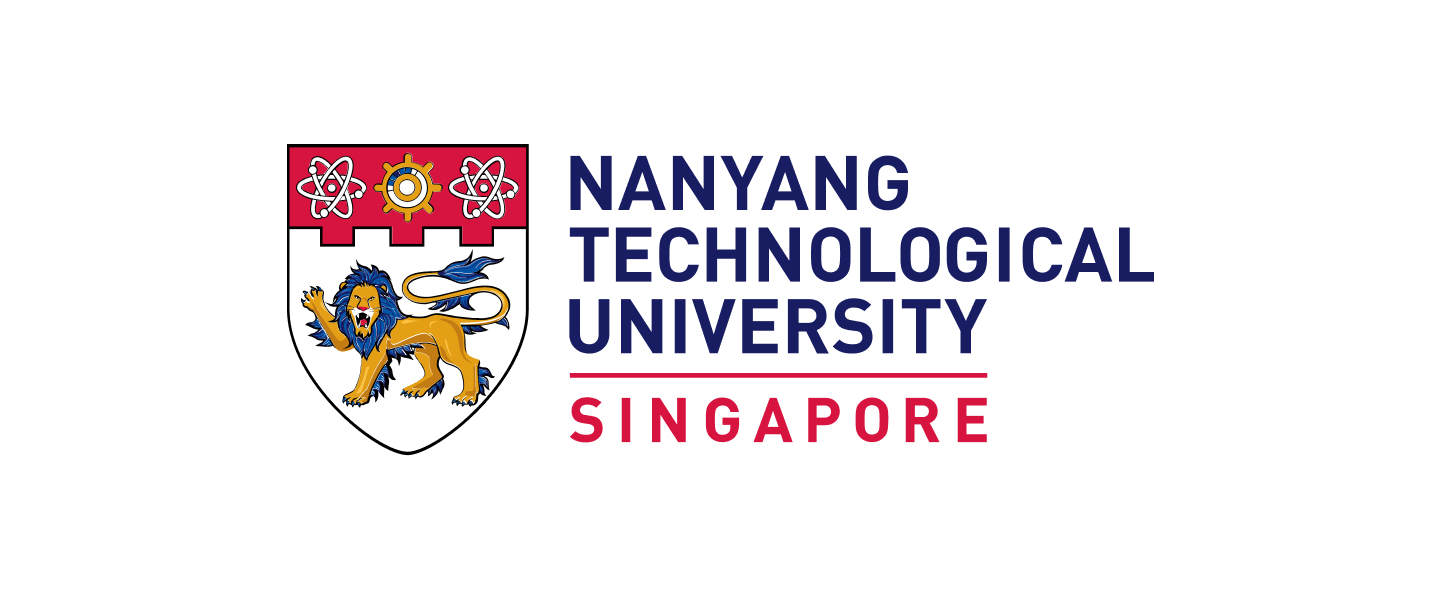
# 



# **SC5010**

**Group Assignment**

|  |  |  |
| --- | --- | --- |
| 1. | Tay Jin Yang | U2140995E |

**Abstract**

For our project, we will be aiming to create a classification machine learning model on R to be able to predict whether the following factors are enough to gain customer satisfaction. And by using this model, we hope that airline companies can better allocate their resources in order to gain customer satisfaction and hence perform better than their competitors. To begin, we first did some data preparation by removing and tunning data such that it can be used for analysis. We then use visual tools available on R to better analyze the correlation between variables. We then use multiple classification machine learning models available and compare them which each other to see which is the most reliable.

**1 Problem Description**

In 2023, the aviation industry is expected to make a return to its original state, after facing unprecedented challenges due to Covid-19. The travel restrictions and bans associated with the pandemic have severely impacted the transportation sector, especially the airline industry, which has literally shut down for nearly 3 years.

Now, aviation is a highly competitive market and airlines must renew themselves to stand out from the competition by providing memorable travel experiences, rather than just being a normal carrier. Improving the quality of customer service is an important aspect of remaining competitive, as it helps to maintain customer loyalty, which will eventually contribute to the long-term growth of the company.

Aim: Help airline companies gain an upper edge against other competitors by improving their passenger satisfaction level, using machine learning and data analysis to find the best area to work on.

**2 Approach (Methodology + Algorithms)**

We will be looking at the dataset for airline passengers’ satisfaction obtained from kaggle, which contains the result from an airline passenger satisfaction survey. The data contains 25 columns variables that can affect a passenger satisfaction level and consist of 103 904 respondents’ response. It also contains the satisfaction level with three possible satisfaction levels, satisfaction, neutral or dissatisfaction. For our project, we will be using R script to view and analyze the correlation between the variables with respect to the satisfaction level. Here is a brief description of the dataset.

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Comments** | **Data Attribute** |
| X | The order of the data | Nominal |
| id | Unique No. for each data | Nominal |
| Gender | Gender of passengers (Female, Male) | Nominal |
| Customer.Type | Customer type (Loyal customer, disloyal customer) | Nominal |
| Age | The actual age of the passengers | Nominal |
| Type.of.Travel | Purpose of the flight of the passengers (Personal Travel, Business Travel) | Nominal |
| Class | Travel class in the plane of the passengers (Business, Eco, Eco Plus) | Nominal |
| Flight.Distance | The flight distance of this journey | Ratio |
| Inflght.wifi.service | Satisfaction level of the inflight wifi service (0:Not Applicable;1-5) | Ordinal |
| Departure.Arrival.time.convenient | Satisfaction level of Departure/Arrival time convenient | Ordinal |
| Ease.of.Online.booking | Satisfaction level of online booking | Ordinal |
| Gate.location | Satisfaction level of Gate location | Ordinal |
| Food.and.drink | Satisfaction level of Food and drink | Ordinal |
| Online.boarding | Satisfaction level of online boarding | Ordinal |
| Seat.comfort | Satisfaction level of Seat comfort | Ordinal |
| Inflight.entertainment | Satisfaction level of inflight entertainment | Ordinal |
| On.board.service | Satisfaction level of On-board service | Ordinal |
| Leg.room.service | Satisfaction level of Leg room service | Ordinal |
| Baggage.handling | Satisfaction level of baggage handling | Ordinal |
| Checkin.service | Satisfaction level of Check-in service | Ordinal |
| Inflight.service | Satisfaction level of inflight service | Ordinal |
| Cleanliness | Satisfaction level of Cleanliness | Ordinal |
| Departure.Delay.in.Minutes | Minutes delayed when departure | Ratio |
| Arrival.Delay.in.Minutes | Minutes delayed when Arrival | Ratio |
| satisfaction | Airline satisfaction level (Satisfaction, neutral or dissatisfaction) | Ordinal |

**3 Data Preparation**

Data Cleaning

We eliminate variables that are irrelevant for an airline company to consider. These are the variables we find irrelevant along with the reasons we have for them.

1. Age - Airline companies cannot control the age of customers wanting to use their airline.
2. X - Irrelevant
3. Id - Irrelevant
4. Type.of.Travel – Airline companies do not have control over the type of travel the customer desire.
5. Gate.location - Airline companies do not have control over gate location as they are usually assigned by airports themselves or on a first come-first serve basis.
6. Departure.Arrival.time.convenient – The preference of a convenient time is subjective to each passengers. In addition, airline company do not have full control over such time as there are also other variables to consider (availability of airplanes, weather, being able to cater a wider spread of timing)

Next, we eliminate data rows with satisfaction level being not applicable as some passengers may only experience certain variables and hence making the comparison unfair.

Feature Engineering

Next, we change categorical data into numerical data so that we can analyze the correlation between data. Here is the data we changed.

1. Satisfaction (Satisfied, Neutral or Dissatisfied) - (1,0)
2. Class (Eco, Eco plus, Business) - (1,2,3)
3. Customer type (Loyal Customer, Disloyal Customer) - (1,0)

Next, we normalize the numeric feature of the data and do a 20-80 split of the data to do testing and training, for both the normalized data and the non-normalized data.

Then, using the normalized data, we use the function corrplot() to plot a Pearson correlation heat map to better visualize the correlation coefficient between the outcome (satisfaction) and each of the predictor.

**4 Data Analysis**

Timeline

Description automatically generated

Figure 1: Pearson correlation heatmap of every variable

Here is what we can infer from our correlation heatmap (what we deemed important):

1. The top 5 variables that have the strongest correlation with passengers’ satisfaction level are the following: Online boarding (0.56), Class (0.47), Inflight entertainment (0.43), Seat comfort (0.38) and Inflight Wi-Fi service (0.37).
2. The variable inflight entertainment has a strong correlation with cleanliness (0.68), followed by on seat comfort (0.61) and food and drink (0.61).
3. The variable seat comfort has a strong correlation with cleanliness (0.7), followed by inflight entertainment (0.61) and food and drink (0.58)
4. There are some variables that have a very weak correlation with satisfaction level as well as other variables and can be removed. (departure delay and arrival delay)

From the analysis above, we have 2 conclusions:

1. We will be filtering out the variables Departure.Delay.in.Minutes and Arrival.Delay.in.Minutes as from the correlation heatmap, these variables does not relate much to the other variables aside each other, which make sense as departing late will lead to arriving late.

Timeline

Description automatically generated



Figure 2: Correlation heatmap to show the variables are isolated.

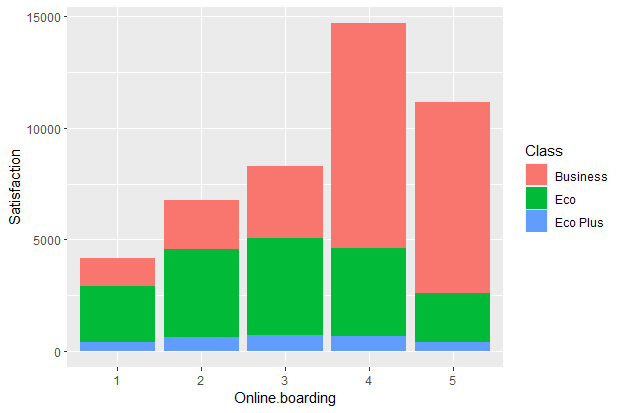
1. When we analyze further on the relation between the strongest correlation variable (Online boarding) and satisfaction level and separate them by class, we realized that the responds from different classes are not evenly spread. As seen from figure 3 below, Economic class contribute the most for satisfaction level 1-3 while business class contribute the most for satisfaction level 4-5. 

Figure 3: Bar chart on Satisfaction level against Online boarding

This implies that business class have better superior online boarding service, which plays a big part in satisfaction level. However, this also implies that the data is not fair as the whole correlation map might have just been an outline of correlation between the business class passengers as they carry more weights. In addition, different classes have different situations, and it makes sense that business class will have better boarding services given that they are paying more. Hence, we concluded that there is a need to split the data among classes as well.

Figure 4: To show the uneven spread in satisfaction between classes.

After splitting the data, we plot each of their correlation heatmap to visually

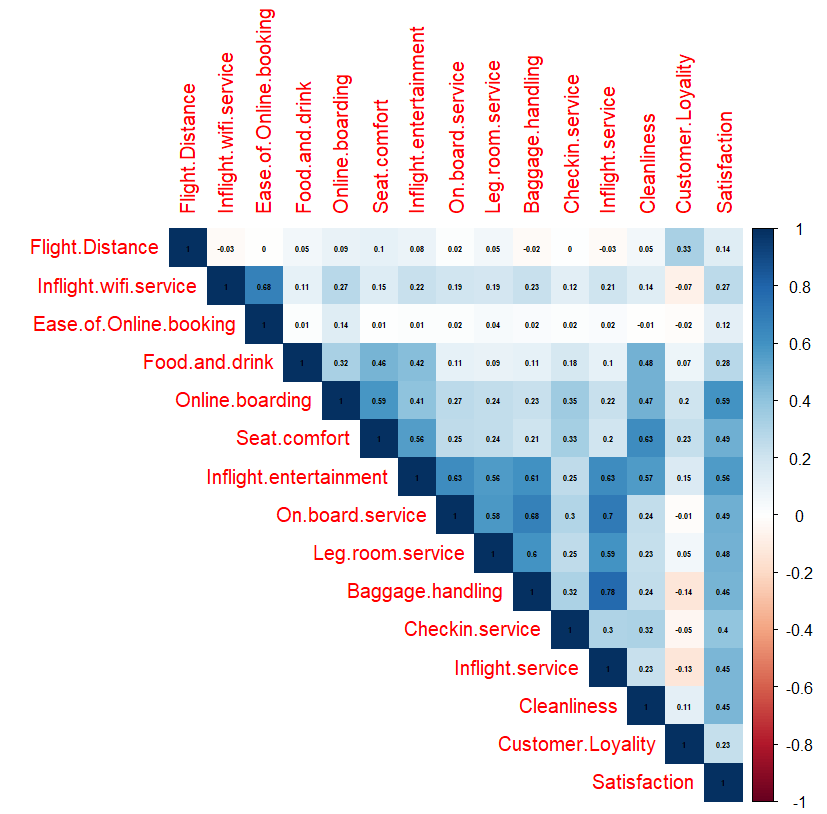


Figure 5: Correlation heatmap for Business class

Based on Figure 5, the top 4 variables that have the strongest correlation with passengers’ satisfaction level are the following: Online boarding (0.59), Inflight entertainment (0.56), Seat comfort (0.49) and on-board service (0.49).

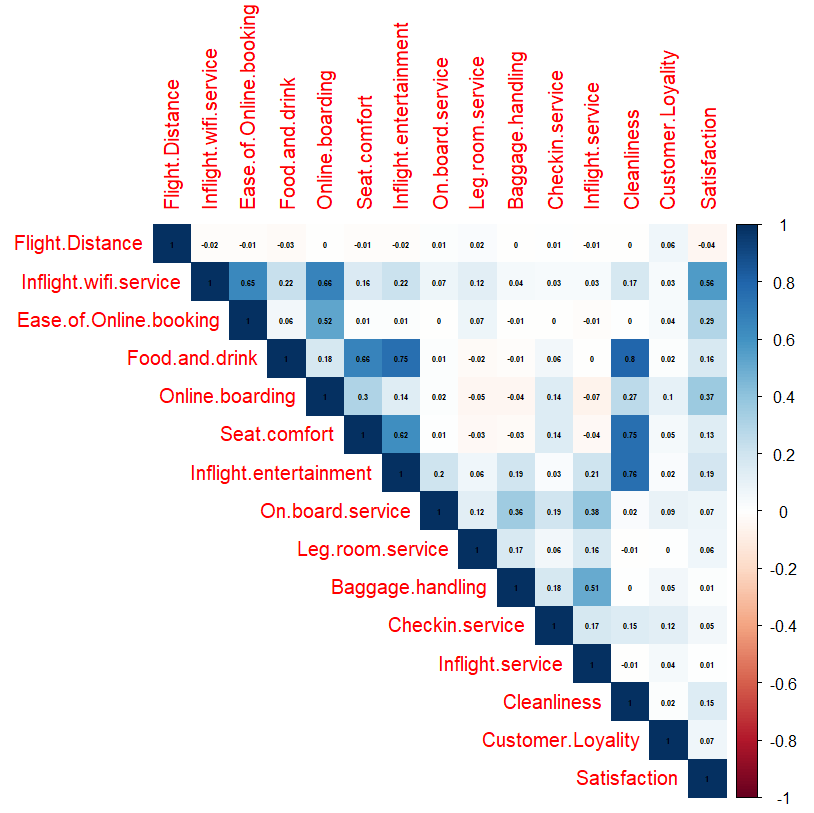


Figure 6: Correlation heatmap for Economy Plus class

Based on Figure 6, the top 4 variables that have the strongest correlation with passengers’ satisfaction level are the following: Inflight wifi service (0.56), Online boarding (0.37), Ease of Online booking (0.29) and Inflight entertainment (0.19).

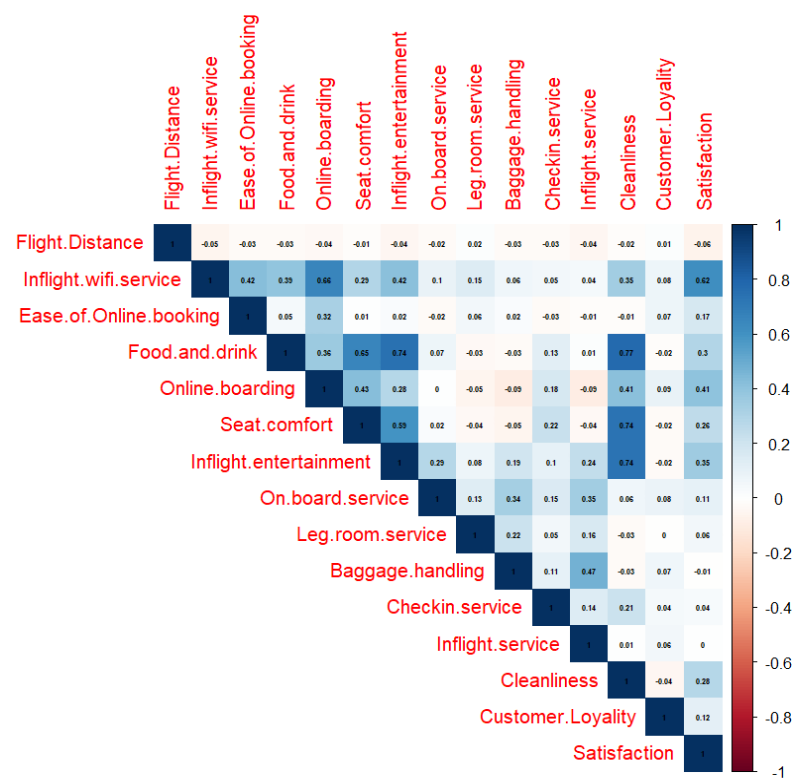


Figure 7: Correlation heatmap for Economy class

Based on Figure 7, the top 4 variables that have the strongest correlation with passengers’ satisfaction level are the following: Inflight wifi service (0.62), Online boarding (0.41), Inflight entertainment (0.35) and food and drinks (0.3).

From these 3 correlation heatmaps (Figure 5,6,7), these are the inferences we deemed important:

1. The economy class and economy plus class are very similar in nature, with the heatmap having almost the same temperature in the same areas, with Inflight wifi service, Online boarding and inflight entertainment being the top few variables.
2. Inflight wifi service is heavily correlated to satisfaction level for economy (0.62) and economy plus (0.56), but so as much for business class (0.27). One possible hypothesis is business class passengers do not rely much on wifi services as there is more infight entertainment available for them. As seen from figure 5, inflight entertainment has a correlation value of 0.56 for business class, which is much higher as compared to economy plus (0.19) and economy class (0.35).
3. Especially for business class, inflight entertainment is closely related to many other variables such as on-board service (0.63), inflight service (0.63), baggage handling (0.61), cleanliness (0.57), seat comfort (0.56) and leg room service (0.56). From this, we can see that business class passengers prioritize services as well as environment during flight and hence, to improve satisfaction level for business class, airline companies can focus on these different areas since inflight entertainment is a strong variable to satisfaction level.

**5 Machine learning**

Now, we will try and use different machine learning models and find the best model to correctly predict the satisfaction level of passengers. In addition, since we are only trying to differentiate those that are satisfied and those that are not, we will be using classification models. Here are the following models we will be using, followed by their pros and cons.

|  |  |  |
| --- | --- | --- |
| **ML model** | **Pros** | **Cons** |
| XGBoost | - Less feature engineering of data required  - Robust against outliers  - Good execution speed | - Harder to tune if too many hyperparameters  - Prone to overfitting  - Weak against sparse and unstructured data |
| Random Forest | - Good performance on imbalanced datasets  - Can handle huge amount of data  - Robust against outliers | - Predictions of trees need to be uncorrelated  - Features must have some predictive power |
| SVM (linear) +SVM (polynomial) | - Outliers have less impact  - Can adjust by changing kernel | - Slow, require large computing process  - Selecting appropriate kernel can be tricky |
| KNN classification | - No assumption about data  - Simple and constantly evolving | - Prone to curse of dimensionality  - Must scale data  - Sensitive to outliers and missing data |
| Logistic Regression | - Feature scaling not required  - Tuning of hyperparameters not required | - Poor performance of non-linear data  - Poor performance if consist of missing data |

Now, we will be applying our machine learning models into 4 different datasets (original data without splitting classes, Business Class passengers, Eco Plus passengers, Eco passengers). Then, we will be comparing their accuracy along with their false positive and false negative percentage. For our situation, it will be better if false negative>false positive as having a false negative can result in excess work that achieve the requirement but a false positive can result in lack of work and does not achieve the requirement.

|  |  |  |
| --- | --- | --- |
| **Applying ML for dataset without splitting classes** | | |
| **ML model** | **Accuracy** | **False Positive / False Negative** |
| XGBoost | 0.9208 | False Positive = 7.9%  False Negative = 8.0% |
| Random Forest | 0.9563 | False Positive = 2.8%  False Negative = 7.2% |
| SVM (linear) | 0.8785259 | False Positive = 8.4%  False Negative = 18.8% |
| SVM (Polynomial, Degree = 2) | 0.9314528 | False Positive = 4.1%  False Negative = 11.8% |
| SVM (Polynomial, Degree = 3) | 0.9381256 | False Positive = 3.5%  False Negative = 11.0% |
| SVM (Polynomial, Degree = 4) | 0.9372157 | False Positive = 11.0 %  False Negative = 12.3% |
| KNN classification (after finding best K) | 0.9391871 | False Positive = 3.4%  False Negative = 10.9% |
| Logistic Regression | 0.8768577 | False Positive = 9.2%  False Negative = 17.9% |

|  |  |  |
| --- | --- | --- |
| **Applying ML for dataset for Business class** | | |
| **ML model** | **Accuracy** | **False Positive / False Negative** |
| XGBoost | 0.9397 | False Positive = 13.6%  False Negative = 5.1% |
| Random Forest | 0.9574 | False Positive = 6.6%  False Negative = 2.8% |
| SVM (linear) | 0.8839458 | False Positive = 16.2%  False Negative = 8.7% |
| SVM (Polynomial, Degree = 2) | 0.9284333 | False Positive = 11.7%  False Negative = 4.3% |
| SVM (Polynomial, Degree = 3) | 0.9355255 | False Positive = 11.6%  False Negative = 3.2% |
| SVM (Polynomial, Degree = 4) | 0.9158607 | False Positive = 17.6 %  False Negative = 2.7% |
| KNN classification (after finding best K) | 0.9393939 | False Positive = 8.3%  False Negative = 4.7% |
| Logistic Regression | 0.8836235 | False Positive = 17.2%  False Negative = 8.2% |

|  |  |  |
| --- | --- | --- |
| **Applying ML for dataset for Economy Plus class** | | |
| **ML model** | **Accuracy** | **False Positive / False Negative** |
| XGBoost | 0.7573 | False Positive = 43.3%  False Negative = 22.9% |
| Random Forest | 0.9513 | False Positive = 1.1%  False Negative = 31.4% |
| SVM (linear) | 0.9285476 | False Positive = 0%  False Negative = 58.0% |
| SVM (Polynomial, Degree = 2) | 0.9345576 | False Positive = 0.7%  False Negative = 48.0% |
| SVM (Polynomial, Degree = 3) | 0.9355593 | False Positive = 0.9%  False Negative = 47.2% |
| SVM (Polynomial, Degree = 4) | 0.9345576 | False Positive = 0.8%  False Negative = 47.4% |
| KNN classification (after finding best K) | 0.9352254 | False Positive = 0.9%  False Negative = 46.3% |
| Logistic Regression | 0.9342237 | False Positive = 0.6%  False Negative = 49.1% |

|  |  |  |
| --- | --- | --- |
| **Applying ML for dataset for Economy class** | | |
| **ML model** | **Accuracy** | **False Positive / False Negative** |
| XGBoost | 0.8976 | False Positive = 100%  False Negative = 9.9% |
| Random Forest | 0.9398 | False Positive = 1.7%  False Negative = 26.7% |
| SVM (linear) | 0.9277108 | False Positive = 1.7%  False Negative = 33.7% |
| SVM (Polynomial, Degree = 2) | 0.937751 | False Positive = 0.5%  False Negative = 33.7% |
| SVM (Polynomial, Degree = 3) | 0.9297189 | False Positive = 0.7%  False Negative = 37.2% |
| SVM (Polynomial, Degree = 4) | 0.9257028 | False Positive = 0.5%  False Negative = 40.1% |
| KNN classification (after finding best K) | 0.9257028 | False Positive = 0.1%  False Negative = 38.4% |
| Logistic Regression | 0.9277108 | False Positive = 1.5%  False Negative = 34.9% |

Summary from the 4 tables above:

1. For all classes dataset, random forest classifier has the best accuracy of 0.9563 with false negative = 7.2% > false positive = 2.8%.

2. For all business class datasets, random forest classifier has the best accuracy of 0.9574 with false negative = 2.8% < false positive = 6.6%. This is not ideal as we would rather have false negative as compared to false positive. However, we saw that all our ML models for business class dataset have false negative<false positive. Hence, the random forest classifier is still the best for business class.

3. For all economy plus class datasets, random forest classifier has the best accuracy of 0.9513 with false negative = 31.4% > false positive = 1.1%.

4. For all economy class datasets, random forest classifier has the best accuracy of 0.9398 with false negative = 26.7% > false positive = 1.7%.

**6 Discussion of Pros and Cons**

Pros - Most of the variables given in the dataset have some predictive power which allows them to have correlation with other variables. As a result, we can also visually analyze the correlation between predictors as well. There is also the abundance of data which helps make our analysis more accurate since passenger satisfaction surveys can be very subjective and a pattern might not be noticeable if the data set is too small.

Cons - There is a huge uneven split of data among the three classes, with most of the data belonging to the business class. This resulted in the dataset of the unsplit classes behaving similar to the business class dataset as seen from comparing figure 1 and figure 5.

**7 Conclusions**

Regarding machine learning, while for the dataset where classes are unsplit has a great accuracy of 0.9563 and a great false negative/false positive ratio, this still should not be used as the data is mostly dominated by business class. In addition, from our data analysis, we see that all three classes have different variables with strong correlation with customer satisfaction level. Hence, there is no ‘one size fit all’ model and airline companies should separate their dataset according to classes and use the ideal models based on each class instead.

**Reference**

Airline passenger satisfaction : The Importance of Customer Services. Moment. (2023, March 15). Retrieved April 16, 2023, from https://moment.tech/2023/03/13/airline-passenger-satisfaction-the-importance-of-customer-services/

<https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction>