A logo with text on it

Description automatically generated

***BA with R – Project Report***

***AIRLINE PASSENGER SATISFACTION***

***Group 17***

**Group Members:**

**Hitarth Parmar**

**Gautam Pai**

**Sachin B Khyadi**

**Pujithaa Vijayanand**

**Introduction**

In the competitive aviation industry, passenger satisfaction is a vital metric that reflects the quality of an airline's services and significantly influences its reputation and profitability. Satisfied passengers are more likely to become loyal customers and advocates for the airline, while dissatisfaction can lead to negative reviews and loss of business.

Understanding the factors that impact passenger satisfaction is essential for airlines to identify areas for improvement and tailor their services to meet customer needs. Factors such as inflight services, flight punctuality, seat comfort, and overall travel experience play a crucial role in shaping customer perceptions.

With the increasing availability of large-scale passenger feedback datasets, machine learning has emerged as a powerful tool to uncover patterns and predict satisfaction levels. This project aims to analyze the key determinants of airline passenger satisfaction using a comprehensive dataset and apply classification techniques to predict whether a passenger is satisfied or not.

By identifying the drivers of satisfaction and areas requiring improvement, this project provides actionable insights for airlines to enhance service quality, improve operational efficiency, and maintain a competitive edge in the market.

**Objective**

The primary objective of this project is to analyze and predict airline passenger satisfaction using machine learning techniques. Specifically, the goals include:

1. **Understanding Passenger Satisfaction Factors:**  
   Identify the key factors influencing passenger satisfaction, such as inflight services, flight punctuality, and seat comfort.
2. **Building a Predictive Model:**  
   Develop a machine learning model to classify passengers as "satisfied" or "not satisfied" based on various demographic, flight-related, and service-related features.
3. **Deriving Actionable Insights:**  
   Provide data-driven recommendations to airlines to improve passenger experiences and operational efficiency by addressing areas of dissatisfaction.
4. **Enhancing Customer Retention:**  
   Leverage insights from the analysis to help airlines foster customer loyalty and maintain a competitive advantage in the aviation industry.

**Data Description**

This dataset captures an airline passenger satisfaction survey, focusing on factors that may influence customer satisfaction. It includes demographic details (e.g., gender, age, customer loyalty), flight details (e.g., travel type, class, distance), and service ratings across various aspects like in-flight Wi-Fi, food, entertainment, and cleanliness. The target variable is customer satisfaction, categorized as “Satisfied,” “Neutral or Dissatisfied.” This dataset is well-suited for exploring which factors are most associated with satisfaction and for building predictive models to assess future customer satisfaction based on survey responses.

This dataset contains a total of 23 variables, consisting of 5 categorical variables such as gender, customer type, type of travel, class and satisfaction and numerical variables containing ratings of various services inflight Wi-Fi, food, entertainment, and cleanliness etc. Additionally, it includes delay times for both departure and arrival.

**Executive Summary**

This project investigates the factors influencing airline passenger satisfaction using machine learning techniques. With increasing competition in the aviation industry, understanding passenger preferences and improving services is crucial for airline success.

The analysis uses the Airline Passenger Satisfaction Dataset, which includes passenger demographics, flight details, inflight service ratings, and satisfaction outcomes. The dataset, consisting of 129,880 records and 24 features, provides a robust foundation for predictive analysis.

Our primary goal is to classify passengers as "satisfied" or "not satisfied" based on various features and identify key drivers of satisfaction. Insights from this study aim to enable airlines to implement data-driven improvements in customer experience, optimizing operations and enhancing customer retention.

**Project Motivation/Background**

Airline passenger satisfaction is a critical determinant of an airline’s reputation, customer loyalty, and overall success. Positive passenger experiences translate into increased customer retention and word-of-mouth promotion, which are vital in the competitive aviation industry.

Given the wealth of data generated through passenger feedback, machine learning offers an efficient way to derive actionable insights from this data. By predicting satisfaction levels and identifying influential factors, airlines can proactively address issues, tailor their services to passenger needs, and enhance overall travel experiences.

This project leverages machine learning classification techniques to explore a dataset of passenger feedback. The insights derived will help airlines prioritize service improvements, ultimately boosting customer satisfaction and maintaining a competitive edge.

**Data Description**

The analysis is based on the **Airline Passenger Satisfaction Dataset**, sourced from Kaggle. The dataset includes **129,880 instances** and **24 features**, capturing a comprehensive snapshot of the factors influencing passenger satisfaction.

Key features include:

* **Passenger Demographics:** Age, gender, and travel class.
* **Flight Details:** Flight distance, departure and arrival delays.
* **Inflight Services:** Ratings for food, entertainment, seat comfort, and inflight service quality.
* **Satisfaction Label:** Indicates whether the passenger was satisfied (1) or not satisfied (0).

The dataset contains a mix of numerical and categorical variables. Numerical features include flight distance and service ratings, while categorical variables cover demographic and travel information.

This dataset provides a robust foundation for machine learning models, enabling us to analyze passenger satisfaction trends and develop predictive insights to aid airline decision-making.

**Data Preprocessing**

**Handling Missing Data:** The "Arrival Delay in Minutes" column had approximately 300 missing values (NAs). Given the large dataset of approximately 100,000 rows, the decision was made to omit these missing values.

**Data Normalization:** The values in the "Age," "Flight Distance," "Departure Delay in Minutes," and "Arrival Delay in Minutes" columns were normalized using a scaling function.

**Handling Categorical Variables:** Categorical variables such as "Gender," "Customer Type," "Type of Travel," and "Class" were converted into dummy variables. Additionally, the "Satisfaction" variable was updated to a binary format, where "Satisfied" was set to 1, and "Neutral or Dissatisfied" were set to 0.

**Exploratory Data Analysis**

* **Loyality vs satisfaction :** As expected, loyal customers were more satisfied.

**A close-up of a white background

Description automatically generated**

* **Age vs satisfaction :** No particular trend. Both male and female had similar satisfaction level.

**A white background with black dots

Description automatically generated**

* **Class vs satisfaction :** Business class travelers were more satisfied with the services. Eco class travelers were not satisfied with the services.

**A close-up of a white background

Description automatically generated**

* Based on the plot below, the “Delay in Arrival” and “Delay in Departure” are correlated.

**A blue dotted line graph

Description automatically generated**

* Below plot shows that passenger travelling for a longer distance were more satisfied with the airline service.

A graph of a plane

Description automatically generated with medium confidence

Below plot shows that passenger satisfaction by Age

A screenshot of a computer code

Description automatically generated



Bar plot for satisfied customers

A screenshot of a computer

Description automatically generated

**A graph of blue bars

Description automatically generated with medium confidence**

**Bar plot for unsatisfied customers**

**A screenshot of a computer

Description automatically generated**

**A graph of a number of red bars

Description automatically generated with medium confidence**

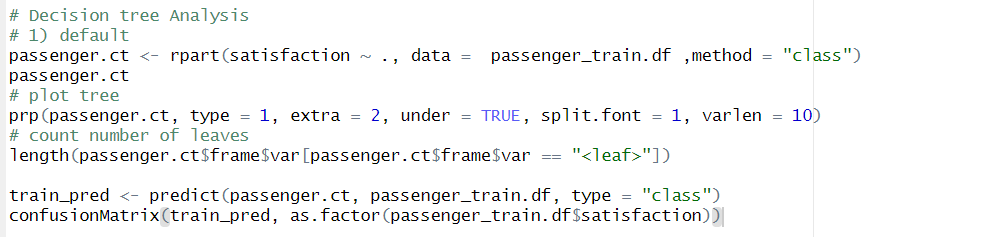
Below table shows the summary statistics of Airline passengers

**A screenshot of a computer screen

Description automatically generated**

**Business Intelligence Models**

**Model 1 – Decision Tree Analysis**

****

**Key Points**

* Tree Creation: A default decision tree is created using the rpart package, predicting satisfaction based on all other variables.
* Visualization: The tree is visualized with prp():
  + type = 1: Displays variable names at splits.
  + extra = 2: Shows class probabilities or proportions at nodes.
  + varlen = 10: Limits variable name lengths for clarity.
* Leaf Count: Counts the terminal nodes (leaves) using length(). Each leaf represents a unique decision path.

**Insights**

* The default tree structure is likely overfitted to the training data because no constraints on minimum splits or complexity are imposed.
* Analyzing leaf count gives a sense of tree complexity. Too many leaves may indicate overfitting.

A diagram of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**2) Cross-Validation and Tree Optimization**

**Key Points**

* **Cross-Validation**: Creates a decision tree with 5-fold cross-validation (xval = 5). This helps evaluate model stability across different data splits.
* **Complexity Parameter (cp)**:
  + printcp() prints a table showing cross-validation error (xerror) at different tree sizes.
  + As the tree grows, xerror decreases initially but plateaus. The optimal cp minimizes this error.
* **Tree Pruning**: The tree is pruned using prune() with the selected cp value (5.9300e-03 in this case). Pruning reduces overfitting by simplifying the tree.

**Insights**

* Cross-validation helps identify the optimal tree complexity.
* Pruned trees balance accuracy and generalizability by removing less important splits.

A computer code with text

Description automatically generated with medium confidence

A computer screen shot of a computer code

Description automatically generated

**3) Predictions and Evaluations**

**Key Points**

* Training Data Prediction: Predictions on the training set are evaluated using a confusion matrix.
* Metrics: The confusionMatrix() function calculates accuracy, sensitivity, and specificity.

**Insights**

* Training accuracy for the pruned tree is expected to be lower than the unpruned tree but more generalizable.
* High accuracy on the training set may indicate overfitting if not validated on unseen data.

A screenshot of a computer

Description automatically generated

**4) Validation Data Classification**

**Key Points:**

* Testing Data Prediction: Predictions are made on the validation dataset using the pruned tree.
* Evaluation: Another confusion matrix evaluates performance on unseen data.

**Insights:**

* A drop in accuracy compared to the training set indicates how well the model generalizes.
* High validation accuracy reflects a good balance between underfitting and overfitting.

**A screenshot of a computer

Description automatically generated**

**Model 2: - Logistic Regression Analysis**

**1 Fetching Data**

Reads the training and test datasets from the specified file paths into data frames (passenger\_train.df and passenger\_test.df).

header = TRUE: Indicates the first row of the CSV file contains column names.

**A close-up of a computer code

Description automatically generated**

**Exploratory Analysis**

Provides a summary of the dataset, including statistics for numerical columns and counts for categorical variables.

**summary(passenger\_train.df)**

**A screenshot of a computer screen

Description automatically generated**

**Handling Categorical Variables**

Iterates over all columns to check for missing (NA) values or empty strings in categorical variables.

Output: Displays the count of NA and empty values for each categorical column.

**A computer screen shot of a computer code

Description automatically generated**

**Removing Missing Values**

Removes rows containing missing (NA) values from the dataset**.**

**passenger\_train.df <- na.omit(passenger\_train.df)**

**Data Analysis : -**

Drops columns X and Arrival.Delay.in.Minutes, as they are either redundant or highly correlated with Departure.Delay.in.Minutes.

passenger\_train.df <- passenger\_train.df[, !colnames(passenger\_train.df) %in% c("X", "Arrival.Delay.in.Minutes")]

**Converting Categorical Variables to Factors**

Converts specific columns into factors for proper handling in machine learning models.

**categorical\_vars <- c("Gender", "Customer.Type", "Type.of.Travel", "Class")**

**passenger\_train.df[categorical\_vars] <- lapply(passenger\_train.df[categorical\_vars], factor)**

**Encoding Target Variable**

Converts the satisfaction column into a binary numeric variable (1 for satisfied, 0 for not satisfied).

**passenger\_train.df$satisfaction <- ifelse(passenger\_train.df$satisfaction == "satisfied", 1, 0)**

**Logistic Regression Model**

Fits a logistic regression model to predict satisfaction based on all other variables.

family = binomial: Specifies a logistic regression model (appropriate for binary classification).

summary(model): Displays model coefficients, significance, and fit statistics.

**A screenshot of a computer

Description automatically generated**

**Prediction and Evaluation**

predict(model, type = "response"): Predicts probabilities of satisfaction for each row in the dataset.

Converts probabilities above 0.5 into the class 1 (satisfied), and the rest as 0 (not satisfied).

**A screenshot of a computer

Description automatically generated**

**Confusion Matrix**

Evaluates the performance of the logistic regression model by comparing predicted classifications with actual values.

A screenshot of a computer program

Description automatically generated

**Model 3: - Neural Network Analysis**

**1. Data Preparation**

**Key Steps:**

1. **Normalization**:
   * A normalize function scales numerical features to a range of 0 to 1.
   * Normalization ensures faster convergence and better model performance by avoiding large value disparities across features.
2. **Categorical to Numeric Conversion**:
   * Converts categorical predictors to numeric or dummy variables using dplyr.
   * Ensures compatibility with the neural network algorithm, which requires numeric inputs.
3. **Target Encoding**:
   * Converts the target variable (satisfaction) into binary format (e.g., 0 and 1), suitable for binary classification.

**Insights:**

* Normalization and preprocessing are crucial for neural network performance as they reduce noise and scaling issues.
* Encoding categorical variables prevents errors and ensures compatibility with the model.

**2. Neural Network Model Training**

**Key Steps:**

1. **Model Definition**:
   * A formula defines the target variable (satisfaction) and predictors.
   * The model uses two hidden layers with 5 and 3 neurons, respectively (hidden = c(5, 3)).
2. **Training**:
   * neuralnet trains the model using backpropagation.
   * linear.output = FALSE ensures the model outputs probabilities, suitable for classification tasks.
3. **Visualization**:
   * The plot(nn\_model) function visualizes the neural network architecture, including input, hidden, and output layers.

**Insights:**

* Hidden layers and neurons determine the complexity of the model. More neurons capture complex patterns but risk overfitting.
* Setting linear.output = FALSE aligns the output with classification tasks, producing probabilities instead of raw values.

**3. Predictions and Evaluation**

**Key Steps:**

1. **Prediction**:
   * The compute() function generates predictions on the test set.
   * Predictions are probabilities, which are converted to binary classes using a threshold of 0.5 (ifelse(nn\_predictions > 0.5, 1, 0)).
2. **Evaluation**:
   * The confusion matrix evaluates the model's performance on the test data.
   * Key metrics include accuracy, precision, recall, and F1-score.

**Insights:**

* A threshold of 0.5 assumes equal importance for both classes. Adjusting the threshold might improve performance for imbalanced datasets.
* Confusion matrix metrics provide insights into misclassification rates and overall model reliability.

**4. General Observations**

1. **Strengths of Neural Networks**:
   * Capable of modeling complex, non-linear relationships between predictors and the target variable.
   * Flexible architecture with customizable layers and neurons.
2. **Challenges**:
   * Computationally expensive, especially for large datasets or deep architectures.
   * Requires careful tuning of hyperparameters (e.g., number of neurons, hidden layers) to avoid underfitting or overfitting.
3. **Performance Indicators**:
   * A high accuracy on the test data indicates a well-trained model.
   * If the model performs well on the training data but poorly on the test data, it may be overfitting.