ITE5424 Big Data Project

Project Topic: Airline Customer Satisfaction

Malika Gupta - N01581424

Problem Statement

Customer satisfaction is essential for the airline industry to maintain a competitive edge and foster customer loyalty. The vast amount of data generated by customer interactions poses significant challenges for processing, analysis, and deriving meaningful insights using traditional methods. This project leverages R technology to manage large volumes of customer feedback and operational data, offering comprehensive analysis and predictions of customer satisfaction levels. By examining the data, we can pinpoint critical areas for service improvement to enhance customer satisfaction and loyalty.

Description of the Dataset

The dataset is sourced from Kaggle. The dataset is provided as a CSV file named Airline_customer_satisfaction.csv. It contains 22 columns and 129,880 rows which collected from airline passengers, focusing on various factors that contribute to their overall satisfaction.

Content of the Dataset:

1. Personal Information:
   - Gender
   - Age
2. Travel Information:
   - Type of Travel: Personal or business purposes.
   - Class: Travel class (Economy, Business, etc.).
   - Flight Distance
3. Service Ratings:
   - Inflight Wifi Service
   - Departure/Arrival Time Convenience
   - 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
   - Cleanliness
4. Operational Data:
   - Departure Delay in Minutes
   - Arrival Delay in Minutes

5. Target Variable:
   - Satisfaction: Overall satisfaction level of the passenger (Satisfied, Neutral, or Dissatisfied).

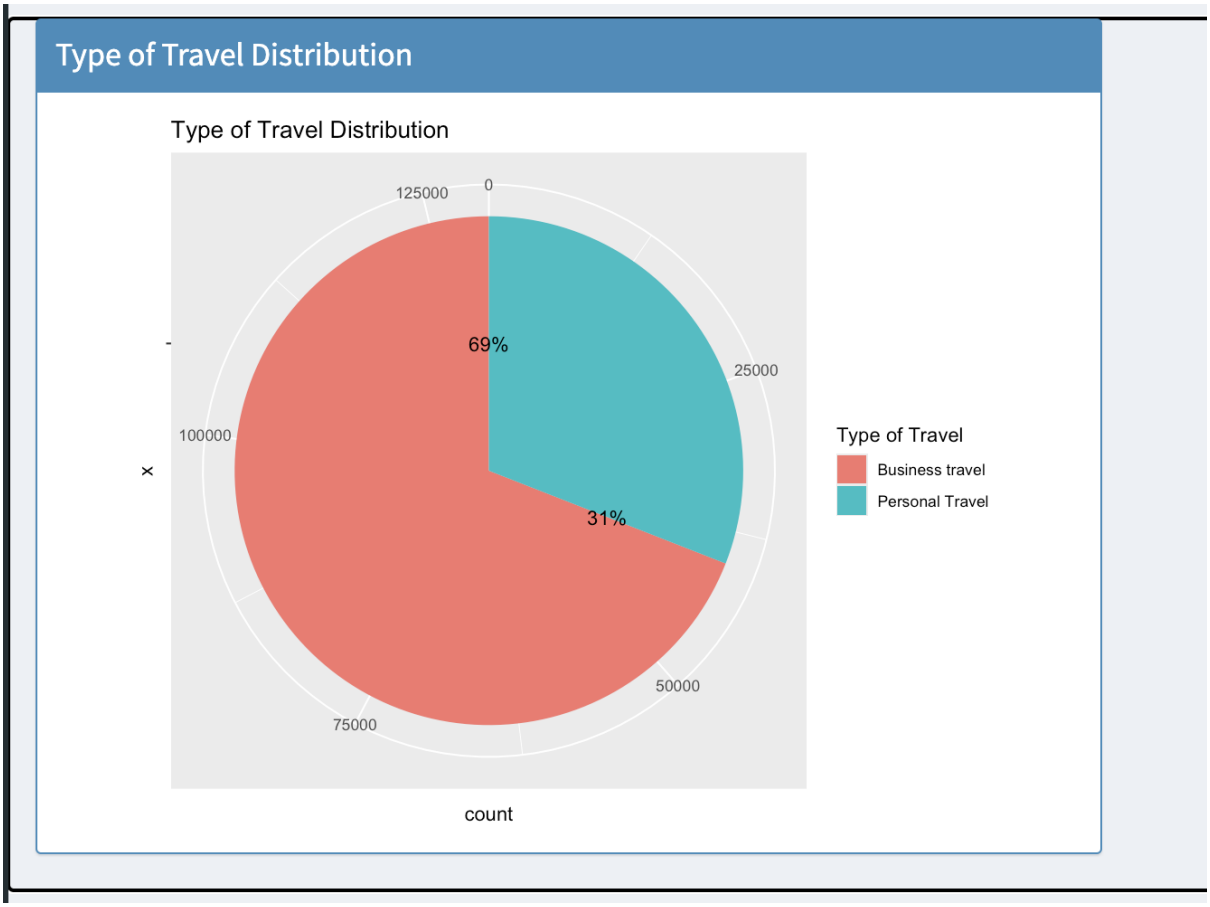
Exploratory and Satisfaction Analysis

Dataset Summary

Exploratory Analysis

```
$Total_Customers
[1] 129880

$Average_Age
[1] 39.4

$Max_Flight_Distance
[1] 6951

$Min_Flight_Distance
[1] 50

$Most_Common_Customer_Type
[1] "Loyal Customer"

$Average_Departure_Delay
[1] 14.71371

$Average_Arrival_Delay
[1] 15.09113
```

Type of Travel Distribution



Distribution of Travel Types

1. Business Travel (Red): 69%
2. Personal Travel (Turquoise): 31%

Interpretation

1. Dominance of Business Travel: The majority of the travel (69%) is for business purposes. This indicates that business travel is more prevalent within the observed population.
2. Significant Personal Travel Segment: Although personal travel comprises a smaller portion (31%) of the total, it still represents a significant segment of the travel population.

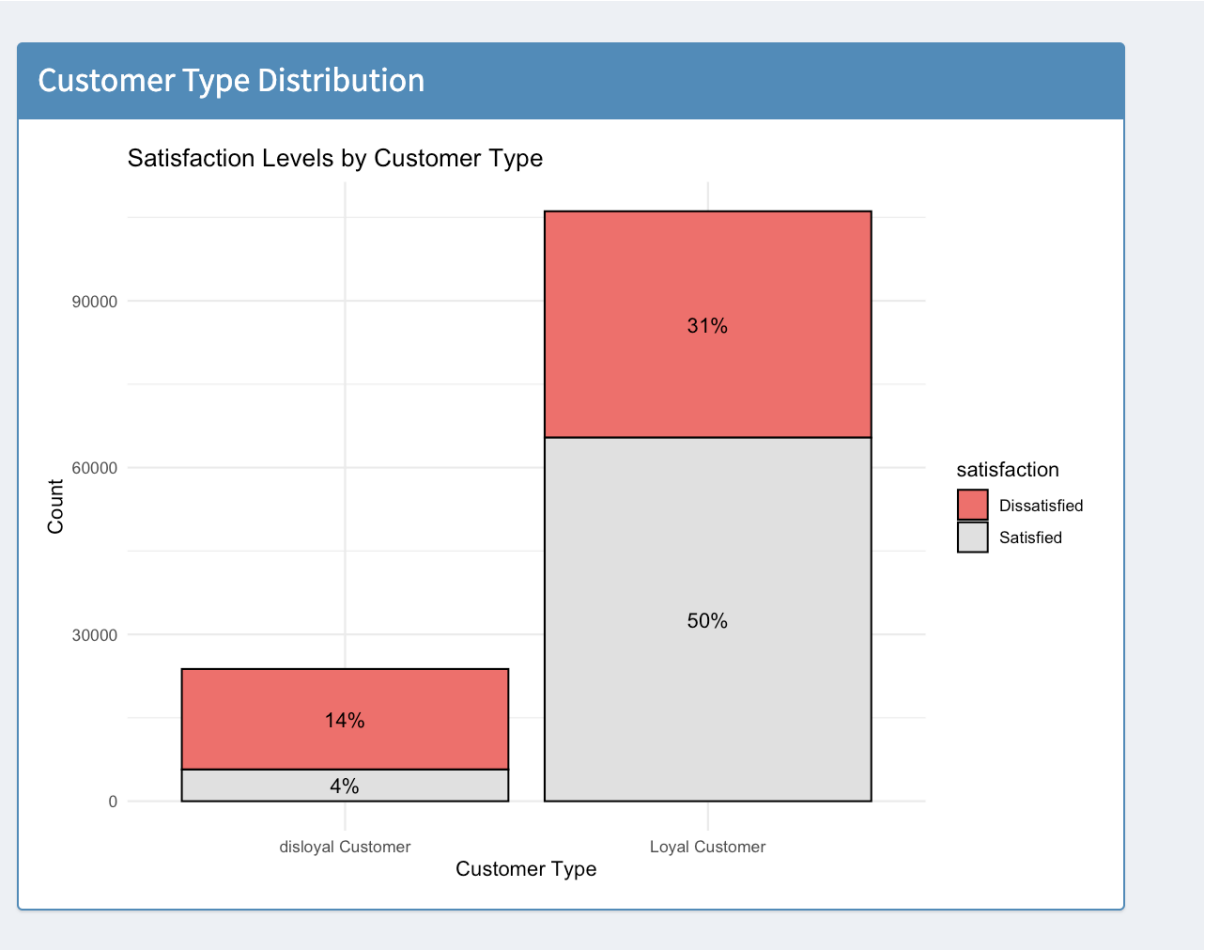
Insights

1. Service Prioritization for Business Travelers: Given that a larger percentage of travellers are traveling for business, it would be beneficial for airlines and travel-related services to prioritize features and amenities that cater to business travellers. This could include options for more flexible travel schedules, enhanced connectivity, and comfortable workspaces.
2. Opportunities in Personal Travel: The 31% of travellers who are traveling for personal reasons still represent a substantial market. Services tailored to personal travellers,

such as family-friendly amenities, leisure packages, and vacation deals, can enhance customer satisfaction and loyalty in this segment.

3. Marketing Strategies: Marketing efforts can be tailored based on these proportions. For instance, campaigns targeting business travellers could focus on efficiency and productivity, while those aimed at personal travellers could emphasize relaxation, family activities, and leisure.

Satisfaction Levels by Customer Type



Distribution

1. Disloyal Customers
    - Dissatisfied (Red): Represents 14% of the total customers.
    - Satisfied (Gray): Represents 4% of the total customers.
2. Loyal Customers:
    - Dissatisfied (Red): Represents 31% of the total customers.
    - Satisfied (Gray): Represents 50% of the total customers.

Interpretation

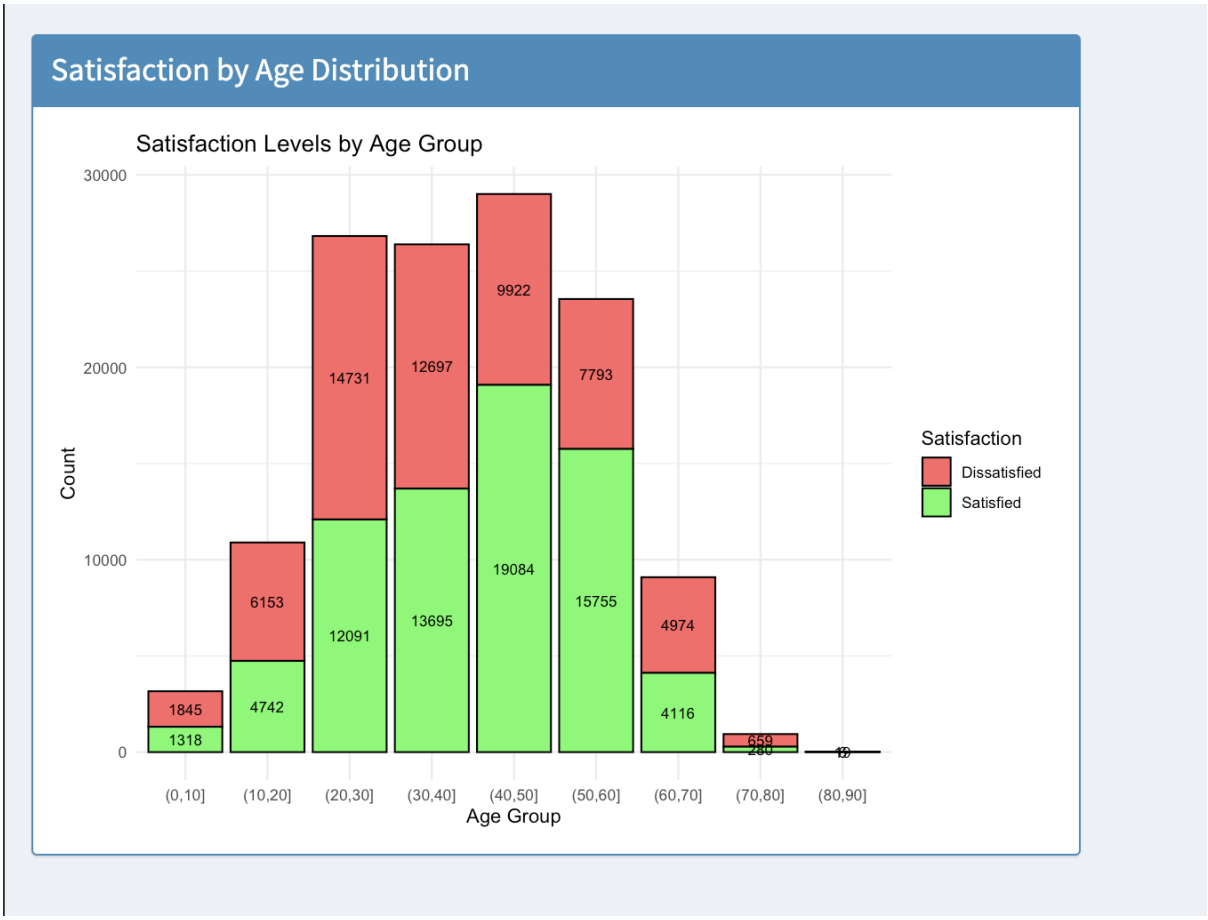
1. Loyal Customers: A significantly higher proportion of loyal customers are satisfied (50%) compared to disloyal customers (4%). However, there is still a notable portion of loyal customers who are dissatisfied (31%).
2. Disloyal Customers: The majority of disloyal customers are dissatisfied (14%), with only a small fraction being satisfied (4%).

Insights

1. Customer Loyalty and Satisfaction: There is a strong correlation between customer loyalty and satisfaction. Loyal customers have a much higher satisfaction rate compared to disloyal customers.
2. Dissatisfaction Rate: The dissatisfaction rate among loyal customers (31%) is lower compared to the dissatisfaction rate among disloyal customers (14%).
3. Target Areas for Improvement: Efforts to improve customer satisfaction should focus on reducing the dissatisfaction rate, especially among loyal customers, as they form a significant portion of the customer base.

In conclusion, the bar chart suggests that building customer loyalty is crucial for achieving higher satisfaction levels. Companies should focus on strategies to convert disloyal customers into loyal ones and address the factors contributing to dissatisfaction among loyal customers to further enhance overall customer satisfaction.

Age Distribution Histogram Analysis



	Ages 0 - 10	Ages 10 -20	Ages 20 - 30	Ages 30 - 40	Ages 40 - 50	Ages 50 - 60	Ages 60 - 70	Ages 70 -80	Ages 80 - 90
Satisfied	1,318	4,742	12,091	13,695	19,084	15,755	4,116	259	0
Dissatisfied	1,845	6,153	12,697	14,731	9,922	7,7933	4,974	659	12

Insights

1. Middle-Aged Satisfaction: The middle-aged groups (40 to 60 years) exhibit the highest levels of satisfaction, suggesting that services are well-tailored to their needs.
2. Room for Improvement: Younger travellers (under 40) and older travellers (over 60) show higher dissatisfaction rates, indicating potential areas for service improvement.
3. Age-Specific Strategies: To improve overall customer satisfaction, targeted strategies should be developed for the under 40 and over 60 age groups.

The chart reveals significant variability in satisfaction levels across different age groups. Middle-aged travellers are generally more satisfied, while younger and older travellers have higher dissatisfaction rates. This suggests a need for age-specific approaches to enhance customer satisfaction, particularly focusing on the needs and preferences of younger and older travellers.

Flight Distance by Class Analysis

Key Observations

1. Flight Distance Distribution:
   i. Business Class:
      - Median flight distance is around 2000 units.
      - The interquartile range (IQR) is from approximately 1000 to 3000 units.
      - There are several outliers beyond 6000 units.
   ii. Eco (Economy) Class:
      - Median flight distance is also around 1500 units.
      - The IQR is slightly narrower than Business, ranging approximately from 1500 to 2500 units.
      - There are a significant number of outliers, similar to Business Class, extending beyond 6000 units.
   iii. Eco Plus Class:
      - Median flight distance is slightly lower, around 1500 units.
      - The IQR is from approximately 1500 to 2500 units.
      - There are outliers, but fewer than in the other classes, going beyond 5000 units.

Comparative Analysis

1. The median flight distances for Eco and Eco Plus classes are very similar, both around 1500 units.
2. Business class has a higher median flight distance compared to Eco and Eco Plus.
3. The range of flight distances (IQR) is largest for Business and Economy, indicating a wider spread of flight distances for these classes.
4. All classes exhibit a significant number of outliers, with Business and Economy showing more extreme values

Interpretation

1. Business Class:
    i. This classes cater to a wider range of flight distances, from shorter to longer flights, as indicated by their similar and broader IQRs.
    ii. The presence of many outliers suggests that both classes are used for both very short and very long flights.
2. Eco (Economy) and Eco Plus Classes:
    i. These classes have similar median flight distances and IQRs, indicating that they cater to similar ranges of flight distances.
    ii. Both classes are used for shorter to medium-length flights more consistently than Business class.
    iii. The presence of fewer outliers in Eco Plus compared to Eco suggests that Eco Plus is slightly less used for extreme flight distances.

In summary, Business class tends to serve longer flights and has more variability in flight distances. Eco and Eco Plus have similar distributions, primarily catering to shorter and medium-length flights, with Eco showing more extreme outliers than Eco Plus.

Average Departure Delay

Distribution

1. Dissatisfied Customers (Red): The average departure delay for dissatisfied customers is approximately 10 minutes.
2. Satisfied Customers (Green): The average departure delay for satisfied customers is approximately 7.5 minutes.

Interpretation

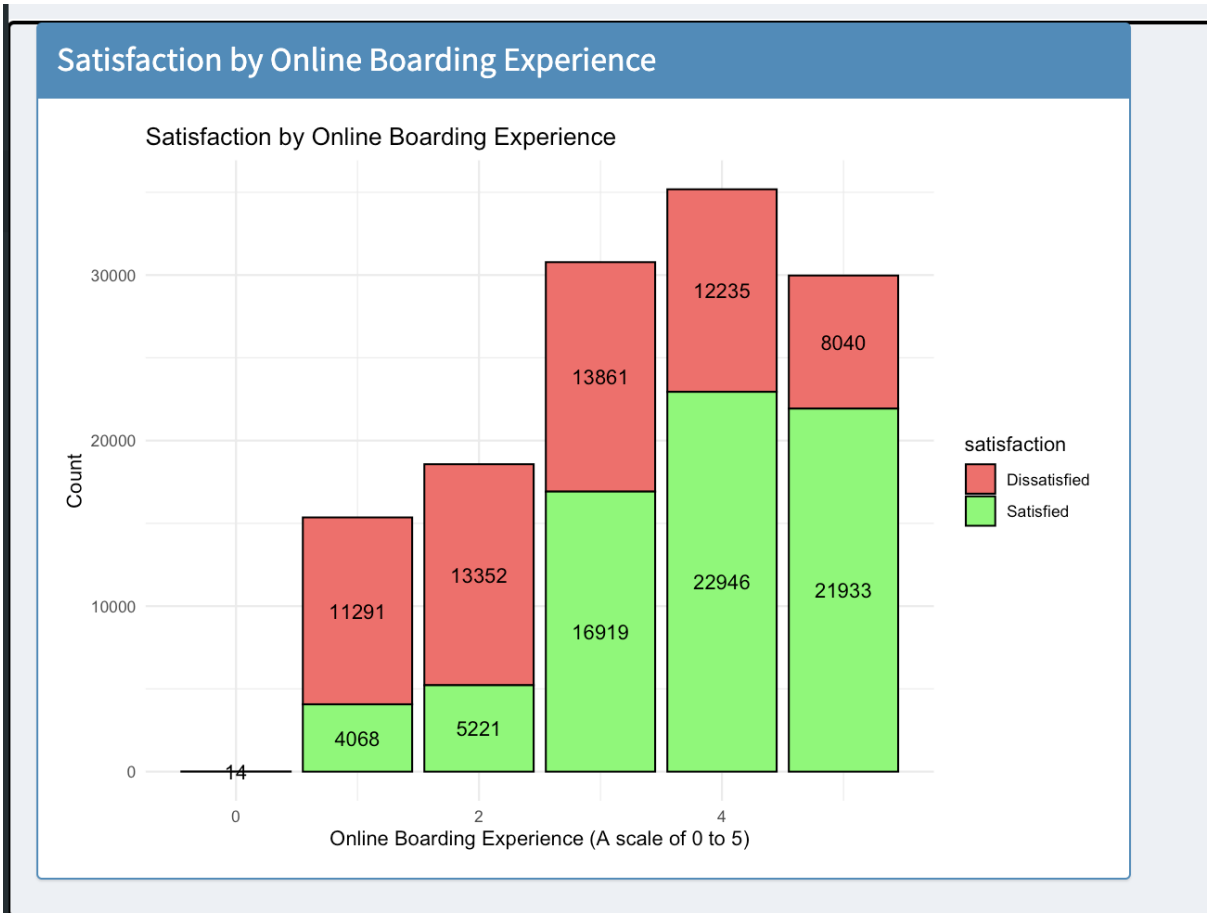
1. Impact of Departure Delay on Satisfaction: There is a clear correlation between shorter departure delays and higher customer satisfaction. On average, satisfied customers experience shorter delays compared to dissatisfied customers.
2. Magnitude of Difference: The difference in average departure delay between satisfied and dissatisfied customers is about 2.5 minutes. While this may seem minor in absolute terms, it is significant enough to impact customer satisfaction levels.

Insights

1. Delay Management: Managing and minimizing departure delays can play a crucial role in improving customer satisfaction. Efforts to reduce delays by even a few minutes can have a noticeable impact on customer perceptions and satisfaction levels.
2. Customer Experience: The data suggests that customers are sensitive to delays, and even small reductions in wait times can enhance their overall experience.
3. Operational Efficiency: Improving operational efficiency to reduce departure delays should be a priority for improving customer satisfaction. This could involve better scheduling, improved turnaround times, and more efficient boarding processes.

In conclusion, the chart highlights the importance of departure delays in determining customer satisfaction. Satisfied customers tend to experience shorter delays, indicating that reducing departure delays is a key strategy for enhancing customer satisfaction. Therefore, airlines and service providers should focus on operational improvements to minimize delays and improve the overall customer experience.

Online Boarding Experience Levels and Satisfaction



Interpretation and Analysis

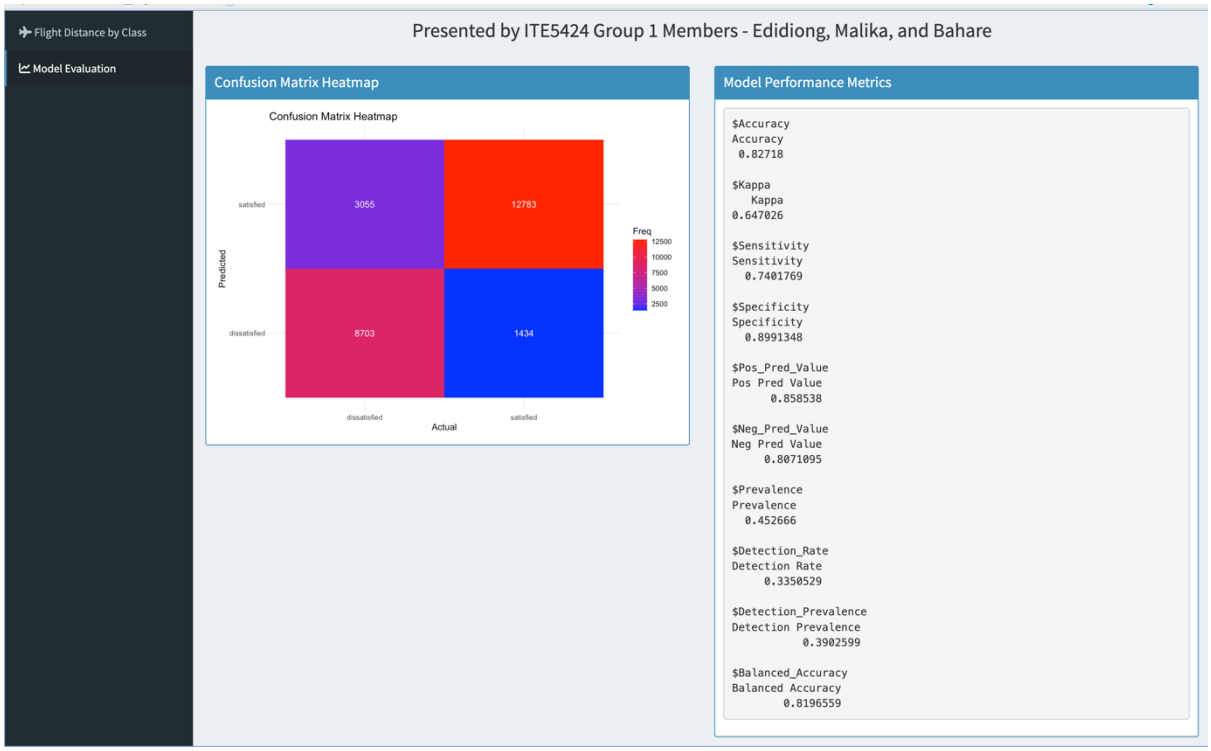
1. Rating 0/5: No customers are satisfied, and only a few are dissatisfied. This indicates a very poor online boarding experience, likely due to a lack of functionality or significant issues.
2. Rating 1/5: A significant number of customers are dissatisfied (11,291), and relatively few are satisfied (4,068). This suggests that at this rating level, customers find the online boarding experience to be inadequate.
3. Rating 2/5: The number of dissatisfied customers (13,352) still exceeds the number of satisfied customers (5,221). However, there is a slight improvement in satisfaction compared to Rating 1.
4. Rating 3/5: This rating shows a notable shift, with more customers satisfied (16,919) than dissatisfied (13,861). This indicates that customers find the online boarding experience to be acceptable at this level.
5. Rating 4/5: A high level of satisfaction is observed, with 22,946 satisfied customers compared to 12,235 dissatisfied customers. This suggests that the online boarding experience is considered very good by most customers at this rating.
6. Rating 5/5: The highest satisfaction level is observed here, with 21,933 satisfied customers and only 8,040 dissatisfied customers. This indicates an excellent online boarding experience.

Key Insights

1. Positive Correlation: There is a clear positive correlation between higher ratings of the online boarding experience and customer satisfaction. As the rating increases from 0 to 5, the proportion of satisfied customers increases significantly, while the proportion of dissatisfied customers decreases.
2. Significant Improvement Threshold: The most critical improvement appears between Ratings 2 and 3, where the number of satisfied customers surpasses the number of dissatisfied customers.
3. Focus on High Ratings: Ensuring that the online boarding experience meets the criteria for higher ratings (4 and 5) can significantly enhance overall customer satisfaction.

The chart indicates that enhancing the online boarding experience to achieve higher ratings is crucial for increasing customer satisfaction. Ratings of 3 and above show a majority of satisfied customers, with the highest satisfaction levels at Ratings 4 and 5. Therefore, airlines and service providers should focus on improving their online boarding systems to achieve higher customer ratings, thereby boosting overall satisfaction and customer loyalty.

Classification and Model Evaluation



Confusion Matrix

True Positives (TP): 8703 (dissatisfied predicted as dissatisfied)
False Positives (FP): 1434 (satisfied predicted as dissatisfied)
False Negatives (FN): 3055 (dissatisfied predicted as satisfied)
True Negatives (TN): 12783 (satisfied predicted as satisfied)

Key Metrics

1. Accuracy: This indicates that 82.72% of the predictions made by the model are correct.

2. 95% Confidence Interval (CI): (0.8225, 0.8318)
   The interval within which the true accuracy of the model is expected to fall with 95% confidence.

3. No Information Rate (NIR): This is the accuracy that would be achieved by always predicting the most frequent class. In this case, if we always predicted "satisfied", we'd be correct 54.73% of the time.

4. P-Value [Acc > NIR]: < 2.2e-16
   This p-value indicates that the accuracy of the model is significantly better than the No Information Rate, suggesting the model has predictive power.

5. Kappa: 0.647
   Kappa statistic measures the agreement between predicted and actual classes, adjusted for chance agreement. A kappa of 0.647 indicates substantial agreement beyond chance.

6. Mcnemar's Test P-Value: < 2.2e-16
   This test evaluates if there are significant differences between the number of false positives and false negatives. A very low p-value suggests a significant difference, indicating the model's predictions are not just a result of random chance.

7. Sensitivity and Specificity:
   i. The model correctly identifies 74.02% of dissatisfied customers.
   ii. The model correctly identifies 89.91% of satisfied customers.

8. Predictive Values:
   i. Of all the customers predicted to be dissatisfied, 85.85% are actually dissatisfied.
   ii. Of all the customers predicted to be satisfied, 80.71% are actually satisfied.

9. Prevalence and Detection:
   i. Prevalence: 45.27% is the proportion of actual dissatisfied customers in the dataset.

   ii. Detection Rate: The proportion of correctly identified dissatisfied customers out of the total dataset is 33.51%.

iii. Detection Prevalence: The proportion of predicted dissatisfied customers out of the total dataset is 39.03%.

10. Balanced Accuracy: 81.97% is the average of sensitivity and specificity, providing a balanced view of the model performance across both classes.

Recommendations

1. Improve Sensitivity: Since the sensitivity is lower than specificity, the airline should consider strategies to reduce false negatives (i.e., dissatisfied customers wrongly predicted as satisfied). This could involve:
   i. Balancing the dataset if it's skewed.
   ii. Exploring different thresholds for classification.
   iii. Using more features or improving feature engineering.

2. Monitor Model Performance: Regularly validate the model to ensure it maintains performance, especially if the underlying data distribution changes.

3. Customer Feedback: Utilize the model to identify and address the needs of dissatisfied customers to improve overall satisfaction.

By focusing on these areas, the model's utility in predicting customer satisfaction can be enhanced, leading to better customer service and retention strategies.