**Airline Passenger Satisfaction Prediction**

Final Report

Team number – 11 (BUAN 6356.004)

Dhruv Rajesh Shah

Prathamesh Prashant Nagraj

Sahithi Venkata Darisi

Shreya Madan Mali

**Contents:**

1. Introduction 3

1.1. Project Motivation 3

1.2. Objective 3

2. Data Source 3

2.1. Data Description 3

3. Exploratory Data Analysis 4

3.1. Customer Satisfaction Distribution 4

3.2. Gender V/S Customer Satisfaction 5

3.3. Customer Type V/S Customer Satisfaction 5

3.4. Travel V/S Customer Satisfaction 6

3.5. Age V/S Customer Satisfaction 7

3.6. Flight Distance V/S Customer Satisfaction 7

4. Data Pre-Processing 8

5. Data Modelling 8

4.1. Model 1 – Logistic Regression 8

4.2. Model 2 – Decision Tree Analysis 13

4.3. Model 3 – Random Forests 17

4.4 Model 4 – Neural Networks 19

6. Model Comparison & Conclusion 21

* **Project Motivation :**

The airlines industry is super competitive and, it works purely on customer satisfaction and customer loyalty. As this is a dynamic industry, understanding and predicting the customer satisfaction becomes crucial. In this project, we have picked up a dataset tried to implement some Business Intelligence techniques to predict the customer satisfaction and how the experience of a customer can be improved to enhance the overall customer satisfaction.

* **Objective :**

The project focusses on predicting the satisfaction of customers with their air travels for a particular airline. The primary objective of this project is to develop a predictive model for Airline Passenger Satisfaction. Leveraging key attributes such as gender, customer type, age, travel purpose, class, flight distance, and satisfaction ratings across various services, the goal is to classify passengers into satisfied (1) or neutral/dissatisfied (0). The model aims to provide airlines with insights for service enhancements, operational efficiency, and strategic decision-making in the highly competitive airline industry. The satisfaction prediction will be based on a scale of 0-5, where 0 represents the least satisfaction and 5 represents the most satisfaction. The model will be trained and validated using a dataset that has undergone processing, resulting in reduced records for more effective analysis.

* **Data Source :**

We shall be using a second-hand dataset from **Kaggle**: **Airline Passenger Satisfaction**

Dataset link: <https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction?select=train.csv>

* **Data Description :**

At a macro level, this Kaggle-sourced dataset encompasses diverse attributes covering passenger demographics, travel context, flight details, satisfaction ratings across various facets of the travel experience, operational information, customer loyalty categorization, and an overall satisfaction metric. These dimensions collectively offer a comprehensive view of factors influencing passenger contentment and dissatisfaction, enabling airlines to glean insights for service enhancements, operational efficiency, and strategic decision-making in the highly competitive airline industry.

The key columns on which the classification will be made are :

1. **Gender**: Gender of the passengers (Female, Male)
2. **Customer Type**: The customer type (Loyal customer, disloyal customer)
3. **Age**: The actual age of the passengers
4. **Type of Travel**: Purpose of the flight of the passengers (Personal Travel, Business Travel)
5. **Class**: Travel class in the plane of the passengers (Business, Eco, Eco Plus)
6. **Flight distance**: The flight distance of this journey
7. **Inflight wifi service**: Satisfaction level of the inflight wifi service
8. **Departure/Arrival time convenient**: Satisfaction level of Departure/Arrival time convenient
9. **Ease of Online booking**: Satisfaction level of online booking
10. **Gate location**: Satisfaction level of Gate location
11. **Food and drink**: Satisfaction level of Food and drink
12. **Online boarding**: Satisfaction level of online boarding
13. **Seat comfort**: Satisfaction level of Seat comfort
14. **Inflight entertainment**: Satisfaction level of inflight entertainment
15. **On-board service**: Satisfaction level of On-board service
16. **Leg room service**: Satisfaction level of Leg room service
17. **Baggage handling**: Satisfaction level of baggage handling
18. **Check-in service**: Satisfaction level of Check-in service
19. **Inflight service**: Satisfaction level of inflight service
20. **Cleanliness**: Satisfaction level of Cleanliness
21. **Departure Delay in Minutes**: Minutes delayed during departure
22. **Arrival Delay in Minutes**: Minutes delayed during Arrival
23. **Satisfaction**: Airline satisfaction level(Satisfaction, neutral or dissatisfaction) – **Target Variable**

The total number of records initially were 103594, post processing the data the number of records have been reduced to 30,000 and 10,000 in Training and Validation datasets respectively, for convenience.

NOTE : Satisfaction levels of input variables are on a scale of 0-5 (0: least satisfied, 5: most satisfied) and the satisfaction levels of target variable is either 1 or 0 (1: satisfied and 0: neutral or dissatisfied). Similarly, many other categorical variables such as gender have been encoded for modelling.

* **Exploratory Data Analysis :**

**Customer Satisfaction Distribution:**

A blue squares with black text

Description automatically generated

The graph shows that there is a higher number of people who are neutral or dissatisfied with their satisfaction level than those who are satisfied. The difference is approximately 10,000 people.

**Gender vs Satisfaction:**

A red and green squares

Description automatically generated

There are more males than females in the “neutral or dissatisfied” category, and more females than males in the “satisfied” category.

**Customer Type vs Satisfaction:**

**A green and red rectangular shapes

Description automatically generated**

There are significantly more satisfied loyal customers than satisfied disloyal customers. Conversely, there are significantly more neutral or dissatisfied disloyal customers than neutral or dissatisfied loyal customers.

**Travel vs Satisfaction**

**A red and green squares

Description automatically generated**

**A close-up of several colored squares

Description automatically generated**

Travel Class Distribution: The highest satisfaction levels are seen in the Business class, followed by Eco Plus and then Eco.

The lowest satisfaction levels are seen in the Eco class. The satisfaction levels in the Business class are almost double that of the Eco class.

**Age vs Satisfaction**

A diagram of a graph

Description automatically generated with medium confidence

The median age for neutral or dissatisfied is higher than the median age for satisfied. The age range for neutral or dissatisfied is larger than the age range for satisfied.

**Flight Distance vs Satisfaction**

**A graph showing a diagram

Description automatically generated with medium confidence**

The graph indicates that satisfied customers tend to have longer flight distances compared to those who are neutral or dissatisfied.

* **Data Pre-processing:**
  + In the downloaded dataset, we removed the 1st and 2nd columns that represented the Sr. No. and IDs respectively as they do not play any role in our analyses.
  + Next, we went on to find any missing values present in the parent dataset and removed them consequently.
  + Similarly, we also checked for duplication of records but found none.
* **Data Modelling :**

1. **Logistic regression**
2. **Decision tree**
3. **Random forest**
4. **Neural network**

**Logistic regression :**

We have used the Logistic regression model to predict customer satisfaction since it is categorical.

A screenshot of a computer

Description automatically generated

A computer screen with blue text

Description automatically generated

Checking the accuracy of the model on validation data :

A screenshot of a computer

Description automatically generated

The accuracy of the model on validation data is 86.96%

Model Selection for full logistic regression :

A computer screen shot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

Checking the accuracy of the model on validation data for backward full logistic regression :

A screenshot of a computer

Description automatically generated

The accuracy of the model on validation data is 86.92%

Inference :

* Key Predictors:

1. Online Boarding:

Coefficient: 0.83

Interpretation: A one-unit increase in the "Online Boarding" variable is associated with an increase of approximately 0.83 in the log-odds of satisfaction, holding other variables constant.

Explanation: When passengers use online boarding services, their chances of being satisfied with their overall experience increase.

Numeric: Passengers using the online boarding facilities are 2.29 times more likely to be satisfied with their air travel.

1. Cleanliness:

Coefficient: 0.2967

Interpretation: A one-unit increase in the "Cleanliness" rating is associated with an increase of approximately 0.2967 in the log-odds of satisfaction, holding other variables constant.

Explanation: The cleaner the airline, the more likely passengers are to be satisfied. A higher cleanliness rating is linked to happier travelers.

Numeric: For cleaner flights, the passengers are 1.34 times more likely to be satisfied.

1. Inflight WiFi Service:

Coefficient: 0.502

Interpretation: Passengers satisfied with inflight WiFi service have approximately 0.50200 times higher odds of overall satisfaction compared to those not satisfied, holding other variables constant.

Explanation: If passengers enjoy the WiFi service during the flight, they are more likely to have an overall positive experience.

Numeric: Passengers are 1.65 times more likely to be satisfied if provided with wifi service.

**Negative Impact Predictors:**

1. Departure Arrival Time Convenience:

Coefficient: - 0.20925

Interpretation: Passengers dissatisfied with departure and arrival time convenience have approximately

0.20925 times lower odds of satisfaction, holding other variables constant.

Numeric: If the Departure/Arrival time is not convenient then passengers are 1.23 times likely to be dissatisfied.

1. Arrival Delay in Minutes:

Coefficient: - 0.38197

Interpretation: For each additional minute of arrival delay, the odds of satisfaction decrease by approximately 0.38197 times, holding other variables constant.

Numeric: If there are arrival delays then passengers are 1.5 times likely to be dissatisfied.

**Decision Tree Analysis :**

Decision tree is a data mining technique that is popularly used for classification and prediction. The tree structure is constructed through a recursive partitioning process, where each internal node represents a decision based on a specific feature, each branch corresponds to a possible outcome of that decision, and each leaf node provides the final classification or regression result.

Default Tree :

A diagram of a tree

Description automatically generated

A screenshot of a computer program

Description automatically generated

The accuracy of the model on validation data is 86.94%

Default tree using Entropy for Splitting :

A diagram of a tree

Description automatically generated

A screenshot of a computer

Description automatically generated

The accuracy of the model on validation data is 86.82% using Entropy.

Deepest Tree:

A black and white image of a grid

Description automatically generated

A screenshot of a computer

Description automatically generated

The accuracy of the model on training data is 100% for deepest tree.

A screenshot of a computer

Description automatically generated

The accuracy of the model on validation data is 91.4% for deepest tree.

We need to prune the tree to get the desired level of accuracy as there is an overfitting problem here. The deepest tree on the training dataset yields 100% accuracy while on validation dataset gives 91.4% accuracy.

Post-pruning :

Finding the point where misclassification rate in the validation dataset is the minimum. At nsplit 149 our tree will show the lowest misclassification rate in validation dataset as per cp table, but the tree will be not visible for explanation or interpretations. Hence, we need to reduce the number of splits but keep them more than the default tree for better accuracy.

A screenshot of a computer program

Description automatically generatedA table of numbers with numbers on it

Description automatically generated

Hence, we choose the nsplit as 15.

A diagram of a network

Description automatically generated

A screenshot of a computer

Description automatically generated

The accuracy of the model on validation data is 88.19% for pruned tree.

Hence, we see that the pruned model has the highest accuracy compared to default tree.

**Random Forests :**

A screenshot of a computer

Description automatically generated

A graph with lines and words

Description automatically generated with medium confidence

Inference :

* The variables mentioned above are the most important feature of the dataset for the pruned tree and default tree. Age, Flight distance, and satisfaction levels for various services (Inflight wifi service, Departure/Arrival time convenient, Ease of Online booking, etc.) are likely to be important factors.
* Comparing the testing accuracy of the four models (default tree, deepest tree, pruned tree, Applying Random Forest), the validation accuracy of the random forest is the highest with a value of 94.37. Hence, we can conclude that the random forest is performing better than the other two models.
* Comparing the Sensitivity and Specificity of the three models (default tree, deepest tree, pruned tree, Applying Random Forest).
* The number of true positive records and true negatives according to the confusion matrix is 3649 and 5046 respectively.
* The number of false positives and false negatives according to the confusion matrix is 555 and 750 respectively.

**Neural Networks :**

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* The neural network uses parallel information processing to extract meaningful information
* The left most nodes (i.e., input nodes) are raw data variables
* The subsequent layers, positioned between the input and output layers, constitute the hidden layers, there is one hidden layer with three neurons (hidden = 3), in the neural network model.
* Neurons apply activation functions (e.g., logistic) to introduce non-linearity for capturing intricate relationships.
* It took 18604 steps to design a neural network and with a error of 1162.212035.
* As per Neural Networks' prediction 40% passengers are satisfied and 60% are dissatisfied but in reality, 44% are satisfied and 56% are dissatisfied.
* Accuracy of Neural Network Model is 89.58%.

**Model Comparison :**

|  |  |  |
| --- | --- | --- |
| **Data Mining Technique** | **Accuracy** | **Error** |
| Decision Tree (Gini index) | 86.94 | 13.06 |
| Decision Tree (Information Gain) | 86.82 | 13.18 |
| Decision Tree (Deeper) | 91.4 | 8.6 |
| Decision Tree (After Pruning) | 88.19 | 11.81 |
| Random Forest | 94.37 | 5.63 |
| Logistic Regression | 86.96 | 13.04 |
| Neural Networks | 89.58 | 10.42 |

**Conclusion :**

* The primary objective of the dataset is to ascertain passenger satisfaction. By using the models developed, we employ these models on the test dataset to check for accuracy. Once the models are constructed using the available data, they can be utilized for forecasting passenger satisfaction.
* The logistic regression model not only provides accurate predictions but also offers actionable insights for strategic decision-making.
* Airlines can allocate resources more effectively by prioritizing areas such as online boarding to maximize customer satisfaction. In conclusion, the random forest and logistic regression analysis serves as a valuable tool for airlines seeking to understand and predict passenger satisfaction. The identified influential factors, especially the role of online boarding, offer practical guidance for implementing targeted improvements and ultimately enhancing the overall travel experience for passengers.