Machine Learning-Based Prediction of Airline Passenger Satisfaction

*Abstract* - The aim of this project is to predict and analyze air passenger satisfaction based on demographic and service factors. Supervised classification methods in the form of Logistic Regression, Random Forest, KNN and XGBoost are utilized for satisfaction level prediction. Exploratory data analysis and correlation analysis are performed before the modeling process to highlight key variables for satisfaction. Missing values are handled properly and features on which the models will be trained are selected. The models are compared on train and test sets based on performance measures like accuracy, F1 score, sensitivity, specificity, and AUC.

Keywords—Airline Passenger Satisfaction, Machine Learning Models, Random Forest, XGBoost, Accuracy, R, Python.

# INTRODUCTION

The airline passengers are asked to complete a survey after every flight, rating the various aspects of services provided. Based on these service ratings and customer attributes, this study aims to frame a predictive model that will be able to predict if a potential customer would be satisfied or not with his/her overall flying experience. The customer satisfaction is categorized into two classes: "Satisfied" and "Neutral or Dissatisfied," making the issue a binary classification problem.

The goal of the project is not only to identify the key drivers of satisfaction, but also to inform customer experience initiatives and operation-level decision-making within airlines through the application of machine learning techniques.

# LITERATURE REVIEW

There is a number of machine learning-based airline passenger satisfaction prediction studies. The majority of the existing research relies on survey data sets and narrows down to using a limited choice of algorithms like Decision Trees or Random Forest. For instance, the data set in this report is in the public domain on Kaggle and had already been investigated with a Random Forest approach and simple performance assessment procedures. But these studies generally do not encompass extensive validation methods and model comparisons of wider range, and therefore may produce biased or slanted conclusions. This project, nonetheless, utilizes a wider variety of supervised learning algorithms such as Support Vector Machines (SVM), Artificial Neural Networks (ANN), Random Forest, and XGBoost.  
 It is supported by cross-validation, dealing with missing data in a proper manner, and feature selection to ensure robustness. The top-performing model among the models tried was XGBoost with 93% accuracy and a wide margin of improvement over previous results. The findings of this research not only confirm the significance of machine learning for the service quality analysis but also present managerial implications to airlines when they pursue customer satisfaction enhancement through data-driven decision-making.monthly.

# METHODOLOGY

1. *Dataset*

The dataset obtained from Kaggle, an open-source data platform. There are 103,594 observations and 24 variables with first variables being customer’s id, and represent different elements of airline passenger experience and satisfaction. The data had already been gathered through post-flight surveys where the passengers gave ratings of the various services offered by the airline.

The data set consists of demographic variables (gender, age, etc.), travel information (flight distance, class of travel, etc.), and service-specific measures (in-flight entertainment, seat comfort, check-in service, etc.) scaled on a 0 to 5 scale—with 0 representing "not applicable" and 1 through 5 representing increasing levels of satisfaction. The target variable is satisfaction, which is a binary categorical variable with two levels: Satisfied and Neutral or Dissatisfied.

There are 5,180 missing values are evenly spread across all columns and reflect potential respondent-level incompleteness. Proper imputation methods exist for such missing values during the preprocessing stage.

A brief overview of the principal variables is as follows:  
• Gender: Passenger's gender (Male, Female)  
• Customer Type: Loyalty status (Disloyal, Loyal)  
• Age: The actual age of the passengers  
• Type of Travel: Personal Travel, Business Travel

• Class: (Business, Eco, Eco Plus)

• Flight distance: The flight distance of this journey

• Service Ratings (e.g., Food and Drink, Cleanliness, Inflight Entertainment, Online Boarding): Satisfaction ratings for individual services (0: Not Applicable; 1–5: Satisfaction level)

• Departure/Arrival Delay in Minutes: Delay times in minutes

• Satisfaction: Binary outcome variable (Satisfied vs. Neutral or Dissatisfied)

1. *Descriptive Statistics*

The first thing to obtain to better understand the data is descriptive statistics. A summary of the descriptive statistics of the numerical values ​​is given in Table 1 below.

A screenshot of a computer

AI-generated content may be incorrect.

Table 1: Descriptive Statistical Summary of Numerical Data

The descriptive statistics of the airline data show the spread, central tendency, and potential skewness of the variables.

Age is distributed highly symmetrically with a mean of 39.35 years and a median of 40 years and does not show any extreme skew. The range is 7 to 85 years, showing wide demographic coverage. Flight Distance also follows a balanced distribution with mean (1189.72) and median (843) very close to one another, although it might have a slight right skew since it has a very long tail (range of 4983 km). All satisfaction ratings like Inflight Cleanliness, Online Boarding Seat Comfort, have scores rated on a 0 to 5 scale.

The distributions of all but a few variables are very slightly right-skewed with mean values very slightly below the medians, showing a slight bias towards a tendency to prefer lower ratings by a small number of travelers. For example, Seat Comfort has a median of 4 and a mean of 3.44. That is, while most passengers rated it positively, some rated it low enough to lower the mean. Departure and Arrival Delays have ranges of outliers (i.e., departure delay up to 1592 minutes) with means (14.75 and 15.18) much greater than medians (both 0). This is indicative of a highly right-skewed distribution with most flights departing and arriving on time but for a few severely delayed flights causing long tails.

Service features such as Inflight Service, Check-in Service, and Cleanliness all possess means and medians within the 3–4 range, which suggests net neutral to positive satisfaction with moderate variability. All of the variables contain 5180 missing values, and they must be included in any further analysis in order to obtain accurate results.

A screenshot of a computer screen

AI-generated content may be incorrect.

Table 2: Descriptive Statistical Summary of Categorical Data

The distribution of categorical variables for the airline dataset shows broad trends across demographic as well as service groups. The Gender category is relatively balanced with 48.24% Female and 46.76% Male. Most customers are Loyal Customers at 77.67% in Type of Customer. The majority of the data is from business travelers. This can create differences in satisfaction levels, for example, business travelers may be more demanding. The Class variable shows that passengers are mostly in Business (45.38%) and Eco (42.77%) class. Eco Plus is the least, occurring only 6.85%. The Satisfaction levels show that there is a difference between the two classes, but it is not too extreme. If we look at the ratios, this difference may have an effect during the training of the model, but it is not an extreme imbalance. In all the categorical variables, around 5% of the values are missing (NA), i.e., normal data missingness in features.

1. *Exploratory Data Analysis*

This section includes 5 research questions analyzing the dataset's structure and relationships between variables.

1. *What is the relationship between gender and satisfaction?*

A graph of a person and person

AI-generated content may be incorrect.

Figure 1: Bar Plot of Customer Satisfaction by Gender

When we look at the gender-based satisfaction, we actually see a similar distribution From here, we can say that gender does not have a direct effect on satisfaction.

A close up of a test

AI-generated content may be incorrect.

Table 3: Chi-Squared Test Result for Gender and Satisfaction

A Chi-squared test was used to quantify the relation between gender and satisfaction. The p-value is much smaller than the critical level of 0.05. Thus, we reject the null hypothesis and conclude that there is a statistically significant association between gender and satisfaction.

1. *What is the relationship between type of travel and satisfaction?*

A graph of travel type

AI-generated content may be incorrect.

Figure 2: Bar Plot of Customer Satisfaction by Type of Travel

Here it can be observed that the majority of the personal travelers fall under the 'dissatisfied or neutral' category, implying that overall travel experience is not being fulfilled to a complete extent. But the business travelers have a much larger proportion of being satisfied. This difference could be because of differences in expectations from travel, needs for service, or class of service that each group is likely to undertake. Business travelers might be experiencing superior services such as business class travel or priority services that could lead to higher satisfaction.

A close up of a test

AI-generated content may be incorrect.

Table 4: Chi-Squared Test Result for Travel Type and Satisfaction

The p-value(< 0.05) indicates a statistically significant relationship between travel type and customer satisfaction.

1. *What is the relationship between class and satisfaction?*

A graph with red and blue squares

AI-generated content may be incorrect.

Figure 3: Bar Plot of Customer Satisfaction by Type of Class

When we look at this plot, we can see that the majority of business class passengers are satisfield, while the majority of eco class passengers are dissatisfied.

A close up of a test

AI-generated content may be incorrect.

Table 5: Chi-Squared Test Result for Type Class and Satisfaction

The p-value(< 0.05) indicates a statistically significant relationship between travel type and customer satisfaction.

A chart with purple squares and numbers

AI-generated content may be incorrect.

Figure 4: Heatmap of Customer Satisfaction by Customer Type and Travel Class

This graph shows the satisfaction level according to customer type and flight class. We can see that the majority of neutral or dissatisfied passengers are loyal customers flying in eco class. We can also see that the majority of satisfied passengers are loyal customers flying in business class. From here, we can actually say that flight class has a very big effect on satisfaction. At the same time, we can say that disloyal customers are generally dissatisfied for each flight class.

A graph of a number of travel

AI-generated content may be incorrect.

Figure 5: Heatmap of Satisfaction by Travel Type and Class

This graph shows the satisfaction percentages according to the type of travel and the flight class. We can say that the majority of the neutral or dissatisfied passengers are those flying in eco class and those traveling individually. Similarly, we can say that the majority of the satisfied passengers are those flying in business class and those traveling for business. From here, we can say that both the flight class and the type of travel are important factors in satisfaction.

1. *Is there a significant association between Inflight Entertainment and overall satisfaction?*

A graph with different colored squares

AI-generated content may be incorrect.

Figure 6: Boxplot Plot of Satisfaction and Inf. entertainment

This boxplot shows the distribution of ‘Infight entertainment’ by customer satisfaction. ‘Inflight entertainment’ is spread out and lower for the 'Neutral or dissatisfied' group. The median is around 3. There is a high spread between the lower and upper quartiles (IQR). This may indicate there are different perceptions among users. Some data is extremely low.

In ‘Infligt entertainment’ scores for 'Satisfied' respondents, the scores look higher and more credible. Median is almost equal to 5, which reflects that seat comfort was rated exceptionally well by the majority of the respondents. Further, the scores are not spread out, i.e., box has low thickness, which means that all the ratings are consistent. Low number of respondents with low ratings.

A close up of a test

AI-generated content may be incorrect.

Table 6: Wilcoxon Test Result for In.entertainmet and Satisfaction

1. *How does customer satisfaction vary across different age segments?*

A graph of different colored bars

AI-generated content may be incorrect.

Figure 7: Boxplot Plot of Satisfaction and Inf. entertainment

The proportion of dissatisfied passengers is higher than that of satisfied passengers in all age groups. Dissatisfaction is particularly pronounced in the youth (0-24) and young adult (24-45) groups. This may indicate that age-specific service improvement strategies are needed to increase customer satisfaction.

A close up of a test

AI-generated content may be incorrect.

Table 7: Chi-Squared Test Result for Age Groups and Satisfaction

Correlation plots are a fundamental component of exploratory data analysis (EDA) because they provide a simple visual indication of the relationships between variability.

A graph of a graph with a bar chart and a red and blue bar chart

AI-generated content may be incorrect.

Figure 8: Correlation plot with other variables with satisfaction

Here in the graph we gladly examined the correlation of all features with the target(satisfaction) variable. And I saw that online boarding was the variable with the highest correlation in general. This shows that improvements in customer-oriented services can make a bigger difference in satisfaction.

1. *Missingness*

We had approximately 7.7% missing values in our dataset. Identifying the missingness mechanism is critical to concluding an effective strategy to replace the missing data. The Little's MCAR test was performed to see if the missing data were Missing Completely at Random (MCAR). The null hypothesis of the test provided is that the data are MCAR. Our test provided a very low p-value (p < 0.001), which is smaller than the alpha. Therefore we rejected the null hypothesis, indicating that the missing data are not MCAR.

A black and yellow lines on a black background

AI-generated content may be incorrect.

Figure 9: Missigness Pattern Plot

To further investigate the mechanism of missingness, we visualized the pattern of missing data using the mice package in R. No evident structure or inter-variable dependency existed in the missingness pattern. Since missing values represented association with observed data but not with missing data, we considered the mechanism of missingness to be Missing At Random (MAR).

We avoided using any deletion-based method such as listwise deletion, which would result in duplicate information loss. Therefore, we applied multiple imputation using mice package and random seed 123 to allow for reproducibility. Then, we applied predictive mean matching (pmm) for continuous data such as Age, Flight.Distance, and ratings for service quality since the distributions were non-normal and skewed. In categorical variables, standard imputation functions of the mice package were used: logreg for binary(e.g, Gender, Customer.Type) and polyreg for multi-class (e.g., Class) variables. These imputation functions ensure the imputed values are appropriate in variable type and reflect the observed patterns in the data.

Hence, the final imputed dataset airline\_data has no missing values and preserves the original distribution characteristics.

1. *Feature Selection*

A graph of flight status

AI-generated content may be incorrect.

Figure 10: Correlation Matrix of Numeric Variables

To identify multicollinearity among the variables, we tested for pair-wise Pearson correlation coefficients. We found an exceptionally high positive correlation (r = 0.97) between Departure Delay in Minutes and Arrival Delay in Minutes. Since high correlation indicates redundancy, we removed Departure Delay in Minutes from the dataset to avoid multicollinearity and model complexity reduction.

Likewise, there was a high correlation (r = 0.72) between Inflight Wifi Service and Ease of Online Booking. Though lower, this level of correlation might also lead to redundancy and lower generalizability of the model. We therefore also removed Inflight Wifi Service from the dataset.  
 These steps enhance model interpretability and lower the risk of overfitting.

1. *Outlier Detection*

A graph of blue and red dots

AI-generated content may be incorrect.

Figure 11: Visualization of Outliers Detected via LOF

To detect local anomalies, we employed the Local Outlier Factor (LOF) method on the numeric features with minPts = 11. Outliers were termed with a value of 1.5. It identified 208 cases as outliers and removed them from the data set.  
 We graphed the outliers on a scatter plot of Age vs Flight Distance so that outliers were clearly separated from the main data cloud. Outliers were later dropped after being identified, and the cleaned data were used for further analysis.

1. *Encoding and Transformation*

In addition, in our project's data preprocessing step, encoding and transformation operations were conducted.

Firstly, ordinal variable class was numerically encoded based on flight class hierarchy (Eco = 0, Eco Plus = 1, Business = 2). Secondly, target variable satisfaction was binary encoded like this: satisfied = 1, neutral or dissatisfied = 0.

One-hot encoding was applied to the categorical variables Customer.Type, Gender, and Type.of.Travel. The first category of each variable was removed to prevent multicollinearity (remove\_first\_dummy = TRUE).  
 Finally, all the numerical variables were normalized using Min-Max scaling, where every feature is scaled to the [0, 1] range. This normalization makes sure that all features are in the same scale, and model performance is improved.

1. *Modelling*

To assess the generalization performance of the model and avoid overfitting, the data were split into training and test sets by stratified sampling. That is to say that 80% of the data were reserved for training and 20% for testing, with the same original distribution as the target variable (satisfaction).

1. *Binary Logistic Regression*

The factors affecting customer satisfaction were investigate using a binary logistic regression model. The model was statistically significant, and several predictors showed meaningful effects.

A screenshot of a computer

AI-generated content may be incorrect.

Table 8: Summary of Logistic Regression Results

The exploratory data analysis (EDA) and logistic regression outcomes are complementary and corroborative in explaining the factors of customer satisfaction. The EDA reported a small, but positive, gender effect on satisfaction supported by the model's observation that male customers have a slightly higher probability of being satisfied than females Travel type analysis revealed that business travelers are more satisfied than personal travelers, a trend strongly backed by the model with a strong negative effect of personal travel on satisfaction. Furthermore, the positive relationship between class and satisfaction in EDA is confirmed by an equally strong positive coefficient for class across regression analysis. Inflight entertainment was still positively associated with satisfaction in exploratory and modeling stages. Age had a negative effect on satisfaction, where older consumers were less satisfied, which was a repeated observation in both analyses. The model also showed that service quality dimensions such as online boarding, in-flight services, and cleanliness have strong positive effects on satisfaction, while arrival delays significantly lower satisfaction. Customer loyalty was also established as a key driver of satisfaction by both the EDA and regression findings. Overall, all the inferences made based on the EDA were confirmed by the logistic regression at a quantitative level, which strongly showed the key drivers of customer satisfaction.

A white rectangular object with black text

AI-generated content may be incorrect.

Table 9: Logistic Regression Model Performance Summary

The model was 86.8% accurate on training data and 86.9% on test data. Training and test performances are in very close resemblance with each other, i.e., the model is not overfitting and is generalizable. Sensitivity (90%) and selectivity (82%) are in good balance. Dissatisfied passengers are predicted with high accuracy. Kappa coefficient is in the range of 0.72-0.73, indicating that the model performs much better than random success.

1. *Decision Tree*

A white rectangular object with black text

AI-generated content may be incorrect.

Table 10: Decision Tree Model Performance Summary

The decision tree is a powerful and interpretable approach to classification problems. It decides by recursively partitioning the data according to important features in a way that tries to optimize classification accuracy. The decision tree model is showing strong and stable performance on both the training and test datasets. The accuracy is high at 88.7% on the training set and 88.9% on the test set, suggesting good generalization with no indication of overfitting. The Kappa statistic, which corrects for agreement by chance, is also strong at 0.77 for training and 0.774 for test data, further supporting the reliability of the model. Sensitivity percentages are also proximate to each other, 85.9% on training and 86.0% on test sets, indicating the model's robustness in identifying the positive cases overall. Specificity, i.e., the model's proficiency in accurately identifying the negative cases, is also high and proximate to both training (90.9%) and test (91.2%) sets. Overall, performance metrics indicate that the decision tree model is accurate and stable and could be an effective predictive model for this data set.

1. *Random Forest*

A white sheet with black text

AI-generated content may be incorrect.

Table 11: Random Forest Model Performance Summary

Random Forest classifier was trained to predict customer satisfaction based on service characteristics. Its accuracy on the test set was 90.7%, which shows high discrimination between dissatisfied and satisfied customers. High specificity (94.4%) shows that the classifier particularly performs well at predicting dissatisfied customers correctly, which may be valuable in customer retention systems.

1. KNN

A white sheet with black text

AI-generated content may be incorrect.

Table 12: KNN Model Performance Summary

The specificity levels are extremely high (96.1% train, 94.7% test), which means that the model is extremely good at correctly identifying the negative class (dissatisfied customers). This implies low false positive rate, which is good if marking a dissatisfied customer as satisfied in error would be business risk.

The KNN model performs equally well on train and test data, with equally balanced sensitivity and specificity, and high overall accuracy. No significant drop in performance from train to test, i.e., low overfitting. But among them all, KNN still may be computationally more costly at prediction time compared to tree-based methods (like Random Forest or XGBoost), especially when dealing with large data.

1. *XGBoost*

A white sheet with black text

AI-generated content may be incorrect.

Table 13: XGBoost Model Performance Summary

The XGBoost model exhibited a very good performance in terms of predicting customer satisfaction. The model achieved 94.8% accuracy on the training dataset and 92.8% on the test dataset, and hence the model generalizes well without overfitting very much. The Cohen's Kappa values were also strong — 0.893 training and 0.853 test — which indicates an enormously high agreement of predicted and actual labels, much higher than would be by chance.

In addition, sensitivity on the test set was 90.3%, and this means the model was quite successful in capturing most of the satisfied customers. Further, specificity was 94.7%, meaning that it was as good in detecting dissatisfied customers with a low false positive rate.

1. *Performance Comparison on Test Data*

A table with numbers and a few black text

AI-generated content may be incorrect.

Table 14: Performance Summary of Machine Learning Models

Overall, the XGBoost model provides a robust and justifiable classification with excellent predictability capacity for both classes. With its high rate of accuracy, train and test reliability, as well as high class discrimination, XGBoost could be among the most accurate models that can forecast customer satisfaction in this study.

A graph with blue and grey bars

AI-generated content may be incorrect.

Figure 12:Plot of Model Comparision by F1 Score

The plot yields a comparison of the F1 score of different machine learning models. It can be seen from the plot that XGBoost had the highest F1 score (~0.92), which represents the best performance in terms of false positives and false negatives.  
 Ensemble models such as SGBoosting, Random Forest, and Bagging Classifier also performed well, confirming the merit of combining various models to improve prediction accuracy and reliability.

Conversely, simpler models such as Logistic Regression and Linear SVC scored the lowest F1 scores, which may indicate limitations in identifying complex, non-linear patterns in the data. While Decision Tree worked fairly well, its ensemble counterparts (Random Forest and Bagging Classifier) worked much better, suggesting that model averaging enhances generalization.

1. *Feature Importance from the best-performing model*

A bar graph with text

AI-generated content may be incorrect.

Figure 13: Feature Importances Based on XGBoost Classifier

The highest ranked characteristic by the XGBoost model was Online boarding, which speaks to the growing role of digital components in services as part of air travel. Very highly ranked were type of travel, class, and type of customer, suggesting that satisfaction levels are quite contingent on context and purpose of trip as well as customer loyalty. These results are consistent with exploratory data analysis trends in which travel and demographic variables were also highly connected with satisfaction. Service-specific factors such as onboard entertainment and ease of web check-in were also among the leading contributors, further reinforcing the finding that in-flight service quality is insufficient but the ease of pre-flight processes also plays a very significant role in the formation of overall satisfaction.

# RESULTS

The analysis began with an exploratory data analysis, which revealed a marginal imbalance in the target variable, with approximately 56% of the customers being categorized as neutral or dissatisfied, and 44% as satisfied. There were some significant relationships between customer satisfaction and attributes such as travel class, travel type, seat comfort, and age groups. For instance, business-class passengers overall were more satisfied than economy-class passengers. Missing values existed for approximately 7.7% of the data, particularly on service factors; missing values were treated carefully with appropriate imputation and feature removal methods.

Following feature selection and preprocessing, features such as Departure Delay in Minutes and Inflight Wifi Service were removed as they were low in terms of their predictive significance and also suffered from quality issues within the data. Local Outlier Factor outlier detection identified 208 outlier observations that were settled for model accuracy and robustness.

The following machine learning algorithms were compared and trained: Decision Tree, Random Forest, K-Nearest Neighbors, Bagging Classifier, and XGBoost. The best performance was recorded in the XGBoost model with test accuracy equal to 92.8%, Cohen's kappa equal to 0.85, sensitivity equal to 90.3%, and specificity equal to 94.7%. These are good balanced classification ability and predictive accuracy. K-Nearest Neighbors and Random Forest models also gave close results with test accuracy levels ranging from 89–90% and the same recall and specificity levels but slightly lower than XGBoost. Bagging Classifier and Decision Tree models performed moderately but not as close to the ensemble models.

Tree model feature importance, especially, pinpointed the main drivers of customer satisfaction, including seat comfort, quality of onboard service, inflight entertainment, and check-in service. The findings are consistent with empirical statistical findings and are self-evident areas for airline service improvement.

# CONCLUSION

This study aimed to examine airline customer satisfaction using a large dataset with over 100,000 observations involving heterogeneous service and demographic factors. After exploratory data analysis, feature selection, and cleaning, various machine learning models were train in the hope of identifying customer satisfaction.

Among all the models that were executed, the best-performing model was XGBoost with maximum accuracy and specificity-sensitivity balance. The ability to manage complex nonlinear interaction and feature interaction helped it perform at the highest level. K-Nearest Neighbors and Random Forest models had good performance but with slightly worse generalization performance.

The research concluded that seat comfort, online boarding, and punctuality aspects are the greatest predictors of satisfaction. These can be utilized by the airlines as top priorities for areas to be improved in customer experience and loyalty. Although the models were acceptable, imputation of missing values and possible outliers was performed in a try to make good predictions. There is other work that can be done using more sources of data, more complex feature engineering processes, and ensemble or hybrid models in a try to try higher accuracy.

Overall the project demonstrates the potential of methodology by data for explanation and forecasting customer satisfaction and which can be used in strategic decision-making in the aviation industry.

# References

[1] Dziura, J. D., Post, L. A., Zhao, Q., Fu, Z., & Peduzzi, P. (2013). Strategies for dealing with missing data in clinical trials: from design to analysis. *The Yale journal of biology and medicine*, *86*(3), 343.

[2] Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In *Machine learning proceedings 1992* (pp. 249-256). Morgan Kaufmann.

[3] teejmahal20. (n.d.). *Airline Passenger Satisfaction* [Data set]. Kaggle. <https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction>