Airlines Satisfaction Prediction

Introduction

Passengers' satisfaction with their air travel is a crucial metric for airlines industry, and customers' evaluations of flight experiences are influenced by a variety of elements such as seat comfort, ease of online reservation, legroom, departure and arrival delay times, and so on. Therefore, my purpose of this project is to establish the classification model to predict passengers' evaluation of flight experience and determine of what are the important factors affecting their satisfaction, which can help airlines understand their advantages and enhance their business from a customer's standpoint.

Data Preparation & Cleaning

The original Kaggle dataset provides us with a training set and a testing set, but I combine them into a single data frame to improve data clarity. The total data set contains a large amount of data (approximately 130K rows) and 25 variables. I first remove the unneeded columns, and then I remove the missing value because it accounts for only 0.3 % of the entries. I also change the value of the 'Satisfaction' variable to a binary value, with 0 representing neutral/dissatisfied and 1 representing satisfied. Finally, there are 129,487 observation columns, 21 feature columns, and 1 binary target column remaining.

Exploratory Data Analysis

After cleaning the data, we can observe some key information according to the different plots:

- *Count plot* (*Figure I*): It shows the distribution of categorial variables. When I rank satisfaction by customer type, disloyal customers have a low level of satisfaction. Furthermore, when I divided satisfaction by travel type, personal travel had significantly lower customer satisfaction. Higher expectations of airline experience could be one of the reasons for the decrease in satisfaction in both cases.
- *Cat plot (bar and barh)*:
    a) from *Figure II*, the number of passengers who are dissatisfied ranges from 7 to 38 years old and from 61 to 79 years old. On the contrary, the number of passengers who are satisfied outnumber those who are dissatisfied in the 39-60 age group.
    b) Look at *Figure III*, the longer the flight distance for business class travel, the more satisfied passengers will be. The distribution of passengers in the two evaluations is nearly equal for other combinations.
- *Correlation Heatmap* (*Figure IV*): It shows the correlation between numerical variables and target variable. It can be clearly seen that the 'Online boarding' has the highest correlation of 50% with our target 'satisfaction' target variable while the 'Gate location' have a low correlation of 0%

Therefore, I decided to drop 'Gate location' from predictors, and there are 21 features and our target variable left.

Prediction Models Building

After setting up the dummy variables, I divide 70% of the data into training sets and 30% into test sets, and then fit six models with our response variable "Satisfaction".

- Logistic Regression(C=1)
- Decision Tree

- Random Forest (n_estimators=10)
  - I set the number of trees in the forest is 10, which is the minimum number can be changed in n_estimators.
- K-Nearest Neighbors(k=9)
  - To find the optimal number of neighbors (K), I use Stratified 5-cross validation to test training set and make prediction for the test set to assess model accuracy when K is between 1 and 10. Finally, based on the *line chart (Figure V)*, I discovered that perhaps K=9 would be a fair balance between two accuracy metrics, which has the highest accuracy in the test set and the second highest accuracy score in the training set.
- Naïve Bayes
  - Its processing time is the fastest of 6 models.
- Support Vector Machines (C=1, kernel='poly')
  - As the C value increases, more decision points the model considers, so the processing time will be very slow. Although the processing time in the SVM model is faster with C=0.001, the accuracy of cross validation in each kernel is decreasing, so I decided to run it with C=1 by default to achieve higher accuracy. Of 4 kernels, SVM model with polynomial kernel gets the highest cross validation accuracy.

Performance Evaluation

Here I conduct Stratified 5-Fold Cross validation across six classification model to look for their accuracy score and use classification report to see the metrics of model, and then we create a bar plot to see the performance of each model with 5 numbers in ***Figure VI***:

Stratified 5-Fold Cross validation (scoring= "accuracy")

- CV accuracy: The random forest model is identified as the best performing with the accuracy of 95.7%

Classification report metrics: (accuracy, weighted average of precision, recall and f1 score)

- Accuracy: The random forest model has the highest accuracy, with 95.5 percent of predictions being correct, while the Naïve Bayes model just shows a test accuracy of 84.8%.
- Sensitivity/recall: when predicting passengers' comment would "Satisfied," proportion we got the highest correct in the random forest model was 95.5% on weighted average. The second are decision trees model, which was correct 94.4% of the time.
- Precision: of all predicted "Satisfied", proportion was also able to capture 95.5% on weighted average. In the random forest model's classification report, we can see that its specificity has 95%, which means that of all actual "Neutral/Dissatisfied", proportion the model matched was 95%.
- F1 score: since the Sensitivity and Precision has highest number in random forest, F1 score will reflect that also be very high, in this case, it is 95.5%.

In general, all the metrics can more than 80% of correct, so I think the six models perform well in predicting passengers' satisfaction, and random forest performs the best of six models with highest test accuracy. But high recall metric will be very important for our business model. If the model predicts actual Satisfied comments as Neutral/Dissatisfied, it will be extremely costly and result in companies losing money because the airline must invest more business in order to improve Passengers' satisfaction. As a result, false negative (predicting Satisfied when

actual Neutral/Dissatisfied) would be predicted to be smaller or non-existent in this scenario, implying that higher sensitivity with lower false negative error would be preferable. The random forest model would be optimal.

Conclusion

Overall, this highly precise classification model can assist airline companies in determining how to improve passenger satisfaction. For example, we can use this random forest model to investigate the important feature that contribute to passenger satisfaction. As shown in the *horizontal bar plot (Figure VII)*, online boarding and inflight Wi-Fi service are critical for passenger satisfaction, so I recommend that airlines develop better Wi-Fi connection software to make it easier to access on-board Wi-Fi. Furthermore, as previously shown in EDA, disloyal customers and personal travel trends to have more Neutral/Dissatisfied comments, but in this plot, I discovered that business class is also a high important feature, so I believe airlines should also pay attention to the ease of online booking, because business travelers prioritize the convenience of booking a business class.

Appendix

Figure I

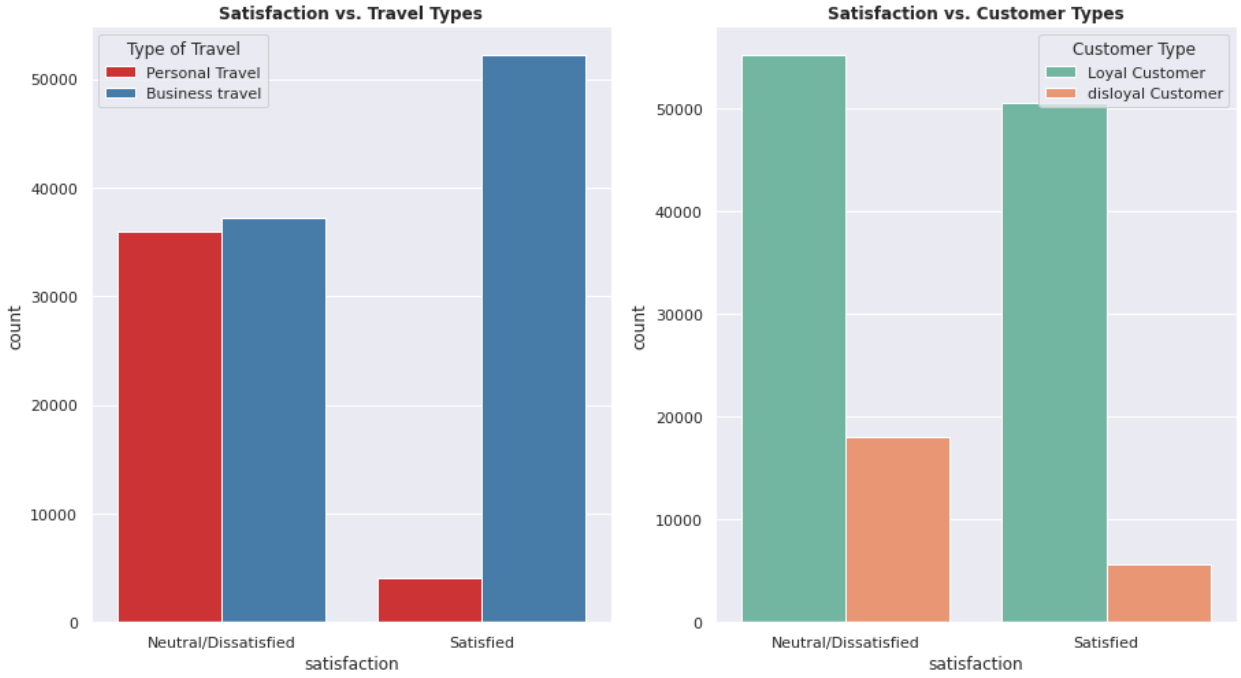


Figure II

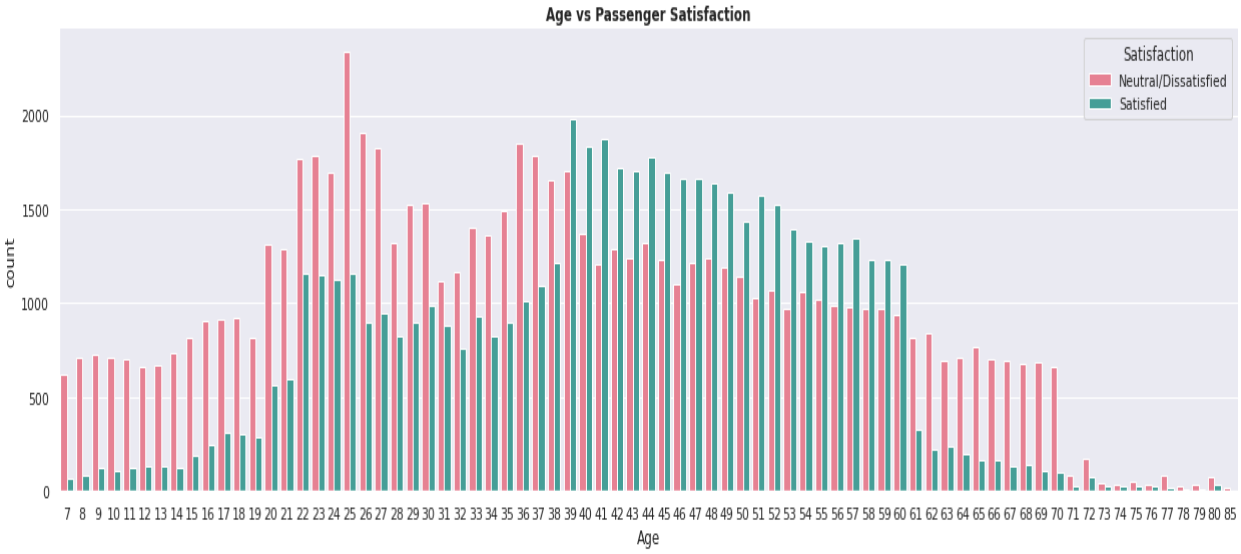


Figure III

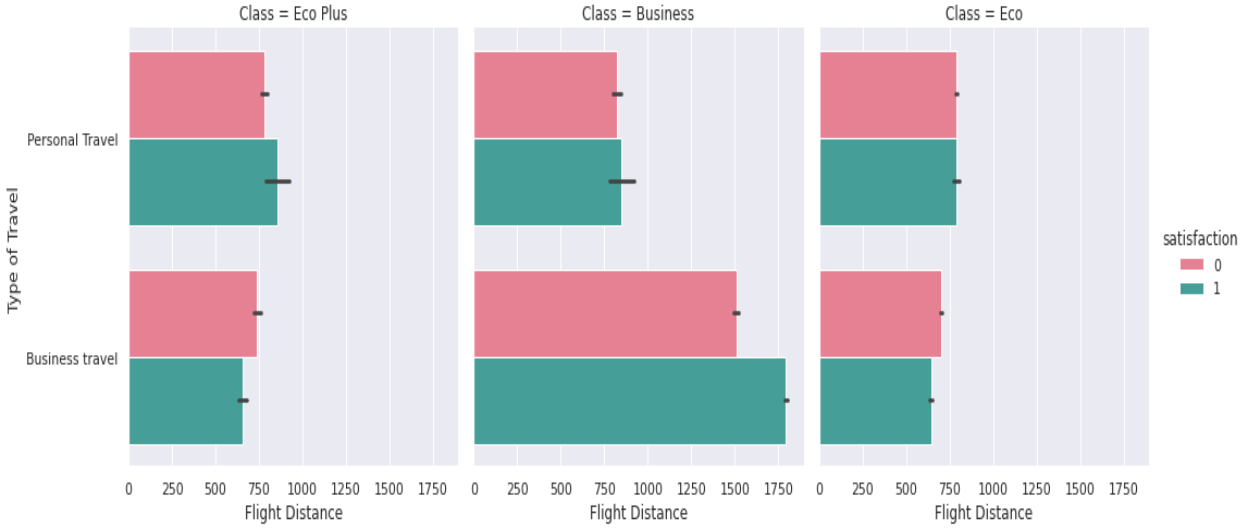


Figure IV

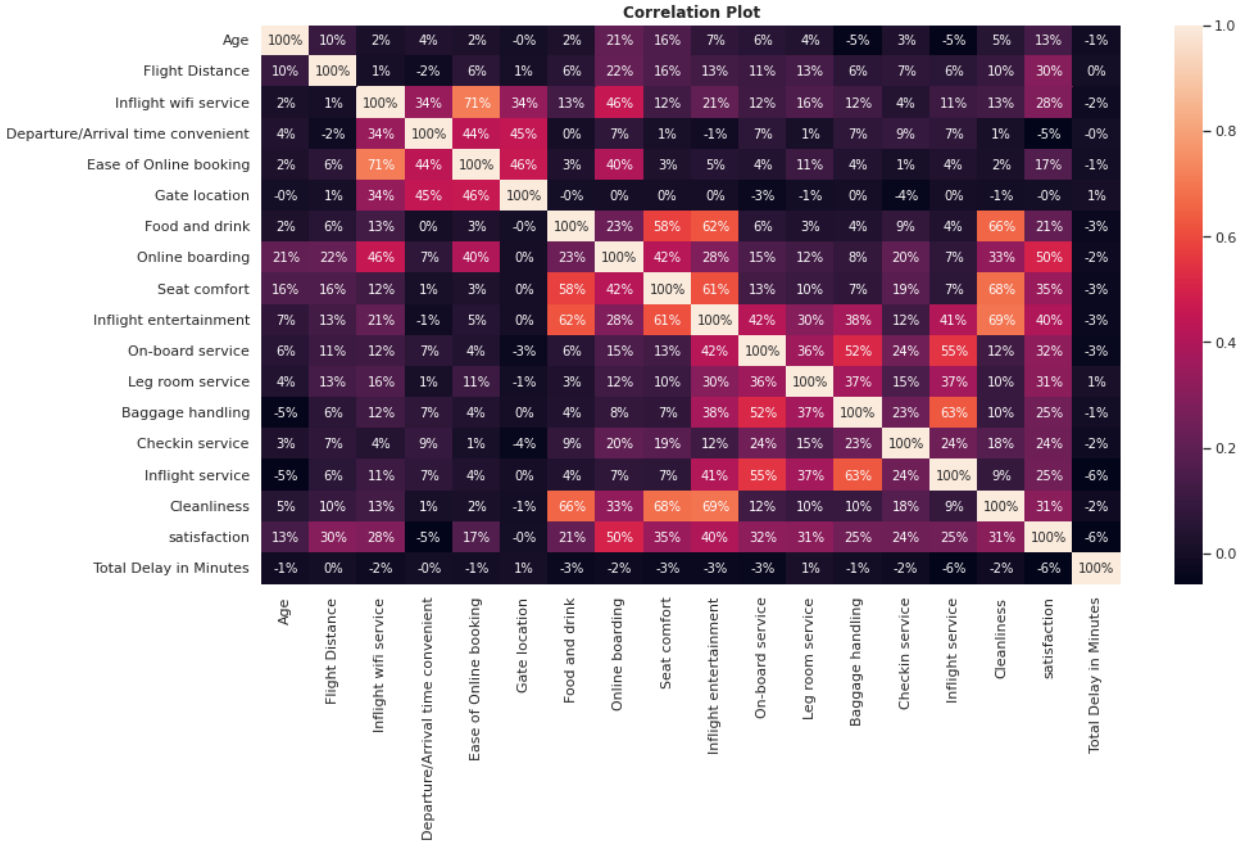


Figure V

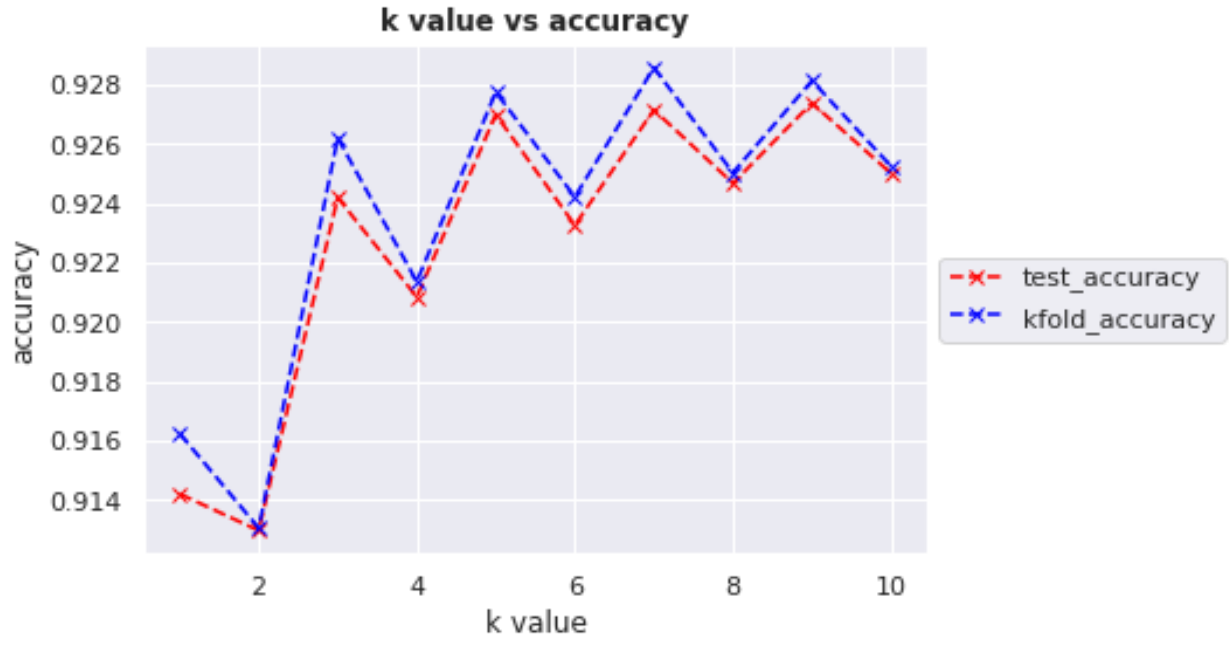


Figure VI

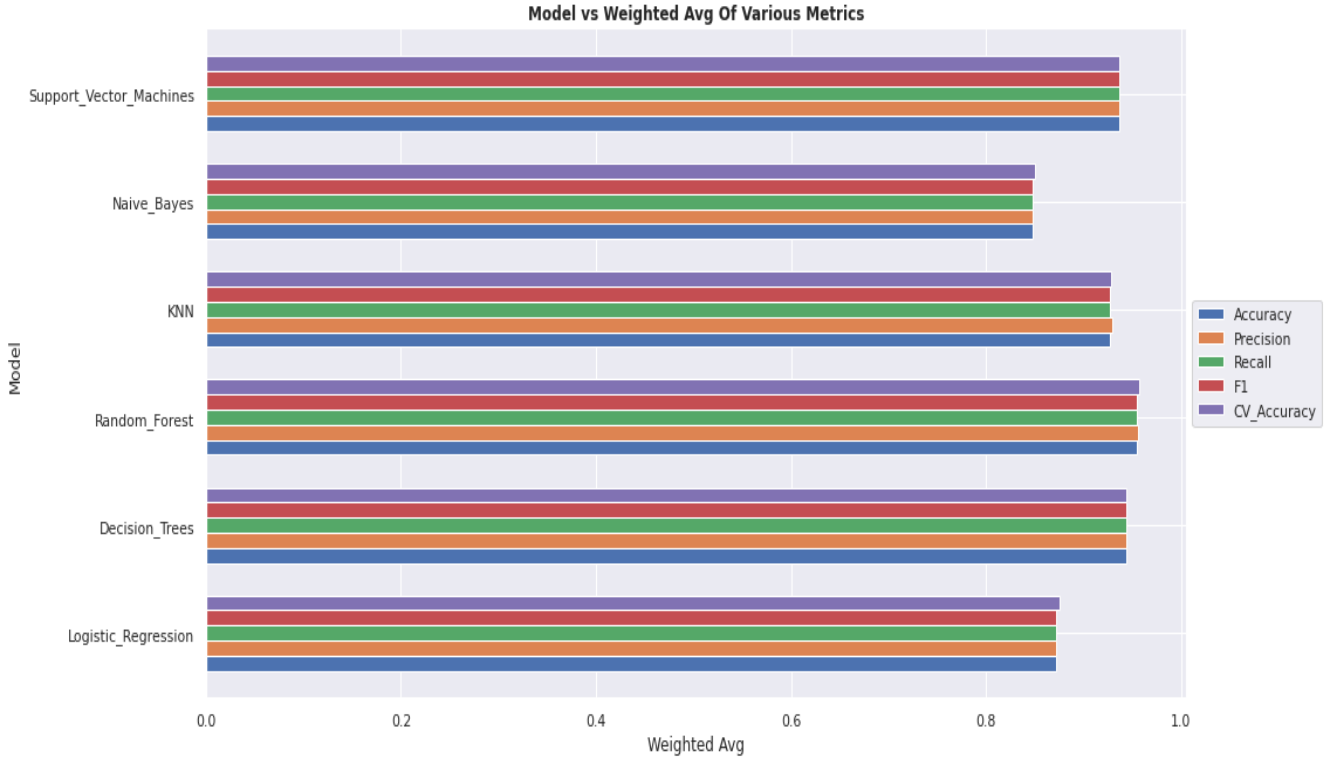
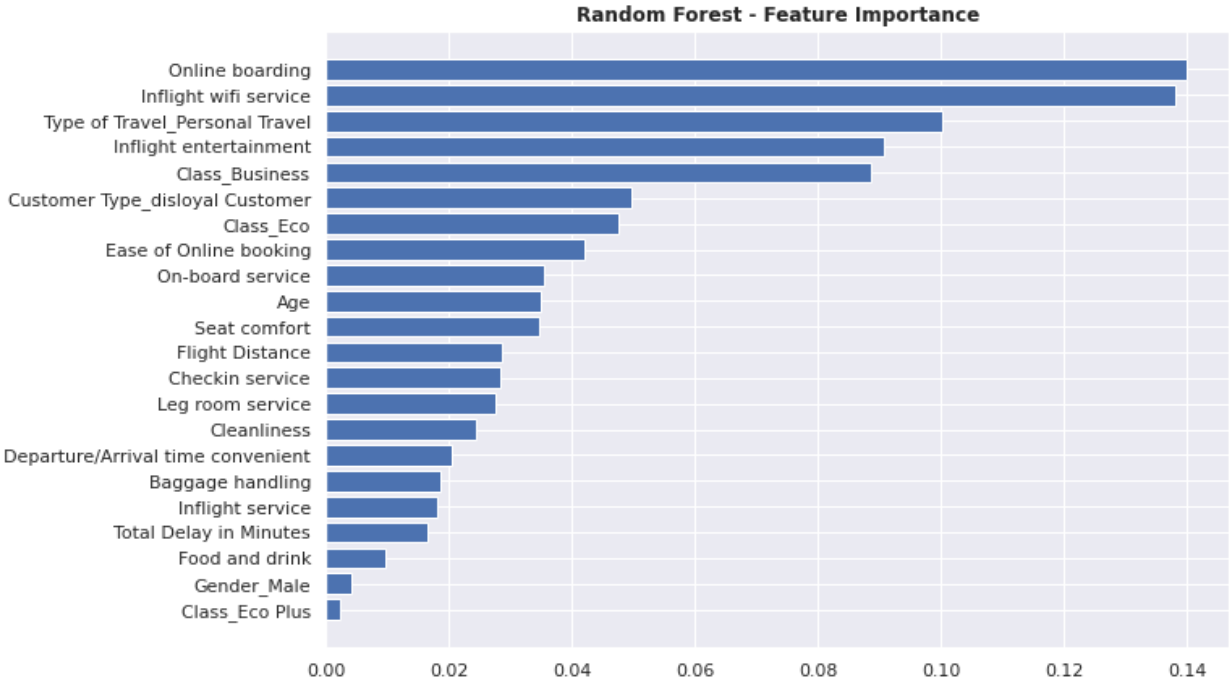


Figure VII


Random Forest - Feature Importance
Online boarding
Inflight wifi service
Type of Travel_Personal Travel
Inflight entertainment
Class_Business
Customer Type_disloyal Customer
Class_Eco
Ease of Online booking
On-board service
Age
Seat comfort
Flight Distance
Checkin service
Leg room service
Cleanliness
Departure/Arrival time convenient
Baggage handling
Inflight service
Total Delay in Minutes
Food and drink
Gender_Male
Class_Eco Plus
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14