Project Title: Building an Optimal Model for Predicting Hotel Booking Cancellations

DSF-PT08: FINAL PROJECT SUBMISSION

Project Owner: James Wachira Muthee

Technical Mentor: Daniel Ekale

Table of Contents

1.Business Problem Definition

2. Business Understanding

3. Data

4. Modelling Tools used

5. Data Understanding

6. Data Preprocessing

7. Modelling

8. Visualizations

9. Model Evaluation

10. Insights

11. Conclusion

12. Future Improvement Strategies

Business Problem Statement

Hotels face difficulties in managing their bookings effectively because they cannot foresee which reservations will be canceled. This unpredictability disrupts their ability to allocate resources, plan efficiently, and maintain profitability. If hotels could forecast cancellations, they would be able to refine their booking strategies, reduce financial losses, and improve the overall guest experience.

Business Objectives

The objective of this project is to develop a predictive model to accurately forecast hotel reservation cancellations so as to enable hotels to:

1. Optimize booking strategies and resource allocation based on cancellation predictions.
2. Minimize revenue loss by proactively managing overbooking and cancellation policies.
3. Improve customer satisfaction by offering targeted incentives to reduce cancellations

Data

The Hotel reservations dataset that I used for this project was obtained from Hotel Reservations Classification Dataset

Tools used

I used inbuilt libraries in Visual Studio Code tool to:

1. Load the dataset
2. Understand the dataset(shape, distribution and summary statistics)
3. Do exploratory data analysis
4. Do Data Preprocessing
5. Develop Models
6. Evaluate Model Performance

Data Understanding

Upon loading the data, the following observations were made:

1. The dataset had 36,275 rows and 19 columns

2. Several columns exhibited a significant number of outliers

3. Some columns displayed a heavy skew in their distributions

4. There were no missing values in the dataset

5. The features were not highly correlated with each other

Data Preprocessing

This is the process of preparing the data before modelling. Steps used in this project are:

1. Converting Categorical variables to numeric variables
2. Applying log transformation to some selected numerical variables with a skewed distribution
3. Adjusting some selected numerical features so that they are on a similar scale
4. Converting the target variable from categorical to Numerical

Modelling

Process of creating models to analyze and interpret data, identify patterns, and make predictions. For this project, 4 models were developed, namely:

1. Logistic Regression Model(Baseline Model)

2. Decision Tree Model(Tuned)

3. Random Forest Model(Untuned)

4. Random Forest Model(Tuned)

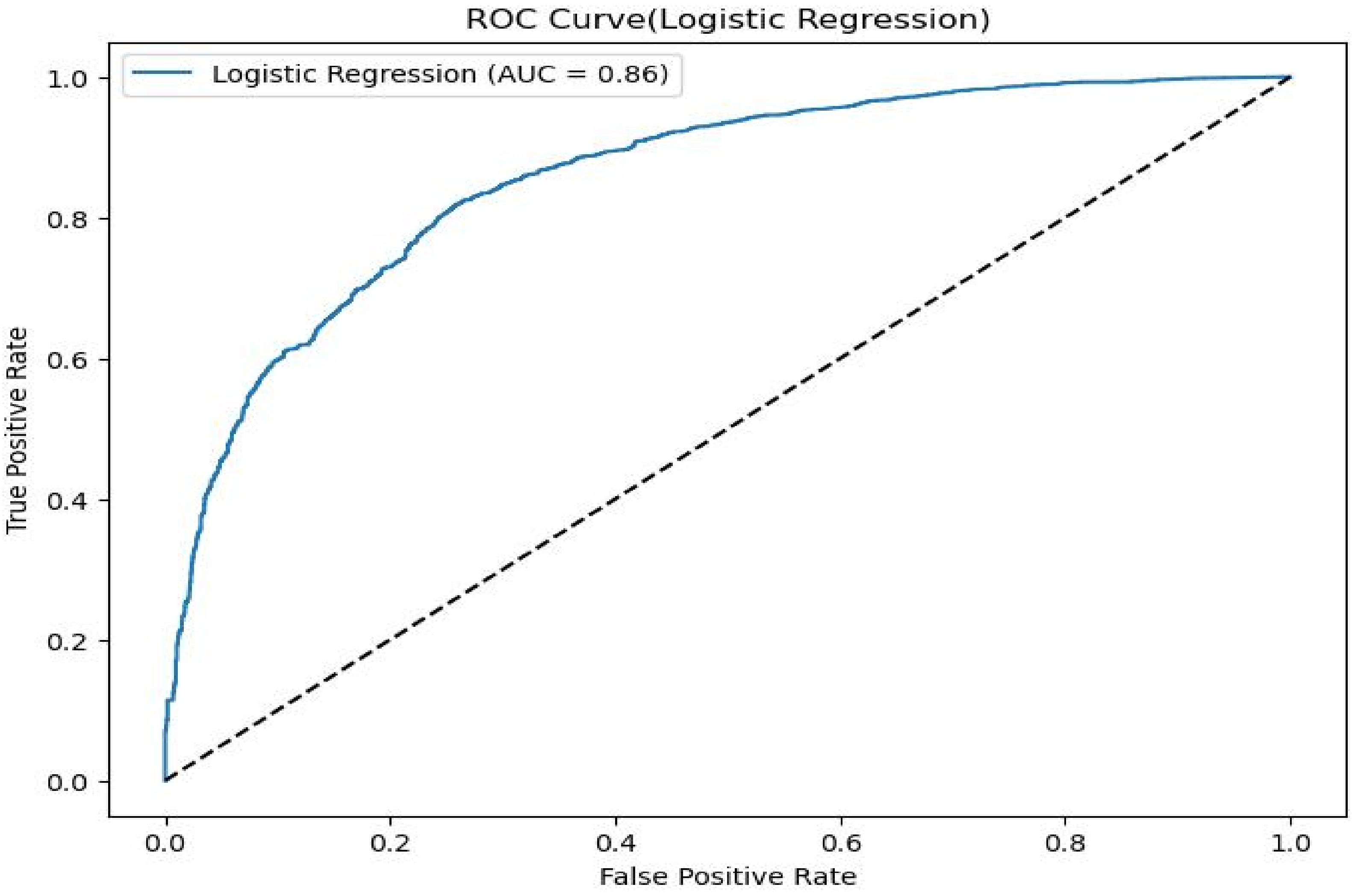
Model Evaluation

The following Metrics were used to evaluate the Models developed:

1. **Accuracy Score**: It represents the proportion of correct predictions made by the model out of all predictions made.

2. **AUC** : Refers to Area under the ROC curve (Receiver Operating Characteristic) Curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR)

Logistic Regression Model-ROC Curve

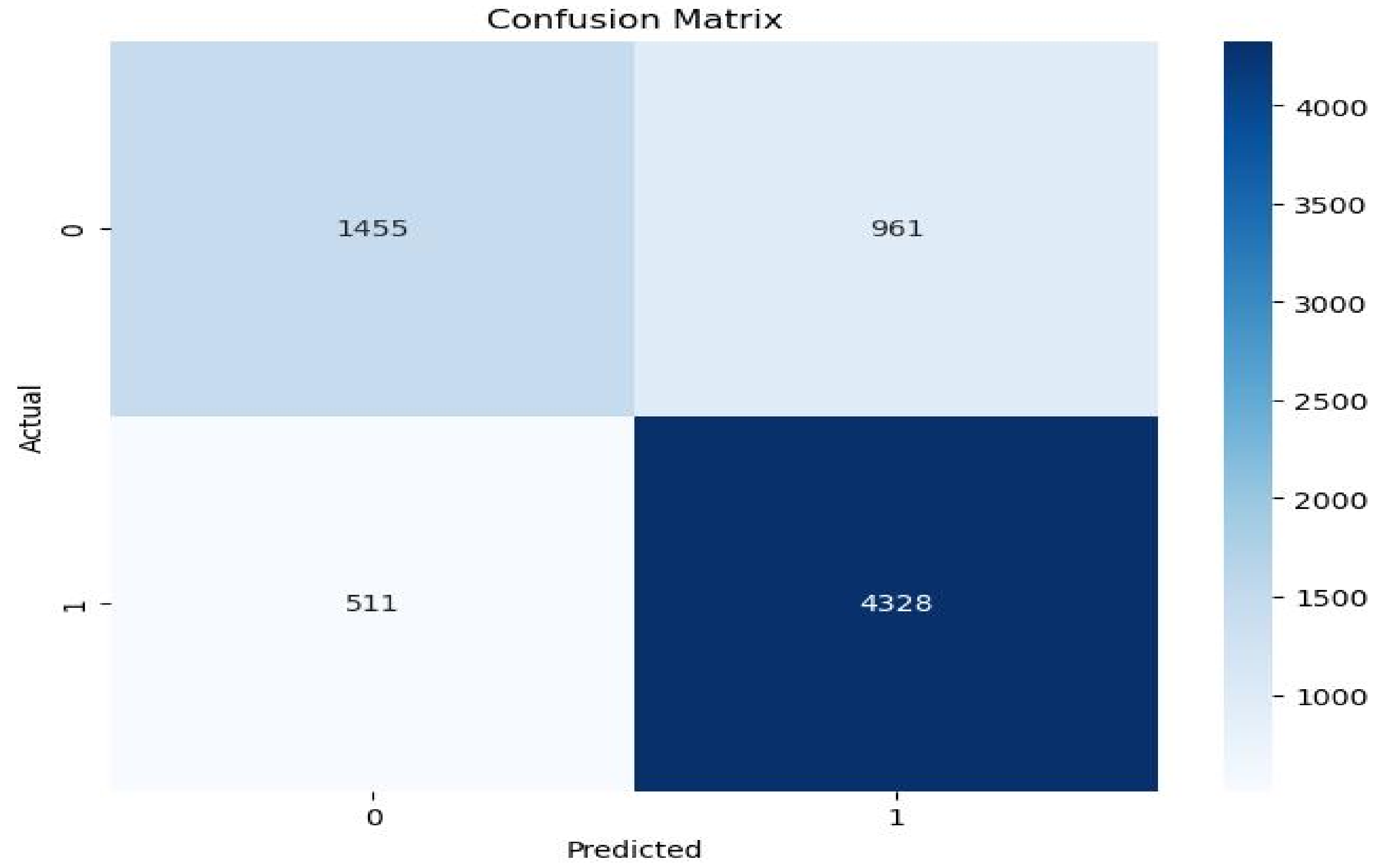


Overall Metric Score

1.Accuracy Score = 80%

2. AUC = 0.86

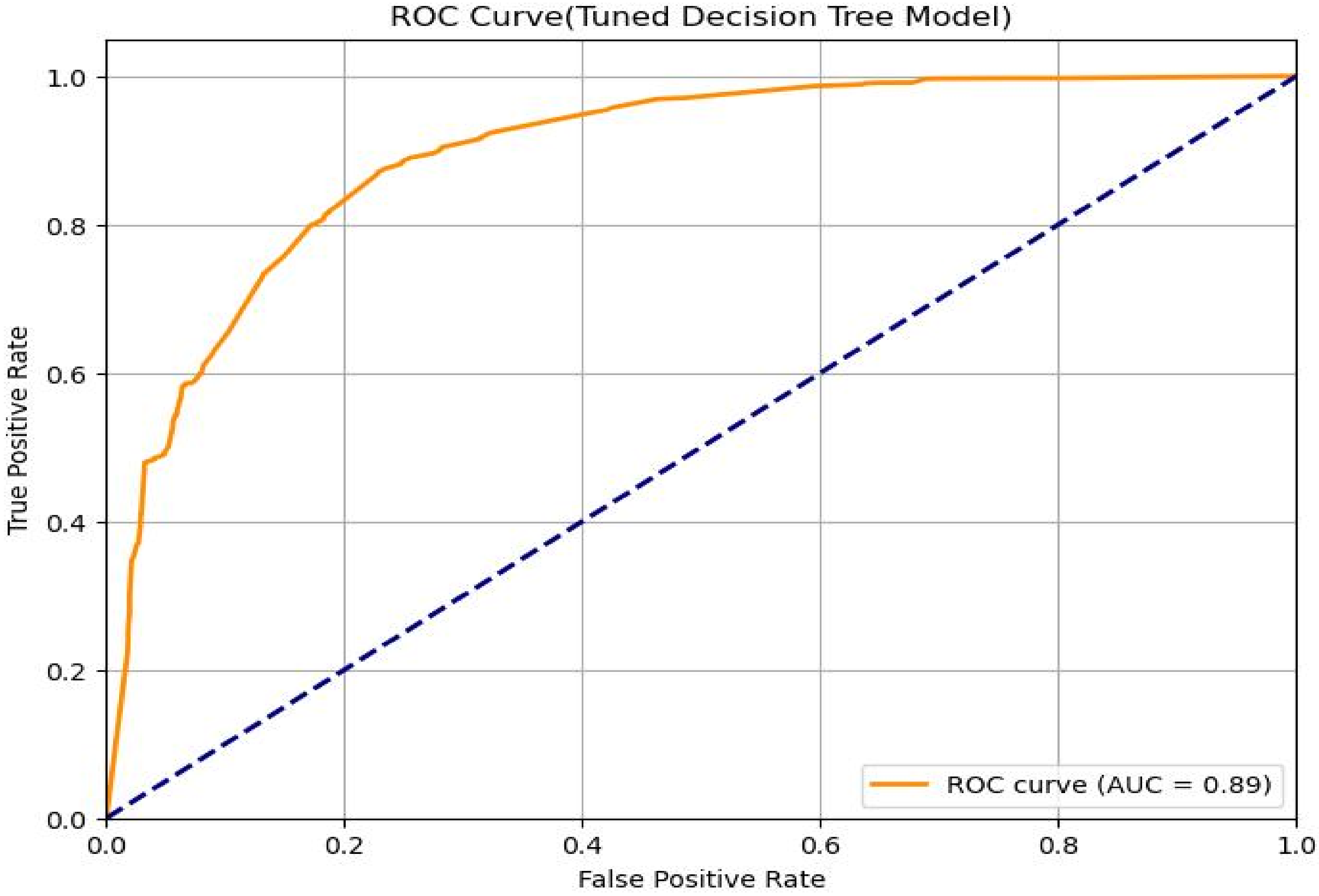
Logistic Regression-Confusion Matrix



Interpretation

- 4,328 bookings correctly predicted as non-cancellations
- 1,455 bookings correctly predicted as cancellations

Decision Tree Model(Tuned)-ROC Curve

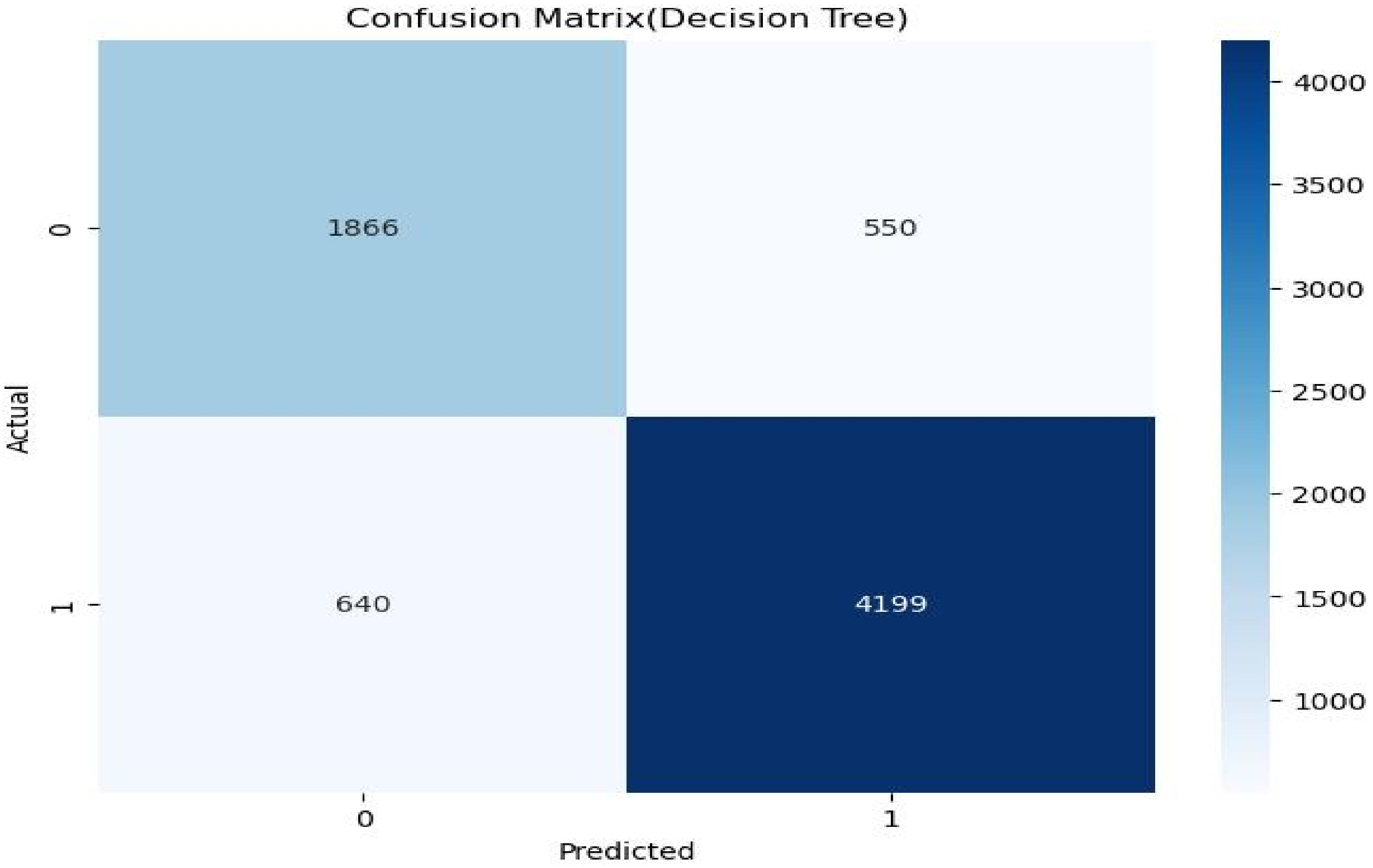


Overall Metric Score

1.Accuracy Score = 84%

2. AUC = 0.89

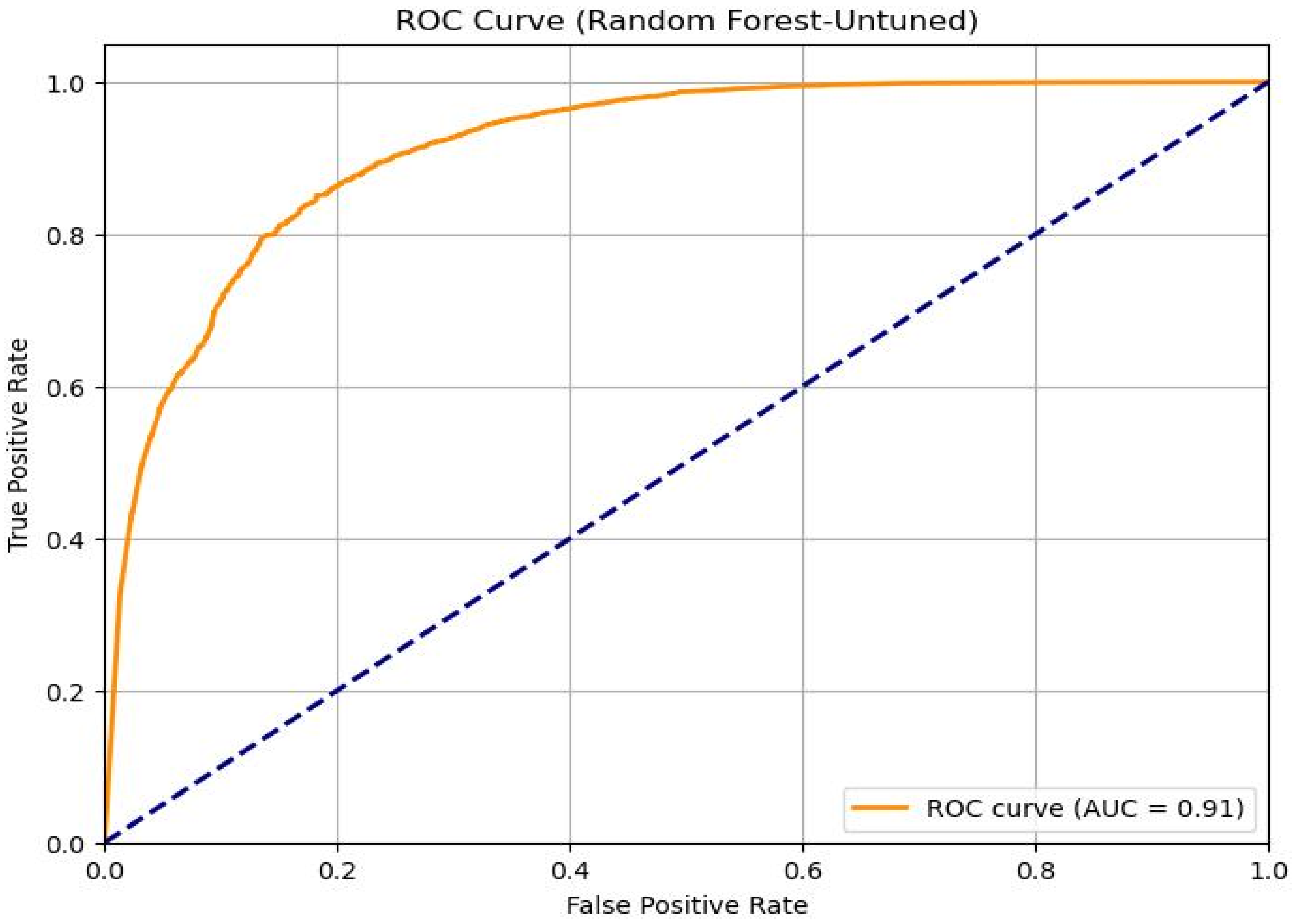
Decision Tree Model(Tuned) - Confusion Matrix



Interpretation

- 4,199 bookings correctly predicted as non-cancellations
- 1,866 bookings correctly predicted as cancellations

Random Forest(Untuned)-ROC Curve

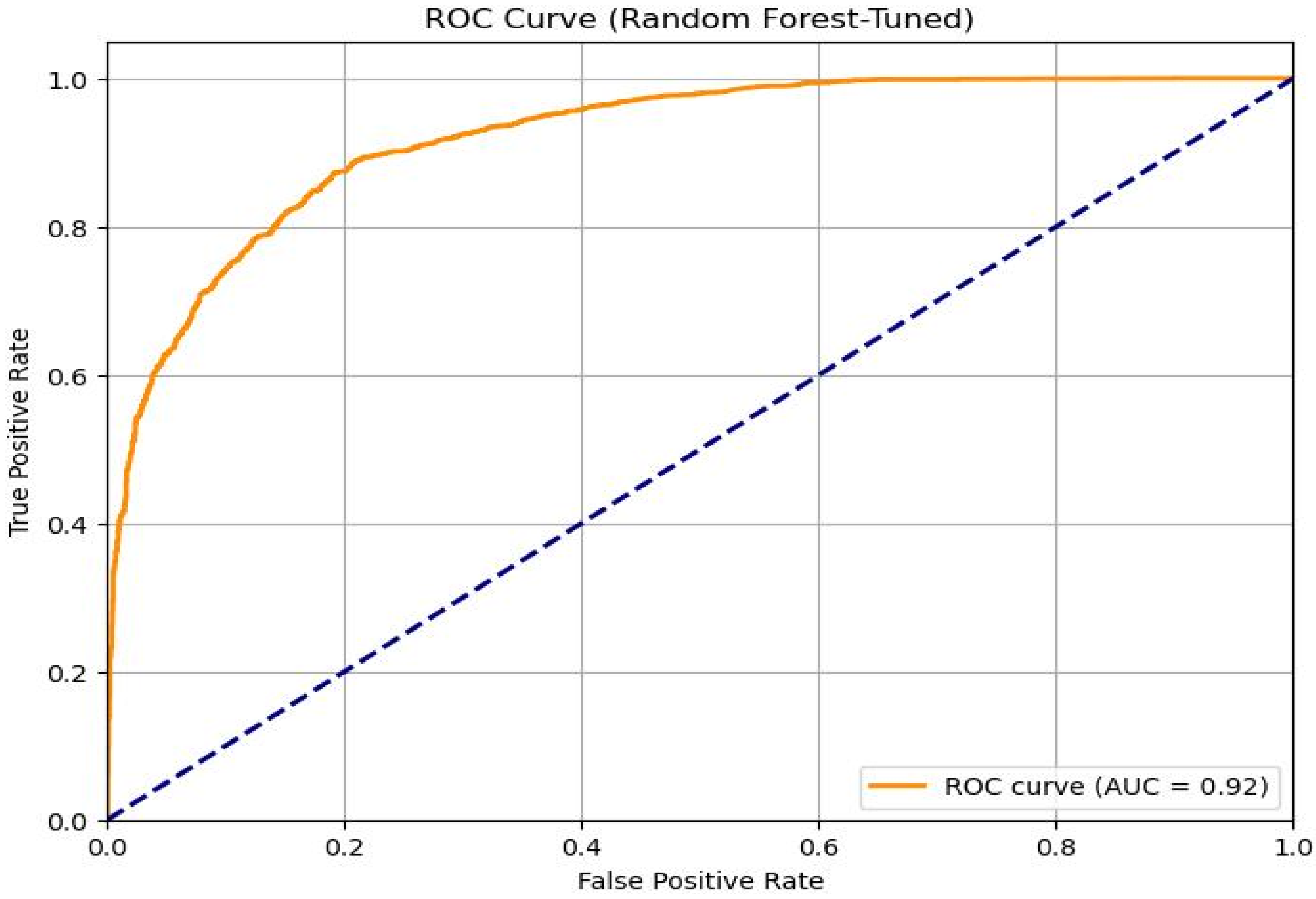


Overall Metric Score

1.Accuracy Score = 85%

2. AUC = 0.91

Random Forest(Tuned)-ROC Curve

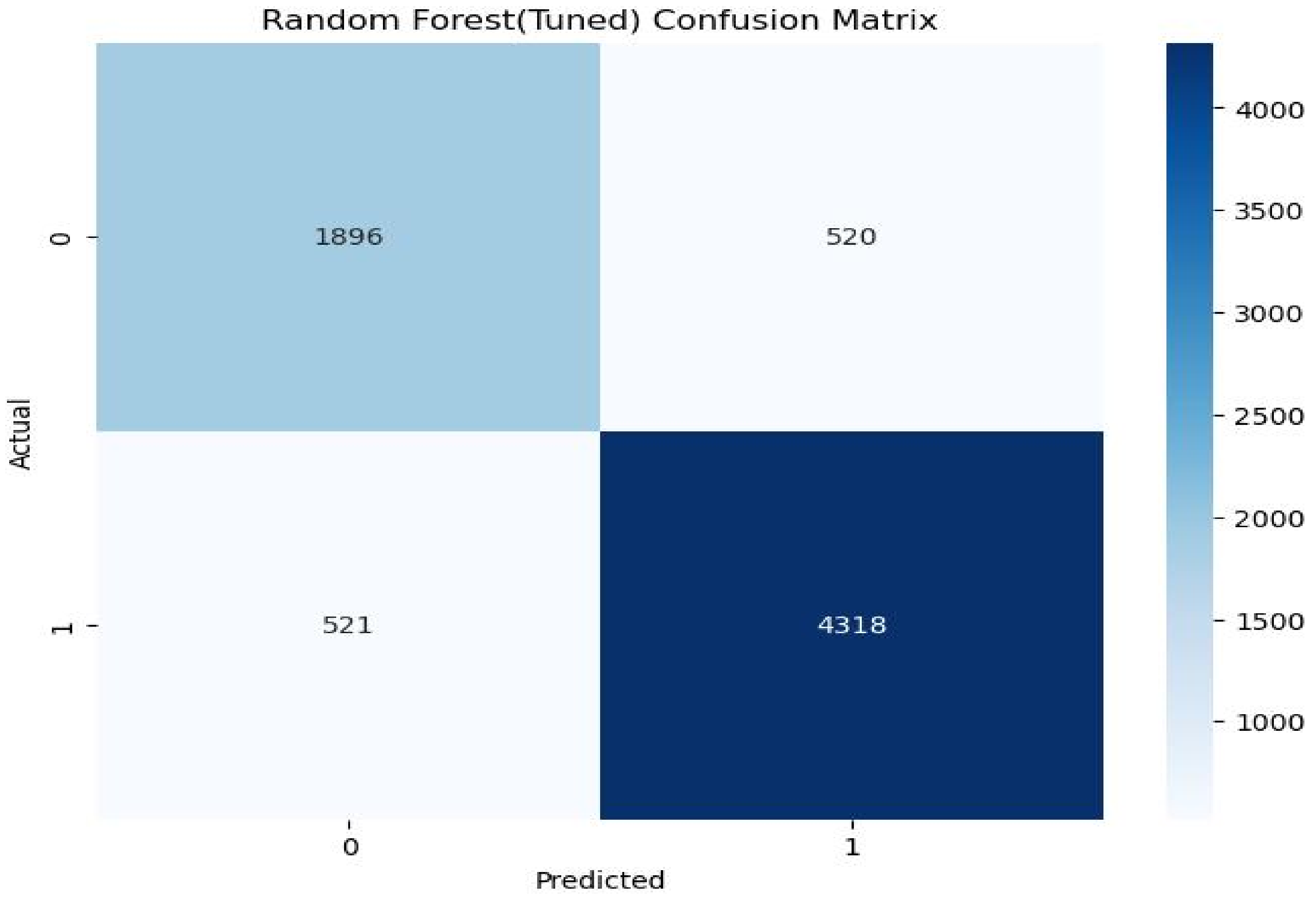


Overall Metric Score

1.Accuracy Score = 86%

2. AUC = 0.92

Random Forest(Tuned)-ROC Curve



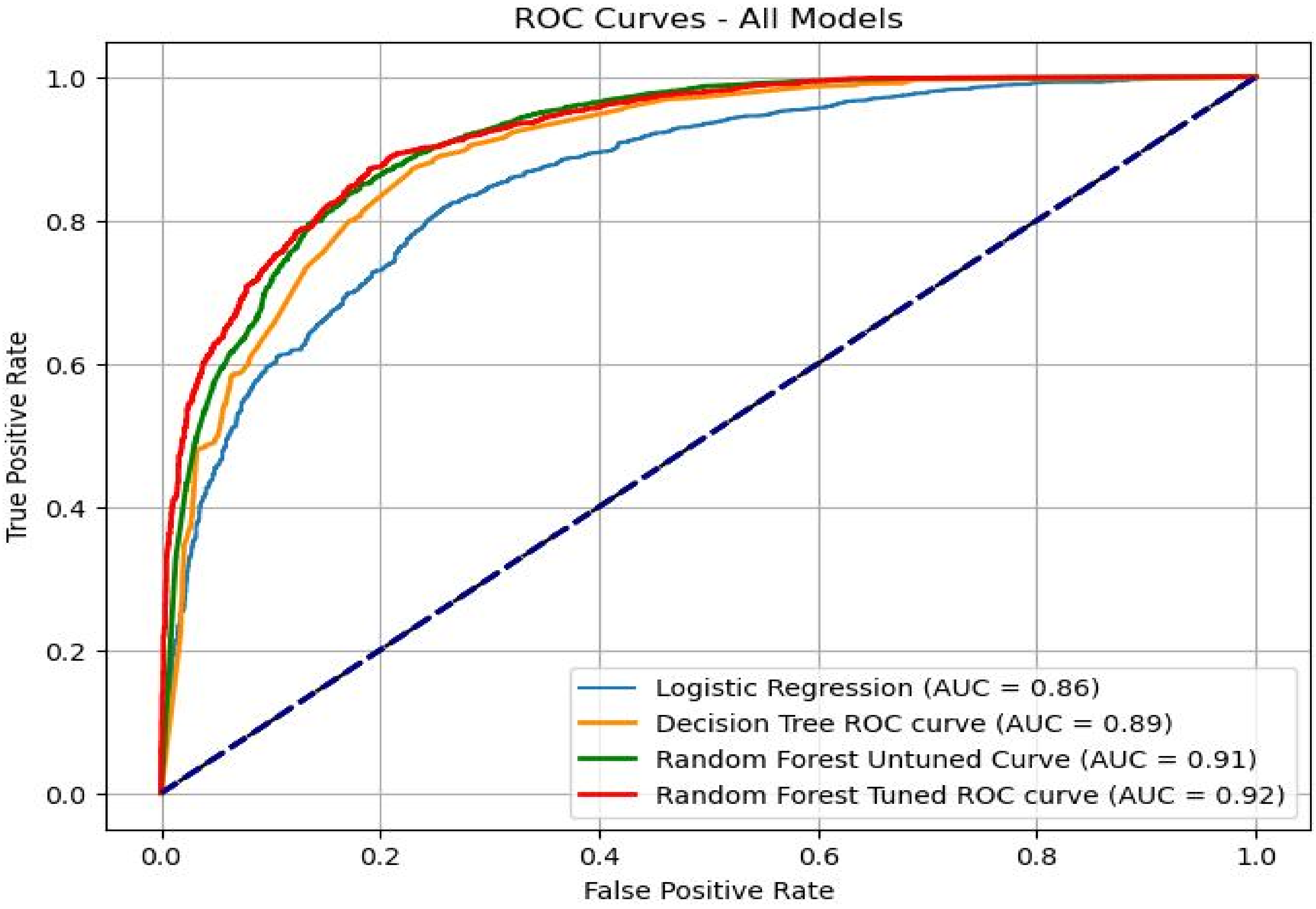
Random Forest(Tuned) Confusion Matrix

Actual \ Predicted	0	1
0	1896	520
1	521	4318

Interpretation

- 4,318 bookings correctly predicted as non-cancellations
- 1,896 bookings correctly predicted as cancellations

Combined ROC Curves-All Models



Best Performing Model

Tuned Random Forest Model with:
1. Accuracy Score = 86%

2. AUC = 0.92

Insights

Upon developing and evaluating the models, I concluded the following:

1**. Tuned Random Forest Model**

Outperforms all other models, with the highest Accuracy Score of 86% and the highest AUC of 0.92.

2. Untuned Random Forest Model

Also performed well, achieving an Accuracy Score of 85% and an AUC of 0.91 even without tuning of parameters

3. Tuned Decision Tree Model

Showed a notable improvement over the baseline Logistic Regression Model, with an Accuracy Score of 84% and an AUC of 0.89, suggesting that decision trees, when optimized, are a reliable choice, though slightly less effective than Random Forest.

4. Logistic Regression Model (Baseline)

With an Accuracy Score of 79% and an AUC of 0.86, serves as a solid starting point but demonstrates lower performance compared to more complex models like Random Forest.

Conclusion

I concluded that the tuned **Random Forest Model** with an **Accuracy Score of 86%** and an **AUC of 0.92** is the best model out of the four models developed.

The overall objective of this project which was to develop a high performing model to predict hotel booking cancellations was therefore met.

Future Improvement Strategies

- **Feature Engineering**: incorporating additional features such as customer demographics, booking patterns could improve model’s performance.

- **Model Ensemble:** Combining the strengths of multiple models through techniques like stacking or boosting could improve performance.

- **Real-Time Predictions**: Implementing this model into a real-time system for predictive booking management could help hotels take proactive measures to minimize cancellations.

Thank You

Q&A