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:

```
1) root 14375 3131 0 (0.7821913 0.2178087)
  2) GENDER=M 6819 1051 0 (0.8458718 0.1541282) *
  3) GENDER=F,U 7556 2080 0 (0.7247221 0.2752779)
    6) IMD_R01_ACT< 0.6304348 5898 1433 0 (0.7570363 0.2429637) *
    7) IMD_R01_ACT>=0.6304348 1658  647 0 (0.6097708 0.3902292)
     14) IMN_HS_PERCENTILE_RANK< 92.5 1166  408 0 (0.6500858 0.3499142) *
     15) IMN_HS_PERCENTILE_RANK>=92.5 492  239 0 (0.5142276 0.4857724)
       30) SP16_ENRL_HRS_AT_CENSUS< 13.5 124   44 0 (0.6451613 0.3548387) *
       31) SP16_ENRL_HRS_AT_CENSUS>=13.5 368  173 1 (0.4701087 0.5298913)
         62) ICN_TFC_INVITED_PREVIOUS=0 230  108 0 (0.5304348 0.4695652)
          124) IMN_HS_PERCENTILE_RANK>=99.5 26    6 0 (0.7692308 0.2307692) *
          125) IMN_HS_PERCENTILE_RANK< 99.5 204  102 0 (0.5000000 0.5000000)
            250) RESIDENT=N 59   22 0 (0.6271186 0.3728814) *
            251) RESIDENT=Y 145   65 1 (0.4482759 0.5517241)
              502) COLLEGE_CODE1=A,E 131   62 1 (0.4732824 0.5267176)
               1004) IMD_R01_ACT< 0.7173913 45   18 0 (0.6000000 0.4000000) *
               1005) IMD_R01_ACT>=0.7173913 86   35 1 (0.4069767 0.5930233) *
              503) COLLEGE_CODE1=B,M,N 14    3 1 (0.2142857 0.7857143) *
         63) ICN_TFC_INVITED_PREVIOUS=[1,1] 138   51 1 (0.3695652 0.6304348) *
```

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.

```
Variables actually used in tree construction:
[1] COLLEGE_CODE1            GENDER                   ICN_TFC_INVITED_PREVIOUS IMD_R01_ACT
[5] IMN_HS_PERCENTILE_RANK   RESIDENT                 SP16_ENRL_HRS_AT_CENSUS

Root node error: 3131/14375 = 0.21781

n= 14375

         CP nsplit rel error  xerror     xstd
1 0.0017566      0   1.00000 1.00000 0.015806
2 0.0014372      7   0.98371 0.99744 0.015791
3 0.0013000      9   0.98084 0.99713 0.015789
```

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.

```
                          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                     14374      15068
RESIDENT                   1   75.021     14373      14993 < 2.2e-16 ***
EVAL_MAJOR                 1   50.348     14372      14943 1.288e-12 ***
COLLEGE_CODE1              5   84.438     14367      14859 < 2.2e-16 ***
LEVEL_GRAD                 1   23.146     14366      14836 1.502e-06 ***
SP16_ENRL_HRS_AT_CENSUS    1   38.820     14365      14797 4.647e-10 ***
ETHNICITY_LOC              8   35.427     14357      14761 2.235e-05 ***
GENDER                     2  292.379     14355      14469 < 2.2e-16 ***
LEVEL                      1    7.717     14354      14461  0.005470 **
ICN_AFRICANAMERICAN        1    3.312     14353      14458  0.068790 .
ICN_TFC_INVITED_PREVIOUS   1    9.949     14352      14448  0.001609 **
IMN_SP16_TERM_UI_GRADED._HRS 1  3.920     14351      14444  0.047708 *
IMN_SP16_TERM_UI_TOTAL_HOURS 1 19.510     14350      14424 1.001e-05 ***
IMN_HS_PERCENTILE_RANK     1   56.642     14349      14368 5.229e-14 ***
IMD_R01_ACT                1   49.047     14348      14319 2.499e-12 ***
BE7_AGE                    6   15.589     14342      14303  0.016137 *
---
```

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.

```
Coefficients:
                                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                                       -3.581306   0.223581 -16.018  < 2e-16 ***
RESIDENTY                                          0.269411   0.047746   5.643 1.67e-08 ***
EVAL_MAJORY                                        0.181633   0.067132   2.706  0.00682 **
COLLEGE_CODE1B                                     0.129673   0.067526   1.920  0.05481 .
COLLEGE_CODE1E                                     0.097750   0.075284   1.298  0.19415
COLLEGE_CODE1M                                    -0.651713   0.308796  -2.110  0.03482 *
COLLEGE_CODE1N                                     0.516791   0.110974   4.657 3.21e-06 ***
COLLEGE_CODE1R                                    -0.369326   0.182830  -2.020  0.04338 *
LEVEL_GRAD                                         0.295539   0.072338   4.086 4.40e-05 ***
SP16_ENRL_HRS_AT_CENSUS                            0.011389   0.010709   1.064  0.28753
ETHNICITY_LOCAlaskan Native or American Indian     0.326770   0.598150   0.546  0.58486
ETHNICITY_LOCAsian                                -0.356174   0.317331  -1.122  0.26169
ETHNICITY_LOCHispanic or Latino(a)                -0.269712   0.308307  -0.875  0.38167
ETHNICITY_LOCMulti-Racial                         -0.169203   0.252899  -0.669  0.50346
ETHNICITY_LOCNative Hawaiian or Other Pacific Islander -0.970414   0.822155  -1.180  0.23787
ETHNICITY_LOCNonresident Alien                    -0.652135   0.308431  -2.114  0.03448 *
ETHNICITY_LOCRace and Ethnicity unknown           -0.209182   0.317632  -0.659  0.51017
ETHNICITY_LOCWhite, not of Hispanic or Latino(a) origin -0.330844   0.300080  -1.103  0.27024
GENDERM                                           -0.707198   0.045502 -15.542  < 2e-16 ***
GENDERU                                           -0.015800   0.602877  -0.026  0.97909
LEVEL                                             -0.216491   0.084167  -2.572  0.01011 *
ICN_AFRICANAMERICANN                               0.433967   0.272958   1.590  0.11186
ICN_TFC_INVITED_PREVIOUS0                          0.129893   0.060971   2.130  0.03314 *
IMN_SP16_TERM_UI_GRADED._HRS                      -0.020834   0.012924  -1.612  0.10695
IMN_SP16_TERM_UI_TOTAL_HOURS                       0.052156   0.013121   3.975 7.04e-05 ***
IMN_HS_PERCENTILE_RANK                             0.008862   0.001722   5.145 2.67e-07 ***
IMD_R01_ACT                                        1.086402   0.155970   6.965 3.27e-12 ***
BE7_AGE(24,32]                                     0.135728   0.108254   1.254  0.20991
BE7_AGE(32,40]                                     0.153726   0.194866   0.789  0.43018
BE7_AGE(40,48]                                     0.362076   0.249187   1.453  0.14622
BE7_AGE(48,56]                                     0.659611   0.328578   2.007  0.04470 *
BE7_AGE(56,64]                                     2.099728   0.660524   3.179  0.00148 **
BE7_AGE(64,72.1]                                   0.841037   1.252566   0.671  0.50193
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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%

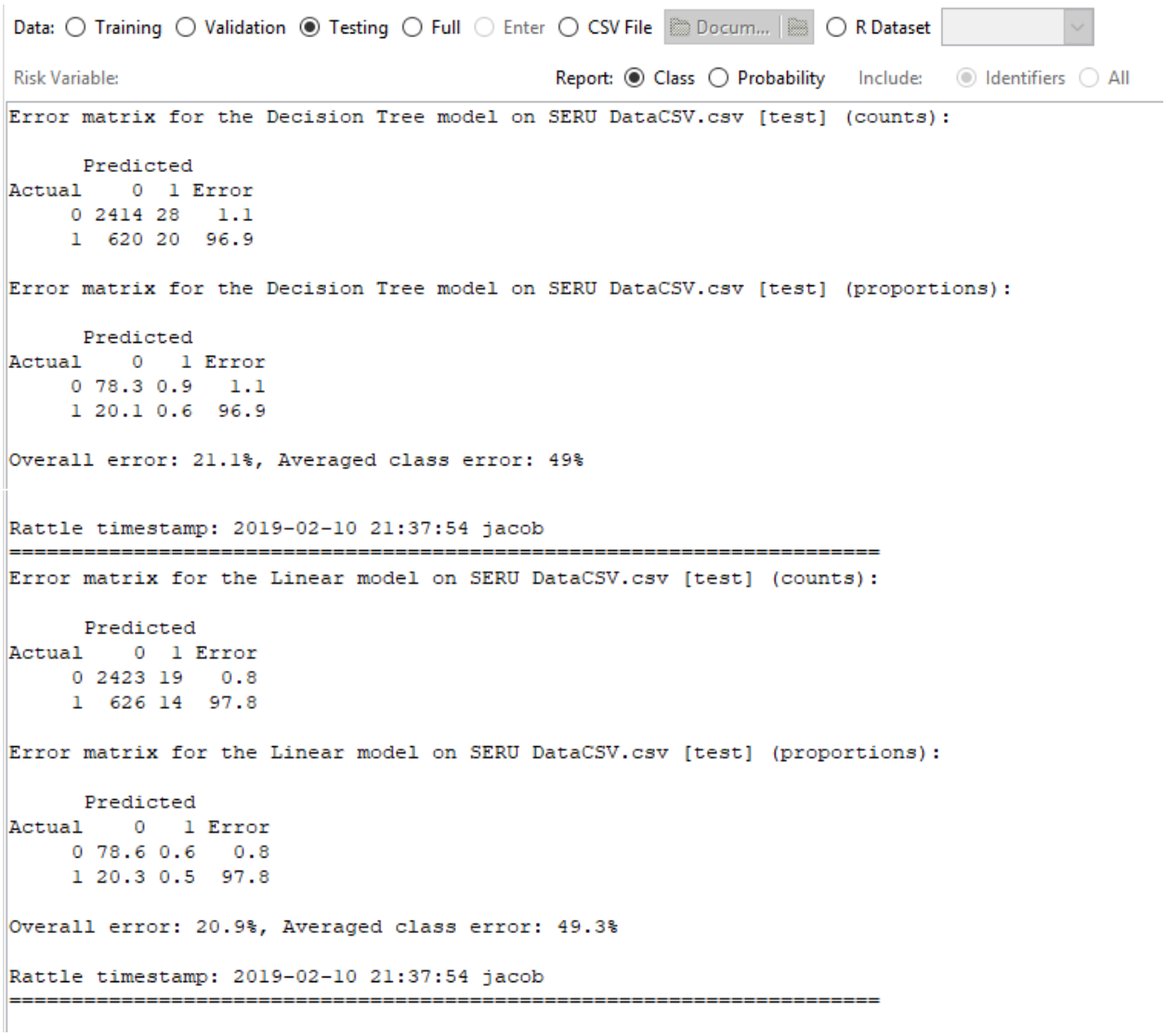
Error = 21.03%

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!