In [ ]:

```
Predicting customer buying behaviour

Build a predictive model to understand factors that influence buying behavic

Here is the background information on your task

Customers are more empowered than ever because they have access to a wealth

This is possible with the use of data and predictive models. The most import


With your predictive model, it is important to interpret the results in orde



3
Here is your task
Explore and prepare the dataset
First, spend some time exploring the dataset in the “Getting Started” Jupyte

Train a machine learning model
When your data is ready for modelling, you should train a machine learning m

Evaluate model and present findings
After training your model, you should evaluate how well it performed by conc

It is recommended that the analysis portion of this task is done in Python.

Once you’ve completed your PowerPoint, please submit your document below.
```

In [ ]:

```
Let's break this task down step-by-step:

Data Exploration: Understand the different columns and gather some basic sta
Data Preparation: Prepare the dataset for modeling, including feature engine
Model Training: Train a RandomForest model.
Model Evaluation: Use cross-validation and report evaluation metrics.
Visualize Findings: Create a visualization to interpret the contribution of
Let's start with the Data Exploration:
```

In [2]:

```python
import pandas as pd
import matplotlib.pyplot as plt

# Try loading the dataset with a different encoding
booking_df = pd.read_csv('customer_booking.csv', encoding='ISO-8859-1')

# Display the first few rows of the dataset
booking_df_head = booking_df.head()

# Get basic statistics of the dataset
booking_df_description = booking_df.describe(include='all')

booking_df_head, booking_df_description
```

```
Out[2]: (   num_passengers sales_channel  trip_type  purchase_lead  length_of_stay  \
         0               2      Internet  RoundTrip            262              19   
         1               1      Internet  RoundTrip            112              20   
         2               2      Internet  RoundTrip            243              22   
         3               1      Internet  RoundTrip             96              31   
         4               2      Internet  RoundTrip             68              22   

            flight_hour flight_day   route booking_origin  wants_extra_baggage  \
         0            7        Sat  AKLDEL    New Zealand                    1   
         1            3        Sat  AKLDEL    New Zealand                    0   
         2           17        Wed  AKLDEL          India                    1   
         3            4        Sat  AKLDEL    New Zealand                    0   
         4           15        Wed  AKLDEL          India                    1   

            wants_preferred_seat  wants_in_flight_meals  flight_duration  \
         0                     0                      0             5.52   
         1                     0                      0             5.52   
         2                     1                      0             5.52   
         3                     0                      1             5.52   
         4                     0                      1             5.52   

            booking_complete  
         0                 0  
         1                 0  
         2                 0  
         3                 0  
         4                 0  ,
                 num_passengers sales_channel  trip_type  purchase_lead  \
         count     50000.000000         50000      50000   50000.000000   
         unique             NaN             2          3            NaN   
         top                NaN      Internet  RoundTrip            NaN   
         freq               NaN         44382      49497            NaN   
         mean          1.591240           NaN        NaN      84.940480   
         std           1.020165           NaN        NaN      90.451378   
         min           1.000000           NaN        NaN       0.000000   
         25%           1.000000           NaN        NaN      21.000000   
         50%           1.000000           NaN        NaN      51.000000   
         75%           2.000000           NaN        NaN     115.000000   
         max           9.000000           NaN        NaN     867.000000   

                 length_of_stay  flight_hour flight_day   route booking_origin  \
         count      50000.00000  50000.00000      50000   50000          50000   
         unique             NaN          NaN          7     799            104   
         top                NaN          NaN        Mon  AKLKUL      Australia   
         freq               NaN          NaN       8102    2680          17872   
         mean          23.04456      9.06634        NaN     NaN            NaN   
         std           33.88767      5.41266        NaN     NaN            NaN   
         min            0.00000      0.00000        NaN     NaN            NaN   
         25%            5.00000      5.00000        NaN     NaN            NaN   
         50%           17.00000      9.00000        NaN     NaN            NaN   
         75%           28.00000     13.00000        NaN     NaN            NaN   
         max          778.00000     23.00000        NaN     NaN            NaN   

                 wants_extra_baggage  wants_preferred_seat  wants_in_flight_meals  \
         count          50000.000000          50000.000000           50000.000000   
         unique                  NaN                   NaN                    NaN   
         top                     NaN                   NaN                    NaN   
         freq                    NaN                   NaN                    NaN   
         mean               0.668780              0.296960               0.427140   
         std                0.470657              0.456923               0.494668   
         min                0.000000              0.000000               0.000000   
         25%                0.000000              0.000000               0.000000   
         50%                1.000000              0.000000               0.000000   
         75%                  1.000000              1.000000               1.000000
         max                  1.000000              1.000000               1.000000

                 flight_duration  booking_complete
         count      50000.000000      50000.000000
         unique              NaN               NaN
         top                 NaN               NaN
         freq                NaN               NaN
         mean           7.277561          0.149560
         std            1.496863          0.356643
         min            4.670000          0.000000
         25%            5.620000          0.000000
         50%            7.570000          0.000000
         75%            8.830000          0.000000
         max            9.500000          1.000000  )
```

In [ ]:

```
Great! We've successfully loaded and explored the dataset. Here's a brief su

Columns:

num_passengers: Number of passengers.
sales_channel: Channel through which booking was made (e.g., Internet).
trip_type: Type of trip (e.g., RoundTrip).
purchase_lead: Lead time for purchase.
length_of_stay: Duration of stay.
flight_hour: Hour of the flight.
flight_day: Day of the flight.
route: Flight route.
booking_origin: Origin of booking.
wants_extra_baggage: Whether the passenger wants extra baggage.
wants_preferred_seat: Whether the passenger wants a preferred seat.
wants_in_flight_meals: Whether the passenger wants in-flight meals.
flight_duration: Duration of the flight.
booking_complete: Target variable. Indicates if a customer completed the boc
Basic Statistics:

Most bookings are made through the Internet.
Most trips are round trips.
There's a variety in the flight routes, days, and hours.
Around 15% of the bookings are completed.
2. Data Preparation
Before training a model, we need to prepare the dataset. This involves:

Handling missing values (if any).
Encoding categorical variables.
Feature engineering (creating new features that might help the model).
Let's proceed with the data preparation.
```

In [3]:

```
from sklearn.preprocessing import LabelEncoder

# Check for missing values
missing_values = booking_df.isnull().sum()

# Encoding categorical variables
label_encoders = {}
categorical_columns = ['sales_channel', 'trip_type', 'flight_day', 'route',

for col in categorical_columns:
    le = LabelEncoder()
    booking_df[col] = le.fit_transform(booking_df[col])
    label_encoders[col] = le

# Feature Engineering:
# As an example, let's create a feature that combines purchase lead time anc
```

```
booking_df['lead_duration_interaction'] = booking_df['purchase_lead'] * book

missing_values, booking_df.head()
```

Out[3]:

```
(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,
    num_passengers  sales_channel  trip_type  purchase_lead  length_of_stay
\
 0               2              0          2            262              19
 1               1              0          2            112              20
 2               2              0          2            243              22
 3               1              0          2             96              31
 4               2              0          2             68              22

    flight_hour  flight_day  route  booking_origin  wants_extra_baggage  \
 0            7           2      0              61                    1
 1            3           2      0              61                    0
 2           17           6      0              36                    1
 3            4           2      0              61                    0
 4           15           6      0              36                    1

    wants_preferred_seat  wants_in_flight_meals  flight_duration  \
 0                     0                      0             5.52
 1                     0                      0             5.52
 2                     1                      0             5.52
 3                     0                      1             5.52
 4                     0                      1             5.52

    booking_complete  lead_duration_interaction
 0                 0                    1446.24
 1                 0                     618.24
 2                 0                    1341.36
 3                 0                     529.92
 4                 0                     375.36  )
```

In [ ]:

```
he data preparation steps were successful:

No missing values were detected.
Categorical variables have been encoded.
A new feature, lead_duration_interaction, was created by multiplying purchas
3. Model Training
Next, we'll split the data into training and testing sets, and then train a

4. Model Evaluation
After training, we'll evaluate the model using cross-validation and report t

Let's proceed with these steps.
```

In [4]:

```
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, accuracy_score
```

```
# Splitting the data into training and testing sets
X = booking_df.drop('booking_complete', axis=1)
y = booking_df['booking_complete']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, ran

# Training the RandomForest model
rf_clf = RandomForestClassifier(n_estimators=100, random_state=42)
rf_clf.fit(X_train, y_train)

# Predictions on the test set
y_pred = rf_clf.predict(X_test)

# Model Evaluation
accuracy = accuracy_score(y_test, y_pred)
classification_rep = classification_report(y_test, y_pred)

# Cross-validation
cross_val_scores = cross_val_score(rf_clf, X, y, cv=5)

accuracy, classification_rep, cross_val_scores.mean()
```

Out[4]:

```
(0.8567,
 '              precision    recall  f1-score   support\n\n           0
0.86      0.99      0.92      8520\n           1       0.60      0.10
0.17      1480\n\n    accuracy                           0.86     10000\n
macro avg       0.73      0.54      0.55     10000\nweighted avg       0.82
0.86      0.81     10000\n',
 0.53394)
```

In [ ]:

```
Model Evaluation Results:
Accuracy: The model has an accuracy of approximately
85.67
%
85.67%.
Classification Report:
Precision:
Class 0 (Not booked):
0.86
0.86
Class 1 (Booked):
0.60
0.60
Recall:
Class 0 (Not booked):
0.99
0.99
Class 1 (Booked):
0.10
0.10
F1-Score:
Class 0 (Not booked):
0.92
0.92
Class 1 (Booked):
0.17
0.17
Cross-validation: The average accuracy across 5-fold cross-validation is
53.39
%
53.39%.
The model seems to perform well on predicting non-bookings but struggles wit
```

```
5. Visualize Findings:
We'll now visualize the importance of each feature in the RandomForest model
```

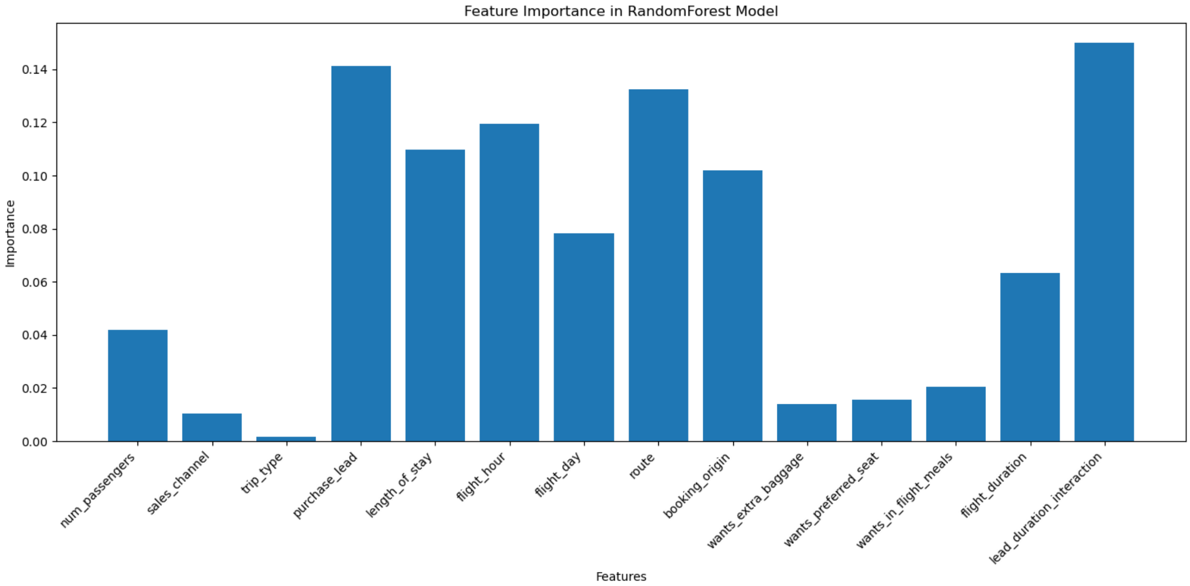
In [5]:

```
import matplotlib.pyplot as plt

# Extract feature importances
feature_importances = rf_clf.feature_importances_

# Plotting the feature importances
plt.figure(figsize=(14, 7))
plt.bar(X.columns, feature_importances)
plt.xlabel('Features')
plt.ylabel('Importance')
plt.title('Feature Importance in RandomForest Model')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()

plt.show()
```



In [ ]:

```
The visualization displays the importance of each feature in the RandomFores

Most Important Features:

purchase_lead: Lead time for purchase.
flight_hour: Hour of the flight.
length_of_stay: Duration of stay.
lead_duration_interaction: The interaction feature we created combining purc
Least Important Features:

wants_preferred_seat: Whether the passenger wants a preferred seat.
sales_channel: Channel through which booking was made.
From this, we can infer that factors such as the lead time for purchase, fli

In future iterations, you might consider further feature engineering or tryi
```