Cloning Machine Intelligence

Student Name: Mridul Gupta
Roll Number: 2015061

BTP report submitted in partial fulfillment of the requirements
for the Degree of B.Tech. in Computer Science & Engineering

on November 19, 2017

BTP Track: Research

BTP Advisor
Dr. Debarka Sengupta

Indraprastha Institute of Information Technology
New Delhi

Student's Declaration

I hereby declare that the work presented in the report entitled **Cloning Machine Intelligence** submitted by me for the partial fulfillment of the requirements for the degree of *Bachelor of Technology* in *Computer Science & Engineering* at Indraprastha Institute of Information Technology, Delhi, is an authentic record of my work carried out under guidance of **Dr. Debarka Sengupta**. Due acknowledgements have been given in the report to all material used. This work has not been submitted anywhere else for the reward of any other degree.

..... **Place & Date: .....**
Mridul Gupta

Certificate

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

..... **Place & Date: .....**
Dr. Debarka Sengupta

Abstract

Machine learning models find their applications in areas where data is both readily available and can be used to solve the given problem. In today's research space, a diversity of such models can be encountered with each one of them having its characteristics of solving the problem. Although these mathematical models are quite adept when applied to the data, the accuracy and confidence of the model are mainly attributed to the data, which in some cases is hard to conjure. This report tries to find new approaches in which a machine learning model can be cloned using a random normal or uniform distribution of data. The approach mainly focuses on generating random data and its corresponding labels for a classification problem using a trained classifier and using the class distribution obtained on this random data to train a similar or different classifier. We validate the ability of this approach by scrutinising how our new model which is fitted with the random data points behaves on the original data. The approach is also verified using different machine learning models with different complexities of data and produces better results for parametric classifiers. The report also argues about how classifiers that perform classification using hyperplane approximation can perform better, thus providing a gateway for machine learning models to learn about the problem from a model rather than the data.

Keywords: **Intelligence Cloning, White Box, Data-less Models**

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor Dr. Debarka Sengupta for providing excellent guidance and being supportive throughout the span of the thesis. Without his patience, critiques and thoughtful insights, the work would've never taken wings. Earnest thanks to my friend Najeeb Khan who has likewise been a noteworthy contributor to this project.

A very special thanks to IIITD for providing an excellent infrastructure, environment and a flexible curriculum to carry out my work at a suitable pace. I am deeply grateful to all the faculty and administrative staff here for their incredibly supportive attitude. I dedicate this work to my parents, who have always been there for me and provided me support both emotionally and fiscally.

Contents

Chapter 1

Introduction

Machine learning has rapidly taken over a substantial part of the industry due to its ability to solve rather complex problems and the sudden surge of data being available because of the internet. With the rate that the application of Machine Learning is growing, it is not far away when this field of study finds widespread application in sensitive areas such as security and medicine (Seurfert et al. and Magoulas et al.)

With the scope of Machine Learning increasing, it becomes necessary to check for the various security issues that can arise when applying a machine learning model to a given problem. Goodfellow et al. have described exploitation of adversarial examples to degrade the confidence of Deep Learning models when predicting the outputs but there has not been any particular study regarding a general overview of a majority of Machine Learning models.

In this report, we observe the tendency of a machine learning model to clone a black box model using nothing but random data on binary classification datasets. Heuristics are devised to come up with novel approaches to change random data in such a way that a model when trained on this data can mimic the performance of the black box model. The report also argues about devising a data-less approach towards learning.

The first chapter discusses the naïve approach of generating random data and using it to judge the performance of cloned model on the original dataset. We also describe the general methodology implemented in the cloning procedure. The second chapter discusses the improvements that can be made to the generation of the random dataset so that the cloned model produces a better accuracy on the datasets into consideration. The third chapter deals with the introduction of equitable class random data and its introduction. We also compare this method with other approaches discussed so far. The fourth chapter discusses the performance of different statistical and Deep Learning models on each of the approaches and provides a hypothesis as to why some models are particularly better at being cloned and why others are not. The report ends with a summary and conclusion along with some future work that we wish to perform to bolster the proposed cloning procedure further.

Chapter 2

Description of Datasets

For the purpose of validating our hypothesis, we used some publicly available datasets from University of California, Irvine (UCI) machine learning repository. The datasets used were of varying size, both in terms of sample size and feature size, so as to further corroborate our findings. The following datasets were used to for binary classification problem

1. **Iris Dataset**: This dataset has 4 features and a total sample size of 100. The dataset is sliced to consider only 2 of the 4 classes
2. **Make Moons Dataset**: This dataset has 2 features and a total sample size of 100. The make moons dataset has a non linear nature.
3. **Breast Cancer Dataset**: This dataset has 30 features with a total sample size of 569 data points.
4. **Heart Dataset**: The Heart dataset has 9 features and 462 data points.
5. **Adult Census Dataset**: The Adult Census dataset has 107 features with 32561 data points.
6. **Bridges Dataset**: This dataset has 29 features with a total of 106 points.
7. **Mushroom Dataset**: This dataset has 112 features with 8124 data points.

The following datasets were used for validating the hypothesis on multiclass classification problem:

1. **Iris Dataset**: This dataset has 4 features and a total sample size of 100. The number of classes in this dataset are 3.
2. **Abalone Dataset**: The Abalone dataset has 10 features and a total of 4177 data points. The number of classes in this dataset is 3.
3. **Car Evaluation Dataset**: This dataset has 16 features with a total of 1728 data points. The number of classes in this dataset is 3.

Chapter 3

Methodology

3.1 Naïve Approach : Generating Random Data from Zero Centered Normal Distribution

We perform the cloning procedure for a model trained on binary classification dataset using any of the statistical methods available. The approach can be easily applied to a model trained on multi-class classification dataset as well as a regression problem by extrapolating the algorithm. The black box model comprises of any machine learning model that has been trained on a dataset. The black box model should produce a good cross-validation accuracy for the original dataset. Cloning of a black box is achieved by predicting the labels on a random dataset or a dataset generated by applying some heuristics. The cloned model is then further trained on this new collection of random data points and their corresponding labels. To determine whether the cloned model is successfully able to mimic the working of the black box we check for the accuracy of the cloned model on the original dataset. Our primary objective is to increase the accuracy of the cloned model on the original dataset and make it as close as possible to the black box classifier.

Under the naïve approach, the random data is generated from a random normal distribution with zero mean and unary standard deviation. Since a dataset comprises of a number of features, each feature value is generated from this normal distribution. The number of features can be easily determined with the help of the black box as it will only accept data of the form [number of samples x number of features]. The algorithm 1 elucidates the generation of random dataset and the cloning procedure.

Algorithm 1: Algorithm for cloning model using random data

Data: blackbox , whitebox, number of features, number of samples

Result: a cloned model

```
data = array of zeros of size [number of features, number of samples];
for each feature in the data do
    feature = normal distribution (mean = 0, standard deviation = 1.0);
    label = blackbox.predict(data);
    whitebox.fit(data, label);
    // Checking the performance of the whitebox on original data
    print whitebox.score(original data, original labels);
```

We performed this experiment in a controlled environment only a certain combination of statistical models as black box and white box (cloned) models. The accuracy score was determined as the ratio of number of correctly classified samples to total number of samples. The observations of the experiment are described in table 3.1

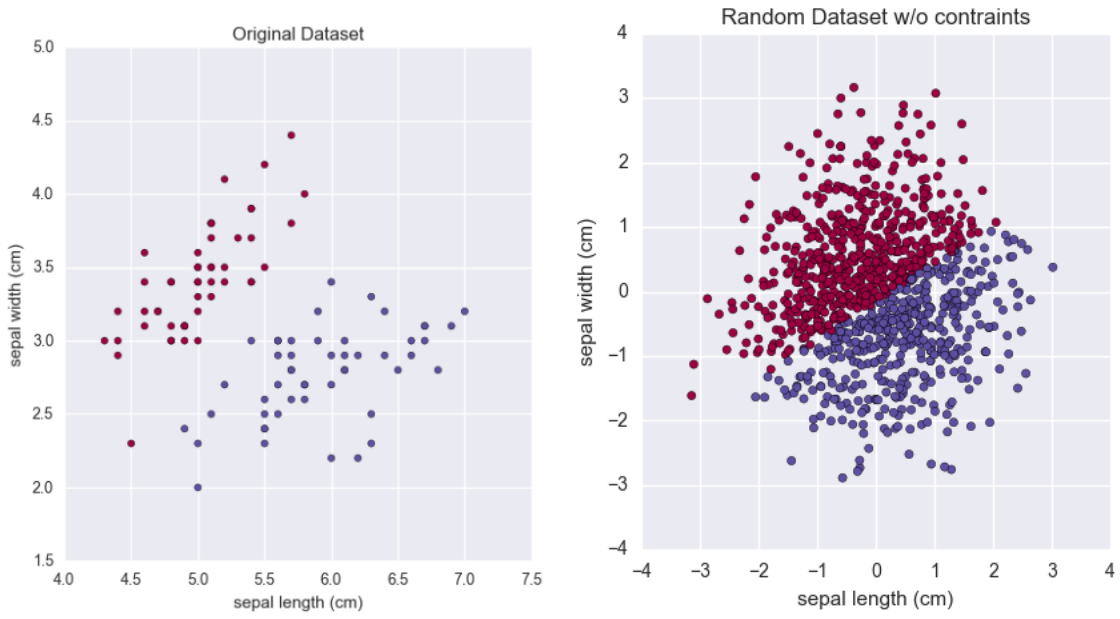
It can be observed from table 3.1 that the black box and white box accuracy has a large difference when the number of features for the dataset increases. The high accuracy for two feature dataset can be attributed to the fact that generating random data from zero centred normal distribution in two dimensions is able to provide sufficient information to the white box about the hyperplane that it successfully approximates the original hyperplane. Approximating the original hyperplane allows the white box to perform accurately on the original dataset.

Dataset Used	Number of Features	Black Box	Black Box Accuracy	White Box	**White Box Accuracy**
Iris	2	Logistic Regression	0.99	Logistic Regression	**0.97**
BUPA	7	Logistic Regression	0.59	Logistic Regression	**0.57**
Heart	9	Logistic Regression	0.73	Logistic Regression	**0.54**
Breast Cancer	30	Logistic Regression	0.95	Logistic Regression	**0.53**
Make Moons[1]	2	Neural Network	0.99	Random Forest	**0.98**

Table 3.1: **Cloned model accuracy on original dataset when trained with random dataset**

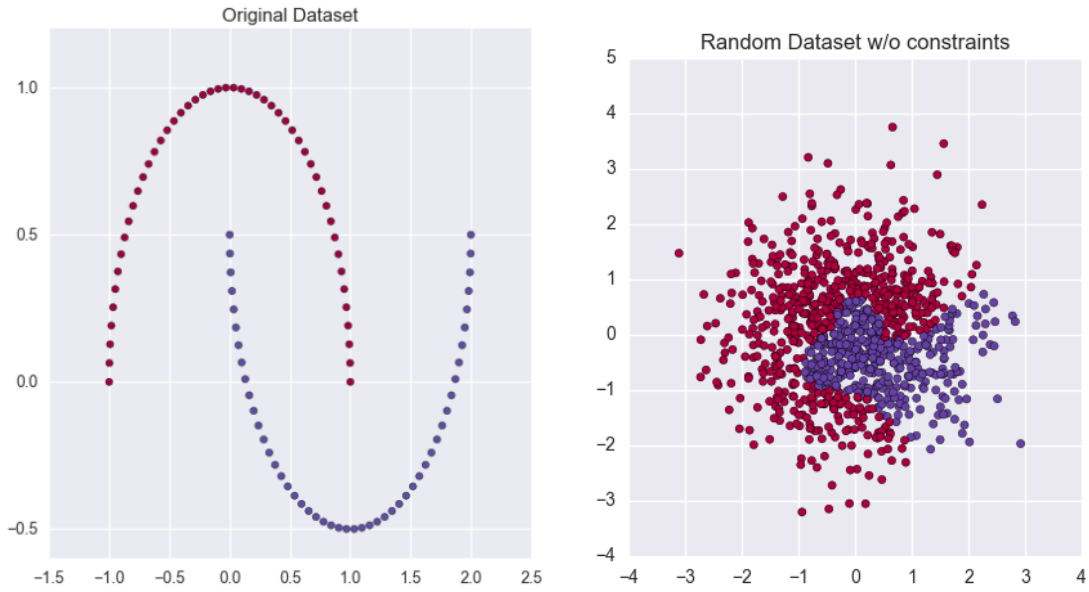
Figure 3.1a shows the original Iris Dataset in 2 dimensions whereas 3.1b shows the randomly generated dataset using algorithm 1. One can observe that in such low dimensions the cloned model (white box) gets enough information from the random dataset to mimic the hyperplane of the black box. The same can be visualized in the Make Moons nonlinear dataset from Figure 3.2a and Figure 3.2b. For higher dimensional data we can observe from the trend of accuracy that this naïve approach tends to perform poorly, only getting around 50 percent of the classification labels correct in its predictions.

[1]Make Moons is a two dimensional non linear dataset. It was evident from its visualisation that a linear classifier would work poorly on it. So we decided to fit it with a non linear network i.e. a single hidden layer neural network with softmax outputs



(a) (b)

Figure 3.1: Iris Dataset in Two Dimensions: (a) Original dataset (b) Randomly generated data



(a) (b)

Figure 3.2: Make Moons Dataset: (a) Original dataset (b) Randomly generated data

3.1.1 Shortcomings of the Naïve approach

Some of the shortcomings of the discussed approach are as follows

1. As visualized earlier, the naïve approach does not scale well when complexity of data concerning the number of features increases. The cloned model accuracy remains lower

even when the black box accuracy is quite high for the same statistical model.

2. Since the random data generated is always derived from zero centered normal distribution, the random data will not accurately work for models which were trained on data which does not have zero mean and unary standard deviation. This schism between the data will widen in the case of nonlinear classifiers in which the hyperplane may change drastically outside the feature values.

3.2 Generating random data under constraints

Machine learning models specifically depend on the data to solve the problem. It is quite common that the data is derived from a natural phenomenon and each of the feature value thus recorded lies under a minimum and maximum value. For example, considering the Iris dataset, the sepal width and sepal length may well be under some specific range. We hypothesize that exploiting this range to generate random data can bolster the performance of our white box even in high dimensions.
Algorithm 2 discusses the generation of random data under constraints. The cloning procedure is rather similar, and the performance of the white box is judged using the same metric.

Algorithm 2: Algorithm for cloning model using random data under constraints

Data: blackbox , whitebox, number of features, number of samples, range of features
Result: a cloned model

```
data = array of zeros of size [number of features, number of samples];
for each feature in the data do
    feature = uniform distribution under constraints(range of feature);
    label = blackbox.predict(data);
    whitebox.fit(data, label);
    // Checking the performance of the whitebox on original data
    print whitebox.score(original data, original labels);
```

3.2.1 Observations and Results for the Constrained Approach

The observations are summarized in table 3.2-3.5. We can observe that the black box, as well as white box, have almost similar accuracy for all of the datasets. If the black box tends to fail on a given dataset and the similar statistical model is used as the white box, then it can be observed that the white box also performs poorly on that dataset. This provides information about the non-augmenting nature of cloning. No additional strength is gained by the white box when we clone it. The observations favor our hypothesis that random data generated under some constraints helps the cloned model in finding the appropriate position of hyperplane and producing fairly positive results regarding accuracy.

Dataset Used	Number of Features	Black Box Accuracy	**White Box Accuracy**
Make Moons	2	0.86	**0.88**
Heart	9	0.75	**0.72**
Bridges	29	0.93	**0.90**
Breast Cancer	30	0.96	**0.94**
Census	107	0.80	**0.76**

Table 3.2: Performance of Logistic Regression as Black Box and White Box model on different datasets

Dataset Used	Number of Features	Black Box Accuracy	**White Box Accuracy**
Make Moons	2	0.94	**0.92**
Heart	9	0.77	**0.71**
Bridges	29	0.91	**0.90**
Breast Cancer	30	0.62	**0.61**
Census	107	0.79	**0.75**

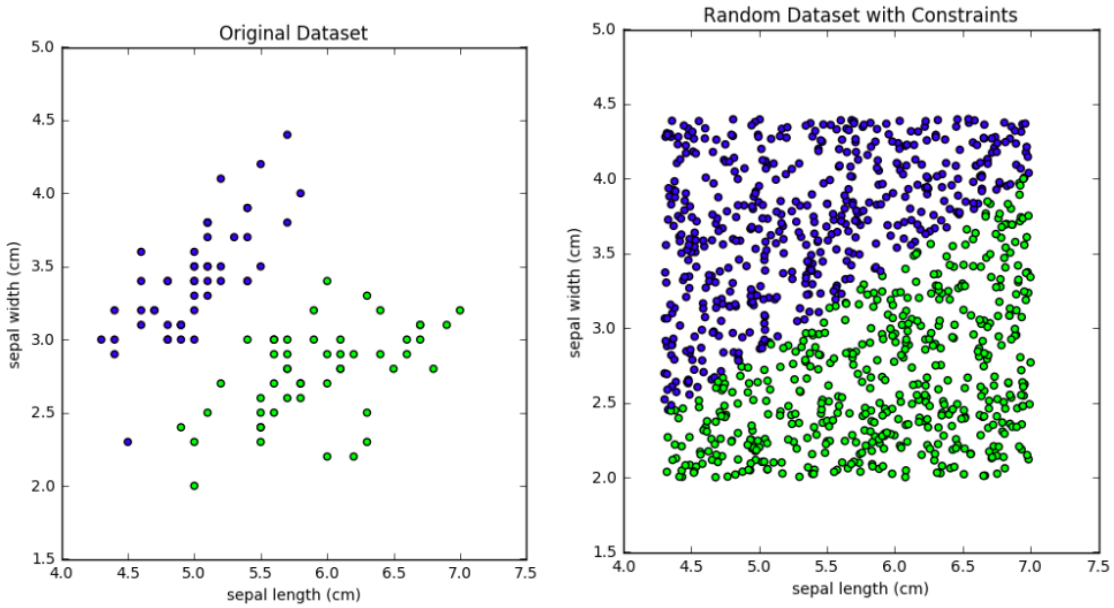
Table 3.3: Performance of Neural Networks as Black Box and White Box model on different datasets

Dataset Used	Number of Features	Black Box Accuracy	**White Box Accuracy**
Make Moons	2	0.99	**0.88**
Heart	9	0.72	**0.55**
Bridges	29	0.98	**0.72**
Breast Cancer	30	0.95	**0.37**
Census	107	0.84	**0.66**

Table 3.4: Performance of K-Nearest Neighbors as Black Box and White Box model on different datasets

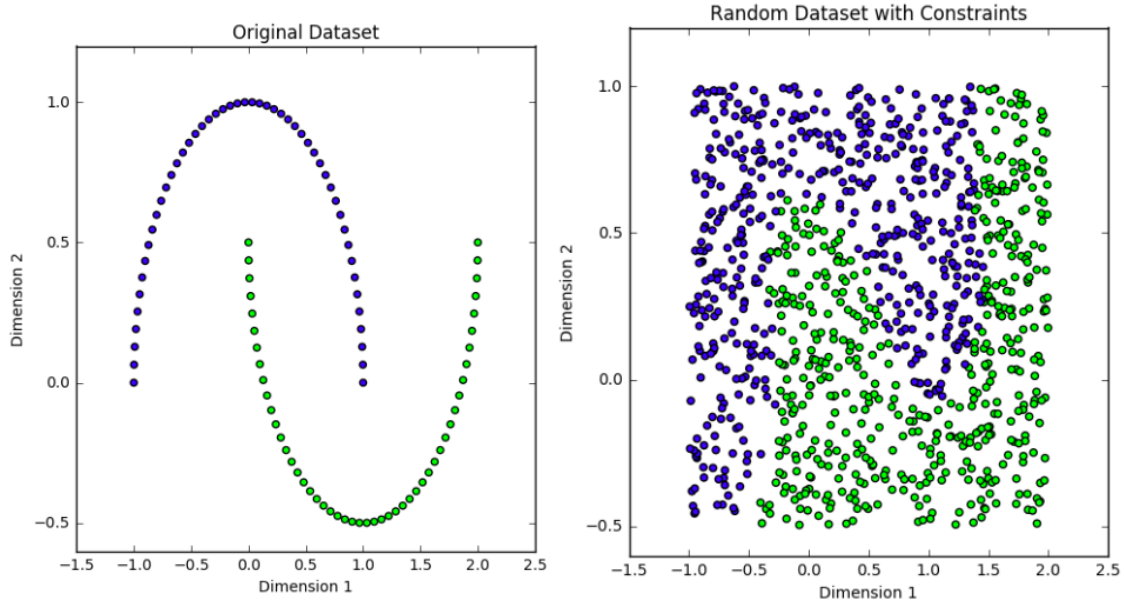
Dataset Used	Number of Features	Black Box Accuracy	**White Box Accuracy**
Make Moons	2	0.99	**0.99**
Heart	9	0.98	**0.72**
Bridges	29	0.98	**0.13**
Breast Cancer	30	0.98	**0.37**
Census	107	0.98	**0.24**

Table 3.5: Performance of Random Forest as Black Box and White Box model on different datasets



(a) (b)

Figure 3.3: Iris Dataset in Two Dimensions: (a) Original dataset (b) Randomly generated data under constraints



(a) (b)

Figure 3.4: Make Moons Dataset: (a) Original dataset (b) Randomly generated data under constraints

Figure 3.3b and Figure 3.4b illustrates the random dataset generated by putting constraints on the data generation procedure. Unlike figure 3.1b and figure 3.2b the hyperplane is more clearly visible and can further provide with increased confidence to the model about the position of the hyperplane.

3.2.2 Shortcomings of constrained random dataset

The only shortcoming of this approach is the **assumption that the feature values lie under a set of constraints**. This assumption may not hold well for dataset where the data is synthetically generated, or the feature can have an infinite domain.

3.2.3 Performance analysis of this approach on different classifiers

Performing this experiment on a single type of linear classifier can only guarantee its effectiveness on models that perform classification using any of the parametric approaches, i.e., by generating the hyperplane. Apart from this, we also need to observe the accuracy of this method on non-parametric models such as Random Forest Classifier. Figure **??** provides some new insights about this approach when it is applied on a dataset with nine features.
We observed in figure **??** that Random Forest Classifier when used as a black box and a model to be cloned there exists a rather gap between the classification accuracy. The black box has a high accuracy whereas the white box performs poorly. This observation shows that such models (due to their non-parametric nature) can be difficult to clone and simple hyperplane approximating procedure cannot guarantee accurate cloning.

[1]Make Moons is a two-dimensional nonlinear dataset. It was evident from its visualization that a linear classifier would work poorly on it. So we decided to fit it with a nonlinear network i.e. a single hidden layer neural network with softmax outputs

Chapter 4

Discussion

The cloning procedure is greatly determined by the dimensionality and type of machine learning based model that is employed as Black Box and White Box. We analyse the cloning procedure and its accuracy using different datasets of varying dimensionality and sizes. We take into account the type of the model by keeping a dataset constant to illustrate how different models responds to the cloning procedure.

4.1 Analysing Binary Classification Dataset

We use binary classification dataset to understand whether the proposed idea can be efficiently used to clone the information contained in the Black Box model into White Box model. The binary classification dataset provides for easy visualisation of hyperplanes involved in separating the data in low dimensions. The visualisation can be used to check whether the White Box can correctly replicate the position of hyperplane as the Black Box model learned it. The results display a further corroboration of this idea in high dimensional binary classification datasets, thus proving the general applicability of this approach. We propose that the random data, when generated under a specified constraint value for each corresponding feature, can efficiently capture the position of the dividing hyperplane in high dimensions. This allows for transferring the captured hyperplane efficiently in a new machine learning model such that it behaves similarly to the original Black Box model.

4.2 Analysing Multi-class Dataset

The multi-class classification is an extension of binary classification with more than two classes expected for a data point. The same approach of generating random data under constraints for each feature can be used to clone the information present in the Black Box model trained on such datasets. The results strengthen random data can also capture that complex hyperplane structures and transferred by training another statistical model on the newly created dataset.

Chapter 5

Conclusion

Generation of random data that can be readily accepted by the black box and allow for its perfect cloning is a rather challenging task. Nonetheless, there are some conclusions to be drawn from the experiments performed.

1. Approximating the hyperplane and its correct position is essential to clone any of the statistical models that perform classification using a parametric approach. The hyperplane provides enough information about the black box that once the white box is trained using a data cleanly separated by it, the black box is essentially cloned.

2. **Clonable v/s Non-clonable Models**

   The large gap between the Black Box and White Box 5-fold cross-validation accuracy of the non-parametric classifiers shows that the hyperplane cloning procedure does not allow models such as these to transfer information from Black Box to White Box, hence frustrating the cloning procedure. The parametric models such as Logistic Regression, Linear Discriminant Analysis etc. gives good results where the Black Box and White Box accuracy is very close for a single dataset.

   The approach to approximate hyperplane makes it difficult to clone non-parametric models such as Decision Trees and Random Forests. Unlike approximating functions, which parametric functions do, non-parametric models use complex decision procedure to perform classification, thus rendering them difficult to be cloned.