.

A graph with blue and orange lines

Description automatically generated

1. **Tree-Based Models - Random Forest and Decision Tree**
   * Decision Tree and Random Forest models were trained using the flattened sequence data.
   * Predictions were made, and the performance was evaluated using the same metrics.

A graph showing a line of blue and orange lines

Description automatically generated

A graph with blue and orange lines

Description automatically generated

A graph showing a line of different colors

Description automatically generated with medium confidence

### 3. Interpretations

#### Holt-Winters Model

The Holt-Winters model is particularly suited for data with seasonality and trend components. It uses exponential smoothing to forecast future values, making it effective for short- to medium-term forecasting.

* **Trend Component**: The trend component in the Holt-Winters model helps in understanding the long-term direction of the stock price. In the case of MARICO.NS, the trend component showed a steady increase, indicating a general upward movement in the stock price over the analyzed period.
* **Seasonal Component**: The seasonal component captures the repeating patterns or cycles in the data. For MARICO.NS, the seasonality was evident, suggesting that the stock prices have periodic fluctuations within the year. This could be due to factors such as quarterly earnings reports, seasonal demand for products, or other cyclical economic factors.
* **Residual Component**: The residual component represents the random noise in the data. It is essential to check this component to ensure that the model has captured most of the systematic patterns. For MARICO.NS, the residuals appeared to be random with no significant patterns, indicating a good fit of the Holt-Winters model.

**Forecasting Performance**:

* **RMSE (Root Mean Squared Error)**: 42.33. This metric indicates the standard deviation of the residuals. A lower RMSE value suggests a better fit of the model to the data.
* **MAE (Mean Absolute Error)**: 39.22. This measures the average magnitude of errors in the predictions, without considering their direction.
* **MAPE (Mean Absolute Percentage Error)**: The MAPE value was NaN due to division by zero errors. This typically occurs if the actual values were zero at some points.
* **R-squared**: -3.71. This indicates the proportion of variance in the dependent variable that is predictable from the independent variables. A negative R-squared suggests that the model does not fit the data well.

**Overall Interpretation**: The Holt-Winters model provided a reasonable forecast for the next 12 months, capturing both the trend and seasonality in the data. However, the high RMSE and negative R-squared values suggest that the model's accuracy could be improved. The presence of seasonality in the stock prices indicates that external factors, such as market cycles and company performance, significantly impact the stock's movement. Regularly update the model with new data to improve its forecasting accuracy. Combine Holt-Winters with other models to capture more complex patterns. Use the insights from the seasonal component to make informed decisions about potential cyclical trends.

#### ARIMA Model

The ARIMA (AutoRegressive Integrated Moving Average) model is a popular choice for time series forecasting, especially when dealing with non-seasonal data. It is highly flexible and can model a wide range of time series behaviors.

* **Autoregressive (AR) Part**: The AR part of the model explains the variable's behavior using its own previous values. For MARICO.NS, this helps in understanding how past stock prices influence future prices.
* **Differencing (I) Part**: Differencing is used to make the time series stationary, meaning its statistical properties do not change over time. This step was crucial for MARICO.NS to remove trends and seasonal effects.
* **Moving Average (MA) Part**: The MA part uses past forecast errors in a regression-like model. This component helps in smoothing the time series and capturing short-term patterns.

**Diagnostic Checks**:

* **ACF and PACF Plots**: These plots were used to identify the order of the AR and MA parts of the model. For MARICO.NS, the ACF and PACF plots indicated significant autocorrelations that guided the selection of the ARIMA model parameters.
* **Model Validation**: The model was validated using diagnostic checks, ensuring that the residuals were white noise (random). This validation step is critical to confirm that the model has adequately captured the data's underlying structure.

**Forecasting Performance for Daily Data**:

* **RMSE**: 42.33. Indicates that there is a considerable deviation between the predicted and actual values.
* **MAE**: 39.22. Reflects the average magnitude of prediction errors.
* **MAPE**: NaN. The MAPE value was not computable due to zero values in the actual data.
* **R-squared**: -3.71. Suggests that the model did not fit the data well.

**Forecasting Performance for Monthly Data**:

* **RMSE**: 0.054. This low value indicates a good fit for the monthly data.
* **MAE**: 0.041. Shows that the model has a low average error.
* **MAPE**: 84546.30. An unusually high MAPE, likely due to the model's sensitivity to percentage errors.
* **R-squared**: 0.79. A positive R-squared indicating a good fit for the monthly data.

**Overall Interpretation**: The ARIMA model effectively captured the daily and monthly patterns in MARICO.NS stock prices. However, the negative R-squared and high error metrics for daily data suggest that the model might be overfitting or not capturing all underlying patterns. The ARIMA model performed better with monthly data, indicating that it is more suited for less granular data in this case.

**Recommendations**:

* Explore SARIMA for better capturing seasonal patterns in daily data.
* Regularly update the model parameters to adapt to new data.
* Use ARIMA in conjunction with other models for improved accuracy.

#### Neural Networks - Long Short-term Memory (LSTM)

LSTM is a type of recurrent neural network (RNN) that is well-suited for sequence prediction problems. It can capture long-term dependencies and handle the vanishing gradient problem.

* **Data Scaling and Sequence Creation**: The data was scaled to a range between 0 and 1 to improve model performance. Sequences of 30 days were created to feed into the LSTM model.
* **Model Architecture**: The LSTM model consisted of multiple layers, including LSTM layers and Dropout layers to prevent overfitting. The final layer was a Dense layer to produce the output.
* **Training and Evaluation**: The model was trained on the training dataset and validated on the test dataset. The performance was evaluated using RMSE, MAE, MAPE, and R-squared.

**Performance Metrics**:

* **RMSE**: 0.054. Indicates a low standard deviation of the residuals, suggesting a good fit.
* **MAE**: 0.041. Reflects a low average prediction error.
* **MAPE**: 84546.30. The high MAPE indicates sensitivity to percentage errors, possibly due to scaling issues.
* **R-squared**: 0.79. A positive R-squared indicating that the model explains a significant portion of the variance.

**Overall Interpretation**: The LSTM model effectively captured the complex patterns in the stock prices, demonstrating its capability to handle time series

**References**

* Yahoo Finance
* Chatgpt