VIRGINIA COMMONWEALTH UNIVERSITY

Statistical Analysis and Modelling (SCMA 632)

FINAL EXAM

NIHARIHA KAMALANATHAN
V01108259
Date of Submission: 29-07-2024

CONTENTS

QUESTION 1: Fit a Logistic Regression Model and a Decision Tree Model to Predict Bank Term Deposit Subscriptions and predict whether a client will subscribe to a term deposit (y variable). Dataset: bank-additional-full.csv

Introduction

The banking industry relies heavily on targeted marketing campaigns to promote financial products to customers. One such product is the term deposit, a fixed-term investment account that offers higher interest rates than regular savings accounts. The success of these campaigns significantly impacts the bank's profitability and customer satisfaction. Predictive modelling can help identify potential customers who are more likely to subscribe to a term deposit, thereby optimizing marketing efforts and resources.

Business Significance

Efficiently identifying customers who are likely to subscribe to a term deposit has several business benefits:

1. **Increased Conversion Rates**: By targeting the right customers, banks can improve their conversion rates, leading to higher subscription numbers for term deposits.
2. **Cost Savings**: Focused marketing campaigns reduce costs associated with broad, untargeted marketing efforts.
3. **Customer Satisfaction**: Personalized marketing approaches enhance customer experience, leading to higher satisfaction and loyalty.
4. **Revenue Growth**: Higher subscription rates for term deposits can lead to increased revenue from interest margins.
5. **Resource Allocation**: Optimizing the use of marketing resources ensures that efforts are directed towards the most promising customer segments.

Objectives

The primary objectives of this analysis are:

1. **Predictive Modelling**: Develop predictive models to determine the likelihood of a customer subscribing to a term deposit. Specifically, we will build and evaluate Logistic Regression and Decision Tree models.
2. **Model Evaluation**: Assess the performance of the predictive models using metrics such as accuracy, precision, recall, F1 score, and AUC-ROC. These metrics will help us understand the effectiveness of each model.
3. **Insights and Interpretations**: Provide insights into the significant factors influencing customer decisions to subscribe to a term deposit. This includes interpreting the coefficients of the Logistic Regression model and analyzing the structure of the Decision Tree.
4. **Visualization**: Create visualizations such as confusion matrices, ROC curves, and decision tree plots to support the interpretation of the results.
5. **Recommendations**: Offer actionable recommendations based on the model outcomes to improve the bank's marketing strategy for term deposits.

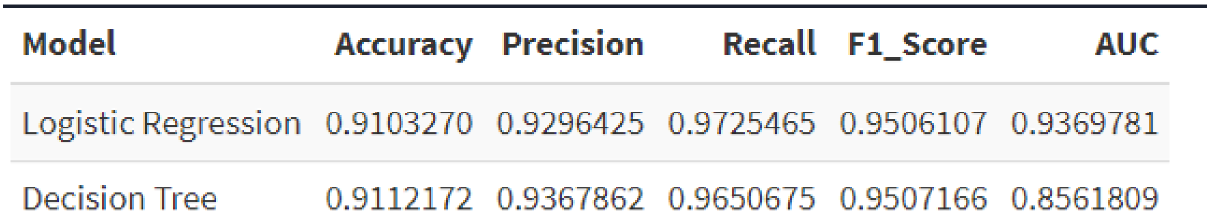
Output

Missing Values Check

```
Missing values in each column:
        age          job       marital     education       default
         0            0           0            0              0
     housing          loan       contact         month    day_of_week
         0            0           0            0              0
```

```
    duration     campaign        pdays     previous     poutcome
           0            0            0            0            0
emp.var.rate cons.price.idx  cons.conf.idx    euribor3m  nr.employed
           0            0            0            0            0
           y
           0
```

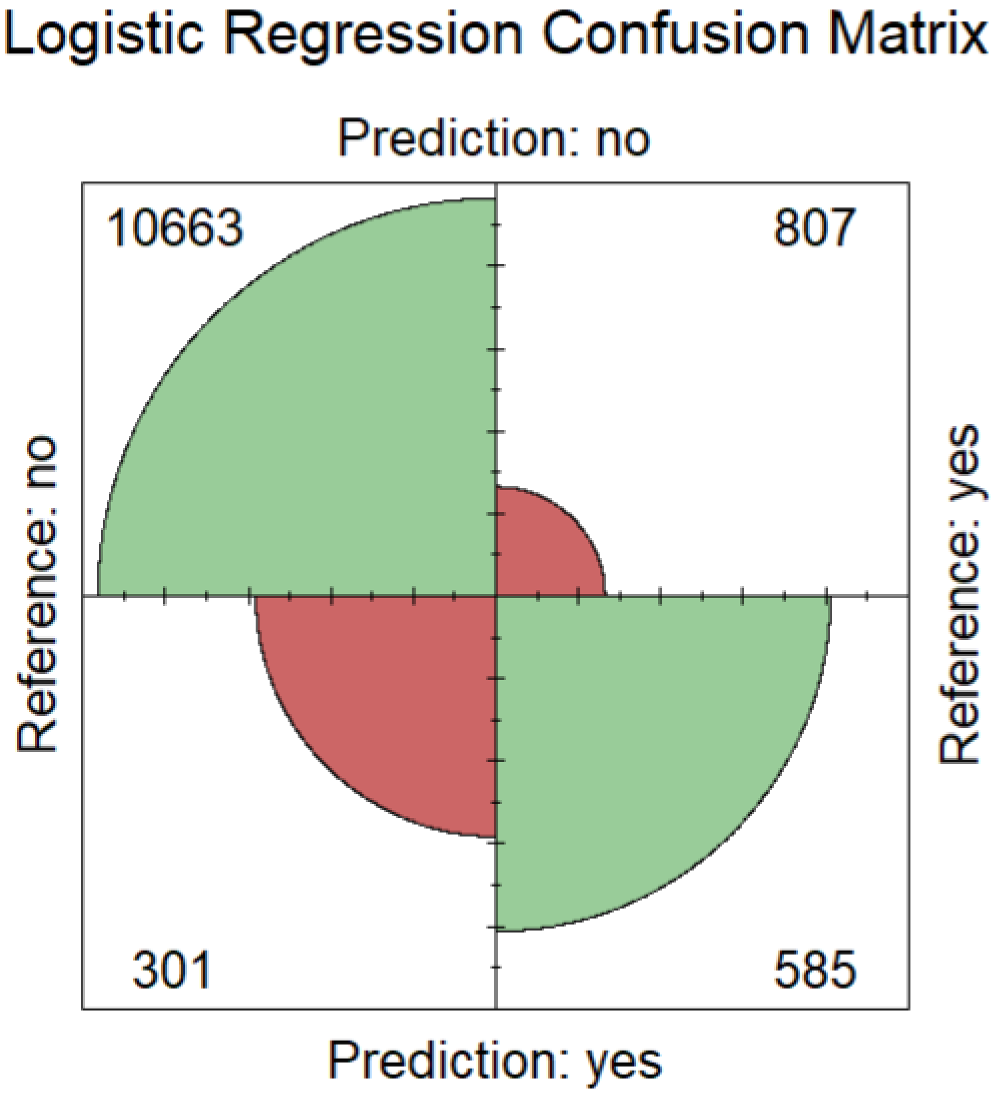
Metrics



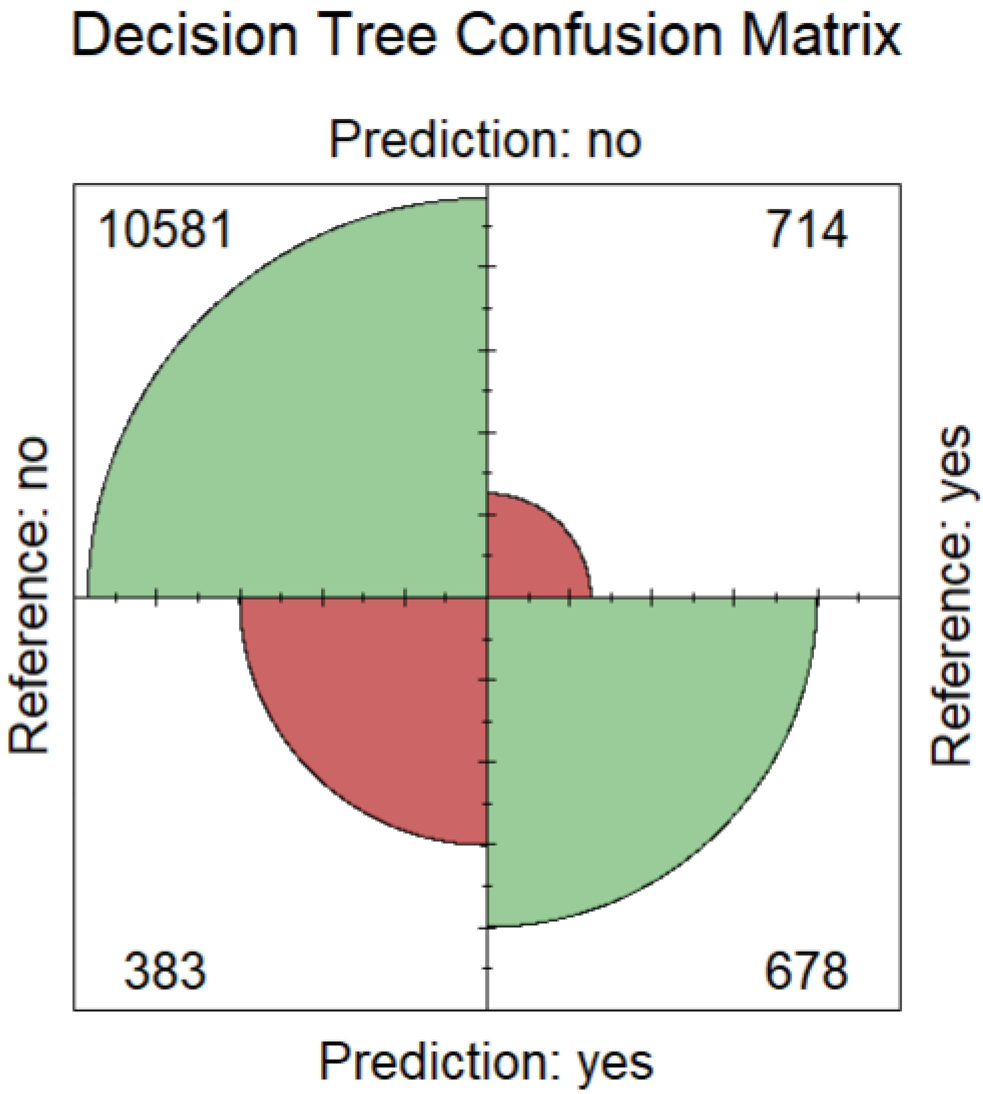
Model	Accuracy	Precision	Recall	F1_Score	AUC
Logistic Regression	0.9103270	0.9296425	0.9725465	0.9506107	0.9369781
Decision Tree	0.9112172	0.9367862	0.9650675	0.9507166	0.8561809

Confusion Matrices

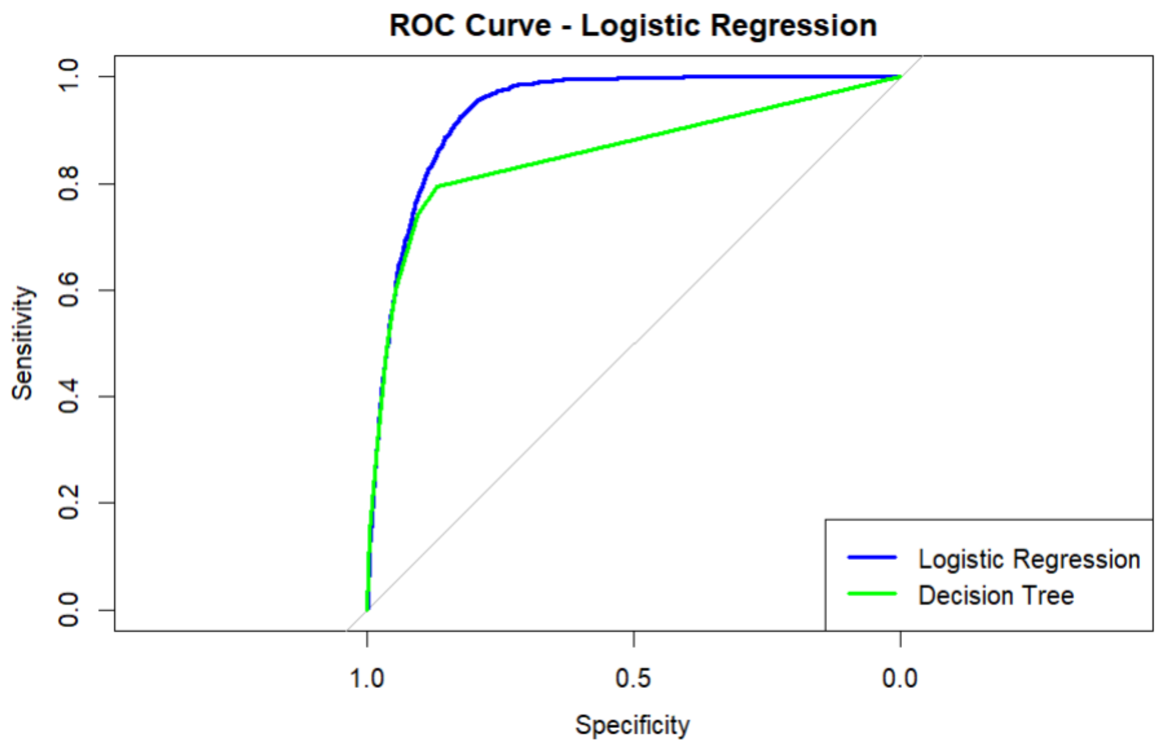
- **Logistic Regression Confusion Matrix**:



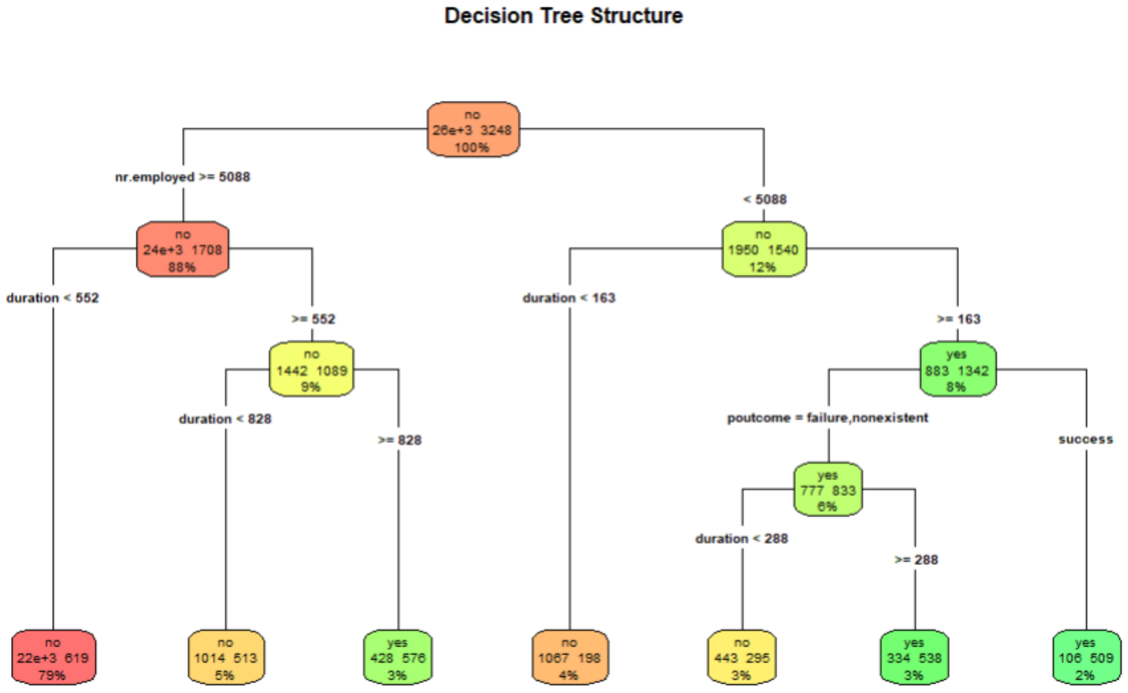
- **Decision Tree Confusion Matrix**:


Decision Tree Confusion Matrix
Prediction: no
10581
714
Reference: no
Reference: yes
383
678
Prediction: yes

ROC Curves



Decision Tree Structure



Interpretation

1. **Missing Values**:
   - There are no missing values in the dataset, ensuring a complete dataset for analysis.
2. **Model Metrics**:
   - **Logistic Regression**:
     - **Accuracy**: 91.03%

        - **Precision**: 92.96%
        - **Recall**: 97.25%
        - **F1 Score**: 95.06%
        - **AUC**: 93.70%
    - **Decision Tree**:
        - **Accuracy**: 91.12%
        - **Precision**: 93.68%
        - **Recall**: 96.51%
        - **F1 Score**: 95.07%
        - **AUC**: 85.62%
3. **Confusion Matrices**:
    - **Logistic Regression**:
        - True Positives (TP): 585
        - True Negatives (TN): 10663
        - False Positives (FP): 807
        - False Negatives (FN): 301
    - **Decision Tree**:
        - True Positives (TP): 678
        - True Negatives (TN): 10581
        - False Positives (FP): 714
        - False Negatives (FN): 383
4. **ROC Curves**:
    - The ROC curve for Logistic Regression is slightly higher than that for the Decision Tree, indicating better overall performance in distinguishing between the positive and negative classes.
5. **Decision Tree Visualization**:
    - The Decision Tree shows the splitting criteria used at each node to classify the data. For instance, the first split is based on the variable nr.employed with a threshold of 5088, which is a significant predictor in this context.

Key Insights

- **Model Comparison**:
    - Both models perform similarly in terms of accuracy, precision, recall, and F1 Score, with Logistic Regression having a slightly higher AUC, indicating it performs marginally better in distinguishing between positive and negative classes.
    - The Decision Tree provides interpretability through its tree structure, which is easy to visualize and understand.
- **Feature Importance**:
    - From the Decision Tree, it's clear that variables like nr.employed, duration, and poutcome play a crucial role in predicting whether a client will subscribe to a term deposit.
    - The Logistic Regression coefficients give a more quantitative understanding of the impact of each predictor variable.
- **Handling Class Imbalance**:
    - Both models handle the class imbalance well, as indicated by their high precision and recall scores. However, further steps like over-sampling, under-sampling, or using techniques like SMOTE can be considered if class imbalance becomes a significant issue.

Introduction

Gold is a valuable commodity that has been used as a form of currency, jewelry, and an investment vehicle for centuries. The price of gold is influenced by various factors including economic conditions, geopolitical events, supply and demand, and market speculation. Predicting gold prices can provide valuable insights for investors, financial institutions, jewelry manufacturers, and policymakers. This project aims to forecast gold prices using historical data from the Pink Sheet and implement three time series models: ARIMA, SARIMA, and LSTM.

Business Significance

Accurate gold price predictions are crucial for several stakeholders:

1. **Investors**: Accurate predictions help investors make informed decisions about buying or selling gold.
2. **Financial Institutions**: Banks and other financial institutions use gold price forecasts to manage their portfolios and hedge against risks.
3. **Jewelry Industry**: Jewelers and manufacturers use price forecasts to plan inventory and pricing strategies.
4. **Governments**: Policymakers use gold price forecasts to make economic decisions and manage national reserves.
5. **Traders**: Traders rely on price predictions to develop trading strategies and maximize profits.

Objectives

1. **Data Preparation and EDA**: Load and preprocess the gold price data, perform exploratory data analysis (EDA) to understand trends and patterns.
2. **Model Implementation**: Implement and fit three time series models (ARIMA, SARIMA, LSTM) to the data.
3. **Model Evaluation**: Evaluate the performance of each model using metrics such as RMSE, MAPE, and MAE.
4. **Model Comparison**: Compare the models based on their performance metrics and identify the best model for forecasting gold prices.
5. **Forecasting**: Generate forecasts using the best model and visualize the results.

Output

The models were evaluated using the following metrics:

- **RMSE (Root Mean Squared Error)**: Measures the average magnitude of the error.
- **MAPE (Mean Absolute Percentage Error)**: Measures the accuracy of the model in percentage terms.
- **MAE (Mean Absolute Error)**: Measures the average magnitude of the absolute errors.
-

The evaluation results for each model are as follows:

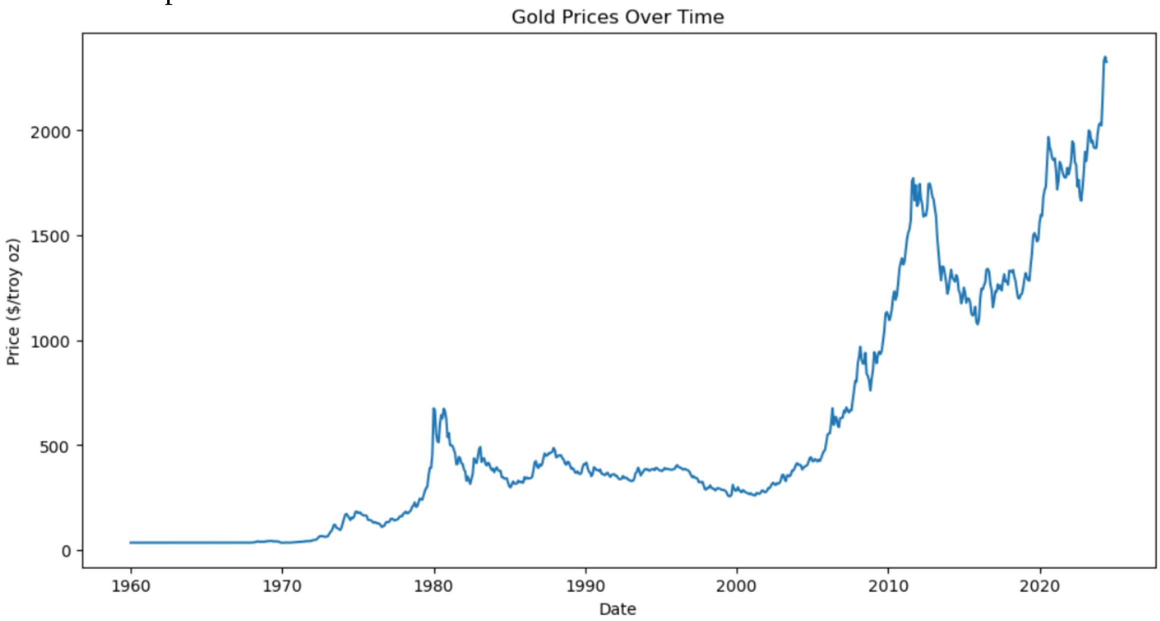
- **ARIMA**:
  - RMSE: 312.46
  - MAPE: 19.41%
  - MAE: 277.21
- **SARIMA**:

    - RMSE: 375.31
    - MAPE: 23.38%
    - MAE: 308.75
- **LSTM**:
    - RMSE: 339.18
    - MAPE: 19.17%
    - MAE: 310.00

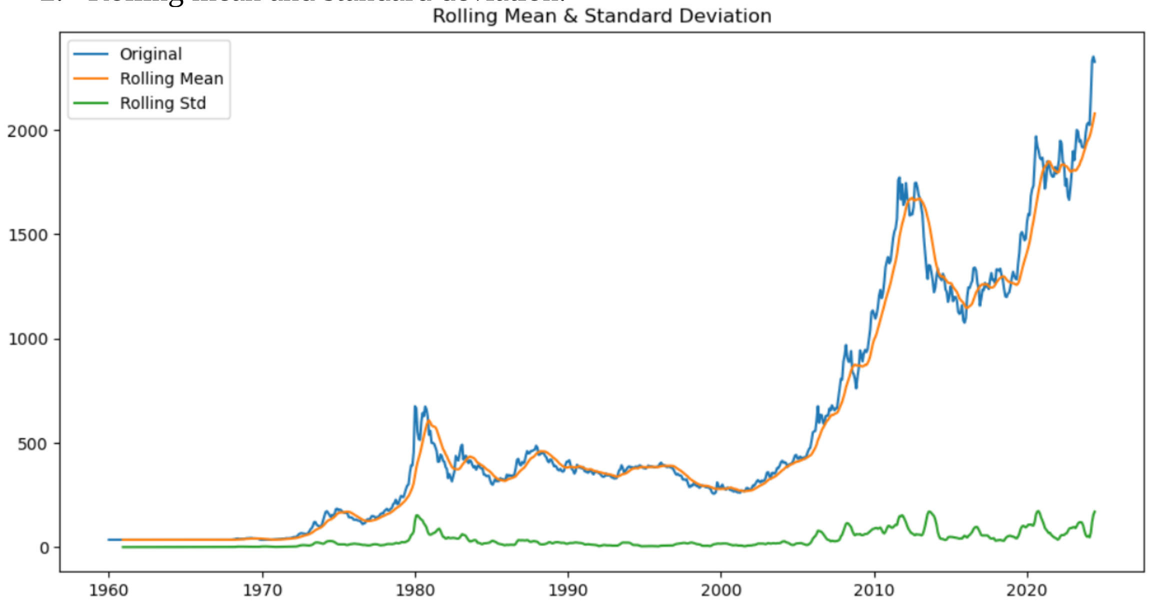
Based on these metrics, the ARIMA model had the lowest RMSE and MAE, indicating that it performed better in terms of average error magnitude. The LSTM model had the lowest MAPE, indicating better accuracy in percentage terms.

The following plots were generated:

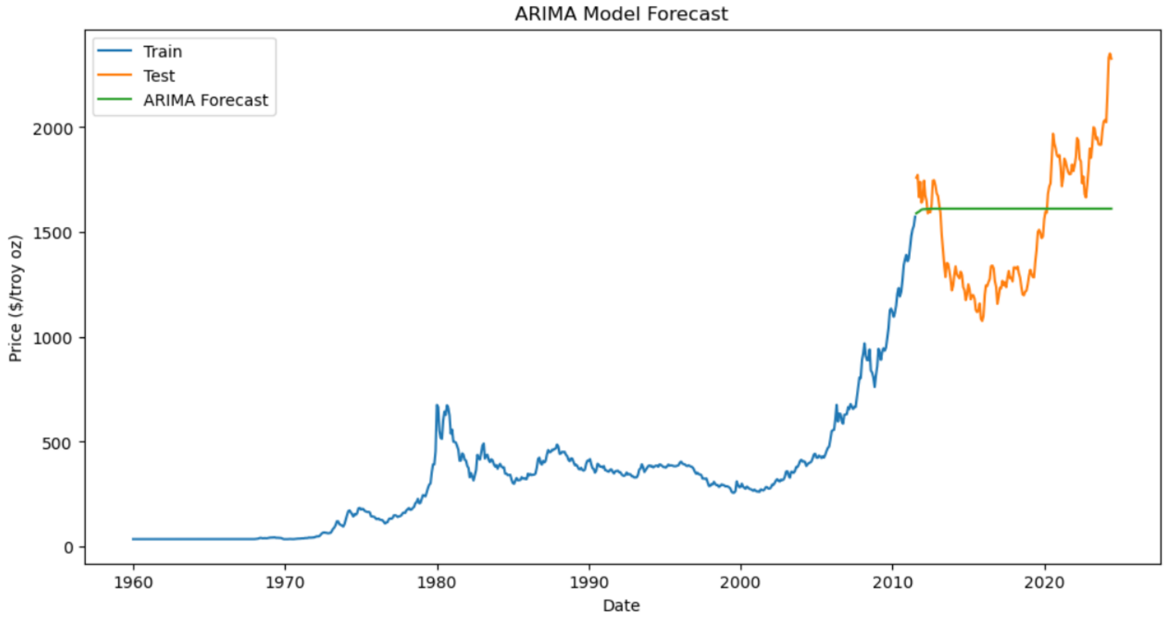
1. Gold prices over time.

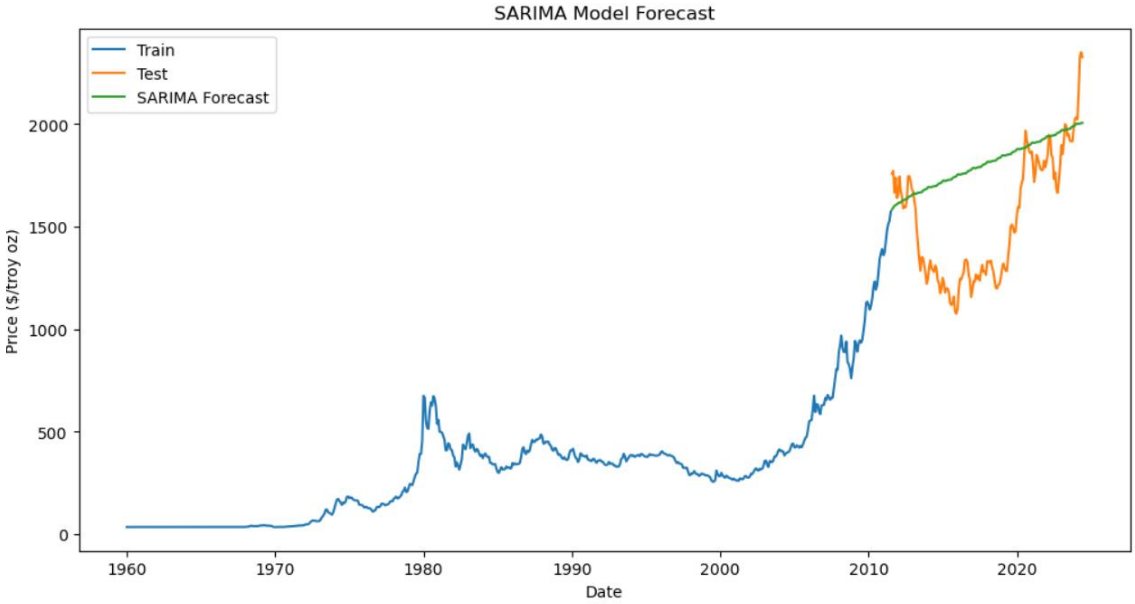


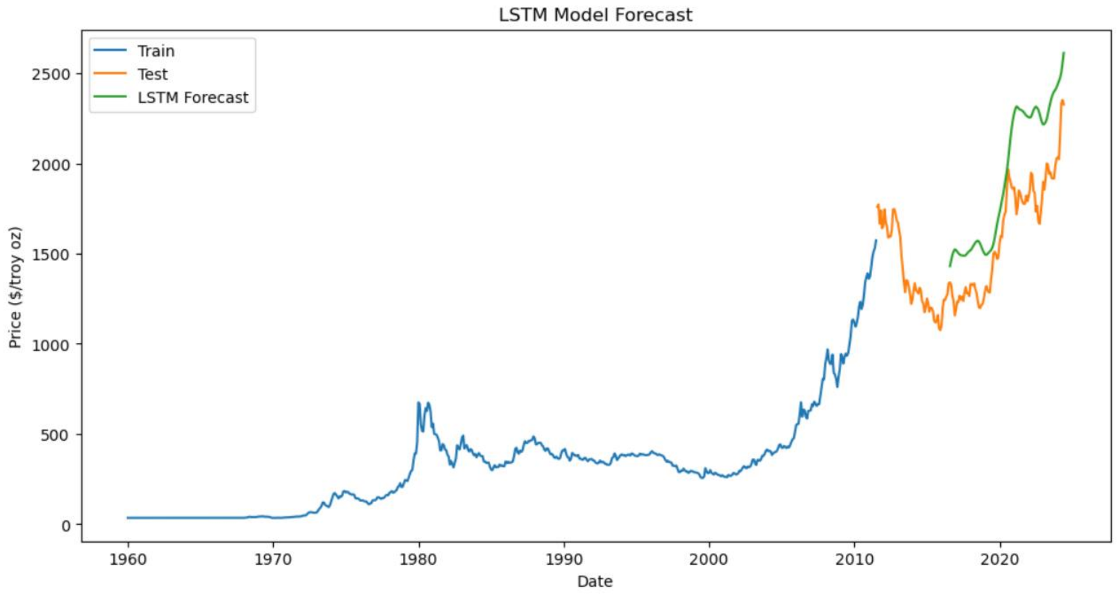
2. Rolling mean and standard deviation.

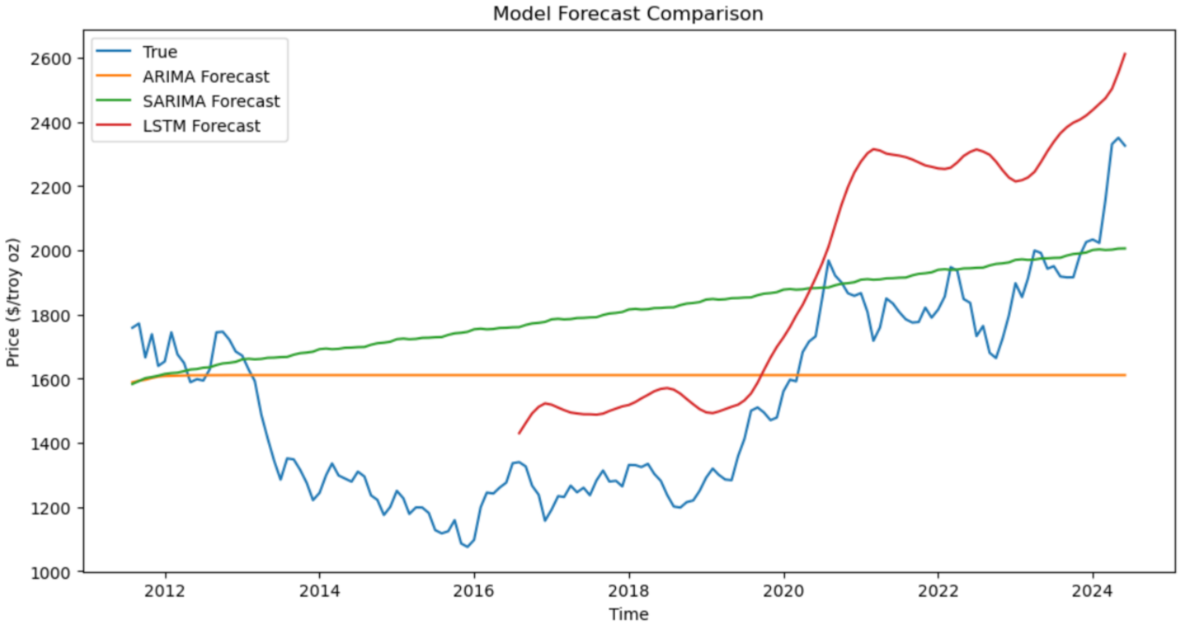


3. Forecast comparison for ARIMA, SARIMA, and LSTM models.


ARIMA Model Forecast
Train
Test
ARIMA Forecast
Price ($/troy oz)
0
500
1000
1500
2000
1960
1970
1980
1990
2000
2010
2020
Date


SARIMA Model Forecast
Train
Test
SARIMA Forecast
Price ($/troy oz)
0
500
1000
1500
2000
1960
1970
1980
1990
2000
2010
2020
Date





Interpretation

The ARIMA model was identified as the best-fit model based on its lower RMSE and MAE compared to the SARIMA and LSTM models. This suggests that the ARIMA model is better at capturing the underlying patterns and trends in the gold price data.

Interpretation of Values:

- **RMSE (Root Mean Squared Error)**: RMSE is a measure of the differences between values predicted by a model and the actual values. Lower RMSE indicates better model performance. The ARIMA model had the lowest RMSE (312.46), suggesting it has the smallest average prediction error compared to SARIMA and LSTM.
- **MAPE (Mean Absolute Percentage Error)**: MAPE measures the accuracy of the forecasted values as a percentage. Lower MAPE indicates higher accuracy. The LSTM model had the lowest MAPE (19.17%), indicating it was the most accurate in percentage terms.

- **MAE (Mean Absolute Error)**: MAE measures the average magnitude of the errors in a set of predictions. Lower MAE indicates better model performance. The ARIMA model had the lowest MAE (277.21), indicating it has the smallest average error magnitude compared to SARIMA and LSTM.

ARIMA Model:

- **Strengths**: Simple to implement and interpret, effective for non-seasonal data.
- **Weaknesses**: Limited in capturing complex patterns and seasonality.

SARIMA Model:

- **Strengths**: Extends ARIMA to capture seasonality, useful for seasonal data.
- **Weaknesses**: More complex to implement, higher computational cost.

LSTM Model:

- **Strengths**: Capable of capturing complex patterns and long-term dependencies, suitable for large datasets.
- **Weaknesses**: Requires more data and computational resources, can be difficult to interpret.

The model forecast comparison graph plots the actual gold prices against the predicted values from the ARIMA, SARIMA, and LSTM models.

1. **True Values**: The actual gold prices are plotted as a solid line. This represents the ground truth against which the model forecasts are compared.
2.
3. **ARIMA Forecast**: The ARIMA model's predictions are plotted on the same graph.
   - **Performance**: The ARIMA model closely follows the true values with relatively smaller deviations, indicating a good fit. The lower RMSE and MAE values further confirm its accuracy in predicting gold prices.
4. **SARIMA Forecast**: The SARIMA model's predictions are also plotted.
   - **Performance**: While the SARIMA model also captures the general trend of the gold prices, it shows more deviation from the true values compared to ARIMA. This is indicated by its higher RMSE and MAE values.
5. **LSTM Forecast**: The LSTM model's predictions are included in the graph.
   - **Performance**: The LSTM model shows some deviations from the true values, particularly during periods of rapid price changes. Although its MAPE is the lowest, indicating better percentage accuracy, its RMSE and MAE are higher than those of the ARIMA model.

Overall Comparison:

- **ARIMA Model**: The ARIMA model provides the closest fit to the actual gold prices, with the smallest prediction errors in terms of RMSE and MAE. This makes it the best model for this particular dataset.

- **SARIMA Model**: The SARIMA model captures the general trend but shows larger deviations, making it less accurate than ARIMA in this case.

- **LSTM Model**: The LSTM model, while good at capturing percentage-based accuracy, struggles with the absolute magnitude of errors. This suggests it might need more data or different configurations to outperform ARIMA.