Airline Passenger Satisfaction

Data Analysis

Group 02

Samujitha Senaratna

Chamodi Siriwardhana

Senuri Perera

Abstract

Table of Content

List of Figures

List of Tables

1. Introduction

The airline sector acts as the vital link for global travel, transporting millions of passengers across diverse destinations. In this intensely competitive environment, where airlines compete for their share of the market, there's a pivotal factor that unequivocally gauges an airline's prosperity: the contentment of its passengers. Given the ever-changing expectations of travelers and the dynamic nature of the industry, there is a crucial need to thoroughly explore what constitutes a genuinely gratifying flying experience.

Airline passenger satisfaction is of paramount importance since satisfied passengers are more likely to become loyal customers. A positive experience encourages passengers to choose the same airline for future travels, fostering customer loyalty. Repeat business is not only economically beneficial for the airline but also contributes to the establishment of a strong and reliable customer base. Also, satisfied passengers often share their positive experiences with others. Positive word-of-mouth serves as a powerful marketing tool, enhancing the airline's reputation.

Developing a model to predict the satisfaction of airline passengers enables airlines to adopt proactive improvement strategies by identifying potential areas of dissatisfaction early on, allowing for targeted enhancements in service quality.

2. Description of the Question

As air travel has become an integral part of modern life, airlines strive to optimize their services to ensure a positive passenger experience. Airlines work to improve their services to provide a satisfying customer experience since air travel has become a necessary component of contemporary living. However, understanding the different factors that affect passenger satisfaction is a complicated task.

The main goal is to develop a predictive model that can accurately anticipate passenger satisfaction level based on these factors. The analysis also aims to offer insight on the key associations between particular characteristics and passenger satisfaction, revealing which features of the travel experience have the most influence. By addressing this issue, the project aims to provide useful information that will enable airlines to make effective choices, which will contribute in improving their services and encouraging higher levels of customer satisfaction.

3. Description of the Dataset

The "Airline Passenger Satisfaction” Dataset is taken from the Kaggle website and contains 103904 observations. There are 25 variables, out of which the response “satisfaction” is a categorical variable with two levels (‘Satisfied’ and ‘Neutral or dissatisfied’).

Variable Name	**Description**
Gender	Gender of the passengers (Female, Male)
Customer Type	The customer type (Loyal customer, disloyal customer)
Age	The actual age of the passengers
Type of Travel	Type of Travel
Class	Travel class in the plane of the passengers (Business, Eco, Eco Plus)
Flight distance	The flight distance of this journey
Inflight wifi service	Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)

Departure/Arrival time convenient	Satisfaction level of Departure/Arrival time convenient
Ease of Online booking	Satisfaction level of online booking
Gate location	Satisfaction level of Gate location
Food and drink	Satisfaction level of Food and drink
Online boarding	Satisfaction level of online boarding
Seat comfort	Satisfaction level of Seat comfort
Inflight entertainment	Satisfaction level of inflight entertainment
On-board service	Satisfaction level of On-board service
Leg room service	Satisfaction level of Leg room service
Baggage handling	Satisfaction level of baggage handling
Check-in service	Satisfaction level of Check-in service
Inflight service	Satisfaction level of inflight service
Cleanliness	Satisfaction level of Cleanliness
Departure Delay in Minutes	Departure Delay in Minutes
Arrival Delay in Minutes	Minutes delayed when Arrival
Satisfaction	Airline satisfaction level (Satisfied, ‘neutral or dissatisfied’)

4. Data Pre-processing

Upon closer examination of the satisfaction variables (‘Inflight wifi service’, ‘Departure/Arrival 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’, ‘Check-in service’, ‘Inflight service’ and ‘Cleanliness’), it was noted that the value zero was present. Given that the satisfaction variables follow an ordinal scale ranging from 1 to 5, the zero value was identified as an anomaly and treated as a missing value. Subsequently, a decision was made to replace the zero value with 1, which corresponds to the lowest level of satisfaction on the ordinal scale. This adjustment ensures the dataset aligns with the intended scale of satisfaction ratings.

Also, There were 310 missing values in the variable, “Arrival Delay in Minutes” and they were imputed by using the median.

5. Results of the Descriptive Analysis

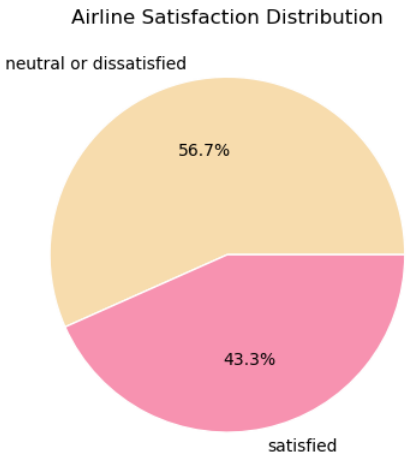


Figure 1: Distribution of Passenger Satisfaction

The response distribution looks very interesting because only less than the half of this airline's passengers are satisfied with the Airline company. service quality, encompassing aspects such as in-flight amenities, punctuality, plays a pivotal role in shaping passenger satisfaction. If any of these components falls short of expectations, it can significantly impact overall satisfaction rates. However, there can be some hidden features associated with this result.

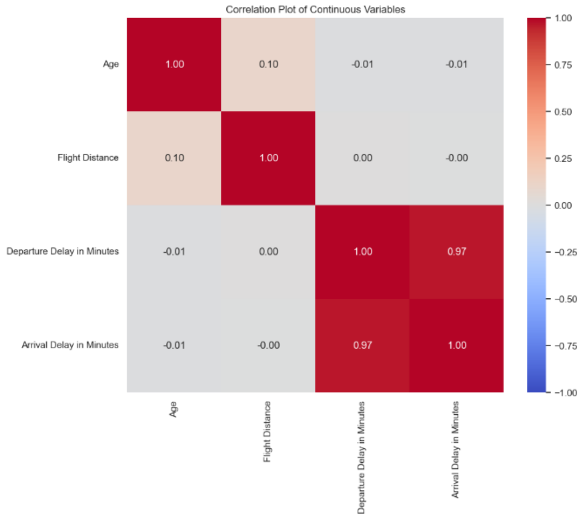


Figure 2: Pearsons Correlation Plot

If the flight of the airline's customers was delayed by a certain amount of time at departure, then the flight will be delayed by about the same amount of time at landing (provided that the aircraft does not accelerate in flight to make up for lost time). This linear association between Arrival delay and the Departure delay can be clearly understood from the Pearsons correlation plot of continuous variables of this dataset. The Pearson’s Correlation Matrix of Continuous Variables provides valuable insights into the relationships between various explanatory variables and AQI values. Based on the observed correlation values, we can draw important conclusions about the associations between these continuous variables. Although Departure delay and Arrival delay showed a strong correlation, all other variables have shown weak associations.

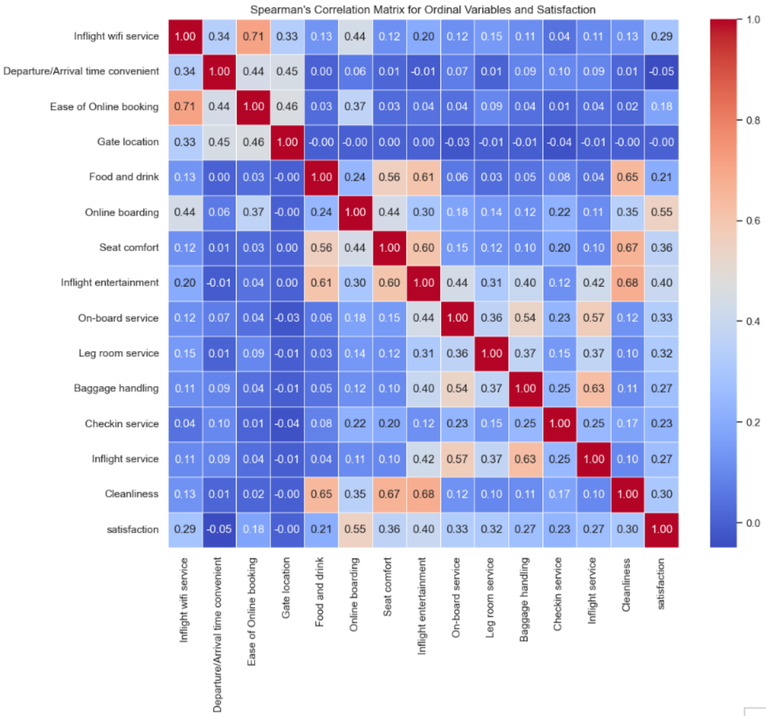


Figure 3: Spearman's Correlation Plot

Spearman's Correlation for Ordinal Variables and for our response satisfaction shows some moderately correlated variables. 'Food and drink', 'Seat Comfort' and 'Inflight entertainment' have shown moderate correlation with Cleanliness. Airlines that prioritize cleanliness in seating areas, tray tables, and overall cabin appearance are likely to extend the same commitment to other aspects, such as providing comfortable seating, quality food and drink options, and a diverse inflight entertainment selection.

When the satisfaction level is divided according to customer class, it can be found that the overall satisfaction of economy class passengers is significantly lower than that of business class passengers. Most business class passengers are satisfied with the service, while most economy class travelers tend to be neutral or dissatisfied with the service.

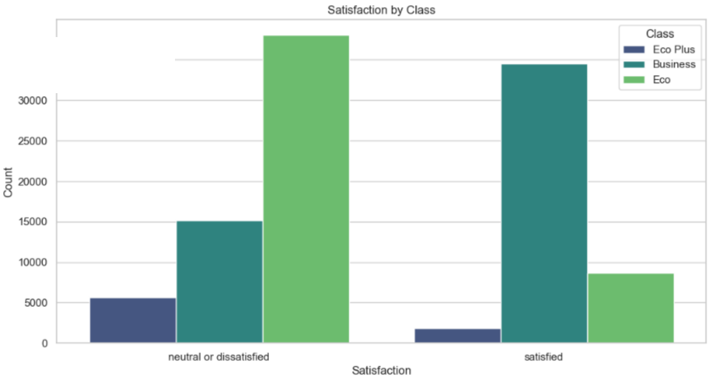


Figure 4: Distribution of Satisfaction by Class

To determine customer loyalty or disloyalty, a company would typically look for specific indicators or variables that reflect a customer's behavior, preferences, or interactions with the company. Therefore, this customer type is highly crucial for the overall satisfaction of the passenger about the airline. Previous research has found that being satisfied as a customer doesn't always mean you're loyal. Despite this, there's a strong connection between customer satisfaction and loyalty. Therefore, analyzing the associations with our response 'satisfaction' by removing the effect of customer type ('Loyal Customer', 'disloyal Customer') would be a smart decision.

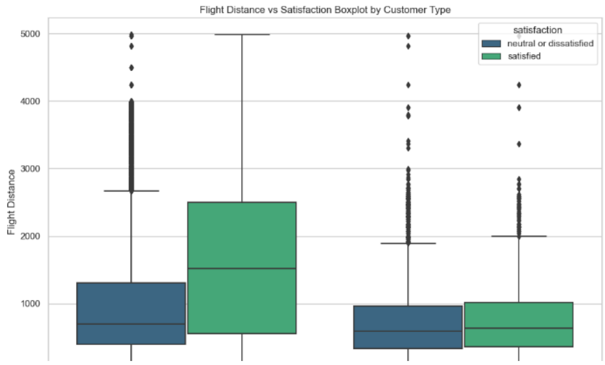


Figure 5: Boxplot of Flight distance Vs. Satisfaction by Customer Type

In both cases (loyal and disloyal), customer satisfaction has increased when the flight distance increases. The major reason for this result is longer flights often come with higher ticket prices. Therefore, customers may associate the increased cost with a higher level of service, amenities, or overall value. As a result, they may feel more satisfied with their experience.

Radar chart shows the average satisfaction score of 14 variables for passengers in customer type. The average scores of loyal customers and disloyal customers on Departure/Arrival time convenient, Seat comfort, online boarding is very different. Loyal customers may have a history of positive experiences with the airline, including comfortable seating arrangements. Therefore, average satisfaction score for seat comfort is higher from loyal customers compared to disloyal ones who may not have experienced the same level of satisfaction.

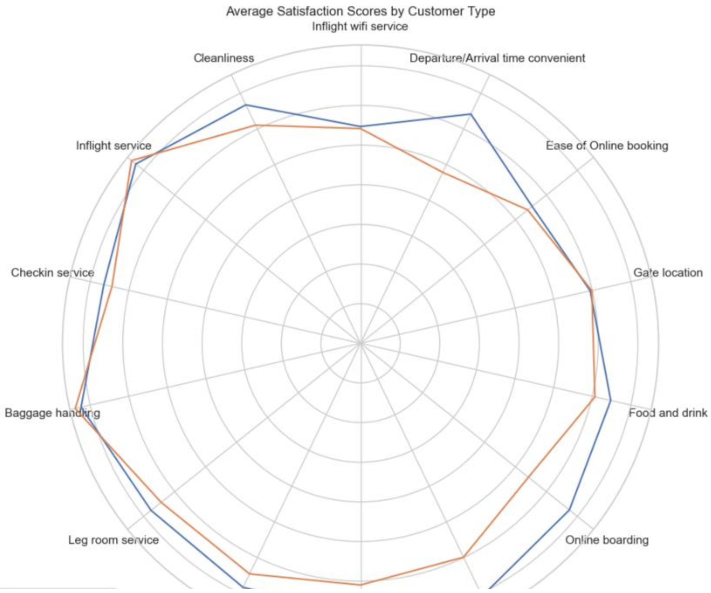


Figure 6: Radar Chart show of Average Satisfaction Scores by customer type

However, there is no significant different in Baggage handling Check-in service, inflight service and gate location. Most of the times these features will be same for both loyal and disloyal groups.

5.1. Further Analysis

Partial Least Squares Analysis

Partial Least Square Regression was performed on the data to identify any clusters among the observations as well as to identify any significantly correlated predictors.

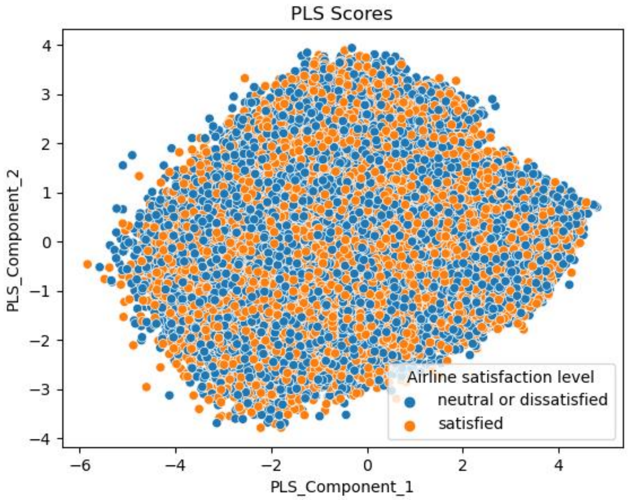


Figure 7: PLS Score Plot

The partial least squares regression model will be fitted with the response variable and the first component accounts for 16.23% of the total variation while the second component explains 6.83% of the variation. However, the score plot does not show distinct groupings of observations.

The graphical representation of loadings indicates the possibility of variable clusters existing within the data. When predictor points appear closely positioned in the loadings plot, it signifies a strong correlation among these predictors. This high correlation among predictors might result in multicollinearity problems when using logistic regression models. Addressing such issues might involve the removal of one or more highly correlated predictors or applying techniques that reduce dimensionality to enhance the model's performance.

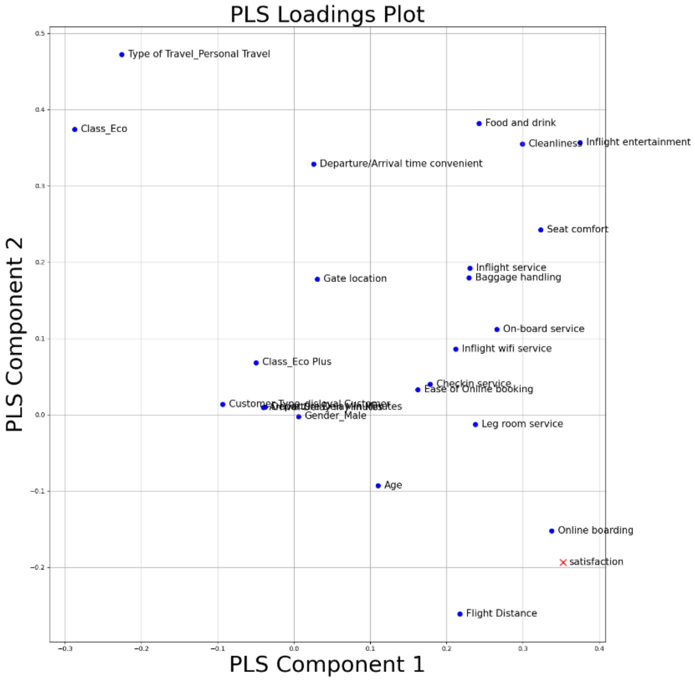


Figure 8: PLS Loading Plot

The proximity of certain predictor variables, such as online boarding, flight distance, and passenger age, to the response variable "satisfaction" in the loadings plot suggests a potential association between these factors and passenger satisfaction.

According to the researchers Online boarding, a convenient method for check-in and boarding without queuing at the airport, has been linked to increased satisfaction in travel experiences (Meidan & Galon, 2010). Studies have highlighted that passengers utilizing online boarding tend to exhibit higher overall satisfaction levels despite other factors like flight delays or cancellations.

And also, Passenger age has also been shown to be correlated with satisfaction. Older passengers tend to be more satisfied with their travel experience than younger passengers. This may be because older passengers are more likely to have realistic expectations about travel and are less likely to be bothered by minor inconveniences. Additionally, older passengers may be more likely to appreciate the value of airline services. For example, a study by Ye and Law (2010) found that older passengers were more satisfied with the price-quality ratio of airline services than younger passengers.

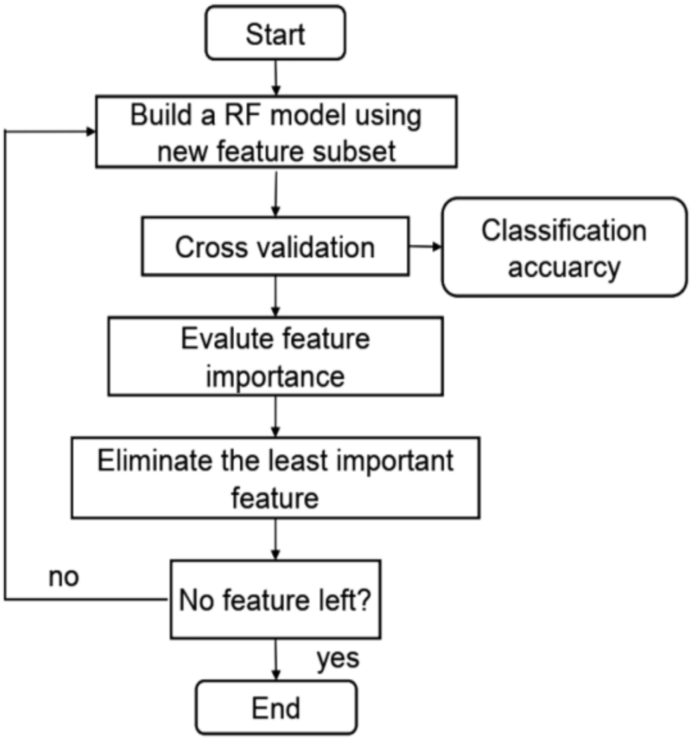
6. Advanced Analysis

Logistic Regression

In the initial phase, a logistic regression model with a ridge penalty was employed. Recognizing the significance of feature selection, Recursive Feature Elimination based on Random Forest (RF-RFE) was used to identify a subset of the most informative features.

Recursive Feature Elimination based on Random Forest (RF-RFE) is a potent approach within statistical learning, blending the ensemble power of Random Forest with iterative feature selection. This methodology employs bootstrap sampling and decision tree construction to create an ensemble of models, with each tree contributing to feature importance scores. Through a backward selection process, features are systematically removed based on their importance, and model performance is evaluated using cross-validation at each iteration. RF-RFE iteratively refines the feature subset, ensuring that the least informative features are progressively eliminated. This technique provides a robust means of identifying the most impactful features for classification accuracy, making it invaluable for enhancing model interpretability and predictive performance in diverse domains.

In our comparative analysis, two passive scenarios were considered: a logistic regression model with a ridge penalty was passively fitted using the features passively selected by RF-RFE, and a separate model was passively fitted using all the features initially considered.


Start
Build a RF model using new feature subset
Cross validation
Classification accuarcy
Evalute feature importance
Eliminate the least important feature
No feature left?
no
yes
End

Using RF-RFE feature selection		Without using RF-RFE feature selection	
	Accuracy		Accuracy
Training Set	0.89	Training Set	0.92
Test Set	0.89	Test Set	0.92

	Accuracy
Training set	0.75
Testing set	0.70

As seen from the table above, the full model performs better than the reduced model. Nevertheless, adherence to the parsimony principle prompts a preference for the reduced model, given its only marginal sacrifice in performance.

KNN Classifier

Subsequently, a K-Nearest Neighbors (KNN) classifier was employed; however, its performance fell considerably short when compared to the preceding logistic regression model.

	Accuracy
Training set	0.91
Testing set	0.90

Random Forest Classifier

A default Random Forest classifier was fitted, revealing a notable improvement in performance compared to previously employed models. The Random Forest's effectiveness underscores its robustness in handling complex relationships within data. The below table shows the results of the default random forest model.

To enhance performance, we conducted hyperparameter tuning on the Random Forest model using a randomizedCV grid search strategy, and the results were promising. The adjustment of these parameters contributed to a notable improvement in overall performance, underscoring the significance of fine-tuning model settings for optimal outcomes.

Optimal Hyperparameters

'n_estimators':10

'max_depth':30

'min_samples_split':5

'min_samples_leaf':4

'max_features': None

	Accuracy
Training set	0.93
Test set	0.92

XG Boost and Ada Boost Classifiers

XGBoost and AdaBoost are both boosting algorithms used in machine learning. XGBoost, short for Extreme Gradient Boosting, is known for its efficiency and speed, employing a gradient boosting framework to build an ensemble of decision trees. On the other hand, AdaBoost, or Adaptive Boosting, focuses on sequentially improving the weaknesses of a base model by assigning different weights to misclassified instances, creating a strong ensemble learner.

XG Boost		Ada Boost	
	Accuracy		Accuracy
Training Set	0.90	Training Set	0.90
Test Set	0.89	Test Set	0.90

The performance of XGBoost and AdaBoost, as evident in the provided table, did not surpass that of the Random Forest model after hyperparameter tuning.

Voting and stacking classifier

Here, a Voting Regressor is constructed by combining the predictions of four different regression models: Random Forest, Logistic regression, XG Boost, and k-Nearest Neighbors (KNN). The Voting Regressor effectively takes a weighted average of the predictions from its constituent models to produce the final prediction.

Voting Classifier		Stacking classifier	
	Accuracy		Accuracy
Training Set	0.85	Training Set	0.87
Test Set	0.83	Test Set	0.86

The outcomes indicate that both Voting Classifier and Stacking classifier models yielded lower performance when compared to the hyperparameter-tuned Random Forest model. So, we can conclude that the best model out of the models that we have fitted is the hyperparameter tuned random forest model.

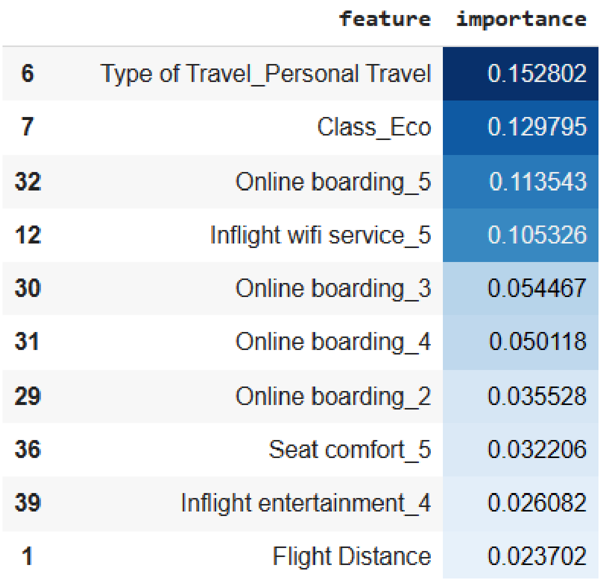
Now, let us see the interpretation of the best model.

Model Interpretation

Interpretability is the degree to which a human can understand the cause of a decision. The higher the interpretability of a machine learning model, the easier it is for someone to comprehend why certain decisions or predictions have been made. The need for interpretability arises from an incompleteness in problem formalization, which means that for certain problems or tasks it is not enough to get the prediction (the what). The model must also explain how it came to the prediction (the why), because a correct prediction only partially solves your original problem.

By uncovering the features and decision-making processes within the model, model interpretation allows airlines to identify specific service aspects that significantly impact passenger satisfaction. This knowledge empowers airlines to make informed enhancements, ensuring a more tailored and satisfactory travel experience for their passengers.

Global Model-Agnostic Methods



	feature	importance
6	Type of Travel_Personal Travel	0.152802
7	Class_Eco	0.129795
32	Online boarding_5	0.113543
12	Inflight wifi service_5	0.105326
30	Online boarding_3	0.054467
31	Online boarding_4	0.050118
29	Online boarding_2	0.035528
36	Seat comfort_5	0.032206
39	Inflight entertainment_4	0.026082
1	Flight Distance	0.023702

An essential aspect of our predictive model's interpretability is the examination of feature importance values.

The Random Forest model highlights that the type of travel, specifically for personal reasons, holds the utmost importance in determining passenger satisfaction. Following closely, the class of travel, particularly in Economy class , emerges as the second most influential factor. Notably, online boarding experiences with a rating of 5 and top-notch in-flight Wi-Fi service with a rating of 5 significantly contribute to overall satisfaction.

The partial dependence plot (short PDP or PD plot) shows the marginal effect one or two features have on the predicted outcome of a machine learning model. Let us look at the PDPs for 3 of the most important features of our Random Forest model.

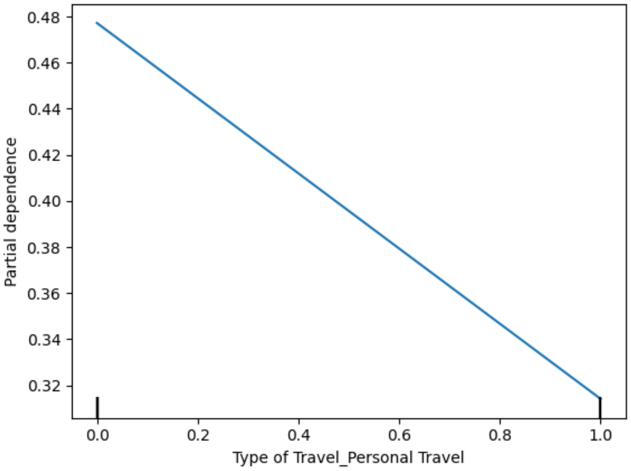


Figure 9: PDP of Type of Travel_Personal Travel

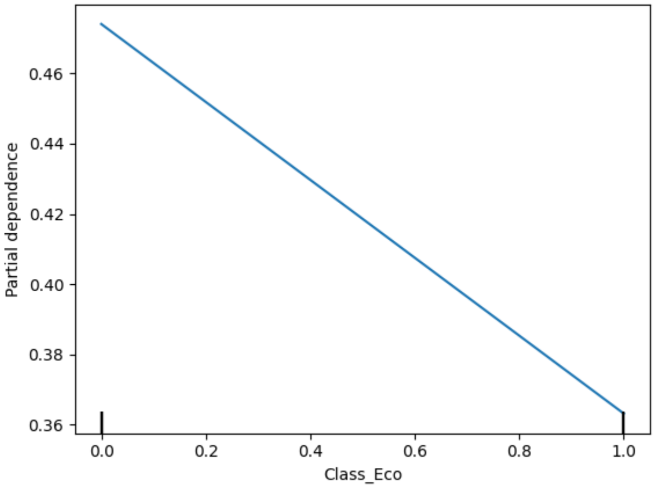


Figure 10 : PDP of Class_Eco

Local Model-Agnostic Methods

Local interpretation methods explain individual predictions. SHAP (SHapley Additive exPlanations) is a method to explain individual predictions. TreeSHAP, a variant of SHAP for tree-based machine learning models such as decision trees, random forest, and gradient

boosted trees. Using Force Plots, we can visualize feature attributions such as Shapley values as "forces". Each feature value is a force that either increases or decreases the prediction. The prediction starts from the baseline. The baseline for Shapley values is the average of all predictions.

Consider an example where the Type of travel is not personal, Class is not economy, Online boarding service was a rating of 4, inflight entertainment was a rating of 5, leg room service was a rating of 5. Note that the random forest model predicts the outcome as 'satisfied'. The force plot for this particular example is as follows.

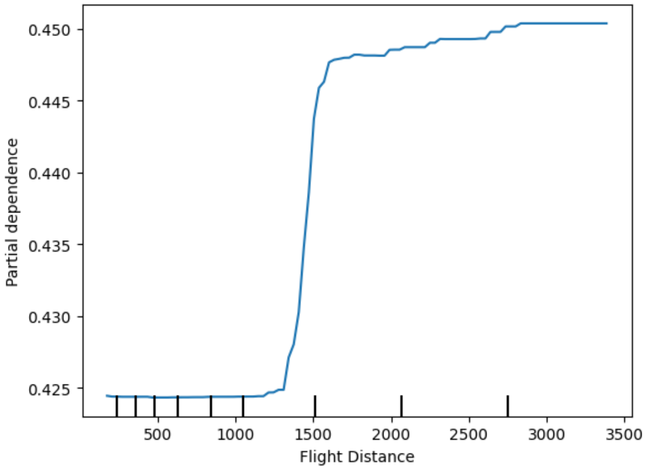


Figure 11: Partial dependency graph of Flight Distance

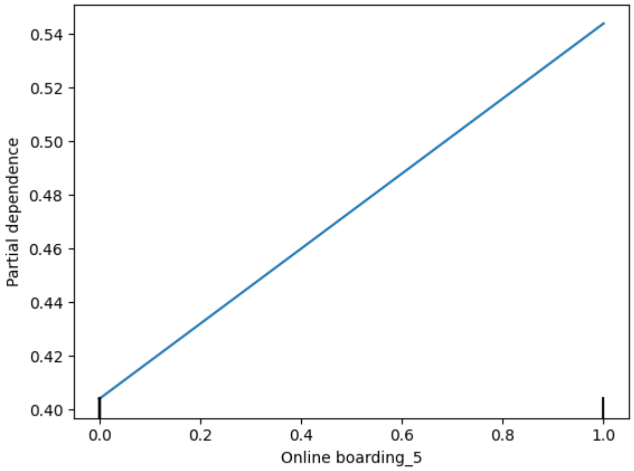


Figure 12: Partial dependency graph of Online boarding_5

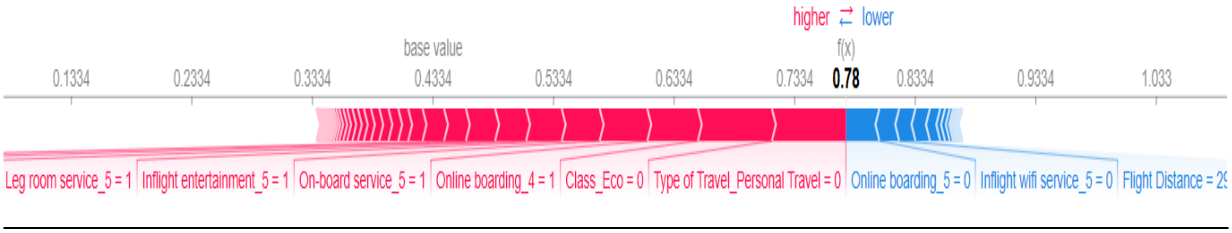


Figure 13: Shapley Force plot

The red bars that represent the mentioned variables and their values have made the predicted probability of being 'satisfied' increase. On the other hand, the blue bars have made the predicted probability of being in the 'satisfied' class decrease.

Issues Encountered and proposed solutions.

Recursive Feature Elimination based on Random Forest (RF-RFE) is a potent approach within statistical learning, blending the ensemble power of Random Forest with iterative feature selection. This methodology employs bootstrap sampling and decision tree construction to create an ensemble of models, with each tree contributing to feature important scores. Through a backward selection process, features are systematically removed based on their importance, and model performance is evaluated using cross-validation at each iteration. RF-RFE iteratively refines the feature subset, ensuring that the least informative features are progressively eliminated. This technique provides a robust means of identifying the most impactful features for classification accuracy, making it invaluable for enhancing model interpretability and predictive performance in diverse domains.

APPENDIX

StatlearningGroup2/Airline-Passenger-Satisfaction (github.com)