Airline Passenger Satisfaction Prediction

Contents

1. Introduction

1.1 Background and Problem Statement

Aviation business has experienced hardest hit during Covid-19 pandemic. Singapore airline operates 0 domestic flights, airline companies faced risk of bankruptcy. With the endemic, people rush to overseas holidays, air travel is expected to surge. What can airlines do to have a solid customer value proposition and a competitive edge?

In the current competitive environment, providing better services in the aviation industry can gain competitive advantages. Aviation companies should understand how their services satisfy customers' needs and should strive to achieve passenger satisfaction. This study examines the parameters that contribute the most to airline passenger satisfaction which can be considered as crucial points by the aviation industry.

1.2 Objectives

Without ML, small group survey with domain knowledge might be applied by business folks. The approach might not have quantitative and significant result and not scalable.

Binary classification models will be applied in this problem to demystify factors that matter to airline passenger satisfaction. The models were then compared to discern the most effective classification of satisfaction of customers based on misclassification rates, kappa coefficients, and sensitivity and specificity via the area under the curve (AUC) from receiver Operator Curves (ROC)

2. Business Use Case/Problem Scope

Before any meaningful and insightful data-driven output can be derived and concluded, it is crucial to clarify our business goals. Our identified mission aims to utilize 4 distinct kinds of machine learning models to help airline companies figure out areas they can improve to **optimize *customer satisfaction.*** We wish to extract the most important features or parameters that affect customer satisfaction for different groups of customers to help airplane companies better understand customer segmentation. We have recoded ed important columns and split the dataset into 18 sub-datasets with three columns “category of customer”, “flight distance” and “delay information” in order to make a targeted prediction within each group.

3. Dataset

Customer satisfaction scores from 120,000+ airline passengers, including additional information about each passenger, their flight, and type of travel, as well as the evaluation of different factors like cleanliness, comfort, service, and overall experience - Link to Dataset

3.1 Variables

The dependent variable Y is satisfaction label

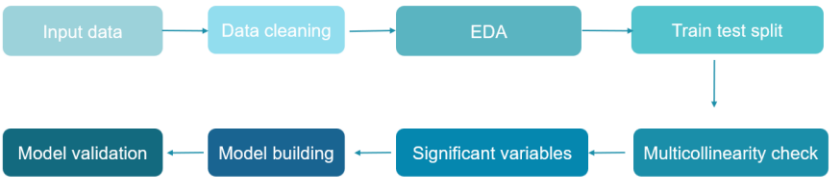
'Neutral or Dissatisfied’: 0

'Satisfied': 1

The independent variables are the 23 variables, customer profile, their flight information and evaluation to airline services.

4. Methodology

Block Diagram



5. Data Cleaning and EDA

5.1 Data Cleaning

5.1.1 Remove unnecessary data points

Variables ID, gender and age have not been used for the model development because gender and age will not have a greater impact on the overall satisfaction prediction.

5.1.2 Recode target variable

Replace the neutral or dissatisfied and satisfied Y variable with binary levels with 0 and 1

5.1.3 Data validation

Survey data is always objective, and requires respondents have correct understanding about score scale, i.e. they should not mark all 4 to 5 for overall unsatisfied nor all 1 to 2 for overall satisfied. Hence, we have checked likewise, and found out no passengers in the dataset have wrong understanding about score scale.

5.1.4 Data imputation

As mentioned above in section 3.1, for survey score fields, 0 means not applicable. In our model construct, score fields are ordinal. Without imputation it would be treated as a lowest score. As such, data imputation is necessary, and naturally score for satisfied and unsatisfied/neutral customers would be different. Hence, we use median by satisfaction group to impute score 0.

5.2 Anomaly Detection Analysis on Miss-matching Scores with Satisfaction

Since we have 12 score features, to visualize the customer clusters, the first step we need to do is doing PCA on our dataset. Our group choose to reduce the dimension into 2.

For the dataset of our project, one of the issues is whether a customer misunderstood the score criteria, one of the extreme scenarios is that a customer is satisfied for the trip, but he/she fill all 1 in the scores or vice versa. So firstly, we need to remove these incorrect data.

From our dataset, we have a binary classification of our customers where we split the customers into the satisfied and dissatisfied (or neural). In the figure 14 attached in Appendix B, we can see the satisfied customers (in green dots) clustered to the left while dissatisfied (or neural) customers clustered to the right. In this figure, we can also see some costumers indeed misunderstood the score criteria, because there are some red dots mixed in the green cluster and vice versa. So, our group tends to exclude this kind of dataset.

To remove the mismatching customers, our group first chose to use k-means for clustering. The k-means clustering result is as figure 15 attached in Appendix B. From the k-means clustering result, we clustered our customer into 2 groups. Our intention is to check whether the actual labels (satisfied/dissatisfied or neutral) is the same with the clustering result. we exclude the kind of data where the two of them are different.

Totally, we have 129878 samples, and the mis-matching data is 36174 samples, which accounts for 36% of our dataset. particularly, 9089 customers are satisfied with the flights while gives a relatively low score.

Another optional clustering model is Gaussian Mixture model, Gaussian Mixture model is more general than k-means, so it may give a better result for the downstream jobs. The 2-D GMM result is as Figure 16 attached in Appendix B.

We also add the covariance vectors of each group. the clustering result is not completely identical. Our group can use the downstream binary classification models to check whether it is better to use the GMM model to exclude mis-matching data. The two clustering models are used by our project to do anomaly detection.

The above anomaly detection method is an additional method for data cleaning, our group eventually won't use this method to clean the mismatching data, since the mis-matching data account for 1 third of the total dataset.

5.3 Analysis of important features affecting Satisfaction

- After the data cleaning the Arrival delay has 99 missing values

```
# Check for null values
df.isnull().sum().sort_values(ascending=False)
```

```
Arrival Delay                     99
Gender                             0
Seat Comfort                       0
survey_min_score                   0
```

Figure 1: Null values in Arrival Delay

We drop all the missing values

- The customers travelling through the airline belong to different classes - Business, Economy and Economy Plus. The ratio of satisfied passengers is very high among the people travelling from business class and the ratio is least for the economy class from the below catplot of satisfaction with respect to count for the variable 'Class'.

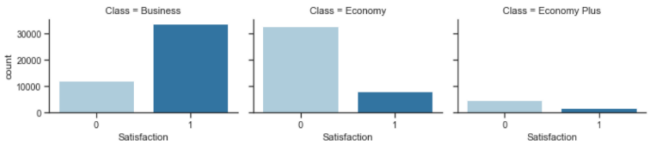
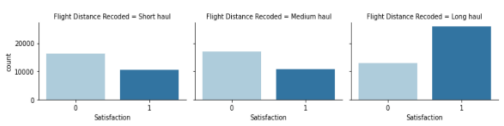


Figure 2: Satisfaction level for different Class

	Flight Distance	Departure Delay	Arrival Delay	Departure and Arrival Time Convenience
count	93417.000000	93417.000000	93417.000000	93417.000000
mean	1295.657257	14.860475	15.274565	3.158237
std	1048.128837	38.597603	39.076623	1.372422
min	31.000000	0.000000	0.000000	1.000000
25%	442.000000	0.000000	0.000000	2.000000
50%	925.000000	0.000000	0.000000	3.000000
75%	1979.000000	12.000000	13.000000	4.000000
max	4983.000000	1592.000000	1584.000000	5.000000

Figure 3 :Statistics for columns to be recoded

- Also, flight distance plays a very important role in customer satisfaction. Hence, we recode flight distance into three categories - Short haul, Medium Haul and Long haul. Flight distance is an important factor which affects customer satisfaction. The flight distance has been recoded as below
  Since flight distance ranges from min 31 km to 4983 and the **minimum distance of flights flying in Japan is between Okinawa and Kume Jima, Japan** we filter out the flights with 'Flight Distance' < 55 miles since they seem to be outliers.



Left – Figure 4: Flight Distance Recoded, Right – Figure 5 : Departure Delay Recoded

- Departure delay is recoded into two categories No Delay and Delay as from the distribution statistics we can see that more than 50% of the flights have delay hence to as to have equivalent distribution we split Departure delay into two categories.
- Arrival Delay was also recoded to No delay as well as Delay categories similarly. We consider four important categories of variables which affect Customer Satisfaction. The variable categories used to split the main dataset into sub datasets are – Class, Flight Distance Recoded, Departure Delay Recoded. Arrival Delay Recoded was excluded from the set of categories used as it was found that there was arrival delay only in cases when the flight had a corresponding departure delay. Arrival Delay variable is dependent on Departure Delay.

6. Model Construct

6.1 Clustering and PCA

PCA always used with clustering to visualise the high dimension feature data. PCA is normally used in exploratory data analysis and for making predictive models. It is commonly used for dimensionality reduction by projecting each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible. The first

principal component can equivalently be defined as a direction that maximizes the variance of the projected data.

The k-means clustering need the distribution of dataset to be identical covariance matrices. So sometimes k-means don't facilitate well. It is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. k-means clustering minimizes within-cluster variances (squared Euclidean distances.

Gaussian mixture model is another useful clustering method which is more general. GMM subjects to a set of mixture models which are probabilistic models for representing the presence of subpopulations within an overall population, without requiring that an observed data set should identify the sub-population to which an individual observation belongs. Our project then builds GMM clustering model on our binary classification dataset to exclude the outliers.

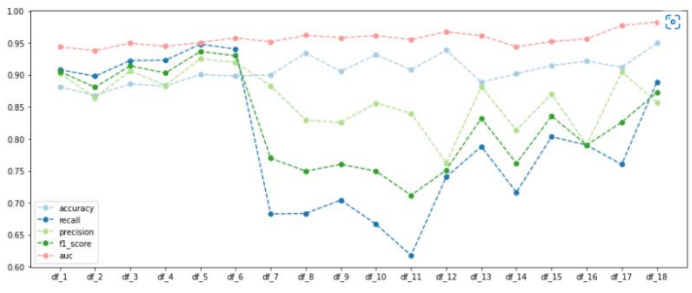
6.2 SVM

In machine learning, support-vector machines are supervised learning models with associated learning algorithms that analyse data for classification and regression analysis. Developed at AT&T Bell Laboratories by Vladimir Vapnik with colleagues. SVMs are one of the most robust prediction methods, being based on statistical learning frameworks or VC theory proposed by Vapnik (1982, 1995) and Chervonenkis (1974). In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. SVM algorithm can also be used in unsupervised learning. The support-vector clustering algorithm, created by Hava Siegelmann and Vladimir Vapnik, applies the statistics of support vectors, developed in the support vector machines algorithm, to categorize unlabeled data.

Our project initially tries to use SVM to do binary classification. however, the training of SVM model is slow because there are many features in our dataset. So, we abandon SVM model for our project.

6.3 Logistic Regression

After we split the dataset into 18 kinds of groups, we built logistic regression models for each of them and generated influential features. We generated a table (figure 8) to highlight every parameter which can measure the performance of each model. From the visualization of each parameter below, we found that group 1(for example, the combination of "Business"," Flight Distance Record","") to group 6 have a relatively better performance while the group 11 had the lowest recall value and f1_score value.



	top1	top2	top3	top4	top5
df_1	Customer Type	In-flight Wifi Service	Online Boarding	Leg Room Service	Check-in Service
df_2	Customer Type	Online Boarding	In-flight Wifi Service	In-flight Service	Check-in Service
df_3	Customer Type	Online Boarding	In-flight Wifi Service	Check-in Service	Cleanliness
df_4	Customer Type	Online Boarding	In-flight Wifi Service	On-board Service	Check-in Service
df_5	Online Boarding	On-board Service	Check-in Service	Cleanliness	In-flight Wifi Service
df_6	Online Boarding	Check-in Service	On-board Service	Cleanliness	In-flight Service
df_7	In-flight Wifi Service	Customer Type	On-board Service	Check-in Service	Online Boarding
df_8	In-flight Wifi Service	Customer Type	Check-in Service	Cleanliness	On-board Service
df_9	In-flight Wifi Service	Type of Travel	Check-in Service	On-board Service	Cleanliness
df_10	In-flight Wifi Service	Type of Travel	Check-in Service	Seat Comfort	On-board Service
df_11	In-flight Wifi Service	Customer Type	Check-in Service	Online Boarding	On-board Service
df_12	In-flight Wifi Service	Type of Travel	Online Boarding	Food and Drink	Baggage Handling
df_13	In-flight Wifi Service	Check-in Service	On-board Service	Online Boarding	Cleanliness
df_14	Type of Travel	In-flight Wifi Service	Check-in Service	On-board Service	Ease of Online Booking
df_15	In-flight Wifi Service	Seat Comfort	Online Boarding	On-board Service	Check-in Service
df_16	In-flight Wifi Service	Type of Travel	Online Boarding	Check-in Service	On-board Service
df_17	In-flight Wifi Service	Type of Travel	Online Boarding	Ease of Online Booking	Departure and Arrival Time Convenience
df_18	In-flight Wifi Service	Online Boarding	In-flight Entertainment	Check-in Service	Seat Comfort

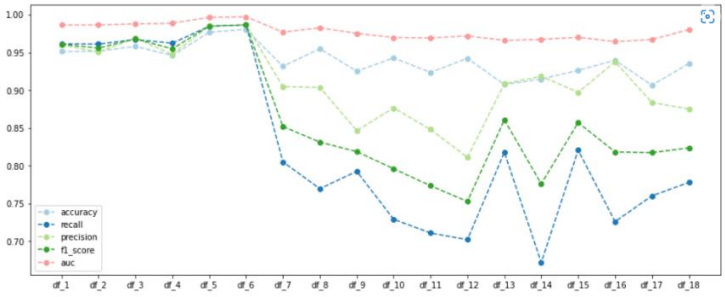
Left- Figure 8: Logistic Regression Model Summary Statistics, right- Figure 9: Logistic Regression Model Feature Importance

We listed the top5 important features for all 18 groups (figure 9), and we found that "In-flight services", "Online Boarding", "Seat comfort" as well as "Check-in Service" are common top features

which related to customer satisfaction. But even for different groups, there will be no obvious differences between top features. Airline companies can easily focus on the mentioned features to fulfil most customers and win their loyalty.

6.4 Random Forest

The random forest algorithm is an extension of the bagging method as it utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees. We apply random forest to each of the 18 datasets. We use the below variables "Type of Travel", "Customer Type", "Class", 'Flight Distance Recoded', 'Departure Delay Recoded', 'Arrival Delay Recoded' as the independent variables.



	top1	top2	top3	top4	top5
df_1	In-flight Entertainment	Online Boarding	Seat Comfort	Cleanliness	In-flight Wifi Service
df_2	In-flight Entertainment	Online Boarding	Seat Comfort	Cleanliness	In-flight Wifi Service
df_3	Online Boarding	In-flight Entertainment	Cleanliness	Seat Comfort	In-flight Wifi Service
df_4	Online Boarding	In-flight Entertainment	Seat Comfort	Cleanliness	In-flight Wifi Service
df_5	In-flight Entertainment	Seat Comfort	Cleanliness	Online Boarding	survey_max_score
df_6	Seat Comfort	In-flight Entertainment	Cleanliness	Online Boarding	Baggage Handling
df_7	In-flight Wifi Service	In-flight Entertainment	Cleanliness	Online Boarding	Food and Drink
df_8	In-flight Wifi Service	In-flight Entertainment	Cleanliness	Food and Drink	Online Boarding
df_9	In-flight Wifi Service	In-flight Entertainment	Cleanliness	Food and Drink	Online Boarding
df_10	In-flight Wifi Service	In-flight Entertainment	Cleanliness	Food and Drink	Online Boarding
df_11	In-flight Wifi Service	Cleanliness	In-flight Entertainment	Food and Drink	Online Boarding
df_12	In-flight Wifi Service	Food and Drink	In-flight Entertainment	Cleanliness	Online Boarding
df_13	In-flight Wifi Service	In-flight Entertainment	Cleanliness	Online Boarding	Food and Drink
df_14	In-flight Wifi Service	In-flight Entertainment	Cleanliness	Seat Comfort	Food and Drink
df_15	In-flight Wifi Service	In-flight Entertainment	Cleanliness	Food and Drink	Online Boarding
df_16	In-flight Wifi Service	In-flight Entertainment	Cleanliness	Online Boarding	Food and Drink
df_17	In-flight Wifi Service	Cleanliness	Food and Drink	In-flight Entertainment	Online Boarding
df_18	In-flight Wifi Service	Cleanliness	In-flight Entertainment	Food and Drink	Online Boarding

Left-Figure 10: Random Forest Model Summary Statistics, right-Figure 11: Random Forest Model Feature Importance

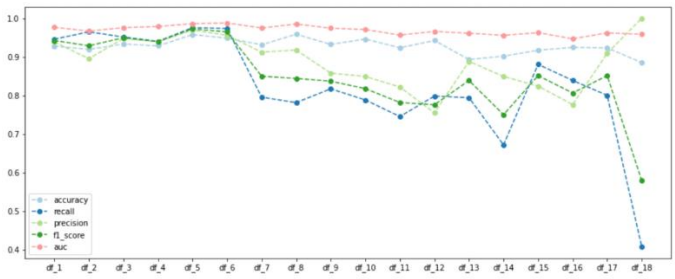
All the statistic scores such as accuracy precision, recall is very high for df_1 to df_6.

F1 score, precision and Recall is less for models df_7 to df_18 as compared to previous models.

- The feature that is **more important** in all the models is **In-flight Wifi service** which appears to be the most important feature for a maximum of the models.
- **Online Boarding is an important feature** for df_3 (**Business Class, Medium haul**, no departure delay) and df_4 (**Business Class, Medium haul**, departure delay) followed by In- Flight Entertainment.
- For datasets df_7 to df_10 and df_13 to df_16 (Passengers belonging to **economy class with flight distance either medium haul or short haul**) have common top three most important features which determine their Satisfaction level – **In-Flight Wifi Service, In-flight Entertainment and Cleanliness** in order from the most important to the third most.
- 4. df_5 and df_6 belong to categories of passengers belonging to Business Class and travelling long haul – **For Long Haul Seat Comfort and Food and Drinks** are of importance

6.5 Decision Tree

Decision tree is an information-based classification model, which split the records based on an attribute test that optimizes criterion, e.g., Gini, entropy etc.
In this study, we applied decision tree model on customer respondent to airline facilities and services, and flight information for the 18 groups we split likewise as logistic regression model and random forest model mentioned above. We noticed that accuracy and AUC is high for all the groups. Besides, we noticed online boarding is the most important feature for 5 out of 6 business class groups. In-flight services are the most important feature for 11 out of 12 economy or economy plus passengers. Details of data model statistics (e.g., accuracy, AUC) and feature importance in Fig 12 and 13.



	top1	top2	top3	top4	top5
df_1	(Online Boarding, 0.431)	(In-flight Wifi Service, 0.187)	(Type of Travel, 0.1448)	(Customer Type, 0.0803)	(In-flight Entertainment, 0.0079)
df_2	(Online Boarding, 0.4148)	(In-flight Wifi Service, 0.1684)	(Leg Room Service, 0.1077)	(Customer Type, 0.073)	(Type of Travel, 0.0579)
df_3	(Online Boarding, 0.5298)	(Type of Travel, 0.1303)	(In-flight Wifi Service, 0.0948)	(Customer Type, 0.0467)	(On-board Service, 0.0408)
df_4	(Online Boarding, 0.4757)	(Type of Travel, 0.1293)	(In-flight Wifi Service, 0.1244)	(In-flight Service, 0.0623)	(Customer Type, 0.0498)
df_5	(In-flight Entertainment, 0.5809)	(In-flight Wifi Service, 0.205)	(Online Boarding, 0.0981)	(Customer Type, 0.0803)	(Seat Comfort, 0.0468)
df_6	(Online Boarding, 0.2873)	(In-flight Service, 0.1092)	(In-flight Entertainment, 0.088)	(In-flight Wifi Service, 0.0817)	(Cleanliness, 0.0587)
df_7	(In-flight Wifi Service, 0.7675)	(Ease of Online Booking, 0.0503)	(Customer Type, 0.0384)	(Type of Travel, 0.0282)	(Check-in Service, 0.0247)
df_8	(In-flight Wifi Service, 0.7517)	(Type of Travel, 0.0525)	(Customer Type, 0.0378)	(Online Boarding, 0.0314)	(Check-in Service, 0.0309)
df_9	(In-flight Wifi Service, 0.7663)	(Ease of Online Booking, 0.0629)	(Customer Type, 0.033)	(Type of Travel, 0.0218)	(Online Boarding, 0.0211)
df_10	(In-flight Wifi Service, 0.7614)	(Type of Travel, 0.0503)	(Ease of Online Booking, 0.0422)	(Customer Type, 0.0378)	(In-flight Service, 0.029)
df_11	(In-flight Wifi Service, 0.8162)	(Type of Travel, 0.0497)	(Ease of Online Booking, 0.0398)	(Baggage Handling, 0.0216)	(Check-in Service, 0.0173)
df_12	(In-flight Wifi Service, 0.8075)	(Type of Travel, 0.0595)	(Ease of Online Booking, 0.0398)	(In-flight Service, 0.0219)	(Check-in Service, 0.019)
df_13	(In-flight Wifi Service, 0.7297)	(Type of Travel, 0.1238)	(Baggage Handling, 0.0556)	(Customer Type, 0.0382)	(Online Boarding, 0.0184)
df_14	(In-flight Wifi Service, 0.7521)	(Type of Travel, 0.0999)	(Baggage Handling, 0.0385)	(Customer Type, 0.0378)	(On-board Service, 0.0356)
df_15	(In-flight Wifi Service, 0.6731)	(Online Boarding, 0.1218)	(Type of Travel, 0.113)	(In-flight Entertainment, 0.0892)	(Customer Type, 0.066)
df_16	(Online Boarding, 0.4197)	(In-flight Wifi Service, 0.3299)	(Type of Travel, 0.1686)	(Leg Room Service, 0.0323)	(Customer Type, 0.0202)
df_17	(In-flight Wifi Service, 0.8814)	(Type of Travel, 0.1094)	(Check-in Service, 0.0032)	(Customer Type, 0.0)	(Flight Distance Recoded, 0.0)
df_18	(In-flight Wifi Service, 0.8647)	(Type of Travel, 0.0353)	(Customer Type, 0.0)	(Flight Distance Recoded, 0.0)	(Baggage Handling, 0.0)

Left-Figure 12: Decision Tree Model Summary Statistics, right-Figure 13: Decision Tree Model Feature Importance

7. Comparison between models

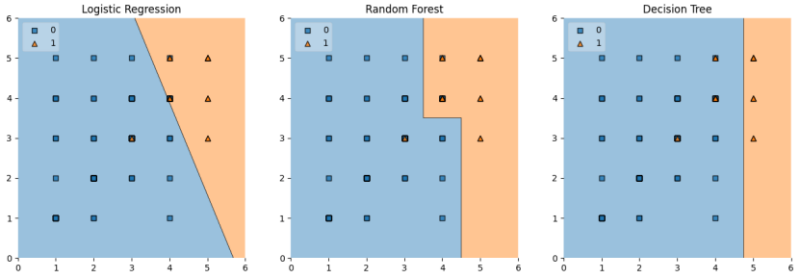


Figure 17: classification comparison between models with the 18th dataset (customers with economy plus, long haul and departure delay)

We choose 2 important features to see the classification result. The vertical axis represents In-flight Wi-Fi Service, and the horizontal axis represents Online boarding. Both axes are scores so we can see an array in the plots. The blue square represents the dissatisfied or neural and the red triangle represents the satisfied. One interesting sign figure in this plot is a small red triangle with a black square background. That means when in-flight Wi-fi Service is 4 and Online boarding is 4, the numbers of customers who are satisfied or dissatisfied (or neural) with the flight are evenly distributed in this point, i.e., the percentages of satisfied or dissatisfied (or neural) are near with each other. We can also see the logistic regression give a smooth split within the data while the random forest gives a zigzag one. That is due to the models' characteristics.

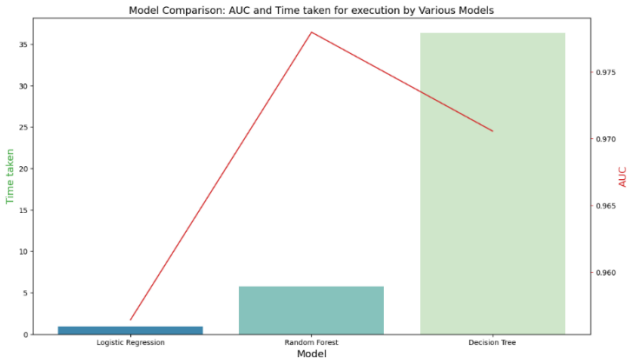


Figure 18: comparison of computing times and AUCs between different models

Our group also gives the comparison bar plot of the overall time taken (total time fitting for 18 split datasets) and comparison line plot of the mean AUC (mean of the model predicted AUC from 18 split dataset) by each model. From the plot, we can see that the time taken for decision tree is much more than either logistic regression or random forest. That is because we use grid search to search

for the best classifiers. Logistic regression model takes least time to fit whist the AUC from random forest is the highest.

8. Summary

1. A common strategy for the Passengers belonging to Economy class/ Economy plus class with flight distance either medium haul or short haul to improve customer satisfaction and to retain the existing customer base would be to target and improve/ maintain **In-Flight Wifi Service, In-flight Entertainment and Cleanliness factors**

2. Strategy for the Business Class passengers travelling for medium haul can be to **streamline the Online boarding process with smoother priority boarding** as the most important parameter which affects satisfaction of Business class passengers appears to be Online Boarding.

3. **For long Haul flight distances food and drink facility as well as seat comfort** for flights must be strategized to be comparable to other airlines so that the airline company can ensure that it provides best service.

Appendix A Dataset

Table 1. Overview of dataset and descriptions

ID	*Unique passenger identifier*
Gender	Gender of the passenger (Female/Male)
Age	Age of the passenger
Customer Type	Type of airline customer (First-time/Returning)
Type of Travel	Purpose of the flight (Business/Personal)
Class	Travel class in the airplane for the passenger seat
Flight Distance	Flight distance in miles
Departure Delay	Flight departure delay in minutes
Arrival Delay	Flight arrival delay in minutes
Departure and Arrival TimeConvenience	Satisfaction level with the convenience of the flight departure and arrival times from 1(lowest) to 5 (highest) - 0 means "not applicable"
Ease of OnlineBooking	Satisfaction level with the online booking experience from 1 (lowest) to 5 (highest) - 0 means "not applicable"
Check-in Service	Satisfaction level with the check-in service from 1 (lowest) to 5 (highest) - 0 means "not applicable"
Online Boarding	Satisfaction level with the online boarding experience from 1 (lowest) to 5(highest) - 0 means "not applicable"
Gate Location	Satisfaction level with the gate location in the airport from 1 (lowest) to 5(highest) - 0 means "not applicable"
On-board Service	Satisfaction level with the on-boarding service in the airport from 1 (lowest) to 5 (highest) - 0 means "not applicable"
Seat Comfort	Satisfaction level with the comfort of the airplane seat from 1 (lowest) to 5(highest) - 0 means "not applicable"
Leg Room Service	Satisfaction level with the leg room of the airplane seat from 1 (lowest) to 5 (highest) - 0 means "not applicable"
Cleanliness	Satisfaction level with the cleanliness of the airplane from 1 (lowest) to 5(highest) - 0 means "not applicable"
Food and Drink	Satisfaction level with the food and drinks on the airplane from 1 (lowest) to 5 (highest) - 0 means "not applicable"
In-flight Service	Satisfaction level with the in-flight service from 1 (lowest) to 5 (highest) - 0means "not applicable"
In-flight WifiService	Satisfaction level with the in-flight Wifi service from 1 (lowest) to 5(highest) - 0 means "not applicable"
In-flightEntertainment	Satisfaction level with the in-flight entertainment from 1 (lowest) to 5 (highest) - 0 means "not applicable"
Baggage Handling	Satisfaction level with the baggage handling from the airline from 1(lowest) to 5 (highest) - 0 means "not applicable"
Satisfaction	Overall satisfaction level with the airline (Satisfied/Neutral or unsatisfied)

Appendix B Modelling

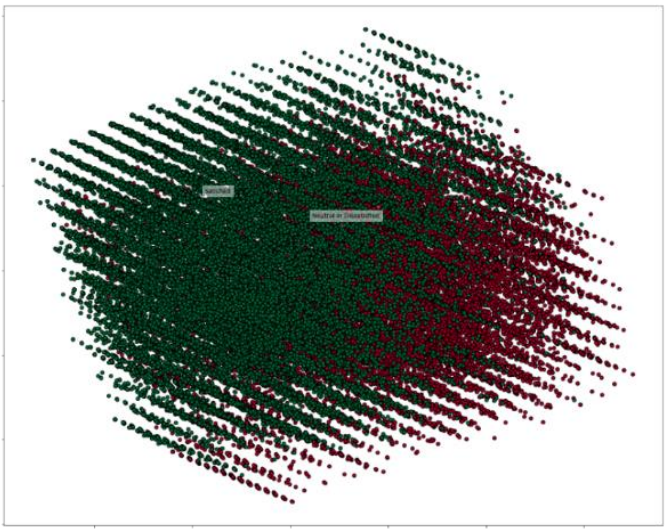


Figure 14: distribution of customers based on their scores

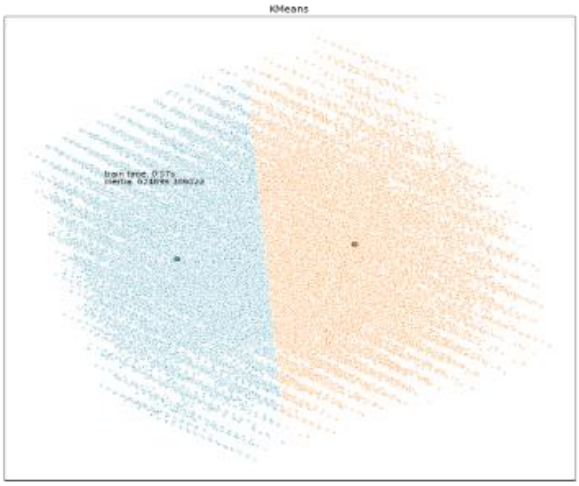


Figure 15: k-means clustering for customers based on their score features

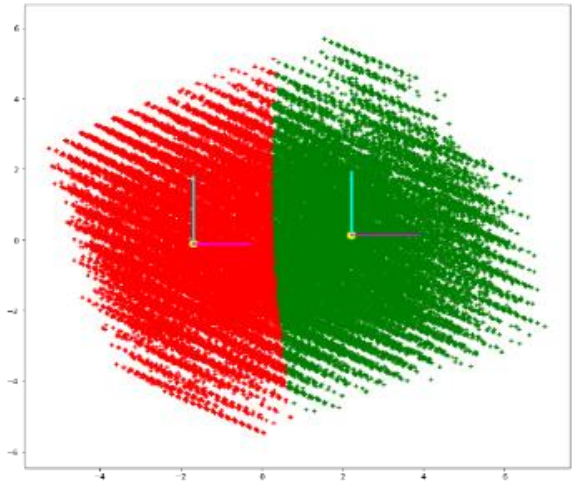


Figure 16 GMM clustering of customers based on their score features