AIRLINE PASSENGER SATISFACTION ANALYSIS

The George Washington University

DATS 6103: Introduction to Data Mining

Team 2: Naiska Buyandalai, Wali Siddique, Sharon Joanna

December 16, 2024

Table of Contents

Introduction

In the highly competitive and customer driven airline industry, passenger satisfaction has become an important metric for assessing service quality. This metric is not only vital for maintaining customer loyalty but also for attracting new clientele.

Therefore, customer feedback plays a pivotal role in shaping service enhancements and influencing operational decisions. Passengers often provide opinions on various aspects of their trip, including seating comfort, food quality, and staff service. Analyzing customer satisfaction data offers invaluable insights into the factors that significantly impact passenger satisfaction, guiding airlines in optimizing their services to meet and exceed customer expectations.

This project utilizes a comprehensive dataset of passenger satisfaction surveys, encompassing demographic information, travel types, service ratings, and overall satisfaction levels. By employing machine learning models, this project aims to predict passenger satisfaction levels and identify key indicators of customer satisfaction. These insights will enable airlines to develop targeted strategies for enhancing the overall customer experience, ultimately improving their competitive position in the market.

Literature Review

This literature review examines three key studies that provide insights into airline passenger satisfaction and its determinants. These studies offer valuable context for our analysis and highlight important factors influencing customer experience in the airline industry.

In-flight Service Quality

Namukasa (2013) examined the influence of airline service quality on passenger satisfaction and loyalty. The study identified pre-flight, in-flight, and post-flight services as key determinants of passenger satisfaction. In-flight services, including seat comfort, food quality, and entertainment, were found to have the strongest impact on overall satisfaction.

Customer Segmentation and Satisfaction

Noviantoro et al. (2022) employed data mining techniques to segment airline passengers based on their satisfaction levels. The study identified distinct customer segments with varying preferences and satisfaction drivers, finding that business travelers and frequent flyers had different satisfaction determinants compared to leisure travelers.

Technology Adoption and Satisfaction

Shiwakoti et al. (2022) investigated the role of digital technologies in enhancing passenger satisfaction. The research highlighted how self-service technologies, mobile applications, and personalized digital experiences significantly influence customer perception and satisfaction levels across different passenger segments.

Dataset Overview

The Airline Passenger Satisfaction dataset, sourced from Kaggle, contains 129,880 rows and 24 columns, offering a comprehensive collection of customer feedback. The dataset contains passenger demographic information and various aspects of the travel experience, such as flight distance, gender, age, class, onboard service, and overall satisfaction.

Column description:

Gender: Gender of the passengers (Female, Male)

Customer Type: The customer type (Loyal customer, Disloyal customer)

Age: The actual age of the passengers

Type of Travel: Purpose of the flight of the passengers (Personal Travel, Business Travel)

Class: Travel class in the plane of the passengers (Business, Economy, Economy Plus)

Flight distance: The flight distance of this journey (miles)

Inflight Wi-Fi service: Satisfaction level of the inflight wife service (0: Not Applicable;1-5)

Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient

Ease of Online booking: Satisfaction level of online booking

Gate location: Satisfaction level of Gate location

Food and drink: Satisfaction level of Food and drink

Online boarding: Satisfaction level of online boarding

Seat comfort: Satisfaction level of Seat comfort

Inflight entertainment: Satisfaction level of inflight entertainment

On-board service: Satisfaction level of On-board service

Leg room service: Satisfaction level of Leg room service

Baggage handling: Satisfaction level of baggage handling

Check-in service: Satisfaction level of Check-in service

Inflight service: Satisfaction level of inflight service

Cleanliness: Satisfaction level of Cleanliness

Departure Delay in Minutes: Minutes delayed when departure

Arrival Delay in Minutes: Minutes delayed when Arrival

Satisfaction: Airline satisfaction level (Satisfaction, neutral or dissatisfaction)

Smart Questions

1. Which demographic groups (e.g., age, gender) report higher satisfaction levels?
2. Which factors are the strongest predictors of low or high satisfaction?
3. How do satisfaction levels vary by flight class (economy, business, first class)?

Data Cleaning

We performed several data cleaning steps to ensure the dataset was suitable for further analysis.

1. **Handling Missing Values:**
   - Survey response columns contained values ranging from 1 to 5, with 0 representing "Not Applicable". All instances of 0 were replaced with NA values to maintain consistency with the dataset's description.
   - All the rows containing NA values were removed, leaving us with 119,204 observations.
2. **Checking for Outliers:**
   - Summary statistics revealed relatively high values with respect to the average value in certain columns, such as maximum age (85) and maximum arrival and departure delays (around 1,590 minutes).
   - These values, while extreme, are realistic in real-world scenarios. For instance, flight delays can sometimes exceed standard durations, and passengers aged 85 are not uncommon. Therefore, we retained these outliers in the dataset.
3. **Data Formatting:**
   - All categorical variables (e.g., "Gender," "Customer Type," "Class") were inspected to ensure consistent formatting and valid categories.
   - Continuous variables (e.g., "Flight Distance," "Departure Delay in Minutes") were checked and no further adjustments were necessary.

Exploratory Data Analysis

To understand the distribution of customer satisfaction, we analyzed the satisfaction levels (“Satisfied” vs “Neutral or Dissatisfied”) with a breakdown by gender.

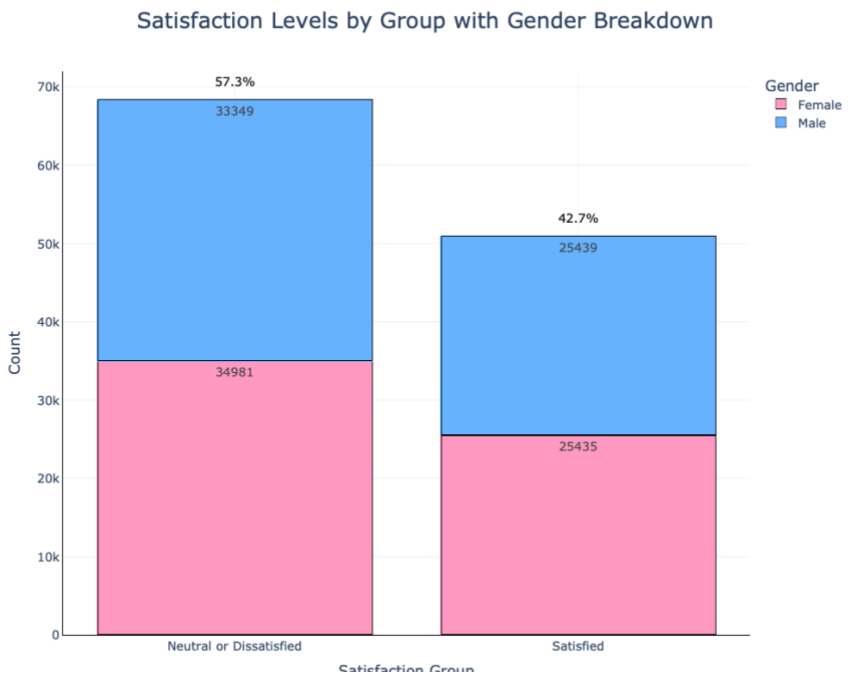


Figure 1. Satisfaction Levels by Group with Gender Breakdown

57.3% of the total responses were neutral or dissatisfied customers and 42.7% of the responses were from satisfied customers.

Furthermore, gender differences in both groups appear minimal. This finding suggests that gender is not one of the main indicators of customer satisfaction.

Analysis on Customer Age and Travel Class

The age distribution analysis reveals key insights into customer satisfaction and flight class preferences.

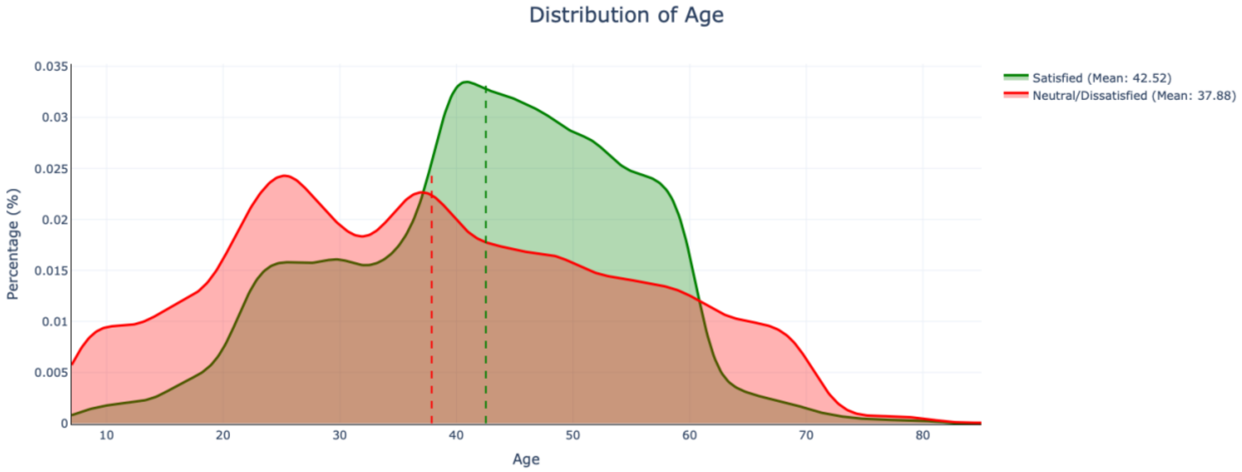


Figure 2. KDE Plot of Age

First, the density plot for satisfaction groups shows that the mean age for satisfied customers (42.52) is notably higher than that of neutral or dissatisfied customers (37.88). Additionally, the age distribution for satisfied customers is concentrated around the middle-aged group (40-60). This suggests that middle-aged individuals may have higher levels of satisfaction, potentially due to factors like comfort and convenience during their travel experience.

To explore this further, the boxplot of age distribution across flight classes highlights that passengers traveling in Business Class tend to have a higher median age compared to those in Eco Plus and

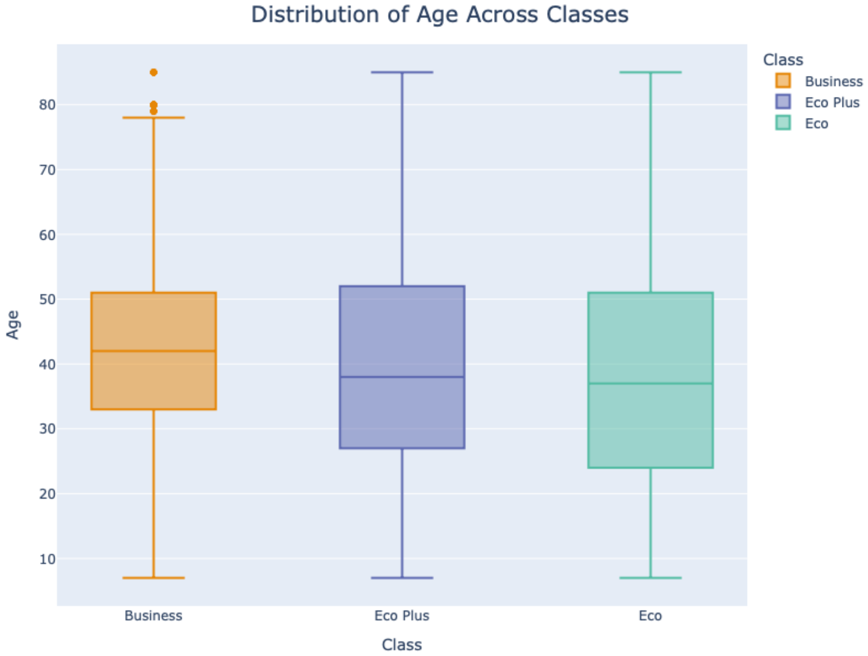


Figure 3. Boxplot of Age Distribution Across Flight Classes

Eco Class. This result aligns with our earlier observation, as middle-aged travelers (who are more likely to fly Business Class) might be business professionals whose needs and expectations are better met.

Analysis on Flight Distance and Travel Class

The analysis of flight distance reveals an interesting trend: more customers are satisfied with longer distance flight, and more customers are neutral or dissatisfied with shorter distance flight. The flight distance distribution plot indicates that satisfied customers tend to travel longer distances, with a mean flight distance of 1,579.92, while neutral or dissatisfied customers have a mean flight distance of 958.64.

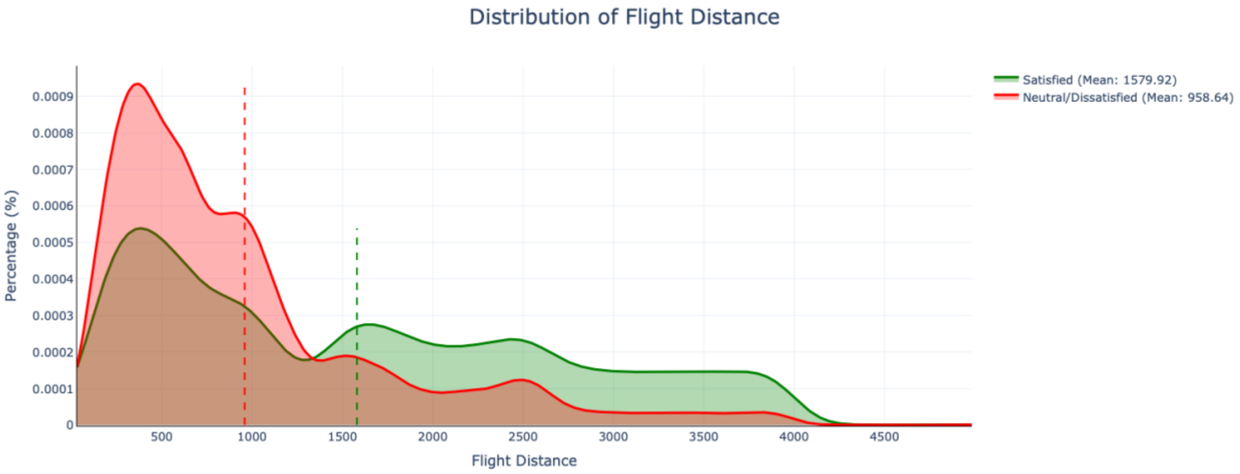


Figure 4. KDE Plot of Flight Distance

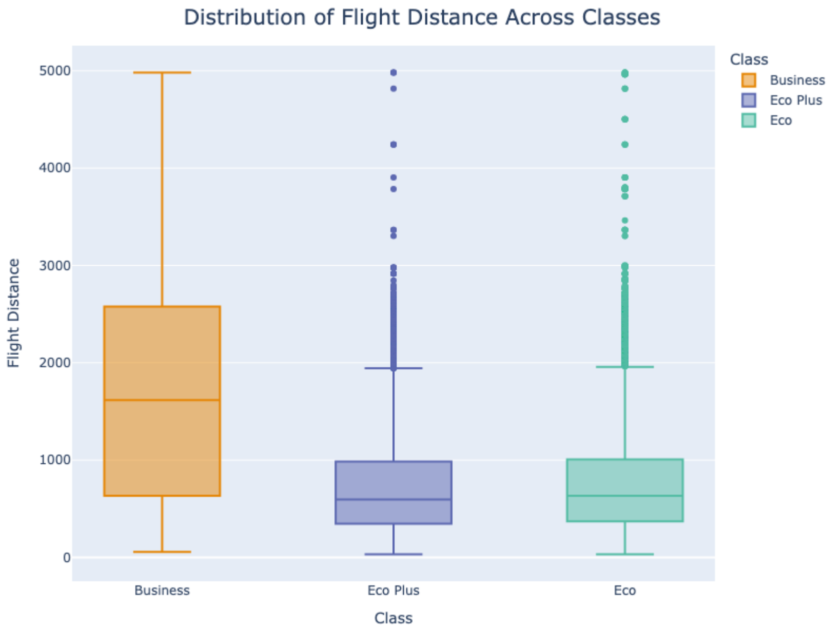


Figure 5. Boxplot of Flight Distance Distribution Across Flight Classes

To further investigate this trend, the boxplot of flight distance across travel classes shows that business class passengers travel significantly longer distances compared to passengers in Economy Plus and Economy classes. This observation suggests a plausible explanation for the higher satisfaction rates on longer flights: passengers on longer flights are more likely to opt for business class, where they receive greater comfort and better services, contributing to higher satisfaction levels.

In contrast, shorter flights where passengers are more likely to fly in economy or lower-tier classes, tend to show higher levels of dissatisfaction, potentially due to limited amenities and comfort.

Analysis on Departure and Arrival Delay

The distribution plots of both departure delay and arrival delay displays the relationship with customer satisfaction. Interestingly, in both plots, the distributions of satisfied customers and neutral or dissatisfied customers overlap significantly, indicating that delays, whether in departure or arrival, do not appear to have a strong impact on customer satisfaction.

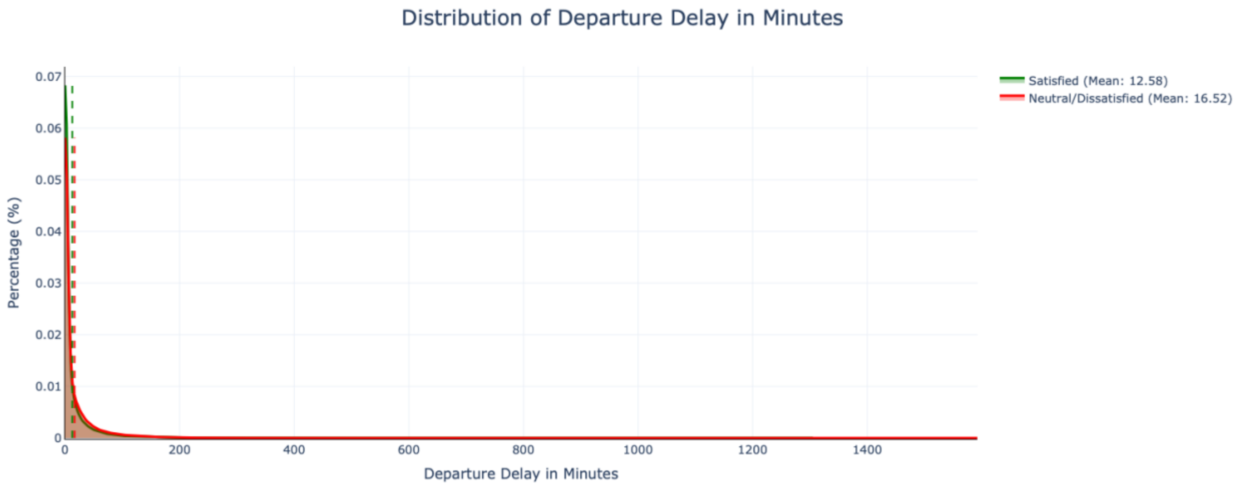


Figure 6. KDE Plot of Departure Delay

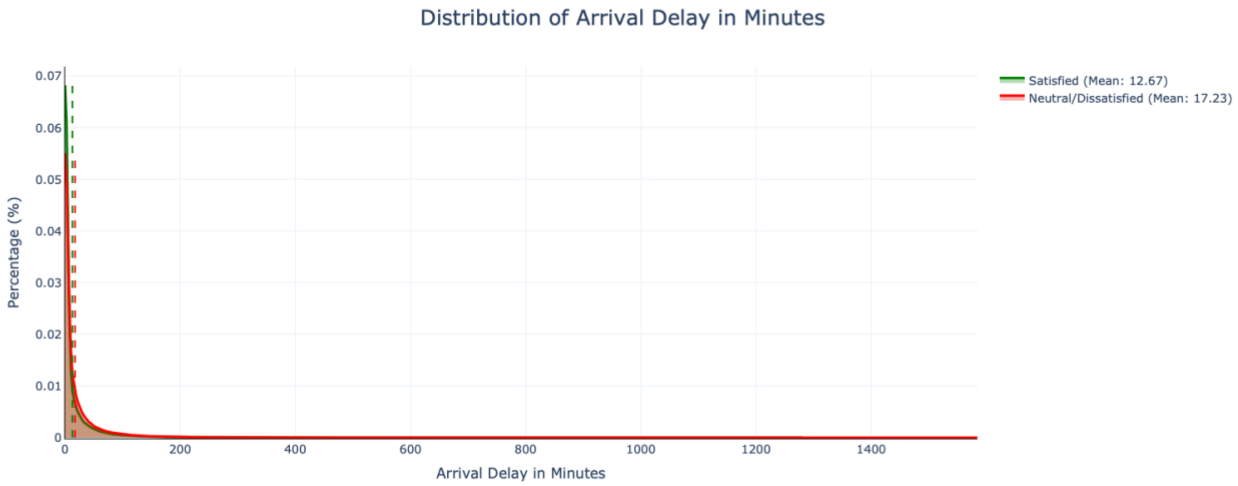


Figure 7. KDE Plot of Arrival Delay

This surprising result suggests that customers' overall satisfaction levels remain relatively unaffected by minor delays in departure or arrival times. The similar patterns observed in both the departure and arrival delay distributions further reinforce this finding, as one would typically expect a close relationship between departure delays and arrival delays due to their interconnected nature.

In essence, these results imply that while delays are often perceived as inconveniences, they may not play a significant role in shaping passengers' overall satisfaction levels, highlighting the importance of other factors such as service quality, comfort, and travel class in influencing customer satisfaction.

Analysis on the Type of Travel and Customer Satisfaction

Among the satisfied customers, a significant majority of 78.8% belong to the Business class, followed by 17.92% from the Economy class, and 4% from the Economy Plus class. In contrast, for the neutral or dissatisfied customer group, the distribution is quite different. 63.43% of these customers belong to the Economy class, followed by 26.73% from the Business class, and 9.83% from the Economy Plus class.

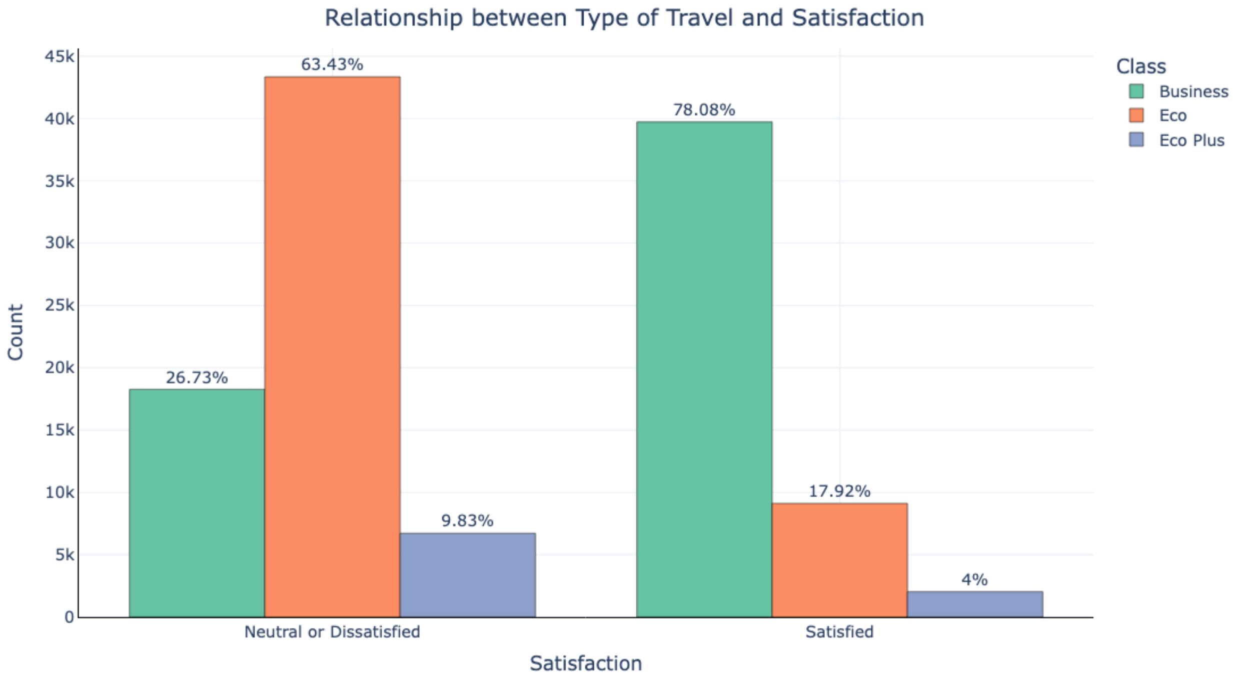


Figure 8. Type of Travel Breakdown for Satisfaction Groups

These findings support our earlier observation regarding flight distance and travel class. Specifically, we noted that customers traveling longer distances were more likely to belong to the Business class, resulting in higher satisfaction rates. The breakdown of travel classes here further strengthens this hypothesis, indicating that the Business class accounts for a substantial proportion of satisfied customers, while dissatisfied customers are primarily concentrated in the Economy class.

This suggests that the type of travel class significantly influences customer satisfaction, with Business class passengers more likely to report higher satisfaction due to longer flight distances and possibly more premium services associated with these flights.

Analysis on Customer Type and Customer Satisfaction

Among Disloyal customers, a significant majority of 81.78% are dissatisfied, while only 18.22% of these customers report being satisfied. In contrast, for Loyal customers, the distribution is more balanced, with 52.63% of them being dissatisfied and 47.37% reporting satisfaction.

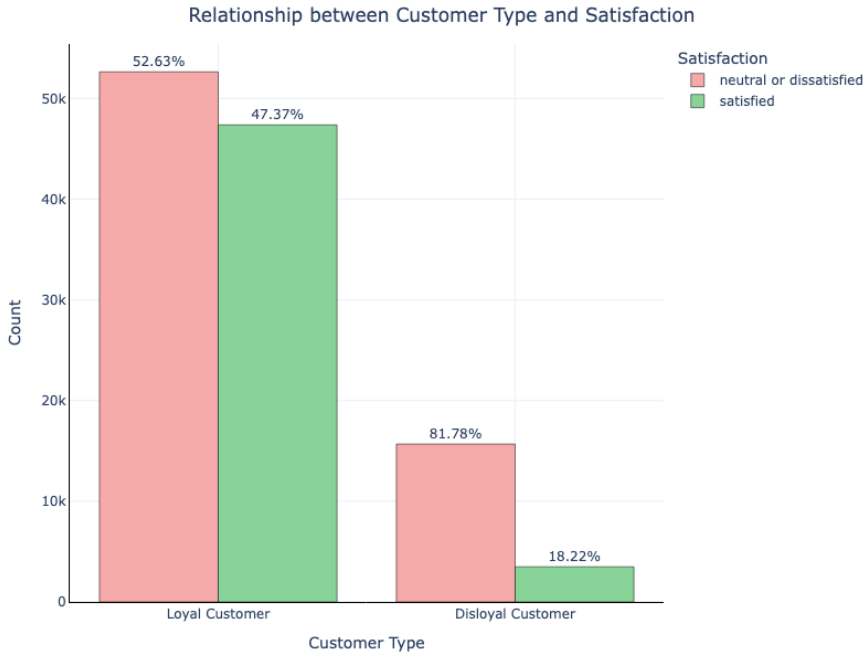


Figure 9. Satisfaction Breakdown Among Customer Type

These findings provide further insights into the relationship between customer loyalty and satisfaction. Disloyal customers tend to be more dissatisfied, indicating that a lack of brand loyalty might correlate with lower satisfaction levels. On the other hand, Loyal customers show a relatively higher level of satisfaction, suggesting that a deeper connection to the brand may contribute to more favorable experiences.

However, even though the percentage of dissatisfied Loyal customers is relatively lower compared to Disloyal customers, 52.63% is still a significant proportion. This suggests that the airline should take this into consideration and focus on reducing the dissatisfied percentage of their Loyal customers. Identifying key indicators that impact customer satisfaction for this group could lead to targeted improvements, ultimately enhancing customer loyalty and satisfaction.

We will further explore this question in the upcoming "Modeling" section, where we aim to identify the key factors influencing customer satisfaction and recommend strategies to address these gaps.

Modeling

Before choosing the machine learning (ML) models to use, it was important to keep the project's scope and objectives in mind. For this project, we focused on two well-known methods for classification: *Logistic Regression*, a simple binary classification model, and *LightGBM*, a gradient-boosted-decision-tree-based classifier.

We selected these two methods for several reasons:

- **Speed**: Both models are fast to train and make predictions, which makes them ideal for quick experimentation.
- **Interpretability**: Logistic Regression is straightforward and easy to interpret, while LightGBM balances performance with a reasonable level of interpretability.
- **Performance**: These methods are known for their accuracy and reliability on a wide range of tasks.
- **Popularity**: LightGBM is widely used in the industry for its ability to handle large datasets efficiently.

In the next sections, we'll explain how each model was applied, along with the data preprocessing steps we used to prepare for training and evaluation.

Logistic Regression

When there is a linear relationship between the features and the target variable, it is often a good idea to start with a simple model. Logistic Regression is a widely used method in such cases, as it effectively captures these linear relationships.

The model works by estimating the probability of the target variable belonging to a specific class based on the input features. This makes it a reliable and interpretable option for predicting class labels, especially when working with new data.

Data Preprocessing for Logistic Regression Model

To prepare the data for Logistic Regression, we needed to ensure the features were in the appropriate format. This involved encoding categorical variables and standardizing numerical features where necessary.

1. **Encoding Categorical Features**:

    - We used *Label Encoding* to transform categorical variables. This approach was chosen over one-hot encoding because many of the categorical features in our dataset had an inherent order. Using one-hot encoding in such cases might cause the model to miss the ordinal relationships and treat all categories equally.
    - The columns we label-encoded include: Gender, Customer Type, Type of Travel, and Class.

- Although Gender is not an ordinal feature, we applied label encoding because it only has two categories. This resulted in values of 0 and 1, effectively representing the presence or absence of a specific gender.
- Features like Seat Comfort, Online Boarding, Food and Drink, and other flight rating columns were left unchanged, as their values (ratings from 1 to 5) were already numeric and required no further transformation.

2. **Standardizing Numerical Features**:

- We standardized the following numeric columns: Age, Flight Distance, Departure Delay in Minutes and Arrival Delay in Minutes.
- For standardization, we used *Scikit-learn's StandardScaler*, which scales these features to have a mean of 0 and a standard deviation of 1. This step ensures all numeric columns are on a similar scale, helping the Logistic Regression model converge faster during training.

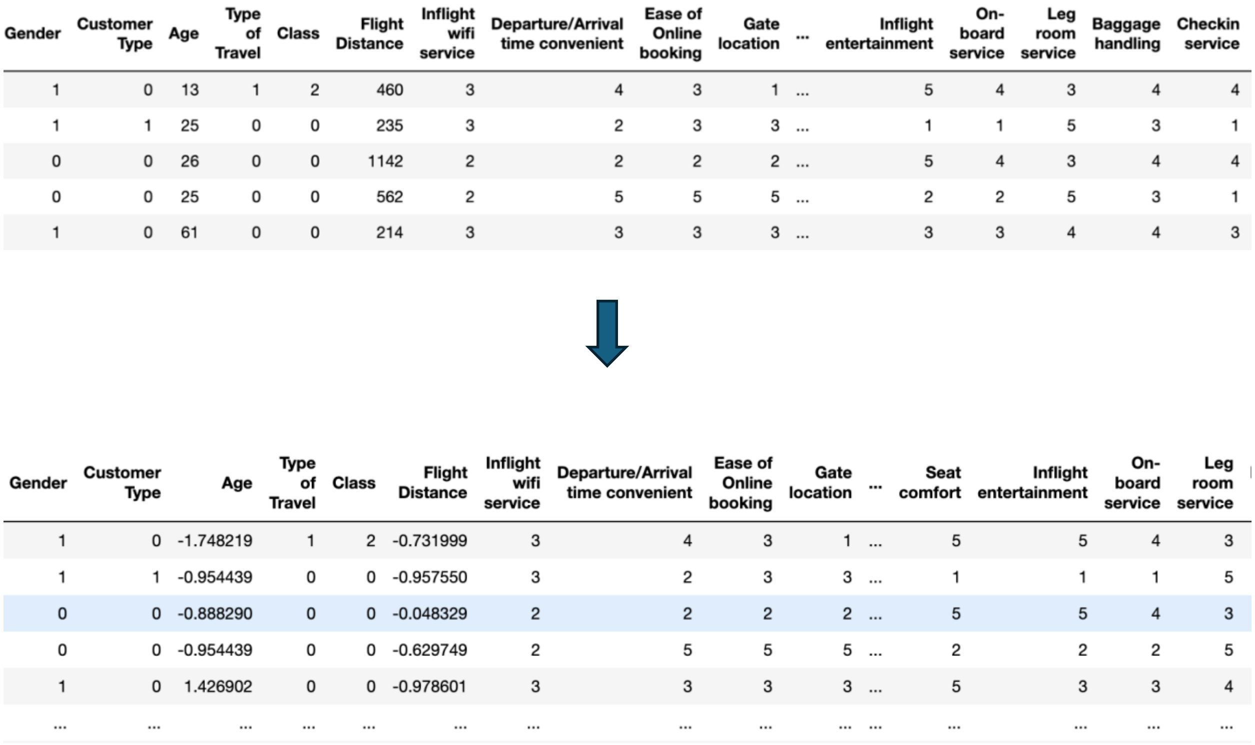
3. **Preparing the Target Variable**:

- Our target variable, Satisfaction, was initially a string with two categories: "Satisfied" and "Neutral or Dissatisfied."
- We converted this column into a binary numeric format, where 1 represents "Satisfied" and 0 represents "Neutral or Dissatisfied."

4. **Train-Test Split**:

- Once the preprocessing was complete, we split the dataset into training and testing sets using *Scikit-learn's train_test_split* function.
- We reserved 30% of the data for the test set to evaluate model performance on unseen data, while the remaining 70% was used for training.
- Although we experimented with a 20-80 split, the 30-70 split produced the most optimal model in terms of test set accuracy.

With these preprocessing steps complete, the data was ready for model training and evaluation. Refer to *figure 10* below to get a sense of how pre-processing changed the data.



Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	Gate location	...	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service
1	0	13	1	2	460	3	4	3	1	...	5	4	3	4	4
1	1	25	0	0	235	3	2	3	3	...	1	1	5	3	1
0	0	26	0	0	1142	2	2	2	2	...	5	4	3	4	4
0	0	25	0	0	562	2	5	5	5	...	2	2	5	3	1
1	0	61	0	0	214	3	3	3	3	...	3	3	4	4	3

Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	Gate location	...	Seat comfort	Inflight entertainment	On-board service	Leg room service
1	0	-1.748219	1	2	-0.731999	3	4	3	1	...	5	5	4	3
1	1	-0.954439	0	0	-0.957550	3	2	3	3	...	1	1	1	5
0	0	-0.888290	0	0	-0.048329	2	2	2	2	...	5	5	4	3
0	0	-0.954439	0	0	-0.629749	2	5	5	5	...	2	2	2	5
1	0	1.426902	0	0	-0.978601	3	3	3	3	...	5	3	3	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Figure 10. Data Preprocessing Result

Logistic Regression Model Evaluation

To evaluate the Logistic Regression model, we adopted a dual approach using two libraries: *Scikit-learn* and *Statsmodels*.

1. **Model Fitting**:

   - We first fitted a Logistic Regression model using *Scikit-learn*. This library is highly efficient for training models and making predictions, making it a great choice for practical purposes.
   - Simultaneously, we fitted the model using *Statsmodels*. The advantage of Statsmodels lies in its detailed statistical output, which makes it easier to interpret the model.

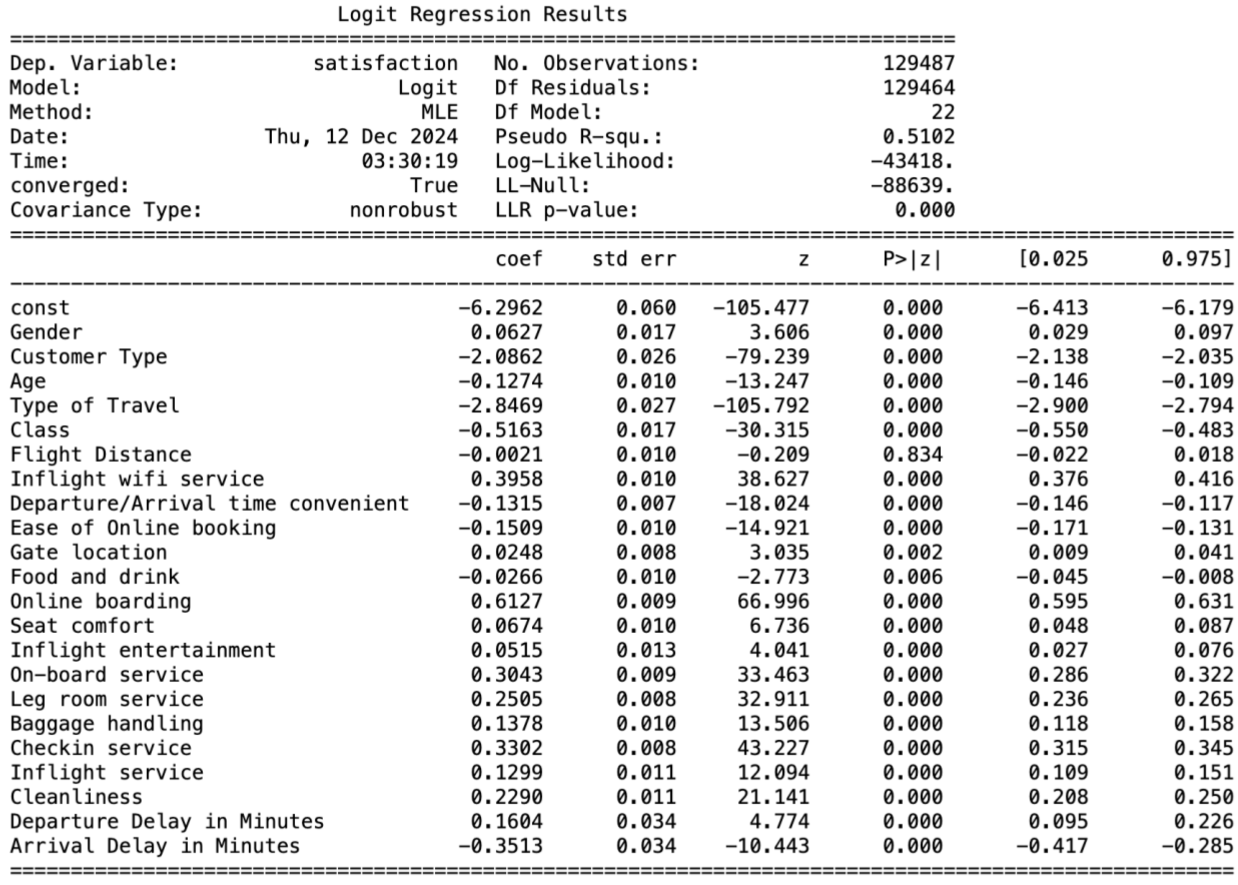
2. **Why Use Statsmodels?**:

   - Statsmodels provides a comprehensive summary of key model metrics, including: *Learned Coefficient Values, P-values,* and *McFadden's Pseudo R-squared Score*
   - These metrics are essential for understanding the relationships between features and the target variable, as well as for assessing model quality.
   - Refer to the figure below for a detailed summary generated from Statsmodels.

3. **Performance Evaluation**:

   - To measure how well the model performed, we employed multiple classification evaluation metrics, each addressing a different aspect of accuracy and model performance.

Logistic Regression Summary



```
                           Logit Regression Results                           
==============================================================================
Dep. Variable:           satisfaction   No. Observations:               129487
Model:                          Logit   Df Residuals:                   129464
Method:                           MLE   Df Model:                           22
Date:                Thu, 12 Dec 2024   Pseudo R-squ.:                  0.5102
Time:                        03:30:19   Log-Likelihood:                -43418.
converged:                       True   LL-Null:                       -88639.
Covariance Type:            nonrobust   LLR p-value:                     0.000
=====================================================================================================
                                        coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------
const                                -6.2962      0.060   -105.477      0.000      -6.413      -6.179
Gender                                0.0627      0.017      3.606      0.000       0.029       0.097
Customer Type                        -2.0862      0.026    -79.239      0.000      -2.138      -2.035
Age                                  -0.1274      0.010    -13.247      0.000      -0.146      -0.109
Type of Travel                       -2.8469      0.027   -105.792      0.000      -2.900      -2.794
Class                                -0.5163      0.017    -30.315      0.000      -0.550      -0.483
Flight Distance                      -0.0021      0.010     -0.209      0.834      -0.022       0.018
Inflight wifi service                 0.3958      0.010     38.627      0.000       0.376       0.416
Departure/Arrival time convenient    -0.1315      0.007    -18.024      0.000      -0.146      -0.117
Ease of Online booking               -0.1509      0.010    -14.921      0.000      -0.171      -0.131
Gate location                         0.0248      0.008      3.035      0.002       0.009       0.041
Food and drink                       -0.0266      0.010     -2.773      0.006      -0.045      -0.008
Online boarding                       0.6127      0.009     66.996      0.000       0.595       0.631
Seat comfort                          0.0674      0.010      6.736      0.000       0.048       0.087
Inflight entertainment                0.0515      0.013      4.041      0.000       0.027       0.076
On-board service                      0.3043      0.009     33.463      0.000       0.286       0.322
Leg room service                      0.2505      0.008     32.911      0.000       0.236       0.265
Baggage handling                      0.1378      0.010     13.506      0.000       0.118       0.158
Checkin service                       0.3302      0.008     43.227      0.000       0.315       0.345
Inflight service                      0.1299      0.011     12.094      0.000       0.109       0.151
Cleanliness                           0.2290      0.011     21.141      0.000       0.208       0.250
Departure Delay in Minutes            0.1604      0.034      4.774      0.000       0.095       0.226
Arrival Delay in Minutes             -0.3513      0.034    -10.443      0.000      -0.417      -0.285
=====================================================================================================
```

Figure 11. Logistic Regression Summary Result

The model summary provides several valuable insights into the relationship between our features and customer satisfaction. Refer to the table in *figure 11* for model summary.

Firstly, most variables in the model appear to be statistically significant, with p-values below the commonly used 0.05 threshold. However, there is one notable exception: Flight Distance, which does not show statistical significance. This suggests that, based on our data, the distance traveled does not meaningfully influence whether a customer was satisfied with their flight.

When looking at the coefficients, we can interpret their impact on the *log(odds)* of customer satisfaction. Features like Inflight WiFi Service**,** Online Boarding**,** Check-in Service**,** and Onboard Service stand out as the most influential in driving customer satisfaction. These variables positively affect the probability of satisfaction, meaning that higher ratings for these services directly correspond to higher overall satisfaction. For example, it makes intuitive sense that customers who rate Online Boarding highly are more likely to have a positive overall experience.

On the other hand, some features have a negative impact on satisfaction. Variables like Customer Type**,** Age**,** Type of Travel**,** Class**,** Ease of Online Booking**,** Food and Drink, and Arrival Delay in Minutes decrease the likelihood of a customer being satisfied. A closer look reveals some interesting patterns: as Age increases, the probability of satisfaction tends to decline. Similarly, the negative coefficient for Class reflects the way we encoded this variable, where Business Class was labeled as 0, Economy Plus as 1, and Economy as 2. The model shows that customers in Economy Class (with a higher label) are less likely to be satisfied compared to those in Business Class.

Overall, many of these results align with logical expectations. Customers who rate essential services such as WiFi and boarding processes positively are naturally more likely to express satisfaction, while factors like travel delays or lower service classes detract from the overall experience.

Evaluation Metrics

To assess the performance of the logistic regression model, we used multiple evaluation metrics, each addressing a specific aspect of accuracy.

The model's Recall and Precision are both around 0.83, indicating that it performs well in predicting both classes accurately. The F1 score, which is the average of Precision and Recall, also aligns at 0.83 as seen in *figure 12*. The confusion matrix, shown in *figure 13* reveals that the model is better at predicting satisfied customers.

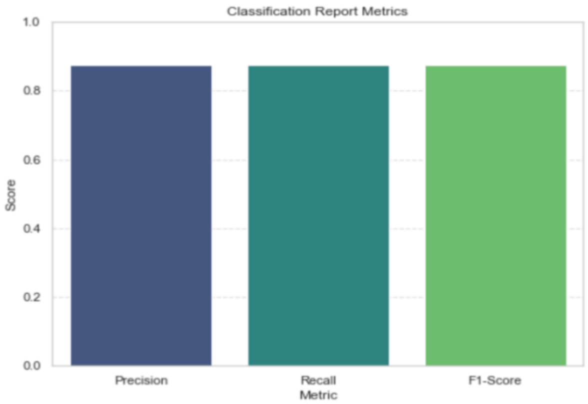


Figure 12. Classification Report Metrics

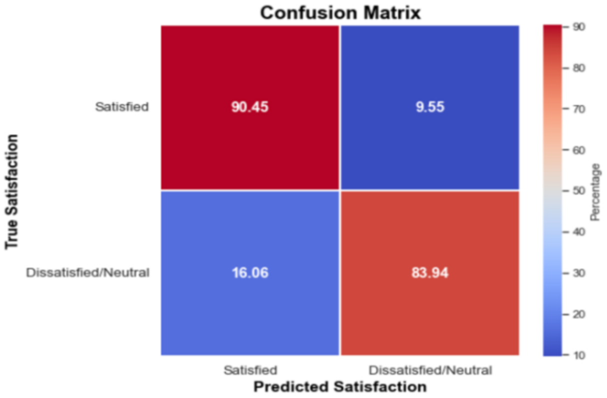


Figure 13. Confusion Matrix

The ROC curve (Receiver Operating Characteristic curve) provides a graphical representation of a model's ability to distinguish between positive and negative classes across different thresholds. From the curve, we can observe that the Area Under the Curve (AUC) is approximately 0.93.

An AUC of 0.93 suggests that the model has a strong ability to discriminate between the classes. In general, an AUC value close to 1 indicates excellent model performance, while a value closer to 0.5 suggests the model is performing no better than random guessing. With an AUC of 0.93, our model is performing well above random chance, demonstrating its effectiveness in distinguishing between satisfied and unsatisfied customers.

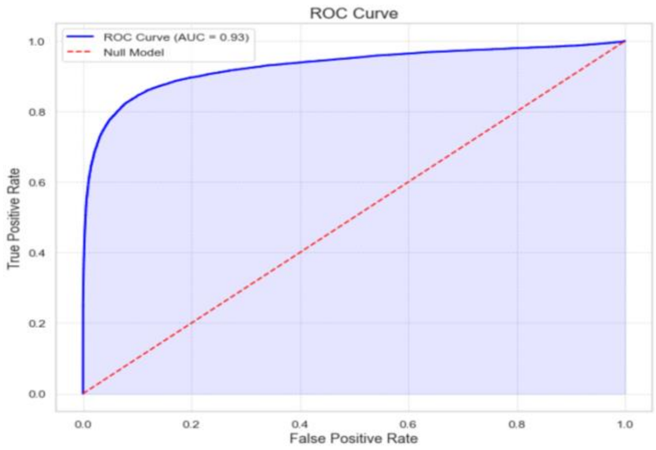
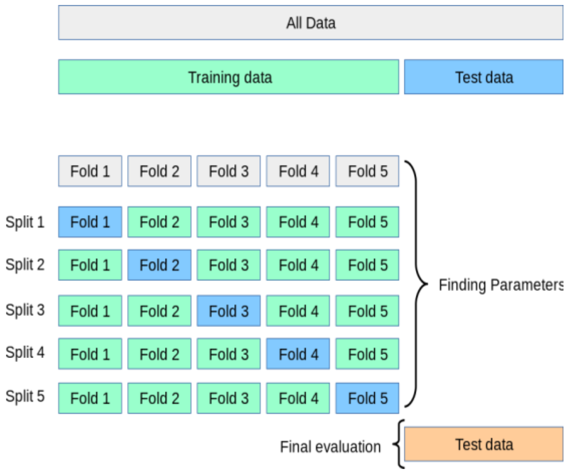


Figure 14. ROC Curve

Cross Validation



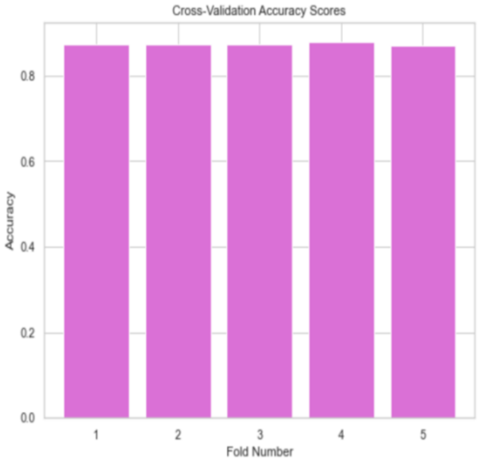


Figure 15. Cross Validation

Figure 16. Cross Validation Accuracy Score

We used 5-Fold Cross Validation to further assess the model's performance and ensure that it wasn't overfitting to the training data. This technique splits the dataset into five equal parts, training the model on four parts and testing it on the remaining part, rotating through all folds. By evaluating the model across multiple subsets of the data, we can assess its generalizability. A visual representation is illustrated in *figure 15*.

The consistent accuracy across all the folds indicates that the model is not overfitting. Overfitting occurs when a model performs well on the training data but poorly on unseen data. Since the accuracy remains stable across the different folds, we can be confident that the model is generalizing well and not memorizing the training data. *Figure 16* is a visual illustration of this generalization.

Optimal Cutoff for Classification

The default cutoff value of 0.5 for predicted probabilities appears to work well in this case. This is because, as shown in *figure 17*, the predicted probabilities clearly separate the classes into two distinct groups: positive and negative. When a predicted probability is greater than 0.5, the model classifies it as positive (e.g., satisfied), and when it is below 0.5, it is classified as negative (e.g., neutral or dissatisfied).

The chart demonstrates that this cutoff effectively divides the predicted probabilities in a way that maximizes the distinction between the two classes, leading to a clear and optimal categorization. Therefore, a threshold of 0.5 is appropriate for our model in this context.

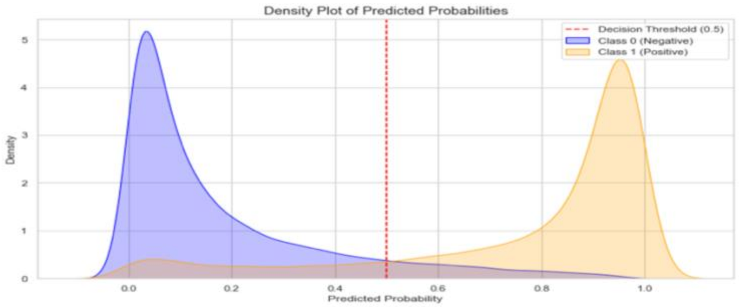


Figure 17. Density Plot of Predictive Probabilities

LightGBM

Model Performance and Insights

LightGBM (Light Gradient Boosting Machine) is an efficient and scalable gradient boosting framework designed for large datasets. It builds weak learners in the form of decision trees using a leaf-wise approach, where the algorithm grows trees by splitting the leaf that results in the highest reduction in loss, rather than level-wise as in traditional tree-building methods. This results in faster convergence and higher accuracy. After training multiple weak learners, LightGBM combines them to create a strong model through an additive process, where each new tree corrects the errors made by the previous ones. This ensemble approach allows LightGBM to effectively handle complex tasks while remaining computationally efficient, making it ideal for large-scale classification and regression problems. *Figure 18* illustrates how this looks.

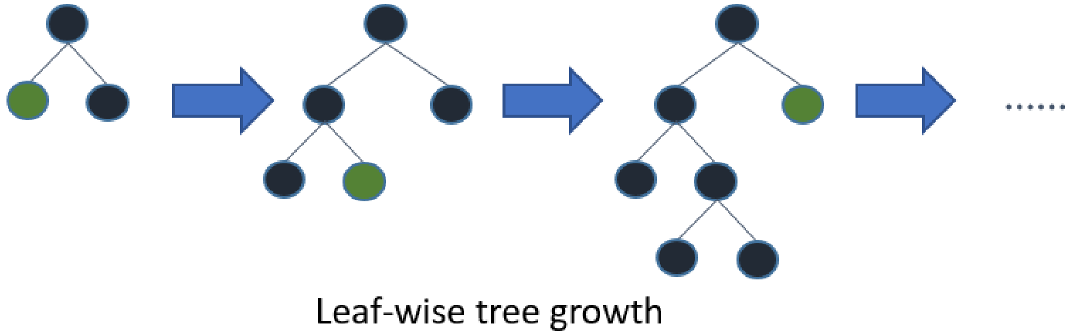


Figure 18. LightGBM

The LightGBM model was trained using the same preprocessed data from the logistic regression model, and it showed significant improvement in performance. The accuracy of the LightGBM model increased to 0.96, up from the previous 0.88 achieved by the logistic regression model, indicating that the LightGBM model is more effective at making correct predictions.

When examining feature importance, both models highlighted Inflight WiFi Services and Online Boarding as key factors in predicting customer satisfaction. This suggests that these features play a significant role in determining the likelihood of a customer being satisfied with their flight, regardless of the model used. *Figure 19* is showing the plot for feature importance.

Additionally, the LightGBM model demonstrated a stronger ability to predict the negative class (i.e., neutral or dissatisfied customers). This is an important improvement, as accurately predicting the negative class is crucial for understanding areas that need improvement in customer satisfaction. This is demonstrated in *figure 20* with the help of a confusion matrix.

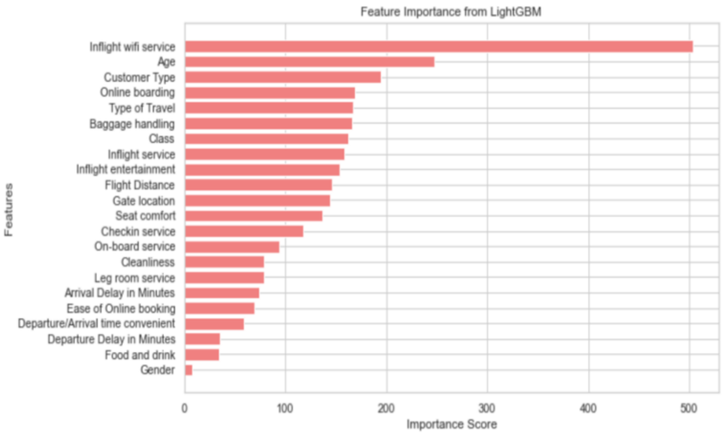


Figure 19. Feature Importance

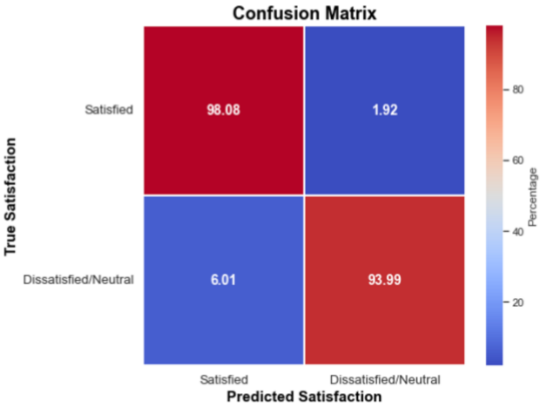


Figure 20. Confusion Matrix

Model Comparison

	Precision	Recall	F1 Score	ROC-AUC	Accuracy
Logistic Regression	**0.83**	**0.83**	**0.83**	**0.93**	**0.88**
LightGBM	**0.96**	**0.96**	**0.96**	**0.96**	**0.96**

In the table in *figure 3.12* we compare different evaluation metrics of both the models. LightGBM outperforms Logistic Regression across all key metrics. It achieves higher precision (0.96 vs. 0.83), indicating it is more effective at correctly identifying positive class instances while minimizing false positives. Similarly, LightGBM excels in recall (0.96 vs. 0.83), demonstrating its better ability to capture all actual positive instances and reduce false negatives. This leads to a higher F1 score for LightGBM (0.96 vs. 0.83), reflecting a more balanced performance in precision and recall. LightGBM also has a superior ROC-AUC score (0.96 vs. 0.93), signifying its stronger ability to distinguish between positive and negative classes. Finally, LightGBM achieves a higher accuracy (0.96 vs. 0.88), showcasing its overall effectiveness in predicting both classes. In summary, LightGBM is the more robust model, outperforming Logistic Regression in terms of precision, recall, F1 score, ROC-AUC, and accuracy.

Conclusion

This project demonstrated the power of data-driven analysis in understanding and predicting passenger satisfaction. By identifying key drivers like inflight WiFi, online boarding, and travel class, airlines can prioritize enhancements in these areas to improve customer experience. While tree-based models achieved higher accuracy, logistic regression provided more interpretability, allowing for flexibility based on business needs. The surprising finding that flight delays had minimal impact on satisfaction challenges conventional assumptions, showcasing the importance of evidence-based decision-making. Overall, this work highlights actionable insights for the airline industry to foster customer loyalty and optimize service delivery.

Key Findings

1. **Balanced Dataset:** The dataset was balanced in terms of the target variable, ensuring fair model training and evaluation across the satisfaction classes.
2. **Influential Factors:** The most important features identified by both logistic regression and tree-based models were Inflight WiFi Service, Online Boarding, and Travel Class. These areas significantly impact passenger satisfaction and provide actionable insights for airlines to focus on.
3. **Minimal Impact of Delays:** Contrary to initial expectations, flight delays (both departure and arrival) had minimal influence on satisfaction levels, as observed during the EDA and confirmed by the models.

4. **Model Performance:** Logistic regression, while more interpretable, provided lower accuracy compared to tree-based models like LightGBM, which delivered higher predictive performance but at the cost of reduced explainability.
5. **Actionable Insights:** Airlines can improve customer satisfaction by enhancing inflight WiFi, streamlining the online boarding process, and providing better travel class experiences.

Limitations

1. **Subjective Ratings:** Passenger-provided ratings for satisfaction-related features may introduce subjectivity or bias, impacting the reliability of conclusions.
2. **Dataset Scope:** The dataset might represent specific demographics or regions, potentially limiting its generalizability to other populations or airline industries.
3. **Static Data:** The dataset does not account for temporal changes, such as post-pandemic travel behavior, which could influence satisfaction levels differently.
4. **Feature Limitations:** Key factors like ticket pricing, weather conditions, and airline-specific policies were not included, limiting the scope of the analysis.
5. **Model Interpretability:** Tree-based models, while accurate, lack interpretability compared to simpler models like logistic regression, which is a challenge for communicating findings to non-technical stakeholders.

Future Research Recommendations

1. **Diverse Datasets:** Include more geographically and demographically diverse datasets to improve the applicability of the findings across regions and airlines.

2. **Temporal Analysis:** Study satisfaction trends over time to assess the impact of evolving airline services or policies on passenger preferences.
3. **Additional Features:** Incorporate variables such as ticket pricing, flight duration, weather conditions, and customer loyalty programs for a deeper understanding of satisfaction.

References

1. Airline passenger satisfaction. Retrieved from https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction
2. Namukasa, J. (2013). The influence of airline service quality on passenger satisfaction and loyalty: The case of Uganda airline industry. *The TQM journal*, *25*(5), 520-532.
3. Noviantoro, T., & Huang, J. P. (2022). Investigating airline passenger satisfaction: Data mining method. *Research in Transportation Business & Management*, *43*, 100726.
4. Shiwakoti, N., Hu, Q., Pang, M. K., Cheung, T. M., Xu, Z., & Jiang, H. (2022). Passengers' perceptions and satisfaction with digital technology adopted by airlines during COVID-19 pandemic. *Future Transportation*, *2*(4), 988-1009.