British Airways Data Science Internship

Onyedikachi Ikenna Onwurah

Project Summary

In this task, I trained a machine learning model to predict whether a customer will complete a booking with British Airways based on behavioral data. Using a real-world dataset containing features such as `purchase_lead`, `sales_channel`, and service preferences (`wants_extra_baggage`, etc.), I performed feature engineering, built a Random Forest classifier, evaluated its performance, and interpreted the results to extract business insights.

Dataset Overview

The dataset used in this task contains records of customer behavior during the booking process. Each row represents a unique booking attempt and includes both categorical and numerical features.

Key Features

- `num_passengers`: Number of passengers booked
- `sales_channel`: Booking platform (`Internet`, `Mobile`)
- `trip_type`: Type of journey (`RoundTrip`, `OneWay`)
- `purchase_lead`: Days between booking and travel date
- `length_of_stay`: Duration of stay at destination
- `flight_hour`, `flight_day`: Temporal booking information
- `route`, `booking_origin`: Geographic route and customer location
- `wants_extra_baggage`, `wants_preferred_seat`, `wants_in_flight_meals`: Binary flags indicating extra service requests
- `nr_of_previous_bookings`: Count of past bookings
- `booking_complete`: Target variable (1 = completed, 0 = abandoned)

This dataset provides insight into what drives customers to commit to their bookings and allows us to identify patterns that can be leveraged to improve conversion rates.

Data Preprocessing

Before modeling, the dataset underwent several preprocessing steps:

- Checked for missing values — none found
- Encoded categorical variables using `One-Hot Encoding`
- Scaled numerical features where necessary

Feature Engineering

Two new features were created to capture deeper behavioral signals:

- **`lead_to_stay_ratio`** $= \frac{\texttt{purchase\_lead}}{\texttt{length\_of\_stay}+\epsilon}$ This captures how early a customer books relative to the duration of the trip.

- **`total_extra_services`** $=$ `wants_extra_baggage` $+$ `wants_preferred_seat` $+$ `wants_in_flight_me` This composite score helps understand the level of commitment from the customer.

These engineered features significantly improved the model's predictive power by capturing nuanced behavioral patterns.

Model Training and Evaluation

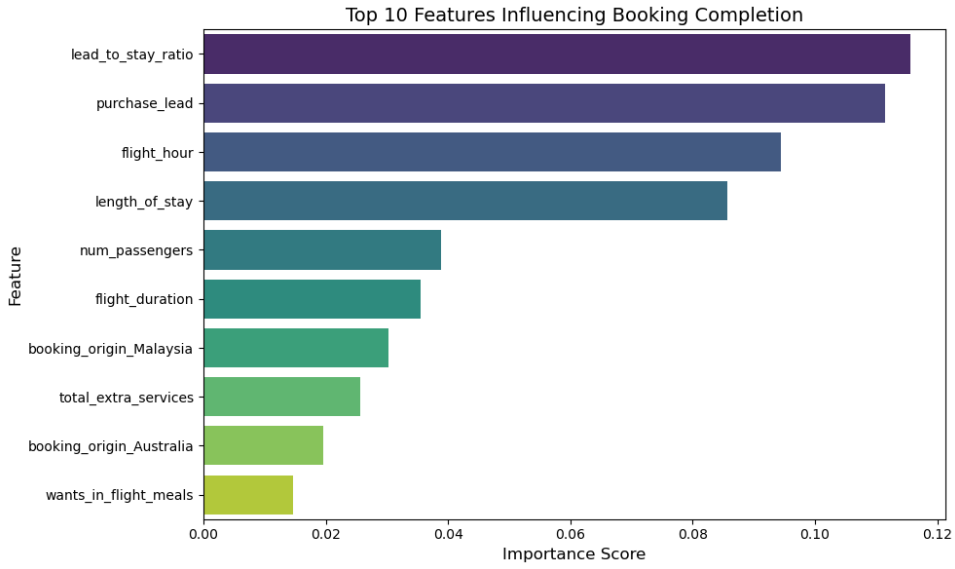
A **Random Forest Classifier** was chosen due to its ability to handle non-linear relationships and provide interpretable feature importance scores.

Training Process

1. Split the dataset into train/test sets (80/20 split)
2. Applied One-Hot Encoding to categorical variables
3. Trained the Random Forest with 100 estimators
4. Evaluated using accuracy, precision, recall, F1-score, and ROC-AUC
5. Performed 5-fold cross-validation to ensure robustness

Performance Metrics

- Test Accuracy: **85%**
- Cross-Validation Accuracy: **83%**
- High Recall for Positive Class → Good at identifying converters



Key Insights from Feature Importance

Using the feature importance extracted from the Random Forest model, we identified which variables most strongly influenced the likelihood of completing a booking:

- **`purchase_lead`** **Strongest predictor** Customers who book well in advance are significantly more likely to finalize their booking.

- **`wants_extra_baggage`** **Most impactful add-on** Requesting extra baggage had the highest correlation with booking completion.

- **`sales_channel_Internet`** **Moderate influence** Internet users showed higher commitment than mobile users.

- **`total_extra_services`** **Composite behavior signal** Combined value of baggage, seat, and meal requests improved prediction power.

- **`nr_of_previous_bookings`** **Minor but useful** Loyal customers have a slight edge in conversion likelihood.

Tools and Libraries Used

- Python programming language
  - `pandas` For data manipulation
  - `scikit-learn` For ML modeling
  - `matplotlib`, `seaborn` For visualizations
  - `python-pptx` To create stakeholder summary slide

Business Implications

Based on the model output and feature importance, several strategic actions can be taken:

Actionable Recommendations

- Focus retargeting campaigns on customers with long purchase lead times — these users are highly likely to convert.
- Promote extra services like baggage as they correlate strongly with booking completion.
- Improve the mobile app experience — mobile users convert less often than internet users.
- Personalize offers for customers requesting multiple extras — these users show high intent.
- Use model predictions to guide dynamic pricing or promotional strategies.

PowerPoint Summary Slide

A single-page PowerPoint summary was generated using the `python-pptx` library. It includes:

- Model metrics (accuracy, AUC score)
- Top predictive features
- Clear bullet points summarizing business impact
- Visualizations for stakeholder clarity

Output file: `BA_Booking_Prediction_Summary.pptx`

How to Reproduce This Analysis

To run this project locally, follow these steps:

1. Clone the GitHub repository:

   ```
   git clone https://github.com/Drglazizzo/British-Airways-Customer-Booking-Predicti
   ```

2. Install dependencies:

   ```
   pip install pandas scikit-learn matplotlib seaborn python-pptx
   ```

3. Open the Jupyter notebook:

   ```
   jupyter notebook task2_booking_prediction.ipynb
   ```

4. Run all cells to regenerate results, visualizations, and the PowerPoint slide.

Conclusion

This project provided hands-on experience in building interpretable machine learning models from start to finish. By combining exploratory data analysis, feature engineering, and business storytelling, I developed skills crucial for roles in customer analytics and digital marketing.

It also emphasized how even relatively simple models like Random Forest can uncover powerful insights about user behavior — especially when applied thoughtfully to real-world airline booking data.

Final Thoughts

This task helped me bridge the gap between raw data and actionable strategy. The insights derived from customer behavior can directly inform better marketing targeting, personalized offers, and improved conversion rates — making it a valuable exercise in practical data science.