BIAS VARIANCE TRADEOFF IN CLASSIFICATION ALGORITHMS ON THE CENSUS INCOME DATASET

[1]PRACHI TARE, [2]SATYAM MISHRA, [3]MUKUL LAKHOTIA,[4]KUSHAGRA GOYAL

[1,2,3,4]Medicaps University, Indore, MP
Email: [1] prachitare11@gmail.com, [2] satyamgusmishra@gmail.com, [3] mukullakh09@gmail.com, [4] goyalkushagra77@gmail.com
Contact: [1]+91-7440854834, [2]+91-9131091600, [3]+91-8989688111,[4]+91-7987663475

Abstract:. Classification algorithms such as Decision Tree and Random Forest Algorithm are one of the most widely applied modeling techniques in the domain of machine learning. Machine learning models backed up by the classification algorithms are extensively used in a plethora of fields such as data analytics, image classification, computer vision, exploratory analysis, game AI etc. In such a situation, the model must have high accuracy, versatility and efficiency for a successful completion of the task in which it is implemented. However, irrespective of algorithm, the metrics by which a model's efficiency is judged is dependent on many factors like confusion matrix, accuracy score and so on. Of all these factors, one has to carefully maintain the balance between bias and variance and optimize model's performance. In this paper, bias and variance will be analyzed on the Census Income dataset with the decision tree and random forest algorithm. Furthermore, the paper will also incline towards the process of regularization and its impact on the balance between bias and variance and to cope up with the possible inconsistencies that might arise due to slight changes in the dataset values. The advantages and disadvantages of fluctuating values of bias and variance respectively depict the extent of model versatility on dataset irrespective of way of dividing the training, testing and validation data. In the paper, the algorithms will classify people into two categories based on the salaries (more than or less than $50K) with various predictor features.

Index terms: Machine Learning, Classification, Random Forest, Decision Tree, Bias, Variance, Loss Function

I. INTRODUCTION

The primary task of classification is to implement our model on the census income dataset and to deal with the target variable in order to categorize people in two groups on the basis of salary and profoundly analyze it. The two algorithms used in this process are namely Decision Tree and Random Forest algorithm. Furthermore, while dealing with datasets in general, there's a probability of encountering few inconsistencies in classification because of some parameters in our machine learning model. For instance, if there is a carton full of pebbles and few people take a guess to predict the number of stones, almost everyone will come up with the different answers. In the domain of machine learning, this paradox is resolved by maintaining a proper balance between variance and bias. In addition to this, the mean or average of output data are also used for generalizing the final verdict. In this manner, we end up with more efficient and accurate analysis.

II. BIAS VARIANCE TRADEOFF & CLASSIFICATION ALGORITHMS

We are given a significantly consistent dataset in order to classify individuals into two groups on the basis of their salary. Thus, we proceed with decision tree classification algorithm. The decision tree model is considered to be the building block of its successor - random forest classification algorithm. By implementing the decision tree, we deal with the bias and variance, which in turn play a vital role in categorizing the training set by making a valid decision tree to make our prediction more accurate. There are many types of decision trees used for classification including the ID3, CART, CHID, Hunt's Decision Tree, C4.5, C5.0 and many others. In decision tree algorithm, the basic process lies on the fact of following a top down approach beginning from the root node and following a specified path which is dependent on the values of attributes (and distributed recursions) based on the Sum of Product (Sop) or Disjunctive Normal Form approach. It finally leads to terminal leaf nodes and thus, the classification is accomplished. The decisive parameters in this classification method are Information Gain, Entropy and Gini Index.

Gain (T, X) = Entropy (T) - Entropy (T, X)

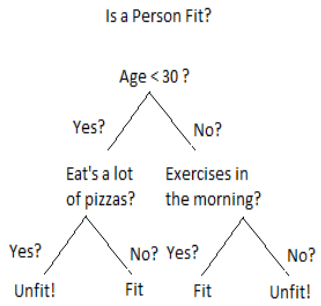

Figure 1: An Intuitive Example of Decision Tree

.

On the other hand, random forest algorithm is a state of the art supervised ensemble learning technique which uses bagging of many decision trees for classifying the target variable. In simpler terms, random forest algorithms uses combinations of decision tree for getting more efficient approach and higher accuracy. Random forest algorithm can be used for predictions via regression as well as categorization via classification. In this algorithm, the combination of decision tree also leads to a very useful criterion of feature importance. The basic principle behind random forest is same as that of decision tree. The only difference between both of these algorithms is the former one takes a combination of many decision trees (*'n_estimators'* hyperparameter of sklearn module in python).

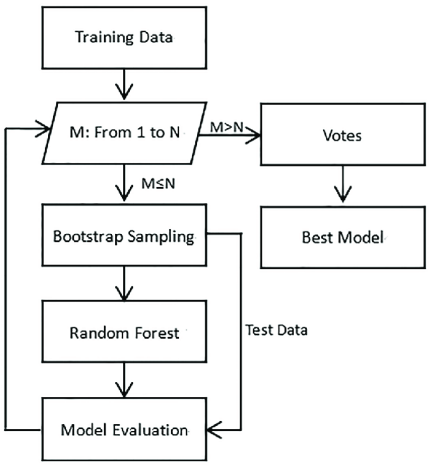

Figure 2: Flow Chart of Random Forest Algorithm

When it comes to the error in prediction and classification, it can be reduced by twitching several parameters within the machine learning model. Bias is one such entity which is used to make the model more efficient by analysing the difference of the exact value of the model and the average predicted value. Higher bias values make a machine learning algorithm faster and makes learning easier but at the same time, it gets more stringent. Decision tree algorithms have relatively lesser bias value as compared to that of linear regression and logistic regression.

In the context of machine learning, variance can be defined as the measure of the deviation in values of target variable if there will be a change in the training data. In simpler words, variance is the change in predictions by the model with change in the dataset. Decision tree algorithm generally has high variance (even higher in non pruned tree). Moreover, high variance leads to more flexibility of the algorithm but the influence of dataset is also increased in such situations.

In general, overfitting is observed when there is low bias with high variance. That means, the machine learning model also considers the noise along with the patterns of the dataset. Decision trees are more likely to experience overfitting of data because of complexities in their structure. On the other hand, underfitting is observed when there is high bias and low variance. That means, the machine learning model is not able to extract the patterns from the supplied data. Underfitting arises whenever there is comparatively less amount of data to train our model or we try to fit in non-linear data within a linear model.

Ideally, a machine learning model should neither be overfit nor underfit. A proper balance must be maintained amongst the bias and variance which make the machine learning model more efficient. This is known as the bias variance tradeoff where the end goal is to minimize the total error and the model complexity for efficient and faster prediction and classification.

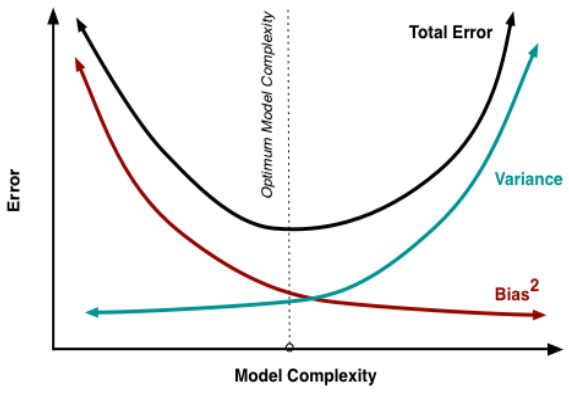

Figure 3: Error Complexity Curve

III. CENSUS INCOME DATASET

The dataset is obtained from the UCI Machine Learning Repository. Sometimes, the Census Income dataset is also referred as the 'Adult' dataset. It has 48,842 instances with 14 attributes. Barry Becker extracted this from the Census database back in 1994. The dataset contains continuous as well as categorical attributes. Our target variable will be the total salary (whether it is more than $50k or not) which can be encoded into a binary categorical predictor variable with proper data preprocessing.

The aforementioned dataset contains 14 columns with missing values and non integer values which need to be preprocessed in order to implement machine learning model. For the non integer values, label encoding and one hot encoding is done via scikit learn, whereas simple imputer (with 'mean' as a parameter) is implemented on the data points with missing values.

Random Forest Model was applied on the dataset with varying number of test sizes - 20%, 25%, 30%, 35%, and 40%. It was observed that the accuracy score was almost the same throughout with minute fluctuations. Furthermore, a bar chart was plotted with the help of matplotlib featuring the decimal places of the accuracy score (mean of the two obtained values) on Y axis and the test size on the X axis. The mean of accuracy score was observed to be 85.27%.

Test Size	Accuracy Score
20.00%	85.68%
20.00%	85.81%
25.00%	85.34%
25.00%	85.52%
30.00%	85.23%
30.00%	85.27%
35.00%	84.70%
35.00%	85.05%
40.00%	84.96%
40.00%	85.15%

Figure 4: Random Forest Scores

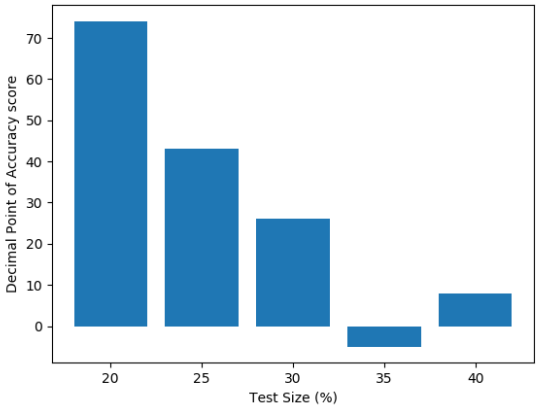

Figure 5: Random Forest Accuracy Plot

It can be comprehended from the above statistics that when random forest classification was implemented on the census income dataset, we get approximately same accuracy irrespective of the size of training data and test data. It only fluctuates within the decimal places.

Decision tree classification was implemented on the same proportions of test and training data as in the random forest classification. Moreover, the bar graph depicts the comparison of 5 values of test percentages. The mean of overall accuracy scores is 81.33%.

Test Size	Accuracy
20.00%	80.88
20.00%	81.06
25.00%	81.34
25.00%	81.58
30.00%	81.07
30.00%	81.32
35.00%	81.4
35.00%	80.99
40.00%	81.44
40.00%	81.1

Figure 6: Decision Tree Scores

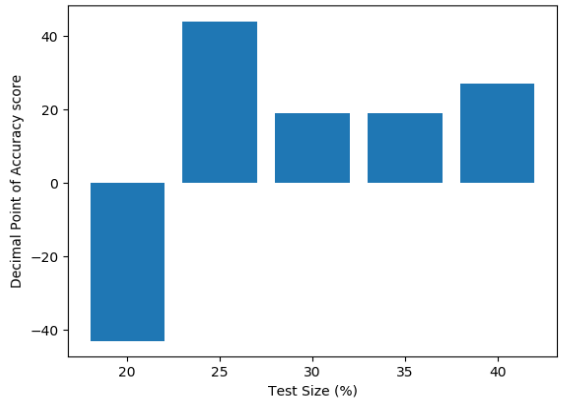

Figure 7: Decision Tree Accuracy Plot

IV. VALIDATION CURVES AND LEARNING CURVES

Validation curve and learning curve are often used to examine the generalization of the model on fluctuating test datasets. Validation curve visualizes the performance over many values for a range of hyperparameters. On the other hand, learning curve is used to determine the effect of number of parameters used for efficiency of model and performance metric. The validation curve for 100 trees in random forest classification is shown below. It is followed by the validation curve of decision tree classifier with max_depth parameter 50. It can be clearly observed that the parameter values for optimum results grows initially and then it either becomes constant or degrades because of hyperparameters. Whereas, learning curve depicts the correlation amongst training and test (cross validation) scores with a number of training set sizes. The learning curves of random forest classification and decision tree classification models are depicted below. The curve is based upon 50 different sizes of training set. In addition to this, the number of trees in the random forest algorithm (for the curve) is 10.

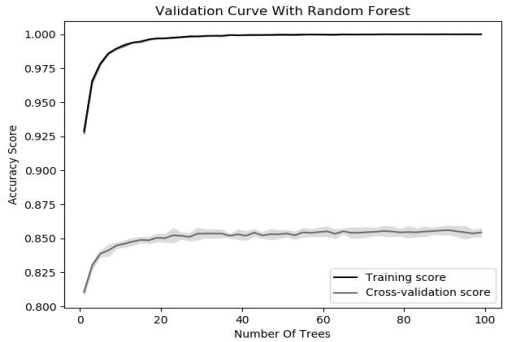

Figure 8: Random Forest Validation Curve

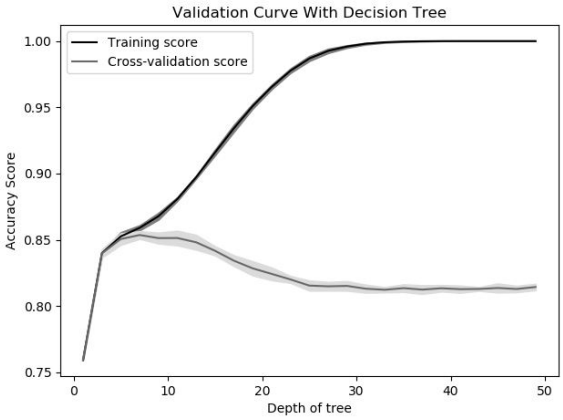

Figure 9: Decision Tree Validation Curve

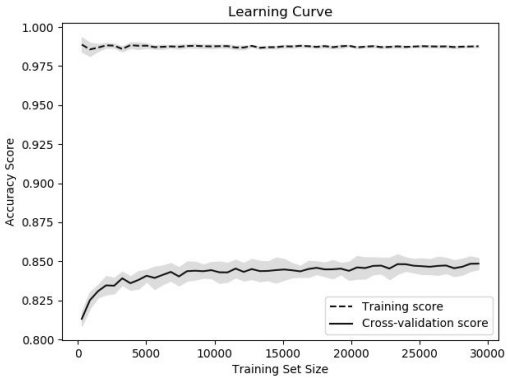

Figure 10: Random Forest Learning Curve

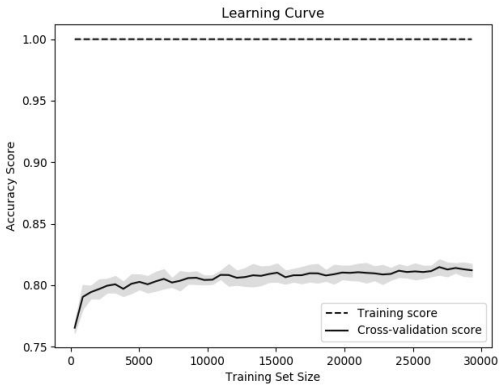

Figure 11: Decision Tree Learning Curve

V. INFERENCE

The inference from the two different models leads to the conclusion that both of these models can be improved and can be more efficient if there would've been more data points in the adult dataset. Moreover, the accuracy scores can be increased further by implementing other machine learning algorithms. In addition to this, it can be concluded that the accuracy scores change but with a lower fluctuation with the change in dataset. This implies

that there is a relatively low variance (consistent model) but is a little bit

inaccurate due to higher bias. Therefore, to gain the optimum efficiency, the model with a slightly more variance can be selected with proper parameter tuning as per the curves plotted so that there is a proper balance between bias and variance. Thus, bias variance tradeoff is at minimum level for maximum efficiency of machine learning model.

CONCLUSION

While applying the decision tree model and random forest model to understand the tradeoffs amongst bias and variance, we come up with the conclusion that the testing accuracy of random forest is more than that of decision tree. Since random forest classification is composed of a number of decision trees (n_estimators), it has a slight edge over decision tree. Apart from accuracy scores, we can conclude that the bias variance tradeoff in this dataset cannot be eradicated completely, but a 'sweet spot' can be found after analyzing the parameters such that the machine learning model gives optimum results.

REFERENCES

[1] Pedro Domingos, "*A Unified Bias-Variance Decomposition and its Applications*," *Proceedings of 17th International Conference on Machine Learning,* pp. 231-238, June 2000

[2] Sisay Menji Bekena, "*Using decision tree classifier to predict income levels*," *Munich Personal RePEc Archive,* paper no. 83406, December 2017

[3] Navoneel Chakraborty, Sanket Biswas, "*A Statistical Approach to Adult Census Income Level Prediction,*" *arxiv.org,* arXiv:1810.10076 *[cs.LG],* version 1, October 2018

[4] S. Deepajothi, Dr. S. Selvarajan, "*A Comparative Study of Classification Techniques On Adult Data Set,*" *IJERT,* vol. 1, issue 8, Oct. 2012

[5] Lichman M., "*Adult Income Dataset*," University of California, Irvine, 2013

[6] Vidya Chockalingam,Sejal Shah, Ronit Shaw, "*Income Classification using Adult Census Data*"

[7] Chet Lemon, Chris Zelazo and Kesav Mulakaluri, *"Predicting if income exceeds $50,000 per year based on 1994 US Census Data with Simple Classification Techniques"*