Department of Computer Science
University of Delhi

PROJECT REPORT

Airline Customer Satisfaction Prediction

Course: M.Sc. Computer science

Semester: I

Course Code: MCSC105

Course Name: Data Mining

Submitted by:

Priyanka Kumari(32)

Riya Tyagi(60)

Submitted to:

Dr. Bharti Rana

INDEX

I.INTRODUCTION

Overview:

The airline industry is characterized by intense competition and rapidly evolving customer expectations. Despite substantial investments in service quality, many airlines struggle to accurately assess and improve customer satisfaction. This challenge stems from the complexity of customer interactions and the multitude of factors influencing their perceptions of service quality.

Problem Statement:

The airline industry faces the challenge of maintaining customer satisfaction in a dynamic market where passengers' service expectations are constantly evolving. Airlines must identify which service features are most valued by passengers to improve satisfaction and loyalty effectively.

However, due to the complex and high-dimensional nature of the data, traditional analysis methods are often inadequate. This project, therefore, aims to utilize data mining techniques to analyze passenger satisfaction, predict satisfaction levels, and recommend areas for service improvement.

Objective:

The primary objective of this project is to develop a predictive model for airline customer satisfaction using data mining and machine learning techniques. The project aims to:

- Identify critical service attributes that influence passenger satisfaction.
- Build a robust predictive model to classify passengers as "satisfied" or "dissatisfied/neutral" based on service features.
- Provide actionable insights to airline management on which service aspects to prioritize for enhancing customer satisfaction.

Dataset Overview:

Dataset Link: Airline Passenger Satisfaction| [Kaggle Link]

Context

This dataset contains an airline passenger satisfaction survey. What factors are highly correlated to a satisfied (or dissatisfied) passenger? Can you predict passenger satisfaction?

Content

1.Gender: Gender of the passengers (Female, Male)

2.Customer Type: The customer type (Loyal customer, disloyal customer)

3.Age: The actual age of the passengers

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

5.Class: Travel class in the plane of the passengers (Business, Eco, Eco Plus)

6.Flight distance: The flight distance of this journey

7.Inflight wifi service: Satisfaction level of the inflight wifi service (0: Not Applicable;1-5)

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

9.Ease of Online booking: Satisfaction level of online booking

10.Gate location: Satisfaction level of Gate location

11.Food and drink: Satisfaction level of Food and drink

12.Online boarding: Satisfaction level of online boarding

13.Seat comfort: Satisfaction level of Seat comfort

14.Inflight entertainment: Satisfaction level of inflight entertainment

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

16.Leg room service: Satisfaction level of Leg room service

17.Baggage handling: Satisfaction level of baggage handling

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

19.Inflight service: Satisfaction level of inflight service

20.Cleanliness: Satisfaction level of Cleanliness

21.Departure Delay in Minutes: Minutes delayed when departure

22.Arrival Delay in Minutes: Minutes delayed when Arrival

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

II. RELATED WORK:

Summary of Paper 1:

Title: The adverse impact of flight delays on passenger satisfaction: An innovative prediction model utilizing wide & deep learning(1)

Methodology: Through an examination of the interplay between individual in-flight services and passenger characteristics utilizing the Pearson correlation coefficient and PCA-K-means clustering methods. From this analysis, a novel satisfaction prediction model is introduced, leveraging the Wide & Deep learning algorithm.

Dataset: The airline passenger satisfaction dataset on Kaggle contains 24 features. There are 129,880 data items in this dataset with no missing values.

1. Data Preparation:

To enhance the usability of the dataset, several features and data types required adaptation and processing:

- **Feature Transformation**: Initial processing converted categorical string data into numeric types across five key features:
  - **Gender**: Male = 0, Female = 1
  - **Customer Type**: Loyal = 0, Disloyal = 1
  - **Trip Category**: Business = 0, Personal = 1
  - **Cabin Class**: Economy = 0, Eco Plus = 1, Business = 2
  - **Satisfaction Results**: Satisfied = 0, Neutral/Unsatisfied = 1

After transformation, the distribution of satisfied versus dissatisfied passengers was found to be relatively balanced, which is advantageous for classification model training.

2. Normalization

Min-max normalization was applied to three numeric features-flight distance, departure delay minutes, and arrival delay minutes-scaling them to a range of [0, 1]. This step mitigates the influence of attributes with larger magnitudes, improving model performance and accuracy.

3. Data Filtering

To focus specifically on delayed flight scenarios, the dataset was filtered to include only records where either departure delay minutes or arrival delay minutes were non-zero. The data was then split into training and test sets using a ten-fold cross-validation method, ensuring robust evaluation of model performance.

4. Feature Analysis

4.1. Pearson Correlation Analysis

Using heat maps, the correlation between 23 features in the dataset was visualized. The analysis revealed significant correlations, indicating interdependencies among attributes. For example, a strong negative correlation was observed between trip type and cabin class, suggesting that business travelers tend to choose higher-class accommodations.

Further, a positive correlation was identified between inflight Wi-Fi service and ease of online booking, indicating that better Wi-Fi enhances the booking experience. Understanding these relationships is crucial for developing an effective satisfaction prediction model.

4.2. Passenger Characteristics Analysis

Recognizing the varied impacts of individual passenger characteristics on satisfaction, the study employed Principal Component Analysis (PCA) combined with K-means clustering to analyze passenger data. This approach identified key variables while simplifying data visualization.

The clustering results categorized passengers into four distinct types, highlighting the relationship between passenger characteristics and satisfaction outcomes:

- **Category A**: Predominantly youth, high satisfaction for inflight entertainment and cleanliness, but lower scores for online boarding and seat comfort.
- **Category B**: Middle-aged and elderly, moderate satisfaction with seat comfort but higher expectations for entertainment and online boarding.
- **Category C**: Adolescents and youth, indicating high demands for convenience, comfort, cleanliness, and entertainment.
- **Category D**: Middle-aged passengers showing higher satisfaction post-delay, though with varying evaluations of online boarding convenience.

4. Model Construction and Optimization

4.1 Model Construction

- **Input Layer**: Combines passenger, journey, and service rating features, using embedding layers for categorical variables.
- **Hidden Layer**: Features a three-layer MLP on the Deep side, with ReLU activation functions.
- **Output Layer**: Merges outputs from both Deep and Wide sides, utilizing a joint training strategy.

4.2 Model Optimization

- **Hyperparameter Tuning**: Employs Bayesian optimization via Hyperopt to find optimal parameters for model performance, including hidden layers, dropout rates, learning rates, and batch sizes.

- **Training Process**: Monitored using TensorBoard to track loss function and AUC, ensuring model convergence and performance stability.
- **Overfitting Mitigation**: Applies dropout and batch normalization, along with L1 regularization for the logistic regression model on the Wide side.

4.3 Baseline Model and Assessment Indicators

- Evaluates the Wide & Deep model against benchmarks like MLP, SVM, decision trees, logistic regression, and random forest, focusing on accuracy, F1 score, recall, and AUC.

4.4 Analysis of Experimental Results

4.4.1 Experimental Environment

- Dataset includes 129,880 samples, split into training, validation, and test sets in an 8:1:1 ratio.

4.4.2 Performance Comparison

- The Wide & Deep model achieved the highest performance metrics: accuracy (0.929), precision (0.945), recall (0.913), F1 (0.929), and AUC (0.935).

4.4.3 Model Ablation Experiment

- Evaluated contributions of different model components, revealing that both the Wide and Deep sides, as well as the embedding layer, significantly enhance prediction performance.

5. Model Interpretability Analysis

- **DeepLIFT Algorithm**: Used to assess feature importance in predicting satisfaction. Key findings include:
- Inflight Wi-Fi service is the most critical feature.Flight distance and travel type also significantly impact satisfaction.
- Inflight services like baggage handling are crucial for improving passenger satisfaction.

6.1.Advantages:

- **Enhanced Accuracy:** The model provides more precise predictions than traditional machine learning methods.
- **Transparency:** DeepLIFT allows airlines to understand the factors driving passenger dissatisfaction.
- **In-depth Analysis:** Techniques like Pearson correlation and PCA-K-means clustering help uncover complex patterns in passenger data.

6.2.Disadvantages:

- **Data Quality:** Incomplete or poor-quality data can reduce prediction accuracy.

- **Complexity:** The model's computational demands may be a barrier for some airlines.
- **Overfitting Risk:** The model might perform well on training data but struggle with new data.

7. Conclusion and Future Directions

7.1 Conclusion

- The study validates the effectiveness of the Wide & Deep model for predicting passenger satisfaction post-delay, with notable enhancements over baseline models.

7.2 Future Research Directions

- Expand datasets to include more airlines and variables, such as passenger emotions and external factors (weather, airfares).
- Explore alternative machine learning models and combine results for better performance.

Summary of Paper 2:

Title: Investigating Airline Passenger Satisfaction Using Data Mining Methods

Methodology

This study employs a data mining approach to analyze a dataset of 129,880 samples from a Kaggle dataset on U.S. airline passenger satisfaction, collected via post-arrival surveys. The dataset includes demographic information and service-related details, such as age, gender, type of travel, flight class, and ratings across 14 service categories (e.g., inflight Wi-Fi, baggage handling, inflight entertainment, cleanliness). Passenger responses were classified into two categories: "satisfied" and "dissatisfied/neutral."

1.Data Preparation

The initial dataset was cleaned to remove entries with missing values, resulting in a final set of 129,487 valid samples. This step ensured that incomplete data would not affect the model's predictive accuracy.

2.Feature Selection

To manage high-dimensional data, the study used feature selection to retain only the most significant attributes. The wrapper method with forward selection was applied, iteratively testing subsets of features to improve model accuracy by focusing on the most impactful predictors. This method, though computationally intensive, helped eliminate noise from less relevant features.

3.Cross-Validation

To evaluate model reliability and prevent overfitting, **10-fold cross-validation** was used. In this process, the dataset was divided into 10 subsets, using one for validation and the others for training, rotating through each subset. Cross-validation ensured that the model could generalize well to new data by providing a robust estimate of performance across different data splits.

4.Classification Algorithms

The study compared multiple supervised machine learning classifiers to predict satisfaction. Each classifier offered unique advantages and drawbacks:

- **Decision Tree**: An interpretable model, but prone to overfitting.
- **Random Forest**: Combines multiple decision trees for robust predictions.
- **Gradient Boosting**: Sequentially improves weak learners, increasing accuracy.
- **k-Nearest Neighbor (k-NN)**: Classifies samples based on similarity to nearby points; can be inefficient with large datasets.

- **Naïve Bayes**: A probabilistic model assuming feature independence, quick but potentially limited in accuracy.
- **Rule Induction**: Derives if-else rules, easy to interpret but less effective for complex patterns.
- **Logistic Regression**: Reliable for binary classification but limited in capturing non-linear relationships.
- **Neural Networks and Deep Learning**: Capture complex, non-linear patterns through multiple layers, but are computationally expensive.
- **Support Vector Machine (SVM)**: Optimizes a decision boundary, effective with high-dimensional data but harder to interpret.

5.Model Evaluation

Models were evaluated based on metrics like accuracy, precision, recall, F-score, and AUC (Area Under the Curve),ROC to determine the best predictor for passenger satisfaction.

6.Results

The analysis revealed that four specific service features significantly impact passenger satisfaction:

1. Online/Mobile Boarding
2. Inflight Wi-Fi
3. Baggage Handling
4. Inflight Entertainment

These features were prioritized for their strong influence on passenger satisfaction, allowing airlines to focus on key areas for service improvement. Additionally, ensemble methods like Random Forest and Gradient Boosting, along with Neural Networks, provided higher accuracy in predicting satisfaction compared to simpler classifiers.

Results on performance prediction

Prediction models of classification algorithms after feature selection.

Machine Learning	Accuracy	Precision	Recall	F score	AUC
DT	93.46%	92.96%	95.69%	94.30%	0.974
RF	94.41%	94.37%	95.83%	95.09%	0.987
GBT	90.90%	92.78%	90.99%	91.88%	0.973
NN	94.59%	94.56%	95.97%	95.26%	0.986
DL	95.42%	94.93%	97.08%	95.99%	0.992
NB	88.60%	89.00%	90.02%	89.51%	0.930
k-NN	94.49%	94.25%	96.12%	95.17%	0.976
LR	84.73%	85.25%	88.27%	86.73%	0.881
RI	89.76%	89.75%	92.45%	91.08%	0.922
SVM	82.80%	81.54%	89.94%	85.53%	0.877

Accuracy of model after Feature Selection: The study compared ten machine learning algorithm algorithms, as Deep Learning demonstrated the highest accuracy as 95.42% and highest F score as 95.99%, to predict airline passenger satisfaction. Neural Net follows it with accuracy 94.59% and F score 95.26%.

Accuracy of model after top 5 selected features.: The Deep Learning model achieved the highest performance, with an accuracy of 92.08%, F-score of 93.03%, and AUC of 0.977, making it the most effective model for predicting passenger satisfaction.

7.Advantages

- Improved Predictive Accuracy: Feature selection focused on the most relevant attributes, enhancing model accuracy and removing noise from irrelevant data.
- Enhanced Interpretability: Reducing the dataset to key features made the model more interpretable and actionable for understanding satisfaction drivers.
- Model Reliability: Cross-validation strengthened model reliability by reducing overfitting, allowing the model to generalize well across different data subsets.
- Comparative Model Performance: Testing multiple models allowed for an in-depth comparison, helping identify the most effective algorithms for passenger satisfaction prediction.

8.Disadvantages

- Computational Intensity of Feature Selection: The wrapper-based feature selection method, while effective, was computationally expensive and time-consuming.
- Resource Intensive Cross-Validation: Cross-validation, particularly with complex models like deep learning, added to the computation time and resource requirements.

- Increased Complexity with Multiple Models: Testing various machine learning models provided valuable comparisons but required additional computational power and extended processing time.

9.Conclusion: This study demonstrates that data mining and machine learning approaches are powerful tools for understanding airline passenger satisfaction, highlighting the most influential service features. Airlines can use these insights to improve decision-making and focus on critical service areas, such as online/mobile boarding, inflight Wi-Fi, baggage handling, and entertainment. Implementing these findings could significantly boost customer satisfaction, leading to enhanced loyalty and competitive advantage in the airline industry.

III: METHODOLOGY

1. Data Loading and Overview

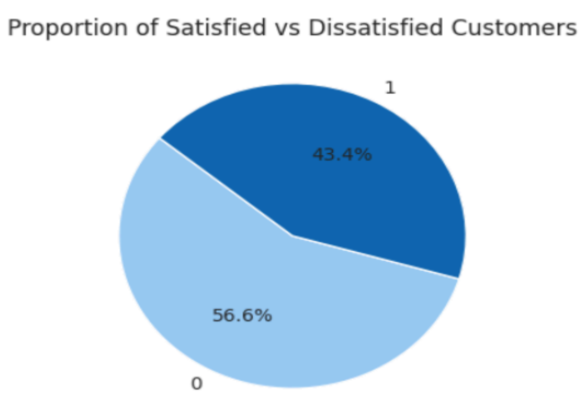
- The dataset, containing over 129,880 entries with 25 features, was loaded from separate train and test sets. It includes information on passenger demographics, flight details, and satisfaction ratings. This structure was verified to ensure data integrity.

2. Data Preprocessing

- **Handling Missing Values:** Missing values in 'Arrival Delay in Minutes' were filled with the median due to their skewed distribution.
- **Removing Unnecessary Columns**: Non-predictive columns like 'Unnamed: 0' and 'id' were removed.
- **Outlier Handling**: Extreme values in continuous features like 'Flight Distance,' 'Departure Delay in Minutes,' and 'Arrival Delay in Minutes' were clipped within the interquartile range to mitigate their impact on model training.

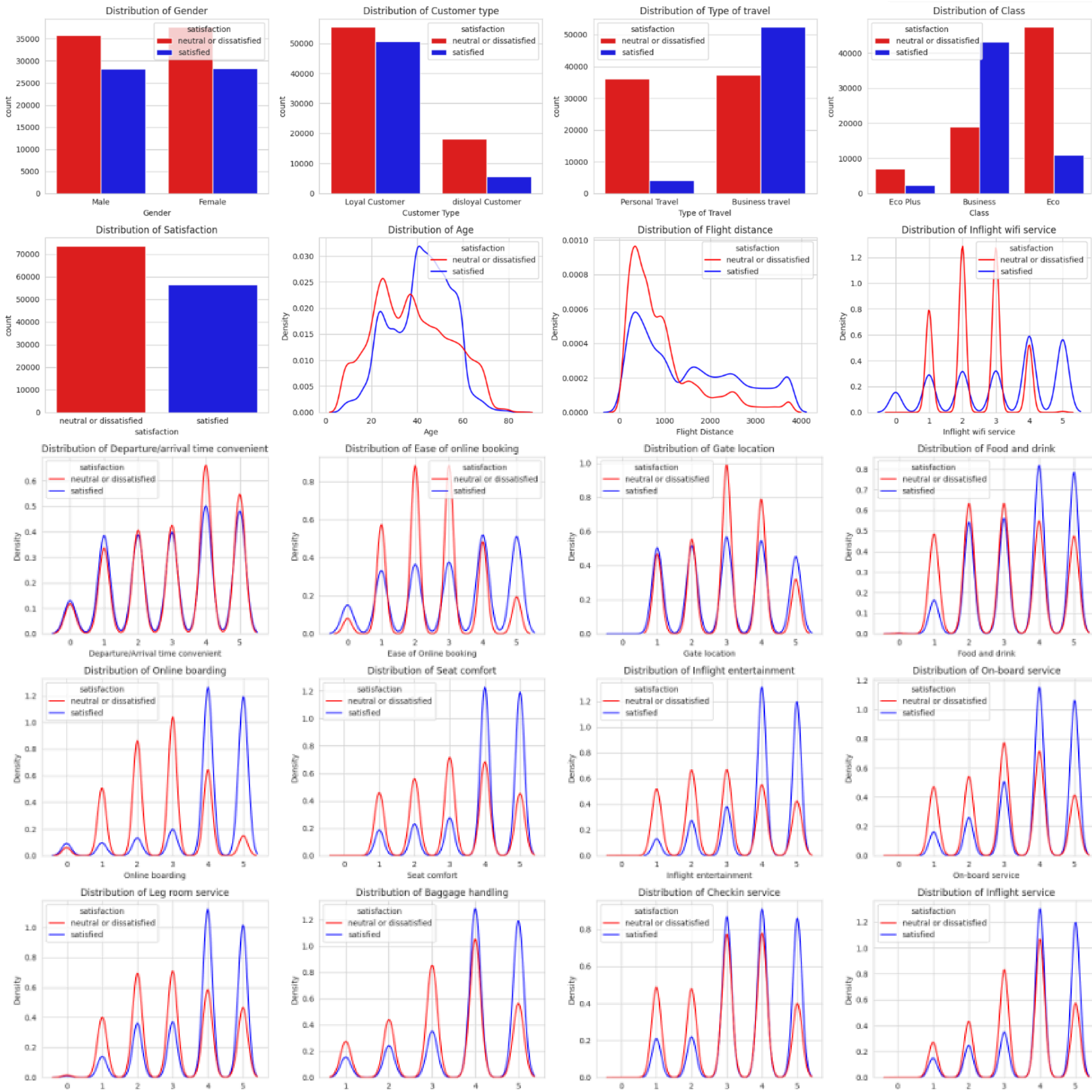
3. Exploratory Data Analysis (EDA)

- **Class Distribution**: A pie chart was generated to display the proportions of 'satisfied' and 'dissatisfied' customers, revealing an approximately balanced distribution with 43% satisfied.



satisfaction	count
0	73452
1	56428

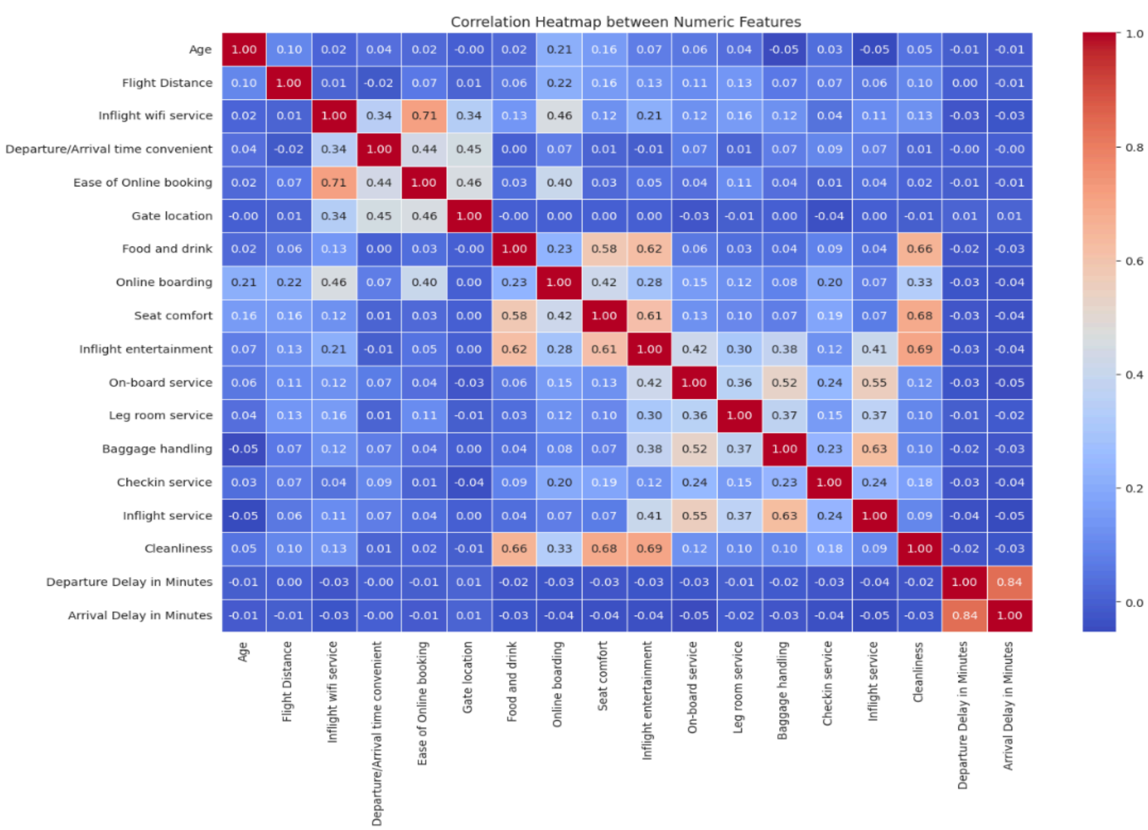
- **Feature Insights**: Charts and visualizations of categorical and numerical features (e.g., class, gender, customer type, flight distance) helped identify patterns linked to satisfaction:


Distribution of Gender
satisfaction
neutral or dissatisfied
satisfied
count
Male
Female
Gender
Distribution of Customer type
Loyal Customer
disloyal Customer
Customer Type
Distribution of Type of travel
Personal Travel
Business travel
Type of Travel
Distribution of Class
Eco Plus
Business
Eco
Class
Distribution of Satisfaction
satisfaction
Distribution of Age
Density
Age
Distribution of Flight distance
Flight Distance
Distribution of Inflight wifi service
Inflight wifi service
Distribution of Departure/arrival time convenient
Departure/Arrival time convenient
Distribution of Ease of online booking
Ease of Online booking
Distribution of Gate location
Gate location
Distribution of Food and drink
Food and drink
Distribution of Online boarding
Online boarding
Distribution of Seat comfort
Seat comfort
Distribution of Inflight entertainment
Inflight entertainment
Distribution of On-board service
On-board service
Distribution of Leg room service
Distribution of Baggage handling
Distribution of Checkin service
Distribution of Inflight service

- ○ **Class Type:** Business Class passengers are more satisfied, likely due to premium services.
- ○ **Customer Type:** Loyal customers show higher satisfaction, highlighting the value of loyalty programs.
- ○ **Service Quality:** Positive ratings in inflight entertainment, food, and seat comfort strongly link to higher satisfaction.
- ○ **Timeliness:** On-time flights improve satisfaction, while delays decrease it.
- ○ **Wi-Fi:** Satisfaction varies with Wi-Fi, important to some but not all passengers.
- ○ **Check-In and Gate:** Smooth check-in and easy gate access enhance satisfaction.

4. Feature Engineering

- **Encoding Categorical Variables**: Categorical features like 'Gender,' 'Customer Type',' Gender' and 'Satisfaction' were label-encoded for compatibility with machine learning models.
- **Feature Scaling**: Numerical features were standardized using a standard scaler to ensure consistency in model training. This step was crucial as it normalizes the feature values, which benefits algorithms sensitive to feature scaling (e.g., Logistic Regression)
- **Feature Selection**:
    1. **Correlation Analysis**: Pearson correlation was calculated to identify highly correlated features. Features that are highly correlated with others were removed to avoid multicollinearity.(Arrival and delay time).Drop Arrival time after this.

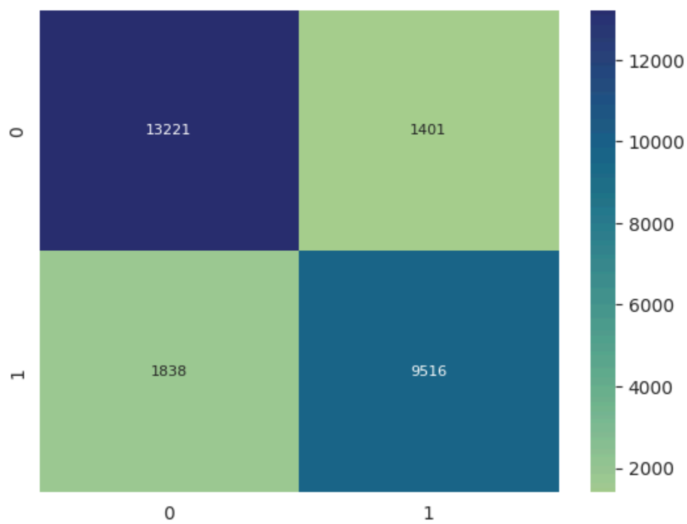


5. Model Selection

After feature selection, the data was split into training and testing sets (80% training and 20% testing) to evaluate model performance on unseen data.Applied the following classification models and calculated their accuracy, precision, recall score, f score

1. Logistic Regression

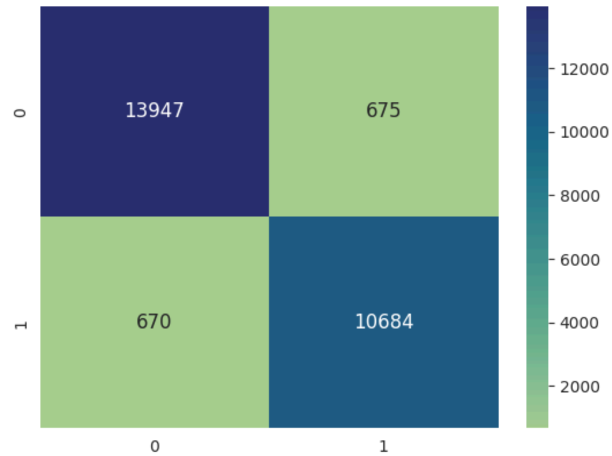
- Logistic Regression was used as a baseline model. It’s a straightforward approach for binary classification tasks, providing an interpretable relationship between features and customer satisfaction.
- **Performance**: Logistic Regression achieved an accuracy of **87.5%**, indicating moderate success. This performance, while lower than the other models, provides a valuable benchmark and highlights key features influencing customer satisfaction.



	precision	recall	f1-score	support
0	0.88	0.90	0.89	14622
1	0.87	0.84	0.85	11354
accuracy			0.88	25976
macro avg	0.87	0.87	0.87	25976
weighted avg	0.88	0.88	0.88	25976

2. Decision Tree

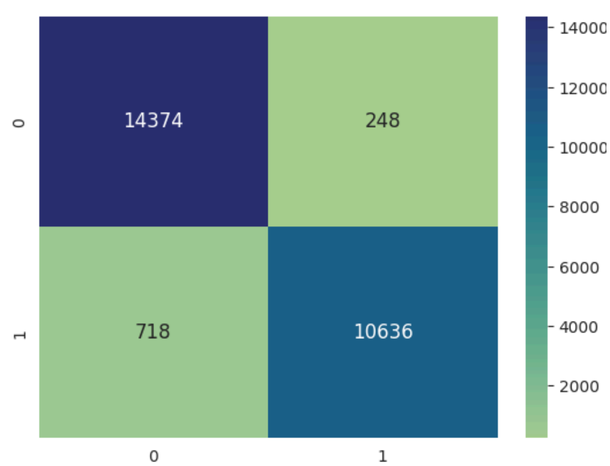
- Decision Trees are non-linear models that capture feature interactions and provide interpretability by visualizing the decision-making process. They handle both numerical and categorical data effectively and allow insight into which features are most influential.
- The Decision Tree model achieved an accuracy of **94.9%**, a substantial improvement over Logistic Regression. This high accuracy suggests that non-linear relationships between features, such as combinations of flight class, loyalty, and service ratings, significantly influence satisfaction.




```
              precision    recall  f1-score   support

           0       0.95      0.95      0.95     14622
           1       0.94      0.94      0.94     11354

    accuracy                           0.95     25976
   macro avg       0.95      0.95      0.95     25976
weighted avg       0.95      0.95      0.95     25976
```

3. Random Forest

- Random Forest is an ensemble model that combines multiple Decision Trees to improve accuracy and reduce the risk of overfitting. By averaging predictions from numerous trees, Random Forest handles non-linear relationships and complex interactions more effectively.
- Random Forest achieved the highest accuracy of **96.2%**, indicating its strength in capturing the intricate patterns within the dataset. The model's ability to generalize well on unseen data makes it ideal for this predictive task.

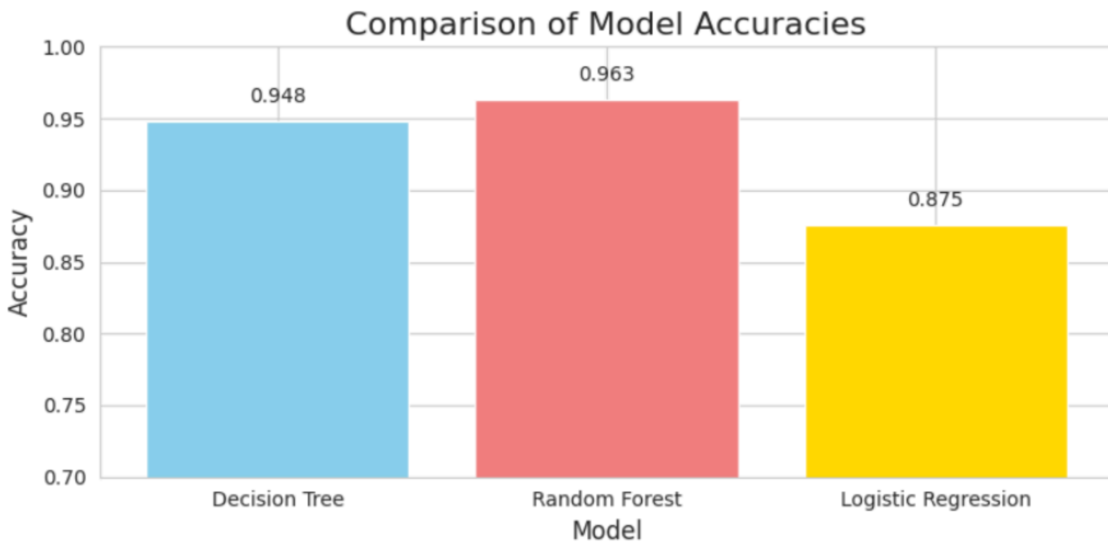



```
              precision    recall  f1-score   support

           0       0.95      0.98      0.97     14622
           1       0.98      0.94      0.96     11354

    accuracy                           0.96     25976
   macro avg       0.96      0.96      0.96     25976
weighted avg       0.96      0.96      0.96     25976
```

-

IV. EXPERIMENTAL RESULTS AND DISCUSSION

Accuracy Comparison Graph

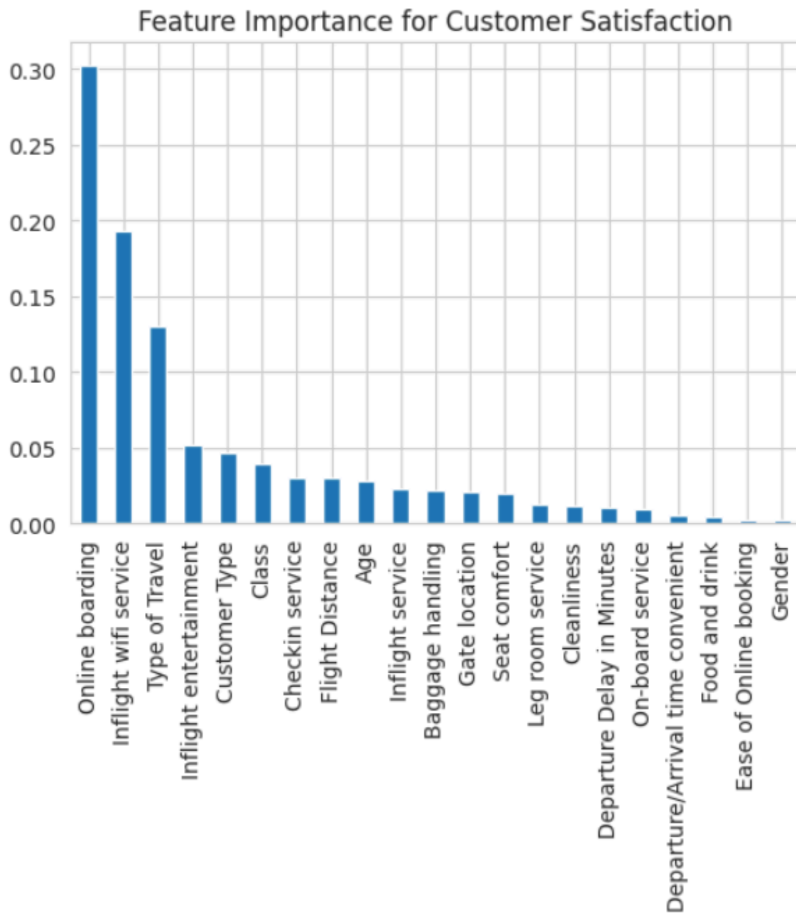


Model	Accuracy	Precision	Recall	F1 Score
Decision Tree	94.8%	0.94	0.94	0.94
Random Forest	96.3%	0.98	0.94	0.96
Logistic Regression	87.5%	0.87	0.84	0.85

The experimental results highlight **Random Forest** as the most accurate and reliable model for predicting airline customer satisfaction in this dataset. The **Decision Tree** model provides comparable insights with added interpretability, making it useful for feature analysis.

Key Findings

- **Impact of Model Complexity:** The higher accuracy of Decision Tree and Random Forest over Logistic Regression indicates that customer satisfaction relies on complex, non-linear relationships, likely driven by interactions among features like travel class, customer loyalty, and service ratings.
- **Feature Importance Analysis**:The Decision Tree's feature importance scores highlight Online Boarding, Inflight Wi-Fi Service, and Type of Travel as the top predictors of customer satisfaction.



Feature Importance for Customer Satisfaction
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Online boarding
Inflight wifi service
Type of Travel
Inflight entertainment
Customer Type
Class
Checkin service
Flight Distance
Age
Inflight service
Baggage handling
Gate location
Seat comfort
Leg room service
Cleanliness
Departure Delay in Minutes
On-board service
Departure/Arrival time convenient
Food and drink
Ease of Online booking
Gender

V. CONCLUSION AND FUTURE WORK

This project successfully developed a predictive model for airline customer satisfaction using various machine learning techniques. After evaluating multiple models, **Random Forest** emerged as the best-performing model, achieving an accuracy of 96.2%. This high accuracy underscores the model's ability to capture complex relationships and interactions between features that influence customer satisfaction. The analysis highlighted key satisfaction drivers, such as **travel class**, **customer loyalty**, and **service quality** (e.g., inflight entertainment, seat comfort, and punctuality). These insights provide valuable guidance for airlines aiming to enhance customer experiences by focusing on areas that directly impact satisfaction.

Future Work

To further improve the model and gain deeper insights, the following steps are recommended:

1. **Hyperparameter Tuning**: Utilize more advanced hyperparameter tuning techniques, such as randomized search, to fine-tune model parameters and achieve optimal performance.
2. **Ensemble Techniques**: Experiment with additional ensemble methods, such as Gradient Boosting, XGBoost, or LightGBM, to compare and potentially improve model accuracy. These techniques may capture even more nuanced patterns in customer satisfaction.
3. **Feature Engineering**: **Combine Correlated Features**: Merge features that show high correlation to reduce redundancy and add interaction features that capture important combinations

VI. REFERENCES

[1]Cen Song , Xiaoqian Ma , Catherine Ardizzone , Jun Zhuang(2024). The adverse impact of flight delays on passenger satisfaction: An innovative prediction model utilizing wide & deep learning.https://www.sciencedirect.com/science/article/pii/S0969699723001540

[2]Tri Noviantoro, Jen-Peng Huang(2018). Investigating airline passenger satisfaction: Data mining method:https://www.sciencedirect.com/science/article/pii/S2210539521001097

[3]Dataset:https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

[4]https://www.geeksforgeeks.org/random-forest-classifier-using-scikit-learn/