Assignment 5 Version 2: A mathematical essay on Random Forest Classifier

S, Karthik
ME18B149 - IDDD Data Science
Indian Institute of Technology Madras
Chennai, India
me18b149@smail.iitm.ac.in

***Abstract*—This document is an overview of the concept of Random Forest Classifiers. It gives a general idea of the topic, along with the mathematical aspects, as well as an application involving it based on data from a real-world problem.**

***Index Terms*—Classification, Random Forest, Machine Learning, Data Science**

I. Introduction

Random forests is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm.

The specific problem discussed is a classification problem which determines the level of safety of the cars based on certain features using the Car Evaluation Dataset. Random forests will be used to analyze the data and build a model which will then be able to predict the category for unknown data. Various insights and conclusions will be made based on the trends followed by the dataset.

This paper is a case study through which the principles of Random forests are implemented. It aims to examine which parameters influence the level of safety of the car by automotive standards.

II. Random Forest

Random Forest, as the name says, is a collection of Decision Trees formed at random. It is said that the more trees it has, the more robust a forest is. Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance. Put simply: random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

Random forests has a variety of applications, such as recommendation engines, image classification and feature selection. It can be used to classify loyal loan applicants, identify fraudulent activity and predict diseases. It lies at the base of the Boruta algorithm, which selects important features in a dataset.

A. Model Structure

Fig 1 shows the model structure of a Random Forest. Random samples of training data are selected to create training samples (subsets) and individual decision trees are trained for those subsets. During prediction, the entry is run through all the decision trees and the outcome classified by most of the trees is selected as the final outcome.

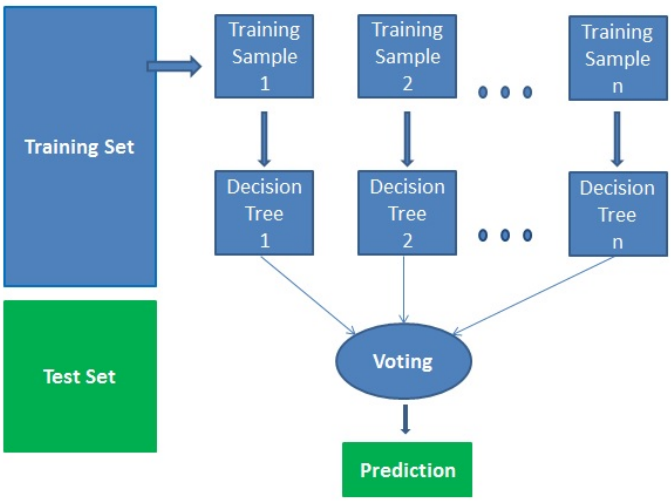

Fig. 1. General structure of a Random Forest Classifier

B. Advantages

- Random forests is considered as a highly accurate and robust method because of the number of decision trees participating in the process.
- It does not suffer from the overfitting problem. The main reason is that it takes the average of all the predictions, which cancels out the biases.
- The algorithm can be used in both classification and regression problems.
- Random forests can also handle missing values. There are two ways to handle these: using median values to replace continuous variables, and computing the proximity-weighted average of missing values.
- You can get the relative feature importance, which helps in selecting the most contributing features for the classifier.

C. Disadvantages

- Random forests is slow in generating predictions because it has multiple decision trees. Whenever it makes a

prediction, all the trees in the forest have to make a prediction for the same given input and then perform voting The model is difficult to interpret compared to a decision tree, where you can easily make a decision by following the path in the tree.on it. This whole process is time-consuming.

- The model is difficult to interpret compared to a decision tree, where you can easily make a decision by following the path in the tree.

D. Finding important features

Random forests also offers a good feature selection indicator. Random forest uses gini importance (discussed in decision trees assignment) or mean decrease in impurity (MDI) to calculate the importance of each feature. Gini importance is also known as the total decrease in node impurity. This is how much the model fit or accuracy decreases when you drop a variable. The larger the decrease, the more significant the variable is. Here, the mean decrease is a significant parameter for variable selection. The Gini index can describe the overall explanatory power of the variables. This score will help to choose the most important features and drop the least important ones for model building.

III. The Problem

A. Overview and objectives

The Car Evaluation Database was derived from a simple hierarchical decision model. The prediction task is to classify a car based on its safety.

In this assignment, we attempt to build a predictive Decision Tree model that answers the question: "what factors of people are more likely to influence the outcome?" using the above data (i.e doors, price, persons, etc).

B. Reading the data

We will be using python throughout this assignment. A single dataset titled "car_evaluation.csv" is provided which contains the details of various parameters of come cars which are shown in detail in Fig 2. Since we are creating a model, we would be splitting the dataset into train and validation sets with a fraction of 10% for validation.

IV. Data cleaning

After reading the data, the next task will be to filter out the attributes or entries that are not required or redundant in this analysis. There are no unfilled values in this dataset which means no entry is being eliminated. Also, all the features are different enough and no feature is redundant in this case which means the dataset will be used completely as it is without any alterations.

V. Exploratory Data Analysis

The complete details of all the attributes are shown in Fig 2.

Our first task is to analyze the data and form visible conclusions and trends from the same with the help of plots.

Variable	Definition	Key
buying	buying price	vhigh, high, med, low
maint	Price of the maintenance	vhigh, high, med, low
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
Target	Target variable to predict	unacc, acc, good, vgood

Fig. 2. Exhaustive list of all the attributes in the given dataset

Fig. 3. Bar chart - Distribution among outcomes

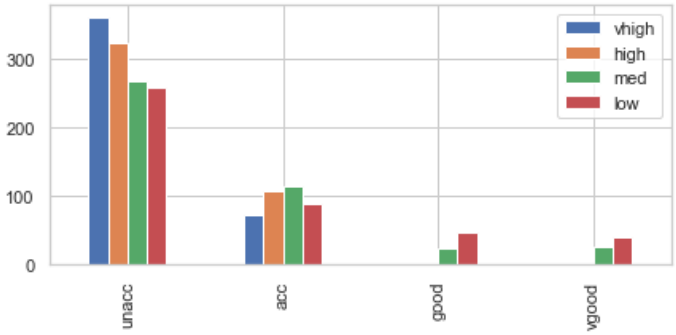

Fig. 4. Bar chart - Classification by price

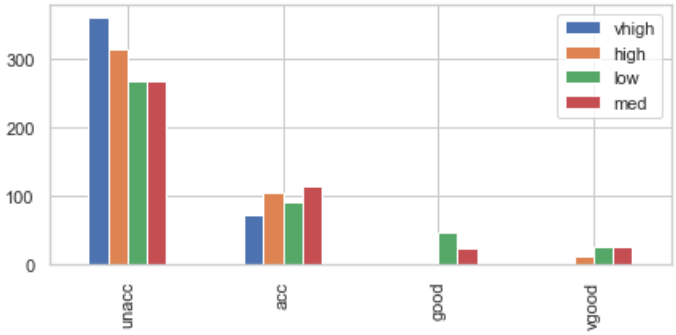

Fig. 5. Bar chart - Classification by maintenance cost

From Fig 3, we can see that the dataset is a highly imbalanced and a skewed one as the number of unacceptable cars are much higher than others.

Fig 4 to Fig 9 shows the classification of the outcomes based on all the features individually. We can see that due to the skewness, every possible value of a feature has maximum fraction of unacceptable cars only for buying and mainteanance

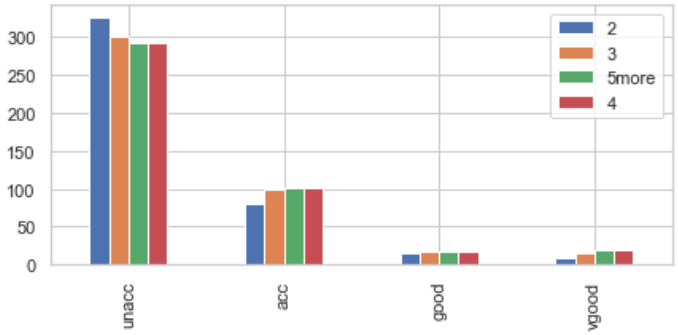

Fig. 6. Bar chart - Classification by number of doors

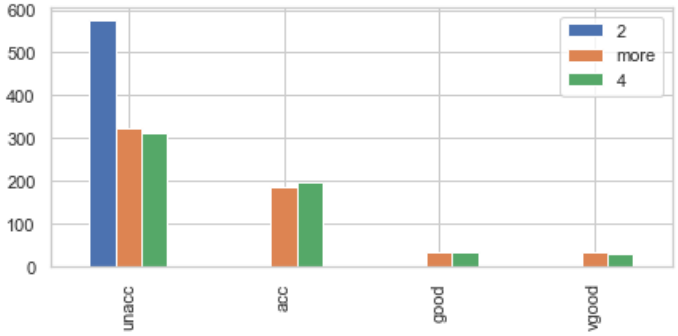

Fig. 7. Bar chart - Classification by capacity

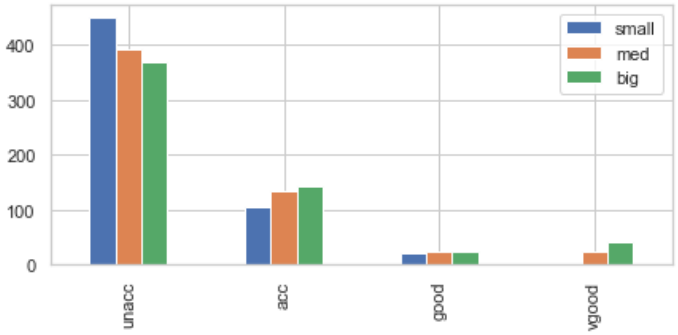

Fig. 8. Bar chart - Classification by lug boot size

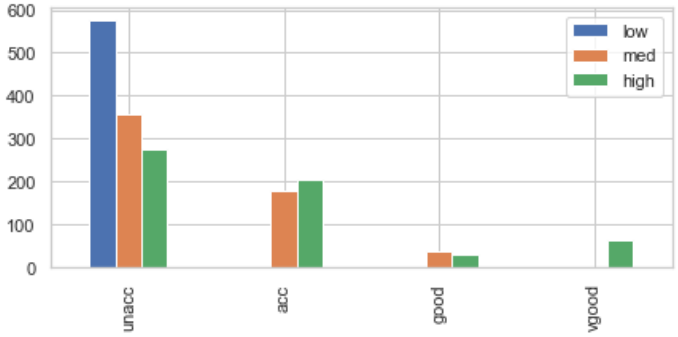

Fig. 9. Bar chart - Classification by safety

costs, number of doors and lug boot size. Hence, we can conclude that these are the features with less weightage.

On the other hand, if we see the safety feature (Fig 9), it is clear that the majority of the unacceptable cars are of low safety which is reasonable as the car has to be safe enough by the automotive standards.

VI. Data preparation

The given dataset is clean in terms of missing data. All the variables are categorical. Hence, label encoding was done for every variable starting with the least possible outcome assigned to 0 and increased subsequently.

TABLE I
SAMPLE ENTRIES OF MODEL COMPATIBLE DATASET

S.No	buying	maint	doors	persons	lug_boot	safety	target
0	3	3	0	0	0	0	unacc
1	3	3	0	0	0	1	unacc
2	3	3	0	0	0	2	unacc
3	3	3	0	0	1	0	unacc
4	3	3	0	0	1	1	unacc

VII. Model preparation, training and validation

Since the datasets are now model compatible, the next step is to create the model. The entire dataset was split into training and validation datasets with a fraction of 10% kept for validation. Validation accuracy, i.e, the fraction of entries for which the predicted and the actual value are the same will be used as the metric to compare between models. 3 parameters (n_estimators, max_depth and max_features) were tuned for a range and graphs plotted between the parameter values and the accuracy. From Fig 10 to Fig 12, the values for the final model was determined as shown in Table II.

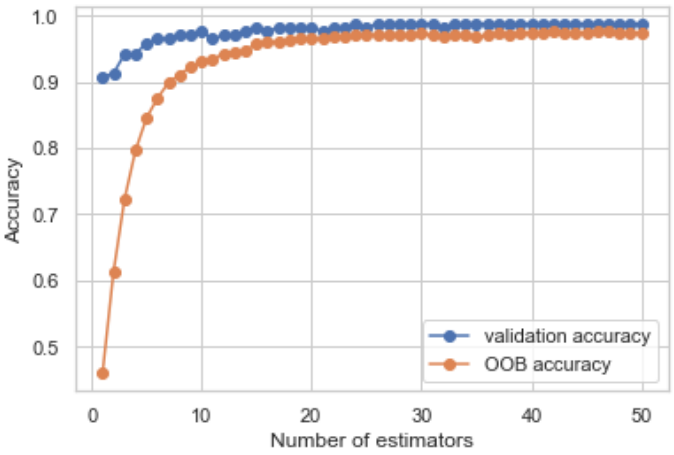

Fig. 10. Plot between accuracy and number of estimators

From Fig 11, one very interesting observation which is atypical of any dataset is that even though the test data WAS NOT used in training, the accuracy proportionally increases therefore unable to determine overfitting nature.

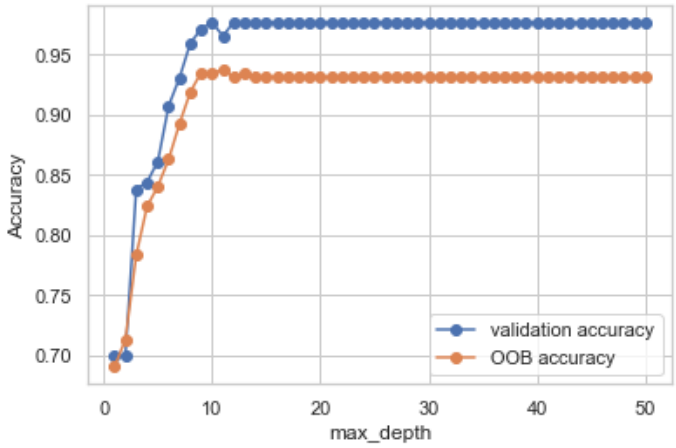

Fig. 11. Plot between accuracy and maximum depth

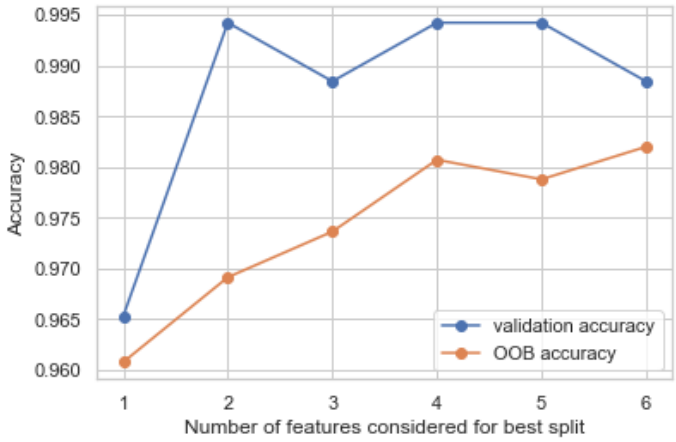

Fig. 12. Plot between accuracy and maximum features

This is one of the advantages of Random Forest over a single decision tree.

A. *Final Model*

Using the above parameters, a Random Forest classifier was created and the trained using the training data. The validation accuracy is determined to be **99.42%** and the Out-of-bag accuracy came out to be **98.07%**.

Also, the confusion matrix which is the plot between the actual outcome and the predicted outcome is plotted for the model and is shown in Fig 13. Only one entry is incorrectly predicted from the test dataset which shows that this is a highly accurate model.

TABLE II
HYPERPARAMETER TUNING RESULTS

S. No.	Parameter	Value
1	n_estimators	30
2	max_depth	10
3	max_features	5

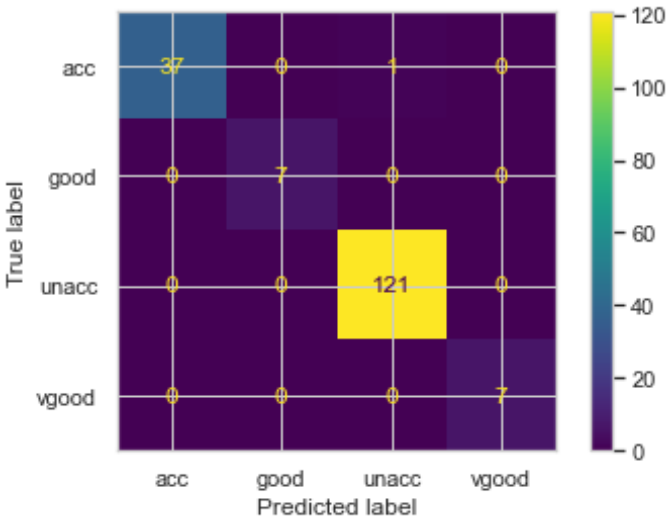

Fig. 13. Confusion matrix - test data

VIII. Grid Search

Since we did manual tuning in the previous analysis and found the most accurate model among them, we have the option to use Grid Search around these parameters to check if the accuracy can be optimized further. Table III gives us the parameters given for the grid search.

TABLE III
HYPERPARAMETER TUNING - GRID SEARCH

S.No	Parameter	Values
1	bootstrap	True
2	max_depth	80, 90, 100, 110
3	max_features	2, 3
4	min_samples_leaf	3, 4, 5
5	min_samples_split	8, 10, 12
6	n_estimators	100, 200, 300, 1000

Upon Grid Search, it was found that the most accurate model is still the same as the one done manually with a classifciation accuracy of **99.42%**. Hence, the final model is still unchanged

IX. Attributes from the confusion matrix

We will try to find some useful parameters for the best model from the results.

1) Precision: Precision can be defined as the percentage of correctly predicted positive outcomes out of all the predicted positive outcomes. It can be given as the ratio of true positives (TP) to the sum of true and false positives (TP + FP).

So, Precision identifies the proportion of correctly predicted positive outcome. It is more concerned with the positive class than the negative class.

Mathematically, precision can be defined as the ratio of TP to (TP + FP). In the best model, it comes out to be 97.68%.

2) Recall/Sensitivity: Recall can be defined as the percentage of correctly predicted positive outcomes out of all the actual positive outcomes. It can be given as the ratio of true positives (TP) to the sum of true positives and false negatives (TP + FN). Recall is also called Sensitivity.

Recall identifies the proportion of correctly predicted actual positives.

Mathematically, recall can be given as the ratio of TP to (TP + FN). In the best model, it comes out to be 100%. This means it is very close to a perfect model.

3) True and False Positive Rate: True positive rate is same as recall (TP/(TP+FN)). False positive rate is given as (FP/(FP+TN)). In this model, True Positive rate is 100% and False Positive Rate is 1.50%.

TABLE IV
BEST MODEL ATTRIBUTES

S.No.	Attribute	Value
1	Classification accuracy	0.9942
2	Classification error	0.021
3	Precision	0.9768
4	Recall or Sensitivity	1
5	True Positive Rate	1
6	False Positive Rate	0.015
7	Specificity	0.8750

X. CONCLUSION

The effects of various factors which could have affected the income of the classification of the car in the Car Evaluation Database were analyzed in great detail. Various factors greatly affected the outcome like safety, number of doors, price, number of passengers.

Various Random Forest Classifier models were constructed on the basis of the known data to determine the right set of hyperparameters which with the best possible model was created with a validation accuracy of 99.42%. Using this model, the confusion matrix was plotted which gave the exact numbers predicted correctly and incorrectly.

Further avenues could be explored using this data where they can be classified using various other classifiers available to study how the parameters might possibly affect each other and also how they affect the income. Also, implementing techniques more advanced than Decision Trees might give a model with better accuracy and predictions which would improve on the current model.

REFERENCES

[1] Dr. Kaushik Mitra, Dr. Ramkrishna Pasumarthy, "Random Forest", EE4708: Data Analytics Laboratory - Week 6

[2] Random Forest — Detailed Overview, Towards Data Science.

[3] Random Forest Classifier - https://scikit-learn.org/