IPSA

INSTITUT POLYTECHNIQUE DES SCIENCES AVANCÉES

SECOND SESSION : Mathematical tools for data science

June 17^{th} 2024

BARSCZUS ROBIN

Table of content

Abstract

In the wake of the global pandemic, the air transport industry has faced many challenges. Improved customer satisfaction strategies are needed to drive recovery and growth. This project is based on a database of air passenger satisfaction data, separated into train and test sets. The aim is to classify passenger satisfaction into three categories: satisfied, neutral or dissatisfied. Using machine learning techniques to analyse the data, this project identifies the key factors influencing passenger satisfaction and develops a predictive model.

Introduction

The global pandemic has had a severe impact on the air transport industry, with passenger numbers falling and customer expectations rising. As airlines strive to recover from this event, it is essential to understand and improve passenger satisfaction. This project focuses on analysing a dataset from a post-pandemic survey, which assesses various aspects of the passenger experience. The dataset has been previously divided into train and test subsets, which facilitates the development of predictive models.

The primary objective of this study is to classify passengers into one of three levels of satisfaction: satisfied, neutral or dissatisfied. To achieve this, we first perform an exploratory analysis of the training set data to discover the underlying patterns and significant predictors of satisfaction. We will then apply machine learning models to classify satisfaction levels.

Through this study, we aim not only to accurately predict passenger satisfaction, but also to highlight critical areas for service improvement, helping airlines to develop more effective strategies to meet changing customer expectations in a challenging economic environment.

1 Data Analysis

For our analysis, we used two primary datasets: the train dataset and the test dataset. These datasets include customer ratings collected from an airline to assess passenger satisfaction.

The datasets are made up of characteristics such as age, gender, flight distance, type of journey, class, etc. Finally, the passenger satisfaction rate, classified into two categories: "satisfied" or "neutral/dissatisfied".

The train data is used to train a Machine Learning model capable of classifying what factors are highly correlated to a satisfied passenger and predicting the latter's satisfaction.

1.1 General satisfaction

First of all, I imported and read the data using pandas.
I have carried out a number of analyses on this data, which will help us to understand the data as a whole.

Firstly, I wanted to point out the number of satisfied passengers compared to the number of dissatisfied or neutral passengers.

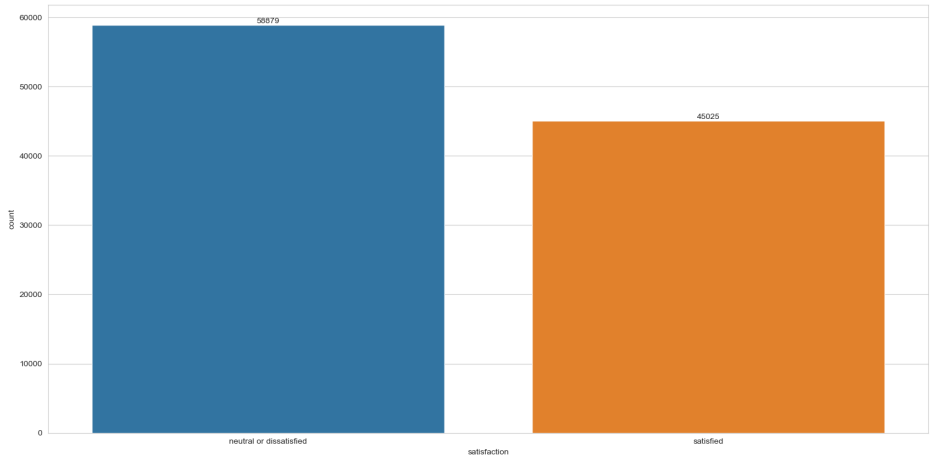

Figure 1: Distribution of satisfied or neutral/dissatisfied customers

As we can see on the previous graph, more people are satisfied (58879). But many people are neutral or unsatisfied (45025). This represents around 43% of customers, which is no mean feat for an airline. Which just goes to show how important our study is.

1.2 Customer gender

Next, we wanted to show the number of female customers and the number of male customers.

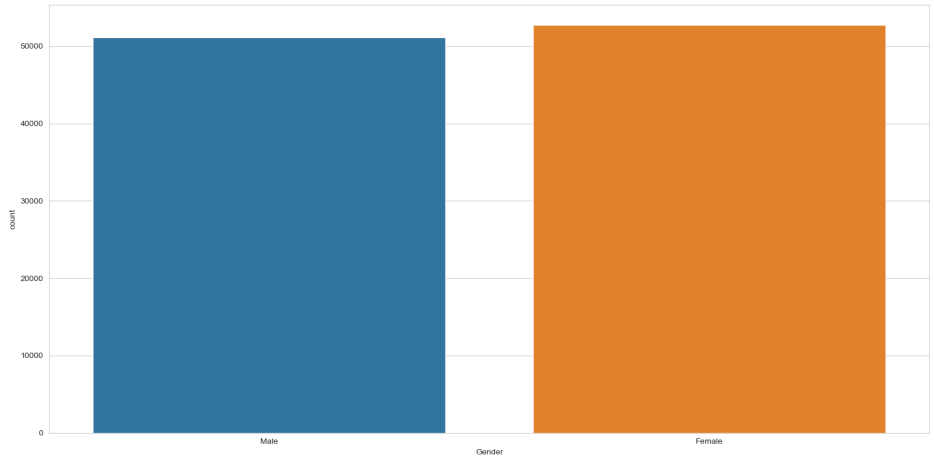

Figure 2: Customer gender

We can see that the distribution of gender is equitable. There are roughly the same number of female customers as male customers.

1.3 Customer satisfaction based on customer type

Here we simply show the satisfaction of customers according to their type, loyal customer and new customer.

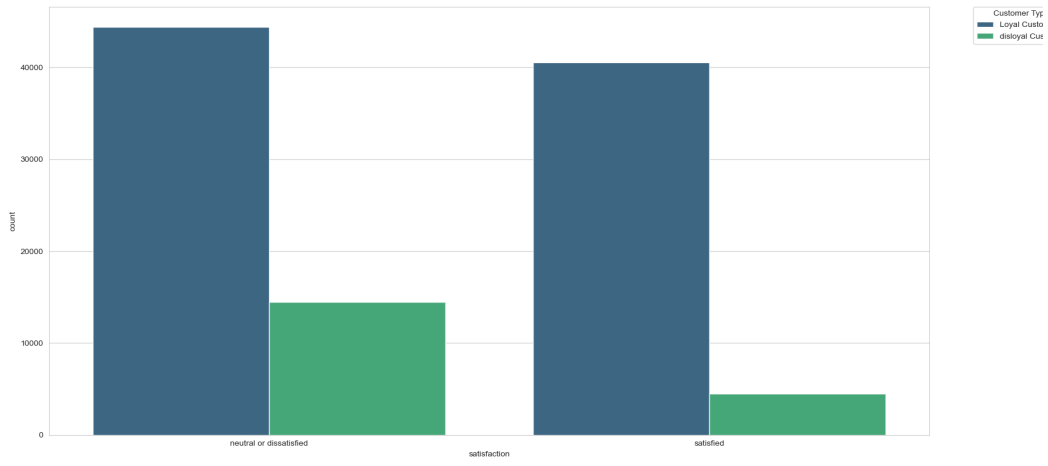

Figure 3: Customer satisfaction based on customer type

We can see that loyal customers are more likely to be dissatisfied, while new customers are evenly split between satisfied and dissatisfied.

1.4 Customer satisfaction base based on type of travel

By correlating customer satisfaction with the type of trip, we can draw the following graph:

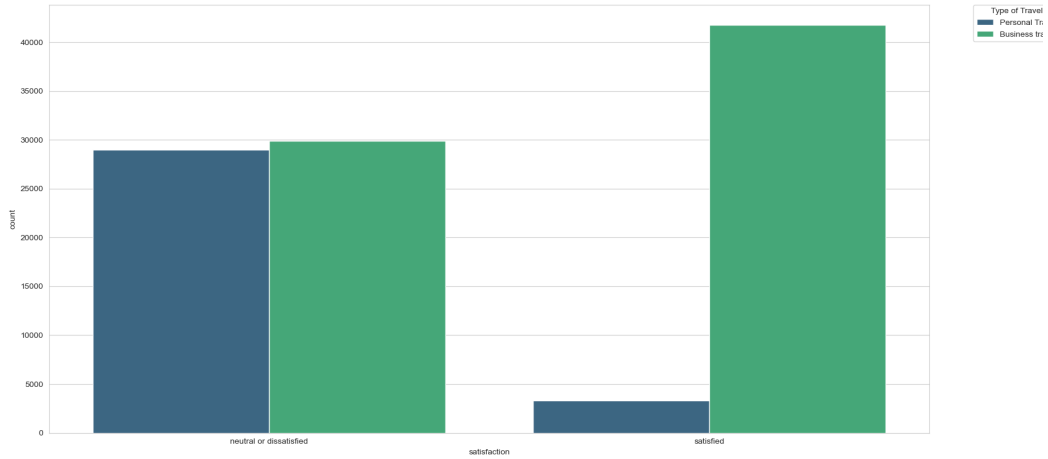

Figure 4: Customer satisfaction base based on type of travel

From this graph we can clearly see that customers are much less satisfied if they are travelling for personal reasons. Conversely, the majority of satisfied customers were travelling for business reasons.

1.5 Customer satisfaction base based on class of travel

We wanted here to show how the class of travel impact the customer's satisfaction. We show this in the following graph:

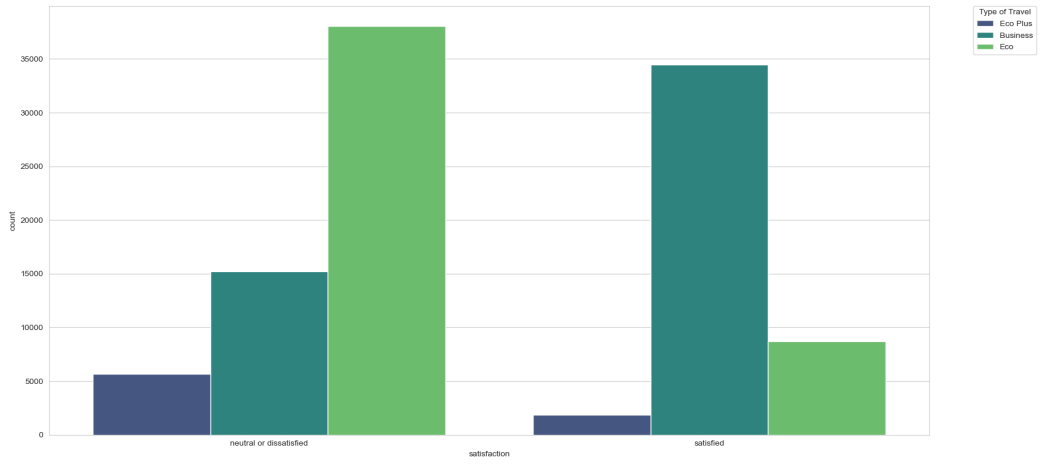

Figure 5: Customer satisfaction base based on class of travel

On this graph, we can clearly see the impact of the class on customer satisfaction. In general, customers travelling in better class will be more satisfied.

1.6 Distribution of customer age

Understanding the age distribution of customers can be useful in adapting services and marketing strategies. The following graph illustrates the age distribution of customers:

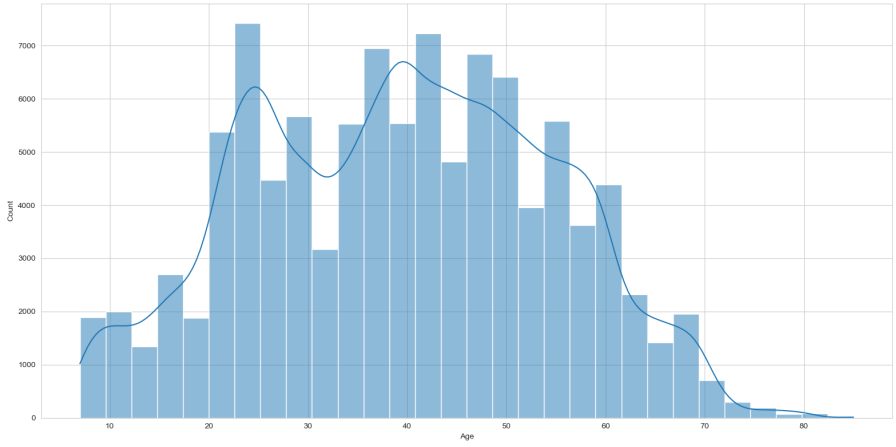

Figure 6: Distribution of customer age

We can see that the age distribution roughly follows a Gaussian distribution, with a majority of customers aged between 25 and 55.

1.7 Customer Profile based Age and Travel Type

Here we analyse customer profiles according to age and type of trip. This tells us something about travellers' preferences and behaviour. The graph below shows customer profiles by age and type of trip:

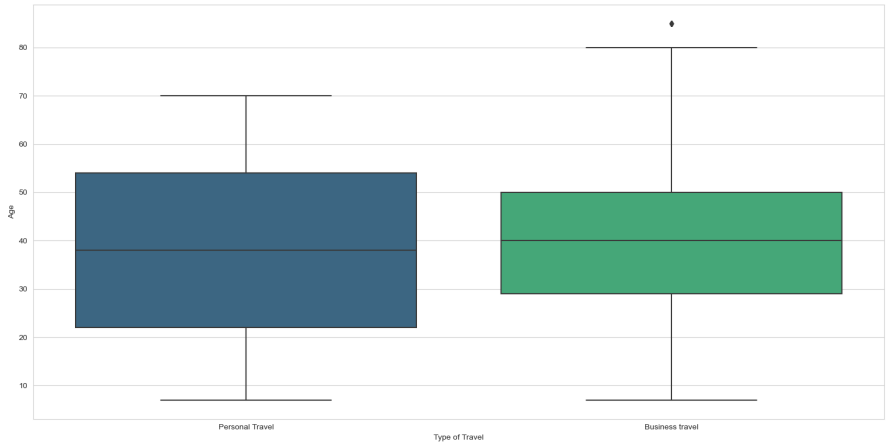

Figure 7: Customer Profile based Age and Travel Type

We can see that the customers travelling for business reasons are tighter around an average age of 40. Those travelling for personal reasons have a slightly lower average age but are also more evenly spread around it.

1.8 Flight distance compare to the Class

Here we show how the distance between flights varies according to class using the following graph:

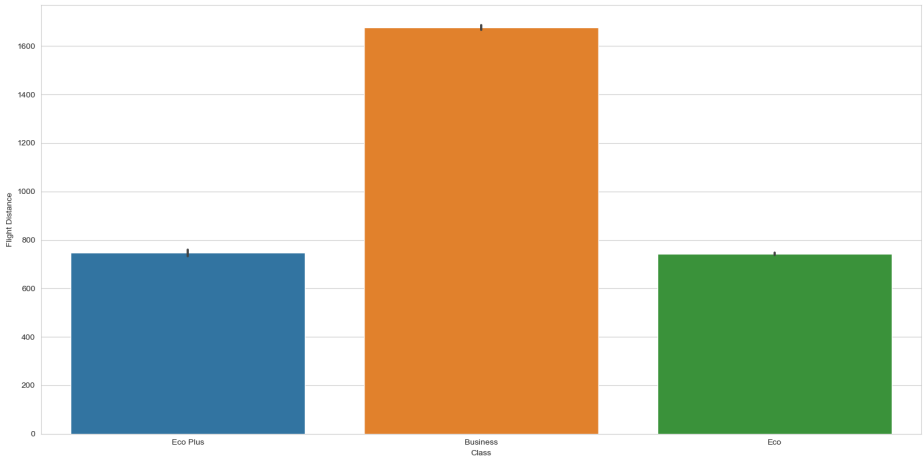

Figure 8: Flight distance compare to the Class

Customers travelling in Business Class travel further than those in Eco or Eco Plus, who travel almost as far.

1.9 Flight distance compare to Customer type

We then wanted to analyse the relationship between flight distance and customer type. The graph below compares the flight distance between different types of customer:

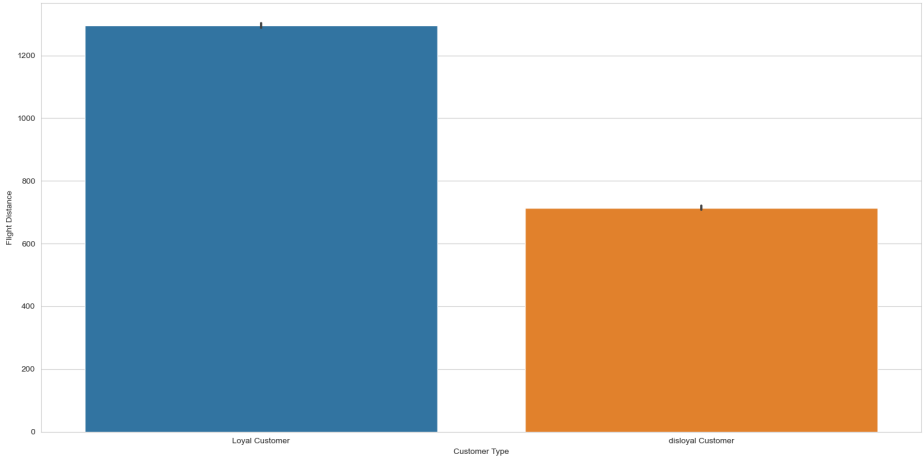

Figure 9: Flight distance compare to Customer type

As we might expect, loyal customers travel further than new customers.

1.10 Heat map

This heat map visualizes the correlation between various airline service factors and customer characteristics, providing insights into how different aspects of the service experience are interrelated.

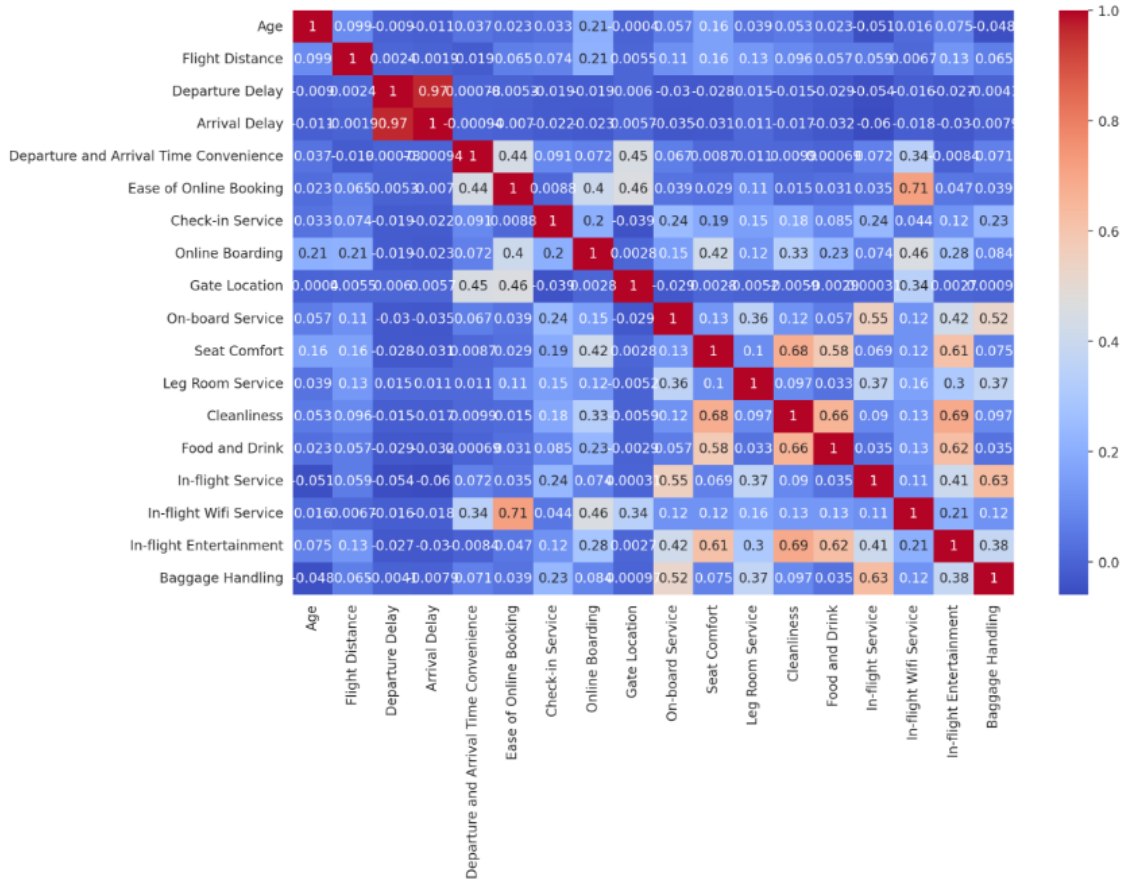

Figure 10: Heat map

This heat map shows the correlation between various factors related to airline services and customer characteristics. Correlation values range from -1 to 1, with values near 1 indicating a strong positive correlation, values near -1 indicating a strong negative correlation. We can therefore notice the following points in this graph:

The correlation coefficient between age and ease of online booking is 0.44, which is rather high. There is therefore a link between age and ease of booking online. The correlation between In-Flight Entertainment and In-Flight Wi-Fi service is 0.41. Improving wifi would therefore theoretically improve in-flight entertainment.

On the other hand, the delay in departure and arrival are linked by a very strong negative correlation (-1.00), which indicates an inverse relationship. When the departure delay increases, the arrival delay unexpectedly decreases, which could be due to compensating factors such as flight speed adjustment.

All these analyses help us to better understand our customers and their expectations. This data can be very useful for the company and for improving customer satisfaction. However, this data does not show us exactly which parameters have the greatest (and least) influence on customer satisfaction. That's why using a machine learning model will help us to understand customer expectations even better.

2 Methodology

2.1 Data Preprocessing

Data preprocessing is an important step in data analysis, especially in machine learning projects where the quality of input data significantly affects model performance. However, in our case the data has already been processed and separated into train data and test data, which greatly facilitates the implementation of a machine learning model.

2.1.1 Handling Missing Values

However, we still had one step before implementing the model, data preprocessing. This allows checking for missing values for each feature in the dataset. The check revealed that the only attribute with missing values was "Arrival time in minutes", with a total of 310 entries missing. To solve this problem, we used an average imputation strategy, which replaces missing values with the average value of the respective column.

After applying mean imputation, a recheck confirmed that there were no more missing values in the dataset.

2.1.2 Checking for Duplicates

The next step in data preparation is looking for duplicate records in the dataset, which could skew the analysis results.

The result indicated that there were no duplicate entries in the train data, confirming that each record represented a unique instance of customer feedback.

The data cleaning process confirmed the effectiveness of the preprocessing steps. After filling missing values and removing duplicates, the dataset was found to be clean and well-structured for the next stages of analysis. This was essential for ensuring the reliability and validity of the machine learning models developed later in the project.

3 Machine Learning Model

For the study, we apply the following models:

- **Logistic Regression:** This model estimates probabilities using a logistic function, making it suitable for scenarios with two possible outcomes.
- **K-Nearest Neighbors (KNN):** non-parametric method that predicts the class of a sample based on the majority class of its closest neighbors.
- **Support Vector Machine (SVM):** SVM is versatile, working with both linear and nonlinear separations.
- **Random Forest:** A set of decision trees that improves forecast accuracy and controls overfitting by averaging the results of individual trees, reducing variance and bias.

3.1 Code Functionality

In machine learning model analysis, the code follows several steps to prepare and evaluate the data. First, we impute data and correct missing values; Numerical columns are populated with median values, while the most frequent values are used for categorical columns. This ensures that all features are complete.

Next, data scaling is applied using StandardScaler. Finally, the models are trained on this processed data and evaluated based on their performance metrics such as precision, recall, and f1 score. This process applies each of the models effectively.

3.2 Performance Analysis

Logistic Regression: Achieved an accuracy of 87%, with balanced precision and recall, showing competent performance in classifying both customer satisfaction levels.

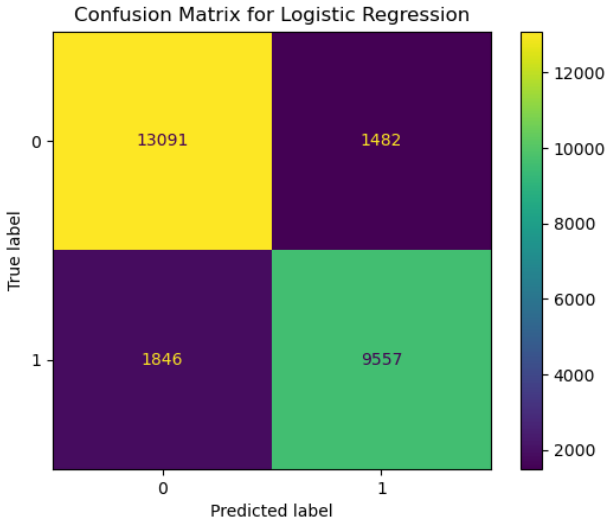

Figure 11: Confusion matrix for logistic regression model

The Logistic Regression model shows reasonable performance with 9,557 true positives, where it accurately predicted satisfied customers, and 13,091 true negatives, correctly identifying those who are neutral or dissatisfied. However, the model also produced 1,482 false positives and 1,846 false negatives. This suggests that while the model is fairly reliable, it has a tendency to misclassified a significant number of customers, particularly by underpredicting customer satisfaction.

K-Nearest Neighbors: Improved accuracy to 92%, reflecting better utilization of the local data structure and effective majority voting.

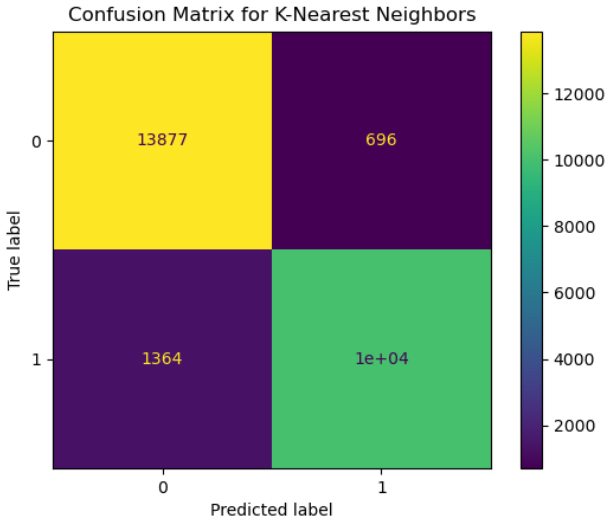

Figure 12: Confusion matrix for K-nearest neighbors model

KNN improves upon the performance of Logistic Regression. It correctly identified 10,000 satisfied customers and 13,877 neutral or dissatisfied customers. The model limited its errors to 696 false positives and 1,364 false negatives. This performance indicates a better handling of class boundaries, making KNN more effective at distinguishing between the two customer states compared to Logistic Regression. The reduced number of misclassifications points to KNN's effectiveness in utilizing the local structure of the data for more accurate predictions.

Support Vector Machine: Enhanced accuracy up to 96% by optimizing the local data structure usage and implementing efficient majority voting.

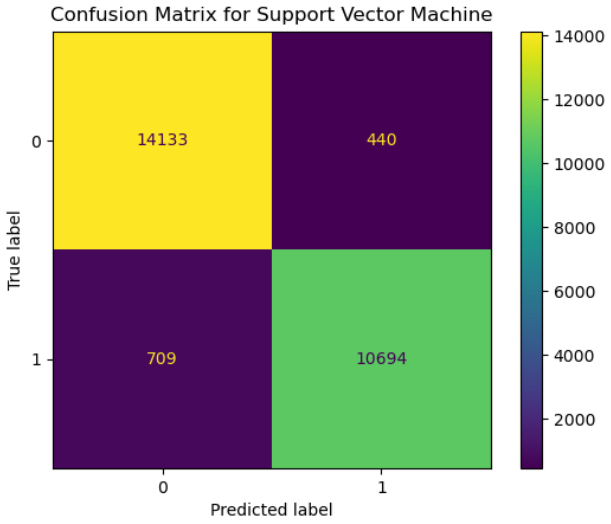

Figure 13: Confusion matrix for support vector machine model

The Support Vector Machine (SVM) model achieved a good level of accuracy. The confusion matrix reveals 10,694 true positives and 14,133 true negatives, suggesting that SVM correctly identified a share of satisfied and neutral/dissatisfied customers. However, the model also generated 709 false positives and 440 false negatives. Despite these classification errors, SVM demonstrates a balanced approach in its predictions.

Random Forest: Exhibited the best performance with an accuracy of 97%, high precision, recall, and the best f1-scores, indicating superior predictive power and robustness.

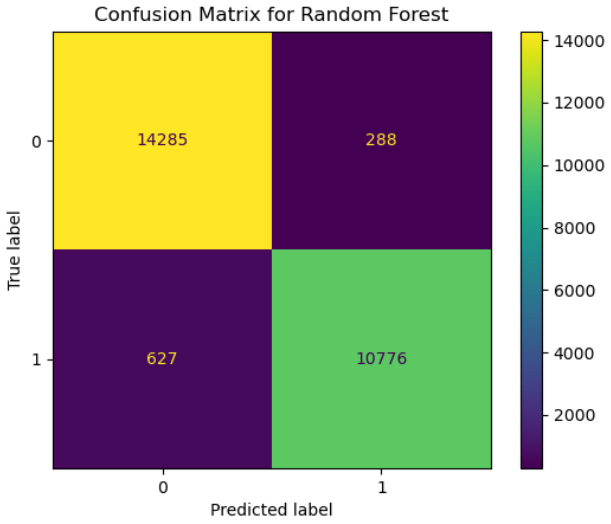

Figure 14: Confusion matrix for random forest model

The Random Forest model exhibits the strongest performance among the three. It boasts 10,781 true positives and 14,282 true negatives, coupled with notably lower false positives and negatives at 291 and 622, respectively. This model's ability to significantly reduce error rates while maintaining high accuracy in identifying both satisfied and neutral or dissatisfied customers highlights its robustness.

The **Random Forest** model is identified as the most performant, making it reliable for predicting customer satisfaction.

4 Model predictions

The final goal of this study is therefore to determine the parameters that will influence customer satisfaction. To do this, we draw graphs showing the impact of different parameters on satisfaction based on logistic regression and random forest models.

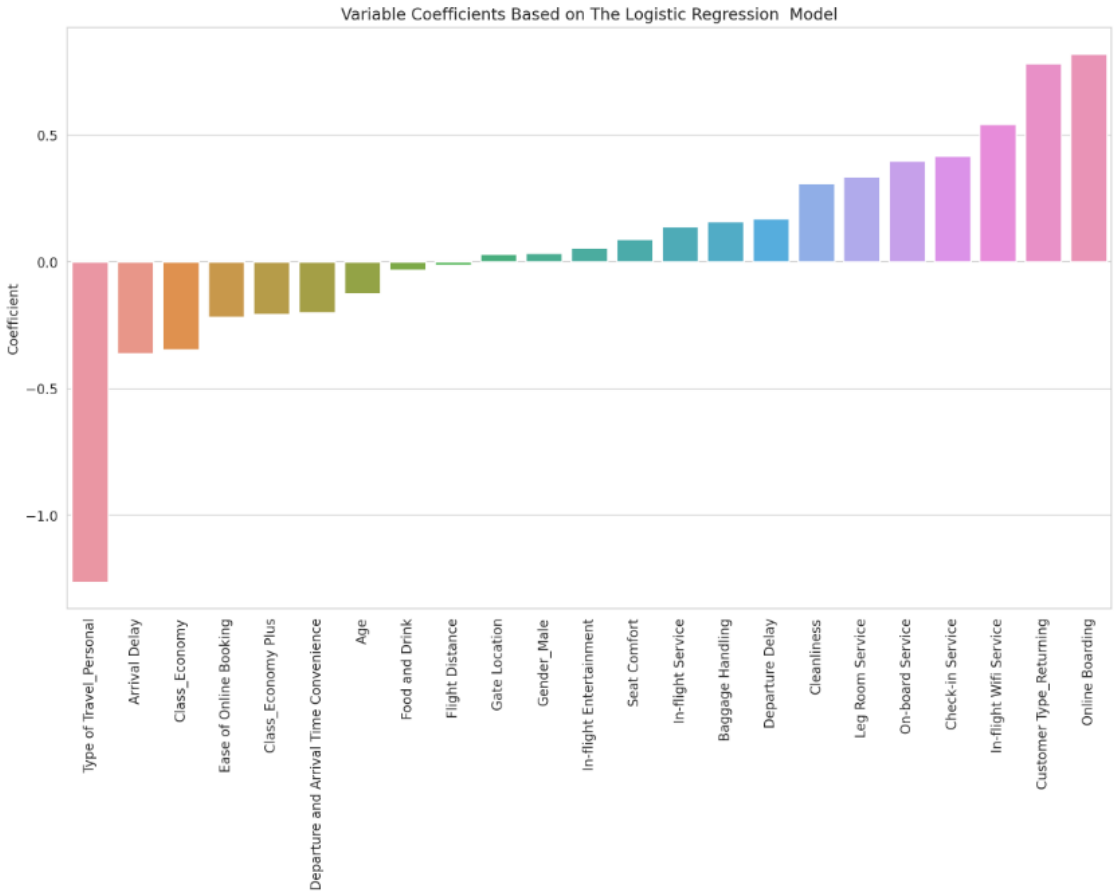

Figure 15: Coefficients based on logistic regression model

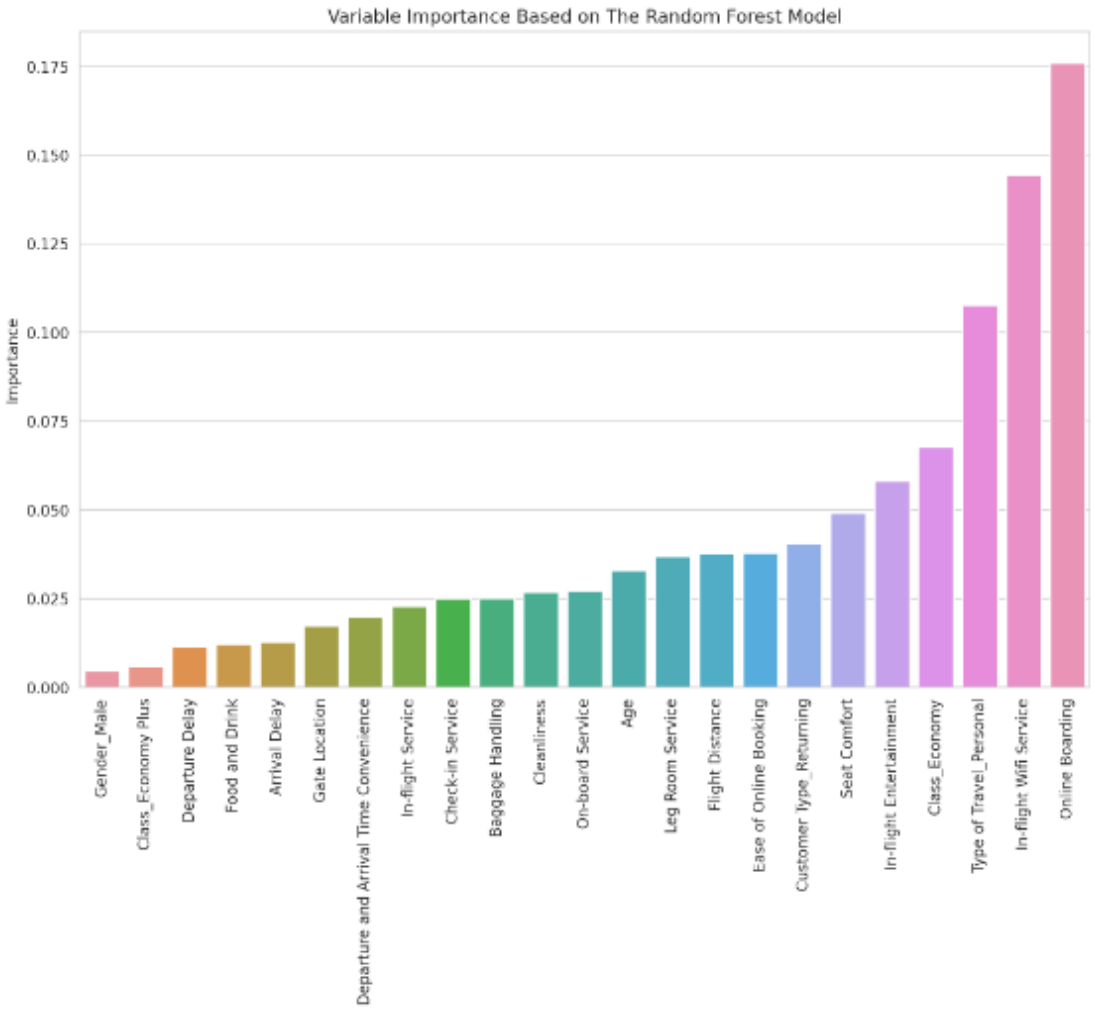

Figure 16: Coefficients based on random forest model

These bar charts from logistic regression and random forest models provide visualizations on which features are most influential in predicting customer satisfaction.
Both charts highlight that while certain features such as "Travel Type" and "Online Boarding" have a crucial impact on satisfaction, their influence can be positive or negative depending on other contextual factors and the how these services are performed. The logistic regression model identifies direct influences on the probability of satisfaction, while the Random Forest model helps prioritize features based on their overall importance to model accuracy.

If we had to pick out the two main points that influence passenger satisfaction, online boarding and on-board wifi service are the two most essential variables.

Conclusion

This research project provided a better understanding of the factors influencing air travel customer satisfaction in the aftermath of the pandemic. By mining a dataset of passenger experiences and applying machine learning techniques, we classified passengers into distinct satisfaction categories and identified the key factors influencing their travel experience.

Throughout the study, we looked at different passenger returns, such as trip type, class and customer characteristics. Our analysis revealed significant differences in satisfaction levels according to these factors, demonstrating the importance of airlines adapting their services to meet customer needs and expectations. The information obtained can help airlines prioritize the most necessary service improvements.

The application of machine learning models, including logistic regression, K-nearest neighbors and random forest, enabled in-depth analysis of the data. The Random Forest model, in particular, performed best, highlighting its robustness and reliability in handling complex datasets with many variables. This model highlighted the importance of specific characteristics.

Features such as online boarding and in-flight Wi-Fi have a significant impact on passenger satisfaction. This suggests that airlines should focus on improving these areas to enhance overall passenger satisfaction.

Ultimately, this project has not only achieved its objective of classifying and predicting passenger satisfaction, but has also enabled airlines to better understand how different service elements influence passenger perception. By focusing on key points, airlines can improve passenger satisfaction and accelerate recovery in the post-pandemic era.