Stefany Laun

Maggie Southwick

Diego Valdes

IST 707

Project Report

6/1/2019

**Introduction**

The American dream, the promise that regardless of one’s social status, rich or poor, black or white, anyone can become a success based on their merits is a core ideal that all citizens take to heart. It is one of the promises absent from many nations; being a key motivation for why so many risk everything to just for the chance to partake in the chance to improve their lives. A cornerstone of that promise is the nation’s education system. One that enshrines students with all the knowledge they will need to become successful regardless of their social status, rich or poor, black or white, they will be judged simply by their merits.

It is common knowledge that there is a divide in the country between the haves and have-nots; a divide that has even pierced the educational system. The knowledge given to students at an affluent school cannot be compared with that received by a child at a poor one. Everything from the condition of the books, the state of school facilities, accessibility of technology, and availability of supplies are all in favor of the haves. How are students that are given a handicap based on their social status expected to compete when they are not provided the same tools to be able to build their merits.

It is different in higher education. Students are reduced to scores on a few sheets of paper, all individuality stripped from them in the spirit of fairness. The haves and have-nots are finally on equal footing. The promise of the American dream can start at a university. While tax dollars contributed to the disparity of the K12 system, every student pays their own way at a university. College is a place where rich or poor, black or white, anyone can become a success based solely on their merits. Or is it?

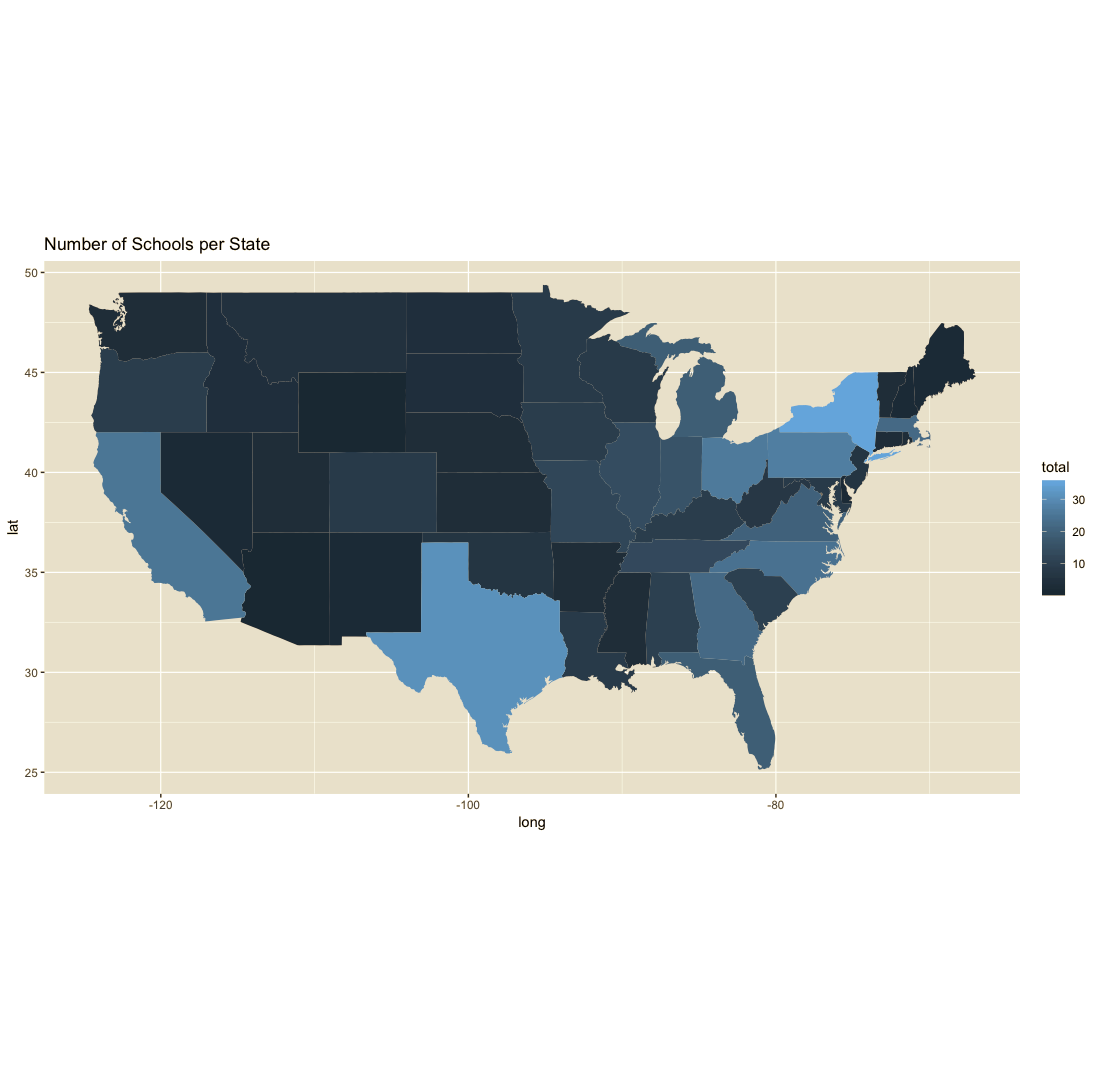
**Analysis**

**About the Data**

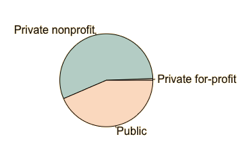
In 2015, the US Department of Education developed a tool called the College Scorecard, which allowed for the comparison of costs and outcomes of studying at a particular university. The tool compares over 7,000 universities on almost 2,000 data points. Public reception to this tool was mixed. Many applauded the sizeable step toward transparency in higher education, while others warned that the way the data was presented could be misleading and understate the importance of financial aid. A group of 7 news articles were analyzed using text mining methods to understand the social context of this dataset. The articles came from a variety of sources, including NPR, Inside Higher Ed and the Wall Street Journal.

Colleges that primarily awarded Bachelor’s degrees without any missing fields were selected for analysis. Extension and online campuses were excluded. This analysis focuses on the relationship between financial aid and student success before, during and after college. To that end, 40 variables for 494 universities were selected to facilitate this analysis. Ultimately, the models in this study focused on 25 of the 40 attributes. These attributes can be classified into 5 groups: university characteristics, financial measures, admission, student progress and employment. Attribute names were substituted with “developer-friendly names” provided in the data dictionary.

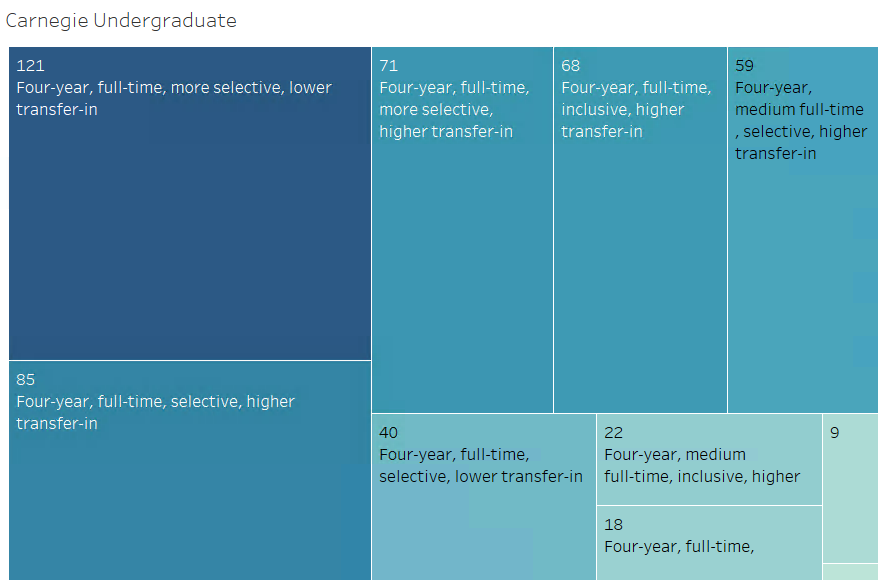
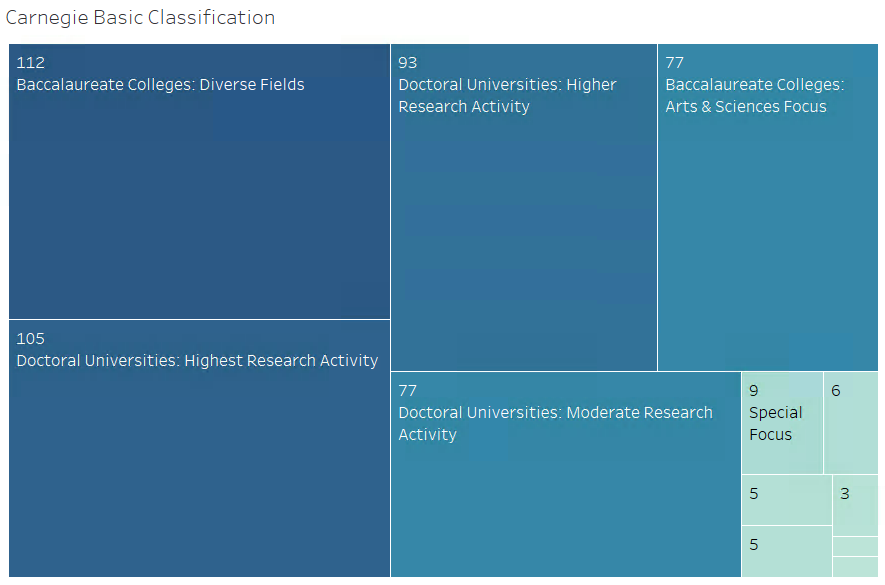
The names of the University were included, and all but two were unique. The two duplicates, both named “Emmanuel College” were modified to include the state abbreviation (i.e “Emmanuel College (GA)” and “Emmanuel College (MA)”). The name attribute was used as a primary key when performing data wrangling during analysis. Colleges from all 50 states, as well as Washington DC were included in the dataset. The states on the Atlantic Coast tended to have the most schools included in their dataset, as did Texas and California. The dataset included less schools from the central states in the US.



Of the 494 schools in the dataset, 276 were privately owned, not for profit schools. There were 3 private for-profit schools, and the remaining 215 were publicly owned.



The dataset also included information about the university’s carnegie classification, which identifies groups of universities that are comparable in terms of research and academic activity. Most universities in the data set were classified either as a Baccalaureate College with a diverse focus, or a Doctoral University with the highest level of research activity. Less common classifications include special focus universities in engineering, music, business, health and faith. The dataset also included the Carnegie undergraduate classification which groups schools according to their student population and selectivity. Most schools were four-year schools with higher selectivity.



The admission rate for each university was expressed as a percentage. The distribution was skewed left; most schools had an admission rate above 50%, while there were some outliers with admission rates under 20%.



Standardized testing scores were provided for each institution. The midpoint SAT scores for Math and Critical Reading were present for every college. SAT Writing scores were more sparsely populated, and therefore was not used frequently in analysis. Scores between the 3 tests were comparable and evenly distributed. Critical reading scores were slightly lower than Math. Since all portions of the SAT are scored on the same 1-800 scale, they could be collectively used to compare SAT performance between schools of different ownership, as shown in the boxplot on the right. Scores for the few private for-profit schools were the lowest. Scores for private nonprofit and public schools were both normally distributed, though private non-profit had a wider spread. The maximum value for private nonprofit schools is about 100 points higher than that of public schools.



The size variable reports the total number of undergraduate degree-seeking students at a university. The distribution is significantly skewed to the right: most schools have small student populations, and the mean is population is lower than the median. This is especially true for private non-profit schools. Almost all private nonprofit schools have undergraduate populations that are less than 10,000 students, though there are several outliers, some of which have populations as high as 50,000. Public schools student populations, however, are more normally distributed. Over 50% of the schools have a population higher than 15,000, and 25% have populations greater than 23,000.

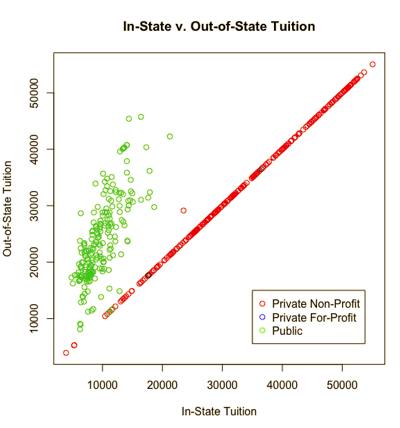


Cost of university attendance was summarized by multiple attributes: attendance.academic\_year, avg\_net\_price, tuition.in\_state and tuition.out\_of\_state. Attendance.academic\_year represents the average total cost of attendance before financial aid, including tuition, room and board, books and program fees. Avg\_net\_price represents the average total cost of those expenses after financial aid. Tuition measures reflect tuition costs only, for those students who live in the same state as the school and qualify for in-state tuition, and those who do not.

The figure below illustrates that the total cost of attendance before aid is does not have a normal distribution: the highest frequency category of payment is between $20k-$30k, while the commonality of price evens out as it gets higher. The cost of a private nonprofit university is, not surprisingly, higher than that of a public university. The average net price after aid has a much more normal distribution, though the distribution is skewed to the right, meaning that most pay less than the mean net price.



In-state and out-of-state tuition are best understood together. For private universities, with few exceptions, there is no difference between the two: tuition is the same whether students live in the same state as the school or not. For public institutions, the cost of out-of-state tuition is invariably higher than in-state. In some cases, it is almost $30,000 higher.



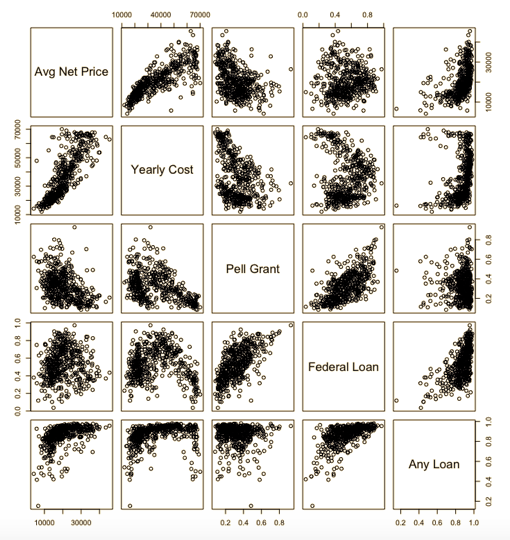
The federal loan rate and pell grant rate fields communicate the percent of the undergraduate student body who received the respective sources of financial aid. Both types of aid are need-based scholarships and awarded by government organizations, rather than the university, The dataset also included the percentage of students with any loan. This includes those loans which are not need based. The figure below illustrates that the percentage of students receiving a federal loan is usually higher than those receiving a pell grant. The distribution of those with any type of loan is skewed significantly to the left. Most students on most campuses cannot pay for school without a loan, regardless of whether they are considered by the government to be low income.



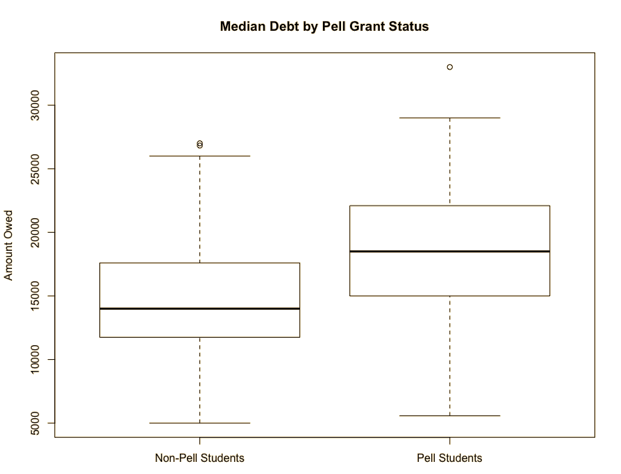
An initial examination of the relationship between cost of attendance and portion of student body with a loan lends interesting insight. The relationships for Pell grants are distinct: as total and net costs increase, the rate of students receiving a Pell grant decreases. This is likely due to the fact that Pell grants are awarded to students from low income families, and while college may be an option, attending an expensive school is not.

Federal loan rate does not have a distinct pattern with average net price, but its relationship with yearly cost is interesting. There appears to be a parabolic relationship: There is a group of schools in which, as yearly cost increases, the percentage of students with a loan increases. This is most relevant in schools with the highest yearly price. There is a second group in which, as cost increases, so too does the percentage of students with a federal loan. This is most common in schools with lower prices. This pattern can most likely be attributed to the same socio-economic reasons observed with pell grants: an extremely high cost of college is prohibitive for those without financial means. In moderately priced colleges, however, tuition payments are more attainable, especially with the assistance of federal loans.

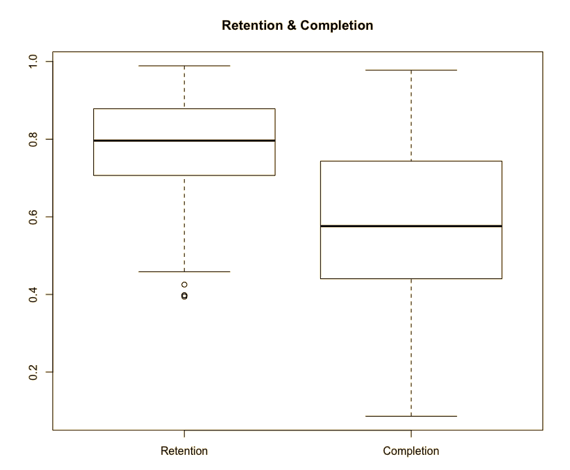
The percentage of students with any loan is highly consistent in its relationship with average yearly cost. Whether the cost is high or low, a high percentage of the student body usually has loans. There are some outliers, especially in those schools with extremely low or extremely high tuition. Its relationship with average net price is similar, though not as strong. There are some schools in the middling range of net price in which about half of the student body has loans and half do not.



The relationship that Pell Grant rates have with university cost is further illustrated in the measures of median debt at the time of leaving the institution (either by graduation or withdrawal). This measure is disaggregated by students who received a pell grant and those who did not. The median amount owed by pell grant students is almost $5,000 higher than those who did not receive a Pell grant. This demonstrates that although these students received additional aid, they still needed to take out more loans than other students.



Retention and completion were used to measure student success while in college. Retention measures the percentage of first time (freshman) full time students who return for a second year. Completion measures the percentage of students who earn their degree in less than or equal to 150% of the expected time to graduation. For students in a 4 year program, this measures whether they graduate in 6 years or less. Retention rates tended to be higher than completion rates. Completion has a much wider spread, with some schools maintaining less than a 20% graduation rate, and with a median slightly higher than 50%.



Median earnings of students 6 years and 10 years after enrolling in the university were used to understand student success after enrollment. For those who graduated 4 years after enrolling, this represents earnings within 2 years and 6 years of graduation. Mean earnings after 10 years are relevant only to those who are employed and not enrolled in another university. There is a steady positive relationship between median wages at 6 and 10 years. Private nonprofit schools have the highest wages. Most schools report median earnings that are less than $50,000 after 6 years and $70,000 after 10 years.



**Models**

Seven articles published at the time the college scorecard was released were analyzed using Text Mining methods to understand public sentiments regarding the data tool. The articles were saved as a plain text file and imported into R studio as a corpus. The text was then cleaned by transforming all text to lowercase, removing numbers and stopwords, removing punctuation and white space, and stemming words (shortening all words to a common root word). Additional common words that did not carry meaning, such as “said”, “use”, “one” and “will” were also removed.

The corpus was transformed into a Term Document matrix, in which each word in the collection of documents was contained in a row, and each article was represented in a column. The values therein represented the number of times a word occurred in a particular document. Since the articles were of similar length, no normalization methods were applied to counteract documents with particularly high or low word counts. The term document matrix was aggregated by term in a new dataframe, such that each term was listed with the total number of times it occurred in all documents. This transformed data frame was used to create a word cloud, which highlights the most common words that were used.

The K-Means Clustering Algorithm was used to cluster data points focused on the admissions phase of the higher education lifecycle and demonstrate the admission rate association with title IV funding, the federal loan rate and pell grant rate based on school ownership (and subsequent access).

Clustering was also used to explore the university data and assign them to groups according to similarity in cost, financial aid and success measures during college. The analysis was based on 7 variables: size, yearly cost, average net price, pell grant rate, federal loan rate, retention rate and completion rate.

Distance measures formed the basis of each clustering model and were generated using the “get\_dist” function from the “factoextra” package in R. Distance was measured using the Euclidean method, as well as the Manhattan distance method. Given that each data point can be plotted on a cartesian plane, the Euclidean distance would measure the length of a straight line between two points, while Manhattan distance would measure the length of the two or more lines, joined at 90 degree angles, which are needed to connect two points.



The resulting distance matrices were used to visualize similarities and differences between groups of documents and form a preliminary impression of how many clusters may be ideal. Additional visualizations rendered using functions from the factoextra package helped to identify the ideal number of clusters as well, and are summarized in the table below.

|  |  |
| --- | --- |
| **Function** | **Parameters** |
| fviz\_nbclust | FUN=hcut  methods=”silhouette” (average silhouette width) |
| clusGap  Fviz\_gap\_stat (clusGap) | FUN=hcut  nstart=25 (generates 25 initial random centroids and chooses the best)  K.max = 10 (maximum number of clusters)  B = 50 (number of samples) |



The Euclidean Distance Matrix pictured above signifies the intersection similar universities with orange cells, and the intersection of dissimilar universities in purple cells. While multiple areas are shaded in orange, indicating a group of similar observations, the colored matrix contains four distinct groups of similar universities, which are outlined in blue. This suggests that 4 clusters may be the ideal number.

The graphs in the image below summarize the findings of the two methods determining the optimal number of clusters. The silhouette method identifies two distinct clusters. While this may be accurate, using only clusters does not meet the needs of this analysis, which would benefit from at least 3 distinct groups of universities to compare. The graph also indicates that using 3-5 clusters would be beneficial, before dropping off significantly at 6 clusters. The clusgap method identifies 4 as the optimal number of clusters, while 5 and 6 are also strong choices. These insights will inform parameter tuning during cluster analysis.



Hierarchical methods were explored initially to group universities together in similar groups. Hclust methods refer to the definition of intercluster distance. Complete linkage is defined by the maximum distance between clusters, while single linkage measures the minimum distance. Ward minimizes the distance within a cluster. Average linkage represents the average distance between each member of each cluster, and centroid measures the distance between the center of the two clusters. The agnes function generated the agglomerative coefficient, which measures the strength of the clustering structure and provided the basis of assessing the best model. The parameters of the tested models are summarized in the table below.

|  |  |  |
| --- | --- | --- |
| **Distance Measure** | **Hclust Methods** | **K-values for dendrogram** |
| Euclidean | Complete, single, ward, average, centroid | 5,4 |
| Manhattan | Complete, single, ward, average, centroid | Not tested |

Partitional clustering methods were then used to discover similarities between schools. The Kmeans function in the cluster package in R was used to this end. This algorithm was tuned with the goal of reducing the distance of each university from the center of the cluster, or the sum of squared error (SSE).

|  |  |
| --- | --- |
| **Distance Measure** | **K-Values** |
| Euclidean | 3,4,6 |
| Manhattan | 3,4,6,8 |

**Association rules mining : During School Success**

Association rules mining was used to identify associations between financial and success measures while students are still in school. The following variables were examined to perform this analysis: carnegie.basic, carnegie.ugrd, state, size, yearly cost of attendance, average net price, pell grant rate, federal loan rate, retention rate, completion rate, and the clusters assigned using retention and completion measures. Each numeric variable was cut in thirds and assigned a discrete label of “low”, “medium” or “high.” Categorical variables were transformed to factors. It was further transformed into a transaction dataset to make the analysis process smoother.

The following parameters were specified to mine for association rules. Each rule set was evaluated based on support, confidence and lift measures.

|  |  |  |  |
| --- | --- | --- | --- |
| Ruleset # | Question to answer | Parameters | {lhs} → {rhs} |
| rules | What are the common associations of universities in general? | supp=0.2 conf=0.9  minlen=2 | Not specified |
| rules2 | Given that a university has a high federal loan rate, what are their other characteristics? | Supp = 0.01  Conf = 0.2  Minlen = 2 | Lhs = “federal\_loan\_rate=high” |
| rules3 | What characteristics are associated with a high federal loan rate? | Supp=0.04  Conf=0.8  Maxlen=5 | Rhs = “federal\_loan\_rate=high” |
| rules8 | Given that a university has a high pell grant rate, what are their other characteristics? | Supp=0.001  Conf=0.01  Minlen=2 | lhs=”pell\_grant\_rate=high” |
|
| rules9 | Given that a university has a low pell grant rate, what are their other characteristics? | Supp=0.001  Conf=0.01  Minlen=2 | lhs=”pell\_grant\_rate=low” |
|
| rules10 | Given that a university has a low federal loan rate, what are their other characteristics? | Supp=0.001  Conf=0.01  Minlen=2 | lhs=”federal\_loan\_rate=high” |
|

**Decision Tree: During School Success**

Decision tree modeling was used to predict success during school, namely retention and completion rates, based on federal loan rates and pell grant rates for a given university. Retention and completion were evaluated separately, but both were predicted using the following attributes: name, size, yearly cost, average net price, pell grant rate, federal loan rate and cluster. Retention and completion rates were cut into 3 categories labeled “low”, “medium” and “high,” The values and number of observations in each category are summarized below, as well as the chance of randomly guessing a value correctly. All other values were left as raw, continuous numbers. The retention data set did not include the completion measure, and vice versa, because the two measures have a strong relationship with each other and would skew results.

**Retention**

|  |  |  |  |
| --- | --- | --- | --- |
| **Label** | **Upper & Lower Values** | **Number of Observations** | **Random Guess Probability** |
| Low | 39.4% - 59.2% | 44 | 8.9% |
| Medium | 59.2%-79.1% | 195 | 39.57% |
| High | 79.1% - 98.9% | 255 | 51.6% |

**Completion**

|  |  |  |  |
| --- | --- | --- | --- |
| **Label** | **Upper & Lower Values** | **Number of Observations** | **Random Guess Probability** |
| Low | 8.5% - 38.3% | 74 | 14% |
| Medium | 38.3% - 68.1% | 253 | 51.2% |
| High | 68.1% - 97.9% | 167 | 33.8% |

The data was randomly divided into testing and training datasets with a user-defined function. The training dataset contained 67% of the data, or 330 observations, while the test set contained the remaining 264 universities. Information Gain and Information Gain Ratio were used to identify the most important variables in predicting each success measure. These variables were tested in rfit2, rfit3, cfit2 and cfit3.

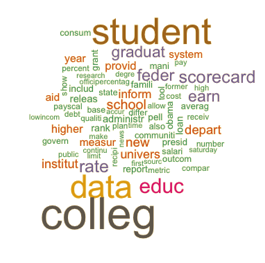
The following parameters were used to build 6 decision trees. Each tree was applied to the testing dataset and evaluated by its accuracy rate in predicting low, medium and high retention and completion rates on both the training and testing datasets.

|  |  |  |
| --- | --- | --- |
| **Tree name** | **Formula** | **Complexity Parameter (cp)** |
| rfit1 | retention ~ . [all] | cp =.0001 |
| rfit12 | retention~. [all] | cp=.007 |
| rfit2 | retention~attendance.academic\_year+federal\_loan\_rate+size+pell\_grant\_rate (gain ratio) | cp=.02 |
| rfit3 | retention~cluster+pell\_grant\_rate+attendance.academic\_year + size (information gain) | cp=.02 |
| cfit1 | completion~ . [all] | cp=.006 |
| cfit2 | completion~attendance.academic\_year + avg\_net\_price + pell\_grant\_rate (gain ratio) | cp=.02 |
| cfit3 | completion~cluster+pell\_grant\_rate+attendance.academic\_year (information gain) | cp=.02 |

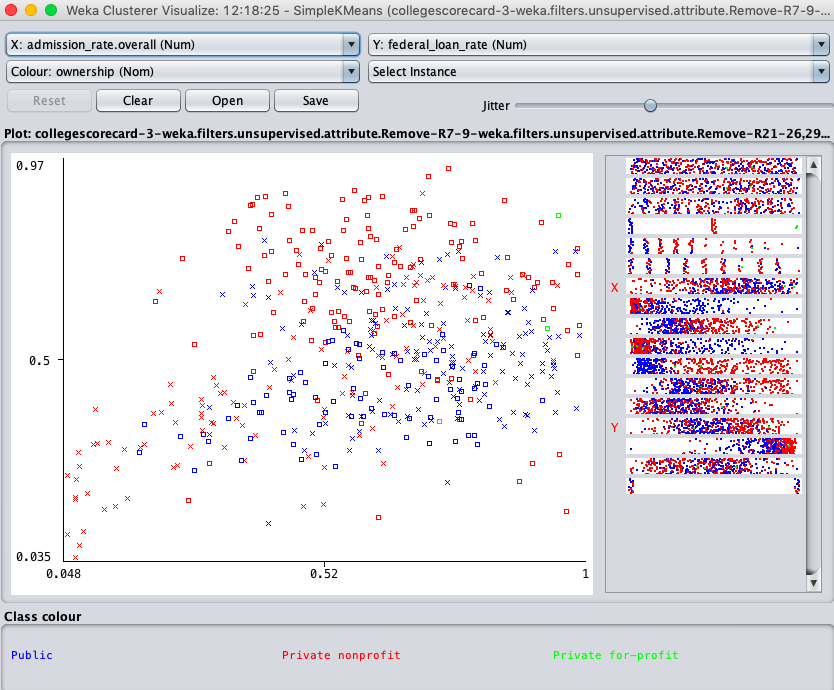
The Naive Bayes, SVM, and Random Forest models were used in an attempt to predict ownership (public, private nonprofit, private for profit) of a student based on their success after leaving the institution. The initial models used 8 variables based on future earnings, loan principals, school State, and SAT scores. These models didn’t perform well, as the State attribute was heavily influencing the predictions. Once it was removed, the models did perform better. Data for private for profit schools was also removed, as there wasn’t enough of samples for the models to be able to accurately predict their results.

**Results**

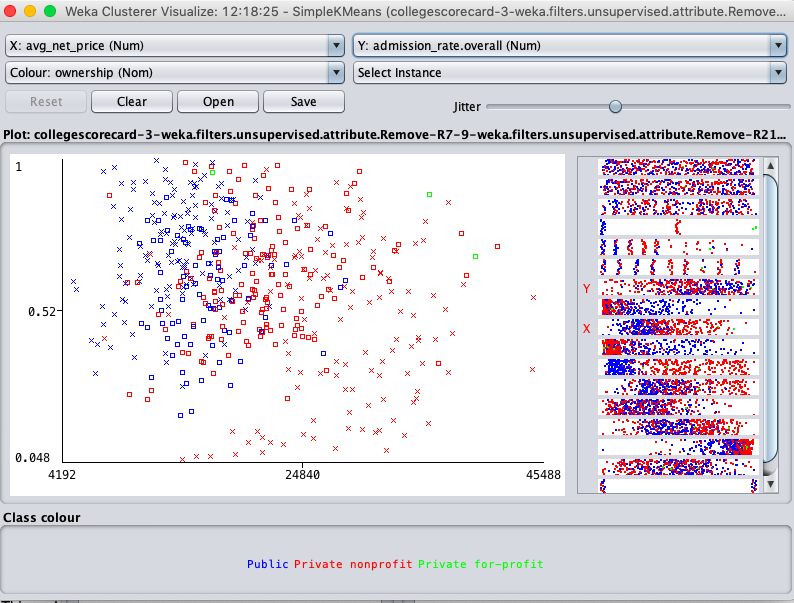
The word cloud produced by the articles is below. Large words such as “ student”, “college”, and “data” were used most commonly. Other prominent words that offer insight into the public reaction to this dataset include “low income” “aid”, “family”, “graduate”, “earn” and “payscale.” The frequency of “low income”, “aid”, and “family” demonstrate that there is a high consideration of the impact that these universities and statistics have on students receiving aid. These students usually come from low income households, as determined by family earnings. The words “graduate”, “earn” and “payscale” show the emphasis this dataset places on outcome measures after graduation. The public is reacting to the data that shows the return on investment from college degrees.



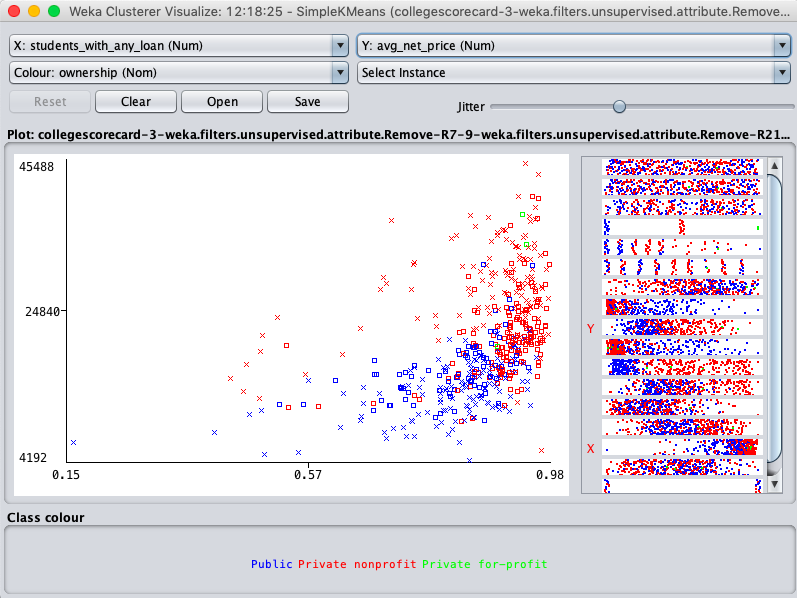
The K-Means Clustering Algorithm returned (when visualizing admission rate on the X axis, federal loan rate on the Y axis, and color as ownership (public or private), that private nonprofit institutions had much higher federal loan rates and were overall more selective than their public counterparts.



This makes sense, as the cost of attendance at a private institution is typically on average far greater than that of a public university. This can be seen in the visualization below.



To further demonstrate the disparity in cost of private versus public universities, the below K-Means Clustering analysis visualization depicts how much higher on average the cost of attendance and subsequent federal loan rate are at a selective private institution is in comparison to its public counterparts.



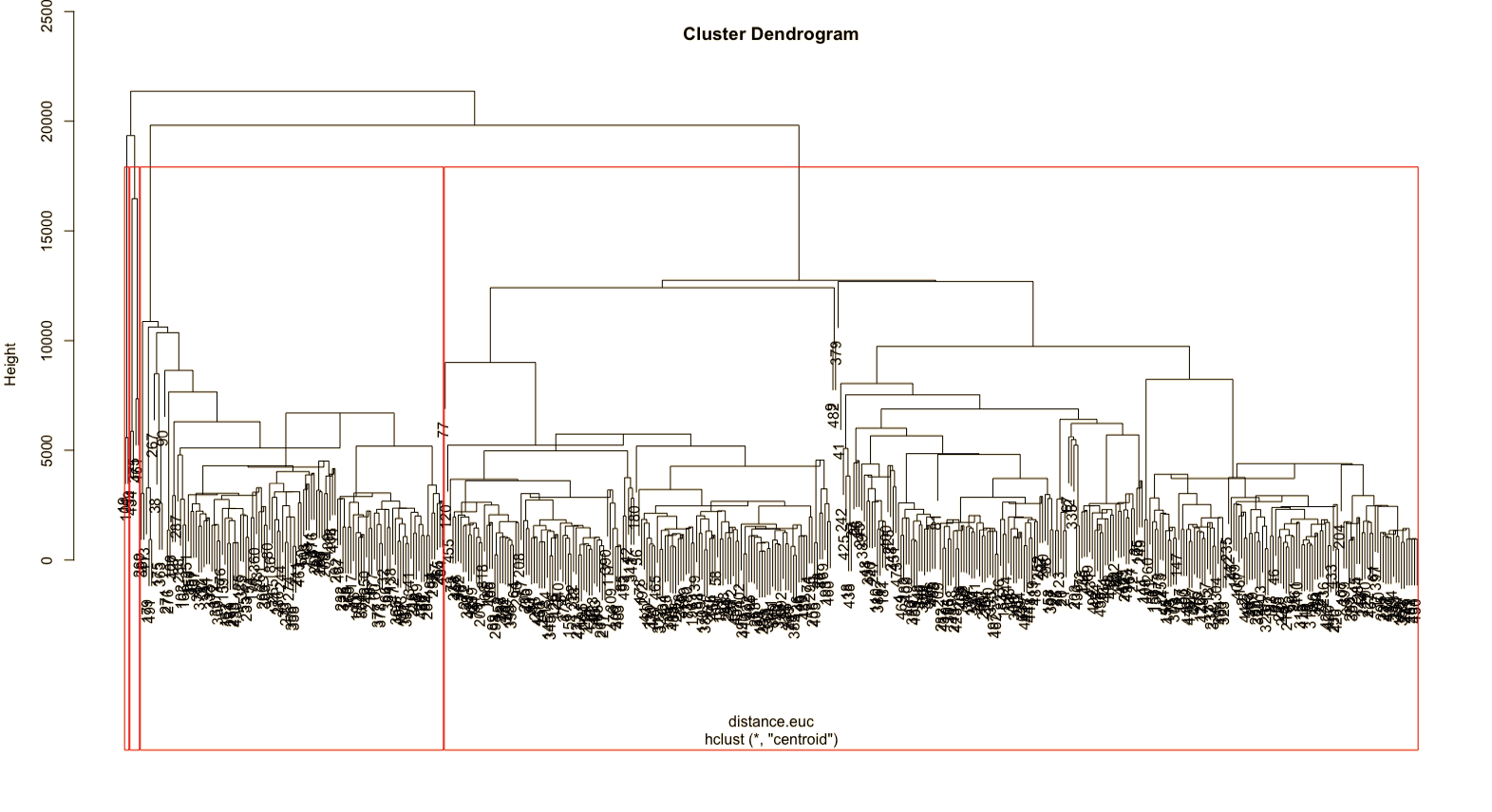
Naive Bayes performed the most poorly for predicting ownership based on future student performance. The kappa score places this model is just slightly better than random guessing. The overall accuracy was computed with a 95% confidence interval (.5047, 0.6697) determining that the lower bound is lower than the no information rate of 0.5822. This also aligns with the kappa score that this model is not better than guessing one ownership 100% of the time.

**Cluster Analysis: During School Success**

The results of comparing distance measures for hierarchical clustering are summarized in the table below. The models built using Ward’s method and Complete linkage had the strongest structures, as indicated by the high value of the agglomerative coefficient, for both Euclidean & Manhattan distance measures. This method was used to evaluate the ideal number of clusters. The agglomerative coefficient was not available for the function used to build models with the centroid clustering method, so this was evaluated visually. This evaluation was especially valuable when building partitional models.

|  |  |  |
| --- | --- | --- |
| Clustering  Methods | Agglomerative Coefficient | |
| Euclidean Distance | Manhattan Distance |
| Single Linkage | .858 | .875 |
| Average Linkage | .9524 | .956 |
| Complete Linkage | .9735 | .98 |
| Ward’s Method | .9958 | .9959 |
| Centroid | N/A | N/A |

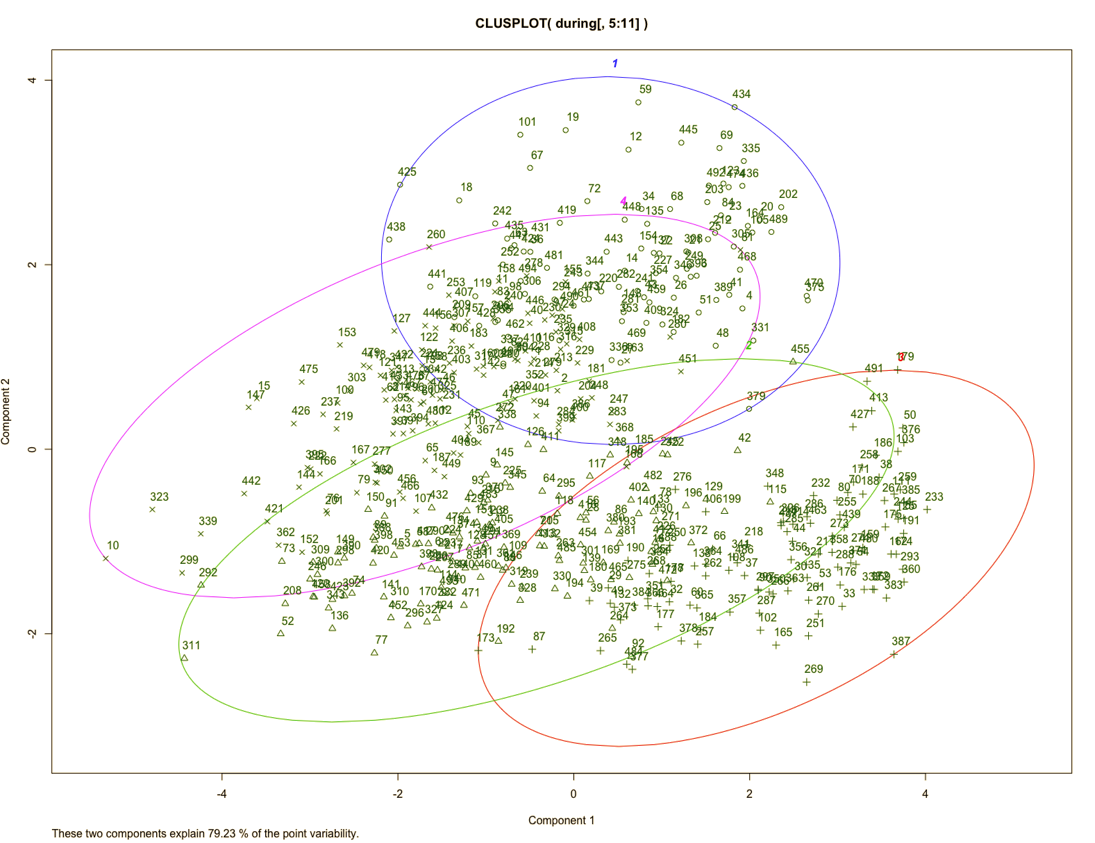
The centroid method was not highly promising. The best model contained three clusters of highly unequal size. Every k-value, including k=2, produced a very small cluster on the left-most side of the dendrogram. This is not the best approach for hierarchical clustering and will be disregarded.



The results of the visual analysis were used to inform testing of partitional clustering parameters, the results of which are summarized in the table below with the optimal k value for each distance measure indicated in bold. The strongest model overall is based on the Euclidean distance measure with k=4 and is visualized in the diagram below. Given the high SSE value, this method of clustering was not used.

**Sum of Squared Error (within cluster)**

|  |  |  |
| --- | --- | --- |
| **K-values** | **Euclidean** | **Manhattan** |
| k=3 | 610,054,907 | 1,111,637,091 |
| k=4 | **422,135,799** | 783,322,118 |
| k=6 | 449,417,967 | **737,723,281** |
| k=8 | Not tested | 740,063,978 |



The models built with Ward’s method and Complete linkage were compared using a tanglegram, pictured below. The models clustered universities similarly: those in the pink, blue and red groups on the left are largely kept together in the dendrogram on the right. The green group has slightly more variation, as indicated by the green lines mixed in with other colors on the right. Overall, however, the models are fairly consistent in the way they group universities together. Since Ward, based on the Manhattan distance measure, has the highest agglomerative coefficient, it will be used for the final cluster assignments.

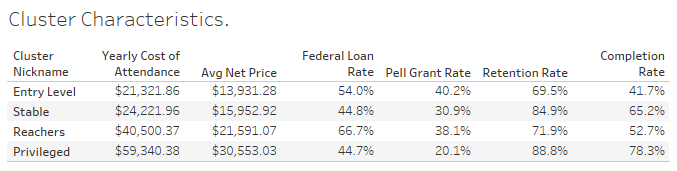


Based on the recommendations of the gap stat function, as well as the visual analysis of the distance matrix, k values of 3-6 were visually evaluated for the Ward and centroid models. Fitness of a model was determined by its ability to divide universities into fairly even groups that would facilitate comparison. Groups should be distinct from each other, and dividing them into small, extremely specific groups would make it difficult to distinguish the remarkable differences between universities. For example, k=3 generates 3 large, general groups, while k=6 generate 6 small specific groups, neither of which would be ideal for comparing groups of universities.



K values of 4 and 5 generate helpful groups that are similar in size, and the right balance of general and specific. K=4 has the advantage of grouping universities along the 2nd split of both tree branches, as indicated by the blue line below. This suggests that it would consistently group universities at the same level of specificity, rather than dividing on lower more specific branches for some groups, as it does when k=5 or k=6. For this reason, k=4 was determined to be the ideal number of clusters.





The characteristics of each cluster were evaluated based on average values of retention rate, completion rate, loan rates and cost of universities. These values are summarized in the table below. Clusters were assigned nicknames based on the values of each variable and are described below.

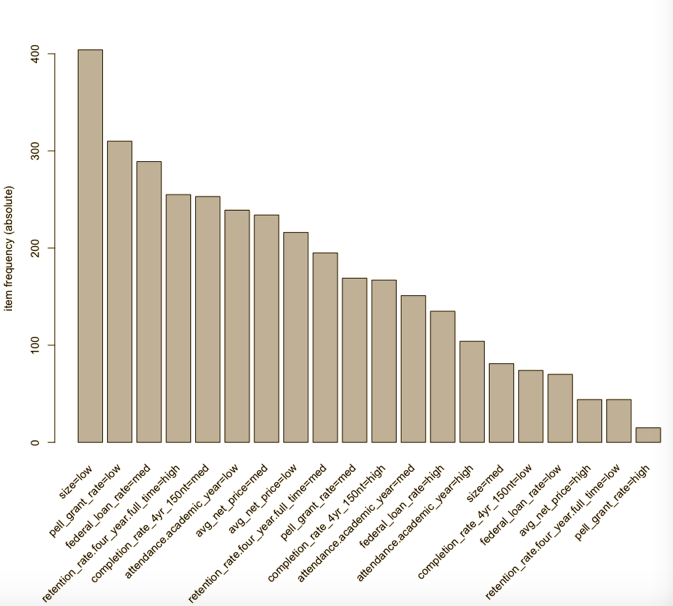
1. Entry level schools have the lowest cost of attendance and average net price. It has some of the highest loan rates, especially pell grants which are awarded to students from low income households. The retention and completion are rates are, on average, the lowest of the four clusters. Students who have lesser means are likely to attend these schools, with poorer results.
2. Stable schools have low costs of attendance, as well as lower federal aid rates. However, their during-college success rates are some of the highest. Schools in this cluster are almost exclusively public schools and are identified as “stable” because they offer a strong return from a relatively small investment. Also, students who attend thse schools tend to be able to afford them, as indicated by low loan rates.
3. Reacher schools are characterized by their high cost of attendance, high loan rates, and low during-school success measures of completion and retention. These schools are a mixture of public and private and are labeled reachers because their student body pays more than they can afford, with a mediocre return.
4. Privileged schools have the highest average cost of attendance, but some of the lowest financial aid rates. This is accompanied by the highest retention and completion rates. Students who attend these expensive schools are able to afford them: even if they do take out some loans, most are not granted need-based loans. These privileged students see a positive return on their significant investment while in school. Almost all of these schools are private institutions.

The clusters are visualized below. The cluster plot is busy, but schools such as Harvard and Stanford stick out in purple, among the Privileged universities, and state schools such as Texas A&M and Ohio State stick out in green, among the stable universities. The entry level group, in red, is comprised of smaller-name schools, as is the Reacher group in blue. There is some overlap between groups. For example, University of Tulsa, a Reacher school, is situated neatly within the Privileged group (circled in black below). It is therefore important to keep in mind that while helpful, these clusters are not perfect. An interactive plot of these university clusters can be found at the link below. <https://public.tableau.com/profile/maggie.southwick#!/vizhome/CostLoans/Clusters>



**Association Rules Mining**

Before beginning association rules mining, an item frequency plot was generated to evaluate the most common observations in the dataset. The top 5 most common observations were low size, low pell grant rate, medium federal loan rate, high retention and medium completion.



|  |  |  |  |
| --- | --- | --- | --- |
| Ruleset # | Question to answer | Parameters | {lhs} → {rhs} |
| rules | What are the common associations of universities in general? | supp=0.2 conf=0.9  minlen=2 | Not specified |
| Rules: | | | |
| {size=low, attendance.academic\_year = high} → {pell\_grant\_rate=low} | | Supp = .21 | confidence=1.0 | lift=1.59 | |
| {size=low, retention\_rate.four\_year.full\_time=high} → {pell\_grant\_rate=low} | | Supp = .33 | confidence=.91 | lift=1.44 | |
| {completion\_rate\_4yr\_150nt=high} => {pell\_grant\_rate=low} | | Supp = .32 | confidence = .95 | lift = 1.52 | |

Insights:

Expensive, small colleges have a low percentage of students with pell grants. Also, high retention and completion are both associated with low pell grant rates.

|  |  |  |  |
| --- | --- | --- | --- |
| Ruleset # | Question to answer | Parameters | {lhs} → {rhs} |
| rules2 | Given that a university has a high federal loan rate, what are their other characteristics? | Supp = 0.01  Conf = 0.2  Minlen = 2 | Lhs = “federal\_loan\_rate=high” |
| rules10 | Given that a university has a low federal loan rate, what are their other characteristics? | Supp=0.001  Conf=0.01  Minlen=2 | lhs=”federal\_loan\_rate=high” |
|
| Rules | | | |
| {federal\_loan\_rate=high} => {completion\_rate\_4yr\_150nt=med} | | Supp = .17 | confidence = .62 | lift = 1.23 | |
| {federal\_loan\_rate=high} => {retention\_rate.four\_year.full\_time=med} | | Supp = .16 | confidence = .57 | lift = 1.44 | |
| {federal\_loan\_rate=high} => {retention\_rate.four\_year.full\_time=low} | | Supp = .06 | confidence = .22 | lift = 2.49 | |
| {federal\_loan\_rate=low} => {retention\_rate.four\_year.full\_time=high} | | Supp = .11 | confidence = .786 | lift = 1.52 | |
| {federal\_loan\_rate=low} => {completion\_rate\_4yr\_150nt=high} | | Supp = .10 | confidence = .714 | lift = 2.11 | |



Insights:

If a school has a high percentage of students with federal loans, their retention and completion rates are 50-60% likely to be medium. It is less common for federal loan rates to be high and have a low retention rate, but the lift for this rule is high, at 2.49, indicating that this rule occurs often, given how uncommon it is for both to occur.

Low federal loan rates are strongly associated with high retention and completion.

|  |  |  |  |
| --- | --- | --- | --- |
| Ruleset # | Question to answer | Parameters | {lhs} → {rhs} |
| rules8 | Given that a university has a high pell grant rate, what are their other characteristics? | Supp=0.001  Conf=0.01  Minlen=2 | lhs=”pell\_grant\_rate=high” |
|
| rules9 | Given that a university has a low pell grant rate, what are their other characteristics? | Supp=0.001  Conf=0.01  Minlen=2 | lhs=”pell\_grant\_rate=low” |
|
| Rules | | | |
| {pell\_grant\_rate=high} => {completion\_rate\_4yr\_150nt=low} | | Supp = .02 | confidence = .73 | lift = 4.89 | |
| {pell\_grant\_rate=high} => {retention\_rate.four\_year.full\_time=low} | | Supp = .16 | confidence = .53 | lift = 5.98 | |
| {pell\_grant\_rate=high} => {retention\_rate.four\_year.full\_time=med} | | Supp = .14 | confidence = .47 | lift = 1.18 | |
| {pell\_grant\_rate=low} => {retention\_rate.four\_year.full\_time=high} | | Supp = .45 | confidence = .72 | lift = 1.38 | |
| {pell\_grant\_rate=low} => {completion\_rate\_4yr\_150nt=high} | | Supp = .28 | confidence = .51 | lift = 1.52 | |

Insights:

If a school has a high percentage of low income students with a pell grant, it is 53%-73% likely that their retention and completion rates will be low. There is a 47% chance that the retention rate will be medium.

If a school has a low percentage of such students, there is a 51% chance that their completion rate will be high, and a 72% chance that their retention rate will be high.



**Decision Tree**

Decision tree models were able to predict a discrete rate of retention and completion with at least 50% accuracy in training scenarios, and at least 69% accuracy in testing scenarios. Having a higher testing accuracy than training accuracy is unexpected, and could be attributed to an imbalanced training and testing set (i.e. the testing set has a disproportionate number of “high” values). Attempts to correct this though different sampling methods produced the same results, and are reported below.

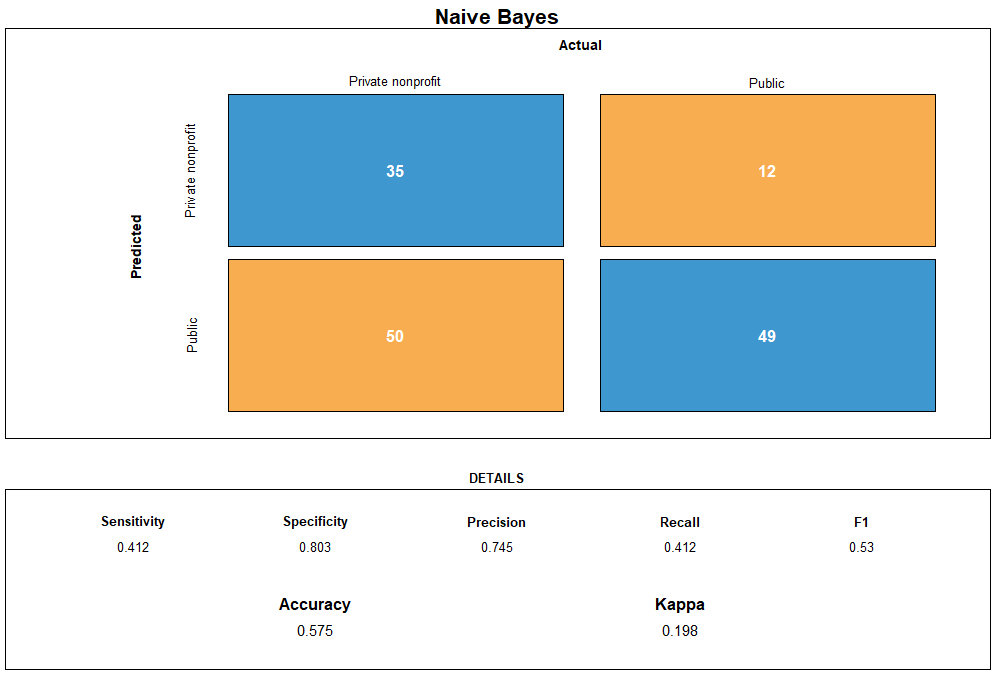
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tree name** | **Formula** | **cp** | **Training Accuracy** | **Test Accuracy** |
| rfit1 | retention ~ . [all] | cp =.0001 | 62% | 71.34% |
| rfit12 | retention ~ . [all] | cp=.007 | 61.49% | 71.95% |
| rit2 | retention~attendance.academic\_year+federal\_loan\_rate+size+pell\_grant\_rate (gain ratio) | cp=.02 | 59% | 75% |
| **rfit3\*** | retention~cluster+pell\_grant\_rate+attendance.academic\_year + size (information gain) | cp=.02 | 59% | 75% |
| cfit1 | completion~ . [all] | cp=.006 | 56.97% | 69.51% |
| cfit2 | completion~attendance.academic\_year + avg\_net\_price + pell\_grant\_rate (gain ratio) | cp=.02 | 49.7% | 71.975% |
| **cfit3\*** | completion~cluster+pell\_grant\_rate+attendance.academic\_year (information gain) | cp=.02 | 52.2% | 73.78% |

Decision tree models for predicting retention rates were more accurate than predicting completion. Retention accuracy was greatly enhanced by focusing on attributes with the highest information gain and gain ratio, which had a testing accuracy rate of 75%. The percent of students who received a pell grant was highly important in this model. If the percentage was less than 30% it was highly likely that retention would be high, and if not high: medium. If pell grant rate was greater than 30%, it was most likely to have no greater than medium retention. If the size of the school was small, it would likely have low retention. If the size of the school was high, it would more likely have high retention. This may be because retention is measured as the percentage of first year students who return for a second term. If the school is larger, it may have a more vibrant culture than perhaps a smaller community college, with a size less than 773.

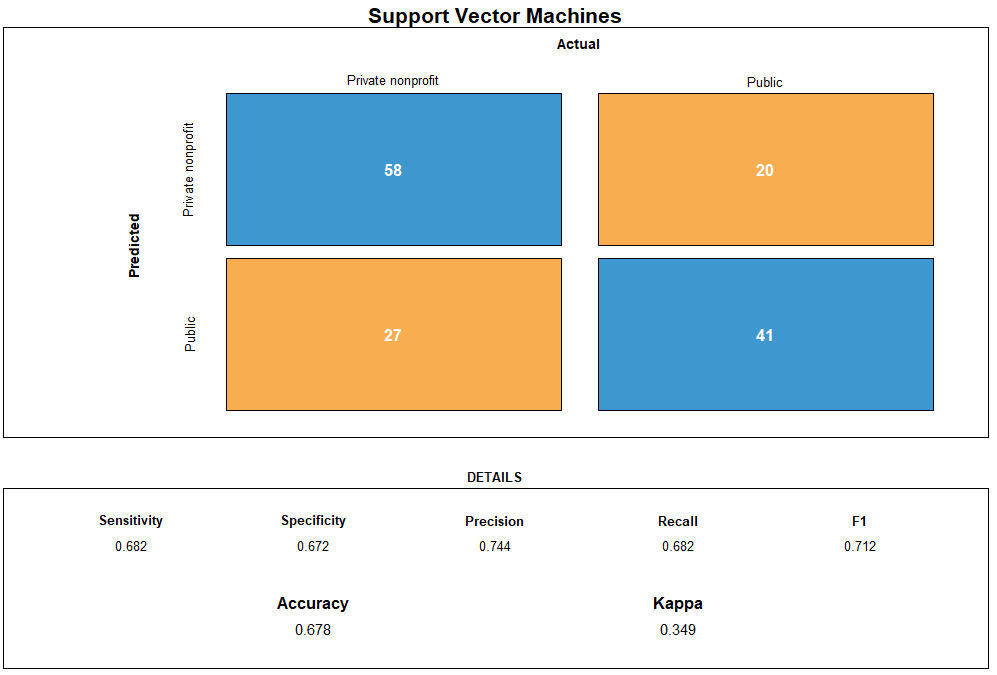


Predicting completion had a slightly lower accuracy rate but was similarly improved using variables with high information gain. Once again, the pell grant rate proved to be most valuable in predicting completion. A school with a pell grant rate less than 23% was highly likely to have a high rate of completion as long as the cluster was not entry level. If it had low financial aid but was an entry level school, it would have medium completion rates. If pell grant rates were greater or equal to 23%, and the yearly cost of the school was greater than $17k, it would also have medium completion rates. Otherwise, if the yearly cost was less than $17k it would most likely have low completion. Finally, if the pell grant rate was greater or equal to 53% the school would most likely have low completion rates.

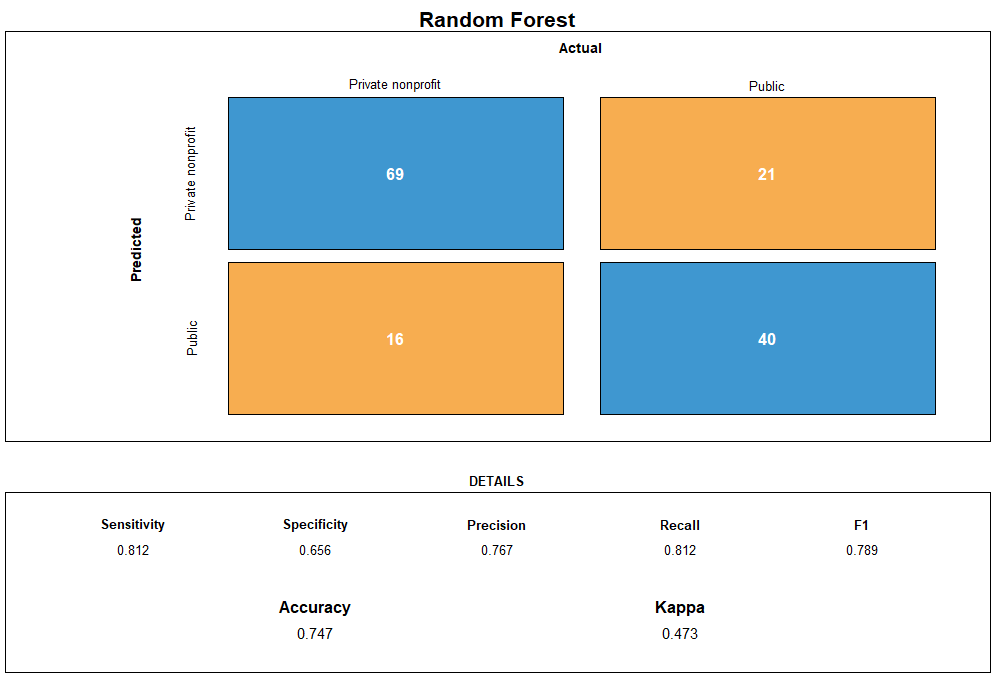




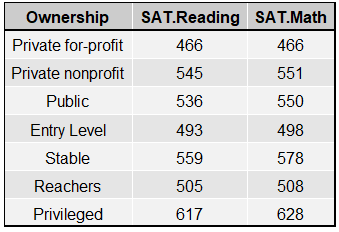
SVM model performed better than Naive Bayes for predicting ownership, but not by much. The kappa score is decent and the accuracy at 95% confidence interval (0.5958, 0.7529) is barely higher than the no information rate at 0.5822. This model is performing slightly better than guessing, but not by much. It does have the ability to be as accurate as 75%. It is possible that with more student focused data, and other variables not used in this analysis, this model could be improved for the predictions.



Random Forest was the best performing model to predict ownership. All of the significant statistics are in desired values. At 95% confidence interval, the lower bound is 69%, which is higher than the previous models did on the test data, and upper bound is 81%. The model’s bounds are considerably higher than the no information rate, meaning it performs better than guessing.



For all three models used to predict ownership, the most significant variables were loan principal, reading SAT score, and math SAT score. There can be significant differences between the types of students attending these universities if looking solely at SAT scores, the difference between students at a privileged and stable school is 50 points.



This observation raises the question: is it the school that is providing the students the avenues to success, or is it the student’s success lifting the reputation of the university, thus attracting more of the types of students who are more likely to succeed for other reasons outside of the school they attend? Would these students be successful at the same pace as their academic peers, regardless of the school name on their diploma?

**Conclusion**

Despite the incredibly strong desire to optimistically conclude that it is, in fact, different in the world of higher education--that the haves and have-nots are finally on equal footing and the promise of the American dream begins with the entrance into a university--this study concludes quite the opposite. From the beginning of the admissions process to retention throughout and ultimate graduation from a university, private or public, it can be clearly seen that the system unfortunately favors the rich and further burdens the poor with strict admissions processes, tough retention and graduation statistics and even more bleak graduate loan payments.

Ultimately, money provides access to better universities and can lead to better success. Private universities on average charge more, have more rigorous admissions standards, and the highest loan rates. High SAT scores are indicative of lower loan rates, either meaning those students are more likely to receive other merit-based scholarships or that they come from families able to pay for rigorous SAT prep courses and subsequently do not need to take out loans to attend college. Additionally, the higher the federal loan rate the higher the admission rate, indicating that the more money one has been provided with the more likely they are to apply and become admitted to a university.

Unfortunately, once the admissions hurdle has been overcome, higher financial aid rates are then associated with universities with low retention and completion rates. The same is true for the reverse: schools with low retention and completion rates are associated with high financial aid rates. While it is not necessarily true that students must pay a tremendous amount to be admitted to a school that will be supportive throughout their academic career, there are far more examples of privileged student success and underprivileged students paying a significant amount not returning successful results.

Interestingly enough, the most powerful predictor of student success for both retention and completion, predicted with 75% accuracy, is pell grant rate: high retention and completion are both associated with low pell grant rates. While pell grants are less common, they hold an incredible amount of significance for the purposes of this study, as they are reserved for students from very low income families. As is to be expected, expensive, small colleges have a low percentage of students with pell grants. If a school has a high percentage of low income students with a pell grant, it is 53%-73% likely that their retention and completion rates will be low, but if a school has a low percentage of such students, there is a 51% chance that their completion rate will be high, and a 72% chance that their retention rate will be high. This is indicative of the fact that students from underprivileged backgrounds are most highly associated with success during school *when they are provided the opportunity to do so.* Opportunity is the most important lesson learned from the work presented herein: the more that a student is truly believed in from the start and the more that is invested that does not come accompanied by crippling debt and disappointment, the more successful we set our future students up to become.