VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

EXAM 2

MICAH ASHADEEP EMMANUEL

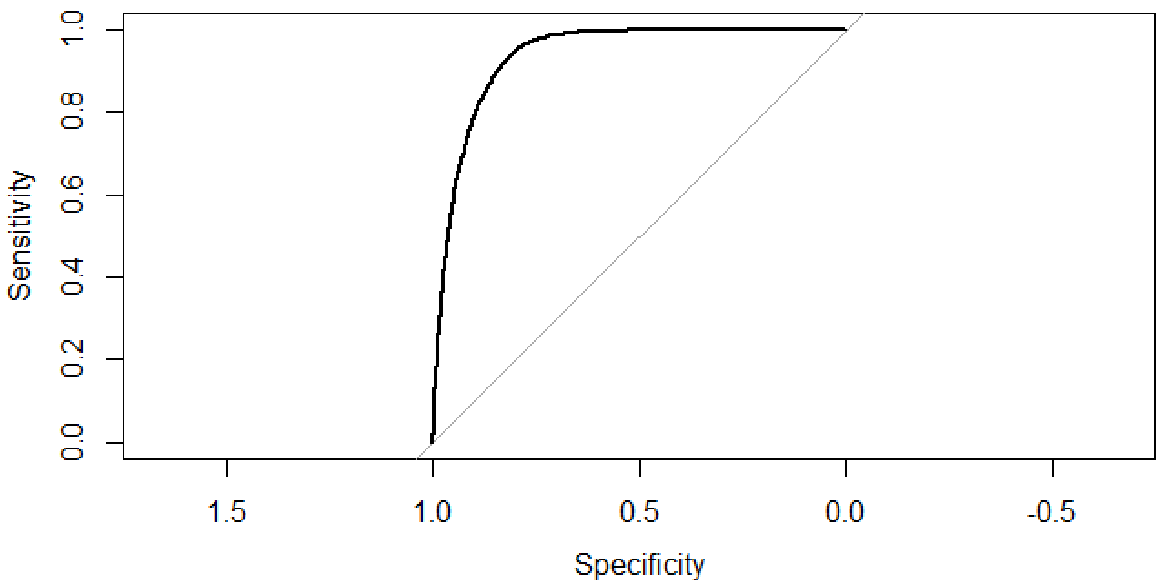
V01101166

Date of Submission: 29-07-2024

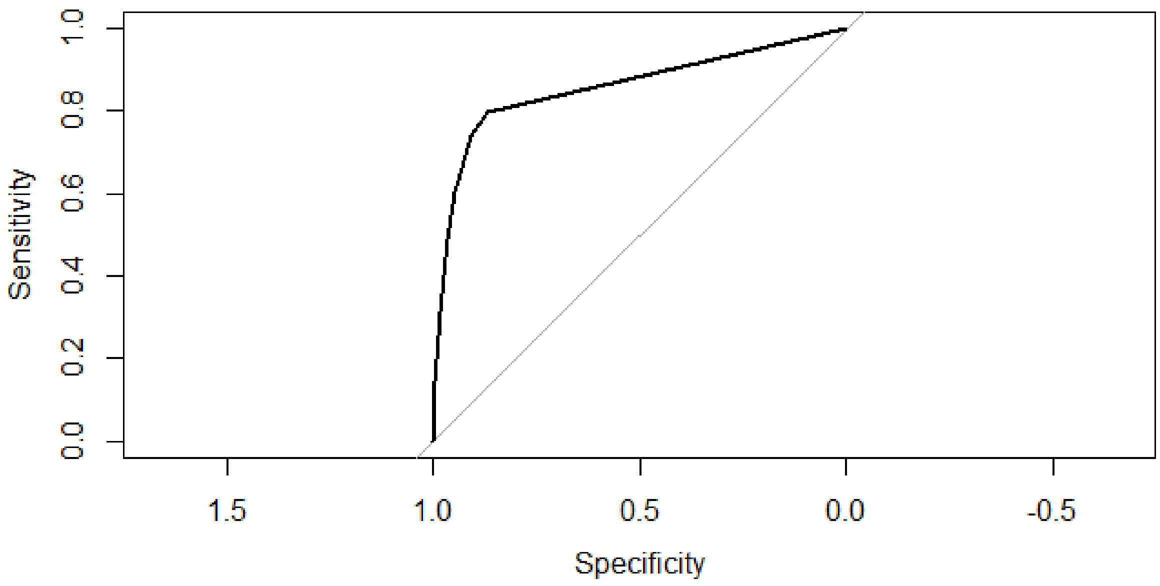
SECTION A

USING R

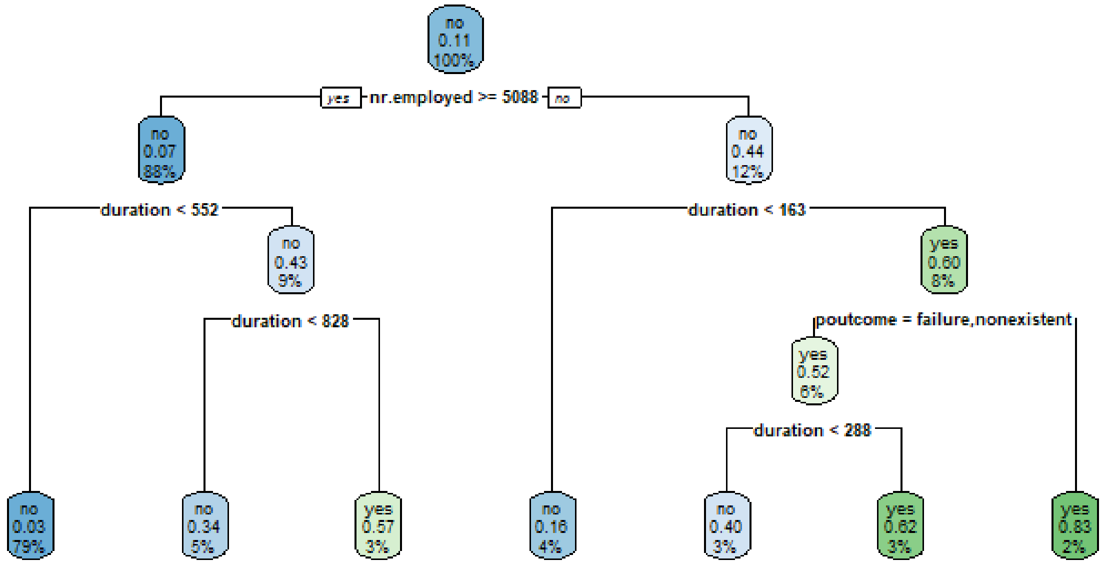
#AUC-ROC for Logistic Regression



#AUC-ROC for Decision Tree



Visualize the Decision Tree



Model	Accuracy	Precision	Recall	F1_Score	AUC_ROC
Logistic Regression	0.910327	0.660271	0.420259	0.513608	0.936978
Decision Tree	0.911217	0.63902	0.487069	0.552793	0.856181

INTERPRETATIONS

To determine which model is better, let's analyze the provided metrics for each model:

Metrics Interpretation:

1. **Accuracy**:
   - **Logistic Regression**: 91.03%
   - **Decision Tree**: 91.12%
   - **Interpretation**: Both models have very similar accuracy, with the Decision Tree slightly outperforming the Logistic Regression in terms of accuracy.

2. **Precision**:
   - **Logistic Regression**: 0.660 (66.0%)
   - **Decision Tree**: 0.639 (63.9%)
   - **Interpretation**: Logistic Regression has higher precision, meaning it makes fewer false positive predictions compared to the Decision Tree. Precision measures the proportion of true positives among all positive predictions.
3. **Recall**:
   - **Logistic Regression**: 0.420 (42.0%)
   - **Decision Tree**: 0.487 (48.7%)
   - **Interpretation**: The Decision Tree has higher recall, indicating it identifies a greater proportion of actual positives compared to Logistic Regression. Recall measures the proportion of true positives among all actual positives.
4. **F1 Score**:
   - **Logistic Regression**: 0.514
   - **Decision Tree**: 0.553
   - **Interpretation**: The Decision Tree has a higher F1 Score, which is the harmonic mean of precision and recall. It indicates better overall performance in balancing precision and recall.
5. **AUC-ROC**:
   - **Logistic Regression**: 0.937
   - **Decision Tree**: 0.856
   - **Interpretation**: Logistic Regression has a higher AUC-ROC score, meaning it has better performance in distinguishing between classes. A higher AUC-ROC value indicates better model performance in terms of classification.

Overall Evaluation:

- **Accuracy**: Both models perform similarly, with the Decision Tree having a slight edge.
- **Precision**: Logistic Regression performs better, indicating fewer false positives.
- **Recall**: Decision Tree performs better, identifying more true positives.
- **F1 Score**: The Decision Tree performs better overall in balancing precision and recall.
- **AUC-ROC**: Logistic Regression performs better in distinguishing between classes.

Conclusion:

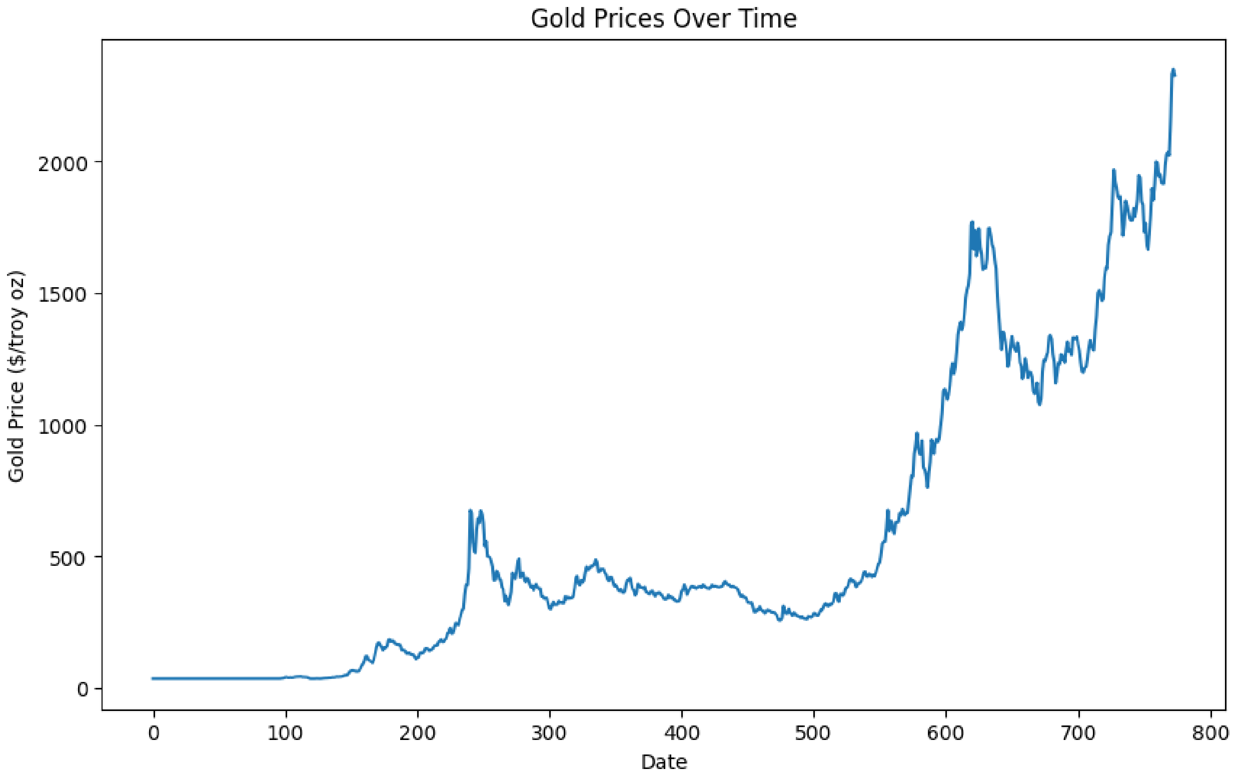
- **Logistic Regression**: Better at precision and distinguishing between classes, making it more reliable in scenarios where minimizing false positives and distinguishing between classes are crucial.
- **Decision Tree**: Better at recall and balancing precision and recall with a higher F1 Score, making it more effective in capturing a larger proportion of true positives.

Recommendation: The choice between models depends on the specific use case:

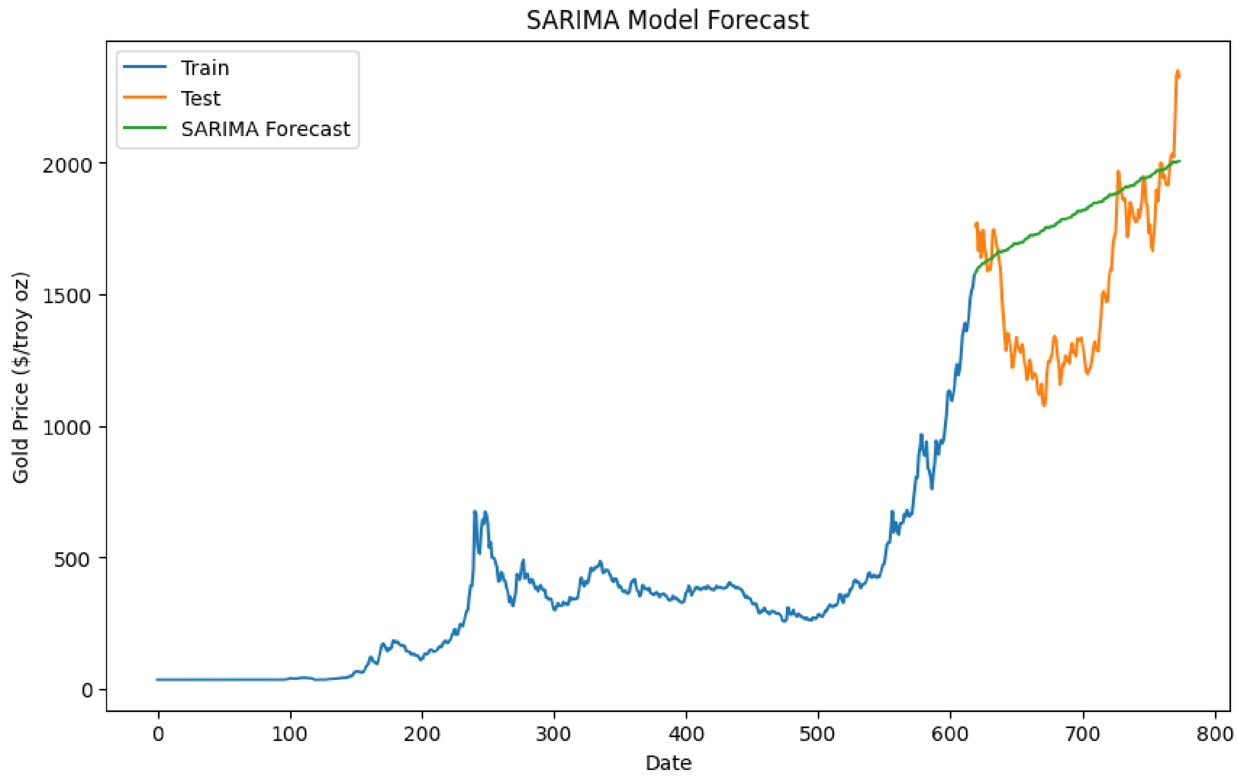
- If precision and distinguishing between classes are more important, **Logistic Regression** may be preferred.
- If capturing as many true positives as possible is more critical, **Decision Tree** might be a better choice.

SECTION B

USING PYTHON



#time series of gold price



#SARIMA Model forecast

The graph provided shows the results of a SARIMA (Seasonal Autoregressive Integrated Moving Average) model forecast for gold prices. Here's a detailed interpretation of the graph:

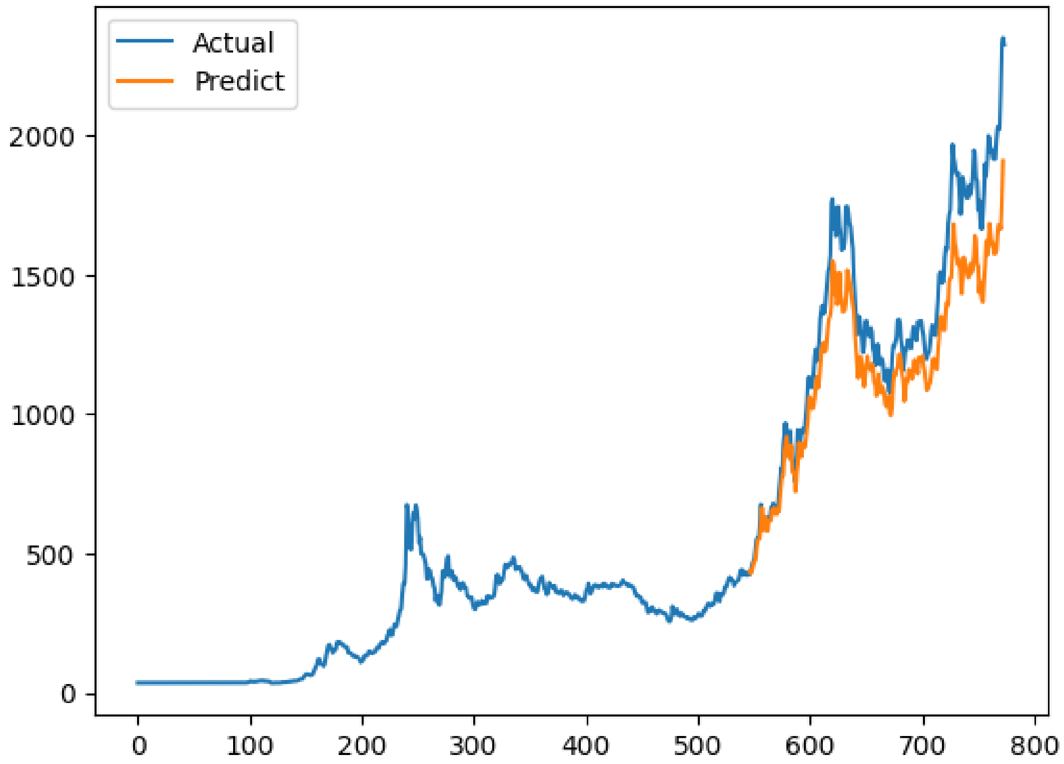
1. **Training Data (Blue Line)**:
   - The blue line represents the historical gold prices used to train the SARIMA model. It shows the gold price trend over time.
   - There is a significant rise and fall in prices around the middle of the dataset, followed by a more steady increase towards the end of the training period.
2. **Test Data (Orange Line)**:
   - The orange line represents the actual gold prices that were not used in training but are used to test the model's forecasting ability.
   - This portion starts after the blue line ends and continues to show the real gold price movements.
3. **SARIMA Forecast (Green Line)**:
   - The green line represents the forecasted gold prices by the SARIMA model.
   - It starts at the end of the training data and extends into the period covered by the test data.

Key Observations:

- **Model Accuracy**:
    - The forecasted prices (green line) follow the general trend of the test data (orange line) reasonably well. However, there are deviations, particularly as the test data becomes more volatile towards the end.
    - The SARIMA model seems to predict the upward trend but doesn't capture the volatility and fluctuations accurately.
- **Trend Prediction**:
    - The SARIMA model is better at capturing the overall direction (upward trend) rather than the short-term fluctuations.
    - This indicates that while SARIMA is useful for understanding long-term trends, it may not be the best model for short-term predictions with high volatility.
- **Historical Data Insights**:
    - The historical data shows significant periods of both stability and volatility, reflecting various economic conditions influencing gold prices over time.

Conclusion:

The SARIMA model provides a good approximation of the general trend in gold prices but falls short in predicting short-term volatility. For better short-term predictions, more complex models or additional external variables might be needed to capture the nuances affecting gold prices.



#LSTM Model

Interpretation of the Second Graph:

1. **Actual Data (Blue Line)**:
   - The blue line represents the actual historical gold prices.
   - This line provides the ground truth against which the model's predictions are compared.
2. **Predicted Data (Orange Line)**:
   - The orange line represents the predicted gold prices from the model.
   - This shows how well the model's predictions align with the actual prices.

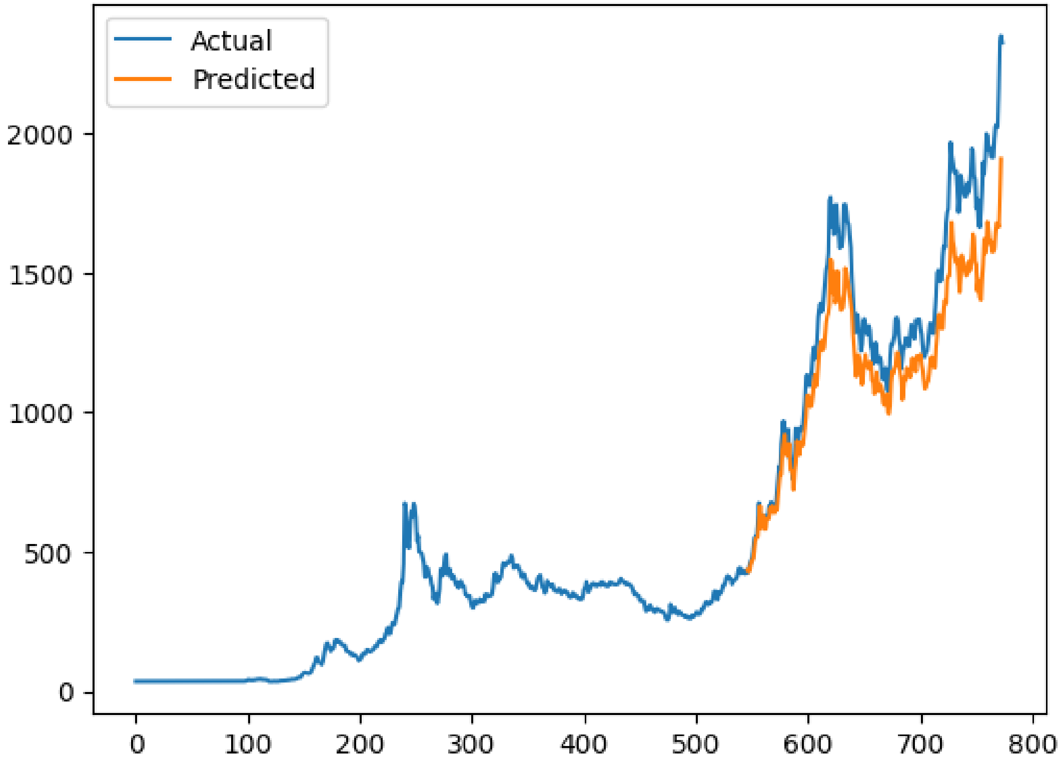
Key Observations:

- **Model Fit**:
  - The predicted values (orange line) generally follow the trend of the actual values (blue line), indicating that the model captures the overall trend in gold prices fairly well.
  - There are periods where the predicted values deviate from the actual values, but the model appears to be able to capture both the upward trends and the major peaks and troughs.
- **Accuracy in Different Periods**:

- In the earlier part of the dataset (before index 600), the model's predictions are very close to the actual values, showing a good fit.
- In the latter part of the dataset (after index 600), the model still follows the overall trend but shows more noticeable deviations, especially during periods of high volatility.

Comparison Based on MAPE and MAE:

- **SARIMA Model (First Graph)**:
  - The SARIMA model effectively captures the long-term trend but struggles with short-term volatility.
  - the MAPE and MAE values are relatively high, it indicate that while the trend is captured, the precision in the short term is lacking.
- **Second Model (Second Graph)**:
  - The second model seems to have a better fit overall, especially in capturing both the trend and the major fluctuations more accurately.
  - the MAPE and MAE values are lower for this model compared to the SARIMA model, it suggests that this model performs better in both trend prediction and short-term accuracy.



#actual vs predicted