**Home Work 2 – Problem (Selected Seed# 28)**

For this homework you will again use the SERU data from the first assignment, this time to build and evaluate some simple predictive models. Begin by describing any data cleaning or transformation steps that you performed.

1. Find the "right" sized decision tree for modeling this dataset. Start by building a tree with no pruning and examine the complexity at various splits. Explain how you choose the best complexity level, and then build the resulting tree. Plot the tree that results from using the best complexity settings. Explain the result. Which features and values are important to the prediction? Do you think these make sense?
2. Now build and examine a logistic regression model. Discuss the coefficients that appear to significantly affect the predicted outcome. Do the magnitudes and signs of these coefficients make sense to you? Why or why not?
3. Finally, fairly evaluate the quality of the predictive results for the two models. Provide the training, validation, and testing set confusion (error) matrices for both. Which is better? Why do you think that is? Are the error rates consistent for the two values of the target? Why?

**Data Transformations**

Converted the missing values for below categorical variables to a ‘N’.

IS\_PARENT\_HIGHER\_ED\_GRAD

AFRICANAMERICAN

AMERINDIAN

ASIAN

DECLINETOSTATE

HISPANIC

NONRESIDENT

PACIFICISLANDER

WHITE

Converted the below attributes to categorical and missing values replaced by 0.

INVITED\_PREVIOUS

Missing values for the below attributes were replaced with the mean/ median value of the attribute.

SP16\_TERM\_UI\_GRADED.\_HRS

SP16\_TERM\_UI\_TOTAL\_HOURS

TOTAL\_UNITS

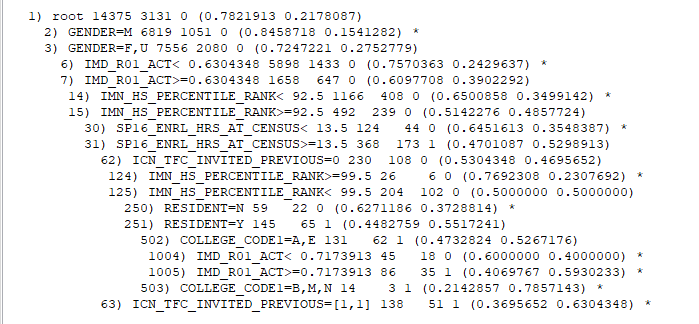
HS\_PERCENTILE\_RANK

Semester.Hours (median)

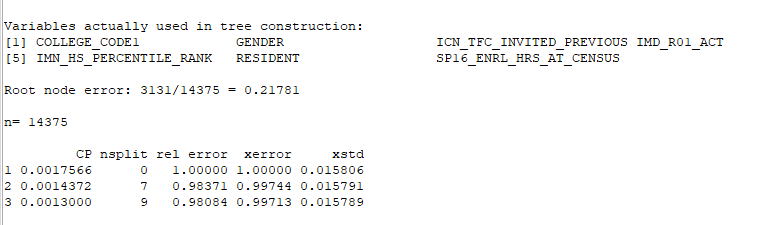
0-1 scaling was done for the ACT score attribute and missing values were replaced with its median. SAT score was ignored since it had several missing values.

**Decision Tree**

After the data transformations, a decision tree was generated with maximum depth. Examining the minimum xerror, the tree was re-generated with a complexity little higher than that of minimum xerror. The corresponding tree is given below:



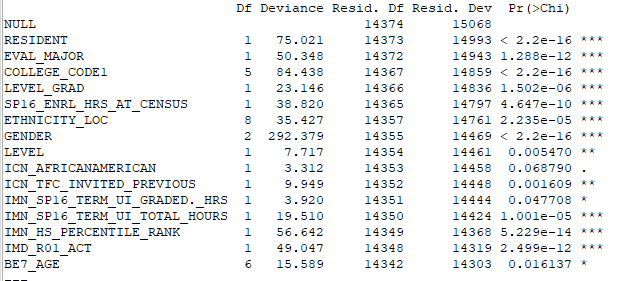
One thing that I noticed was that some of the features got repeated multiple times in the tree. For example, IMD\_R01\_ACT and IMN\_HS\_PERCENTILE\_RANK got repeated twice. The significant variables for this tree are GENDER, RESIDENT, IMD\_R01\_ACT, IMN\_HS\_PERCENTILE\_RANK, SP16\_ENRL\_HRS\_AT\_CENSUS, ICN\_TFC\_INVITED\_PREVIOUS and COLLEGE\_CODE1. The tree has 10 leaf nodes, in which none of them are pure nodes, since we pruned the tree for getting a better and less complex model.



Based on the wrong predictions from the tree, the error for this model was found to be 21.78%

**Linear Logistic Regression**

The significane of the attributes are given below. The \*\*\* ones are highly significant and how it’s significance varies is based on the corresponding attribute coefficient value from the next table.



Here RESIDENT is a significant attribute. Since this has a positive coefficient value for ‘Y’, it means that the student who is a resident is more probable to respond to the survey. In similar fashon, we can define the significance as mentioned in the examples below:

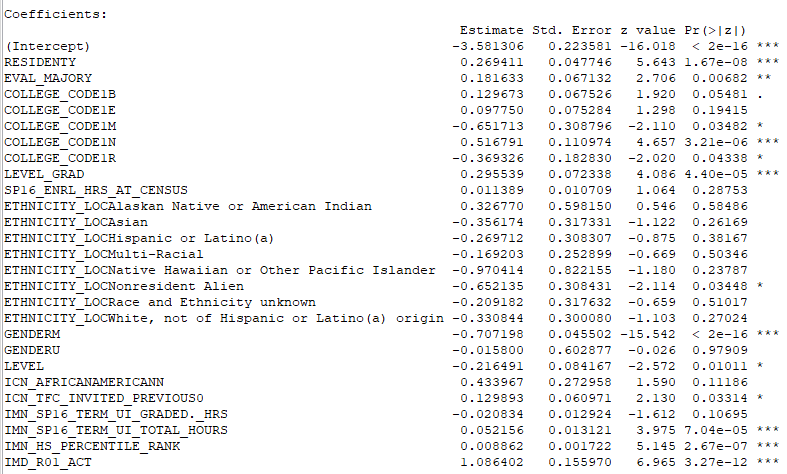
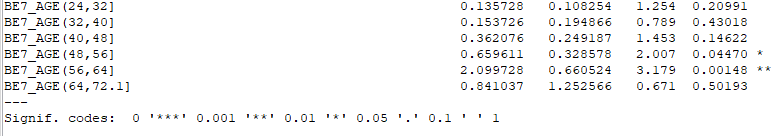
EVAL\_MAJOR – Students who got the additional evaluation questions are more probable to respond to the survey.

COLLEGE\_CODE1 – Students who are with ‘College of Nursing’ are more probable to respond, while the students in ‘Carver College of Medicine’ is less likely to respond to the survey

LEVEL\_GRAD – More the senior, more likely to respond to the survey.

GENDER – Male students are less likely to respond to the survey.

HS\_PERCENTILE\_RANK – Students who have higer rank in HS are more likely to respond.

**Evaluation**

Based on the evaluation of the Decision Tree model given below, few of the relevent attributes can be calculated as shown here:

Accuracy = (2414+20)/ (2414+20+28+620) = 78.97%

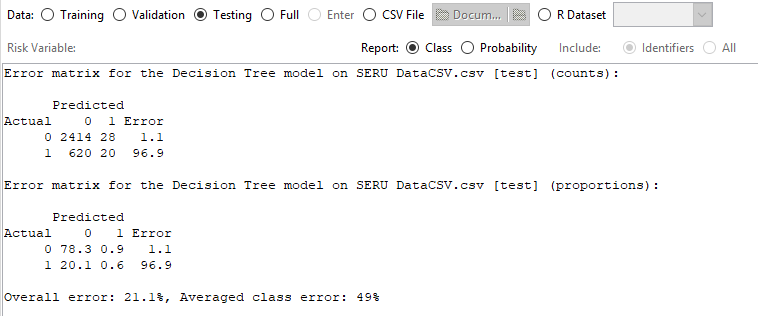
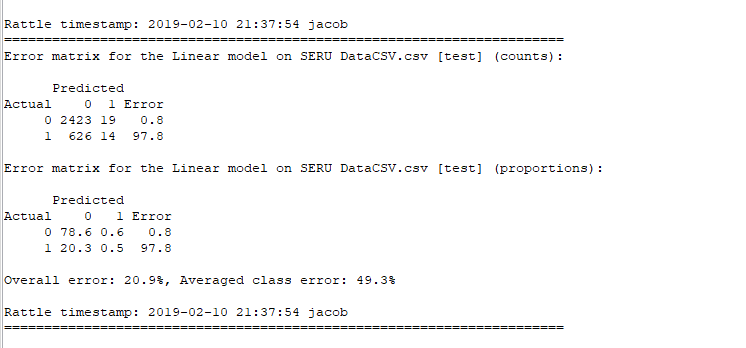
Error = 21.03%

Sensitivity or Recall = 20/ 640 = 3.13%

Specificity = 2414/ 2442 = 98.85%

PPV or Precision = 20/ 48 = 41.67%

NPV = 2414/ 3034 = 79.56%

Based on the error comparisons with both the models, it looks like the logistic regression model did a better job with lesser overall error. The model was applied on unseen data with the ‘TEST’ partition for getting the above results.

Thank you!