Project Report: Airline Passenger Satisfaction Prediction

Anjali Jaiswal 234161002

***Abstract*—This project aims to predict passenger satisfaction with airline services using machine learning techniques. We explore a dataset containing demographic information, travel details, and satisfaction ratings of airline passengers to identify factors influencing satisfaction levels. Various machine learning models, including Random Forest, Logistic Regression, MLP Classifier, and Gradient Boosting Classifier, are trained and evaluated on selected features to predict passenger satisfaction accurately. The results demonstrate the effectiveness of these models in predicting satisfaction levels and provide valuable insights for improving airline services.**

I. Introduction

The airline industry is highly competitive, with airlines constantly striving to enhance customer satisfaction and loyalty. Understanding factors that influence passenger satisfaction is crucial for airlines to improve their services and maintain a loyal customer base. In this project, I aim to predict passenger satisfaction with airline services using machine learning techniques.

The dataset used in this project contains information about airline passengers, including their demographic information, travel details, and satisfaction ratings. By analyzing this data and building predictive models, I seek to identify the key factors that contribute to passenger satisfaction and develop a model that can accurately predict satisfaction levels based on various attributes.

This project has several objectives:

- Explore the dataset to understand its structure and characteristics.
- Preprocess the data to handle missing values, encode categorical variables, and prepare it for modeling.
- Perform feature selection to identify the most relevant features for predicting passenger satisfaction.
- Train and evaluate machine learning models, including Random Forest, Logistic Regression, MLP Classifier, and Gradient Boosting Classifier.
- Analyze the performance of the models and identify the most effective approach for predicting passenger satisfaction.

By achieving these objectives, I aim to provide insights into the factors that influence passenger satisfaction in the airline industry and develop a predictive model that can assist airlines in improving their services and enhancing customer experience.

II. Dataset Description

The dataset consists of two files: `train.csv` and `test.csv`. The training dataset (`train.csv`) contains 103,904 rows, while the test dataset (`test.csv`) contains 17,965 rows. Each row represents a passenger's feedback and demographic information after a flight.

A. Features

The dataset includes the following features:

- **id**: Unique identifier for each passenger.
- **Gender**: Gender of the passenger (*Male* or *Female*).
- **Customer Type**: Type of customer (*Loyal Customer* or *Disloyal Customer*).
- **Age**: Age of the passenger.
- **Type of Travel**: Purpose of travel (*Personal Travel* or *Business Travel*).
- **Class**: Class of travel (*Business*, *Eco*, or *Eco Plus*).
- **Flight Distance**: Distance of the flight traveled by the passenger.
- **Inflight wifi service**: Satisfaction rating for inflight Wi-Fi service.
- **Departure/Arrival time convenient**: Satisfaction rating for departure/arrival time convenience.
- **Ease of Online booking**: Satisfaction rating for ease of online booking.
- **Gate location**: Satisfaction rating for gate location convenience.
- **Food and drink**: Satisfaction rating for food and drink quality.
- **Online boarding**: Satisfaction rating for online boarding process.
- **Seat comfort**: Satisfaction rating for seat comfort.
- **Inflight entertainment**: Satisfaction rating for inflight entertainment.
- **On-board service**: Satisfaction rating for on-board service.
- **Leg room service**: Satisfaction rating for leg room service.
- **Baggage handling**: Satisfaction rating for baggage handling.
- **Check-in service**: Satisfaction rating for check-in service.
- **Inflight service**: Satisfaction rating for inflight service.
- **Cleanliness**: Satisfaction rating for cleanliness.
- **Departure Delay in Minutes**: Delay in departure time in minutes.
- **Arrival Delay in Minutes**: Delay in arrival time in minutes.
- **Satisfaction**: Overall passenger satisfaction rating (*satisfied* or *neutral or dissatisfied*).

This dataset provides a comprehensive set of features to analyze various aspects of the flight experience and predict passenger satisfaction levels.

III. Data Preprocessing

- I started by loading the dataset and performing exploratory data analysis to understand its structure and characteristics.
- Missing values were handled using appropriate techniques such as imputation.
- Categorical variables were encoded using label encoding or one-hot encoding.
- Feature selection was performed to identify relevant features for modeling.

IV. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial initial step in any data analysis task as it helps in understanding the dataset's structure, characteristics, and relationships between variables. In this section, we will explore the dataset containing information about airline passengers' demographics, travel details, and satisfaction ratings through various visualization techniques and statistical summaries.

A. Histograms

Histograms provide a visual representation of the distribution of numerical variables in the dataset. By examining histograms, we can identify the central tendency, dispersion, and skewness of the data. For example, analyzing the histogram of the "Age" feature can reveal the age distribution of passengers in the dataset.

B. Bar Charts

Bar charts are useful for visualizing the frequency distribution of categorical variables. By plotting bar charts for features such as "Gender" and "Customer Type," we can observe the distribution of passengers based on these demographic categories. This helps in understanding the composition of the dataset in terms of gender and customer type.

C. Box Plots

Box plots are effective for visualizing the distribution of numerical variables across different categories. For instance, plotting box plots for "Flight Distance" by "Class" allows us to compare the distribution of flight distances for different travel classes. This can help identify any significant differences or outliers in flight distance based on class.

D. Scatter Plots

Scatter plots are valuable for exploring relationships between two numerical variables. For example, plotting a scatter plot of "Flight Distance" versus "Arrival Delay in Minutes" can help identify any correlation or patterns between the distance traveled and arrival delays.

E. Frequency Distribution

Frequency distribution charts, such as count plots, provide insights into the distribution of categorical variables. By plotting the frequency distribution of passenger satisfaction levels, we can visualize the proportion of satisfied and dissatisfied passengers in the dataset. This helps in understanding the overall satisfaction levels and identifying potential areas for improvement.

Through EDA, we aim to gain a comprehensive understanding of the dataset, uncover patterns and relationships between variables, and identify potential insights that can guide further analysis and model development.

V. Model Training and Evaluation

A. Random Forest Classifier

- Trained a Random Forest Classifier and selected features.
- Evaluated the model's performance using metrics such as accuracy, precision, recall, and F1-score.

B. Logistic Regression

- Utilized Logistic Regression to build a predictive model.
- Scaled the features and trained the model on the selected features.
- Evaluated the model's performance and compared it with other models.

C. MLP Classifier (Neural Network)

- Implemented an MLP Classifier for complex modeling.
- Trained the neural network model using the scaled features.
- Evaluated the performance of the model and analyzed the results.

D. Gradient Boosting Classifier

- Employed Gradient Boosting Classifier to leverage ensemble learning.
- Trained the model on the selected features and evaluated its performance.
- Compared the results with other models to determine the most effective approach.

VI. Feature Selection Using Random Forest

Feature selection is a crucial step in building predictive models, as it helps identify the most relevant features that contribute to the model's performance while eliminating irrelevant or redundant ones. In this project, I employed the Random Forest algorithm for feature selection.

A. Methodology

Random Forest is an ensemble learning technique that combines multiple decision trees to create a robust model. One of the advantages of Random Forest is its ability to rank features based on their importance in predicting the target variable. This importance score is calculated based on the

decrease in impurity (e.g., Gini impurity) caused by each feature when used in the decision trees.

To perform feature selection using Random Forest, I followed these steps:

1) **Training the Random Forest Model**: I trained a Random Forest classifier using the entire set of features in the dataset.
2) **Feature Importance Calculation**: After training the model, I extracted the feature importance scores from the Random Forest.
3) **Selecting Important Features**: I selected the top-performing features based on their importance scores. These features were considered to have the most significant impact on the target variable.

B. Feature Importance

The feature importance analysis conducted using the Random Forest Classifier revealed the following significant features, ranked by their importance scores:

Feature	Importance
Online boarding	0.175
Inflight wifi service	0.128
Type of Travel	0.105
Class	0.097
Seat comfort	0.055
Inflight entertainment	0.043
Customer Type	0.039
Leg room service	0.039
Ease of Online booking	0.038
Age	0.032
Flight Distance	0.031
On-board service	0.029
Cleanliness	0.028
Inflight service	0.027
Check-in service	0.024
Baggage handling	0.023
Unnamed	0.019
Departure/Arrival time convenient	0.015
Gate location	0.015
Food and drink	0.012
Arrival Delay in Minutes	0.012
Departure Delay in Minutes	0.010
Gender	0.004

These features play a crucial role in predicting passenger satisfaction, with "Online boarding" exhibiting the highest importance score, followed by "Inflight wifi service" and "Type of Travel".

VII. Predictive Models on Selected Features

Feature Selection and Scaling

Features are selected based on importance scores, with a threshold of 0.03, and then scaled using standardization.

Model Architecture

1) **Logistic Regression:** No explicit architecture.
2) **MLP Classifier:**
   - Two hidden layers:
     - First hidden layer: 100 neurons
     - Second hidden layer: 50 neurons
3) **Gradient Boosting Classifier:**
   - Ensemble of decision trees:
     - Number of trees: 100
     - Maximum depth of each tree: 3

Training

All models are initialized and trained using the scaled features of the training data.

Prediction

Predictions are made on the test data using each trained model.

Performance Evaluation

Performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix are computed to evaluate the performance of the logistic regression model.

VIII. Results and Discussion

A. Random Forest

After training the Random Forest Classifier and making predictions on the test dataset, the following evaluation metrics were computed:

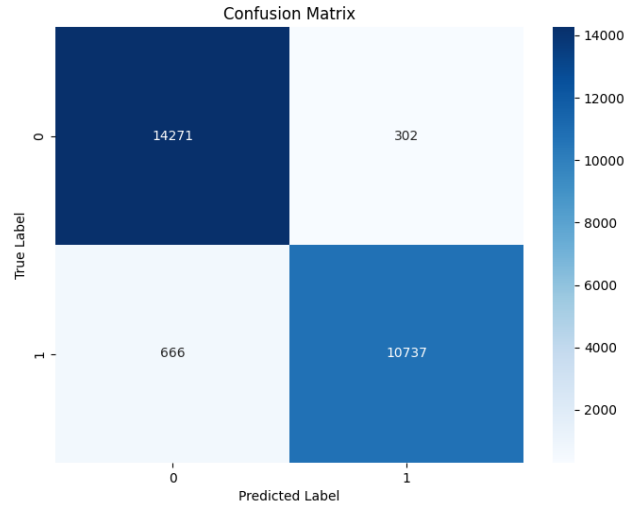
	Precision	Recall	F1-Score	Support
0	0.96	0.98	0.97	14573
1	0.97	0.94	0.96	11403

- **Accuracy:** 0.96
- **Macro Avg (Precision/Recall/F1-Score):** 0.96
- **Weighted Avg (Precision/Recall/F1-Score):** 0.96

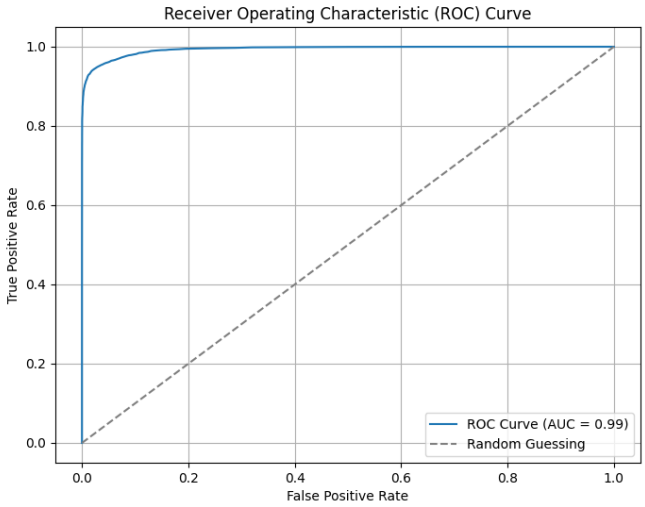
The feature importance scores obtained from the Random Forest model provided valuable insights into the dataset. By analyzing these scores, I identified the most relevant features for predicting passenger satisfaction. These features were then used for building subsequent predictive models such as logistic regression, MLP classifier, and gradient boosting classifier.

B. Logistic Regression

The logistic regression model achieved an accuracy of 85.91%, indicating that it correctly classified 85.91% of the instances in the test set. The precision score, which measures the proportion of true positive predictions among all positive predictions, was found to be 85.89%. Moreover, the recall score, representing the proportion of true positive instances that were correctly classified, also stood at 85.91%. The F1-score, which combines precision and recall into a single metric, was calculated to be 85.89%. These results indicate that the logistic regression model performed reasonably well in predicting the target variable based on the selected features.

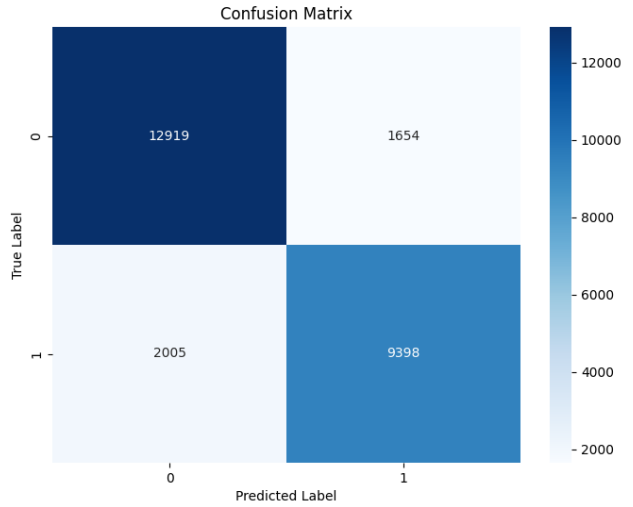


(a) Confusion Matrix

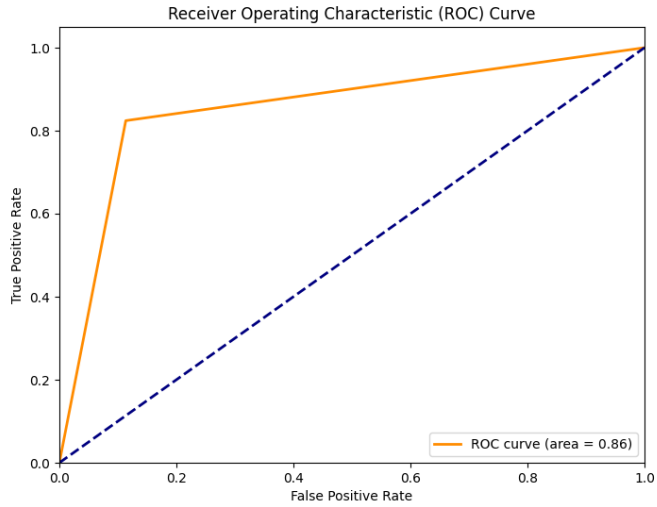


(b) ROC Curve

Fig. 1: Confusion Matrix and ROC Curve of Random Forest



(a) Confusion Matrix



(b) ROC Curve

Fig. 2: Confusion Matrix and ROC Curve of Logistic Regression

C. MLP Classifier

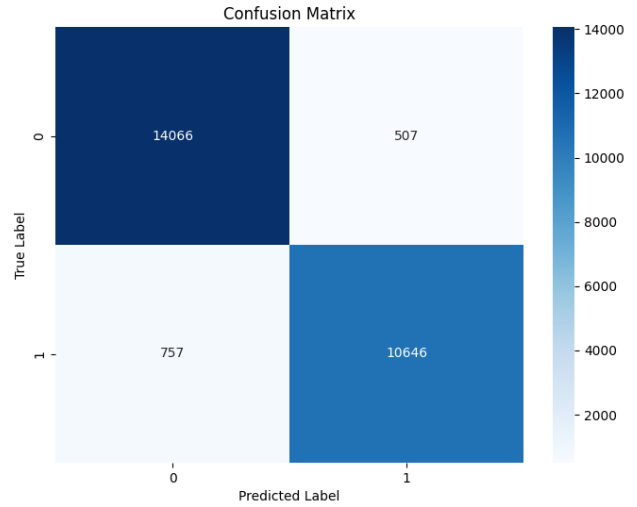
The MLP classifier exhibited outstanding performance in predicting the target variable using the selected features. With an accuracy of 95.13%, the model demonstrated its ability to correctly classify instances in the test set. Furthermore, achieving precision, recall, and F1-score values of approximately 95.14%, 95.13%, and 95.13% respectively, underscores the model's effectiveness in accurately identifying positive instances while minimizing false positives and false negatives. Overall, the MLP classifier demonstrated strong predictive capabilities and robust performance, highlighting its suitability for the task at hand.

D. Gradient Boosting Classifier

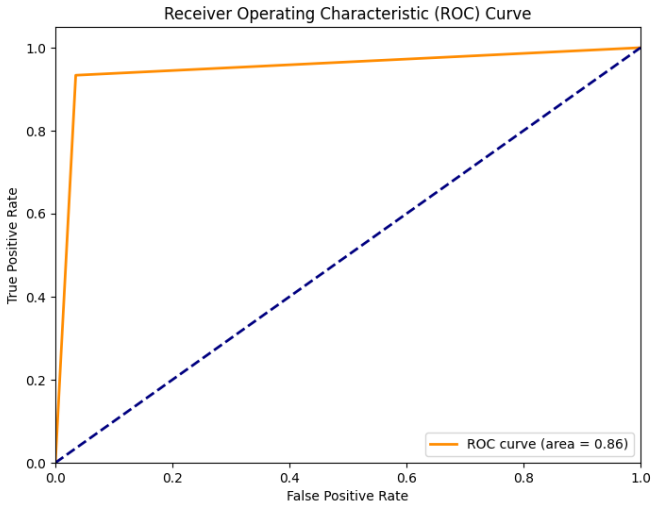
The Gradient Boosting Classifier displayed impressive performance metrics when applied to the dataset. With an accuracy score of 93.10%, the model showcased its ability to accurately classify instances in the test set. Precision, recall, and F1-score values, all hovering around 93.11%, indicate the model's effectiveness in correctly identifying positive instances while maintaining a balance between precision and recall. Overall, these results underscore the Gradient Boosting Classifier's strong predictive capabilities and its potential suitability for the task at hand.

IX. Conclusion

In conclusion, this project aimed to develop and evaluate machine learning models for predicting customer satisfaction based on various features. The dataset comprised essential features such as gender, age, flight distance, and service ratings, among others.

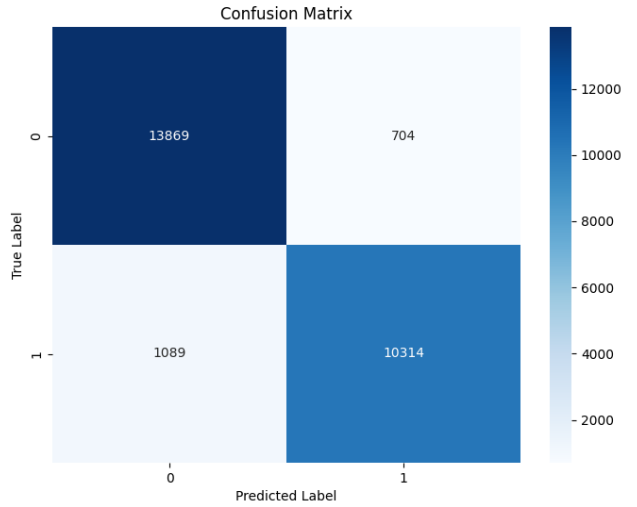


(a) Confusion Matrix

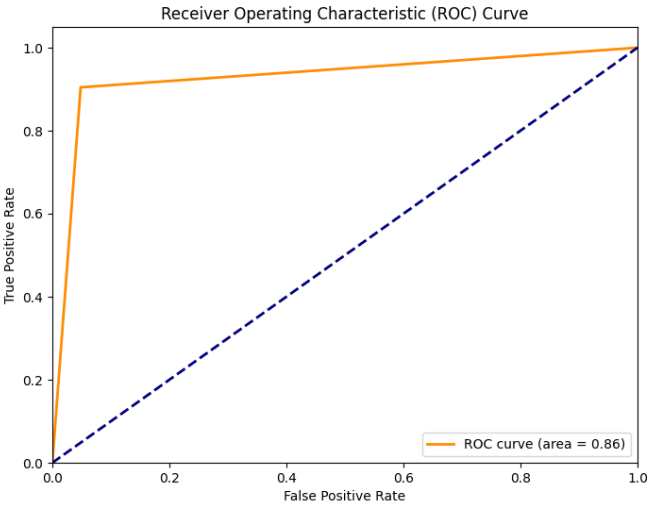


(b) ROC Curve

Fig. 3: Confusion Matrix and ROC Curve of MLP Classifier



(a) Confusion Matrix



(b) ROC Curve

Fig. 4: Confusion Matrix and ROC Curve of Gradient Boosting Classifier

Initially, feature selection was performed using the Random Forest algorithm, which identified key predictors contributing to customer satisfaction. Subsequently, logistic regression, MLP classifier, and Gradient Boosting Classifier models were trained and evaluated on the selected features.

The logistic regression model demonstrated an accuracy of 85.91%, while the MLP classifier achieved a higher accuracy of 95.13%. Additionally, the Gradient Boosting Classifier exhibited an accuracy of 93.10%. These models showcased robust performance metrics, indicating their efficacy in predicting customer satisfaction.

The features, including 'Online boarding', 'Inflight wifi service', 'Type of Travel', 'Class', 'Seat comfort', 'Inflight entertainment', 'Customer Type', 'Leg room service', 'Ease of Online booking', 'Age', and 'Flight Distance', were identified as relevant for assessing customer satisfaction using a Random Forest algorithm. These features were further validated by employing other machine learning models and evaluating their performance metrics. This comprehensive approach ensures the robustness and reliability of the selected features in capturing the key factors influencing customer satisfaction in airline services.

Overall, the project highlights the significance of feature selection and model evaluation in developing effective predictive models for customer satisfaction. The results obtained underscore the potential of machine learning techniques in improving customer service and enhancing overall business performance.

X. Future Work

- Further fine-tuning of hyperparameters for each model could potentially improve performance.
- Exploring additional features or engineering new features may enhance the predictive power of the models.
- Deploying the trained model into a production environment for real-time predictions could be a valuable next step.