CLASSIFICATION ALGORITHMS(Problem 2)

Nitin K (2017csb1093)

30.10.2020

CSE FINAL YEAR, UG

IIT ROPAR

INTRODUCTION

This a report for the Decision Tree algorithm. We are going to analyse how the algorithm performs on the given dataset(Adult Dataset).

Algorithms Details:

I have Taken the Algorithm from the books. That is why I am providing the reference only as the algorithm has already been discussed in the class.

Decision Tree

Reference book for the Algorithm is **Data Mining and Analysis: Fundamental Concepts and Algorithms, Mohammed J. Zaki and Wagner Meira Jr.**

DATA

DataSet Descriptions

Here We have Provided DataSet:

- **Adult DataSet (https://www.kaggle.com/qizarafzaal/adult-dataset)**

```
                 39         77516  ...            0            40
count  32560.000000  3.256000e+04  ...  32560.000000  32560.000000
mean      38.581634  1.897818e+05  ...     87.306511     40.437469
std       13.640642  1.055498e+05  ...    402.966116     12.347618
min       17.000000  1.228500e+04  ...      0.000000      1.000000
25%       28.000000  1.178315e+05  ...      0.000000     40.000000
50%       37.000000  1.783630e+05  ...      0.000000     40.000000
75%       48.000000  2.370545e+05  ...      0.000000     45.000000
max       90.000000  1.484705e+06  ...   4356.000000     99.000000
```

RESULTS

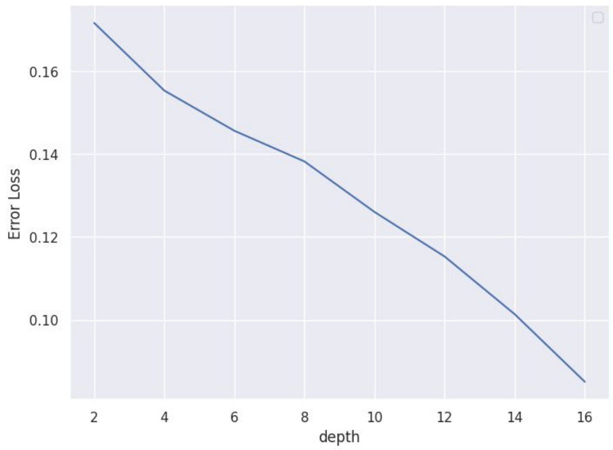


Fig:1.1(Train)

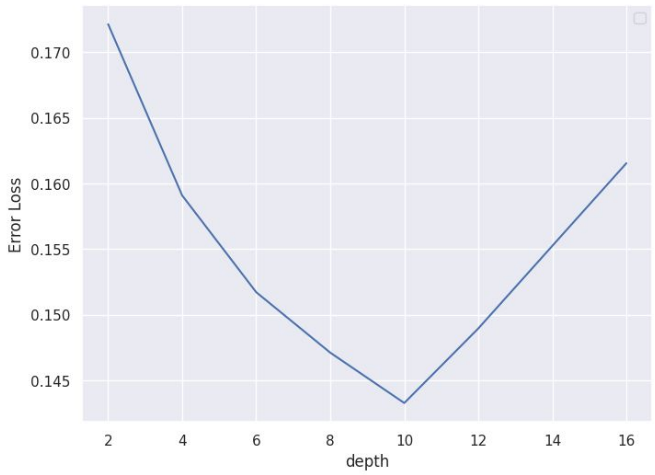
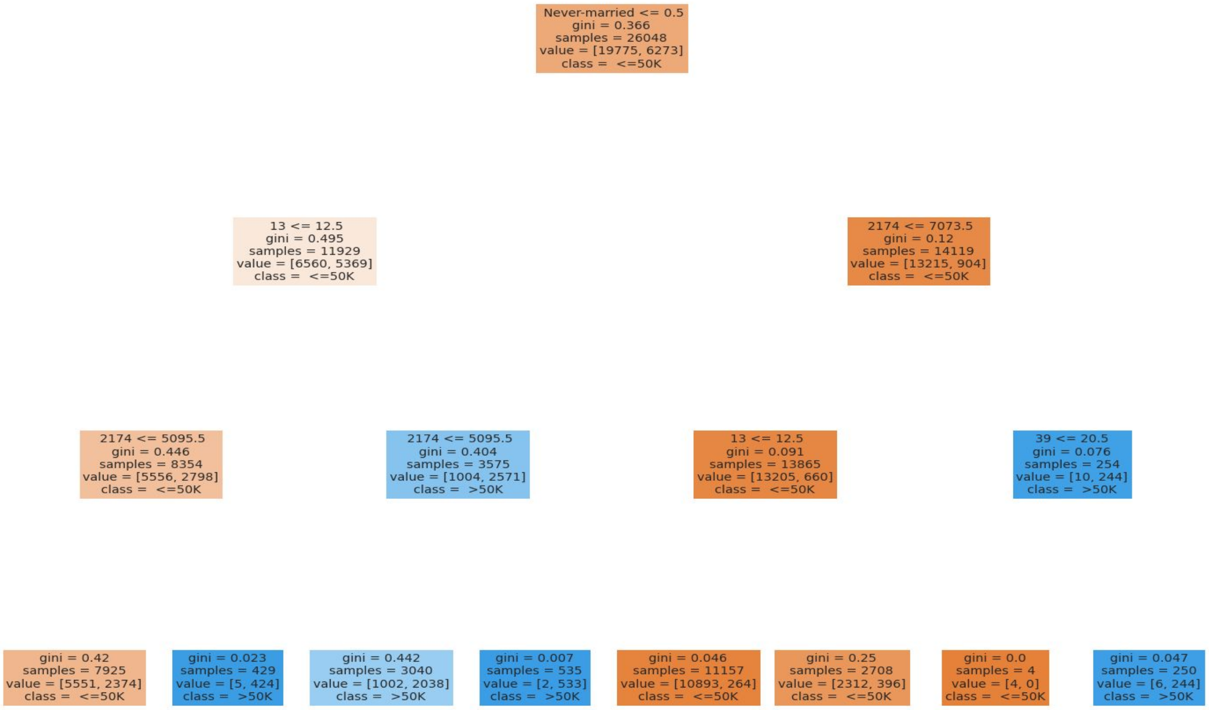


Fig:1.2(Test)

Fig:1.1 is the loss on training(80%) dataset whereas **Fig:1.2** is the loss on test(20%) dataset. From both the figures we can see that before depth = 10 both train and test loss are decreasing but after **depth**=**10** the loss for the test dataset started to increase. It means that after **depth** = **10** the Decision tree model started to overfit on train data as it is continuous decreasing after **depth** = **10**. Hence **depth** = **10** is the optimal decision tree to get the best result for the provided dataset.

Below I am providing a decision tree but this is not optimal this is only for view reference for **depth** = **3**.



Decision tree for depth = **3**

I am not providing any decision tree diagram for depth = 10 because the nodes of the tree would be very unclear in the image.

CONCLUSION

Usually the dataset has a large number of features which result in a large number of splits which may give complex tree and it may also lead to overfit the data as we see in our dataset that after depth = 10 the decision tree started to overfit on the train dataset.So it is really necessary to know when we are going to stop and this can be done by using pruning method which removes the branches that make use of low importance features. It reduces the complexity of the tree.

As our data contains both numerical and categorical value and the decision tree is performing really better on such attributes dataset with **85.7%** accuracy. So from this we can say that it handles such a dataset very well.