

Contents

[Introduction 2](#_Toc152612397)

[Critical Business Questions 3](#_Toc152612398)

[Data Collection Process 3](#_Toc152612399)

[Methodology 3](#_Toc152612400)

[Missing Values and Planned Measures 4](#_Toc152612401)

[Data Preprocessing Steps 5](#_Toc152612402)

[Data Summary and Visualizations 5](#_Toc152612403)

[Data Analysis 7](#_Toc152612404)

[Results 10](#_Toc152612405)

[Conclusion and Recommendations 11](#_Toc152612406)

[Conclusion 12](#_Toc152612407)

# Table of Figures

[Figure 1 Customer satisfaction by class 6](#_Toc152612436)

[Figure 2 Customer satisfaction based on flight distances 7](#_Toc152612437)

[Figure 3 Customer satisfaction based on delays 7](#_Toc152612438)

[Figure 4 Customer satisfaction and inflight entertainment 8](#_Toc152612439)

[Figure 5 Feature Importance from RFC 8](#_Toc152612440)

[Figure 6 Top Feature Importance from RFC   9](#_Toc152612441)

[Figure 7 Confusion matrix from RFC 9](#_Toc152612442)

[Figure 8 Frequency from Logistic Regression Analysis 10](#_Toc152612443)

[Figure 9 Confusion matrix and Accuracy from Logistic Regression Analysis 11](#_Toc152612444)

# Introduction

In the aviation industry, providing an exceptional travel experience is not just about getting passengers from one destination to another, it is about creating memorable and enjoyable journeys for everyone. As one of the industry leaders in aviation, we are committed to continuously improving our services and ensuring that our passengers' needs and preferences are met to the highest standards. In this project, a detailed analysis has been carried out based on the available data for customer satisfaction in terms of various inflight/service factors of the airline. Using the analysis of customer satisfaction survey data retrieved from kaggle.com we will answer two critical business questions: **What are the most important factors affecting customer satisfaction on a flight?** and **can we predict if a customer will be satisfied with the airline services based on current services offered by the airlines?** These questions are important for the airline to answer because it will help them understand what service improvements need to be made, identify the best practices for resource allocation, and overall make better informed data-driven business decisions. We will be working with a dataset that contains 129,880 passenger or customer responses to an airline satisfaction survey.

The selected dataset includes data on the following variables:

* Age
* Gender
* Type of Travel
* Class
* Flight Distance
* Inflight Wi-Fi Service
* Departure/Arrival Time Convenient
* Ease of Online Booking
* Gate Location
* Food & Drink
* Online Boarding
* Seat Comfort
* Inflight Entertainment
* On-Board Service
* Leg Room Service
* Baggage Handling
* Check-In Service
* Inflight Service
* Cleanliness
* Departure Delay in Minutes
* Arrival Delay in Minutes
* Satisfaction

# Critical Business Questions

1. What are the most important factors affecting customer satisfaction on a flight?
2. Can we predict if a customer will be satisfied with the airline services based on current services offered by the airlines?

# Data Collection Process

Data was downloaded from Kaggle.com. Python was used to carry out a Random Forest Classifier to find out the relevant importance of the various factors in the satisfaction of the airline's current passengers. Based on the analysis, a detailed proposal with various improvement options will be presented to the management.

# Methodology

To carry out the Random Forest Classifier for predicting customer satisfaction following steps will be followed:

1. Loading the Dataset:

Firstly, import the dataset into Excel/XLMiner. Examine and clean the data, handling missing values, and outliers, create usable data using dummy variables. Ensure that the target variable "Satisfaction" is binary, where "satisfied" is coded as 1 and "neutral or dissatisfied" is coded as 0.

1. Data Exploration:

Perform exploratory data analysis (EDA) to understand the distribution of variables. Then, create visualizations such as histograms, bar charts, and correlation matrices to gain insights into the dataset.

1. Loading the Dataset into Python:

The code begins by using pd.read\_csv to load the dataset. Information regarding airline passenger satisfaction can be found in the dataset.

1. Exploring the Dataset:

A preview of the data is given by displaying the first few rows of the dataset using print(pdf. head()). print(df.isnull()). To find any missing values in the dataset, use sum().

1. Handling Missing Values:

The code divides the goal variable (y) and the features (X), assuming that satisfaction' is the target variable.

One-hot encoding is used in the features to transform the categorical variables to numerical representation (pd.get\_dummies(X)).

Using SimpleImputer, missing values in the dataset are imputed using the mean technique. By doing this, the dataset is guaranteed to include no NaN values.

1. Splitting the Dataset:

Train\_test\_split is used to divide the dataset into training and testing sets.

Training a Random Forest Classifier:

A random state of 42 and 100 trees is used to initialise a RandomForestClassifier. Next, using model.fit(X\_train, y\_train), the model is trained on the training set.

1. Feature Importance:

Model.feature\_importances\_ is used to compute each variable's feature importance in forecasting customer satisfaction.

1. Visualizing Feature Importance:

The feature importance values are then displayed using a matplotlib (plt. barh) horizontal bar plot. Validate the model's ability to predict customer satisfaction.

1. Displaying Top Features:

The feature importance DataFrame (importance\_df) is sorted in descending order to display the top features influencing customer satisfaction.

It seems that the model does a good job of forecasting customer satisfaction based on the present services provided by the airline, based on the accuracy metrics and classification report. Summarize the findings and highlight the key factors that significantly influence customer satisfaction based on the logistic regression model. Use visualizations and insights to present the results effectively.

1. Recommendations:

Based on the identified important factors, provide recommendations for the airline to improve customer satisfaction. Discuss potential strategies for enhancing the passenger experience.

For second code, the methodology used in excel miner. For the limitation of excel miners capacity to 6000 rows, we have used first 6000 row for carrying our logistic regression analysis. We have selected the satisfaction as the output variable and other variables as input. We used 60/40 training/validation with rescaling data to normalization. From running the model, we have received the results from the miner.

# Missing Values and Planned Measures

It has been noticed that there are some missing values in various rows for various columns. To combat this situation, two strategies can be taken. If the missing values are relatively few and randomly distributed, removing rows with missing data might be considered but with care because removing too many rows may lead to loss of valuable information. Replacing missing values with a calculated or estimated value can also be considered.  Common imputation methods include mean, median, or mode imputation for numerical data, or using the most frequent category for categorical data. It should be mentioned that imputation introduces assumptions about the data. For the simplicity of the project and having a big dataset with few missing data points that is also randomly distributed, it is rational to remove the whole row with any missing data point. To do so, a simple code in R can produce a filtered dataset. The code is {df <- na.omit(df)}.

# Data Preprocessing Steps

To pre-process our data in Excel, we first formatted the dataset into a table. We converted our target variable, “Satisfaction”, into a binary format where "satisfied" is coded as 1 and "neutral or dissatisfied" is coded as 0 using the function =IF(Y2="neutral or dissatisfied",0,1). We did this to fit the model requirement, to simply analysis, and to ensure that we will be able to compute performance measures like accuracy, precision, recall, and F1 score in our logistic regression model. Afterwards, we used domain knowledge, EDA, and feature selection techniques (e.g., correlation analysis) to select the most relevant variables that will be of importance in our analysis.

# Data Summary and Visualizations

In this section, some of the insights of correlated variables with customer satisfaction and the critical business question have been visualized to show their relation and importance.

A diagram of a pie chart

Description automatically generated

Figure 1 Customer satisfaction by class

Figure 1 allows us to view how many customers responded with 'Satisfied' based on which type of class they flew with. This table can be filtered to narrow down specific age groups and we can use it to predict customer satisfaction based on which class the age group booked.

A graph with blue bars and white text

Description automatically generated

Figure 2 Customer satisfaction based on flight distances

An analysis of Figure 2 shows that there is a correlation between customer dissatisfaction and shorter flight distances. This reveals the necessity for strategic improvements in our short-flight experience to improve overall satisfaction.

A graph with blue and orange bars



Figure 3 Customer satisfaction based on delays

Figure 3 allows us to visualize the average delay times for our satisfied and dissatisfied customers. This information can be analyzed and used for strategic planning of how to deal with potential delays in the future.

A graph of a service rating

Description automatically generated with medium confidence

Figure 4 Customer satisfaction and inflight entertainment

Figure 4 shows us the ratings of a few of our in-flight services organized on whether they were overall satisfied or unsatisfied. This information can help us project what overall satisfaction level future airline customers will report based on these variables in-flight.

# Data Analysis

The results of the Random Forest Classifier detailed in the first code are as followed:

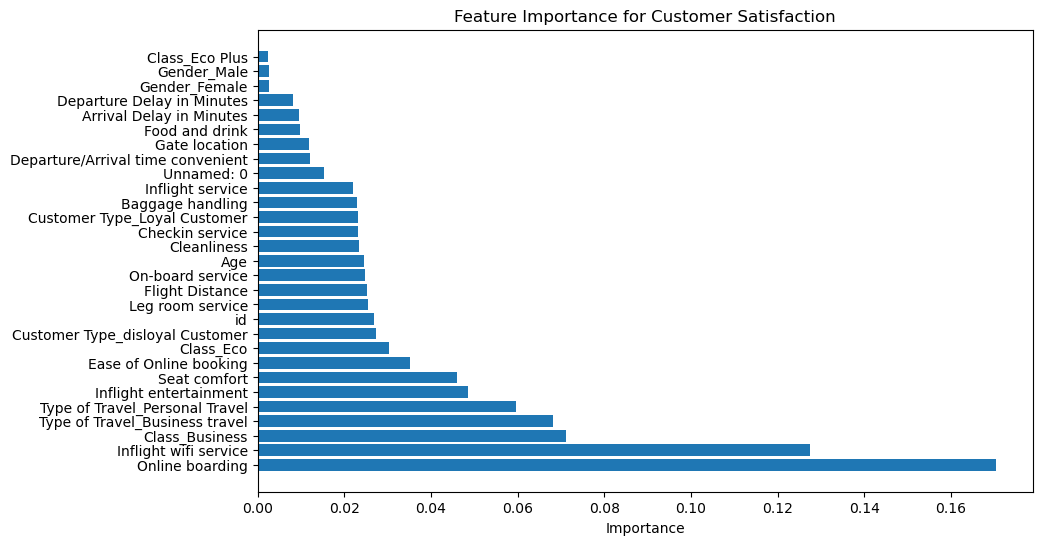


Figure 5 Feature Importance from RFC

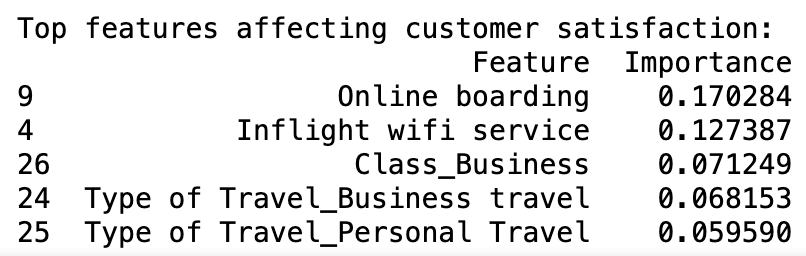


Figure 6 Top Feature Importance from RFC

The result of the second code is as followed:

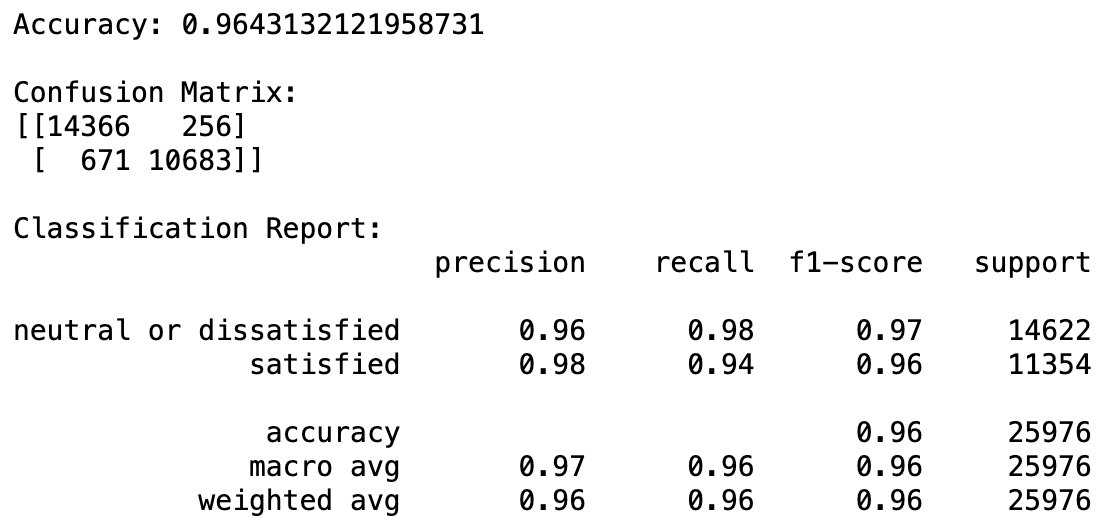


Figure 7 Confusion matrix from RFC

Alternatively, we have conducted a logistic regression analysis to classify the model. This logistical regression model contains only 6,000 rows because our dataset is too large to process through Excel/XLMiner, but the results still provide interesting results that mirror the above. The result of the logistical regression is below:

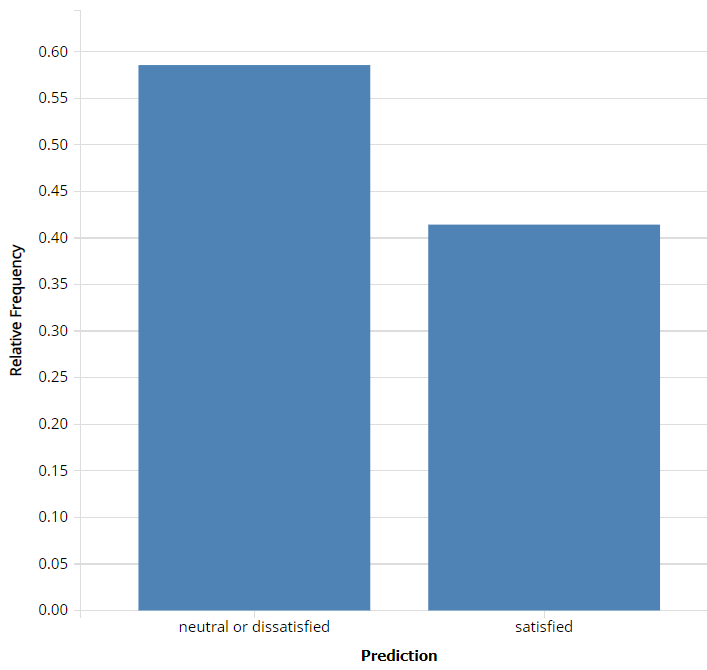


Figure 8 Frequency from Logistic Regression Analysis

Figure 8 shows that satisfied frequency is about 4. The accuracy is about 88%. So, RFC model is more appropriate and performing better on that case.

|  |  |  |  |
| --- | --- | --- | --- |
| **Confusion Matrix** | | |  |
| **Actual\Predicted** | **neutral or dissatisfied** | **satisfied** |  |
| **neutral or dissatisfied** | 1219 | 123 |  |
| **satisfied** | 186 | 871 |  |
|  |  |  |  |
| **Error Report** | | | |
| **Class** | **# Cases** | **# Errors** | **% Error** |
| **neutral or dissatisfied** | 1342 | 123 | 9.165424739 |
| **satisfied** | 1057 | 186 | 17.59697256 |
| **Overall** | 2399 | 309 | 12.88036682 |
|  |  |  |  |
| **Metrics** | |  |  |
| **Metric** | **Value** |  |  |
| **Accuracy (#correct)** | 2090 |  |  |
| **Accuracy (%correct)** | 87.11963318 |  |  |
| **Specificity** | 0.908345753 |  |  |
| **Sensitivity (Recall)** | 0.824030274 |  |  |
| **Precision** | 0.876257545 |  |  |
| **F1 score** | 0.849341784 |  |  |
| **Success Class** | satisfied |  |  |
| **Success Probability** | 0.5 |  |  |

Figure 9 Confusion matrix and Accuracy from Logistic Regression Analysis

# Results

In this section, we will provide an explanation of our results from our Random Forest Classifier and Accuracy/Classification Report we conducted on the dataset. We will also discuss how these results answer our critical business questions.

What are the most important factors affecting customer satisfaction on a flight?

From the Random Forest Classifier, we conducted using Python we found that according to our dataset of 100,000+ responses to an airline satisfaction survey, the top five factors that affect whether a customer will be satisfied or not are Online Boarding quality, Inflight Wi-Fi Service quality, if they are flying in Business class, and whether they are traveling for business or personal reasons. The least important factors affecting customer satisfaction are if they are flying Eco Plus, Gender, and Arrival/Departure Delay times.

Can we predict if a customer will be satisfied with the airline services based on current services offered by the airlines?

In the second code we ran on the dataset, we received results on the accuracy of the Feature Importance for Customer Satisfaction ranking and the models’ prediction accuracy overall.

Accuracy: The model has an overall accuracy of roughly 96.43%, which is substantially high. This indicates that, on the test set, the model accurately predicts consumer satisfaction or dissatisfaction around 96.43% of the time.

Confusion Chart:

10,683 cases were True Positives (TP) where the model accurately predicted, "satisfied."

14,366 cases of True Negative (TN) occur when the model accurately predicted "neutral or dissatisfied."

256 cases were classified as False Positives (FP) when the model predicted "satisfied" but the actual label was "neutral or dissatisfied."

False Negative (FN): 671 instances were classified as False Negatives (FN) when the model predicted "neutral or dissatisfied" but the actual label was "satisfied."

Classification Report:

Precision: 98% of the predictions for "satisfied" and 96% of the predictions for "neutral or dissatisfied" were made with accuracy. The precision of the favorable forecasts is represented by their accuracy.

Recall: 94% and 98%, respectively, of the predictions made for "satisfied" and "neutral or dissatisfied" were correct. Recall is a measure of the model's capacity to record all successful cases.

F1-score: Both classes had high F1-scores, which are the harmonic means of recall and precision. This suggests that recall and precision are well-balanced.

These results indicate that, given the features employed in the model, it performs well in forecasting customer satisfaction and has a high accuracy.

# Conclusion and Recommendations

From our analysis of the customer satisfaction survey dataset, we can conclude that the top features affecting customer satisfaction are identifiable and the prediction model can accurately predict satisfaction levels. To improve overall customer/passenger satisfaction we have made the following recommendations to the airline:

**Increase Investment in Enhancing Online Booking & Inflight Wi-Fi Technology**

From the Random Forest Classifier analysis, we found that the top two features that influence customer satisfaction are Online Booking and Inflight Wi-Fi quality. Investment of time and effort into improving these two technologies can look like optimizing the mobile app and website interface, increasing advertisements of the features to customers, or developing additional related technologies and services.

**Reevaluate Booking Class Options and Features**

Another interesting insight from the Random Forest Classifier is that customers who flew in Business Class were in the top factors affecting customer satisfaction, while customers flying in Eco Plus were the lowest factor affecting customer satisfaction. This stark contrast is evidence that the airline should reevaluate and determine what adjustments need to be made in the Class system as a whole entity. Possibly the company could fully eliminate the ‘Eco Plus’ class offering and instead focus its’ resources on enhancing the Business class experience.

**Tailor Services to the Traveller**

Lastly, to improve customer satisfaction we recommend that the airline tailor their services to the type of travel their customers are experiencing. Recognizing that personal and business travelers have different expectations for their journeys, the airline may consider investing in amenities aligned with these expectations, such as establishing business meeting lounges or family-call zones. Acknowledging and accommodating the diverse purposes of travel can significantly contribute to improved customer satisfaction.

# Conclusion

The Random Forest Classifier analysis of a vast airline satisfaction survey dataset has yielded critical insights into the determinants of customer satisfaction. Key factors, including Online Boarding and Inflight Wi-Fi quality, Business class travel, and the purpose of travel, emerged as significant influencers. The predictive model exhibited an impressive 96.43% accuracy, showcasing its reliability in anticipating customer satisfaction levels. The recommendations derived from the analysis suggest targeted investments in enhancing online booking and inflight Wi-Fi technology, a reassessment of Booking Class options with a potential focus on eliminating Eco Plus, and tailoring services to different traveler types. By implementing these strategic measures, the airline stands to enhance overall customer satisfaction, with a particular emphasis on aligning services with customer expectations and preferences.