Can we predict airline customer satisfaction?

ORIE 5741

Freda Jia
Peter Wu
Summer Xiao

Freda is responsible for sections 1, 2.1, 2.2, 4.1, and relevant work of the report. Peter is responsible for sections 3, 4.2, and relevant work of the report. Summer is responsible for 4.3, 5, and relevant work of the report. Everyone contributes equally to section 2.3 and the video presentation.

https://github.com/Peta0228/Airline-Satisfaction-Prediction

May 10, 2024

1 Introduction

In today's highly competitive airline industry, ensuring customer satisfaction is paramount for airlines to maintain a competitive edge and foster customer loyalty. Understanding the factors that influence customer satisfaction during air travel is crucial. However, traditional ways of collecting and analyzing feedback through customer surveys may not be enough. It is possible to use data analysis and machine learning to bolster the process. We want to use past surveys and build predictive models to offer a more efficient and proactive approach to understanding and predicting customer satisfaction.

2 Data

2.1 Data Overview

The dataset titled "Airline Passenger Satisfaction" is sourced from Kaggle. It comprises a training dataset, which accounts for 80% of the complete dataset; a testing dataset, constituting 20% of the total dataset and utilized for estimation purposes.

Comprising 25 columns, the training dataset encompasses 103,904 entries, while the testing dataset comprises 25,976 entries. Each entry within the dataset corresponds to a unique passenger journey and encompasses comprehensive information, including gender, customer type, age, type of travel, class, flight distance, and diverse satisfaction ratings encompassing various aspects of the journey. These aspects include inflight wifi service, departure/arrival time convenience, online booking experience, seat comfort, cleanliness, among others.

Of particular significance is an additional column dedicated to the satisfaction level, indicating whether passengers rated their experience as satisfactory, neutral, or dis-satisfactory. This column serves as the target variable for prediction in our analysis.

2.2 Exploratory Data Analysis

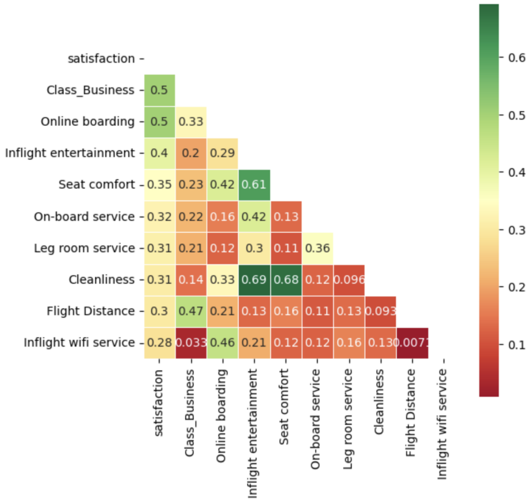

Figure 1: Heatmap of the correlation matrix

A heatmap of the correlation matrix has been generated to explore the relationships between variables, with a specific focus on the top 10 variables that are most strongly correlated with 'satisfaction' in the training dataset, showing in Figure 1. Each label on the rows and columns represents a distinct variable. The values in each cell of the heatmap indicate the correlation coefficient between pairs of variables. The accompanying color scale ranges from red (indicating a low correlation, closer to -1), through yellow (indicative of a neutral correlation, close to 0), to green (showing a high correlation, closer to 1).

2.3 Feature Engineering

Some categorical variables are converted to numerical. We encoded gender, customer type, type of travel, class, and satisfaction into binary values. Specifically, for the 'Class' column, which denotes the travel class of passengers, we utilized the one-hot encoding technique. This process resulted in the creation of three new binary variables: $Class\_Business$, $Class\_Eco$, and $Class\_EcoPlus$. Each of these variables represents a different travel class, with a value of 1 indicating the presence of that class and 0 otherwise. For the target variable 'satisfaction', 'satisfied' is encoded as 1, and both 'neutral' or 'dissatisfied' are converted to 0.

Additionally, during our exploration of the dataset, we observed a strong correlation between the 'Departure Delay in Minutes' and 'Arrival Delay in Minutes' variables. Given their close relationship and the redundancy in the information they provide, we made the decision to drop the 'Arrival Delay in Minutes' column from our dataset. This reduces the feature correlations.

3 Principal Component Analysis (PCA)

3.1 Dimensionality Reduction & Unsupervised Clustering

PCA is useful in reducing the dimensionality of dataset. With 24 features and more than 100k records, the model is likely to overfit without PCA. With the correlations between the variables, PCA becomes even more necessary. Figure 2 explained displayed the cumulative percentage of explained variance on the top 10 principal components (PCs), in combination they explained 77.43% of the variance on training data.

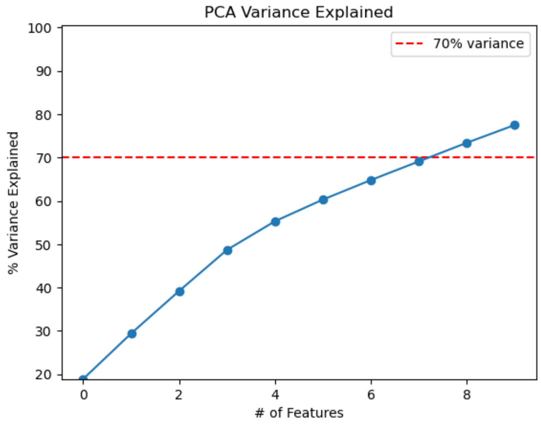

Figure 2: Cumulative Variance Explained by PCA

One advantage of PCA is we can visualize the PCs, and thus inferring their representation. Recall that PCs are linear combinations of features, and so the laoding and score in a PC can be meaningful to interpret. PCA is an unsupervised learning method, and loading can be treated as characteristics of a PC cluster, while score being how close a data point is to a cluster. We visualize on the first two PCs.

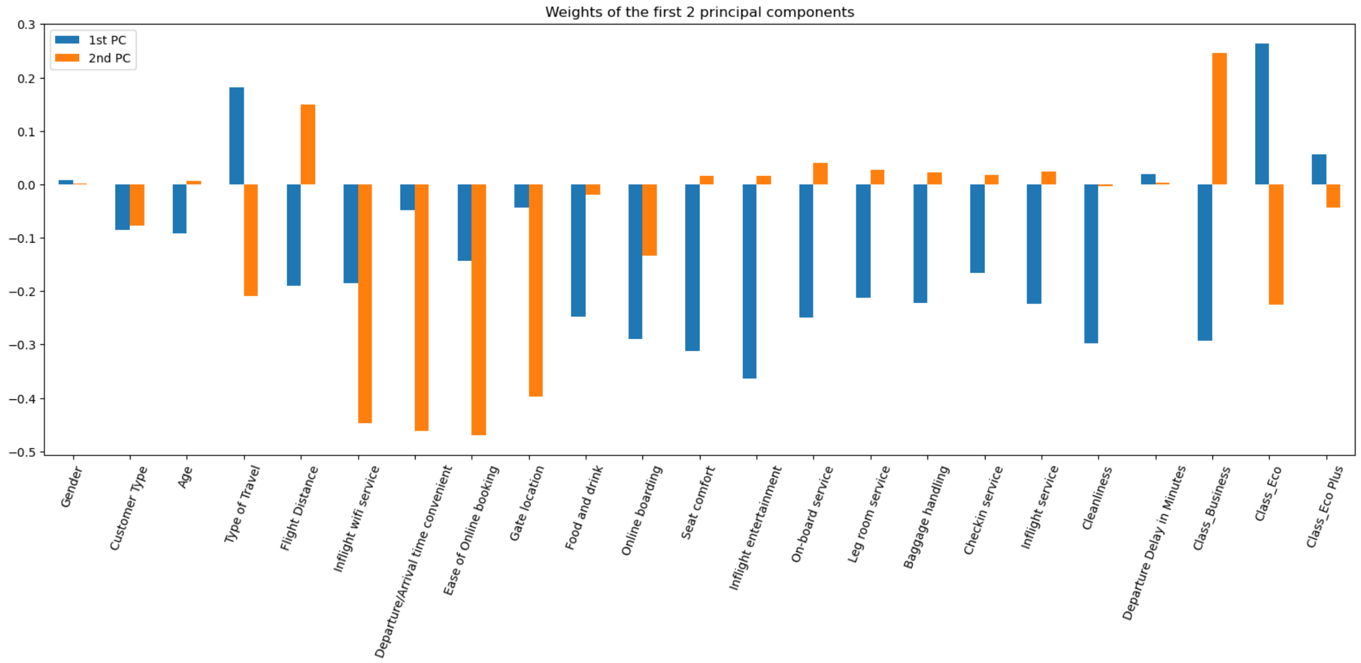

Figure 3: Feature Weights on PC1 and PC2

Figure 3 shows the weights of all features on the first 2 PCs. Blue bars are loading of PC1, and orange bars are loading of PC2. PC1 captures the characteristics of customers predominantly engaged in personal travel. These customers are likely not to travel in Business Class and show a preference for more economical options, such as traveling in Economy Class. The service quality features such as Seat Comfort and Inflight Entertainment tend to have negative coefficients, suggesting dissatisfaction or lower expectations in these areas. This group can be described as non-business, economically conscious, and discerning in certain service aspects.

PC2, in contrast, reflects the traits of customers who primarily travel for business purposes. This group is associated with longer flight distances and a preference for Business Class. Unlike the first group, these customers place a higher emphasis on efficiency, as indicated by positive coefficients for Departure/Arrival time convenience and Inflight WiFi service/ease of online booking. They are characterized as business travelers who prioritize efficient service and convenience, albeit with less emphasis on leisure-oriented features.

4 Methodology

4.1 Logistic Regression

For logistic regression (LogReg), we conducted 5-fold cross-validation(CV) to determine the optimal regularization parameter C for both L1 and L2 regularization, shown in Table 1. In LogReg, C controls the regularization strength. A small C imposes stronger regularization, penalizing large coefficients to prevent overfitting. Conversely, a large C reduces regularization, allowing the model to closely fit the training data. Balancing C is essential for optimal generalization performance. Based on all our accuracy results in Table 2, the highest testing accuracy is achieved in **L1 without PCA**.

C	**L1 Regularization**	**L2 Regularization**
Without PCA	10	0.1
With PCA (10 PCs)	10	0.1
With PCA (12 PCs)	0.1	1
With PCA (8 PCs)	0.01	0.001

Table 1: Optimal regularization parameter (C)

Model	**L1 Regularization**		**L2 Regularization**	
	Training	Testing	Training	Testing
Without PCA	87.49%	87.16%	87.49%	87.12%
With PCA (10 PCs)	84.67%	84.36%	84.68%	84.35%
With PCA (12 PCs)	84.94%	84.63%	84.94%	84.62%
With PCA (8 PCs)	83.62%	83.16%	83.61%	83.20%

Table 2: Accuracy Table for LogReg

4.1.1 Logistic Regression Without PCA

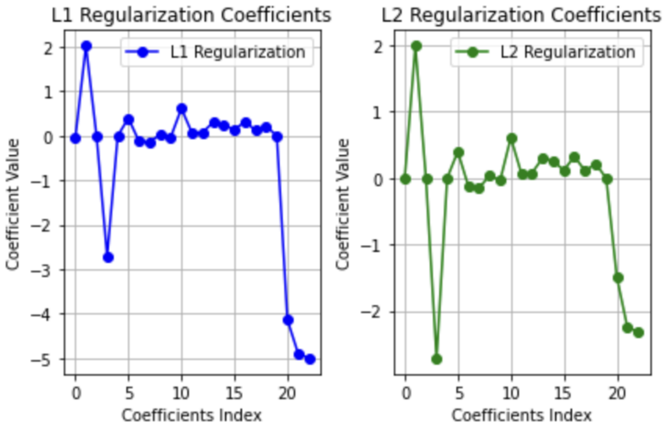

Figure 4: Coefficients from LogReg without PCA

In LogReg without PCA, notably, both regularizations produced comparable results. As shown in Figure 4, the L1 regularization coefficients graph displays a characteristic pattern where many coefficients converge to zero, typical of L1 regularization. Initially, coefficients cluster around zero, with exceptions showing notable positive (e.g., Customer Type) or negative influences (e.g., Type of Travel). Towards the end, coefficients for attributes like Class_Business, Class_Eco, and Class_Eco Plus sharply drop to -5, indicating substantial negative impacts or elimination.

In L2, the coefficients exhibit a more balanced distribution around zero compared to L1. L2 regularization ensures a more uniform balance of feature influence, mitigating the risk of overfitting. Table 3 shows the significant coefficients.

4.1.2 Logistic Regression With PCA

The performance of LogReg with PCA and those without is comparable in accuracy. We also tried to adjust the number of PCs utilized in the LogReg models. When employing 12 PCs, we observed

Coefficients	**L1 Regularization**	**L2 Regularization**
Customer Type	2.029	2.096
Type of Travel	-2.713	-2.830
Satisfaction	-4.109	-1.591
Class_Business	-4.848	-2.293
Class_Eco	-4.965	-2.395

Table 3: Significant coefficients

they explained 84.03% variance whereas using 8 PCs explained 69.08%.

The performance of LogReg with L1 and L2 regularization is remarkably similar. This similarity could be attributed to several factors. Firstly, most features in our dataset are discrete, which may lead to a less pronounced difference in the impact of regularization. Second, the presence of correlated features can influence how L1 and L2 regularization behave. In scenarios where features exhibit high correlation, the regularization penalties may affect the coefficients similarly in both L1 and L2. Meanwhile, LogReg is inherently a simpler model compared to more complex algorithms. Consequently, the regularization effect on model performance may not variate significantly between L1 and L2 .

4.2 Support Vector Machine (SVM)

Model	Training	Testing
Without PCA	67.07%	66.96%
With PCA	92.39%	92.28%

Table 4: Accuracy Table for SVM

We used a nonlinear transformation on the dataset called Radio Basis Function (RBF) kernel, to improve the linear decision boundary used by vanilla SVM. Overall, the **with PCA SVM** produced the hightest accuracy.

4.2.1 Improvement from PCA

Table 4 shows a significant accuracy improvement from applying PCA. One possible explanation is due to the high dimensionality of the features, like some other machine learning methods, SVM could suffer from the curse of dimensionality. As the number of dimensions increase, the data that SVM needs to generalize accurately also increase, and likely SVM will have much more supporting vectors than in a low-dimension setting. However, it is unusual to have that many data points near/on the margin. In our dataset' situation, not as many customers will be on the edge deciding their satisfaction. Arguably, most of them are either satisfied, or not. Thus the high dimensionality can easily lead our SVM's overfitting, and PCA becomes necessary to address the issue.

4.2.2 Radio Basis Function (RBF) Kernel

The RBF Kernel is used to implicitly map the input data into a higher-dimensional space where it is more separable. RBF provides good generalization for SVM, as it has no strong assumption about the data distribution and let data points to have similar predictions when they are close to each other. This creates a robust and complex decision boundary for SVM.

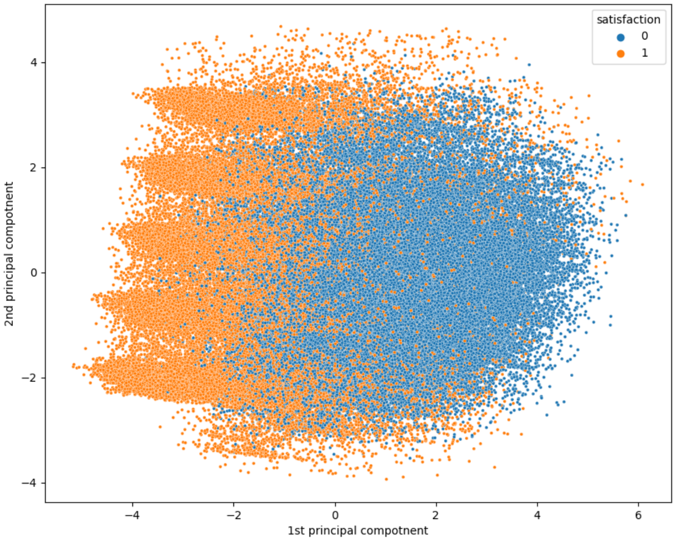

Figure 5: Satisfaction Distribution on PC1 and PC2

Figure 12 shows the distribution of target variable on PC1 and PC2. Clearly, the decision boundary is non-linear, which is why PCA and RBF kernel in combination should work well.

4.3 Ensemble Methods: Random Forest and XGBoost

We implemented Random Forest (RF) and XGBoost with parameter tuning and 5-fold CV.

4.3.1 Model Implementation and Parameter Tuning

RF builds multiple decision trees with data bootstrapping and feature subsampling, and aggregating results to get higher accuracy and stability. Five parameters are tuned: the number of trees `n_estimators`, the maximum number of features per split `max_features`, maximum tree depth `max_depth`, the minimum number of samples required to split a node `min_samples_split`, and the minimum number of samples per leaf `min_samples_leaf` 5. A subset of all possible parameter combinations are tried to greedily achieve the best fine-tuning.

XGBoost, or Extreme Gradient Boosting, is another powerful ensemble decision trees method that combines weak learners into a much stronger one. It achieves this by iteratively fitting new weak learners to the residual errors of the previous learners and then updating the prediction model. We also tune five parameters for our XGBoost model: the number of trees `n_estimators`, learning rate `learning_rate`, maximum tree depth `max_depth`, data subsample ratio per tree building `subsample`, and the column subsample ratio per tree building `colsample_bytree` 5. Again, our limited attempts give promising results.

Random Forest	Grid Value	**XGBoost**	Grid Value
`n_estimators`	100, **200**, 300	`n_estimators`	100, 200, **300**
`max_features`	4, **5**	`learning_rate`	**0.05**, 0.1, 0.2
`max_depth`	10, **20**	`max_depth`	3, 6, **9**
`min_samples_split`	**10**, 20	`subsample`	0.5, 0.7, **0.9**
`min_samples_leaf`	**50** , 100	`colsample_bytree`	0.5, **0.7**, 0.9

Table 5: Ensemble Methods Parameter Tuning (* indicates best set of parameters)

4.3.2 Model Evaluation

As shown in Table 6, **XGBoost** performs slightly better in testing accuracy than RF.

Model	Training	Testing
Random Forest	94.74%	94.59%
XGBoost	98.30%	96.46%

Table 6: Accuracy Table for Ensemble Methods

We also look at the ROC-AUC score (area under the ROC curve) and confusion matrix on the test set. ROC-AUC is 0.94 and 0.96 respectively for the fine-tuned RF and XGBoosting. In Figure 6 and 7, both models demonstrated great prediction results. We anticipate air companies to focus more on reducing **False Negatives** (FN) in which dissatisfied customers are mistakenly predicted as satisfied, as failing to identify dissatisfied group undermines the efforts on enhancing customer satisfaction. Meanwhile, mis-classifying satisfied customers as dissatisfied is less detrimental. Considering this, XGBoost might be more sutiable as it has fewer FN than RF.

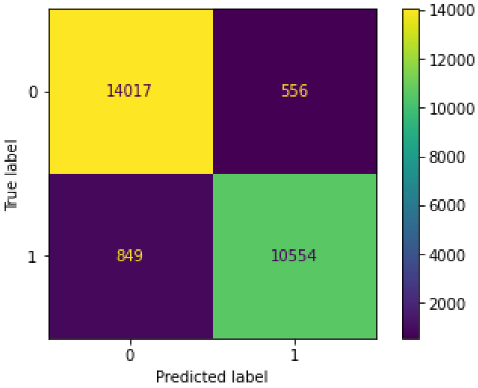

Figure 6: Fine-tuend RF

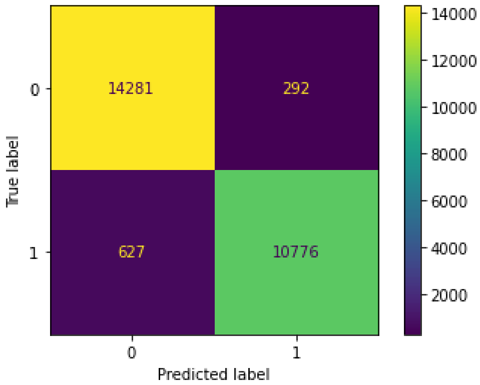

Figure 7: Fine-tuend XGBoost

4.3.3 Feature Importance

Feature importance helps in understanding factors that are crucial to high satisfaction. We started with Mean Decrease Impurity (MDI). The higher the MDI, the more important a variable is in decreasing node impurity. We also applied SHAP (SHapley Additive exPlanations) values to interpret the impact of each feature on the prediction outcome [1]. SHAP values on RF shows that dissatisfied customers are not happy with their online boarding experience, check-in service, cleanliness, and inflight entertainment, and are likely to be on a personal trip. In contrast, good online boarding service and inflight wifi service make business-class travelers satisfied. We draw such customer profiles by analyzing figure 8 and 9.

[1] `https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability`

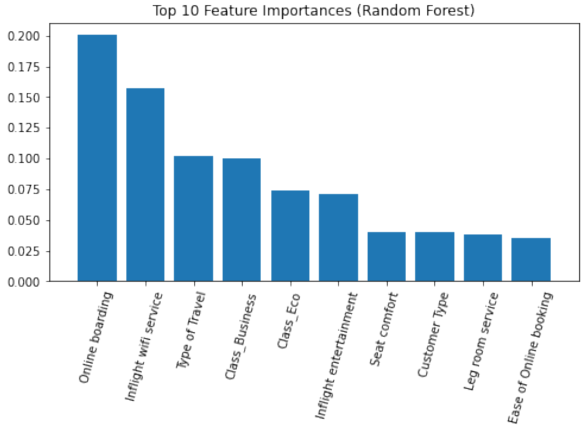

Figure 8: RF Model Feature Importance

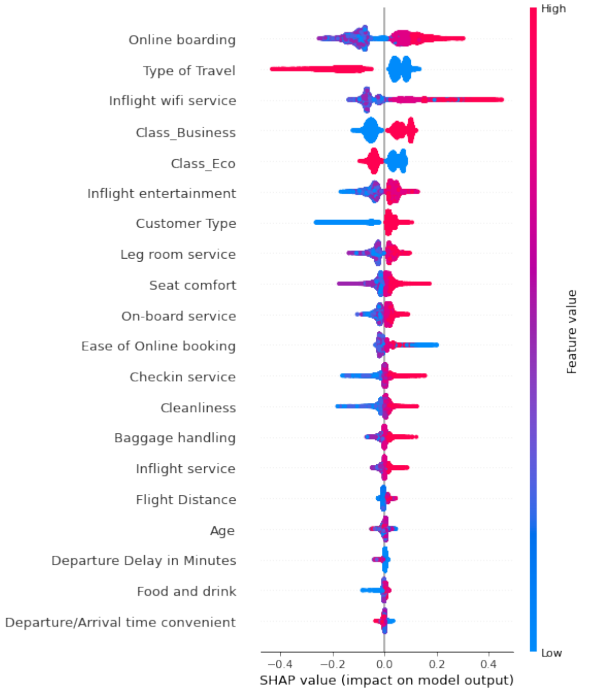

Figure 9: RF Model SHAP Values

5 Conclusion

In conclusion, we are able to achieve promising results in predicting customer satisfaction for airline companies. All of our final models show at least 85% training and testing accuracy, among which XGBoost achieved the highest in both metrics. In general, ensemble methods and SVM with Kernel (non-linear) work better than LogReg.

Admittedly, there are a few limitations to our project. First, we do not know how many airline companies provide the dataset we use. If the data primarily comes from a limited number of airlines or specific regions, the models' applicability to other contexts is compromised. Second, despite efforts to optimize model complexity through PCA and ensemble methods, the risk of overfitting remains given the high dimensionality of our data. Finally, our dataset only has survey data and misses information from text forms such as customer feedback where advanced techniques such as Natural Language Processing can be used for better analysis.

Several enhancements can be made to our project. First, we can refine our LogReg model using feature selection based on the regularization results to better balance complexity and performance. Although our ensemble methods perform quite well for now, we could try different parameters to improve. As we discover business and leisure travelers behave differently through PCA and SHAP analysis, we could target different types of travelers for prediction.

We recognize that our project might produce a Weapon of Math Destruction (WMD). The complex algorithms we use have inherent opacity and relatively unclear decision-making processes. It would be difficult to challenge or rectify biased conclusions or inaccurate implications for our prediction results. If the training data are not representative of a diverse customer base, these issues could be perpetuated and amplified, leading to unfair treatment of some customer groups. Therefore, we are against using this model for critical decision-making within airline companies. Our goal is to provide insights into improving customer satisfaction without risking significant operational repercussions.

6 Appendix

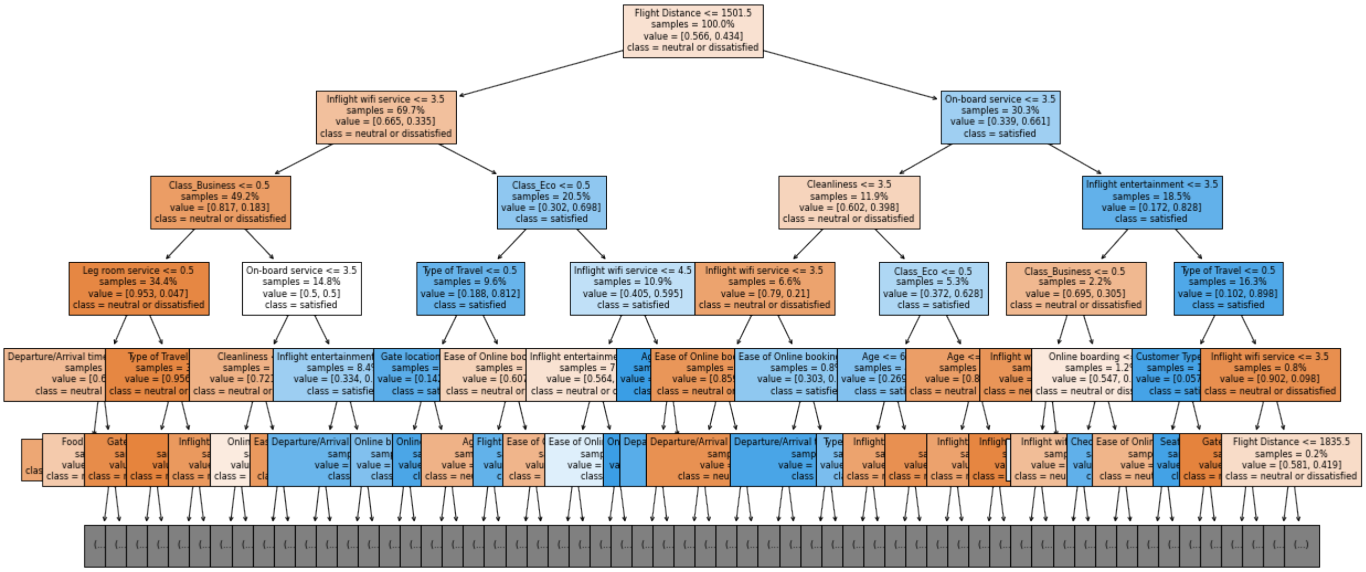

Figure 10: RF Model Tree Visualization

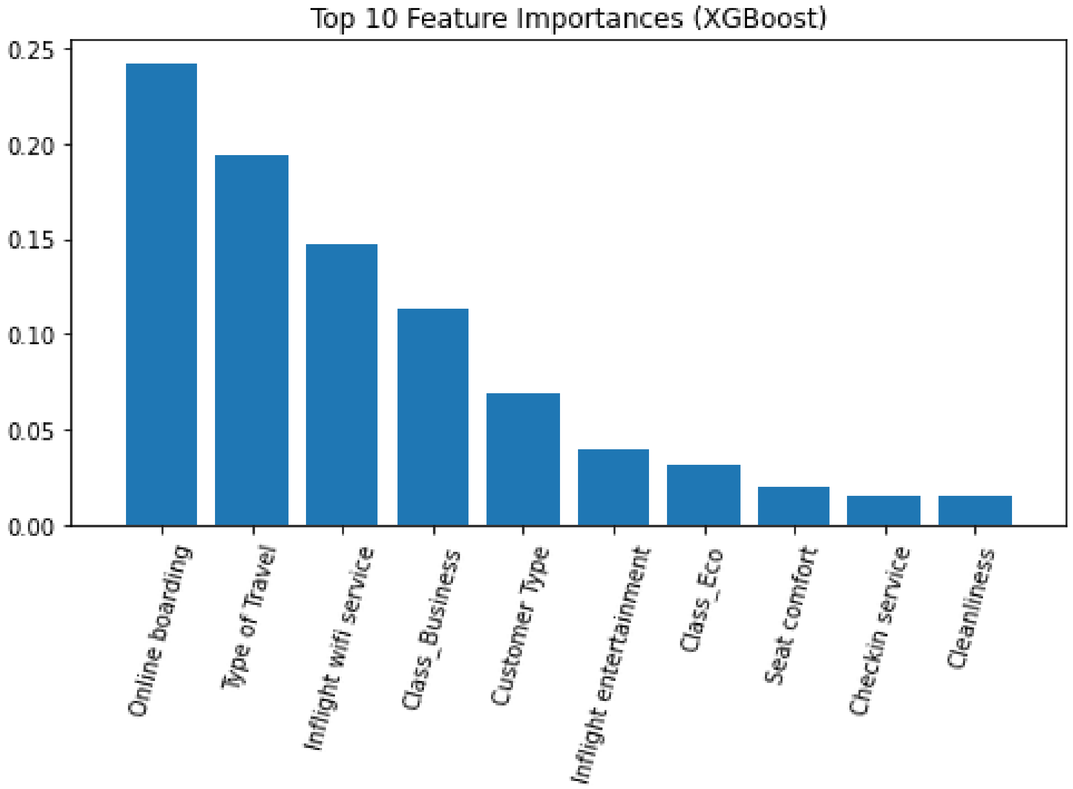

Figure 11: XGBoost Model Feature Importance

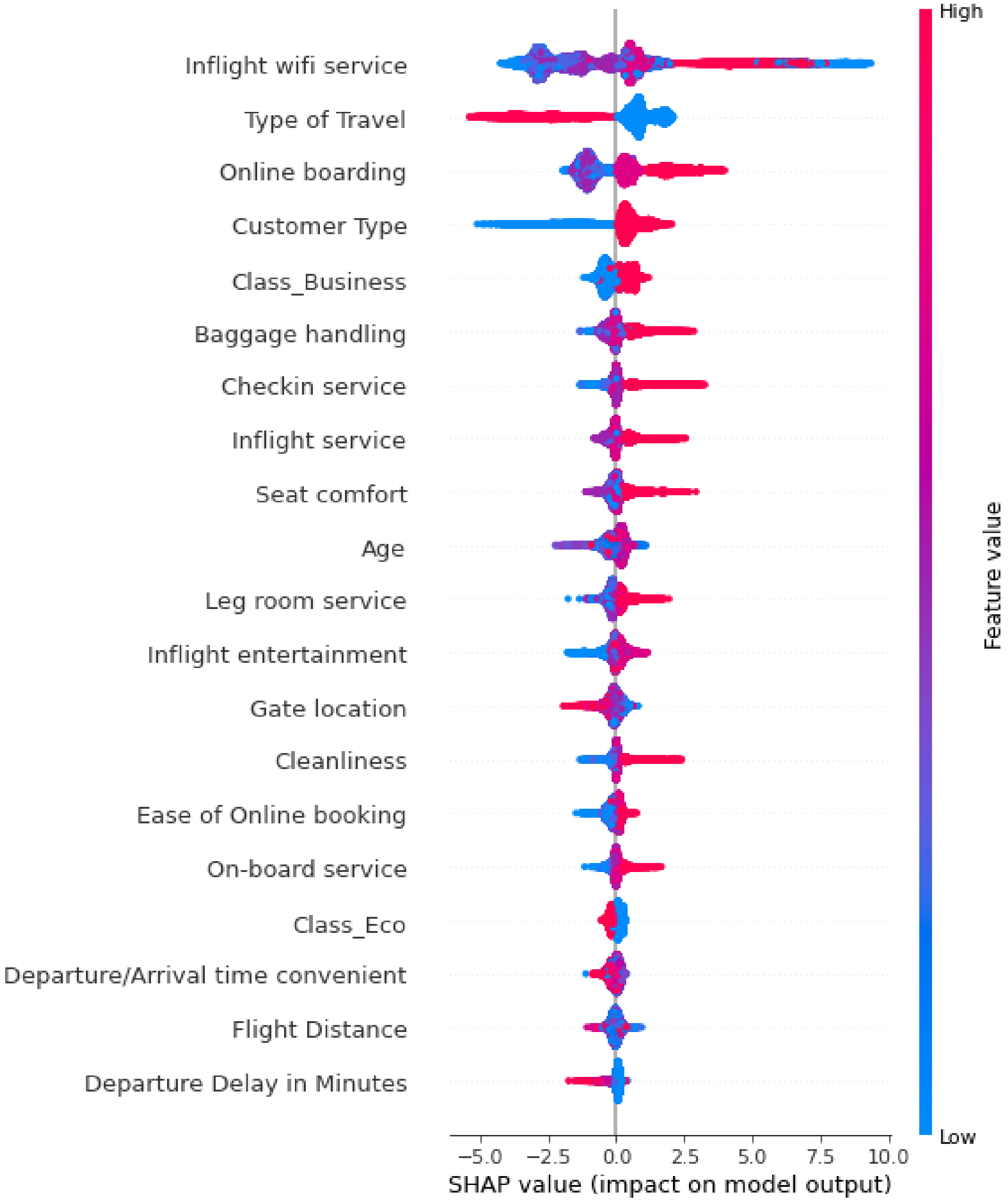

Figure 12: XGBoost Model SHAP Values