Flight Satisfaction Prediction

FMT

Domain

Air travel, Transport, Consumer Satisfaction

Business Context

Customer satisfaction is an important factor for a business. Happy customers bring value and revenue. For the airline industry, cost and customer satisfaction plays a pivotal role. It’s important that customers have an excellent experience every time they travel.
The data we have at hand is of passengers and their feedback regarding their flight experience. Each row is one passenger. Apart from the feedback from the customers across various attributes(15 in total) like food, online_support, cleanliness etc, we have data about the customers' age, loyalty to the airline, gender and class.
The target column is a binary variable which tells us if the customer is satisfied or neutral/dissatisfied.
We will analyse various features to understand which features are contributing to the satisfaction of the customers.

Objective

The task at hand is to analyze reasons for customers' satisfaction or dissatisfaction.

And finally, we build a model to predict customer satisfaction using all or some of the data we have

Dataset description

Flight_data.csv : (90917, 9)

1. **ID** : Unique id for a passenger
2. **Gender:** Gender of the passenger
3. **CustomerType:** Type of customer
4. **Age**: Age of the passenger
5. **TypeTravel** : Type of travel (Personal /Business)
6. **Class** : Trvale class
7. **Flight_Distance** : Distance of the journey
8. **DepartureDelayin_Mins** : Delay in departure
9. **ArrivalDelayin_Mins**: Delay in arrival

Survey_data.csv : (90917, 16)

1. **Id** : unique id for a passenger
2. **Satisfaction:** satisfaction of the passenger
3. **Seat_comfort**: how is seat comfort
4. **Departure.Arrival.time_convenient** : feedback on time convenient
5. **Food_drink** : feedback on food and drink services
6. **Gate_location** : feedback on gate location
7. **Inflightwifi_service**: feedback on inflight internet service
8. **Inflight_entertainment** : feedback on inflight internet service
9. **Online_support** : feedback on online support from the airline
10. **Ease_of_Onlinebooking**: feedback on booking facilities
11. **Onboard_service** : feedback on onboarding services
12. **Leg_room_service**: how is legroom in the flight
13. **Baggage_handling**: feedback on handling baggage
14. **Checkin_service** : feedback on check-in services
15. **Cleanliness**: feedback on cleanliness and hygiene
16. **Online_boarding**: feedback on online boarding services

Steps

1. Import the necessary liberraries and read the provided CSVs as dataframe and perform the below steps
    - a. Check a few observations and shape of both the dataframes
    - b. Join the two dataframes using the 'id' column as the primary key [Hint - you can use join, merge, or concat function]
    - c. Check for missing values. Impute the missing values if there is any
    - d. Perform bivariate analysis and check for the correlation
    - e. Encode all the categorical “Feedback columns” like Seat_comfort', 'Departure.Arrival.time_convenient', 'Food_drink', etc
    - f. Display countplot of feedback attribute with respect to Customer Satisfaction and mention your insight
    - g. Print the average feedback score

2. Print the number of people who are more than just satisfied with the "Inflight_entertainment" and yet were dissatisfied overall
3. Print the number of people who are more than just satisfied with the "Inflight_entertainment" and were satisfied overall
4. Create a new column which is the mean of 'Ease_of_Onlinebooking', 'Online_boarding', 'Online_support' and name it "avg_feedback_of_online_services"

5. Segregate the dependent column ("Satisfaction") from the data frame. And split the dataset into training and testing set (70:30 split)
6. Build a decision tree classifier and print confusion matrix for the test data

7. Cross validation:
    - a. Perform cross validation on a decision tree algorithm and print the score.
    - b. Perform cross validation on a random algorithm and print the score.

8. Apply Grid Search CV to get the best hyper parameters for the random forest model
9. Build a random forest model using the above hyper parameters and check the accuracy and confusion matrix
10. Print feature importance of the Random Forest model
11. Build a Pipeline to automate and simplify the above model building process

Further Explore:

1. Cluster different segments of customers to get more insights about their behaviours.
2. Create new features and select the best features to improve your model further.
3. Check for outliers and impute them
4. Apply PCA to reduce the complexity of the model

Learning Outcomes

- Decision Tree
- Random Forest
- Cross validation
- Grid Search
- Exploratory Data Analysis
- PCA
- Pipeline