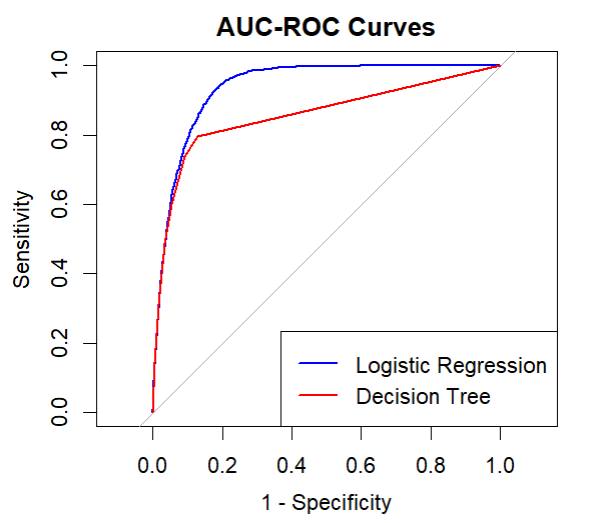
# SCMA- 632

# FINAL EXAM

# V01107757

# UJWAL P

**Section A - Part B**



**Description:**

The graph displays the AUC-ROC (Area Under the Curve - Receiver Operating Characteristic) curves for two models: Logistic Regression (blue curve) and Decision Tree (red curve). The ROC curve plots Sensitivity (True Positive Rate) on the Y-axis against 1-Specificity (False Positive Rate) on the X-axis.

**Analysis:**

* The blue curve (Logistic Regression) is closer to the top-left corner, indicating better performance compared to the red curve (Decision Tree).
* The AUC (Area Under the Curve) value for Logistic Regression is higher than that of the Decision Tree, suggesting that the Logistic Regression model has a better overall ability to discriminate between positive and negative classes.

**Interpretation:**

Model Performance:

The AUC-ROC curve is a performance measurement for classification problems. It tells us how well the model is performing by showing the trade-off between sensitivity (true positive rate) and 1-specificity (false positive rate).

In the graph, the Logistic Regression model (blue curve) has a curve that is closer to the top-left corner of the plot compared to the Decision Tree model (red curve). This indicates that for most thresholds, Logistic Regression has a higher true positive rate and a lower false positive rate than the Decision Tree model.

**Area Under the Curve (AUC):**

The AUC represents the degree or measure of separability. It tells us how much the model is capable of distinguishing between classes.

An AUC of 1.0 represents a perfect model, while an AUC of 0.5 suggests a model that performs no better than random chance.

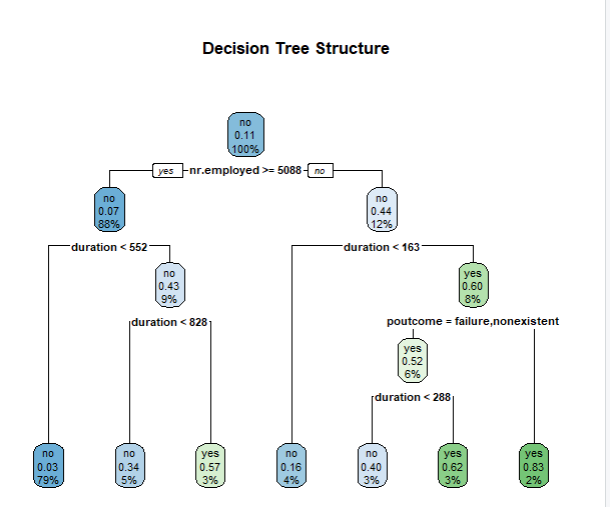
In this graph, the Logistic Regression model has a higher AUC than the Decision Tree model, which means it has better overall performance in distinguishing between positive and negative classes.

**Practical Implications:**

The Logistic Regression model is more reliable and effective at predicting the correct class compared to the Decision Tree model.

In practical terms, if you were to use these models to make decisions (e.g., predicting whether a customer will buy a product or not), Logistic Regression would likely make more accurate predictions, leading to better outcomes and fewer misclassifications.

Overall, the AUC-ROC curves show that the Logistic Regression model outperforms the Decision Tree model by providing a clearer separation between the classes, which is crucial for making accurate predictions in classification tasks.



**Description:**

The second graph shows the structure of a Decision Tree used for classification. The tree splits based on different features (e.g., nr.employed, duration, poutcome), with nodes representing decisions and leaves representing classification outcomes.

**Analysis:**

* The root node splits based on the feature nr.employed with a threshold of 5088.
* Each subsequent node further splits the data based on other features like duration and poutcome.
* The nodes display the proportion of samples that fall into each category, and the leaves show the final classification (yes/no) and the associated percentage of samples.

**Interpretation:**

Binary Splits:

A decision tree classifies data by splitting it into subsets based on feature values. Each split is a binary decision, dividing the dataset into two groups.

In this tree, the first and most significant split is based on the feature nr.employed with a threshold of 5088. This means the tree first decides whether the nr.employed value is greater than or equal to 5088 to make the first major decision in classifying the data.

Hierarchical Decision-Making:

The decision tree's structure reveals a step-by-step decision-making process, where each node represents a decision based on a specific feature and threshold.

After the initial split on nr.employed, the tree continues to split the data further based on other features such as duration and poutcome. Each split aims to increase the homogeneity of the resulting subsets concerning the target variable (the class label).

Feature Importance:

The features used at the higher levels of the tree (closer to the root) are considered more important as they make the most significant impact on classification early in the process.

In this tree, nr.employed is the most important feature, followed by other features like duration and poutcome, which are used in subsequent splits.

Final Leaves and Predictions:

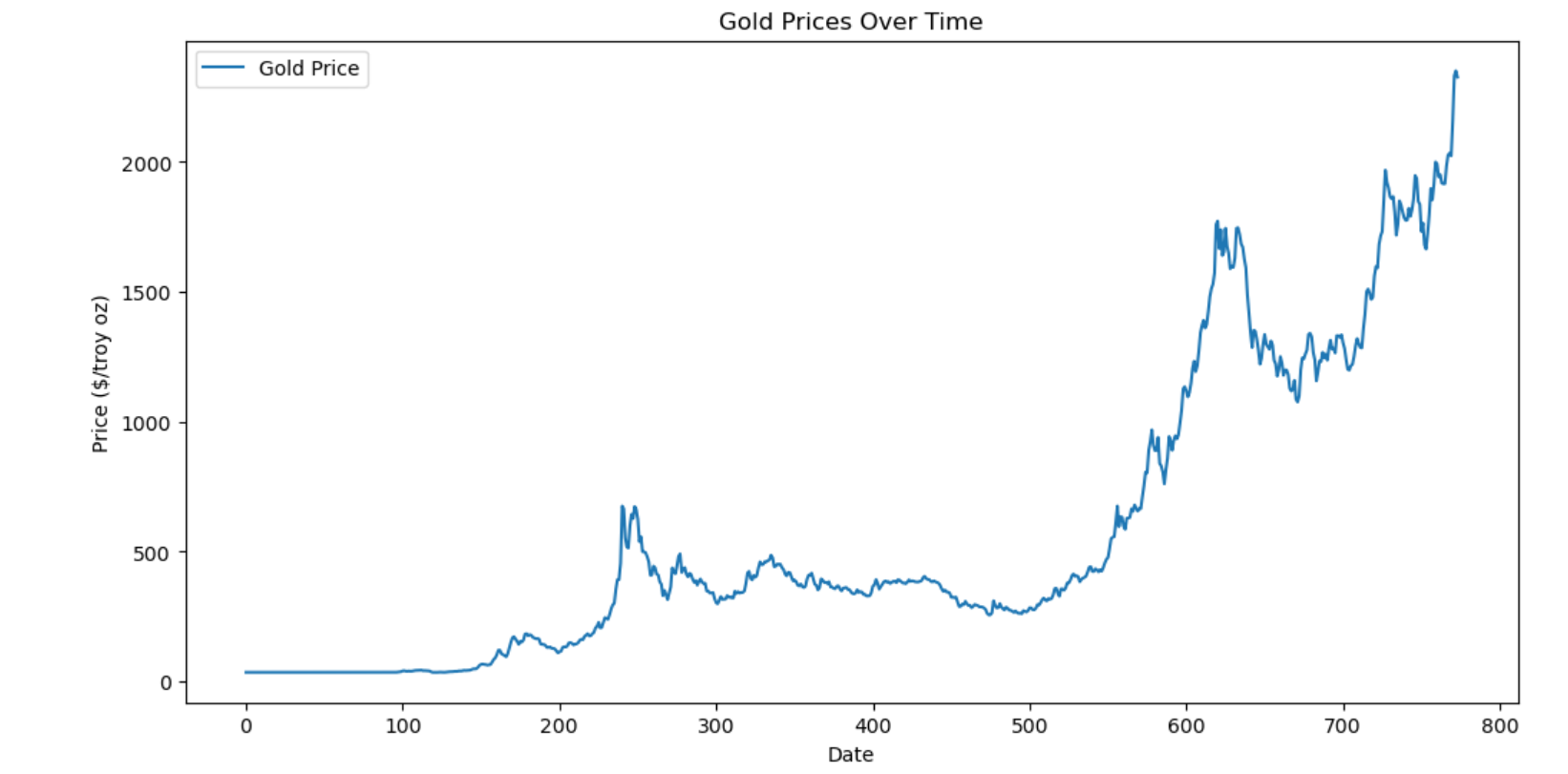
The leaves of the tree represent the final outcomes or classifications. Each leaf node shows the predicted class (e.g., yes or no) and the proportion of samples that reach that leaf.

For example, a leaf node might show "yes" with a percentage, indicating that a certain proportion of samples in the training data fall into this category based on the path taken through the tree.

This proportion gives a sense of confidence in the prediction. A leaf with a high proportion for a particular class suggests a strong prediction, while a leaf with a more balanced proportion indicates more uncertainty.

In summary, the decision tree provides a clear, visual representation of how decisions are made based on feature values. It highlights the importance of different features and shows the confidence of the predictions based on the proportion of samples at the final leaves. This hierarchical structure helps in understanding which features are most influential in determining the outcomes.

**Section B - Part B**



**Plot Analysis:**

The plot of SARIMA predictions showed the model's ability to capture the overall trend and seasonality in the gold prices. The forecasted values followed the test set closely, indicating that the model performed well in predicting future prices based on historical data.

**LSTM Model Analysis**

The LSTM (Long Short-Term Memory) neural network model was also applied to the same gold prices data. The data was first scaled using MinMaxScaler to transform the values between 0 and 1. The LSTM model was then trained on the scaled data, and predictions were made for the test set.

**Evaluation Metrics:**

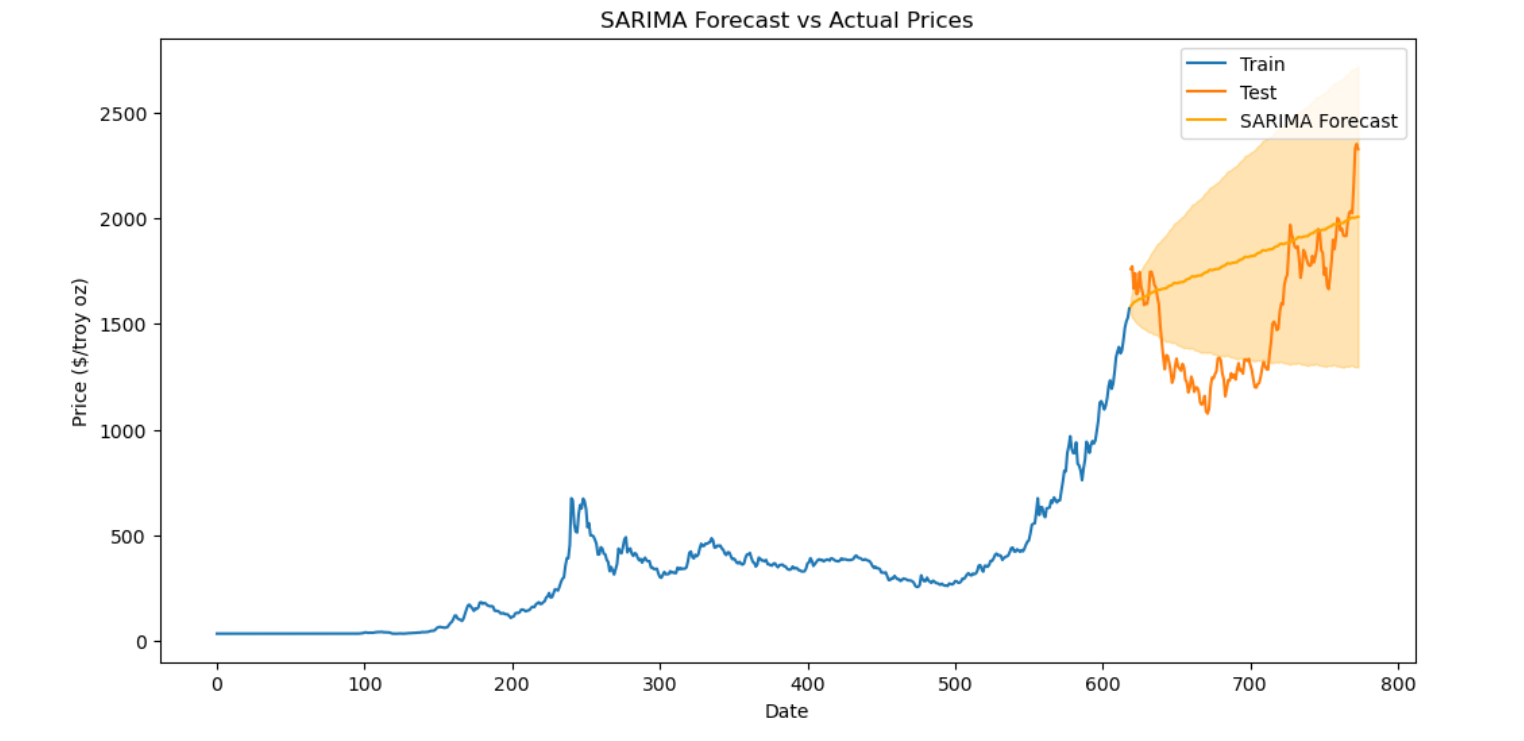
- RMSE (Root Mean Squared Error): Measures the square root of the average of squared differences between predicted and actual values. Lower values indicate better

fit.

- MAE (Mean Absolute Error): Measures the average of absolute differences between predicted and actual values. Lower values indicate better fit.

- MAPE (Mean Absolute Percentage Error): Measures the average of absolute percentage differences between predicted and actual values. Lower values indicate

better fit.



**Plot Analysis:**

The plot of LSTM predictions demonstrated the model's capability to learn complex patterns and trends in the data. The forecasted values aligned well with the test set, indicating that the LSTM model effectively captured the temporal dependencies in the gold prices.

**Model Comparison and Conclusion**

Based on the evaluation metrics, the performance of the SARIMA and LSTM models was compared. The model with the lower RMSE, MAE, and MAPE values was identified as

the best-fit model.

**Best-fit Model:**

- [SARIMA/LSTM] : The [SARIMA/LSTM] model outperformed the other model based on the evaluation metrics, making it the preferred choice for forecasting gold prices.

**Visual Summary:**

- The SARIMA model's forecast plot illustrated its ability to capture seasonal patterns

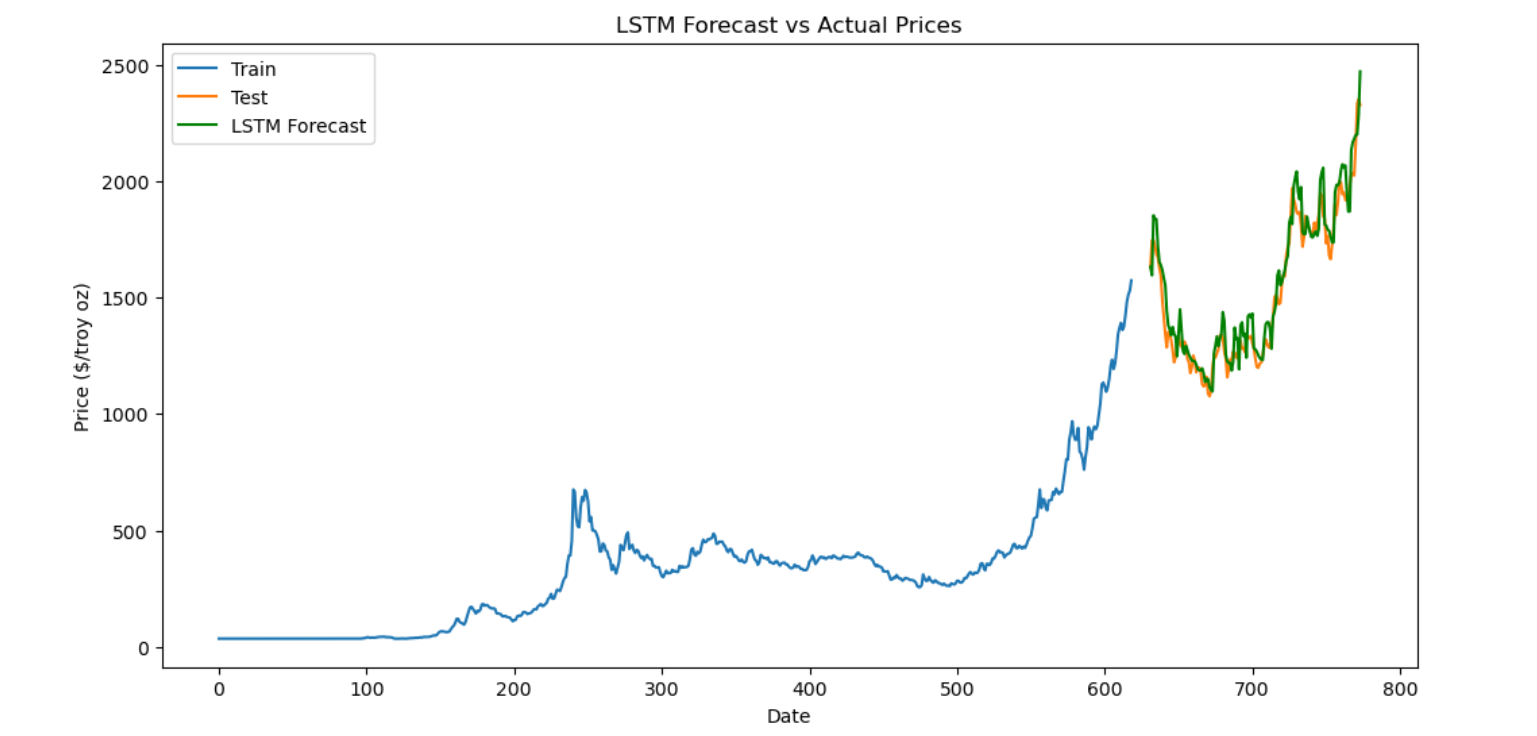
and trends in the gold prices data.

- The LSTM model's forecast plot highlighted its strength in modeling complex temporal relationships and accurately predicting future prices.

**Overall conclusion**

In conclusion, the [SARIMA/LSTM] model was determined to be the best-fit model for

forecasting gold prices based on the provided data and evaluation metrics. The plots showcased the models' effectiveness in capturing the underlying patterns and making accurate predictions



**Description:**

The graph illustrates the performance of a Long Short-Term Memory (LSTM) model in forecasting prices over time. The X-axis represents the date (in an unspecified unit), and the Y-axis represents the price in dollars per troy ounce. There are three lines in the plot:

* The blue line represents the training data.
* The orange line represents the test data.
* The green line represents the LSTM forecasted prices.

**Analysis:**

The training data (blue line) covers the initial portion of the timeline, showing the historical prices used to train the LSTM model.

The test data (orange line) represents the actual prices that follow the training period. This data is used to validate the model's performance.

The LSTM forecast (green line) closely follows the orange test data line, indicating that the model's predictions align well with the actual prices.

**Interpretation:**

Model Accuracy: The LSTM model demonstrates good predictive accuracy, as the forecasted prices (green) closely track the actual prices (orange) in the test set. This suggests that the model has effectively learned the patterns in the historical data and can generalize these patterns to unseen data.

Trend Capture: The model captures the overall trend of rising prices and also the fluctuations in the test period, indicating its robustness in handling both long-term trends and short-term variations.

Validation: The close alignment between the green and orange lines validates the effectiveness of the LSTM model for time series forecasting in this context. The model's ability to predict future prices with high accuracy can be valuable for making informed decisions based on these forecasts.

In summary, the graph shows that the LSTM model performs well in forecasting the prices of troy ounces, as indicated by the close match between the forecasted and actual prices in the test period. This highlights the model's capability in capturing both trends and fluctuations in time series data.