STA363SecAFinalProject

Rickey Huang

5/9/2021

Contents

List of Figures

List of Tables

Abstract

Section 1: Data and Motivation

Section 1.1: Goal of the Analysis

This is a Project exploring the cutomers' satisfaction of the US Ariline. The main goal of the analysis is both association and prediction. To be more specific, this project wants to understand which variables and how these variables affect the satisfaction of passengers who take a flight of US Airlnes. The data set used in this project is the result from a survey taken by passengers of US Airline.[1] In the survey, basic information of the passengers, like gender, age, and the flight information, like customer type, class, departure/arrival time were collected. Other than these objective data, some subjective scores are also asked in the survey, like ease of online booking, gate location, food and drink, and so on. All these scores are scaled in the range from 1 to 5.

Section 1.2: The Data Set

As taking a closer look at the data set, the Airline data set has 36728 observations and 23 variables. Among these variables, the response variable is *"satisfaction"* which is a categorical variable with two levels "satisfied" and "dissatisfiaction". In the rest of the 35 variables, 4 of them are numerical type data, which are *"Age"*, *"Flight.Distance"*, *"Departure.Delay.in.Minutes"*, and *"Arrival.Delay.in.Minutes"*, while the rest are all categorical variable variables. *"Age"* represents the age of the passengers who take the flight. *"Flight.Distance"* is the distance of the flight. *"Departure.Delay.in.Minutes"* is the time of delay for departure in minutes, and *"Arrival.Delay.in.Minutes"* is the time of dealy for arrival time in minutes. *"Gender"* of the passengers is a binary categorical variable. *"Type.ofTravel"* is also a binary categorical variable which has two levels "Personal Travel" or "Bussiness Travel". *"Class"* means the travel class in the plane in three levels "Bussiness", "Eco", and "Eco Plus". The variables like *"Inflight.wifi"* and *"Ease.of.Online.booking"* are categorical variables evaluating variables scored by passengers from 1 to 5, where 1 means least satisfied, while 5 means most satisfied. The original dataset can be accessed using the url: https://kaggle.com/teejmahal20/airline-passenger-satisfaction.

Section 2: Data Cleaning

Section 2.1: Missing Data

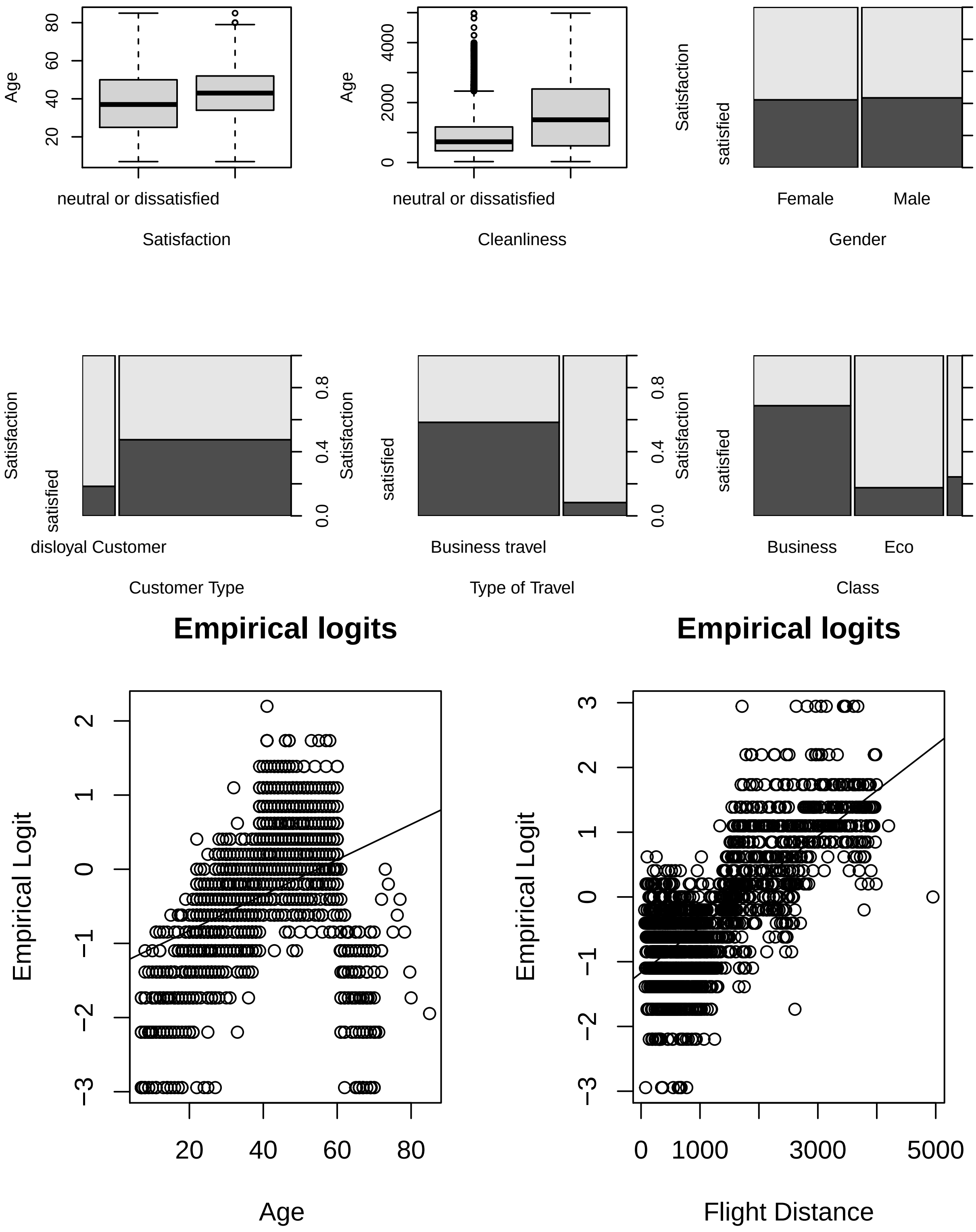
After checking any missing data in the data set, the Airline data is found to be completely observed, which means there is no missing data. Hence no imputation are required here.

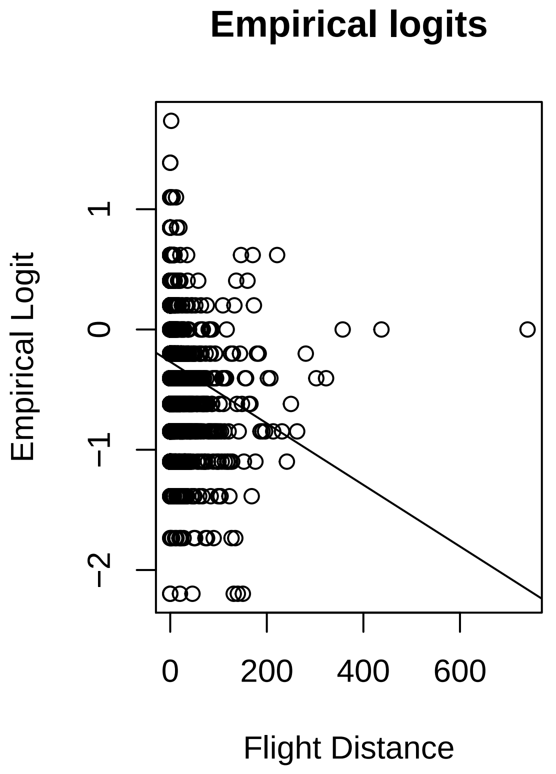
Section 2.2: Editing the Variables

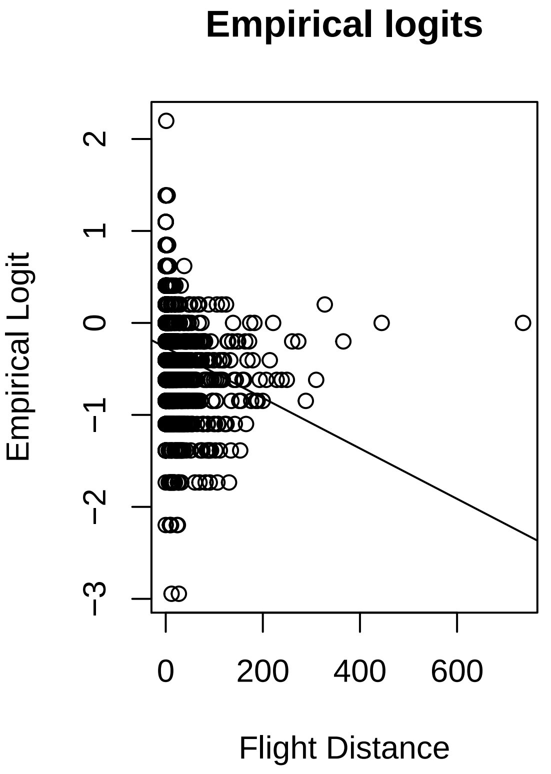
Since the response variable *"satisfaction"* is a binary categorical variable, and for the convenience of analysis with methods like logistic regression, a new variable *"satisfactin.num"* is created based on the original variable which include the same information as *"satisfaction"* but is recorded in a different way. "1" in *"satisfaction.num"* means "satisfied" in *"satisfaction"*, while "0" represents "neutral or dissatisfied".

[1] "Airline passenger satisfaction," 2020, online, Internet, 9 May 2021., Available: https://kaggle.com/teejmahal20/airline-passenger-satisfaction.

Section 3: Method 1:







Section 4: Method 2: Tree & Forest Models

Section 3.1: Introduction

Section 4.1.1 Reasons for considering another method

Even though the logistic regression model with BSS in the method 1 does the shrinkage and gives a regression line that reveals the association and makes prediction, the model is restricted by this single regression line. A transformation of the variables or a change in the form of the regression line would change the model a lot. Therefore, the tree model and forest model should be considered sicne they are not sensitive to the form of regression lines, while to tree model and forest model are not useing all of the variables in the model for the prediction, they also do the selection as the BSS does.

Section 4.1.2: How could the tree and forest model answer the reseach quesiton

Since the response variable *"satisfaction"* is a categorical data, and the goal of the project is to understand which variables affect the satisfaction (association) and how this variables affect the satisfaction (prediction). Both the classification tree model and classifcation random forest model are created in the method 1. Since the tree model has a clear and direct visualization to understand the association between variables, which can do the association job. The classification random forest model using bootstrapped samples with random subset of variables to train the model and make the prediction, so it could be a technique to sovle the prediction requirement. A combination of the tree model and forest model will be counted as the method to solve the research question here.

Section 4.2: Data Visualization (EDA)

Since the tree method and the forest model don't need to decide the type of the model like regression models, in which the form of the regression line need to be decide first after analyzing the correlation of the variables, the detailed relationship among variable are not necessary here for the tree and forest model. Instead, the distribution of the response variable is explored.

In order to explore the distribution of the response variable, a bar chart Figure 1 is created, because the response variable is a categorical variable. From the bar chart we can see that there are 20990 observations are at the “neutral or dissatisfied” level, while 15738 rows are at the “satisfied” level. Since there is not a level that has an extremely low frequency comparing to the other, the tree and forest model are safe to be used here.

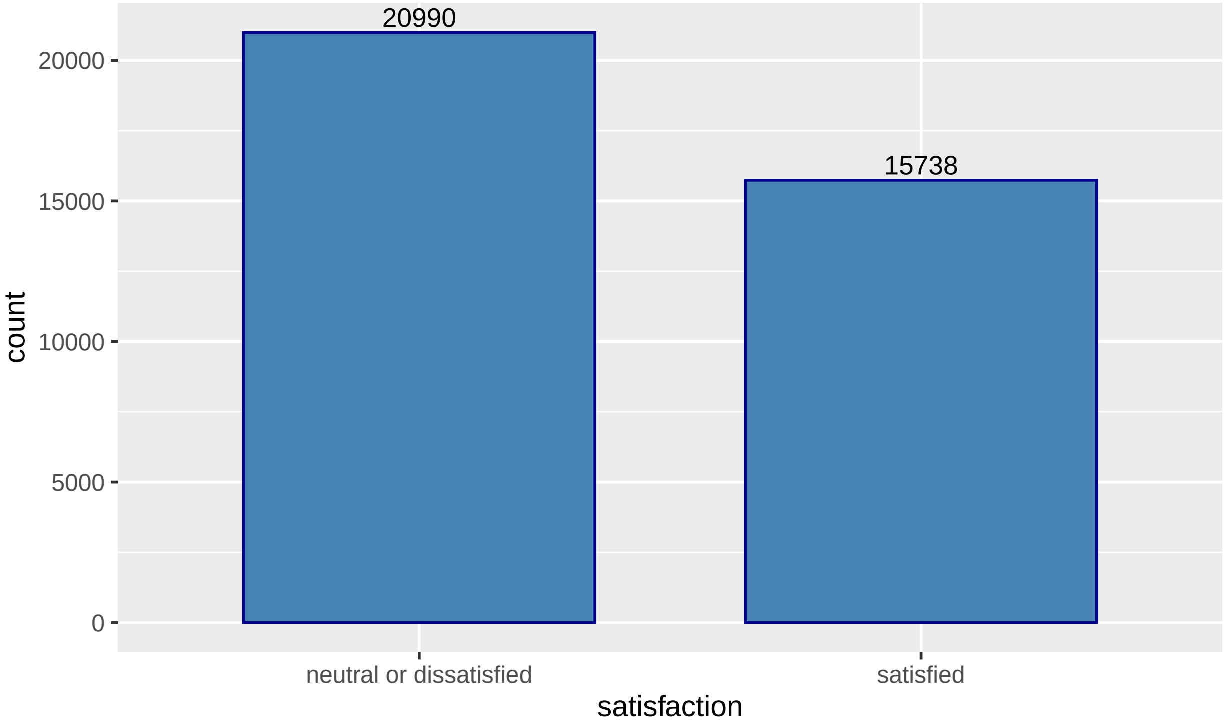


Figure 1: Distribution of the response variable

Section 4.3: Method

Section 4.3.1: Tree Model

Section 4.3.1.1: The Full Tree Model

First, a full tree model with all variables in the data set is trained. The fullTree model has a Root Node Error (RNE) of 0.4285, and the RNE in the classification tree model represents the Classification Error Rate (CER) of the root node. the the test CER for each split can be calculated using the Formula 1, which is the RNE times the percent change from RNE by each split. Since we want a low error rate, we want the to have a the test CER that is only a small portion of the RNE. The cp plot is created in Figure 2. it depicts how the change in the cp affects the test CER, and the number of splits are shown on the top of the plot. Hence, a number of split between 1 and 18 would create a comparatively small CER and with a appropriate number of splits.

$$test\ CER = RNE \times xerror \tag{1}$$

Section 4.3.1.2: Methodology for Grwoing and Pruning the Tree Model

In order to grow the classification tree model, the Gini Index of trees are computed and compared, and the fullTree model is created by minimizing the Gini Index, since a smaller Gini Index implies a more stable model. The Gini Index could be calculated using the Formula 2, where $|T|$ is the number of leafs in the tree,

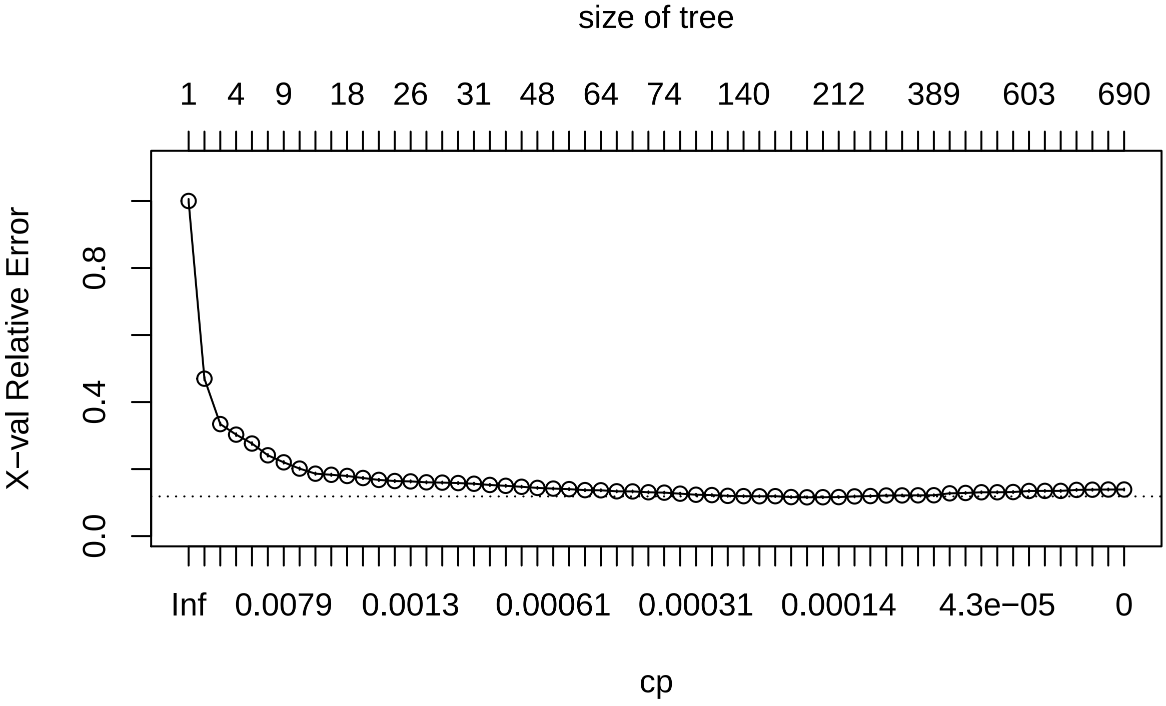


Figure 2: cp plot for fullTree

Table 1: cp table for the fullTree (excerpts)

CP	nsplit	rel error	xerror
7.8155e-03	8	0.238658	0.22010
5.8775e-03	12	0.197547	0.20142
3.5583e-03	14	0.185792	0.18630
3.1770e-03	15	0.182234	0.18300
3.1135e-03	17	0.175880	0.17950

and $G(Leaf_l)$ is the Impurity Score at a certain leaf, which can be computed using the Formula 3. In this formula, n represents the number of levels in the response variable.

$$Gini\ Index = \sum_{l=1}^{|T|} \frac{n_l}{n} G(Leaf_l) \tag{2}$$

$$G(Leaf_i) = 1 - \sum_{j=1}^{m} \hat{p}^2_{(Y=level_j, Leaf_i)} \tag{3}$$

In order to prune the tree, the a penalty term is added to the classification tree model with an optimal tuning parameter α. Then, the pruned tree model would be created by minimizing the penalized Gini Index.

Secction 4.3.1.3: Pruning the fullTree Model to create the Tree Model

Since the whole cp table is too long, only part of the cp table of fullTree is shown in the Table 1. To find an optimal α, the percent change from the RNE is compared, and a 14-split tree is chosen here since with further split, the percent change from the RNE is only 0.003 or smaller, so the further splits is not worthy. Hence, we have the optimal $\alpha = 0.185792$.

The Pruned Tree is shown in the Figure 3. This is a detailed visualization of the Tree model since the portion of data in each leaf is shown in this figure. Also a simplified visualization that is much easier to be explained is shown in the Figure 4. The darker a certain color is, the prediction is more stable.

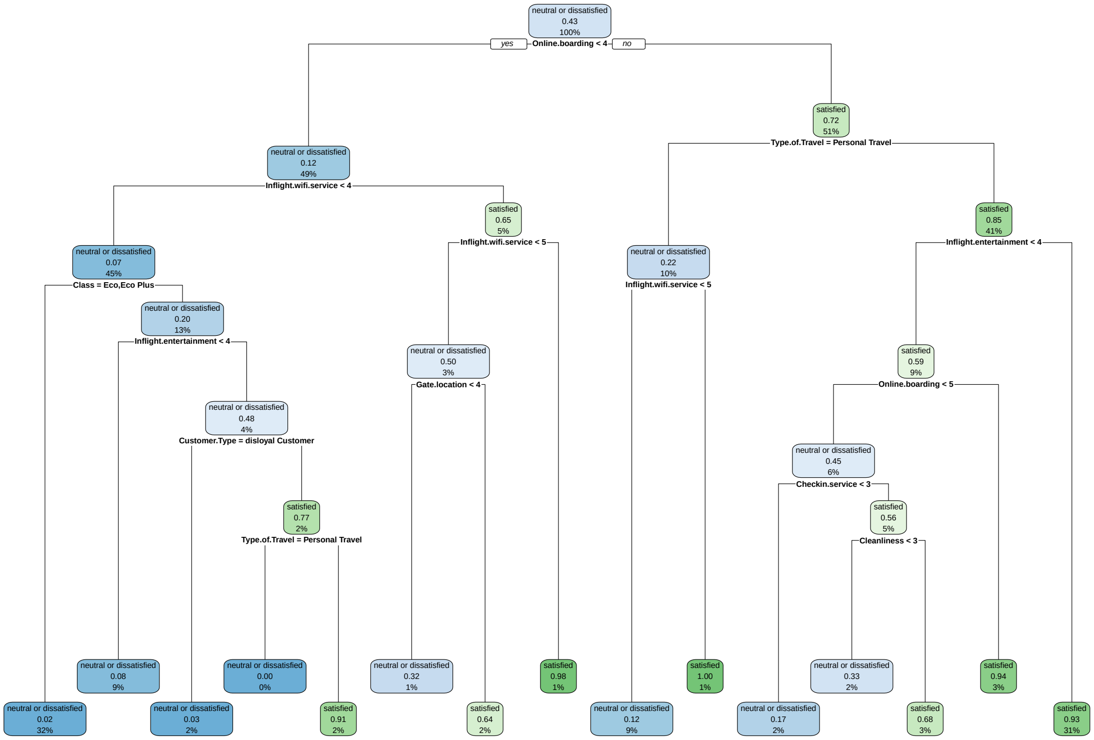


Figure 3: The Detailed Visualization for Tree

Conclusions

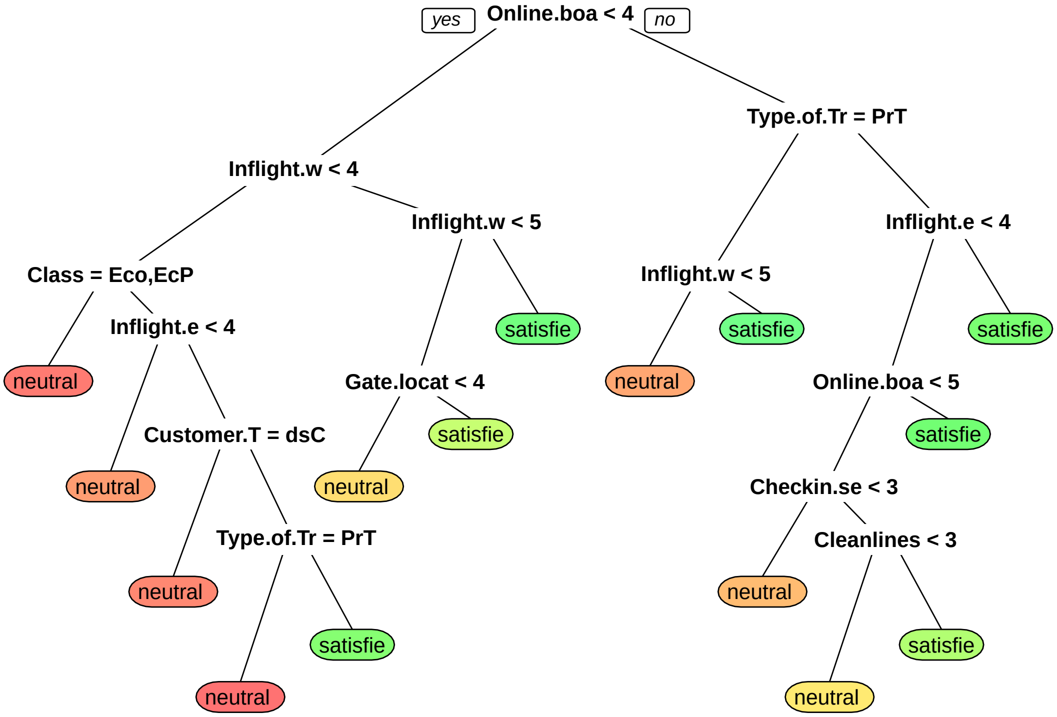


Figure 4: The Concise Visualization for Tree

Works Cited Page

"Airline passenger satisfaction," 2020. Online. Internet. 9 May 2021.. Available: https://kaggle.com/teejmahal20/airline-passenger-satisfaction.