1 Abstract

This project uses a structured airline passenger satisfaction survey dataset to build and evaluate machine learning models that predict whether a passenger is satisfied or neutral/dissatisfied. Using demographic features, travel information, and detailed service ratings, we train several supervised models including Logistic Regression, Random Forest, XGBoost. Among all models evaluated, XGBoost achieved the best overall predictive performance, demonstrating that passenger satisfaction can be reliably inferred. The result of this research provides an important basis for airlines to enhance customer experience and allocate service resources efficiently.

2 Introduction

Passenger satisfaction is a key driver of an airline to enhance competitiveness. Understanding which factors influence satisfaction is crucial to improve service quality, reduce customer churn, and strengthen loyalty programs. This project focuses on the following predictive task: "Given structured passenger survey input, predict whether a customer is 'satisfied' or 'neutral/ dissatisfied.'" We implement and compare multiple supervised machine learning models to identify the most effective method for predicting customer satisfaction.

- Input: The model utilizes a comprehensive set of features derived from the passenger survey, including Demographic Attributes {gender, age}, Flight Characteristics {class, type of travel, flight distance} and Service Ratings {inflight wifi service, departure/arrival time convenient, ease of online booking, gate location, food and drink, online boarding, seat comfort, etc.}.
- Output: We then use a {logistic regression, random forest, XGBoost} to output a binary classification label indicating the customer's overall satisfaction level {satisfied, neutral/dissatisfied}.

3 Relative Work

Research on airline passenger satisfaction has expanded significantly in recent years, with an increasing focus on applying machine learning techniques to identify the key drivers of customer experience. While early studies primarily relied on service-quality frameworks, more recent work has shifted toward predictive modeling and data-driven analysis.

A notable stream of research compares the performance of standard classification algorithms for predicting passenger satisfaction. Prior studies have evaluated models such as Naive Bayes, K-Nearest Neighbors, Decision Trees, Support Vector Machines, and Random Forests, with the goal of identifying the most accurate predictive approach (Nurdina & Puspita, 2023; Hayadi et al., 2021). While these studies provide useful benchmarking results, they generally offer limited interpretability regarding the underlying drivers of satisfaction. Other work has introduced hybrid modeling pipelines that integrate feature selection and classification methods. For example, Jiang et al., in 2022 proposed an RF-RFE-LR pipeline that integrates Random Forest feature selection, Recursive Feature Elimination, and Logistic Regression, demonstrating the value of combined approaches for improving predictive performance.

Beyond prediction, a growing research emphasizes feature importance analysis to pinpoint the most influential factors affecting passenger satisfaction. Findings from these studies commonly highlight variables such as type of travel, inflight Wi-Fi quality, customer loyalty status, and online boarding efficiency (Mirzahossein & Rezashoar, 2025). Building on this interpretability-focused direction, our work applies SHAP analysis, which provides fine-grained, instance-level explanations of how individual features contribute to model predictions. More recently, causal inference frameworks have emerged in airline satisfaction research, highlighting a shift toward methods that identify the true causal impact of digital service improvements on passenger satisfaction (Mirthipati, 2024).

Based on prior work, our study evaluated multiple machine learning models and compared feature selection techniques. Rather than focusing solely on predictive accuracy or traditional feature importance measures, we introduced XGBoost and SHAP to achieve both strong predictive performance and clear, interpretable insights

into the factors shaping passenger satisfaction.

4 Dataset and Features

We base our analysis on the Airline Passenger Satisfaction dataset from Kaggle (https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction). The dataset provided was already split into a training set (103,904 examples) and test set (25,976 examples), contain a total of 129,880 passenger survey records, both with 24 columns and describe passenger's demographic attributes, flight characteristics, service ratings, and satisfaction outcome.

Before preprocessing, we performed several inspection and validation steps to ensure the dataset was correctly loaded and internally consistent. We first examined the raw data using .head( ) to verify column names, data types, and example entries. We also checked dataset dimensions train_df.shape, test_df.shape and confirmed that both training and test sets contained the same feature columns.

To prepare the dataset for modeling, we applied several preprocessing steps. Starting with the removal of non-informative columns. We dropped the 'id' and 'satisfaction' columns from the feature matrix. The 'id' variable is a unique identifier assigned to each passenger and does not contain any statistical information related to predicted satisfaction. The 'satisfaction' column is the target variable we aim to predict. Allowing the model to access it as an input feature would result in data leakage.

We also handled missing values. A small number of records had missing values in 'Arrival Delay in Minutes'. We imputed these missing entries using the median of the training set and test set respectively. Median imputation is preferred over mean imputation in this context because flight arrival delays exhibit a highly right-skewed distribution with occasional extreme outliers which heavily influence the mean. In contrast, the median is a robust statistic that better represents the central tendency of skewed variables.

Next, we needed to encode our nominal and ordinal categorical features. We applied three distinct encoding strategies to ensure that all variables are represented in a numerically meaningful way for supervised learning algorithms. Specifically, we used One-Hot Encoding, Ordinal Encoding and Label Encoding.

- *One-Hot Encoding* was applied to nominal features (Gender, Type of Travel). The first category was dropped to avoid redundancy.
- *Ordinal Encoding* was used for ordinal features (Customer Type, Class), where categorical levels reflect a natural ordering (e.g., disloyal Customer < loyal Customer). This encoding preserves rank information, allowing models to exploit relative differences across categories.
- *Label Encoding* was applied to the target variable (satisfaction). This converts the two satisfaction categories into binary numeric labels, enabling supervised learning algorithms to process it correctly.

Lastly, we addressed feature normalization by applying Z-score normalization using StandardScaler. This standardized continuous variables such as Age, Flight Distance, Departure Delay, and Arrival Delay. This prevents features with large magnitudes from dominating the optimization process and improves model stability for algorithms such as logistic regression.

This dataset provides more than 100,000 training samples, offering a high degree of variability. This makes it well-suited for supervised learning, allowing models to learn complex nonlinear relationships between service attributes, operational performance, and customer satisfaction outcomes.

5 Methods

The airline customer satisfaction dataset that we used for this analysis requires a classification based model. This is because we are trying to determine if a passenger fits in one of two categories "satisfied" or "neutral/dissatisfied". This narrowed down the type of models that we wanted to analyze for this data.

After our data was sufficiently cleaned, we wanted to test three different models to determine which would result

in the best accuracy. First, we fit a logistic regression model to our dataset which is used to model the probability of an outcome using a sigmoid function: $\sigma(x) = \frac{1}{1+e^{-x}}$

Some key properties of this function include that as x approaches positive infinity, $\sigma(x)$ approaches 1 and as x approaches negative infinity, $\sigma(x)$ approaches 0. Lastly, when x equals zero $\sigma(x)$ is 0.5. These properties show how this function can be used to find the probability of a number as its results fall between 0 and 1.

We used a logistic regression as our base function, which used an L_2 penalty term to regularize and keep the model from being too complex. This penalty term is proportional to the sum of the squared coefficients. This specific method of regularization changes the weights of the coefficients, but it keeps all of the original features in the model. We also wanted to compare this to a LASSO regression. This uses an L_1 penalty term, and it is proportional to the sum of the absolute values of the coefficients, which encourages the model to shrink less important feature coefficients to zero. Applying this type of penalty allows for feature selection, as features with a coefficient of zero are not used in the model.

We also wanted to test an additional feature selection technique, RFECV, using logistic regression as our base function. This technique recursively removes the least important features from a dataset to identify the optimal subset that gives best performance. It incorporates cross-validation at each iteration and evaluates performance on different folds.

Once we had all three versions of our model, we compared their performance by evaluating their accuracy, precision, and recall. We found that our base model, with the L_2 penalty term, had the best results, but we wanted to conduct additional testing against different types of classification models. We specifically wanted to test models that do not assume linear features, as assuming that the relationship between predictors and log-odds is linear is one of the limitations of logistic regression. We did so by implementing random forest and XGBoost.

Random forest is a model which uses a large number of decision trees during training and combines their predictions to improve accuracy and reduce overfitting. Each tree is trained on a bootstrap sample of the data, and each tree learns based off of errors from its stumps. The final decision, in this case classifying if a passenger is satisfied or not, is made via majority vote of all trees in the model.

XGBoost was the last type of model that we fit our data to. It also uses trees and boosts by fitting a tree to residuals in each step. However, the way it does so is different. It uses trees built sequentially, instead of independently, with each new tree correcting the errors of the previous trees. This is another powerful classification tool that we wanted to analyze for our data.

Once we finalized our logistic regression, random forest, and XGBoost models, we found their accuracy, precision, and recall, as these are the best metrics to track performance of binary categorical models. We know that the best models have higher accuracy, precision, and recall. Therefore, this is what we looked for in our evaluation.

We found the best model and implemented SHAP analysis, in order to be able to more easily interpret our model and get real insights about our different features. SHAP is a method for interpreting machine learning models by assigning each feature a contribution value to the prediction. Based on the SHAP values, we can quantify how much a feature participates in the prediction.

Together, these methods allowed us to find the model which would make the best predictions for our data as well as understand the key features that impact these predictions. This is useful because it allows airlines to predict if customers will be satisfied with a combination of features, as well as understand which features have the biggest impact on satisfaction. They can then make data driven insights based on these features to improve their passenger experience.

6 Experiments/Results/Discussion

As explained in the methods section, we first fit a model for a base logistic regression, a LASSO regression, and used RFECV analysis on our original logistic regression model. To do so we had to set hyperparameters for each

step. For the base logistic regression, with an L_2 penalty, we used max iterations of 1000 to minimize the loss function. For LASSO we specified that we wanted to test 10 different inverse regularization strengths C, that we wanted to test each of these 10 Cs on 5 cross validation folds, we wanted to use an L_1 penalty to select our features, and again we wanted to use a maximum number of iterations of 1000. Lastly, for RFECV we once again used a maximum iteration of 1000, we specified that we wanted to only remove one feature every time that the model was evaluated, and that we wanted to have 5 cross validation folds each time that we removed a feature.

When we scored each of these models' predictions we found that our base logistic regression was the best model as it had the highest accuracy, precision, and recall scores of the three as seen in the figure below.

Base Logistic Regression:

- Accuracy: 0.8713427779488759
- Precision: 0.8680485800383526
- Recall: 0.8336402701043585

LASSO:

- Accuracy: 0.8711887896519864
- Precision: 0.8678659483152223
- Recall: 0.8334648776637726

RFECV:

- Accuracy: 0.8704188481675392
- Precision: 0.8664174341205434
- Recall: 0.8332894852231869

Logistic regression assumes a linear relationship between predictors and log-odds, which might not be optimal for our data. Because of this, we want to compare logistic regression to random forest and XGBoost before we interpret our features. Since our base model using L_2 penalty is the only one that preserves all of the original features, like random forest and XGBoost, we used this type of model for the comparison.

For the random forest model, we set the random state hyperparameter to 42. This is an arbitrary hyperparameter that sets the seed for the random number generator so that we can reproduce our specific results. For XGBoost we wanted to evaluate the model's performance using logarithmic loss, which is common for binary classification models like the ones we are fitting. We also set the hyperparameter for the label encoder to false since we already encoded our data during processing. We ran each type of model, logistic regression, random forest, and XGBoost through GridSearch in order to find the best hyperparameters for the model, besides the ones already listed.

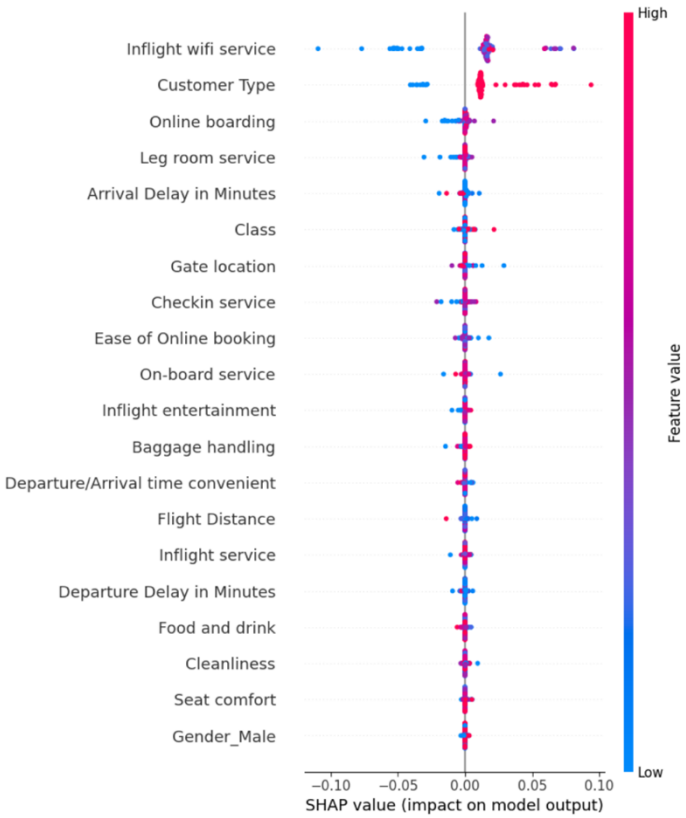
Once all of the models that we wanted to test were created, we found the best hyperparameters for each one. For logistic regression we found that the best inverse regularization strength was 1 for our data set ($C = 1$). Random forest had best parameters that indicated a maximum depth of 20 for each of the decision trees, splits based on a minimum of 2 nodes, and 200 different trees. Lastly, our model using XGBoost had the best hyperparameters of learning rate of 0.1, maximum depth of 6 for each boosted tree, and 200 different boosted trees. Once these models were trained with these hyperparameters in mind, we compared their performances. The output below showcases our findings.

	Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC	Training Time (s)
0	Logistic Regression	0.871535	0.868042	0.834166	0.850767	0.925630	23.95
1	Random Forest	0.963197	0.972416	0.942910	0.957435	0.993972	280.82
2	XGBoost	0.964275	0.973682	0.944138	0.958682	0.995279	32.19

This table shows us many key takeaways from each of our models. Most importantly, we found that XGBoost was the best model for our dataset. It had the highest scores for accuracy, precision, recall, ROC-AUC, and F1 Score. It also took considerably less time to train than random forest. Because of this, it is the model we recommend using to get the best predictions for airline passenger satisfaction.

Being able to make predictions using this model is valuable, but it is not enough. We also need to evaluate the features used to make these predictions. Airlines do not only want to know if customers will be satisfied by inputting a set of features, they also want to understand what features are driving the biggest impact and if this impact is positive or negative. This will allow them to make informed business decisions to improve their satisfaction rates. To do this, we retrained the XGBoost model using its best parameters and then performed SHAP analysis. The SHAP analysis output can be seen below.

Based on the SHAP values, we can quantify how much a feature participates in the prediction. Below are our top 5 most important features ranked from highest to lowest magnitudes, followed by the magnitude of their impact on model output:



a. Inflight wifi service (0.027167): Higher wifi service reviews, from customers, lead to higher overall passenger satisfaction

b. Customer type (0.019843): Being a loyal customer positively impacts overall passenger satisfaction, in comparison to disloyal customers.

c. Online boarding (0.002438): Higher online boarding satisfaction levels leads to higher overall passenger satisfaction

d. Leg room service (0.001400): Higher leg room satisfaction leads to higher overall passenger satisfaction

e. Arrival delay in minutes (0.001272): Lower arrival delay in minutes leads to higher overall passenger satisfaction

Based on the SHAP analysis, several focused strategies can help airlines improve passenger satisfaction. First, airlines should upgrade inflight Wi-Fi service and offer tiered options. Second, enhancing loyalty programs with better benefits and point incentives can boost membership conversion. Third, optimizing the online boarding process through simpler check-in flows can create a smoother journey. Fourth, improving seat comfort and legroom by adjusting cabin seat layout and offering more extra-legroom options. Finally, stronger on-time performance management can help reduce dissatisfaction when delays occur.

7 Conclusion/Future Work

This study developed and evaluated multiple machine learning models to predict airline passenger satisfaction. After systematically comparing Logistic Regression, Random Forest, and XGBoost, XGBoost demonstrated the strongest performance with the highest accuracy, precision, and recall. To enhance interpretability, we incorporated SHAP analysis to identify the features that most influence satisfaction, making the model both accurate and actionable for decision-making. Overall, the findings show that combining high-performing predictive models with interpretable analytics can generate meaningful, data-driven strategies. However, a limitation is that the current framework focuses solely on satisfaction, which does not always translate directly into profit gains for airlines. As future work, we will explore modeling approaches that connect service improvements to financial outcomes, enabling predictions of both satisfaction and their economic impact.

8 Contributions

Julia Petty	Looked through the initial code for errors, and suggested additional potential adjustments as our project progressed; Wrote the Experiments/Results/Discussion part of the report and edited the Methods section to flow with it; Presented the comparisons of the logistic regression, random forest, and XGBoost models as well as a deep dive into the specifics of our final XGBoost model.
Vishnu Akumalla	Created the first version of code; Corrected errors and refined the code based on suggestions from team members; Presented the base logistic regression, hyperparameter tuning, and feature importance parts of the final presentation; Worked on the Methods section of the report.
Yunxuan Hu	Read and collect the relative literature; Finished the conclusion and future work section; Made recommendation according to the results; Presented the literature and recommendation.
Junlin Kou	Read the code and finished the Abstract, Introduction, Data and Features sections of the report; Presented the research question and importance of the question, background of dataset and data preprocessing; Revised and improved the formatting of the report.

9 Reference

Hayadi, B. H., Kim, J. M., Hulliyah, K., & Sukmana, H. T. (2021). Predicting airline passenger satisfaction with classification algorithms. International Journal of Informatics and Information Systems, 4(1), 82-94.

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), 11174.

Mirthipati, Tejas. "Enhancing airline customer satisfaction: A machine learning and causal analysis approach." arXiv preprint arXiv:2405.09076 (2024).

Mirzahossein, Hamid, and Soheil Rezashoar. "Feature Importance Analysis of Optimized Machine Learning Modeling for Predicting Customers Satisfaction at the United States Airlines." Machine Learning with Applications (2025): 100734.

Nurdina, A., & Puspita, A. B. I. (2023). Naive Bayes and KNN for Airline Passenger Satisfaction Classification: Comparative Analysis. Journal of Information System Exploration and Research, 1(2).