**Report for Flight Passenger Satisfaction**

**PROBLEM STATEMENT**

Primarily objective of data is to predict the satisfaction of passenger for flight service. We will try to use this data to create a model which tries to predict the passenger satisfaction.

We have worked on the following parts:

1. Understanding problem statement.
2. Data Mining.
3. Data Cleansing.
4. Exploratory Data Analysis.
5. Sampling (Train & Test Data).
6. Model Deployment & Predictions.
7. Feature Selections
8. Result based on inputs.

**APPROACH TAKEN**

The data was in the form of (titanic.xls) format. The file contains 11 attributes and the target variable is Survived. Based on 10 independent variables we have to prediction about Survival.

The data is divided in to two files. One is train file and test file, which contains below 24 attributes and the target variable is Satisfaction. Based on following attributes we have to predict about passenger satisfaction

1. id: Unique id number to each passenger.
2. Gender: Gender of the passengers (Female, Male)
3. Customer Type: The customer type (Loyal customer, disloyal customer)
4. Age: The actual age of the passengers
5. Type of Travel: Purpose of the flight of the passengers (Personal Travel, Business Travel)
6. Class: Travel class in the plane of the passengers (Business, Eco, Eco Plus)
7. Flight distance: The flight distance of this journey
8. Inflight wifi service: Satisfaction level of the inflight wifi service (0: Not Applicable;1-5)
9. Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient
10. Ease of Online booking: Satisfaction level of online booking
11. Gate location: Satisfaction level of Gate location
12. Food and drink: Satisfaction level of Food and drink
13. Online boarding: Satisfaction level of online boarding
14. Seat comfort: Satisfaction level of Seat comfort
15. Inflight entertainment: Satisfaction level of inflight entertainment
16. On-board service: Satisfaction level of On-board service
17. Leg room service: Satisfaction level of Leg room service
18. Baggage handling: Satisfaction level of baggage handling
19. Check-in service: Satisfaction level of Check-in service
20. Inflight service: Satisfaction level of inflight service
21. Cleanliness: Satisfaction level of Cleanliness
22. Departure Delay in Minutes: Minutes delayed when departure
23. Arrival Delay in Minutes: Minutes delayed when Arrival
24. Satisfaction: Airline satisfaction level (Satisfaction, neutral or dissatisfaction)

Firstly, we have tried understanding the data, what we found:

1. Dealing with missing data:

The missing values in the data were all focused in the Arrival Delay in Minutes column. We have imputed the nulls in Arrival Delay in Minutes column by use of mean value. We have dropped the unnamed:0 columns where the values are 0, since there are numerous activities in the data attributed to these might act as an outlier. Similarly, we have dropped id column, as id is unique for each passenger.

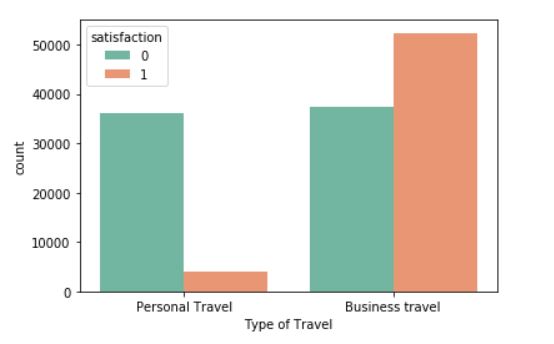
1. Data Cleaning:

Firstly, we have merged both the files for better results and understanding. Then we have performed cleaning of data which contains variable transformation. We have converted non-numeric columns into numeric for ease in understanding.

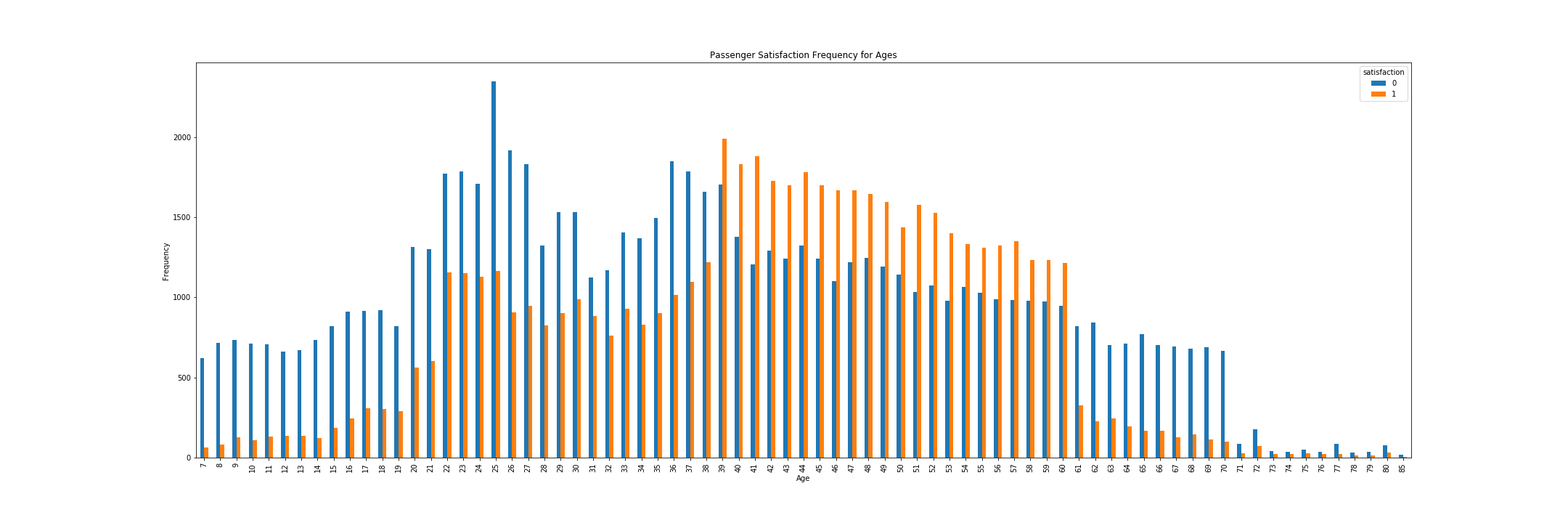
1. Exploratory Data Analysis:

After cleaning of the data, we try to get insights by using visualization techniques with the help seaborn and matplotlib library. Some of insights are as follows:

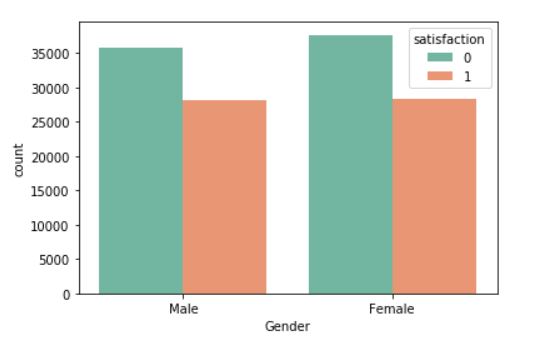
* Type of Travel and Satisfaction graph:



* + Business travel is generally satisfied with the flight but personal travel is generally neutral or dissatisfied with the flight.
  + Personal travel is %90 neutral or dissatisfied with the flight. This statistic is interesting.
* Age and Satisfaction Graph:



* + 0-30 years old is generally neutral or dissatisfied with the flight.
  + 30-60 years old is generally satisfied with the flight.
  + 65-85 years old passengers is neutral or dissatisfied.
* Gender and Target variable graph:



* + Females and males are satisfaction probability almost equal. Male %44, Female %43 satisfaction

1. Sampling:

After performing visualization over the cleansed data. We have split the given data into two different parts with the help of scikit-learn library. The ratio for split criteria is 70:30. Out of 100% data, 70% of data goes in training of model and 30% data goes for testing of predictions.

1. Model Building & Predictions:

We have implemented various machine learning algorithms such as Logistic Regression, Decision Tree and Adaptive Boosting in order to identify the most optimal model. For each of these models the Confusion Matrix, Accuracy, Sensitivity, Specificity and F1score were calculated. Further we found importance and results based on the models we have used.

1. Feature Selection:

Lastly, for feature selection, we have used various feature selection techniques to understand the significant attributes from the data. We have used Decision Tree and Recursive Feature Elimination method to choose the most important attributes of the data.

1. Interpretation of Results:

The results for different algorithms were inferred from the data were:

**INTERPRETATION OF RESULTS**

**LOGISTIC REGRESSION: -**

**Results for Logistic Regression**

* Confusion Matrix:
  + The matrix represents 19631 true positive values which means that all of these records were correctly predicted as leading to a Confirmation.
  + 14168 are true negative values which means that all of these records were correctly predicted as not leading to a Confirmation.
  + 2817 are false positive values which means that all of these records were predicted as leading to a Confirmation but did not actually result in a Confirmation.
  + 2348 are false negative values which means that all of these records were predicted as not leading to a Confirmation but actually result in a Confirmation.
  + Accuracy: The accuracy 86.74 % is used as an appropriate representation of the model.
  + Precision: The score 0.8931 implies that if a model predicts a class for the record as confirmed there is an 89.31% likelihood that the model is correct.
  + Recall: The score 0.8745 implies that the model can recall 87.45 % of instances of a particular class.

**Confusion Matrix for Logistic Regression: -**

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 19631 | 2817 |
| 1 | 2348 | 14168 |

**Sensitivity: 87.45%**

**Specificity: 85.78%**

**False Positive Ratio: 14.21%**

**Precision: 89.31%**

**Overall Accuracy: 78.07%**

**DECISION TREE: -**

**Results for Decision Tree:**

* Confusion Matrix:
  + The matrix represents 20830 true positive values which means that all of these records were correctly predicted as leading to a Confirmation.
  + 15849 are true negative values which means that all of these records were correctly predicted as not leading to a Confirmation.
  + 1136 are false positive values which means that all of these records were predicted as leading to a Confirmation but did not actually result in a Confirmation.
  + 1149 are false negative values which means that all of these records were predicted as not leading to a Confirmation but actually result in a Confirmation.
  + Accuracy: The accuracy 94.13% is used as an appropriate representation of the model.
  + Precision: The score 0.9477 implies that if a model predicts a class for the record as confirmed there is an 94.78% likelihood that the model is correct.
  + Recall: The score 0.9482 implies that the model can recall 94.82% of instances of a particular class.

**Confusion Matrix for Decision Tree: -**

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 20830 | 1136 |
| 1 | 1149 | 15849 |

**Sensitivity: 94.82%**

**Specificity: 93.24%**

**False Positive Ratio: 6.75%**

**Precision: 94.77%**

**Overall Accuracy: 94.13%**

**ADAPTIVE BOOSTING: -**

**Results for Adaptive Boosting:**

* Confusion Matrix:
  + The matrix represents 21014 true positive values which means that all of these records were correctly predicted as leading to a Confirmation.
  + 15734 are true negative values which means that all of these records were correctly predicted as not leading to a Confirmation.
  + 1251 are false positive values which means that all of these records were predicted as leading to a Confirmation but did not actually result in a Confirmation.
  + 965 are false negative values which means that all of these records were predicted as not leading to a Confirmation but actually result in a Confirmation.
  + Accuracy: The accuracy 94.31 % is used as an appropriate representation of the model.
  + Precision: The score 0.9560 implies that if a model predicts a class for the record as confirmed there is an 95.60% likelihood that the model is correct.
  + Recall: The score 0.9438 implies that the model can recall 94.68% of instances of a particular class.

**Confusion Matrix for Adaptive Boosting: -**

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 21014 | 1251 |
| 1 | 965 | 15734 |

**Sensitivity: 94.38%**

**Specificity: 94.22%**

**False Positive Ratio: 5.77 %**

**Precision: 95.60%**

**Overall Accuracy: 94.31%**

**Conclusion:**

Comparing the accuracy of the different models, Decision Tree is the best among the classification models. Recursive Feature Elimination Method highlights the importance of predictor’s online boarding, Inflight wifi service, Type of Travel, Inflight Entertainment, Age, Gender, and Customer Type. After analyzing all the models we can conclude that predictor’s Online boarding, Inflight wifi service, Type of Travel did played a major role for Flight Passenger Satisfaction

We can conclude that with the help of voting classifier our model estimates 95.39% correctly. Our model works well. We can conclude the results generated by Decision Tree are best compare to other models.