CS 334 Final Project
May 7th, 2024

Improving Airline Customer Satisfaction: Identification of Influential Factors Based on Decision Tree and Logistic Regression

Tung Dinh, Anji Ni
Emory University

Abstract

In the competitive airline industry, ensuring customer satisfaction is crucial for growth and profitability. This study uses machine learning models, specifically decision trees and logistic regression, to pinpoint key factors that affect passenger satisfaction. We analyze the Airline Passenger Satisfaction dataset, which includes a variety of demographic, travel, and service evaluation metrics, to predict how satisfied passengers will be. Our models undergo thorough hyperparameter tuning to optimize their accuracy, and are evaluated using metrics such as the F1 score, AUC, and AUPRC to confirm their reliability. Unlike conventional methods that rely on initial feature selection, our approach directly trains models and identifies the most impactful features through the best estimator's results. Additionally, we perform subgroup analysis to examine how these factors vary among different groups of passengers, defined by seat class and travel purpose, offering specific insights to enhance service delivery. This methodology not only simplifies the analysis process but also provides detailed insights into the drivers of passenger satisfaction.

Introduction

Customer satisfaction is vital for maintaining loyalty and securing a competitive edge in the airline industry. As airlines aim to improve passenger experiences, understanding the detailed factors that contribute to satisfaction is crucial. This study tackles the challenge of pinpointing these factors using predictive analytics, a technique still emerging in real-time airline decision-making. Our approach avoids traditional preliminary feature selection, instead opting for direct model training followed by the extraction of feature importance from the best-performing estimator. This method not only simplifies the modeling process but also enhances the accuracy of identifying key influencers of satisfaction. The significance of this research is its capacity to provide insights that airlines can leverage to customize their services according to the specific preferences of different passenger segments, ultimately enhancing customer satisfaction.

Background

Advances in predictive analytics have spurred interest in using machine learning to boost customer satisfaction within the airline sector. A prominent study by Habbal typifies the traditional approach to predictive modeling, aiming to find the optimal model for predicting customer satisfaction using a Kaggle dataset.[1] Typically, studies in this topic used multiple machine learning models, including KNN, Logistic Regression, Random Forest, and Naïve Bayes to predict the satisfaction level of flight passengers based a series of features that evaluate their in-light and pre-flight experiences.[2-4] These studies compare the prediction accuracy across different models and aim to find the best-performing one. The current literature, however, mainly interprets their results from exploratory data analysis or results from a single machine learning model. The results can be insightful in helping the airlines to improve their services but may lack comprehensiveness and validity because of multiple influencing factors.

In contrast, our method diverges from this standard model optimization and subsequent feature significance assessment. We directly train decision tree (DT) and logistic regression (LR) models using the extensive Airline Passenger Satisfaction dataset, bypassing initial feature selection to immediately use the best estimator's results for identifying the most significant features. This streamlined approach facilitates a more direct identification of satisfaction drivers. Additionally, our study expands to examine how these factors differ among various passenger subgroups defined by seat class and travel purpose. This targeted subgroup analysis aids in pinpointing critical service elements, providing direct, actionable insights for service improvement in the airline industry.

Methods

Model Description. 1) Decision tree classifier first builds a tree by repetitively splitting the observations in training data based on their values for different features. The algorithm aims to minimize the entropy/gini index of the target labels in each resulting node. Impurity-based feature importance is generated during this training process, which represents the predictive power and magnitude of path influence of a feature. Given a decision tree, a prediction is made by tracking a path according to the feature values of the input observation. 2) Logistic regression classifier is trained by recursively updating the coefficients in a linear combination of the input features to find the optimal coefficient values that minimize a loss function. By applying a sigmoid function, the output can be interpreted as the probability of the target label to be true. The value of the coefficient represents the expected change in the log odds of the target value per unit change in the corresponding feature. Therefore, these coefficients can reflect the association between a feature and the target value. 3) Bagging is an ensemble technique that aggregates the results of different classifiers trained on bootstrapped sub-sample from the training set. In this experiment, bagging is applied to both Decision Tree and Logistic Regression classifiers.

Hyperparameter Tuning. We use GridSearchCV for hyperparameter tuning to optimize each model's performance. The parameters tested and the best parameters selected are as follows:

1) Decision Tree Classifier:
Parameters tested:
Criterion: ['gini', 'entropy']
Max depth: [3, 5, 10, None]
Min samples leaf: [1, 2, 4]
Best parameters: 'criterion': 'gini', 'max_depth': None, 'min_samples_leaf': 4
2) Logistic Regression:
Parameters tested:
C (Regularization strength): [0.01, 0.1, 1, 10, 100]
Solver: ['liblinear', 'lbfgs']

Penalty: ['l1', 'l2']
Class weight: [None, 'balanced']
Best parameters: 'C': 0.01, 'class_weight': None, 'penalty': 'l1', 'solver': 'liblinear'

Experiment Procedures

Introduction to the dataset. Our data is obtained from a well-structured dataset from Kaggle, which is specifically tailored for predicting airline passenger satisfaction. The dataset is pre-split into training and testing sets, containing 104,000 and 26,000 entries respectively, across 26 distinct attributes. These attributes include a mix of categorical and numerical data types that describe various aspects of a passenger's flight experience. Categorical attributes include Gender (Female, Male), Customer Type (Loyal customer, disloyal customer), Type of Travel (Personal Travel, Business Travel), and Class (Business, Eco, Eco Plus). These categorical variables are crucial for capturing the demographic and travel preferences of the passengers. Numerical attributes include Age, Flight Distance, and various satisfaction levels (ranging from 0 for Not Applicable to 5 for maximum satisfaction) for services such as Inflight wifi, Online booking, Seat comfort, Inflight entertainment, and others. Additionally, time-related attributes like Departure Delay in Minutes and Arrival Delay in Minutes are included to assess their impact on overall satisfaction.

Preprocessing. 1)Handling Missing Values: We begin by identifying and removing any records with missing values (N/A) to maintain the integrity of our analysis. 2) Encoding Categorical Variables: To make the categorical data amenable to machine learning algorithms, we apply one-hot encoding. This process transforms categorical variables into a form that could be provided to ML algorithms to do a better job in prediction. 3) Normalizing Numerical Variables: All numerical features are scaled using StandardScaler. This normalization ensures that each feature contributes equally to the analysis and improves the convergence during the model training phase. 4) Target Variable Transformation: The satisfaction attribute, which is the target variable, is converted from a categorical (Satisfaction, neutral, or dissatisfaction) into a binary format where 'neutral or dissatisfied' is mapped to 0 and 'satisfied' is mapped to 1. This binary format is particularly useful for the classification models employed in this study. By meticulously preparing and preprocessing the data, we ensure that the dataset is primed for effective model training and subsequent analysis. This step is crucial as it directly influences the performance and accuracy of the predictive models that follow.

Subgroup division. The observations in the original dataset are classified into six subgroups according to the passenger’s purpose of travel and seat class. The purpose of travel can be either Business or Personal in the dataset, while the seat class may take one value from Business, Economy, and Economy Plus.

Exploratory data analysis. The sub-group training sets possess two potential issues. First, there is data imbalance in the training sets (Figure 1); second, some of the sub-group training sets only have less than 10% of the observations in the original data set. These two problems may lead to overfitting and failure to learn the characteristics of the minority class. Therefore, minority oversampling is applied to all the sub-group training sets, and ensemble method (bagging) is applied to both types of classier.

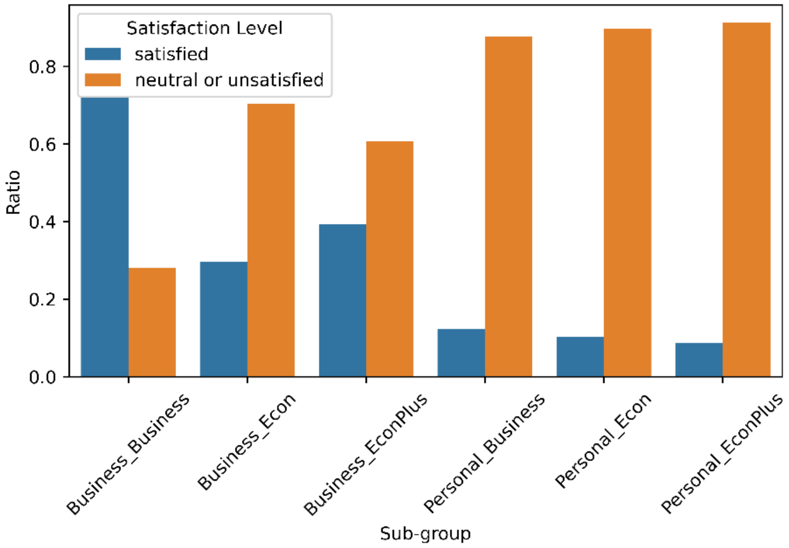

Figure 1. data imbalance of each sub-group

Model Training and Evaluation. An ensemble of 20 DT or LR models (bagging) is trained on the complete dataset and each of the six subgroups respectively using the best parameters identified. Their performances are evaluated using ROC AUC, AUPRC, and F1 Score.

Feature Importance Analysis. Feature importance is obtained from the weight of the feature in the DT or LR model to identify factors that most significantly influence passenger satisfaction. For the Decision Tree, feature importance is represented as the average of the feature weight across 20 classifiers in the ensemble. For Logistic Regression, the averaged values of the coefficients are used as the measure of feature importance.

Results Visualization. We use matplotlib to plot the evaluation results and feature importances, providing clear visual insights into the models' performance and the impact of different features on predictions. The plots facilitate understanding of the contributions of individual features towards the model's decision-making process.

Results and Discussions

The performances of classifiers trained in this experiment are summarized in Table 1 and Table 2. On the original data set, both DT and LR classifier perform well despite a lower F1 score of LR classifier. For the sub-group data, however, the performance of classifiers start to diverge from each other. Overall, all the DT classifiers have acceptable performance, although the F1 scores of the last three sub-groups are lower than that of first three sub-groups; on the other hand, the LR classifiers perform worse than their DT counterparts, and the last three classifiers have F1 score lower than 0.5. A closer examination of the metrics revealed that the AUPRC is significantly lower than AUC for the last three subgroups, suggesting that the model does poorly in terms of precision and recall balance. The evidence implies that the data imbalance is not erased after the adjustment using minority oversampling, which can be caused by overfitting replicated minority class and poor generalizability of the model that follows. Moreover, the trend that LR classifiers perform worse than the DT classifier may suggest that the choice of regularization parameters, which are specific to LR classifier, may adversely influence the performance.

Table 1. Metrics assessing performances of DT and LR trained on complete data set.

	Decision Tree	Logistic Regression
AUC	0.971	0.925
AUPRC	0.956	0.931
F1	0.944	0.851

Table 2. Metrics assessing performances of DT and LR trained on sub-group training sets.

	Decision Tree					
	Business_Business	Business_Econ	Business_EconPlus	Personal_Business	Personal_Econ	Personal_EconPlus
AUC	0.994	0.982	0.980	0.984	0.974	0.987
AUPRC	0.997	0.966	0.970	0.881	0.849	0.908
F1	0.982	0.905	0.921	0.740	0.712	0.785
	Logistic Regression					
	Business_Business	Business_Econ	Business_EconPlus	Personal_Business	Personal_Econ	Personal_EconPlus
AUC	0.938	0.889	0.901	0.751	0.778	0.808
AUPRC	0.978	0.841	0.881	0.567	0.634	0.613
F1	0.901	0.751	0.794	0.319	0.366	0.313

The effect of imbalance data can be demonstrated more straightforwardly by presenting the ROC curves (Figure 2). Since the ROC curve measures the balance between false positive rate (FPR) and true positive rate (TPR) at different threshold values, a preferred binary classifier should have a ROC curve close to the top-left corner of the graph. As represented by Figure 2, all the ROC curves for DT classifiers are closer to the top-left corner compared to LR classifiers, while the ROC curves for the last three groups are obviously flatter. Therefore, due to the validity of the classifiers, we decided to focus our analyses on the first three subgroups as well as the general trend produced by the complete data set.

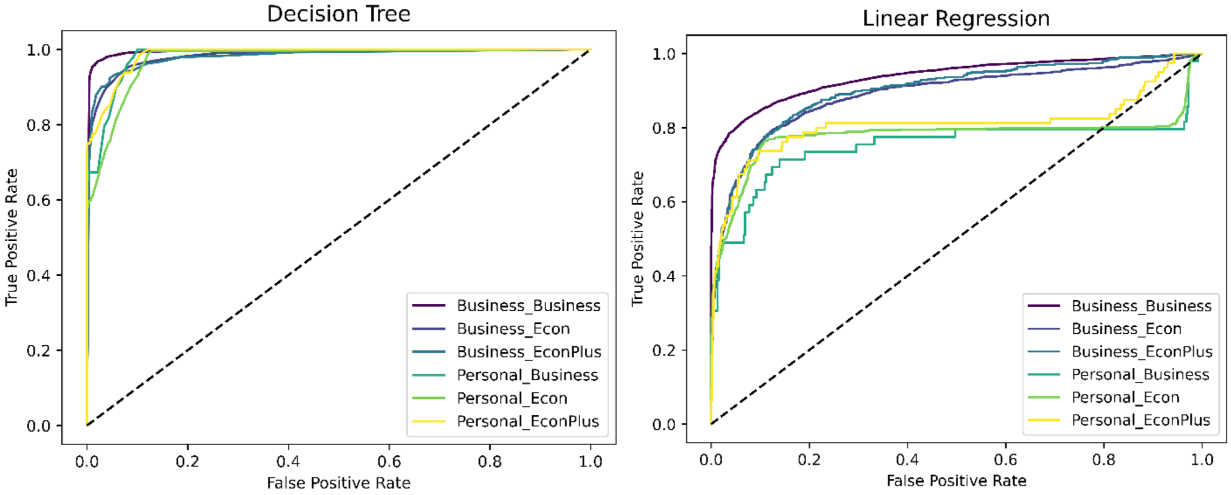

Figure 2. ROC curves of DT and LR classifiers trained on each sub-group data.

According to the feature importances acquired from the DT classifier trained on complete data set (Figure 3), the convenience of pre-boarding services (Online boarding and check-in services) and in-flight entertainments (in-flight Wi-Fi and entertainment systems) predominantly influence the satisfaction of passengers. Although the LR classifier agrees with the DT classifier on the importance of pre-boarding services and in-flight entertainments, it ranks the influence of cleanliness and duration of flight delay significantly higher. It should also be noticed that both the classifiers trained on the complete data set indicate that type of travel and seat class have an impact on the satisfaction level of the passengers, which motivated us to dive deeper into the sub-group analysis.

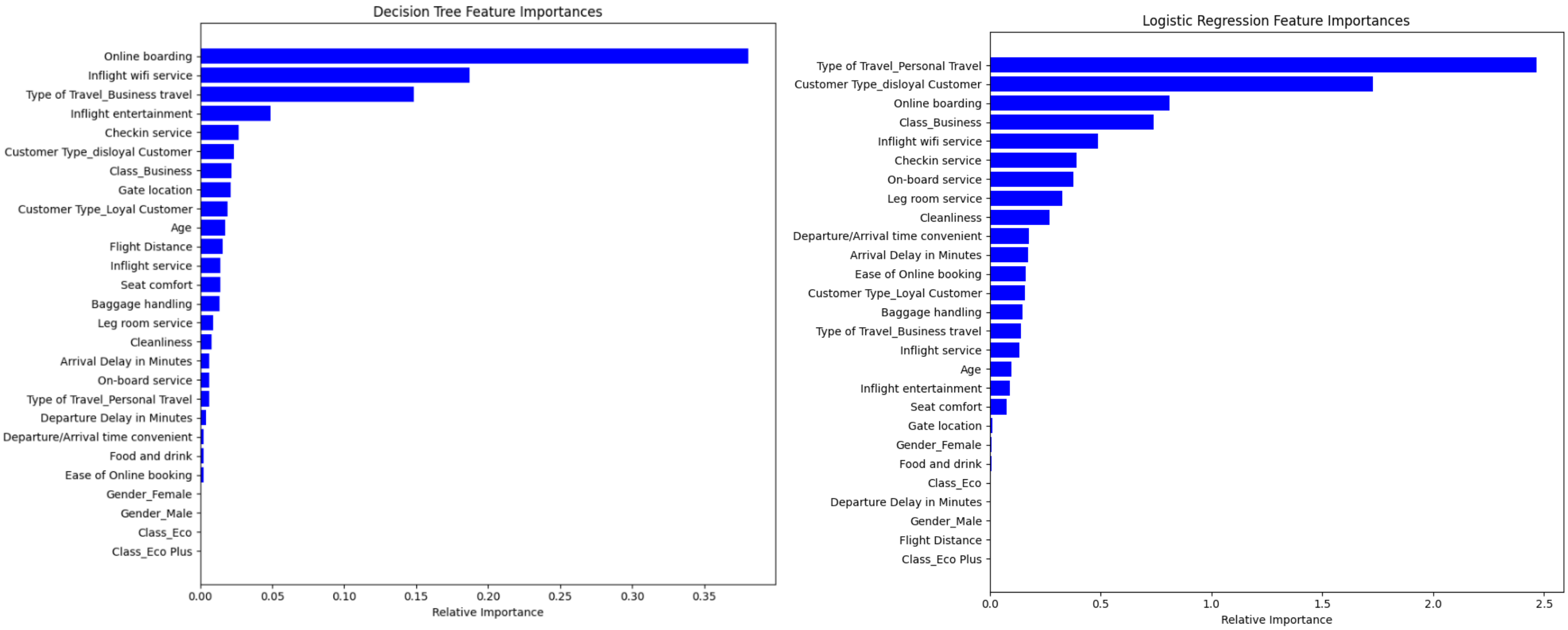

Figure 3. Feature importances acquired from DT and LR classifiers trained on the complete data set.

In Figure 4, group-wise feature importances generated by DT classifiers and LR classifiers are summarized in two heat maps. There are several important observations made from these two graphs. Firstly, from the heat map of DT classifiers, it can be concluded that in-flight service is influential for the satisfaction level across the whole passenger body; on the other hand, passengers who travel for business purposes have a more comprehensive criteria to assess their flight experiences, and the general trend about pre-flight services, cleanliness, and delay time are mainly contributed by these passengers.

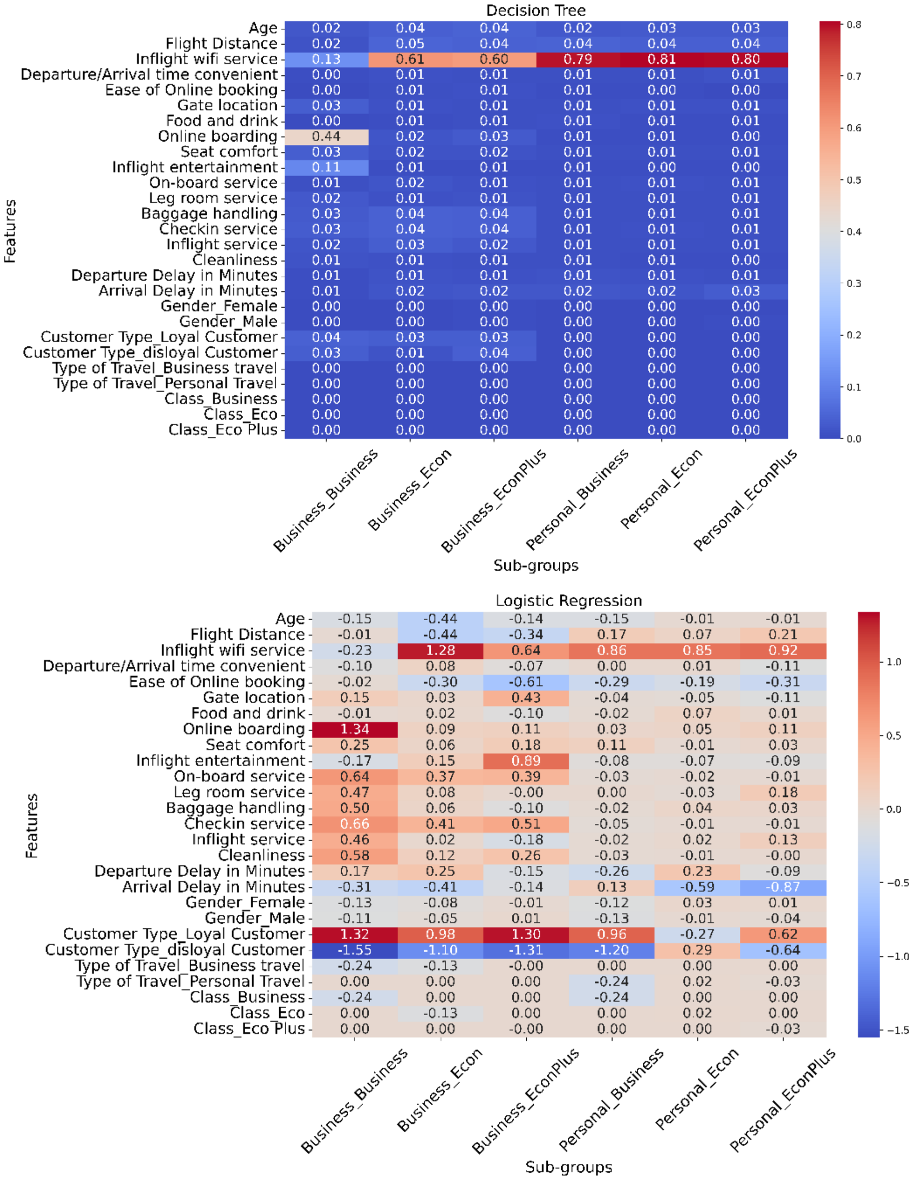

Figure 4. Feature importances from DT and LR classifiers trained on sub-group data.

The impurity-based feature weights of RT classifiers successfully select features that have significant influence on the satisfaction level, but they are unable to indicate the direction of such an influence. By contrast, the sign of feature weights in LR classifiers can be used to study if a feature influences the target value positively or negatively. From the second heat map, it can be observed that most of the services have positive coefficients, suggesting that improving these services may increase the likelihood that a passenger will feel satisfied about their flight experiences. However, an interesting observation is that the classifier conveys an unexpected message that on-board Wi-Fi service actually adversely influences the satisfaction level of passengers in business class who travel for business purposes.

Considering all the observations, we propose several suggestions for airlines to improve their flight experience. First, airlines should consider equipping all their airplanes with free on-board Wi-Fi services and updating their websites/mobile apps to facilitate easier online check-in processes. In addition, airlines should pay more attention to the business class experiences, including cleanliness, leg room, and entertainment system. Although not fully controlled by airlines, delays due to unnatural causes, such as understaffing or mechanical failure of planes, should be minimized.

In future, this project could be improved in several aspects. First, more analyses should be done to identify if the poor performances of classifiers on the last three subgroups derive from overfitting or underfitting, and the magnitude of regularization should be adjusted accordingly. In addition, data of satisfaction level specific to airlines could be incorporated into the dataset to produce more insights into how each airline should customize its services.

Contribution

Tung Dinh: Responsible for data preprocessing, training Decision Tree and Logistic Regression on the whole dataset, and extracting most important features based on trained models. Presenting and reporting the results. Responsible for writing the introduction, background, and most of the experiment procedure section.

Anji Ni: responsible for all the works involved in subgroup analyses; responsible for presenting and reporting the corresponding results; responsible for writing part of the method and experiment procedure section; responsible for writing all the results and discussions section.

Code

https://github.com/viettungggg/airline-customer-satisfaction

Reference

1. Habbal, M. R. (2022). Predicting & Optimizing Airlines Customer Satisfaction Using Classification (Graduate Capstone Project, Rochester Institute of Technology). Rochester Institute of Technology Capstone Repository.
2. Hulliyah, K. (2021). Predicting airline passenger satisfaction with classification algorithms. IJIIS: International Journal of Informatics and Information Systems, 4(1), 82–94. https://doi.org/10.47738/ijiis.v4i1.80
3. Jiang, X., Zhang, Y., Li, Y., & Zhang, B. (2022). Forecast and analysis of aircraft passenger satisfaction based on RF-RFE-LR model. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-14566-3
4. Nurdina, A., & Puspita, A. B. (2023). Naive Bayes and Knn for airline passenger satisfaction classification: Comparative analysis. Journal of Information System Exploration and Research, 1(2). https://doi.org/10.52465/joiser.v1i2.167