Airline Satisfaction - Classification

1. Problem Statement

The Airline Company seeks to identify factors influencing client satisfaction to predict customer sentiment accurately. They would aim to tailor personalized offers based on individual preferences and ratings, enhancing customer retention by addressing needs effectively and fostering loyalty.

2. Proposed Solution

This paper aims to identify the optimal supervised machine learning model and its tuned hyperparameters for accurately predicting customer satisfaction as “satisfied” or “dissatisfied.” Additionally, an unsupervised clustering technique will segment customers into four distinct clusters based on key attributes, enabling a deeper understanding of customer profiles and preferences. This dual approach would help in developing targeted strategies for enhancing customer engagement and satisfaction.

3. Exploratory Data Analysis

3.1 Understanding the Data

Data Source: https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

The data set has 23 features, and close to 26,000 instances of data. This is a good set for most model prediction and building. This involves a mix of 4 categorical values, and 18 were numeric.

3.2 Data Cleaning

The data was cleaned for null values – 83 in total. The rows with null values, were simply dropped as they represented less than 0.1% of the total dataset, making the drop insignificant for model performance.

3.3 Data Exploration

3.3.1 Target Analysis

The target predictor was mostly balanced, with close to 12k instances of “satisfied” and 14k instances of “dissatisfied or neutral” prediction, therefore no advanced techniques of SMOTE or other balancing methods, needed to be applied.

3.3.2 Univariate Analysis

I first examined the target continuous numerical. Age distribution was almost normally distributed, with most of the values falling between the age of 25 to 60. The flight distance distribution showcased that, for most passengers the data was right skewed, with most values shown for lower distances (less than 1000kms). There were only a few instances of departure delay, with most having 0 (more than 99% of the dataset).

When looked at the discrete numerical data analysis, it was found that, most data were well distributed with some skews noted. Overall, most ratings were between 3 to 4.

3.3.3 Multivariate Analysis

There were some variables, which showcased high correlation. For dimension reduction, I checked variables with a high correlation and dropped values which showcased that they were not a good predictor. This dropping was made by checking feature distribution with the help of a box-plot graph and scatter plots for some. As a rule, I have performed this analysis for all variable pairs, with a correlation of more than 40%.

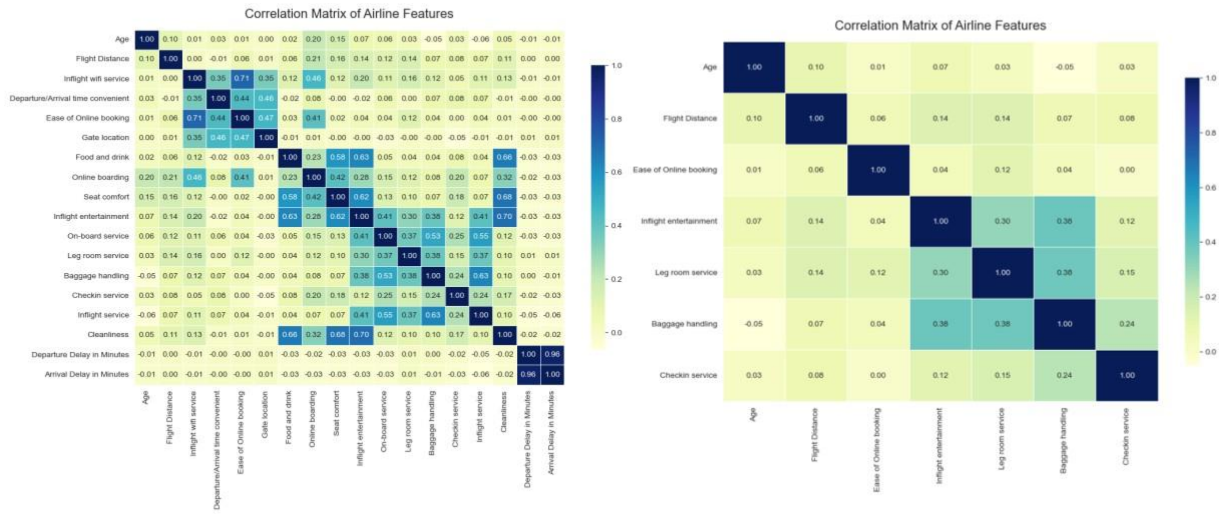


Figure 1 Correlation Matrix, before (left) and after(right) feature selection and reduction.

3.3.4 Feature Encoding

For the use of model building, I have performed dummy encoding for all categorical features. I have made sure to drop the first column for each to prevent high correlation. This resulted in the final encoded data frame to be generated. This encoded data frame has a total of 12 features. Post this, label encoding was performed for the target variable.

4. Supervised Models Evaluation and Interpretation

4.1 Building Models

The Models were built by first taking the data and scaling it. Once the data was scaled, the dataset was divided for a train and test split of 80:20. The training set was then further refined with the help of cross validation implementation with 5 folds. This was again paired with a grid-search method to obtain the optimum parameters.

Once the optimum parameters were found for each model, those values were used to create new models final models and used for prediction on the test set. This was followed by appropriate metrics evaluation such as accuracy, f1, precision, recall, and ROC Curve.

4.2 Comparative Analysis and Model Selection

4.2.1 Comparison Analysis of All Models

In this section, I have consolidate the performance metrics of all evaluated models. The primary focus is on accuracy as a measure of overall correctness, but we also consider F1, precision, and recall to understand the balance between identifying satisfied customers (precision) and not missing them (recall). Selecting the ideal model has direct business implications: an airline aiming to improve customer satisfaction must identify dissatisfied passengers accurately to implement interventions, while correctly recognizing satisfied customers helps validate successful initiatives.

Summary of Model Performance:

Model	Accuracy	F1	Precision	Recall	AUC for ROC Curve
Decision Tree	91.67	0.90	0.91	0.90	0.98
Random Forest	90.75	0.89	0.91	0.88	0.98
Bagging	90.81	0.89	0.91	0.88	0.97
Boosting (AdaBoost)	90.8	0.88	0.91	0.86	0.97
Gaussian Naive Bayes (GNB)	81.08	0.78	0.79	0.76	0.88
Linear Discriminant (LDA)	82.72	0.80	0.80	0.79	0.89
Quadratic Discriminant (QDA)	82.85	0.80	0.80	0.79	0.90
K-Nearest Neighbours (KNN)	90.75	0.89	0.91	0.88	0.96

While F1, precision, and recall were discussed qualitatively for GNB and QDA, their primary reported metric was accuracy. Both models underperformed relative to tree-based and KNN models, making them less ideal for final selection.

Analysis:

The Decision Tree model, after careful tuning, achieved the highest accuracy (91.67%), outperforming all other models. It also maintained strong F1, precision, and recall scores, indicating that it successfully identifies satisfied customers with minimal misclassifications. The ensemble methods (Random Forest, Bagging, Boosting) and the KNN model offered robust alternatives, but none matched the Decision Tree's top accuracy. While they demonstrate resilience and a balanced trade-off between precision and recall, their slight drop in accuracy could mean missing opportunities to correctly classify customers—an essential factor for strategic business decisions.

Gaussian Naive Bayes and Quadratic Discriminant Analysis, though faster and simpler, lagged significantly behind in accuracy (around 81–83%) and thus present a lower value proposition from a business perspective. Their lower predictive performance may lead to more misclassifications, reducing the reliability of insights derived from them.

4.2.2 Choosing the Best Model

In an airline industry context, identifying passenger satisfaction accurately is critical for targeting improvement initiatives, personalizing services, and maintaining a competitive edge. A model with superior accuracy ensures fewer errors in classifying satisfaction levels, enabling the airline to take targeted actions to enhance the travel experience.

The selection criterion prioritizes accuracy as the primary measure. We only consider trading a minimal accuracy drop if there were a substantial gain in F1 (indicating a far better balance between precision and recall). However, in this analysis, the Decision Tree model not only leads in accuracy but also maintains high precision and recall. This negates the need to consider a model with lower accuracy for a marginally improved F1.

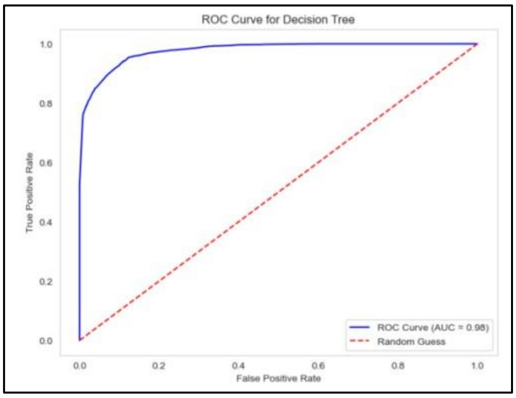
Selected Model: Decision Tree

Rationale:

Highest Accuracy (91.67%): Ensures that the maximum number of passengers are correctly classified, minimizing both false positives and false negatives.

Strong Balanced Metrics (High F1, Precision, and Recall): Confirms the model's consistent performance in identifying both satisfied and dissatisfied passengers.

Strong Area under curve for ROC: When looking at the ROC curve, it was close to the top-left corner, indicating high True Positive Rates (TPR) and low False Positive Rates (FPR) across thresholds. The AUC (Area Under the Curve) value of 0.98 confirms a very strong ability to distinguish between classes, meaning the model is highly effective in classification tasks.



Business Impact: Reliable predictions allow the airline to confidently trust the model's classifications. This can guide effective loyalty programs for satisfied customers and targeted interventions for those showing dissatisfaction, ultimately improving overall satisfaction ratings and potentially increasing market share.

In conclusion, the Decision Tree model emerges as the best choice, striking an optimal balance between model complexity, interpretability, and predictive prowess. Its superior performance and robust metrics stand to deliver the most tangible business benefits.

4.3 Unsupervised Model with K-means Clustering

4.3.1 Model Architecture

K-Means is an unsupervised machine learning algorithm used for partitioning a dataset into a predefined number of clusters (k). The algorithm iteratively assigns data points to clusters and updates cluster centroids to minimize the Within-Cluster Sum of Squares (WCSS). It works by randomly initialising centroids and then computing nearby distances to map data for each cluster and keep updating the centroid, based on a distance metric.

4.3.2 Finding Optimal Value for K

To find the optimal value for K, there was testing made for 1 to 10 K-values and comparing their silhouette scores and within cluster sum of squares. This resulted in the following graphs:

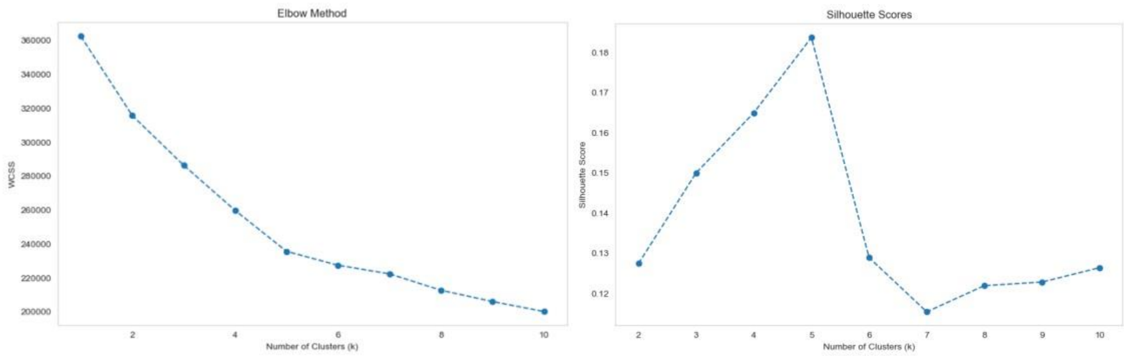


Figure 2 Elbow Diagram and Silhouette scores for K-means hyper parameter tuning

It was found that the optimal value for k is 4. This would result in the highest silhouette score (helps show distance between each cluster) and the result in the WCSS score was not that much. Our goal for K-means is to reduce the amount of within sum of squares and make sure that each cluster is as spread apart from the other as possible.

5 Business Application

A Decision Tree model can be applied for the Airline Company for making predictions on whether a customer will be satisfied or not. This could then result in finding the key features which resulted in their satisfaction and dissatisfaction. The decision tree, based on our model would perform close to a 90% accuracy for the unknown customer data.

The Decision Tree Model plotted also identified key features that are the key decision makers for whether a customer is satisfied or not. These key features identified showcased the following insights:

5.1 Insights from Decision Tree Analysis on Customer Satisfaction

1. Personal Travel Customers Are More Likely to Be Satisfied

Implication: Airlines should enhance their offerings for personal travellers, such as targeted promotions or personalized services, to further improve satisfaction levels in this segment.

2. Disloyal Customers Report Higher Satisfaction

Implication: This indicates that customers who fly with multiple airlines find the current airline's performance better. To capitalize on this, the airline can focus on loyalty programs and incentives that encourage repeat travel while emphasizing its standout features.

3. Check-in Service Is a Critical Factor

Implication: Streamlining the check-in process, such as through digital kiosks or efficient staffing, can significantly improve satisfaction. Airlines should consider reducing wait times and enhancing the ease of check-in experiences.

4. Younger Customers Are More Satisfied

Implication: Younger travellers may appreciate modern conveniences or affordability. Airlines can tailor marketing campaigns and onboard experiences to maintain their loyalty, such as Wi-Fi services or eco-friendly initiatives.

5. Legroom Is a Key Feature for Satisfaction

Implication: Since many customers value legroom, especially in economy class, airlines should consider optimizing cabin layouts to offer more space or advertising flights with superior legroom to attract this demographic.

6. Inflight Entertainment and Baggage Handling Are Less Influential

Implication: While important, these features may not be primary drivers of satisfaction. Airlines can allocate resources strategically, focusing on areas that directly impact satisfaction while maintaining adequate standards for these services.

7. Long-Distance Travelers Tend to Be Less Satisfied

Implication: To improve satisfaction among long-haul travellers, airlines can focus on comfort and convenience, such as better seating, meal quality, and in-flight amenities. Addressing pain points in long-distance travel can increase overall satisfaction.

5.2 Cluster Descriptions for Airline Satisfaction Analysis

The clustering performed with the help of k-means clustering helped us find the following insights and make the appropriate grouping:

Cluster 0: Premium Frequent Flyers

These customers are generally middle-aged (average age: 44.1) and frequently fly long distances (average flight distance: 1888.53 km). They rate inflight entertainment, legroom, baggage handling, and inflight service highly (all above 4.3), indicating high satisfaction with premium services. The near absence of personal travel and eco-class preferences suggests they are mostly loyal business travellers. Their arrival delays are minimal, reflecting efficient travel experiences.

Cluster 1: Occasional Personal Travelers
This group comprises younger travellers (average age: 37.11) with shorter flight distances (average 758.23 km). They primarily travel for personal reasons (90%) and mostly choose economy class (90%). Ratings for services like inflight entertainment and legroom are moderate (around 3.2), indicating a focus on cost over luxury. They face slightly higher delays (14.76 minutes) but maintain overall satisfaction.

Cluster 2: Disloyal Economy Travelers

Primarily composed of younger customers (average age: 29.9) who are disloyal (100%) and predominantly travel in economy class (57%). Their flight distances are short (706.04 km), and service ratings are average (around 3.1-3.7). This group likely includes budget-conscious or first-time travellers. Arrival delays are slightly higher at 15 minutes, potentially impacting satisfaction.

Cluster 3: Dissatisfied Middle-Aged Flyers

This cluster consists of middle-aged travellers (average age: 44.1) who exhibit the lowest satisfaction levels across all service categories (ratings between 2.38 and 2.72). They take medium-length flights (1173.51 km) and have the highest arrival delays (17.73 minutes). The mix of personal and non-personal travel suggests they may lack clear loyalty to the airline, with dissatisfaction stemming from poor service experiences.

5.3 Limitation for the model:

The model will only perform this well, if we are expected to see a similar set of result for our outcome/new dataset as well. For example, we saw that most flyers flew for shorter distances in the EDA. If a similar pattern is observed for the new user data, the model would most likely perform well in those cases only. The model was not tested on isolated data, or outlier focused dataset.