**Project Report: Airline Passenger Satisfaction**

# **Business Problem, Motivation and Setting**

One of the most important factors influencing the revenue of an airline is passenger satisfaction. For the airline companies, it has become increasingly difficult to retain loyal/repeating passengers. In this academic project, we worked on a business case of an Airline by applying business analytics models to mine the data retrieved from the database. Our aim is to identify the customer satisfaction level based on their rating on various aspects of the airline experience so that we can try to improve the satisfaction level of the passengers to retain them for long term.

# **Data Description, Data Processing and Feature Engineering**

Our dataset contains information about the Airline passengers rating who have traveled with the airline. Among all the travelers, some passengers end up rating the airline in its different aspects, while others just travel and leave no reviews. Using this dataset in our project, we aim to create a prediction model which could identify the user's satisfaction and improve it for the next time.

Customer satisfaction scores are obtained from around 130k airline passengers, including additional information about each passenger, their flight, and type of travel, as well as their evaluation of different factors like cleanliness, comfort, service, and overall experience. The data represents airline satisfaction data consisting of 129,880 observations and 24 attributes.

We first perform exploratory data analysis to find out the underlying data structure in the raw data and then perform data cleaning and manipulation by converting categorical variables by performing one-hot encoding and replacing the missing values and removing duplications and reducing the dimension of the data. Once we obtain the clean data, we can input this data into the machine learning models so that we get higher chances of predicting the right output values to the real-world future scenarios based on past data and offer recommendations to the business at the end of the project.

After making sure that there are no duplicate instances of data, we found missing values in the ‘Arrival Delay’ column. We imputed the missing values with the mean of the column. The following is the table depicting the variables used in our analysis and their description:

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Variable** | **Description** |
| 1 | ID | Unique passenger identifier |
| 2 | Gender | Gender of the passenger (Female/Male) |
| 3 | Age | Age of the passenger |
| 4 | Customer Type | Type of airline customer (First-time/Returning) |
| 5 | Type of Travel | Purpose of the flight (Business/Personal) |
| 6 | Class | Travel class in the airplane for the passenger seat |
| 7 | Flight Distance | Flight distance in miles |
| 8 | Departure Delay | Flight departure delay in minutes |
| 9 | Arrival Delay | Flight arrival delay in minutes |
| 10 | Departure and Arrival Time Convenience | Satisfaction level with the convenience of the flight departure and  arrival times |
| 11 | Ease of Online Booking | Satisfaction level with the online booking experience |
| 12 | Check-in Service | Satisfaction level with the check-in service |
| 13 | Online Boarding | Satisfaction level with the online boarding |
| 14 | Gate Location | Satisfaction level with the gate |
| 15 | On-board Service | Satisfaction level with the on-boarding service |
| 16 | Seat Comfort | Satisfaction level with the comfort of the airplane |
| 17 | Leg Room Service | Satisfaction level with the leg room of the airplane seat |
| 18 | Cleanliness | Satisfaction level with the cleanliness of the airplane |
| 19 | Food and Drink | Satisfaction level with the food and drinks on the airplane |
| 20 | In-flight Service | Satisfaction level with the in-flight service |
| 21 | In-flight Wi-Fi Service | Satisfaction level with the in-flight Wi-Fi |
| 22 | In-flight Entertainment | Satisfaction level with the in-flight entertainment |
| 23 | Baggage Handling | Satisfaction level with the baggage handling |
| 24 | Satisfaction | Overall satisfaction level with the airline |

# **Data Visualization**

To broaden our understanding of the dataset, we made use of the ‘matplotlib’ library and visualized the behavior of variables as follows:Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

To understand the correlation between the variables, we made use of the heat map.

Chart

Description automatically generated

The variables ‘Departure Delay’ and ‘Arrival Delay’ were found to be highly correlated. Hence, we dropped Arrival Delay from our analysis and proceeded further to mitigate endogeneity.

**Machine Learning Models**

After data processing, we went further in our analysis to observe which factor or combination of factors could determine the overall satisfaction of the customer. We applied multiple machine learning models to predict satisfaction, in such a business case we used – Logistic Regression, Decision tree, Bagging, and Radom Forest Algorithms.

**Model Building & Model Evaluation**

The performance of the above-mentioned models is determined by evaluating the Confusion matrix and the Area under the curve of ROC.

# **Logit (Logistic Regression) Model**

We built a logistic regression model to know the basic classification of what the model can perform without any hyperparameters and listed the output of the model below. From the results we can say that this model seems to be performing a decent job without overfitting the data in hand along with a good F1 score.

**Chart

Description automatically generatedConfusion matrix and AUC:**

**Chart, treemap chart

Description automatically generated**

1) Test Accuracy = 0.8533004824966636

2) Train Accuracy = 0.854426063619165

3) Precision Score = 0.8137464788732395

4) Recall score: 0.8570072386377121

5) Accuracy score: 0.8533004824966636

6) F1 score: 0.8348167841867993

# **Decision Tree Model**

We wanted to test how different classification models perform on the same data and built the Decision Tree model using the entropy as the impurity measure since the variables are of the categorical measure. The results obtained from this basic Decision Tree model was matching with one of the draw backs of this model where the model was overfitting the data. To generalize this model we used a range of hyperparameters namely, max\_depth to restrict the tree growth to a max depth of 8 using range (1,8) and max\_leaf\_nodes to restrict the maximum number of features that are taken into the account for splitting each node to 8 using range (2,8) and minimum number of samples to 8 using range (2,8). After building a model using the grid search method, we found that the overfitting was reduced and model was more generalized to the data for the best hyper tuning parameters of max\_depth: 4, max\_leaf\_nodes: 7, min\_samples\_split: 2. The results obtain are as below.

**Confusion matrix and AUC:**

Chart

Description automatically generated

Initial output of the Decision Tree -

1) Test Accuracy = 0.9281387947849297

2) Train Accuracy = 1.0

To address the overfitting issue, we use the gird search method, and the following results are as below:

1) Test Accuracy = 0.870059542141464

2) Train Accuracy = 0.8714197720973206

3) Precision Score = 0.8950083752093803

4) Recall score: 0.7925714963806811

5) Accuracy score: 0.870059542141464

6) F1 score: 0.8406809528304856

# **Decision Tree with Bagging**

After obtaining the results from the decision tree we wanted to improve the accuracy of the model without overfitting, hence we choose an Ensemble method with includes a decision tree with bagging concept. The above model there could be a probability of overfitting, but in this case, we use 200 fully grown Decision trees as this is a minimum number to use when we are using a random forest or an Ensemble method to get good performance measures, each tree has a sample size of 100 records but with different data records with replacement. The output obtain from each of these trees are combined and are evaluated based on the majority voting in this case to provide one single output. The performance measure of the model is as below.

**Confusion matrix and AUC:**

Chart

Description automatically generatedChart

Description automatically generated

1) Test Accuracy = 0.8953136228313315

2) Train Accuracy = 0.8967838444278234

3) Precision Score = 0.9040420013916124

4) Recall score: 0.8479886080455679

5) Accuracy score: 0.8953136228313315

6) F1 score: 0.875118635765239

# **Random Forest**

Finally, we wanted used a random forest method with parameters of 200 Decision trees and each tree with a sample size of 100 with different records for each tree. The output obtained from each of these trees are combined and are evaluated based on the majority voting in this case to provide one single output. With the rest of the hyperparameters set to their default values, we observed that this model is performing well. The results are obtained as below.

**Confusion matrix and AUC:**

Chart

Description automatically generatedChart, treemap chart

Description automatically generated

1) Test Accuracy = 0.898265065188379

2) Train Accuracy = 0.8991266663733556

3) Precision Score = 0.9199244201198853

4) Recall score: 0.8377239824374035

5) Accuracy score: 0.898265065188379

6) F1 score: 0.8769020557729332

7) AUC score: 0.9569954699677659

# **Conclusion**

Of all the models, Random Forest model was chosen as the best performing model based on the AUC-ROC. The importance of the top ten variables in calculating whether the passenger is satisfied or not are as follows:

Graphical user interface, application

Description automatically generated

By managing/improving these factors, the airline can improve the overall satisfaction level of a passenger.

# **References**

* <https://www.mavenanalytics.io/data-playground>
* Software we used is <https://jupyter.org/>.