Airline Booking Predictive Model

Presented By: Simidola Lawani

Data

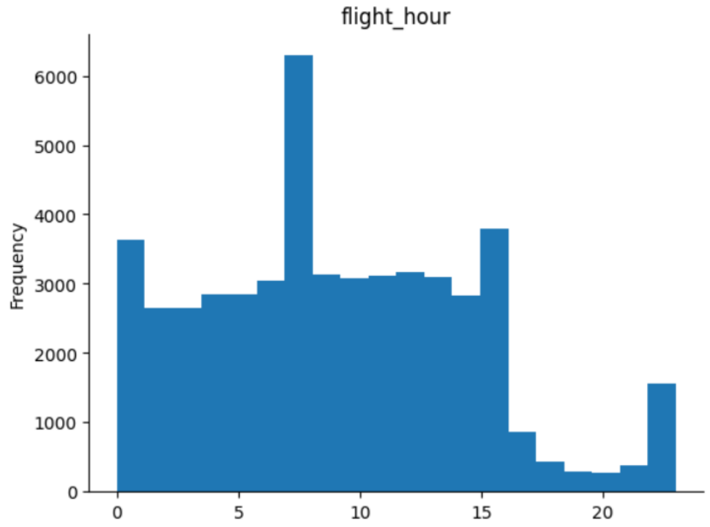
50000 data entries.

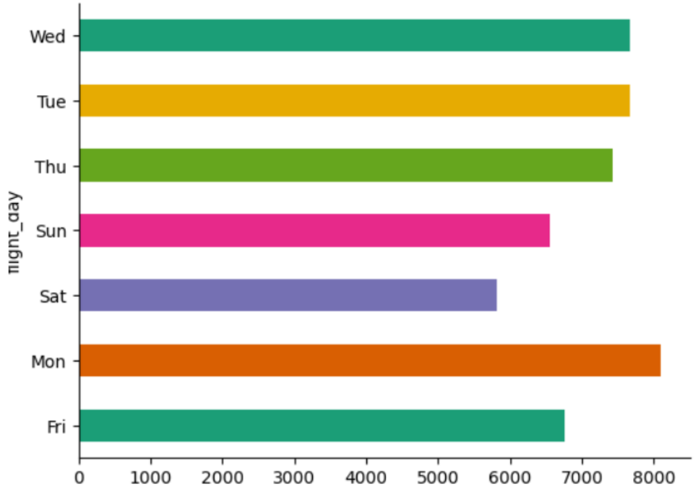
No null values

Predicting if customers will likely complete bookings or not

Exploratory Data Analysis

- With over 8000 bookings, Mondays appears to have the highest bookings of the days of the week with 8am being the most preferred time by customers.



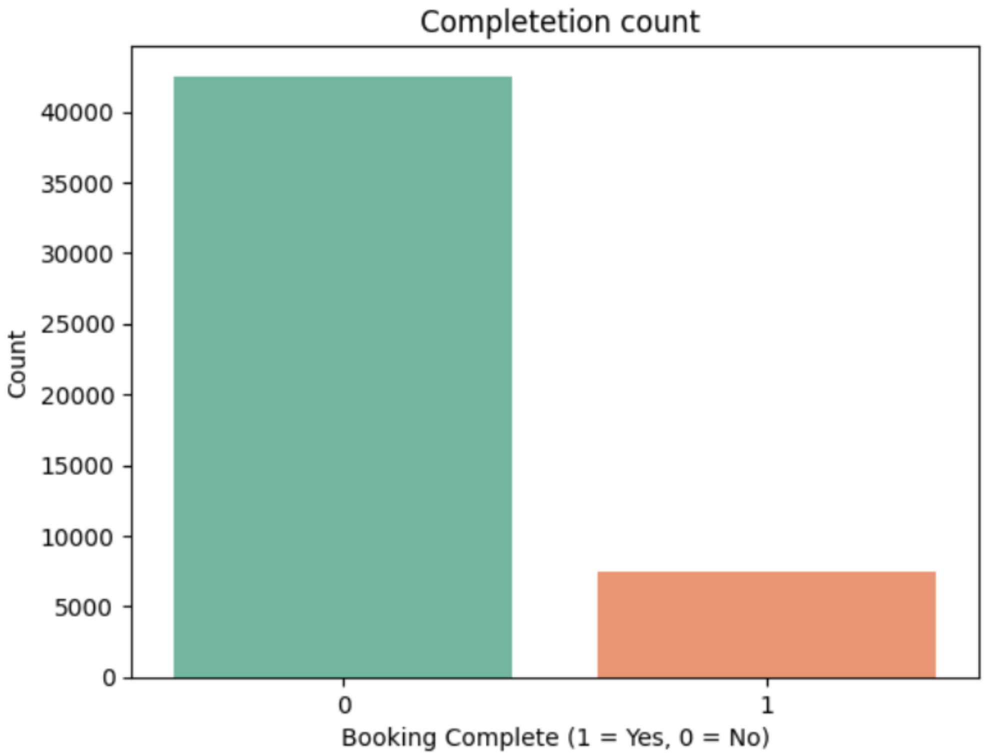


Exploratory Data Analysis

- It is apparent that there are more uncompleted bookings then there are completed bookings from the data visualization on the right. The reason why needs further investigation.

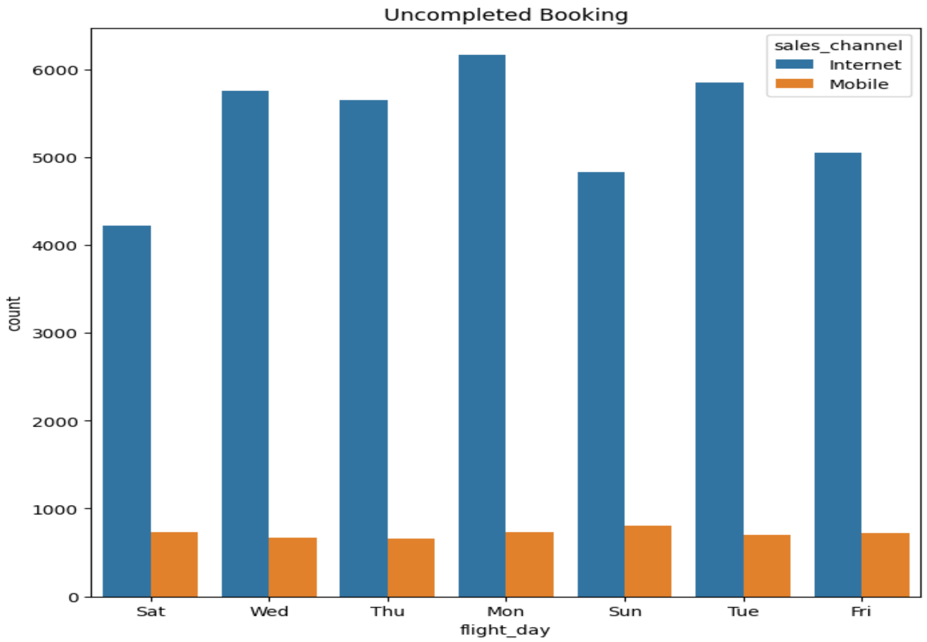
Uncompleted bookings: 42522

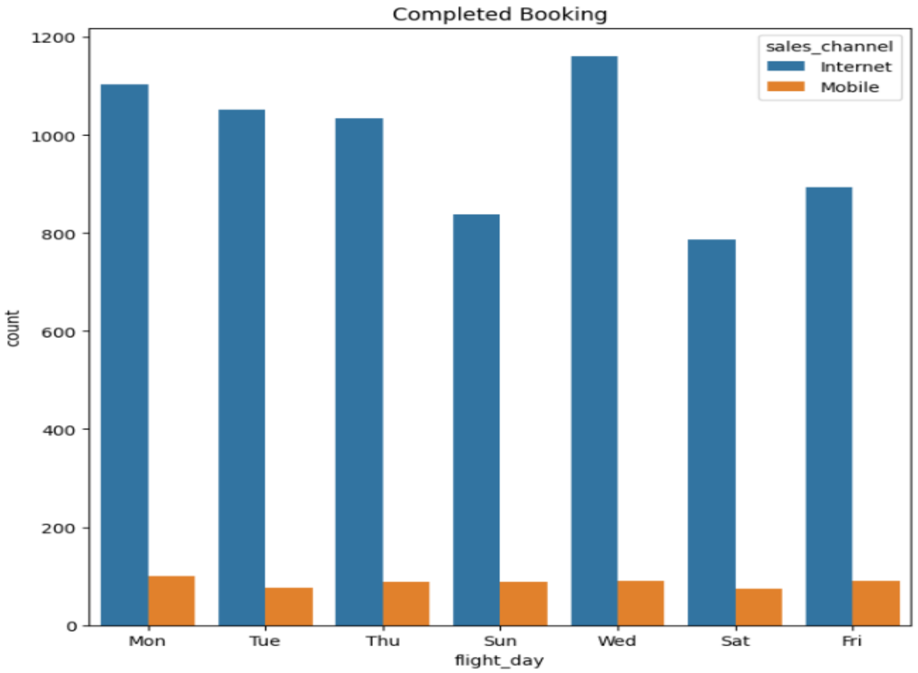
Completed bookings: 7478



Exploratory Data Analysis

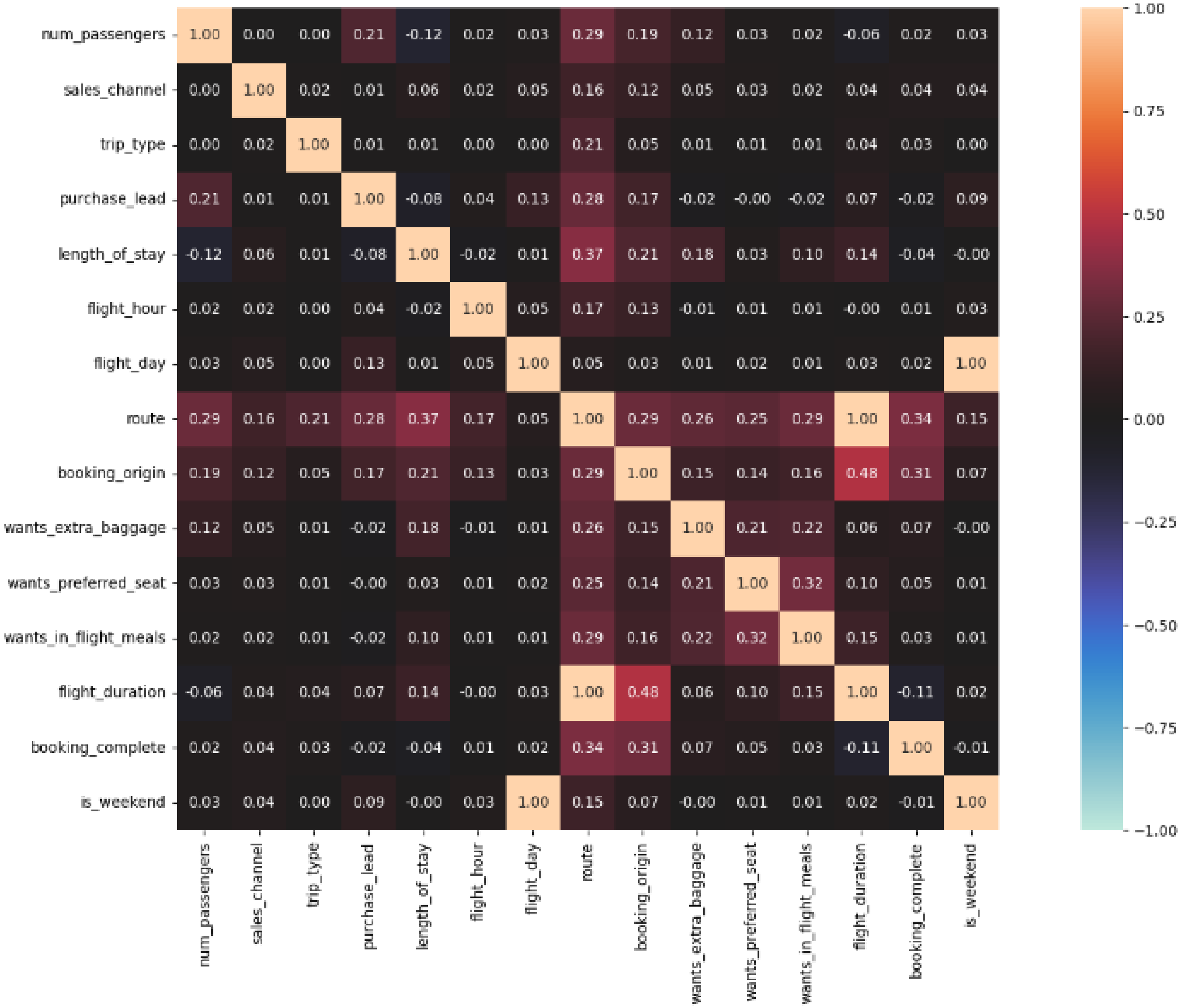
- It appears we have more completed bookings on Wednesday with data entries of about 1100 customers using the internet sales channel.
- Uncompleted bookings appears to be slightly more on the mobile sales channel which could be a pointer that more marketing needs to be carried out to attract users from that channel.





Correlated Variables)Columns)

- Introduced “is_weekend” by adding Sat and Sun columns together so it expected to see a correlation between “flight_day” and “is weekend”.
- Flight duration and route are also strongly correlated as one will expect.
- Booking_origin and flight_duration are also slightly correlated.



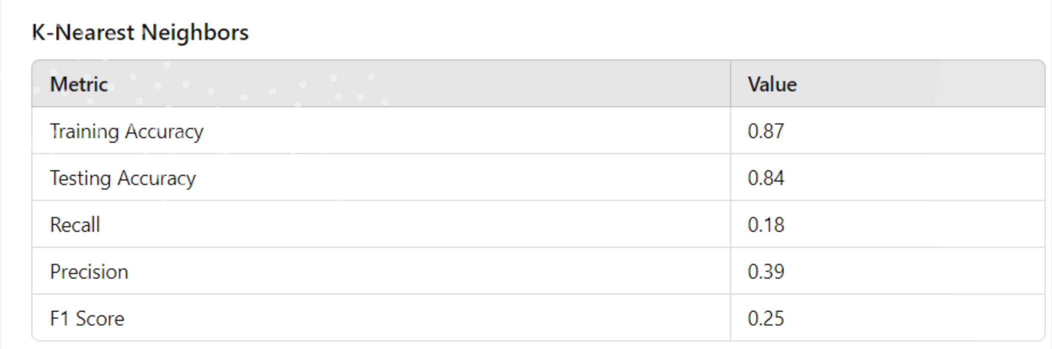
	num_passengers	sales_channel	trip_type	purchase_lead	length_of_stay	flight_hour	flight_day	route	booking_origin	wants_extra_baggage	wants_preferred_seat	wants_in_flight_meals	flight_duration	booking_complete	is_weekend
num_passengers	1.00	0.00	0.00	0.21	-0.12	0.02	0.03	0.29	0.19	0.12	0.03	0.02	-0.06	0.02	0.03
sales_channel	0.00	1.00	0.02	0.01	0.06	0.02	0.05	0.16	0.12	0.05	0.03	0.02	0.04	0.04	0.04
trip_type	0.00	0.02	1.00	0.01	0.01	0.00	0.00	0.21	0.05	0.01	0.01	0.01	0.04	0.03	0.00
purchase_lead	0.21	0.01	0.01	1.00	-0.08	0.04	0.13	0.28	0.17	-0.02	-0.00	-0.02	0.07	-0.02	0.09
length_of_stay	-0.12	0.06	0.01	-0.08	1.00	-0.02	0.01	0.37	0.21	0.18	0.03	0.10	0.14	-0.04	-0.00
flight_hour	0.02	0.02	0.00	0.04	-0.02	1.00	0.05	0.17	0.13	-0.01	0.01	0.01	-0.00	0.01	0.03
flight_day	0.03	0.05	0.00	0.13	0.01	0.05	1.00	0.05	0.03	0.01	0.02	0.01	0.03	0.02	1.00
route	0.29	0.16	0.21	0.28	0.37	0.17	0.05	1.00	0.29	0.26	0.25	0.29	1.00	0.34	0.15
booking_origin	0.19	0.12	0.05	0.17	0.21	0.13	0.03	0.29	1.00	0.15	0.14	0.16	0.48	0.31	0.07
wants_extra_baggage	0.12	0.05	0.01	-0.02	0.18	-0.01	0.01	0.26	0.15	1.00	0.21	0.22	0.06	0.07	-0.00
wants_preferred_seat	0.03	0.03	0.01	-0.00	0.03	0.01	0.02	0.25	0.14	0.21	1.00	0.32	0.10	0.05	0.01
wants_in_flight_meals	0.02	0.02	0.01	-0.02	0.10	0.01	0.01	0.29	0.16	0.22	0.32	1.00	0.15	0.03	0.01
flight_duration	-0.06	0.04	0.04	0.07	0.14	-0.00	0.03	1.00	0.48	0.06	0.10	0.15	1.00	-0.11	0.02
booking_complete	0.02	0.04	0.03	-0.02	-0.04	0.01	0.02	0.34	0.31	0.07	0.05	0.03	-0.11	1.00	-0.01
is_weekend	0.03	0.04	0.00	0.09	-0.00	0.03	1.00	0.15	0.07	-0.00	0.01	0.01	0.02	-0.01	1.00

Modelling

- Splitted data into training data(80%) and testing data(20%).
- Booking_complete was our target variable.
- Used ColumnTransformer for different preprocessing steps to categorical and numerical columns, expanding the dataset with one-hot encoded variables for categorical columns.
- Pipeline Explanation: The Pipeline ensures that scaling and logistic regression are combined into a single step, making it easier to maintain and evaluate the model. The StandardScaler normalizes the features, which is crucial for logistic regression as it improves convergence and model performance.
- Model Evaluation: The evaluation metrics (accuracy, recall, precision, F1 score) provide a comprehensive understanding of the model's performance. Accuracy is the overall correctness, recall measures how many actual positives were identified, precision calculates the proportion of positive identifications that were actually correct, and the F1 score is the harmonic mean of precision and recall.
- Feature Importance: By extracting and visualizing the coefficients from the logistic regression model, we got insights into which features are most influential in predicting the target variable. The absolute value of the coefficients indicates the importance.

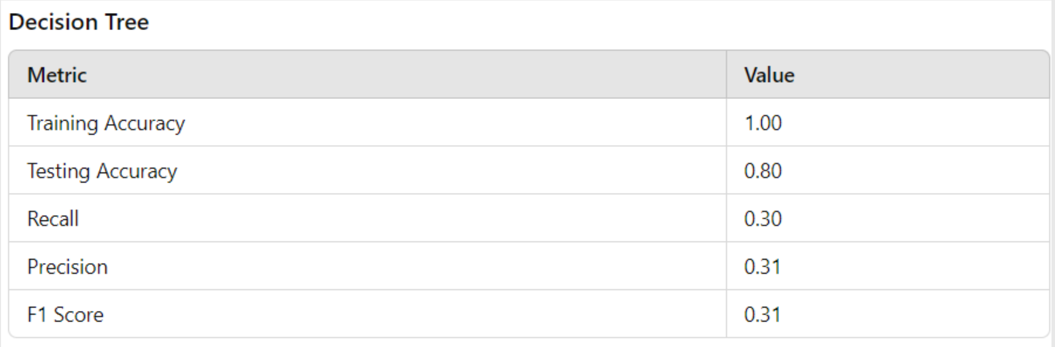
Models Results

Comparison of different models performance, which helps in quickly identifying which model performs better based on various metrics. In this group decision tree did extremely well with 100% on training and 80% on testing data.



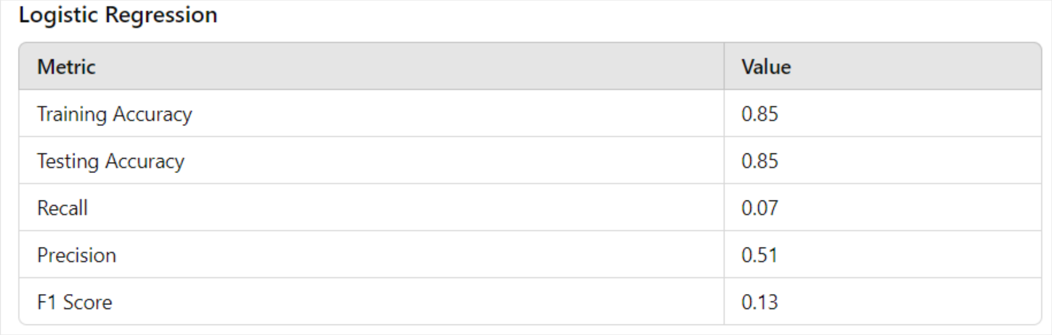
K-Nearest Neighbors

Metric	Value
Training Accuracy	0.87
Testing Accuracy	0.84
Recall	0.18
Precision	0.39
F1 Score	0.25



Decision Tree

Metric	Value
Training Accuracy	1.00
Testing Accuracy	0.80
Recall	0.30
Precision	0.31
F1 Score	0.31

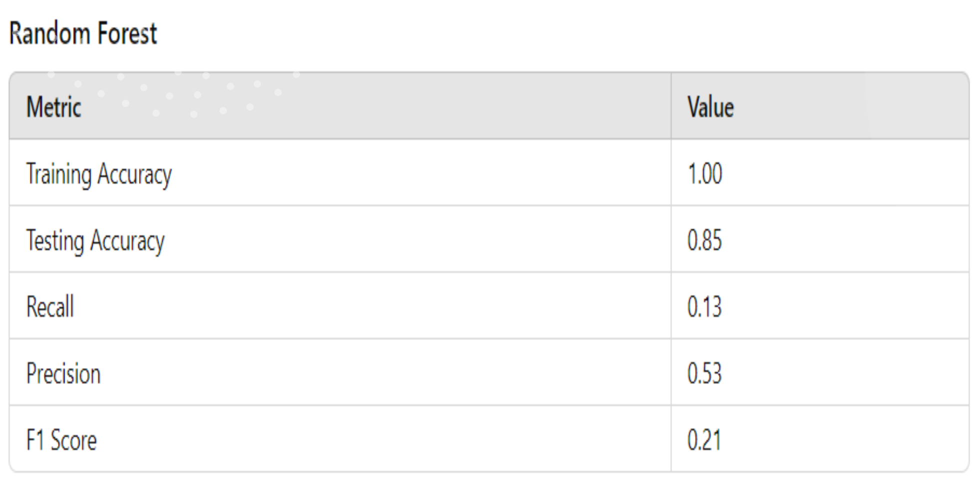


Logistic Regression

Metric	Value
Training Accuracy	0.85
Testing Accuracy	0.85
Recall	0.07
Precision	0.51
F1 Score	0.13

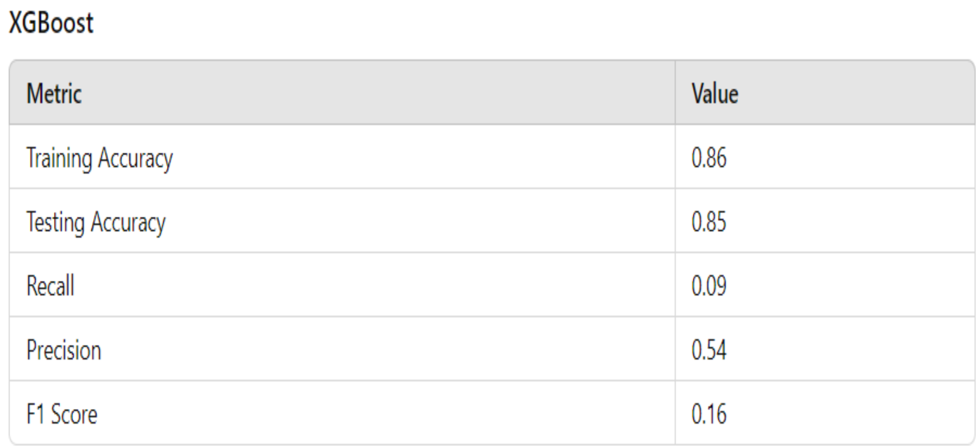
Models Results

- Radom forest prediction on the testing data set increased by 5% compared to that of the decision tree performance on testing data which was 80%. Xgboost also had 85% on the testing dataset.



Random Forest

Metric	Value
Training Accuracy	1.00
Testing Accuracy	0.85
Recall	0.13
Precision	0.53
F1 Score	0.21



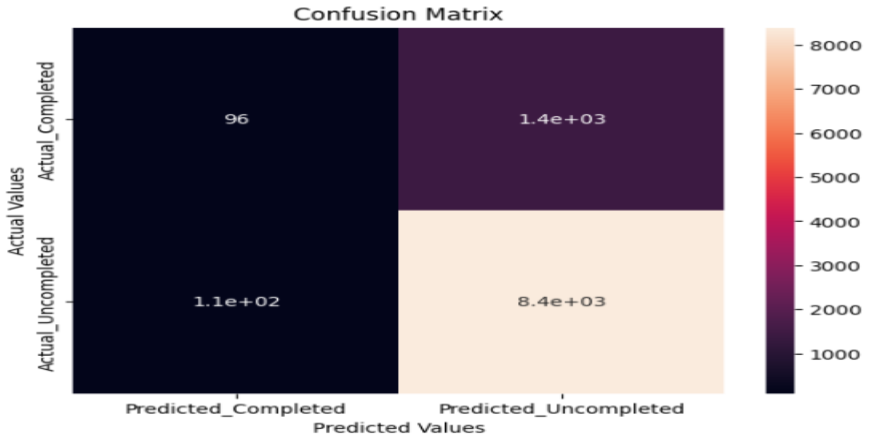
XGBoost

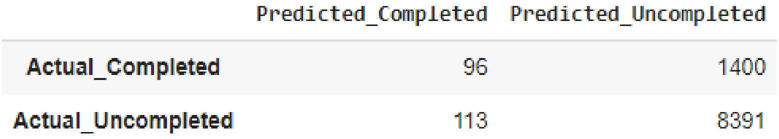
Metric	Value
Training Accuracy	0.86
Testing Accuracy	0.85
Recall	0.09
Precision	0.54
F1 Score	0.16

Model Results

From the initial exploratory data analysis that was carried out we observed that there are less people completing their bookings at a very high proportion, therefore it is expected that our models classifies less predicted completed bookings and more predicted uncompleted books.

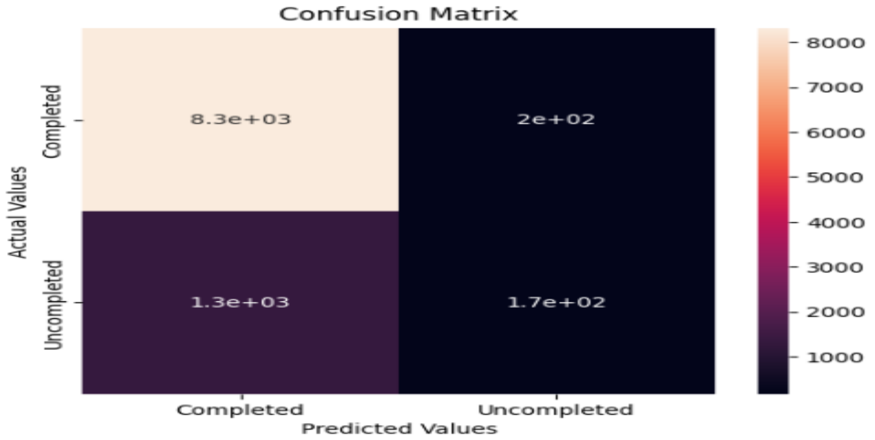
- Logistic Regression

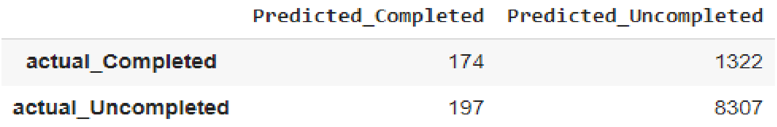




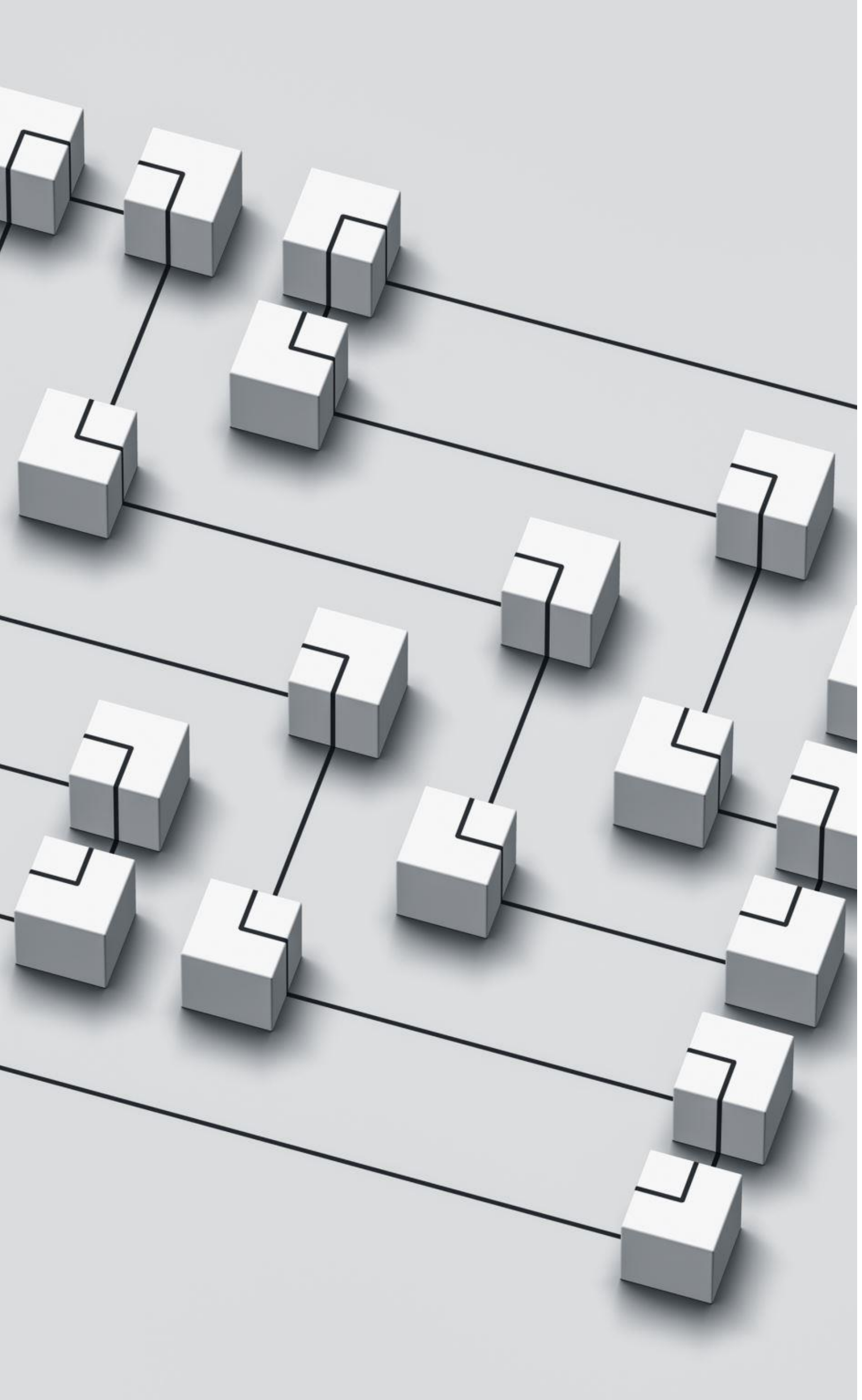
	Predicted_Completed	Predicted_Uncompleted
Actual_Completed	96	1400
Actual_Uncompleted	113	8391

- Random Forest





	Predicted_Completed	Predicted_Uncompleted
actual_Completed	174	1322
actual_Uncompleted	197	8307



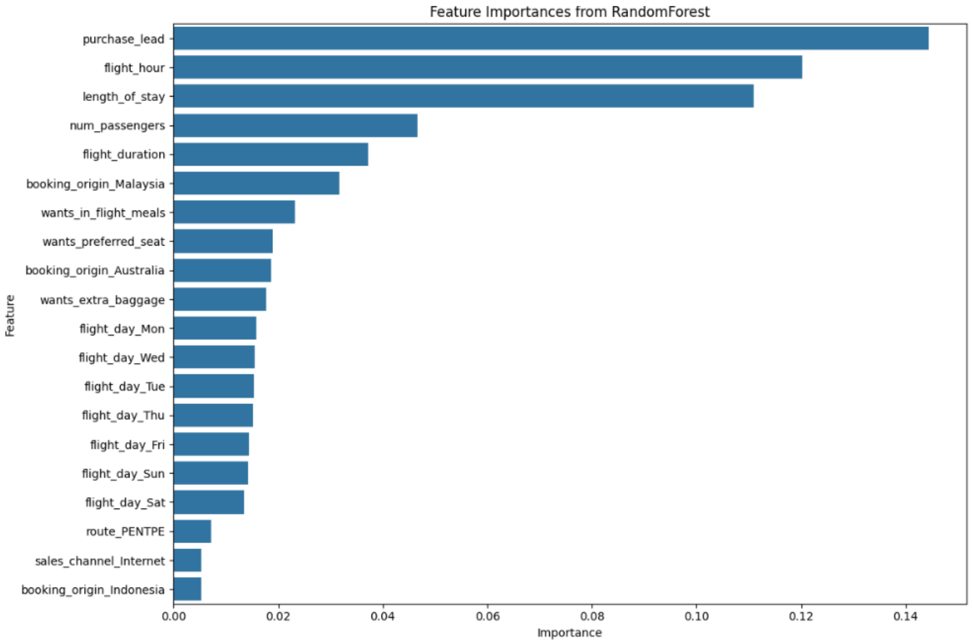
Finding Feature Imprtances

- Data Preprocessing: The ColumnTransformer used here applies different preprocessing steps to categorical and numerical columns, expanding the dataset with one-hot encoded variables for categorical columns.
- Model Training: The RandomForestClassifier is trained within the pipeline, allowing it to work directly with the transformed features.
- Feature Name Extraction: The function get_transformed_feature_names traverses the ColumnTransformer to fetch the transformed feature names, ensuring they match the expanded feature set after preprocessing.
- Feature Importances: Feature importances are extracted from the trained model and matched with the correct feature names, allowing for a comprehensive analysis of which features are most significant.Visualization: The bar plot visualizes the feature importances, highlighting the most impactful features for better interpretability.

Feature Importances

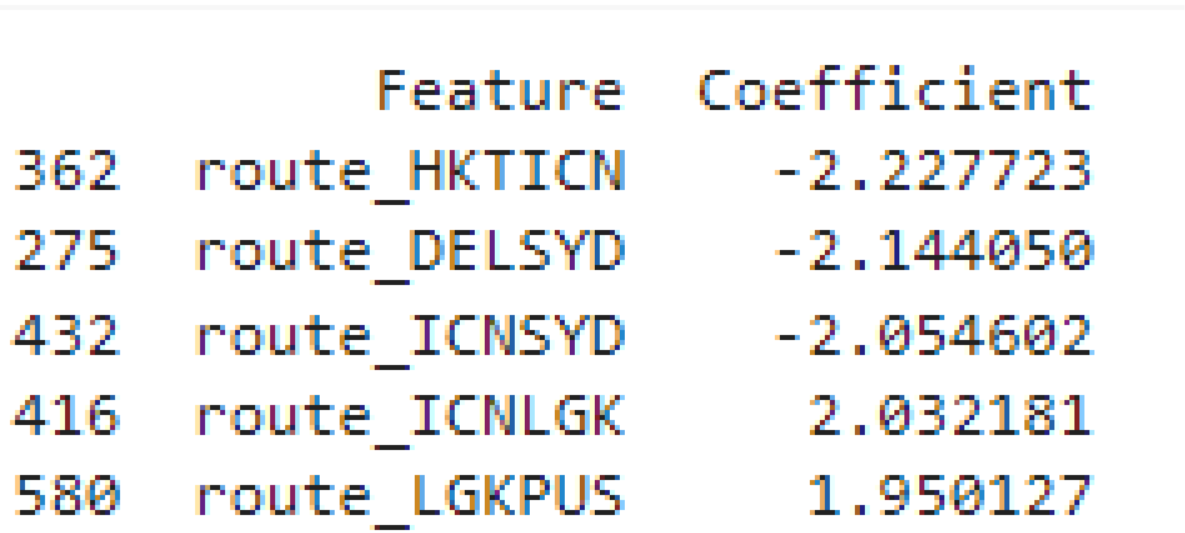
The horizontal bar chart "Feature Importances from RandomForest" on the right displays various features and their corresponding importance scores.

- Purchase Lead: The feature with the highest importance is "purchase_lead."
- Flight Hour of Day: The second most important feature is "flight_hour_of_day."
- Length of Stay: Following that, "length_of_stay" plays a significant role.
- Number of Passengers: The importance of the "num_passengers" feature is also notable.
- Flight duration is fifth important to the model prediction.



Feature Importances

- By extracting and visualizing the coefficients from the logistic regression model, we got insights into which features are most influential in predicting the target variable. The absolute value of the coefficients indicates the importance. Route appears to be the most importance with this method.



```
            Feature  Coefficient
362  route_HKTICN    -2.227723
275  route_DELSYD    -2.144050
432  route_ICNSYD    -2.054602
416  route_ICNLGK     2.032181
580  route_LGKPUS     1.950127
```

Next Steps

- Optimize Model Parameters: Tune hyperparameters for better accuracy.
- Explore More Features: Investigate new features to enhance the model.
- Business Application: Use insights for marketing and customer engagement strategies