Final Project: Analysing and Visualizing CollegeScores2yr Dataset in RStudio

George Mason University
STAT-515-006 | Prof. Richard Sigman

John Victor Kanaparthy (G01283591)
Sahith Reddy Battapuram (G01337411)
Venkata Abhilaash Annamreddi (G01336658)
George Mason University
Fairfax, Virgina
kvictor2@gmu.edu
sbattapu@gmu.edu
vannamre@gmu.edu

Abstract

Schools invest a lot of time and effort to make sure their students have the best environment to grow. They maintain the data of the students to analyze few parameters such as scores, fee, their origin, etc. In this project, we chose to work on a dataset about US colleges and universities. Using this dataset, we analyzed various parameters and chose to focus on the Completion Rate and Median Debt of students who complete the program. We opted to go with Classification Tree, Random Forests and Multiple Linear Regression to build predictive models. By using these models, we tried to understand what factors in general would affect the college graduation rate of the students and to understand what the financial condition of the students is.

Keywords – Schools, parameters, Completion Rate, Median Debt, Random Forests, Multiple Linear Regression, predictive models

I. PROJECT DESCRIPTION

Every year, after completing school, Americans opt to do Higher Education in various colleges and universities. They take various factors into consideration while choosing the university, which influence their decision like the amount to spend on their education, how far should be the university from home, etc. Also, after joining few students change their minds and transfer from one college to another and few drop from their high school and start a career.

Considering all these factors, the Administration has launched a new College Scorecard to give credible and unbiased information about college performance to address the lack of information regarding college quality and expenses. By observing this information, students and their parents can correctly decide their choice of preference to select the university.

This research aims to examine the data, and how it can be utilized to assess an institution's impact on a subset of performance indicators. To create predictive models, we used Classification Tree, Random Forests and Multiple Linear Regression. We tried to identify what elements in general might affect a student's college graduation rate, as well as what their financial situation is, by applying these models.

II. DATASET

Dataset Name: CollegeScores2yr
This dataset has 1141 observations and 37 variables

Source: This is the built in R dataset in the Lock5Data package.

Description: The dataset chosen for the project is CollegeScores2yr which is the subset of the variables in the full College Scorecard and contains only the schools that primarily grant associate degree i.e., (MainDegree = 2).

The types of variables in the dataset are:

- UNITID (Nominal Data) – School ID
- INSTNM (Nominal Data) – Name of the School
- CITY (Nominal Data) – Location of the School
- ACCREDAGENCY (Nominal Data) – Accreditation Agency
- MAIN (Ordinal Data) – Flag for main campus
- PREDDEG (Ordinal Data) – Predominant undergrad degree
- HIGHDEG (Ordinal Data) – Highest Degree
- CONTROL (Ordinal Data) – School is controlled by (Private, Profit, Public)
- REGION (Nominal Data) – Region of country (Midwest, Northeast, Southeast, Territory, West)
- LOCALE (Nominal Data) – Locale (City, Rural, Suburb, Town)
- LATITUDE (Interval) – Latitude
- LONGITUDE (Interval) – Longitude
- ADM_RATE (Ratio) – Admission Rate
- ACTCMMID (Ratio) – Median of ACT Scores
- ACTENMID (Ratio) – Midpoint of the ACT English score
- ACTMTMID (Ratio) – Midpoint of the ACT math score
- ACTWRMID (Ratio) – Midpoint of the ACT writing score
- SAT_AVG (Ratio) – Average combined SAT scores
- DISTANCEONLY (Nominal Data) – Only online (distance) programs

- UGDS (Ratio) – Enrollment of undergraduate certificate/degree – seeking students
- UGDS_WHITE (Ratio) – Total share of enrollment of undergraduate degree – seeking students who are white
- UGDS_BLACK (Ratio) – Total share of enrollment of undergraduate degree – seeking students who are black
- UGDS_HISP (Ratio) – Total share of enrollment of undergraduate degree – seeking students who are Hispanic
- UGDS_ASIAN (Ratio) – Total share of enrollment of undergraduate degree – seeking students who are Asian
- UGDS_OTHERS (Ratio) – Total share of enrollment of undergraduate degree – seeking students of all other categorization
- PPTUG_EF (Ratio) – Share of undergraduate, degree/certificate - seeking students who are part – time
- TUITIONFEE_IN (Ratio) – In – state tuition fee
- TUITIONFEE_OUT (Ratio) – Out – of – state tuition fee
- TUITFTE (Ratio) – Net Tuition revenue per FTE student
- INEXPFTE (Ratio) – Instructional spending per FTE student
- AVGFACSAL (Ratio) – Average monthly salary for full – time faculty
- PFTFAC (Ratio) – Full time faculty percent
- PCTPELL (Ratio) – Percent of students receiving Pell grants
- C150_4 (Ratio) – Completion Rate
- PAR_ED_PCT_1STGEN (Ratio) – First generation students' percent
- DEBT_MDN (Ratio) – Median debt for students who complete program
- FAMINC (Ratio) – Average Family Income
- MD_FAMINC (Ratio) – Median Family Income

Units of analysis:

The analysis was done on these fields:

- C150_4
- DEBT_MDN

Data Inspection and preparation of analysis dataset:

The dataset was first inspected for null values and found out that the data contained NA values. Also, there are few columns in which the data is considered sensitive which are replaced with the character privacy suppressed. All these values are removed using the function **na.omit()** in RStudio.

There is a field in the dataset with the name UGDS_OTHERS which gives the information about the total enrollment share of undergraduate degree – seeking students of all other categorization who are not White, Black, Hispanic, and Asian. We extracted the values to this field by using the excel formula:

1 – (UGDS_WHITE + UGDS_BLACK + UGDS_HISPANIC + UGDS_ASIAN).

III. EXPLORATORY ANALYSIS AND RESEARCH QUESTIONS

An EDA is a thorough examination which uncovers the underlying structure of a dataset, and it is important to do before building models because it gives us the trends, patterns, and relationships that are not readily available. Reliable conclusions can't be made by just looking the dataset. Instead, we must look at it carefully and analyze methodically through analytical lens. By doing this, we can rectify errors, delete the records which disturb the data and understand the relationships between the key attributes.

The Figure 1 represents the missing map which shows us if there are any null values in the dataset. The map shows the null values in the form of a yellow line if they are present and if we use the function: na.omit(), all the null values get eliminated and the map displays the empty black screen. Amelia library is used in generating this map. Instead of analyzing to check for the null values, it is easy and better to visualize using this map.

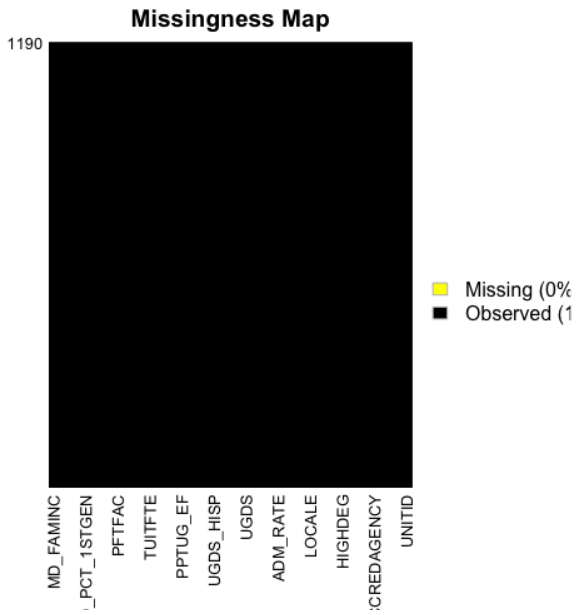

Figure 1

The Figure 2 represents the histogram, which gives the count of ADM_RATE (Admission Rate). CONTROL (School is controlled by) is used as the filled color to see which sector performed well in the admission rate. The profit variable count was very less when compared to other two categories.

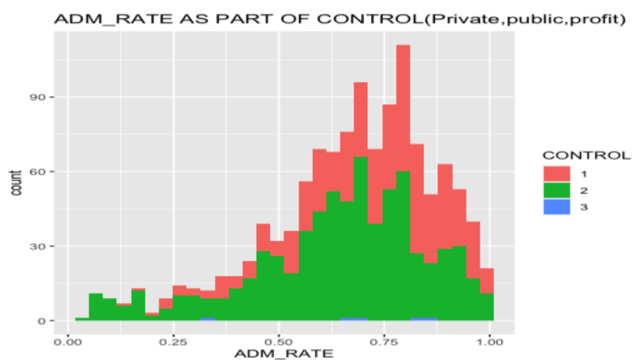

Figure 2

The Figure 3 represents the boxplot to check for outliers. It is drawn by considering the variable TUITFTE (Net Tuition Revenue per FTE Student). REGION (Region of Country) is used as a filled in color to see which region has outliers and which region is good.

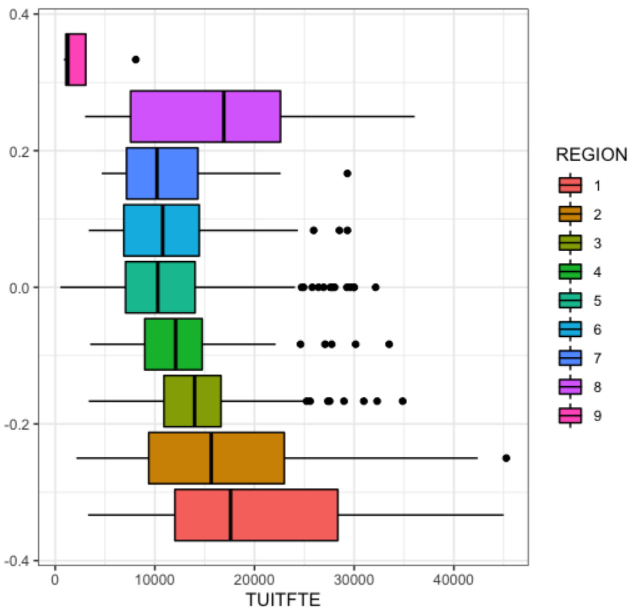

Figure 3

The Figure 4 represents the scatter plot which is drawn between two variables MD_FAMINC (Median Family Income) and DEBT_MDN (Median Debt for students who complete program). HIGHDEG (Highest Degree) is used as the filled color to observe how the two fields performed in the various levels of highest degree achieved by students.

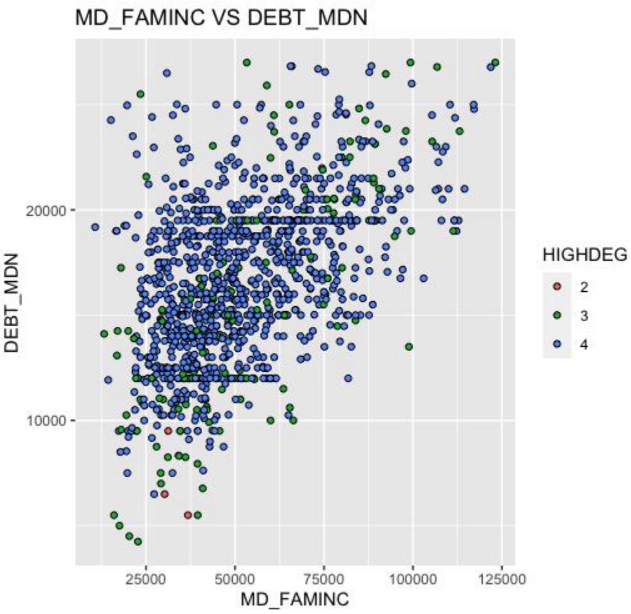

Figure 4

The Figure 5 represents the scatter plot which is drawn between two variables SAT_AVG (Average Combined SAT Scores) and TUITIONFEE_OUT (Out – of – state tuition fee). PREDDEG (Predominant undergrad degree) is used as the filled color to observe how the two fields performed in the various levels of predominant degree. From the graph, we can observe that 2 has very less values.

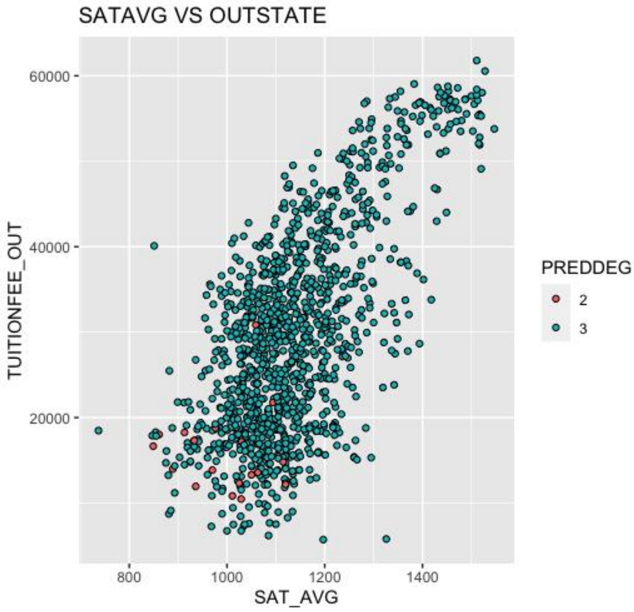

Figure 5

The Research questions which we formulated are:

- Can a model be built to predict the debt of students?
- What would be the key variables in the debt of student model
- Can a model be built to predict the completion rate (Graduation rate) of students?
- What would be the key variables in the completion rate of students

Solving these questions will help us understand what are the key components that could lead to student debt and what are the success metrics for a student to graduate. Using these metrics, we can understand what educational institutions need to focus on to help students with the career.

IV. DATA ANALYSIS

Methods and Software used:

Excel and RStudio were primarily used in this project. Excel was used to clean the dataset, i.e., to remove NA and privacy suppressed values. Whereas RStudio was used for data visualization and to build regression models.

For regression models, Classification Tree, Random Forest, and Multiple Linear Regression were used.

Results:

When working on building the models, the strongest correlation for DEBT_MDN (Median debt for students who complete program) was TUITFTE (Net Tuition Revenue per FTE student), and the highest correlation for C150_4 (Completion Rate) was SAT_AVG (Average combined SAT scores). However, varimp plot was used post the creation of Random Forest Models to get a better idea on the importance of each predictor variables.

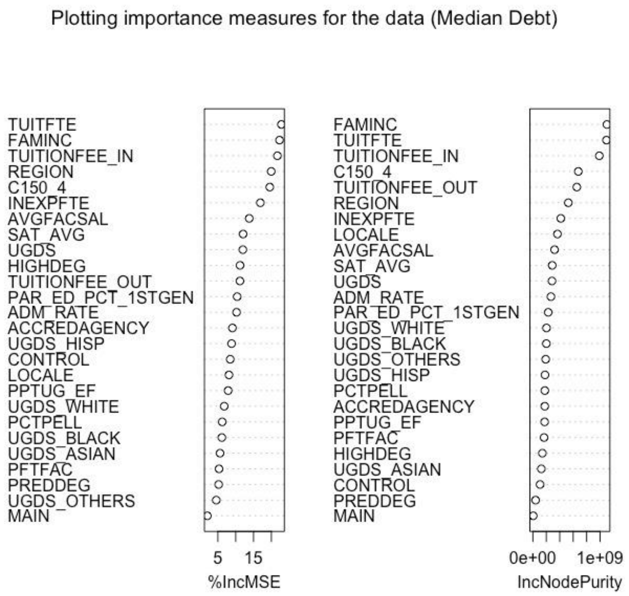

Figure 6

Figure 6 represents the varimp plot for Median Debt to plot the important measures/predictors.

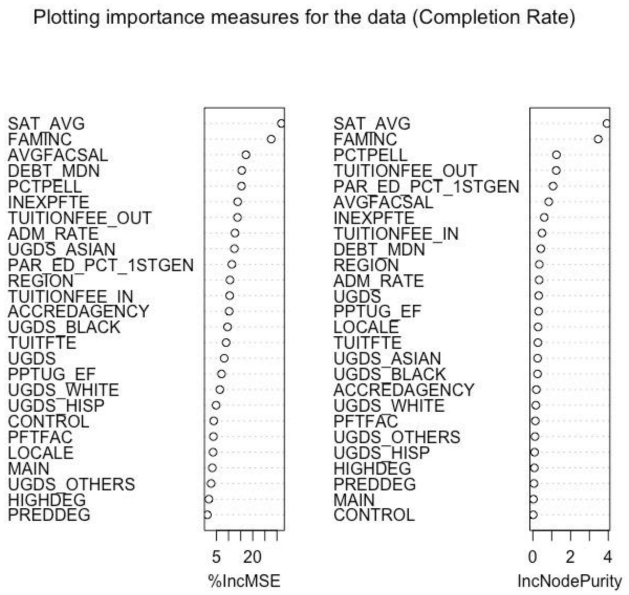

Figure 7

Figure 7 represents the varimp plot for Completion Rate to plot the important measures/predictors.

Building models for DEBT_MDN:

```
Call:
 randomForest(formula = DEBT_MDN ~ ., data = data2, importance = TRUE,      subset = train)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 8

          Mean of squared residuals: 6944314
                    % Var explained: 55.74
```

Figure 8

Figure 8 represents that Random Forest model was built using all the available predictor variables.

```
Call:
 randomForest(formula = DEBT_MDN ~ TUITFTE + FAMINC + TUITIONFEE_IN +      REGION + C150_4 + INEXPFTE + AVGFACSAL + SAT_AV
G, data = data2,      importance = TRUE, subset = train)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 2

          Mean of squared residuals: 7167432
                    % Var explained: 54.31
```

Figure 9

Figure 9 represents that another Random Forest Model was built by using top 8 most important predictors.

```
Call:
lm(formula = DEBT_MDN ~ ., data = data2)

Residuals:
    Min      1Q  Median      3Q     Max
-9681.7 -1633.6   -10.4  1490.9 10245.2

Coefficients: (2 not defined because of singularities)
                                                              Estimate
(Intercept)                                                  -1.961e+04
ACCREDAGENCYAccrediting Commission of Career Schools and Colleges  3.911e+03
ACCREDAGENCYAssociation for Bibical Higher Educaiton         -2.344e+03
ACCREDAGENCYHigher Learning Commission                       -1.749e+03
ACCREDAGENCYMiddle States Commission on Higher Education     -6.306e+03
```

Figure 10(a)

```
PCTPELL                                                      1.85e-05 ***
C150_4                                                       1.55e-12 ***
PAR_ED_PCT_1STGEN                                            0.532702
FAMINC                                                       1.38e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2713 on 1138 degrees of freedom
Multiple R-squared:  0.5606,    Adjusted R-squared:  0.5409
F-statistic: 28.47 on 51 and 1138 DF,  p-value: < 2.2e-16
```

Figure 10(b)

Figure 10(a) represents that Multiple linear regression model was built using all the variables from the dataset. The data list was too long to provide here. So, there is a limited data in the Figure 10(a)

```
Call:
lm(formula = DEBT_MDN ~ TUITFTE + FAMINC + TUITIONFEE_IN + REGION +
    C150_4 + INEXPFTE + AVGFACSAL + SAT_AVG, data = data2)

Residuals:
     Min       1Q   Median       3Q      Max
-10883.3  -1791.7     30.8   1818.0  12308.4

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.092e+04  1.207e+03  17.333  < 2e-16 ***
TUITFTE       7.629e-02  2.285e-02   3.339 0.000866 ***
FAMINC        5.730e-02  6.582e-03   8.705  < 2e-16 ***
TUITIONFEE_IN 5.751e-02  1.081e-02   5.319 1.25e-07 ***
REGION2       2.411e+02  4.052e+02   0.595 0.551925
REGION3       4.214e+02  4.276e+02   0.985 0.324589
REGION4      -8.247e+02  4.652e+02  -1.773 0.076510 .
REGION5      -6.768e+02  4.263e+02  -1.587 0.112669
REGION6      -1.439e+03  4.958e+02  -2.902 0.003780 **
REGION7      -2.094e+03  6.652e+02  -3.148 0.001683 **
REGION8      -1.139e+03  4.604e+02  -2.474 0.013520 *
REGION9      -7.179e+03  1.569e+03  -4.575 5.27e-06 ***
C150_4        7.287e+03  1.081e+03   6.742 2.44e-11 ***
INEXPFTE     -8.720e-02  1.416e-02  -6.158 1.01e-09 ***
AVGFACSAL     5.929e-02  6.482e-02   0.915 0.360504
SAT_AVG      -1.223e+01  1.574e+00  -7.772 1.67e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3002 on 1174 degrees of freedom
Multiple R-squared:  0.4449,    Adjusted R-squared:  0.4379
F-statistic: 62.74 on 15 and 1174 DF,  p-value: < 2.2e-16
```

Figure 11

Figure 11 represents that Multiple linear regression model was built using top 8 most important predictors.

```
Regression tree:
rpart(formula = DEBT_MDN ~ ., data = data2, method = "anova",
    cp = 0.001)

Variables actually used in tree construction:
[1] ADM_RATE      AVGFACSAL     C150_4        FAMINC        HIGHDEG       REGION        TUITFTE       TUITIONFEE_IN
[9] UGDS_BLACK

Root node error: 1.906e+10/1190 = 16016806

n= 1190

         CP nsplit rel error  xerror     xstd
1  0.229640      0   1.00000 1.00124 0.038884
2  0.067238      1   0.77036 0.82180 0.036012
3  0.050862      2   0.70312 0.79831 0.035355
4  0.041679      3   0.65226 0.72855 0.033812
5  0.034222      4   0.61058 0.69596 0.033459
6  0.023207      5   0.57636 0.66873 0.032008
7  0.017700      6   0.55315 0.64539 0.032202
8  0.013470      7   0.53545 0.61436 0.030563
9  0.013161      8   0.52198 0.61571 0.030697
10 0.012604      9   0.50882 0.61375 0.030687
11 0.011540     10   0.49621 0.60663 0.029912
12 0.009118     11   0.48467 0.58962 0.029546
```

Figure 12

Figure 12 represents classification tree using all the variables from the dataset

```
#Results:                                         MSE      RMSE
#RandomForest Test-Set MSE                        7313246  2704.3013885
#RandomForest Test-Set MSE(8 most-important)      8177772  2859.680
#Multiple Linear Regression                       7037914  2652.907
#Multiple Linear Regression(8 most-important)     8890161  2981.637
#Classification Tree Test-Set MSE                 9899079  3146.28
```

Figure 13

Figure 13 represents the mean square error and root mean square error for each model.

Based on the above values, Multiple linear regression turned out to be the best model with the highest accuracy.

Building models for C150_4:

```
Call:
 randomForest(formula = C150_4 ~ ., data = data2, importance = TRUE,      subset = train)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 8

          Mean of squared residuals: 0.006233175
                    % Var explained: 77.82
```

Figure 14

Figure 14 represents that Random Forest model was built using all the available predictor variables.

```
Call:
 randomForest(formula = C150_4 ~ SAT_AVG + FAMINC + AVGFACSAL +      DEBT_MDN + PCTPELL + INEXPFTE + TUITIONFEE_OUT + ADM_R
ATE,      data = data2, importance = TRUE, subset = train)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 2

          Mean of squared residuals: 0.00650161
                    % Var explained: 76.87
```

Figure 15

Figure 15 represents that another Random Forest Model was built by using top 8 most important predictors.

```
Call:
lm(formula = C150_4 ~ ., data = data2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.28571 -0.04215  0.00312  0.04037  0.45665

Coefficients: (2 not defined because of singularities)
                                                                                       Estimate
(Intercept)                                                                           -2.059e-01
ACCREDAGENCYAccrediting Commission of Career Schools and Colleges                     -1.941e-02
ACCREDAGENCYAssociation for Bibical Higher Educaiton                                   2.901e-02
ACCREDAGENCYHigher Learning Commission                                                 1.598e-01
ACCREDAGENCYMiddle States Commission on Higher Education                               2.204e-01
ACCREDAGENCYNew England Commission on Higher Education                                 1.723e-01
ACCREDAGENCYNorthwest Commission on Colleges and Universities                          1.730e-01
ACCREDAGENCYSouthern Association of Colleges and Schools Commission on Colleges        1.599e-01
ACCREDAGENCYTransnational Association of Christian Colleges and Schools                1.179e-01
ACCREDAGENCYWestern Association of Schools and Colleges Senior Colleges and University Commission  1.975e-01
MAIN1                                                                                  7.254e-02
PREDDEG3                                                                              -1.211e-01
HIGHDEG3                                                                              -7.842e-02
HIGHDEG4                                                                              -6.412e-02
CONTROL2                                                                               6.741e-02
```

Figure 16(a)

```
AVGFACSAL                                                 0.000438 ***
PFTFAC                                                    0.846521
PCTPELL                                                   0.604079
PAR_ED_PCT_1STGEN                                         0.065245 .
DEBT_MDN                                                  1.55e-12 ***
FAMINC                                                    3.99e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07371 on 1138 degrees of freedom
Multiple R-squared:  0.8246,    Adjusted R-squared:  0.8167
F-statistic: 104.9 on 51 and 1138 DF,  p-value: < 2.2e-16
```

Figure 16(b)

Figure 16(a) represents that Multiple linear regression model was built using all the variables from the dataset. The data list was too long to provide here. So, there is a limited data in the Figure 16(a)

```
Call:
lm(formula = C150_4 ~ SAT_AVG + FAMINC + AVGFACSAL + DEBT_MDN +
    PCTPELL + INEXPFTE + TUITIONFEE_OUT + ADM_RATE, data = data2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33592 -0.04895 -0.00111  0.04860  0.53696

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -4.954e-01  5.470e-02  -9.056  < 2e-16 ***
SAT_AVG         6.552e-04  4.157e-05  15.761  < 2e-16 ***
FAMINC          1.691e-06  2.228e-07   7.590 6.47e-14 ***
AVGFACSAL       1.129e-05  1.481e-06   7.626 4.95e-14 ***
DEBT_MDN        6.010e-06  7.690e-07   7.815 1.21e-14 ***
PCTPELL         7.829e-03  3.379e-02   0.232  0.81684
INEXPFTE       -1.017e-07  3.895e-07  -0.261  0.79402
TUITIONFEE_OUT  9.602e-07  3.319e-07   2.893  0.00388 **
ADM_RATE       -9.613e-03  1.501e-02  -0.640  0.52203
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08314 on 1181 degrees of freedom
Multiple R-squared:  0.7684,    Adjusted R-squared:  0.7668
F-statistic: 489.8 on 8 and 1181 DF,  p-value: < 2.2e-16
```

Figure 17

Figure 17 represents that Multiple linear regression model was built using top 8 most important predictors.

```
Regression tree:
rpart(formula = C150_4 ~ ., data = data2, method = "anova", cp = 0.001)

Variables actually used in tree construction:
[1] ACCREDAGENCY ADM_RATE     DEBT_MDN     FAMINC       INEXPFTE     PPTUG_EF     REGION       SAT_AVG
[9] UGDS

Root node error: 35.247/1190 = 0.02962

n= 1190

          CP nsplit rel error  xerror     xstd
1  0.4387488      0   1.00000 1.00227 0.037301
2  0.1064798      1   0.56125 0.58328 0.024447
3  0.0770152      2   0.45477 0.47437 0.023290
4  0.0337671      3   0.37776 0.40351 0.022057
5  0.0258632      4   0.34399 0.35662 0.019190
6  0.0177088      5   0.31813 0.34443 0.019387
7  0.0132890      6   0.30042 0.32085 0.019082
8  0.0130657      7   0.28713 0.30787 0.019383
9  0.0111897      8   0.27406 0.30434 0.019317
10 0.0108198      9   0.26287 0.29377 0.018369
11 0.0104839     10   0.25205 0.29133 0.018387
12 0.0078618     11   0.24157 0.28593 0.018786
13 0.0067473     12   0.23371 0.27333 0.018268
14 0.0066530     13   0.22696 0.26746 0.017642
15 0.0056770     15   0.21365 0.26671 0.017773
16 0.0043362     16   0.20798 0.26600 0.018077
17 0.0042216     17   0.20364 0.26058 0.018037
18 0.0041471     18   0.19942 0.26153 0.018098
19 0.0040498     19   0.19527 0.26187 0.018101
20 0.0033482     20   0.19122 0.26072 0.018122
21 0.0027178     21   0.18787 0.25136 0.017849
```

Figure 18

Figure 18 represents classification tree using all the variables from the dataset

```
#Results:                                         MSE      RMSE
#RandomForest Test-Set MSE                        0.00592  0.07699
#RandomForest Test-Set MSE(8 most-important)      0.00642  0.08015
#Multiple Linear Regression                       0.00519  0.07208
#Multiple Linear Regression(8 most-important)     0.00686  0.08282
#Classification Tree Test-Set MSE                 0.00744  0.08628
```

Figure 19

Figure 19 represents the mean square error and root mean square error for each model.

Based on the above values, Multiple linear regression turned out to be the best model with the highest accuracy.

V. CONCLUSIONS / FURTHER ANALYSIS

Conclusions:

While working on this project, we were able to explore what factors influenced a student’s graduation rate and what variables could be a probable cause for a student to go into debt.

These were the factors influencing a student’s graduation rate as per the data and analysis/models made:
Their SAT score, their Average Family income, Average monthly salary for full-time faculty (if the monthly salary of

the faculty is less there may be various reasons like faculty's experience may be less and students can't follow them), their Median debt after completing the program, and their Percent who receive pell grants.

These were the factors influencing the median debt as per the data and analysis/models made:
Their net tuition revenue, their average family income, their in-state tuition fee, their region of country, and their completion rate.

Further Analysis:

This dataset can be used for a lot more research and exploration using a variety of additional methods. A further study could be done on the effect of the faculty salary since this was one of the important variables which showed a high correlation rate with student graduation rate. To gain a better grasp of this dataset, future work could entail working with a wider range of data.

V. REFERENCES

[1] LANDER, J. A. R. E. D. P. (2021). R for everyone. ADDISON-WESLEY

[2] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: With applications in R. Springer.

[3] Data Profiles - Department of Education Open Data Platform. https://data.ed.gov/dataset/college-scorecard-all-data-files-through-6-2020/resources. Accessed 14 Nov. 2021.