Airline Passenger Satisfaction Classification

Aryan Pakdel
Birmingham City University
Aryan.pakdel@mail.bcu.ac.uk

Abstract— *Passenger satisfaction is a key metric in the aviation industry, directly influencing customer loyalty and the overall reputation of airlines. This project explores the use of machine learning to predict airline passenger satisfaction based on historical data. The dataset includes demographic details, flight-related information, and service ratings, with the target variable indicating satisfaction status. Through exploratory data analysis, significant patterns were identified, including correlations between delays, service quality, and satisfaction levels. Three machine learning models—Random Forest, Logistic Regression, and XGBoost—were trained and evaluated. The XGBoost model outperformed others, achieving an accuracy of 96% and an AUC score of 0.92 on the test set. These findings underscore the potential of machine learning in delivering actionable insights to improve customer experience and operational efficiency in the airline industry. Future directions include integrating additional features and deploying the model in real-time systems.*

Keywords: Airline Passenger Satisfaction, Machine Learning, Classification, XGBoost, Random Forest, Logistic Regression, Predictive Analytics

1. Introduction

Passenger satisfaction is a critical factor in the aviation industry, influencing customer retention, brand reputation, and overall business success. Understanding the drivers of satisfaction allows airlines to refine their services and address customer concerns effectively. As passenger data becomes more readily available, machine learning offers a powerful approach to uncovering patterns and predicting satisfaction levels.This project explores the use of machine learning techniques to classify airline passengers as satisfied or dissatisfied based on historical data. The dataset includes a range of features such as demographic details, flight-related metrics, and service ratings, providing a holistic view of factors influencing satisfaction. The target variable is binary, indicating whether a passenger was satisfied with their flight experience.

To achieve our objective, we implemented and evaluated three machine learning models: Random Forest, Logistic Regression, and XGBoost. These models were selected for their established efficacy in classification tasks and their complementary strengths in handling different types of data. The dataset was preprocessed and split into training, validation, and test sets to ensure unbiased model evaluation. This study focuses on building a robust predictive model to classify passenger satisfaction accurately. By comparing model performances and analyzing key features, we aim to provide valuable insights into the factors that influence satisfaction in the aviation sector. The results of this analysis demonstrate the potential of data-driven methods in addressing industry challenges and improving decision-making processes.

2. Dataset Overview

The dataset used in this project provides detailed information about airline passengers, with features describing various aspects of their flight experience and demographics. Initially, the data was provided in two separate files: one for training and one for testing. To ensure consistency and allow for a custom train-test split, these two datasets were merged, resulting in a single dataset with a total shape of **129,880 rows and 25 columns**.

Before preprocessing, two irrelevant columns, **Unnamed: 0** and **id**, were removed as they provided no meaningful information for the analysis or modelling process. This left a refined dataset ready for further analysis.

3. Data Exploratory Analysis

The The dataset consists of **129,880 entries** and **23 columns** after preprocessing. Each column represents a feature describing passenger demographics, flight details, or service ratings, with the target column, **satisfaction**, indicating passenger satisfaction. Here are the key details:

Data Types:

Numerical Features: 17 columns, including Age, Flight Distance, and various service ratings on a scale of 1 to 5.

Categorical Features: 5 columns, including Gender, Customer Type, Type of Travel, Class, and satisfaction.

Missing Values:

The Arrival Delay in Minutes column contains 393 missing values, which will require imputation or handling during preprocessing.

The numerical features in the dataset provide valuable insights into the distribution and range of various attributes related to the passengers' flight experiences. Below is a summary of the key numerical columns:

Age:

The mean age of passengers is 39.43 years, with a wide range spanning from 7 to 85 years. The standard deviation of 15.12 indicates a fairly diverse age distribution.

Flight Distance:

The average flight distance is 1190.32 km, with a minimum of 31 km and a maximum of 4983 km. The large standard deviation (997.45 km) suggests significant variation in flight distances.

Service Ratings:

Several features, such as Inflight Wifi Service, Ease of Online Booking, and Food and Drink, represent passenger ratings on a scale from 0 to 5. The mean values for these features are generally around 3, indicating that most passengers rated the services neutrally to positively, with Inflight Wifi Service having a mean of 2.73 and Ease of Online Booking scoring slightly higher at 2.76. Ratings tend to have moderate variability with standard deviations around 1.3.

Delays:

The Departure Delay has a mean value of 14.71 minutes, while the Arrival Delay shows a mean of 15.09 minutes. There is significant variation, with delays ranging from 0 to 1592 minutes. The standard deviation for both delays is high, indicating occasional extreme delays that can skew the overall distribution.

Percentiles:

At the 25th percentile, most ratings and features (except for delays) are clustered around the lower end, with values near 2 for most service-related ratings. By the 75th percentile, these features shift toward higher ratings, reflecting more positive experiences (values approaching 5).

The target variable, **satisfaction**, represents the passenger's satisfaction with their flight experience, classified into two categories: "satisfied" and "neutral or dissatisfied." Below is the distribution of these two classes:

Neutral or Dissatisfied: **73,225** passengers (56.55%)

Satisfied: **56,262** passengers (43.45%)

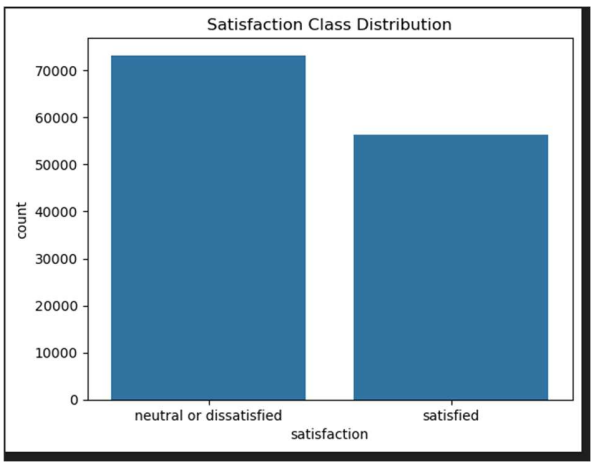

Figure 1: Satisfaction Class Distribution (Value)

This indicates that the dataset is slightly imbalanced, with a higher proportion of passengers categorized as "neutral or dissatisfied" compared to those who are "satisfied."

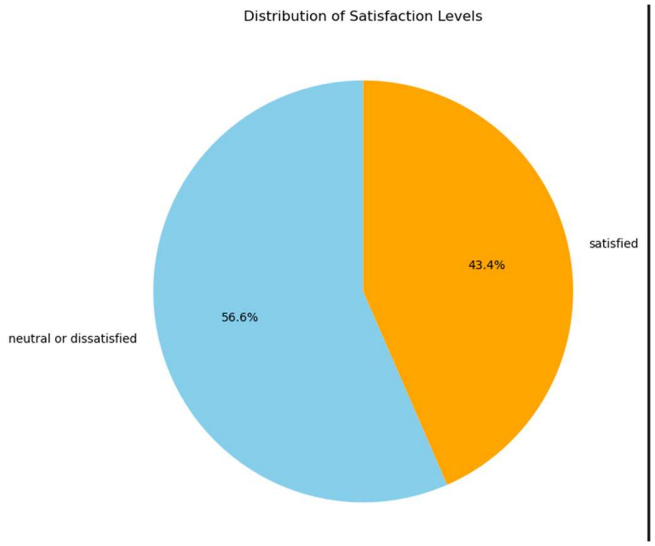

Figure 2: Satisfaction Class Distribution (Percentage)

Age vs. Satisfaction:

A crosstab analysis between **Age** and **Satisfaction** reveals interesting patterns in how satisfaction levels vary across different age groups:

- The highest number of **neutral or dissatisfied** passengers is observed in the **25-year-old** age group, suggesting that younger passengers may be less satisfied with their flight experience.
- On the other hand, the **39-year-old** age group has the highest number of **satisfied** passengers, indicating that middle-aged passengers may generally have more positive experiences during their flights.
- Additionally, passengers aged between **39 to 60 years** tend to exhibit the highest overall satisfaction levels, with a notable concentration of satisfied passengers in this age range.

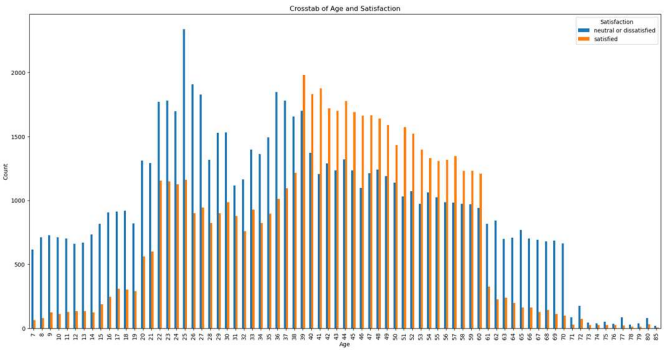

Figure 3: Crosstab of Age and Satisfaction

A crosstab analysis between **Gender** and **Satisfaction** shows that the distribution of satisfaction levels is quite similar across both genders. For both Male and Female passengers, the number of neutral or dissatisfied individuals is higher than the number of satisfied individuals.

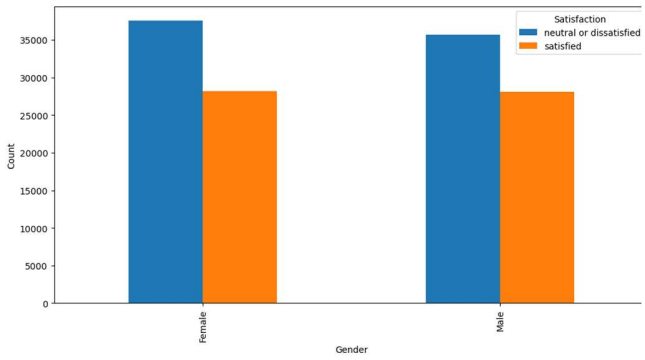

Figure 4: Crosstab of Gender and Satisfaction

A correlation analysis of the numerical features reveals a very high positive correlation between **Departure Delay in Minutes** and **Arrival Delay in Minutes**, with a correlation coefficient of **0.97**. This strong correlation indicates that passengers who experience significant delays at departure tend to have similarly high delays upon arrival.

This relationship is expected, as departure delays often cascade throughout the flight, resulting in a corresponding increase in arrival delays. Such a high correlation suggests that these two features are likely capturing similar information about the flight experience.

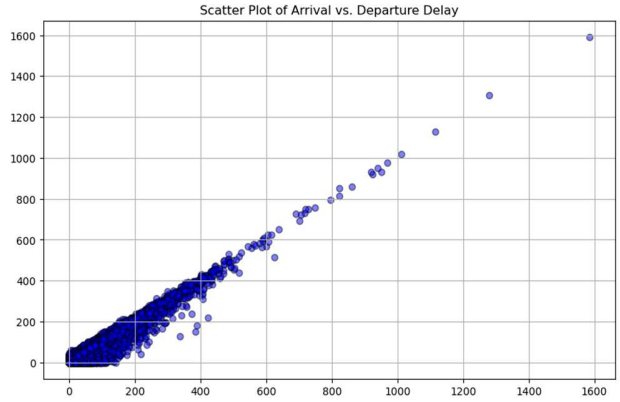

Figure 5: Scatter Plot of Arrival vs. Departure Delay

4. Feature Engineering

During the feature engineering process, we ensured the dataset was prepared for model training by addressing a few key aspects:

Duplicates: There were no duplicate entries in the dataset, ensuring the integrity of the data.

Missing Values: We removed rows with missing values, specifically in the Arrival Delay in Minutes column. Since the number of missing values was relatively small and their removal was not expected to significantly impact the model, this approach was deemed appropriate.

Categorical Data: For the categorical features (such as Gender, Customer Type, Type of Travel, and Class), we applied One-Hot Encoding to convert these variables into numerical representations. This encoding method allows the model to handle categorical features effectively by creating binary columns for each category.

5. Modeling Approach

The dataset was split into training, validation, and test sets to ensure effective model evaluation and generalization.

- **Training Set:** 82,871 samples
- **Validation Set:** 20,718 samples
- **Test Set:** 25,898 samples

Three models, Random Forest, Logistic Regression, and XGBoost, were trained using their default hyperparameters to gain an initial performance overview. The goal was to identify the best-performing model for further optimization. The results on the validation set were as follows:

Random Forest: 96.19% accuracy
Logistic Regression: 83.49% accuracy
XGBoost: 96.26% accuracy

XGBoost and Random Forest demonstrated similarly high accuracy, while Logistic Regression performed noticeably worse. These findings suggested that XGBoost was the best-performing model, making it the primary candidate for further hyperparameter tuning.

For **Random Forest** hyperparameter tuning, a GridSearchCV approach was used to optimize the model's performance by searching over the following parameter grid:

- **n_estimators**: [100, 200, 300] – Number of trees in the forest.
- **max_depth**: [None, 10, 20, 30] – Maximum depth of each tree.
- **min_samples_split**: [2, 5, 10] – Minimum number of samples required to split an internal node.
- **min_samples_leaf**: [1, 2, 4] – Minimum number of samples required to be at a leaf node.

After performing the GridSearchCV, the optimal parameters were **max_depth=30**, **min_samples_leaf=1**, **min_samples_split=2**, and **n_estimators=300** resulting in an accuracy of **96.28%** on the validation set , slightly improving upon the initial default model accuracy of **96.19%.**

The best **XGBoost** parameters (**colsample_bytree=0.8**, **learning_rate=0.1**, **max_depth=6**, **n_estimators=200**, **subsample=1.0**) achieved an accuracy of **96.36%** on the validation set, compared to the default XGBoost accuracy of **96.26%**, showing a slight improvement like Random Forest after hyperparameter tuning.

6. Evaluating and Results

The **Random Forest** model was evaluated on the validation set using multiple metrics to assess its performance. The confusion matrix showed that out of the 20,718 instances, the model correctly classified 11,411 neutral or dissatisfied passengers (True Negatives) and 8,518 satisfied passengers (True Positives). However, it misclassified 254 neutral or dissatisfied passengers as satisfied (False Positives) and 535 satisfied passengers as neutral or dissatisfied (False Negatives).

	Precision	Recall	F1-Score	Support
0 (Neutral or Dissatisfied)	0.96	0.98	0.97	11665
1 (Satisfied)	0.97	0.94	0.96	9053
Accuracy			0.96	20718
Macro avg	0.96	0.96	0.96	20718
Weighted avg	0.96	0.96	0.96	20718

Table 1: Random Forest Classification Report

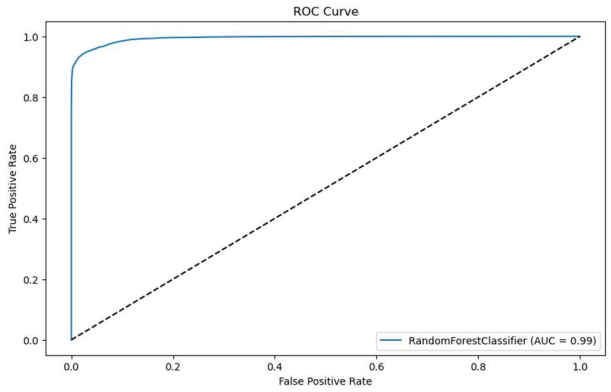

Figure 6: ROC Curve For Random Forest Classification

The classification report revealed strong performance, with an accuracy of 96%. The model achieved a precision of 0.96 for neutral or dissatisfied passengers and 0.97 for satisfied passengers, meaning it rarely misclassified positive instances. Recall was high for both classes, with 0.98 for neutral or dissatisfied and 0.94 for satisfied, indicating the model was able to identify most of the true positives for each class. The F1-scores for both classes were 0.97 and 0.96, respectively, suggesting a good balance between precision and recall. The ROC AUC score was 0.994, which indicates that the Random Forest model has excellent discriminatory power between the two classes. An ROC AUC score close to 1.0 indicates a model that performs very well, making it highly suitable for classifying passenger satisfaction levels.

The **Logistic Regression** model achieved an accuracy of 83% on the validation set. The confusion matrix showed that it correctly identified 9,834 neutral or dissatisfied passengers and 7,464 satisfied passengers. However, it misclassified 1,831 neutral or dissatisfied passengers as satisfied, and 1,589 satisfied passengers as neutral or dissatisfied. The model showed decent precision and recall, with 0.86 precision for neutral or dissatisfied and 0.80 for satisfied. The ROC AUC score of 0.896 indicates reasonable performance but is lower than Random Forest's score, suggesting that Logistic Regression is less effective in distinguishing between the classes compared to other models.

	precision	recall	f1-score	support
0	0.860	0.840	0.850	11665
1	0.798	0.824	0.811	9053
accuracy	0.830			20718
macro avg	0.829	0.832	0.830	20718
weighted avg	0.840	0.830	0.834	20718

Table 2: Logistic Regression Classification Report

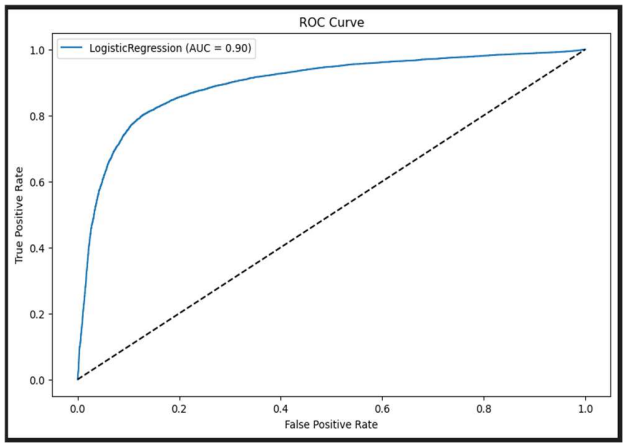

Figure 7: ROC Curve For Logistic Regression

The **XGBoost** model delivered an impressive accuracy of 96% on the validation set. The confusion matrix shows it correctly classified 11,409 neutral or dissatisfied passengers and 8,534 satisfied passengers. It misclassified 256 neutral or dissatisfied passengers as satisfied, and 519 satisfied passengers as neutral or dissatisfied.

	precision	recall	f1-score	support
0	0.960	0.982	0.971	11665
1	0.971	0.942	0.956	9053
accuracy	0.960			20718
macro avg	0.965	0.962	0.963	20718
weighted avg	0.960	0.960	0.960	20718

Table 3: XGBoost Classification Report

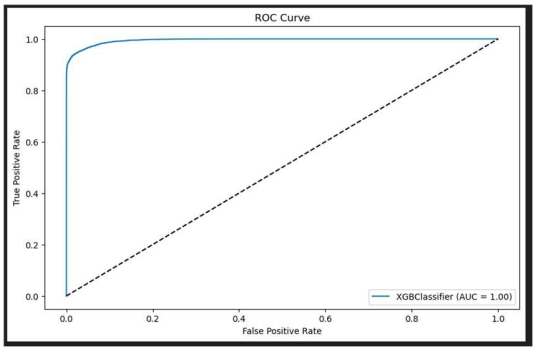

Figure 8: ROC Curve For XGBoost

The model achieved high precision and recall, with a precision of 0.96 for neutral or dissatisfied and 0.97 for satisfied. The ROC AUC score of 0.995 indicates excellent performance, significantly outperforming both Random Forest and Logistic Regression in distinguishing between the classes. This suggests that XGBoost is the best model for this classification task.

7. Evaluating on Unseen Data

The **XGBoost** model, identified as the best-performing algorithm during the evaluation phase, was applied to the test dataset for final validation. The test dataset, which contained 25,898 samples, demonstrated the model's robustness and consistency in predictions.

The evaluation results were as follows:

- Accuracy: 96.15%
- ROC AUC Score: 0.9938

Class	Precision	Recall	F1-Score	Support
0 (Neutral/Dissatisfied)	0.96	0.98	0.97	14,668
1 (Satisfied)	0.97	0.94	0.95	11,230
Overall Accuracy	-	-	**0.96**	25,898
Macro Average	0.96	0.96	0.96	25,898
Weighted Average	0.96	0.96	0.96	25,898

Table 4: XGBoost Classification Report On Test Set

These results affirm the model's consistency and its suitability for deployment in practical scenarios, such as helping airlines analyze and predict customer satisfaction trends to improve their services. The evaluation confirmed that the model is both accurate and reliable, with strong potential for real-world applications in the airline industry.

8. Conclusion

The airline passenger satisfaction classification project provided valuable insights into customer satisfaction trends and the effectiveness of various machine learning models for predictive analysis. Through thorough data exploration and feature engineering, key patterns were uncovered, such as the significant influence of age and gender on satisfaction levels, and the high correlation between arrival and departure delays.

	Model	Accuracy	Precision	Recall	F1-Score	AUC
0	Random Forest	0.961917	0.971044	0.940904	0.955736	0.994161
1	Logistic Regression	0.834926	0.803012	0.824478	0.813604	0.896033
2	XGBoost	0.962593	0.970876	0.942671	0.956566	0.995149

Table 4: Comparison Between All Models

The Random Forest, Logistic Regression, and XGBoost models were evaluated, with XGBoost emerging as the most effective model, achieving an accuracy of 96.3% and an AUC-ROC score of 0.9951 on the validation dataset. Hyperparameter tuning further enhanced performance for both Random Forest and XGBoost, while Logistic Regression displayed limited improvements. This project underscores the importance of leveraging advanced models like XGBoost for classification problems where high accuracy is critical. The findings could potentially help airlines improve service quality and enhance the overall customer experience by focusing on key satisfaction factors identified during the analysis. The robust methodology and results validate the potential of machine learning in solving real-world classification problems efficiently.

Referencing

[1]Kaggle. (n.d.). *Airline Passenger Satisfaction Classification*. Available at: https://www.kaggle.com/code/sharatsk/classification-airline-passenger-satisfaction

[2]Kaggle. (n.d.). *Flight Passenger Satisfaction: EDA and Prediction*. Available at: https://www.kaggle.com/code/chandrimad31/flight-passenger-satisfaction-eda-and-prediction

[3]Kaggle. (n.d.). *Airline Passenger Satisfaction Dataset*. Available at: https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

[4]OpenAI. (2024). *ChatGPT (December 2024 version)*. Available at: https://openai.com/chatgpt [Accessed 9 Dec. 2024].