Car Evaluation ML Project

Justin Agudah, Romelo Seals
Department of Electrical Engineering and Computer Science
University of Missouri-Columbia
Columbia, MO 65201
(joagnp,rsb3x)@umsystem.edu

Abstract

The Car Evaluation dataset represents several different aspects of the car. We propose that the attributes of the car determine how much someone will buy a certain car. By utilizing the decision tree algorithm, splitting the data, and creating dummy data, we plan on seeing what criteria must be met for a car to be considered unacceptable, acceptable, good, and very good.

GitHub: https://github.com/jagudah/CarEvaluationML

1 Introduction

In today's world, various automobile companies, and manufacturers are always competing against each other within the automobile industry, making sure their automobiles surpass the automobiles of their competitors in terms of safety, fuel efficiency, capacity, comfort, technological advancement, and other features, all while selling them at a reasonable price.

However, the term “reasonable” differs from consumer to consumer, especially when it comes to buying automobiles. How does a consumer truly know if they are getting their money’s worth when deciding which car to buy? What are the main attributes of an automobile that automobile companies and manufacturers research and develop so they appeal more to customers? These questions emphasize the importance of the price and technical evaluations of automobiles because the rising prices of automobiles mean that better methods are needed to determine a car’s price and technical characteristics accurately. That is why we need better methods, in this case, machine learning algorithms, to estimate a car’s value, which is why we’ll be using classification algorithms for this ordeal, more specifically, decision trees.

Proposal

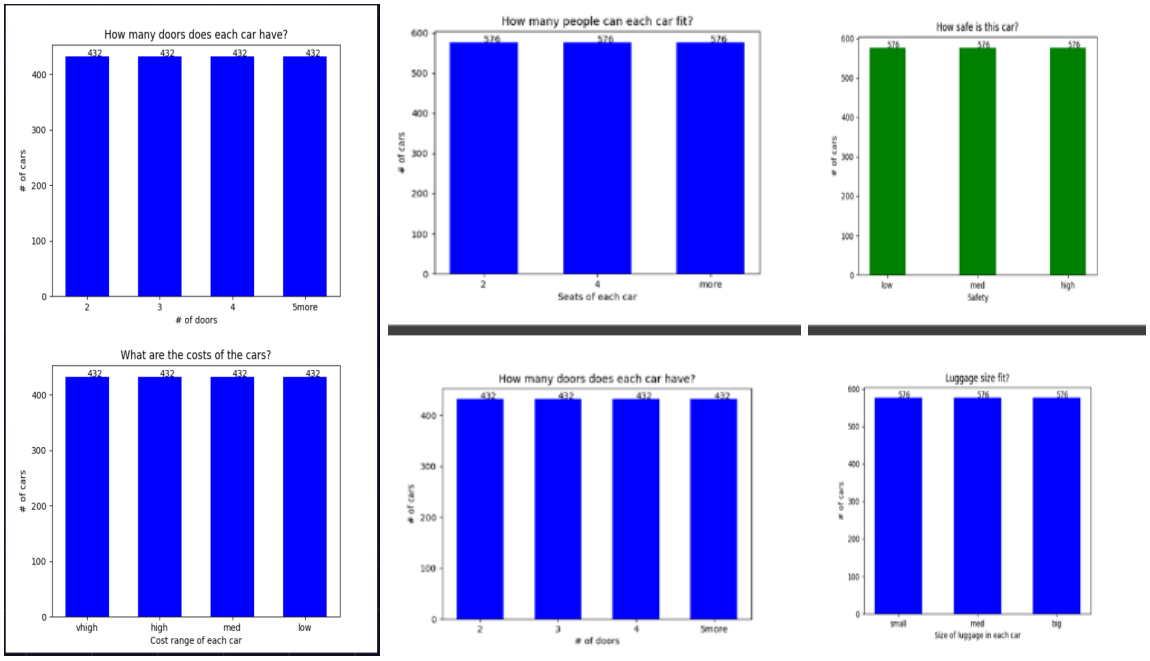
The real-world task and importance lie in the evaluation of the price and safety of cars because the rising prices of automobiles mean that better methods are needed to determine a car’s price and technical characteristics accurately. The machine learning method we plan on utilizing is decision tree classification and knn-classification for our project. The approach we plan on handling our tasks individually while having weekly meetings to see each other’s progress and work together. The timeline will consist of the individual's work time and weekly meetings, and the project should be completed in 3-4 weeks. The agile sprint framework was followed having weeks of sprints being our main plan.

Data Set Information

Car Evaluation Database was derived from a simple hierarchical decision model originally developed for the demonstration of DEX, M. Bohanec, V. Rajkovic: Expert system for decision making. Sistemica 1(1), pp. 145-157, 1990.). The model evaluates cars according to the following concept structure:

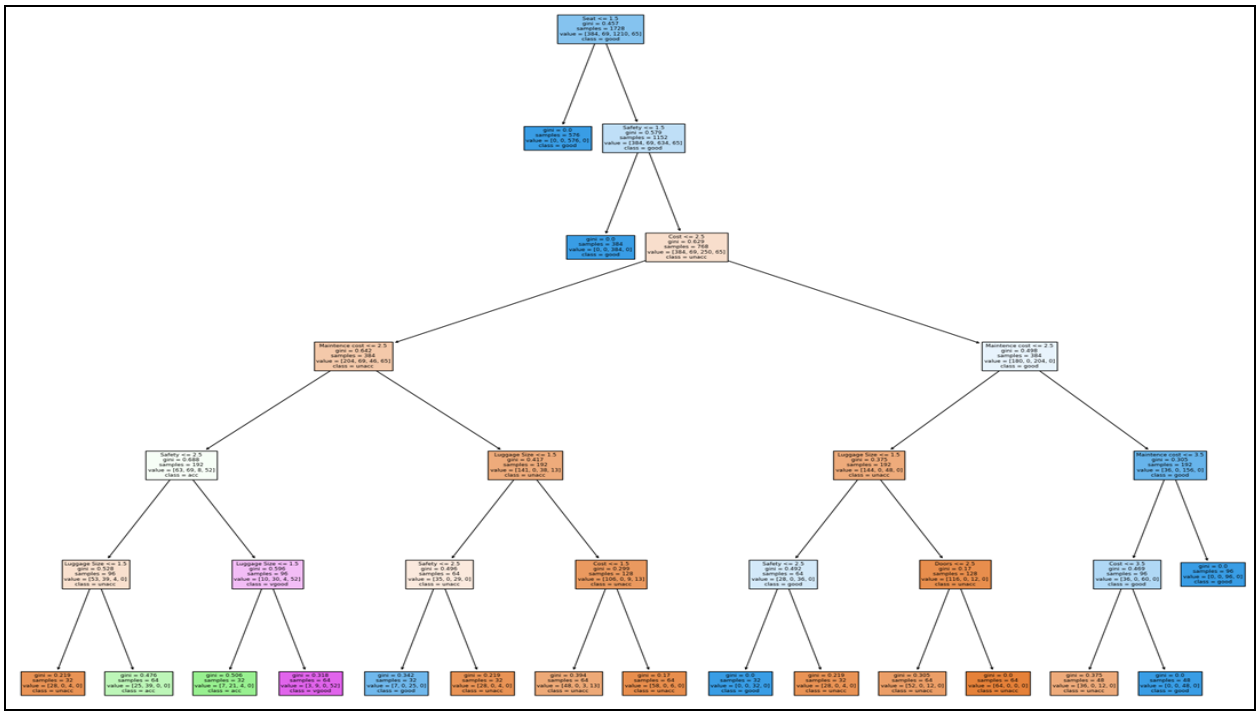
CAR	car acceptability	
. PRICE	overall price	
. . buying	buying price	v-high, high, med, low
. . maint	price of the maintenance	v-high, high, med, low
. TECH	technical characteristics	
. . COMFORT	comfort	
. . . doors	number of doors	2, 3, 4, 5-more
. . . persons	capacity in terms of persons to carry	2, 4, more
. . . lug_boot	the size of luggage boot	small, med, big
. . safety	estimated safety of the car	low, med, high

Input attributes are printed in lowercase. Besides the target concept (CAR), the model includes three intermediate concepts: PRICE, TECH, COMFORT. Every concept is in the original model related to its lower level descendants by a set of examples (for these examples sets see [Web Link]). The Car Evaluation Database contains examples with the structural information removed, i.e., directly relates CAR to the six input attributes: buying, maint, doors, persons, lug_boot, safety. Because of known underlying concept structure, this database may be particularly useful for testing constructive induction and structure discovery methods. In the images on the next page, it shows that the distribution of the characteristics is all equal, there's an equal amount of doors for each car, and so on



Formatting the data

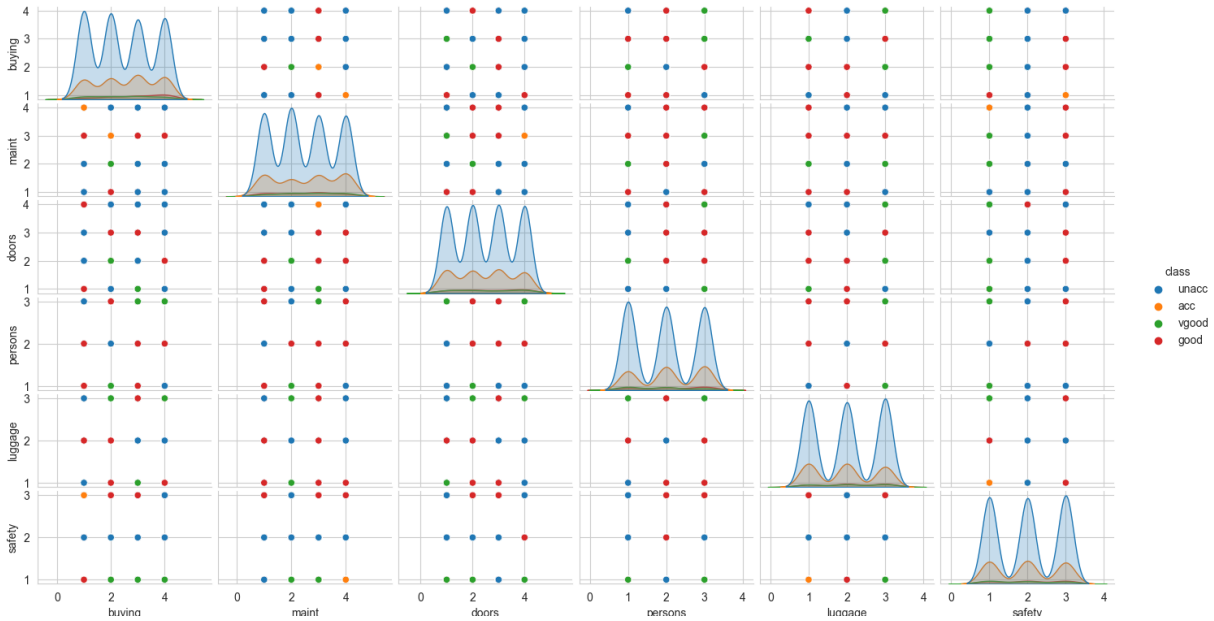
In order to run the classification algorithms, we had to turn the .txt file given to us by the dataset website into a CSV so we can run the pandas library on the CSV to make a data frame and 2D array. First, originally the plan created an array of objects known as cars object which contained all of the attributes that each car has. As the research went on, we realized that the idea did not work for the classification algorithms we were trying to run. So we went the long way of converting the data into an array of car objects into a 2D array of these attributes and then converting the 2D array into a data frame.



Classification

The two classifications we utilized were the decision tree classification and the knn-classification. On the decision tree classification, we used the entire dataset to create the decision tree, converting each one of the attributes of the cars into numbers, we were able to figure out which one of the attributes affected the class the car belongs to. A quick run-through of what the attributes were translated into, low represents 1, so if the buying price was low, that was a 1 on the buying price, if the people fitting into the car was 2, then this was translated into 1 as well since this was the lowest number the attribute could represent. Then running the algorithm, we noticed having too many levels made the classification illegible and figured out ways to lower the tree's height so the tree is more readable. For testing, we created another file that will ask what kind of car the person is asking for and then spit out the classification of the car. The next classification we ran was the knn-classification, for this classification, we did split the data in two but first, we randomized the data to see if we can get a higher guess rate after

realizing the inability to guess well when randomized we continued to use the dataset given then ran through the knn-classification. We first created a scatter plot which was very uninteresting and did not change anything about our findings then ran the knn-classification. As we ran testing to the predictions we found that 9 nearest neighbors were what we needed to have the highest prediction rate of 70-75%. The image below represents the scatter plot created.



Conclusion

According to the results from our classification algorithms, we concluded that the values for each attribute are evenly distributed among cars in the bar graphs, showing us a uniform distribution between the data sets. We can also conclude that there is a positive and direct correlation between a car's technical characteristics and price. Overall, more accurate and precise methods are needed to accurately determine the cost, comfort, and safety of various vehicles.

Acknowledgments

The first data set used in this project was obtained from https://archive.ics.uci.edu/ml/datasets/Car+Evaluation. This data set was created by Mark Bohanec and donated by Mark Bohanec and Blaz Zupan. Car Evaluation Database was derived from a simple hierarchical decision model.

References

- Bohanec, M. (1997, June 1). *Car Evaluation Data Set*. UCI Machine Learning Repository: Car Evaluation Data Set. https://archive.ics.uci.edu/ml/datasets/Car+Evaluation