British Airways passenger booking project report
Springboard Data Science Track

Overall and problem:

This dataset provides comprehensive information about customers' preferences and behaviors related to airline holiday bookings. With detailed attributes covering various aspects of the booking process, this dataset is ideal for analyzing and understanding customer choices and patterns in the airline industry. The business problem solved here is British Airways. The goal is to use a trained model developed on the provided dataset to predict consumers' sales channels to book airline vacations.

Dataset:

https://www.kaggle.com/datasets/manishkumar7432698/airline-passangers-booking-data?select=Passanger_booking_data.csv

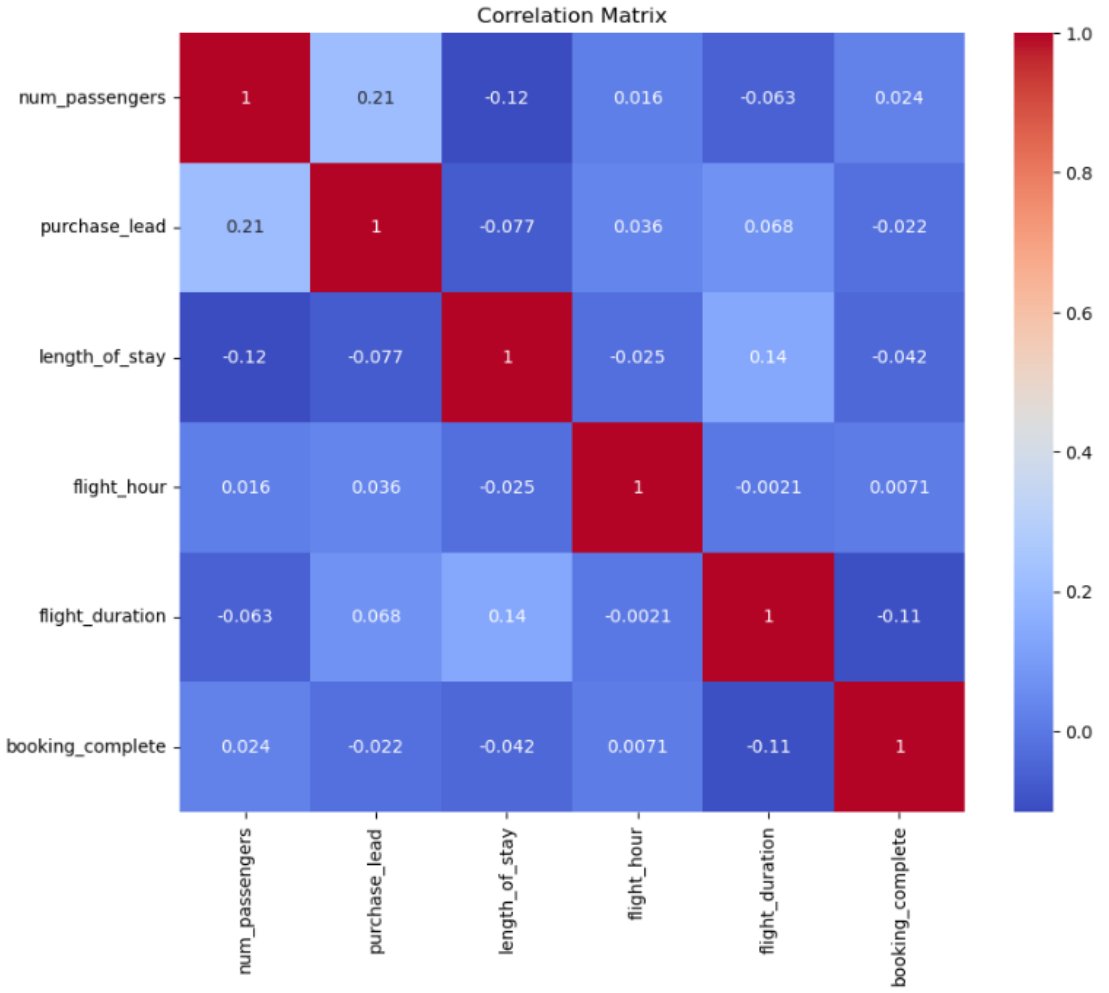
Preprocessing and Training Data:

First, we create dummy variables for trip_type. Add the dummies back to the data frame and remove the original column for trip_type. Then we standardize the magnitude of numeric features using a scaler. And finally, we split into testing and training datasets. After this series of operations, we found that the current dataset doesn't have any categorical data.

```
Index(['num_passengers', 'flight_hour', 'wants_extra_baggage',
       'wants_preferred_seat', 'wants_in_flight_meals', 'flight_duration',
       'booking_complete', 'CircleTrip', 'OneWay', 'RoundTrip',
       ...
       'Timor-Leste', 'Tonga', 'Tunisia', 'Turkey', 'Ukraine',
       'United Arab Emirates', 'United Kingdom', 'United States', 'Vanuatu',
       'Vietnam'],
      dtype='object', length=922)
```

EDA:

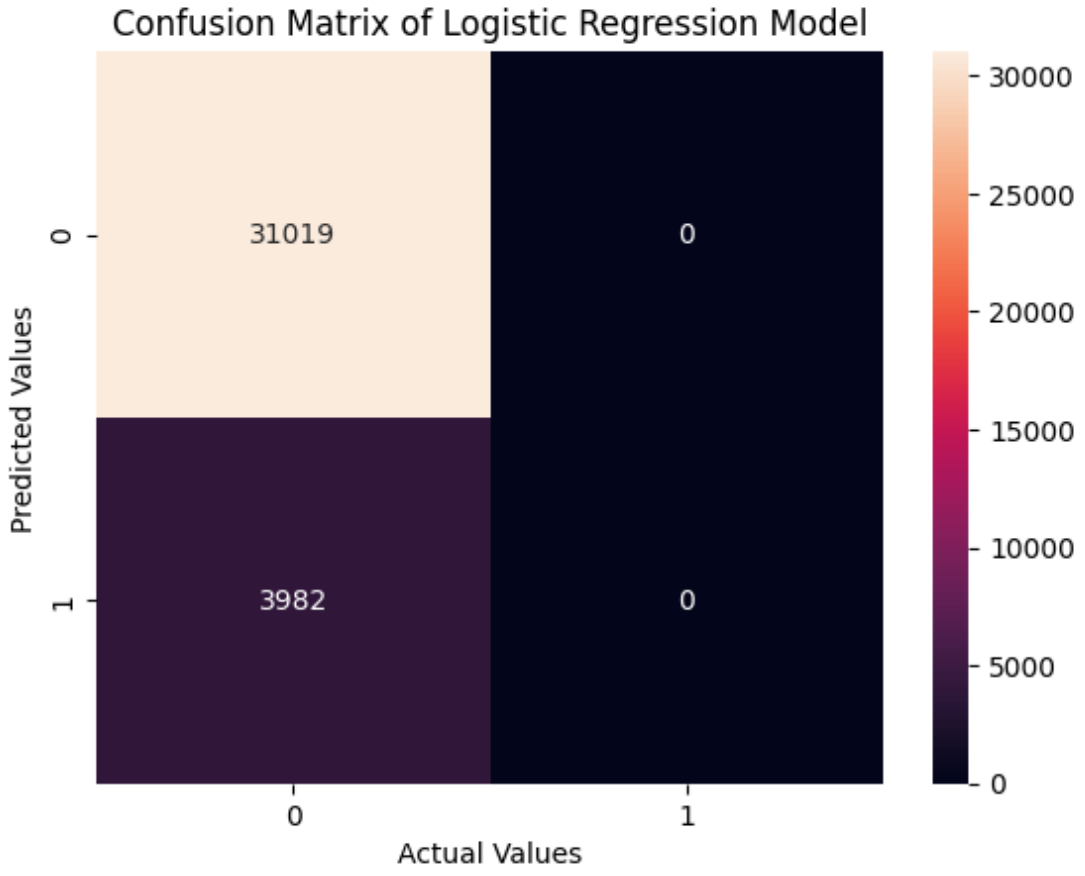
The heat map can most intuitively see the relationship between various variables.

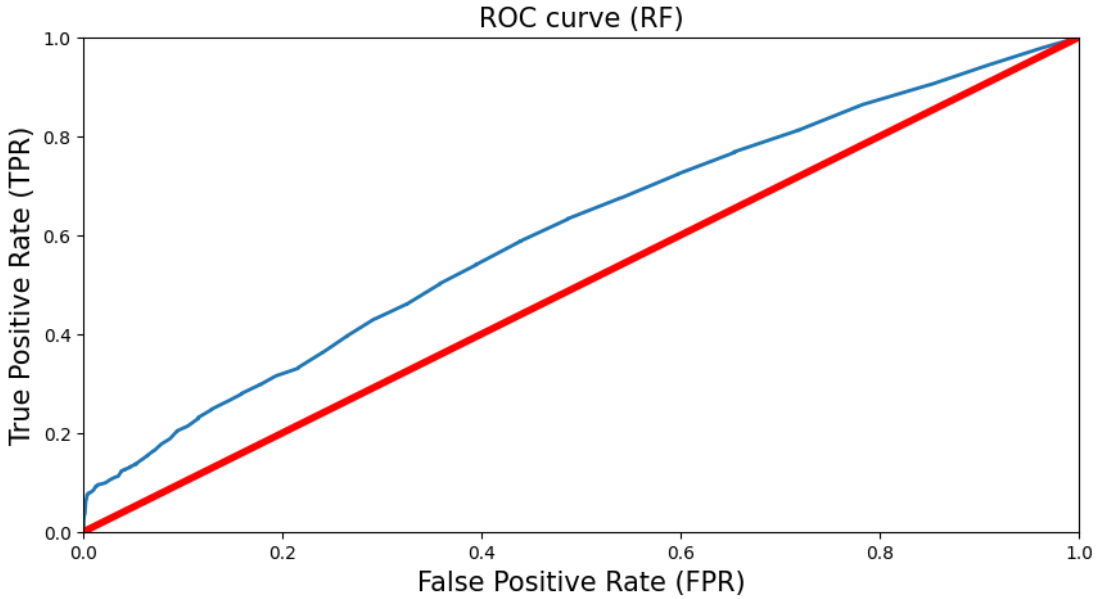


Modeling:

We used four different models to test the accuracy. Logistic Regression(Test Accuracy: 0.8908739417372176). K-Nearest neighbor (KNN)(Test Accuracy: 0.896273581761216). Random Forest(Test Accuracy: 0.892940470635291). Naive Bayes(Test Accuracy: 0.8908739417372176).

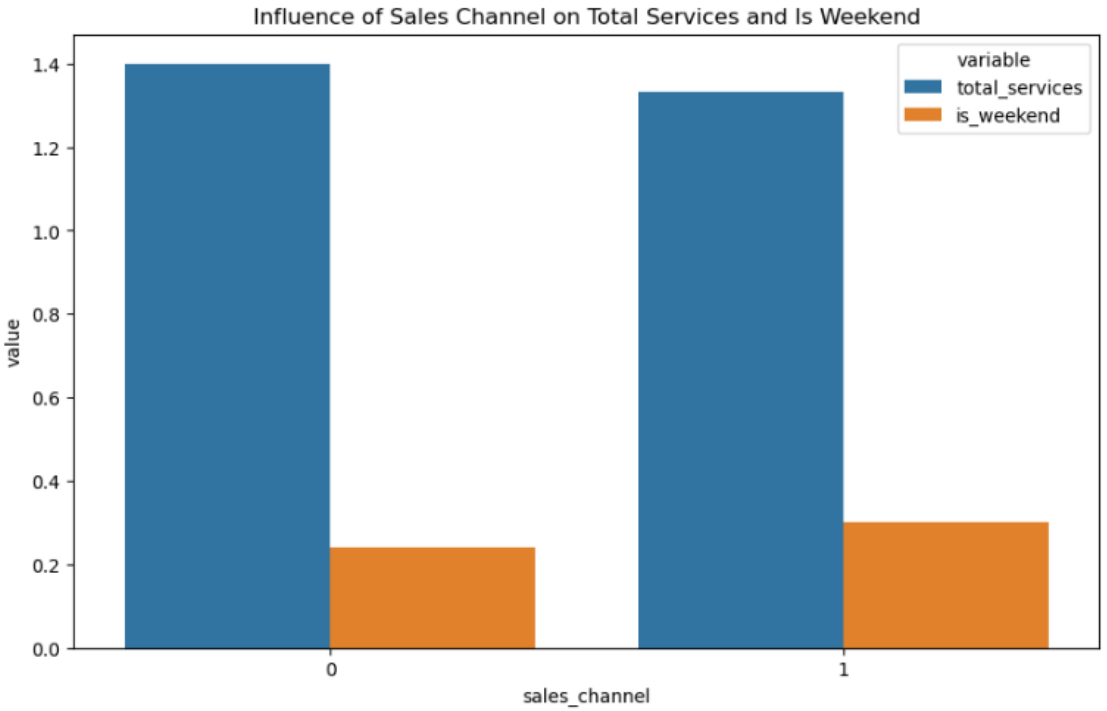
As we can see here the K-Nearest neighbor is the best-performing model. (Test Accuracy: 0.896273581761216). The most influential features are purchase_lead, route, flight_hour, length_of_stay, and booking_origin.

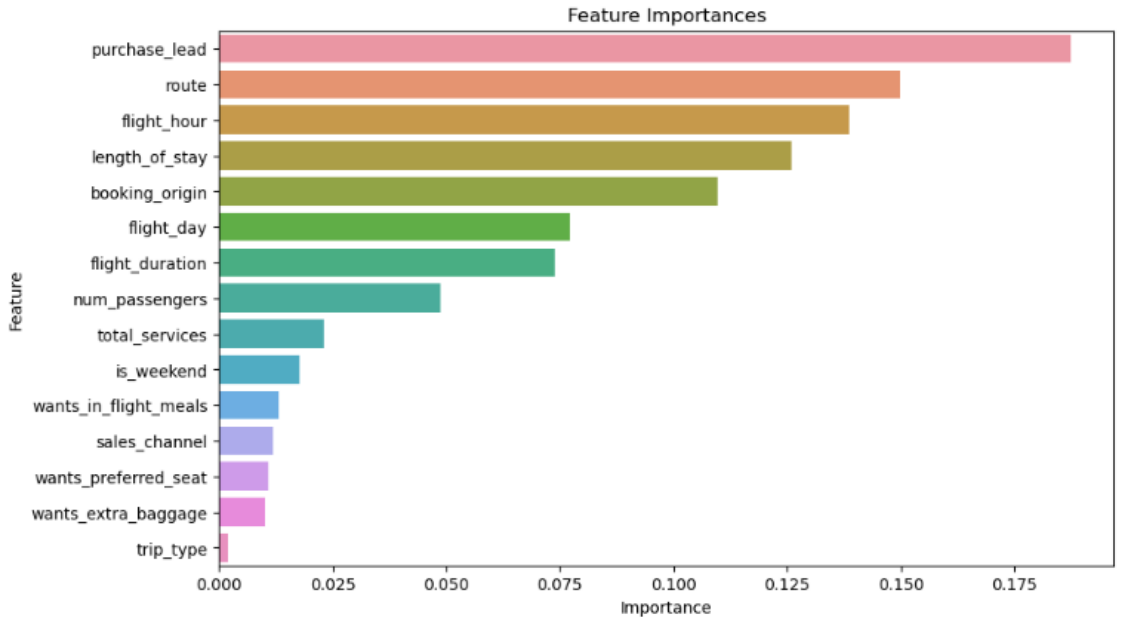




The above is the ROC-AUC Score and the ROC curve. Since I don't understand ROC-AUC Score and ROC curve very well, I didn't have a deep understanding after outputting the image.

Some predictions:





·British Airways may need to look at the density of bookings on weekends and analyze the exact times of popular flights.

·Because it can be seen from the specific data that the average number of additional services and the preference for weekend flights change with the change of booking lead time. Booking channels play an important part in this.

·The average values of variables like total_services and is_weekend vary between different sales channels.

·Based on the above speculation, the completion rate increases in direct proportion to the number of services and booking channels. From 10.68% for 0 services to 18.59% for 3 services.

Future Improvements:

I think we should continue to develop the convenience of mobile phone booking channels, although the data can show that the vast majority of users still insist on using the Internet to book tickets. But perhaps British Airways could benefit more from mobile bookings by increasing the number of total services and weekend flight preference, two key values.