Definition

Project Overview

Retention rate is the percentage of a school’s first-time, first-year undergraduate students who continue at that school the next year (FAFSA, 2016). Retention rates forecast a lot of detail about colleges and universities; they are strongly considered in national and international rankings. Why is that? Bottom-line, retention rates tell whether the transitioning young adult is happy and surviving in his/her new academic setting. Low retention rates signal likely problems within the college or university. Some of the possible problems include poor housing options, poor dining services, financial aid shifts, rigorous courses, overpopulated classes, or a weak advising system. Though factors such as homesickness and family matters affect all colleges, retention rates can give an incoming student perspective on the value a college or university places on you (Unigo, 2016).

Retention rates are also a priority for academic institutions. Increases in retention rates vastly reduce the costs of recruiting efforts to procure sustainable enrollment numbers. Higher retention rates mean less seats to fill next year. Also, most state governments have shifted their funding criteria from enrollment numbers to performance-based metrics, such as retention rate and graduation rate (Sousa, 2015).

Graduation rate is the percentage of a school’s first-time, first-year undergraduate students who complete their program within 150% of the published time for the program (FAFSA, 2016). For instance, a qualifying student enrolled in a four-year program must graduate within 6 years. Graduation rates answer two questions for an incoming student: How many stay and how many get out with a degree? These are two very good questions every student should know about his/her institution of choice (Engelmyer, 2016). Graduation rates can serve as one measure of accountability and transparency for a college or institution. Low graduation rates may mean students aren’t getting the support they deserve, whether that be financially or academically (Sealy-Morris, 2016).

Both graduation rates and retention rates serve as key statistics when incoming students prepare to choose colleges or universities. This makes the data important to students and academic institutions. The College Scorecard is a website sponsored by the Department of Education that makes it easier for students to search information about colleges or universities. Every school that allows students to accept Federal financial aid must submit data to the college scorecard. The project is designed to increase transparency, putting power in the hands of students and families to compare colleges and see how well schools are preparing their students to succeed (Dept. of Education). It is not specified whether the dataset has been used for machine learning analysis previously. A link to the dataset is provided in the appendix [m]. The public raw data file on this website is procured for analysis of the graduation and retention rates (Data.gov, 2016).

Problem Statement

What university-level features predict the presence of a strong retention rate and/or graduation rate? The raw datasets from the College Scorecard website was downloaded to answer this question. Possible features will be extracted from the dataset and checked for missing values to create a finalized dataset to perform the analysis. The finalized dataset includes all feature variables, target variables (graduation rates and retention rates), and their observations (colleges or universities) without any missing values. Different metrics and calculations will be performed on the dataset.

Metrics

In the visualization section, the correlation coefficient and coefficient of determination will be used on a plot of the graduation rate versus the retention rate. Based on descriptive hypothesis, the graduation rates will increase as the retention rates increase. The correlation coefficient is a measure of linear correlation between graduation rates and retention rates. The coefficient of determination is the proportion of the variance in the graduation rates that is predictable from the retention rates.

For comparing the performance of the models, the coefficient of determination, or score, will compare the performance of the regression models. This metric will determine which model illustrates more of the variance in the target variables that is predictable from the feature variables.

To quantify the “goodness” of a clustering quantity, the mean silhouette coefficient from each data point will provide a scoring method of a given clustering. The number of clusters that provides the highest silhouette coefficient will be used in the clustering method.

Analysis

Data Exploration

The raw datasets zip file was downloaded from the College Scoreboard website (Data.gov, 2016). The dataset from the most recent year, 2013 is used as the raw data file. The data dictionary and full data documentation files helped interpret the variables and figure out which variables are potential feature and target variables. The original raw dataset has 7804 institutions or observations and 1729 variables. Explanations of variables can be found in the documentation included with the dataset.

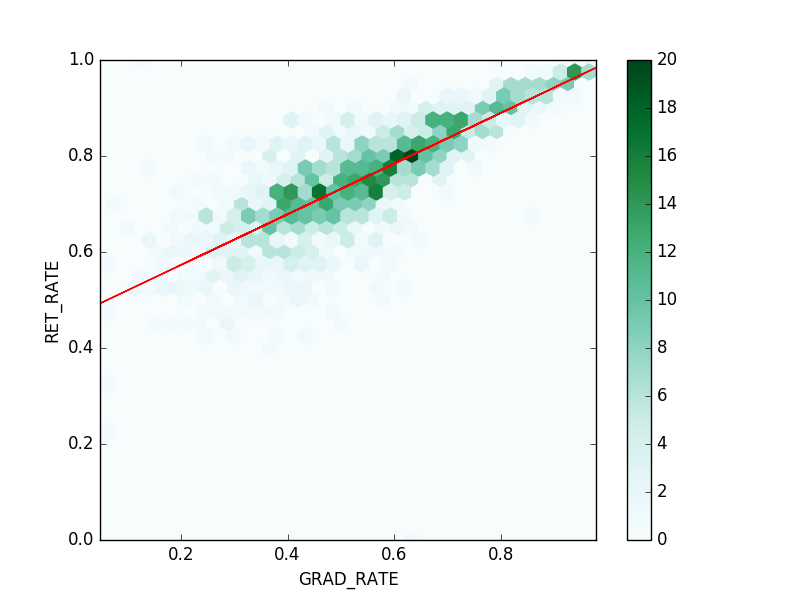
The observations in the dataset are post-secondary institutions of learning in the United States. The Student Right to Know Act requires all schools that award federal student aid dollars to supply the federal government with certain information (Engelmyer, 2016). Data is also obtained from federal financial aid data and tax information data. Each institution is listed along with many demographic variables indicating place and setting. There is information about each school such as the URL, Carnegie classifications, and revenue; also there are variables detailing earnings and employment of post-secondary students over spans of time. Many variables also detail the academics and admissions of each school, for example the SAT scores and percentiles. The other categories represented are the school costs for students, student body demographics, financial aid data, completion metrics, repayment data, and outcomes for Title IV students. Examples of data points are given in the appendix [n]. Only 11 of the 1729 variables are shown.

There are many abnormalities and characteristics about the input data that needs to be addressed. There are 1729 variables; most of them do not contain information helpful to analysis. Many of the useful variables have missing values; the institutions with missing values in feature or target variables will be removed from the analysis. Some of the elements are protected for privacy purposes by the National Student Loan Data System (NSLDS) and Treasury. Those variables are shown as Privacy Supressed and must be deleted.

In the final processed dataset, the mean and median of graduation rate are 0.542370 and 0.533600, respectively. The mean and median of retention rate are 0.753840 and 0.757200, respectively. Mean and medians were taken to observe if the graduation rate and retention rate are possibly normally distribution. Judging by the proximity of the means and medians, there is a possibility the rates are normally distributed. 1246 of the institutions are predominantly bachelor degree granting and 1075 institutions have graduate degree programs. More than half of the colleges and universities are private, nonprofit. The southeast region is represented the most in the final dataset and most of the institutions are in large cities. The average admission rates and SAT score are 0.6414 and 1058, respectively. The average cost of attendance is 32,177 dollars. The detailed processing of the dataset is in the Methodology section.

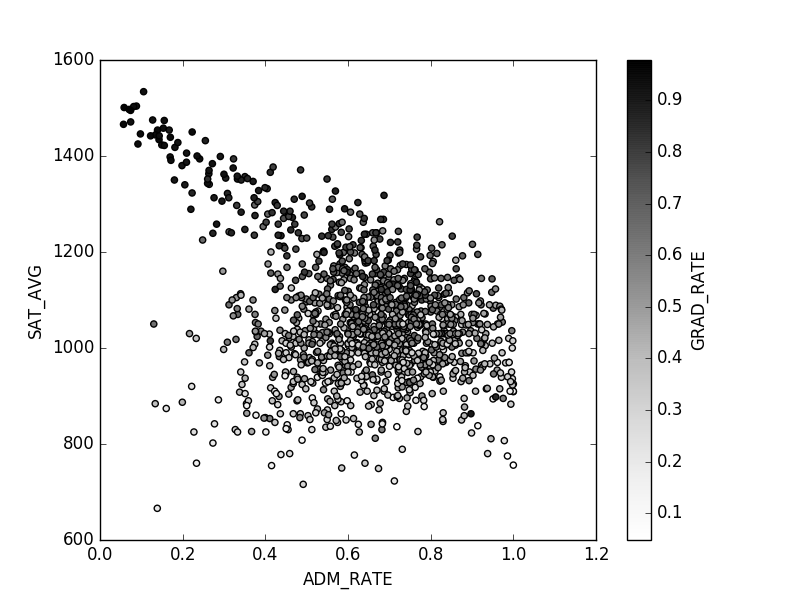
Exploratory Visualization

Figure 1: Hexagon Bin of Graduation Rate Versus Retention Rate



Heat bars indicate number of institutions, or data points, within each hexagon.

From observing the hexagon bin, graduation rates and retention rates appear to have correlation. As the graduation rates increase, the retention rates tend to increase. The correlation coefficient is 0.79, which is high. 63% of the variance in graduation rates is also predictable from retention rates. Therefore, university level features that predict strong graduation rates can possibly predict strong retention rates, and vice versa.



Heat Index indicates the graduation rate at each institution

Figure 2: Admission Rate Versus SAT Average plotted with a Graduation Rate Heat Index

Above is a scatter plot of admissions rate versus SAT average. Naturally, the harder it is to get into a school, the higher the SAT average. Graduation rate is visible through the colors of the graph. Graduation rate is easier to depict than retention rate because it has a higher variance. Through observing the above visualization, there appears to be multiple clusters. Clustering analysis will be performed as a consequence of the above graph.

Algorithms and Techniques

The goal is to be able to use the features to predict graduation and retention rates. This is a supervised learning problem, since the features is used to approximate the target variables. This is also a regression problem, since both target variables are continuous. Since there are a variety of variable types for the features, it is most advantageous if the supervised learner requires little data preparation. It needs to be able to handle numerical and categorical data. Therefore, Decision Tree Regressor is used. Not only does it fit all of the above requirements, it is simple to understand and interpret. Two supervised Decision Tree Regression learners are fitted onto the data predicting each target variable. To protect from overfitting, a K-Fold cross validation technique is used on each learner. Folds will be made and tested against the model trained from the other folds. A score will be procured from each learner and the average will be computed further.

Based off explanatory visualizations, the data may produce clusters that could make predicting graduation rates and retention rates more effective. The first step will be to observe if a Gaussian Mixture Model (GMM) clusters the data. The GMM is a soft clustering method; it assumes all data points are derived from a given number of clusters that have Gaussian distributions. It then uses the Expectation-Maximization algorithm to place each cluster’s center in the most optimal position that maximizes the probability of the data, given the cluster centers. The advantages of a GMM are that it can work with any distribution, the clusters can overlap, and it has a monotonically non-decreasing likelihood. Specifically, a Dirichlet Process GMM will be used. The Dirichlet process effectively fits only enough clusters to accurately describe the data. It operates best under an unknown amount of clusters.

The dataset will be separated based on cluster designation. There are 14 features in the final dataset. The curse of dimensionality states the amount of data needed for retaining statistical power grows exponentially with features. The next step is to use a Principal Component Analysis (PCA) on the original features and each cluster’s features. A PCA is resourceful when there are many features, but it is most likely a smaller number of features are actually driving the patterns. The PCA creates composite features that try to minimize information loss when projected onto them. Principal components containing the most information are always ahead of the other components, and the information does not overlap. Using a PCA will reduce the dimensions or features and possibly understand some driving patterns in the data. Models with varying amounts of principal components will be trained onto each cluster; once the variance and information begins to plateau, the number of additional components will stop. Each cluster will have its own PCA analysis; dissimilar clusters may have dissimilar principal components.

Once the PCA is done to the features and the features on each cluster, supervised Decision Tree regression learners will be trained onto each cluster using its principal components to predict each target variable. To protect from overfitting, a K-fold cross validation technique will be applied to each regression. For each cluster, the data will be split into a training set and a testing set. Folds will be made and tested against the model trained from the other folds. scores will be taken from the regression analysis.

Benchmark

The benchmark for measuring performance is the coefficient of determination, or score. The score from the initial decision tree regressors fitted onto the model without clustering and/or principal components will be compared to the final decision tree regressors fitted onto the model with clustering and/or principal components.

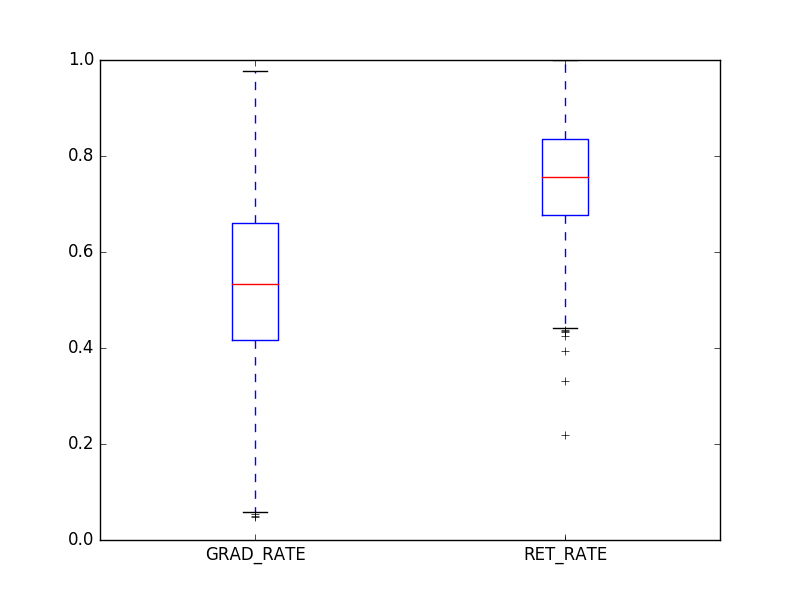
Methodology

Data Preprocessing

Processing the data corresponding to the features was done to ensure analysis could be performed on each variable. Colleges or universities that were not the main campus were not included. This prevents universities like Kaplan or DeVry from being counted more than once. Institutions whose predominant degree awarded was not classified or entirely graduate were not included. Observations that did not grant degrees were not included. US Service Schools were not included. Schools with no locale were not included. Institutions that did not have a documented admission rate were removed. Schools that did not have an SAT equivalent average were removed. Colleges or universities not currently operating were removed. Observations without an average cost of attendance or percent of faculty that is full-time were not included. Institutions that did not report a percentage of undergraduates above 25 were removed. Schools that did not report percentages of types of parent education were not included.

The data was also processed for the target variables. Two variables had to be merged for both retention rates and graduation rates: the retention rates for four year institutions and less than four year institutions and the graduation rates for four year institutions and less than four year institutions [c]. Institutions that did not have a retention rate and graduation rate were removed from the final analysis.

Figure 3: Box Plot of the Outliers for Graduation Rate and Retention Rate



Outliers from graduation rate and retention rate were also removed. Outliers were classified as points outside the inner quartiles by a range of 1.5 times the interquartile range. The outliers are from institutions who have uncharacteristically low graduation or retention rates. The final dataset has 1278 institutions with 14 features and 2 target variables.

Implementation

For graduation rate and retention rate, a decision tree regressor was fitted onto the dataset for nine folds using the K-fold cross validation [d]. The average coefficient of determinations for graduation rate and retention rate are 0.6022 and 0.4770, respectively.

A Dirichlet process Gaussian mixture model was applied to the dataset to determine how many clusters are in the dataset [e]. Based on the highest silhouette score of 0.5012, the optimal number of clusters is 2. The data was split into two clusters. Cluster 1 has 763 institutions and cluster 2 has 515 institutions.

In each cluster, a decision tree regressor was fitted onto the cluster, predicting the target variables, and tested using the K-fold cross validation technique. Seven folds were used on the first cluster [h]. The average coefficient of determinations for graduation rate and retention rate are .5261 and 0.4339, respectively. Five folds were used on the second cluster [h]. The average coefficient of determinations for graduation rate and retention rate are 0.5506 and 0.3840, respectively.

A PCA was done on each cluster’s features to reduce the dimensionality. In the first cluster, almost 99.99 percent of the variance is explained in one component [f]. In the second cluster, more than 99.95 percent of the variance is explained in one component. The features of the data were reduced to one component in each cluster and a decision tree regressor was fitted onto each cluster, predicting the target variables, and tested using the K-fold cross validation. Seven folds were used on the first cluster [i]. The average coefficient of determinations for graduation rate and retention rate are 0.1227 and -0.1107, respectively. Five folds were used on the second cluster [i]. The average coefficient of determinations for graduation rate and retention rate are -0.1889 and -0.7809.

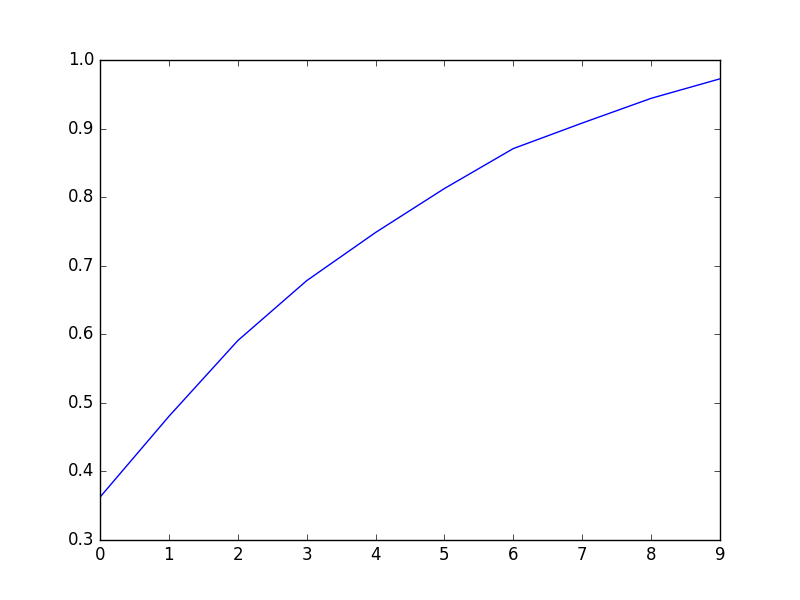
Refinement

Tuning of all models was performed in the hopes of finding better models that predict the target variables. A grid search was performed on decision tree regressors for the following parameters: maximum depth, minimum samples to split, and the minimum sample leaves. The maximum depth parameter denotes the maximum depth of the tree; it allows the tree to only grow to the specified depth. The default is to split the leaves until all leaves are pure; the grid searched ranged from 1 to 21. The minimum samples to split parameter is the minimum number of samples required to split an internal node. The default is 2; the grid search ranged from 2 to 10. The minimum sample leaves parameter is the minimum samples required to be at a leaf node. The default is 1; the grid search ranged from 1 to 10 (Scikit, 2014).

A K-fold cross validation technique was used for each combination of parameters. The final parameters that gave the highest average score were recorded and delivered in the appendix. Nine folds were used. In the decision tree regressors fitted onto the dataset for both target variables, there was a significant increase in the scores. For graduation rate, the coefficient of determination increased from 0.6022 to 0.8137; for retention rate, the coefficient of determination is now 0.7621, previously recorded as 0.4770 [j].

A PCA analysis was done on the data’s features to possibly reduce the dimensionality while preserving the predictors of variance in the target variables. First, each feature was normalized to have a mean of 0 and a standard deviation of 1. This means that each institution’s data (with j the institution and k the feature), the data is transformed by:

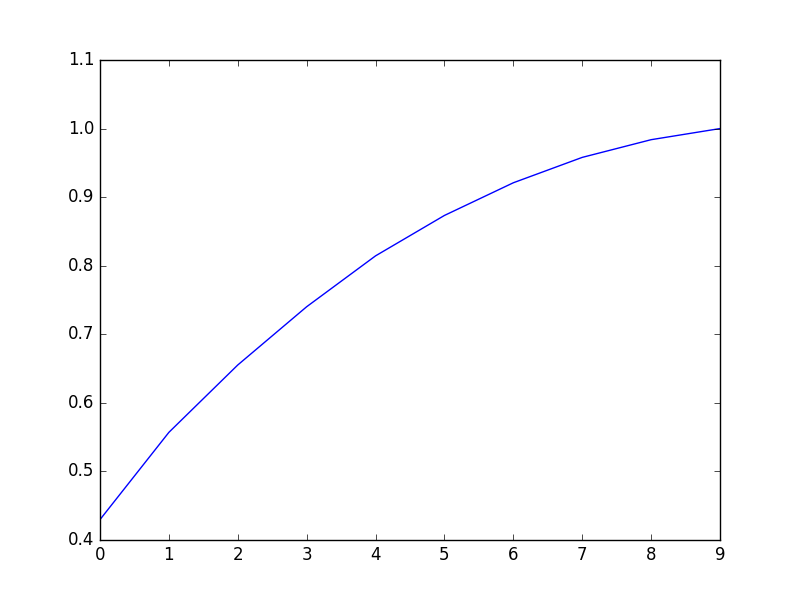
where is the average of feature k in the dataset and is the standard deviation of feature k in the dataset.

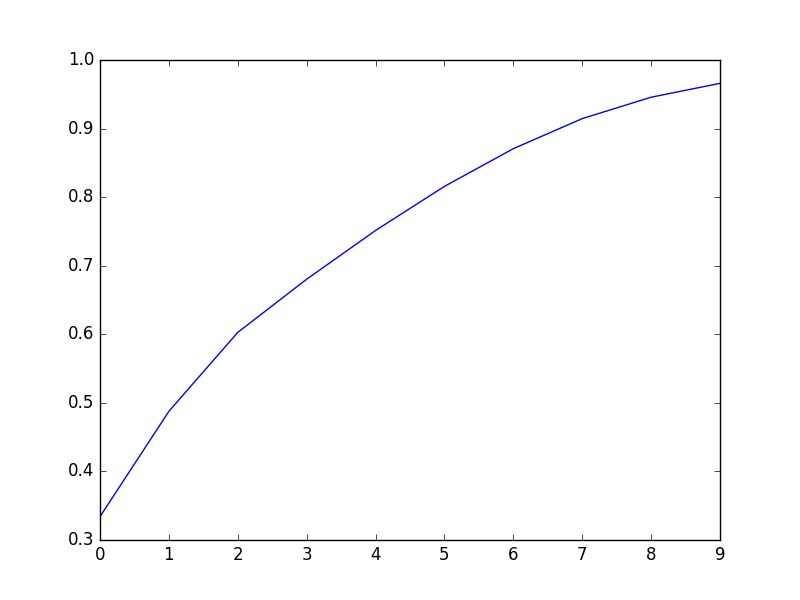
This graph shows the cumulative sum of the explained variance at each component of the PCA for 10 components. Unlike some PCA, there is not a distinct component that is observed to have distinctly more explained variance than its latter. Therefore, the chosen number of components is the least amount of components to explain at least 90% of the variance; in this case, it is 8. The data is transformed to the 8 components and decision tree regressors were fitted onto the components for each target variable. Each decision tree regressor was tuned through the aforementioned grid search to find the maximum through the optimal parameters. For graduation rate and retention rate, the coefficient of determinations were 0.7551 and 0.6729, respectively, using the principal components.

The grid search on the clustered dataset also provided better results than the initial un-tuned decision tree regressors on the clusters. Seven folds were used for cluster 1 and five for cluster 2. For example, after tuning decision tree regressors for both target variables onto cluster 1, the score for graduation rate moved from .5261 to .8295; retention rate moved from 0.4339 to 0.7265 [k]. For cluster 2, the graduation rate increased from 0.5506 to 0.8385, and retention rate increased from 0.3840 to 0.7662 [l].

A PCA analysis was done on each cluster’s features to reduce the dimensionality. This time, each feature for each cluster was first normalized to have a mean of zero and a deviation of 1. This means that for each institution’s data (with i the cluster number, j the institution, and k the feature number), the data is transformed by:

where is the average of feature k in cluster i and is the standard deviation of feature k in cluster i.

This graph shows the cumulative sum of the explained variance at each component of the PCA for 10 components in cluster 1. Unlike some PCA, there is not a distinct component that is observed to have distinctly more explained variance than its latter. Therefore, the chosen number of components is the least amount of components to explain at least 90% of the variance; in this case, it is 7. The data is transformed to the 7 components and decision tree regressors were fitted onto the components for each target variable. Each decision tree regressor was tuned through the aforementioned grid search on cluster 1 to find the maximum through the optimal parameters. For graduation rate and retention rate in cluster 1, the coefficient of determinations were 0.7965 and 0.6243, respectively, using the principal components.



This graph shows the cumulative sum of the explained variance at each component of the PCA for 10 components in cluster 2. Unlike some PCA, there is not a distinct component that is observed to have distinctly more explained variance than its latter. Therefore, the chosen number of components is the least amount of components to explain at least 90% of the variance; in this case, it is 8. The data is transformed to the 8 components and decision tree regressors were fitted onto the components for each target variable. Each decision tree regressor was tuned through the aforementioned grid search on cluster 2 to find the maximum through the optimal parameters. For graduation rate and retention rate in cluster 2, the coefficient of determinations were 0.7159 and 0.5444, respectively, using the principal components.

Results

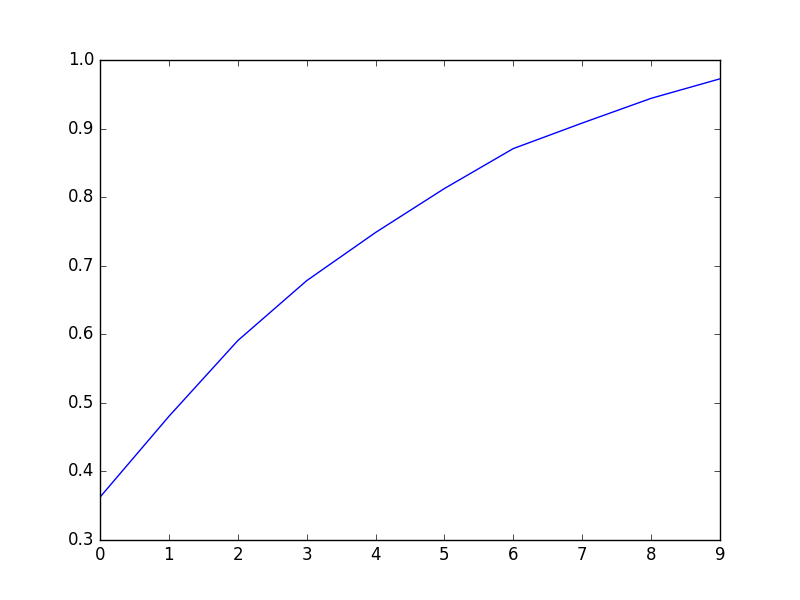
Model Evaluation and Validation

The refined models show substantial improvement upon the initial models. There are four complete decision tree regressor models for each target variable. After the refinement process, decision tree regressors were placed on the dataset, components of the dataset, clusters, and components of the clusters. For each target variable, the decision tree regressor did not show substantial increase in with effective clustering. For graduation rate, the refined decision tree regressor on the dataset produced an average score of 81.37%, while the clusters produced scores of 82.95% and 83.85%. For retention rate, the refined decision tree regressor on the dataset produced an average score of 76.21%, while the clusters produced scores of 72.65% and 76.62%.

Table 1: Comparing and contrasting for all Decision Tree Regressors



Similar outcomes are also seen in the comparison of the coefficients of determination on the component analysis seen in the chart above right. This suggests relatively small differences, if any, in predicting the target variables between the full dataset and clustered data. Therefore, following the principle of Occam’s Razor, clustering is not necessary in the final model.

 To evaluate whether the features or components should be used from the full dataset, observing the principle components and features must be done. Of the fourteen features, five are multi-categorical features. This makes fitting components onto the feature variables difficult even with normalization.

This is observed and explained through graph of the cumulative variance explained by the components. In a PCA, the expectation is to see initial components explain high amounts of variance, then a significant drop in the amount of variance explained by the latter component. This is not observable in this PCA; there is no abrupt change in amount of explained variance because the data is not suitable for PCA. Therefore PCA is not useful for the final model.

The final models are the decision tree regressors of the target variables, fitted onto the features with refined parameters for optimal prediction performance [j]. Refinement of models has increased the score by more than 0.2 for both target variables through the grid search.

Solutions received when re-implementing the final models differ from the grid search solutions. Although the grid search has allowed the comparision of 1,890 different models with different parameters, scores are not based on the testing set. A 9-fold cross validation technique was used on the models three times with three different random states using the parameters found in the grid search. The solutions are shown below:

Table 2: 9-Fold Decision Tree Regressors Placed with 3 Different Random States



These solutions show that more than 72% of the variation in graduation rate, on average, can be predicted from the features in final model. Also, roughly 66% of the variation in retention rate can be predicted from the features in the final model. The final model is reasonable and aligns with solution expectations. The parameters of the models are appropriate and play a crucial role in the model’s capabilities [j]. The parameters allow the features to predict more than 10% extra variance in the graduation rate and almost 20% more variance in the retention rate.

The final models have been tested numerous times using K-fold cross validation. This allows the models to be tested on unseen data nine times. Based on the table above, the models are robust, with minimal discrepancies between the averages when using different random states. Small changes in the training data do not greatly affect the results. This makes the results highly trustable.

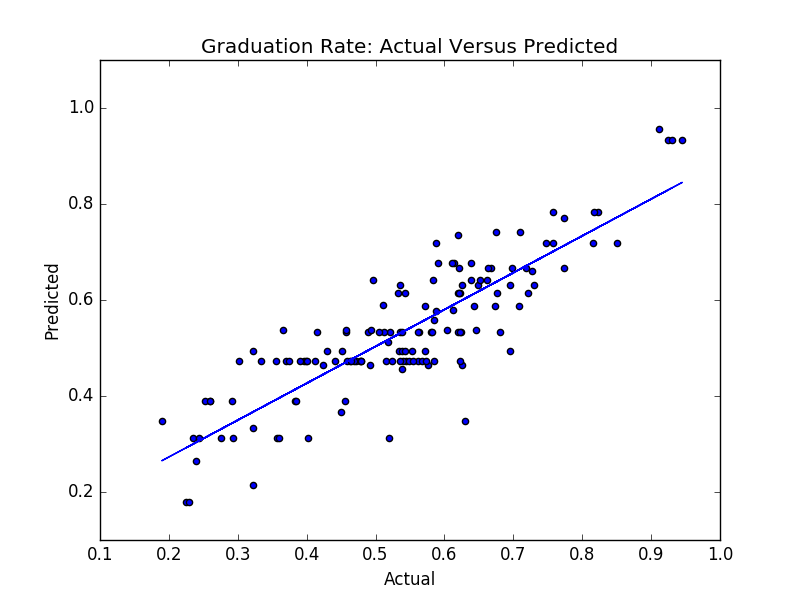
Justification

In the benchmark, I set out to compare different models to find the optimal model. The best model, based on the results, is the simplest model: decision tree regressors fitted onto the features [b] without clustering and principal components for the target variables [a]. Tuning of parameters allowed for more accuracy predicting the variance in the target variables [j].

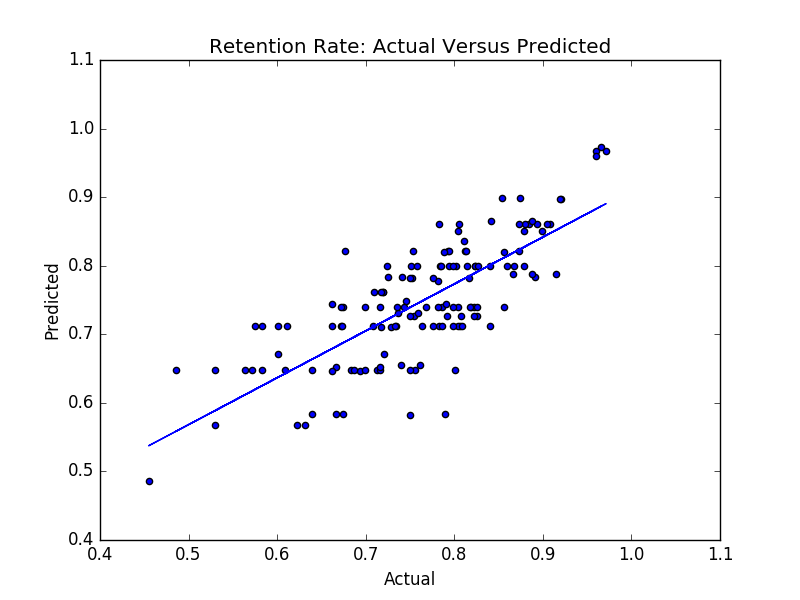
More than 72% of the variance in graduation rate is predictable through the features using the optimal parameters. Those parameters are a maximum depth of 6, a minimum samples split of 2, and a minimum sample leaf of 8. Roughly 66% of the variance in retention rate is predictable through the features using the optimal parameters. Those parameters are a maximum depth of 6, a minimum samples split of 2, and a minimum sample leaf of 6.

Conclusion

Free-Form Visualization

The dataset was randomly split into non-overlapping training and testing subsets covering and of the dataset, respectively. A decision tree regressor for graduation rate was fitted onto the training dataset. Predictions for graduation rate on the testing subset were made. The graph shown to the left is a scatter plot of the actual graduation rate of the testing subset versus the predicted graduation rate of the testing subset. A line was also regressed onto the graph. The regressed line fits the data well; the coefficient of determination is more than 72%.

A decision tree regressor for retention rate was also fitted onto the training dataset. Predictions for retention rate on the testing subset were made. The graph shown below is a scatter plot of the actual retention rate of the testing subset versus the predicted retention rate of the testing subset. A line was also regressed onto the graph. Although the coefficient of determination is just approximately 54%, the regressed line fits the data well.



Reflection

A student’s success is critical when he/she chooses to participate in post-secondary education. This education is not mandatory; adults spend a lot of time and take many risks when deciding to pursue a degree. Most leave with a burden of debt. Therefore, it is critical to know the likelihood of success at any institution.

Retention rate is a meaningful category to any future student. It lets you understand how much the university values its transitioning students in their first year. This project effectively uses many variables to predict the variability of retention rate. Four models were trained and refined to produce a high predictability of the variance in the retention rate. The final model for retention rate fits all of my expectations of the project.

Naturally, the graduation rate can’t be more than the retention rate. I predicted the variance of the retention rate to be more difficult to predict than the variance of the graduation rate, and it was. The main difference between retention rate and graduation rate is that graduation rate gives a macro-perspective of each student until their completion: the value a university places on its student throughout their studies. Four models were trained and refined to produce a high predictability of the variance in the graduation rate. The final model for graduation rate fits all of my expectations of the project.

The four complete models for graduation rate and retention rate were a decision tree regressor using the features, a decision tree regressor using the components, a decision tree regressor placed on each cluster using the features, and a decision tree regressor placed on each cluster using the components. When creating more complex models, the goal is to have improvement upon results. Based on the initial screening, it was determined there was a possibility of clustering. Modeling with clusters increases complexity, therefore the complex models only replace the simpler ones if it enhances results. Based on the results for graduation rate and retention rate, clustering did not improve results enough to warrant extra complexity. The two models with clustering were thus not the final model.

When refining the component analysis, it can be observed that some features have much higher values, thus much higher variance. For example, the average cost of attendance will have a substantially higher variance than the admission rate. This was solved through normalizing each feature; this way the variance fluctuations will not hinder the PCA. But, after solving this problem, it is still noticeable that the explained variance does not behave similar to a PCA. This problem is created by the multi-categorical features. As a result, the PCA was absolved, which left the simplest model as the final model for graduation rate and retention rate: the decision tree regressor using the features.

After refinement of the decision tree regressor to predict the retention rate, the features are able to predict roughly 66% of the variance in retention rate. After refinement of the decision tree regressor to predict the graduation rate, the features are able to predict more than 72% of the variance in graduation rate. These models should definitely be used in a general setting to show colleges probable controllable factors that would improve their graduation rates and retention rates. These models should also be used in a general setting for students to observe which colleges have key features that may potentially improve their chances of graduating.

There were many aspects of this project which were interesting. Most of the research hindered upon which models to implement. Documents on different types of PCA, for example, were read when the analysis did not produce wanted results. Even though the PCA ended in failure after the normalization process, many techniques were read and learned. The entire project was difficult. It took a lot of time, effort, and sacrifice.

Improvement

One aspect which could be improved are the features. As mentioned in the Project Overview, there are many factors that play a key role in the retention and graduation rates. Some considerable features which possibly would lead to an even further increase in predicting the variance of the target variables are housing options, dining services, financial aid shifts, rigorous course details, population of classes, or advising system details. This information was not gathered during the database’s creation.

I do not believe any further improvements could be made on the algorithms or techniques used in the project; if I felt there were, I would go back and use them. There were no algorithm or techniques researched that I did not eventually figure out how to implement. I would use my solution as a final benchmark and I believe a better solution exists, if some of the aforementioned features are procured when creating the next database.

Appendix

1. The two target variables(labels) and their specifications are:
2. graduation rate (GRAD\_RATE):
3. retention rate (RET\_RATE):
4. The fourteen feature variables(labels) are:
5. Predominant degree rewarding (PREDDEG): 1-predominantly certificate-degree granting; 2-predominantly associate’s-degree granting; 3-predominantly bachelor’s-degree granting
6. Highest degree awarded (HIGHDEG): 1-Certificate degree; 2-Associate degree; 3-Bachelor’s degree; 4-Graduate degree
7. Ownership (CONTROL): 1-Public; 2-Private nonprofit; 3-Private for-profit
8. Region of United States (region): 1-New England; 2-Mid East; 3-Great Lakes; 4-Plains; 5-Southeast; 6-Southwest; 7-Rocky Mountains; 8-Far West; 9-Outlying Areas
9. Location type (LOCALE): 11-Large city; 12-Midsize city; 13-Small city; 21-Large Suburb; 22-Midsize suburb; 23-Small suburb; 31-Fringe town; 32- Distant town; 33-Remote town; 41-Fringe rural; 42-Distant rural; 43-Remote rural
10. Admission rate (ADM\_RATE):
11. Average SAT equivalent (SAT\_AVG):
12. Average cost of attendance (COSTT4\_A):
13. Percent of faculty that is full-time (PFTFAC):
14. Percent of undergraduates above 25 (UG25abv):
15. Percent of first generation college students (PAR\_ED\_PCT\_1STGEN):

1. Percent whose parents highest level of ed. Middle School (PAR\_ED\_PCT\_MS):

1. Percent whose parents highest level of ed. High School (PAR\_ED\_PCT\_HS):

1. Percent whose parents highest level of ed. Post-Secondary (PAR\_ED\_PCT\_PS):

1. The two variables in the raw data for retention rates RET\_FT4 and RET\_FTL4 are the retention rates for four year institutions and retention rates for less than four year institutions, respectively. The RET\_RATE variable is created to merge the variables together.

The two variables in the raw data for graduation rates C150\_4 and C150\_L4 are the graduation rates for four year institutions and graduation rates for less than four year institutions, respectively. The GRAD\_RATE variable is created to merge the variables together.

1. Here is the table for the scores of the initial 9-Fold Decision Tree Regressor applied to both target variables:

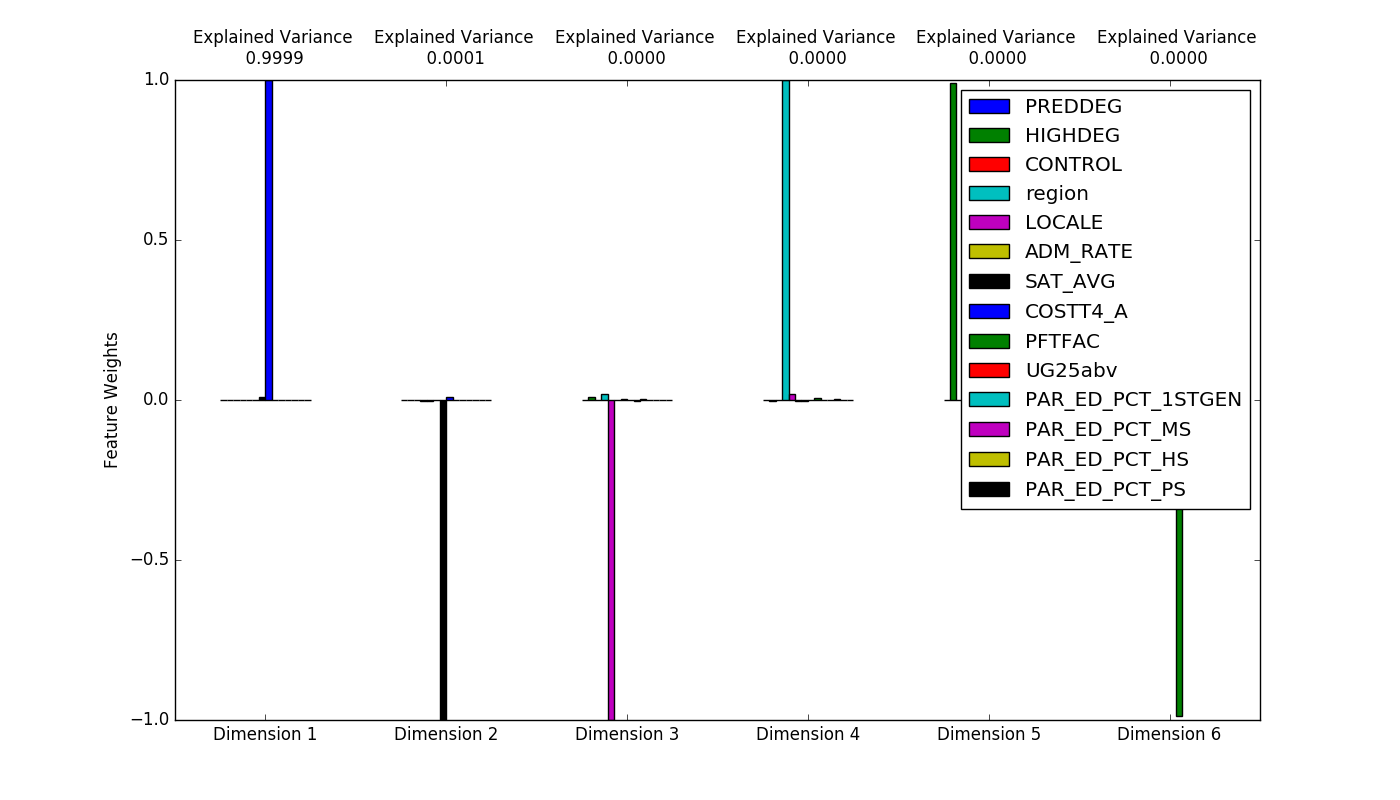
|  |  |  |
| --- | --- | --- |
|  | 9-Fold Decision Tree Regressor | |
| Graduation Rate | Retention Rate |
| 1 | 0.59418367 | 0.58268413 |
| 2 | 0.63746321 | 0.49345408 |
| 3 | 0.65875113 | 0.47510737 |
| 4 | 0.41228717 | 0.38143848 |
| 5 | 0.59694033 | 0.47565701 |
| 6 | 0.65862856 | 0.51768145 |
| 7 | 0.63179317 | 0.53337029 |
| 8 | 0.71688467 | 0.43801567 |
| 9 | 0.51283934 | 0.39555947 |
| Average | 0.602196805 | 0.476996439 |

1. These are the silhouette scores for 20 different cluster sizes:

|  |  |  |  |
| --- | --- | --- | --- |
| Silhouette Scores for Different Clusters | | | |
| 2 | 0.501175 | 12 | -0.08223 |
| 3 | 0.291302 | 13 | -0.17313 |
| 4 | 0.20563 | 14 | -0.1863 |
| 5 | 0.205721 | 15 | -0.18307 |
| 6 | 0.174167 | 16 | -0.20676 |
| 7 | 0.009391 | 17 | -0.20557 |
| 8 | 0.027655 | 18 | -0.23023 |
| 9 | -0.0256 | 19 | -0.24815 |
| 10 | -0.00548 | 20 | -0.26438 |
| 11 | -0.04699 | 21 | -0.26595 |

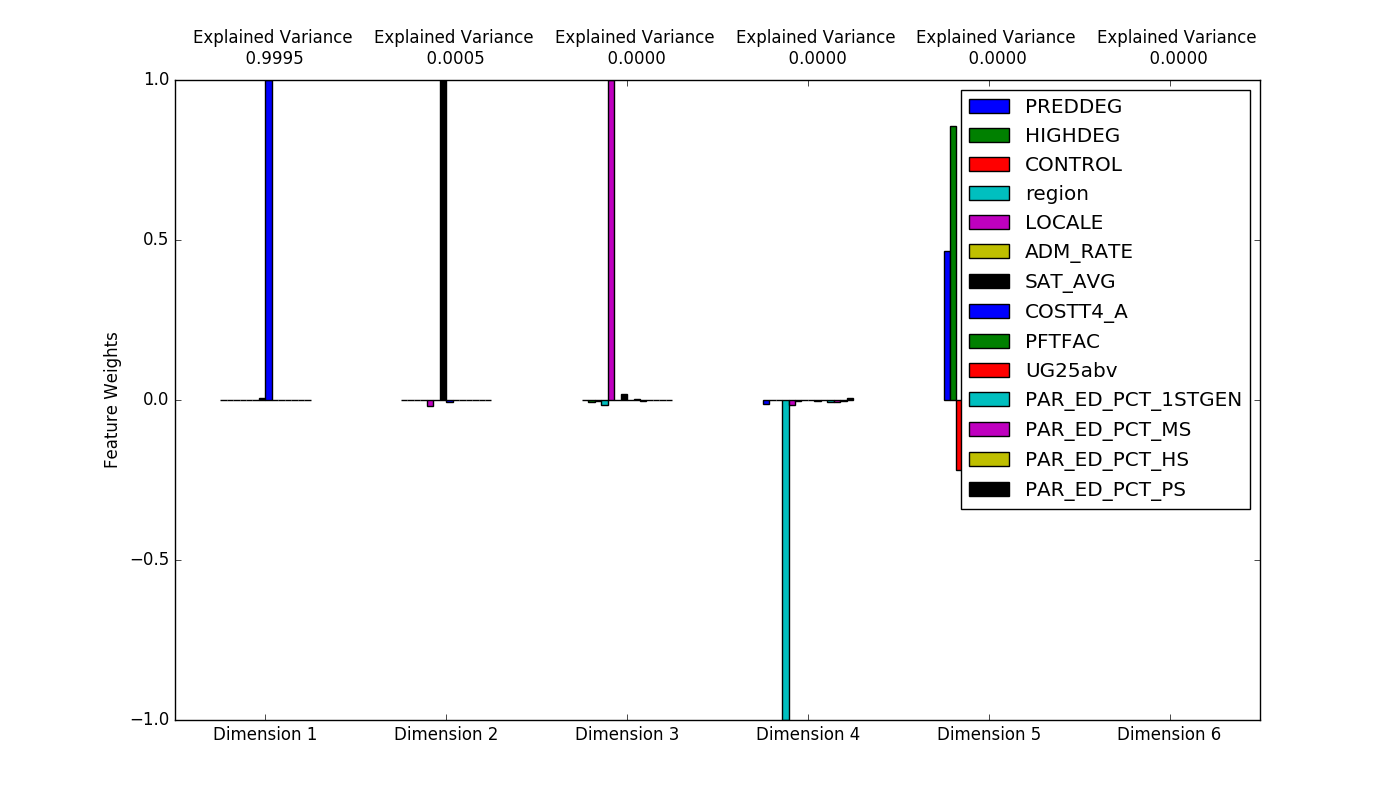
1. A table and plot of the explained variances for Cluster 1 in the PCA:

|  |  |
| --- | --- |
| PCs | Explained Variance |
|
| 1 | 9.999e-1 |
| 2 | 9.096e-5 |
| 3 | 8.560e-7 |
| 4 | 3.531e-8 |
| 5 | 1.479e-9 |
| 6 | 6.221e-10 |



1. A table and plot of the explained variances for Cluster 2 in the PCA:

|  |  |
| --- | --- |
| PCs | Explained Variance |
|
| 1 | 9.995e-1 |
| 2 | 4.541e-4 |
| 3 | 3.598e-6 |
| 4 | 1.568e-7 |
| 5 | 9.042e-9 |
| 6 | 2.211e-9 |



1. Here is the table for the scores of the 7-Fold Decision Tree Regressor applied to both target variables for Cluster 1 and the 5-Fold Decision Tree Regressor applied to both target variables for Cluster 2:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Cluster 1 | 7-Fold Decision Tree Regressor | |  |  |  |  |
| Graduation Rate | Retention Rate |  | Cluster 2 | 5-Fold Decision Tree Regressor | |
| 1 | 0.48909929 | 0.58535788 |  | Graduation Rate | Retention Rate |
| 2 | 0.61504394 | 0.45964725 |  | 1 | 0.50820159 | 0.5010814 |
| 3 | 0.521503 | 0.28448031 |  | 2 | 0.40618846 | 0.62171043 |
| 4 | 0.35134839 | 0.31162202 |  | 3 | 0.54833542 | 0.22274195 |
| 5 | 0.52618471 | 0.48442375 |  | 4 | 0.70164426 | 0.42147134 |
| 6 | 0.63279792 | 0.62281622 |  | 5 | 0.58881893 | 0.152979 |
| 7 | 0.54643418 | 0.28886603 |  | Average | 0.550637733 | 0.383996824 |
| Average | 0.526058777 | 0.433887636 |  |  |  |  |

1. Here is the table for the scores of the 7-Fold Decision Tree Regressor applied to both target variables on Cluster 1’s component and the 5-Fold Decision Tree Regressor applied to both target variables on Cluster 2’s component:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Cluster 1 PCA | 7-Fold Decision Tree Regressor | |  |  |  |  |
| Graduation Rate | Retention Rate |  | Cluster 2 PCA | 5-Fold Decision Tree Regressor | |
| 1 | 0.31155881 | 0.17308876 |  | Graduation Rate | Retention Rate |
| 2 | 0.35447499 | 0.25230293 |  | 1 | -0.1583703 | -0.71051193 |
| 3 | 0.16666315 | -0.33970223 |  | 2 | -0.19307675 | -0.67414804 |
| 4 | -0.18011306 | -0.30331437 |  | 3 | -0.06170919 | -0.7938103 |
| 5 | 0.29592702 | -0.17485488 |  | 4 | -0.17967272 | -0.55735118 |
| 6 | -0.03586246 | 0.01531856 |  | 5 | -0.3516758 | -1.16872756 |
| 7 | -0.05374022 | -0.39796904 |  | Average | -0.188900953 | -0.780909801 |
| Average | 0.122701174 | -0.110732897 |  |  |  |  |

1. Parameters and scores for the initial un-tuned Decision Tree Regressors and the initial tuned Decision Tree Regressors:



1. Parameters and scores for the un-tuned Decision Tree Regressors and the tuned Decision Tree Regressors for Cluster 1:



1. Parameters and scores for the un-tuned Decision Tree Regressors and the tuned Decision Tree Regressors for Cluster 2:



1. Link to the dataset: <https://catalog.data.gov/dataset/college-scorecard>
2. Examples of data points:



Bibliography

Data.gov. "Data Catalog." College Scorecard\_Raw\_Data.zip. N.p., n.d. Web. 22 July 2016.

Dept. of Education. "Data Documentation for College Scorecard." College Scorecard. US Department of Education, 2016. Web. 25 Aug. 2016.

Engelmyer, Lynell. "2 Key Statistics For Comparing Colleges: Graduation Rate and Retention Rate Explained - College Raptor Blog." College Raptor Blog. N.p., n.d. Web. 20 July 2016.

FAFSA. "What Are Graduation, Retention, and Transfer Rates?" What Are Graduation, Retention, and Transfer Rates? FAFSA, n.d. Web. 20 July 2016.

Scikit-learn Developers. "Sklearn.tree.DecisionTreeRegressor." Sklearn.tree.DecisionTreeRegressor — Scikit-learn 0.17.1 Documentation. N.p., 2014. Web. 09 Aug. 2016.

Sealy-Morris, Carrie. "Are College Graduation Rates Important?" Value Colleges. N.p., n.d. Web. 20 July 2016.

Sousa, Terianne. "Student Retention Is More Important Than Ever." Higher Ed Live. N.p., 09 Sept. 2015. Web. 20 July 2016.

Unigo. "What Are Freshman Retention Rates and Why Do They Matter? | Unigo." What Are Freshman Retention Rates and Why Do They Matter? | Unigo. N.p., n.d. Web. 20 July 2016.