Predicting US Airline Passenger Satisfaction: A Data-Driven Approach.

Hrayr Muradyan, Azin Piran, Sopuruchi Chisom, Shengjia Yu.

2024-12-07

Table of contents

Summary

In this project, we created a classification model that uses a decision tree algorithm to predict airline customers satisfaction based on a variety of factors such as in-flight service quality, seat comfort, and demographic information. Customers were classified as either satisfied (positive ratings) or neutral/dissatisfied (negative ratings). Our final decision tree model performed well on an unknown test dataset, with an overall accuracy of 95.02%, a recall of 92.61%, a precision of 95.86%, and an F1-score of 94.21%.Based on the confusion matrix (Figure 1), the model correctly identified 24,683 cases while misclassifying 1,293 instances. There were 839 false negatives (satisfied consumers categorized as indifferent or dissatisfied), and 454 false positives (neutral or unsatisfied customers classified as satisfied). The imbalance between false negatives and false positives shows that the model favours overall accuracy while having slightly lesser sensitivity for spotting satisfied consumers.These findings show that the decision tree model accurately detects consumer satisfaction patterns, making it a useful tool for examining key aspects that influence satisfaction. However, the relatively high frequency of

false negatives may lead to an underestimation of satisfied consumers, potentially leading in ineffective compensatory measures. Similarly, false positives may result in consumer unhappiness being disregarded. Despite these limitations, the model provides a solid foundation for customer satisfaction analysis and could be used as an effective decision-making tool by airlines. Future work should strive to improve the model's sensitivity and specificity while also validating its performance across a wide range of customer segments, aircraft routes, and service circumstances.

Introduction

In the highly competitive field of air transport management, passenger satisfaction plays a critical role in making customer loyalty, providing operational insights, enhancing financial performance, and ensuring compliance with regulations and rankings (Sadegh Eshaghi et al, 2024). While there are numerous studies held on factors influencing customer satisfaction like service quality (???Namukasa, 2023), it is very important to be able to predict the customer satisfaction with high accuracy for understanding how to improve and make better decisions. In this study, we aim to create a reliable predictive model that will predict US airline passenger satisfaction with high performance.

Methods

Dataset Overview

The dataset includes a variety of features such as: - **Flight Information**: Flight distance, departure delay, arrival delay, etc. - **Passenger Demographics**: Age, gender, travel class, etc. - **Flight Service Quality**: Satisfaction ratings on aspects such as in-flight entertainment, seat comfort, and food quality.

The goal is to predict the **passenger satisfaction** (either satisfied or dissatisfied) based on these features.

Data Exploration and Preprocessing

Target Variable Analysis

The target variable, 'satisfaction', consists of two categories: 'neutral or dissatisfied' and 'satisfied'. The distribution of this target variable was checked to assess class balance:

- **Class Distribution**: The ‘satisfaction’ column is slightly imbalanced, with 56.5% of the observations labeled as ‘neutral or dissatisfied’ and 43.5% as ‘satisfied’. This is considered a minor imbalance and should not significantly affect model performance.
- **Data Type and Category Check**: The ‘satisfaction’ column was confirmed to be of type `object` ???, indicating that it is categorical. There are only two unique categories in this column, as expected.

???? we should have a graph here

Data Splitting

To ensure that the model is evaluated on unseen data, the dataset was randomly split into training and testing subsets:

- **Train-Test Split**: The data was split into 80% training and 20% testing, yielding 103,904 observations in the training set and 25,976 observations in the test set. The split was done using a random seed to ensure reproducibility.

Exploratory Data Analysis (EDA)

The purpose of this Exploratory Data Analysis (EDA) is to gain a deeper understanding of the Airline Passenger Satisfaction dataset, identify potential issues, and uncover insights that could influence passenger satisfaction. Below is a summarized analysis of key findings from the EDA process.

Distribution of Numerical Variables

Histograms were plotted for all numerical features, revealing that most of the numerical variables, except for age, are right-skewed. This suggests that these features do not follow a normal distribution and may require transformation. ???

Correlation Analysis

A correlation matrix was computed for numerical variables to check for multicollinearity. Notably, `departure_delay_in_minutes` and `arrival_delay_in_minutes` were found to be highly correlated, suggesting redundancy. One of these features can be dropped to avoid multicollinearity.

A heatmap visualization of the correlation matrix highlighted these relationships: ???

Feature-Target Relationships

- **Numerical Features**: Density plots were generated to examine how numerical features correlate with the target variable `satisfaction`. These plots indicated that no single feature strongly separated the two classes, suggesting that complex interactions between features may be necessary for predictive modeling. ???
- **Ordinal Features**: Count plots were used to analyze ordinal features, such as `seat_comfort`, `on_board_service`, and `inflight_entertainment`. These features showed very few observations with values of 0, which will need to be addressed during preprocessing.

The EDA revealed that the target variable is slightly imbalanced, while many numerical variables are right-skewed and may require transformations. There is high correlation between certain features, suggesting redundancy, and no single feature strongly separates the target variable, indicating the need for more complex models. Missing values were addressed by removing redundant columns, and the dataset is now ready for further preprocessing and modeling.

Preprocessing

A feature was removed due to missing values and high correlation with another feature, making it redundant.

2. **Encoding Categorical Variables**
   Categorical variables were encoded for machine learning algorithms. One feature was converted into a binary variable, and other categorical variables were one-hot encoded to transform them into numerical format.

3. **Scaling Numerical Features**
   Numerical features were standardized or scaled to ensure they have similar ranges. This helps prevent dominance of variables with larger ranges in machine learning models. Different scaling methods were applied based on the type of data:

   - One-hot encoding for categorical variables.
   - Min-Max scaling for ordinal variables.
   - Standard scaling for numerical variables.

4. **Column Transformer Setup**
   A `ColumnTransformer` was used to apply appropriate transformations to each feature type, ensuring the data is ready for machine learning models. Categorical variables were one-hot encoded, ordinal variables were min-max scaled, and numerical variables were standardized.

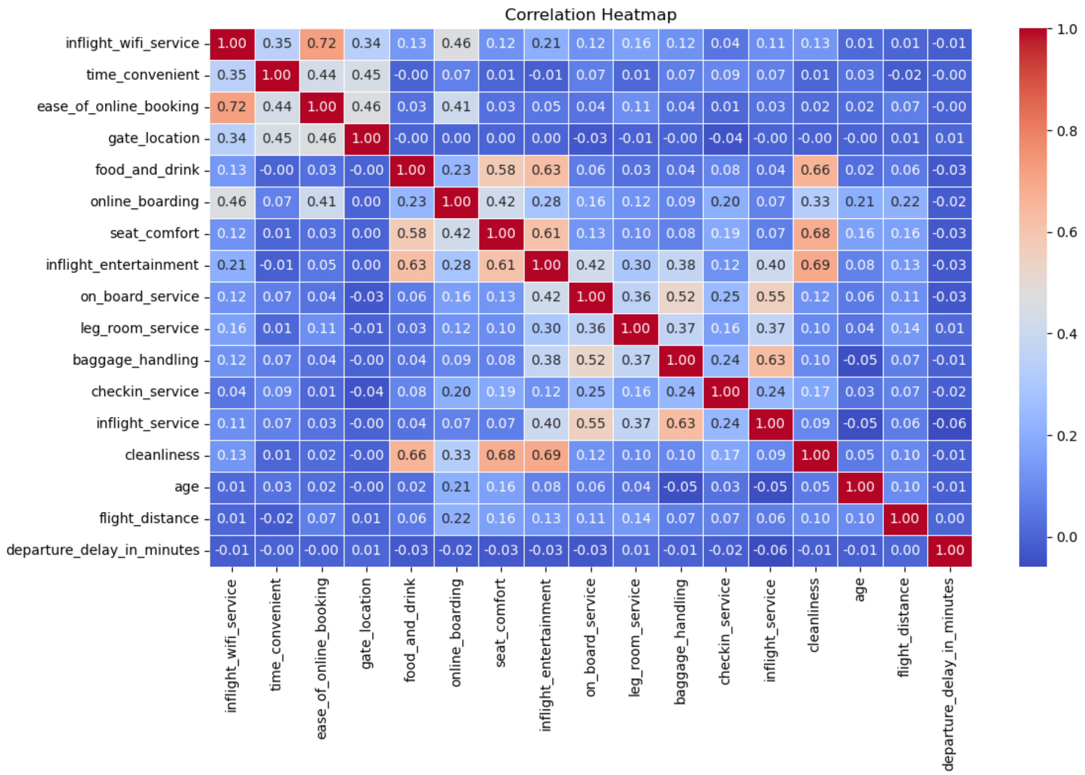
This preprocessing pipeline ensures the dataset is properly prepared for analysis and modeling, with consistent scaling and encoding applied across the features.

Analysis

Results and Discussion

To determine each feature's role in forecasting consumer satisfaction, we examined its correlations with the target variable. No features showed significant correlation with satisfaction(Figure 2), which might result in near-perfect class separation, implying that satisfaction is impacted by numerous variables rather than just one. Many features were kept in the model because they provided unique information. However, multicollinearity was found between arrival_delay_in_minutes and departure_delay_in_minutes (correlation coefficient = 0.97, Figure 1). To solve this, we removed the arrival_delay_in_minutes feature, as both perhaps provided the same information.We chose the Decision Tree classifier since it outperformed alternative models, including Dummy and Logistic Regression, on all validation measures. The Decision Tree has the best validation_accuracy (94.5%) and validation_f1 (0.945), demonstrating an excellent balance of precision and recall.To further improve the decision tree, we used grid search with 10-fold cross-validation to optimize the max_depth hyperparameter, testing depths of 10, 12, 15, and 18.The ideal depth of 15 resulted in a mean cross-validation accuracy of 95.2%, which balanced model complexity and performance. While Decision Trees are prone to overfitting, rigorous hyperparameter tweaking reduced this problem, providing strong generalization on test data.

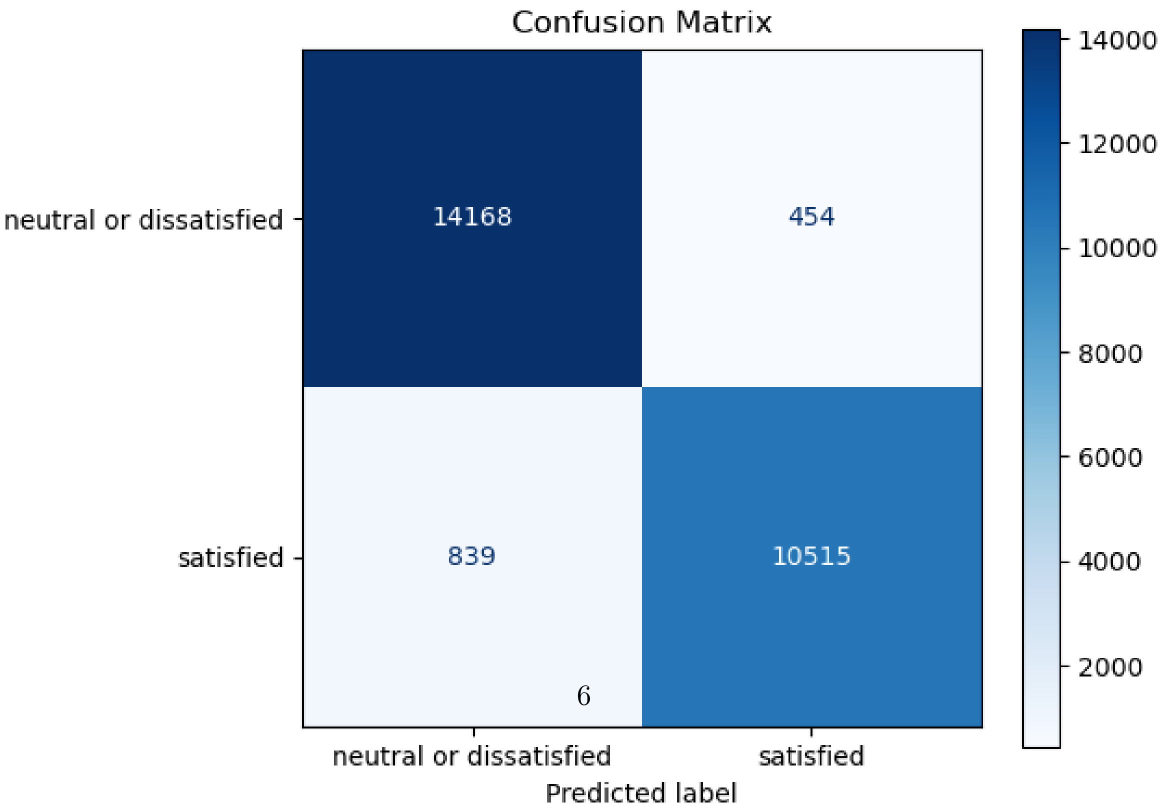
The resulting decision tree model performed well on the test dataset, with an overall accuracy of 95.0% and an F1-score of 94.2%. The confusion matrix (Figure 3) showed that the model successfully identified 24,683 of 25,976 samples. However, there were 839 false positives (satisfied customers misclassified as dissatisfied) and 454 false negatives (unsatisfied customers misclassified as satisfied). While the model is effective at identifying pleased customers, decreasing false negatives remains a top focus since they represent missed chances to boost customer retention.In the future, analyzing misclassifications may uncover trends or interactions that were previously undetected, leading to improved model performance. Testing ensemble approaches, such as Random Forest or Gradient Boosting, could help reduce overfitting and better leverage feature interactions. Including probabilistic outputs would allow stakeholders to assess prediction confidence levels, enabling more targeted interventions for high-risk consumers. Adjusting class weights in ensemble models could further minimize bias.



Correlation Heatmap

	inflight_wifi_service	time_convenient	ease_of_online_booking	gate_location	food_and_drink	online_boarding	seat_comfort	inflight_entertainment	on_board_service	leg_room_service	baggage_handling	checkin_service	inflight_service	cleanliness	age	flight_distance	departure_delay_in_minutes
inflight_wifi_service	1.00	0.35	0.72	0.34	0.13	0.46	0.12	0.21	0.12	0.16	0.12	0.04	0.11	0.13	0.01	0.01	-0.01
time_convenient	0.35	1.00	0.44	0.45	-0.00	0.07	0.01	-0.01	0.07	0.01	0.07	0.09	0.07	0.01	0.03	-0.02	-0.00
ease_of_online_booking	0.72	0.44	1.00	0.46	0.03	0.41	0.03	0.05	0.04	0.11	0.04	0.01	0.03	0.02	0.02	0.07	-0.00
gate_location	0.34	0.45	0.46	1.00	-0.00	0.00	0.00	0.00	-0.03	-0.01	-0.00	-0.04	-0.00	-0.00	-0.00	0.01	0.01
food_and_drink	0.13	-0.00	0.03	-0.00	1.00	0.23	0.58	0.63	0.06	0.03	0.04	0.08	0.04	0.66	0.02	0.06	-0.03
online_boarding	0.46	0.07	0.41	0.00	0.23	1.00	0.42	0.28	0.16	0.12	0.09	0.20	0.07	0.33	0.21	0.22	-0.02
seat_comfort	0.12	0.01	0.03	0.00	0.58	0.42	1.00	0.61	0.13	0.10	0.08	0.19	0.07	0.68	0.16	0.16	-0.03
inflight_entertainment	0.21	-0.01	0.05	0.00	0.63	0.28	0.61	1.00	0.42	0.30	0.38	0.12	0.40	0.69	0.08	0.13	-0.03
on_board_service	0.12	0.07	0.04	-0.03	0.06	0.16	0.13	0.42	1.00	0.36	0.52	0.25	0.55	0.12	0.06	0.11	-0.03
leg_room_service	0.16	0.01	0.11	-0.01	0.03	0.12	0.10	0.30	0.36	1.00	0.37	0.16	0.37	0.10	0.04	0.14	0.01
baggage_handling	0.12	0.07	0.04	-0.00	0.04	0.09	0.08	0.38	0.52	0.37	1.00	0.24	0.63	0.10	-0.05	0.07	-0.01
checkin_service	0.04	0.09	0.01	-0.04	0.08	0.20	0.19	0.12	0.25	0.16	0.24	1.00	0.24	0.17	0.03	0.07	-0.02
inflight_service	0.11	0.07	0.03	-0.00	0.04	0.07	0.07	0.40	0.55	0.37	0.63	0.24	1.00	0.09	-0.05	0.06	-0.06
cleanliness	0.13	0.01	0.02	-0.00	0.66	0.33	0.68	0.69	0.12	0.10	0.10	0.17	0.09	1.00	0.05	0.10	-0.01
age	0.01	0.03	0.02	-0.00	0.02	0.21	0.16	0.08	0.06	0.04	-0.05	0.03	-0.05	0.05	1.00	0.10	-0.01
flight_distance	0.01	-0.02	0.07	0.01	0.06	0.22	0.16	0.13	0.11	0.14	0.07	0.07	0.06	0.10	0.10	1.00	0.00
departure_delay_in_minutes	-0.01	-0.00	-0.00	0.01	-0.03	-0.02	-0.03	-0.03	-0.03	0.01	-0.01	-0.02	-0.06	-0.01	-0.01	0.00	1.00

Figure 1: Heatmap of Correlations Between Predictors for Airline Customer Satisfaction Dataset



Confusion Matrix

	neutral or dissatisfied	satisfied
neutral or dissatisfied	14168	454
satisfied	839	10515

References

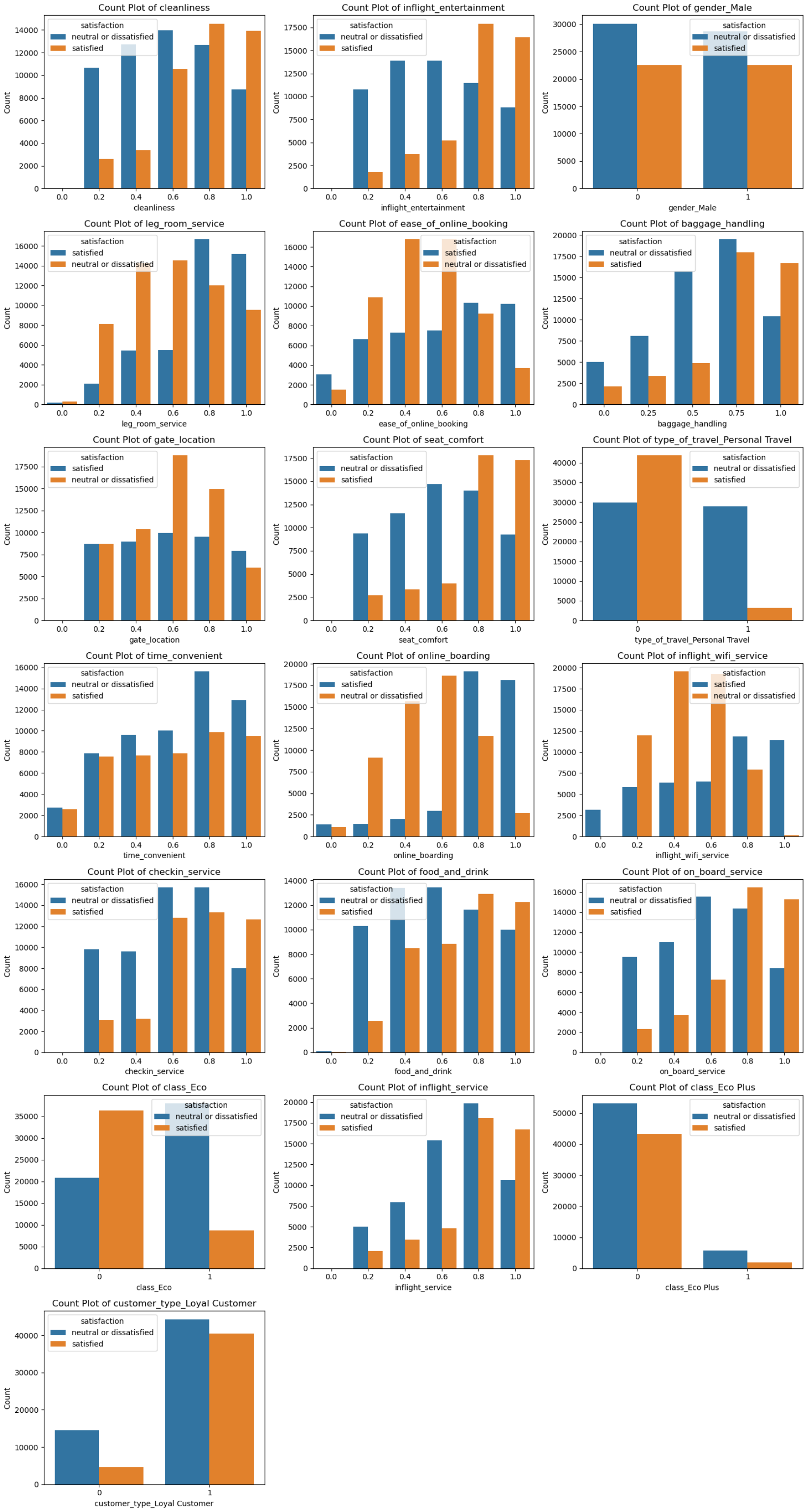


Figure 2: Distribution of Categorical Features Across Customer Satisfaction Levels