Summary Report of British Airways customer bookings

Problem Statement

Our head has tasked us with predicting the flight cancellation rate to better understand customer behavior and improve operational efficiency. The goal is to identify key factors influencing whether a booking is completed or canceled, allowing the airline to implement targeted strategies to reduce cancellations.

Dataset Overview

The dataset provided contains 50,000 entries with 14 features related to customer bookings. Key features include the number of passengers, sales channel, trip type, purchase lead time, length of stay, flight timing, route, booking origin, and various customer preferences. The target variable, booking_complete, indicates whether a booking was completed (1) or canceled (0). The dataset contains both categorical and numerical variables, requiring preprocessing before model development.

Methodology Summary

Exploratory Data Analysis (EDA):

- Conducted a comprehensive EDA to understand the distribution and relationships within the data.
- Visualized key variables such as purchase_lead, length_of_stay, and booking_origin to identify trends and patterns.
- Analyzed the impact of customer preferences (e.g., extra baggage, preferred seat) on booking completion.

For detailed visualizations and insights, refer to the accompanying Python notebook.

Model Development:

Developed several classification models, including Logistic Regression, Random Forest, and XGBoost, to predict booking completion.

Due to the class imbalance (85% cancellations vs. 15% bookings), the F1 score was used as the primary evaluation metric rather than accuracy, ensuring better assessment of the minority class.

Addressing Class Imbalance:

Employed the imblearn library to create a balanced Random Forest model using techniques like SMOTE and Random Under-Sampling.

This approach significantly improved the F1 score for the minority class (class 1: booking completed), enhancing the model's ability to correctly predict actual bookings.

This methodology provided a robust framework for understanding and predicting flight cancellations, allowing for data-driven decisions to improve booking retention.

Insights Report

Booking Completion Rate:

Only 15% of customers completed their bookings, highlighting a significant challenge in retaining customers during the booking process.

Trip Type & Cancellation Trends:

Round Trip is the most popular type of trip among customers. However, it also has a high cancellation rate, indicating that customers may reconsider their plans after initially opting for a round trip.

Customers with longer lengths of stay are more likely to cancel their bookings, suggesting that extended trips might be subject to greater uncertainty or changes in plans.

Flight Duration & Popular Routes:

The majority of customers prefer short-haul flights, typically lasting between 6 to 13 hours.

The route from Auckland (AKL) to Kuala Lumpur (KUL) is the most popular, with a 9-hour flight being the preferred choice among travelers on this route.

Flight Day Patterns:

Monday has the highest number of flights lasting between 8 to 13 hours, indicating a preference for starting the week with medium-length travel.

Tuesday, Wednesday, and Thursday have a moderate number of flights, slightly higher than Fridays. In contrast, weekend flights are less frequent, possibly due to fewer business-related trips.

A detailed heatmap of flight distribution by day and duration is available in the Python notebook.

Geographical Insights:

The countries with the most significant number of flyers are China, Australia, Indonesia, India, and Malaysia. These regions represent major markets for British Airways, likely due to their large populations and strong travel demand.

Model Building Process Summary

Data Preprocessing

Categorical variables were mapped to numerical values using label encoding to make them suitable for machine learning models.

Model Selection:

Several classification models were employed, including XGBoost, Random Forest, and Logistic Regression. These models were chosen for their robustness and ability to handle structured data efficiently.

Evaluation Metric:

Due to the significant class imbalance (15% bookings completed vs. 85% cancellations), the F1 score was prioritized over accuracy as the evaluation metric. This approach ensured that the model's performance on the minority class (booking completions) was accurately measured.

Addressing Class Imbalance:

To handle the class imbalance, the imblearn library was used. Specifically, the SMOTE (Synthetic Minority Over-sampling Technique) was applied to oversample the minority class.

A balanced Random Forest model was then trained on the adjusted dataset, which led to an improvement in the model's ability to predict the minority class.

Final Model Performance:

After addressing the class imbalance and applying SMOTE, the final model achieved a ROC-AUC score of 78%, indicating a good ability to distinguish between bookings completed and canceled.

Data Collection Critique

The dataset provided for this analysis is extensive, with 50,000 entries and 14 features, covering various aspects of customer bookings. However, there are some limitations to the data that could impact the accuracy and generalizability of the models built. First, the data does not capture customer demographics, which are often crucial in understanding booking behavior. Factors like age, income level, and travel purpose could provide deeper insights into why certain bookings are canceled. Additionally, the dataset lacks information on external factors such as seasonal trends, economic conditions, or promotional campaigns, all of which can significantly influence booking decisions. The reliance on internal booking data alone may lead to models that are less effective in real-world scenarios where these external

factors play a role. Finally, while the dataset includes information on flight routes and times, it does not account for flight availability or pricing, which are critical variables in predicting customer decisions. Expanding the dataset to include these missing elements would likely enhance the predictive power of the models.

Feature importance analysis

The feature importance analysis provides valuable insights into the factors that most influence whether a flight booking is completed or canceled. The most critical factor is **purchase_lead**—the number of days between the booking and the travel date—with an importance score of **0.163**. This finding suggests that the timing of a booking is crucial, as customers who book well in advance or closer to their travel date may have different commitment levels, influencing the likelihood of completing the booking.

Following closely are **booking_origin** and **route**, with importance scores of **0.157** and **0.153**, respectively. These factors highlight the significance of geographical and route-specific trends in booking behavior. Certain regions and routes may have more volatility in bookings, possibly due to varying levels of demand, competition, or external factors like travel restrictions.

Length_of_stay and **flight_hour** also show substantial influence, suggesting that longer trips and specific flight times might correlate with higher cancellation rates. These factors likely reflect the complexity and planning required for certain travel plans, which can increase the likelihood of changes or cancellations.

On the other hand, customer preferences such as **wants_extra_baggage**, **wants_in_flight_meals**, and **wants_preferred_seat** have lower importance scores, indicating that these factors, while relevant, are less critical in predicting whether a booking is completed. Overall, the analysis emphasizes the importance of timing, location, and travel logistics over specific customer preferences in determining booking outcomes.