Assignment 5
A Mathematical Essay on Random Forest

Pankaj Bhardwaj
Cyber Physical Systems
Indian Institute of Technology Madras
Chennai, India
bs20b024@smail.iitm.ac.in

***Abstract*—This paper intends to analyze a dataset related to car evaluations and make predictions about whether customers are likely to purchase a car based on specific features within the dataset. The dataset is balanced and contains both alphanumeric categorical attributes and a multi-class target variable.Random forest models are predictive algorithms known for their effectiveness in handling high-dimensional feature spaces. This article delves into the mathematical foundation of random forests and their potential for optimizing models. It also covers different metrics for evaluating their performance and suggests methods for enhancing their effectiveness. .**

***Index Terms*—Random Forest, Ensemble Models, confusion matrix, accuracy, precision, recall, F1 score, Category Encoders, Bagging**

I. Introduction

The rapid advancement of modern technology has greatly simplified and accelerated people's lives. Transportation, in particular, has witnessed significant progress. Nowadays, individuals have their own means of getting around, and cars have emerged as one of the most dependable and universally embraced modes of transportation worldwide. Each car model comes with its unique set of features, and the variations in these features make the car-buying decision a complex one for consumers. This challenge leads to our problem statement: assessing cars based on diverse features and predicting people's decisions when it comes to purchasing a vehicle. Essentially, this problem is a classification task.

To address this problem, we employ the Random Forest Algorithm. This algorithm is a refinement of the decision tree algorithm, constructing multiple decision trees and deriving the best outcome from them. The fundamental task involves building decision trees, which seek to identify commonalities among data points by recursively dividing a dataset into smaller groups at each iteration. Ultimately, it determines the answer by considering the majority vote from all the trees.

Assessing and forecasting the choice of purchasing cars is a multi-class classification challenge. We are provided with data on diverse car features. These features are analyzed and used to draw meaningful conclusions and identify correlations. The problem revolves around six distinct features, and the objective is to determine the optimal choice within the context of these features.

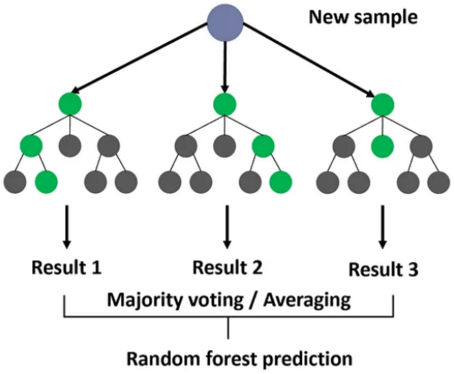

Fig. 1. Random Forest

II. Random Forest

Random Forests are ensemble models that employ the combination of numerous weak decision trees, each of which possesses limited predictive capability, to construct a unified model that delivers high performance. In this section, we explore the process of building and the functioning of a decision tree. We then proceed to grasp the aggregation of multiple such trees into a Random Forest and the rationale behind their exceptional performance. Random Forest is an ensemble machine learning algorithm that combines the predictions of multiple decision trees to improve accuracy and reduce overfitting. It operates through the following mathematical principles:

A. Bagging (Bootstrap Aggregating)

Random Forest builds multiple decision trees using bootstrap samples, which are randomly selected subsets of the training data with replacement. If you have a dataset with N samples, a Random Forest typically uses N samples for each tree, represented as $\{(X_i, y_i)\}$, where X_i is a feature vector and y_i is the label.

B. Feature Randomness

Random Forest introduces randomness in the selection of features for each tree's splits. At each split, a random subset of features $\{F_i\}$ is considered. This reduces the correlation between trees and promotes diversity.

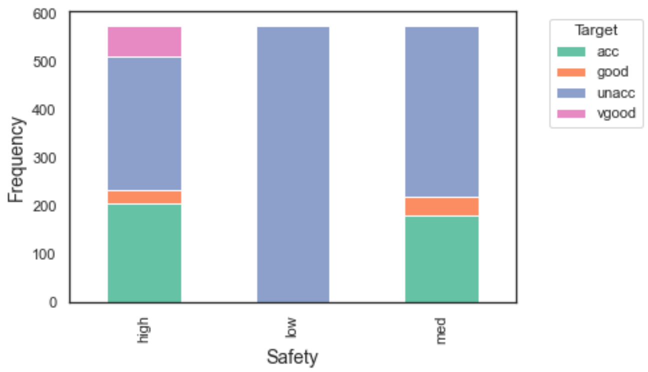

Fig. 2. Target Frequency Vs for Level of Safety

C. Building Decision Trees

For each decision tree, Random Forest uses a random subset of data and a random subset of features to create the tree. The tree is grown by recursively splitting the data at the node that results in the best split, considering only the random subset of features.Decision trees are represented and constructed using various mathematical formulas and concepts. Here are two of the key mathematical elements used in decision trees:

1) Gini Impurity: It is used in classification problems and is defined as:

$$Gini(D) = 1 - \sum_{i=1}^{c} (p_i)^2$$

where D is the dataset, c is the number of classes, and p_i is the proportion of samples in class i.

2) Entropy: Entropy: Another measure of impurity used in classification problems:

$$H(D) = -\sum_{i=1}^{c} p_i \log_2(p_i)$$

D. Aggregating Predictions

Once all individual decision trees are built, predictions are made on new data points by running them through all the trees. For classification tasks, a majority vote is taken among the trees to make the final prediction. For regression tasks, the predictions from all trees are averaged to obtain the final prediction.

III. Data

The car-evaluation dataset provided is well-structured and involves assessing the wisdom of buying a car based on a range of automobile characteristics. These features encompass factors like the buying cost, maintenance cost, the number of doors, passenger capacity, luggage space, and safety rating. While there's a relatively even distribution of these attributes, buyers tend to have preferences for one group of features over another.

TABLE I
Variables

Variable Name	Description	Key
buying	Buying Price	vhigh, high, med, low
maint	Price of the maintenance	vhigh, high, med, low
doors	Number of doors	2, 3, 4, 5more
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
target	Target Variable	unacc, acc, good, vgood

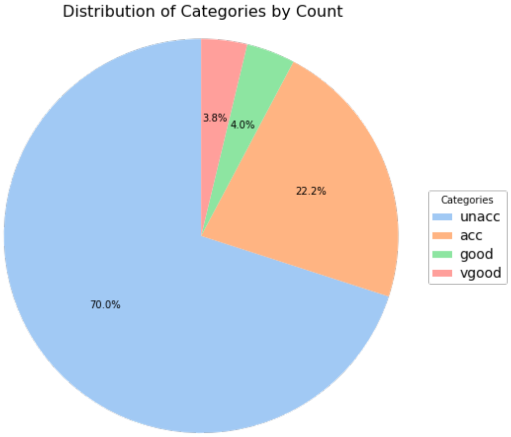

Fig. 3. Countplot for Target Variable

The input variables in this dataset fall into various categories, including both numerical and alphabetic values, and they exhibit a clear ranking or order. The various input features and their definitions are given in Table I. The target variable, or the variable to be predicted is the class that the car belongs to: unacc, acc, good, vgood. To make these variables usable, they have been transformed into ordinal numeric values, which represent the relative rank or level of each attribute. Importantly, it's worth noting that this dataset doesn't contain any missing values, which means that no additional data preprocessing was necessary.

IV. The Problem

For the using given dataset, we try to predict the target variable given input variables. The steps followed for the same are:

- Data cleaning and imputation
- Data visualization and exploratory analysis
- Building random forest models and tuning hyperparameters
- Using the final model to make predictions on the test data
- Identifying the most important features using the random forest

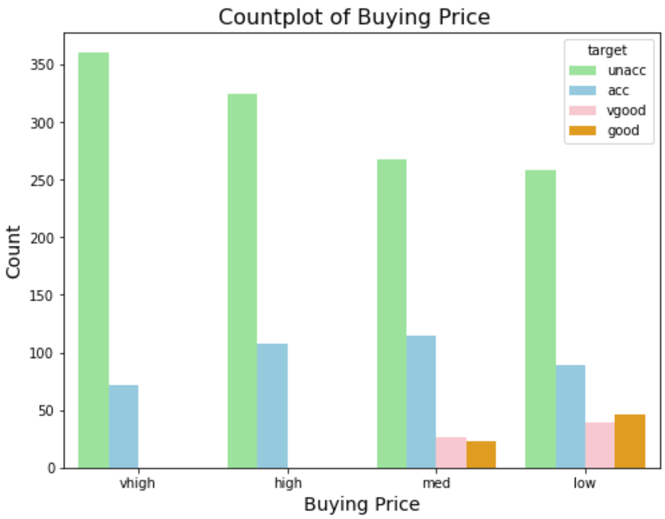

Fig. 4. Countplot for Buying Price

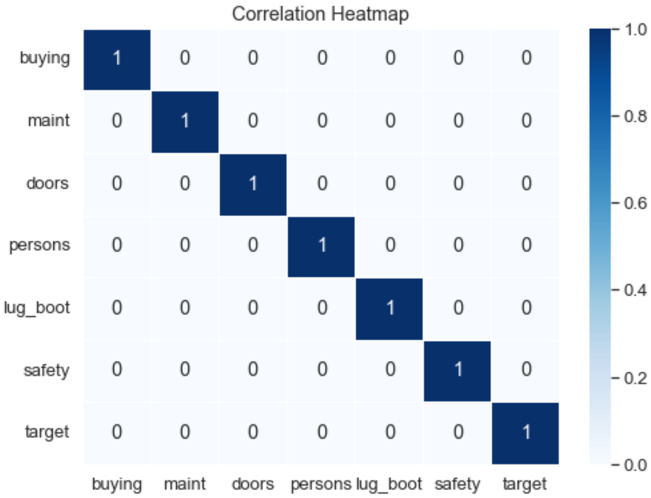

Fig. 5. Correlation Map

A. Visualization

The data we received is mostly clean, but some data preprocessing is required to ensure a stable model training process and improve generalization performance. The dataset was visualized through histograms and count plots, which illustrated the relationship between the decisions and each of the individual features. These visualizations revealed that the values of the features are distributed relatively evenly, but when it comes to the decision, there is a noticeable bias towards the ”unaccepted” category. The dataset contains a total of 1727 data points, and there are no missing values in the dataset. It is split into two subsets: a training dataset and a test dataset, with the training dataset comprising 70% of the total data, and the test dataset making up the remaining 30%. The distribution of data points within the training dataset for each input feature is detailed in Table 1. In particular, when examining the plot of buying price in relation to the decision, a clear pattern emerges. People are more inclined to view the decision to buy as a good one when the purchase price is low. Conversely, when prices are on the expensive side, a majority of customers tend to categorize buying as an unacceptable choice. The maintenance cost of the car also plays a pivotal role in shaping the decision to buy. If the maintenance cost is exceptionally high, it is typically not considered a good decision. Conversely, when the maintenance cost is low, people are more inclined to regard it as a very good decision.

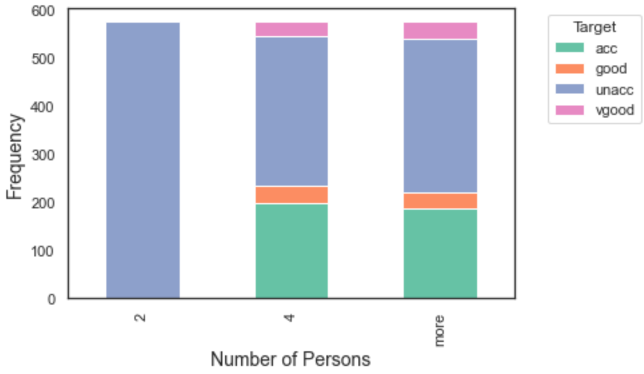

Fig. 6. Target Frequency Vs Number of Persons

B. Applying Random Forest

We constructed a predictive Multi-Class Random Forest model using the provided data. Upon examining the correlation map, it’s evident that there are no significant correlations among the features. As a result, we can use all of these features for model training without concern for multicollinearity or strong inter-feature dependencies. The dataset was initially divided into two subsets: a training dataset and a test dataset. In the training dataset, a Random Forest classification model was constructed using the popular scikit-learn library. Subsequently, this model was employed to make predictions on the test dataset. Impressively, the model achieved an accuracy rate of 97%, indicating that it accurately predicted 97% of the input samples correctly. In essence, this demonstrates the model’s proficiency in making accurate classifications.

In addition to accuracy, another essential performance metric is the macro average F1 Score, which measures the model’s ability to balance precision and recall across different classes. The F1 Score is a crucial indicator of a model’s overall performance. In this case, the macro average F1 Score was found to be 0.96, showcasing the model’s competency in achieving a good balance between precision and recall for the various classes in the dataset. To ensure that the model could effectively process the categorical data within the input variables, I applied one-hot encoding using category encoders. This approach allowed me to capture complex relationships within the data and effectively classify instances into one of the four target categories, enhancing the model’s predictive performance.

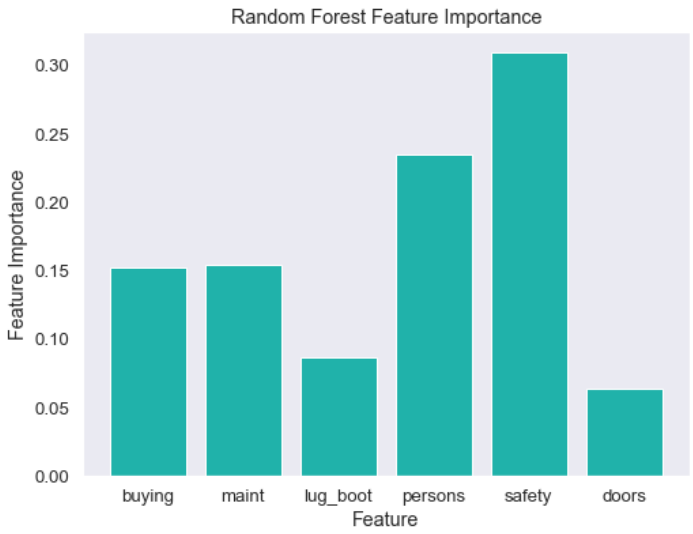

Fig. 7. Feature Importance Plot

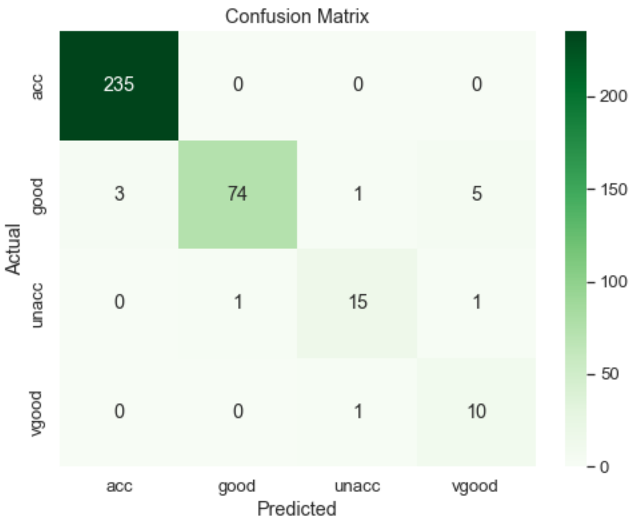

Fig. 8. Confusion Matrix

C. *Observations and Insights*

The model's performance was evaluated using a test dataset derived from the same source data. The accuracy of the model on this test dataset was determined to be 97%, indicating that it correctly predicted 97% of the input instances. Additionally, the macro average F1 Score was computed that came cout to be 0.96. The number of people a car can accommodate significantly influences the buying decision. If a car has only two seats, it is deemed unacceptable. On the other hand, cars with four or more seats are considered acceptable, good, or very good choices. The number of doors, however, does not appear to have a substantial impact on people's decisions when purchasing cars.

The most critical factor that greatly influences decision-making is the safety level. If a car's safety rating is low, it is unequivocally an unacceptable choice. Conversely, the safer the car, the more likely it is to be seen as a favorable option for purchase. The resulting confusion matrix for the model is shown in Fig 7. It is evident that the model exhibits minimal overfitting to the training data. Additionally, the model does not exhibit a bias toward any specific class, and despite the significant class support imbalance, its performance remains consistent. It's worth noting that even without data augmentation, the random forest model performs at least as effectively as a decision tree with data augmentation. In summary, people generally view it as the best decision to buy a car that is affordable, has lower maintenance costs, features four or more doors, offers a high level of safety, and can accommodate at least four people. In my classification model, I obtained a confusion matrix that provided valuable insights into its performance. To visually represent this performance evaluation,A visualization plot of the confusion matrix using sns was made

Conclusion

A thorough exploratory data analysis was carried out on the provided car-evaluation dataset, and a predictive classification model was built using the Random Forest Classification Algorithm. The EDA revealed that most people tend to prefer buying cars with lower costs rather than expensive ones. Additionally, features such as lower maintenance costs, larger luggage space, and increased seating capacity are viewed positively when considering car purchases. Conversely, selecting a car with a low safety rating is seen as a non-negotiable aspect.

The Random Forest model demonstrated an impressive accuracy of 97%, indicating that the Decision Tree algorithm is well-suited for handling multi-class classification tasks. In the future, further enhancements can be made by exploring advanced classification algorithms like XGBoost to improve the model's efficiency and performance. This development benefits both companies and consumers, allowing companies to assess the potential success of new cars with specific features and enabling consumers to gauge the popularity of specific cars among other buyers.

Acknowledgment

I would like to thank professor Dr. Gaurav Raina and IIT Madras for providing me this opportunity to work on this project. I would also like to express my gratitude towards my classmates and friends who helped me with this project.

References

[1] V. Y. Kulkarni and P. K. Sinha, ”Pruning of Random Forest classifiers: A survey and future directions,” 2012 International Conference on Data Science Engineering (ICDSE), 2012, pp. 64-68

[2] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani,“Classification” in An Introduction to Statistical Learning, New York,Springer, 2013

[3] ”Liu, Y. H. (2019). Python machine learning by example: Implement machine learning algorithms and techniques to build Intelligent Systems (2nd ed.). Packt Publishing Ltd.”

[4] ”What is a Random Forest? — TIBCO Software,”