IE 583 HOMEWORK – 3

DEEPAK-GEORGE THOMAS

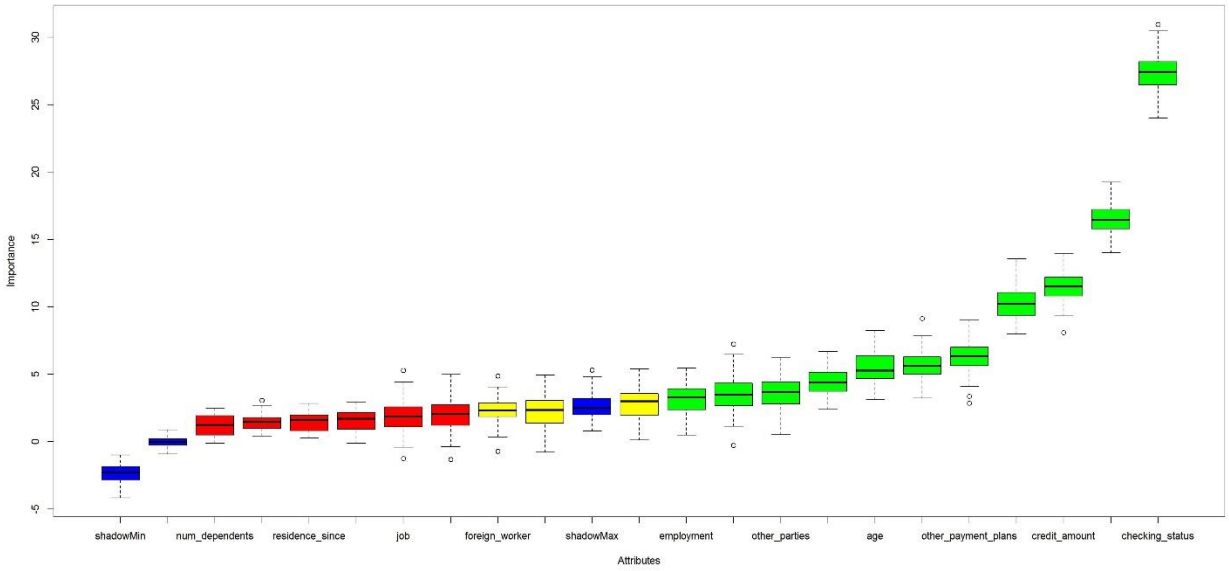
976938993

M.S(MECHANICAL ENGINEERING)

1. *Attribute Selection*

In order to perform attribute selection, the package Boruta was implemented. This package is used to determine all pertinent attributes in a dataset. This algorithm behaves as a wrapper which wraps around an algorithm called Random Forest (*Kursa et al.,* Feature Selection with the Boruta Package).

The results for the credit dataset are as follows –



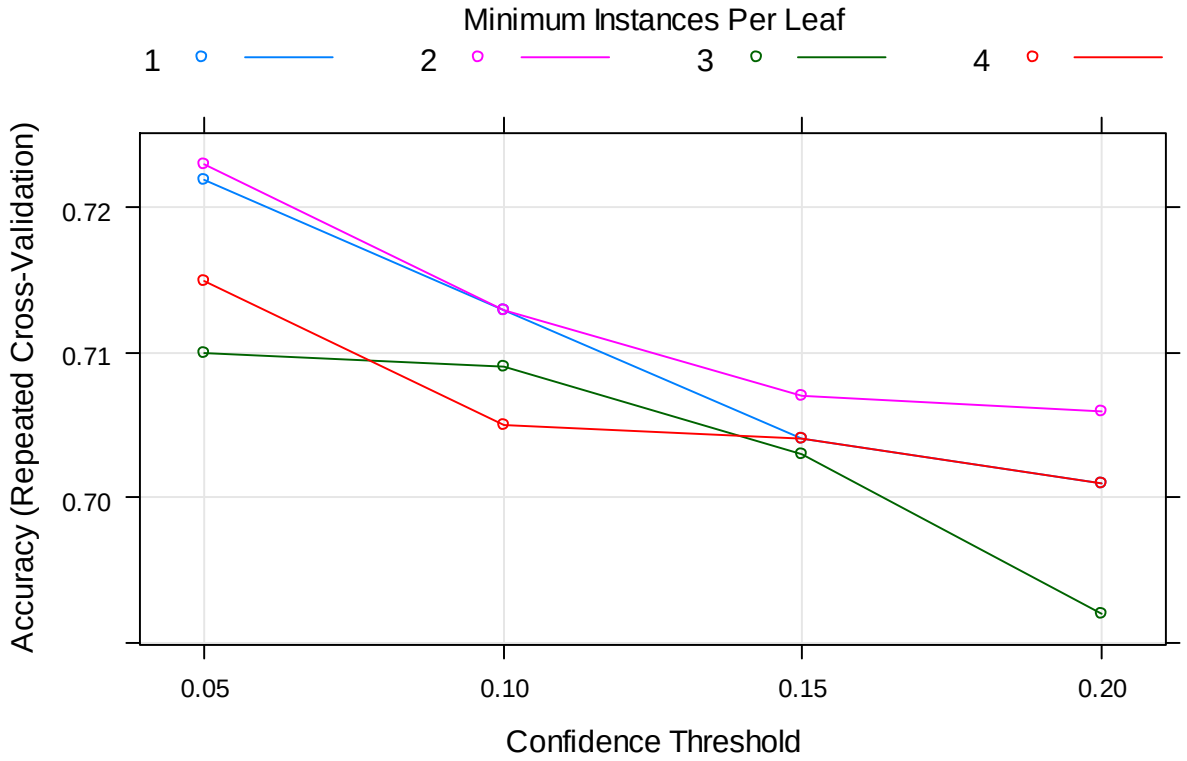
The good attributes came out to be employment, other parties, age, other payment plans, credit amount and checking status, installment commitment, employment, purpose, age, credit history and duration. The bad attributes were number of existing_credits, job, num_dependents, own_telephone, personal_status and date of residence. The tentative attributes came out to be foreign_worker, housing, property_magnitude.

With the absence of Baruto, the maximum accuracy came out be 0.72. After feature selection using Baruto, the maximum accutacy obtained was 0.718, which was very surprising as it should have been higher as compared to other earlier model. But after taking into account the tentative models, we got an increased accuracy of 0.745.

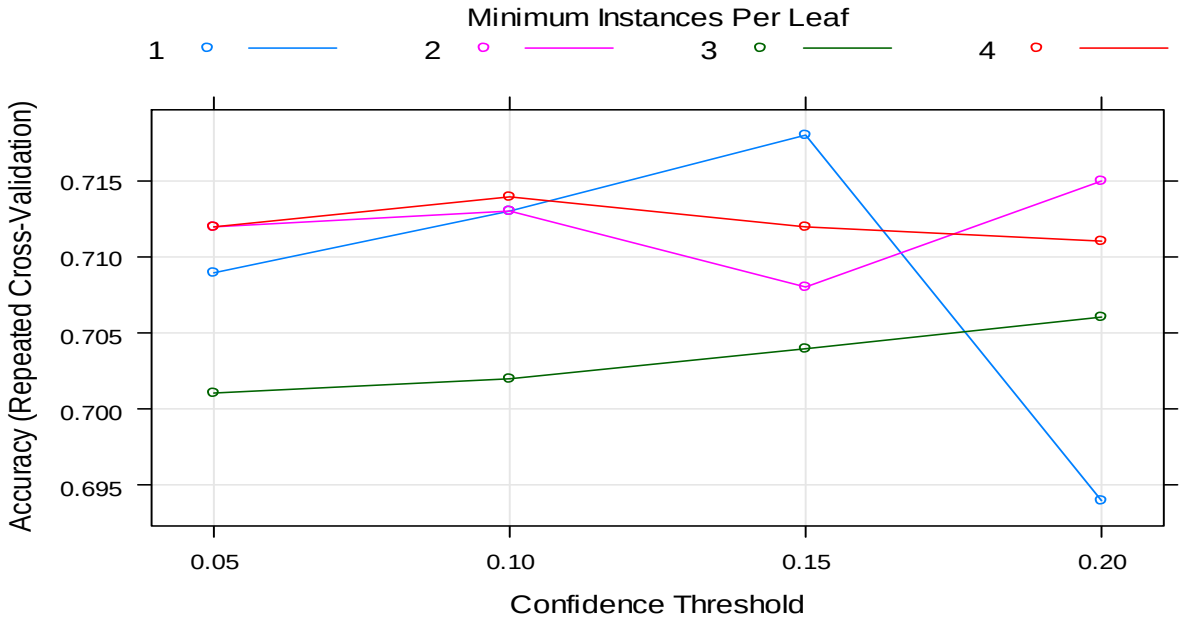
2. *Attribute Construction*

This feature was implemented by making 2 changes –

1) The attributes termed important by Boruta were classified as factor variable where required. The highest accuracy (0.722) was obtained with a confidence of 0.05 and 2 instances per leaf.



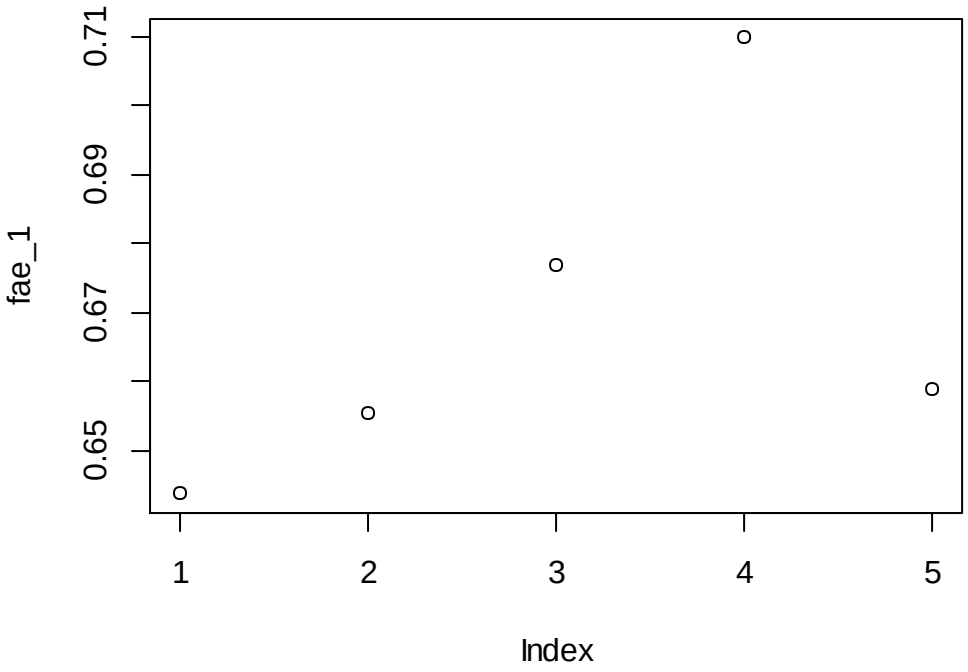
2) A new attribute which was made by combining credit history and credit amount. Attributes with a history of more than 6 months and a credit amount of less than 500 was classified as GOOD.



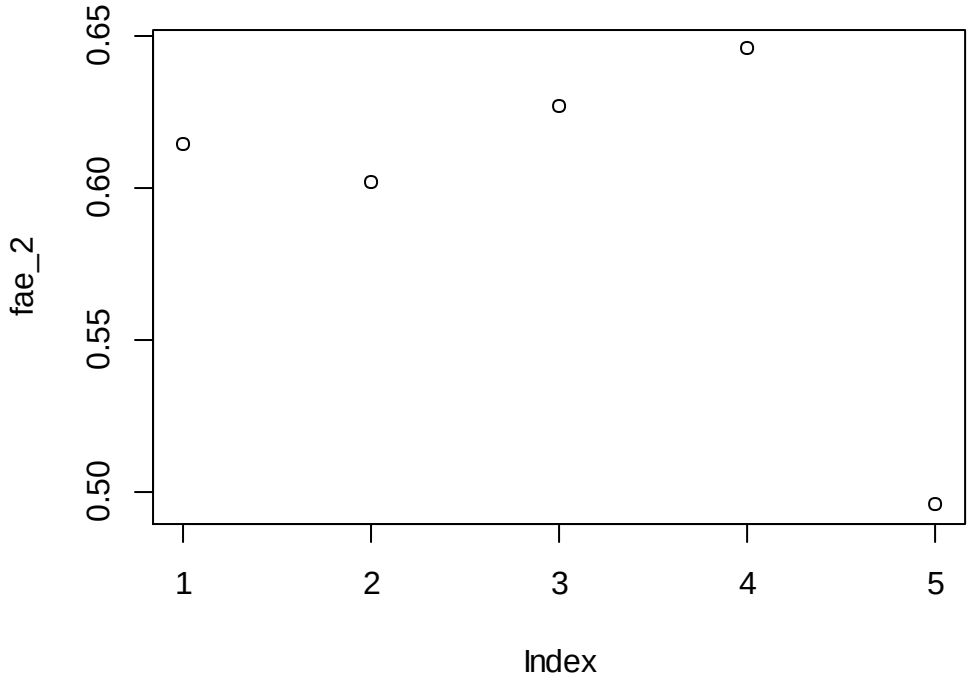

```
The highest accuracy was obtained with a confidence threshold of 0.15 a
nd one number of leaf.
```

3) An accuracy of .722 was obtained after implementing change 1 and 0.718 after implementing the second change.

3. *Biased random sampling*
   a. Currently, we have 300 bad instances and 700 good instances. This leads to data imbalance. In order to mitigate the issues due to this form of imbalance, we artificially balance the data using upSampling and down Sampling.
   b. However, the accuracy for this method cannot be determined the way it usually is done using CV. This is because the test data should never be artificially resampled (to prevent overfitting). But when we perform CV on our training data, all the folds have resampled data. Therefore, in this case we manually create the folds and then train the data.
   c. In the case of upSampling, the bad instances are increased to 700 from 300. This gives a ratio of 700:700 for bad and good instances. In this case, we obtain a mean accuracy of 66.8%.

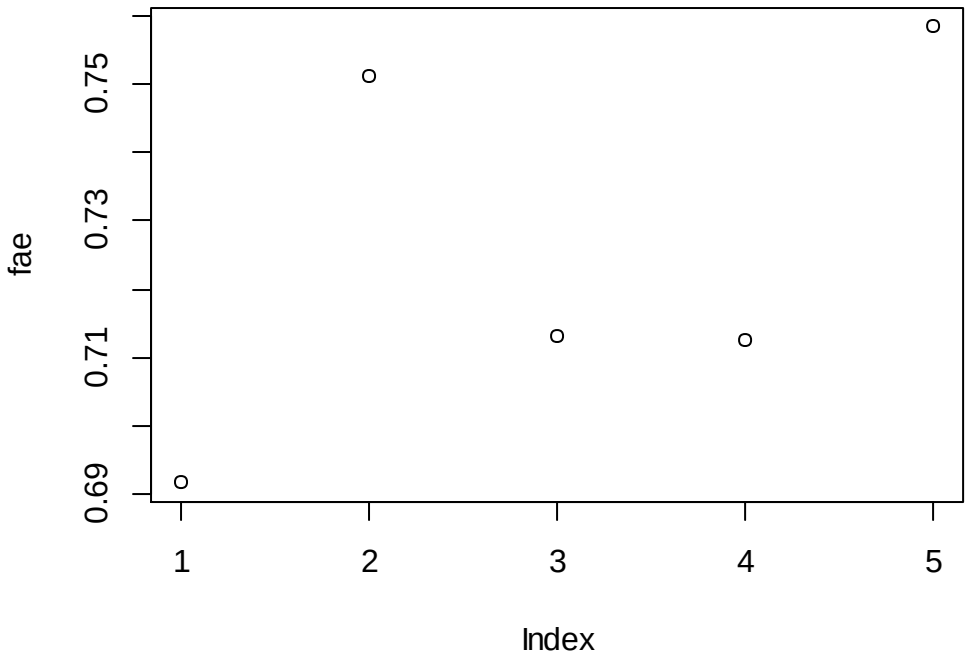


d. In the case of downSampling, the good instances are reduced to 300. This increases the proportion of bad instances and the final data set contains 300 good and 300 bad instances. The mean accuracy comes to be 59.7%.



4. SMOTE

a. Herein, synthetic instances are generated using a nearest neighbor approach as opposed to resampling.
b. In this artificially generated dataset, the ratio between good and bad classes are 900:1200.
c. As in the above case, the usual method of CV cannot be implemented here as the test fold also will contain artificially rebalanced data. Therefore, folds are manually generated.
d. SMOTE gives a mean accuracy of .723. As expected, this methods outperforms biased resampling techniques like upSampling and downSampling.


fae
0.69
0.71
0.73
0.75
1
2
3
4
5
Index