A Mathematical Essay on Random Forests

Aryan Pandey
Department of Ocean Engineering
Indian Institute of Technology, Madras
Chennai, India
na19b030@smail.iitm.ac.in

***Abstract*—This document is an overview of the mathematical aspects of Random Forests as well as its application on a sample data set. The algorithm has been applied on a data set of cars and the prediction task is to classify a car based on its safety**

I. Introduction

Classification is a two-step process, learning step and prediction step, in machine learning. In the learning step, the model is developed based on given training data. In the prediction step, the model is used to predict the response for given data. Decision Tree is one of the easiest and popular classification algorithms to understand and interpret.

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data).

In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

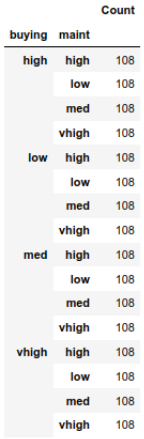
In this paper, the aim is to apply Decision Tree to classify a car based on its safety. The paper systematically goes through first the mathematical details of Decision Tree, the nature of the data set which has been given to us, then the problem we have in hand and how it has been solved, and finally the conclusions which were drawn. Useful insights and figures have been presented whenever necessary.

II. Datasets

The dataset given has the following details of multiple cars with the aim of classifying it based on its safety. It has the details of buying price, price of maintenance, number of doors, seating capacity, size of luggage boot and the estimated safety of the car.

The dataset consists of purely categorical columns in which the columns related to buying price, price of maintenance, and estimated safety of the car are categorised into very high, high, medium and low. The target condition of the car has 4 categories which are very good, good, acceptable and unacceptable.

Fig.1 shows that all these columns have a similar split across the categories. This similar observation can be made across all columns except for the target column where it is unbalanced.



		Count
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

Fig. 1. Dataset has similar splits

Fig.2 shows us the correlation of the features with each other as well as the correlation of the features with the target variable. We can see that the features have no correlation with each other whereas some features like safety and seating capacity have a good correlation to the target.

III. Models

This section discusses the mathematical and conceptual aspects of the Random Forest algorithm.

A. Intuition

Random forest is a supervised learning algorithm. It has two variations – one is used for classification problems and other is used for regression problems. It is one of the most flexible and easy to use algorithm. It creates decision trees on the given data samples, gets prediction from each tree and selects the best solution by means of voting. It is also a pretty good indicator of feature importance. Random forest algorithm combines multiple decision-trees, resulting in a forest of trees, hence the name Random Forest. In the random forest classifier, the higher the number of trees in the forest results in higher accuracy. Before understanding the working of the random forest we must look into the ensemble technique. Ensemble simply means combining multiple models. Thus a collection

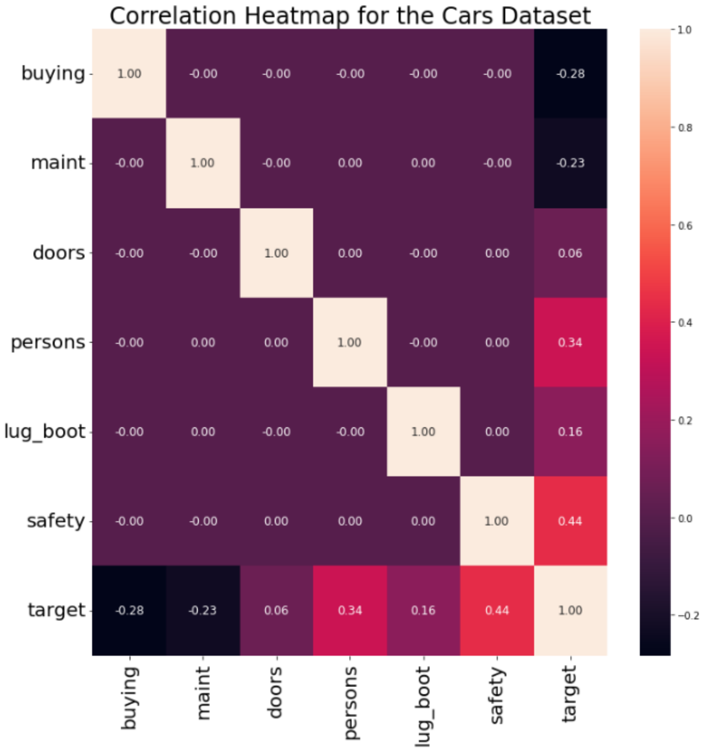


Fig. 2. Correlation Heatmap of features

of models is used to make predictions rather than an indivic model. Ensemble uses two types of models:

- Bagging– It creates a different training subset fi sample training data with replacement the final outpu based on majority voting. For example, Random For
- Boosting– It combines weak learners into strong learr by creating sequential models such that the final mc has the highest accuracy. For example, ADA BOOST, BOOST

As mentioned earlier, Random forest works on the Bagg principle. Now let's dive in and understand bagging in de Bagging Bagging, also known as Bootstrap Aggregation is the ensemble technique used by random forest. Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now each model is trained independently which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as aggregation.

B. The Algorithm

Steps involved in the random forest algorithm:

- In Random forest n number of random records are taken from the data set having k number of records.
- Individual decision trees are constructed for each sample
- Each decision tree will generate an output.
- Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

Going into more detail, we will explain the steps. In the first stage, we randomly select "k" features out of total m features and build the random forest. In the first stage, we proceed as follows:-

- Randomly select k features from a total of m features where k < m
- Among the k features, calculate the node d using the best split point
- Split the node into daughter nodes using the best split
- Repeat 1 to 3 steps until l number of nodes has been reached
- Build forest by repeating steps 1 to 4 for n number of times to create n number of trees

In the second stage, we make predictions using the trained random forest algorithm

- We take the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome.
- Then, we calculate the votes for each predicted target.
- Finally, we consider the high voted predicted target as the final prediction from the random forest algorithm.

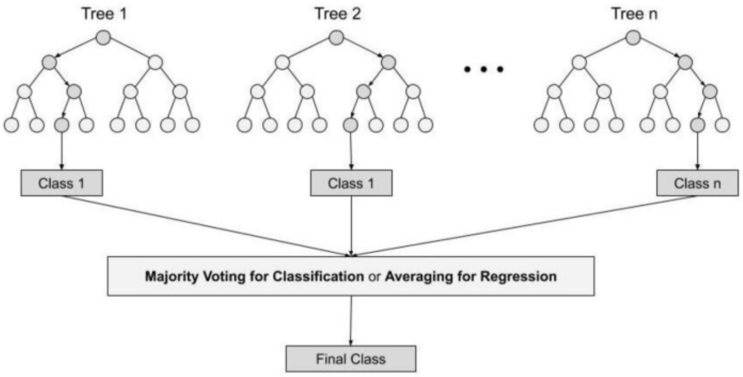

Tree 1
Tree 2
Tree n
Class 1
Class 1
Class n
Majority Voting for Classification or Averaging for Regression
Final Class

Fig. 3. Depiction of Random Forests

C. Assumptions

A random forest's assumptions are the same as the assumptions an individual decision tree makes which are:

- In the beginning, a part of the training set is considered as the root.
- Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model.
- Records are distributed recursively on the basis of attribute values.
- Order to placing attributes as root or internal node of the tree is done by using some statistical approach.

D. *Advantages*

- One of the most accurate learning algorithms available
- It can handle many predictor variables
- Provides estimates of the importance of different predictor variables
- Maintains accuracy even when a large proportion of the data is missing

E. *Disadvantages*

- Can overfit datasets that are particularly noisy
- For data including categorical predictor variables with different number of levels, random forests are biased in favor of those predictors with more levels
- Therefore, the variable importance scores from random forest are not always reliable for this type of data

IV. Modelling

In this section, we will explore how the model was applied to the dataset at hand and what we can infer from it. The categorical features were first encoded ordinally to ensure that the ranking of the categories was maintained in the encodings. We then split the dataset into a train and validation split of an 85:15 ratio and train the Random Forest Classifier.

We obtain an accuracy of 97% and an F1 score of 0.9 on the dataset. The confusion matrix on the validation set is given in Fig. 3.

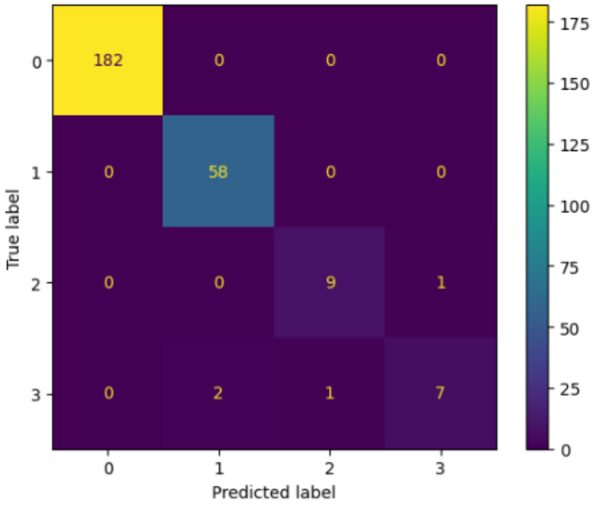


Fig. 4. Confusion Matrix of the Model Predictions

We can also find out the first few layers of the decision tree. This has been represented in Fig.4. As you can see the safety is the primary feature that is taken into account when plotting the tree.

V. Conclusions

Safety was one of the key factors in predicting the target variable for the car. The model is doing a good job of segregating all the categorical features into the different classes. This shows that the dataset has a good amount of separability and the misclassification rate is very low.

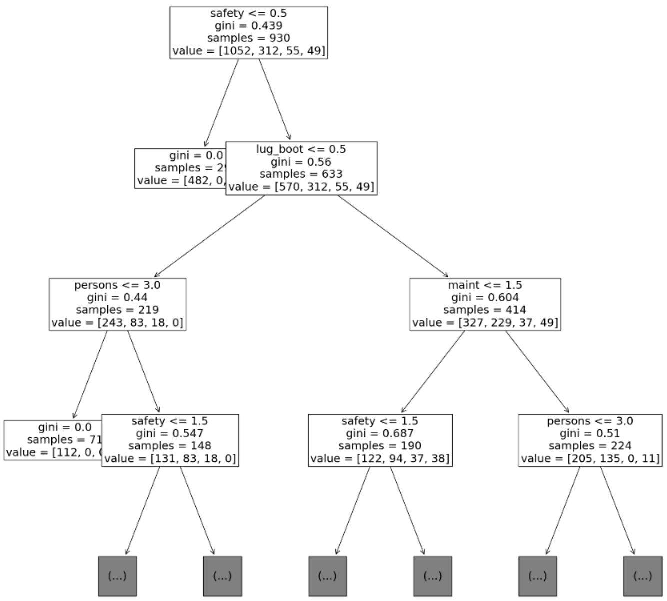


Fig. 5. Plotting the first few layers of the tree