Airline Passenger Satisfaction

By
Vishal Wadagavi
26th Nov 2021

Airline Passenger Satisfaction

Problem statement:

Context
This dataset contains an airline passenger satisfaction survey. What factors are highly correlated to a satisfied (or dissatisfied) passenger?
Can you predict passenger satisfaction?

Content

- Gender: Gender of the passengers (Female, Male)
- Customer Type: The customer type (Loyal customer, disloyal customer)
- Age: The actual age of the passengers
- Type of Travel: Purpose of the flight of the passengers (Personal Travel, Business Travel)
- Class: Travel class in the plane of the passengers (Business, Eco, Eco Plus)
- Flight distance: The flight distance of this journey
- Inflight wifi service: Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)
- Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient
- Ease of Online booking: Satisfaction level of online booking
- Gate location: Satisfaction level of Gate location
- Food and drink: Satisfaction level of Food and drink
- Online boarding: Satisfaction level of online boarding
- Seat comfort: Satisfaction level of Seat comfort
- Inflight entertainment: Satisfaction level of inflight entertainment

Airline Passenger Satisfaction

- On-board service: Satisfaction level of On-board service
- Leg room service: Satisfaction level of Leg room service
- Baggage handling: Satisfaction level of baggage handling
- Check-in service: Satisfaction level of Check-in service
- Inflight service: Satisfaction level of inflight service
- Cleanliness: Satisfaction level of Cleanliness
- Departure Delay in Minutes: Minutes delayed when departure
- Arrival Delay in Minutes: Minutes delayed when Arrival
- Satisfaction: Airline satisfaction level(Satisfaction, neutral or dissatisfaction)

Airline Passenger Satisfaction

Approach:

1. Data Gathering:
    - Dataset is fetched from a data source called Kaggle.
    - Data is already in the form of csv format, so it can be directly read using pandas.
2. Data Cleaning:
    - There are missing values which needs replacement or features itself needs to be dropped.
    - There are no duplicates in data set so there is no need of dropping data points.
    - Couple of features have high variation between 99%le and max values. Outliers needs to be treated.
    - Defining target features and converting it to binary values (0/1).
    - Scaling numerical features with Standardscaler()
    - Encoding Categorical features using labelencoder or onehotencoder based on type of feature.
    - Numerical features which have value_counts less than 10 needs to be treated as categorical features.
    - Feature dropping if there is high correlation coefficients between features.

Airline Passenger Satisfaction

3. Data Exploration:
    - Exploring 5 points summary of all the features to know how data is varying from regular interval of percentiles.
    - Heat map of correlation between features are plotted using which dropping of features is decided.
    - Distribution plots and boxplots are plotted in order to explore the outliers.
    - Bi-variate analysis is done to observe how categorical feature values are influencing to the target variable.
4. Model Building:
    - Pre processed numerical and categorical features are concatenated to form final data frame which will be used for model building.
    - Final data frame is split into independent and target features.
    - And also data is split into train and test data.
    - Train data set is used for the training the model and test data is used for validating it.
    - LogisticRegression model needs to be built for classification and also SVM algorithm needs to be used in which parameters are tuned to define the kernel which is more accurate in classification. .

Airline Passenger Satisfaction

5. Evaluation:
    - Built models are evaluated against test data.
    - Model comparison is done using confusion matrix and Classification report.
    - In this problem, it is important to identify non-satisfied customers.
    - In order to reduce false classification of non-satisfied customers, precision needs to be increased.