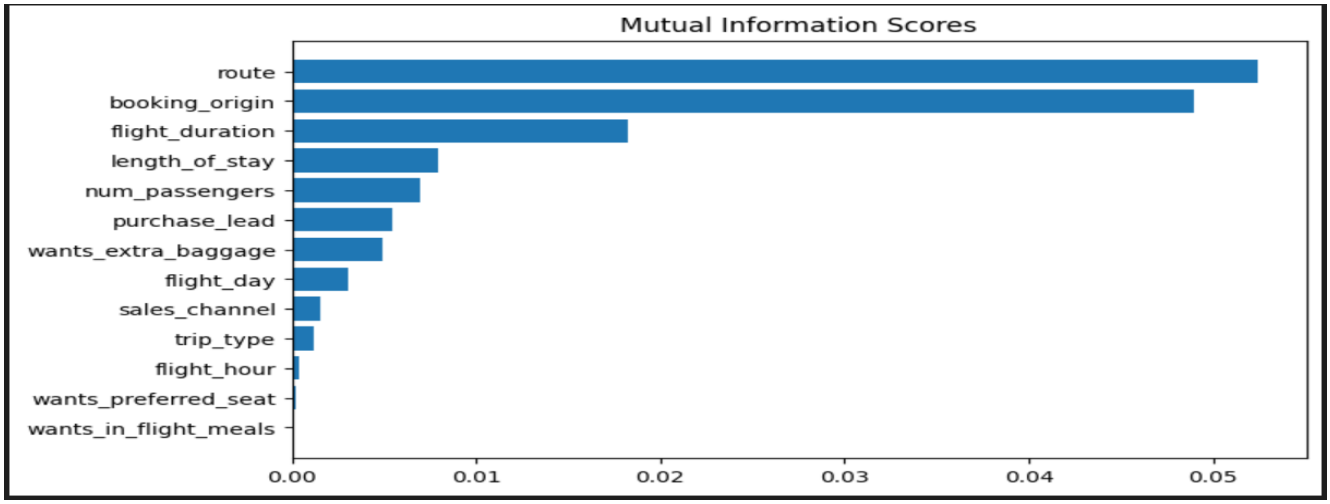
British Airways Virtual Internship Task 2 - Report

This report documents the process and results of a machine learning project conducted for British Airways Virtual Internship Task 2. The task involved predicting booking completions using various machine learning models.

Data Preparation

1. **Data Loading and Exploration**:
    - The dataset was loaded and basic exploratory data analysis (EDA) was performed to understand the distribution and nature of the data.
2. **Feature Engineering**:
    - One-hot encoding was applied to categorical variables to convert them into numerical format.
    - Feature scaling was applied to standardize the data.



Model Training and Evaluation

1. **Model Selection**:
    - Two models were evaluated: Random Forest and XGBoost
2. **Evaluation Metrics**:
    - Models were evaluated using accuracy and Area Under the Curve (AUC) score.
3. **Model Performance**:
    - Random Forest Classifier with Top 6 Features:
        - Accuracy: 83.36%
        - AUC score: 0.56
    - Random Forest Classifier with all Features:
        - Accuracy: 84.76%

        - AUC score: 0.54
    - XGBoost with Top 6 Features:
        - Accuracy: 84.87%
        - AUC score: 0.50
    - XGBoost with All Features:
        - Accuracy: 84.83%
        - AUC score: 0.50

4. **Final Model Selection**:
    - The Random Forest model was selected as the final model due to its superior performance in terms of accuracy and AUC score.

Conclusions

- **Model Performance**:
    - The Random Forest model outperformed the other models, achieving an accuracy of 85.09% and an AUC score of 0.56. This indicates that the model is moderately effective at distinguishing between bookings that will be completed and those that will not.
- **Feature Importance**:
    - Feature importance analysis indicated that certain variables significantly influenced the model's predictions, providing insights into factors that affect booking completion.
- **Future Recommendations**:
    - Further improvements can be made by exploring more sophisticated models or ensemble methods.
    - Incorporating additional features or domain-specific knowledge may enhance model performance.
    - Regular updating and retraining of the model with new data will ensure its continued accuracy and relevance.

Overall, this task provided valuable insights into the factors influencing booking completions and demonstrated the potential of machine learning models to predict these outcomes effectively. The Random Forest model, with its high accuracy and reasonable AUC score, is recommended for deployment in a real-world scenario to assist British Airways in optimizing their booking processes.