

# Airlines Booking Prediction Project

We will approach this problem using complete MLDLC (Machine Learning Development Life Cycle).

## Step-1: Framing the Problem

Q1) What is the actual problem?

A1) Our problem is to predict whether a customer will book a flight on the basis of his/her preferences, trip specifications.

Q2) Who is our target audience?

A2) Airline Companies (Example, British Airways) who are having sufficient knowledge of customer preferences, most preferred trip details and are willing to classify whether the customer will book the flight or not on the basis of a particular set of attributes.

Q3) What are the basic characteristics of ML Model for this problem?

A3) To solve this problem I will be using supervised machine learning and training will be done in offline mode(without the use of any API). This will be a classification problem.

Q4) What do I know about data?

A4) I will use the data from Kaggle ('Passanger\_booking\_data.csv'). It has data of 50,002 customers. As input I have 13 columns and as output I have 'booking\_complete'. All the columns have numerical data except ['sales\_channel', 'trip\_type', 'flight\_day', 'route', 'booking\_origin'] as it is categorical data.

Q5) What will be my approach to solve this issue?

A5) I will follow all the steps of ML Development Life Cycle till 'Model Deployment'.

## Step-2: Data Gathering

Data is used from **Kaggle**. Provided By **Manish Kumar**.

Source: [https://www.kaggle.com/datasets/manishkumar7432698/airline-passangers-booking-data?select=Passanger\\_booking\\_data.csv](https://www.kaggle.com/datasets/manishkumar7432698/airline-passangers-booking-data?select=Passanger_booking_data.csv)

Data Sample:

	num_passengers	sales_channel	trip_type	purchase_lead	length_of_stay	flight_hour	flight_day	route	booking_origin
22674	3	Internet	RoundTrip	76	19	1	Wed	PENPVG	Malaysia
44936	2	Mobile	RoundTrip	248	6	11	Thu	DMKPVG	China
18599	2	Internet	RoundTrip	60	24	13	Sat	MELMNL	Australia
36193	1	Internet	RoundTrip	2	5	15	Fri	COKPER	Australia
45386	2	Internet	RoundTrip	99	6	2	Thu	DPSKIX	Japan

booking_origin	wants_extra_baggage	wants_preferred_seat	wants_in_flight_meals	flight_duration	booking_complete
Malaysia	1	0	0	5.33	0
China	0	0	0	5.33	0
Australia	1	0	0	8.83	0
Australia	0	0	0	5.62	0
Japan	1	0	0	7.00	0

About Data:

Inputs:

- 1) num\_passengers: Number of passengers associated with each booking.
- 2) sales\_channel: How customers reached to the website.
- 3) trip\_type: Whether booking is for one-way trip or round(two-way) trip.
- 4) purchase\_lead: Duration between booking date and travel date.
- 5) length\_of\_stay: Duration of holiday stay.
- 6) flight\_hour: Specifies hour of flight.
- 7) flight\_day: Specifies day of week for flight.
- 8) route: Specifies flight route.
- 9) booking\_origin: Source of booking.
- 10) wants\_extra\_baggage: Whether customer desires extra baggage allowance.
- 11) wants\_preferred\_seat: Whether customer desires preferred seats.
- 12) wants\_in\_flight\_meals: Whether customer desires meal during the flight.
- 13) flight\_duration: Duration of the flight.

Output:

booking\_complete: Whether customer successfully booked the flight or not.

### Step-3: Data Preprocessing

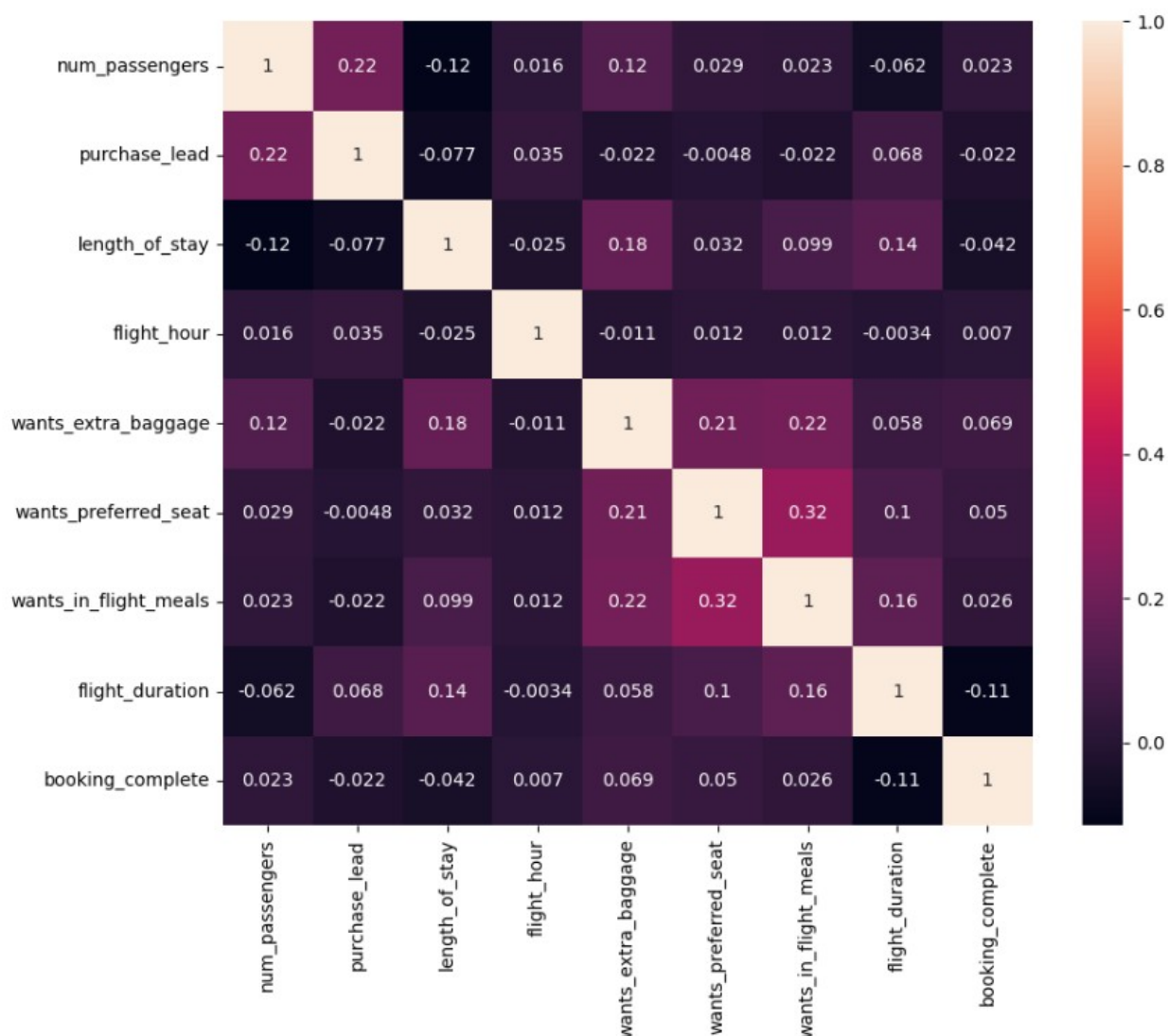
- To save memory I will reduce the datatypes to int8, int16 and float16, and convert object to category. Using this step I reduced the memory usage from 5.3+ MB to 923.1+ KB (Almost 82.58% Decrease).
- I also dropped 719 duplicate values.
- Since 99% of customers are willing to book 'RoundTrip'. So, I dropped other rows.
- There are almost 104 countries and very few passengers are there from these countries. So, I aggregated countries who have less than 1,000 passengers to 'Other'.
- I did not use 'trip\_type' as all are 'RoundTrip' and I did not use 'route' as it has 799 unique values.

## Step-4: Exploratory Data Analysis (EDA)

Since, This is my streamlit project (project with proper GUI) so I have not explored such steps here to see complete understanding, data analysis, visualizations and experimentation on this dataset you can see my notebook on Kaggle.

Link: <https://www.kaggle.com/code/uditkishoregagnani/airline-booking-prediction>

Example: correlation matrix of input variables



Other Observations:

- Number of Complete Bookings that availed all added benefits: 1598
- Number of Complete Bookings that did not avail any added benefits: 1102
- Number of Incomplete Bookings that desired to avail all added benefits: 6952
- Number of Incomplete Bookings that desired to not avail any added benefits: 9155

## Step-5: Feature Engineering and Selection

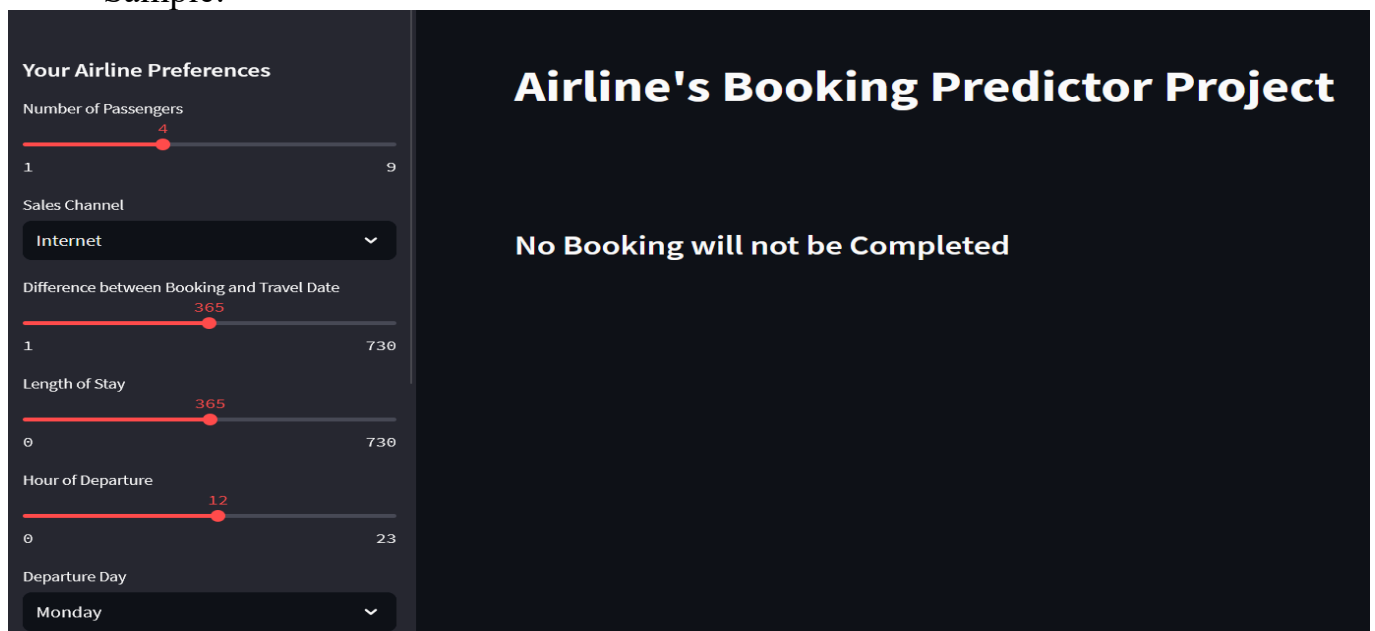
- I used Pipeline to do feature engineering tasks (in this case OneHotEncoding and PowerTransformation).
- Using this pipeline I fit\_transform my training data and transform my test data.
- Using PCA I extracted 13 best parameters for my classification model.

## Step-6: Model Training, Evaluation and Selection

- To solve this particular problem I decided to go with Stacking.
- Here I used 5 different algorithms (SGDClassifier, AdaBoostClassifier, SVC, LogisticRegression as 4 estimators and GradientBoostingClassifier as final estimator).
- I did try various other models (Everything is in my Kaggle Notebook).
- This model gave me minimum 84.60% accuracy everytime I used it and that too with 5 cross validations.

## Step-7: GUI for This Model

Sample:



Since, 'Sales Channel', 'Departure Day', 'Booking Origin', Extra Benefits ('Extra Baggage', 'Preferred Seats', 'In Flight Meals') are categorical variables so they are select boxes and rest all input variables are sliders.

I have improved this project a bit as the model is predicting 'booking\_completed' value as 1 or 0 and so I created if, else condition to show the booking will be completed or not.

## Step-8: Scope of Improvement / Future Enhancements

- In future a bigger project can be created for an Airline Company like (British Airways, Qatar Airways) in form of complete automated system at the back-end of the website where the user will enter all the details to see if desired flights are available or not and the employees of the company will get a response whether or not this customer will book the seat.
- Also this model can be deployed on Cloud.
- From technical perspective I can try various hyper-parameters on each algorithm and find the best model/ stack for this problem. But this is a very expensive approach in terms of time and computation.