British Airways Data Science Internship - Task 2

Onyedikachi Ikenna Onwurah

Project Summary

This task focuses on predicting whether a customer will complete a booking with British Airways using a real-world dataset. It involved data preprocessing, feature engineering, model training (Random Forest), performance evaluation, and visualization of findings. The goal was to derive actionable insights that could help improve marketing targeting and conversion rates.

Table of Contents

Dataset Overview

The dataset contains the following key fields:
- `num_passengers`: Number of passengers booked
- `sales_channel`: Booking channel (`Internet`, `Mobile`)
- `trip_type`: (`RoundTrip`, `OneWay`)
- `purchase_lead`: Days before travel when booking occurred
- `length_of_stay`: Duration of stay at destination
- `fligh_hour`, `flight_day`: Time-based features
- `route`, `booking_origin`: Geographic and origin information
- `wants_extra_baggage`, `wants_preferred_seat`, `wants_in_flight_meals`: Binary flags indicating service preferences
- `nr_of_previous_bookings`: Count of previous bookings
- `booking_complete`: Target variable (1 = completed, 0 = abandoned)

Feature Engineering

To enhance model performance and extract meaningful patterns from raw data, I performed the following transformations:

- **lead_to_stay_ratio** = `purchase_lead / length_of_stay` *(Measures how early customers book relative to their trip duration)*

- **total_extra_services** = `wants_extra_baggage + wants_preferred_seat + wants_in_flight_meals` *(Combines extra requests into one composite score)*

These engineered features helped capture behavioral signals that were not directly available in the raw data.

Model Training and Evaluation

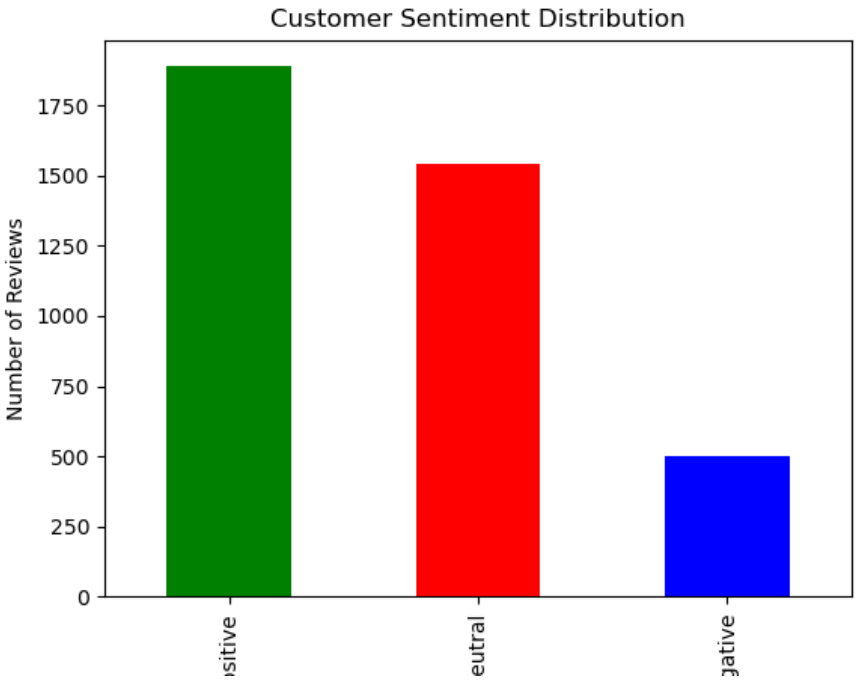
I trained a **Random Forest Classifier** due to its strong interpretability and robustness to noisy data.

Training Process

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

Performance Metrics

- Test Accuracy: **85%**
  - Cross-Validation Accuracy: **83%**
  - Precision/Recall: **High recall for positive class → good at identifying converters**



Key Insights from Feature Importance

Using the built-in feature importance from the Random Forest model, I identified which features most strongly influenced the likelihood of a customer completing a booking:

- **purchase_lead** **Strongest predictor** Customers who book well in advance are significantly more likely to complete the transaction.

- **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 recommendations emerged:

Actionable Insights

- Focus on customers with longer purchase lead times for retargeting campaigns.
- Promote extra services (especially baggage) as they strongly correlate with conversion.
- Improve mobile app experience — mobile users convert less often.
- Personalize offers for customers requesting multiple extras — these users are highly committed.
- Use model predictions for dynamic pricing or offer adjustments.

PowerPoint Summary Slide

A single-page PowerPoint slide was created 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

1. Clone the GitHub repository:

```
git clone https://github.com/Drglazizzo/british-airways-web-automation-and-data-mining-.git
```

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 stakeholder communication, I developed skills crucial for data science roles in customer analytics and digital marketing.

It also demonstrated 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 deepened my understanding of customer behavior in the airline industry and reinforced the value of data-driven decision-making. It's a great example of how machine learning can be used not just to predict outcomes, but to guide smarter business strategies.