British Airways – Forage Internship

By Andrew Sale

January 9th 2023

Executive Summary

After analyzing 50,000 flight searches we built a model for predicting whether the booking was completed. The model allowed us to identify the key features that determine the booking completion rate. These were purchase lead, departure time, length of stay, the flight duration, whether the search is made from Australia, Malaysia, or elsewhere, and the number of passengers.

Contents

Introduction

We are provided with details of 50,000 flight searches with British Airways (BA), containing information such as route, number of passengers, baggage, and day of week of departure. It also includes whether the search was converted into a complete booking.

The task is to determine which features of this search makes a user more likely to complete the booking.

Once the airline knows the importance of each feature, they can use this knowledge to focus their efforts to acquire more bookings, perhaps through focused advertising or adjusting the user experience accordingly.

Methodology

The process involved several key steps:

1. Exploratory data analysis. We visually inspected each feature to see if it is likely to influence booking completion.
2. Data preprocessing. Data was split for training and testing; categorical variables were handled with binning and one-hot encoding.
3. Modelling. Random grid searches were performed for random forest models and a histogram-based gradient boosted classification tree.
4. Model analysis. The receiver operating characteristic curves were the principal means for comparing model performance, along with model accuracy

For each step we used a jupyter notebook and python. Pandas was used to manage the dataset, sci-kit learn was used for modelling and analysis.

Results

Exploratory data analysis

Numerical variables

First, we looked at the distributions of the numerical variables, comparing the distributions when the booking was completed with those when it was not. Figure 1 shows that in most cases there is little difference between the distributions, suggesting that these variables will not be useful. There are some exceptions:

- Length of stay: Bookings for shorter trips tend to be more likely to be completed.
- Flight duration: Bookings for the longest flights are less likely to be completed.
- Extras: If extra baggage, preferred set, or in-flight meals are requested, then flights are slightly more likely to be booked.

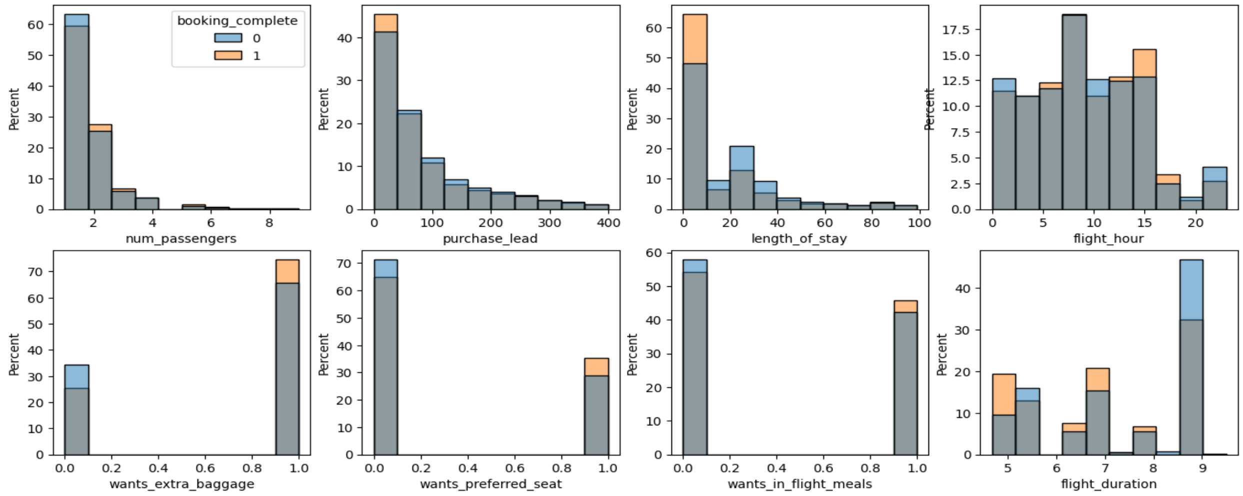

Figure 1: Comparison of distributions of numerical variables.

Booking origin

When it comes to the country in which the booking is made, there is a notable variation in completion rates, with some countries, such as Malaysia or Vietnam, having substantially higher completion rates (over 25%), while others, such as Australia and New Zealand, having much lower rates (around 5%). See Figure 2 for the variation among the 20 most popular booking origins.

Route

As is to be expected, since some booking origins have substantially higher completion rates than others, this naturally means that some routes have higher completion rates, as shown in Figure 3 for the 40 most searched routes.

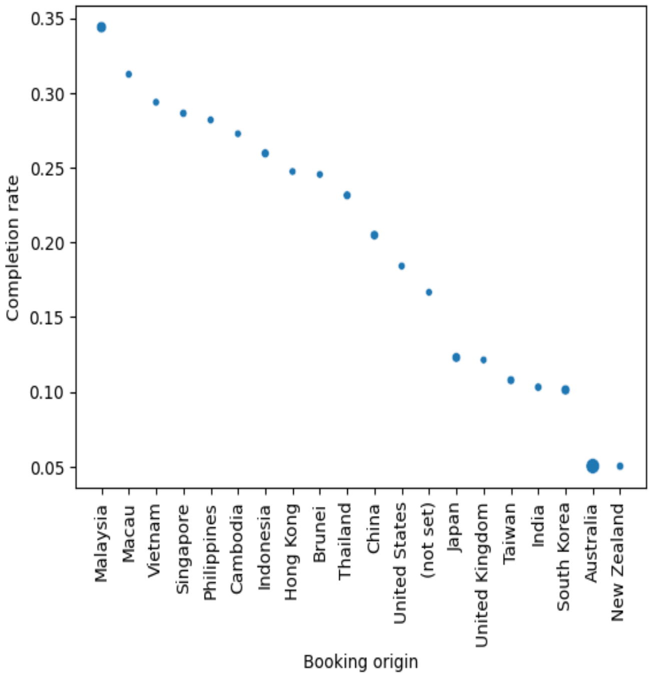

Figure 2: Completion rate by location of booking origin. The size of the dots represents the number of searches made from that location.

Booking origin and route correlation

For each route we looked at the different booking origins. Under the assumption that the two most frequent origins correspond to the departure and destination countries, we determined that only around 1 in 20 (5.5%) searches come from a country that is not involved in the route requested. We decided that this was low enough to disregard the route altogether in the modeling stage and instead allow it to focus on booking origin.

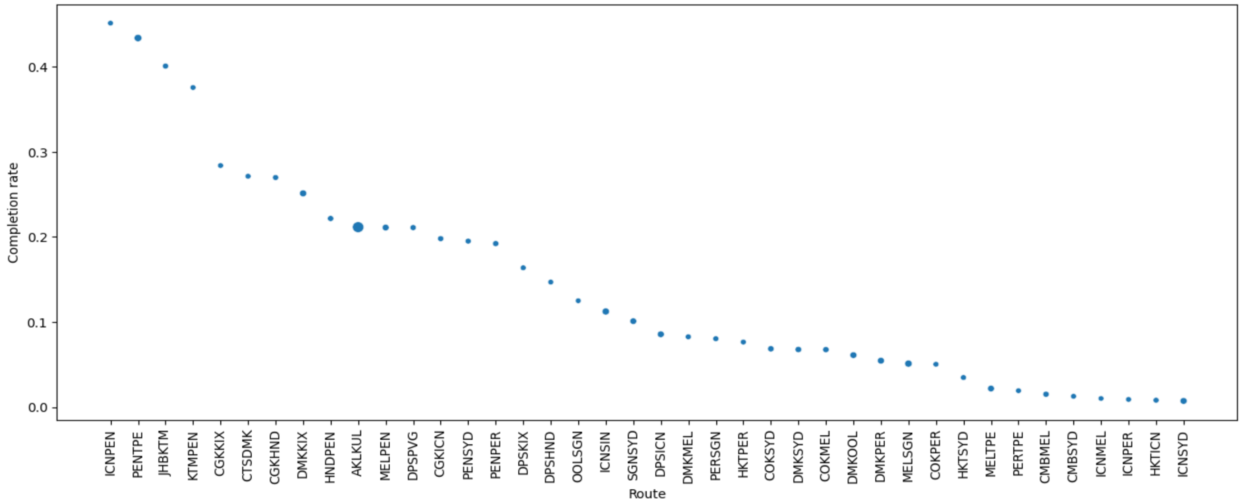

Figure 3: Completion rate of searches by route. Dot sizes represent number of searches made for that route.

Model performance

We performed a random grid search for a random forest classifier and a histogram-based gradient boosted tree classifier (both scikit-learn implementations).

Hyperparameters:

For the random forest, we varied:

- n_estimators: the number of trees used (geometric distribution, $p = 0.004$),
- max_depth: the max depth of each tree (uniform distribution for integers in $[2{,}100]$),
- min_samples_split: the number of samples required to split an internal node (uniform distribution for integers in $[2{,}40]$),
- max_samples: the proportion of samples used for each tree (beta distribution, $\alpha = 8, \beta = 3$).

The best performing model had:

n_estimators = 423, max_depth = 69, min_samples_split = 4, max_samples = 0.886.

For the gradient boosted classifier, we varied:

- n_iter: the maximum number of iterations (geometric distribution, $p = 0.01$),
- max_depth: the max depth of each tree (uniform distribution for integers in $[2{,}100]$),
- max_leaf_nodes: the maximum number of leaves per tree (uniform distribution for integers in $[20{,}60]$),
- l2_regularization: the regularization constant (beta distribution, $\alpha = 1, \beta = 6$)
- learning_rate: multiplication factor (beta distribution, $\alpha = 1.003, \beta = 20$).

The best performing model had:

max_iter = 792, max_depth = 17, max_leaf_nodes = 44, l2_regularization = 0.0414, learning_rate = 0.345.

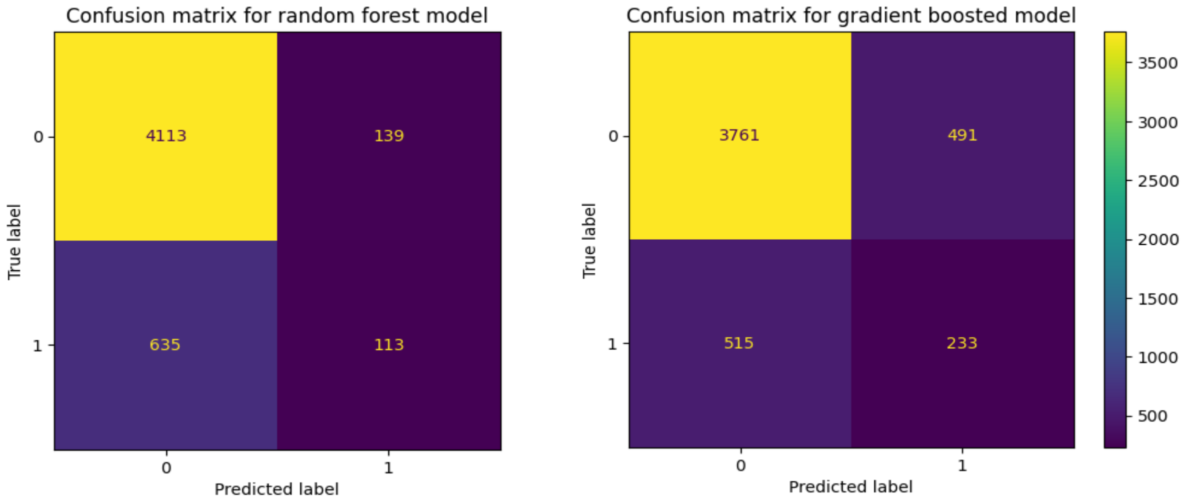

Figure 4: Confusion matrices for two optimum models.

Model evaluation

Confusion matrices for the two optimum models are shown below in Figure 4. We can see that the random forest model has higher overall accuracy, and performs very well at labelling incomplete bookings. However it has *low recall*: it fails to accurately label most of the complete bookings as such.

The gradient boosted model has slightly higher (but still low) recall, while its performance on incomplete bookings is worse.

Figure 5 shows the received operator curves (ROC) for the two optimum models. The greater the area underneath the curve, the better the model. We see that the random forest curve (blue) always lies above that of the gradient boosted model, indicating better performance.

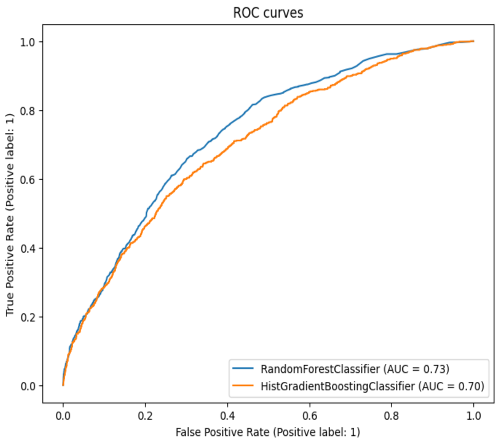

Figure 5: Receiver operator curves for each model.

The following table compares some metrics for each model. The highlighted scores are the better of the two.

	Random forest classifier	Gradient boosted classifier
Accuracy	84.5%	79.9%
ROC Area under curve	0.733	0.703
Recall	0.151	0.311
Precision	0.448	0.322

Feature importance

The random forest model allows us to identify which features have greatest impact on the outcome: whether a booking is completed or not. It counts how many times each feature is used to split a node in one of the trees. The more times it is used, the more important it is presumed to be.

Figure 6 shows the top 20 features, ordered by importance in the random forest model. The feature *purchase_lead* is a clear winner, with *flight_hour* and *length_of_stay* tied for second.

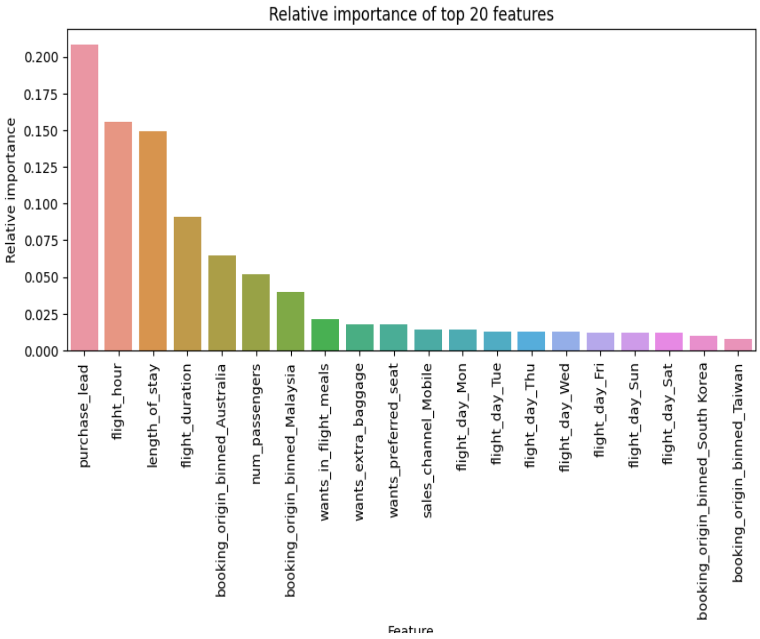

Figure 6: Feature importance.

We deduce the important features are:

- Purchase lead,
- Departure time,
- Length of stay,
- Flights duration,
- Whether the booking is made from Australia, Malaysia, or another country,
- Number of passengers.

Discussion

While the models produced were not very good at predicting whether a booking is completed or not, this is likely a difficult problem and we cannot expect an accurate model. For example, a user may search for the flight one day when comparing prices, and then the next day return to book it, with exactly the same search parameters. With this in mind, a new feature that records whether the user is returning to make the same search would help model performance. Nonetheless, we were still able to identify the important features.

Conclusion

The following table summarizes the important features and what conditions are necessary for higher completion rates.

Feature	**Condition for higher completion rates**
Purchase lead	Shorter lead time
Departure time	Afternoon departure
Length of stay	Shorter stay
Flights duration	Under 7 hours
Whether the booking is made from Australia, Malaysia, or another country	Malaysia
Number of passengers	More than 1