Montclair State University

Data Mining for Business Analytics

Stanislav Mamonov, Ph.D

Jaykumar Raichura

Airline Passengers Satisfaction

Our group developed a model that identifies and predicts the satisfaction of passengers using key attributes from the dataset for preliminary purposes by using its demographic data recorded from surveys and feedback. This insight will assist the airline company to improve its customer's experiences, acquire new & retain customers. Enhancing customer's satisfaction will increase the airline company's viability within the aviation industry, ensuring future growth from its competitiveness and reputation. It is important to know the wants and needs of a customer to achieve passenger satisfaction.

Customer satisfaction is a key element for businesses as it serves as a continuing effect of service quality improvement. The final classification algorithms for this research include Logistic Regression, Random Forest Classifier, Decision Tree, KNN and ADA boost.

Dataset

Our dataset is provided by the Kaggle website and consists of airline passengers' experiences while flying and their satisfaction. The dataset contains 25 attributes and a range index of 25,976 entries.

Each variable will be examined to understand the distributed overall satisfaction.

- Satisfaction: Airline satisfaction level (Satisfaction, neutral or dissatisfaction).
- ID: Unique ID number of passenger
- Gender: Gender of the passengers (Female, Male)
- Customer Type: The customer type (Loyal customer, disloyal customer)
- Age: The actual age of the passengers
- Type of Travel: Purpose of the flight of the passengers (Personal Travel, Business Travel)
- Class: Travel class in the plane of the passengers (Business, Eco, Eco Plus)
- Flight distance: The flight distance of this journey
- Inflight Wi-Fi service: Satisfaction level of the inflight Wi-Fi service (1-5)
- Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient
- Ease of Online booking: Satisfaction level of online booking
- Gate location: Satisfaction level of Gate location
- Food and drink: Satisfaction level of Food and drink
- Online boarding: Satisfaction level of online boarding
- Seat comfort: Satisfaction level of Seat comfort
- Inflight entertainment: Satisfaction level of inflight entertainment

- On-board service: Satisfaction level of On-board service
- Leg room service: Satisfaction level of Leg room service
- Baggage handling: Satisfaction level of baggage handling
- Check-in service: Satisfaction level of Check-in service
- Inflight service: Satisfaction level of inflight service
- Cleanliness: Satisfaction level of Cleanliness
- Departure Delay in Minutes: Minutes delayed when departure
- Arrival Delay in Minutes: Minutes delayed when Arrival

Airline Passengers Evaluation Data Analysis & Feature Selection

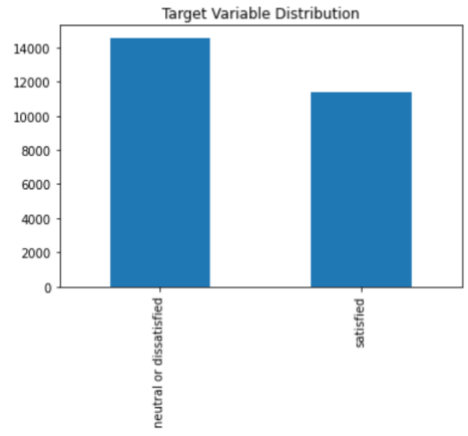
Our target variable, Satisfaction, is our outcome or dependent variable. A binary nominal data type of 1 for satisfied passengers (11,403) and 0 for neutral or dissatisfied passengers (14,573)

- Categorical Variables (Binary & Ordinal): Gender, Customer Type, Type of Travel, Class, Satisfaction, Inflight WIFI service, Departure/Arrival time convenient, Ease of Online booking, Gate location, Food and Drink, On-board service, Seat comfort, Inflight entertainment, Online boarding, Leg room service, Baggage handling, Checking service, Inflight service, Cleanliness.
- Numerical Variables: ID, Unnamed, Age, Flight Distance, Departure Delay in Minutes, Arrival Delay in Minutes.

The dataset is almost equally balanced, see Table 1. Neutral or Dissatisfied represents 56.10% (14,573) of the total data and Satisfied represents 43.89% (11,403) of all data records respectively.

Table 1.

```
neutral or dissatisfied    14573
satisfied                  11403
Name: satisfaction, dtype: int64
```



Data Preprocessing

Data cleaning is a vital role in deriving the output of a machine learning model. Our data preprocessing is the first step in the data mining and data analysis process that takes raw data and transforms it into a format that is simple to understand and analyze.

- Data transforming - Converted text values into numerical format by dividing categorical features into binary and non- binary. For example: Gender, Customer Type, Type of Travel, Class, and Satisfaction.
- Our dataset indicated there were missing data or null values for Arrival Delay in Minutes (83). We imputed the missing values using mean mode.
- Detecting and filtering outliers - proceeded to impute outliers. Our dataset entries decreased from 25,000 to 15,000+, total of entries is 15,324
- Removed features that do not contribute to our predictive modeling. For example, ID and Unnamed columns.

We used pandas for data handling, scikit-learn for performance metrics and matplotlib for visualization. The pandas package offers methods that assisted us summarizing data. The method *mean, std, min, max,* and *median* to learn the characteristics of each variable. Further options like null and sum highlighted the variables that contained null values. Matrix correlations using method *corr* to visualize positive and negative relationships with the target variable.

Exploratory Data Analysis

Data visualization role is an important key in developing and understanding the data before the model implementation. Summarizing the variables, we analyzed their skew and metrics relative to each other. Satisfaction levels are justifiably distributed.

To understand the relationship between the feature variables and the target variable we leverage the data and examine the distribution of satisfaction within the attributes by finding its positive and negative correlation.

Correlation Matrix – shows all pair correlations between variables, these pairs can have positive and negative correlation.

Table 2.

```
satisfaction                      1.000000
Online boarding                   0.494526
Type of Travel                    0.453268
Inflight entertainment            0.398951
Seat comfort                      0.346275
On-board service                  0.321491
Cleanliness                       0.314350
Leg room service                  0.309629
Flight Distance                   0.295292
Inflight wifi service             0.280395
Baggage handling                  0.252415
Inflight service                  0.245573
Checkin service                   0.241495
Food and drink                    0.216868
Customer Type                     0.179632
Ease of Online booking            0.157709
Age                               0.121697
id                                0.010965
Gender                            0.007335
Unnamed: 0                        0.001511
Gate location                    -0.016627
Departure Delay in Minutes       -0.051635
Arrival Delay in Minutes         -0.060960
Departure/Arrival time convenient -0.064798
Class                            -0.443715
Name: satisfaction, dtype: float64
```

Correlated features (positive relationship with our target variable): Online Boarding, Type of Travel, Inflight Entertainment, Seat Comfort, On-board Service, Cleanliness, Leg Room Service, Flight distance, Inflight WIFI Service, Baggage Handling, Inflight Service, Checking Service, Food and Drink, Customer type, Ease of Online Booking, Age, and Gender.

Uncorrelated features (negative relationship with our target variable): Gate Location, Departure Delay in Minutes, Arrival Delay in Minutes, Departure/Arrival Time Convenient, Class.

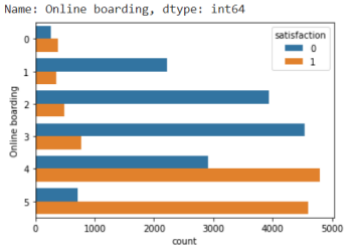
Features and Target Variable

We determined the relative importance of the variables with passengers' satisfaction.

Online Boarding – The satisfaction of the passengers for online boarding are higher on level 4 followed by level 5. We visualize that on level 3 the total amount of unsatisfied passengers is higher.

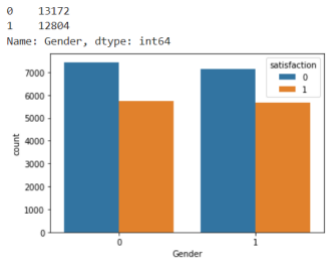
Table 3.

```
4    7706
3    5313
5    5307
2    4429
1    2569
0     652
Name: Online boarding, dtype: int64
```



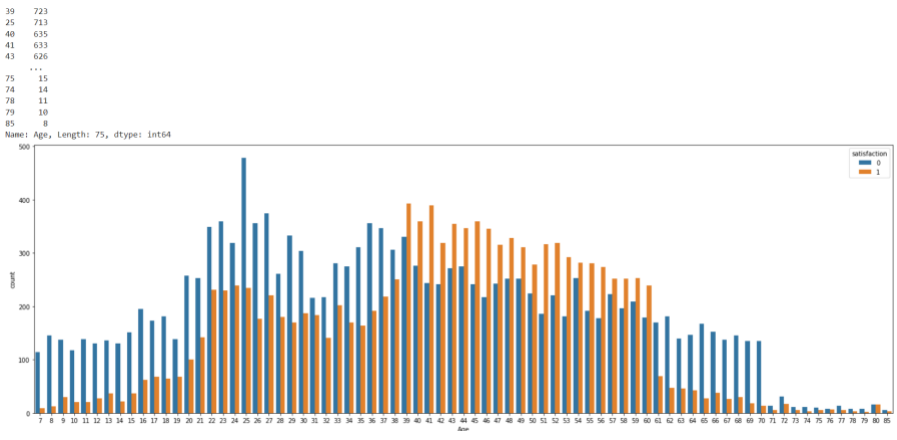
Gender – The satisfaction level of the gender is distributed at the same level. Female (0) and Male (1). The total amount of dissatisfaction are on the higher level in comparison to satisfaction.

Table 4.



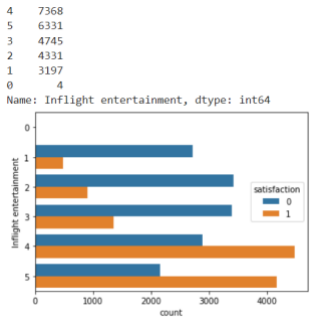
Age – The satisfaction of passengers depending on the passengers age fluctuates. From age 7 to 38 years old and from 61 to 79 years old are for highly dissatisfied passengers at comparison of satisfied passengers within the age range of 39 to 60 years old.

Table 5.



Inflight Entertainment – The satisfaction level on Inflight Entertainment is higher on level 4 followed by level 5.

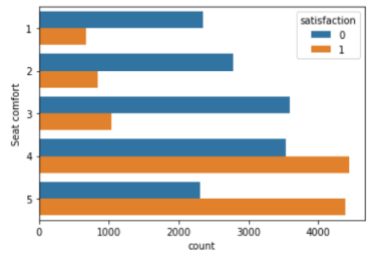
Table 7.



Seat Comfort – The satisfaction passenger level of seat comfort is higher on level 4 followed by level 5.

Table 8.

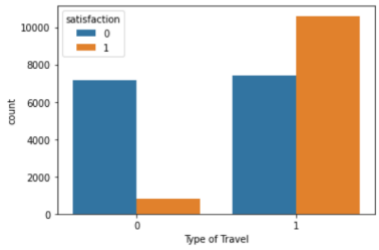
```
4    7991
5    6688
3    4632
2    3632
1    3033
Name: Seat comfort, dtype: int64
```



Type of Travel – The satisfaction level of the type of traveling is higher on Business Travel (1) than Personal Travel (0).

Table 9.

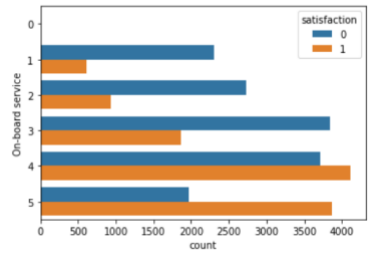
```
1    18038
0     7938
Name: Type of Travel, dtype: int64
```



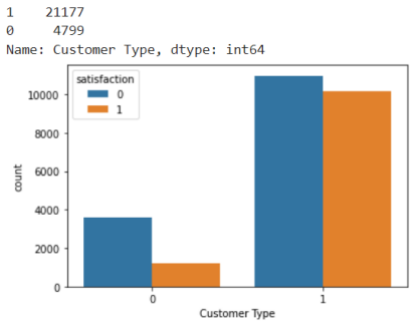
On-board Service – The satisfaction level of On-board Service is higher on level 4, followed by level 5. We can visualize that level 3 the total number of unsatisfied passengers is higher than satisfied passengers.

Table 10.

```
4    7836
5    5844
3    5709
2    3670
1    2915
0       2
Name: On-board service, dtype: int64
```



Customer Type –The satisfaction level on Customer Type – Loyal Customer (1) is significative lesser than unsatisfied customers.



Business Problem

Our report on the Evaluation Data Analysis reveals that level 1, 2 and 3 from categorical variables (ordinal) are often under unsatisfied customers. Their satisfaction is less likely to be fulfilled and seems to be linked to customer type – loyalty. Customer loyalty will increase the possibilities of customer satisfaction. Customer will need a positive experience such as great customer service (on board service), online booking, seat comfort, inflight entertainment, and cleanliness.

Per our Correlation Matrix the highest positive variable 'Online Booking' – airlines should focus on improve a user-friendly customer interface site as business travelers enjoys the comfort and pleasure while flying.

Model Evaluation

Our group will be evaluating the performance of our models: Logistic Regression, Decision Tree Classifier, Random Forest and ADABoost. Finding the best algorithm for predicting the satisfaction of the passengers, predicting the best model, and evaluating with the best recall and precision.

- Accuracy - Tells the accuracy on our predictive model is.
- Recall - It is the ability of a classification model to identify all relevant instances. Tell us how often the model predicts 'Yes'.
- Precision - It is the ability of a classification model to return only relevant instances. Tell us how often our model was correct.
- F1 score - It is a single metric that combines recall and precision using mean. The higher the value of F1 score, the better the model is considered.
- ROC Curve and AUC - The higher the AUC value, the better the model is at distinguishing between 'correct' and 'not correct'. The ideal AUC is equal to 1.

Target Variable: Satisfaction

Independent Variables: Gender, Customer Type, Age, Type of Travel, Class, 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, Checking service, Inflight service, Cleanliness, Departure Delay in minutes, Arrival Delay in minutes.

We proceeded to split, train and test the data. We used sklearn function to split the training data in two datasets, test size 0.7 and random state 101 and test set for evaluation of chosen model.

Logistic Regression- Utilized to forecast a binary outcome and classify observations. The target variable is a binary variable with two classes 0 and 1. The results as follows:

Accuracy	Precision	Recall	F1 score
0.88	0.88	0.89	0.88

Decision Tree - The most well-known classification technique. It's a tree-like model used for classification and regression. The results as follows:

Accuracy	Precision	Recall	F1 score
0.93	0.93	0.92	0.92

Random Forest Classifier - Composes an ensemble through random combinations of the hyperparameters used in a random forest model to optimize the solution for the model. The number of trees are 50,70,100 and 150. The results as follows:

Accuracy	Precision	Recall	F1 score
0.95	0.93	0.96	0.94

KNN (k-nearest neighbors)- Simple classification. For this model we removed the outliers and normalized the dataset using Mix Max scaling and then rescaling the data. The results as follows:

Accuracy	Precision	Recall	F1 score
0.90	0.88	0.93	0.90

Summary

The best model based on ROC AUC is Random Forest with Randomized Search of 99%, followed by Random Forest 99% and KNN 94%. The best model based on accuracy is the Random Forest Classifier with 95%, followed by Decision Tree 93% and KNN with 90%.

Our optimal model is Random Forest with Randomized Search, the accuracy, precision, recall and F1 scores outperformed all other models, and the AUC is close to 1 indicating it's the closest to a perfect model.

Metric	Logistic	kNN	Decision Tree	Random Forest	RandomForest with Randomised Search	AdaBoost
Precision	0.8817	0.8804	0.9304	0.9372	0.9422	0.9450
Recall	0.8905	0.9369	0.9309	0.9636	0.9655	0.9339
F1-Score	0.8861	0.9078	0.9307	0.9502	0.9537	0.9394
AUC	0.94	0.97	0.94	0.99	0.99	0.98

Overall, the customer satisfaction prediction model is proposed using customer evaluation data for the airline through a process of variable selection and through data preprocessing and correlation analysis. As a result, the optimal test AUC of 99% is calculated in the Random Forest with Randomized Search. The possibility of a customer satisfaction prediction model is confirmed.

References

Klein, TJ. (2019) Airline Passenger Satisfaction [Dataset]
https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

McKinney,W. (2021) Python for Data Analysis (Second Edition). O'Reilly