Constructing a Prediction Model for College Rankings: What Determines the Ranking of an Institution?

Jake Leem and Jaechul Jung

Abstract

The purpose of this study is to create a prediction model for college ranking with variables that are perceived to be important in student admission. Different prediction models were constructed to compare their accuracy of prediction. We realized that variables SAT and ACT scores, tuition and admission all demonstrated a linear relationship with college rankings. Clustering was performed to verify if the colleges could be clustered by a certain variable and supervised learning procedures were undertaken for cross-validation and verification of the model's external validity. The results of this study shows that rankings could be most accurately predicted with the multiple-regression model that includes variables average SAT scores, tuition, and admission rate. The model demonstrated better out of sample performance upon the application of a GAM model.

Introduction

College ranking is often considered important when prospective students choose their college of interest. Often colleges with high rankings are very selective, requiring students with high SAT or ACT scores and a solid application essay. Our research group was curious on whether we could generate an estimation model that accurately predicts the contemporary college rankings with different variables that are considered important upon student enrollment such as tuition, SAT or ACT scores, admission rate, etc. In other words, what variables determine colleges to be classified as a top-tier ranking college? Also, could we accurately predict the college ranking with certain variables?

Methods

The data were taken from three different sources. Data for general college information regarding mean SAT or ACT scores and admission rates were taken from the US Department of Education college scorecard data[1]. Data for college rankings were taken from U.S. News & World Report Historical Liberal Arts College and University Rankings[2]. Data for college tuition were taken from College Affordability and Transparency List Explanation Form[3].

Upon garnering different datasets, we tidied the data into a single file under the name of the institution and removed any observations that had unspecified values (Null, ?, or blank space).

We sought if there were any pre-existing patterns within our variables prior to constructing prediction models using the pairs function. We then constructed a simple regression model and slowly added complexity by including other variables that we deem to be important in determining the ranking of the institution. The simple regression model with admission rate only was initially used to test the widely accepted notion that selectivity leads to higher ranking. Later, multi regression models were used to test if the results of the first model still persisted even with the inclusion of other predictor variables.

We then performed cross-validation to see how accurately each model performs on an independent sample by comparing its RMSE, MAE, and RSquared values. We also introduced the Generalized Additive model to add flexibility to the linear regression, as it allows for changes in the variables to correspond with non-linear changes in the outcome, thereby improving the prediction of the model.

Results

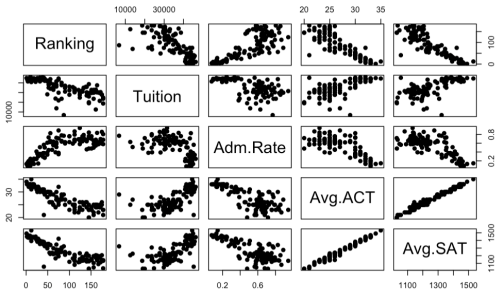


Figure 1. Finding pre-existing patterns using pairs function.

All of the variables of our choice seem to share some degree of linear relationships with our dependent variable college ranking. In particular, average ACT and SAT scores demonstrate almost perfect linearity. Upon this data we thought linear models would be sufficient enough to create the prediction model.

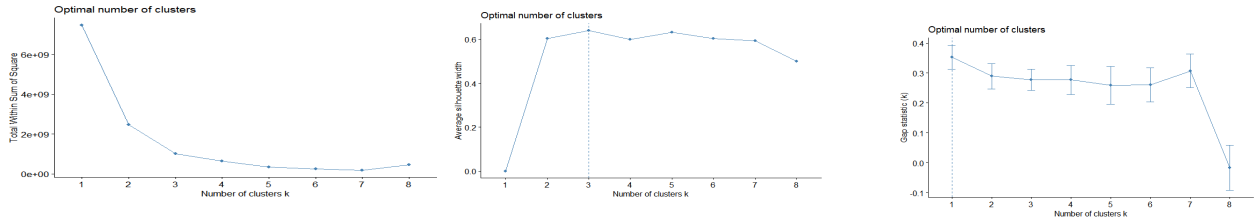
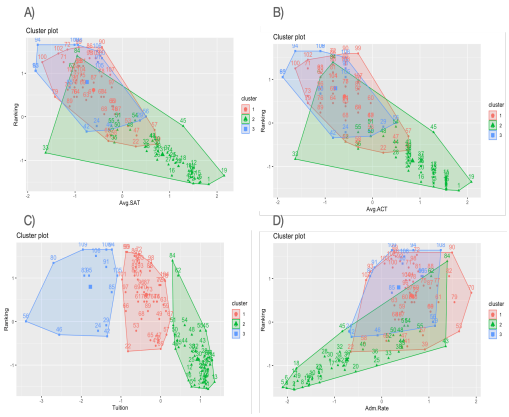


Figure 2. Finding the optimal k value for clustering. 1) Elbow, 2) Silhouette, 3) Gap Statistics

K-means algorithm is used to minimize the total withinness. None of the methods listed above are perfect, however, we came up with the ideal number of clusters to the point where one additional cluster does not contribute as much to the improvement in withinness. It appears that 3 clusters are ideal for all three methods used above.

Figure 3. Clustering Results. Cluster plot of A) Average SAT Score, B) Average ACT Score, C) Tuition, D) Admission Rate



The clustering suggests that clusters created by SAT and ACT scores don't vary; the centroids seem to be almost exactly the same. This suggests that SAT and ACT scores are valued almost as equal to one another in determining the school ranking. Clusters for tuition and admission rates demonstrate that generally schools with high tuition and low acceptance rates rank high, as schools with high rankings were clustered together. Although there might be some data points that don't necessarily reflect this trend, this suggests that variables "tuition," and "admission rates," are also effective in predicting school rankings.

```
Call:
summary.resamples(object = resamps)

Models: LM1, LM2, LM3, LM4, LM5, LMFinal
Number of resamples: 35

MAE
             Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
LM1      21.83834 26.91144 29.51845 29.50355 31.40555 38.58545    0
LM2      21.07920 28.30496 30.53095 30.97524 35.14304 41.74450    0
LM3      15.00459 18.66589 19.93664 20.11262 21.40526 27.72096    0
LM4      14.41317 17.16320 19.42725 19.46247 22.07714 26.03685    0
LM5      14.86631 18.20198 20.01353 20.37142 22.53229 25.90875    0
LMFinal 14.31805 17.00347 18.95357 19.22481 21.69245 24.87898    0

RMSE
             Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
LM1      27.76644 34.33379 37.42018 37.30859 40.15577 45.58000    0
LM2      24.90204 33.56173 36.33168 38.78226 44.90210 52.92995    0
LM3      20.81742 24.79788 26.77038 27.71049 30.57866 38.61322    0
LM4      17.82610 23.24019 26.56855 26.38206 29.47381 33.57030    0
LM5      18.73480 24.31169 26.63528 26.80444 29.68018 33.67186    0
LMFinal 17.82805 23.30077 25.98210 25.97786 29.04749 32.99411    0

Rsquared
              Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
LM1      0.3497451 0.4964930 0.5682690 0.5541180 0.6140040 0.7625582    0
LM2      0.2662357 0.3985095 0.5884609 0.5414984 0.6684061 0.8925696    0
LM3      0.5629295 0.7087326 0.7708822 0.7548426 0.8047945 0.8659959    0
LM4      0.6775119 0.7280329 0.7765620 0.7752623 0.8138241 0.8943105    0
LM5      0.6745820 0.7269613 0.7664116 0.7680192 0.8210908 0.8818806    0
LMFinal 0.6752795 0.7345763 0.7908100 0.7820381 0.8240174 0.8937054    0
```

Figure 4. Cross-validation. LM1 (Ranking ~ Admission Rate), LM2 (Ranking ~ Tuition), LM3 (Ranking ~ Average ACT + Average SAT), LM4 (Ranking ~ All Variables), LM5 (Ranking ~ Tuition + Avg.ACT + Adm.Rate), LMFinal (Ranking ~ Tuition + Avg.SAT + Adm.Rate)

After training each model using 7 repeats of 5 fold-cross validation, we noticed that a multiple regression model factoring variables Tuition, Avg.SAT, and Adm.Rate gives us the highest R-Squared value and the lowest MAE and RMSE, suggesting that LMFinal has the best out-of-sample performance. This is most likely due to the correlation between variables SAT and ACT, leading to issues of multicollinearity when both of these variables are factored upon construction of prediction models.

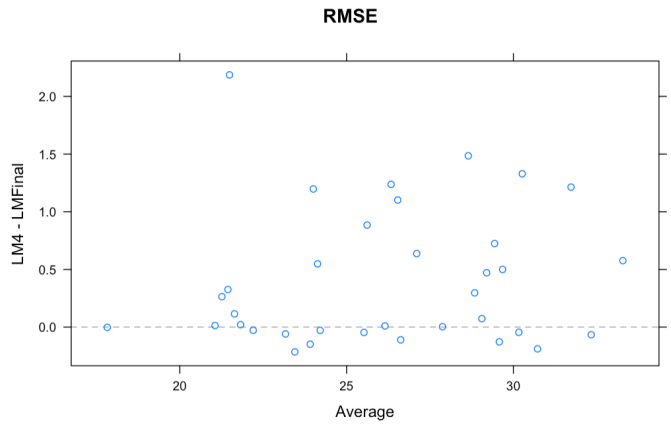


Figure 5. Bland-Altman plot between LM4 and LMFinal.

LM4 and LMFinal both demonstrated similar values of mean R-Squared, MAE, and RMSE values. Therefore, we used the Bland-Altman plot to compare models that were similar. Since we used 7 repeats of 5 fold-cross validation, there are 35 data-points plotted. Each point represents a paired difference for RMSE. Y-axis depicts the paired difference, while the x-axis depicts the average value of RMSE. We reassured that the performance of the models is nearly equal as the plot is a random scatter. Although LM4 that includes all of the predictor variables demonstrate similar performance with that of LMFinal, it is however better to omit the variable Avg.ACT that demonstrates multicollinearity with Avg.SAT.

Conclusion and Discussion

Tuition, Admission rate, ACT & SAT scores were all significant variables in predicting the ranking of a liberal arts college throughout the United States.

Our initial hypothesis that low admission rate corresponds to high college ranking was verified upon simple regression. All of the predictor variables of our choice demonstrated linear relationship with the dependent variable school ranking while variables Avg.SAT and Avg.ACT scores demonstrated almost perfect linear correlation. Upon constructing different multiple regression models, we noticed the model that includes all of the predictor variables has a great out of sample performance. However, the positive coefficient of the variable Avg.ACT and infinitesimally small coefficient of Adm.Rate seemed unparalleled to previously observed effects (upon simple regression) and the correlation we observed

earlier; rise in ACT scores and decrease in admission rate should improve school rankings, but it was suggesting otherwise[Appendix I]. We believe inclusion of variables Avg.SAT and Avg.ACT that are highly linearly related to one another have led to such misleading variable effects.

Upon exclusion of Avg.ACT, we realized that the model had better out of sample performance and compared to the previous model. The effect of the variables were also congruent with our expectation; higher SAT scores, lower admission rate, higher tuition contributed to higher college ranking. We extended our analysis by comparing the final multiple-regression model by constructing a GAM model with the same choice of predictor variables. As evidenced by the Bland-Altman plot, the GAM model provides slightly better predictions. Although the data points look fairly random, LMFinal appears to have more errors than the GAM model, since there are more data points above the 0 line [Appendix II].

In summary, college rankings could easily be predicted with variables that most people believe to be important in determining the rank of an institution. Although some may believe that college ranking is determined upon holistic analysis of different categories, it was surprising to see such strong correlation between college ranking with the most obvious variables. Further study can be done to find other variables that could improve our prediction model or find a more solid dataset that includes more college data as large numbers of college observations have been lost upon data cleaning processes.

Sources

1. College Scorecard Data. (n.d.). College ScoreCard, US Department of Education. https://collegescorecard.ed.gov/data/documentation/

2. Andrew G. Reiter, "U.S. News & World Report Historical Liberal Arts College and University Rankings," http://andyreiter.com/datasets/

3. Office of Postsecondary Education, College Affordability and Transparency List Explanation Form, 2014–15. October 19, 2020, from https://catalog.data.gov/dataset/college-affordability-and-transparency-list-explanation-form-201415

Appendix

```
Call:
lm(formula = Ranking ~ Adm.Rate, data = Tidy2)

Residuals:
    Min      1Q  Median      3Q     Max
-85.888 -23.685  -4.256  22.357  82.533

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -13.007      9.544  -1.363    0.176
Adm.Rate     187.411     16.615  11.280   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37.5 on 107 degrees of freedom
Multiple R-squared:  0.5432,    Adjusted R-squared:  0.5389
F-statistic: 127.2 on 1 and 107 DF,  p-value: < 2.2e-16


Call:
lm(formula = Ranking ~ Tuition, data = Tidy2)

Residuals:
     Min       1Q   Median       3Q      Max
-133.053  -24.623   -0.332   24.018   88.951

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.521e+02  1.623e+01   15.54   <2e-16 ***
Tuition     -4.727e-03  4.513e-04  -10.47   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38.99 on 107 degrees of freedom
Multiple R-squared:  0.5063,    Adjusted R-squared:  0.5017
F-statistic: 109.7 on 1 and 107 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Ranking ~ Avg.ACT + Avg.SAT, data = Tidy2)

Residuals:
     Min       1Q   Median       3Q      Max
-123.205  -13.762   -2.691   11.834   68.567

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 583.1453    55.1981  10.565  < 2e-16 ***
Avg.ACT       1.7083     4.7393   0.360  0.71922
Avg.SAT      -0.4307     0.1419  -3.035  0.00303 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27.81 on 106 degrees of freedom
Multiple R-squared:  0.7512,    Adjusted R-squared:  0.7466
F-statistic: 160.1 on 2 and 106 DF,  p-value: < 2.2e-16


Call:
lm(formula = Ranking ~ Tuition + Avg.ACT + Avg.SAT, data = Tidy2)

Residuals:
    Min      1Q  Median      3Q     Max
-91.622 -10.131  -1.893  11.345  61.685

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.694e+02  5.175e+01  11.003  < 2e-16 ***
Tuition     -1.633e-03  4.066e-04  -4.015 0.000112 ***
Avg.ACT      3.653e+00  4.460e+00   0.819 0.414614
Avg.SAT     -4.164e-01  1.328e-01  -3.135 0.002231 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.01 on 105 degrees of freedom
Multiple R-squared:  0.7844,    Adjusted R-squared:  0.7782
F-statistic: 127.3 on 3 and 105 DF,  p-value: < 2.2e-16
```

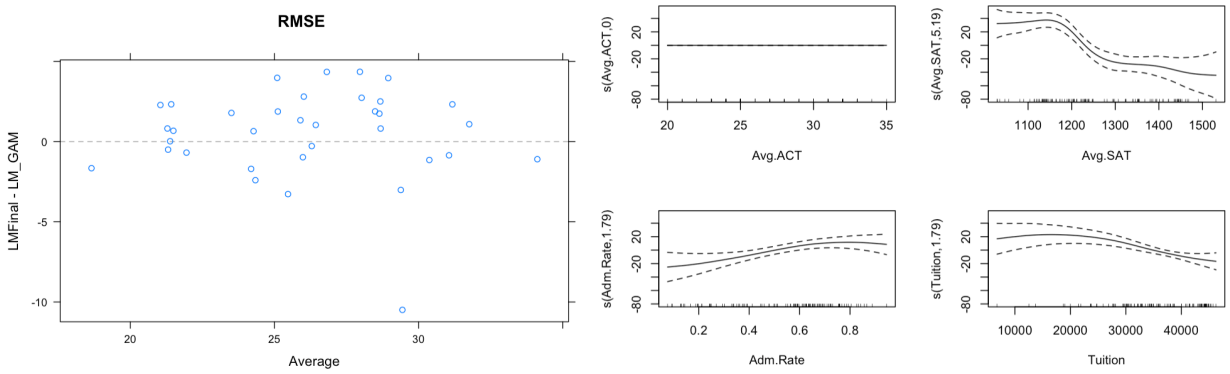
```
Call:
lm(formula = Ranking ~ Tuition + Adm.Rate + Avg.ACT + Avg.SAT,
    data = Tidy2)

Residuals:
    Min      1Q  Median      3Q     Max
-84.041 -13.102  -1.093   9.511  52.809

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.800e+02  6.268e+01   7.658 1.02e-11 ***
Tuition     -1.519e-03  4.003e-04  -3.795 0.000249 ***
Adm.Rate     4.171e+01  1.726e+01   2.416 0.017418 *
Avg.ACT      3.538e+00  4.361e+00   0.811 0.419079
Avg.SAT     -3.637e-01  1.317e-01  -2.762 0.006791 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.43 on 104 degrees of freedom
Multiple R-squared:  0.7958,    Adjusted R-squared:  0.788
F-statistic: 101.3 on 4 and 104 DF,  p-value: < 2.2e-16
```

Appendix I. Regression output for different linear models constructed.



Appendix II. BlandAltman plot of RMSE (GAM vs Final Multiple-Regression Model) and Explanatory Variable vs Prediction Graph