How to classify a Pulsar?

Name: Eunbin Ko

1. Introduction

Pulsars are a type of neutron star that emit beams, which give scientific interest. Here, there is data for detecting Pulsars with 8 different variables. The Class is binary classification with 0 being negative, which means the detected star is not Pulsar and 1 being positive, which means detected star is Purlar. For this project, classification tree, logistic regression, boosting method, and random forest methods are used to fit and train the model, then confusion matrix is used to compare the testing error of these models. Then the best methods can be selected among all. By using the model interpretability with variable importance measures and partial dependence plots, the most important variables can be classified.

2. Datasets

'HTRU2 Data Set' from UCI website ‘https://archive.ics.uci.edu/ml/datasets/HTRU2’ contains 8 predictor variables and one class variable, which has been set as Y. Each predictor variable corresponds to:

1. Mean of the integrated profile.
2. Standard deviation of the integrated profile.
3. Excess kurtosis of the integrated profile.
4. Skewness of the integrated profile.
5. Mean of the DM-SNR curve.
6. Standard deviation of the DM-SNR curve.
7. Excess kurtosis of the DM-SNR curve.
8. Skewness of the DM-SNR curve.

The sample view of data frame after the cleaning shows below.

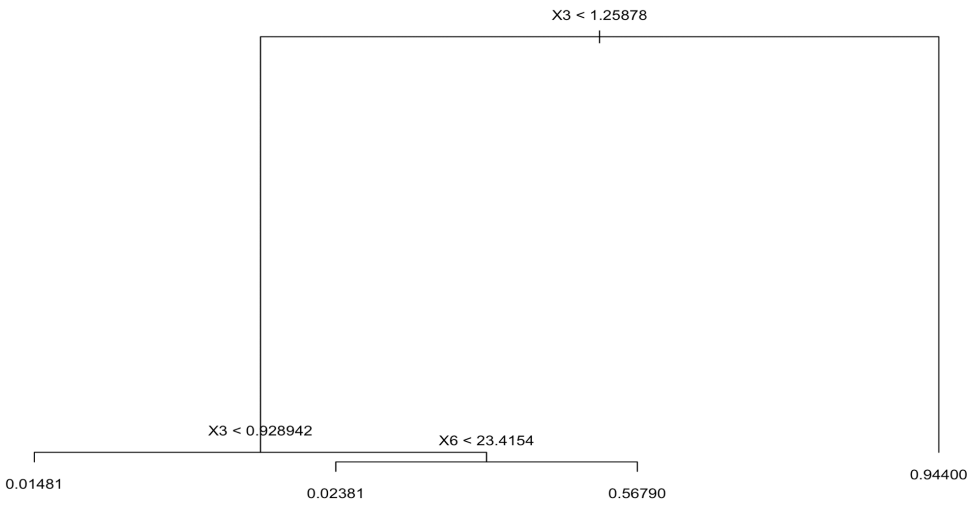
	Y <int>	X1 <dbl>	X2 <dbl>	X3 <dbl>	X4 <dbl>	X5 <dbl>	X6 <dbl>	X7 <dbl>	X8 <dbl>
1	0	102.50781	58.88243	0.46531815	-0.5150879	1.677258	14.86015	10.576487	127.39358
2	0	103.01562	39.34165	0.32332837	1.0511644	3.121237	21.74467	7.735822	63.17191
5	0	93.57031	46.69811	0.53190485	0.4167211	1.636288	14.54507	10.621748	131.39400
7	0	130.38281	39.84406	-0.15832276	0.3895404	1.220736	14.37894	13.539456	198.23646
8	0	107.25000	52.62708	0.45268802	0.1703474	2.331940	14.48685	9.001004	107.97251
11	0	133.25781	44.05824	-0.08105986	0.1153615	1.632107	12.00781	11.972067	195.54345

Then, separate this cleaned data into training and testing dataset. Here, the training dataset has randomly sampled and contains half of the size of the original data. Then the testing dataset is the remaining half of the dataset.

Remark: Since generating training and testing data is different every time due to random sampling, there could have different results but will not vary too much.

3. Classification tree

We will use several methods to determine the most important variables in the dataset. First, let's use a tree method to see the result. Below is the plot of fitted tree.



From the tree, we can see that the variables actually used are “X3” and “X6”, which are “Excess kurtosis of the integrated profile” and “Standard deviation of the DM-SNR curve”, respectively. If the value of “X3” is greater than 1.26, it is most likely assumable that detected star is a Pular, since the value of node is close to 1. When “X3” is less than 0.93, then the detected star is most likely not a Pulsar since the value of node is close to 0. When “X3” is between 0.93 and 1.26, then we consider the “X6”. If “X6” is smaller than 23.42, then we still consider the detected star is not a Pulsar. And if “X6” is greater than 23.42, then there is about half of the percentage that the detected star is a Pulsar.

There are 4 terminal nodes in the tree, and residual mean deviance is 0.01967 = 175.9 / 8944, and test error is 0.01849794. Since both values are very small and negligible, the result of above tree can consider as very accurate.

4. Comparison between Logistic regression, Boosting, Bagging, and Random Forest

Boosting, bagging, and random forest methods are applied to the training dataset for evaluation. By testing each of the performance, the accuracy of each different methods will be shown. Also, by comparing these results with logistic regression, we can know what is the best approach to this dataset.

First, boosting method has the misclassification error of 0.0203, and the confusion matrix is shown below:

boosting_class	0	1
0	8101	132
1	50	666

Next, bagging method has the misclassification error of 0.0190, and the confusion matrix is shown below:

bagging_class	0	1
0	8086	105
1	65	693

Then, random forest has the misclassification error of 0.0187, and the confusion matrix is shown below:

rf_class	0	1
0	8097	113
1	54	685

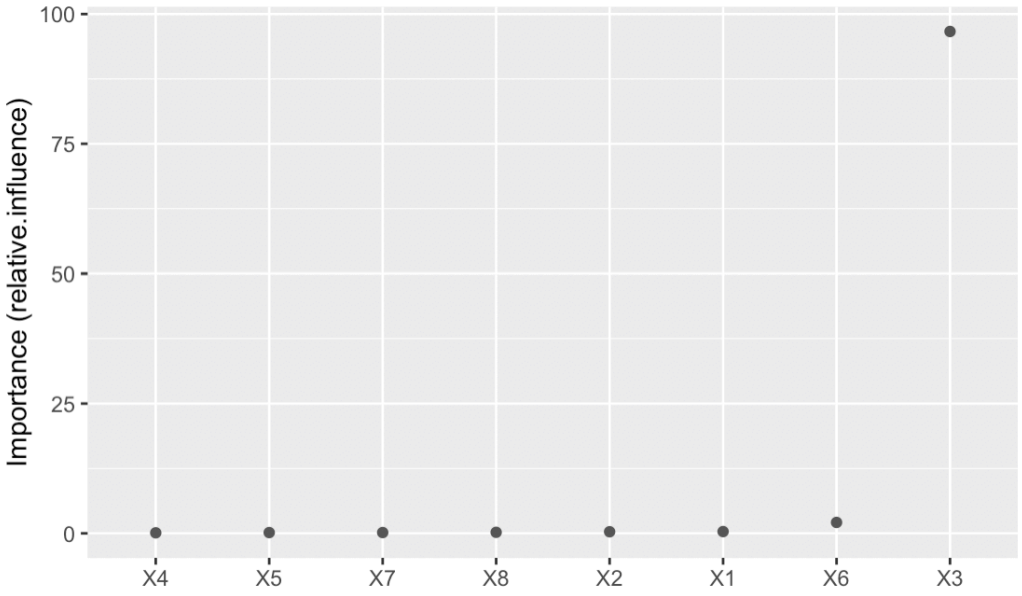
Lastly, compare these results with logistic regression, which has the misclassification error of 0.020. Below is the confusion matrix:

logistic_class	0	1
0	8097	54
1	125	673

According to the result above, all of the methods have great performance and misclassification errors are very small. Comparing with logistic regression, boosting, bagging, and random forest all have better prediction than regression analysis, and among them random forest has the best performance in terms of misclassification error rate. In fact, all methods are all reliable prediction.

5. Find most important variables using boosting method

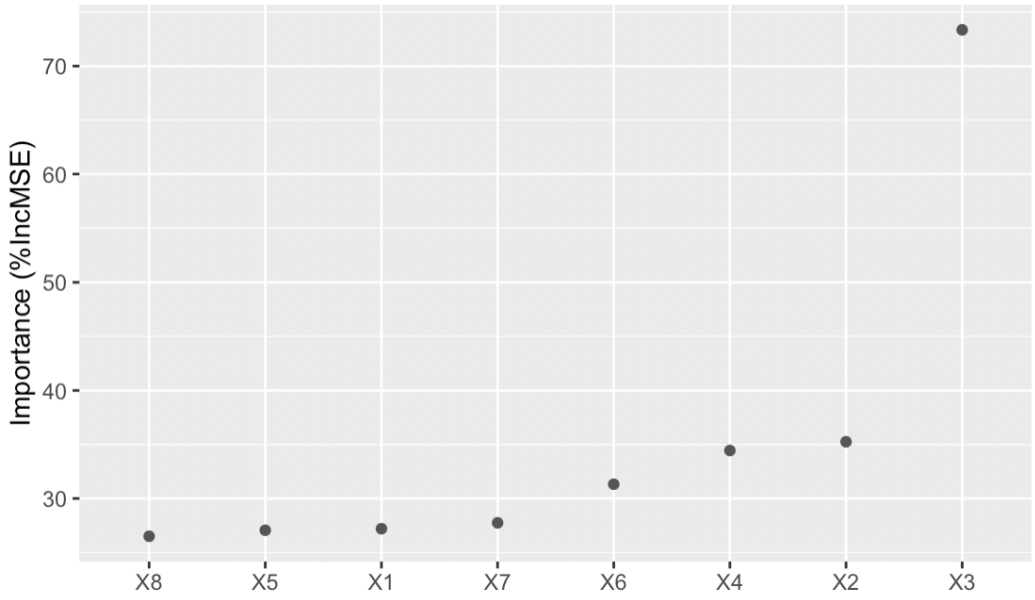
From boosting method, the most important variables can be found through plotting the variable importance plot. Below is the result plot:



According to the above variance importance plot and relative influence values, "X3", which is the excess kurtosis of the integrated profile is the most important variables among all other variables. It is prominently defined due to high relative influence value on "X3". The second important variable is "X6", which is the standard deviation of the DM-SNR curve though its relative influence is far less than "X3".

6. Find most important variables using random forest

From random forest method, the most important variables can be found through plotting the variable importance plot. Below is the result plot:

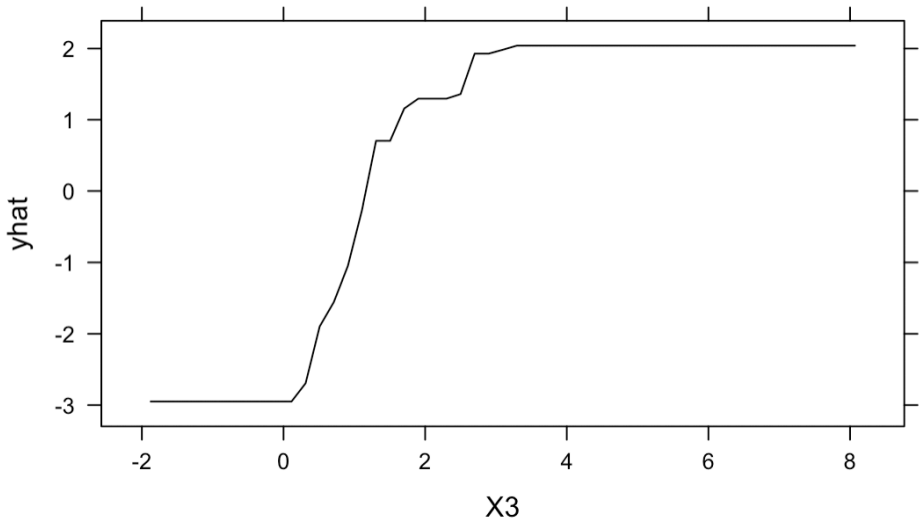


The first most important variable is “X3”, which is same as the boosting method, but the second important variable is different from boosting method. Here, the second most important variable is “X2”, that is the standard deviation of the integrated profile, which is different from the boosting method.

7. Partial Dependence Plot from boosting method

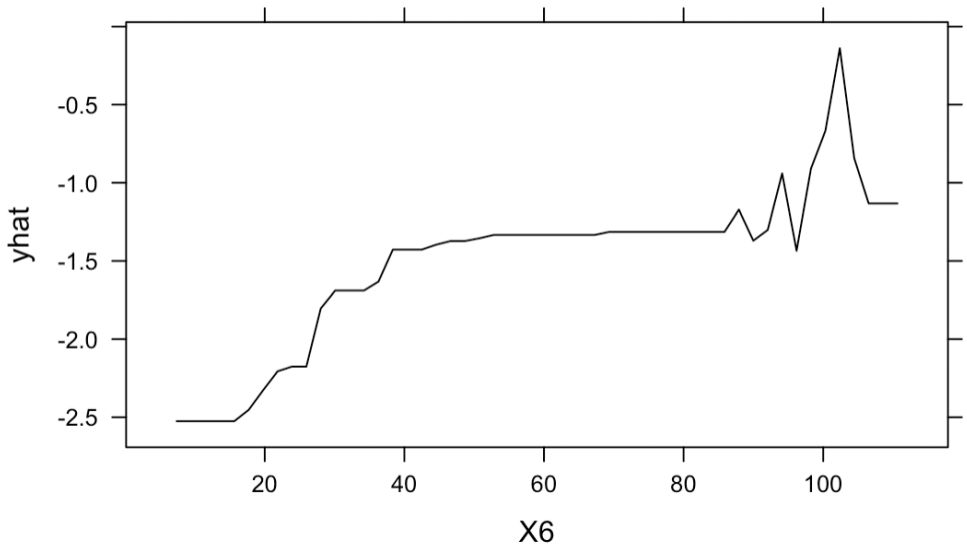
Since random forest and boosting method give the different important variables, the partial dependence plot for the first two most important variables are fitted.

First see the boosted tree method’s partial dependence plot for variable “X3”, which is the excess kurtosis of the integrated profile.



From above partial dependence plot of "X3", it shows that predicted variable will not change when "X3" is less than 0 or "X3" is greater than 3. However, there is rapid increase in predicted variable when "X3" is about between 0 and 2.

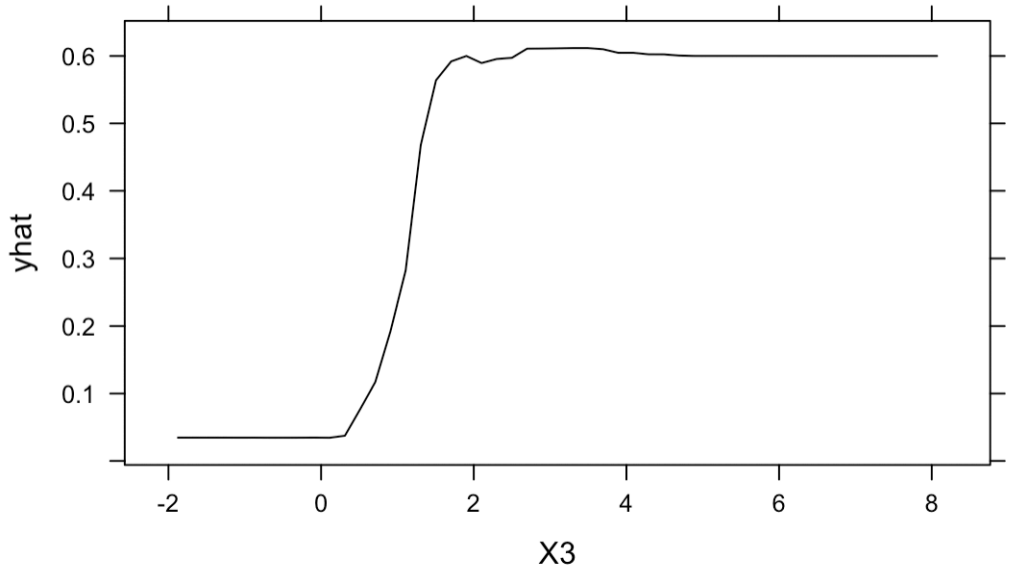
Next see the boosted tree method's partial dependence plot for variable "X6", which is the Standard deviation of the DM-SNR curve.



From above partial dependence plot of “X6”, it shows that predicted variable will increase until “X6” is about 40, then it will remain the same until about 85, then there is some up and downs with a peak at around 100.

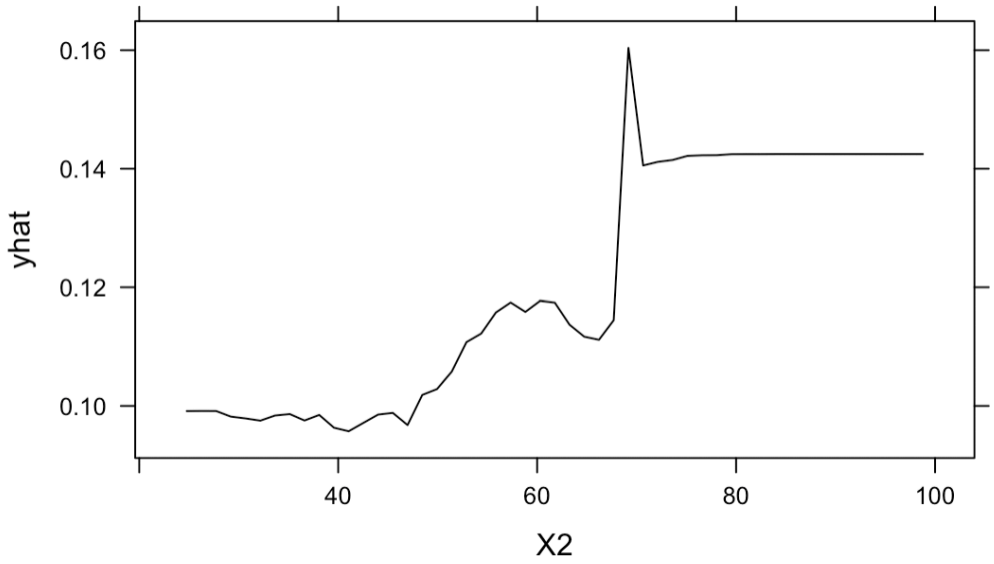
8. Partial Dependence Plot from random forest

Now see the partial dependence plot using the random forest for variable “X3”, which is “Excess kurtosis of the integrated profile”.



This above partial dependence plot is fairly similar to the boosting method’s. The range of “X3” that affect the predicted variable is little narrower than the boosting method. When “X3” is between about 0.5 and 1.5, the predicted variable will increase rapidly. Otherwise, predicted variable will not change.

Next see the random forest’s partial dependence plot for variable “X2”, which is “Standard deviation of the integrated profile”.



This partial dependence plot of "X2" shows that it has a large peak at around 70. After this peak, the predicted variable decrease rapidly and then remains the same after all. Therefore, when "X2" is at 70, it affects the prediction the most.

9. Conclusion

There are four testing method being used in this project. They are classification tree, logistic regression, boosting method, and random forest. The testing error for different models are all very small and negligible, but among all, random forest shows the best performance. Throughout 8 different variables for detecting Pulsar, the most important variable is "X3", which is the excess kurtosis of the integrated profile. This variable can be used as the only predictors to predict whether the detected star is Pulsar, since the importance of this variable is prominently high comparing to rest of the predictors. The result of second important variable is different when using different method of classification. For random forest, the second important variable is shown as "X2", which is the standard deviation of the

integrated profile. When using the boosting method, the second most important variable is then “X6”, which is the Standard deviation of the DM-SNR curve. We can suggest to scientists to use the excess kurtosis of the integrated profile as the main source of classifying Pulsar.