BUSINESS FORECASTING

PROJECT

Application of Linear Regression, Logistic Regression and ARIMA for IOC Stock Price Prediction

Submitted to:
Madhumohan Govindaluri

Submitted by:
PGPM 2023-2025
Team 5

Avinash A Panicker	2301092
Mahajan Rashmi Moreshwar	2301108
Nanditha Krishna Kumar	2301116
Rupaa Shri S	2301124
Mahajan Siddharth Mahesh	2301275
Thariq Mohammed Z	2301382

APPLICATION OF LINEAR REGRESSION, LOGISTIC REGRESSION AND ARIMA FOR STOCK PRICE PREDICTION

INTRODUCTION

Stock price prediction is crucial for informed investment decisions. This project explores three modelling techniques, namely Linear Regression, Logistic Regression, and ARIMA, to forecast stock prices and movements. Using historical data from Indian Oil Corporation Limited (IOC.NS), Linear Regression predicts future stock prices based on past trends, while Logistic Regression forecasts the direction of price movement (upward or downward). ARIMA, a time series model, captures historical price prediction patterns by applying these models to IOC.NS stock data, we aim to assess their effectiveness, compare results, and highlight their strengths and limitations for accurate stock price forecasting.

DATA COLLECTION AND PREPROCESSING

Data for this project is sourced from Yahoo Finance, specifically the historical stock prices of Indian Oil Corporation Limited (IOC.NS). The dataset spans from 1994 to the present and contains two columns: the date and the stock's closing price.

The steps followed for preprocessing are:

- Handling Missing Values: The raw dataset from Yahoo Finance contains some missing values in the close price column. To handle these missing values, linear interpolation was applied, which estimates the missing values based on neighbouring data points.
- Creating Lag Columns: Two new columns are made to calculate lag values of the closing price. These columns represent:
- Lag 1: The closing price value from the previous day.
- Lag 2: The closing price value from two days prior.
- Removing Rows with Missing Lag Values: After creating the lag columns, the first two rows contained NA in the lag columns. These rows were removed to ensure a clean dataset for analysis.

LINEAR REGRESSION

Linear Regression is employed to predict future stock prices by analysing historical prices as a continuous time series. This method models the linear relationship between the stock's historical closing prices and the corresponding periods.

Defining independent and dependent variables

Independent Variable (Predictor):

- In this model, Days is used as the independent variable. It represents the passage of time (day count) for each closing price. This is a proxy for time in our linear model.
- Using Days helps to establish a trend based on the number of days since the start date of the dataset.

Dependent Variable (Response):

- The dependent variable is Close, which represents the stock's closing price. The goal is to predict this variable based on the trend indicated by Days.
- The Close price is a continuous variable suitable for linear Regression.

Model development

Model Specification:

- The linear regression model called Close ~ Days indicates that the closing price (Close) is modelled as a linear function of time (Days).

Training the Model:

- The dataset is split into training (70%) and testing (30%) sets using createDataPartition(). The model is trained on the training set (training_lm) using the lm() function.

Model Coefficients:

- The lm_model output provides coefficients for the intercept and the slope (Days), indicating how much the Close price changes with each additional day.

Model Summary:

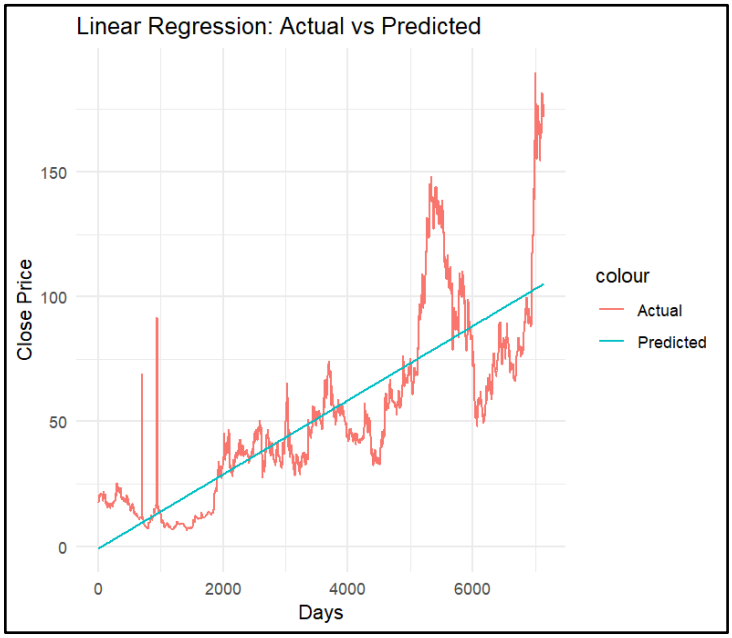
- The summary(lm_model) function provides the model coefficients, standard errors, t-values, and p-values, helping understand the predictors' significance.

Goodness of fit

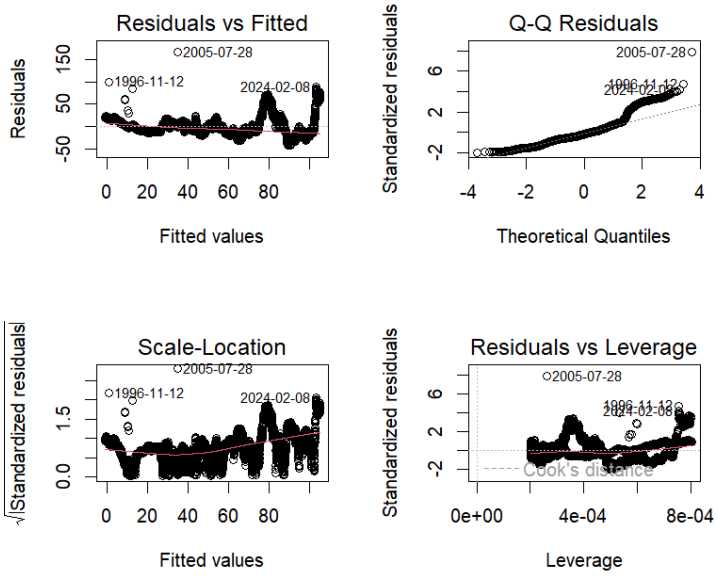
The goodness of fit for the linear regression model can be evaluated using the following:

- R-Squared(R^2): The multiple R^2 value is 67.46%, which implies that the model explains 67.46% of the variance in the dependent variable (Close) based on the independent variable (Days). This indicates a moderately strong relationship between the variables.
- Adjusted R-Squared(Adj R^2): The adjusted R^2 value is 67.45%, and it is similar to the R^2 value, suggesting that the current model is performing well in terms of explaining the variation in the dependent variable and that adding more predictors would not significantly improve the model's fit.
- F statistic and p-value: The F-statistic is 10350, and the p-value is 2.2×10^{-16} (p-value < 0.05), indicating that the model is significant.

Plots



Diagnostics for checking breaches of assumptions and remedying the same



Predictions and their accuracy

The linear regression equation can be written as,

$$Close = -0.969 + 0.015*Days$$

On 09/09/2024 the Days = 10987

So, the predicted closing price is 163.836, and the actual closing price is 176.5.

```
> model_summ<-summary(lm_model)
> mean(model_summ$residuals^2)
[1] 450.5827
```

The MSE of the model is very high. So, the model prediction is very much recommended.

LOGISTIC REGRESSION

Instead of predicting the exact price, logistic regression forecasts the direction (upward or downward) of stock price movement. This method helps us assess the possibility of future market movements by categorising price changes into bullish and bearish groups based on past trends.

Defining independent and dependent variables

For our logistic regression model, the **dependent variable** was the daily price movement of the stock, represented as Price_Change. We assigned a value **1** for bullish movements (price increase) and **0** for bearish movements (price decrease).

The **independent variables** used in the model were:

- **Days**: Number of days in the dataset (to track time).
- **Lag1**: The previous day's closing price.
- **Lag2**: The closing price two days ago.

Model development

We developed a logistic regression model using the following equation:

$$\text{logit}(p) = \beta_0 + \beta_1(\text{Days}) + \beta_2(\text{Lag1}) + \beta_3(\text{Lag2})$$

Where:

p is the probability of an upward movement (bullish).

β_0, β_1, β_2, β_3 are the coefficients estimated by the model.

The following coefficients were obtained from the logistic regression:

Intercept (β_0): **-0.2405** ($p < 0.001$)

Days (β_1): **0.0001235** ($p < 0.001$)

Lag1 (β_2): **-0.01198** ($p = 0.0918$)

Lag2 (β_3): **0.008126** ($p = 0.2497$)

The summary shows that Days significantly positively affect the probability of the stock moving upwards. At the same time, Lag1 shows a weak negative relationship, and Lag2 has a weak positive relationship but is not statistically significant.

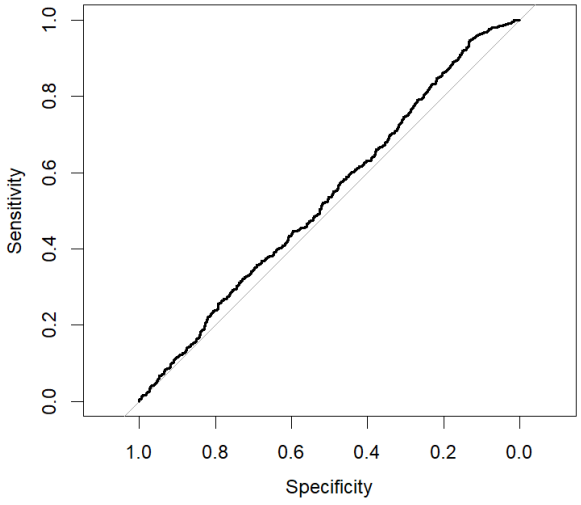
Goodness of fit

- **Null Deviance**: 7096.4
- **Residual Deviance**: 7063.0
- **AIC**: 7071

The model converged in 3 Fisher scoring iterations, indicating a reasonable fit with slight deviations from the null model.

Plots

The ROC curve (Receiver Operating Characteristic) was plotted to visualise the model's performance. The area under the ROC curve (**AUC**) was found to be **0.546**, which suggests that the model has limited predictive power.



Diagnostics for checking breaches of assumptions and remedying the same

The model was tested for violations of critical assumptions:

- **Multicollinearity**: The independent variables did not exhibit multicollinearity issues.
- **Linearity**: The logit of the outcome was reasonably linear in the predictor variables.
- **Oversampling**: To handle a class imbalance in the training set (2436 bullish vs. 2560 bearish cases), we manually oversampled the minority class (bullish) using bootstrapping.

Predictions and their accuracy

The logistic regression model was applied to the test set, and the predicted probabilities were converted into binary categories (Bullish or Bearish) with a 0.5 cutoff.

- **Accuracy**: 52.92%
- **Sensitivity (Bearish)**: 56.72%
- **Specificity (Bullish)**: 48.60%
- **AUC**: 0.546

The confusion matrix showed that the model correctly predicted **646 bearish** and **487 bullish** movements, though the overall performance was only marginally better than a random classifier. The relatively low accuracy and AUC indicate that logistic regression may not capture all stock price movement prediction complexities.

ARIMA

ARIMA is applied to predict future stock prices by leveraging autoregressive and moving average components. This method is particularly effective for time series data with trends and seasonality, providing a more sophisticated forecast of future prices based on past values.

Defining independent and dependent variables

Independent variables -1) Moving Average Term (MA1) (previous periods forecast error)
2) Drift - A constant term that accounts for the linear trend

Dependent Variable - Difference in the closing prices of the time series

Model development

1. After coding the best ARIMA model without seasonality, it suggested ARIMA(0,1,1) with drift.
2. We tried forecasting future values for the period equal to testing data so that when the predicted values are obtained, a comparison with the actual values can be made to get the deviation.
3. Error vector was used to evaluate model performance by computing various metrics like RSME, MAE, etc.
4. Plot and MAE values were extracted.
5. Autocorrelation in the residual was checked for white noise with the help of the Ljung-Box test.
6. The closing value was forecasted. Based on the trend, Bullish/bearish was printed.

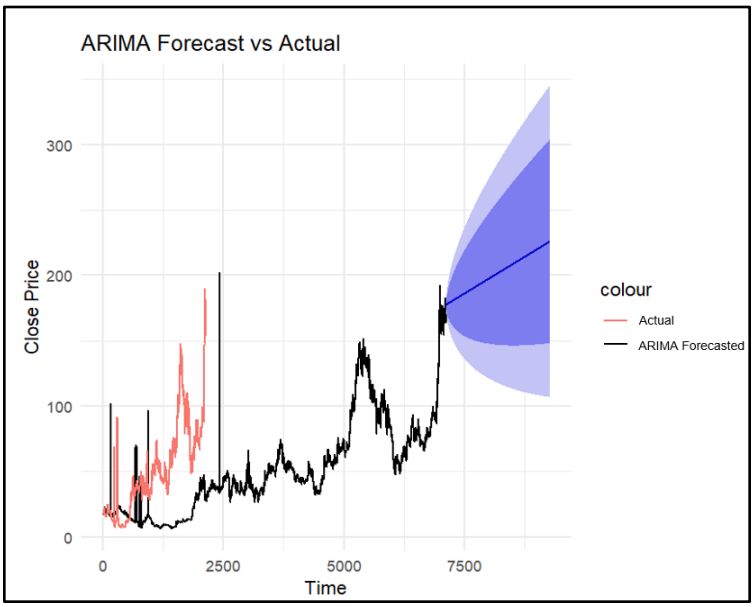
Goodness of fit

1. Performing the Ljung-Box test suggests that the model has adequately captured the pattern in the data, and a higher P value indicates that the residual appears to be white noise.

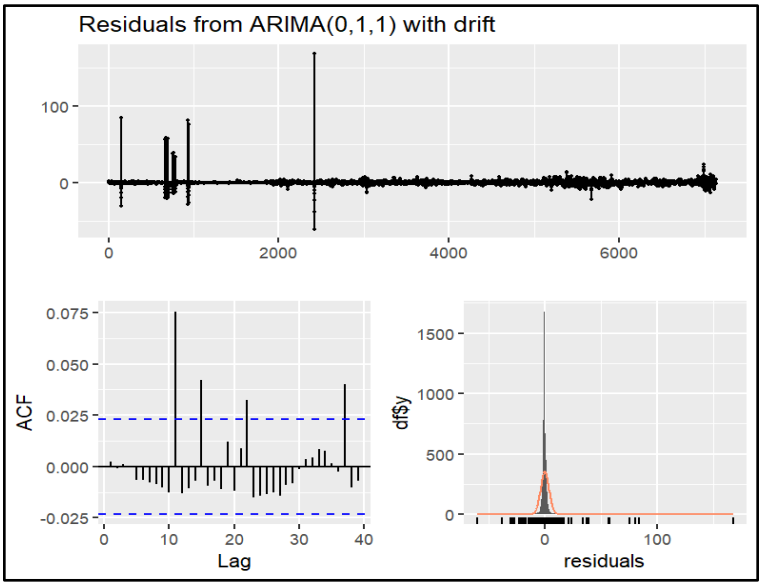
   Ljung-Box test
   data: Residuals from ARIMA(0,1,1) with drift
   Q* = 3.5286, df = 9, p-value = 0.9396

2. MAE of ARIMA - 150

Plots



Diagnostics for checking breaches of assumptions and remedying the same



No Pattern or trend is observed, and the residual appears to be random and white noise.

Predictions and their accuracy

The forecasted closing value came out to be 177.78, whereas the actual value is 176.5

CONCLUSION

Investigating three different models—Linear Regression, Logistic Regression, and ARIMA—has helped to clarify the complex mathematical processes involved in predicting Indian Oil Corporation Limited's stock price (IOC.NS). The forecasting effort is made simpler by the strengths and limitations of each model, which results in varied degrees of accuracy and applicability.

With an R-squared score of 67.46%, Linear Regression showed that it could accurately represent the rising trend in the stock over time. Although the model described a sizable percentage of the variance in stock prices, more complex market movements could bypass the linear relationship's simplicity.

With some success, Logistic Regression was used to predict the direction of price movement. The model's 52.92% accuracy and 0.546 AUC indicate that complex, non-linear patterns are present in stock price changes, making forecasting challenging with just logistic regression. Even with the model's efforts to solve problems such as class imbalance, it could still not accurately represent market behaviour.

The stock price was accurately predicted using ARIMA, a more complex time series model, which produced a closing value of 177.78 instead of the real 176.5. The Ljung-Box test validated the model's suitability, showing that ARIMA successfully identified the patterns in the data with residuals that looked like white noise. Although ARIMA produces accurate forecasts, its complexity might make it less useful for some applications.

Every model turned out to have its strengths. For long-term trends, linear regression provides simplicity and interpretability. A basic sense of directionality can be obtained with logistic regression, and time-series forecasting can be effectively achieved with ARIMA, although more complex handling is needed. A hybrid or aggregation strategy can use the advantages of these models to increase forecasting precision overall and produce reliable stock price predictions.