Harnessing Regression Techniques to Predict College Quality

Jacob Mazurkiewicz

Department of Computer Science & Mathematics, Duquesne University

NE-COSC-423-523-CPMA-582-01 SP24: Machine Learning

March 17, 2024

Every year, the website *Collegescorecard.ed.gov* collects data points on U.S.-based Universities such as admission rates, Student-to-faculty ratio, and the median salary of graduating students (*Data Home: College Scorecard*). To improve Duquesne University and the university searching experience for college-seeking students, a regression model was developed to estimate the average "Score" of Universities based on controllable university features like faculty salary. The calculated "Score" is a combination of successful outcomes for students such as a high median salary after graduation, low debt, and the likelihood of securing employment. The "Score" is a proxy to estimate the overall quality of a university. An emphasis on controllable features ensured that any patterns detected in the analysis could be actionable for a university seeking to potentially their score and quality. For example, features such as geographic location were ignored because these cannot be changed by a university. Features like the number of programs offered, however, were included because these can be altered by the University. The resulting regression model can serve two beneficial purposes. Firstly, the University can ascertain which of the controllable variables most contributes to positive outcomes for students. This insight into informative and uninformative features can also allow Universities to allocate resources most effectively when optimizing for target metrics and making changes to better themselves. Secondly, this regression model can be used as a tool by students to predict their potential for successful outcomes based on the qualities of the University they are examining. In this way, they can make informed decisions about their future by measuring the controllable metrics of the University and their impact on their predicted success. This model, if deployed correctly, can serve both academic institutions and student populations in highlighting transparency in university decision outcomes on student prosperity.

Background:

The goal of this analysis is to use the College Scorecard data to determine actionable improvements not only for Duquesne University but also for student populations as a whole. Using machine learning techniques to uncover valuable information in the data can lead to powerful outcomes for students and colleges. The target variable of interest in this analysis is "Score", a derived metric. The score is derived from four variables. 'MD_EARN_WNE_P10' which measures the median earnings of students working and not enrolled 10 years after entry, 'COUNT_WNE_P10'which measures the number of students working and not enrolled 10 years after entry, 'GRAD_DEBT_MDN_SUPP' which measures the median debt of completers, suppressed for n=30, and 'UGDS' which measures the undergraduate enrollment. The score is calculated by taking the ratio of MD_EARN_WNE_P10 and GRAD_DEBT_MDN_SUPP and adding it to the ratio of UGDS and COUNT_WNE_P10 (This ratio is also weighted by half so that it is only half as important in calculating the final score compared to the salary to debt ratio). This "Score" variable captures both the earning-to-debt ratio of graduated students as well as their potential to secure employment. These were determined to be the key factors in assessing the quality of a given University. The "Score" is also normalized, with each component variable being standardized through Z-score standardization.

The Dataset used for this analysis and model is a combination of every year in the College Scorecard database for all U.S. Universities since 2006. The website describes their data as "Institution-level data files for 1996-97 through 2021-22 containing aggregate data for each institution. Includes information on institutional characteristics, enrollment, student aid, costs, and student outcomes" (*Data Home: College Scorecard*). Many of the fields for certain

universities have been marked with a privacy seal, preventing them from being included in any model or analysis. The data was also restricted to Universities offering bachelor's degree programs or higher to filter out lower-cost specialty colleges like beauty schools and police academies and produce recommendations more relevant to Duquesne University

Regression is a supervised machine-learning technique used to predict continuous values (Anwar, 2021). It uses features of the dataset, or characteristics, such as University Admission Rate or Undergraduate Student to Faculty Ratio, to predict a continuous target variable, such as Undergraduate Salary upon completion of a degree. By finding patterns between the features of data and the target variable, it can be used to predict future targets when fed unseen feature data. Regression requires labeled data, or data in which the target variable is clearly defined, giving it belonging to “supervised” learning (Anwar, 2021). Several different types of regression techniques, such as boosted decision trees and linear regression, were employed in this analysis to predict the target variables. Specifically, 4 model types were used: Random Forests, Linear Regression, ‘Extreme’ Boosted trees, and Support Vector Machines.

Random Forests are an ensemble of decision trees that use a majority voting system to predict an average value. It uses decision thresholds of the features to inform its decision (Rosidi, 2022). Random Forests combine several “weak learners” into a more powerful model by using dozens or hundreds of trees.

Linear Regression draws a mathematical relationship between the features and outcome variables and seeks to capture this relationship with a formula of features and weights for each feature plus an intercept value (Anwar, 2021). Linear Regression gives clear values for the weights of its features, making it a very transparent model.

XGBoost, or extreme gradient boosted trees, are an ensemble decision tree method that uses the residual error from previous trees to iteratively build a more accurate tree. XGBoost is a scalable implementation of the gradient-boosted trees algorithm. XGBoost differs from Random Forest in that Random Forests builds trees in parallel and takes a majority vote while XGBoost builds trees sequentially, using the gradient of the loss function to iteratively improve each tree, then taking the decision of all trees (*Introduction to boosted trees*).

Support Vector Machines seek to create a hyperplane that captures the pattern of data to make regression predictions. SVMs, for short, use an 'error boundary' which is a width on either side of a regression line or hyperplane. Only the data points within the width of the error boundary inform the line or hyperplane of the best fit, giving way to them "supporting" the line or hyperplane (Rosidi, 2022).

Materials & Methods:

The Materials used for this model are as follows. All the code was compiled within Visual Studio Code and logged in a GitHub repository. The Python Pandas library was used to read the data and manipulate it into data frames suitable for analysis. The Scikit-Learn (sklearn) Python library was used extensively for building models. The scaling object, used to standardize numerical features, Standard Scaler from sklearn was used, along with training and testing splitting functions to divide the dataset along with a grid search module to obtain the optimal hyperparameters for the models. Three of the regression models, random forest, support vector machines, and linear regression, were all implementations from sklearn as well. The data visualization library Seaborn was used to create relevant graphics and depictions of the dataset. The open-source boosted trees library XGBoost was also utilized.

To build this regression model, a full dataset was compiled of every year in the college scorecard archives since 2006. This master dataset was constructed by concatenating successive pandas data frames and subsetting them with the 'columns of interest' for the analysis.

The data was then explored through visualization and other methods of analysis. The missing value breakdown of the various columns was checked and the distributions of certain features were visualized. Basic correlations between the features and target variables were explored through scatterplots.

Next, the dataset was engineered for a model. All the NaN values, of which there is a significant amount, were imputed using various methods. One method of choice for imputation was using the mean value of the column grouped by the university so that if any one value existed across all the years recorded for a university, the mean of those values or value would be imputed. Following this, any row containing a NaN feature was dropped. Another method used was the scikit-learn implementation of K-Nearest-Neighbors Imputation, which uses supervised learning to use other column values in a row to predict the value of the NaN. Similarly, the scikit-learn implementation of the iterative imputer, which also uses supervised learning to predict the NaN values but with regression rather than nearest neighbors, was tried. Missing values were also imputed using XGBoost's built-in NaN-handling capability. The method with the best results was chosen to be used for further analysis.

Next, the data was split into features and labels and divided among test and training sets. The training features were scaled to standardize their numeric representations to ensure consistent model interpretation. Next, the four regression models (random forest, support vector machine, linear regression, and XGBoost) were instantiated and fit to the training data and each

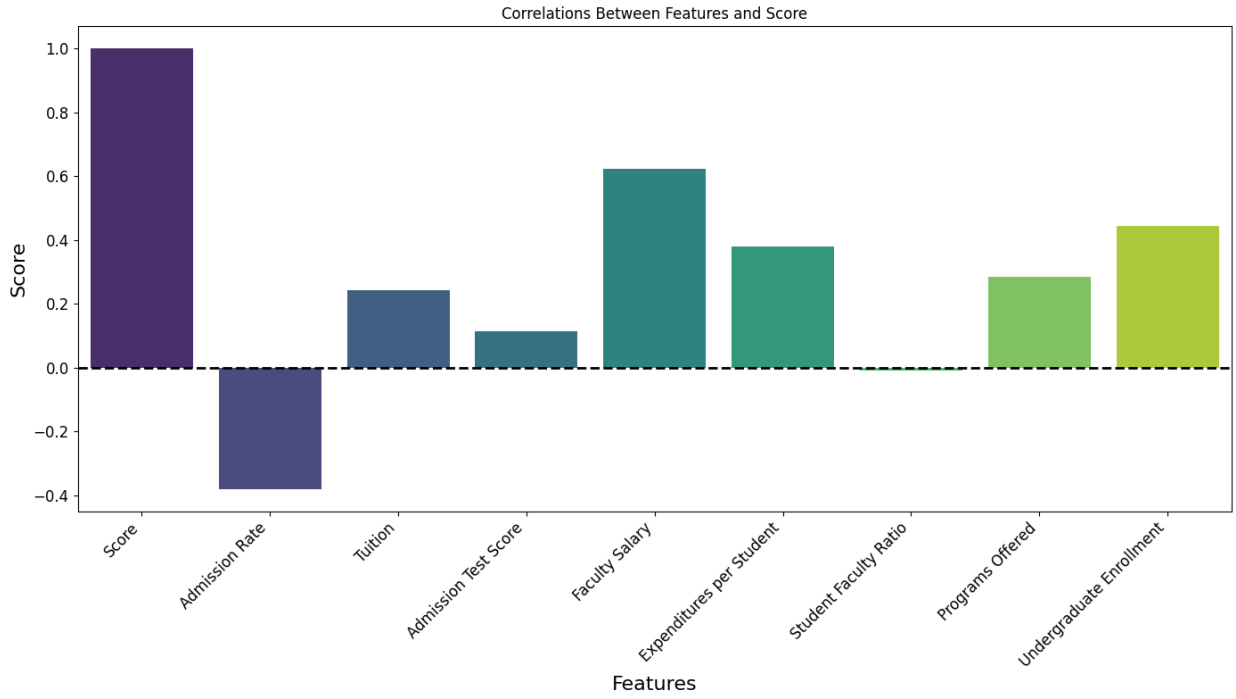
target variable. To determine the best hyperparameters to fit each model, a grid search using the GridSearchCV module from scikit-learn was used. 3 parameter grids specifying common values for various hyperparameters such as “learning rate” for XGBoost were constructed and fit to the Random Forest, XGBoost, and Linear Regression Models. After determining the best parameters for each model, they were re-fit with the optimal conditions to reduce error.

The Mean Squared Error of each model was calculated by using the testing data and the predicted values. Among these, the best model was used to predict the scores for all universities based on the input features. The importance of each feature, as well as their correlations with “Score”, were also computed to gain insight into the model and the features that contributed to the ranking.

Results:

Figure 1

Feature Correlations with College “Score”



Note: Expenditures per Student are only for instructional costs, this does not include amenities or scholarships. Admission Test Score also measures the standardized testing benchmarks needed for admission into the university.

Figure 2

Feature Importances from Random Forest Model

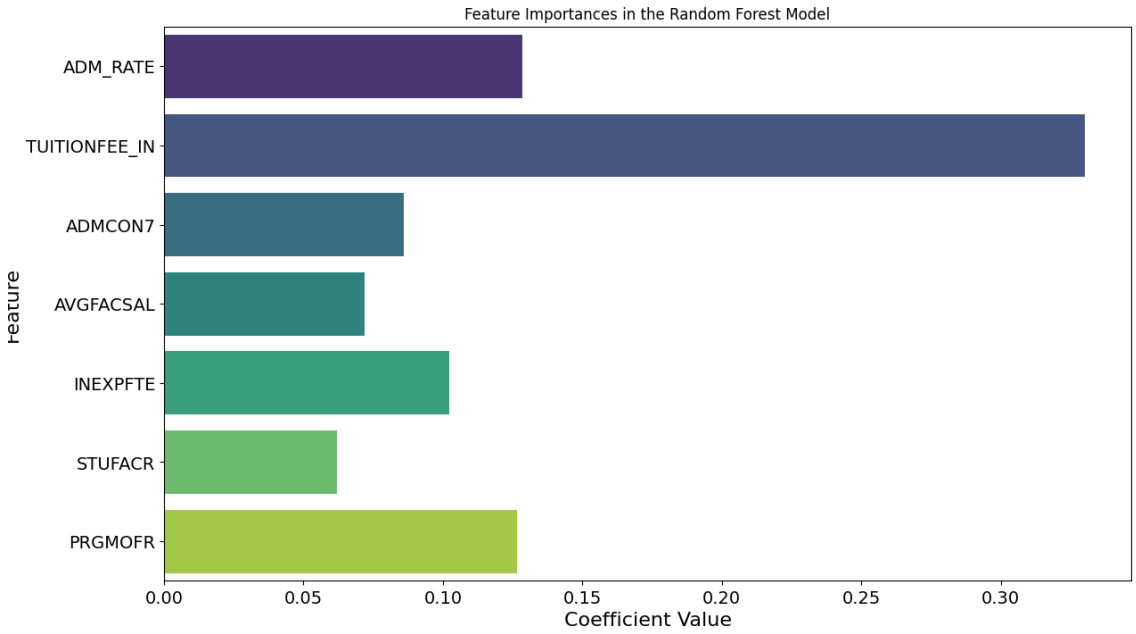


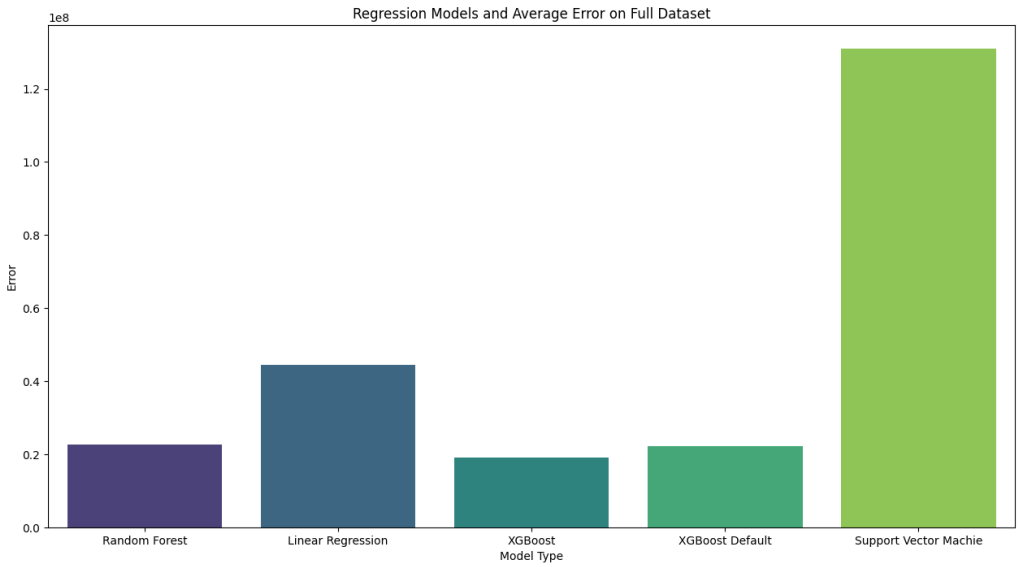
Table 1

Results of Grid Search for Best Parameters

	Model Type	Best Parameters for Model
0	Random Forest	{'max_depth': 20, 'min_samples_split': 2, 'n_estimators': 150}
1	XGBoost	{'learning_rate': 0.2, 'max_depth': 7, 'n_estimators': 200}
2	Regression	{'fit_intercept': True, 'positive': False}

Figure 3

Total Mean Square Errors for Various Regression Model Types



Note: The XGBoost Model with the tuned hyperparameters had the least amount of error

Figure 4

The top 30 College by "Score"

	University	Score
1622	Philadelphia College of Osteopathic Medicine	27.612006
1464	New York Law School	21.985230
1278	Meharry Medical College	15.215886
2249	University of California-San Francisco	11.351107
1757	Rosalind Franklin University of Medicine and S...	9.192727
2166	Thomas Edison State University	8.544719
867	Harvard University	8.290914
2522	University of Texas Southwestern Medical Center	7.803762
2210	United States Merchant Marine Academy	7.793420
1663	Princeton University	7.784344
1827	Salus University	6.848989
1265	Massachusetts Institute of Technology	6.133028
2673	Western University of Health Sciences	5.628585
2243	University of California-Hastings College of Law	5.415169
641	Duke University	5.255792
1736	Rice University	5.178262
1833	San Diego Mesa College	5.145596
336	California Institute of Technology	4.840023
1277	Medical University of South Carolina	4.803608
1651	Pomona College	4.593472
2742	Yale University	4.546301
1853	Santa Ana College	4.497976
757	Foothill College	4.357876
2171	Thunderbird School of Global Management	4.355766
772	Franklin W Olin College of Engineering	4.315261
318	CUNY Graduate School and University Center	4.274395
52	Amherst College	4.224116

Figure 5

The bottom 30 College by "Score"

	University	Score
619	Design Institute of San Diego	-1.670560
211	Beulah Heights University	-1.671849
2102	The Art Institute of Washington-Dulles	-1.675536
2691	Westwood College-Northlake	-1.680420
255	Broadview University-Layton	-1.681999
2686	Westwood College-Dupage	-1.685207
2679	Westwood College-Annandale	-1.686072
1422	National American University-Lewisville	-1.688532
1063	International Academy of Design and Technology...	-1.688861
579	Daymar College-Owensboro	-1.703345
256	Broadview University-Orem	-1.703725
714	Everglades University	-1.705976
577	Daymar College-Louisville	-1.709544
254	Broadview University-Boise	-1.711361
253	Broadview Entertainment Arts University	-1.733738
1432	National American University-Wichita West	-1.734598
1251	Martin University	-1.737139
1409	National American University-Burnsville	-1.737605
1405	National American University-Austin South	-1.740813
576	Daymar College-Bellevue	-1.748213
109	Arkansas Baptist College	-1.749723
1061	International Academy of Design and Technology...	-1.755220
2136	The North Coast College	-1.763799
2687	Westwood College-Ft Worth	-1.772347
862	Harrison College-Grove City	-1.796971
1955	Southwest University of Visual Arts-Tucson	-1.833089
1954	Southwest University of Visual Arts-Albuquerque	-1.833501
2684	Westwood College-Dallas	-1.894454
2688	Westwood College-Houston South	-1.926935
67	Apex School of Theology	-1.939774

Figure 6

Error of XGBoost Model with and without Best Parameters found from Grid Search

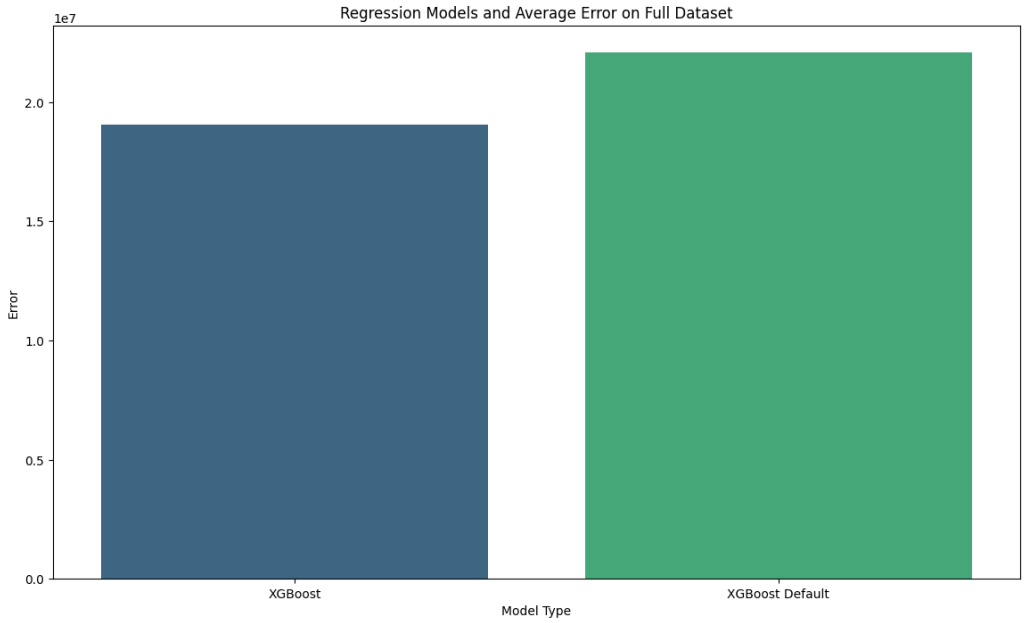


Figure 7

The Score of Local Pittsburgh Colleges

```
University of Pittsburgh-Pittsburgh Campus has a mean 'Score' rank of: 1027 out of 2748 universities.
Carnegie Mellon University has a mean 'Score' rank of: 225 out of 2748 universities.
Duquesne University has a mean 'Score' rank of: 827 out of 2748 universities.
```

Table 2

Total Not a Number (NaN) values for Full Dataset

```
df.isna().sum()
ADM_RATE                15903
TUITIONFEE_IN            7297
IRPS_NRA                15430
ADMCON7                  8226
AVGFACSAL                2108
PFTFAC                   4078
UGDS                     4733
TRANS_4                  8461
INEXPFTE                  255
OPENADMP                 6615
BOOKSUPPLY               7842
ROOMBOARD_OFF           11452
OTHEREXPENSE_OFF        11450
OTHEREXPENSE_FAM        11376
STUFACR                  9442
IRPS_NRA                15430
INSTNM                      0
PRGMOFR                 44368
WDRAW_ORIG_YR2_RT        7415
PCT75_EARN_WNE_P10      24696
COUNT_WNE_P10           24683
MD_EARN_WNE_P10         24683
GRAD_DEBT_MDN_SUPP       3473
Year                        0
dtype: int64
df.shape
(44801, 24)
```

Figure 8

Scatterplot Produced in Exploratory Data Analysis Stage

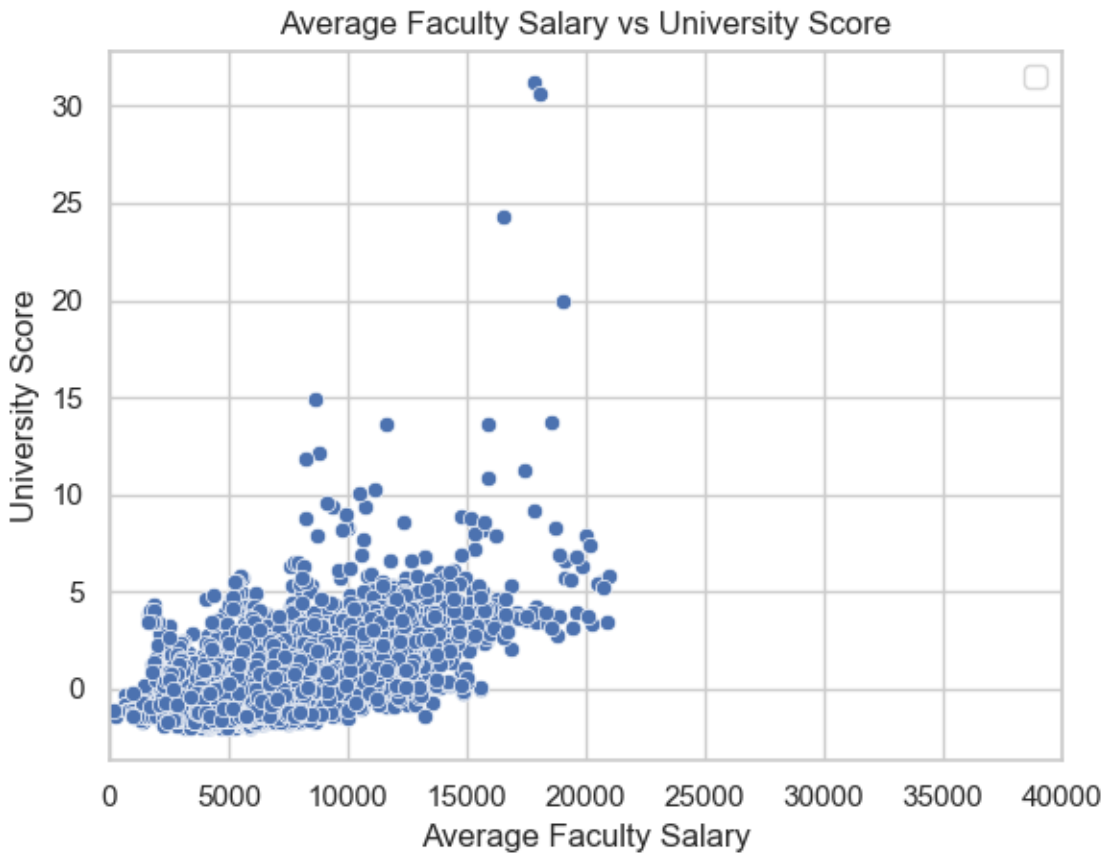


Table 3

Top 30 Universities Features and Duquesne's

Column1	Top 30 Universities	Duquesne
Programs Offered	25.19	15.345
Engineering Degrees	14.20%	0.30%
Average Faculty Salary	$11,459.15	$8,658.50
Admission Rate	20%	73.50%
Endowment Excluding Outliers	$282,327,970.86	$247,335,875.00
Instructional Expenses per Student	$27,037.00	$10,748.12

Discussion:

After building the various models, a variety of conclusions about the models, the data, and their implications were found. To begin, the largest challenge encountered while performing analysis on this dataset was the presence of missing values. Large swaths of data were marked with a “Privacy Suppressed” label indicating that, for privacy reasons, the data could not be obtained for that column and row. This, along with many blank values, led to a staggering amount of Not a Number (NaN) values within the dataset. As seen in Table 2, some columns contained up to 44,000 missing values. This issue proved especially prevalent for smaller universities where it can be inferred that data collection methods are not as robust or encouraged compared to large, established institutions. Additionally, to ensure model efficacy, universities that did not have any data for any of the target variables were dropped from the analysis. It was determined that imputing target variables would lead to noise, so any University that did not have target variables recorded was excluded.

After attempting various imputation methods, the following strategy resulted in the lowest error. All rows containing missing values for the target variables were dropped from the dataset. Next, all feature column NaNs were imputed using the scikit-learn iterative imputer module. If any NaNs remained following this, the rows were simply dropped from the analysis. Several imputation techniques were experimented with to retain the core information of the data and future research should be done into the optimal methods for data of this structure. The mean of each feature column, grouped by university, was attempted. KNN-supervised imputation, which uses a K-Nearest-Neighbors search to predict the NaN value from the other row values,

was also implemented. Finally, the XGBoost built-in NaN imputer was deployed. None of these three imputation methods were as effective in reducing the error as the first one described above. After building a random forest decision tree model to predict the score of each university, the features and splits were examined to determine which features contributed most strongly to the predictions. As Figure 2 depicts, the number of programs offered, the total tuition, and the admission rate were the most important features in predicting the score. This means these features, on average, were higher in the decision tree splits marking them as more definitive predictors of "Score". Therefore, these features may be the most important for Universities to examine when attempting to improve the quality of their establishment by improving their “Score”.

The correlations of the features and the “Score” were also determined to assess whether positive or negative values for the features contribute to a better score. As seen in Figure 1, features such as high tuition rates, high faculty salary, high undergraduate enrollment, high faculty salary, high expenditures per student, and high amounts of programs offered all correlate with a better score. Low admission rates correlate with a higher score, and student to faculty ratio is a negligible predictor. This intuitively makes sense, as Universities with more resources routinely perform better. They can pay more to their faculty, resulting in attracting higher quality instructors with greater teaching abilities. They limit which students they admit to those with high academic potential, allowing for a more elite student body. They offer a wide range of programs and invest in them with money, all leading to attracting more talented students and developing them better within the classroom. These qualities allow students to enter the workforce with better skills which will help them command higher salaries and gain employment, improving the score of their alma matter. While these universities generally have a

higher tuition rate, it can be inferred that students attending these universities, because they are talented, garner more scholarships which can reduce their debt as well, improving their university scores.

When examining the top 30 colleges ranked by score as depicted in Figure 4, medical specialty schools and law schools scored highest most likely because they offer high salaries to graduates and have direct pipelines to gain employment. Unsurprisingly, Ivy League institutions like Harvard University and Princeton University, as well as elite tech colleges like MIT, are included in the high score category. These universities maintain a reputation for their elite academic rigor, and the findings show that the payoff of this can be quantified.

When examining the bottom 30 colleges, design and theology schools are commonly found. These schools and their trades, while meritorious, do not generally command high salaries in the U.S. workforce (*Top 26 theology careers [+ salary info]* 2024). Due to their graduates receiving lower salaries, and often not gaining employment due to a lower number of jobs in the U.S. economy, their scores are generally lower. This shows that, when striving for financial success, art and theology schools are not the best avenues.

Specifically, to improve Duquesne, the model recommends the following actions. It should be noted that among the top 30 universities, 3 were excluded from the metrics seen in Table 3 due to their unusually high endowments. These universities (Harvard, Princeton, MIT) could skew financial recommendations like faculty salary as they have a disproportionate amount of wealth that is not seen in most universities. Excluding these outliers, as seen in Table 3, the endowments among Duquesne and the Top 30 universities are similar enough for proper analysis.

Duquesne, to improve its quality and score, should implement more programs as a whole, offering a wider array of disciplines to study. Specifically, most of the top universities have robust engineering degrees as seen in Table 3, which Duquesne would be wise to explore adding to their curriculum. Engineers command high salaries and are in demand in the workforce, which would certainly improve a university "Score" if these students secure prosperous careers following graduation (*11 in-demand engineering jobs (plus salaries and duties)*). Duquesne should also spend more money on their faculty salaries and instruction for students. As seen in Table 3, they proportionally spend less on these items compared to top universities. Spending more money on faculty salaries and resources for professors can attract higher quality talent that can teach and prepare students better for life after graduation, increasing their odds of landing a high paying job and securing employment. Increased instructional expenses can also give them better skills in the workplace which can command them higher salaries and allow them to better manage their debt following graduation. Perhaps Duquesne should redirect money away from certain areas like building new structures to these areas to improve these metrics. Finally, Duquesne should raise their admission standards to lower their admission rate. This can allow Duquesne to be more selective with their students and admit higher quality talent. By becoming more exclusive, Duquesne can raise their standards and ensure that interested students of high caliber view them as a worthy destination. By acting on these suggestions, Duquesne is likely to increase their "Score" which can be indicative of the overall quality of a university.

Choosing these features to predict score proved difficult when building the model, as the desire was to only select features that could be controlled by the University while remaining informative of the target variable. All of these features, such as average faculty salary or tuition, can be influenced by university policy decisions. Another challenge in choosing these features

was the presence of NaN values for some informative features. Several informative features needed to be excluded from the model altogether due to an overwhelming amount of missing data. The models generated generally suffered from high bias, but more features could not necessarily be added to improve this with the amount of missing data for some of them. Therefore, a balance between including more features and ensuring that they were informative needed to be achieved to optimize for low error.

To obtain the optimal hyperparameters for each model type, a grid search was employed. As seen in Table 1, the optimal parameters for the Linear Regression, Random Forest, and XGBoost models are shown. The support vector machine was excluded from the grid search as it routinely performed worse than the other models. Grid search is a compute and time-intensive process and obtaining the ideal parameters for models that were deemed inferior in this context was decided to *not* be a worthwhile investment.

The Grid Search significantly improved the performance of our models, however. As seen in Figure 6, the error of XGBoost reduced dramatically when fit with the ideal hyperparameters derived from the grid search.

The best-performing model was the XGBoost implementation, closely followed by the random forest model. The accuracy of the models was measured using the predicted target variables compared to the actual targets, using the mean square error to assess model competency. Both of these models employed decision trees to reach their predictions. Due to the varying contexts of the variables and the likely non-linear relationship between some of them and the target variable, threshold-based models like decision trees seemed to perform better compared to linear models. The linear models had, on average, higher errors compared to the decision tree-based models. For this reason, the XGBoost model, fitted with the data from the

best imputation technique described earlier, was used as the official model for specific predictions.

Conclusion:

Regression techniques can be applied to large datasets to predict continuous values and derive important features, both of which can be used to influence real-world decisions. After obtaining and combining a large dataset from *Collegescorecard.ed.gov*, several regression models were built to predict the overall quality of a university from a set of feature variables that could be controlled with University policy, like average faculty salary or acceptance rate. The quality "Score" was derived from the salary-to-debt ratio and the enrollment-to-employment ratio. The results showed that variables like high faculty salary, high tuition, and offering a wide range of programs are associated with student success. Schools with high acceptance rates are associated with less positive student outcomes. To optimize their score, Universities can offer more programs in a variety of disciplines, make their admission process more selective, reallocate more resources to instructional expenses and experiences for students in the classroom, increase the faculty salaries and attract higher talent professors, and attract more students. This will improve their score, and therefore, potentially increase positive outcomes for their students

Specifically, the model recommends that Duquesne increase the number of their programs and expand their engineering school. It also recommends proportionally spending more money on staff salary and instructional expenses within the classroom to better develop student’s skills. Finally, by raising admission standards Duquesne can be more selective and attract more

talented high school students. By following these actions, Duquesne can increase their “Score” and overall quality.

The data contained large amounts of missing data, mostly due to “Privacy Suppressed” labels indicating privacy concerns. Large amounts of data, particularly for small universities, were also missing. To impute this data, a combination of iterative techniques and dropping rows was used to construct a dataset that contained as much of the original information as possible.

After building and evaluating multiple models, the XGBoost model, fitted with optimal hyperparameters, proved to be the best in predicting the college score. This model was used to predict the scores of all remaining available universities and rank them. The University with the greatest score was the Philadelphia College of Osteopathic Medicine while the school with the lowest score was the Apex School of Theology. Ivy League and Medical schools were common in the high score category while Design and Theology Schools were common in the low score category, indicating their respective potential for graduates to command high salaries, little debt, with high employability.

References

Anwar, A. (2021, June 7). *A beginner's Guide to Regression Analysis in machine learning*. Medium. https://towardsdatascience.com/a-beginners-guide-to-regression-analysis-in-machine-learning-8a828b491bbf

Data Home: College Scorecard. Data Home | College Scorecard. (n.d.). https://collegescorecard.ed.gov/data

Introduction to boosted trees. Introduction to Boosted Trees - xgboost 2.0.3 documentation. (n.d.). https://xgboost.readthedocs.io/en/stable/tutorials/model.html

Mazurkiewicz, J. (n.d.). *College Improvement GitHub Repository*. GitHub. *Code used for this Project*. https://github.com/jakemaz66/CollegeImprovement

Rosidi, N. (2022, February 2). *Overview of machine learning algorithms: Regression*. Medium. https://towardsdatascience.com/overview-of-machine-learning-algorithms-regression-e0f5510e84c

Top 26 theology careers [+ salary info]. University of San Diego Online Degrees. (2024, March 6). https://onlinedegrees.sandiego.edu/theology-careers/

11 in-demand engineering jobs (plus salaries and duties). (n.d.). https://www.indeed.com/career-advice/finding-a-job/in-demand-engineering-jobs