Airline Passenger Prediction

Denny Alvito Ginting A.K.A dendeni

TABLE OF CONTENT

BUSINESS BACKGROUND

BACKGROUND

I am a Data Scientist at Japan Airline Department

Japan Airline works at giving flight transportation for our customer satisfaction. In this task, I am going to build a Machine Learning Model to predict whether a passenger satisfied or not from the airline services

OBJECTIVES

This project has an objective to predict whether a passenger satisfied or not while using the airline services. This feedback could be useful for further flight on the Airline

OUTPUT

Satisfaction Prediction

From its given Feature

PROJECT LIMITATION

This project only works on limited features such as "Inflight Wifi Services", Airline that support "Eco Plus" Travel, since not all Airline have these features

ANALYTIC APPROACH

Machine Learning

Supervised Learning. Binary Classification to determine whether a passenger satisfied or not

Performance Measures

Classification Report by prioritizing f1-score

DATA UNDERSTANDING AND EXPLORATORY DATA ANALYSIS

DATA COLLECTION

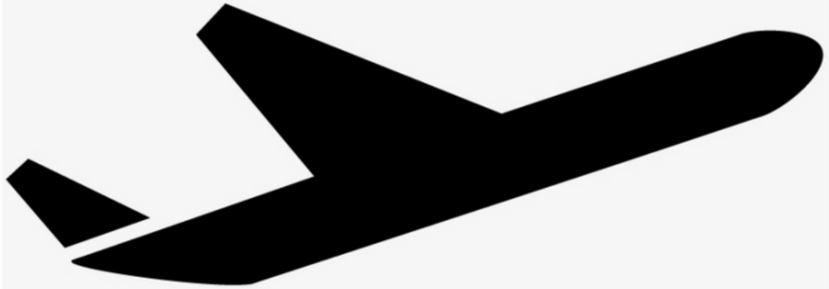
Kaggle

The datasets is provided with
129880 rows and 25 columns

COLUMN VARIABLES

- Unnamed: 0
- id (Passenger's ID)
- Gender (Male or Female)
- Customer Type (Loyal Customer or disloyal Customer)
- Age
- Type of Travel (Business or Personal Travel)
- Class (Business, Eco, or Eco Plus)
- Flight Distance
- Inflight Wifi Service
- Departure/Arrival Time Convenient
- Ease of Online Booking
- Gate Location
- Food and Drink
- Online Boarding

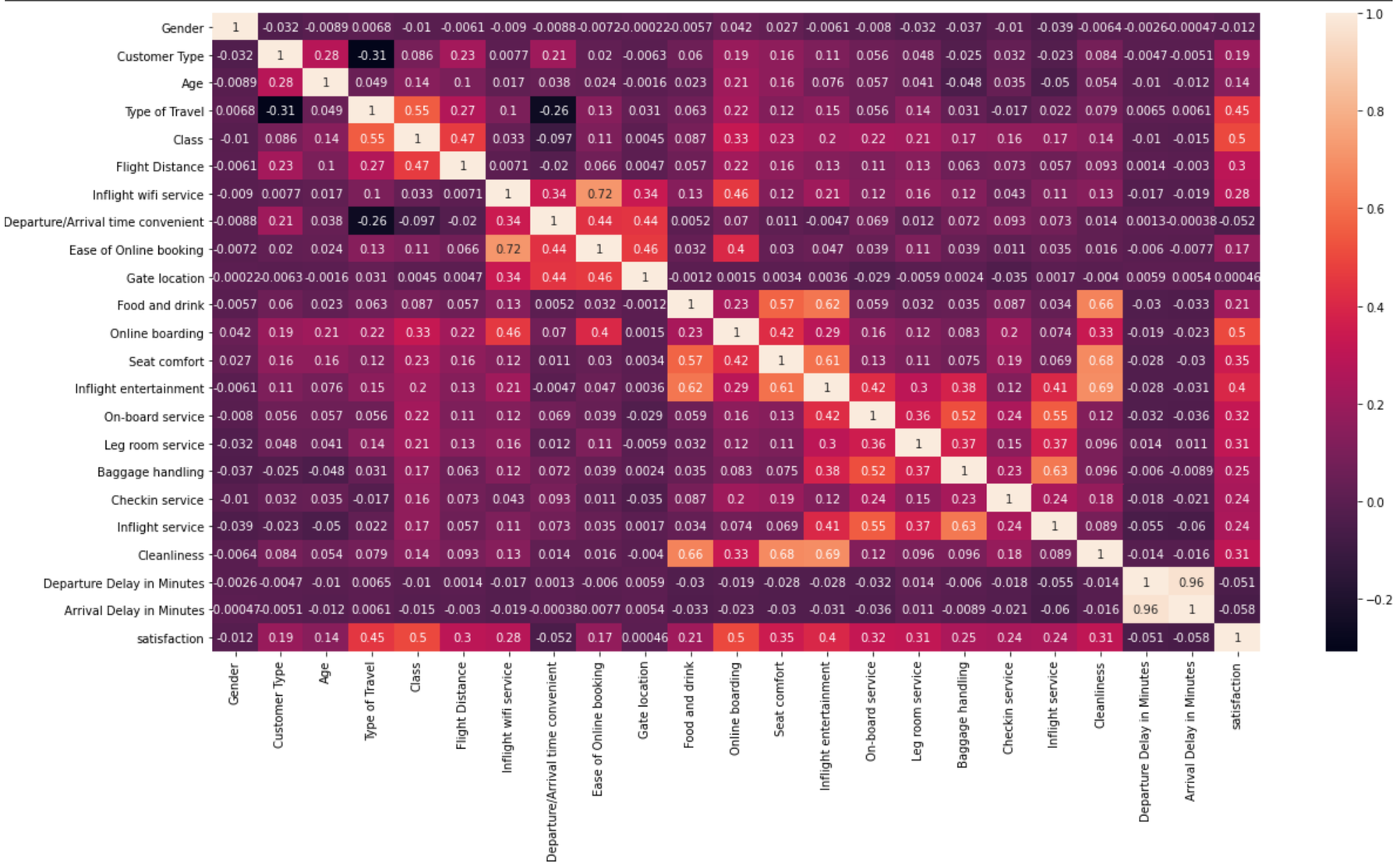
- Seat Comfort
- Inflight Entertainment
- On-board Service
- Leg room Service
- Baggage Handling
- Checkin Service
- Inflight Service
- Cleanliness
- Departure Delay in Minutes
- Arrival Delay in Minutes
- Satisfaction (Satisfied or Neutral/Dissatisfied)



DATASETS DIVISION

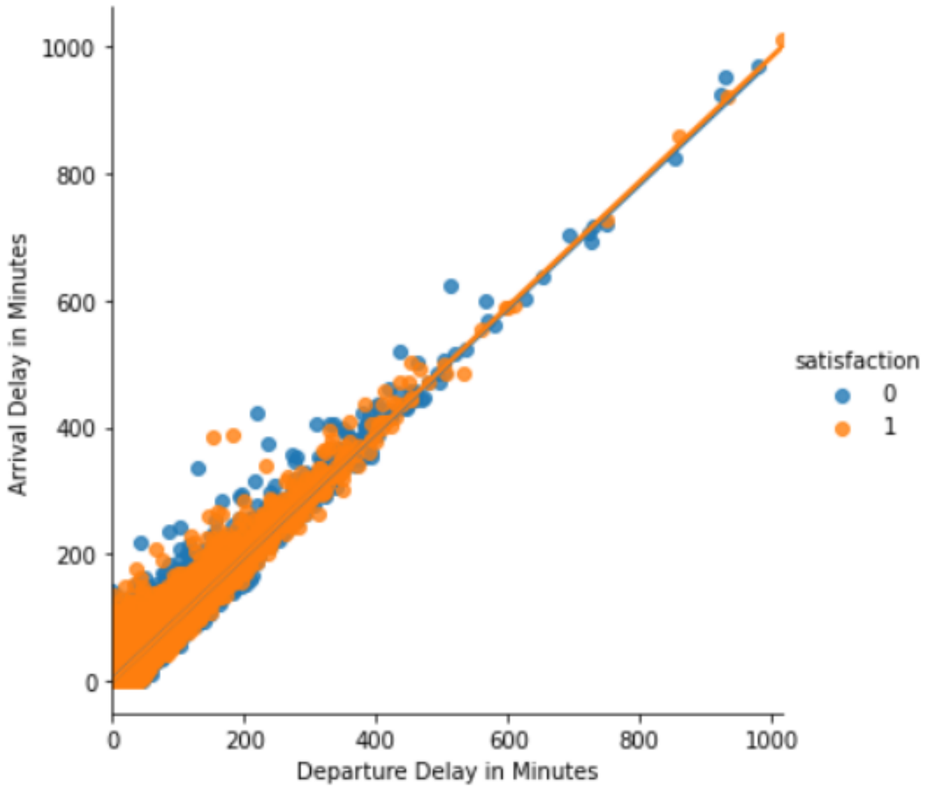
80% 20%

The datasets divided with 80%
as development and 20% as
future predict data



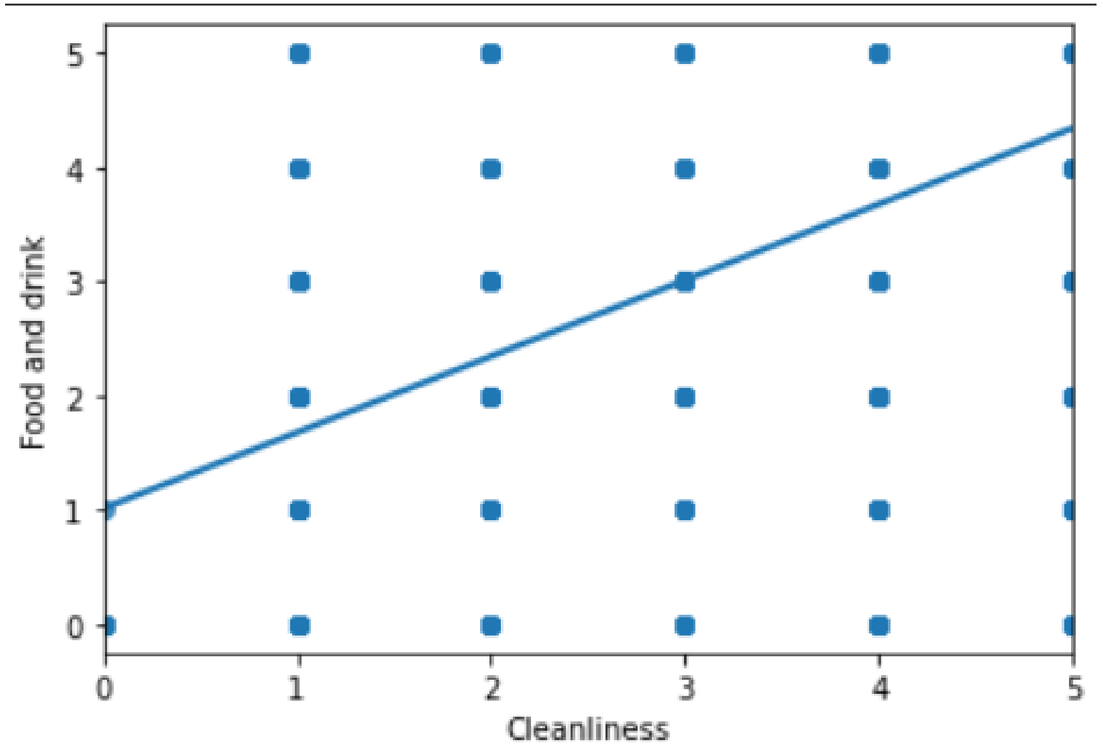
- From the heatmap above, we can see a lot of features that correlate with target variable (satisfaction), for example, Online Boarding, Type of Travel, Class, Inflight Entertainment, Seat Comfort.

DEPARTURE AND ARRIVAL DELAY ON SATISFACTION



- From the plot beside, we can see that Departure Delay in Minutes and Arrival Delay in Minutes has a strong correlations. This does make sense since in real case, people who tends late on depature most likely also late on arriving.
- The plot also tell us about the satisfaction, both satisfaction are in the same range, from this plot, we hardly see if these features have strong correlation with satisfaction because people are satisfied also neutral/dissatisfied by looking at this plot. This plot also proves the correlation matrix which showed that both of these features have low correlation with target variable

CLEANLINESS WITH FOOD AND DRINK



- From the plot above, it tell us that Cleanliness have a positive correlation with Food and drink. This does make sense since people want a food and drink while flying in airline but also want to keep cleanliness so that they can feel comfortable while eating and drinking
- For some reason, the plot looks like that because of these features actually a rating system (0 - 5), so that's why it looks like having gap, but overall, the plot shows some proof to our correlation matrix

AUDIENS

Profesional muda yang memiliki kelebihan pendapatan, dan pensiunan yang mencari hobi.

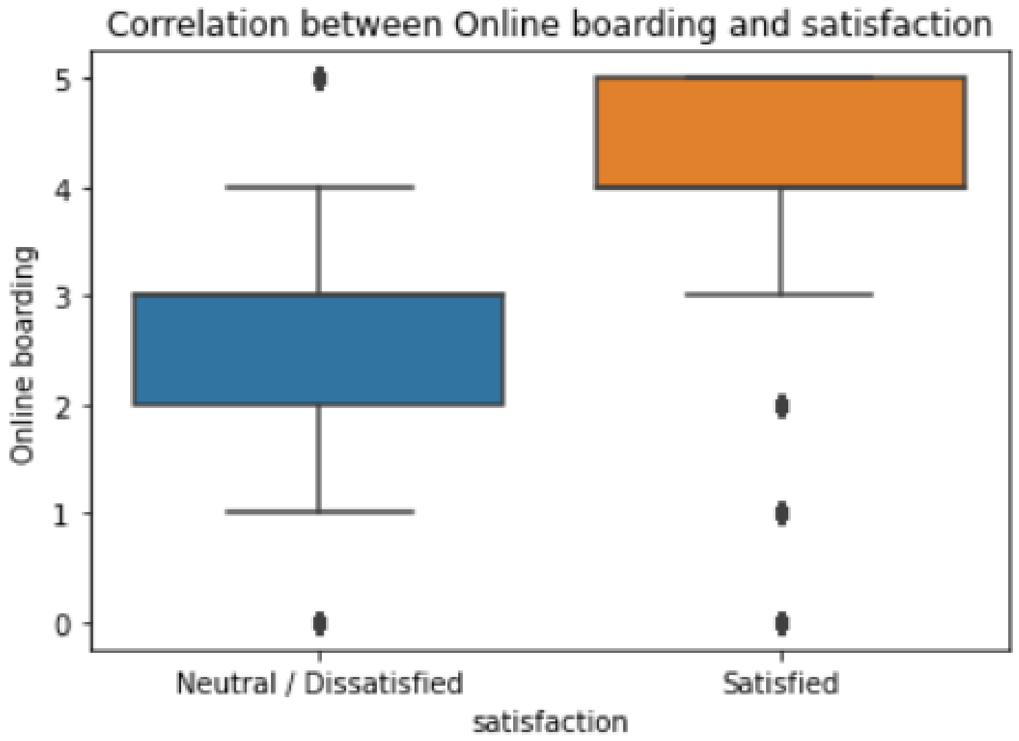
Profesional Muda

QR8 akan memberi mereka kesempatan untuk menemukan seniman baru, membangun koneksi, dan belajar dari mereka.

Pensiunan

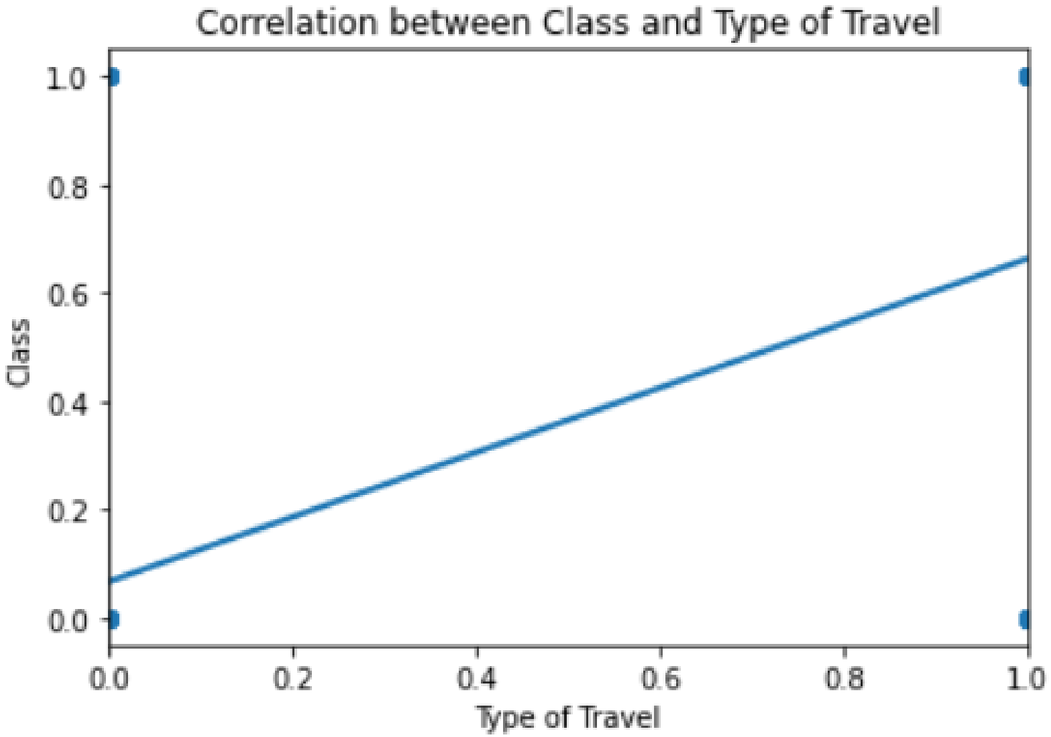
Pensiunan sering mencari hobi baru. QR8 akan memberi mereka sarana untuk belajar di rumah.

SATISFACTION ON ONLINE BOARDING



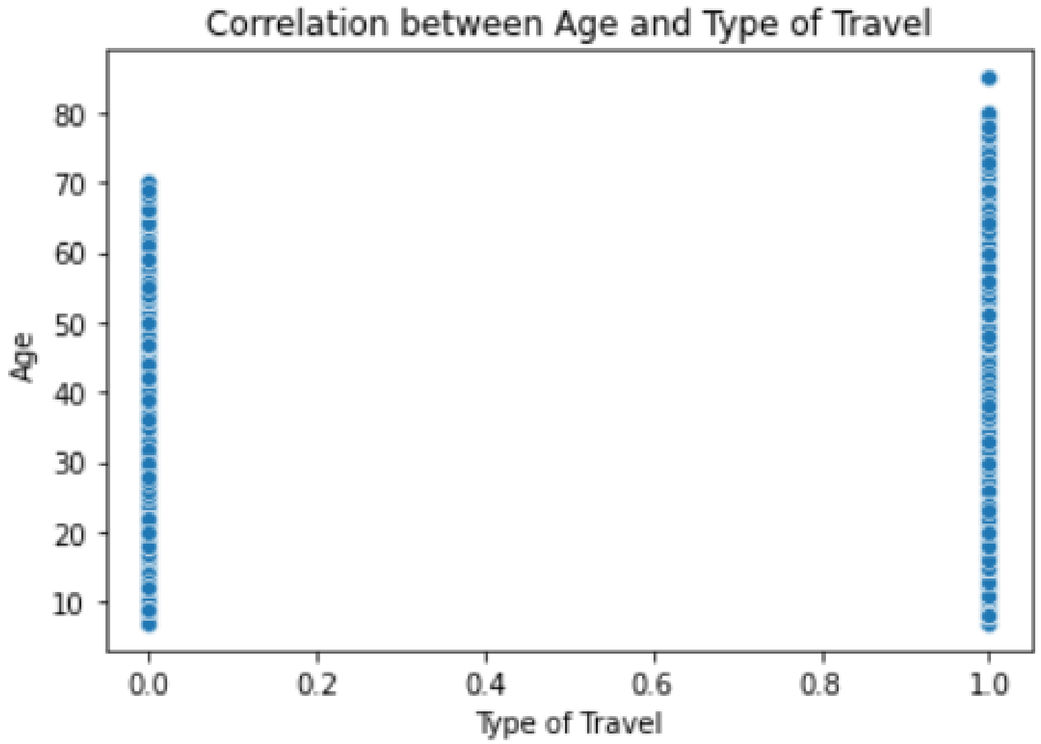
- People who are having a good online boarding experience tends to be more satisfied than people who have decent online boarding experience. The plot also related to real world case since people will be more satisfied if they get better experience in online boarding.

CLASS AND TYPE OF TRAVEL CORRELATION



- Type of Travel and Class have a good positive correlation based on the plot, we hardly see the scatterplot since these datas are categorical variable. However, this plot is also related to real life case, people that have Business class more likely have a type of travel: Business Travel .

AGE AND TYPE OF TRAVEL CORRELATION

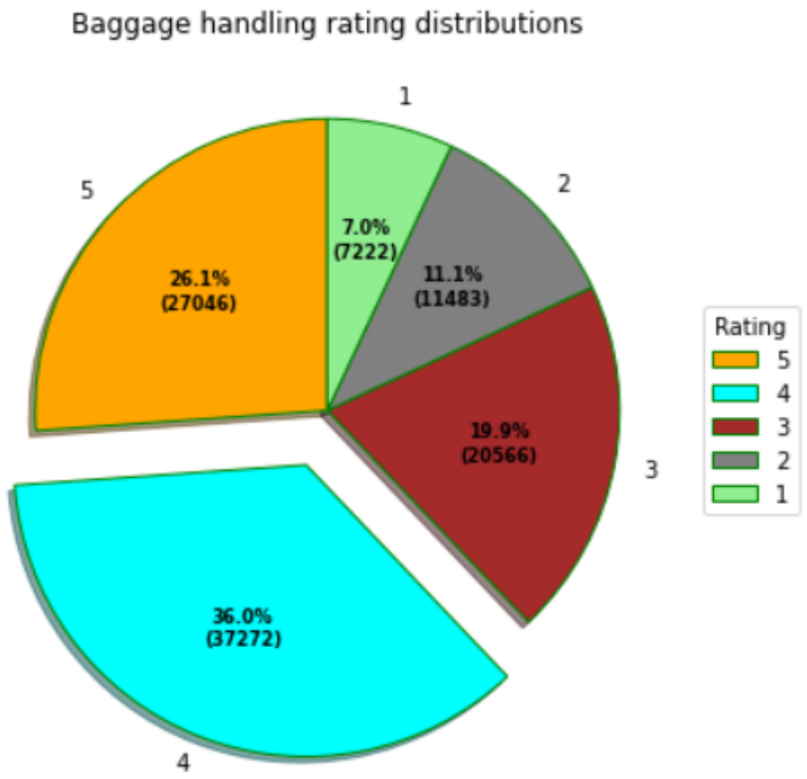


- From the graph, it seems like both type of travel (Economy and Business) have the same range age of people. However, people that have age more than 70 tends to have a Business Travel, this is possible since people that have that range of age need more comfortable place which we normally found in Business Travel.

EASE ONLINE BOARDING AND WIFI SERVICE CORRELATION

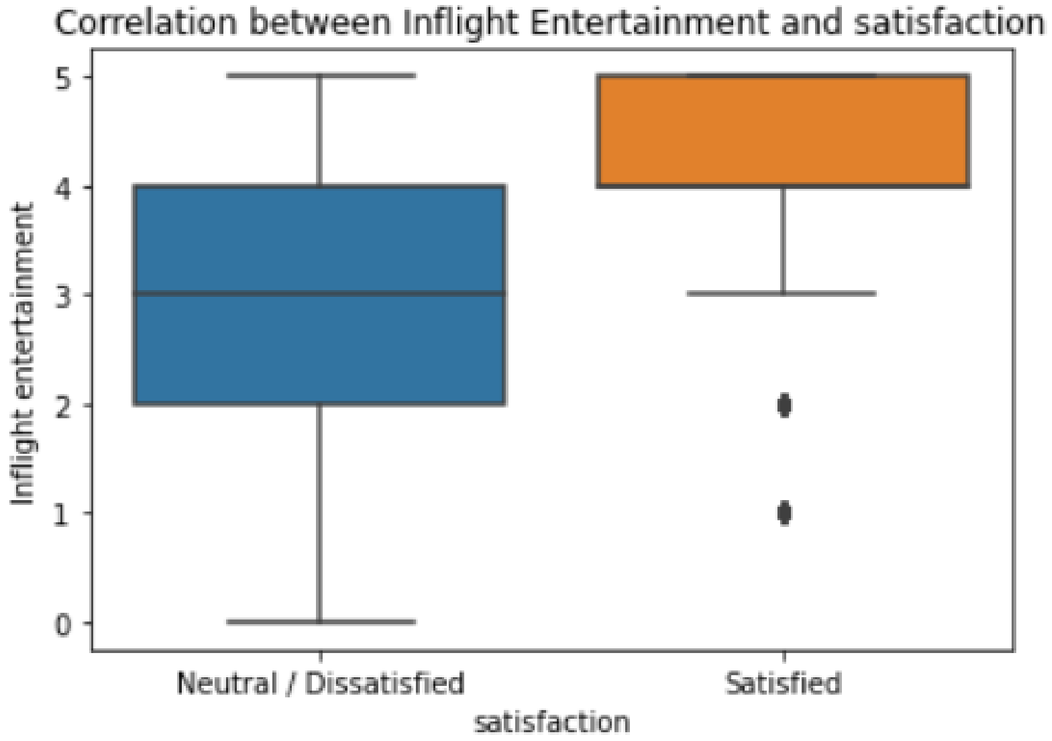
- The plot shows that the increasing Inflight wifi service affects the increasing Ease of Online Booking. This is make sense because people that have good experience on wifi service will getting more ease on booking the schedule online. That is also why the satisfaction also increasing with the increasing of the wifi service.

BAGGAGE HANDLING RATING DISTRIBUTIONS



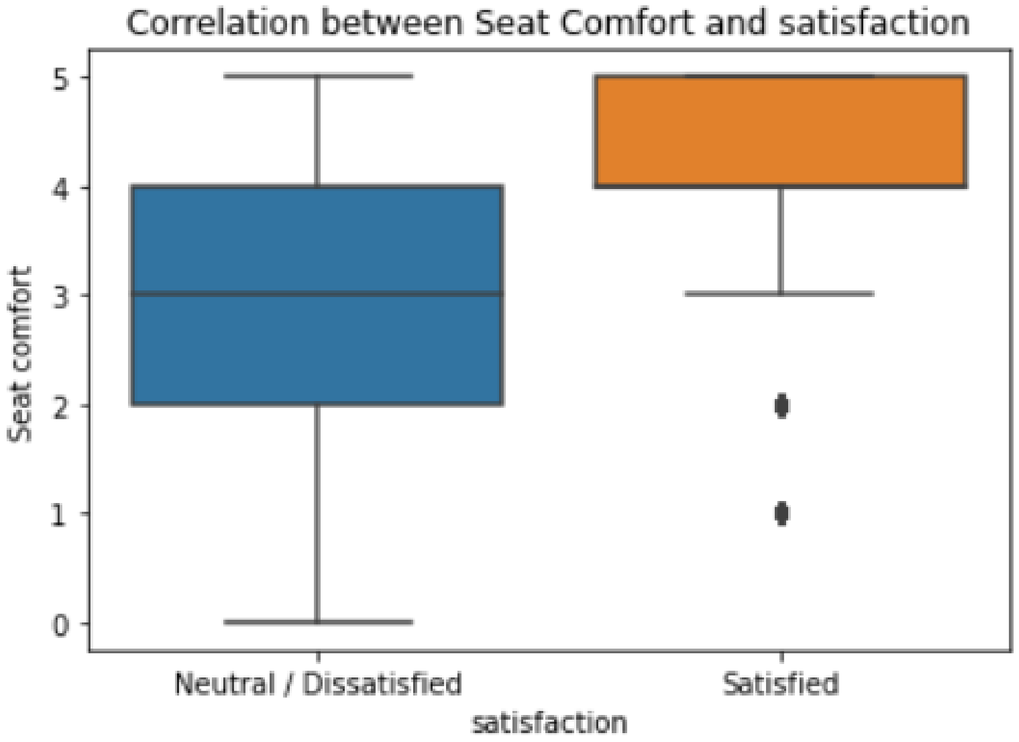
- From the pie chart, we can see that most of passenger experience good baggage handling service since the chart shows most rating lies in rating 4 and 5.

INFLIGHT ENTERTAINMENT ON SATISFACTION



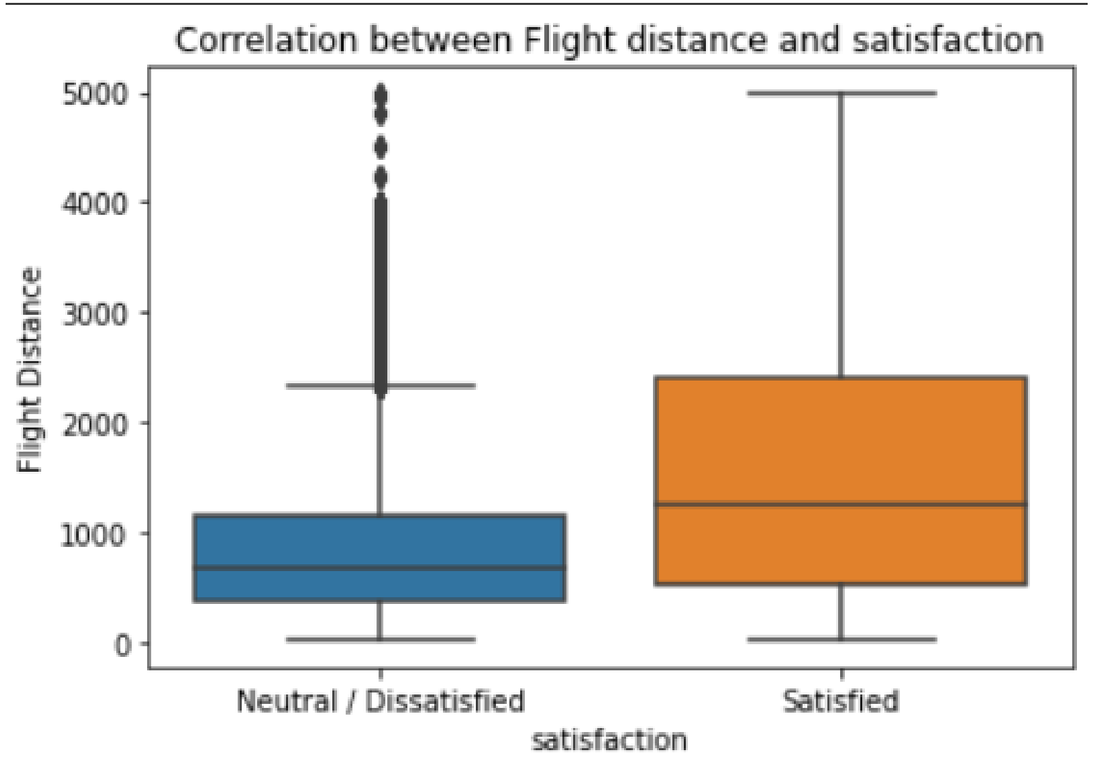
- From the plot, we can see that some people feel satisfied if they get better inflight entertainment. However, people will feel neutral/dissatisfied if they get lesser inflight entertainment. This does make sense since people will feel bored while inflight, better entertainment will satisfy the passenger. We also see that some of people also feel satisfied even thought they get lesser inflight entertainment. This is also possible since maybe passenger fell asleep or doing other work during flight so they don't really need entertainment.

SEAT COMFORT ON SATISFACTION



- Seat Comfort plot also shows us that people will more satisfied if they get more comfortable seat. However, some of people doesn't bother if they get less Seat Comfort. This thing is possible since Comfortable is relatively different for each person

FLIGHT DISTANCE ON SATISFACTION



- From the plot, we can see that most passenger that travels more than approximately more than 1300 will feel more satisfied. But, since the Neutral/Dissatisfied Distributions a little bit scattered, it also means that there are also a lot of people that don't feel satisfied enought while travel in further distance

DATA PREPROCESSING

01

Removing Missing Values

02

Handling Outliers

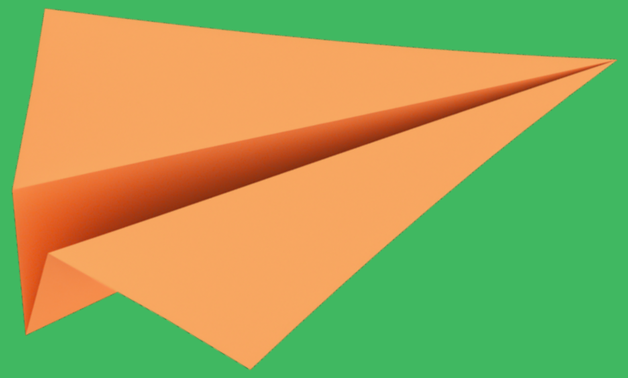
By dropping the outliers on the feature: Arrival Delay in Minutes and Departure Delay in Minutes

03

Encoding Variables

Change categorical variable to numerical variable

METHODOLOGY



MODELLING

Baseline Model

Using LogisticRegression with model score of 0.876708626148737 3

TESTING A BASIC MODEL

Testing 6 Models

```
{'Ada Boost': 0.9308826936442969,
 'Decision Tree': 0.9462506757278555,
 'KNN': 0.7423739284886863,
 'Random Forest': 0.9630473395629006,
 'SGDClassifier': 0.6455710865703915,
 'Support Vector Machine': 0.6691636419800757}
```

XGBOOST

0.9596107807552707

- After we try to use XGBoost model, Random Forest shows slightly better performance than this model.

CATBOOST

0.9654413468221484

From score above, we can see that CatBoostClassifier gives a great performance to our model.

MODEL EVALUATION

	precision	recall	f1-score	support
0	0.96	0.98	0.97	14767
1	0.97	0.95	0.96	11131
accuracy			0.97	25898
macro avg	0.97	0.96	0.96	25898
weighted avg	0.97	0.97	0.97	25898

- We achieved 97% of f1-score, which is great.

PREDICT ON UNSEEN DATASETS

	precision	recall	f1-score	support
0	0.96	0.98	0.97	14528
1	0.97	0.95	0.96	11365
accuracy			0.96	25893
macro avg	0.96	0.96	0.96	25893
weighted avg	0.96	0.96	0.96	25893

- We still achieved 97% of f1-score, it means that this model can predict well on unseen data

BUSINESS CASE STUDY

Business Case Study

By using Machine Learning Model, The Airline Department can improve their services by looking at the passenger satisfaction, this could also help the airline increasing their popularity through the passenger in the future




CONCLUSION AND RECOMMENDATION

CONCLUSION

In order to achive the Business Objective, the model is being tested and validated by some of models to train the datasets. After some training and validating, it was found out that CatBoostClassifier model seems to fulfill my model needs.
This model give about 97% of f1-score, means that this models can predict well enough for the unseen datasets

RECOMMENDATION

Get more correlation and more detailed data on passenger's data so that the model could predict more detailed on passenger's satisfaction.

Using New features such as passengers' income could be a great help for the model

CONTACT ME

08118493003

08118493003

https://github.com/DnYAlv/Scholarship-ML-Final-Project

THANK YOU