Classification of the Car Evaluation dataset

ABHISHEK NAIK[1,*] **AND PRATIK JAIN**[2,*]

[1] *School of Informatics and Computing, Bloomington, IN 47408, U.S.A.*
[2] *School of Informatics and Computing, Bloomington, IN 47408, U.S.A.*
[*] *Corresponding authors: ahnaik@indiana.edu, jainps@iu.edu*

Compiled April 27, 2017

The Car evaluation dataset consists of 1728 instances and 6 attributes. We train our classifier using a part of these records and then test it on the test dataset consisting of the remaining records. On the basis of this classification, we aim to determine the class of the new Acceptability attribute.

Keywords: Data Mining, CSCI-B565, Classification, Car Evaluation, Random forest, decision trees, Mosaic plots, Box plots

INTRODUCTION

In this project, we work upon the Car Evaluation dataset in order to predict its acceptability factor. This dataset consists of 1728 instances with 6 attributes. We start off with the preprocessing of the data wherein we convert the categorical attributes into numerical attributes. Post this, in order to gain further insight about the dataset, we proceed with the visualization of the dataset by plotting Box plots and Mosaic plots for various comibinations of attributes. These plots helped us understand the relationships and patterns existing in the dataset. Then we classify the dataset using Random Forest and Decision trees. We achieve accuracies close to 98 percent in both the cases. In the real world, we intuitively know few facts like the safer a car is, the more acceptable it would be; the higher the maintenance and cost of a car, the lesser its preference. By analyzing the dataset our aim is to justify these beliefs. Along with all these, we also aim to get a good understanding of some classifiers like the Decision Trees and Random Forest.

DATASET SELECTION AND DESCRIPTION

For this project, we are analyzing the UCI Car Evaluation dataset (https://archive.uci.edu.ics/ml/datasets/car+evaluation). This dataset has 1728 instances, consisting of 6 attributes - buying, maintenance, doors, persons, luggage boot and safety. Based on these 6 attributes, we would be predicting 4 types of class values - unacceptable, acceptable, good and very good. Besides our personal interest in cars, the dataset is of the perfect size in the sense that it is neither too heavy, nor too small. It is uniform as well and its attributes have been explained well enough for us to do a good and detailed analysis. This dataset also gives us good insight as to what factors are important for concluding how acceptable a car is. Lastly, the data is good enough to be worked upon and represented as a matrix of numbers or strings.

Table 1. Class Distribution

Class	N	N[%]
unacc	1210	70.023
acc	384	22.22
good	69	3.993
v-good	65	3.762

The description for this dataset can be given as follows:

Number of attributes: 6

Attribute values:
Buying: v-high, high, med, low
Maint: v-high, high, med, low
Doors: 2,3,4,5-more
Persons: 2,4,more
`lug_boot`: small, med, big
safety: low, med, high

Missing attribute values: None
Table 1 gives the Class Distribution.

VISUALIZATION

As a part of preprocessing, for making the data uniform, we converted the data so that v-high was mapped to 4, high to 3, med to 2 and low to 1. We just converted all the categorical values into numeric ones. Furthermore, in order to get a good visualization of the data, we selected two plots - Box plot and Mosaic plot. Descriptions of the same, as well as their role in

our project, have been highlighted below.

Box Plot

Box plots [1] are used to display groups of numerical data by means of their quartiles. They also denote variability outside the upper and lower quartiles by means of whiskers. They thus enable us to represent outliers as well. In our project, we found the box plot between various pairs of attributes to find the correlation amongst them. By looking at the box plot between acceptability and safety, shown in Figure 1, we can conclude that as the safety of the car increases, the average acceptance also increases. Similarly, as shown in Figure 2, the box plot between buying cost and acceptability shows that as the cost increases, the acceptability decreases.

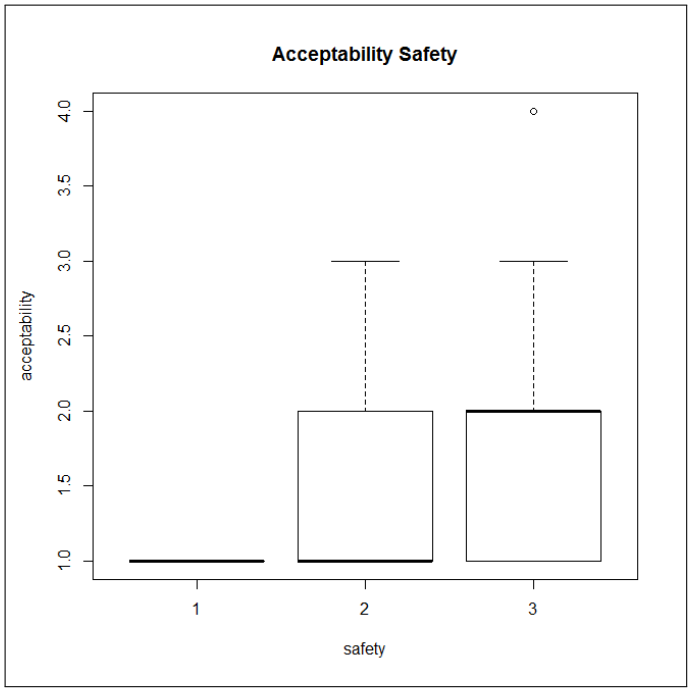

Fig. 1. Acceptability, safety and maintenance Mosaic plots

Mosaic Plot

Mosaic plots [2] are used for data visualization from a qualitative perspective. In case of these plots, the area of the tiles is directly proportional to the number of observations for that particular category. The variables that are displayed may be either categorical or ordinal. The minimum number of variables is 2 without any upper limit, but using many variables makes the visualization cramped. The drawback of Mosaic plots is that it is not possible to plot a confidence interval.

In our project, we have plotted the Mosaic plots of acceptability with all the other pairs of attributes. Figures 3, 4 and 5 are 3 of these plots. The first figure is the Mosaic plot of acceptability, safety and maintenance. It shows that cars that have low safety are unacceptable. Also, the cars with medium safety and high maintenance are not considered to be good or very good in terms of acceptability. These observations thus confirm to our general understanding that less safe vehicles should not be acceptable. The second figure is the Mosaic plot of persons, acceptability and buying. It shows that two-people cars are found to be unacceptable, which goes with the generic global trend. The third figure

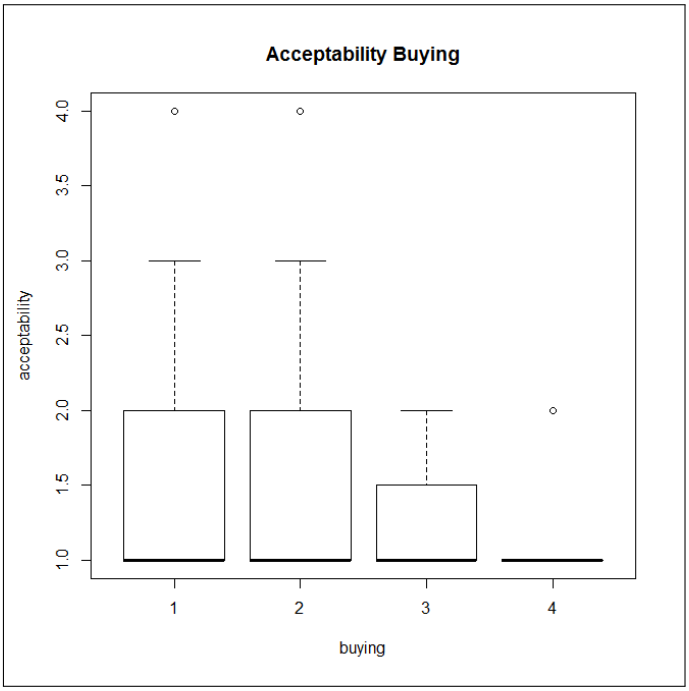

Fig. 2. Acceptability, safety and maintenance Mosaic plots

is the Mosaic plot of acceptability, maintenance and buying. This graph shows that when the maintenance is very high, the car is generally unacceptable; and the acceptability is never very good or even good. The number of cars that have high maintenance and either a high or very high buying cost is less, which justifies the general fact that such cars are generally not preferred.

CLASSIFIERS USED

In our project, we have divided our dataset into two separate parts - one for the training and the other for testing. We created these parts by randomly sampling the dataset in hand and creating two equal halves of the data. We used the following classifiers:

Decision Tree

Decision Trees form a tree-like graph to aid the process of decision making. In a decision tree, an *internal node* represents a test to be carried out, *branches* denote the outcomes of the test, *leaf nodes* represent the class label and the *paths* denote the classification rules. They are easy to understand since they follow a linear approach. They are advantageous because they have outcomes (branches) even for sparse data [3].

In our project, for classifying the dataset, we used the *rpart* library for creating a decision tree classifier. The plot for this classifier is as shown in Figure 6 which shows that maximum records can be split on the basis of the safety and buying cost attributes. In all our subsequent runs, we got an average error rate of 0.027.

Random Forest

Random Forests, also known as Random Decision Forests, are an *ensemble learning* method. They create many decision trees during training and subsequently use it for classification and regression. We chose Random Forest since they avoid over-fitting caused by Decision Trees [4]. It makes use of bootstrap samples

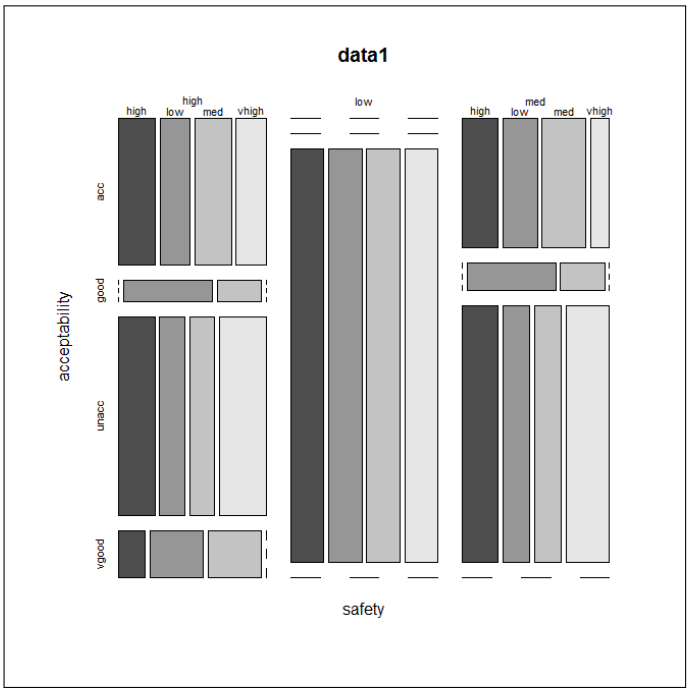

Fig. 3. Acceptability, safety and maintenance Mosaic plots

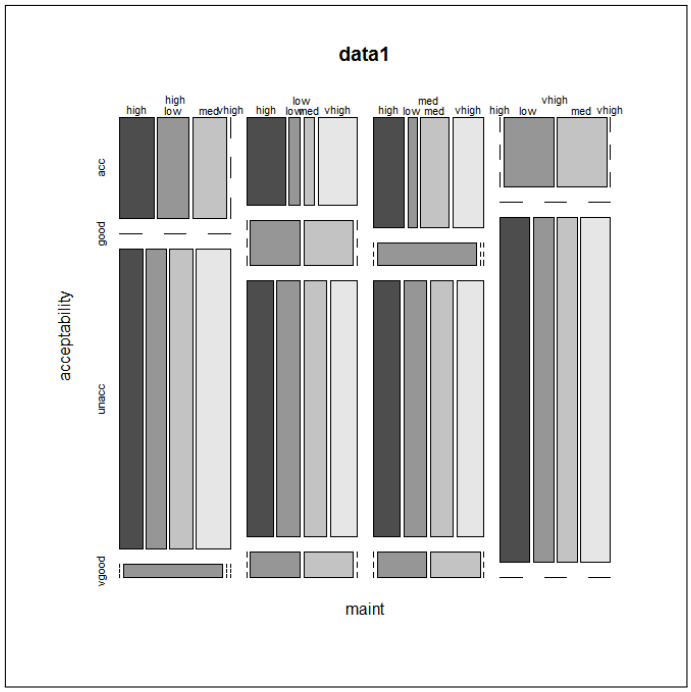

Fig. 5. Acceptability, maintenance and buying mosaic plots

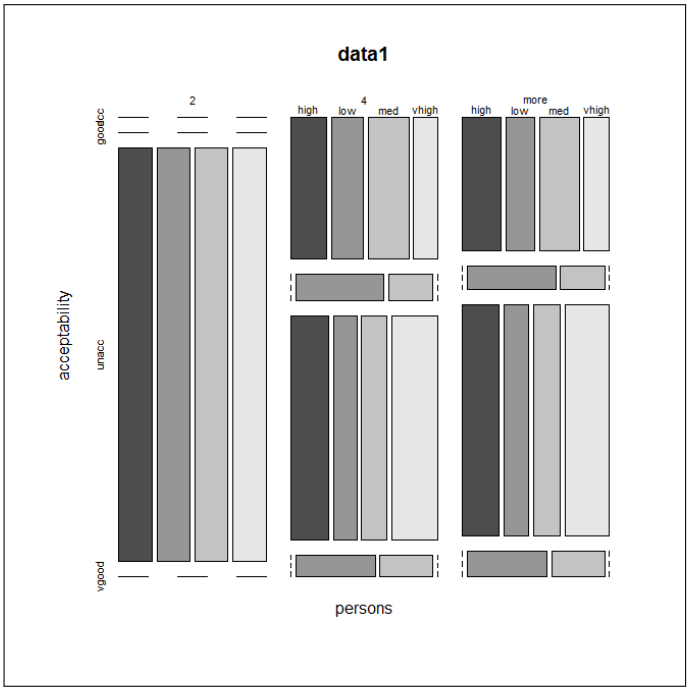

Fig. 4. Persons, acceptability and buying mosaic plots

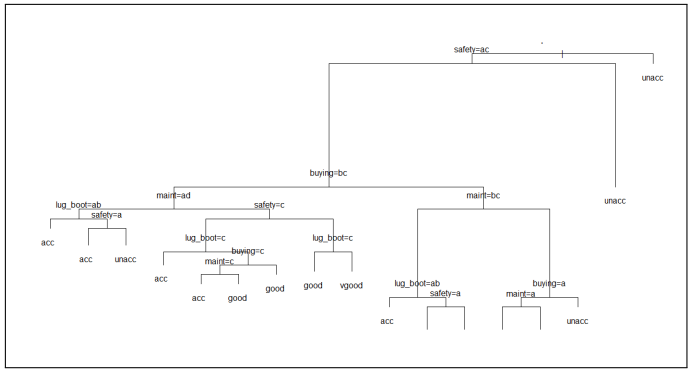

Fig. 6. Decision Tree Classifier

of the training data along with randomized feature selection. A prediction is then made by either finding out the majority of the votes or averaging the predictions of the ensemble. It usually delivers highly accurate performance [5].

We have used the Random Forest library for creating this classifier and observed an average error rate of 0.022. Figure 7 shows the error rates for the various number of trees used while classifying the dataset.

When we started off with the project, we had taken equal size subsets for both the training as well as testing purposes. Gradually, we varied the proportions of the training-to-testing data and observed the results. The results showed us that as the proportionality of the training data increased, the classifier accuracy also increased in both the cases.

The entire code for the project has been included in a file named 'project.r'.

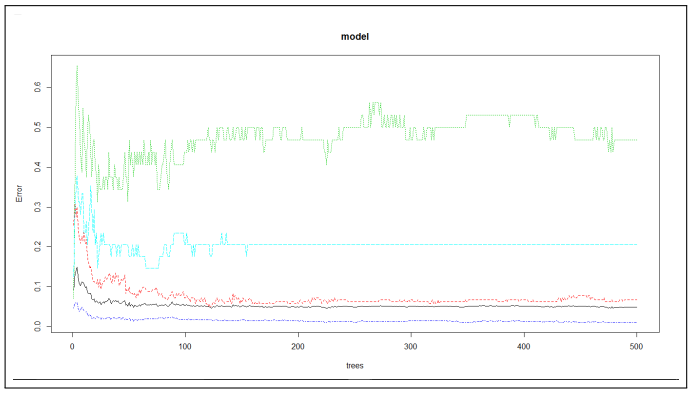

Fig. 7. Random Forest Classifier

CONCLUSION

In this project, we studied the Car Evaluation project and analyzed and analyzed its various attributes and their correlations. We initially took the Car Evaluation dataset and preprocessed it by taking various actions like converting the categorical attributes into numerical attributes. We then visualized the dataset by plotting various graphs. After this, we created classifiers to determine the unknown attribute 'Acceptability'. We achieved high accuracy for both the classifiers.

REFERENCES

[1] Wikipedia, "Box plot," Web page, online; accessed 26-Mar-2017. [Online]. Available: https://en.wikipedia.org/wiki/Box_plot

[2] Wikipedia, "Mosiac plot," Web page, online; accessed 26-Mar-2017. [Online]. Available: https://en.wikipedia.org/wiki/Mosaic_plot

[3] Wikipedia, "Decision tree," Web page, online; accessed 24-Mar-2017. [Online]. Available: https://en.wikipedia.org/wiki/Decision_tree

[4] Wikipedia, "Random forest," Web page, online; accessed 25-Mar-2017. [Online]. Available: https://en.wikipedia.org/wiki/Random_forest

[5] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: http://dx.doi.org/10.1023/A:1010933404324