Task 1 : Lounge Eligibility Model - What I Worked On

For Task 1, I worked on building a **simple and practical lounge eligibility model** to understand how many passengers might use different airport lounges in the future.

Since future flight schedules and passenger details are not always fixed, the focus of this task was to create a model that works **without relying on exact data**, but still makes sense from a business point of view.

How I approached the task

Instead of going flight by flight, I grouped flights in a **high-level and reusable way**:

- Short-haul routes
- Long-haul routes

This approach keeps the model easy to apply and flexible for future planning.

What I built

I created a **lookup table** that estimates the **average percentage of passengers** eligible for each lounge tier:

- Tier 1 (highest level access)
- Tier 2 (First Lounge access)
- Tier 3 (Club Lounge access)

The percentages are **average-based estimates**, meant only for planning and comparison, not exact passenger counts.

Lounge Eligibility Lookup Table

This table shows the estimated average percentage of passengers eligible for each lounge tier based on route type.

Percentages are based on high-level assumptions and are used only for planning purposes.

Route Type	Tier 1 (%)	Tier 2 (%)	Tier 3 (%)
Short Haul	2%	7%	18%
Long Haul	3%	11%	26%

Notes		
	Short Haul	Typically fewer premium and high-loyalty passengers.
	Long Haul	Higher share of premium cabins and frequent flyers.

Assumptions I used

To keep the model realistic:

- Long-haul routes were assumed to have **more premium and frequent flyers**
- Short-haul routes were assumed to have **fewer lounge-eligible passengers**
- Percentages were kept **conservative** to avoid overestimating demand

Why this model works

I also added a **justification section** to clearly explain:

- Why the flights were grouped this way
- How the assumptions were made
- How the same model can be reused even if schedules change

Because the model uses **broad categories**, it can be applied to **future or unknown schedules** without redoing everything.

Justification for Lounge Eligibility Model

Question	Answer
How did you choose to group the flights?	Flights were grouped into short-haul and long-haul routes to keep the model simple and easy to apply across different schedules.
Why do these groupings make sense?	Route type usually affects the mix of passengers, as long-haul flights tend to carry more premium and frequent-flyer customers than short-haul flights.
What assumptions did you make?	Lounge eligibility percentages were estimated using average values, with the assumption that long-haul routes have a higher share of premium and loyalty passengers, while short-haul routes have fewer.
How can your model apply to future or changing flight schedules?	Because the model is based on broad route categories rather than specific flights or aircraft, it can be easily reused for future or unknown schedules.

Justification Sheet

Tools I used

- Microsoft Excel (main output)

Final outcome (Conclusion)

The final result is a **clean, easy-to-understand model** that helps estimate lounge demand using simple assumptions and clear logic. This task helped me understand how data modeling can support real business decisions, even when information is limited.

Task 2: Predicting Customer Buying Behaviour

Introduction

In today's competitive airline industry, customers have access to extensive information and make purchase decisions well before travel. As a result, airlines must adopt proactive strategies to identify and target potential customers early in the buying cycle.

This task focuses on using customer booking data and machine learning techniques to predict whether a customer will complete a holiday booking. The objective is to build a predictive model, evaluate its performance, and interpret key factors influencing customer buying behaviour.

Dataset Overview

The dataset contains historical customer booking information, including travel details, booking behaviour, and customer preferences. These variables provide insights into customer intent and purchasing patterns.

- Each row represents a customer booking instance
- The target variable is **booking_complete**, where:
  - 0 = Booking not completed
  - 1 = Booking completed

	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
0	2	Internet	RoundTrip	262	19	7	Sat	AKLDEL	New Zealand	1	0	0	5.52	0
1	1	Internet	RoundTrip	112	20	3	Sat	AKLDEL	New Zealand	0	0	0	5.52	0
2	2	Internet	RoundTrip	243	22	17	Wed	AKLDEL	India	1	1	0	5.52	0
3	1	Internet	RoundTrip	96	31	4	Sat	AKLDEL	New Zealand	0	0	1	5.52	0
4	2	Internet	RoundTrip	68	22	15	Wed	AKLDEL	India	1	0	1	5.52	0

Sample view of the customer booking dataset

Data Preparation

Data preparation is a critical step to ensure model accuracy and reliability.

The following steps were performed:

- Missing values were identified and removed
- Categorical variables were converted into numerical form using label encoding
- The flight_day column, originally in text format, was encoded numerically
- All features were validated to ensure compatibility with machine learning algorithms

These steps ensured the dataset was clean, consistent, and suitable for model training.

Check Missing Values & if there are missing values simply drop

```
[8]: df.isnull().sum()

[8]: num_passengers          0
     sales_channel           0
     trip_type               0
     purchase_lead           0
     length_of_stay          0
     flight_hour             0
     flight_day              0
     route                   0
     booking_origin          0
     wants_extra_baggage     0
     wants_preferred_seat    0
     wants_in_flight_meals   0
     flight_duration         0
     booking_complete        0
     dtype: int64
```

Checking the Final Data that all are in numeric (d-type = int , float)

```
[18]: df.dtypes

[18]: num_passengers          int64
      sales_channel           int64
      trip_type               int64
      purchase_lead           int64
      length_of_stay          int64
      flight_hour             int64
      flight_day              int64
      route                   int64
      booking_origin          int64
      wants_extra_baggage     int64
      wants_preferred_seat    int64
      wants_in_flight_meals   int64
      flight_duration       float64
      booking_complete        int64
      dtype: object

      All columns are Numeric [Ready for ML]
```

Label encoding instead of hard-coded mapping

```
[15]: from sklearn.preprocessing import LabelEncoder
      le = LabelEncoder()
      df["flight_day"] = le.fit_transform(df["flight_day"])
```

Encode other Categorical varaibles (d-type = object)

```
[16]: categorical_cols = ["sales_channel","trip_type","route","booking_origin"]
      for col in categorical_cols:
          df[col] = le.fit_transform(df[col])
```

Data cleaning and encoding process

Model Selection and Training

A **Random Forest Classifier** was selected for this task due to its ability to:

- Handle complex, non-linear relationships
- Perform well on structured tabular data
- Provide feature importance for model interpretability

To address class imbalance in the dataset, balanced class weights were applied during training. The dataset was split into training and testing sets using an 80–20 ratio, while preserving class distribution.

Random Forest model training

Train Random Forest Model

```
[23]:
from sklearn.ensemble import RandomForestClassifier
rf = RandomForestClassifier(
    n_estimators=200,
    random_state=42,
    class_weight="balanced"
)
rf.fit(X_train, y_train)

[23]:
▾ RandomForestClassifier
▶ Parameters
```

Model Evaluation

1 Evaluation Metrics

The trained model was evaluated using multiple metrics to provide a balanced assessment of performance.

- **Accuracy:** 85.25%
- **Balanced Accuracy:** 53.68%

Although overall accuracy is high, balanced accuracy provides a more realistic evaluation due to class imbalance. The model performs very well in identifying non-booking customers but finds it challenging to correctly predict booking customers.

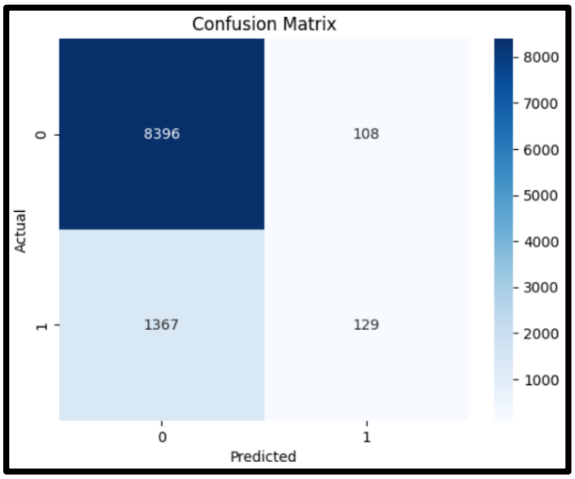
```
Accuracy: 0.8525
Balanced Accuracy: 0.5367650202987208

Classification Report:

              precision    recall  f1-score   support

           0       0.86      0.99      0.92      8504
           1       0.54      0.09      0.15      1496

    accuracy                           0.85     10000
   macro avg       0.70      0.54      0.53     10000
weighted avg       0.81      0.85      0.80     10000
```

Model evaluation metrics

2 Confusion Matrix Analysis

The confusion matrix provides a detailed breakdown of prediction results:

- The model correctly classifies most non-booking customers
- A large number of booking customers are misclassified as non-booking
- This behaviour highlights the challenge of predicting relatively rare booking events



Confusion matrix showing prediction performance

3 Cross-Validation

To ensure model stability and generalization, **5-fold stratified cross-validation** was performed.

- **Mean Cross-Validation Accuracy:** 85.33%

This result indicates consistent performance across multiple data splits and confirms that the model is not overfitting.

Cross-Validation (Stratified)

[27]:

```
skf = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)

cv_scores = cross_val_score(rf, X, y, cv=skf)
print("Mean CV Accuracy:", cv_scores.mean())
```

```
Mean CV Accuracy: 0.8533200000000001
```

Cross-validation results

Feature Importance Analysis

1 Feature Importance Table

Feature importance analysis was conducted to understand how each variable contributes to the predictive power of the model.

The most influential features include:

- Purchase lead time
- Route
- Booking origin
- Flight hour
- Length of stay

These features indicate that **booking timing and travel context** play a more significant role than ancillary preferences.

	Feature	Importance
3	purchase_lead	0.170446
7	route	0.154131
8	booking_origin	0.153575
5	flight_hour	0.125817
4	length_of_stay	0.125377
6	flight_day	0.080631
12	flight_duration	0.076710
0	num_passengers	0.043478
11	wants_in_flight_meals	0.020390
9	wants_extra_baggage	0.017607

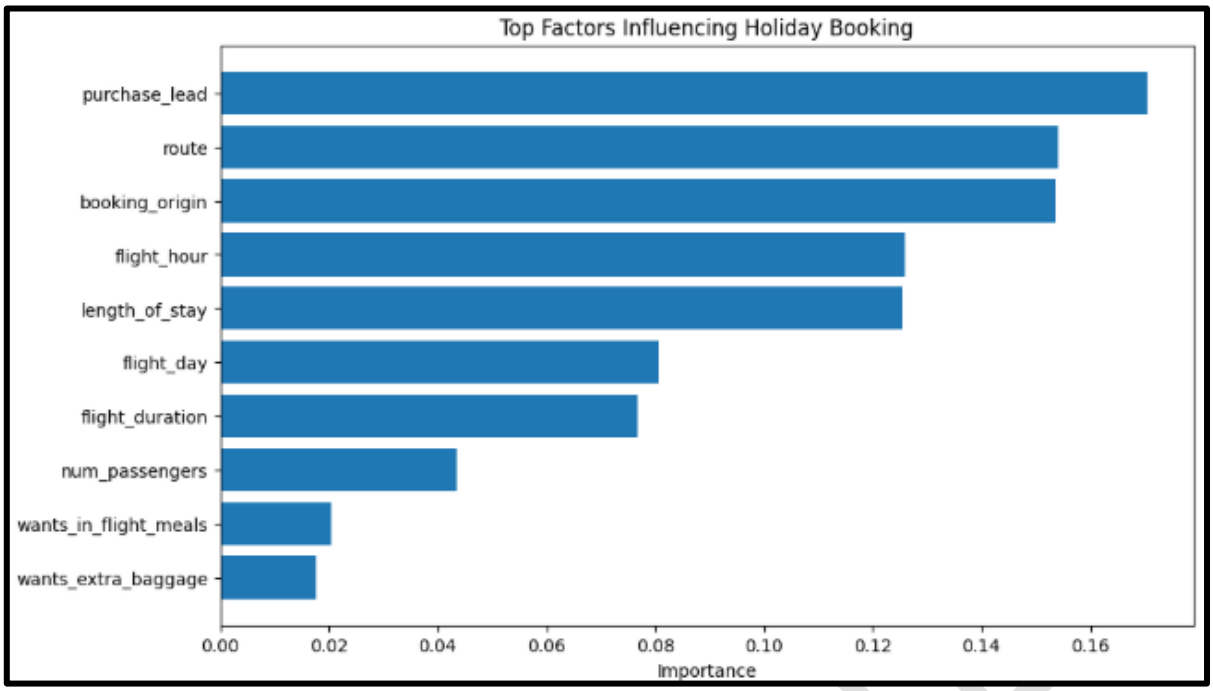
Top features influencing booking behaviour

2 Feature Importance Visualization

A horizontal bar chart was used to visually represent feature importance for easy interpretation.

Key observations:

- purchase_lead is the strongest predictor of booking behaviour
- Route and origin significantly influence booking decisions
- Ancillary services such as meals and baggage have relatively low impact



Visualization of key factors influencing holiday bookings

Business Insights

Based on the analysis, the following business insights were derived:

- Customers who plan earlier are more likely to complete holiday bookings
- Travel-related factors such as route and booking origin strongly affect buying behaviour
- Ancillary preferences have limited influence on booking decisions
- Predictive insights can be used to target high-intent customers before travel

These insights can support proactive marketing strategies and personalized offers.

Conclusion

This task successfully demonstrated the use of machine learning to predict customer buying behaviour using historical booking data. While predicting booking completion remains challenging due to class imbalance, the Random Forest model provided valuable insights into the key drivers of customer decisions.

The feature importance analysis highlights actionable factors that airlines can leverage to improve customer acquisition strategies and enhance early engagement.