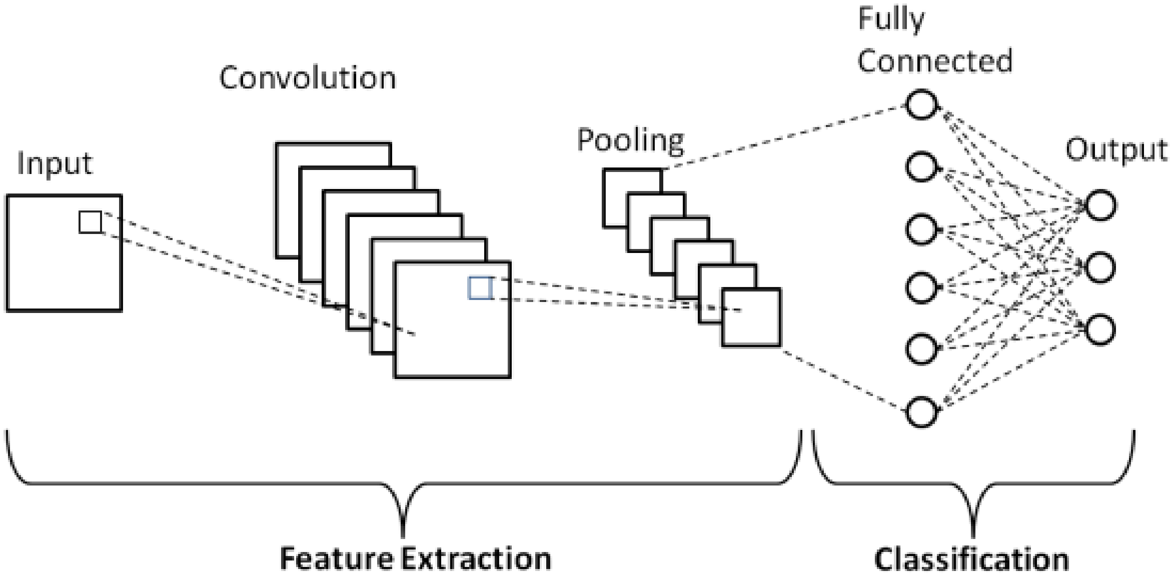
the Comparison between Machine Learning Approaches and Statistical Model

Dongzhou Huang*and Weicheng Qian†

April 30, 2020



*Department of Statistics, Rice University
†Department of Statistics, Rice University

1 Introduction

Machine learning approaches attract lots of people's interest by their amazing ability in improving prediction accuracy. Over the past decade, machine learning approaches have been widely adopted in a number of fields to improve accuracy. Traditional fields contain image recognition, natural language processing and artificial robots. With the increasing influence of machine learning approaches, they have been applying in some other fields such as medicine, biology, chemical engineering and even history and art. At this point, machine learning approaches have been a tide for academic research and industrial implement are will be lasting for the following decades.

On the contrary, statistical model, the previous star, is no longer popular in analysing data. Machine learning approaches are gradually encroaching upon the territories belonging to statistical model. With doubt on the validity of p-value, people lose confidence on statistical model, and turn to machine learning approaches, even with the risk of losing interpretability. More and more students in statistics change their research interest and begin to study machine learning approaches, which reflects the change of hot topic of research.

Does this phenomenon means that statistical model should be abandoned into the garbage cans? Are machine learning approaches always superior to statistical model? As students in statistics, we are really concerned for the comparison between machine learning approaches and statistical model. It is our responsibility to make some contributions to those problems, even our answer may not be accepted.

Throughout the history, many scientists may come up with their own answers. There are three paper in the literature about comparison between machine learning approaches and statistical model. For example, the authors of [4] apply eight different models (five machine learning approaches containing back propagation neural network and support vector machine and three statistical models containing ARIMA and space-time model) to the 2-minute travel speed data collected from three Remote Traffic Microwave Sensors located on a southbound segment of 4th ring road in Beijing City and conclude that machine learning approaches outperform two traditional statistical models on the prediction performance. The authors of [2] compare traditional statistical and machine learning methods in the field of landslide susceptibility modeling. Without sufficient geotechnical data, it is impossible to conduct physically-based model. The authors apply seven models (including logistic regression, support vector machine, random forest and boot strap aggregated classification trees) to geotechnial data and draw the conclusion that random forest and boot strap aggregated classification trees have overall best prediction performance. These two paper both show machine learning approaches are better that statistical models. However, the third paper [5] provides a new idea, that is, combining both machine learning approaches and statistical model to increase prediction accuracy. The authors use machine learning approaches such as a minimum classification error (MCE) and support vector machine (SVM) to exploit automatically prior knowledge obtained from the speech database, which are implemented to assist the voice activity detectors based on statistical models to improve prediction performance.

However, in the above paper, the criterion used to judge which model is better is prediction accuracy, which ignore the biggest difference between machine learning approaches and statistical model. One of the advantages of statistical model is that we can do statistical inference to interpret the results and to select variables based on population distribution. For example, in linear regression, relying on t-test, we can easily rule out irrelevant variables if the corresponding parameters are not significant. Moveover, some diagnosis tools are devised to check the validity of the statistical model. In contrast, machine learning approaches do not possess such advantages. Although principal component analysis is the common way to select variables for machine learning approaches, it is well accepted that method based on correlations is problematic and would draw ridiculous conclusions. Machine learning approaches are more likely to a black box, of which we only know the input and the output but have little knowledge about the inside structure. As a result, to explain how they work is hard and to understand why they work is impossible.

Keeping such difference in mind, we are curious about whether the interpretability of statistical model will improve prediction accuracy. In order to make interpretability play its role in prediction, we design two kinds of experiments.

In the first experiment, we first generate determined variables $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and noise variables $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_m)$. And then, based on $\mathbf{X}$, we use logistic model to generate label variable Y, i.e.,

$$\mathbb{E}\left[Y = 1 | \mathbf{X}\right] = g\left(\beta \cdot \mathbf{X}\right), \tag{1}$$

where β is the parameter vector selected by us and $g(x) = 1/(1 + e^{-x})$ is the logistic function. At this point, we obtain feature data $(\mathbf{X}, \mathbf{Z})$ and label data Y. If we apply logistic regression to the data set, thanks to statistical inference, we can easily deduce that Z is irrelevant to Y and should be ruled out. And the prediction accuracy are expected to be high, since the true model and the applied model are coincident. However, as for machine learning approaches, lacking of inference tools, the influence of noise variables cannot be eliminated completely. As a result, we expect that the prediction accuracy of machine learning approached cannot be better than that of logistic regression.

In the second experiment, we also use the same determined variables $\mathbf{X}$ and noise variables $\mathbf{Z}$. But the label variable is different and is generated by

$$\mathbb{E}\left[Y = 1|\mathbf{X}\right] = h\left(\beta \cdot \mathbf{X}\right), \tag{2}$$

where β is the parameter vector selected by us and $h(x) = (2\pi)^{-1}\tan^{-1}(x) + 1/2$. In this case, we do not expect the performance of logistic regression would be good. The reason is that the statistical inference may not be valid if the true model is different with the applied model. And we expert the performance of machine learning approaches may be better. This experiment design is motivated by [1] and [3]. The reason is that machine learning approaches rely less on the structure of the input data. By implementing appropriate machine learning technique, the prediction accuracy is expected to be improved remarkably.

The rest of the section is devoted to four main models we apply to our constructed data sets. they are logistic regression, decision tree (DT), random forest (RF) and convolutional neural networks (CNN). Since logistic regression is a statistical model, which is familiar to students in statistics, we skip the introduction. Thus, we will focus mainly on decision tree, random forest and CNN.

The first machine learning model we used is the decision tree (DT). Decision tree is a non-parametric classification method, which uses a set of rules to predict that each observation belongs to the most commonly occurring outcomes of the training data including class labels, event outcomes and resource costs. The model has a flowchart structure in which the internal node functions as the test of an attribute (variable), each branch indicates the results of the test, and each leaf node represents the class label. The whole path form root to leaf nodes represents the classification rules. Since Decision Tree suffers from the issue of overfitting from time to time, we can use k-fold cross-validation, which randomly partitions the data set into folds of similar size, to see if the tree needs any pruning which can create a more robust model as well as make it more interpretable for us.

Next, we applied the random forest (RF) model to our data. Random forest is similar to the decision tree method in that it builds trees, hence the name "random forest". This is an ensemble learning method which creates a multitude of decision trees (in our case 500 trees) at training time,and outputting the class that occurs most frequently among them. A great amount of relatively uncorrelated models (trees) functioning as a committee will outperform any of the individual constituent models. Another advantage that random forest has over decision trees is the element of randomness which it avoids the issue of overfitting that decision trees might have.

Finally, we used the convolutional neural networks (CNN), which it is a deep learning model or a multi-layer perceptron similar to artificial neural network. This model motivated by image analysis is basically a hierarchical model, which the raw data can be RGB imagines and raw voice data, etc. The CNN structure utilizes procedures like convolution, pooling and non-linear activation function projection to stack its layers. It extracts high level information from raw data input layer and abstract the information layer by layer, which this step is called "feed-forward". In the convolution layer, it calculates the dot product of input and filter weight matrix and output it as the result. The pooling is used to reduce the feature space's dimension with the loss of its depth. In the non-linearity layer, it uses activation functions like RELU or SoftMax instead of traditional Sigmoid or Tan-H activation function. For every negative input, the RELU function will return as zero. And for all the positive inputs, RELU will return as the same value. In the fully connected layer, it compiles the data extracted by the previous layers to form the final output as the objective function. By calculating the error or loss between the predictions and actual values, CNN utilizes back-propagation algorithm to back-forward the error or loss from the last layer to the first layer. After updating the parameters in every layer, the CNN iterates the back-propagation again till the model converges to achieve the purpose of model training.

2 Methodology and Data Description

In the first experiment (see (1)), we generate two sets of data, and we call them L1 and L2. The first set L1 contains two determined variables ($X_1 \sim Unif[1, 10]$ and $X_2 \sim Bernoulli(0.5)$) and three noise variables ($Z_1 \sim \mathcal{N}(1, 1)$, $Z_2 \sim Poisson(5)$ and $Z_3 \sim Bernoulli(0.7)$), and the label variable Y are generated according to (1). The second set L2 contains more variables, fourteen determined variables and seventeen noise variables, and they are summarized in Table 1.

In the second experiment, we also generate two data sets T1 and T2. The determined variables and noise variables in T1 and T2 are the same to those in L1 and L2 respectively. The only difference is that the label variable Y is generated according to (2), not (1).

Table 1: Variable Description

$X_1, \ldots, X_6$: Normal distribution	X_7, X_8 : Categorical distribution
$X_9, \ldots, X_{12}$: Bernoulli distribution	X_{13}, X_{14} : Uniform distribution
$Z_1, \ldots, Z_7$: Normal distribution	Z_8, Z_9 : Categorical distribution
$Z_{10}, \ldots, Z_{14}$: Bernoulli distribution	Z_{15}, Z_{16} : Poisson distribution
Z_{17} : Uniform distribution	Y: generated according to (1)

3 Models Comparison and Results Discussion

3.1 Model analysis of L1

We first apply logistic regression to data set L1. We use stepwise variable selection based on AIC criterion, and X_1, X_2 and Z_2 are selected as the explanatory variables. The result surprises us. To explain it, we observe that the inference for logistic regression is based on asymptotic distribution, that means, when the sample size tends to infinity, the distribution of test statistic converges to chi square distribution. The size of L1 is only 35000, much less than infinity, that is why there exists error in the inference.

We then apply machine learning approaches (decision tree, random forest and CNN) to L1, and we draw the plot of decision tree (Figure 1) and the plots of ROC curve for each model (Figure 2). And the prediction performance of each model is summarize in Table 2.

As shown in Figure 1, the decision tree select correct variables (X_1 and X_2) as explanatory variables, which shows the algorithm of decision tree is quite robust and principal component analysis is valid when the number of variable is small.

As shown in Figure 2 and Table 2, the prediction performance of each model is quite closed, since the prediction accuracy of each model is nearly 78%. This result contradicts our previous expectation among these models, that is, the interpretability in statistical model would not assist to increase the prediction performance. On the other hand, machine learning approaches are no better than statistical model for this kind of data set, even a little bit worse.

And a phenomenon attracts our interest, that is, the prediction performance of decision tree is quite close to other models. According to previous experiences, decision tree is normally worse that random forest or CNN. However, in our data set, there are few variables, so that decision tree can also capture the essence of the data.

Figure 1: Decision Tree of L1

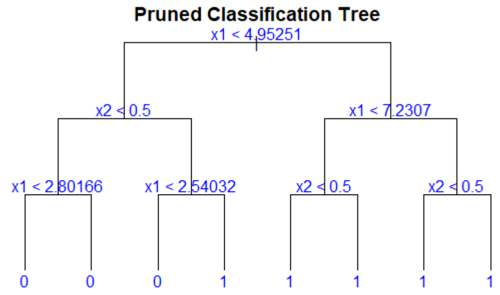
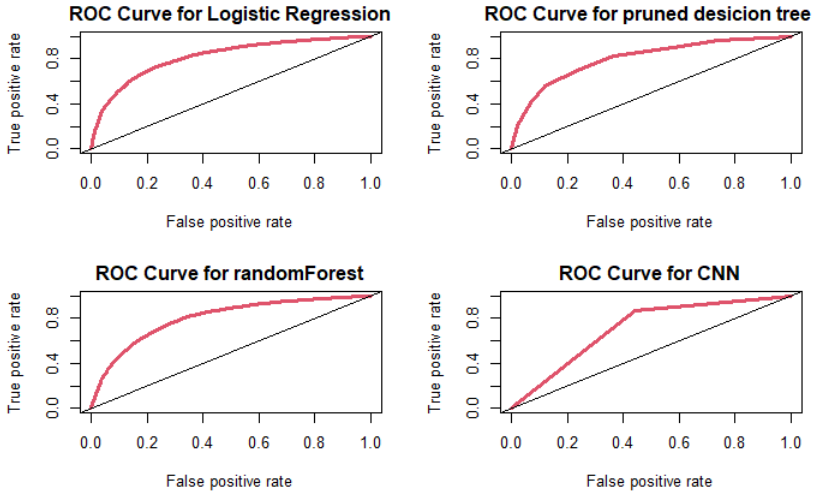


Figure 2: ROC curves for L1



3.2 Model analysis of L2

For data set L2, we also apply logistic regression, decision tree, random forest and CNN and draw ROC curve of each model (Figure 3). And the prediction performance of each model is summarize in Table 3.

In contrast to the previous data set L1, we can see the prediction performance of decision tree on data set L2 is significantly worse than other models. The possible explanation is that when the number of variables increases, the algorithm of decision tree would not extract sufficient variable for training. Actually, only variables $X_1, X_2, X_3, X_4, X_5, X_6$ and X_{12} are used to construct the tree and other $X's$ are discarded. Since decision tree only exploit a part of information, the prediction is not as accurate as other models.

Moveover, the performance of models other than decision tree on data set L2 is better that that on L1. A possible reason is that with the increase of variables, more information are provided to train

Table 2: Prediction Performance of each Model in L1

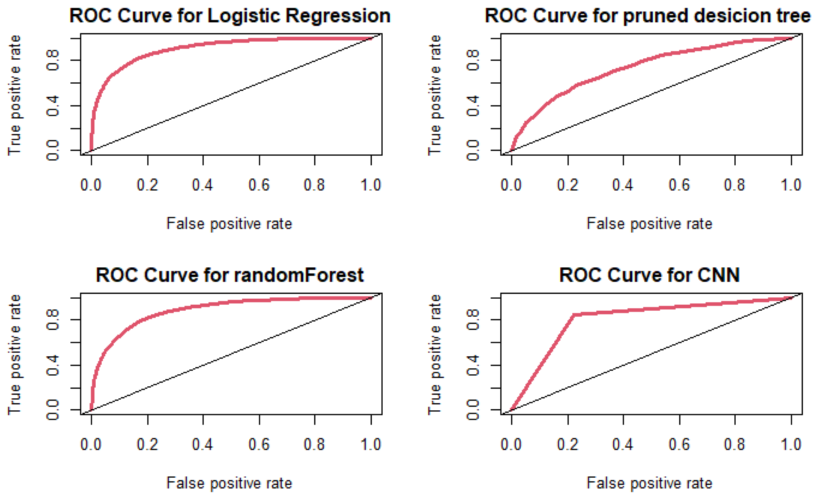
Model	Accuracy	Kappa	AUC
Logistic Regression	78.02%	42%	81.88%
Pruned Decision Tree	77.37%	46%	80.39%
Random Forest	78.03%	43%	80.30%
CNN	77.97%	43%	70.15%

Table 3: Prediction Performance of each Model in L2

Model	Accuracy	Kappa	AUC
Logistic Regression	82.64%	65%	90.83%
Pruned Decision Tree	67.20%	34%	73.72%
Random Forest	80.94%	62%	88.88%
CNN	81.58%	63%	81.56%

the models, that means, higher accuracy.

Figure 3: ROC curves for L2



3.3 Model analysis of T1

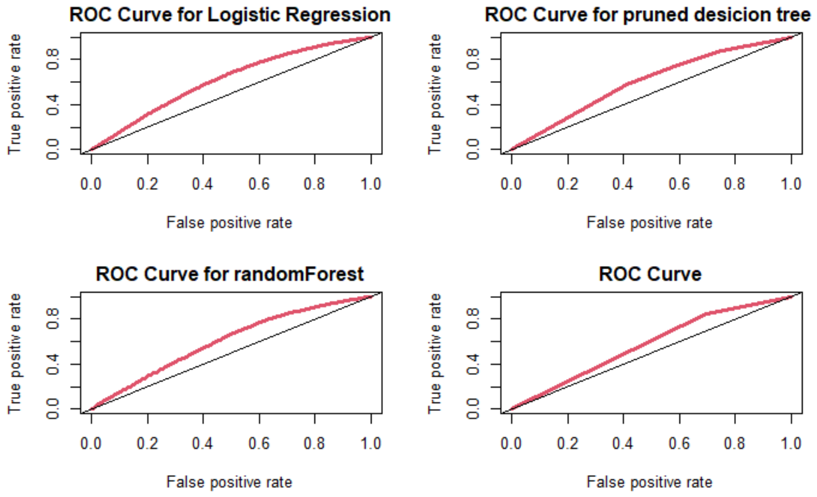
For data set T1, we also apply logistic regression, decision tree, random forest and CNN and draw ROC curve of each model (Figure 4). And the prediction performance of each model is summarize in Table 4.

Table 4: Prediction Performance of each Model in T1

Model	Accuracy	Kappa	AUC
Logistic Regression	62.77%	17%	61.87%
Pruned Decision Tree	62.55%	14%	60.35%
Random Forest	62.59%	16%	60.51%
CNN	62.71%	16%	57.62%

As shown Figure 4 and Table 4, we can see the prediction performance of all the models is not ideal. As for logistic regression, we expect that result is not good, since the true model is not logistic model. However, we are very surprised that the performance of machine learning approaches is even worse. This phenomenon indicates that machine learning approaches do not outperform logistic regression, otherwise, the prediction accuracy of machine learning approaches would be far better than logistic regression on this type of data.

Figure 4: ROC curves for T1



3.4 Model analysis of T2

For data set T2, we also apply logistic regression, decision tree, random forest and CNN and draw ROC curve of each model (Figure 5). And the prediction performance of each model is summarize in Table 5.

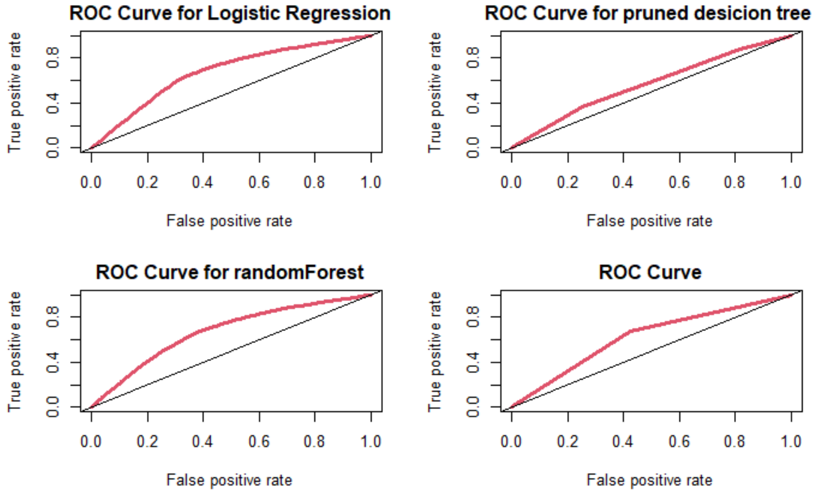
We observe that the prediction performance of decision tree is extremely worse that other models. The reason is explained in previous cases. And again, in contrast to performance of models on T2,

Table 5: Prediction Performance of each Model in T2

Model	Accuracy	Kappa	AUC
Logistic Regression	64.83%	30%	67.76%
Pruned Decision Tree	55.13%	11%	56.68%
Random Forest	64.43%	29%	67.43%
CNN	63.46%	27%	63.37%

the accuracy of most models is increasing.

Figure 5: ROC curves for T2



4 Conclusion

- The Interpretability of statistical model does not help to increase prediction performance.
- Even for the data set where which is suitable to apply logistic regression, the prediction performance of machine learning approaches are no better than that of logistic regression.
- Machine learning approaches and statistical model are equally effective.
- With increase of variables, the performance of decision tree becomes worse while the prediction accuracy of other models is increasing.

References

[1] George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[2] JN Goetz, Alexander Brenning, H Petschko, and P Leopold. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Computers & geosciences*, 81:1–11, 2015.

[3] Norbert Jankowski and Marek Grochowski. Comparison of instances seletion algorithms i. algorithms survey. In *International conference on artificial intelligence and soft computing*, pages 598–603. Springer, 2004.

[4] Han Jiang, Yajie Zou, Shen Zhang, Jinjun Tang, and Yinhai Wang. Short-term speed prediction using remote microwave sensor data: machine learning versus statistical model. *Mathematical Problems in Engineering*, 2016, 2016.

[5] Jong Won Shin, Joon-Hyuk Chang, and Nam Soo Kim. Voice activity detection based on statistical models and machine learning approaches. *Computer Speech & Language*, 24(3):515–530, 2010.