Assignment 4: A mathematical essay on decision trees

Sukriti Shukla
Department of Engineering Design
Indian Institute of Technology Madras
Chennai-600036, India
ed20b067@smail.iitm.ac.in

***Abstract*—In this study, we utilize a Decision Tree Classifier to assess the safety of cars based on a variety of factors such as buying price, maintenance cost, and seating capacity. Through the application of exploratory data analysis and feature selection techniques, we establish a nexus between the safety ratings of cars and other attributes like boot-luggage capacity. The classifier serves as a robust tool for modeling these relationships and offers valuable insights into the factors that significantly influence car safety.**

***Index Terms*—Car Safety, Decision Tree Classifier, Exploratory Data Analysis, Feature Importance**

I. Introduction

This study aims to understand the factors influencing car safety using machine learning techniques, specifically utilizing the Car Evaluation Database. Factors such as buying price, maintenance cost, seating capacity, and luggage boot size are examined to understand their impact on the safety classification of cars.

Decision Trees form the backbone of our analytical methodology. Utilizing a hierarchical, tree-like model of "if-else" conditions, this machine learning approach excels at both classification and regression tasks. In the current study focused on classification, the algorithm evaluates feature-based questions starting at a root node and traverses down the tree. This leads to a leaf node, which provides the predicted class label. The flexibility of Decision Trees makes them apt for handling multiple types of classification outcomes.

Our primary goal is to use the Car Evaluation Database to classify cars based on their safety levels. This comprehensive dataset offers a variety of features, including buying price, maintenance cost, and estimated safety ratings. One of the central challenges is to discern which features significantly contribute to the safety classification of cars. Understanding these key features is not only vital for optimizing the model's predictive performance but also for gaining valuable insights into what factors most influence car safety.

The next section provides an in-depth look at the dataset and its essential attributes. This is followed by a section focused on seeing the foundational aspects of Decision Trees. In the fourth section, findings and observations obtained from the data and model analyses are presented. Lastly, the fifth section offers a summary of the study's key contributions and discusses potential directions for future research.

II. Decision Tree

Decision Trees are a class of supervised machine learning models designed for both classification and regression tasks. Mathematically, a decision tree aims to partition the feature space $\mathcal{X}$ into disjoint regions $R_1, R_2, \ldots, R_J$ and then assign a label c_j to each region R_j.

A. Tree Structure

A Decision Tree is a hierarchical structure consisting of nodes and edges. The root node at the top of the tree represents the entire dataset or the complete feature space $\mathcal{X}$. As one moves down the tree, internal nodes define conditions on the features X_i, effectively partitioning the space into subsets. These conditions are often inequalities like $X_i < s$ or $X_i \geq s$. Leaf nodes or terminal nodes contain the final outcome, either as a class label in a classification task or a continuous value in regression.

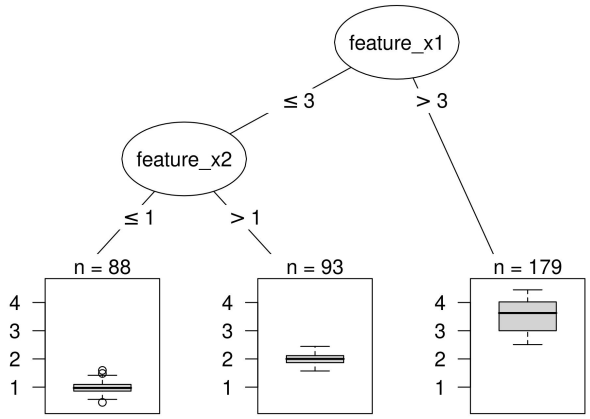

Fig. 1. Example of a decision tree

B. Splitting Criteria

Determining the optimal feature X_i and the best splitting point s is crucial for the effectiveness of the tree. This is usually done by minimizing a loss function or criterion

$C(R_1, R_2, \ldots, R_J)$. In classification tasks, impurity measures like Gini impurity or entropy are commonly used:

$$\text{Gini}(t) = 1 - \sum_{i=1}^{k} p_i^2$$

$$\text{Entropy}(t) = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

Here, p_i is the proportion of samples of class i in node t.

For regression trees, the mean squared error (MSE) provides a good splitting criterion:

$$\text{MSE}(t) = \frac{1}{|t|} \sum_{i \in t} (y_i - \bar{y})^2$$

C. Pruning

Decision Trees are prone to overfitting, especially when they are deep. Pruning techniques like reduced error pruning or cost complexity pruning help to simplify the tree. Cost complexity adds a regularization term to the loss function:

$$C_\alpha(T) = C(T) + \alpha|T|$$

Here, $|T|$ is the number of terminal nodes, and α is a complexity parameter.

D. Feature Importance

Decision Trees naturally perform feature selection. The importance of a feature X_i is often quantified as the weighted sum of the impurity decreases ΔC it brings about:

$$I(X_i) = \sum_{t \in \text{nodes}} \Delta C(t) \times \frac{|t|}{|T|}$$

E. Interpretability

One of the major advantages of Decision Trees is their transparency and ease of interpretation. Each decision at an internal node can be easily understood, allowing for clear interpretability and the ability to trace back the reasoning behind each prediction.

III. Dataset

The dataset used in this study is the Car Evaluation Database, which is tailored for assessing car safety based on multiple features. The dataset comprises a total of 1728 entries, each with 7 features including buying price, maintenance cost, number of doors, seating capacity, luggage boot size, and an estimated safety rating.

The target variable, named 'Target,' aims to classify the safety of the car into one of four categories: unaccountable, accountable, good, and very good.The dataset was preprocessed to handle categorical variables through encoding techniques. This made it compatible for machine learning algorithms that require numerical input features.

TABLE I
Description of Dataset Features

Feature	Description
Buying	Buying price of the car
Maint	Maintenance cost
Doors	Number of doors
Persons	Seating capacity
Lug Boot	Size of the luggage boot
Safety	Estimated safety rating

IV. The Problem

The primary objective of this study is to construct a classifier that predicts the safety of a car. Prior to fitting decision trees to the dataset, an in-depth data analysis is undertaken. The trained model will subsequently be evaluated on an unseen dataset.

A. Data Analysis and Feature Engineering

The dataset was observed to be clean, free from any missing (NaN) values, alleviating concerns related to data imputation. A count plot was employed to inspect class distribution in the target variable. The plot revealed a pronounced class imbalance, with most samples falling under the 'Unacceptable' and 'Acceptable' categories.

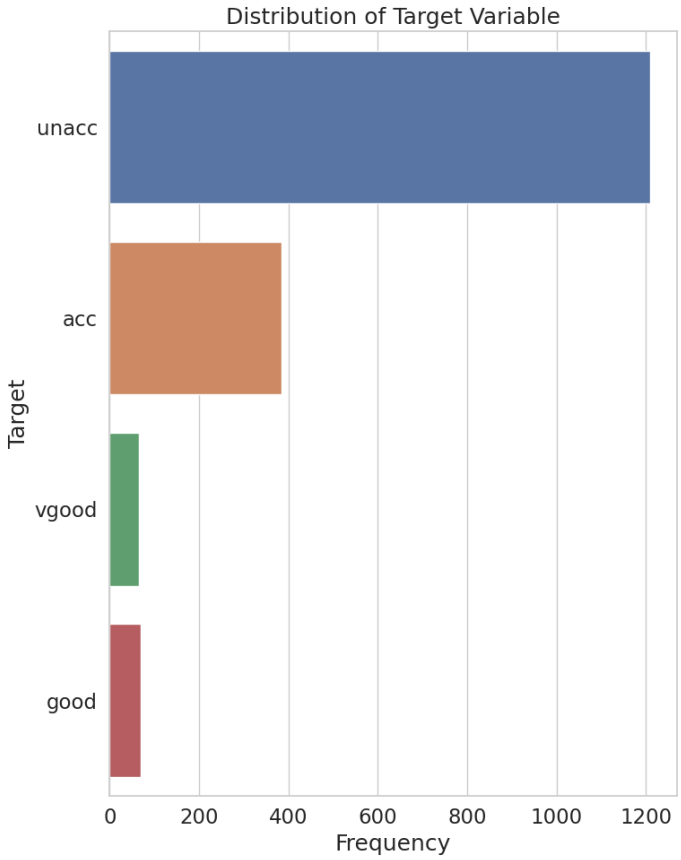

Fig. 2. Target Value Distribution

B. Exploratory Data Analysis

In this section, we delve into the relationships between various features and the target variable. A series of count plots were generated to visualize these relationships.

1) Maintenance Price vs Target: The trends observed are as follows:

- As maintenance costs rise, the car's safety rating generally decreases.
- Notably, cars with high to very high maintenance costs do not fall under the 'good' or 'very good' safety categories.

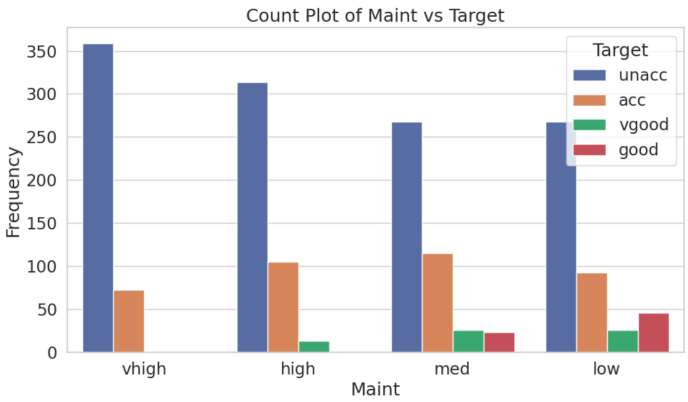

Fig. 3. Count plot of maintenance price vs target

2) Buying Price vs Target: The insights from this feature are surprising:

- Cars with higher buying prices tend to have poorer safety ratings.
- This could be attributed to luxury and ultra-luxury cars, which may compromise on safety for speed and design.

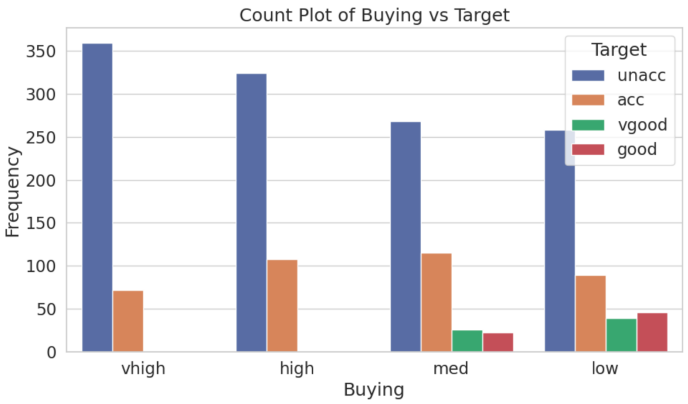

Fig. 4. Count plot of buying price vs target

3) Number of People vs Target: The trends here are:

- Cars with reduced seating capacity generally have lower safety ratings, possibly due to their sporty nature.
- Cars with a capacity of 4 or more show a similar distribution in safety ratings.

Fig. 5. Count plot of persons vs target

4) Luggage Boot Capacity vs Target: The observations include:

- Cars with small luggage boots never fall under the 'very good' category.
- A mixed distribution is observed for cars with medium to large luggage spaces.
- A general trend suggests that larger luggage spaces correlate with higher safety ratings.

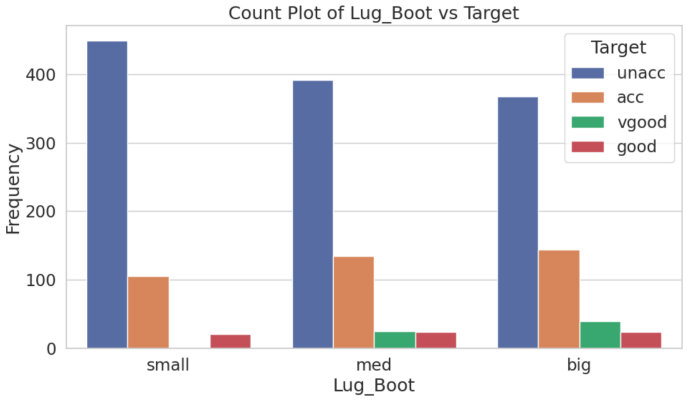

Fig. 6. Count plot of luggage boot vs target

5) Safety vs Target: The relationship between the estimated safety rating and the target variable reveals some key insights:

- Cars with low safety estimates predominantly fall under the 'Unacceptable' category in the target variable.
- Conversely, cars with high safety estimates are more likely to be classified as 'Good' or 'Very Good' in the target safety rating.
- Interestingly, a moderate safety estimate does not guarantee a moderate safety rating, suggesting other features also play a significant role.

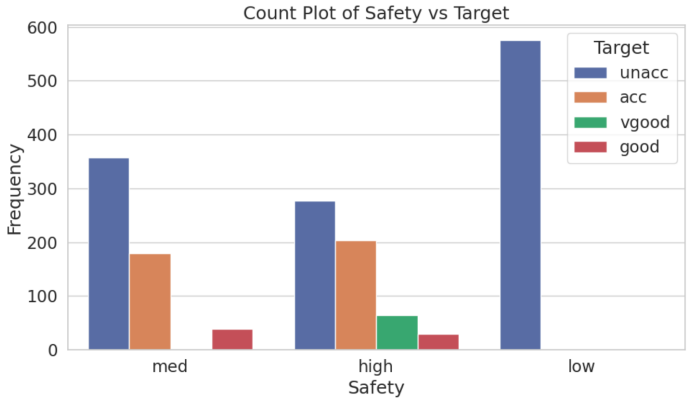

Fig. 7. Count plot of safety vs target

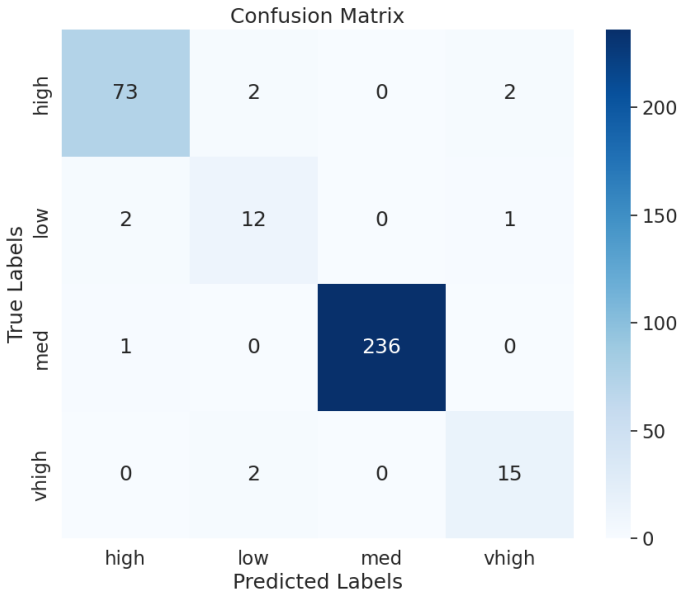

Fig. 8. Confusion Matrix of the Decision Tree Model

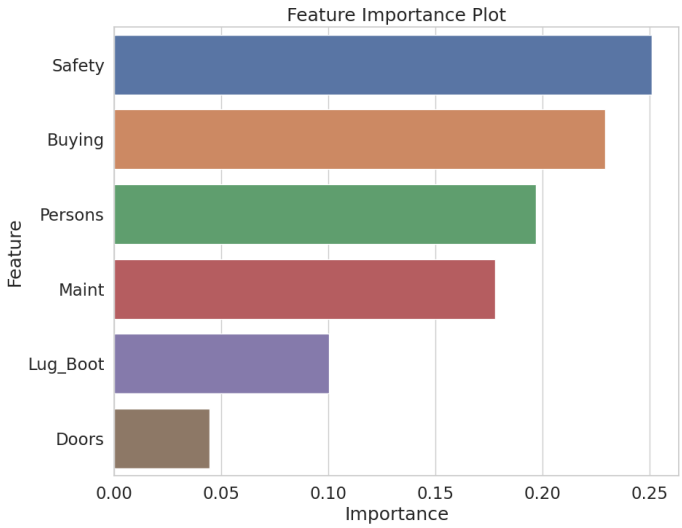

Fig. 9. Plot of feature importance

C. Data Preprocessing: Label Encoding

Before fitting the decision tree model, it was necessary to convert categorical variables into a format that could be provided to machine learning algorithms. Label encoding was performed on the ordinal and categorical features. In this process, each unique category value is assigned an integer, starting from 0. For example, for the 'Buying Price' feature, 'Low' might be encoded as 0, 'Medium' as 1, 'High' as 2, and 'Very High' as 3. This transformation enables the algorithm to handle these features effectively.

D. Model

We used a Decision Tree Classifier as our main tool for predicting car safety. The model considered various aspects like the cost of the car, its maintenance price, and estimated safety levels. A confusion matrix and Feature Importance plot were used to evaluate and understand the model's performance. The Decision Tree Classifier yielded an accuracy of approximately 97.1% on the test dataset, a remarkable achievement. The model excels in predicting the "vgood" (very good) class, boasting a Precision, Recall, and F1-score of 1.00. For other classes, the model performs quite well, but there is room for enhancement, particularly in the "acc" (accountable) category.

V. Conclusion

The aim of this research was to categorize car safety based on different features. Using a Decision Tree, we were able to make meaningful predictions and also identify which features matter the most. One notable point was the imbalance in the dataset, which is an area that can be improved for better model accuracy.

A. Avenues for further research

This study serves as a starting point and there are multiple directions for further research:

- Techniques like SMOTE could be used to address the issue of class imbalance.
- Adding more features like the brand or age of the car might give us a fuller picture.
- Comparing the Decision Tree model with other algorithms could offer more insights.
- Testing the model's predictions in real-world scenarios would be an important next step.

References

[1] Bishop, Christopher M. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

[2] Quinlan, J. Ross. *Induction of Decision Trees*. Machine Learning, 1(1), 81–106, 1986.

[3] Safavian, S. Rasoul and Landgrebe, David. *A Survey of Decision Tree Classifier Methodology*. IEEE Transactions on Systems, Man, and Cybernetics, 21(3), 660–674, 1991.
[4] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*. Springer, New York, 2001.