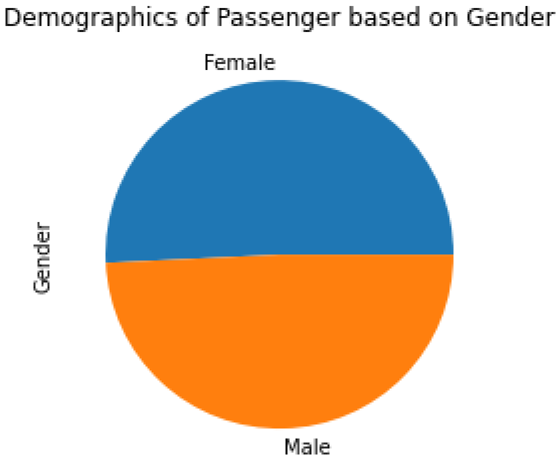
Summary for Report on Airline Passenger Satisfaction

Prepared by Aashish Panta

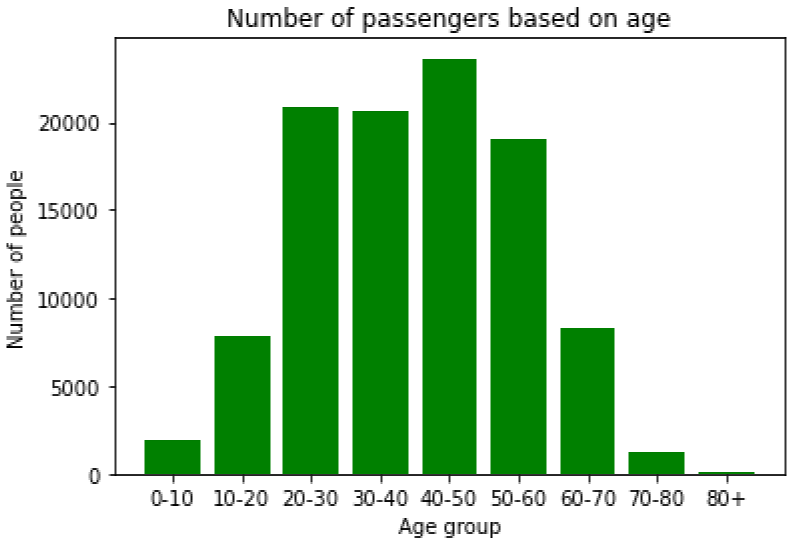
Airline Passenger Dataset:

The dataset is divided into two sets of training and testing data. Training size is of almost 104,000 instances whereas testing data has around 26,000 instances. These datasets consist of several important features like gender of the passenger, age, type of travel, travel class, satisfaction level of booking, check-in service, seat comfortness, cleanliness, inflight wifi service and so on, and based on these factors, passengers determined whether they were satisfied with the airline or not. Most of the features have a score of 0-5 where 5 is excellent, 1 is bad and 0 is missing data. We can actually work on this dataset as a classification problem and make a machine learning model to predict whether a passenger will classify the airline as ‘satisfied’ or ‘neutral or dissatisfied’ based on the opinion of the airline features.

Gender: This dataset consists of surveys from 52.5k female and 51k male population. Since the number of male and female is balanced in this case, it is safe to assume that there is no sample bias based on gender.



Age: The given dataset has a survey of people from all age groups. People belonging to age group 40-50 have the highest number of travels. The median age in the given dataset is 40 years whereas the median age of the US population is 38.4 years.



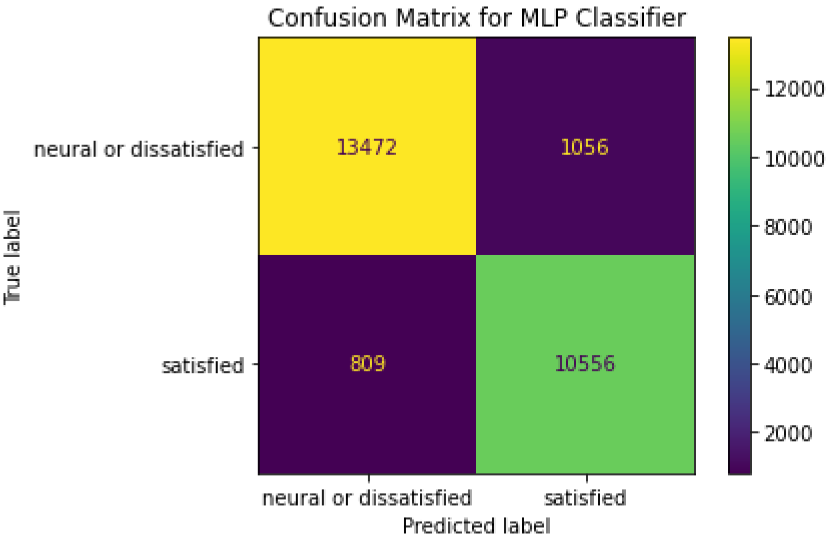
This also shows that our data sample is a very close representation to the overall population nationwide and is unbiased.

Data Preprocessing:

Some features of the data like gender, class, type of travel and customer type were in string format. These strings were temporarily changed to integer format. For example, in the Gender column, 'Male' was changed to 0 and 'Female' to 1.

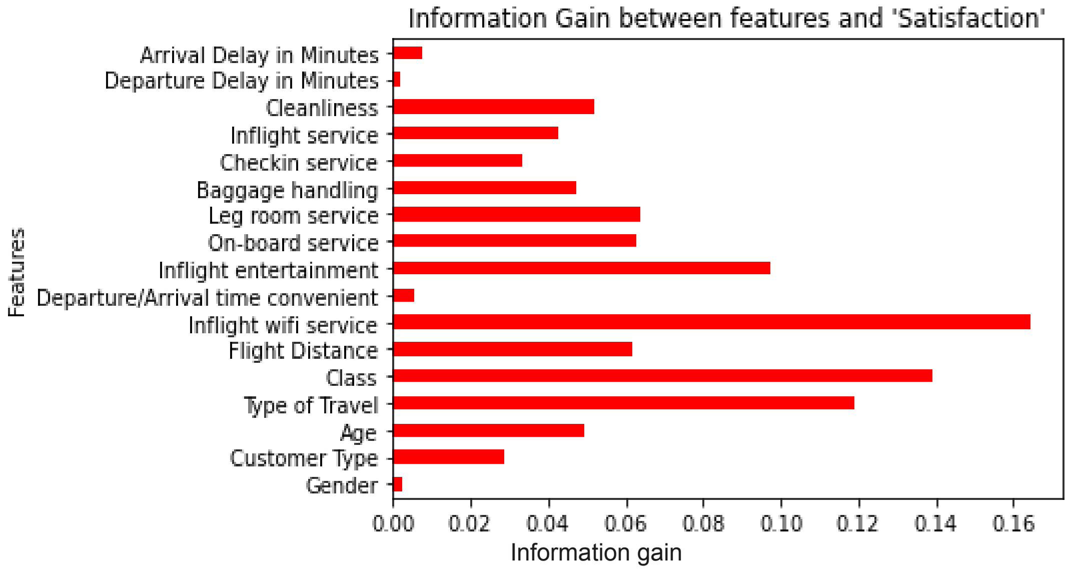
Classification:

This modified dataset was then fed into different classification models. K-Nearest Neighbor, Multi-layer Perceptron Classifier, GaussianNB and Logistics Regression were used for classification purposes. Among



all these, we achieved the highest accuracy of 93.18% using MLP classifier setting the parameters hidden_layer_sizes=(10,10,10) and max_iter=500. We achieved the second highest accuracy of 84.33% using GaussianNB. KNN was the worst performer among these with an accuracy of 74%. The n_neighbors parameter was set to 7 in this case.

Findings: We used feature_selection library to calculate the information gain among the different features and satisfaction, and found out that ‘Inflight Wifi Service’ was most closely related to the satisfaction of the airlines. The other most important features were ‘Class’, ‘Type of Travel’ and ‘Inflight Entertainment’ respectively. Also, gender and delays in arrival and departure were found to have the least effect in how passengers vote overall satisfaction of the airlines.



Challenges: Although the dataset was not very large in size, running the model using several permutations of parameters was pretty time-consuming. It took more than 5 hours to run GridSearchCV with 8 possible models.