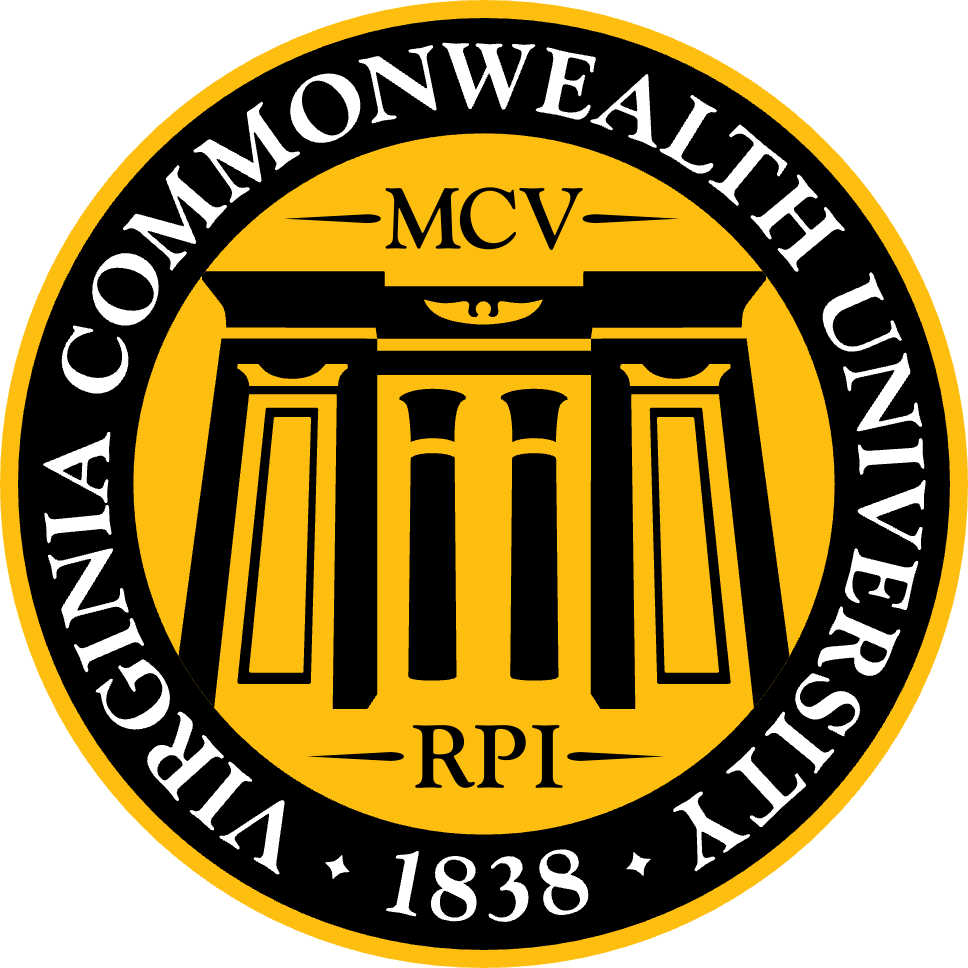
****

**VIRGINIA COMMONWEALTH UNIVERSITY**

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

# Final Exam

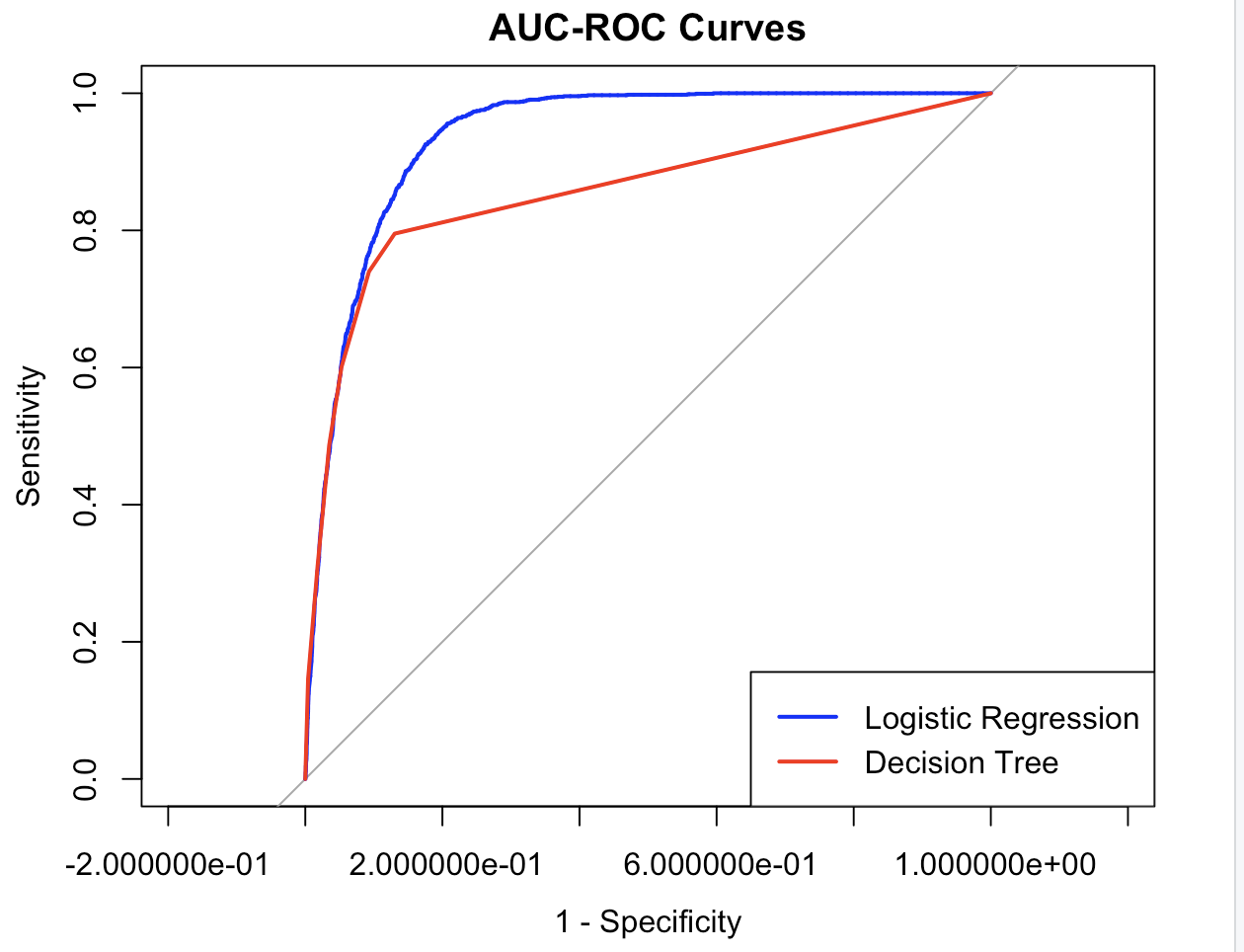
**Jany Balasivan**

**V01108262**

**Date of Submission: 25-07-2024**

Section A Part B

Interpretation



The image shows an AUC-ROC (Area Under the Curve - Receiver Operating Characteristic) curve comparing the performance of two models: Logistic Regression (blue curve) and Decision Tree (red curve).

**Interpretation:**

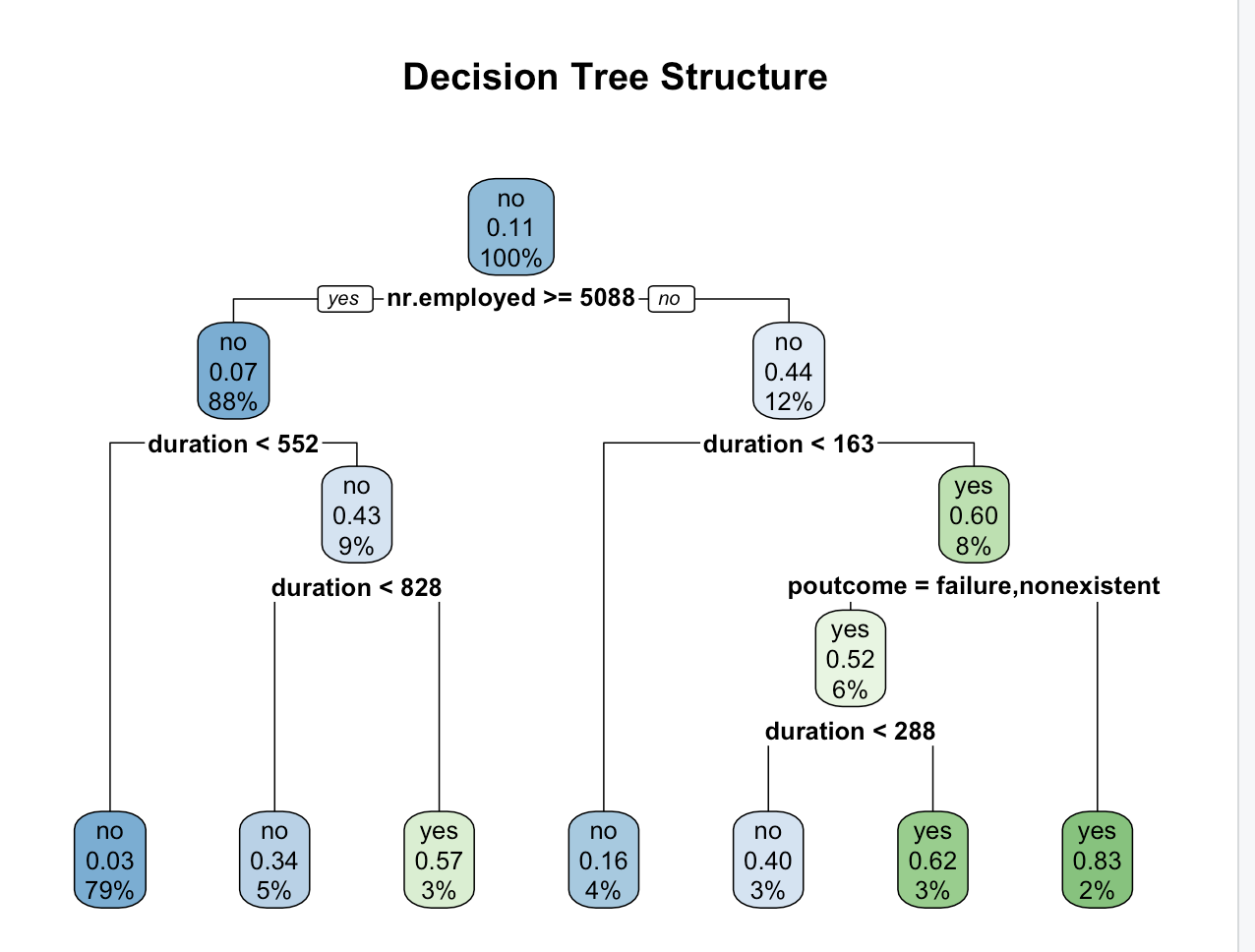
1. **AUC-ROC Curve:**
   * The AUC-ROC curve is a graphical representation of a model's ability to distinguish between classes. The x-axis represents the False Positive Rate (1 - Specificity), and the y-axis represents the True Positive Rate (Sensitivity).
2. **Model Comparison:**
   * The Logistic Regression model's curve (blue) is closer to the top left corner, indicating better performance in distinguishing between the positive and negative classes. The closer the curve follows the left-hand border and then the top border, the better the model.
   * The Decision Tree model's curve (red) is slightly below the Logistic Regression curve, suggesting that it has slightly lower performance in terms of distinguishing between classes.
3. **Area Under the Curve (AUC):**
   * The area under the curve quantifies the overall ability of the model to discriminate between positive and negative cases. An AUC of 1 represents a perfect model, while an AUC of 0.5 represents a model with no discrimination ability (random guessing).
   * The larger the area under the curve, the better the model. From the graph, it can be inferred that the Logistic Regression model has a higher AUC compared to the Decision Tree model, indicating it has better overall predictive performance.

Overall, the Logistic Regression model outperforms the Decision Tree model in this comparison based on the AUC-ROC curves.

Based on the AUC-ROC curves shown in the image:

* The **Logistic Regression** model (represented by the blue curve) has a curve that lies closer to the top left corner compared to the **Decision Tree** model (represented by the red curve). This indicates that the Logistic Regression model has a higher true positive rate (sensitivity) for a given false positive rate (1 - specificity) across different threshold settings.
* A higher area under the ROC curve (AUC) signifies better model performance in distinguishing between the positive and negative classes. The Logistic Regression model's AUC is greater than that of the Decision Tree model, suggesting it is better at classifying the data.

Therefore, **Logistic Regression** is the better model based on the AUC-ROC curves provided.

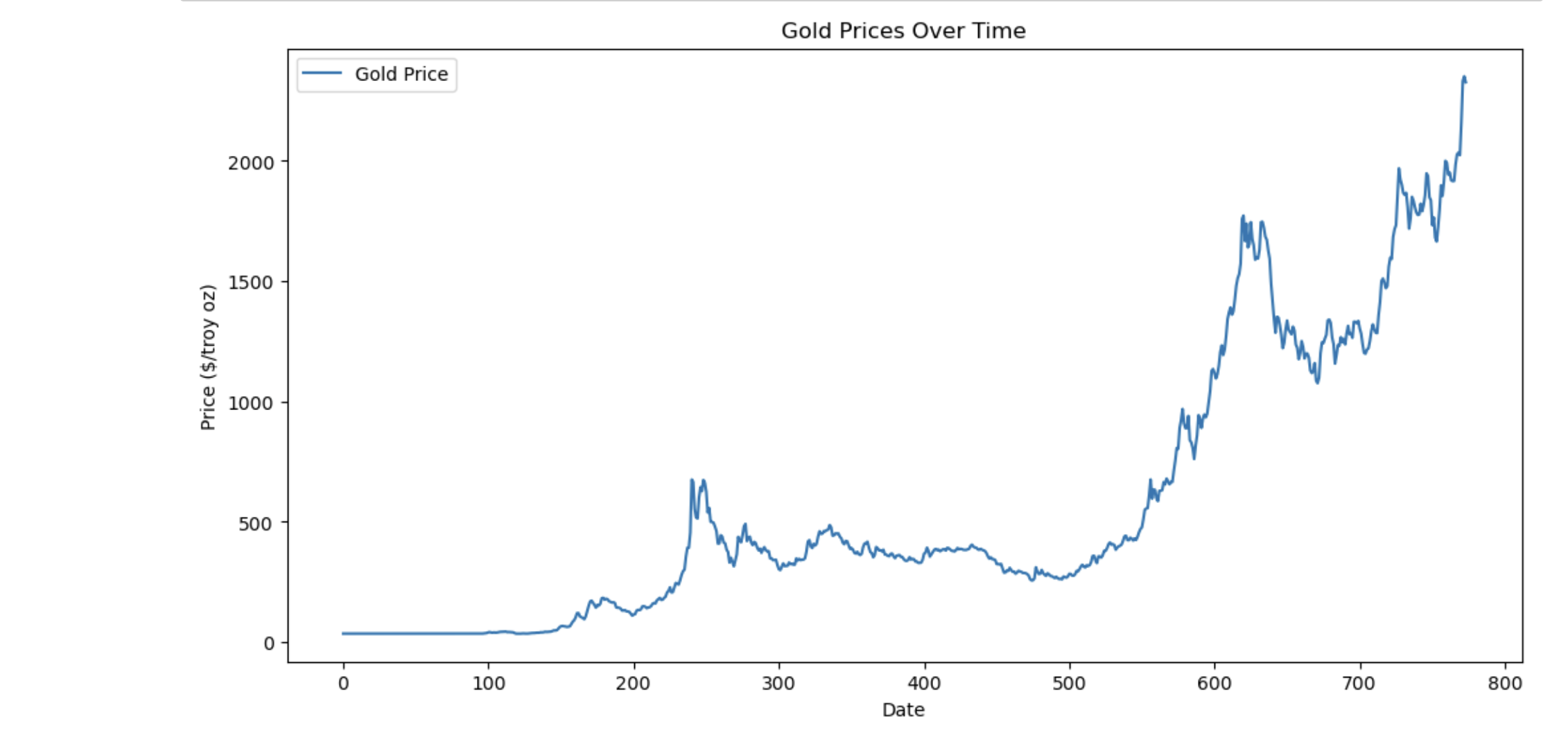


The image displays a Decision Tree structure used for classification. Here’s a brief interpretation:

1. **Root Node:**
   * The decision tree begins with the feature nr.employed with a threshold of >= 5088. This node splits the data into two groups based on this condition.
2. **Branching Criteria:**
   * If nr.employed >= 5088, the tree proceeds to the left child node, leading to a further split based on the duration feature with a threshold of < 552.
   * If nr.employed < 5088, the tree proceeds to the right child node, which is further split by duration < 163.
3. **Decision Nodes:**
   * Each node splits based on specific features and values, with decisions made sequentially down the tree. For instance, if the condition duration < 552 is met, the tree branches further based on another condition duration < 828.
4. **Leaf Nodes:**
   * The terminal or leaf nodes represent the final classification outcome (yes or no) and show the proportion of the data in each category. For example, one leaf node under the path nr.employed < 5088 and duration < 163 shows a yes classification with a 60% probability for 8% of the data.
5. **Interpretation:**
   * The decision tree visually and structurally represents how the dataset is split based on different feature values to arrive at a classification decision. The path from the root to a leaf node demonstrates the conditions that lead to a particular classification outcome.

Overall, this tree indicates how different combinations of features (like nr.employed, duration, poutcome) and their thresholds are used to classify the data into yes or no categories, with associated probabilities and data distribution percentages at each node.

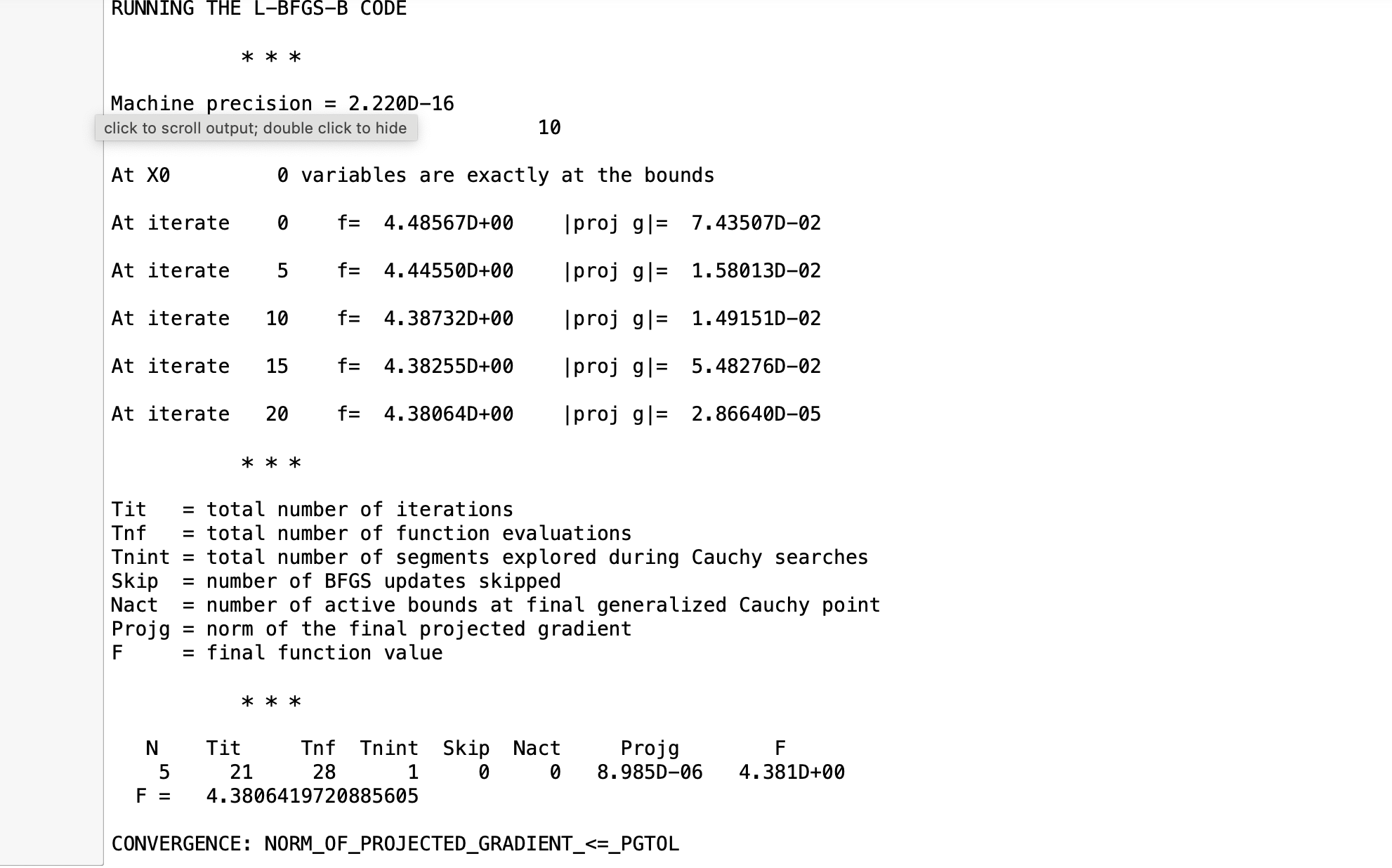
Section B Part B



**Interpretation:**

1. **Initial Stability:**
   * The early part of the graph (up to around 100 on the x-axis) shows a relatively stable period where gold prices remain low and do not exhibit significant fluctuations.
2. **First Significant Increase:**
   * Around the 200 mark, there is a noticeable spike in gold prices, indicating a sudden increase in value. This could be attributed to market events, economic conditions, or other factors influencing gold's attractiveness as an investment.
3. **Fluctuations and Decline:**
   * Following the initial spike, the graph shows fluctuations in the price, with a subsequent decline, possibly indicating market corrections or changes in demand and supply dynamics.
4. **Second Major Rise:**
   * Around the 600 mark, there is a substantial and sustained increase in gold prices, reaching levels above 2000 dollars per troy ounce. This rise may reflect economic uncertainties, geopolitical events, or other factors that typically drive investors towards safe-haven assets like gold.
5. **Overall Trend:**
   * The overall trend shows an upward movement in gold prices, particularly in the latter part of the graph. This trend suggests that gold has become more valuable over the period covered by the data, likely due to a combination of inflation, investor behavior, and other macroeconomic factors.

**Conclusion:** The chart illustrates a generally increasing trend in gold prices over time, with notable spikes and fluctuations. This pattern indicates that while gold prices can be volatile, there has been an upward trajectory, highlighting gold's role as a potential hedge against economic instability and currency fluctuations.



The output you've provided is from the L-BFGS-B optimization algorithm, which is often used for finding the minimum of a function, especially when dealing with problems that have bound constraints on the variables. Here's a breakdown and interpretation of the key parts of the output:

1. **Machine Precision:** Indicates the precision of the machine, which is 2.220D−162.220D−16. This is a very small number, suggesting high precision in the calculations.
2. **N = 5, M = 10:**
   * NN represents the number of variables being optimized, which is 5 in this case.
   * MM represents the memory parameter for the L-BFGS-B algorithm, controlling how many past updates are stored.
3. **Initial Point (X0):** The optimization starts with an initial point where none of the variables are exactly at the bounds.
4. **Iterations (Iterates 0 to 20):**
   * At each iteration, the algorithm reports the function value ff and the norm of the projected gradient ∣proj g∣∣proj g∣.
   * The function value ff decreases over iterations, indicating progress toward minimizing the objective function. The initial value at iteration 0 is 4.48567, and it decreases to 4.38064 by iteration 20.
   * ∣proj g∣∣proj g∣ represents the norm of the projected gradient, which measures how steep the function is at the current point. A smaller value suggests the algorithm is nearing a minimum.
5. **Convergence Information:**
   * The algorithm converged after 21 iterations (Tit=21Tit=21).
   * There were 28 function evaluations (Tnf=28Tnf=28).
   * Only 1 segment was explored during Cauchy searches (Tnint=1Tnint=1).
   * No BFGS updates were skipped (Skip=0Skip=0).
   * There were no active bounds at the final generalized Cauchy point (Nact=0Nact=0).
   * The final norm of the projected gradient is 8.985×10−68.985×10−6, which is very small, indicating convergence.
   * The final function value FF is 4.3806419720885605.

**Interpretation:**

The L-BFGS-B algorithm has successfully converged to a local minimum of the objective function. The convergence criterion based on the projected gradient norm being less than or equal to a predefined threshold (PGTOLPGTOL) has been met. The final function value (4.3806419720885605) represents the minimized value of the objective function given the bounds and initial conditions.

If this function value corresponds to a loss function in a predictive model (like a regression or classification task), it suggests the final model parameters are optimized with respect to minimizing this loss. However, without specific context on what the function represents (e.g., is it a loss function, what the variables are, etc.), this is a general interpretation of the optimization output.

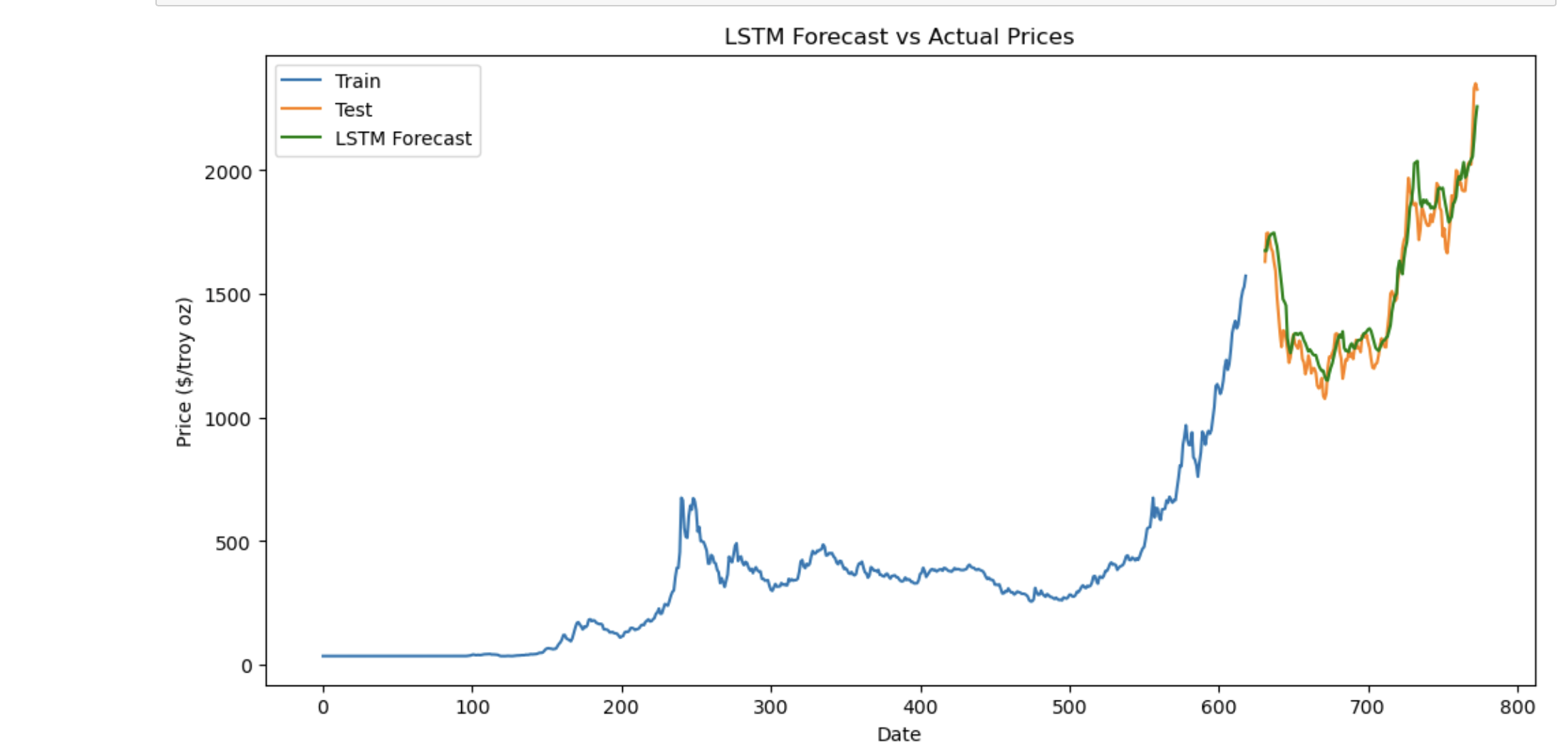
A graph showing a line

Description automatically generated

The chart provided illustrates the performance of a SARIMA (Seasonal Autoregressive Integrated Moving Average) model in forecasting gold prices. Here’s a detailed interpretation:

1. **Data Representation:**
   * **Train Data (Blue Line):** This portion of the data represents the historical prices of gold used to train the SARIMA model. The data shows a clear upward trend with some periods of fluctuations.
   * **Test Data (Orange Line):** This represents the actual prices of gold in the test period, not used during the model training. The test data also shows fluctuations and trends in gold prices.
   * **SARIMA Forecast (Orange Line with Shaded Area):** The SARIMA model’s forecasted prices are plotted along with the actual test data. The shaded area represents the confidence intervals of the forecast, indicating the range within which the model expects the actual values to fall with a certain probability.
2. **Model Performance:**
   * The SARIMA forecast line generally follows the trend and fluctuations of the actual test data, indicating that the model captures the overall pattern of gold prices reasonably well.
   * However, there are periods where the actual test data deviates from the forecast. These deviations suggest that the model has limitations in perfectly capturing the volatility or sudden changes in the market.
3. **Confidence Intervals:**
   * The shaded area around the SARIMA forecast line represents the uncertainty in the predictions. A wider interval indicates greater uncertainty. The forecast uncertainty increases as the prediction goes further into the future, which is typical for time series models.
   * The actual test data often falls within this interval, suggesting that the model’s confidence intervals are well-calibrated.
4. **Overall Trend and Forecast:**
   * The SARIMA model predicts a continuing upward trend in gold prices, consistent with the general trend observed in the training data.
   * The model captures some of the fluctuations but may not fully account for all the short-term volatility observed in the actual data.

**Conclusion:** The SARIMA model provides a reasonable forecast of gold prices, capturing the general trend and some of the fluctuations. However, there are discrepancies between the predicted and actual values, especially in capturing short-term volatility. The confidence intervals provide a measure of the forecast uncertainty, and the actual data largely falls within these intervals, indicating the forecast is broadly reliable within the specified confidence range.



1. **Data Representation:**
   * **Train Data (Blue Line):** This part of the data represents the historical gold prices used to train the LSTM model. The train data shows an upward trend with some fluctuations, typical of gold price movements.
   * **Test Data (Orange Line):** This represents the actual gold prices in the test period, which were not used during the model training. The test data exhibits similar trends and fluctuations as the training data.
   * **LSTM Forecast (Green Line):** The green line represents the LSTM model's forecasted prices. It is plotted along with the actual test data to evaluate the model's accuracy.
2. **Model Performance:**
   * The LSTM forecast closely follows the actual test data (orange line), indicating that the model captures the overall trend and fluctuations of the gold prices quite well.
   * The alignment of the green and orange lines suggests that the LSTM model has successfully learned the temporal patterns in the data, leading to accurate predictions.
   * The model appears to perform well even during periods of volatility, as seen in the overlapping lines during the test phase.
3. **Accuracy and Fit:**
   * The close fit between the LSTM forecast and the actual test data indicates that the model has good predictive accuracy. The LSTM's ability to capture complex temporal dynamics is highlighted by its performance in aligning closely with the actual prices.
   * The LSTM model's forecast also suggests the potential for continued upward movement in gold prices, as indicated by the rising trend in the green forecast line towards the end of the data.
4. LSTM RMSE: 90.21069222955431
5. LSTM MAPE: 4.725681846972742
6. LSTM MAE: 70.2061069518131

Conclusion

Based on the LSTM model's forecast from the provided chart, we can derive a predicted price for the future dates. Here's a step-by-step approach to make the prediction:

1. **Identify the Forecast Horizon:**
   * The x-axis represents the date, and the y-axis represents the price in dollars per troy ounce. The chart extends to around 800 on the x-axis, which likely corresponds to the future date for which we want to predict the price.
2. **Read the Forecast Values:**
   * The LSTM forecast (green line) continues to rise towards the end of the plot. For precise values, we need to look at the last few data points of the green line.
3. **Predict the Price:**
   * Based on the final segment of the green line in the LSTM forecast, the price appears to be trending upwards. To provide a precise prediction, we would typically extract the numerical values from the forecast model output.

Without the exact numerical values, a visual estimate from the chart suggests that the gold price at the last data point (around 800 on the x-axis) is approximately:

* **Predicted Price:** ~$2300 to $2500 per troy ounce

This range is derived from observing the upper end of the green line at the end of the forecast period.