Intro to Data Science Capstone Report

Team Boeing 737

Amanpreet Singh Saimbhi
as15798

Pratap Narra
pn2158

1. Introduction

Running an airline company is extremely challenging due to stringent government approval processes and involves huge initial investments. Keeping in mind the profound amount of work required to start an airline, it is better to do some research prior on the customer satisfaction levels on various parameters, this not only helps new players but also the existing ones to improve their customers' flying experience. Therefore, using the concepts learned in this course we have decided to explore the Airline Passenger Satisfaction Dataset which has around 104k customer satisfaction ratings on various aspects of the flight, like food, seat comfort, legroom, onboard wifi, etc.

2. Dataset

The dataset has 104k customers' satisfaction ratings on different aspects of the flight journey, this is obtained from kaggle[dataset]. The dataset has 103904 rows and 25 columns. The dataset has 310 NaN values in total, but there are no NaN values in the columns that we worked on for most of the project. The missing values in the column Delay in Arrival Time have been filled with the mean value of the column. Fortunately, the dataset does not have any other column having missing values. Further exploration for dimensionality reduction has been done in the Classification section. Each row of the dataset represents a survey response from a customer and their additional information.

The first column has just the row number, further, the columns are:

1) **id**: id of the customer
2) **Gender**: Gender of the passengers (Female, Male)
3) **Customer Type**: The customer type (Loyal customer, disloyal customer)

4) **Age**: The actual age of the passengers
5) **Type of Travel**: Purpose of the flight of the passengers (Personal Travel, Business Travel)
6) **Class**: Travel class in the plane of the passengers (Business, Eco, Eco Plus)
7) **Flight distance**: The flight distance of this journey
8) **Inflight wifi service**: Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)
9) **Departure/Arrival time convenient:** Satisfaction level of Departure/Arrival time convenient
10) **Ease of Online booking:** Satisfaction level of online booking
11) **Gate location**: Satisfaction level of Gate location
12) **Food and drink**: Satisfaction level of Food and drink
13) **Online boarding**: Satisfaction level of online boarding
14) **Seat comfort**: Satisfaction level of Seat comfort
15) **Inflight entertainment:** Satisfaction level of inflight entertainment
16) **On-board service**: Satisfaction level of On-board service
17) **Leg room service**: Satisfaction level of Leg room service
18) **Baggage handling**: Satisfaction level of baggage handling
19) **Check-in service**: Satisfaction level of Check-in service
20) **Inflight service**: Satisfaction level of inflight service
21) **Cleanliness**: Satisfaction level of Cleanliness
22) **Departure Delay in Minutes**: Minutes delayed when departure
23) **Arrival Delay in Minutes**: Minutes delayed when Arrival
24) **Satisfaction**: Airline satisfaction level(Satisfaction, neutral or dissatisfaction)

3. Hypothesis Tests

In this section, we explored different hypothesis questions that might be helpful for an airline company and answered them. For all these questions we set the alpha level to 0.05. Since the sample size is quite large we have a large power.

Q1. Is cleanliness satisfaction gendered, i.e do female and male passengers find cleanliness on the flight different?

We aim to find out if male and female passengers view the cleanliness of flight differently, since we are comparing the cleanliness satisfaction ratings, the **u-test** is an appropriate test.

We start by assuming that the null hypothesis “Male and female passengers have similar satisfactory ratings”, and perform the u-test. Firstly we made two lists one with cleanliness satisfaction ratings of males and the other with that of females. This part of the dataset doesn’t have any NaNs.

The resultant p-value is **0.02,** which is less than the alpha, so the probability of the data given chance alone is very less so we reject the null hypothesis**.**

Q2. Do younger and older passengers rate seat comfort differently?

As again we are comparing the seat comfort satisfaction ratings, we used the **u-test**. Firstly we find the median age of passengers and split the data frame into two halves based on the median. The median age of passengers is 40. There are no Nans in this part of the data. Then two lists are made which have the seat comfort ratings of younger and older passengers. Figure 1 shows the age distribution in the dataset.

We assume the null hypothesis “Younger and older passengers rate the seat comfort similarly”. After performing the u-test, the p-value comes out to be ~0, so as this is less than alpha, we reject the null hypothesis.

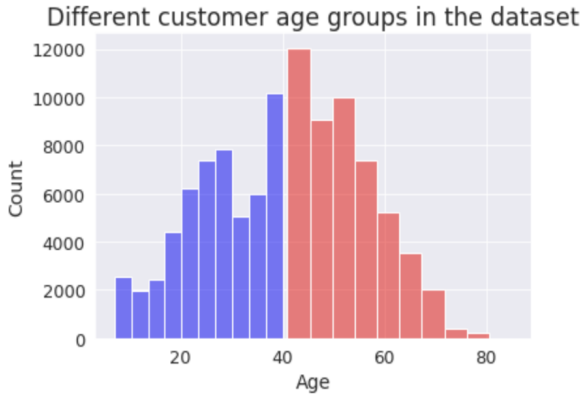

Figure 1

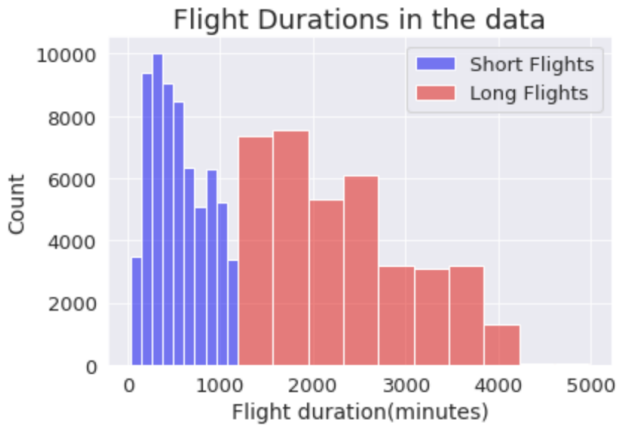

Figure 2

Q3. Do passengers traveling on long flights vs ones traveling on short flights(duration) rate the food differently?

We start by finding the mean flight distance and split the data frame into two halves, the mean is 1189 minutes. We assume the null hypothesis of both long and short-rate food satisfaction similarly. In order to compare the distributions instead of just medians, we have used the **ks-test**. Figure 2 shows the flight duration distribution in the dataset.

The p-value turned out to be 2.590992504354077e-53, so we reject the null hypothesis.

3. Prediction

3.1 Question

Can we predict the delay in arrival time using the delay in departure time while controlling for confounds?

3.2 Approach

The goal is to utilize the correlation matrix for feature selection and verification. Subsequently, we will use the data for the delay in arrival time and the delay in departure time to train a regression model and check if the model is able to make predictions reasonably.

3.3 Analysis and Results

We start by loading the entire dataset in a data frame and then we analyze the properties of the data using the Pandas in-built functions. We notice that an unnecessary column has been created in the data frame so we drop that, and additionally the column ‘id’ is of no use in our analysis so we drop that as well. We note that there are “310” missing values in the column “Arrival delay in minutes.”, we replace the missing values with the mean value of the entire column. Next, we have

columns having values that are categorical strings. We map all the values in columns - 'Customer Type', 'Type of Travel', ‘Class’, 'satisfaction', and ‘Gender’ to their respective numeric values. For example: “'Business travel': 0, 'Personal Travel': 1”. With this step, our preprocessing of the data frame has been completed.

As the first step, we implement and analyze the correlation matrix to understand the correlation between the data. It can be observed from the matrix that both the columns being discussed have a high correlation and more importantly they have a really weak correlation with all the remaining columns. Therefore, it makes sense for us to implement a linear regression model as we do not have to control for any “known” confounds.

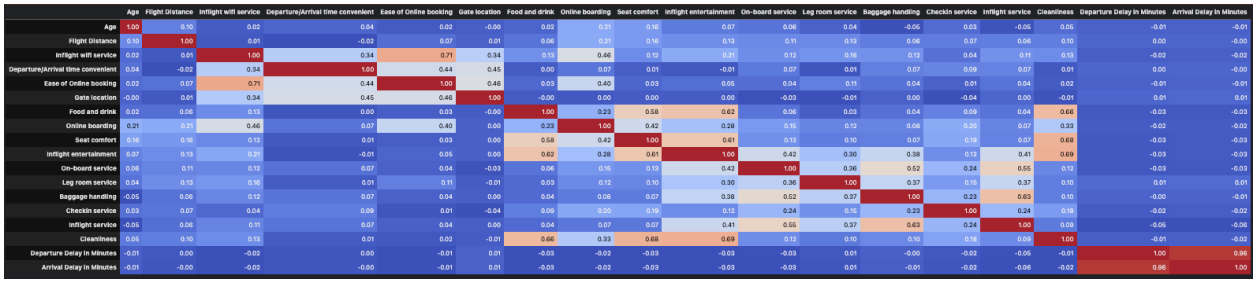

Figure 3

As the next step, we define the variable X which is our predictor, with the values of the column “delay in departure time” and the variable Y which is the predicted variable, with the values of the column “delay in arrival time”. We then use X and Y to train the linear regression model. Regression models with Lasso and Ridge regression have also been trained, and the value of epsilon is selected optimally using grid search. It can be seen from table 1 that they perform similarly, but regularized models, in general, will be able to generalize better to the unseen data. With some exploration, we were able to achieve an R2 score of **0.936** on the test set. This concludes that we are able to predict the delay in arrival time using the delay in departure time, moreover, it can be visualized with the best-fit plot in figure 4.

Table 1

Classifier	**R2**
Linear Regression	0.9361956634850266
LR with Lasso	0.9361956625449127
LR with Ridge	0.9362394366935345

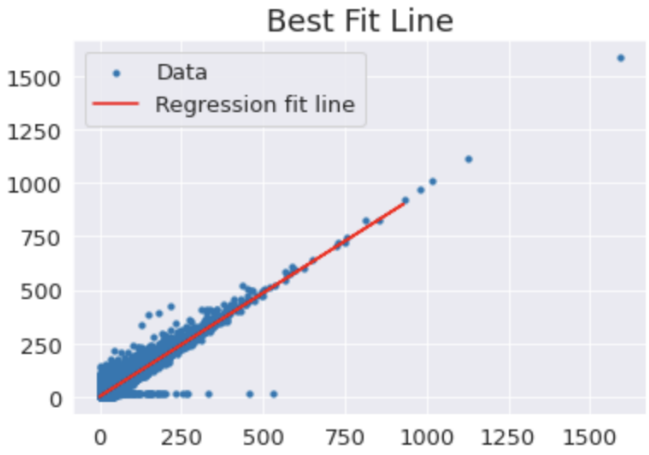

Figure 4

4. Classification

4.1 Question

Will the passenger be satisfied or dissatisfied given all the information we have in the dataset?

4.2 Approach

The goal is to get the most suitable X training data and Y corresponding labels from the data and train various classification models. Using several metrics and visualizations, we will be selecting the model that gives the best results and answers our posed question.

4.3 Analysis and Results

The preprocessing of data has been performed in a similar manner to the preprocessing done for the “prediction” section. In the subsequent step, we keep the values of the column ‘Age’ as the Y (labels), and the remaining columns from the data frame as the X. We then split the data with an 80/20 train/test split. Now, to make sure that the data is not far spread out, we used the “fit_transform()” method to scale the data and also learn the mean and variance of the training set.

At this point, we were curious to see if there is any separation possible between the “satisfied” and “dissatisfied” data points. To visualize that, PCA has been used for both 2 components and 3 components. It can be seen from figures 5, and 6 that separation is visible, and it becomes more distinct with the addition of components.

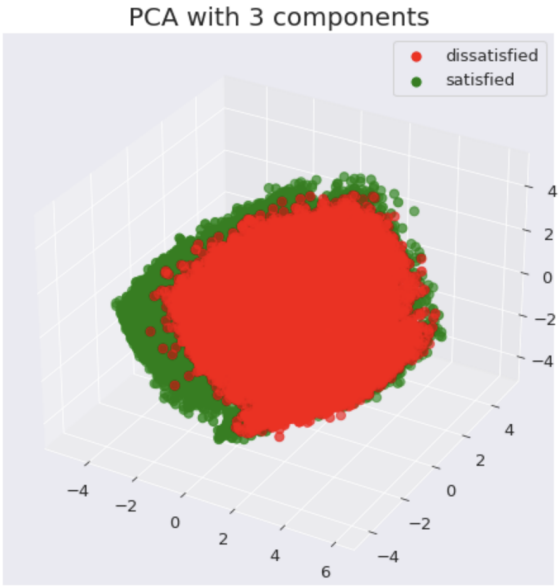

Figure 5

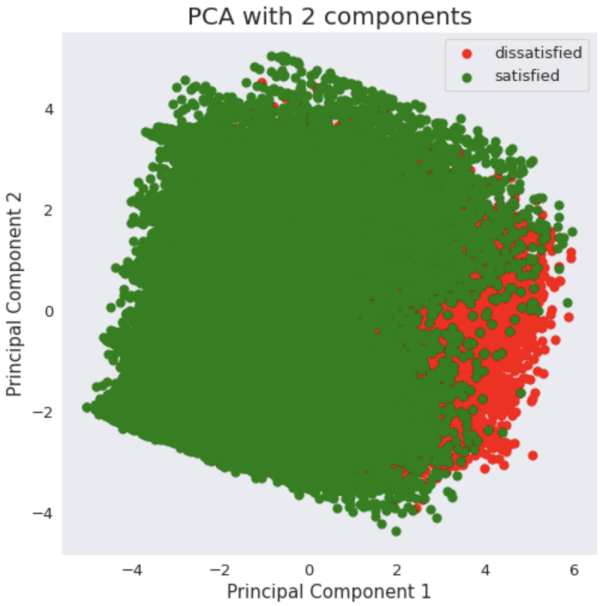

Figure 6

We want to implement a clustering algorithm before diving into classification to check if there are any clusters forming naturally. Therefore, with the help of PCA, we explain 80% of the variance. We implemented a Scree plot to see how many principal components are needed. As it can be observed the line is bent at 1, thus, we need at least 2 principal components.

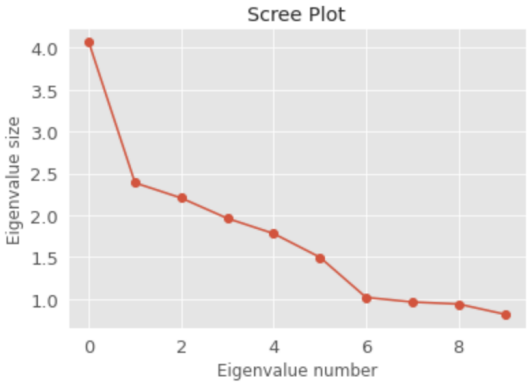

Figure 7

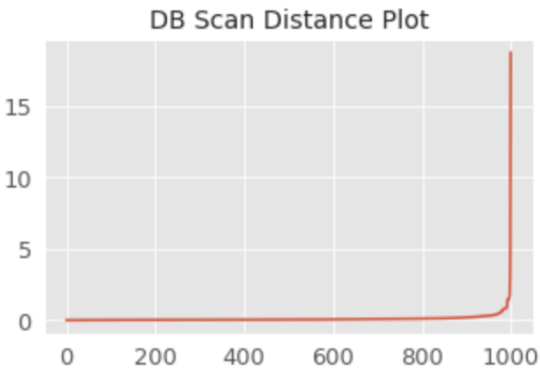

Figure 8

We select DB Scan as our clustering algorithm as it does a great job of filtering out noise and incorporates density while forming clusters. Now, we observe that on the entire dataset it is becoming extremely challenging to visualize the information on the 2-d axis, and also to form clusters. For the sake of visualization, we reduce the sample size and run the clustering algorithm on it. With the help of the distance plot, we select the optimal epsilon value to be 1 and keep the minimum samples in a cluster to be 4. There are 2 clusters have been formed as seen in figure 9.

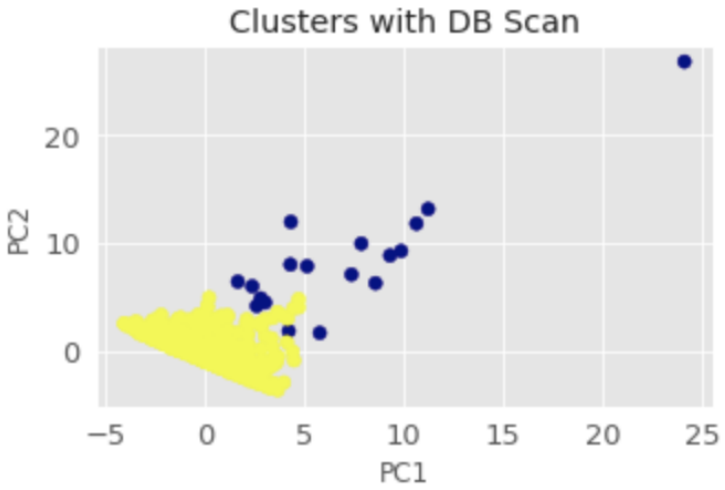

Figure 9

We finally move on to train the classification models on all the columns, as we observed from the dimensionality reduction exploration that higher dimensions are needed for better separation. We will be tracking various metrics and visualizing multiple plots to compare the models and verify if they are able to achieve the desired results. Hyperparameter tuning has also been performed to achieve the best performance with respect to each model.

Table 2

Classifier	**R2**	**AUC**
Logistic Regression	0.87	0.92
Random Forest	0.93	0.97
ADA Boost	0.93	0.97

It can be seen from table 2 that the task of classification can be performed with reliable accuracy. We can answer the question and establish that for new passenger

flight information, based on their experience, we will be able to tell whether they will be satisfied or not. We select ADA Boost as the best classifier that is giving AUC of **0.97** and an R2 of **0.93**. ROC plots can also be seen for various classifiers in figure 10..

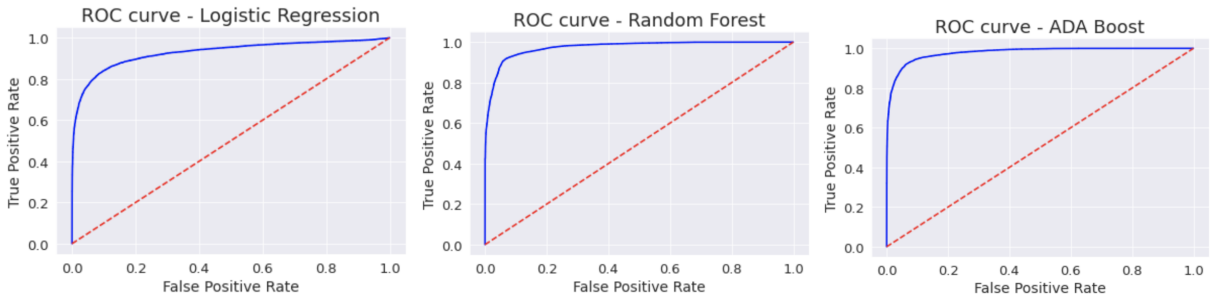

Figure 10

5. Classification

5.1 Takeaways

This dataset can help us answer a plethora of questions and we have answered some of them in our analysis.

- Cleanliness satisfaction rate is gendered.
- The younger and older customers rate the seat satisfaction rate differently.
- The flight duration has an effect on the food and drink satisfaction rate.
- Delay in arrival time can be predicted using the delay in departure time.
- It can be classified whether a passenger will be satisfied or dissatisfied with the overall airline experience with the other information available in the dataset.
- Also, we would like to mention one thing that did not work as well. The age of the passenger can not be predicted reliably given other information.

5.2 Limitations

The dataset has relatively few columns, and there could be multiple other factors responsible for the satisfaction of a passenger. We have relied on correlation for regression analysis, but as we know that correlation is not causation, there could be “unknown” confounding variables.

5.3 Ideal Dataset

In an ideal Airline Passenger Satisfaction dataset, there should be additional columns. For example- the duration of the flight. Also, the average rating criteria for each column should be present, because then only we will be able to truly assess, on average if a person gives a 4.5/5 rating, then for them to give a 3 is more significant than someone whose average rating is 3.

5.4 Extra Exploration

We aim to explore if whether the age of the customer affects the customer type, so we start by making two data frames one for loyal and the other for disloyal customers. And we also assume the null hypothesis “Ages of both customer types are similar”. Since here we are comparing the ages of the customer, it makes sense to reduce the data to its mean, so we used the t-test for this part. Also, there are no NaN values present in this part of the dataset. After the t-test, the p-value comes out to be ~0, so we reject the null hypothesis. Therefore the alternative hypothesis is “Ages of both customer types are different”.