CAPSTONE PROJECT

BA - 723

ABSTRACT

Predicting Airline Passenger Satisfaction using various machine learning techniques

Raghav Gupta
301272406

Table of Contents

Executive Summary

1) Executive Introduction

Air travel has become an essential part of modern life, connecting people and places in ways never seen before. The airline industry's success is determined by more than merely reaching destinations; it is determined by the contentment of its passengers. This report digs into the dynamic world of airline passenger pleasure, shining light on the industry's vital position and revealing key insights that can affect its direction.

2) Executive Objective

This capstone project's primary goal is to research and evaluate the elements that have a substantial impact on passenger satisfaction in the airline business. This study tries to find patterns, trends, and correlations that shed light on the causes of passenger satisfaction by conducting a complete assessment of current literature, collecting pertinent data, and employing rigorous analysis methodologies. Furthermore, the study intends to provide strategic recommendations to airlines for improving their services and overall customer experience.

3) Executive Model Description

The path to decoding the intricate landscape of airline passenger satisfaction was guided by a systematic model selection process. Our approach involved a step-by-step evaluation of different predictive models, ultimately leading us to the model that offered the most profound insights into the factors shaping passenger contentment. Each predictive model, including logistic regression, decision tree, random forest, and gradient boosting, underwent a rigorous evaluation. We meticulously applied each model to our dataset, carefully examining its performance metrics and its ability to capture the intricate nuances of passenger satisfaction.

4) Executive Recommendations

Personal Travel: Experiences Created for Individual Journeys
Recognize the significance of personal travel and how it affects passenger pleasure. Create experiences that appeal to individual interests and provide flexibility, therefore boosting the leisure side of travel.

Elevating Comfort and Productivity is a Class Distinction:

Using the information about travel class preferences to our advantage. Enhance "Business Class" offers with amenities that increase productivity, while focusing on providing "Economy Class" passengers with affordability and comfort.

Meeting the Needs of Business Travellers: Convenience and Connectivity Recognize the impact of the "Business Travel" category. Create services that are tailored to the specific needs of business travelers, with an emphasis on convenience, connectivity, and efficient experiences.

Excellence in In-Flight Wi-Fi: Constant Connectivity for Everyone
Highlight the significance of in-flight Wi-Fi service. Invest in strong and dependable connectivity solutions to meet passengers' demand for seamless online experiences during flights.

Online Boarding: Simplify the Boarding Process with Online Boarding Efficiency. Streamline "Online Boarding" procedures for the convenience of passengers. To improve the pre-flight experience, prioritize user-friendly processes, clear directions, and easily available boarding information.

Introduction

1) Background

Passenger satisfaction is critical in determining the direction of the airline sector. Understanding the intricate aspects that lead to passenger pleasure has become critical with global connectivity and growing customer expectations. This study digs into the complex landscape of airline passenger satisfaction, with the goal of elucidating the factors that determine traveler experiences. We hope to provide a comprehensive grasp of important variables and their impact on satisfaction levels by evaluating key variables and their impact on satisfaction levels. This analysis provides not only a view into the complicated tapestry of air travel experiences but also a potential path for improving customer journeys and fostering growth in the airline industry.

2) Problem Statement

Passenger satisfaction is a critical factor in an airline's success and reputation in the ever-changing world of air travel. Meeting and exceeding passenger expectations has become a complex challenge as various aspects like as in-flight amenities and online experiences come into play. The task is to identify and analyze the primary drivers of passenger pleasure, as well as to comprehend their interconnection. This study seeks to answer the following overarching question: What are the underlying factors that significantly influence airline passenger satisfaction, and how can these insights inform strategic decisions for enhanced customer experiences? This study aims to provide significant insights that can influence the industry by conducting an in-depth investigation of these critical elements.

3) Objectives and Measurement

The primary objective of this research is to find the best model with the highest accuracy rate for predicting passenger satisfaction.

Moreover, this research will help to answer the following questions:

1. How well can we predict passenger satisfaction based on historical data of customer feedback and reviews?

2. Can we predict passenger satisfaction levels based on a combination of flight-related factors (e.g., ease of online booking, online boarding, check-in services) and demographics-related factors (e.g., age, type of travel, gender)?
3. What are the key factors that significantly impact airline passenger satisfaction?
4. What are the primary reasons behind customer complaints or negative feedback regarding airline services?
5. What amenities or services do passengers value the most in terms of enhancing their satisfaction?

4) Assumptions and Limitations

Some of the assumptions and limitations for this analysis are as follows:-

1) Various information like carrier, cost, travel date, and flight capacity are not given in the dataset. These variables might have had an impact on the satisfaction of passengers.
2) The dataset lacks a specified timeframe for data collection, leaving us uncertain whether the information was gathered prior to or post the pandemic. Given the significant shifts in passenger preferences caused by the pandemic, this temporal context is crucial for accurate analysis.
3) Because of demographic and regional differences, the data obtained for analysis may suffer from sampling bias, as survey respondents may not fully represent the entire spectrum of airline customers.
4) The study is vulnerable to self-reporting bias, as respondents' impressions may be influenced by recollection bias or social desirability bias.

5) "Data on departure and arrival delays were adjusted in this study using capping and flooring techniques." Delays longer than 30 minutes were capped, and delays less than a certain threshold were adjusted to 0 minutes. This strategy tries to address extreme values while focusing on the significant impact of long delays on passenger displeasure, in line with industry insights."

Data Sources

1) Data Set Introduction

This study's principal data source is Kaggle, a well-known site that contains a wide range of datasets and tools. The datasets from Kaggle provide a solid foundation for our investigation of passenger satisfaction in the airline industry. My dataset has 103904 rows and 25 columns with each column containing varied information related to passengers. These observations aid us in constructing a model that aligns with the goals of this project. Out of 25 columns 23 columns were used, which is explained in the Exclusions section of this report. Moreover, out of 23 columns 19 columns are categorical and 4 are numerical.

2) Exclusions

Out of 25 columns in the dataset, 2 columns were excluded from the analysis.

1. **Id Column:** Because of its inherent purpose as a unique identifier for individual data entries, the dataset's 'id' column was removed from our analysis. Because it has no substantive significance to the assessment of passenger satisfaction and the underlying influencing factors, its deletion simplifies our emphasis on the essential variables of interest.

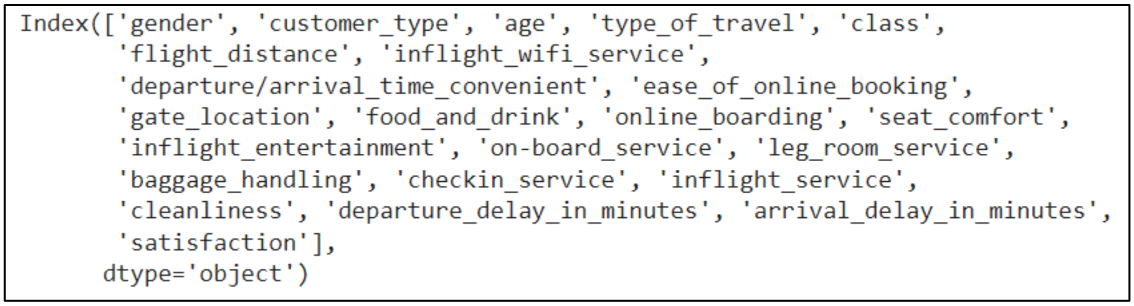
2. **Unnamed 0 Column**: The 'Unnamed 0' column was also removed from the dataset. This column appears to serve as an index or identifier, with no apparent significance to the analysis of passenger pleasure or its associated variables. Our study is streamlined by eliminating this column, allowing us to focus primarily on variables that add meaning to our understanding of passenger contentment in the airline business.

3) Initial Data Cleansing or Preparation

A major alteration was performed during the earliest stages of data preparation to improve the dataset's consistency and accessibility. Several processes were engaged in this change to streamline column names for optimal analysis:

1. **Column Naming Convention:** To ensure consistency and convenience of reference, all column names were transformed to lowercase letters.
2. **Eliminating Spaces:** To replace spaces within column names, underscores were introduced. This change ensures that column names are consistent and adhere to the best standards for variable naming.

This rigorous data transformation not only improves the dataset's general readability and manageability but also lays the groundwork for smooth analytical processes. It represents our dedication to data hygiene and systematic organization, both of which are critical to obtaining accurate and informative results.

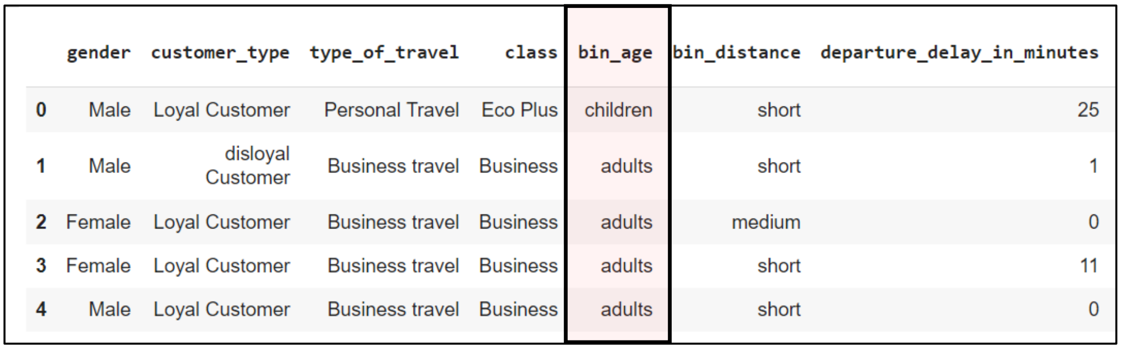


```
Index(['gender', 'customer_type', 'age', 'type_of_travel', 'class',
       'flight_distance', 'inflight_wifi_service',
       'departure/arrival_time_convenient', 'ease_of_online_booking',
       'gate_location', 'food_and_drink', 'online_boarding', 'seat_comfort',
       'inflight_entertainment', 'on-board_service', 'leg_room_service',
       'baggage_handling', 'checkin_service', 'inflight_service',
       'cleanliness', 'departure_delay_in_minutes', 'arrival_delay_in_minutes',
       'satisfaction'],
      dtype='object')
```

fig., 1

Moreover, binning was applied to both the 'age' and 'flight distance' variables during the initial stages of dataset preparation. Binning is the process of categorizing continuous variables into intervals, which allows for the orderly study of trends and patterns. The following precise actions were taken:

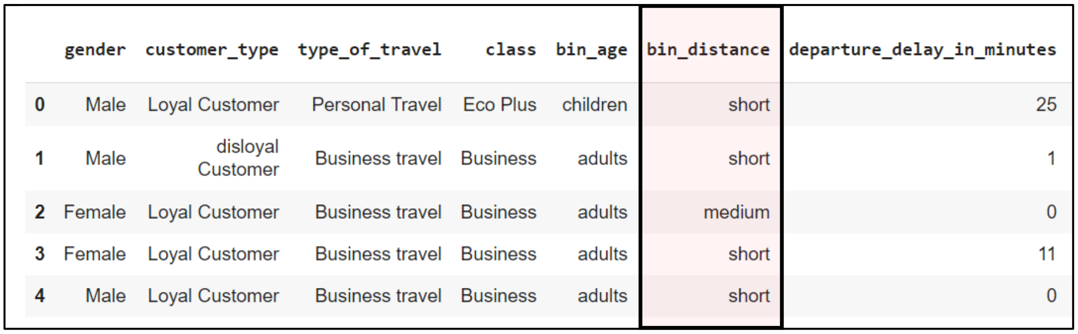
1. **Age Binning:** The 'age' variable, which represents the ages of the passengers, was separated into distinct 4 age groups ‘children(0-14)’, ‘youth(15-24)’, ‘adult(25-65)’, ‘senior(66-90)’. This category facilitates finding age-related trends in passenger satisfaction, which helps with result interpretation. These groups were considered as per age groups used generally by the Canadian Government.



	gender	customer_type	type_of_travel	class	bin_age	bin_distance	departure_delay_in_minutes
0	Male	Loyal Customer	Personal Travel	Eco Plus	children	short	25
1	Male	disloyal Customer	Business travel	Business	adults	short	1
2	Female	Loyal Customer	Business travel	Business	adults	medium	0
3	Female	Loyal Customer	Business travel	Business	adults	short	11
4	Male	Loyal Customer	Business travel	Business	adults	short	0

fig., 2

2. **Flight Distance Binning:** The variable 'flight distance', which represents trip distance, was split into 3 bins ‘short(0-700)’, ‘medium(701-3000)’, and ‘long(3001-5000)’. This segmentation allows for an analysis of how satisfaction levels differ depending on the length of the travel.

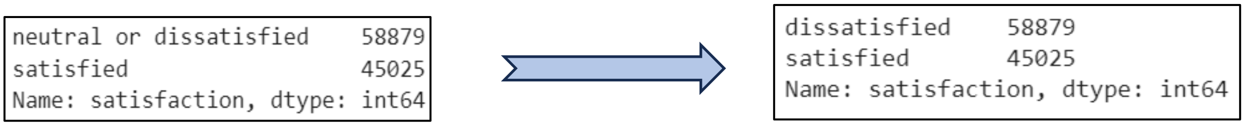


	gender	customer_type	type_of_travel	class	bin_age	bin_distance	departure_delay_in_minutes
0	Male	Loyal Customer	Personal Travel	Eco Plus	children	short	25
1	Male	disloyal Customer	Business travel	Business	adults	short	1
2	Female	Loyal Customer	Business travel	Business	adults	medium	0
3	Female	Loyal Customer	Business travel	Business	adults	short	11
4	Male	Loyal Customer	Business travel	Business	adults	short	0

fig., 3

Using binning techniques simplifies the analysis of these continuous variables and improves the capacity to discover correlations between passenger satisfaction and other variables.

In the final phase, the target variable underwent a transformation where the values 'satisfied' and 'dissatisfied or neutral' were modified to 'satisfied' and 'dissatisfied' respectively.



```
neutral or dissatisfied    58879
satisfied                  45025
Name: satisfaction, dtype: int64
```

```
dissatisfied    58879
satisfied       45025
Name: satisfaction, dtype: int64
```

4) Data Dictionary

Sr. No	Attribute Name	Description
1	Gender	The gender of the passenger.
2	Customer Type	Indicates whether the passenger is a loyal customer or a first-time traveler.
3	Bin Age	The age of the passenger reflects a demographic factor as per the bins mentioned above.
4	Type of Travel	Specifies if the travel is for business or personal.
5	Class	The class of travel – Economy, Business, or Eco Plus.

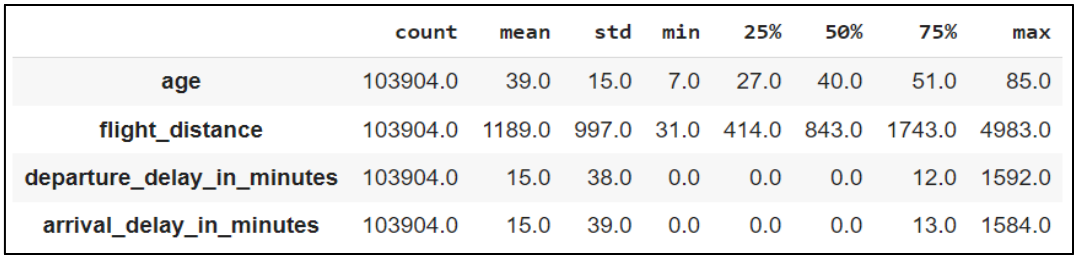
6	Bin Flight Distance	The distance of the flight in above mentioned bins in kilometers.
7	Inflight Wi-Fi Service	Ratings (0: Not Applicable; 1-5)) for the quality of in-flight Wi-Fi.
8	Departure/Arrival Time Convenience	Ratings for the convenience of departure and arrival times.
9	Ease of Online Booking	Ratings for the ease of booking tickets online.
10	Gate Location	Ratings for the convenience of gate locations.
11	Food and Drink	Ratings for the quality of food and beverages.
12	Online Boarding	Ratings for the ease of online boarding procedures.
13	Seat Comfort	Ratings for the comfort of seats.
14	Inflight Entertainment	Ratings for the quality of in-flight entertainment.
15	On-board Service	Ratings for the quality of on-board services.
16	Leg Room Service	Ratings for the legroom comfort.
17	Baggage Handling	Ratings for the efficiency of baggage handling.
18	Check-in Service	Ratings for the quality of check-in services.
19	Inflight Service	Ratings for the in-flight service quality.
20	Cleanliness	Ratings for the cleanliness of the aircraft.
21	Departure Delay in Minutes	The delay of time at departure in minutes.
22	Arrival Delay in Minutes	The delay of time at arrival in minutes.
23	**Satisfaction**	**The binary outcome variable indicates whether the passenger was 'satisfied' or 'dissatisfied'.**

Data Exploration

1) Techniques

Various strategies are used in data exploration to understand the underlying patterns, relationships, and properties of a dataset. Here are some of the most common data exploration techniques:

Descriptive Statistics: Basic statistical metrics, such as mean, median, mode, and standard deviation, provide an early grasp of the data's central tendencies and distribution.



	count	mean	std	min	25%	50%	75%	max
age	103904.0	39.0	15.0	7.0	27.0	40.0	51.0	85.0
flight_distance	103904.0	1189.0	997.0	31.0	414.0	843.0	1743.0	4983.0
departure_delay_in_minutes	103904.0	15.0	38.0	0.0	0.0	0.0	12.0	1592.0
arrival_delay_in_minutes	103904.0	15.0	39.0	0.0	0.0	0.0	13.0	1584.0

fig., 4

1. Data Visualizations

Data visualization is an effective tool for presenting complex information in an understandable and accessible manner. The following are some popular data visualization techniques for exploratory analysis:

- Violin Plot
- Box Plots
- Pie Charts
- Bar Charts

Some important insights from EDA are as follows:

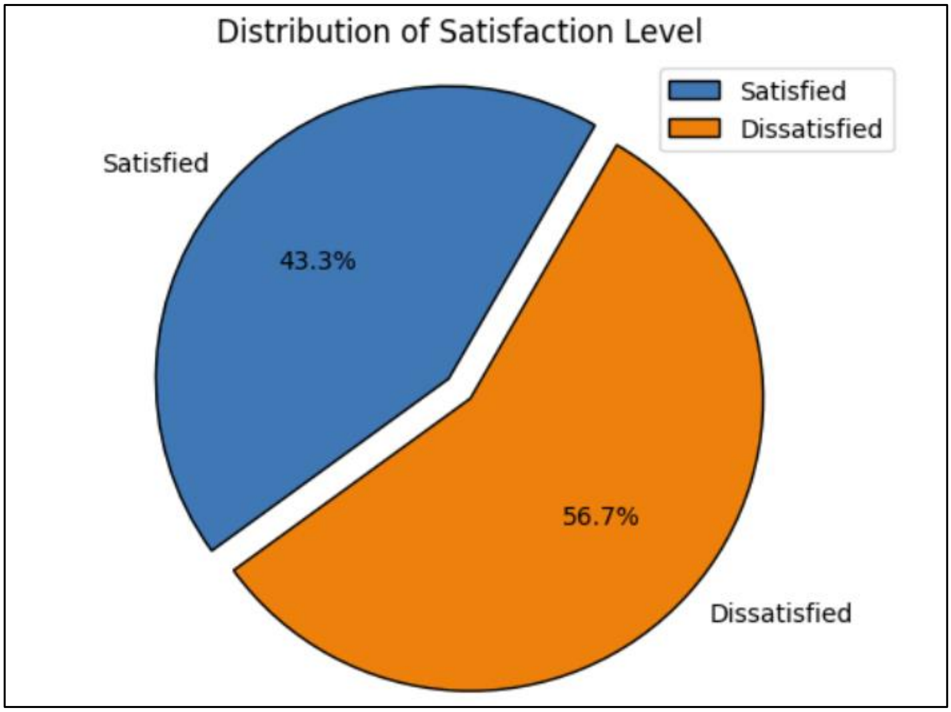


fig., 5

In accordance with fig., 5 it's evident that the distribution of the target variable exhibits an imbalanced nature. Within this distribution, the proportion of satisfied passengers stands at 43.3%, while dissatisfied passengers account for a higher share of 56.7%.

This imbalanced distribution holds significance in the context of the analysis. When a dataset's target variable is imbalanced, meaning that one class greatly outweighs the other in terms of frequency, it can pose certain challenges during model training and analysis. For instance, a machine learning model trained on an imbalanced dataset might struggle to accurately predict the minority class due to the inherent bias towards the majority class.

In the case of this analysis, the higher prevalence of dissatisfied passengers underscores the potential impact of class imbalance on the interpretation of results. Addressing this imbalance involve employing technique such as SMOTE to handle imbalanced datasets. By acknowledging and managing this imbalance, the analysis can yield more accurate insights into factors influencing passenger satisfaction within the airline context.

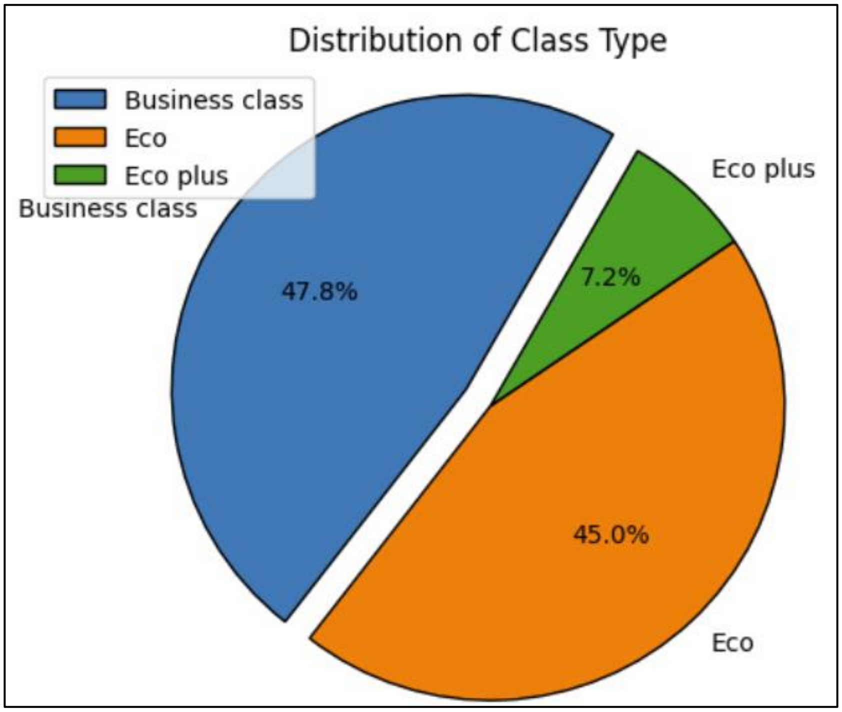


fig., 6

Fig., 6 depicts that majority of travellers i.e., approximately 48% are from business class, while Eco class travellers are 45%. Eco plus travellers are just approximately 7%.

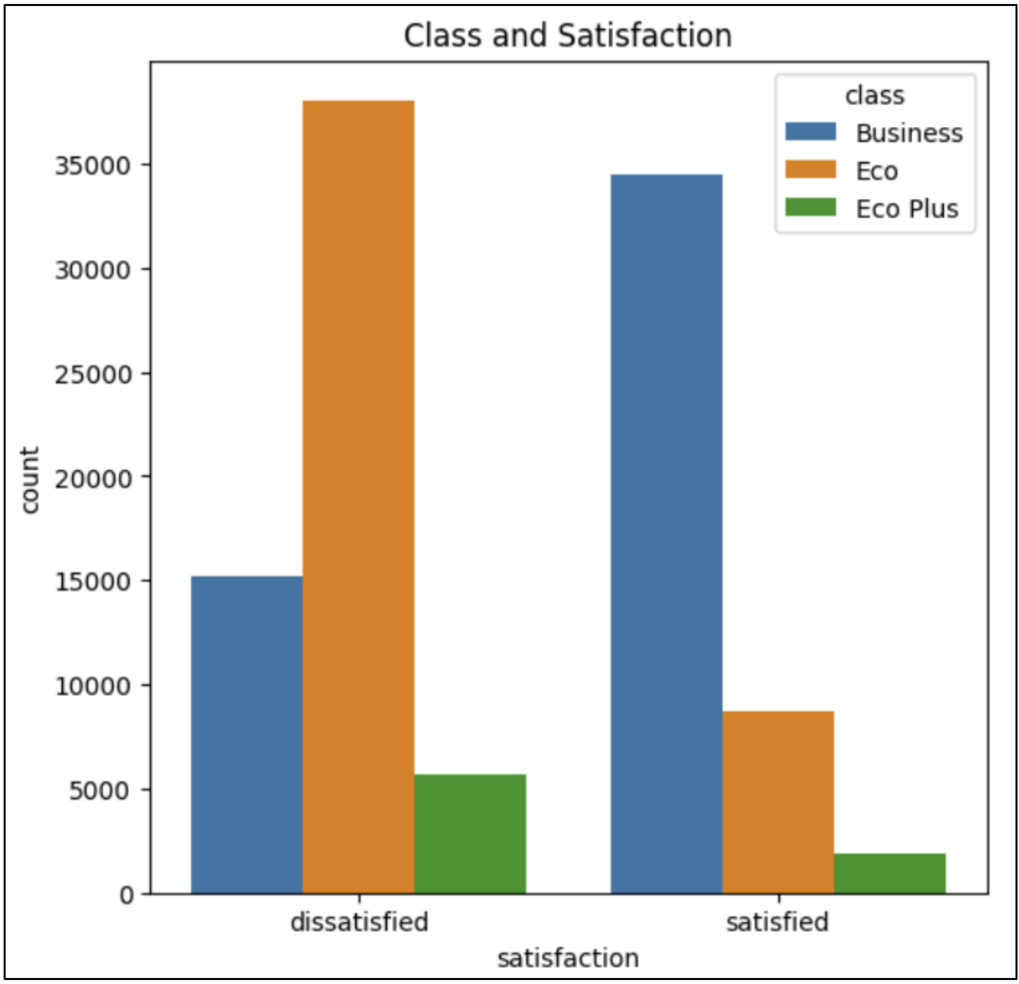


fig., 7

As depicted in the figure, 7 a significant portion of satisfied individuals are affiliated with the Business Class, while a notable proportion expressing dissatisfaction belongs to the Eco class. This observation underscores the need for the airline company to enhance services for Eco-class passengers, considering their apparent discontent. Moreover, the figure illustrates a near-equal count of passengers opting for both Eco class and Business class, highlighting a balanced distribution of travelers as shown in the figure, 6.

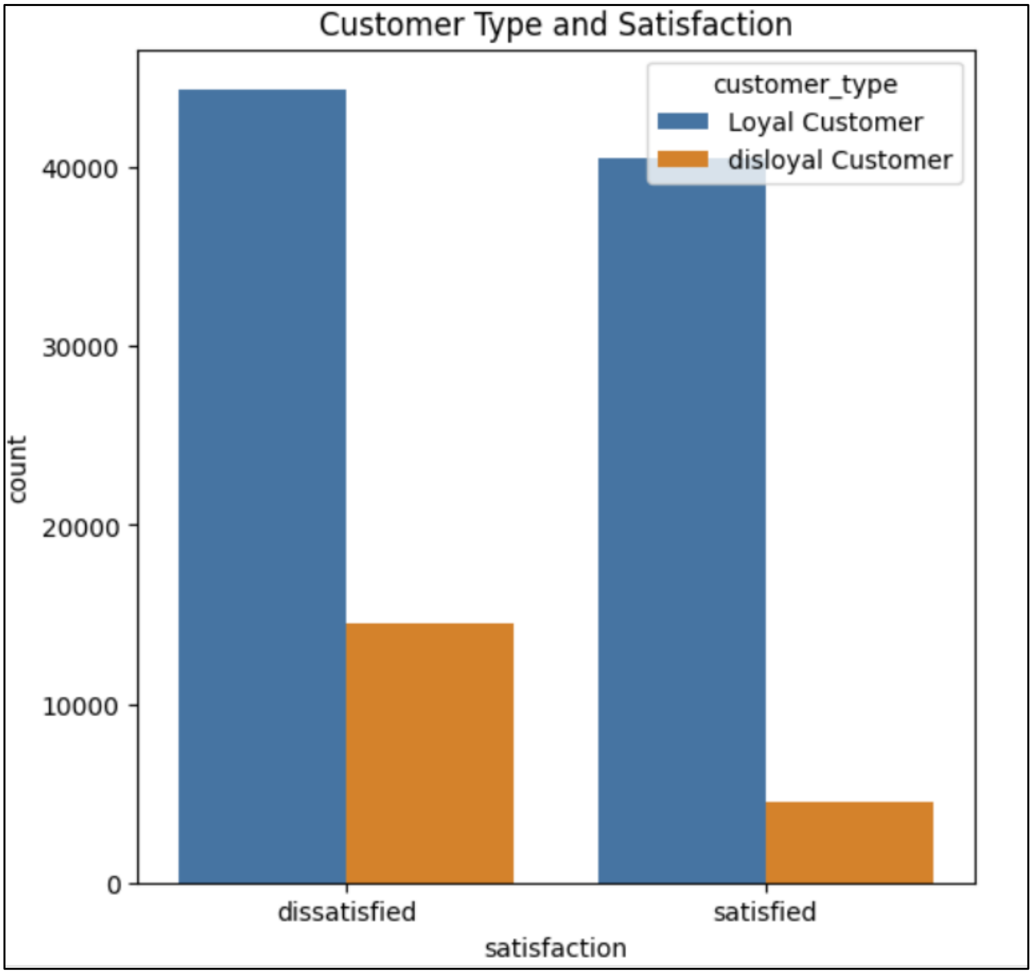


fig., 8

As per fig., 8 even loyal customers are dissatisfied with the airline company, which is too bad for their business. Loyal customers who are satisfied with the company's services are lesser than those who are dissatisfied. The company should make sure that they work well on their services to have a satisfied loyal customer rather than just a loyal customer.

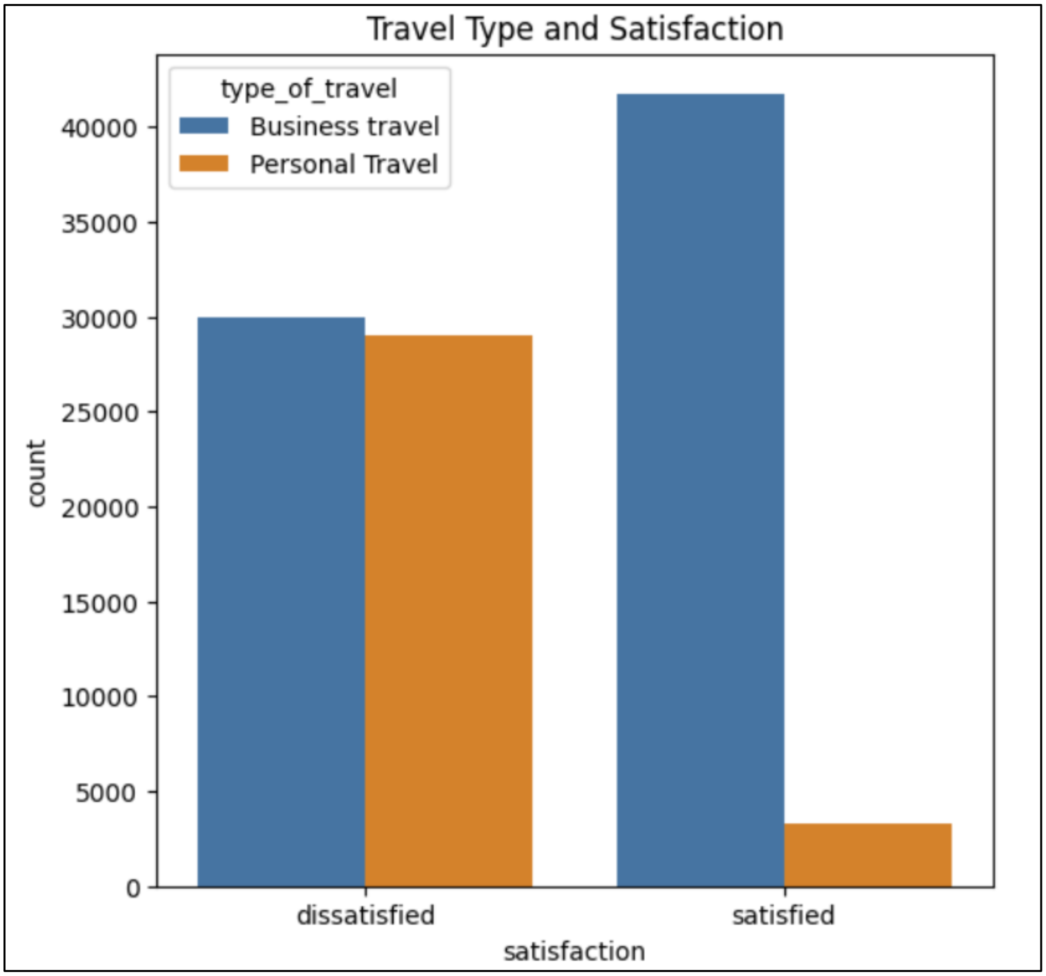


fig., 9

As per fig., 9 it can be depicted that people who traveled for business purposes are highly satisfied rather than people who traveled for personal reasons such as leisure. Personal reasons passengers are not really satisfied with the airline services.

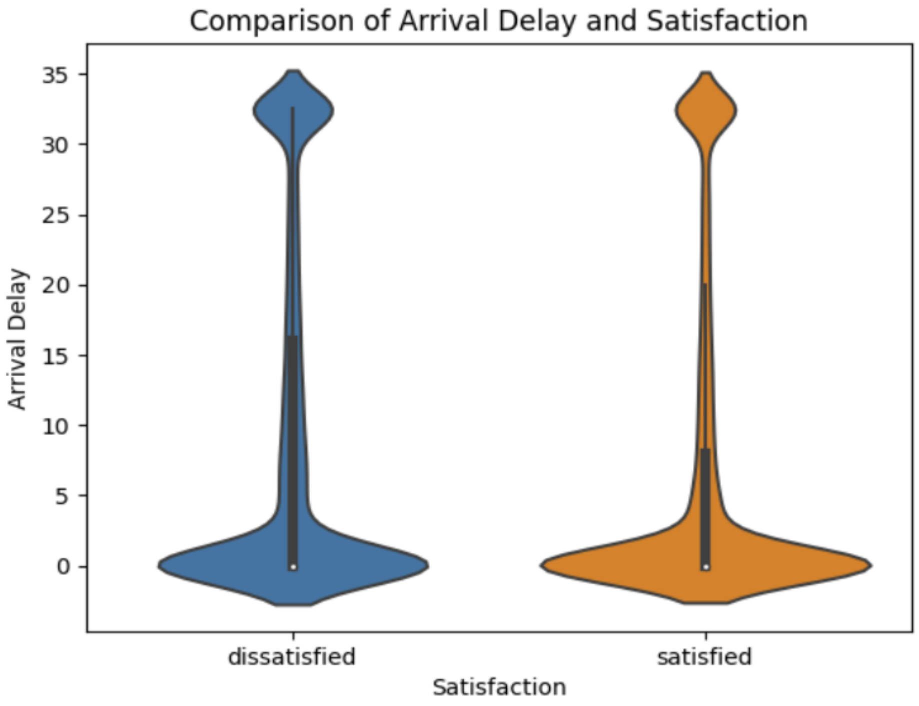


fig., 10

After an analysis of the arrival delay and satisfaction columns, it becomes apparent that a higher level of dissatisfaction is observed when an arrival delay surpasses the range of 5 to 8 minutes fig., 10.

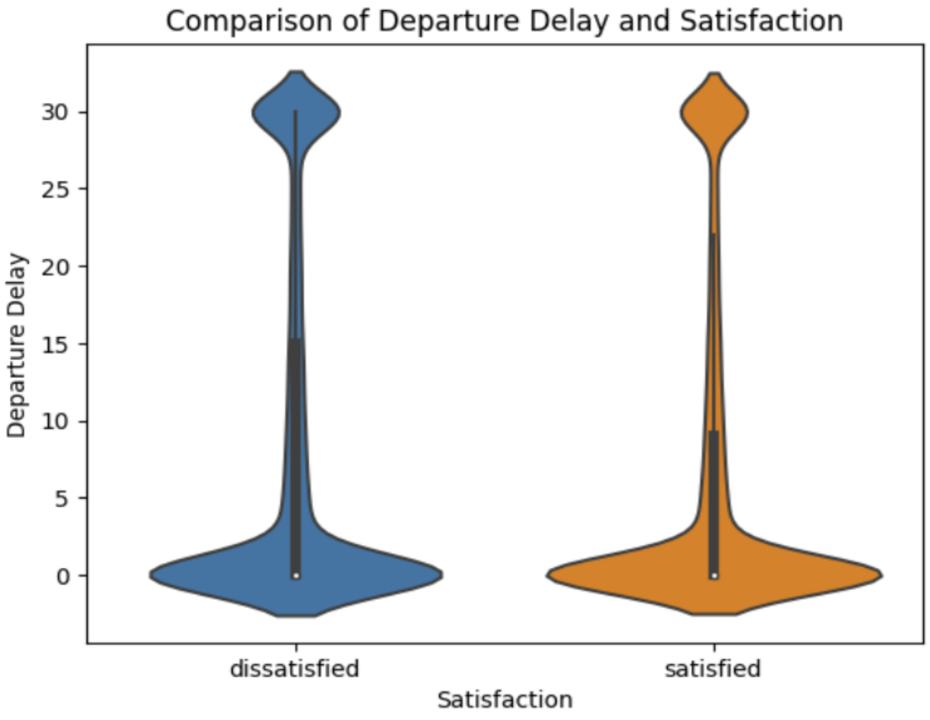


fig., 11

After an analysis of the departure delay and satisfaction columns, it becomes apparent that a higher level of dissatisfaction is observed when a departure delay surpasses the range of 9 to 12 minutes fig., 11.

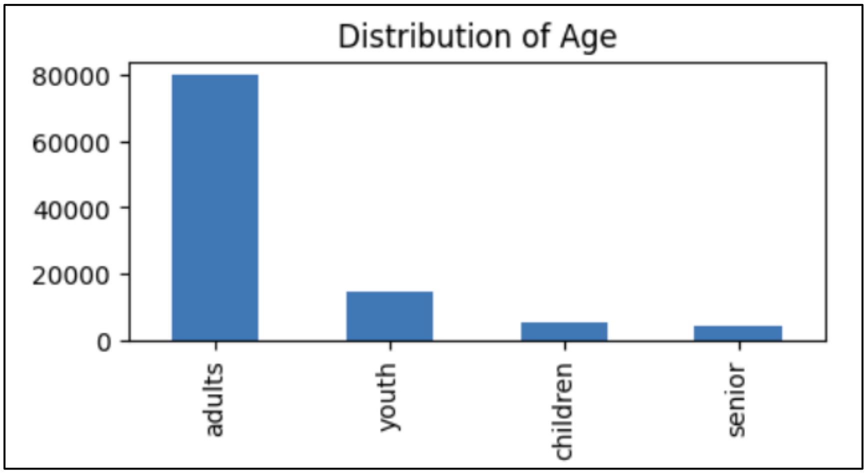


fig., 12

As per fig., 12 most of the passengers that travel belong to the age category of 24 years to 64 years as per our binning.

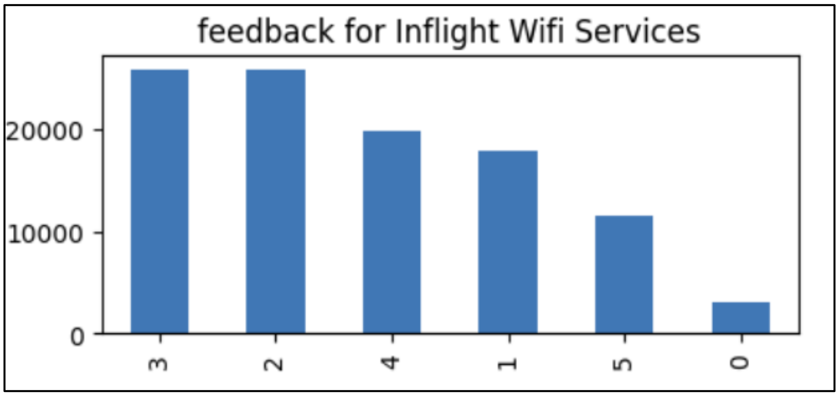


fig., 13

Fig., 13 tells us that most people are neither highly satisfied nor highly dissatisfied with the inflight Wi-Fi services provided by the airlines.

2. Outlier Treatment

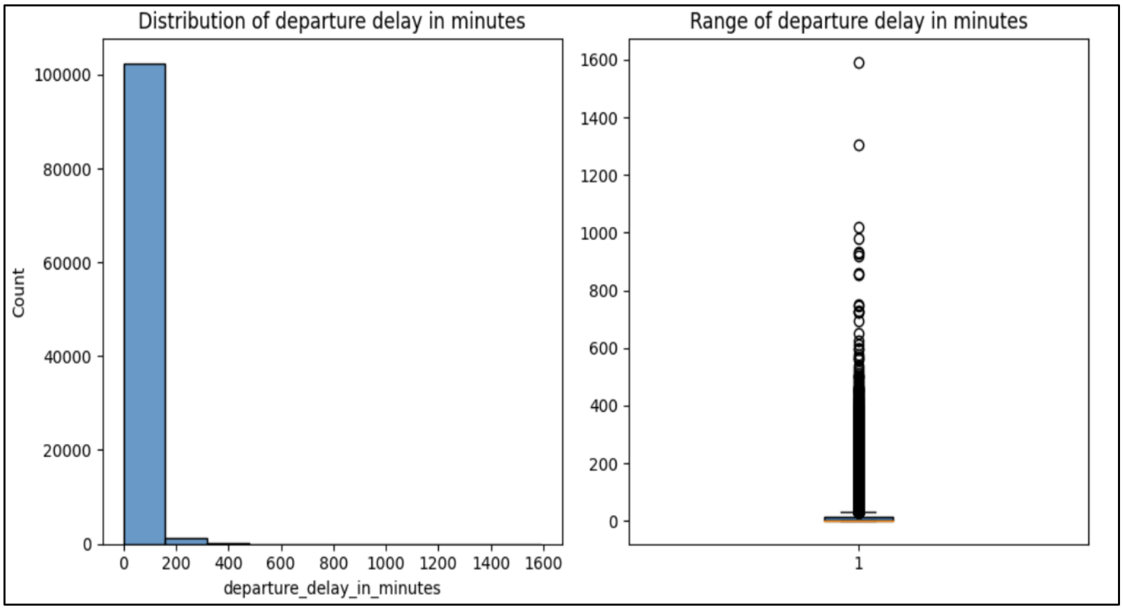


fig., 14

As per fig., 14 we can see there are so many outliers in the departure delay in minutes variable. The maximum value of the variable is 1592. Due to this, it is skewed to the right as its skewed value is 6.7338. The cap and floor method has been used for the treatment of the outliers. IQR method was used to set upper bound values and lower bound values and all the values above and below the upper bound and lower bound values are treated accordingly.

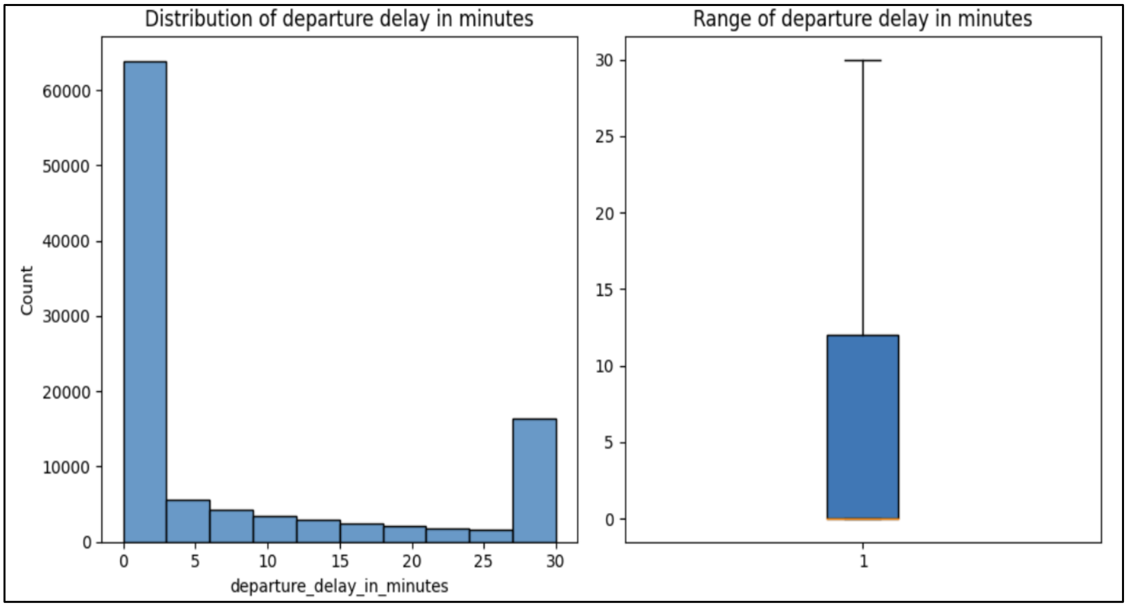


fig., 15

Fig., 15 is after the treatment of the outliers where we can see that the box plot has been changed and now there are no outliers as compared to the previous. After the outlier treatment, the skewness of the variable comes down to 1.2155.

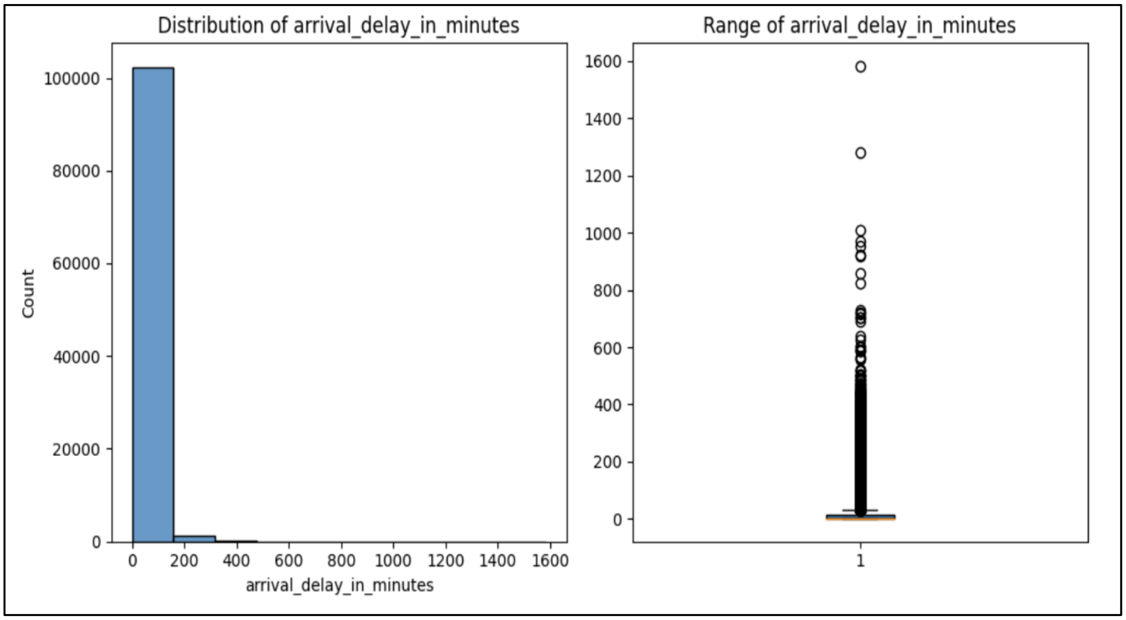


fig., 16

As per fig., 16 we can see there are so many outliers in the arrival delay in minutes variable. The maximum value of the variable is 1584. Due to this, it is skewed to the right as its skewed value is 6.6051. The cap and floor method has been used for the treatment of the outliers. IQR method was used to set upper bound values and lower bound values and all the values above and below the upper bound and lower bound values are treated accordingly.

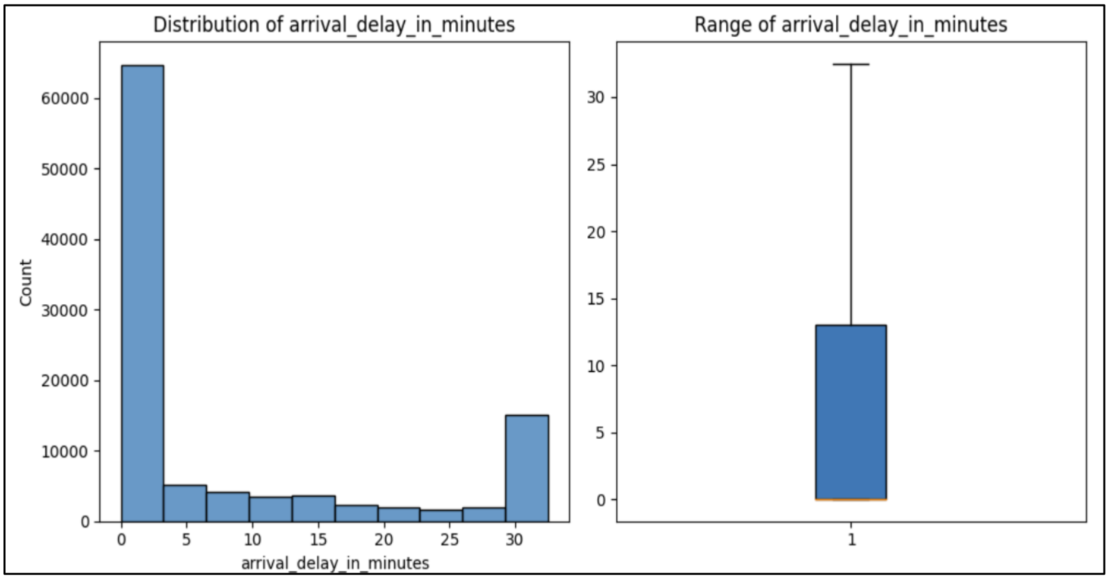


fig., 17

Fig., 17 is after the treatment of the outliers where we can see that the box plot has been changed and now there are no outliers as compared to the previous. After the outlier treatment, the skewness of the variable comes down to 1.2353.

3. Correlation Analysis

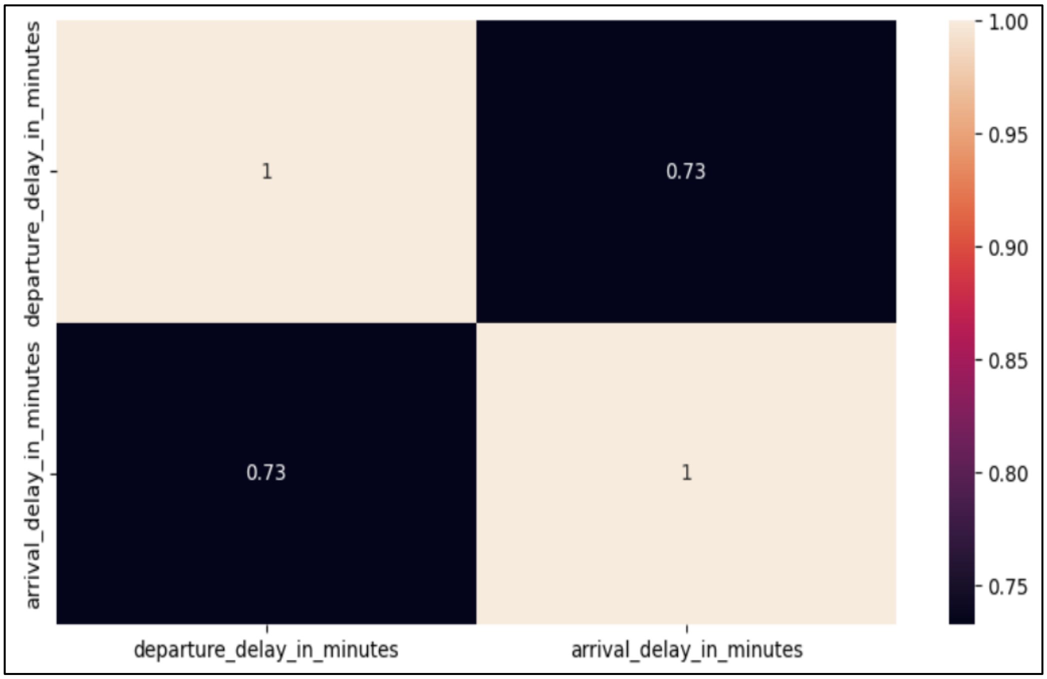


fig., 18

Correlation analysis is a basic technique for determining the degree and direction of a relationship between two or more variables. It aids in understanding how changes in one variable are related to changes in another.

The correlation coefficient is a statistical measure that assesses the level of relationship between two variables. It has a value between -1 and 1, with -1 indicating a high negative correlation, 0 indicating no correlation, and 1 indicating a significant positive correlation.

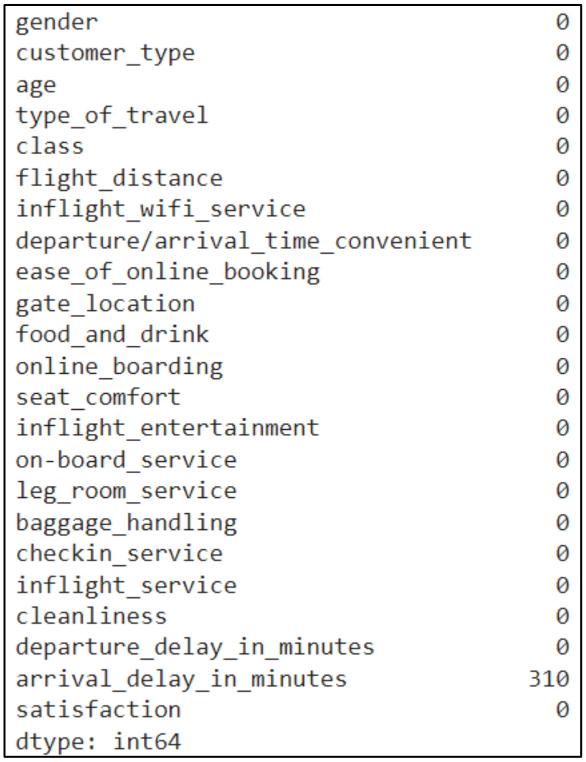
Fig., 18 indicates that the correlation value of 0.73 between the variables "arrival delay in minutes" and "departure delay in minutes" has a significant positive linear relationship. This figure indicates that an increase in arrival delay corresponds to a significant rise in departure delay, and vice versa. This positive association means that flights with longer arrival delays tend to have correspondingly longer departure delays as well. Even though being positively correlated with each other we are not dropping any of the variables as we would like to know how delays either in arrival or departure affect the satisfaction of the passengers.

2) Data Cleansing

Data cleansing sometimes referred to as data cleaning or data pretreatment, is an important phase in the data analysis process. It entails locating and correcting flaws, inconsistencies, and inaccuracies in a dataset to assure its quality, dependability, and suitability for analysis. This is how data cleansing is commonly performed:

Handling Missing Values: fig., 19

We came into a case where the "arrival delay in minutes" variable had 310 missing values. To solve this problem, we used a popular imputation approach to replace the missing values with the median of the available data in the same variable. This method preserves the dataset's integrity and offers a plausible estimate for missing values based on the data distribution's central tendency. The median imputation approach is especially useful when dealing with skewed or outlier-prone data since it is less susceptible to extreme values than the mean imputation method. By utilizing the median to fill in the missing values, the dataset keeps its completeness, which is critical for correct analysis and interpretation.



```
gender                                0
customer_type                         0
age                                   0
type_of_travel                        0
class                                 0
flight_distance                       0
inflight_wifi_service                 0
departure/arrival_time_convenient     0
ease_of_online_booking                0
gate_location                         0
food_and_drink                        0
online_boarding                       0
seat_comfort                          0
inflight_entertainment                0
on-board_service                      0
leg_room_service                      0
baggage_handling                      0
checkin_service                       0
inflight_service                      0
cleanliness                           0
departure_delay_in_minutes            0
arrival_delay_in_minutes            310
satisfaction                          0
dtype: int64
```

fig., 19

Removing Duplicates: We're happy to report that our dataset has no duplicate records. The absence of duplicate entries is advantageous since it improves data accuracy and eliminates the possibility of skewed analysis results. This clean dataset serves as a solid foundation for our analysis, guaranteeing that each observation is unique and distinct, allowing us to draw accurate conclusions from our research.

3) Summary

The exploratory study dug into many components of the dataset to identify insights and trends linked to passenger satisfaction in the setting of an airline. Several significant findings emerged:

Class Imbalance: An imbalanced scenario was identified by the distribution of the target variable. Satisfied travelers made up 43.3% of the dataset, whereas unsatisfied passengers made up a larger 56.7%. This imbalance may have an impact on model training and analysis, which should be considered in the following stages.

Arrival and Departure Delays: There was a high positive correlation (0.73) found between arrival and departure delays in minutes. This association implies that higher arrival delays frequently correspond to longer departure delays, emphasizing the interdependence of these two factors in aircraft operations.

Initial Data Preprocessing: To improve data quality, data cleansing procedures were used. The median of the available data was used to impute missing values in the "arrival delay in minutes" variable (310 missing values), contributing to data completeness.

Outlier treatment: Outliers in the variables "departure delay in minutes" and "arrival delay in minutes" were detected and handled using the Interquartile Range (IQR) method. Outliers were defined as extreme values that fell outside the range of Q1 - 1.5 * IQR to Q3 + 1.5 * IQR and were controlled using appropriate procedures. This procedure ensured data integrity and improved analysis accuracy by reducing the impact of outliers on the results.

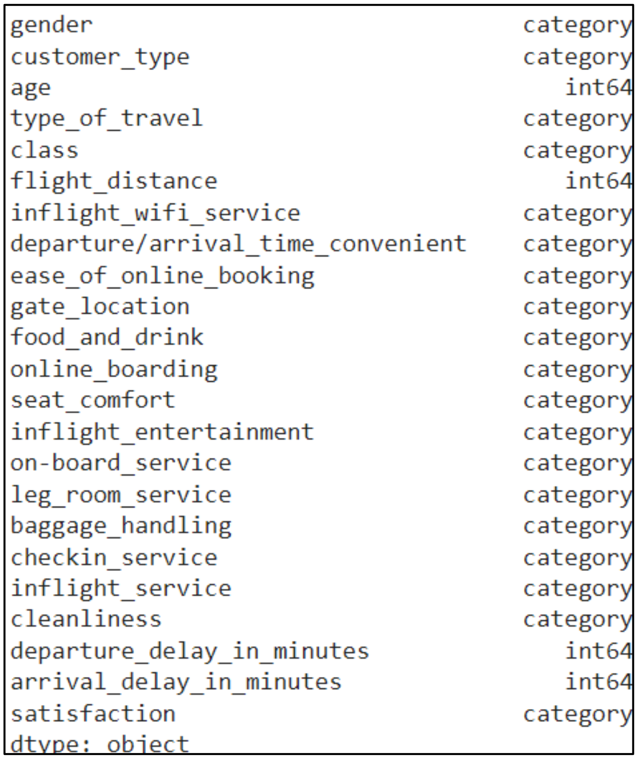
Overall, the exploratory analysis gave important insights into the properties of the dataset and indicated possible topics for additional exploration. The findings emphasize the necessity of taking class imbalance into account, mitigating delays, and adapting services based on class type to improve overall passenger satisfaction and flight operations.

Data Preparation and Feature Engineering

1) Data Preparation Needs

1. Data Type Conversion/Transformation

We completed the required data type conversion for specified variables in our dataset. We've transformed the variables related to passenger feedback, such as 'Inflight Wi-Fi service', 'Departure/Arrival time convenient', 'Ease of Online booking', 'Gate location', 'Food and drink', 'Online boarding', 'Seat comfort', 'Inflight entertainment', 'On-board service', 'Leg room service', 'Baggage handling', 'Check-in service', 'Inflight service', and 'Cleanliness', into categorical variables, fig., 20



```
gender                                category
customer_type                         category
age                                      int64
type_of_travel                        category
class                                 category
flight_distance                          int64
inflight_wifi_service                 category
departure/arrival_time_convenient     category
ease_of_online_booking                category
gate_location                         category
food_and_drink                        category
online_boarding                       category
seat_comfort                          category
inflight_entertainment                category
on-board_service                      category
leg_room_service                      category
baggage_handling                      category
checkin_service                       category
inflight_service                      category
cleanliness                           category
departure_delay_in_minutes               int64
arrival_delay_in_minutes                 int64
satisfaction                          category
dtype: object
```

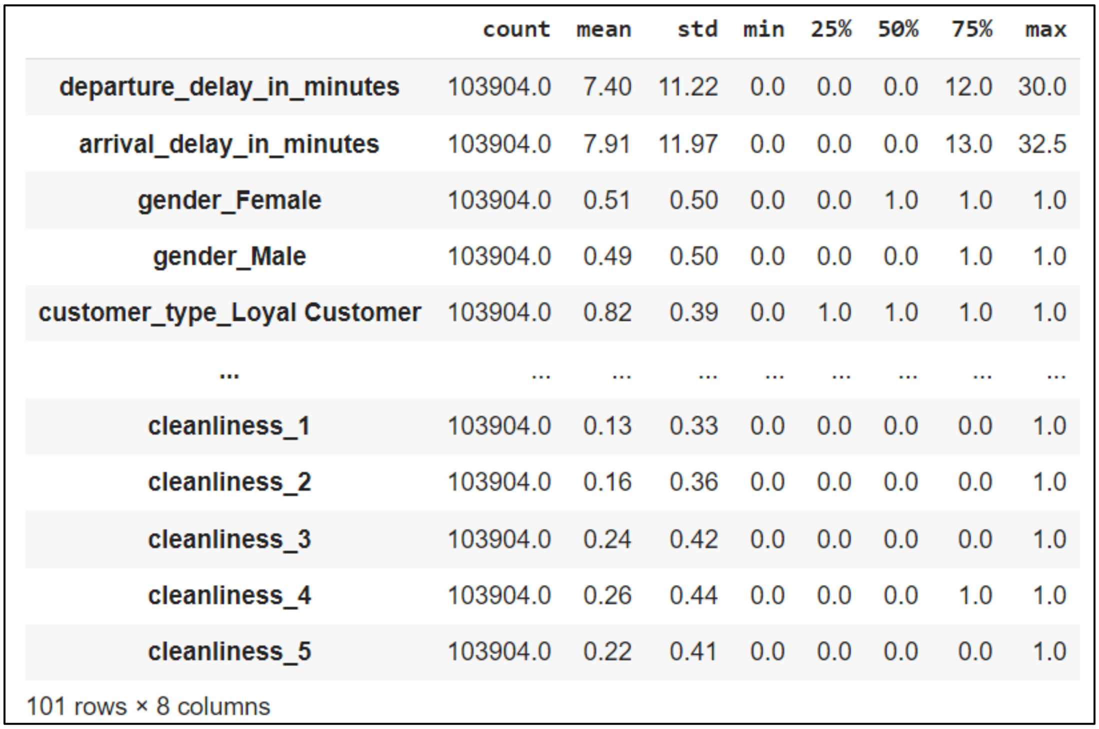
fig., 20

This modification conforms the data types to the nature of these variables as feedback properties. Converting to a categorical format allows for more accurate analysis, modeling, and visualization. Categorical variables enable meaningful feedback rating grouping and interpretation. This careful data preparation ensures that our dataset is properly formatted for in-depth investigation and analysis of passenger feedback and satisfaction.

2. Excluded Columns

Certain columns, such as 'id' and 'Unnamed 0', were purposefully left out of our research. These columns are generally made up of unique identifiers or indexing values that have little bearing on the analysis of passenger satisfaction or associated variables. By removing these columns, we may focus our analysis on the relevant variables that have a direct impact on our research aims. This strategic decision assures that our analysis is efficient, meaningful, and in line with our investigation objectives, fig., 21.

3. Standardization



	count	mean	std	min	25%	50%	75%	max
departure_delay_in_minutes	103904.0	7.40	11.22	0.0	0.0	0.0	12.0	30.0
arrival_delay_in_minutes	103904.0	7.91	11.97	0.0	0.0	0.0	13.0	32.5
gender_Female	103904.0	0.51	0.50	0.0	0.0	1.0	1.0	1.0
gender_Male	103904.0	0.49	0.50	0.0	0.0	0.0	1.0	1.0
customer_type_Loyal Customer	103904.0	0.82	0.39	0.0	1.0	1.0	1.0	1.0
...	...	...	...	...	...	...	...	...
cleanliness_1	103904.0	0.13	0.33	0.0	0.0	0.0	0.0	1.0
cleanliness_2	103904.0	0.16	0.36	0.0	0.0	0.0	0.0	1.0
cleanliness_3	103904.0	0.24	0.42	0.0	0.0	0.0	0.0	1.0
cleanliness_4	103904.0	0.26	0.44	0.0	0.0	0.0	1.0	1.0
cleanliness_5	103904.0	0.22	0.41	0.0	0.0	0.0	0.0	1.0

101 rows × 8 columns

fig., 21

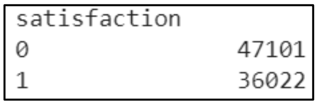
We choose to use the min-max scaler to normalize two variables: "arrival delay in minutes" and "departure delay in minutes." The fact that all other variables had previously been turned into a range of 0 to 1 after being converted into dummy variables influenced this decision which is discussed in detail in the Data Preparation Needs and Engineering (Section 3.0). By using the min-max scaler on "arrival delay" and "departure delay," we ensure that these variables are likewise on a comparable scale, allowing for fair and consistent comparisons between all our features. This standardization improves the performance and interpretability of our analysis and models.

4. Imputations

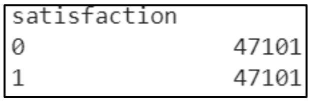
We discovered missing values in the "arrival delays in minutes" column during our data preparation process. We used the dependable approach of median imputation to remedy this. This entailed replacing missing values in the "arrival delays in minutes" column with the column's median value. The median was chosen because of its capacity to withstand the impact of outliers, ensuring that extreme numbers do not adversely affect the imputed findings. This is being discussed more in Data Exploration (Section 2.0).

2) Oversampling

We used the Synthetic Minority Over-sampling Technique (SMOTE) to correct the imbalance in our target variable, where 'satisfied' accounts for 36,022 cases and 'dissatisfied' accounts for 47,101 instances. Using the 'auto' setting in SMOTE was our special method.



```
satisfaction
0               47101
1               36022
```



```
satisfaction
0               47101
1               47101
```

SMOTE's 'auto' mode dynamically decides the number of synthetic instances to generate for the minority class, aiming for a more equal distribution. This necessitated building synthetic instances of the 'dissatisfied' class to align its representation with that of the 'satisfied' class in our scenario.

We hoped to effectively counteract the class imbalance by using the 'auto' mode in SMOTE, allowing our research to produce more accurate insights and forecasts. This technique ensures that both classes are treated similarly, increasing the dependability of our data, and allowing for a more comprehensive knowledge of the elements impacting customer pleasure in the context of an airline.

3) Feature Engineering

1. Creation of Bins

We understood the significance of transforming continuous variables into meaningful categorical segments as part of our entire feature engineering strategy. This procedure, known as binning, required categorizing numerical factors such as age and flying distance into separate intervals or "bins." We hoped to simplify complex data distributions and discover useful patterns that might be buried by the variables' continuous nature.

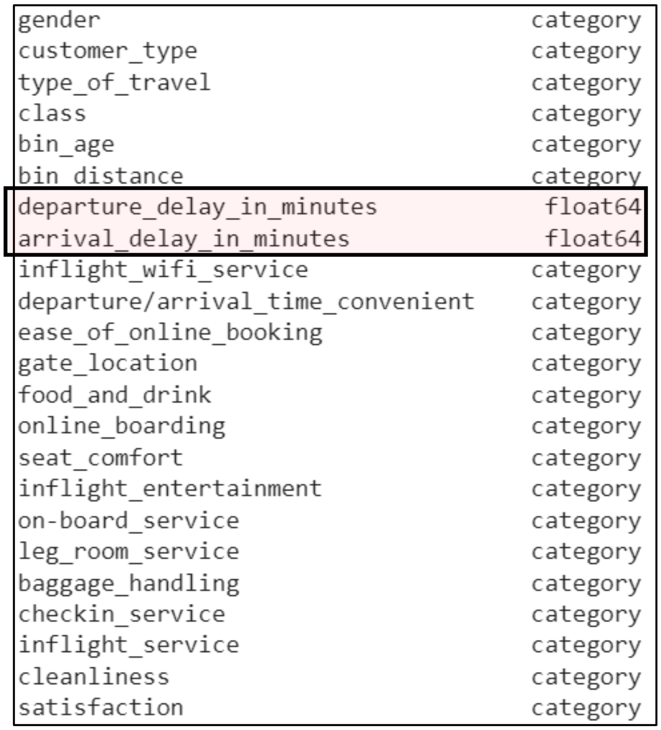
For example, the variable "age" was divided into specified age groups, while the variable "flight distance" was divided into intervals based on how much distance was travelled. This modification enabled us to evaluate trends and patterns within these intervals, revealing insights that would have been difficult to discover with raw continuous values.

We got two major benefits by using binning: improved interpretability and the ability to capture nonlinear interactions that may occur within these variables.

Our strategic application of this technique leads to a more thorough analysis by providing a clearer picture of how these variables influence passenger pleasure. This methodical approach to feature engineering improves our research and allows for more accurate model construction and interpretation.

2. One-hot encoding

We needed to convert categorical variables into a format suitable for analysis and modeling during the data preprocessing step. One-shot encoding developed as a critical approach for accomplishing this change. This procedure entails expressing categorical variables as binary columns, with each distinct category forming its own column with binary values (0 or 1).



```
gender                                category
customer_type                         category
type_of_travel                        category
class                                 category
bin_age                               category
bin_distance                          category
departure_delay_in_minutes             float64
arrival_delay_in_minutes               float64
inflight_wifi_service                 category
departure/arrival_time_convenient     category
ease_of_online_booking                category
gate_location                         category
food_and_drink                        category
online_boarding                       category
seat_comfort                          category
inflight_entertainment                category
on-board_service                      category
leg_room_service                      category
baggage_handling                      category
checkin_service                       category
inflight_service                      category
cleanliness                           category
satisfaction                          category
```

fig., 22

All the categorical variables which are mentioned in the fig., 22 are used for one-hot encoding apart from numeric variables which are marked in red. For instance, variables like "class" and "type of travel" were encoded using one-hot encoding.

The "class" variable, representing passenger class categories such as Business, Eco, and Eco Plus, was transformed into distinct binary columns for each class. Similarly, the "type of travel" variable, encompassing categories like Business travel and Personal travel, was also encoded using this technique. Same technique was used for each categorical variable. One-hot encoding not only allows us to keep categorical data intact, but it also allows machine learning algorithms to effectively use these variables in our research. This transformation ensures that our research is reliable and fair while capturing the subtleties of categorical attributes. We improve the quality of our dataset for later analysis, modelling, and decision-making by including one-hot encoding.

Model Exploration

1) Modeling introduction

Our modelling technique was a meticulously planned procedure aimed at unravelling the complexities of passenger satisfaction in the airline business. We attempted to deliver in-depth insights and practical recommendations for improving passenger experience by seamlessly integrating data preparation, feature engineering, model selection, and hyperparameter tuning. Our method included a variety of methods, including Logistic Regression, Decision Trees, Random Forest, and Gradient Boosting. We were able to acquire a wide range of data patterns and interactions because to our diversified choices. Grid search was used to fine-tune hyperparameters for Decision Trees. Grid search evaluated hyperparameter combinations systematically, assuring optimal model performance.

Model performance was assessed using ROC-AUC Score, and accuracy metrics. We were able to glean practical insights from the models' outputs because to our emphasis on interpretability. A feature significance analysis was performed within the Random Forest model to find variables influencing passenger satisfaction. This research provided a clear picture of which elements were important in driving satisfaction levels.

Our modeling strategy was designed to provide airline companies with actionable insights for improving passenger satisfaction. We facilitated strategic decision-making to create improvements and competitive advantage by leveraging the power of data-driven modelling.

We hoped to provide these firms with a detailed grasp of passenger satisfaction patterns, supported by accurate projections and educated recommendations, using this comprehensive modelling technique. This strategy is aimed at producing practical improvements that meet the needs of passengers and contribute to a better airline experience.

2) Modeling Techniques

For all the modeling techniques we employed, we partitioned our dataset into training and testing subsets. This division adhered to an 80/20 split ratio for both the training and testing data, respectively.

1. Logistic Regression

Using the Logistic Regression technique, our examination of passenger satisfaction delves into the complexities of the airline sector. This model, noted for its interpretability and predictive capacity, allowed us to get significant insights into the elements that influence passenger satisfaction. Because the outcome is a probability, the dependent variable has a range of 0 to 1. A logit transformation is performed to the odds in logistic regression, which is the probability of success divided by the probability of failure.

We chose Logistic Regression because of its transparency and ability to quantify the influence of each characteristic. Because of the model's simplicity, we were able to simply interpret the coefficients and understand how each variable affected passenger pleasure.

Model performance was assessed using criteria such as ROC-AUC Scores and accuracy matrix. The coefficients linked with each feature revealed the size and direction of their impact on predicting passenger satisfaction.

Model Performance

Accuracy: 0.936

Confusion Matrix: In percentages fig., 23

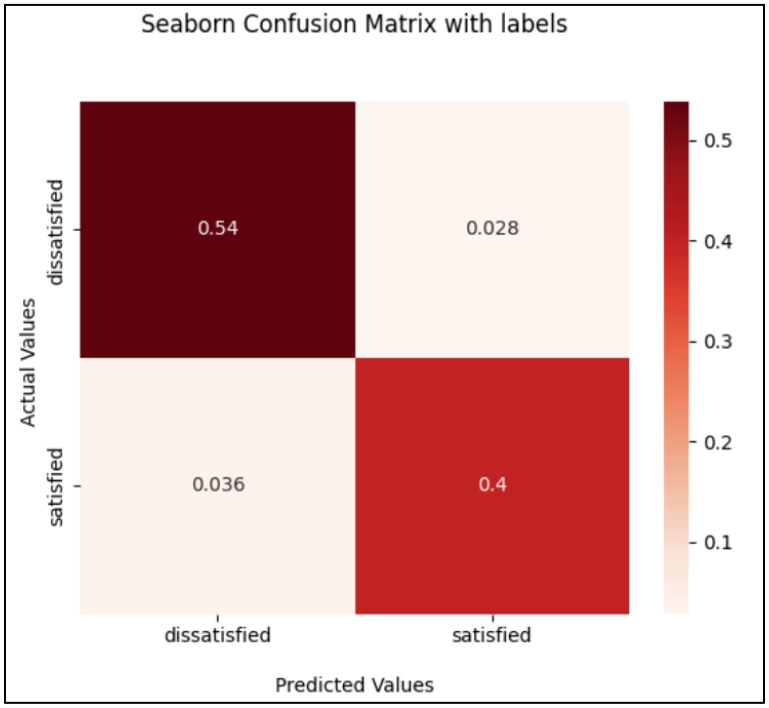


fig., 23

The predictive performance of Logistic Regression models can be broken down as follows:

Predicted unsatisfied travellers: The Logistic Regression model correctly predicted 54% of the unsatisfied travellers. This means that the program properly identified and predicted 54% of the passengers who reported being dissatisfied.

Predicted Satisfied Passengers: The model, on the other hand, predicted that 40% of the passengers would be satisfied. This means that the model accurately categorized and forecasted 40% of all guests who reported being satisfied with their trip.

ROC – AUC Score: 0.9798 fig., 24

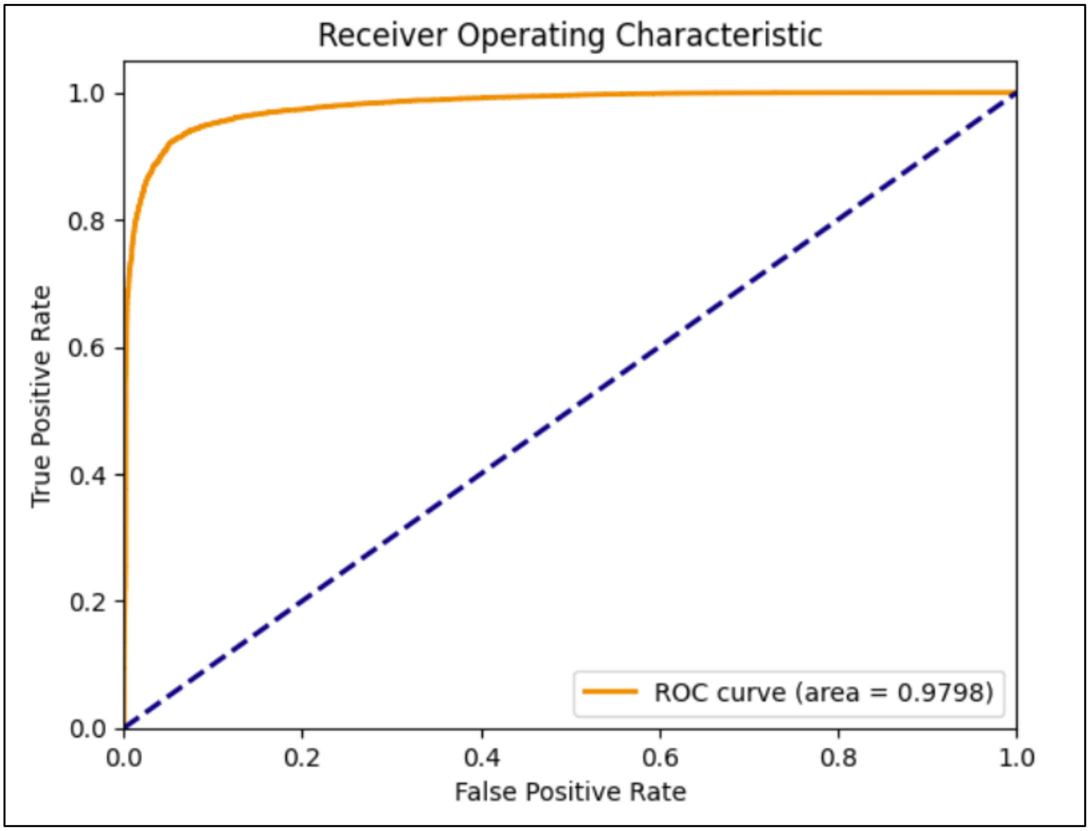
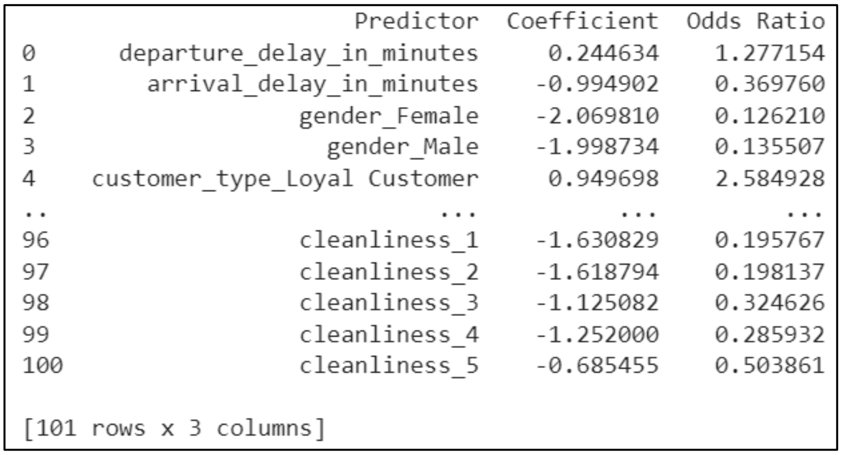


fig., 24

Logistic Regression has a ROC-AUC of 0.9798. This score assesses the model's ability to differentiate between positive and negative classifications. A higher score implies that the model is better at appropriately rating positive cases higher in terms of expected probabilities than negative examples. A score of 0.9798 indicates an excellent performance in classifying passenger pleasure in this scenario.

Coefficients and Odds Ratios: The dataset has 101 rows, as shown in fig., 25 In the next analysis, we will look at how odds ratios and coefficients are interpreted for two traits, while the explanations for the remaining variables will follow the same path for explanation. This concentrated approach enables us to give in-depth insights while keeping the overall dataset display concise.



```
                        Predictor  Coefficient  Odds Ratio
0      departure_delay_in_minutes     0.244634    1.277154
1        arrival_delay_in_minutes    -0.994902    0.369760
2                   gender_Female    -2.069810    0.126210
3                     gender_Male    -1.998734    0.135507
4    customer_type_Loyal Customer     0.949698    2.584928
..                            ...          ...         ...
96                  cleanliness_1    -1.630829    0.195767
97                  cleanliness_2    -1.618794    0.198137
98                  cleanliness_3    -1.125082    0.324626
99                  cleanliness_4    -1.252000    0.285932
100                 cleanliness_5    -0.685455    0.503861

[101 rows x 3 columns]
```

fig., 25

Predictor: Departure Delay in Minutes

The coefficient is 0.244634.

The odds ratio is 1.277154.

Explanation: The probability of a passenger being satisfied increase by approximately 1.277 times for every unit increase of departure delay in minutes.

This suggests that as departure time grows, so does the likelihood of a passenger being satisfied with their trip. It's worth noting that, while the coefficient represents the change in log-odds, the probabilities ratio provides a more natural understanding of the impact on the odds of satisfaction.

Predictor: Customer Type - Loyal Customer

The coefficient is 0.949698.

The odds ratio is 2.584928.

Explanation: When a passenger is designated as a "Loyal Customer," their chances of being satisfied improve nearly 2.585 times. This shows that devoted consumers are more likely than other customer categories to have a positive satisfaction experience. The positive correlation shows that being a loyal customer increases your chances of satisfaction.

2. Random Forest

The Random Forest technique, recognized for its ensemble of decision trees that collectively produce solid predictions, was used to analyze passenger satisfaction. We were able to use this modelling technique to uncover complicated linkages within the data and provide accurate estimates for passenger satisfaction levels.

To improve the performance of our Random Forest model, we used many estimators, with the value set at 500. This strategy guarantees that the model benefits from the combined insights of many decision trees, resulting in strong and accurate predictions of passenger pleasure.

Random Forest employs an ensemble of decision trees, each of which has been trained on a different subset of the data. This variety reduces overfitting and improves the model's capacity to capture complicated patterns.

Random Forest excels in making accurate predictions by combining the results of several decision trees. This ensemble strategy ensures that the model generalizes effectively to new data, avoiding the drawbacks of relying too heavily on individual trees.

Model Performance

Accuracy: 0.96

Confusion Matrix: In percentages fig., 26

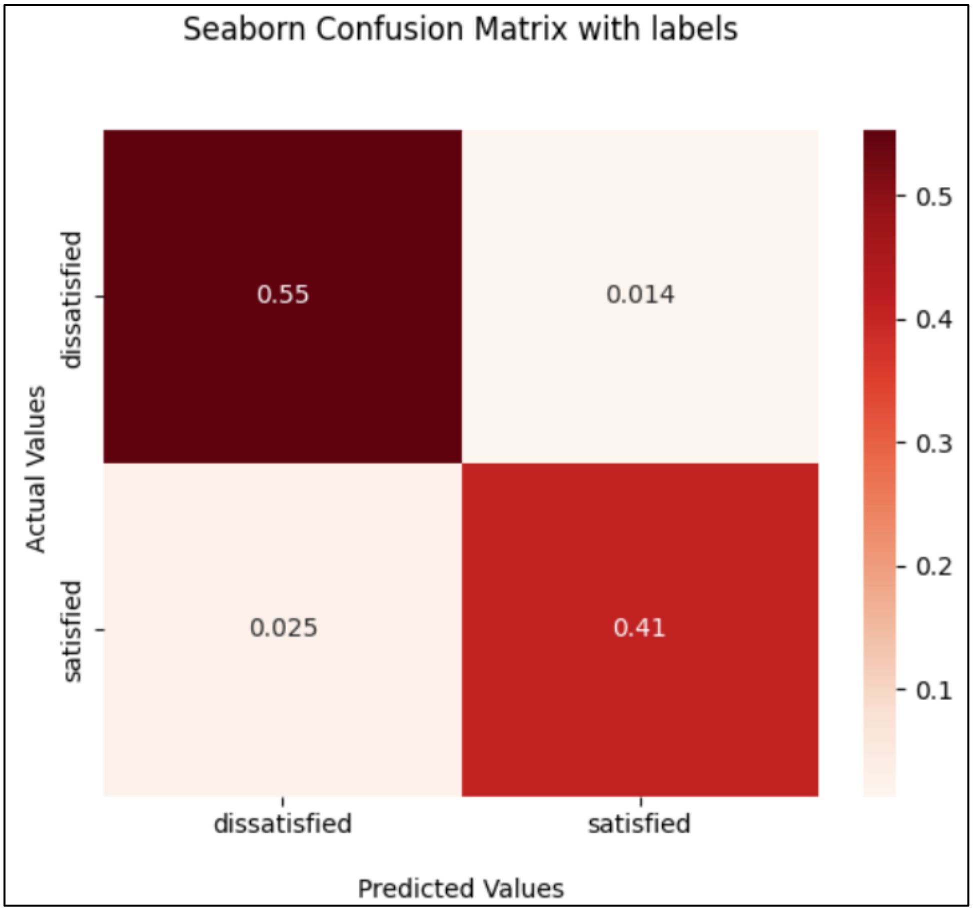


fig., 26

Predicted Dissatisfied Passengers: The Random Forest model successfully predicted 55% of the passengers who were actually dissatisfied with their experience. This means that the program properly identified and predicted 55% of the passengers who reported being dissatisfied.

Predicted Satisfied Passengers: In contrast, the model predicted that 41% of passengers would be satisfied with their journey. In other words, the model correctly identified and predicted 41% of all passengers who reported being satisfied with their trip.

ROC – AUC Score: 0.9935 fig., 27

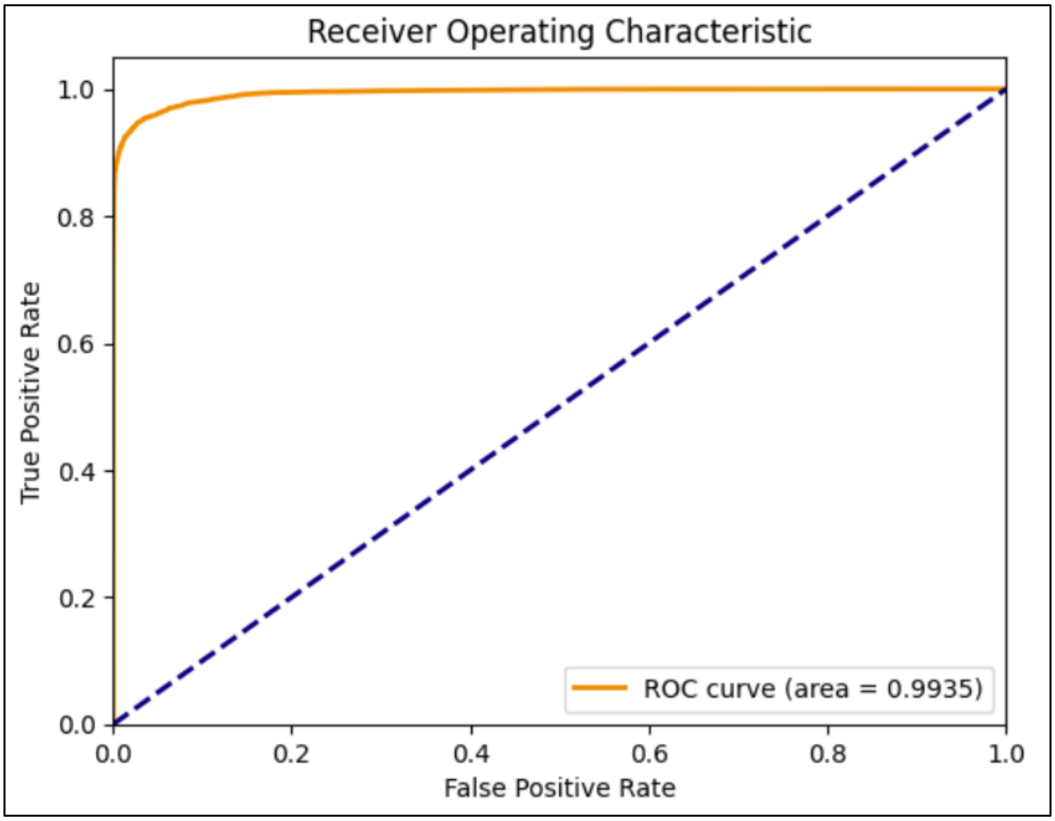
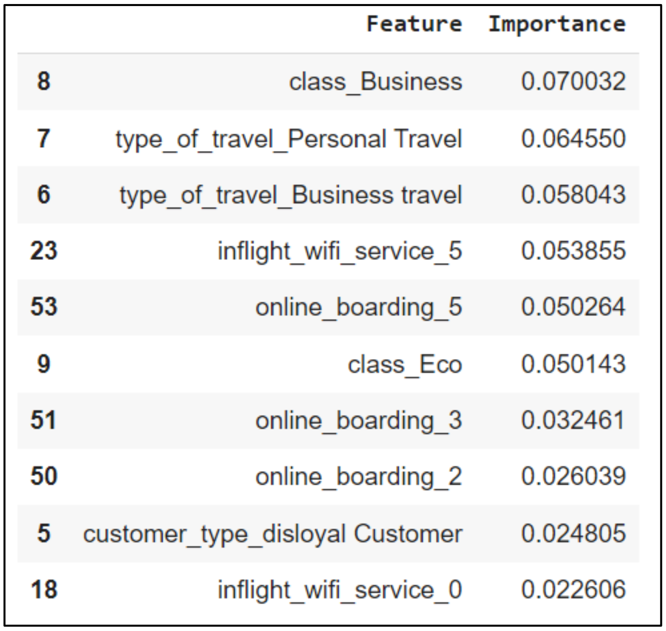


fig., 27

Feature Importance: The concept of feature importance is critical in our study. It entails quantifying the impact of individual characteristics on forecasting passenger pleasure. This quantification assists us in determining which variables have the greatest impact on correct forecasts and which have the least impact.



	Feature	Importance
8	class_Business	0.070032
7	type_of_travel_Personal Travel	0.064550
6	type_of_travel_Business travel	0.058043
23	inflight_wifi_service_5	0.053855
53	online_boarding_5	0.050264
9	class_Eco	0.050143
51	online_boarding_3	0.032461
50	online_boarding_2	0.026039
5	customer_type_disloyal Customer	0.024805
18	inflight_wifi_service_0	0.022606

fig., 28

As per fig., 28 we can see our top 10 variables which are important for the airline companies. These variables play an important role in deciding the passenger's satisfaction. Here is the explanation of some of them.

According to feature importance analysis, the most influential aspects include the passenger's travel class (with a concentration on Business class), the purpose of travel (Business or Personal), and the quality of in-flight Wi-Fi connectivity. Furthermore, the simplicity of online boarding and the passenger's class (Economy) are important predictors of passenger satisfaction. These findings highlight the crucial factors that affect passenger experiences and influence overall satisfaction levels.

In our case, if a passenger is travelling in Business class (class_Business), on a Personal Travel (type_of_travel_Personal Travel), rates the inflight Wi-Fi service at

5 (inflight_wifi_service_5), and has a smooth online boarding experience (online_boarding_5), the passenger is more likely to be satisfied.

However, if a person travels in Economy (class_Eco), the goal of trip is Business (type_of_travel_Business travel), and the in-flight Wi-Fi service or online boarding experience is ranked lower, the model may forecast a lower likelihood of satisfaction for that passenger.

3. Decision Trees

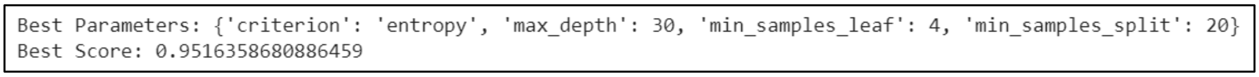
Decision Trees are an effective technique for analyzing passenger satisfaction dynamics. These models work by segmenting data into hierarchical nodes based on certain features and then predicting passenger satisfaction levels.

Decision Trees make decisions by asking questions based on features in a step-by-step procedure. Each feature is assessed to discover the most relevant splits, allowing the model to distinguish between satisfied and dissatisfied passengers.

The model ranks each feature in terms of its efficacy in differentiating between satisfaction levels. The model identifies critical factors that have a significant impact on passenger satisfaction estimates via subsequent splits.

Decision Trees can record nonlinear relationships in data, allowing them to understand intricate interactions between different qualities.

To optimize the model's performance, we used GridSearchCV in conjunction with Decision Trees. GridSearchCV entails systematically trying various hyperparameter combinations to determine the ideal configuration that produces the best results. GridSearchCV allowed us to fine-tune parameters like maximum depth, minimum samples per leaf, and splitting criteria in the context of Decision Trees. This rigorous method guaranteed that the Decision Tree model was fully optimized, resulting in precise and trustworthy estimates of passenger pleasure fig., 29



```
Best Parameters: {'criterion': 'entropy', 'max_depth': 30, 'min_samples_leaf': 4, 'min_samples_split': 20}
Best Score: 0.9516358680886459
```

fig., 29

Model Performance

Accuracy: 0.951

Classification summary: In percentages fig., 30

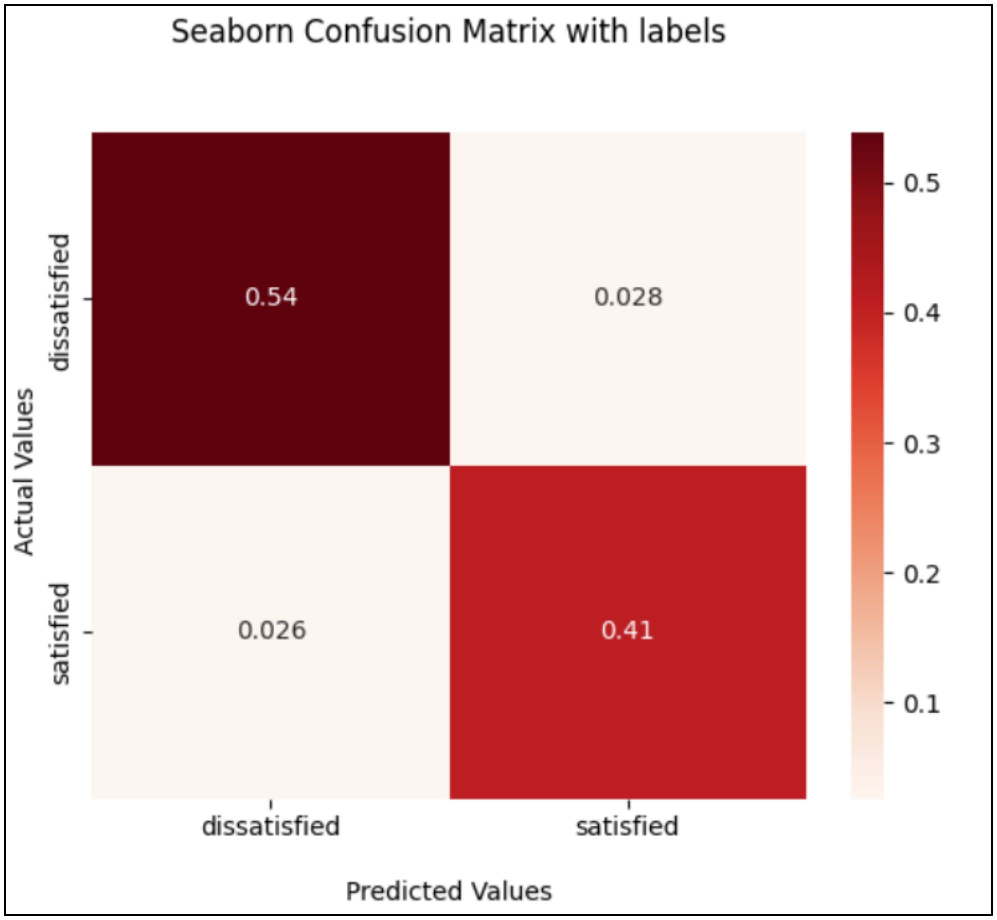


fig., 30

Predicted Dissatisfied Passengers: The Decision Trees model correctly predicted 54% of the passengers that were dissatisfied with their trip. This means that 54% of the passengers who reported being dissatisfied were accurately recognized and predicted by the algorithm.

Predicted Satisfied Passengers: In contrast, the model predicted that 41% of passengers would be satisfied with their journey. In essence, the algorithm correctly categorized and predicted 41% of all guests who reported being satisfied with their trip.

ROC-AUC Scores: 0.9447 fig., 31

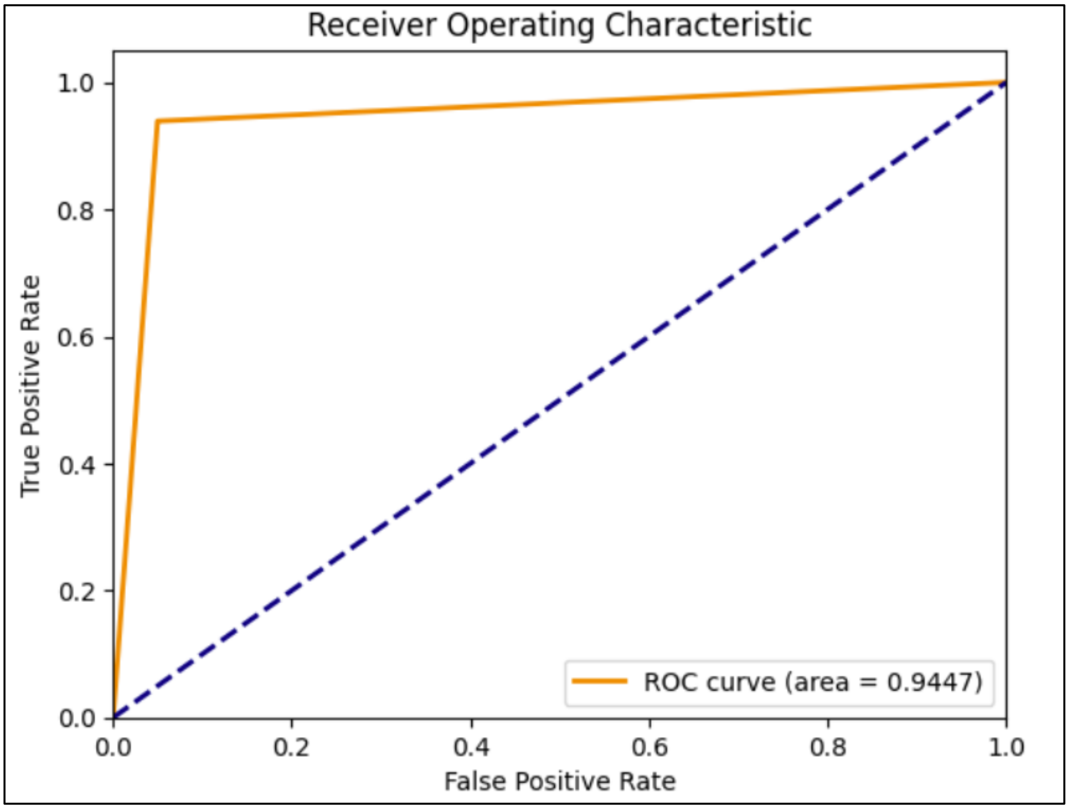


fig., 31

Features Selection: The "Business Class" feature is the first split in the Decision Tree. This means that the model first differentiates between passengers travelling in Business Class and those who are not. Following this break, the tree splits into two alternative routes based on the "Inflight Wi-Fi Services" and "Online Boarding" scores. fig., 32

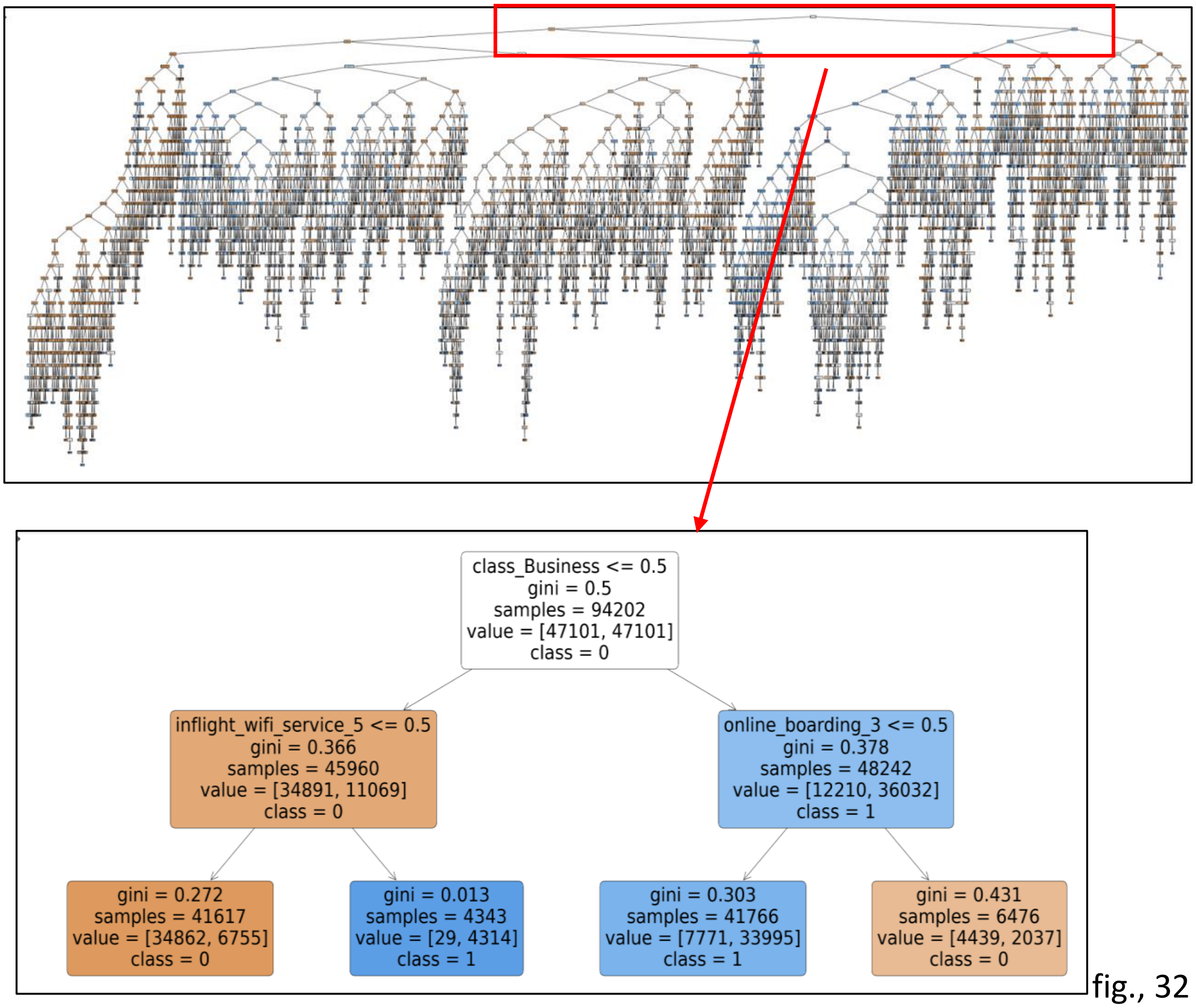


fig., 32

4. Gradient Boosting

The Gradient Boosting algorithm is a reliable method for uncovering complex patterns in passenger satisfaction data. This method entails integrating the strengths of multiple models iteratively to generate a refined ensemble that outperforms its separate components. Gradient Boosting incorporates techniques to prevent overfitting, ensuring that the ensemble model generalizes well to new data. The model analyzes the relevance of each trait, allowing us to determine which attributes are most important in making correct predictions.

To control model complexity and improve generalization, regularization approaches are used.

During the modelling process, we deliberately tweaked the Gradient Boosting model's hyperparameters to improve its prediction performance. The hyperparameter selection process was thorough to achieve high accuracy while avoiding overfitting. We specifically put the number of estimators option to 500, indicating that the ensemble will include 500 decision trees. This large number of trees enables the model to catch a wide range of patterns in the data, resulting in more accurate predictions. Furthermore, a learning rate of 0.1 was chosen to limit each individual tree's contribution, preventing over-reliance on any tree and enabling the model to generalize effectively. Finally, we used the max depth argument to limit the depth of each decision tree to three. This restriction on tree depth guarantees that the model captures the most important traits without becoming overly complicated, improving its ability to generalize to new data. This set of hyperparameters is a well-informed strategy that maximizes the model's predictive skills while keeping it adaptable to unknown events.

Model Performance

Accuracy: 0.957

Confusion Matrix: In percentages fig., 33

<u>Dissatisfied Passengers (55% Predicted):</u> The Gradient Boosting algorithm correctly predicted 55% of the passengers that were actually dissatisfied with their experience. This means that the model properly recognized and predicted

dissatisfaction in 55% of the passengers who reported being dissatisfied. This is consistent with the model's ability to detect displeasure signals.

Predicted Satisfied Passengers (41%): The model projected that 41% of passengers will be satisfied with their journey. The model correctly categorized and predicted that 41% of all passengers who reported being satisfied were satisfied. While this percentage is slightly smaller than that of unhappy passengers, it still demonstrates the model's ability to identify satisfied passengers.

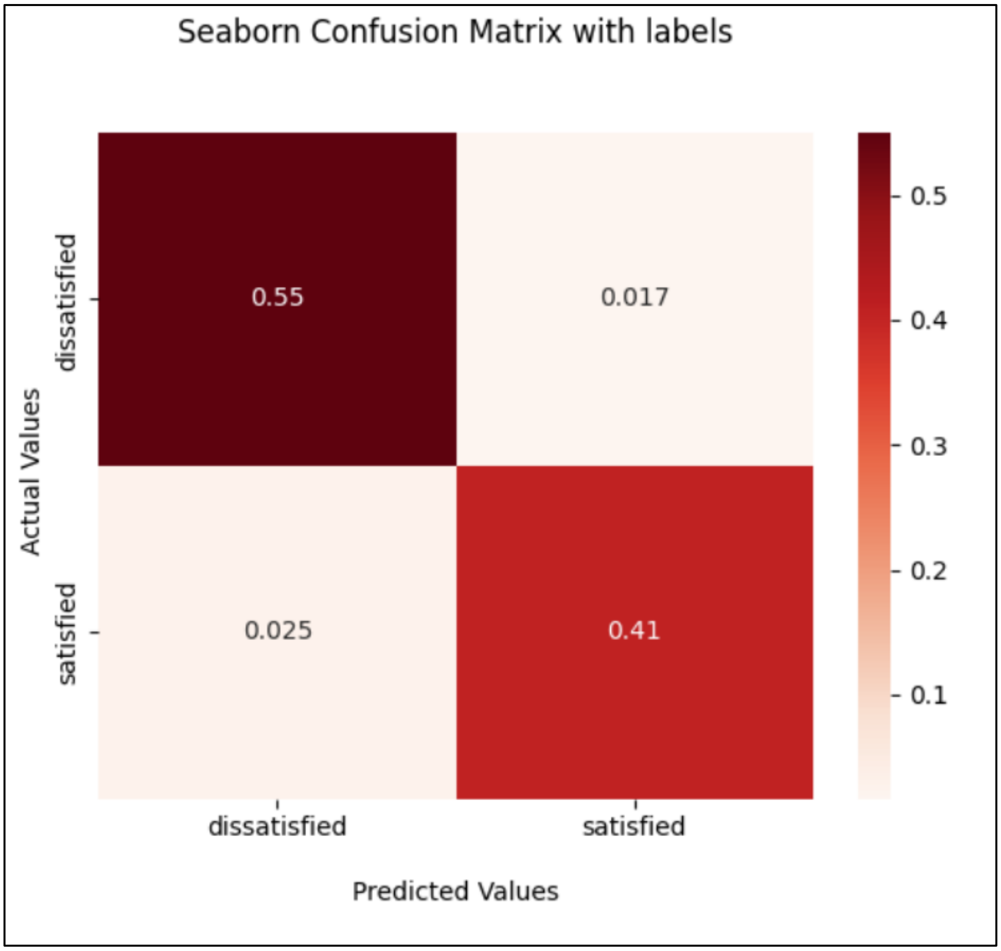


fig., 33

ROC – AUC Score: 0.9932 fig., 34

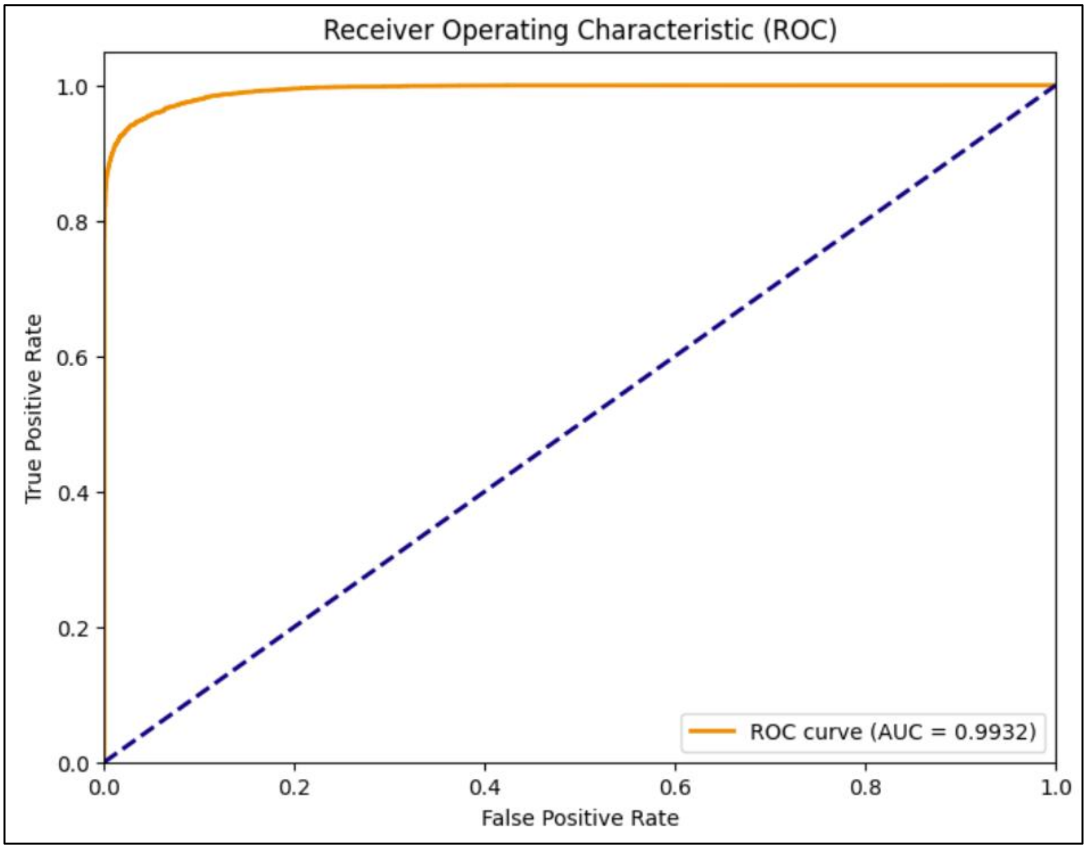


fig., 34

3) Model Comparison

During the evaluation of our prediction models, we used two critical metrics, accuracy, and ROC-AUC score, to allow for a thorough comparison of their performance fig., 35. The proportion of correct predictions in respect to the total number of forecasts is measured by accuracy, a fundamental metric. This statistic provides an overall picture of the model's ability to forecast passenger satisfaction levels. However, given the inherent class imbalance in our dataset, where satisfied and dissatisfied passengers may be represented unequally, we also used the ROC-AUC score.

The ROC-AUC score, or Receiver Operating Characteristic - Area Under the Curve, provides an in-depth look at a model's ability to discern between levels of satisfaction. This metric demonstrates the model's ability to accurately rank and predict a wide range of passenger satisfaction scenarios by showing the True Positive Rate versus the False Positive Rate. When dealing with imbalanced classes or evaluating the model's performance across various classification thresholds, ROC-AUC becomes extremely useful.

We gained a thorough understanding of our models' capabilities by combining these metrics. While accuracy provided a general indicator of overall correctness, the ROC-AUC score shed light on their discriminatory abilities, allowing us to evaluate their performance beyond simple class predictions. These measures, when combined, led our decision-making process, assisting us in discovering the model that not only properly forecasts satisfaction levels but also excels at distinguishing between them.

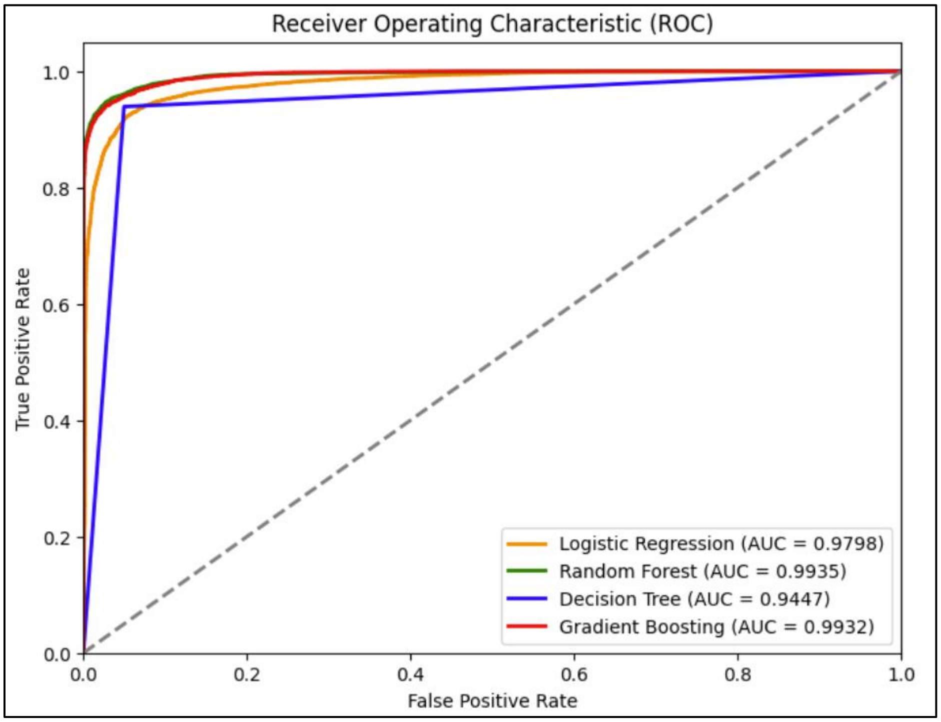


fig., 35

Model	Accuracy	ROC – AUC Score
Random Forest	96 %	0.9935
Gradient Boosting	95.7%	0.9932
Decision Trees	95.1%	0.9446
Logistic Regression	93.6%	0.9798

We compared four algorithms in our model evaluation: Random Forest, Gradient Boosting, Decision Trees, and Logistic Regression. The accuracy and ROC-AUC scores of each model were critical in determining its prediction ability.

The Random Forest excelled in accurate predictions and class distinction, with an amazing 96% accuracy and a high ROC-AUC score of 0.9935.

Gradient Boosting demonstrated good prediction and class separation capabilities with a 95.7% accuracy and a ROC-AUC score of 0.9932.

Decision Trees achieved 95.1% accuracy and a ROC-AUC of 0.9446, demonstrating strong prediction and reasonable class separation.

Logistic Regression obtained 93.6% accuracy with a ROC-AUC of 0.9798, exhibiting strong prediction and class separation abilities.

These ratings jointly influenced our model selection process, providing insight into their strengths for precise and effective prediction of passenger pleasure.

4) Model Selection

Following a thorough review of our models and consideration of their numerical performance indicators, we determined that the Random Forest model was the best fit. The Random Forest model had a remarkable accuracy of 96%, suggesting its strong ability to reliably forecast passenger satisfaction levels. Furthermore, its remarkable ROC-AUC score of 0.9935 demonstrated its exceptional ability to discern between satisfied and dissatisfied travellers.

While other models performed well, the Random Forest model continuously outperformed them in terms of accuracy and ROC-AUC score. The Random Forest technique is an appealing alternative for our passenger satisfaction prediction challenge due to its high accuracy and great class separation.

The Random Forest model was chosen to get the maximum prediction accuracy and reliability. It also demonstrates our dedication to make sound decisions based on both practical application and quantitative analysis.

5) Model Theory

The Random Forest algorithm is a versatile and effective ensemble learning method that may be used for classification as well as regression. During training, it builds many decision trees and generates predictions using a voting or averaging technique. The name "Random Forest" refers to the randomness included by the modelling process. The algorithm operates as follows:

- **Building a Tree Forest:** Multiple decision trees are constructed, each trained on a different portion of the training data and with a different set of features. This randomness promotes variation among the various trees, reducing overfitting and improving the model's generalizability.

- **Aggregation Predictive:** The projected classes of all individual trees are aggregated for classification problems using a majority voting process. To provide the final prediction for regression tasks, the projected values of all trees are summed.
- **Minimizing Overfitting:** The randomization supplied during tree construction keeps the model from fitting noise in the data, resulting in less overfitting. When many trees are combined, the model benefits from their collective wisdom, which frequently results in more accurate predictions.
- **Importance Feature:** Random Forest determines the relevance of each feature by calculating how much the model's accuracy decreases when a specific feature is removed. This feature importance data assists in determining which attributes contribute significantly to the predictions.
- **Evaluation of Out-of-Bag Items:** During the training of each tree, a subset of data (about one-third) is not used. Without the requirement for a separate validation set, this "out-of-bag" data can be used to evaluate the model's performance.
- **Noise Tolerance:** The averaging and voting procedures of Random Forest reduce the impact of noisy or irrelevant variables, resulting in more consistent and reliable predictions.

In conclusion, Random Forest is a powerful modelling tool that leverages the characteristics of several decision trees to make accurate forecasts. Its capacity to control overfitting, handle multiple data formats, and deliver feature importance insights makes it a popular choice for addressing challenging classification and regression jobs in a variety of industries.

6) Model Assumptions and Limitations

Assumptions:

Tree Independence: The Random Forest model assumes that the individual decision trees in the ensemble are built independently. This enables the aggregation process to function properly, utilizing the diverse insights from each tree.

No Multicollinearity: While Random Forest is resistant to feature multicollinearity, it is still advantageous to have features that contribute unique information to minimize repetition.

Random Sampling: The model implies that the random subsampling of data for each tree accurately represents the underlying population distribution.

Limitations:

Complexity: Because of the amalgamation of several trees, the Random Forest model can be complex and difficult to grasp, despite its usefulness. When compared to simpler models, extracting insights regarding the links between features and predictions can be less apparent.

Memory and computation: Building many decision trees and combining their findings can be memory and computationally demanding, especially when dealing with huge datasets.

Overfitting Risk: While Random Forest lowers overfitting by design, there is still a risk of overfitting if the number of trees in the ensemble is too large or the depth of the trees is not well regulated.

Impact of Data Balance: Random Forest may not handle class imbalance as well as other specialized strategies, such as boosting algorithms.

Tuning Hyperparameters: Choosing the best hyperparameters for the model, such as the number of trees or maximum depth, necessitates careful thought and experimentation.

Extrapolation Limitations: When dealing with data that is outside the range of the training data, the model's prediction ability may suffer.

Handling Categorical Variables: Random Forest can handle categorical variables, but it may not always successfully capture complicated relationships between them, especially when there are several layers.

7) Model Sensitivity to Key Drivers

We included a wide range of factors in our dataset, including passenger demographics, journey details, service experiences, and overall satisfaction. The Random Forest model is responsive to the following main drivers:

Demographics: Gender, Customer Type, Mode of Travel, and Class are all demographic factors that influence passenger satisfaction. The intricate interconnections between these variables and how they influence the chance of satisfaction are captured by our Random Forest model.

Travel Details: Age and Flight Distance have been discretized into Bin Age and Bin Distance. Our model investigates how various age groups and travel distances affect passenger pleasure. This investigation determines whether certain age groups or distances are more likely to result in satisfaction.

Delays in Departure and Arrival: Departure and arrival delays in minutes are critical in determining how delays effect satisfaction. Our approach considers whether passengers become satisfied or not as delays lengthen. This insight assists us in determining the point at which dissatisfaction becomes noticeable.

Experiences in Customer Service: Inflight Wi-Fi Service, Ease of Online Booking, Food and Drink, Online Boarding, Seat Comfort, Inflight Entertainment, and other factors provide insight into how various service experiences contribute to customer satisfaction. Our approach identifies which components of the journey have the greatest impact on satisfaction.

Overall Service Level: Attributes such as Leg Room Service, Baggage Handling, Check-in Service, In-Flight Service, and Cleanliness represent the overall quality of service that customers get. Our Random Forest model identifies how these characteristics interact.

We may use the Random Forest model to determine the relative importance and sensitivity of each attribute in driving passenger satisfaction. The model's feature importance analysis ranks these features, allowing us to learn which elements are most important in predicting passenger pleasure. This knowledge enables us to make informed decisions about how to devote resources and efforts to improve the passenger experience and, ultimately, satisfaction levels.

Conclusion and Recommendations

1) Impact on Business Problem

Our detailed study of modern modelling approaches yielded useful insights in this endeavour to improve passenger pleasure and upgrade our services. Our efforts have resulted in a conclusive recommendation that has the potential to significantly impact our business situation. The Random Forest approach, which we advocate, has shown exceptional ability to predict passenger satisfaction levels with high accuracy.

The strategic use of the Random Forest approach has enormous promise for our airline. We got an in-depth grasp of the aspects that significantly influence passengers' contentment levels by researching into the primary determinants of passenger satisfaction. The model's sensitivity analysis reveals the complicated interaction of elements that impact passengers' impressions, from demographics to service experiences.

The ramifications of this forecasting ability go far beyond simple analysis. We can make informed judgments to improve the overall passenger experience if we have a clear understanding of passenger preferences. Our ability to predict and address pain areas, prioritize enhancements, and effectively deploy resources has considerably improved. This translates into tailored efforts that respond to our passengers' unique demands, resulting in improved satisfaction rates, stronger loyalty, and a competitive edge in the airline market environment.

Furthermore, the Random Forest model's insights provide a road map for strategic changes. We may direct our attention where it counts most by focusing on the traits that have the most influence. This enhancement not only improves passenger experiences but also increases operational efficiency.

The recommended model's scope encompasses several aspects of our organization. The Random Forest model has the ability to alter the way we approach customer satisfaction, from marketing efforts that resonate with certain passenger demographics to operational adjustments that minimize delays and improve services.

As we adopt this data-driven approach, we enter an era in which passenger happiness is a measurable and achievable outcome rather than a goal. We guide towards a future of happier passengers, enhanced brand reputation, and sustained growth by leveraging the power of the Random Forest model.

Finally, our journey through predictive analysis has revealed a road to meaningful transformation. We go on a path that combines business objectives with customer requirements, ultimately revolutionizing the passenger experience and setting new standards of excellence in the aviation sector, with the Random Forest model at the helm.

2) Recommended next steps

Implementation of Insights: Convert the Random Forest model's insights into actionable strategies. To steer your operational and service improvements, focus on the traits identified as significant drivers of passenger happiness.

Tailored Marketing Campaigns: Create marketing campaigns that speak to certain passenger segments identified in the model as influential. Create customised message that corresponds to the preferences and demands of various consumer types.

Operational Adjustments: Use the model's findings to create operational changes that reduce delays, increase service quality, and improve the overall travel experience. You can reduce passenger dissatisfaction by addressing pain areas.

Integration of the Feedback Loop: Integrate the model's predictions and insights into your feedback collection process. Use this data to validate and fine-tune your efforts, so generating a continual feedback loop for continuous progress.

Real-time Monitoring: Using the model's predictions, implement real-time monitoring of passenger satisfaction. This proactive strategy enables you to deal with problems as they develop and maintain continuously high levels of satisfaction.

Employee Training: Share your findings with your team to ensure a unified effort in providing great service. Train personnel to prioritize the features that are most important to passengers.

References

Jiang, X., Zhang, Y., Li, Y., & Zhang, B. (2022, July 1). *Forecast and analysis of aircraft passenger satisfaction based on RF-RFE-LR model*. Nature News. https://www.nature.com/articles/s41598-022-14566-3

16, J., & Wickramasinghe, S. (2021, July 16). *Bias & variance in Machine Learning: Concepts & Tutorials*. BMC Blogs. https://www.bmc.com/blogs/bias-variance-machine-learning/

Klein, T. (2020, February 20). *Airline passenger satisfaction*. Kaggle. https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction/code?select=train.csv

Government of Canada, S. C. (2017, May 8). *Age categories, Life Cycle Groupings*. Government of Canada, Statistics Canada. https://www.statcan.gc.ca/en/concepts/definitions/age2

11. correlation and regression: The BMJ. The BMJ | The BMJ: leading general medical journal. Research. Education. Comment. (2020, October 28). https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression

ChatGPT. (n.d.). https://chat.openai.com

Sklearn.preprocessing.MinMaxScaler. scikit. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

Satpathy, S. (2023, July 24). *Smote for imbalanced classification with python*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/

Saini, A. (2023, August 2). *Gradient boosting algorithm: A complete guide for beginners*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/#undefined