Assignment 5: a mathematical essay on random forest classifier

Krishna Sah Teli
Cyber-Physical Systems (I2MP)
Indian Institute of Technology (IITM)
Madras - 600036, India
ge22m018@smail.iitm.ac.in

***Abstract*—In this paper, a CART has been studied and its implementation using random forest has been done exploring evaluation of car based on its safety measures. It explains about the different parameters related to car which play crucial role in understanding its safety. The correlation between the safety and other parameters has been depicted graphically with the help of data visualization tools. Even though the price of car is high if its safety feature is not satisfactory then it will not be categorized as of good quality. Data analysis shows that the maintenance cost increases as the surge of buying price of car and the capacity of car also does contribute in its grading. It also shows that the number of doors increase with the increase in capacity of car and decrease with that of its price. It also allows us to visualize that the price of car does not make it worthy if safety parameter is ignored, the reason of for low price car becoming more popular and accessible. With the help of a classification random forest machine learning model, the car evaluation dataset has been implemented to build a classification model and its accuracy, precision and recall properties have been explored. Using two different sets of estimators, the predictive models have been built and study shows that the accuracy of the model increases with the increase in number of estimators up to certain number. In this revised paper, I made modification in abstract and Introduction followed by conclusion. I have also improve the data visualization to get more insight about our problem.**

***Index Terms*—Random Forest, Data Visualization, Correlation, Recall, number of estimators, Precision**

I. Introduction

In this era of technology, the technology has been growing rapidly and its applications have great impact on our daily life and working place. It has been playing great role in modernizing every aspects of our life from industrial to personal point of view. It has been possible because of data. Data is the fuel for the technology. Many industries and companies came to know the importance of data during the covid-19 time as they were not having tangible access other than online to reach out to customers. By preprocessing and feature engineering techniques, raw data can be converted into meaningful data which later used for training different machine learning models or developing artificial intelligence based devices. ML and data science have geared up the revolutionary changes in the field of technology. It has not only contributed in the technical field but we can see its impact on every field from agriculture to astronomy. Thus, it has become the most trending topic in current time period. Ml-model can predict, estimate, recognize, detect, etc. based on the data and type of training model. In this paper, car evaluation data has been used to classify the car quality based on its safety measures. It also explains the correlation between the data features such as number of doors, capacity of car, luggage size, cost price, maintenance price, safety with respect to its grading index. With the help of data analysis and visualization, ground truth has been explored followed by development of classification model.

To build the classification model, random forest algorithm, a supervised machine learning model [1] has been implemented. A random forest algorithm is a combination of multiple decision trees models. It is applicable for both classification as well as regression problems. Thus, it is known by CART. Since its implementation is easy and more flexible, it is more common ML model. It creates number of decision trees on the given data samples and select the best predictive solution by means of voting. It has been found out that higher the number of trees more the accuracy of model. It has several application in industries from small scale company to big company. for example, it has wide range of applications such as intrusion detection [5] , medical checkup [4], [6], [7], weather forecasting, developing AI-based devices, e-commerce stack fall down prediction, banking, scientific studies, etc.

The main focus in this paper is to demonstrate how a safety of car contributes in its evaluation. It correlates the car evaluation to its other features such as buying cost, maintenance cost, number of doors, capacity, etc. With the help of visualization tools and techniques, the correlation between the features of car have been illustrated in graphical form. It also illustrates implementation of random forest algorithm over the data to develop a classification model. It has been found out that the accuracy of the model is (94.74%) with 10 estimators where as the accuracy goes up to 96.49% when the number of estimators increases to 100.

The rest of the paper has been organized as follows: section II outlines the data set related to our problem, section III explains the mathematics behind the random forest and tools used for developing our model. section IV Modelling explains about the random forest based trained model and section V concludes the paper with key contributions made and directions for future works.

II. Data set

The source of the car evaluation data is kaggle which is famous for the data science and storage of data. This data has totally 7 columns namely 'buying', 'maint', 'doors', 'persons', 'lug_boot', 'safety', 'decision'. All the features have been taken into consideration for training the model. 'buying' shows the car price and it has three different values namely low, medium, high, and very high where as 'maint' explains about the maintenance cost of the car and it has also values similar to 'buying'. Similarly, 'persons' shows the capacity of car, and lug_boot tells about the size of luggage which can be accommodated in car. Further, 'safety' attribute represents the safety index of car such as low, medium, high, and very high and based on it, car is evaluated into 'decision' attribute such as acceptable, unacceptable, good, and very good.

With the data analysis and data visualization tools, the correlation among the data features has been studied and correlated to the safety and car evaluation. The fig.1 shows about 70% of cars are categorized as unacceptable and out of remaining 22.2% of cars are classified as just acceptable, only 3% are good, and 4% are considered as very good of quality.

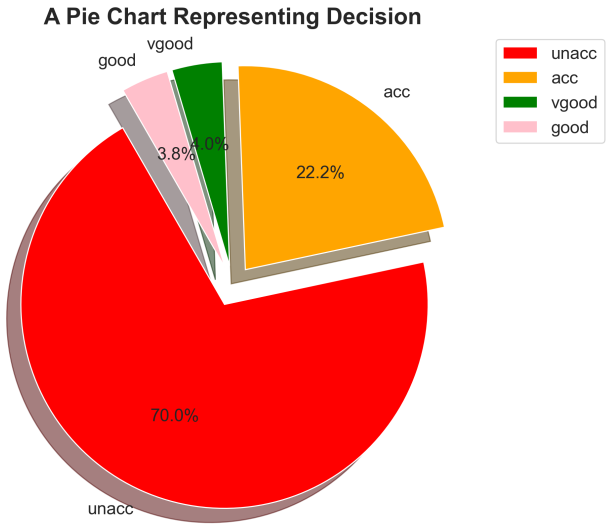

Fig. 1. Distribution of decision features

Similarly, the fig.2 shows about 80% of the cars are classified into unacceptable category regardless of their buying price. However, some of the low price cars fall into good and very good category whereas there is no possibility for high price or very high price cars. It shows that the price of car does not show the quality of car.

The fig. 3 illustrates that the maintenance cost of the car is proportional to its buying cost. It may be because the spare part of costly car is also costly. It does not tell that if the maintenance cost is expensive, then the car should be of good safety quality. Even though the price and maintenance cost is low, one can be good quality of car.

Likewise fig.3, figure 4 also describes that the capacity of car helps in evaluating the quality as if the capacity is 2

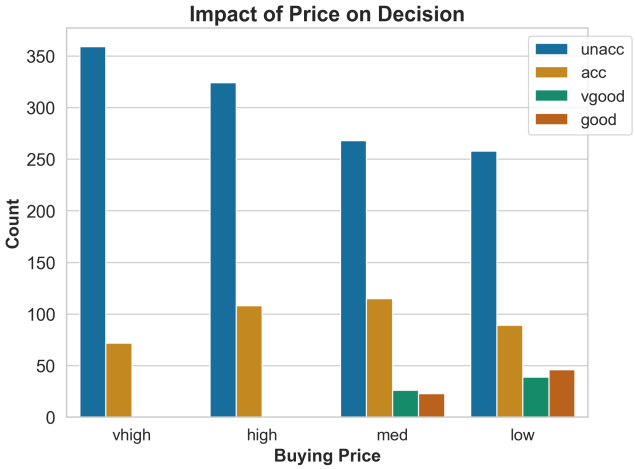

Fig. 2. Showing the prices of cars

persons then it is categorized as unacceptable regardless of other facilities. If the capacity is more than or equal to 4, then there is about 50% of chance to fall into unaccepted and 10% chance to fall into good or very good. It also shows that about 30% of cars are just classified into acceptable.

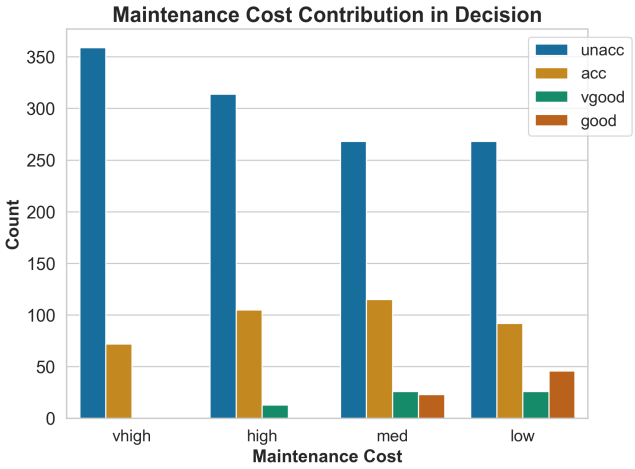

Fig. 3. Maintenance cost of car

The figure 5 depicts that the if the safety is compromised then it will be put into unacceptable category otherwise there is possibility of labelled as good or very good. Similarly, 30% of highly safe cars are discarded because of some other reasons, it may be capacity is less, maintenance cost is high. It shows that if the safety is medium then there is no chance of getting very good labelled but there is more than 50% chance to fall into unacceptable category.

The below figure 6 illustrates that the evaluation of cars based on the number of doors present in it. This figure clearly shows that number of doors is not so important parameter to evaluate the car quality as regardless of door counting, distribution is almost similar. However, if the doors is 2, it shows that the car size is small ans its capacity is less thus

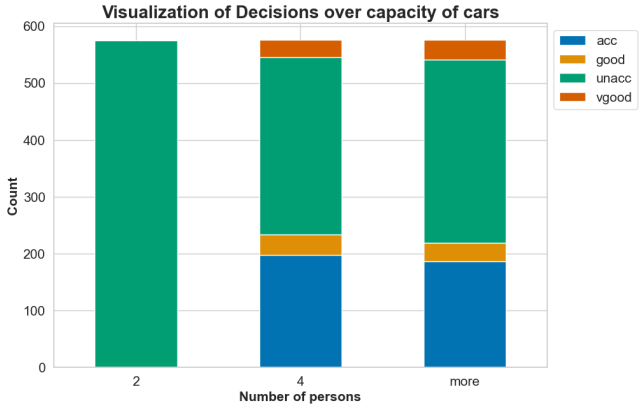

Fig. 4. Distribution of capacity of cars

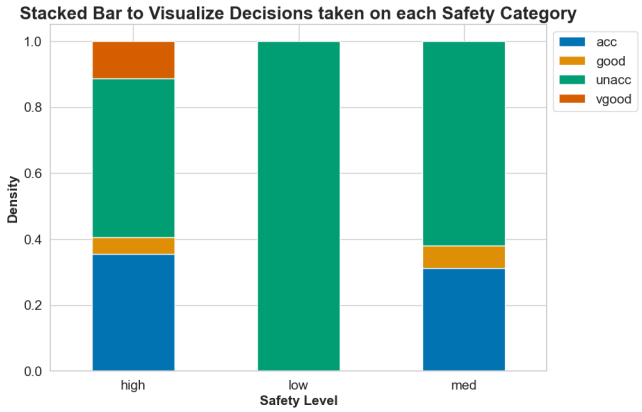

Fig. 5. Safety of cars with respect to decision

there is comparatively more probability of getting unaccepted.

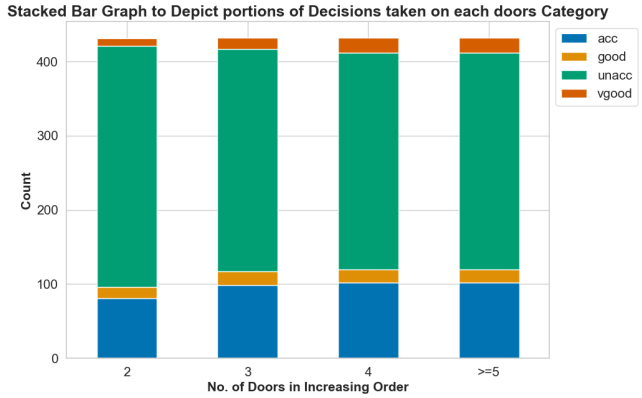

Fig. 6. Car evaluation based on the number of doors

Here the figure 7 depicts the relation between the luggage size and decision taken over car evaluation. It shows that more number of cars has facility to hold big size of luggage and more probability of getting classified into either acceptable, good, or very good. However, there is more probability of falling into unacceptable regardless of luggage size holding capacity.

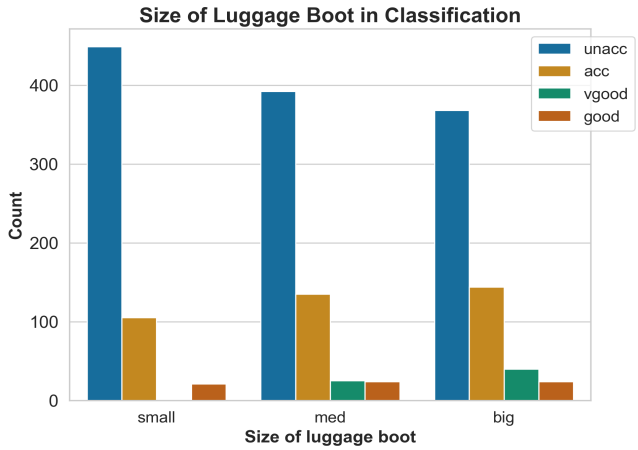

Fig. 7. Car evaluation based on the number of doors

III. Random Forest Classifier

A random forest is a supervised learning algorithm which is used for both classification and regression tasks [1]. It builds number of decision trees on the data samples and takes their majority vote for classification and average in case of regression. Here decision tree has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes. A decision tree starts with a root node, which does not have any incoming branches. The leaf nodes represent all the possible outcomes within the dataset.

Random forest learning employs an ensemble technique which helps in combining multiple decision tree models. Ensemble involves two different methods i.e Bagging and Boosting. Bagging creates a different training subset from sample training data with replacement and the final output is based on majority voting. Boosting combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. So, the random forest algorithm is based on bagging where as ADA Boost, XGBoost are based on boosting.

A. 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. Steps Involved in Random Forest Algorithm

Random forest algorithm can be divided into two stages.

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:-

1. Randomly select k features from a total of m features where k $<$ m.

2. Among the k features, calculate the node d using the best split point.

3. Split the node into daughter nodes using the best split.

4. Repeat 1 to 3 steps until l number of nodes has been reached.

5. 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.

1. We take the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome.

2. Then, we calculate the votes for each predicted target.

3. Finally, we consider the high voted predicted target as the final prediction from the random forest algorithm.

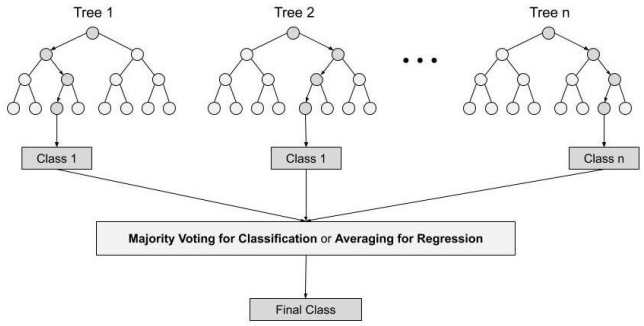

Fig. 8. Random Forest Combination of Decision Trees

C. Feature Selection with Random Forest

Random forests algorithm can be used for feature selection process. This algorithm can be used to rank the importance of variables in a regression or classification problem.

We measure the variable importance in a dataset by fitting the random forest algorithm to the data. During the fitting process, the out-of-bag error for each data point is recorded and averaged over the forest.

The importance of the j^{th} feature was measured after training. The values of the j^{th} feature were permuted among the training data and the out-of-bag error was again computed on this perturbed dataset. The importance score for the j^{th} feature is computed by averaging the difference in out-of-bag error before and after the permutation over all trees. The score is normalized by the standard deviation of these differences.

Features which produce large values for this score are ranked as more important than features which produce small values. Based on this score, we will choose the most important features and drop the least important ones for model building.

IV. Modelling

Random forest is a combination of multiple decision trees which is applicable for developing both classification and regression models [1]. With the help of bagging techniques, all the decision trees are combined and final model is developed with majority of votes. In the experiments, two different sets of number of estimators have been used. First set of 10 estimators and accuracy of model was found 94.74%. With the increase in number of estimators, the model accuracy also increases which has been demonstrated when the another set of 100 estimators are used and model accuracy was found 96.49%. By applying the correlation technique, a variable "doors" was found insignificant. So, another model without "doors" attribute was trained and its accuracy found to be improved. In order to evaluate, confusion matrix has been used and f-score, recall and precision have been found out for decision attribute. Finally, decision tree of the model has been generated to visualize the top-down splitting as shown in below figures. This paper helps to future data science enthusiast learn more about random forest algorithm and its implementation.

Acknowledgement

It is my pleasure to express with deep sense of gratitude to Professor Gaurav Raina, for his guidance, continual encouragement, understanding.

I would like to extend my gratitude to IIT Madras, for providing with an environment to work in and for his inspiration during the tenure of the course.

Conclusion

In this paper, random forest algorithm has been used to evaluate the car based on safety measures from the given data set. Data analysis and visualization has been implemented visualize the correlation among the variables and cause-effect. It has been found out that the accuracy of model is 94.74% with 10 estimators and 96.49% with 100 estimators. The fact is that car evaluation more sensitive towards the safety measures as if safety is compromised then car will be categorized as unacceptable. There are some features such as doors, and luggage size do not have so much impact on car evaluation. This paper also helps new researcher to get more information about random forest algorithm and its related terminologies.

References

[1] A. Chaudhary, S. Kolhe, and R. Kamal, "An improved random forest classifier for multi-class classification," Information Processing in Agriculture, vol. 3, no. 4, 2016, doi: 10.1016/j.inpa.2016.08.002.

[2] C. G. Siji George and B. Sumathi, "Grid search tuning of hyperparameters in random forest classifier for customer feedback sentiment prediction," International Journal of Advanced Computer Science and Applications, vol. 11, no. 9, 2020, doi: 10.14569/IJACSA.2020.0110920.

[3] N. Kandhoul, S. K. Dhurandher, and I. Woungang, "Random forest classifier-based safe and reliable routing for opportunistic IoT networks," International Journal of Communication Systems, vol. 34, no. 1, 2021, doi: 10.1002/dac.4646.

[4] S. B. O. Macaulay, B. S. Aribisala, S. A. Akande, B. A. Akinnuwesi, and O. A. Olabanjo, "Breast cancer risk prediction in African women using Random Forest Classifier," Cancer Treat Res Commun, vol. 28, 2021, doi: 10.1016/j.ctarc.2021.100396.

[5] H. Alqahtani, I. H. Sarker, A. Kalim, S. M. Minhaz Hossain, S. Ikhlaq, and S. Hossain, "Cyber intrusion detection using machine learning classification techniques," in Communications in Computer and Information Science, 2020, vol. 1235 CCIS. doi: 10.1007/978-981-15-6648-6_10.

[6] R. S. Rao and A. R. Pais, "Detection of phishing websites using an efficient feature-based machine learning framework," Neural Comput Appl, vol. 31, no. 8, 2019, doi: 10.1007/s00521-017-3305-0.

[7] A. R. Chowdhury, T. Chatterjee, and S. Banerjee, “A Random Forest classifier-based approach in the detection of abnormalities in the retina,” Med Biol Eng Comput, vol. 57, no. 1, 2019, doi: 10.1007/s11517-018-1878-0.
[8] X. Wang et al., “Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier,” BMC Med Inform Decis Mak, vol. 21, no. 1, 2021, doi: 10.1186/s12911-021-01471-4.

Source code link: https://drive.google.com/drive/folders/1HHrtAAO8ZGeHG0aGXi0Ccw4rMhAlsojs?usp=share_link