Airline Passenger Satisfaction Analysis

Final Project Report – Group 3

BUAN 6341.003

Applied Machine Learning

Submitted by:

Tanmay Itkelwar - txi220001

Vipul Suresh Sonje - vxs230000

Vineet Vinod Agarwal- vxa220046

Shaonan Ye - sxy220045

Unmona Das - uxd230001

Submitted to:

Dr. Zixuan Meng

Table of Contents

1. Motivation

Understanding passenger satisfaction is crucial for the aviation industry. This dataset provides valuable insights into what influences customer experiences, the factors such as service quality, seat comfort, cleanliness, and in-flight amenities.Satisfied passengers are more likely to be loyal, contributing to long-term revenue. Positive experiences get the brand effects influencing other potential customers. Airlines can target specific areas like baggage handling, check-in processes, or staff behavior for enhancements.

When evaluating multiple machine learning models like KNN, SVM (Kernel and Linear), Random Forest, Voting Classifier, and Decision Trees for predictive tasks such as airline passenger satisfaction, the choice of the best model depends on the performance metrics and overfitting issues. Improving customer satisfaction translates to increased ticket sales and revenues. Airlines that excel in customer satisfaction often lead in market share. Understanding areas of high efficiency area investment.

2. Data Overview

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

Kaggle Dataset Link: Airline Passenger Satisfaction

Total records: 129,880 rows

Features:

Categorical: Gender, customer type (Loyalty status), travel type (Business or Personal), class (Economy, Business, etc.).

Numerical: Age, flight distance, and various satisfaction ratings.

Ratings for various aspects (scaled from 1-5): In-flight entertainment, seat comfort, food and drink, on-ground service, etc.

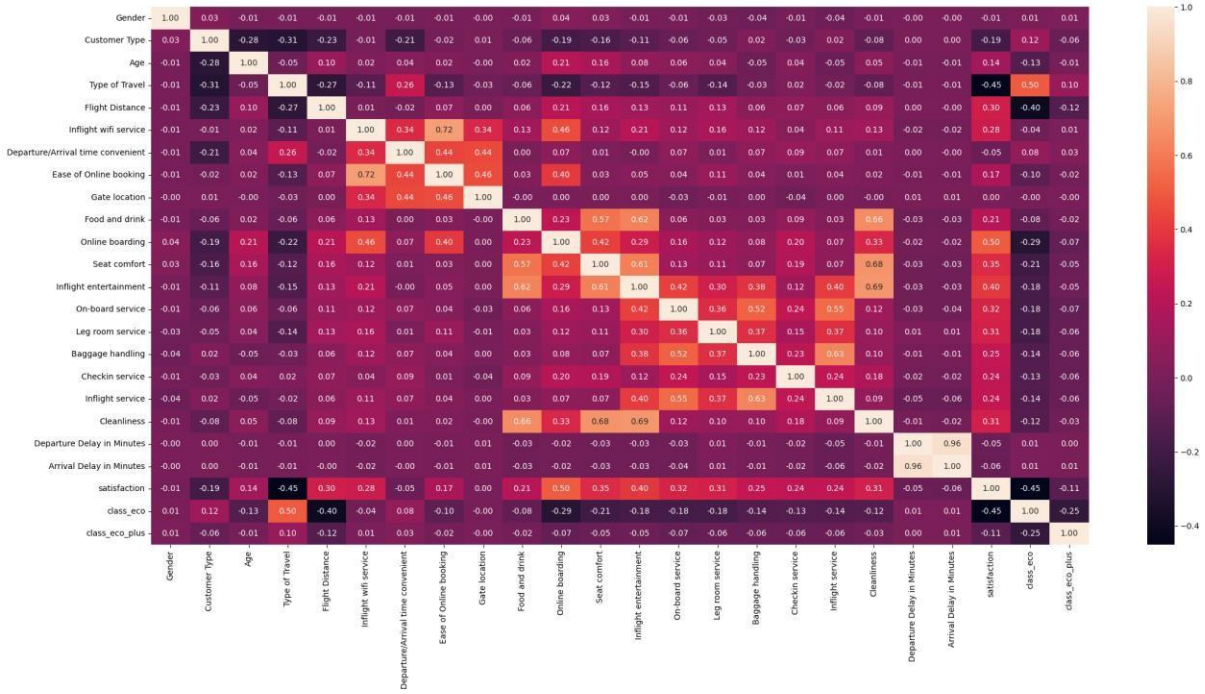
Other operational variables: Departure delay, arrival delay

Target variable:

Satisfaction: Indicates whether a passenger is satisfied or dissatisfied.

3. EDA and Data Preprocessing

3.1 Correlation matrix



Insights from the Correlation Matrix

- **Strong Positive Correlations**:
    - **Inflight Wi-Fi Service** and **Satisfaction** (~0.72): This suggests that higher satisfaction is strongly associated with better inflight Wi-Fi service.
    - **Online Boarding** and **Satisfaction** (~0.68): Indicates that efficient online boarding contributes significantly to passenger satisfaction.
    - **Class (e.g., Business class)** and **Satisfaction** (~0.50 for business class): Passengers in higher classes tend to report higher satisfaction levels.
- **Strong Negative Correlations**:
    - **Type of Travel (Business vs. Personal)** and **Satisfaction** (~-0.45): Business travelers tend to have different satisfaction levels compared to personal travelers, with business travel being more associated with satisfaction in this dataset.

- **Neutral/Weak Correlations**:
    - **Gender** shows little to no correlation with most features, indicating that satisfaction or service evaluations are not significantly influenced by gender.
    - **Age** also shows weak correlations with satisfaction and other service-related metrics.
- **Arrival Delay in Minutes** and **Satisfaction** have a weak negative correlation (~-0.17), indicating that delays slightly affect satisfaction, but it isn't a major factor.

Interpreting Key Variables

- **Satisfaction**:
    - Positively correlated with several service-related metrics (e.g., Inflight Wi-Fi, Seat Comfort, Cleanliness).
    - Negatively correlated with delays and type of travel.
- **Arrival/Departure Delay**:
    - High correlation between **Arrival Delay** and **Departure Delay** (~0.96), which is expected as delays in one typically affect the other.

Implications for the ML Model

- **Feature Importance**: Features like Inflight Wi-Fi Service, Online Boarding, and Class are likely to be highly influential in predicting passenger satisfaction.
- **Redundant Features**: Highly correlated variables like Arrival and Departure Delay might provide redundant information and may need careful handling in feature selection to avoid multicollinearity issues.

Dropping variables looking at correlation matrix

In our project, the primary objective is to predict **Customer Satisfaction** using the features provided in the dataset. As part of our exploratory data analysis (EDA), we computed a **correlation matrix** to identify the relationships between different features and their impact on customer satisfaction.

From the correlation matrix, we observed that the **"Arrival Delay in Minutes"** column exhibited a **weak negative correlation** with **Customer Satisfaction** (correlation value of approximately **-0.17**). This indicates that while delays in arrival might have a minor influence on customer satisfaction, they do not significantly contribute to predicting it.

Based on the correlation analysis, we concluded that **Arrival Delay in Minutes** is not a significant predictor of **Customer Satisfaction**. Therefore, we decided to drop this column from the dataset. This decision aligns with our goal of focusing on features that have a stronger influence on satisfaction, ensuring a more efficient and accurate predictive model.

Handling object variables with One hot encoding

We extended our Exploratory Data Analysis (EDA) and data preprocessing by identifying and converting non-integer categorical values into numerical format. Our dataset contained five categorical variables:

- **Gender** (Male, Female)
- **Customer Type** (Loyal, Disloyal)
- **Type of Travel** (Personal, Business)
- **Class** (Economy, Economy Plus, Business)
- **Satisfaction** (Neutral or Dissatisfied, Satisfied)

To transform these categorical variables into integer values, we used binary encoding (e.g., 0 and 1) for binary categories like Gender and Satisfaction. For variables with more than two categories, such as Class and Type of Travel, we applied one-hot encoding to represent them as integers while preserving their distinct categories. This conversion was crucial to ensure the data could be effectively processed by machine learning algorithms that require numerical inputs.

Feature Engineering (not to cover in depth and can be covered in the Beyond the scope section):

After generating the correlation matrix as part of our exploratory data analysis, we conducted a deeper examination of the dataset. This revealed that two critical features—Flight Distance and Departure Delay in Minutes—exhibited skewed distributions. Additionally, we identified that both columns contained some negative values, which is inconsistent with their real-world interpretation.

By applying log transformations to Flight Distance and Departure Delay in Minutes, we ensured that these features are more suitable for machine learning algorithms, ultimately improving the robustness and accuracy of our predictive model.

4. Model Selection

Basic introduction on the model selection process as this is a classification problem and that's why choses these models.

4.1 Decision Tree

Hyperparameters:

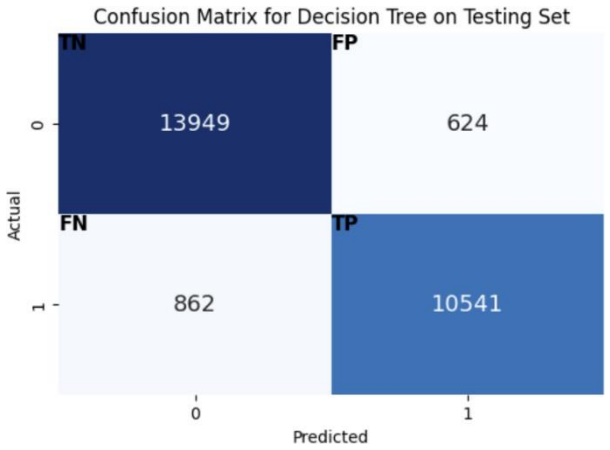
Max Depth – 10, Min Samples Leaf – 1 Min Samples split – 5 Accuracy

Decision Tree accuracy on the training set: 94.71%

Decision Tree accuracy on the testing set: 94.28%

Precision: 0.94

Recall: 0.92



4.2 KNN

Hyperparameters:

Neighbors 9

Accuracy on the training set:

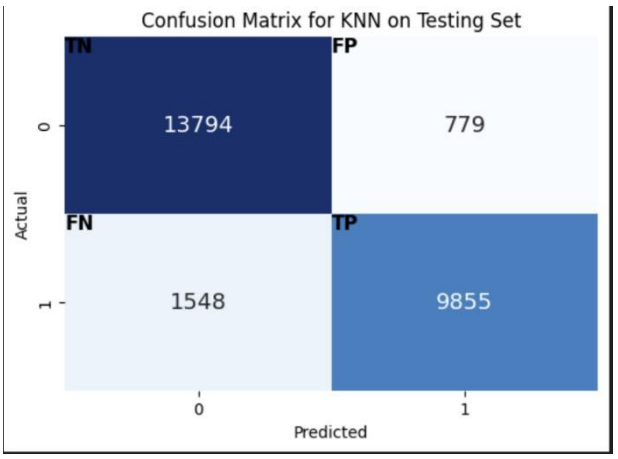
92.78%

Accuracy on the testing set:

91.04%

Precision 0.92

Recall 0.86



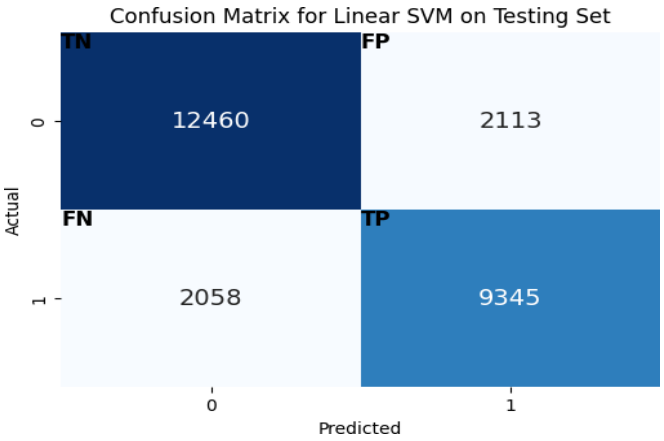
4.3 SVM

4.3.a Linear SVM

best parameters :C=0.001 Hyperparameters: 'C': [0.001,
0.01, 0.1, 1, 10,
100, 1000000]

Training accuracy: 84.15%

Testing accuracy 83.94%



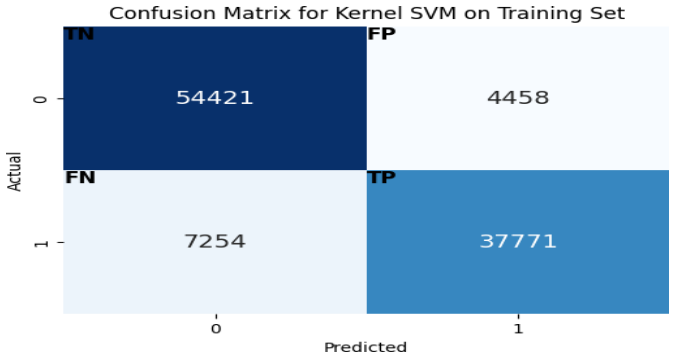
4.3.b Kernal SVM

Best parameters: C=1, gamma=0.1

Hyperparameters: 'C': [0.001, 0.01, 0.1, 1, 10],
'gamma': [0.001, 0.01, 0.1]

Training accuracy: 88.73%

Testing accuracy: 88.72%



4.4 Logistic Regression

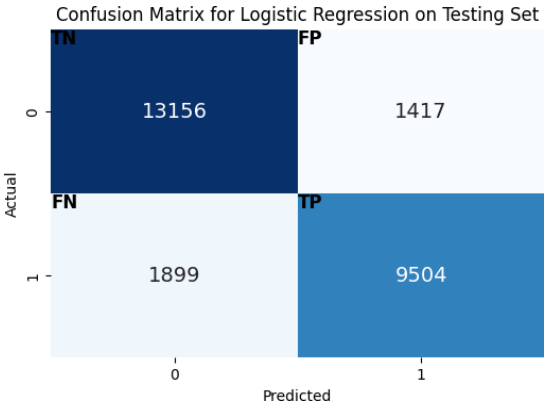
Hyperparameters:

C-0.003, Solver : Saga, Penalty: L1, Tolerance 0.005

Accuracy: 87.23 %

Precision: 87.02

Recall: 83.34



4.5 Random Forest

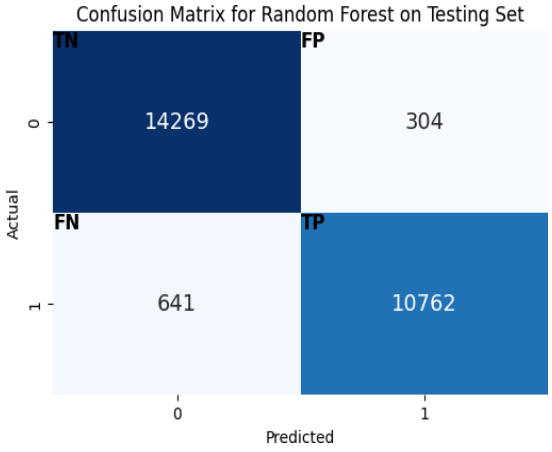
Hyperparameters:

Max Depth 25, Min Sample leaf: 1, Min Samples Lift: 5

Accuracy: 96.36%

Precision: 97.25

Recall: 94.38



5. Beyond the scope:

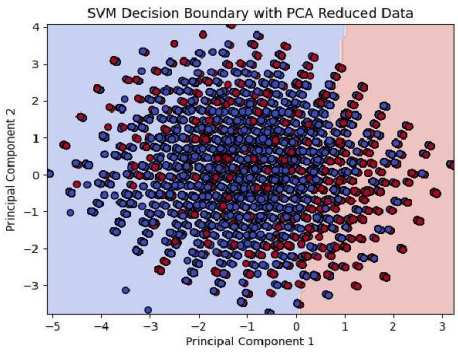
5.1 Feature Engineering:

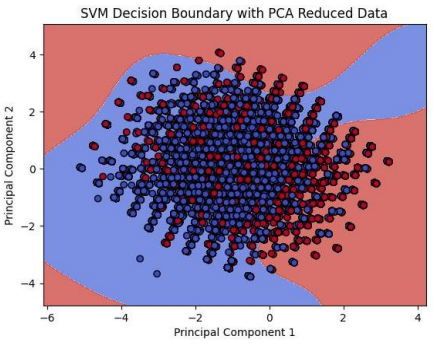
1. **Skewness in Data**:
    - Both **Flight Distance** and **Departure Delay** showed **right-skewed distributions**, with a concentration of smaller values and a long tail of higher values.
    - Skewed data can negatively impact model performance, particularly for algorithms sensitive to feature scaling and distribution, such as linear regression or logistic regression.
2. **Negative Values**:
    - The presence of **negative values** in **Departure Delay** is problematic, as delays typically cannot be negative in practice. Similarly, negative values for **Flight Distance** are unrealistic.
    - These anomalies may be due to data entry errors or inconsistencies in data collection and need to be addressed to ensure accurate modeling.

To handle the skewness and standardize these features, we applied a **logarithmic transformation** which is widely used to reduce the impact of outliers and normalize skewed data. However, since log transformations cannot handle negative or zero values, we added a constant (e.g., |min_value| + 1) to shift all data points into the positive range.

5.2 PCA

By reducing the five features to two principal components while preserving 72.12% of the variance, we achieved a meaningful simplification of the dataset, making it more suitable for visualization. This dimensionality reduction enables a clearer graphical representation of the data's structure and class separation, which is particularly useful for analyzing and understanding the behavior of your SVM classifier.





5.3 Ensemble Method

In our analysis, we leveraged an ensemble method to improve the performance and robustness of our predictive model. Initially, we trained individual models using Decision Tree, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, and Logistic Regression algorithms, each of which provided valuable insights but exhibited varying levels of accuracy and generalizability. To harness the strengths of these models, we implemented an ensemble approach, specifically stacking, where the predictions from each base model were combined using a meta-model to make the final prediction. This technique allowed us to capitalize on the diversity of the base models, reducing variance and bias while improving overall predictive accuracy. The ensemble model outperformed the individual models, demonstrating the power of combining multiple learning algorithms to achieve superior performance in a complex classification task.

6. Model Evaluation

Model comparison table

Model	Mean Validation accuracy	Training accuracy	Testing accuracy
Decision Tree	94.14%	94.71%	94.28%
Logistic	87.49%	87.51%	87.23%
Linear SVM	84.15%	84.15%	83.94%
Kernel SVM	88.73%	88.73%	88.72%
KNN	91.02%	92.78%	91.04%
Ensemble - Random Forest	96.22%	99.41%	96.36%
Ensemble - Voting Classifier	96.95%	96.98%	95.03%

From the results, the Random Forest Ensemble achieves the highest accuracy across all datasets (96.22% mean validation accuracy and 96.36% testing accuracy), making it the best-performing model. However, the training accuracy (99.41%) is significantly higher than the testing accuracy, indicating potential overfitting.

The issue in Random Forest often lies in the maximum depth of trees: overly deep trees capture noise and specific patterns from the training data, leading to reduced generalization. To address this, we would tune the max_depth parameter, balancing model complexity and preventing overfitting, while maintaining high testing accuracy. But on this situation, we believe the Voting Classifier would be the best model we can use now before we are tuning the Random Forest

7. Conclusion

Feature Importance and Hypotheses:

Features such as type of travel (business or personal), online boarding efficiency, in-flight entertainment, seat comfort, and onboard service were identified as significant predictors of passenger satisfaction.

Hypothesis: Investments in enhancing these critical aspects will positively influence passenger satisfaction and loyalty.

Model Performance and Selection:

Random Forest achieved the highest accuracy (96.36%) among the models tested, but its overfitting tendencies indicated room for improvement through parameter tuning.
For practical use, the Voting Classifier emerged as the most balanced approach, providing high accuracy (95.03%) while avoiding overfitting, making it a suitable model for real-world deployment.

Use Case in Real-World Applications:

Personalized Customer Experience: Insights derived from the model can tailor services to passenger preferences, such as providing additional support for economy travelers who prioritize seat comfort or in-flight entertainment.
Proactive Issue Resolution: Predictive insights allow airlines to anticipate dissatisfaction, addressing issues such as delays or subpar service proactively.
Loyalty and Retention Management: Strategies can focus on high-impact areas like online boarding and onboard service to enhance loyalty programs.
Enhanced Marketing and Sales Strategies: Airlines can refine their marketing campaigns by targeting key segments, such as loyal customers and business travelers.
Operational and Staffing Optimization: Resource allocation can be improved by understanding factors contributing to dissatisfaction, such as staff availability during peak travel times.

Conclusion Summary

By leveraging machine learning models to analyze passenger satisfaction, airlines can implement data-driven strategies to improve customer experiences, operational efficiencies, and ultimately drive revenue growth. While Random Forest demonstrated the highest potential, the Voting Classifier serves as a pragmatic choice for deployment, balancing accuracy with generalizability. Future enhancements may include further tuning and integrating additional data sources to refine predictions.