Optimizing Student Institution Pairings

Bilal Ali
bilalali@vt.edu

Anna Aquino
nmap@vt.edu

Daniel Ballance
dmb2011@vt.edu

Abstract

This project provides data-driven insights into the relationships between key data points regarding institutions of higher learning and the outcomes students achieve after graduation. This analysis is focused on enabling college-bound high school graduates to make higher-quality decisions regarding their choice of college based on outcomes such as debt repayment rates, completion rates, and post-graduation earnings. These insights are achieved through regression-based machine learning algorithm analysis, yielding strong correlations with family income and high earnings and that, overall, obtaining an undergraduate degree bolstered earning potential later in life.

1. Introduction

The National Center for Education Statistics reports that there were 7,236 post-secondary education institutions within the United States, making potential students' institution selection process one that would be overwhelmed by choice [1]. Students, families, and other stakeholders within the educational arena need tools to help narrow the field of possible options. To refine the pool of possible institutions, it is required to analyze both data about the institution itself and the outcomes of its graduates. Such analysis will yield possibilities bounded by both the constraints of a given student's requirements of an institution and the restraints placed on students by institutions.

Given the large number of both institutions in the United States and the possible "variables" that are required to make an informed decision, a data-driven approach to solving this problem is highly appropriate. To attempt to manually wrangle all of this information is impractical, while solely relying on anecdotal information from friends and family can be biased and incomplete. Using a data-driven approach, machine learning algorithms can handle the sheer amount of information available and process it into a usable form. Such an approach enables a user to take advantage of the large amounts of available information while augmenting any other selection processes they choose to use (e.g., family recommendations, athletic affiliations, etc.).

The design of this problem must be tightly controlled to prevent its scope from expanding beyond the intent of this project. Overwhelmingly vast and complex answers exist to the simple question, "What is the best institution of higher learning?", along with equally vast and complex data. Therefore, the scope of this project shall be limited to a bachelor's degree focus using as generic-as-available data to keep applicability to a given student as wide as possible, while focusing on earnings-related outcomes for graduates.

2. Background/Related Work

To get a greater sense of understanding of the dataset, features were first explored through different visualizations. For instance, the feature for mean student earnings eight years post-graduation was explored in two different geographical representations. Figure 1 represents that data segmented by state, where as Figure 10 (can be found in the appendix) displays that data by city. This was a great aided in understanding the data within a national context.

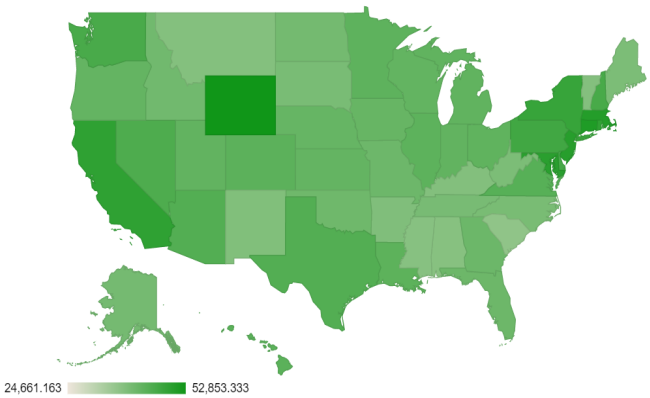

Figure 1. 8 Year Post Grad Alumni Mean Income Distribution

3. Approach

3.1 Scale of Data

To address this institution selection problem, the United States Department of Education's (DoE) College Scorecard from data.gov was selected for its robust dataset [2]. Table 1 provides the scale of the data available.

Table 1. Consolidated Dataset Statistics

Attribute	Value
Initial Number of Rows	124,699
Initial Number of Features	1,777
Post-Secondary Education Intuitions Represented	7,236

This raw dataset greatly exceeds the requirements for this project and provides data that is diverse and large enough to more than adequately address the stated problem. As enumerated in its data dictionary, the dataset contains institution classification; student demographic, financial, and generational information; and institution admissions and post-graduation data from 1996 through 2016 (with variations in year-to-year feature availability as data collection evolved over time). These features span across Boolean, numerical, and categorical data points.

3.2 Dataset Exploration

To understand the data available, initial dataset exploration began with its provided data dictionary. Within the dictionary, each of the dataset's 1,777 features were clearly defined and sorted into specific categories: root, school, academics, admissions, student, cost, aid, repayment, completion, and earnings. Further, each of these categories were applied to 7,236 unique institutions, including trade schools, associate's degree-focused community colleges, and full-fledged universities.

The dataset's challenges became evident in even reviewing the just dictionary's features: not all features were present for all years. The

2.33-gigabyte dataset itself was divided into yearly comma-separated values (CSVs) where feature presence was not uniform. Additionally, the institutions included were not present across all years; schools would appear and disappear across years as schools were founded, closed, or experienced fluctuating accreditation statuses.

Investigation into the yearly data subsets yielded additional challenges: in addition to expected missing (or "NaN") data points, significant portions of data were suppressed for privacy reasons. Reviewing the available data, the majority of features provided real-valued data vice classification or otherwise encoded data. Many of the features are variations of a common theme; for example, a given institution's alumni student loan repayment rates are provided in one-, three-, five-, and seven-year post-graduation cohorts. So while a single "type" of information is available in the dataset, it is often presented in several different ways. Figure 2 provides an example of data availability by cohort for a given feature.

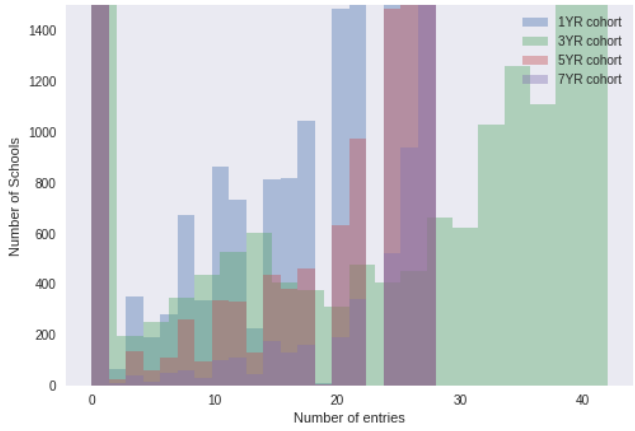

Figure 2. Data Entry Distribution

4. Dataset Preprocessing

Dataset preparation was conducted in five distinct phases: ingestion, inspection, reduction, aggregation, and imputation.

4.1 Ingestion

As the dataset was split across 20 individual CSVs in addition to the data dictionary itself, the first preprocessing requirement was to ingest the entirety of the dataset into a Jupyter Notebook. This was accomplished by developing three key methods: a helper method to load an individual CSV, a primary method to load all CSV files leveraging the helper method within a for loop, and an additional similar method to load the contents of the data dictionary. Both the dataset itself and the data dictionary were saved as pandas dataframes for follow-on analysis.

4.2 Inspection

During initial program development, inspection of the created dataframes revealed that data from years 1999-00 and onward contained at least over 1,000 features, suggesting a significant change in the DoE's data collection methodology; years 1996-97, 1997-98, and 1998-1999 contained fewer than 900 features and also represented the oldest data within the entire dataset, thus of least value for analysis. Additionally, multiple CSVs contained completely, or high amounts of, NaN (to include privacy suppressed data) features or rows.

4.3 Reduction

Having established that:

1) Specific CSVs are of far less value for analysis (i.e., the early-year CSVs);
2) The scope of this project would be limited to institutions that primarily awarded bachelor's degrees (i.e., the most common situation for a recent college-bound high school graduate); and
3) The dataset as a whole had multiple NaN value-related problems to solve,

The next step in preprocessing was to reduce the dataset with these issues in mind. This was executed in a three-stage approach: combining the many CSVs into a single file for analysis, retaining only predominantly bachelor's degree-awarding institutions, and reducing the overall number of NaN values within the dataset.

4.3.1 Combining CSVs into a Singular Dataset

To translate the 20 individual CSVs into a manageable dataset, the aforementioned loader methods were modified to read each CSV from local storage, evaluate its number of features (ignoring CSVs with fewer than 900 features), and add it to a single pandas dataframe. When concatenating this new data with the existing dataframe, an additional "YEAR" feature was added to retain from which specific CSV the data was provided. Once all data was added to the dataframe, it was retained in memory and exported to local storage as a single CSV for ease of ingestion for follow-on analysis over the period of the project.

4.3.2 Retaining Predominantly Bachelor's Degree-Awarding Institutions

Exploring the data dictionary revealed that each year's CSV contained a feature that labeled each institution by its predominantly-awarded degree type: certifications, associate's degrees, bachelor's degrees, or postgraduate degrees (i.e., master's degrees and beyond). This already-encoded data point enabled rapid dataset manipulation to retain only institutions that the DoE identified as predominantly bachelor's degree awarding, thus directly enabling the identification of institutions within the project's scope. In its implementation, this discrimination reduced the number of institutions within the dataset by 41.6-percent (75,628), from 181,591 to 105,963.

4.3.3 NaN Value Reduction

To address the overall number of NaN values, the loader methods were modified to leverage pandas' built-in dataframe creation parameters to drop all rows and features that were completely NaN. While effective at "clear cutting" NaN values from the dataset, it did not address discovered cases where a single institution (reflected in a single row) or single feature had only a few valid data points with the vast majority being NaN.

To handle these specific low-value cases, "NaN threshold" values were developed during dataset exploration to determine the most effective percentage of actual data an institution or feature must contain to be retained in the dataset. Knowing this would not eliminate all NaN values within the dataset, it would handle the vast majority, leaving any remainder to be addressed through follow-on manipulation and processing.

After the compression/aggregation step detailed below, features were removed if NaN values exceeded 5-percent of the total data points. In its implementation, this discrimination reduced the number of features within the dataset by 13.8-percent (238), from

1,726 to 1,531. Institutions were removed if NaN values exceeded 30-percent of the number of features within the dataset to further reduce the number NaN values. In its implementation, this discrimination reduced the number of institutions within the dataset by 12.3-percent (3,727), from 30,335 to 26,608 (Note: this large number of institutions is due to one institution having many instantiations, up to one for each yearly CSV in the raw dataset).

4.4 Aggregation

With the dataset largely free of unwanted data and NaN values, the dataset still contained multiple "duplicate" institution entries corresponding to each year's data for that institution. To adequately scope the size of this dataset to the requirements of this project, the dataset required compressing multi-year institutional data into a single value for each feature. This would enable follow-on analysis that would capture the general nature of an institution with respect to a given feature, thus falling with the scope of developing and using generalized institutional metrics. This aggregated data was then saved as the dataset to be used for analysis.

4.5 Imputation

Despite these multiple processing steps, the compressed dataset still retained a small number of NaN values. Knowing that most machine learning algorithms cannot tolerate any missing data points, missing values were imputed, using scikit-learn's Imputer, with the mean for the continuous variables and the median for the categorical variables. This resulted in the best representative sample for each institution vice imputing across all institutions. The later would have resulted in less-accurate imputations. This action resulted in a 100-percent complete dataset ready for analysis. Additionally, the resulting dataset was exported to local storage as a CSV for ease of follow-on analysis (submitted with this report).

5. Data Analysis

5.1 Suitability of Algorithm Types

With a solidified analytical dataset prepared, the suitability and appropriateness of various machine learning algorithms could be considered. The most likely to be successful and useful models were initially determined to be supervised and regression-based vice classification-based models. This is due to the fact the vast majority of data is highly diverse, real-valued data. For example, post-graduation earnings data could theoretically have over 100,000 unique values within that single feature alone. With that in mind, the majority of machine learning work focused on regression-based supervised models; however, specific tests scenarios were developed to exercise both unsupervised (i.e., k-means) and classification-based supervised (i.e., Naive Bayes) models.

5.2 Maintaining the Data Dictionary

Early in the data preprocessing and analysis phases, it was quickly determined that a customized and up-to-date data dictionary would be required to handle the multitude of tests, changes, and tweaks to come. This was especially true was each available feature in the dataset was mapped to either non-predictive, predictive, or target data "bins" and further sorted and reduced to key predictive and target features used by the employed machine learning algorithms.

To maintain a data dictionary that reflected the contents of the dataset and could readily provide useful groupings of features for analysis, a modified version of the data dictionary was developed and ingested alongside the dataset itself (submitted with this report). New methods were developed to update the contents of the data dictionary on-demand against the features present in the dataset, ensuring that the two were constantly aligned throughout analysis.

6. Algorithm Analysis and Findings

6.1 Unsupervised Analysis: k-Means

In the analytical process, the first step was to conduct k-means cluster analysis to determine if there were patterns in the data that would enable better or more precise supervised learning decisions in subsequent analysis. In preparation from clustering, 6 target features and 6 predictive feature pairs (12 predictive features in total) were selected based on common sense relationships. Those features are enumerated in Table 2:

Table 2. k-means Target/Predictive Features Pairings

Target Feature	Predictive Features
First-time, full-time four-year institution student completion rate (C100_4)	1. Average incoming student SAT score 2. Median student debt at graduation
First-time, full-time four-year institution student transfer rate (TRANS_4)	1. Institution admissions rate 2. Average student age at entry
Students not working 10 years after enrollment (COUNT_NWNE_P10)	1. Average debt of non-graduates 2. Midpoint of ACT cumulative score
Worker median earnings 10 years after enrollment (MD_EARN_WNE_P10)	1. Average incoming student SAT score 2. Institution admissions rate
Seven-year graduate loan repayment rate (COMP_RPY_7YR_RT)	1. Median student debt at graduation 2. Percentage of student who took loans
Seven-year non-graduate loan repayment rate (NONCOM_RPY_7YR_RT)	1. Median student debt at graduation 2. Percentage of student who took loans

In preparation for this clustering, the target data required an additional encoding step to "classify" its data into discrete groups that enabled clustering. Across all target features, five discrete "classes" were developed (valued 0-4) based on the range of values in each feature. These ranges are reported in the k-means Implementation notebook (submitted with this report).

In analyzing average student Standard Aptitude Test (SAT) scores against median student debt with school completion rate classifications, it becomes immediately apparent that the does not lend itself to k-means clustering, as seen in Figure 3 below; the classifications are deeply intermixed, which is a known struggle for clustering algorithms.

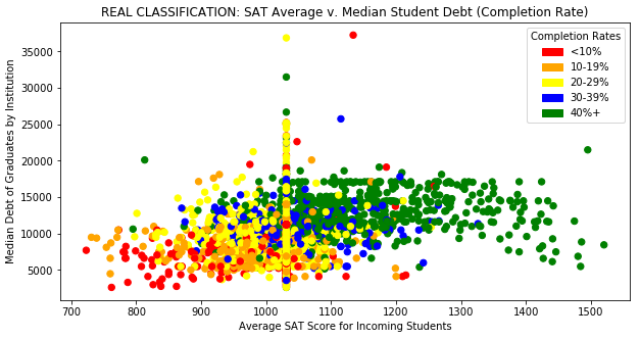

Figure 3. REAL CLASSIFICATION: SAT Average v. Median Student Debt (Completion Rate)

The ill-fit of such intermixed data in a clustering algorithm is readily apparent when reviewing the output of the k-means algorithm; it attempts to identify proper clusters within the dataset, but is unable to derive any meaningful insights with the provided data. The results of clustering the above data is provided in Figure 4 below. Additional data visualization pairings are available in the Appendix.

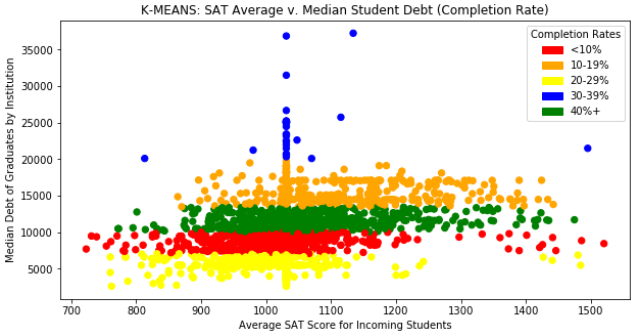

Figure 4. K-MEANS: SAT Average v. Median Student Debt (Completion Rate).

6.2 Supervised Analysis: Multilayer Perceptron (MLP)

To analyze the success of neural networks, MLP was applied against the dataset. The primary advantage of MLP over other supervised learning models is that it is able to distinguish data that is not linearly separable, which could lead to better successes with this specific data.

As with k-means, additional preprocessing was required to implement MLP. For consistency the same encoding methodology was used for target data, reducing all entries to a 0-4 value. Further, the predictive data required manipulation: as MLP requires input data to be scaled, the values were scaled using scikit-learn's StandardScaler (see MLP Implementation notebook for scaling and encoding details, submitted with this report).

For each target feature, 5-fold cross validation was conducted against a MLP classifier using rectifier linear units ("relu") as the activation function and (100, 100, 100, 100, 100) for hidden layers. A "None" random state was utilized. The results of the algorithm are provided in the table below (the raw results of this analysis have been submitted with this report).

Table 3. MLP Average Results by Target Feature

Target Feature	Precision	Recall	F1-Score
C100_4	0.56	0.54	0.55
TRANS_4	0.68	0.73	0.70
COUNT_NWNE_P10	0.72	0.72	0.71
MD_EARN_WNE_P10	0.70	0.67	0.67
COMP_RPY_7YR_RT	0.70	0.69	0.69
NONCOM_RPT_7YR_RT	0.81	0.82	0.81

As seen in the table, the MLP classifier provided accuracy scores ranging from 0.55 (C100_4) to 0.81 (NONCOM_RPY_7YR_RT). When compared to the input data and number of data points within each class, there appears to be a direct relationship between the uneven distribution of data points within classes and accuracy of the classification. For example, in a randomly selected fold from the C100_4 analysis, the largest and smallest support values differed by only 121, while in NONCOM_RPY_7YR_RT, that difference was 376. Further investigation into this behavior could warrant a different encoding scheme or further optimization of the MLP classifier.

6.3 Supervised Analysis: Naïve Bayes

In further classification-based analysis, Naive Bayes classification was tested. The target feature detailed if the alumni of a given institution were earning above or below the average income eight years after graduating. Since there were hundreds of features, the number of predictive features was reduced. This reduction was based on a correlation matrix, detailing the relationships between features (see Figure 5 below), developed during dataset exploration. This matrix yielded the features that were most impactful to the target feature; the remaining predictive features were discarded for this analysis.

The remaining data proved interesting: institutions with higher math standardized test scores produced alumni who had a higher mean income. A possible reason for this phenomena is that these students may have been seeking degrees in higher paying fields (e.g., STEM-based professions). Understanding the possible impact of selection bias in this analysis, this result would make for a great topic for follow-on investigation.

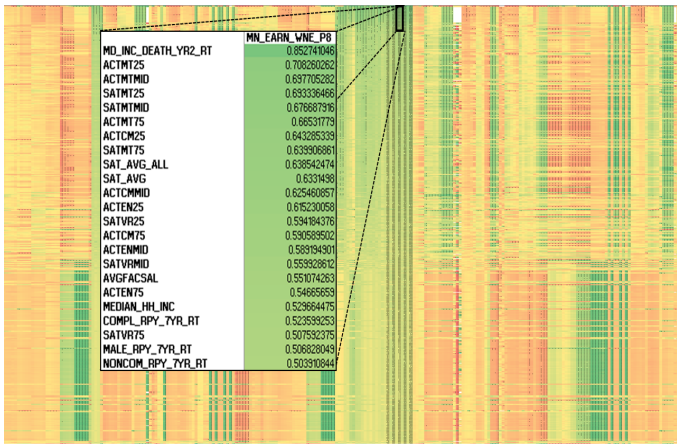

Figure 5. Post Grad Earnings Correlation Matrix

6.4 Regression Models

The goal to predict the continuous values of college graduate mean earnings using the plethora of features lends itself to a supervised learning problem. In addition to using the features to approximate the targeted earnings variables, it is also a regression problem, since the target variables are continuous.

To select a prediction model, various regression models were trained, tested, and compared. Grid Search was implemented to determine the hyperparameters for each of the models that give the least prediction error. The performance measure used to compare the various prediction models was Root Mean Squared Error (RMSE). RMSE is the squared error of the various input points from the provided model. Lower RMSE values suggest the points are closer to the model and thus a better fit, whereas a higher value suggests that the points are far apart from our model. RMSE scores were compared for each model relative to other model scores.

In addition to the model selection, numerous factors, to include multiple test ratio sizes, target earnings fields, and feature field selections were considered. Evaluations were run several times to narrow the comparative factors to three test sizes: 1) 20-percent, 2) 30-percent, and 3) 40-percent; against two target fields: 1) mean earnings of students working and not enrolled 10 years after entry, 2) mean earnings of students working and not enrolled 9 years after entry; and two sets of features: 1) 23 hand-selected features and 2) all features except for other earning target fields. The regressors above were executed within pipelines with various scalers, none of which outperformed the independent model and were removed from final evaluation. The 23 selected features are a small set of student representative attributes and combine tuition, student, and financial aid related features. These features and their descriptions may be found in Table 4:

Table 4. Predictive Features used in Regression Model Evaluation

Feature	Description
COMP_ORIG_YR2_RT	Percent completed within 2 years at original institution
COMP_ORIG_YR4_RT	Percent completed within 4 years at original institution
DEBT_MDN	The median original amount of the loan principal upon entering repayment
DEP_INC_AVG	Average family income of dependent students in real 2015 dollars.
ENRL_ORIG_YR2_RT	Percent still enrolled at original institution within 2 years
ENRL_ORIG_YR4_RT	Percent still enrolled at original institution within 4 years
GRAD_DEBT_MDN	The median debt for students who have completed
IND_INC_AVG	Average family income of independent students in real 2015 dollars.
OVERALL_YR2_N	Number of students in overall 2-year completion cohort
OVERALL_YR3_N	Number of students in overall 3-year completion cohort
OVERALL_YR4_N	Number of students in overall 4-year completion cohort
OVERALL_YR6_N	Number of students in overall 6-year completion cohort
OVERALL_YR8_N	Number of students in overall 8-year completion cohort
PAR_ED_PCT_1STGEN	Percentage first-generation students
PPTUG_EF	Share of undergraduate, degree-/certificate-seeking students who are part-time
TUITIONFEE_IN	In-state tuition and fees
TUITIONFEE_OUT	Out-of-state tuition and fees
UG25ABV	Percentage of undergraduates aged 25 and above
UGDS	Enrollment of undergraduate certificate/degree-seeking students
UGDS_NRA	Total share of enrollment of undergraduate degree-seeking students who are non-resident aliens
WDRAW_DEBT_MDN	The median debt for students who have not completed
WDRAW_ORIG_YR2_RT	Percent withdrawn from original institution within 2 years
WDRAW_ORIG_YR4_RT	Percent withdrawn from original institution within 4 years

One important note is that tree based models are not designed to work with very sparse features. All remaining nulls in the dataset were imputed using the median strategy.

The five regression models evaluated were: Linear Regression, Random Forest Regressor, Decision Tree Regressor, Ridge Regressor, and Support Vector Regression (SVR).

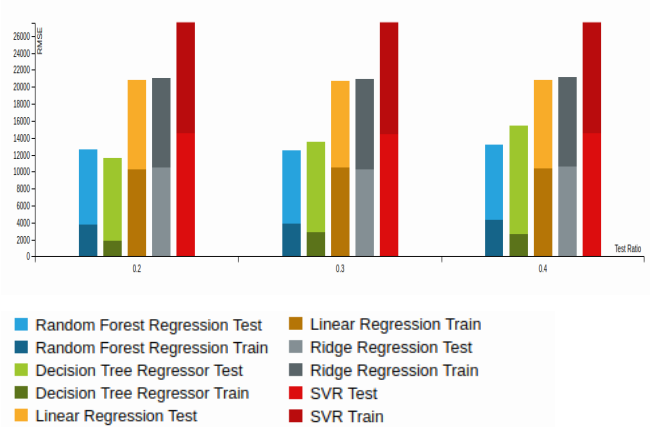

Figure 6. RMSE Training and Test Scores of 5 Regression Models

A subset of results is shown in Figure 6 above, where the train and test RMSE scores are depicted across each of the test ratio sizes for each model. It presents the best RMSE values from the different techniques used in the model evaluation using the 23 selected features and targeting the Mean earnings of students working and not enrolled 10 years after entry (i.e., MN_EARN_WNE_P10).

The Linear Regression, Ridge, and SVR regression errors were nearly equivocal between the training and test predictions, which indicates underfitting. These models indicate high bias and poor predictions. Given that Ridge regression is an extension of Linear regressions, but with coefficient constraints, it was expected for Ridge to outperform Linear and it did not. Linear Regression requires a linear relationship between the features and target variable. Based on the relationship plots in Figure 9: Relationships of features to Mean earnings of students working and not enrolled 10 years after entry (MN_EARN_WNE_P10), where there are few linear relationships to the target, the higher error rate makes sense.

Although the lowest error rate is seen in the Decision Tree regression model with the 0.2 test ratio, the RMSE for the training set is significantly lower than the testing RMSE, which indicates overfitting. The Decision Tree regression is built on nonlinear models. It uses entropy as a metric of determining whether certain features provided added information to the model, breaking the data down into smaller and smaller subsets.

The best predictive model is accurate on unobserved data. The lowest errors were obtained from Random Forest regression on average across the test ratios. Since Random Forest requires multiple iteration of Decision Trees, it makes sense that Random Forest should outperform Decision Trees. Random Forest Regression should reduce the overfitting, since it removes features that are mostly duplicated by other features. However, it can lend to a belief that a feature is a strong predictor than others in the same group deemed less important, when they actually are very closely related to the response variable. It was the best performer in predicting earnings in both of the feature size tests and in two of the three test ratio sizes.

The importance of features is assigned by the model, based on which features were used to make splits in the decision trees. The lower RMSE score is likely because Random Forest regression reduces variance by performing this feature selection. Figure 7 below visualizes the linear relationship between the Actual and Predicted values of the Random Forest regression using 0.2 test ratio with the 23 selected features. Unsurprisingly, the high training to low testing split of the data resulted in the lowest error rates since there is more data to train upon.

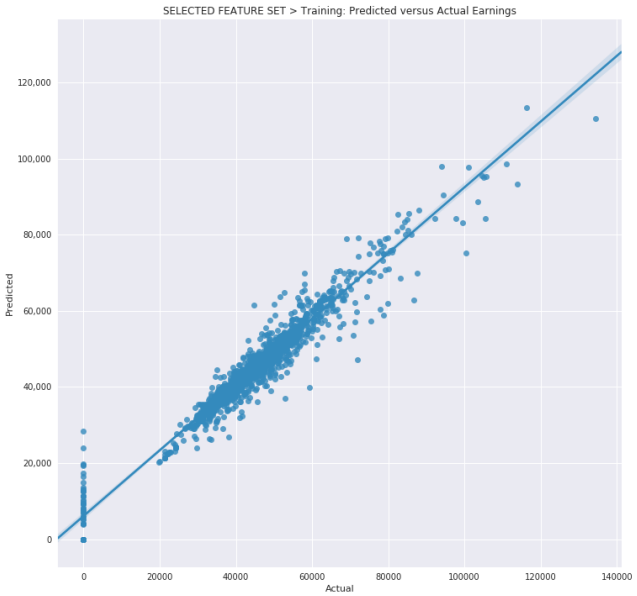

Figure 7. Random Forest Regression Training: Actual v. Predicted Mean Salary after 10 Years

The testing results in Figure 8: Random Forest Regression Testing: Actual v. Predicted Mean Salary after 10 years show that outliers, which Random Forest is sensitive to, may have influenced the prediction error rate.

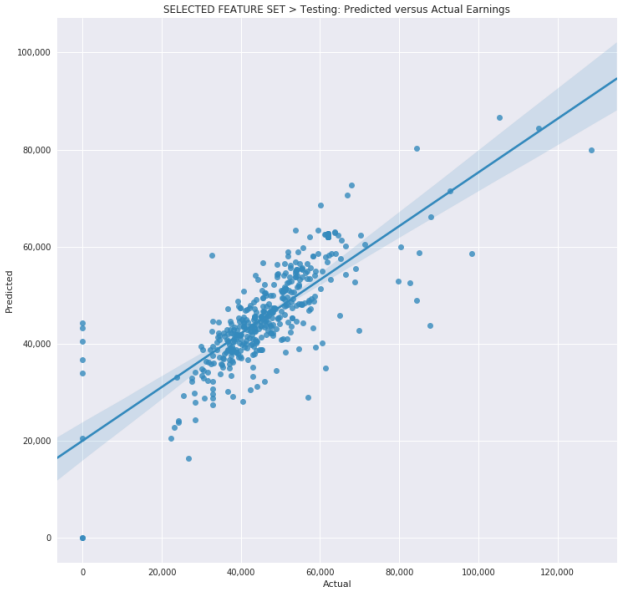

Figure 8. Random Forest Regression Testing: Actual v. Predicted Mean Salary after 10 years

Table 5. Top 10 Features by Random Forest Regression Importance

SCORE	FEATURE	DESCRIPTION
0.16	DEP_INC_AVG	Average family income of dependent students in real 2015 dollars
0.12	OVERALL_YR8_N	Number of students in overall 8-year completion cohort
0.09	DEBT_MDN	The median original amount of the loan principal upon entering repayment
0.06	WDRAW_DEBT_MDN	The median debt for students who have not completed
0.06	ENRL_ORIG_YR2_RT	Percent still enrolled at original institution within 2 years
0.05	TUITIONFEE_OUT	Out-of-state tuition and fees
0.05	TUITIONFEE_IN	In-state tuition and fees
0.04	WDRAW_ORIG_YR2_RT	Percent withdrawn from original institution within 2 years
0.04	OVERALL_YR6_N	Number of students in overall 6-year completion cohort
0.03	UGDS_NRA	Total share of enrollment of undergraduate degree-seeking students who are non-resident aliens

Table 5: Top 10 features by Random Forest Regression Importance shows the top 10 features by importance as determined by the Random Forest regression model using the 23 selected features. Table 7 in the appendix lists the top 15 features by Random Forest Regression importance using the full feature set. In both results, the repayment features are expected, since the ability to repay a loan should be highly dependent on the payers' post graduation income. The students' family income level were differentiating factors within the repayment features. Interestingly, the most important repayment feature in the full features set was differentiated by gender, the five-year repayment rate for males. Unsurprisingly, the admissions features, such as SAT and ACT scores, were also ranked high to moderate features.

The goal to predict the continuous values of college graduate mean earnings using the plethora of features lends itself to a supervised learning problem. In addition to using the features to approximate the targeted earnings variables, it is also a regression problem, since the target variables are continuous. The anticipated strong predictors included repayment, entrance test scores, and familial income factors, whereas the debt factors were unexpectedly less influential.

7. Conclusion

This analysis used a data-driven approach via machine learning to aid in a general college-bound high school graduate's undergraduate education institution selection process. Intentionally focused on "big picture" metrics in service to a widely-applicable result, this analysis utilized a large, highly-diverse dataset from American institutions to deliver an ultimately regression-based results to serve as a general decision-making aid.

This project required extensive data manipulation to prepare it for analysis. Such manipulation was the primary challenge in conducting analysis; the size, scope, and diversity of the dataset required tightly confining the target analysis and in-depth analysis of available data to differentiate useful features from information better suited for a different stated problem.

College Scorecard data may serve to inform prospective students, analyzing the data is a challenge due to the high sparsity. Modeling with the entire feature space was not possible and imputation alone poses issues with data integrity. Data preprocessing was tailored accordingly to facilitate a populated, but not skewed data set. There is an abundance of potential in modeling this data set due to its breadth and scope; attributes of the data that also influenced us to narrow our scope of analysis.

A diverse collection of machine learning algorithms was used, spanning classification- and regression-based as well as unsupervised algorithms. Ultimately, regression-based algorithms proved to be best suited for the dataset; classification-based algorithms would require extensive re-encoding of target information. Unsupervised clustering also did not perform well as the data did not present strong potential classes in plotted groupings.

Overall, the research revealed that the average family income positively correlates to the median earnings of graduate in either public private institutions. Expectedly, the cohort completion rates also indicate a high importance to mean earnings, aligning with studies that show that obtaining a bachelor's degree can increase lifetime earnings considerably relative to not attaining a degree.

Current methods of decision making based on college admissions information and published college rankings could be augmented with analytical methods along with the data from the College Scorecard data set to better tailor college selection to student attributes.

Interesting future steps would include employing methods such as LASSO regression to help identify parameters affecting success rates and expand beyond earnings as a indicator of success. Engineering new features, such as 'value' comparing the tuition cost to staff earnings, may also contribute to greater model accuracy and insights.

8. References

[1] U.S. Department of Education, National Center for Education Statistics. (2016). *Digest of Education Statistics, 2015* (NCES 2016-014), Table 105.50

[2] "College Scorecard Data." College Scorecard, collegescorecard.ed.gov/data/.

[3] Carnevale, A., Rose, S., Chea, B. "The College Payoff: Education, Occupations, Lifetime Earnings." Georgetown University, Center on Education and the Workforce, August 2011. https://www2.ed.gov/policy/highered/reg/hearulemaking/2011/collegepayoff.pdf

9. Appendix

Table 6. Supporting Report Content

File(s)	Description
K-means Implementation.ipynb	Evaluate K-Means model
MLP Implementation.ipynb	Evaluate MLP model
Naive Bayes Implementation.ipynb	Evaluate Naive Bayes model
Regression_Models.ipynb	Evaluate Supervised Learning Regression Models
Processed_data/Project_Data_Dictionary.csv	Custom data dictionary
Processed_data/data_for_model_eval.csv	Output of preprocessing steps; used as initial input to model evaluations
Results/	Output from modelling steps
Visualizations/	Visual graph objects
Exploration/	Notebooks modeling our data exploration (load/read/encode)
Preprocess/	Notebooks that combine multiple years of data together. Our main preprocessing steps can be found here

Table 7. Top 15 Features from Full Feature Set by Random Forest Regression Importance

SCORE	Features	DESCRIPTION
0.04	MALE_RPY_5YR_RT	Five-year repayment rate for males
0.03	SATMT75	75th percentile of SAT scores at the institution (math)
0.03	RPY_5YR_N	Number of students in the 5-year repayment rate cohort
0.03	NONCOM_RPY_3YR_RT_SUPP	3-year repayment rate for non-completers, suppressed for n=30
0.03	NONCOM_RPY_3YR_N	Number of students in the 3-year repayment rate of non-completers cohort
0.03	NONCOM_RPY_1YR_N	Number of students in the 1-year repayment rate of non-completers cohort
0.03	MD_INC_RPY_1YR_N	Number of students in the 1-year repayment rate of middle-income (between $30,000 and $75,000 in nominal family income) students cohort
0.03	LO_INC_RPY_5YR_RT	Five-year repayment rate by family income ($0-30,000)
0.03	IND_RPY_5YR_N	Number of students in the 5-year repayment rate of independent students cohort
0.03	COMPL_RPY_5YR_RT	Five-year repayment rate for completers
0.03	C150_4	Completion rate for first-time, full-time students at four-year institutions (150% of expected time to completion)
0.02	SAT_AVG	Average SAT equivalent score of students admitted
0.02	RPY_7YR_RT	Fraction of repayment cohort who are not in default, and with loan balances that have declined seven years since entering repayment, excluding enrolled and military deferment from calculation. (rolling averages)
0.02	HI_INC_YR4_N	Number of high-income (above $75,000 in nominal family income) students in overall

		4-year completion cohort
0.02	HI_INC_RPY_3YR_RT	Three-year repayment rate by family income ($75,000+)
0.02	ACTCMMID	Midpoint of the ACT cumulative score
0.01	WDRAW_2YR_TRANS_YR6_RT	Percent who transferred to a 2-year institution and withdrew within 6 years
0.01	UNKN_2YR_TRANS_YR8_RT	Percent who transferred to a 2-year institution and whose status is unknown within 8 years
0.01	SATVR25	25th percentile of SAT scores at the institution (critical reading)
0.01	PELL_RPY_7YR_RT	Seven-year repayment rate for students who received a Pell grant while at the school
0.01	PELL_ENRL_ORIG_YR6_RT	Percent of students who received a Pell Grant at the institution and who were still enrolled at original institution within 6 years
0.01	NOTFIRSTGEN_RPY_7YR_RT	Seven-year repayment rate for students who are not first-generation

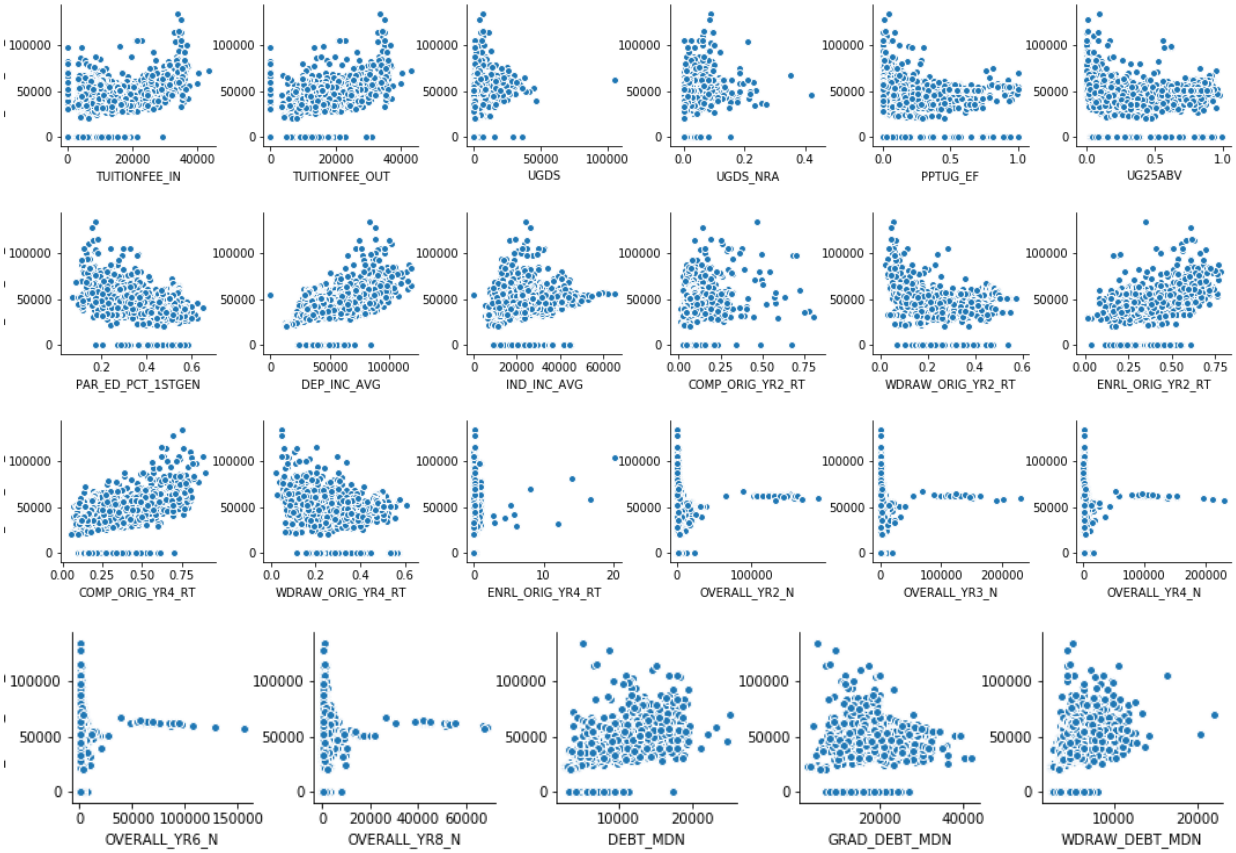

Figure 9. Relationships of features to Mean earnings of students working and not enrolled 10 years after entry (MN_EARN_WNE_P10)

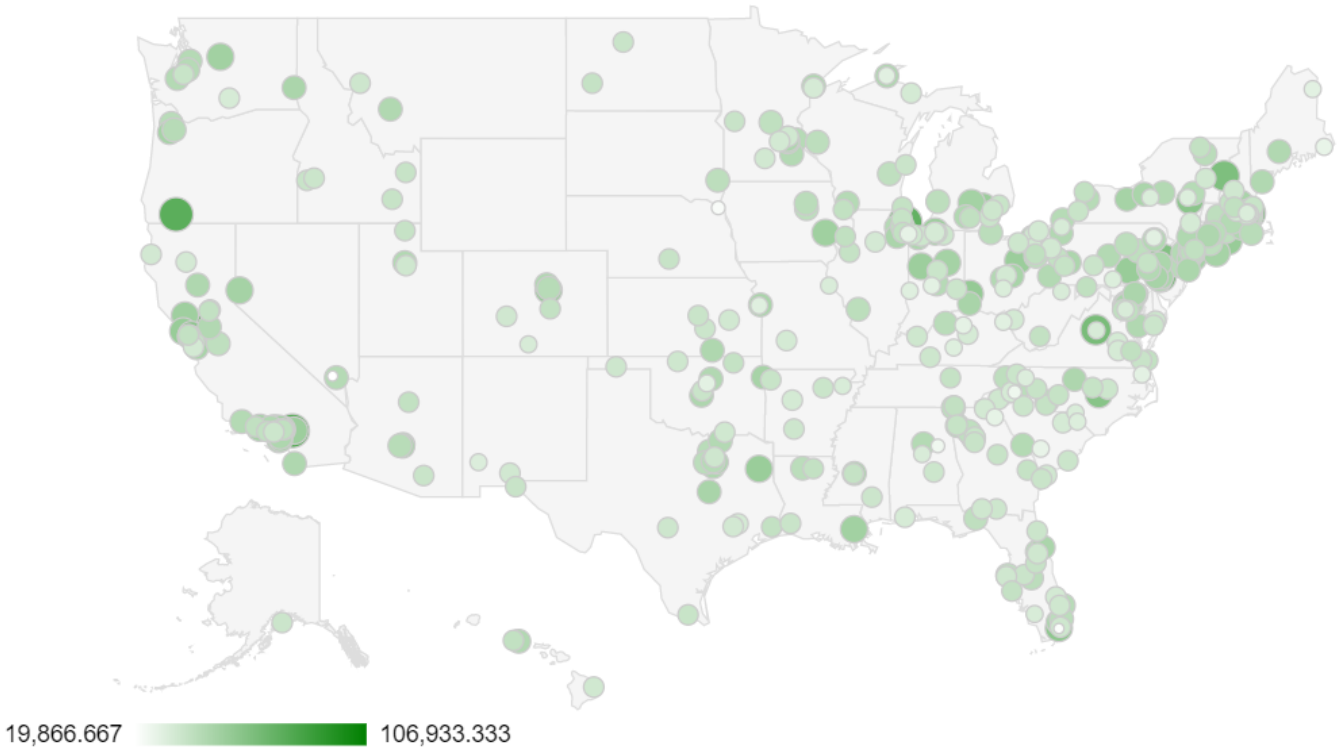

Figure 10. Mean Income Distribution of Alumni 8 Years Post Grad by City

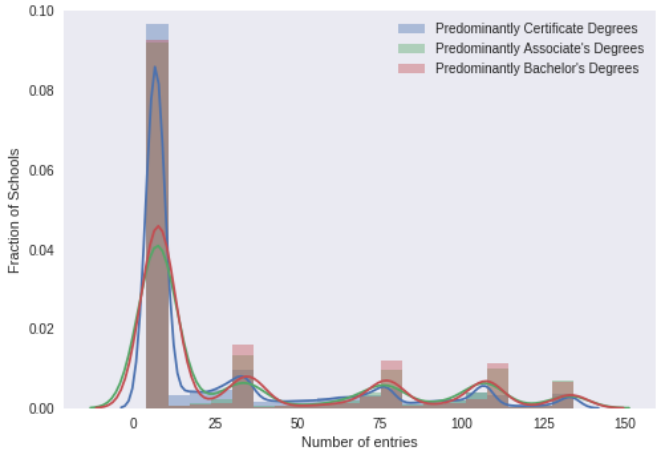

Figure 11. Repayment Data Histogram by Degree

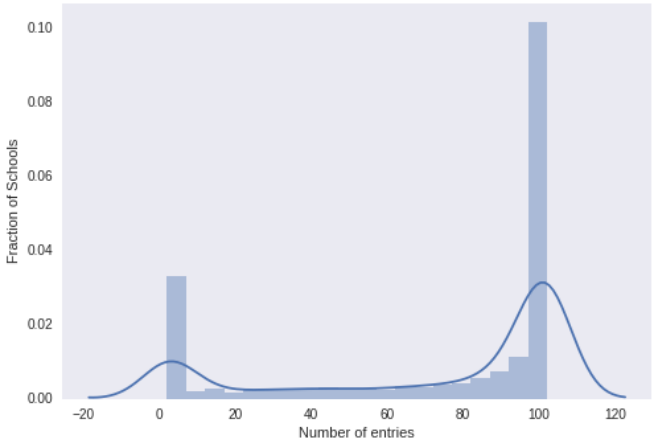

Figure 12. All Repayment Data Histogram

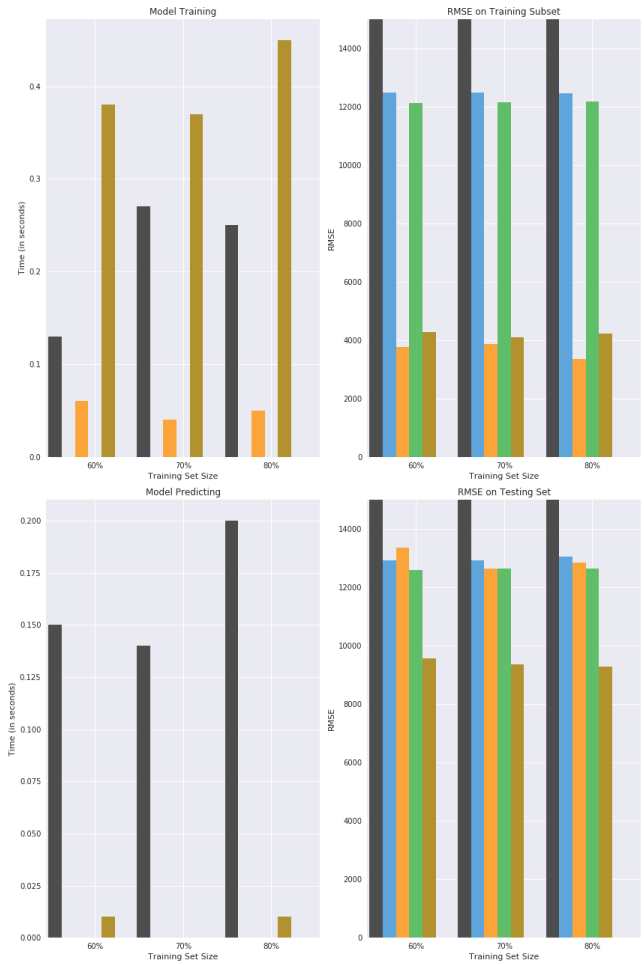

Figure 13. Supervised Model Mean 9 Years Earnings

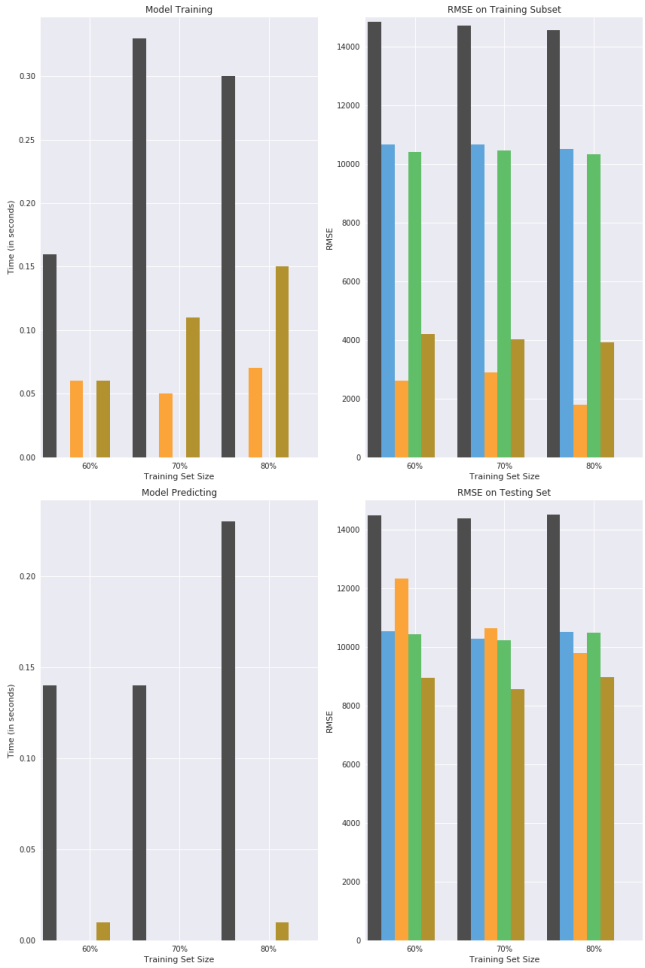

Figure 14. Supervised Model Mean 10 Years Earnings

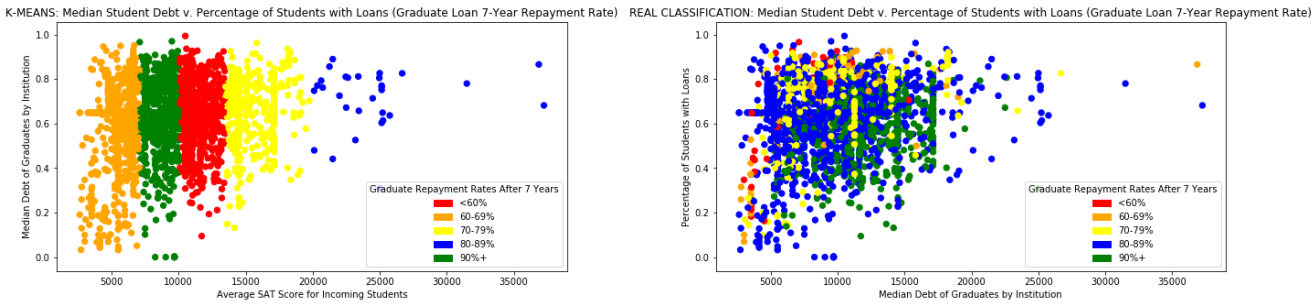

Figure 15. K-Means Graduation Repayment Rates after 7 Years VS Real Classification

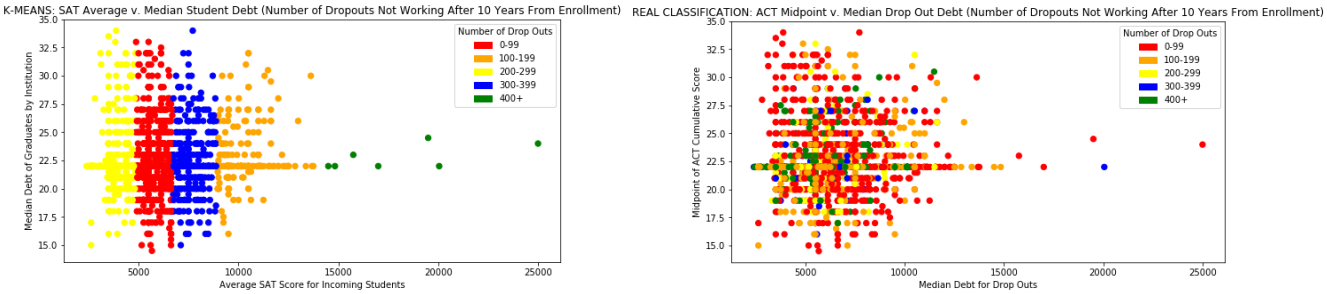

Figure 16. K-Means Number of Dropouts VS Real Classification

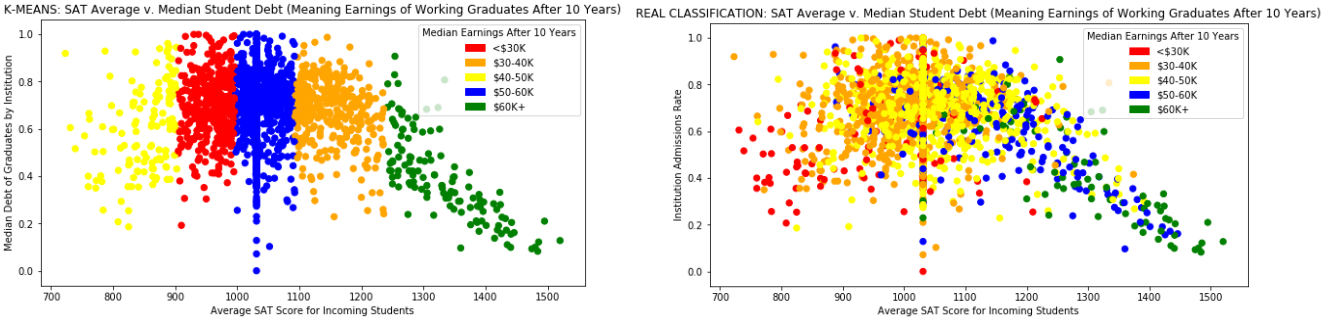

Figure 17. K-Means Median Earnings After 10 Years VS Real Classification

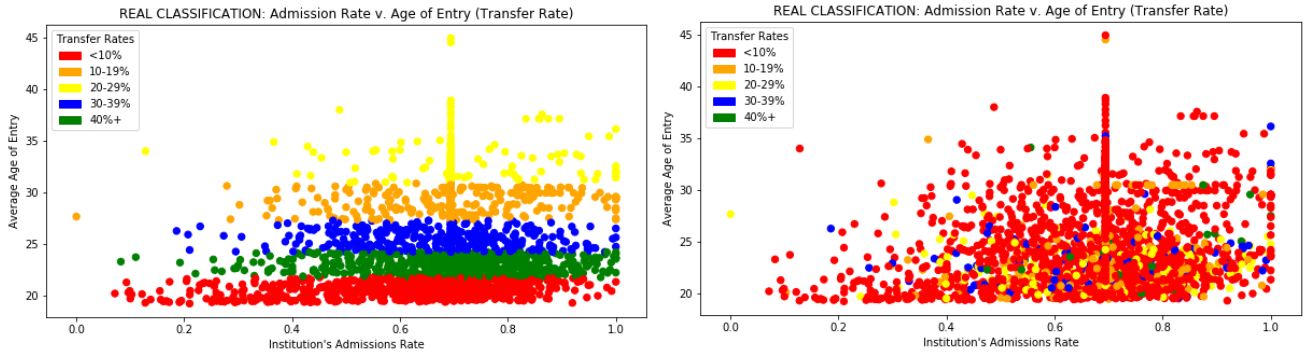

Figure 18. K-Means Transfer Rate VS Real Classification