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

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

2024-12-07

Table of contents

Predicting US Airline Passenger Satisfaction: A Data-Driven Approach

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.

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

Data Validation

To ensure the integrity and quality of the dataset, a series of validation steps were performed:

- **Correct Data File Format**: The data files were loaded in `.csv` format, compatible with Python's pandas library. The training and test datasets were successfully merged, with no inconsistencies in file format. The combined dataset consists of 129,880 rows and 24 columns. ???
- **Correct Column Names**: The column names were standardized by converting them to lowercase and replacing spaces or dashes with underscores for consistency. A specific renaming step was applied to the 'departure/arrival_time_convenient' column, renaming it to 'time_convenient'.
- **Empty Observations**: The dataset was checked for empty rows using a schema validation function. No empty rows were found, ensuring data completeness.
- **Missingness**: Missing data was examined for each column. A threshold of 5% missing data per column was established. The 'arrival_delay_in_minutes' column had a minor amount of missing data (approximately 0.3%), which is within the acceptable range.
- **Correct Data Types**: Each column's data type was validated to ensure consistency. All columns had the expected data types (e.g., integer for numerical columns, string for categorical variables). The target variable, 'satisfaction', was confirmed as a categorical variable.
- **No Duplicate Observations**: Duplicates were checked based on the unique 'id' column and across all columns. No duplicates were found, indicating that each observation is unique.

- **Outliers and Anomalous Values**: Each numerical column was checked for outliers and anomalous values. Variables such as 'flight_distance', 'age', and 'delay times' were inspected, with no extreme outliers detected. Range constraints for certain features were defined (e.g., 'age' should be between 0 and 100, 'flight_distance' should be positive).
- **Correct Category Levels**: The categorical variables, such as 'customer_type', 'class', and 'satisfaction', were validated to ensure no incorrect category levels. The 'customer_type' column had a case mismatch ("disloyal customer"), which was corrected to "Disloyal Customer" for consistency.

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.

Modeling

Results and Interpretation

Conclusion

Summary

Limitations and Future Work

References

List any research papers, articles, or sources that you referred to while doing the analysis.