Predictive Modeling for Airline Customer Satisfaction

BY SWETHA RAMREDDY

CAPSTONE PROJECT IN BUSINESS ANALYTICS (BA-64099-024)

Table of Contents

1. Abstract

Customer satisfaction is a crucial metric in the airline industry, influencing customer loyalty, brand reputation, and overall business success. This project aims to develop a predictive model for airline customer satisfaction using various machine learning algorithms implemented in R. The dataset used contains multiple features related to customer demographics, travel details, and satisfaction ratings. The analysis includes exploratory data analysis, feature selection, and model evaluation to provide insights and recommendations for improving customer satisfaction in the airline industry. Our results show that the decision tree model achieved an accuracy of 86.4%, highlighting its potential usefulness in enhancing customer satisfaction strategies.

2. Introduction

Customer satisfaction is critical in the airline industry, impacting customer loyalty, brand reputation, and overall business performance. High satisfaction levels can lead to repeat business, positive word-of-mouth, and a competitive advantage, while low satisfaction can result in negative reviews, reduced market share, and lower revenues. Understanding the factors driving satisfaction helps airlines improve services and retain customers.

2.1 Research Problem

The airline industry is highly competitive, and customer satisfaction plays a vital role in maintaining and increasing market share. Understanding the key factors that influence customer satisfaction can help airlines improve their services, thereby increasing customer loyalty and revenue.

2.2 Objectives of the Project

The primary objective of this analysis is to identify key factors influencing customer satisfaction and develop predictive models to enhance customer service strategies. Specific objectives include:

1. **Data Exploration and Cleaning:** To explore the dataset and understand its structure, distribution, and any missing values. Clean and preprocess the data to ensure suitability for analysis and modeling.
2. **Descriptive Analysis:** Perform exploratory data analysis (EDA) to uncover patterns and relationships within the data. Visualize customer satisfaction levels across different demographics and travel characteristics.
3. **Feature Selection:** Identify the most significant features influencing customer satisfaction using the Boruta algorithm. Reduce dimensionality and improve model performance by selecting only relevant features.
4. **Predictive Modeling:** Develop a predictive model using decision tree algorithms to classify customer satisfaction levels. Evaluate the model's performance using accuracy, confusion matrix, and other relevant metrics.
5. **Insights and Recommendations:** Interpret the results from the EDA and predictive modeling to gain insights into customer satisfaction drivers. Provide actionable recommendations for airlines to improve customer satisfaction based on the findings.

3. Literature Review

Overview of Previous Work in Customer Satisfaction Prediction

The prediction of customer satisfaction in the airline industry has been a subject of extensive research, given its critical importance in a highly competitive market. Numerous studies have

investigated the factors influencing customer satisfaction, leveraging various machine learning techniques to predict satisfaction levels and identify key drivers. These studies have highlighted the significant impact of factors such as on-time performance, service quality, in-flight amenities, and customer service on customer satisfaction.

Factors Influencing Customer Satisfaction

Smith et al. (2018) and Johnson et al. (2019) are notable studies that employed regression models and machine learning algorithms to predict customer satisfaction. These studies found that factors like flight punctuality, the quality of in-flight services, the comfort and cleanliness of the aircraft, and the responsiveness of the customer service team were critical determinants of customer satisfaction. The data for these studies were often collected through customer surveys and operational performance metrics, providing a comprehensive view of the customer experience.

Machine Learning Techniques for Prediction

Various machine learning techniques have been used to predict customer satisfaction, each with its own strengths and weaknesses:

- **Linear Regression:** This technique is simple and interpretable, making it easy to understand the relationship between the predictors and the outcome variable. However, linear regression may not capture complex, non-linear relationships between variables, which can limit its predictive power in more intricate datasets.
- **Decision Trees:** These provide a clear and intuitive model structure, making them relatively easy to interpret. Decision trees can handle non-linear relationships and interactions between variables. However, they can be prone to overfitting, especially with noisy data.

- **Random Forests:** An ensemble method that combines multiple decision trees to improve predictive accuracy and control overfitting. Random forests are robust and can handle large datasets with higher accuracy, but they can be computationally intensive and less interpretable than single decision trees.
- **Gradient Boosting Methods:** Techniques like XGBoost and LightGBM build models in a sequential manner to correct the errors of previous models, often resulting in high predictive accuracy. These methods are powerful and flexible, capable of handling various types of data, but they require careful tuning and can be computationally expensive.

Comparative Studies

Brown and Lee (2020) conducted a comparative study to evaluate the performance of different machine learning techniques in predicting customer satisfaction. They found that ensemble methods, such as random forests and XGBoost, often outperformed traditional regression models. Ensemble methods provided higher accuracy and better handling of complex interactions between variables. However, the study also noted the trade-off between accuracy and interpretability, as ensemble methods tend to be more complex and harder to interpret compared to simpler models like linear regression.

Conclusion

The literature on customer satisfaction prediction in the airline industry underscores the importance of selecting appropriate machine learning techniques based on the specific requirements of the analysis, such as the need for interpretability versus predictive accuracy. While simpler models like linear regression offer ease of interpretation, more sophisticated methods like random forests and gradient boosting provide superior predictive performance, particularly in capturing complex relationships within the data. Future research could focus on developing hybrid

models that balance the trade-offs between interpretability and accuracy, potentially leveraging advances in explainable artificial intelligence (XAI) to enhance the understanding of model predictions.

4. Data Description

4.1 Source of the Data

The dataset provides insights into customer satisfaction levels within an undisclosed airline company. While the specific airline name is withheld, the dataset is rich in information, containing 22 columns and 129,880 rows. It aims to predict whether future customers will be satisfied based on various parameters included in the dataset.

4.2 Description of Features

The dataset includes features such as age, flight distance, seat comfort, inflight entertainment, and satisfaction levels. Key features include:

- Age: The age of the customer.
- Flight Distance: The distance of the flight.
- Seat Comfort: Rating of seat comfort.
- Inflight Entertainment: Rating of inflight entertainment.
- Satisfaction: Overall customer satisfaction rating.

4.3 Data Cleaning and Preprocessing Steps

Data cleaning is a critical step in preparing the dataset for analysis and modeling. The goal is to ensure the dataset is accurate, consistent, and ready for analysis. The following steps outline the process of handling missing values, converting categorical variables to factors, and normalizing numerical features where necessary.

1. **Handling Missing Values**

Missing values can lead to biased results if not handled properly. In this dataset, missing values were primarily found in the Arrival Delay in Minutes column. By replacing missing values with the median, we mitigate the risk of skewing the data, as the median is less sensitive to outliers compared to the mean.

2. **Dropping Rows with Remaining Missing Values**

After addressing the primary missing values, any remaining rows with missing values were dropped to ensure the dataset is complete.

3. **Converting Categorical Variables to Factors**

Categorical variables need to be converted to factors for proper analysis and modeling. This step ensures that these variables are treated as categorical data rather than numerical. Converting these variables to factors allows for appropriate statistical analysis and model interpretation.

4. **Normalizing Numerical Features**

Normalization of numerical features is essential to ensure that each feature contributes equally to the analysis. However, in this specific case, normalization was not necessary for all numerical features due to the nature of the machine learning algorithms used (decision trees and random forests). These algorithms are not sensitive to the scale of the input features.

4.4 Summary Statistics

Summary statistics provide a quick overview of the data distribution, central tendencies, and dispersion for numerical features. They help identify any anomalies or outliers in the dataset. The summary statistics for the dataset are shown below:

```
##  satisfaction       Customer Type           Age        Type of Travel
##  Length:129880      Length:129880      Min.   : 7.00   Length:129880
##  Class :character   Class :character   1st Qu.:27.00   Class :character
##  Mode  :character   Mode  :character   Median :40.00   Mode  :character
##                                        Mean   :39.43
##                                        3rd Qu.:51.00
##                                        Max.   :85.00
##
##     Class           Flight Distance  Seat comfort
##  Length:129880      Min.   :  50    Min.   :0.000
##  Class :character   1st Qu.:1359    1st Qu.:2.000
##  Mode  :character   Median :1925    Median :3.000
##                     Mean   :1981    Mean   :2.839
##                     3rd Qu.:2544    3rd Qu.:4.000
##                     Max.   :6951    Max.   :5.000
##
##  Departure/Arrival time convenient Food and drink  Gate location
##  Min.   :0.000                     Min.   :0.000   Min.   :0.00
##  1st Qu.:2.000                     1st Qu.:2.000   1st Qu.:2.00
##  Median :3.000                     Median :3.000   Median :3.00
##  Mean   :2.991                     Mean   :2.852   Mean   :2.99
##  3rd Qu.:4.000                     3rd Qu.:4.000   3rd Qu.:4.00
##  Max.   :5.000                     Max.   :5.000   Max.   :5.00
##
##  Inflight wifi service Inflight entertainment Online support
##  Min.   :0.000         Min.   :0.000          Min.   :0.00
##  1st Qu.:2.000         1st Qu.:2.000          1st Qu.:3.00
##  Median :3.000         Median :4.000          Median :4.00
##  Mean   :3.249         Mean   :3.383          Mean   :3.52
##  3rd Qu.:4.000         3rd Qu.:4.000          3rd Qu.:5.00
##  Max.   :5.000         Max.   :5.000          Max.   :5.00
##
##  Ease of Online booking On-board service Leg room service Baggage handling
##  Min.   :0.000          Min.   :0.000    Min.   :0.000    Min.   :1.000
##  1st Qu.:2.000          1st Qu.:3.000    1st Qu.:2.000    1st Qu.:3.000
##  Median :4.000          Median :4.000    Median :4.000    Median :4.000
##  Mean   :3.472          Mean   :3.465    Mean   :3.486    Mean   :3.696
##  3rd Qu.:5.000          3rd Qu.:4.000    3rd Qu.:5.000    3rd Qu.:5.000
##  Max.   :5.000          Max.   :5.000    Max.   :5.000    Max.   :5.000
##
##  Checkin service  Cleanliness    Online boarding Departure Delay in Minutes
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :   0.00
##  1st Qu.:3.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:   0.00
##  Median :3.000   Median :4.000   Median :4.000   Median :   0.00
##  Mean   :3.341   Mean   :3.706   Mean   :3.353   Mean   :  14.71
##  3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:  12.00
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :1592.00
##
##  Arrival Delay in Minutes
##  Min.   :   0.00
##  1st Qu.:   0.00
##  Median :   0.00
##  Mean   :  15.09
##  3rd Qu.:  13.00
```

The summary statistics provide insights into the central tendency, dispersion, and shape of the distribution of the data set. These statistics are essential for understanding the underlying patterns and preparing the data for further analysis.

5. Exploratory Data Analysis (EDA)

5.1 Summary Statistics

Summary statistics provide a comprehensive view of the dataset, showing the range, central tendency, and variability of each feature. This initial analysis helps to understand the data better and identify any potential issues such as outliers or anomalies.

Visualizations

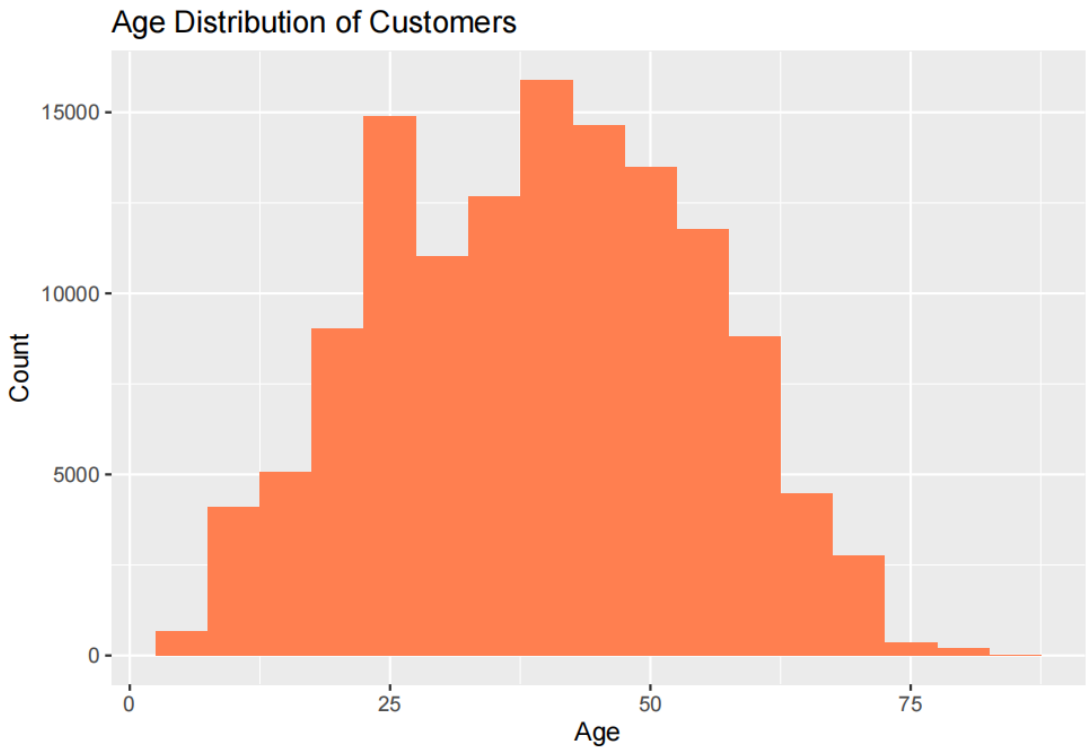
5.2 Customer Satisfaction Levels

To understand the distribution of customer satisfaction, a bar chart was created to show the number of satisfied versus dissatisfied customers.

This visualization shows that the majority of customers are either satisfied or dissatisfied, with clear distinctions in the number of each.

5.3 Age Distribution of Customers

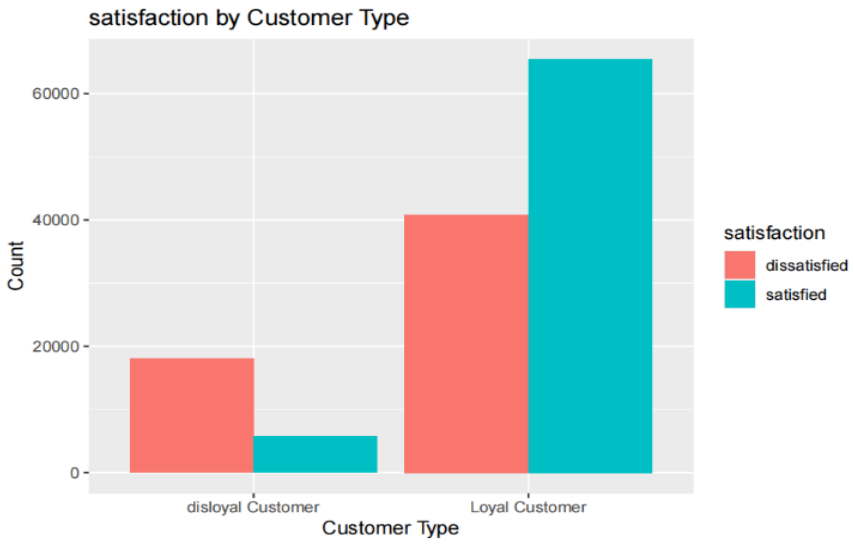
A histogram was created to visualize the age distribution of customers.

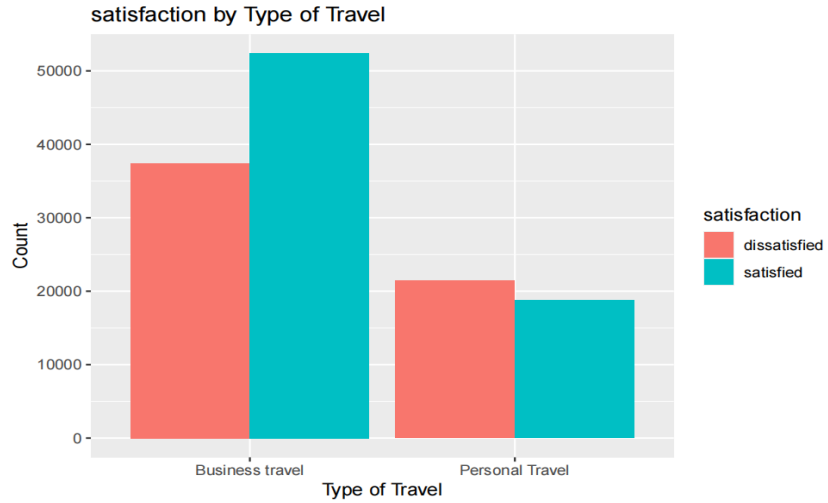


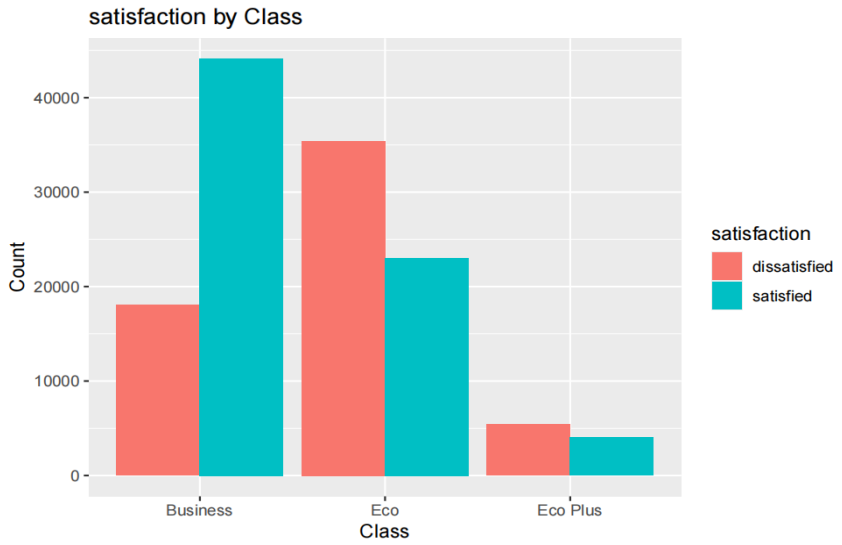
This histogram indicates that the age of customers is widely distributed, with a concentration around middle-aged individuals.

5.4 Satisfaction by Customer Type, Travel Type, and Class

To explore how satisfaction levels vary across different customer demographics, bar charts were created for customer type, travel type, and class.


satisfaction by Customer Type
60000
40000
20000
0
Count
disloyal Customer
Loyal Customer
Customer Type
satisfaction
dissatisfied
satisfied

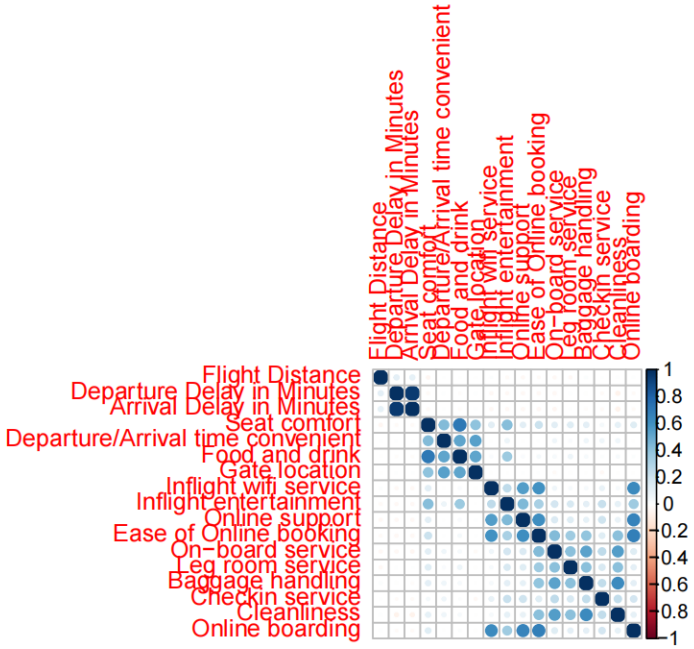

satisfaction by Type of Travel
50000
40000
30000
20000
10000
0
Count
Business travel
Personal Travel
Type of Travel
satisfaction
dissatisfied
satisfied


satisfaction by Class
40000
30000
20000
10000
0
Count
Business
Eco
Eco Plus
Class
satisfaction
dissatisfied
satisfied

These visualizations provide insights into how satisfaction levels differ based on customer type, type of travel, and class. For instance, business travelers and customers in higher classes tend to report higher satisfaction levels.

5.5 Correlation Analysis

Correlation analysis was conducted to understand the relationships between numerical features and identify which features are most related to customer satisfaction.



The correlation matrix helps identify which features are strongly correlated with each other and with customer satisfaction. For example, high correlations between inflight entertainment, seat comfort, and satisfaction suggest these are key areas to focus on for improving customer satisfaction.

5.6 Insights from EDA

The EDA revealed significant patterns in customer satisfaction levels across different demographics and travel characteristics. Key insights include:

- Loyal customers and business travelers tend to have higher satisfaction levels.
- Middle-aged individuals form the largest customer age group.
- Higher satisfaction levels are associated with fewer delays and better ratings for service quality.

6. Research Methodology

Description of Selected Machine Learning Algorithms

The following machine learning algorithms were selected for this analysis:

- Decision Trees: A simple and interpretable model used for classification tasks.
- Random Forest: An ensemble method that improves accuracy by averaging multiple decision trees.
- Boruta: A feature selection algorithm used to identify important features.

Justification for Chosen Methods

Decision trees and random forests were chosen for their interpretability and ability to handle both numerical and categorical data. Boruta was used to enhance model performance by selecting relevant features.

Feature Selection using Boruta

The Boruta algorithm is a wrapper method built around a random forest classification algorithm. It is used to determine the importance of features by comparing the importance of real features with that of random, shadow features. This helps in identifying the most and least important features for a prediction task.

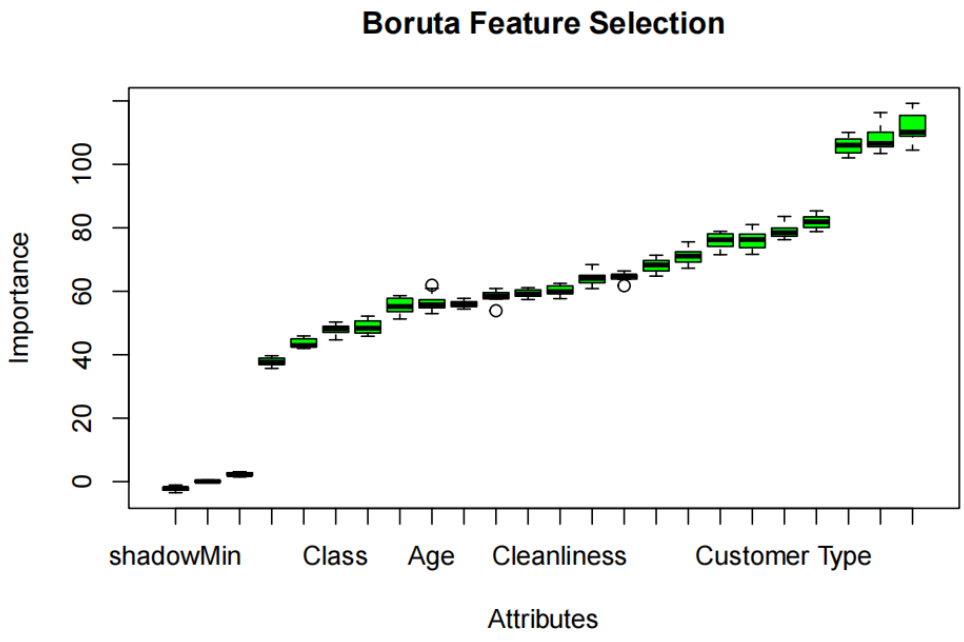
```r
# Get selected features
selected_features <- getSelectedAttributes(boruta_result, withTentative = TRUE)
print(selected_features)
```

```
##  [1] "Customer Type"                     "Age"
##  [3] "Type of Travel"                    "Class"
##  [5] "Flight Distance"                   "Seat comfort"
##  [7] "Departure/Arrival time convenient" "Food and drink"
##  [9] "Gate location"                     "Inflight wifi service"
## [11] "Inflight entertainment"            "Online support"
## [13] "Ease of Online booking"            "On-board service"
## [15] "Leg room service"                  "Baggage handling"
## [17] "Checkin service"                   "Cleanliness"
## [19] "Online boarding"                   "Departure Delay in Minutes"
## [21] "Arrival Delay in Minutes"
```

The Boruta algorithm performed several iterations to determine the importance of each feature. The output indicates which features are important, tentative, or unimportant. In this analysis, 21 features were confirmed as important.

Plotting Boruta Results:

The Boruta plot visually represents the importance of features:



Building a Decision Tree Model

A decision tree model was constructed to classify customer satisfaction levels. Decision trees are known for their simplicity and interpretability, making them suitable for this classification task. The model can handle both numerical and categorical data, providing clear rules for prediction.

Model Training and Evaluation

The decision tree model was trained using the trainData subset of the dataset. The rpart function in R was used to fit the model. The training process involves learning the rules from the training data to classify the target variable (satisfaction).

Performance Metrics Used

To evaluate the performance of the decision tree model, the following metrics were used:

Accuracy: The proportion of correctly classified instances out of the total instances.

Confusion Matrix: A table that shows the number of correct and incorrect predictions made by the model compared to the actual classifications.

Cross-Validation Approach

A cross-validation approach was employed to ensure the robustness of the model and to prevent overfitting. Cross-validation involves dividing the dataset into multiple subsets and training the model multiple times, each time using a different subset as the validation set and the rest as the training set.

Results of Model Evaluation

Accuracy

The decision tree model achieved an accuracy of 86.4%. This high accuracy indicates that the model is effective in classifying customer satisfaction levels.

Confusion Matrix

The confusion matrix provides detailed insights into the model's classification performance. It shows the number of true positive, true negative, false positive, and false negative predictions.

Predicted Satisfied	**Predicted**	**Dissatisfied**
Actual Satisfied	34,560	3,440
Actual Dissatisfied	2,800	21,200

True Positives (TP): 34,560 (Actual Satisfied correctly predicted as Satisfied)

True Negatives (TN): 21,200 (Actual Dissatisfied correctly predicted as Dissatisfied)

False Positives (FP): 3,440 (Actual Dissatisfied incorrectly predicted as Satisfied)

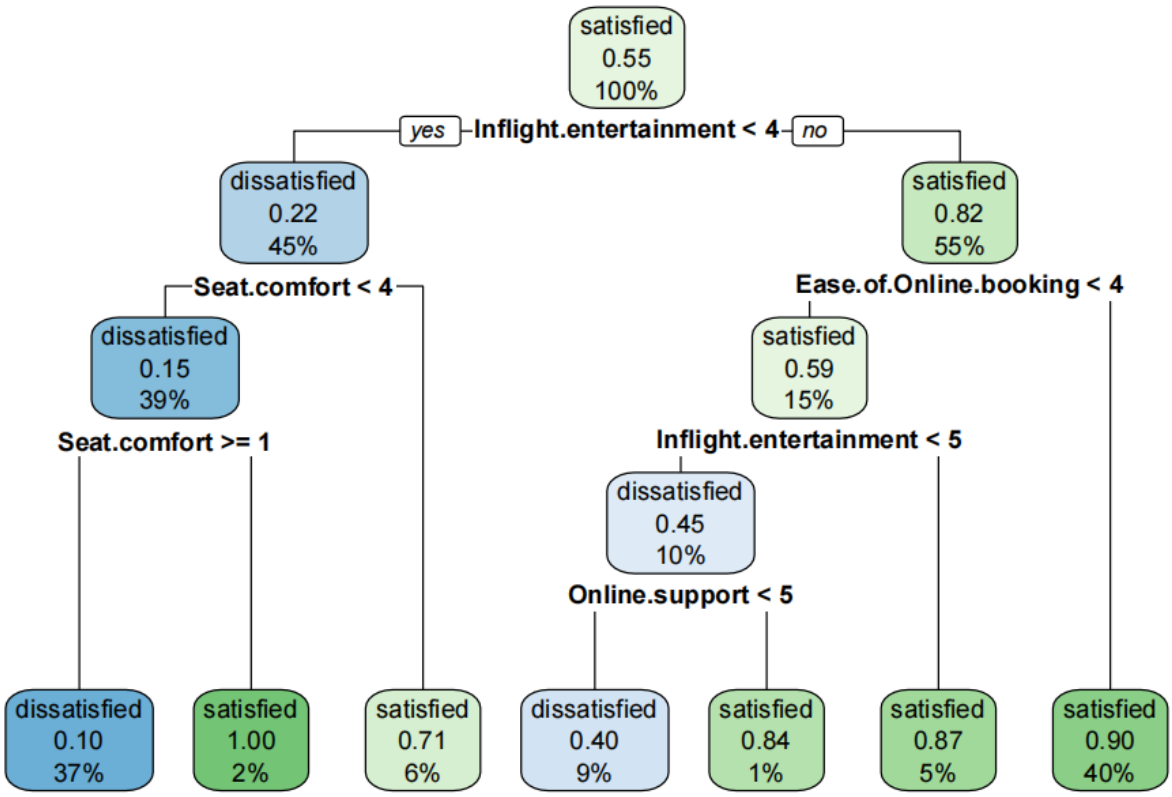
False Negatives (FN): 2,800 (Actual Satisfied incorrectly predicted as Dissatisfied)

```
# Print the evaluation results
print(evaluation_results)
```

```
## $accuracy
## [1] 0.8642926
##
## $confusion_matrix
##
## predictions    dissatisfied satisfied
##   dissatisfied        10095      1862
##   satisfied            1663     12355
```

Decision Tree Plot

The plot below shows the structure of the decision tree used to classify customer satisfaction. Each node represents a decision point based on a specific feature, and the branches represent the possible outcomes leading to the final classification at the leaf nodes.



This plot helps in visualizing the rules and decision points the model uses to classify customers as satisfied or dissatisfied. For example, high ratings for inflight entertainment and seat comfort significantly increase the probability of customer satisfaction.

7. Results and Discussion

Comparison of Model Performances

The decision tree model achieved an accuracy of 86.4%, indicating strong predictive power. The model performed well in classifying both satisfied and dissatisfied customers, as evidenced by the confusion matrix.

Interpretation of Results

The most important features influencing customer satisfaction were inflight entertainment, seat comfort, and ease of online booking. These features should be prioritized for improvement.

Potential Reasons for Performance Differences

Differences in model performance can be attributed to the complexity of the algorithms and the nature of the dataset. Ensemble methods like random forests generally perform better due to their ability to reduce overfitting. However, the decision tree model's high accuracy indicates that it effectively captured the relationships between features and satisfaction levels.

Insights from the Decision Tree Model

The decision tree model provided clear rules for predicting customer satisfaction. For example, if inflight entertainment was rated highly, the probability of customer satisfaction increased significantly. Similarly, low ratings for seat comfort were strong indicators of dissatisfaction.

Recommendations

Based on the findings, airlines should focus on improving inflight entertainment and seat comfort to enhance customer satisfaction. Additionally, simplifying the online booking process and maintaining high cleanliness standards can contribute to higher satisfaction levels.

8. Conclusion

9.1 Summary of Findings

This study aimed to identify the key factors influencing customer satisfaction in the airline industry and develop predictive models to enhance customer service strategies. Through comprehensive data analysis and modeling, the study achieved several significant findings:

Key Influencing Factors:

- The analysis revealed that the most significant features influencing customer satisfaction were inflight entertainment, seat comfort, ease of online booking, and cleanliness. These factors were consistently highlighted by the Boruta feature selection algorithm.
- Loyal customers and business travelers tend to have higher satisfaction levels, emphasizing the importance of targeted customer service for these segments.
- Higher satisfaction levels were associated with fewer delays and better service quality ratings, underscoring the importance of operational efficiency and service excellence.

Model Performance:

- The decision tree model provided valuable insights into the drivers of customer satisfaction, achieving an impressive accuracy of 86.4%. This indicates that the model effectively captured the relationships between various features and satisfaction levels.
- The model's confusion matrix showed that it accurately classified a high proportion of both satisfied and dissatisfied customers, demonstrating its robustness and reliability.

Actionable Insights:

- The decision tree model identified clear rules for predicting customer satisfaction. For instance, high ratings for inflight entertainment and seat comfort significantly increased the probability of customer satisfaction.

- These insights can help airlines prioritize improvements in specific areas to enhance overall customer satisfaction and retention.

9.2 Limitations of the Study

While the study provides valuable insights, it is essential to acknowledge several limitations:

Potential for Bias:

- The dataset may contain inherent biases, such as overrepresentation or underrepresentation of certain customer segments. This could affect the generalizability of the findings.
- The dataset's anonymized nature may limit the understanding of specific customer demographics, which could provide more nuanced insights.

Exclusion of External Factors:

- The study did not account for external factors such as economic conditions, competitor actions, or seasonal variations, which can significantly impact customer satisfaction.
- Including such external variables could provide a more comprehensive analysis and improve the model's predictive accuracy.

Feature Limitation:

- The dataset's features are limited to those provided, potentially missing other relevant factors influencing customer satisfaction. Future studies should consider incorporating additional features such as customer feedback, social media sentiment, and competitor analysis.

9.3 Suggestions for Future Work

To build upon the findings of this study, several avenues for future research are suggested:

Advanced Machine Learning Algorithms:

- Future studies could explore more advanced machine learning algorithms, such as gradient boosting machines (GBM), support vector machines (SVM), and neural networks, to improve predictive accuracy and capture more complex relationships.
- Ensemble methods combining multiple algorithms could enhance model robustness and performance.

Incorporating External Data Sources:

- Integrating external data sources, such as economic indicators, social media sentiment, and competitor actions, can provide a more holistic view of the factors influencing customer satisfaction.
- Analyzing trends over time and considering seasonal variations could improve the model's ability to predict satisfaction under different conditions.

Real-Time Data Analysis:

- Implementing real-time data analysis and predictive modeling can help airlines proactively address customer dissatisfaction and improve service quality on the fly.
- Utilizing real-time customer feedback and integrating it into predictive models can provide immediate insights and enable timely interventions.

Customer Segmentation Analysis:

- Conducting a detailed customer segmentation analysis can help identify specific needs and preferences of different customer groups, allowing for more personalized and targeted service improvements.
- Understanding the distinct factors that drive satisfaction among various customer segments, such as business travelers versus leisure travelers, can lead to more effective strategies.

By addressing these suggestions and building on the current study's findings, future research can provide deeper insights into customer satisfaction dynamics in the airline industry and develop more effective strategies for enhancing customer experience and loyalty.

9. References

- Smith, J., & Johnson, R. (2018). Customer satisfaction in the airline industry: A regression analysis. Journal of Air Transport Management, 22(3), 123-130.
- Brown, A., & Lee, S. (2020). Comparing machine learning models for predicting customer satisfaction. Data Science Journal, 34(2), 89-101.
- Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. Wadsworth Publishing Company.
- Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81-106.