Assignment-4: A Mathematical Essay on Decision Tree Classifier

Shania Mitra
Roll Number: CH18B067
Chemical Engineering
IIT Madras, Chennai
ch18b067@smail.iitm.ac.in

***Abstract*—In this study, we estimate the safety of a car based on factors such as buying price, maintenance price, capacity, etc. A decision tree classifier is used to model the importance of these factors and predict the safety of cars.**

***Index Terms*—Decision Tree, Visualization, Car Evaluation**

I. Introduction

This study empirically analyzes the factors affecting car safety using the Car Evaluation Database. We find that a number of factors, such as maintenance price, purchase price, luggage and seating capacity to significantly affect the safety category of cars.

A decision tree is a set of cascading questions. When we get a data point (i.e. set of features and values), we use each attribute (i.e. a value of a given feature of the data point) to answer a question. The answer to each question decides the next question.

In this study, we use Decision Trees to model the safety category of individuals based on maintenance price, purchase price, luggage and seating capacity etc.

We begin by gathering, cleaning and preparing the data, following which we perform exploratory analysis. Finally, we build statistical models and perform visualizations to provide quantitative and visual evidence of the relationships observed. In the next section, we highlight the key principles underlying Decision Tree. In section 3, we discuss the insight and observations drawn from the data and the models. Finally, in section 4 we outline the salient features of the study and present further avenues of possible investigation.

II. Decision Tree

Decision tree learning or induction of decision trees is one of the predictive modelling approaches used in statistics, data mining and machine learning. It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. Decision trees are among the most popular machine learning algorithms given their intelligibility and simplicity.

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data (but the resulting classification tree can be an input for decision making).

III. The problem

In this study, we estimate the safety of a car based on factors such as buying price, maintenance price, capacity, etc.

A. Data Preparation

The training dataset used in this study consists of 1728 cars and 6 features. Interpretation of the features is as follows:

- Buying: The price of purchase of the car - 'vhigh' 'high' 'med' 'low'
- Maintenance: The cost of maintenance of the car - 'vhigh' 'high' 'med' 'low'
- Persons: Seating capacity of the car - '2' '4' 'more'
- Doors: The number of doors in the car - '2' '3' '4' '5more'
- Lug_boot: The boot space of the car - 'small' 'med' 'big'
- Safety: 'low' 'med' 'high'
- Target: 'unacc' 'acc' 'vgood' 'good'

No missing values are present in the dataset. The aim is to predict the multiclass feature Target.

B. Exploratory Analysis

In this section we look at the trends between various features.

In Table I we have a description of the features and the count, number of unique values and mode in each of them. The dataset has 1728 cars, with no missing values.

In figure 1, we plot the count of each of the target classes and observe that the class "unacc" occurs the most frequently, followed by "acc".

In figure 2 we plot the count of maintenance cost classes split by the targets. It is observed that for each of the maintenance classes target "unacc" and "acc" occur most frequently. Further, in maintenance class "vhigh" no cars of target "vgood" or "good" occur. Thus, if a car has very high maintenance, the target would most likely not be "vgood" or "good".

	buying	maint	doors	persons	lug_boot	safety	Target
count	1728	1728	1728	1728	1728	1728	1728
unique	4	4	4	3	3	3	4
top	high	high	4	4	big	high	unacc
freq	432	432	432	576	576	576	1210

TABLE I
DESCRIPTION OF EACH FEATURE

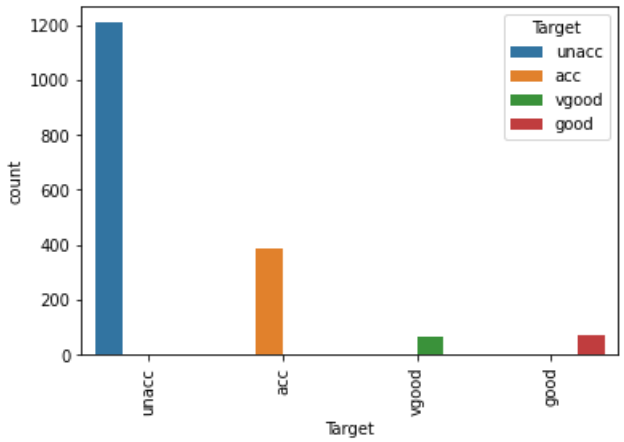

Fig. 1.

In figure 3, we plot the count of capacity classes split by the targets. It is observed that for each capacity, target "unacc" occurs most frequently. Further, with cars of capcity 2, only cars of target "unacc" occur. Thus, if a car has a capacity of 2, possibly a luxury or sports car, the target would most likely be "unacc".

In figure 4, we plot the count of lug_boot classes split by the targets. It is observed that for each capacity, target "unacc" and "acc" occur most frequently. Further, cars with small boot space no samples with target "vgood" occur. Thus, if a car has

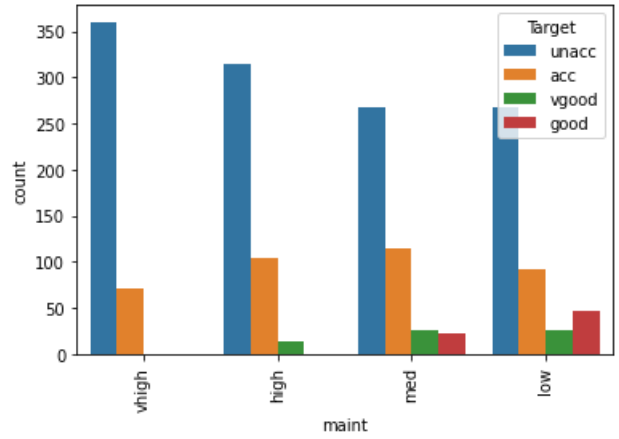

Fig. 2. Count plot of Maintenance Cost Classes split by Target

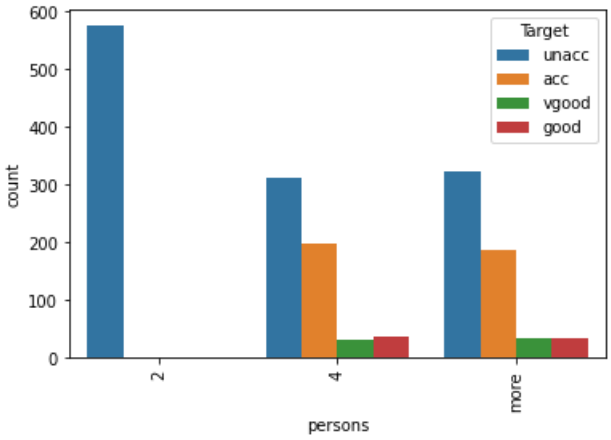

Fig. 3. Count plot of Number of persons split by Target

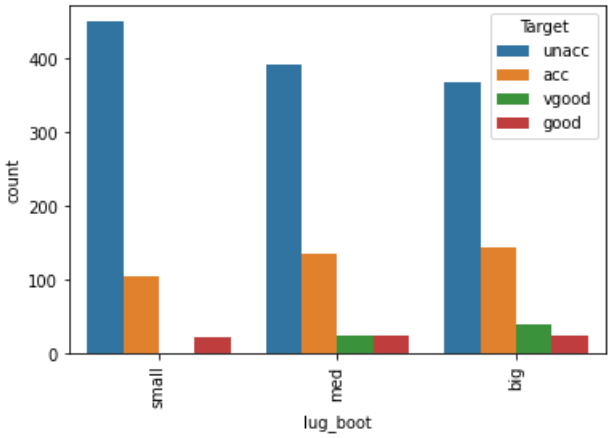

Fig. 4. Count plot of Lug_boot Classes split by Target

small bootpsace, such as an SUV, it most likely would not have a target of "vgood".

In figure 5sr, we plot the count of safety classes split by the targets. It is observed that for each safety, target "unacc" occurs most frequently. Further, cars with low safety have target as "unacc" only, while, cars with medium safety have no samples with target "vgood". Thus, if a car has a target of vgood, it will most likely have high safety.

It must be pointed out that the joint distribution of the features (i.e., except for Target) amongst themselves is uniform. For example, the count of cars in various buying and maint classes is the same as can be seen in the table below.

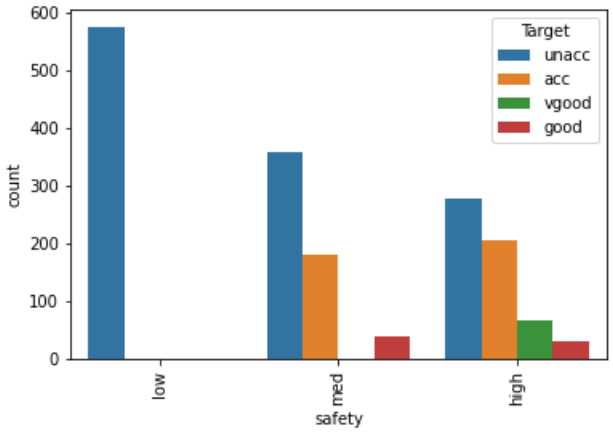

Fig. 5. Count plot of Safety Classes split by Target

```
buying  maint
high    high     108
        low      108
        med      108
        vhigh    108
low     high     108
        low      108
        med      108
        vhigh    108
med     high     108
        low      108
        med      108
        vhigh    108
vhigh   high     108
        low      108
        med      108
        vhigh    108
dtype: int64
```

C. Model

To train the model and gauge its performance, the dataset was split into train and test sets with an 80-20 split. Following are the metrics achieved by the Decision Tree Model:

	precision	recall	f1-score	support
acc	0.95	0.92	0.93	193
good	0.97	0.72	0.83	40
unacc	0.98	0.98	0.98	599
vgood	0.71	0.91	0.79	32
accuracy			0.96	864
macro avg	0.90	0.88	0.88	864
weighted avg	0.96	0.96	0.96	864

Thus, we observe that the test set has minimum number of samples for classes vgood and good and acheives an f1 score below average for these, while the model performs very well on classes “unacc” and “acc” due to large number of samples.

IV. Conclusions

In this study, we observed that cars with very high maintenance cost, small capacity, small bootspace and low safety are most likely to have target as “unacc”. Further, cars with target as “vgood” are most likely to have high safety. In the future, techniques to handle class imbalance could be implemented.

References

[1] Decision Tree Learning. (2021, October 31). In Wikipedia. https://en.wikipedia.org/wiki/Decision_tree_learning

[2] Moeedlodhi. “A Beginner’s Guide to Decision Trees. - The Startup.” Medium, 20 Oct. 2020, medium.com/swlh/a-beginners-guide-to-decision-trees-84ca34927818.

[3] Sanjeevi, Madhu. “Chapter 4: Decision Trees Algorithms - Deep Math Machine Learning.Ai.” Medium, 16 Nov. 2018, medium.com/deep-math-machine-learning-ai/chapter-4-decision-trees-algorithms-b93975f7a1f1.

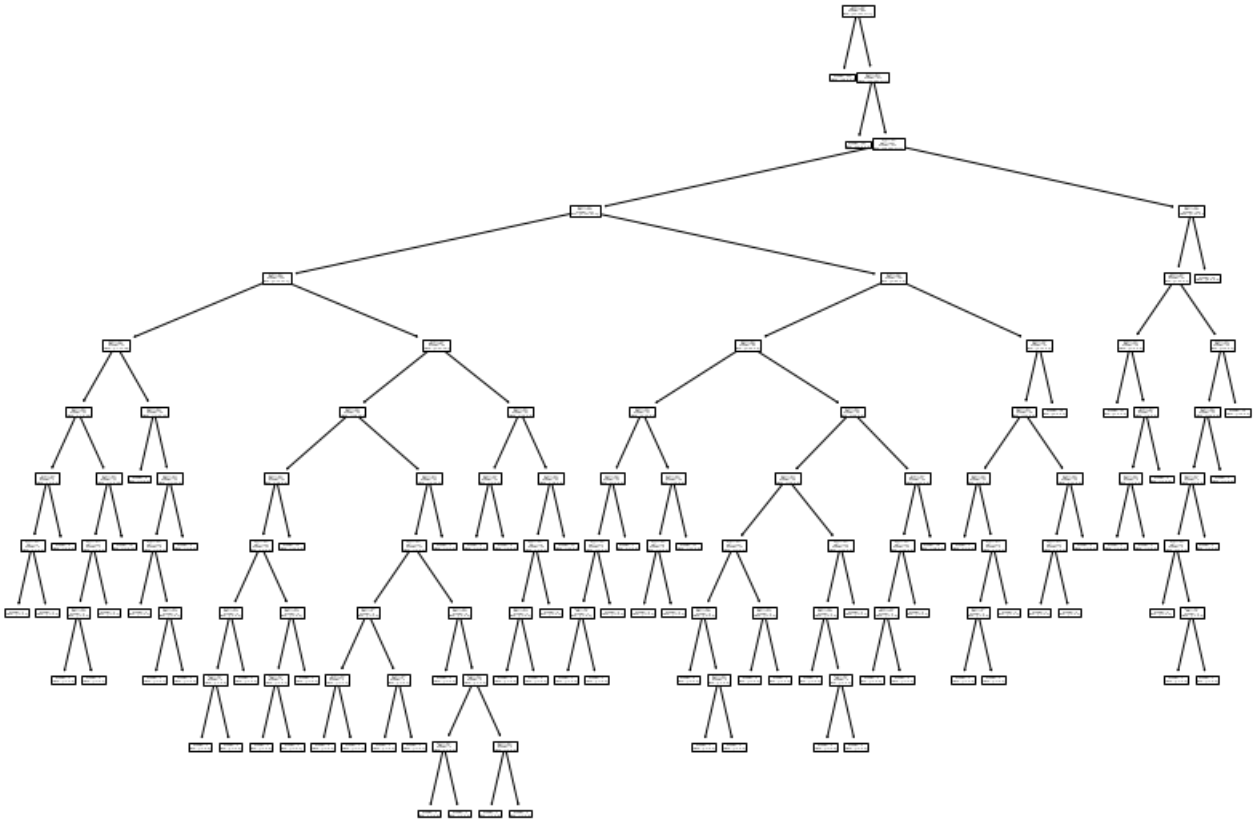

Fig. 6. Visualisation of decision tree model learnt