Nathaniel Jones
Classification Project

Stat4310: Data Mining
Dr. Vanderheyden

Introduction:

Today, a student may be eligible for a portion of the Pell grant if their income, or their parents' income, in the case of dependent students, is less than $50,000. The portion the student receives increases as the amount of income decreases. A maximum amount of $6495 is granted to those students whose household income is less than or equal to $20,000. For these reasons, receiving a Pell grant is associated with low-income students. I researched into the CollegeScorecard dataset and created the indicator variable, "PELL_CAT," by using the proportion of Pell-receiving students ("PCTPELL") to categorize institutions as either "majority Pell" (>50% of students receiving a Pell grant) or "minority Pell" (≤50% of students receiving a Pell grant). I created this indicator variable to determine where lower-income students attended more frequently and the outcomes at these institutions.

Key Findings and Results:

Best Model: *{Accuracy/AUC – Training:* 97.8%/99.9%, *Testing:* 86.2%/93.5%, *Validation:* 88.1%/93.4%*}*

The model I determined to be the best was a random forest on the standardized dataset that was reduced to the top 10 most independent variables. Figure one displays the features that were most important in classifying. The most important of these features to this model was the percentage of Federal Loan Borrowers at the institution, and the next most important was the agency that accredits the institution. With only 138 incorrect classifications in total (Figure 2), this model correctly classified:

- 537 majority Pell institutions
- 325 minority Pell institutions

The following figures display the important features to this model:

Figure 1: Feature Importance Plot for the Random Forest Classifier

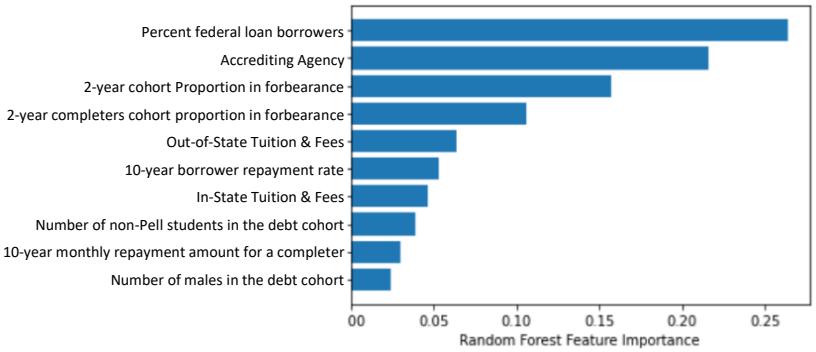
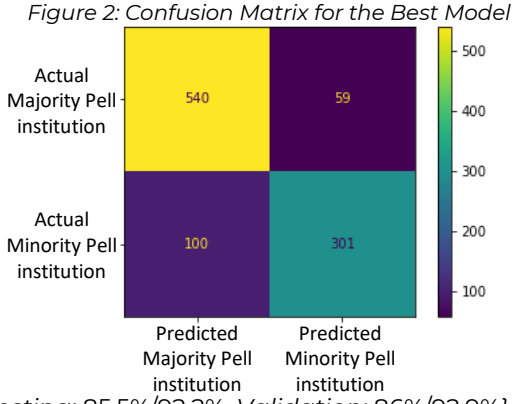


Figure 2: Confusion Matrix for the Best Model



Next Best Model: *{Accuracy/AUC – Training:* 86.2%/93.9%, *Testing:* 85.5%/92.2%, *Validation:* 86%/92.9%*}*

The next best model used the method k-nearest neighbors on the standardized-transformed dataset. This model correctly classified 511 majority Pell institutions and 320 minority Pell institutions, while misclassifying a total of 169 incorrect classifications across both groups. Figure 3 displays the results of this classifier using the validation dataset. The red and blue dots are the correctly classified Majority and Minority Pell schools while the black and dark brown dots are the incorrect classifications.

Figure 3: Scatterplot of the Monthly Average Faculty Salary by the percent of Federal Loan Borrowers Stratified by Correct/Incorrect Classification.

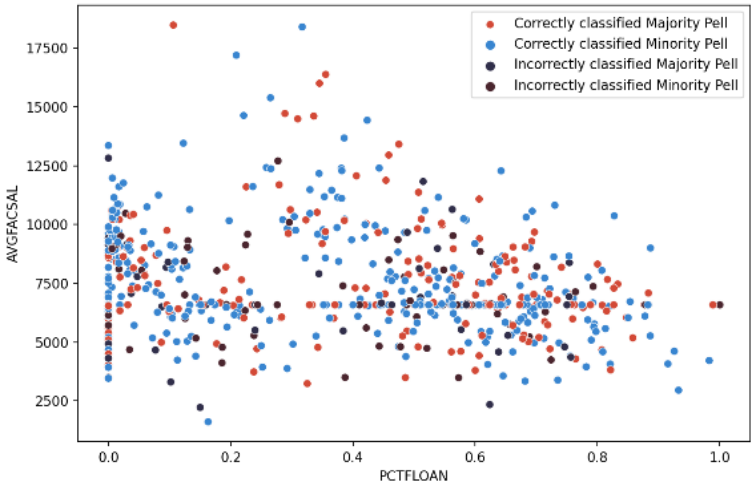
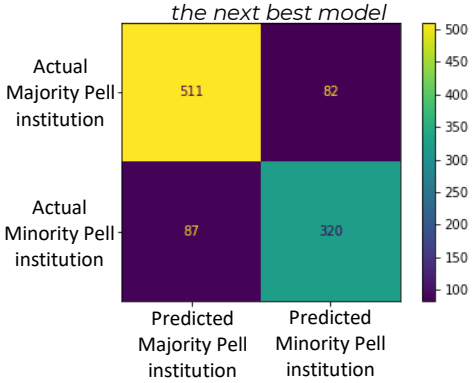


Figure 4: Confusion Matrix for the next best model



Background, and Data Source:

The CollegeScorecard dataset is released by the U.S. Department of Education through the Integrated Post-Secondary Education Data System (IPEDS). IPEDS surveys post-secondary institutions, the IRS, and FSA annually and collects the responses to 13 different surveys about the institutions; a portion of these responses make up the CollegeScorecard dataset. This data consists of variables related to post-secondary institutions, such as student demographics, admission rates, SAT/ACT scores, costs related to attendance, academic outcomes, and loan outcomes. IPEDS has released 23 years of post-secondary data. This project will focus on the most recent release, which is the 2017-18 school year.

Problem Statement:

The goal of this project is to create a model that classifies institutions as either "Majority Pell" or "Minority Pell". Explicitly, I want to answer the question,

"Which institutional features are associated with the majority Pell schools?"

By answering this question, we can better understand the most prevalent attributes of schools with a high proportion of students receiving a Pell grant. Gaining the understanding of these attributes will aid in detecting problems, trends, and stratifications at institutions where low-income students are in high attendance.

Key Response Variable, Data Metrics, Concerns, and Preparation:

I created the variable "PELL_CAT" to indicate whether an institution is a "Majority Pell" institution (0) or a "Minority Pell" institution (1). This variable was created based on the percentage of students receiving a Pell grant ("PCTPELL") by categorizing institutions with greater than 50% of their student population receiving a Pell grant as "Majority Pell" institutions, and institutions with 50% or less of their student population receiving a Pell grant as "Minority Pell" institutions. Out of 6,806 observations across 2,384 variables, 3,439 Majority Pell institutions, 2,575 Minority Pell institutions, and 792 other institutions had missing values for the variable, "PCTPELL." Institutions with missing "PCTPELL" values were dropped from the dataset, along with institutions located anywhere other than a U.S. state or Washington D.C. This left the dataset with 5,789 observations, of which 3,435 institutions were labelled "Majority Pell" and 2,444 were labelled "Minority Pell."

Other concerns with the data included:

Concern	Solution
➢ 1,419 columns contained missing values for every observation.	✓ These columns were dropped.
➢ 16 variables related to an institution's URL, ALIAS, FSA ID, or description of offered programs.	✓ These columns were dropped.
➢ 155 variables were pooled and/or suppressed versions of other variables in the dataset.	✓ These columns were dropped.
➢ Of the remaining 783 variables, 523 contained missing values.	✓ These columns were imputed on the median at a tolerance level of 42.8% missing values.
➢ 4.18% of feature pairings were highly correlated with each other.	✓ Dimension reduction and variable selection were used to reduce multicollinearity.

The final dataset contained a total of 5,789 observations across 446 variables. Of the variables, 3 were institutional IDs, 225 were categorical data, and 221 were numeric values.

Feature Engineering and Methods:

During the cleaning and structuring phase, I broke down the 446 variables into groups based on their definitions and/or variable type. I formed the variables into seven subgroupings which include: ID variables, variables that related to a Classification of Instructional Program (CIP) code, variables related to institutional demographics, variables related to an institution's student population demographics, variables specifying the institution's geolocation, variables specifying the academic outcomes at each university, and variables describing an institution's typical loan outcome proportions, counts, and rates.

Variable Transformation:

The CollegeScorecard dataset included the variable, "T4APPROVALDATE," which indicates the date an institution entered the Title 4 Program. Initially, this variable consisted of only one column with values in "MM/DD/YYYY" format. I split this column into three separate columns, each representing the month, day, and year, and then I created a variable to indicate the season an institution entered the Title 4 Program. Many variables in the dataset can be used to create variables of interest. For example, the variable "BBRR2_FED_UG_DFLT" indicates the loan default rate for undergraduate students 2 years after exiting their school, and "DBRR2_FED_UG_N" indicates the number of undergraduate students in the 2-year post-exit cohort. The product of these two variables is the number of students in the post-exit cohort that defaulted on their federal undergraduate loan within 2 years of exiting their school.

One-Hot Encoding and Target Encoding:

I used One-Hot Encoding to create dummy variables for the following ordinal variables:

- "ST_FIPS"
- "ICLEVEL"
- "ACCREDCODE"
- "PREDDEG"
- "HIGHDEG"
- "CURROPER"
- "CONTROL"
- "MAIN"
- "HCM2"
- "month"
- "OPENADMP"
- "OPEFLAG"
- "HSI"
- "HBCU"
- "ANNHI"
- "PBI"
- "TRIBAL"
- "NANTI"

I then dropped the first level from each variable and the original variable from the dataset, leaving only variables indicating the name, city location, ZIP code, and accreditation agency name for an institution. I chose to target encode these variables, as well as the variable I created, "Season."

Dimension Reduction and Variable Selection:

I found during the pre-stages of modelling that around 4% of my feature pairings were highly correlated. Many of these pairs were perfectly correlated to another variable in a similar grouping as described above. Initially, I encoded the institution's geolocation, the name of the school, and the name of the agency that accredits the school. I used dummy variables for the remaining categorical features in my dataset since a majority of those remaining were ordinal. I then performed five transformations for each of the variables in my dataset. This resulted in nearly 1,000 additional variables in my dataset and increased the percentage of highly correlated feature pairings to 6.8%.

This method resulted in poor model performance and training speed. I replanned my method and backtracked to before One-Hot encoding the categorical features. A new plan of action was formed, I decided to perform principal component analysis on individual data subgroupings to reduce the number of variables in my dataset. I individually performed this step on the groupings of variables related to CIP codes, academic outcomes, and loan outcomes

Principal Component Analysis:

Variables Related to the Classification of Instructional Program (CIP) Code

There were 228 variables corresponding to program offerings of the institutions.
Of those, 190 were three level categorical variables, with 0 indicating that the institution does not offer the program, 1 indicating that the institution offers the program in-person and online, and 2 indicating that the institution only offers the program online. The additional 38 variables corresponded to the proportion of degrees/certificates an institution awards for a particular program they offer.

Figure 5: Line plot of Explained Variances ratio by number of components of Program Offerings

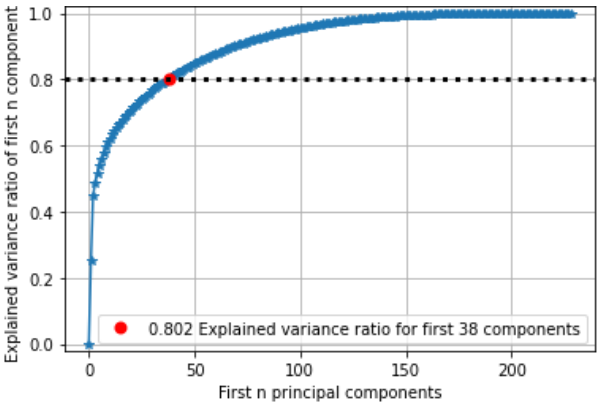
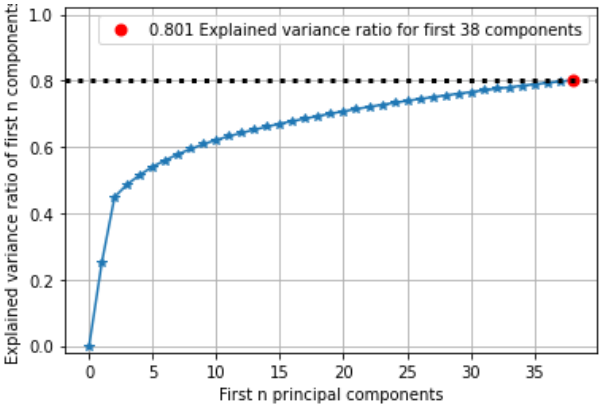


Figure 6: Line plot of Explained Variances ratio by number of components of Program Offerings (Zoomed X-Axis)



From the plot above, we can see that the first 38 principal components reached an explained variance ratio of 0.802 (Figure 5 and 6). Notably, there were 38 principal components that accounted for 80.1% of the explained variance ratio. It was interesting because my dataset included 38 different CIP code programs. With just the top five components, 54.1% of the explained variance ratio was achieved with the variables associated with program offerings for the CIP codes 52 (Business, Management, and Marketing Service programs), 51 (Nursing and health Professional), and 12 (Personal and Culinary Services).

Variables Related to the Loan Outcomes

There were 64 variables that described different median loan outcomes or adjusted cohort counts for the loan outcomes of students after they exit their institution. These variables are included for 1-, 2-, 4-, 5-, 10-, or 20-year cohorts. It was found that 3 components accounted for 81% of the explained variance (Figure 7 and 8). The variable that best represented each of the principal components were selected as the representative.

- The principal loan amount accumulated by undergraduate students who received an award in the 2-year loan outcome cohort best explained the first principal component.
- The loan repayment rate of undergraduates who received an award 4-years after graduating best represented the second component.
- The number of parents in the 1-year loan outcome cohort who received a PLUS loan best represented the third component.

Figure 7: Line Plot of Explained Variances Ratio by Number of Components of Loan Outcomes

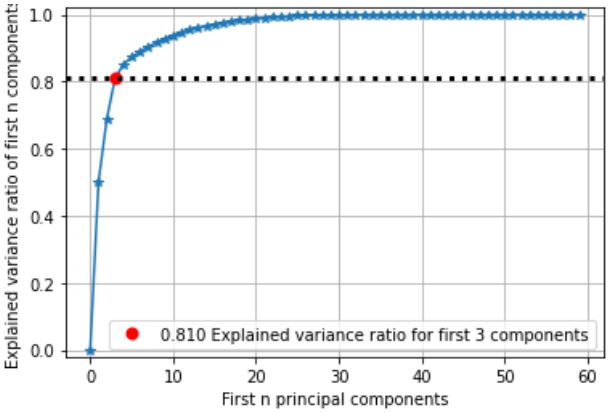
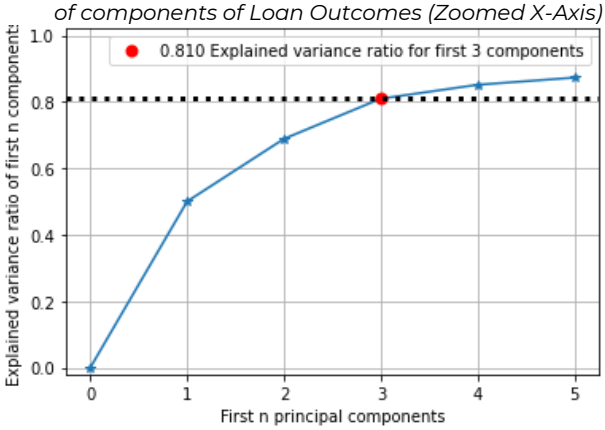


Figure 8: Line plot of Explained Variances ratio by number of components of Loan Outcomes (Zoomed X-Axis)



Variables Related to Academic Outcomes:

There were 28 variables related to academic outcomes. A student may exit their institution within either 6 years (150% completion time) or 8 years (200% completion time). As a student exits the school they attend, they leave with one of three outcomes, either they leave with an award/certificate/degree, they transfer to another school, or they withdraw from their school without receiving an award. The CollegeScorecard dataset records these three outcome groups for both completion time cohorts, plus an additional outcome including students still enrolled at the same institution after 8 years without receiving an award. For each of these outcome groups, four variables are recorded for combinations of full-time/part-time and first-time/not first-time students. During the initial attempt at modelling the dataset, these variables produced correlation problems, I decided to use PCA to find the variables that have the highest proportion of the explained variance in the target variable.

Figure 9: Line Plot of Explained Variances Ratio by Number of Components of Academic Outcomes

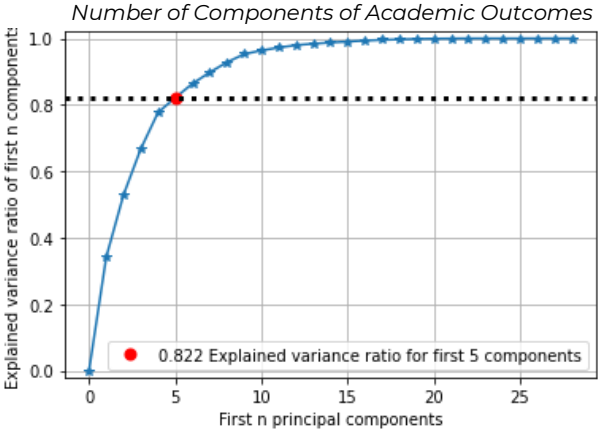
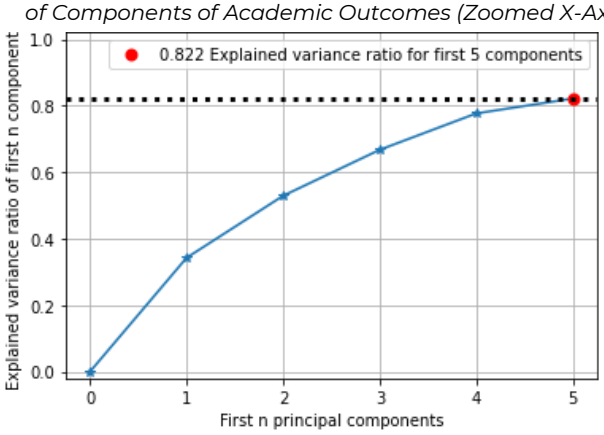


Figure 10: Line Plot of Explained Variances Ratio by Number of Components of Academic Outcomes (Zoomed X-Axis)



From the model, 82% of the explained variance was gained from the first five components (Figure 9 and 10). The most important feature for each of the first five components were selected as the representative. These features are listed below in order:

- The proportion of Full-time students' who graduated within 8 years (PC1).
- The proportion of Full-time students' who transferred within 8 years (PC2).
- The adjusted count of the Full-time, First-time student 8-year outcome cohort (PC3).
- The proportion of Full-time students' who is still enrolled without receiving an award after 8 years (PC4).
- The proportion of Full-time, First-time students that graduated within 8 years (PC5).

Interestingly, the proportion of Full-time students who are still enrolled without receiving an award after 8 years was an important feature. I fall into this category and earlier exploration into my data found that the median proportion of students that are still enrolled without receiving an award after 8 years is 0.3%.

Model Development:

I created models for the original dataset, a standardized version of the original dataset, and another standardized version of the original dataset with various variable transformations. For each dataset I ran a principal component analysis, logistic regression, XGBoost algorithm, K-nearest neighbors, and random forest. In addition, I used a function in sklearn that computes the mutual information between two random variables and returns the variables that are closest to zero. The mutual information between dependent variables will return higher values. Truly independent variables will return values equal to zero. The top ten variables closest to zero were selected for each variation of the three datasets and analyzed.

Analyses on the Original Dataset

Principal Component Analysis.

The first analysis considered all variables from the original data set and found that the first 3 components accounted for 83% of the explained variance. Figure 11 below visually displays the difference in the explained variance for each of the three components. The first principal component accounted for 54.2% of the explained variance while the other two components explained less than 15% each. Figure 12 displays a scatterplot with the first component on the vertical axis and the third component on the horizontal axis. From this angle we can begin to see a pure clustering of majority Pell schools (in grey) in the top right quadrant of the plot.

Figure 11: Cumulation of 83% Explained Variance by the First Three Principal Components for the Full Original Dataset

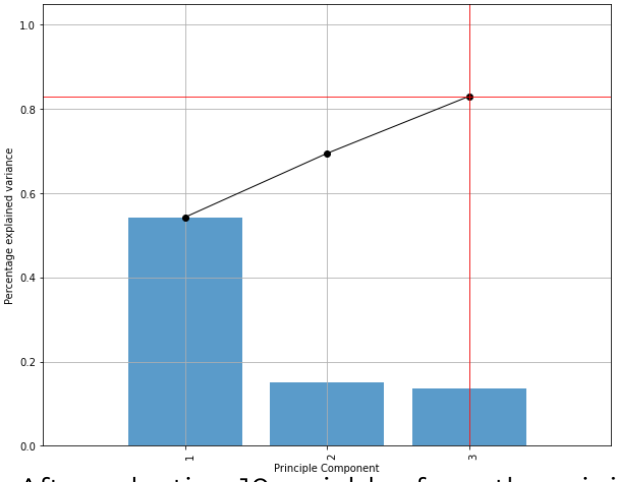
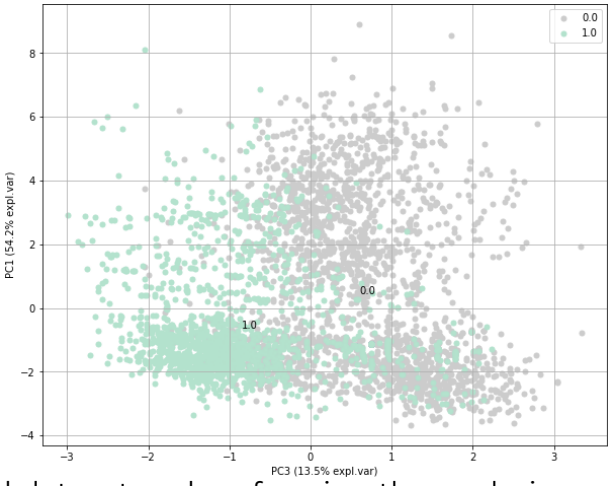


Figure 12: Scatterplot for the Explained Variance of the Third Principal Component by the First for the Full Original Dataset



After selecting 10 variables from the original dataset and performing the analysis again, I found that the first five components in the model accumulated 99.3% of the explained variance. Figure 13 shows that the first component made up nearly 80% of the explained variance, which was greater than four times the cumulative sum of the next four components. The features that are most important to each of the components are the in-state tuition and fee (PC1), the typical amount of debt accumulated by an undergraduate student that received an award (PC2), the out-of-state tuition and fees (PC3), the average salary of the faculty (PC4), and the number of students in the high-income student debt cohort (PC5). Figure 14 on the next page displays the first component on the horizontal axis and the third component on the vertical axis. Moving across the plot from left to right, we can see that most of the minority Pell schools (in seafoam) are clustered close to the origin. Figure 15 displays a different angle of the components where the horizontal axis is the second component, and the vertical axis is the first component. This view of the principal components displays a region that is purely made up of majority Pell schools.

Figure 13: Cumulation of 99.35% Explained Variance by the First Five Principal Components for the Top Ten Most Independent Variables

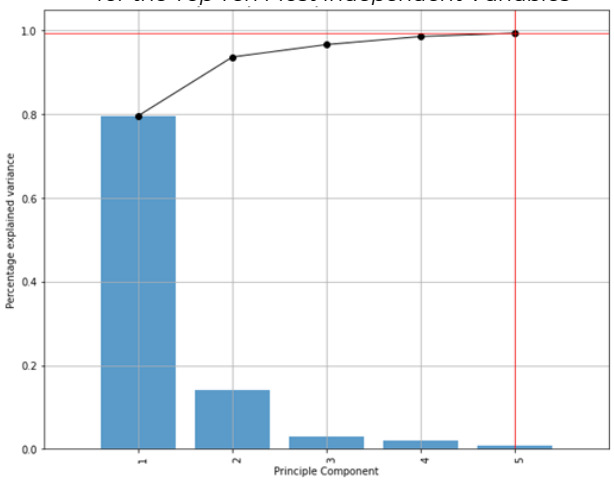


Figure 14: Scatterplot for the Explained Variance of the First Principal Component by the Third for the Top Ten Most Independent Variables

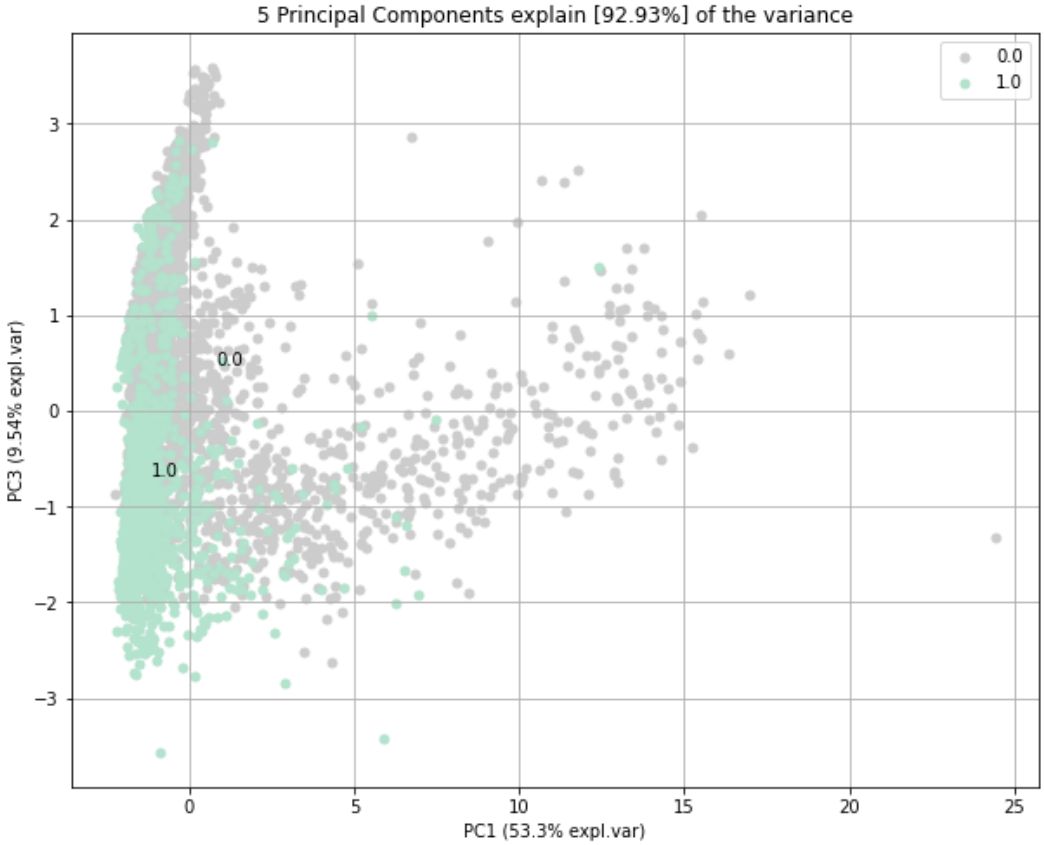
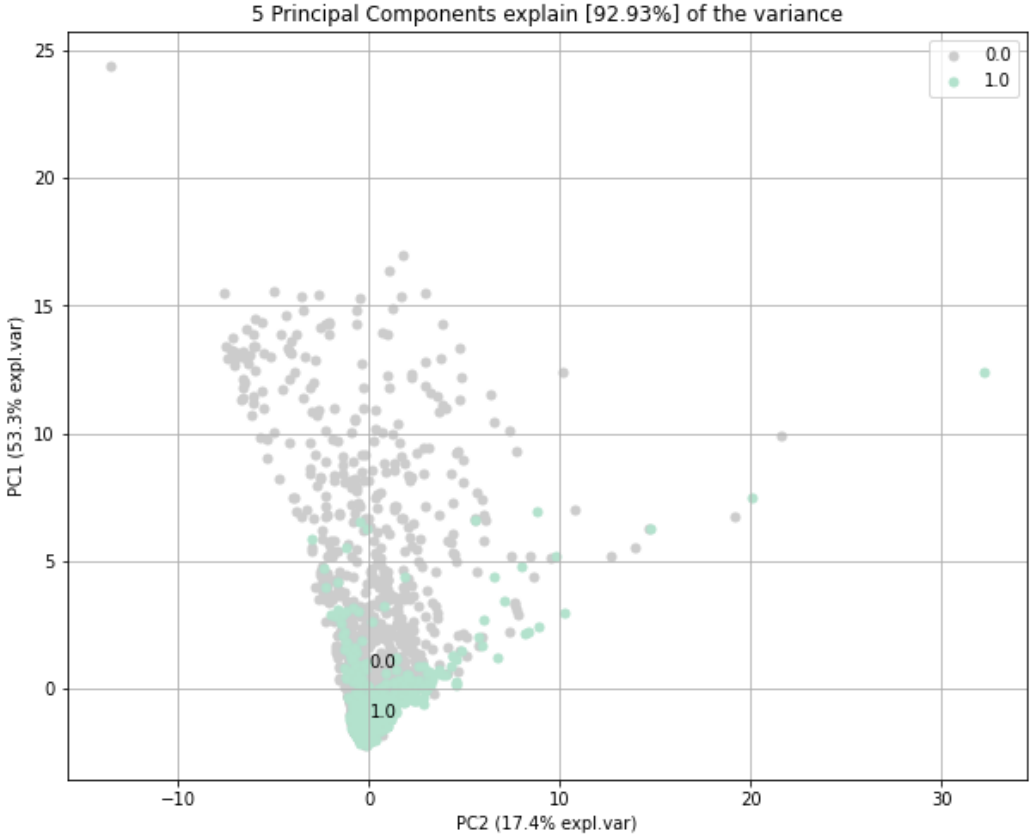


Figure 15: Scatterplot for the Explained Variance of the First Principal Component by the Second for the Top Ten Most Independent Variables



Logistic Regression.

For the original dataset, I performed a logistic regression using an L1 penalty, a value of 0.01 for the C parameter, and a max iteration of 10. The training set produced an accuracy of 58%, the test set produced an accuracy of 59.6%, and the validation set produced an accuracy of 56.4%. In addition, the area under the curve (AUC) for the test and validation sets were 67% and 62%, respectively, while the training set scored an AUC of 61.1%. I then looked at the relation between the correctly and incorrectly predicted institutions. The model showed that it can predict majority Pell schools correctly (555 vs 38), but not minority Pell institutions (35 vs 372, shown in Figure 16). Since the model poorly predicted minority Pell institutions, I concluded that it would not be a good model for classifying majority and minority Pell institutions.

Figure 16: Confusion Matrix for the Full Original Dataset

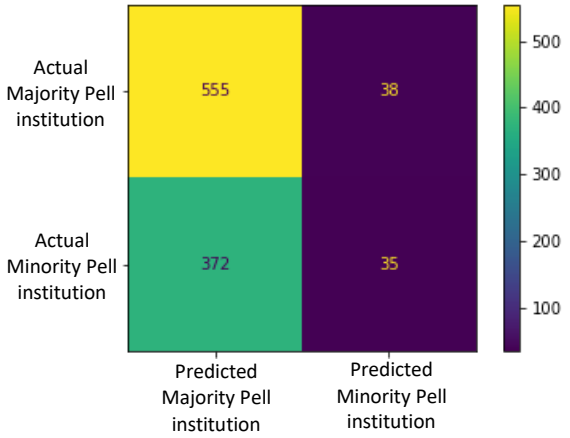
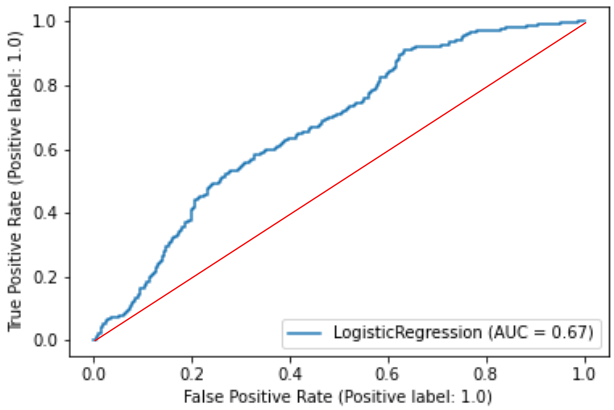


Figure 17: The Test set AUC Plot for the Correct/Incorrect Ratio for the Full Original dataset



I then reduced number of variables in this dataset using the entropy gain between two random variables and found the 10 most independent variables of the set. I performed a logistic regression on these 10 variables using an L2 penalty, a value of 0.1 for the C parameter, and a max iteration of 50. The training set produced an accuracy of 58.3%, the test set produced an accuracy of 59.2%, and the validation set produced an accuracy of 54.8%. In addition, the area under the curve (AUC) for the test and validation sets were 62.3% and 59.1%, respectively, while the training set scored an AUC of 61.5%. Although this model correctly labelled more minority Pell institutions, it did worse at labelling the majority Pell institutions (Figure 18). Therefore, this model was also not fit for classifying majority and minority Pell institutions.

Figure 18: Confusion Matrix for the Top Ten Most Independent

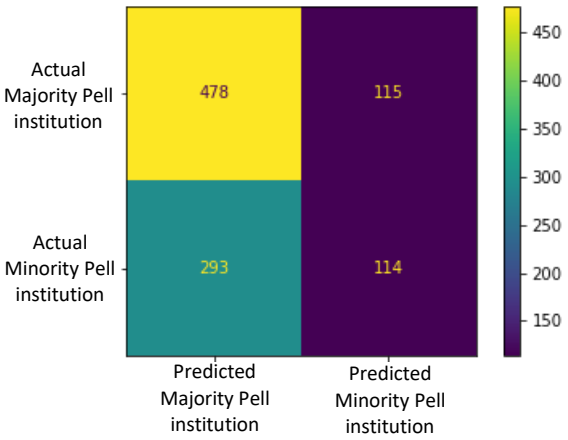
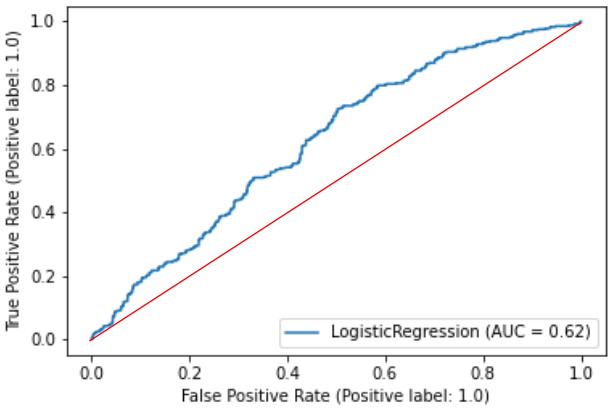


Figure 19: AUC Plot for the Correct/Incorrect Ratio for the Top Ten Most Independent Variables



Logistic Regression with Principal Component Analysis.

I then performed a convolution of a principal component analysis (PCA) with a logistic regression to explore the predictive ability of the principal components in the original dataset. Since PCA requires standardization and I did not standardize the original dataset, I was not hopeful in finding a better model for this dataset.

With all the variables, an L2 penalty, a max iteration of 10, and the value 0.01 for the C parameter, the model produced 51% accuracy on the training set, 52.8% accuracy on the test set, and 52.3% on the validation set. For the test and validation sets, the model performed better in terms of AUC (65.8% and 61.6% respectively) than logistic regression alone. However, analyzing the relation between correctly and incorrectly predicted institutions revealed that this method performed very well at predicting minority Pell institutions, but poorly predicted majority Pell institutions (Figure 20).

Figure 20: Confusion Matrix for the Full Original Dataset

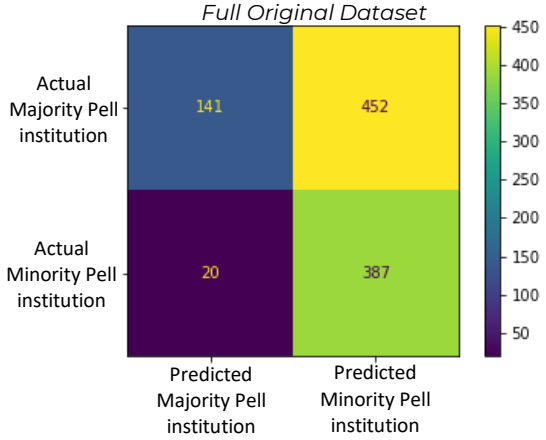
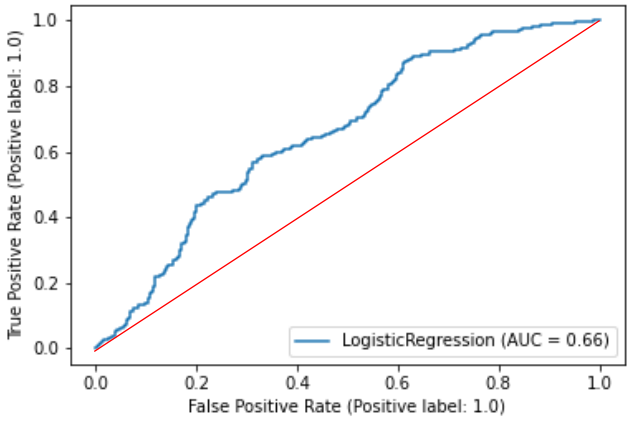


Figure 21: AUC Plot for the Correct/Incorrect Ratio for the Full Original Dataset



Using only the 10 most independent variables from the original dataset, a second logistical regression with PCA found double the amount of correctly labelled majority Pell institutions compared to the previous model. The parameters of the model included an L1 penalty, a value of 0.001 for the C parameter, and a max iteration of 100. Although it correctly predicted a high amount of both majority and minority Pell institutions, this model incorrectly predicted an equally high amount of majority Pell institutions (Figure 22). This model produced accuracies of 64.1%, 66.2, and 64.1% for the training, testing, and validation sets, respectively. In addition, the AUCs produced were 71.3%, 72.3%, and 71% for the training, testing, and validation sets, also respectively. Since this model made an equal amount of correct majority Pell predictions as it did incorrect majority Pell predictions, it cannot be considered a good classifier for this data.

Figure 22: Confusion Matrix for the Top Ten Most Independent variables

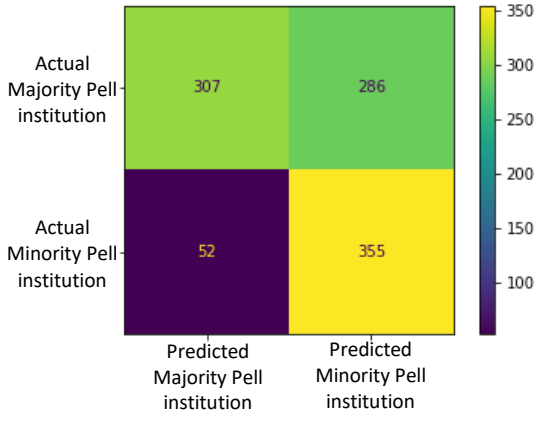
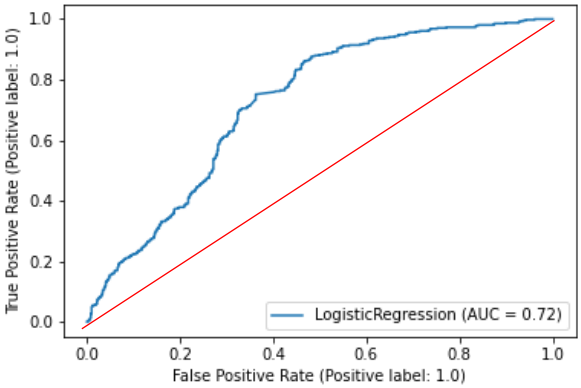


Figure 23: AUC Plot for the Correct/Incorrect Ratio for the Top Ten Most Independent Variables



XGBoost.

The next analysis used an XGBoost algorithm with a learning rate of 0.1, a max depth of 2, a gamma value of 0.5, and 400 estimators to classify the institutions. This analysis was conducted over the full dataset and the dataset reduced to the top ten most independent variables. The parameters listed above were found to produce the best version of the models for both datasets.

For the full original dataset, the model achieved accuracies of 92.6%, 87.4%, and 89.3%, respectively for the training, testing, and validation datasets. It also produced AUCs of 98.9%, 96%, and 96.7%, again respectively. As depicted in figure 24 below, the model produced the highest ratio of correct to incorrect predictions for both majority (494 correct vs. 99 incorrect) and minority (375 correct vs. 32 incorrect) Pell institutions. Since this model produces a high number of correct predictions and a low number of incorrect predictions, I concluded that this model can be considered the best classifier for the original data.

Figure 24: Confusion Matrix for the Full Original Dataset

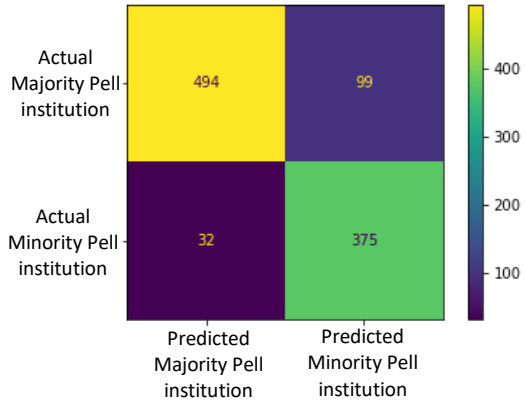
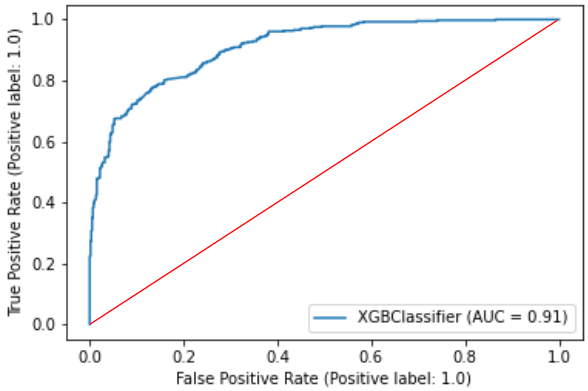


Figure 25: AUC Plot for the Correct/Incorrect Ratio for the Full Original Dataset



The dataset containing only the 10 most independent variables from the original dataset produced the best model for classification based off the original data. The model achieved accuracies of 83.2%, 78.8%, and 80.6%, respectively for the training, testing, and validation datasets. It also produced AUCs of 94.9%, 91%, and 91.7%, again respectively. This model produced a high number of correctly labelled majority and minority Pell institutions but still incorrectly labelled majority Pell schools at a significant rate. Even though this model gets the job done, the goal of this research is to classify the majority Pell schools. On this basis, a model that incorrectly predicts majority Pell schools does not satisfy the goal. Therefore, this model may not be selected as the best classifier for the original dataset.

Figure 26: Confusion Matrix for the Top Ten Most Independent Variables

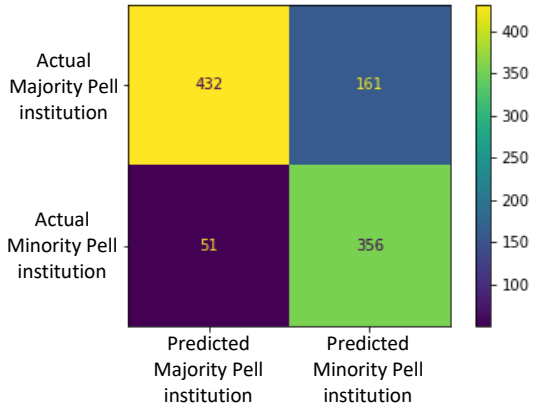
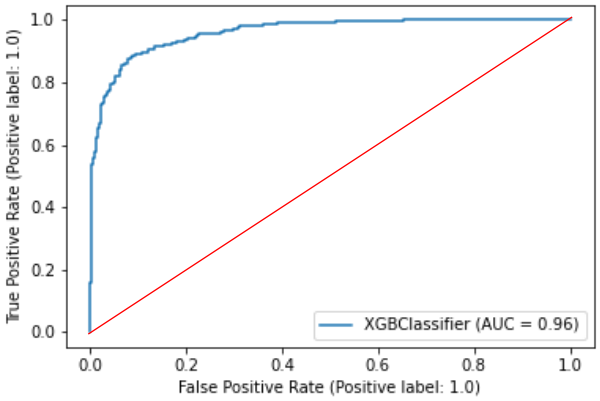


Figure 27: AUC Plot for the Correct/Incorrect Ratio for the Top Ten Most Independent Variables



Random Forest.

The next method used to classify the original dataset was random forest with the entropy criterion, a max depth of 6, a minimum leaf size of 6, and 700 estimators for both the full and reduced datasets.

For the full original dataset, the model achieved accuracies of 97.8%, 84.3%, and 83.2%, respectively for the training, testing, and validation datasets. It also produced AUCs of 99.9%, 93.5%, and 93.4%, again respectively. These results were nearly equivalent to the results of the XGBoost analysis over the reduced dataset, but in this case more minority Pell institutions were incorrectly predicted (57 incorrect in this method vs. 32 incorrect in XGBoost). For this reason, I considered this model to be slightly lesser in predictive ability than the XGBoost model over the reduced dataset.

Figure 28: Confusion Matrix for the Full Original Dataset

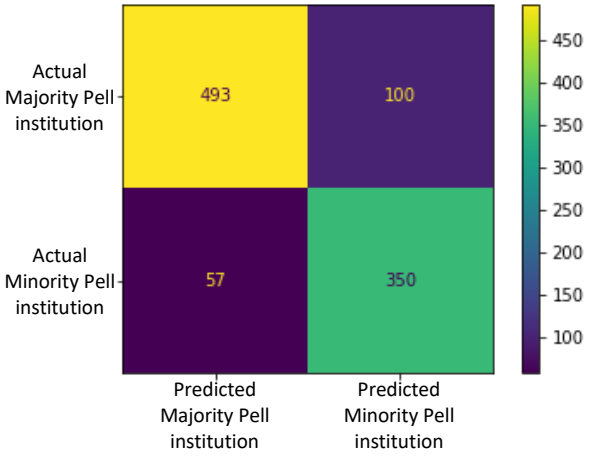
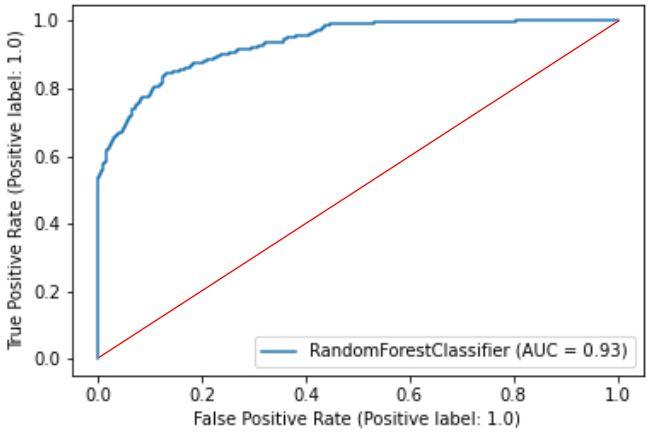


Figure 29: AUC Plot for the Correct/Incorrect Ratio for the Full Original Dataset



The model for the reduced original dataset achieved accuracies of 97.8%, 84.3%, and 83.9%, respectively for the training, testing, and validation datasets. It also produced AUCs of 99.9%, 93.5%, and 93.4%, again respectively. Even though this model produced similar scores for the accuracy and AUC, it produced twice as many incorrect predictions of minority Pell institutions and fewer incorrect predictions for majority Pell institutions. Since this model would accurately predict the highest number of majority Pell institutions, it can be considered a good classifier. However, since it did not produce the highest ratio between correct and incorrect predictions, I considered the XGBoost method to be the better of the two.

Figure 30: Confusion Matrix for the Top Ten Most Independent Variables

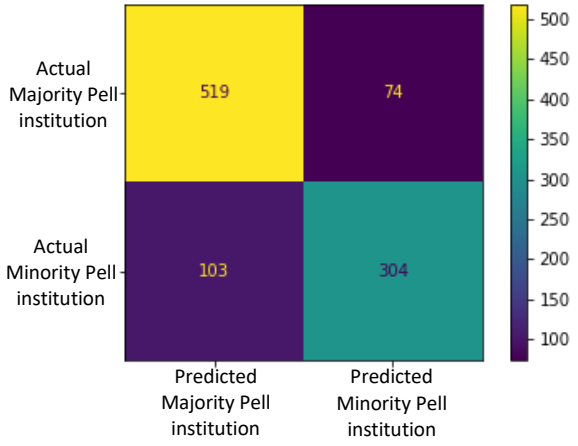
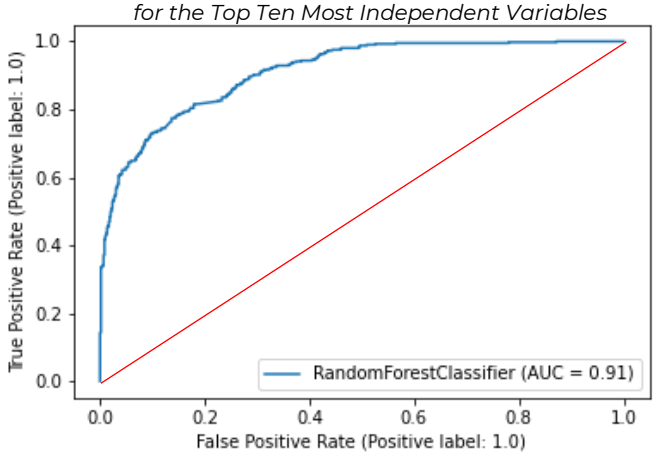


Figure 31: AUC Plot for the Correct/Incorrect Ratio for the Top Ten Most Independent Variables



Analyses on the Standardized Dataset

Principal Component Analysis.

The next analysis considered the standardized version of the original data set. It was found that the first 80 components accounted for 80.2% of the explained variance ratio (Figure 32). The most important feature for the first principal component was typical amount of debt a Pell student accumulates while attending school while the second component was best explained with the number of students in the debt cohort. Figure 33 displays a scatterplot of the 3rd and 8th principal components. The most important features to these components were the number of full-time, first-time students (3rd component) at the university and the far west region of the U.S. (8th component). A lot of overlap in the majority and minority Pell schools are given by the first and second component view, but the view in Figure 33 begins to show group differences.

Figure 32: *Line plot of Explained Variances Ratio by Number of Principal Components of the Standardized Dataset*

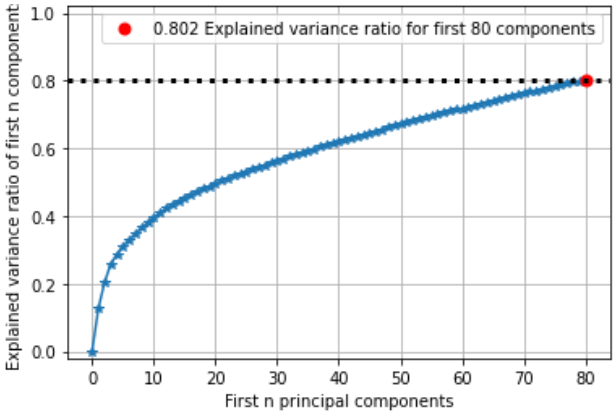
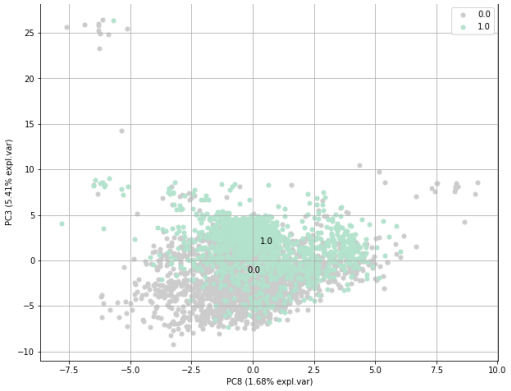


Figure 33: *Scatterplot for the Explained Variance of the Second Principal Component by the First for the Full Standardized Dataset*



I then used the mutual information gain between the variables of the standardized dataset to choose the 10 most independent. I found that 4 components achieve 82.8% of the explained variance (Figure 34). The four most important features for these components are the monthly payments a student who received an award would pay on their loans over ten years (PC1), whether the school is a for-profit private school or not (PC2), the 4-year loan repayment rate for undergraduates that received an award (PC3), and the number of students that are not first-generation in the debt cohort (PC4). Figure 35 displays the first component on the horizontal axis and the second component on the vertical axis. This display shows a large clustering of minority Pell schools (in seafoam) in the upper-right quadrant of the field and that majority Pell schools (in grey) show a positive relationship between the first two components.

Figure 34: *Line Plot of Explained Variances Ratio by Number of Components of the Top Ten Most Independent Variables*

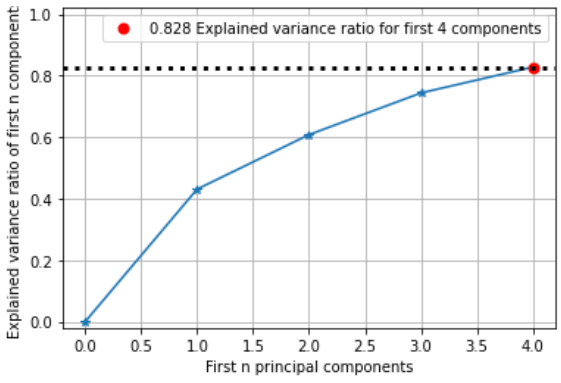
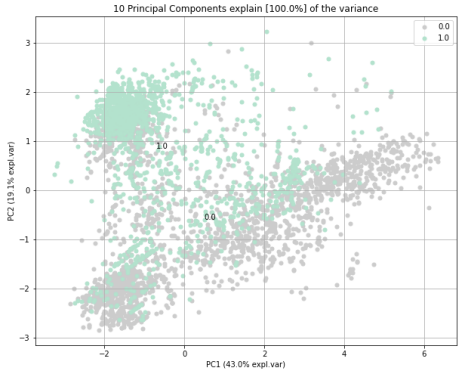


Figure 35: *Scatterplot for the Explained Variance of the Second Principal Component by the First for the Top Ten Most Independent Variables*



Logistic Regression.

For the standardized dataset, I performed a logistic regression using an L1 penalty, a value of 0.05 for the C parameter, and a maximum iteration of 500. This model achieved accuracies of 87.8%, 87.5%, and 86.4%, respectively for the training, testing, and validation datasets. It also produced AUCs of 95.3%, 94.5%, and 94.8%, again respectively for the training, testing, and validation datasets. This model produced a significantly different result than the previous logistic regression on the original dataset. Figure 19 displays the confusion matrix for number of correct and incorrect predictions of majority (510 correct, 83 incorrect) and minority Pell institutions (357 correct, 50 incorrect). These results were comparable to the method I considered the best on the original dataset (XGBoost) but there were fewer misclassified majority Pell institutions and more misclassified minority Pell institutions.

Figure 36: Confusion Matrix for the Full Standardized Dataset

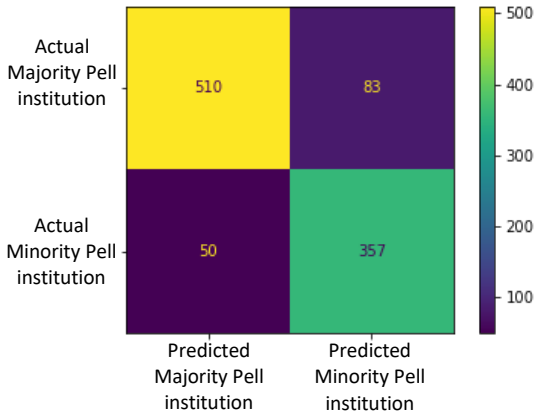
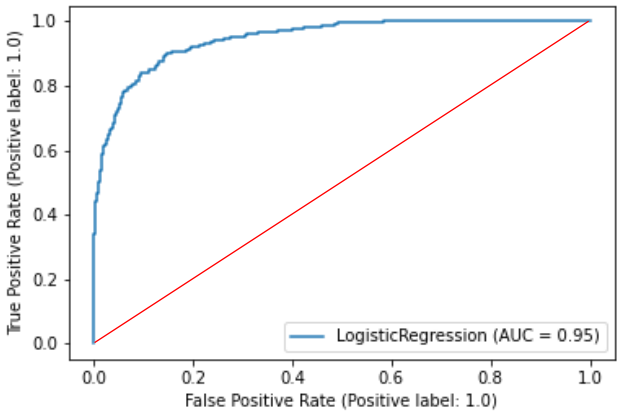


Figure 37: AUC Plot for the Correct/Incorrect Ratio for the Full Standardized Dataset



I then performed another logistic regression on the reduced standardized dataset using an L1 penalty, a value of 0.01 for the C parameter, and a maximum iteration of 50. This model achieved accuracies of 82.7%, 84.6%, and 84.9%, respectively for the training, testing, and validation datasets. It also produced AUCs of 90.5%, 91.3%, and 91.0 %, again respectively for the training, testing, and validation datasets. Similar to the model created on the full dataset, this model performed really well at classifying majority Pell institutions but was not as accurate at predicting minority Pell institutions. For this reason, I concluded that this model was not as good as the model created above.

Figure 38: Confusion Matrix for the Top Ten Most Independent Variables

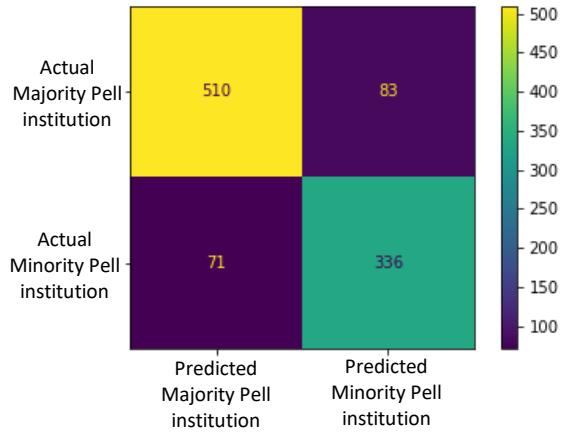
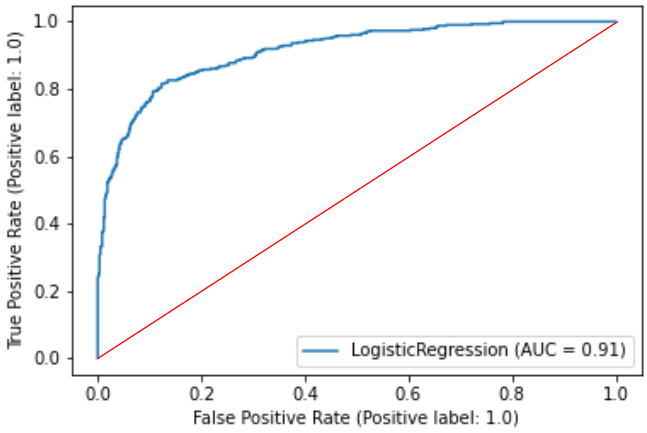


Figure 39: AUC Plot for the Correct/Incorrect Ratio for the Top Ten Most Independent Variables



Logistic Regression with Principal Component Analysis.

For the standardized dataset, I performed a principal component analysis of the standardized dataset and then a logistic regression using an L1 penalty, a value of 1 for the C parameter, and a maximum iteration of 200. This model achieved accuracies of 89.1%, 86.9%, and 87%, respectively for the training, testing, and validation datasets. It also produced AUCs of 96.2%, 94.6%, and 94.6%, again respectively for the training, testing, and validation datasets. The results of this model were very different from the results we found on the original dataset. This model correctly labelled 511 majority Pell institutions while incorrectly labeling 49 minority Pell institutions as majority Pell. In contrast, this model labelled 358 minority Pell institutions correctly while misclassifying 82 majority Pell institutions as minority Pell. Since the number of misclassifications were tolerable, I considered these results sufficient in classifying majority Pell institutions.

Figure 40: Confusion Matrix for the Full Standardized Dataset

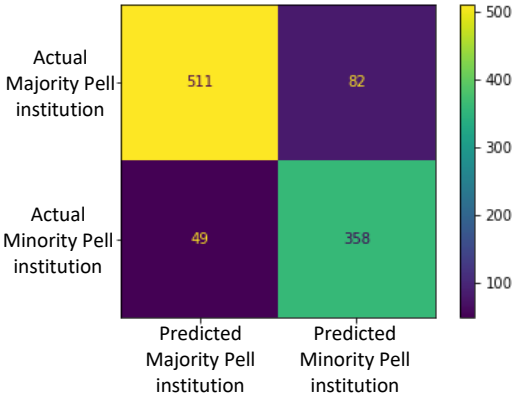
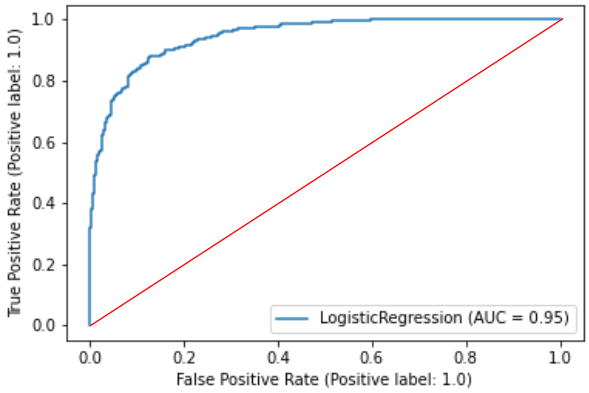


Figure 41: AUC Plot for the Correct/Incorrect Ratio for the Full Standardized Dataset



I then performed this method again on the reduced version of the standardized dataset using an L1 penalty, a value of 0.01 for the C parameter, and a maximum iteration of 200. This model achieved accuracies of 79.8%, 81%, and 80%, respectively for the training, testing, and validation datasets. It also produced AUCs of 87.8%, 89.5%, and 87.6%, again respectively for the training, testing, and validation datasets. This version of this method did not perform as well as the one above. In addition to the accuracy and AUC decreasing, this model misclassified more majority Pell institutions as minority Pell and misclassified more minority Pell institutions as majority Pell. For this reason, I considered other models to be better at classifying my data.

Figure 42: Confusion Matrix for the Top Ten Most Independent Variables

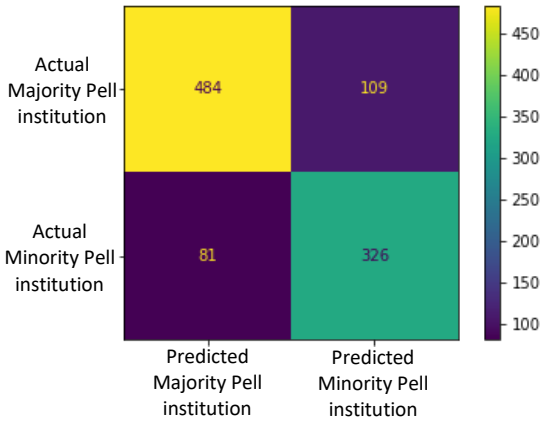
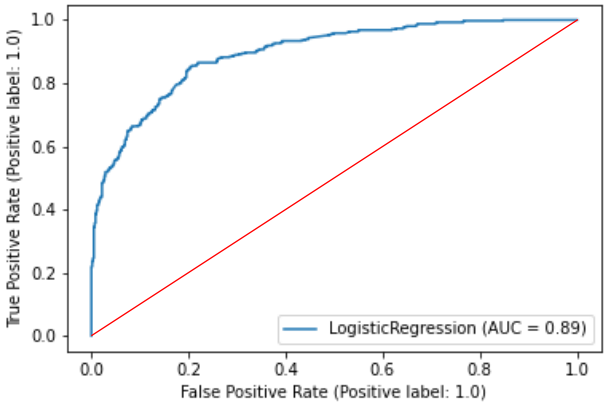


Figure 43: AUC Plot for the Correct/Incorrect Ratio for the Top Ten Most Independent Variables



K-Nearest Neighbors with Principal Component Analysis.

The next analysis used PCA with a k-nearest neighbors with a Euclidean metric, a distance weight, a value of 55 for the number of neighbors in a cluster, and a maximum leaf size of 5. This model achieved accuracies of 100%, 83.1%, and 81.7%, respectively for the training, testing, and validation datasets. It also produced AUCs of 100%, 91%, and 90.4%, again respectively for the training, testing, and validation datasets. I was concerned this model overfits the data, but the confusion matrix in figure 19 shows that this model performed well at classifying the majority and minority Pell institutions correctly. The misclassification this model produced was nearly equivalent. This model incorrectly labelled 82 majority Pell institutions as minority Pell and 87 minority Pell institutions as majority Pell. I considered these test results to be tolerable and I considered this model a good classifier.

Figure 44: Confusion Matrix for the Full Standardized Dataset

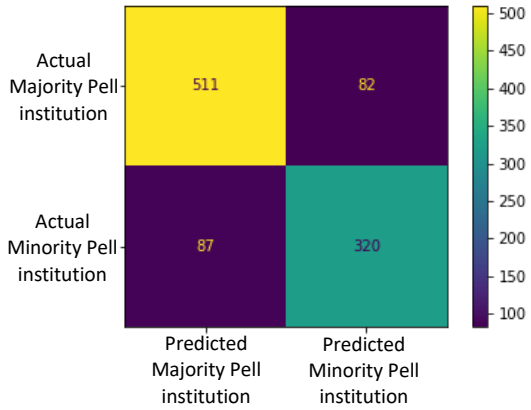
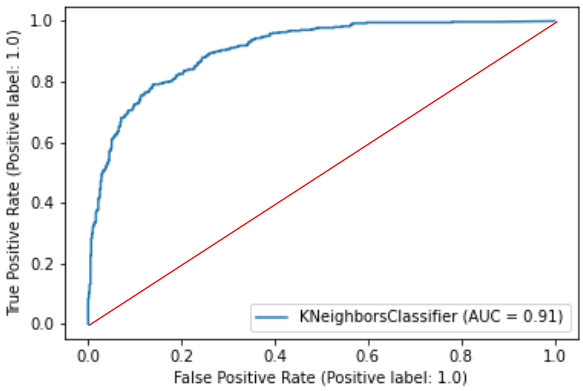


Figure 45: AUC Plot for the Correct/Incorrect Ratio for the Full Standardized Dataset



I then performed the same method on the reduced version of the standardized dataset using a Euclidean metric, a distance weight, a value of 65 for the number of neighbors in a cluster, and a maximum leaf size of 35. This model achieved accuracies of 100%, 84.6%, and 85.8%, respectively for the training, testing, and validation datasets. It also produced AUCs of 100%, 92%, and 92.5%, again respectively for the training, testing, and validation datasets. In comparison to the model created on the full standardized dataset, this model was better at classifying majority Pell institutions but worse at classifying minority Pell institutions. I was still concerned about overfitting but the confusion matrix for the test set showed that this model correctly labelled 534 majority Pell institutions while incorrectly labelling 96 minority Pell institutions as majority Pell. For this reason, I considered this model to be a good classifier.

Figure 46: Confusion Matrix for the Top Ten Most Independent Variables

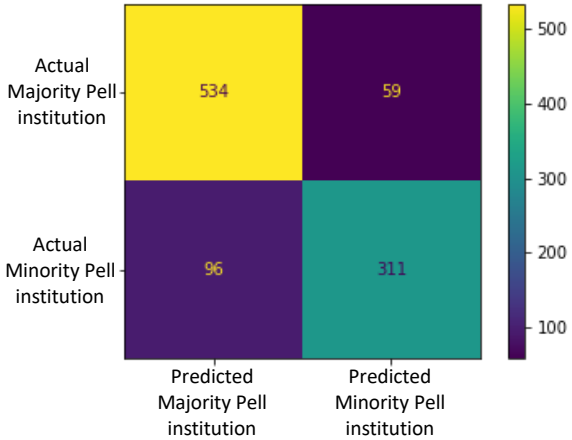
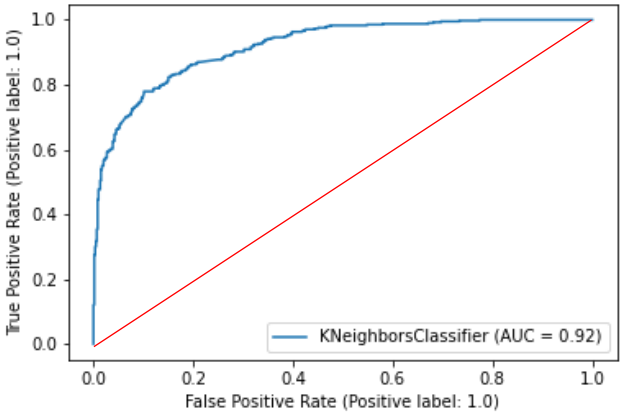


Figure 47: AUC Plot for the Correct/Incorrect Ratio for the Top Ten Most Independent Variables



XGBoost.

For the full standardized dataset, I performed the method XGBoost using a learning rate of 0.1, a max depth of 2, a gamma value of 0.5 and 400 estimators. This model achieved accuracies of 92.63%, 87.4%, and 89.3%, respectively for the training, testing, and validation datasets. It also produced AUCs of 98.9%, 96%, and 96.7%, again respectively for the training, testing, and validation datasets. This model performed extremely well at classifying minority Pell institutions correctly (375 correct, 32 incorrect) and produced tolerable errors in labelling majority Pell institutions (494 correct, 99 incorrect). Since this model produced a tolerable number of misclassifications, I considered it to be good at classifying majority Pell schools.

Figure 48: Confusion Matrix for the Full Standardized Dataset

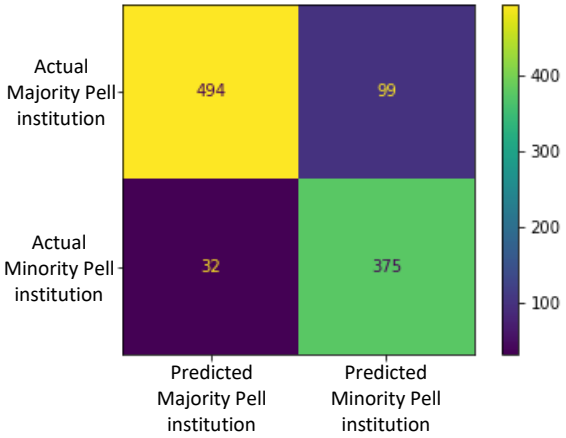
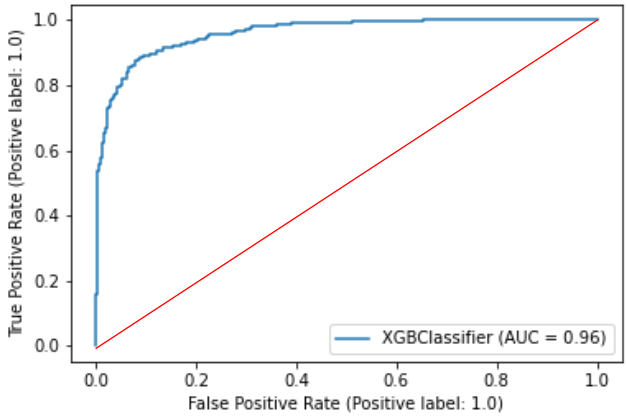


Figure 49: AUC Plot for the Correct/Incorrect Ratio for the Full Standardized Dataset



I then performed this method on the reduced standardized dataset using a learning rate of 0.1, a max depth of 6, a gamma value of 0.5 and 400 estimators. This model achieved accuracies of 94.1%, 80.8%, and 80.5%, respectively for the training, testing, and validation datasets. It also produced AUCs of 99.5%, 91.7%, and 92.4%, again respectively for the training, testing, and validation datasets. This model did significantly worse than the model performed on the full dataset. Since it classified more majority and minority Pell institutions incorrectly, I considered it to be not a good classifier for my dataset.

Figure 50: Confusion Matrix for the Top Ten Most Independent Variables

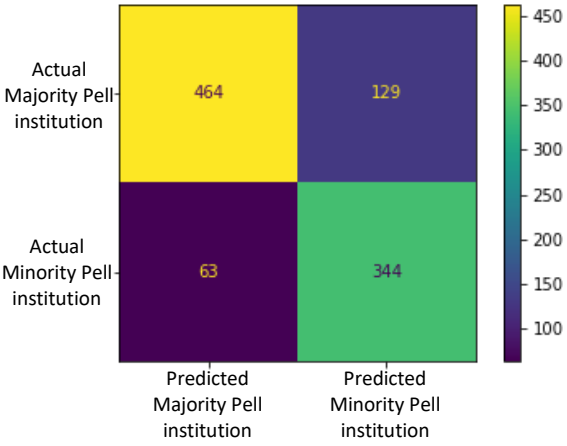
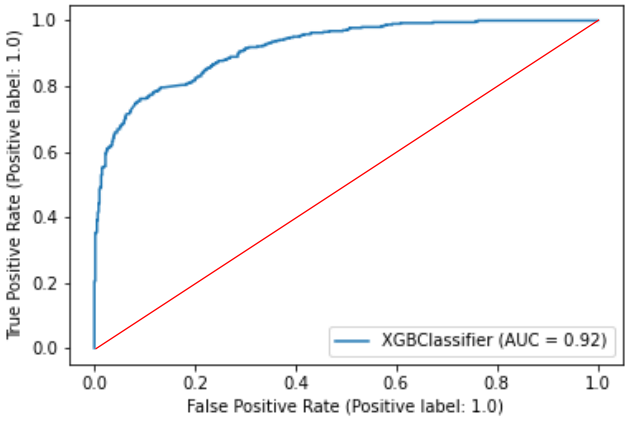


Figure 51: AUC Plot for the Correct/Incorrect Ratio for the Top Ten Most Independent Variables



Random Forest.

The next method I used to classify my data was random forest. For the full standardized dataset, I used the entropy criterion, a maximum depth of 6, a minimum leaf size of 6, and 700 estimators. This model achieved accuracies of 97.8%, 84.3%, and 83.9%, respectively for the training, testing, and validation datasets. It also produced AUCs of 99.9%, 93.4%, and 93.4%, again respectively for the training, testing, and validation datasets. This model did well at classifying minority Pell institutions (350 correct, 57 incorrect) but misclassified 101 majority Pell institutions. Since the goal is to classify majority Pell institutions, I did not consider this model to be the best classifier.

Figure 52: Confusion Matrix for the Full Standardized Dataset

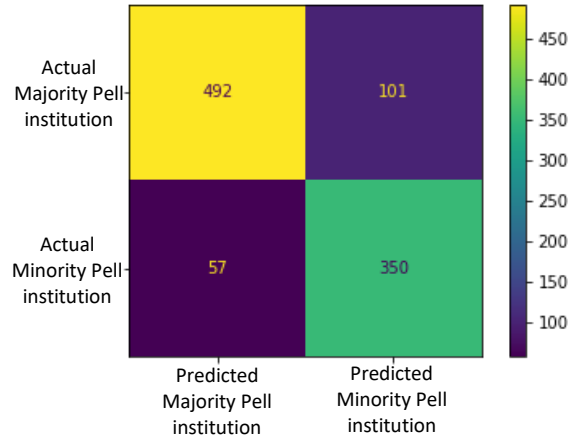
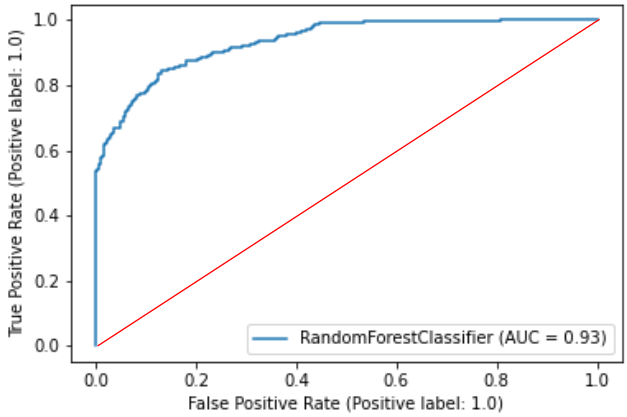


Figure 53: AUC Plot for the Correct/Incorrect Ratio for the Full Standardized Dataset



For the reduced standardized dataset, I performed another random forest using the entropy criterion, a maximum depth of 6, a minimum leaf size of 6, and 700 estimators. This model achieved accuracies of 97.8%, 86.2%, and 88.1%, respectively for the training, testing, and validation datasets. It also produced AUCs of 99.9%, 93.5%, and 93.4%, again respectively for the training, testing, and validation datasets. Since we are trying to classify the majority Pell schools and this model produced the best correct to incorrect ratio and the lowest misclassification of majority Pell institutions, I considered this method to be the best at producing a model that classifies the target variable.

Figure 54: Confusion Matrix for the Top Ten Most Independent Variables

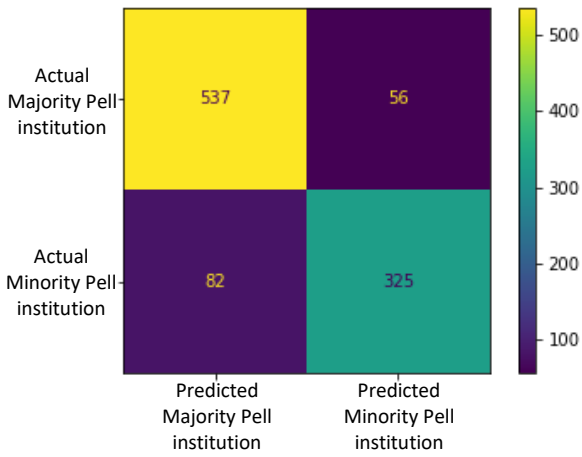
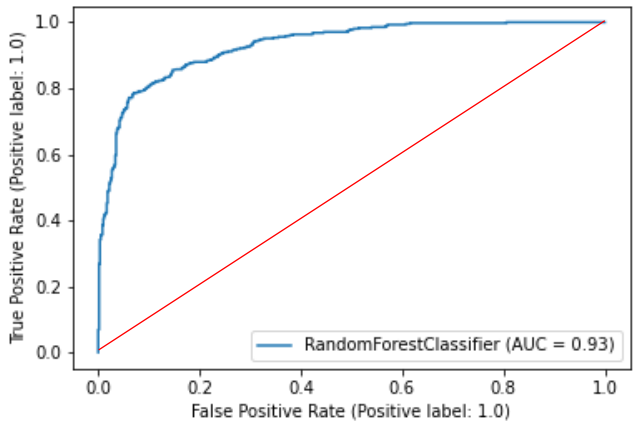


Figure 55: AUC Plot for the Correct/Incorrect Ratio for the Top Ten Most Independent Variables



Analyses on the Standardized Dataset with Variable Transformations

Principle Component Analysis.

This principal component analysis considered the standardized dataset with the addition of various transformed variables. Similar to the analysis for the standardized data, this analysis found that the first 80 components accumulate 80.2% of the explained variance ratio (Figure 56). Of the 80 components, the first four explain 29.05% of the total explained variance, and the most important features to these components were the institution's out-of-state tuition and fees binned into 10 bins (PC1), the number of students who withdrew from their school in the debt cohort (PC2), whether the institution was a for-profit, private school or not (PC3), and the percent of Full-time students that received an award within 8 years (PC4). Figure 58 displays the first component on the horizontal axis and the fourth component on the vertical axis. This plot shows majority Pell institutions (in grey) are more spread out than minority Pell institutions (in seafoam). In particular, the upper right quadrant contains mostly majority Pell institutions while the minority Pell institutions are clustered around the vertical origin.

Figure 56: Line plot of Explained Variances Ratio by Number of Components of the Standardized-Transformed Dataset

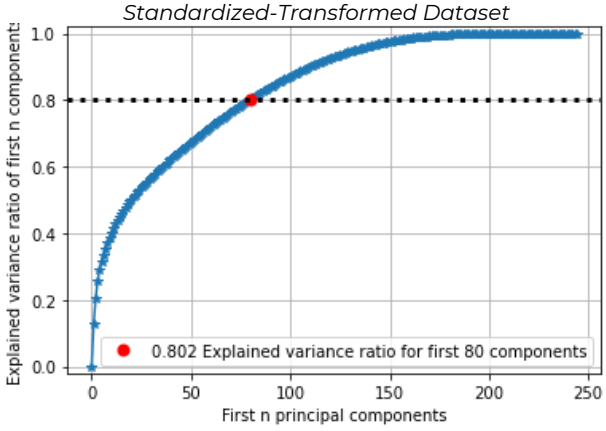


Figure 57: Cumulation of 29% Explained Variance by the First Four Principal Components for the Full Standardized-Transformed Dataset

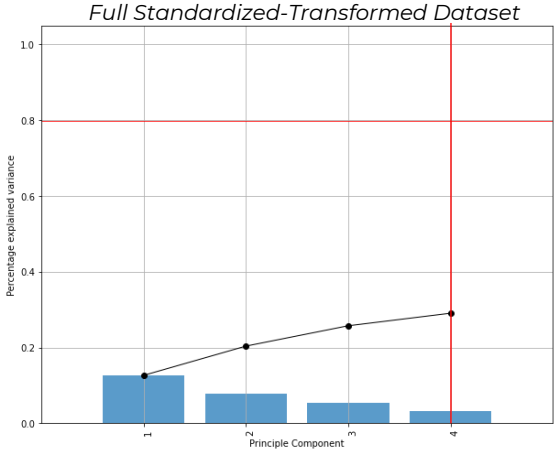
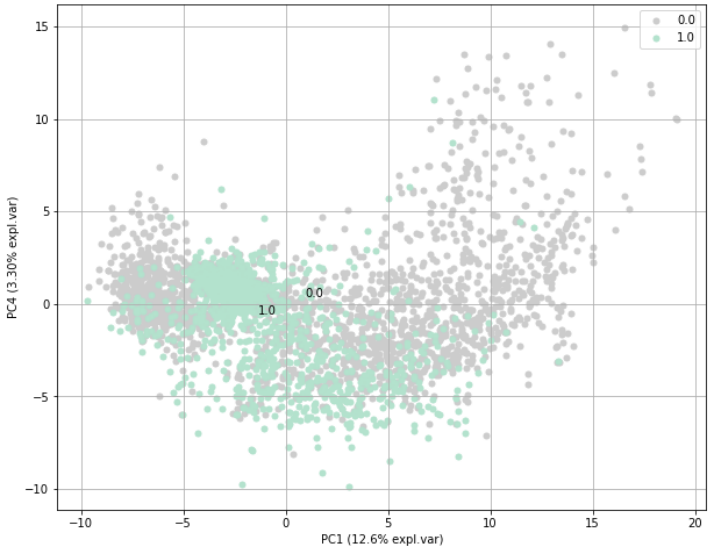


Figure 58: Scatterplot for the Explained Variance of the Second Principal Component by the First for the Full Standardized-Transformed Dataset



For the final principal component analysis, I reduced the standardized dataset with the transformed variables by finding the 10 most independent variables. Of the explained variance, the first three components held 83.6% of the explained variance ratio. These components are the out-of-state tuition and fees for an institution (PC1), the percentage of federal loan borrowers (PC2), the typical amount of debt accrued by a student who withdraws from the institution (PC3), the city that the institution is in (PC4), and the median salary a faculty member at the institution is paid (PC5). Figure 61 displays the second component on the horizontal axis while the vertical axis displays the fifth component. The left side of the graph contains most of the minority Pell institutions (in seafoam) densely clustered around the origin of the vertical axis. Majority Pell institutions (in grey) appear on both sides of the field, but the right side contains institutions that are mostly majority Pell.

Figure 59: Cumulation of 83% Explained Variance by the First Five Principal Components for the Top Ten Most Independent Variables

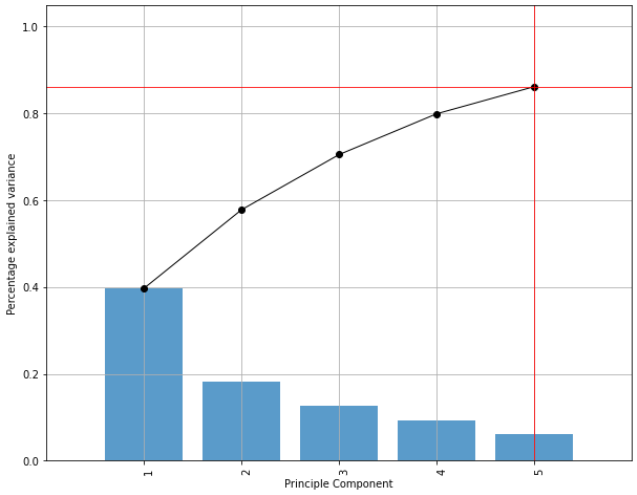


Figure 60: Line plot of Explained Variances Ratio by Number of Components of the Top Ten Most Independent Variables

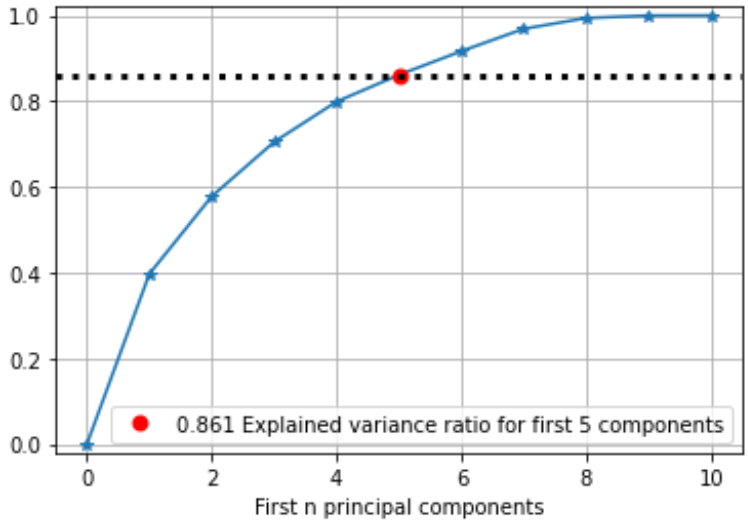
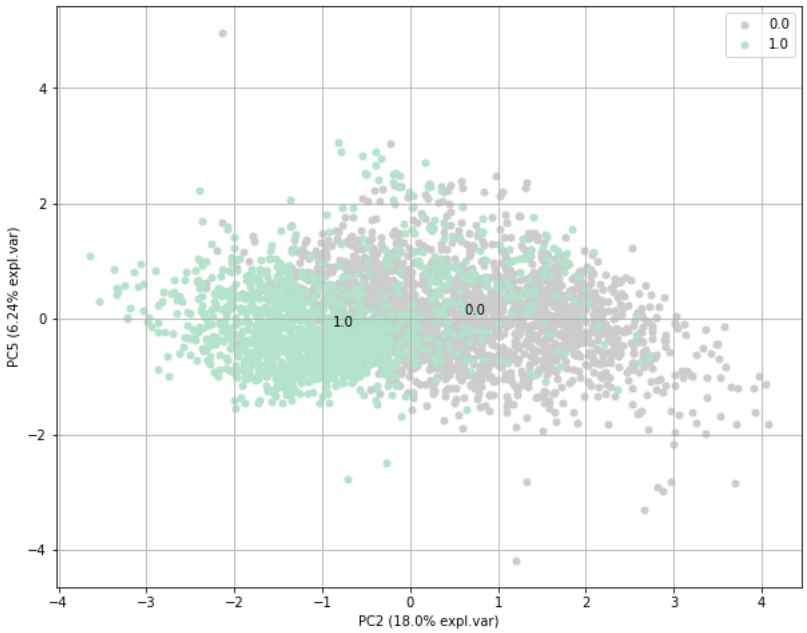


Figure 61: Scatterplot for the Explained Variance of the Second Principal Component by the First for the Top Ten Most Independent Variables



Logistic Regression.

For the standardized dataset with variable transformations, I performed a logistic regression using an L1 penalty, a value of 0.1 for the C parameter, and a maximum iteration of 100. This model achieved accuracies of 83.2%, 82%, and 84%, respectively for the training, testing, and validation datasets. It also produced AUCs of 92.4%, 92.2%, and 92.6%, again respectively for the training, testing, and validation datasets. In comparison to the other logistic regression sections, this model performed better at correctly classifying the institutions. 514 Majority Pell institutions were correctly classified while 79 majority Pell institutions were incorrectly classified as minority Pell. In addition, 349 minority Pell institutions were correctly classified with only 58 incorrectly classified as majority Pell. I considered this error rate to be tolerable and concluded that this model was sufficient in classifying the institutions.

Figure 62: Confusion Matrix for the Full Standardized-Transformed Dataset

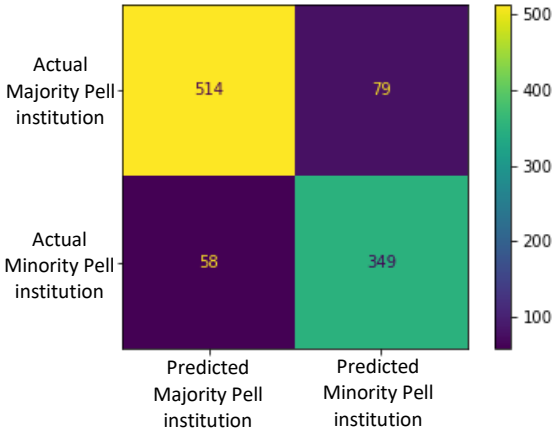
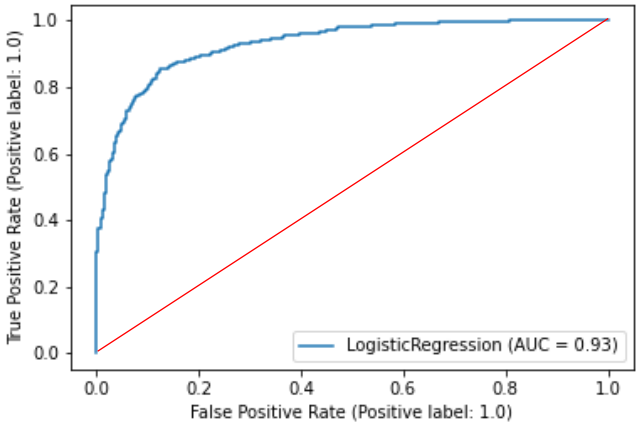


Figure 63: AUC Plot for the Correct/Incorrect Ratio for the Full Standardized-Transformed Dataset



Another logistic regression was performed for the reduced version of the standardized dataset with variable transformations using an L2 penalty, a value of 0.005 for the C parameter, and a maximum iteration of 1. This model achieved accuracies of 80.8%, 82.3%, and 80.9%, respectively for the training, testing, and validation datasets. It also produced AUCs of 89.2%, 90.4%, and 89.1%, again respectively for the training, testing, and validation datasets. This version of the model did not perform any better than the model produced on the full dataset. Instead, this model misclassified more minority and majority Pell institutions. I concluded that this model was sufficient in classifying the institutions but was not the best model found during this project.

Figure 64: Confusion Matrix for the Top Ten Most Independent Variables

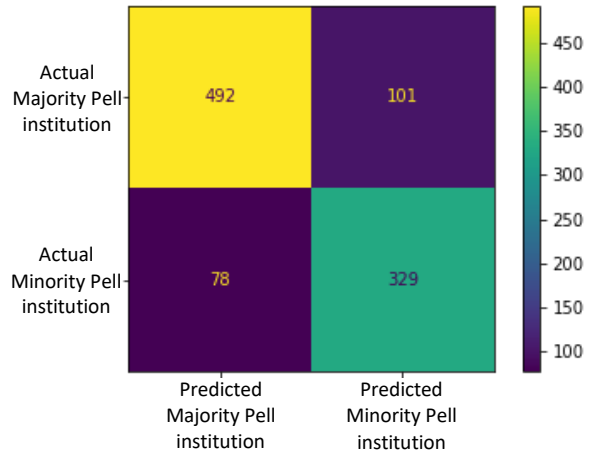
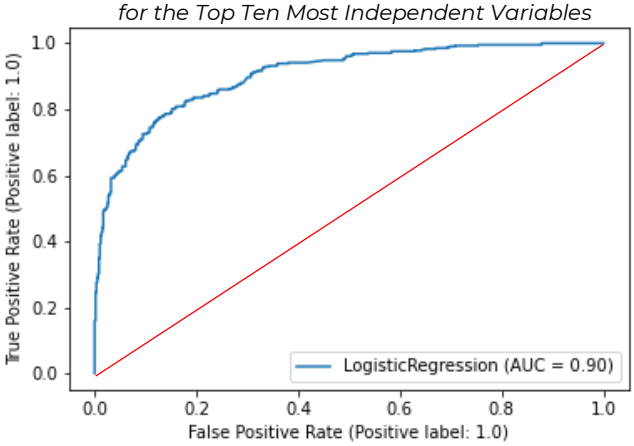


Figure 65: AUC Plot for the Correct/Incorrect Ratio for the Top Ten Most Independent Variables



Logistic Regression with Principal Component Analysis.

The next analysis used PCA with logistic regression using an L2 penalty, a value of 0.01 for the C parameter, and a maximum iteration of 200. This model achieved accuracies of 87.8%, 87%, and 87%, respectively for the training, testing, and validation datasets. It also produced AUCs of 95.6%, 94.6%, and 94.9%, again respectively for the training, testing, and validation datasets. This model produced 510 correctly labelled majority Pell institutions while only misclassifying 83 as minority Pell. In addition, this model correctly classified 359 minority Pell institutions while incorrectly classifying 48 as majority Pell. I concluded that these errors were tolerable and that this model performs well at classifying the data.

Figure 66: Confusion Matrix for the Full Standardized-Transformed Dataset

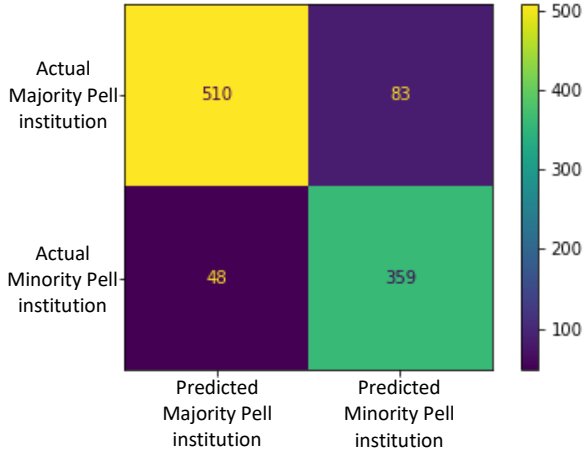
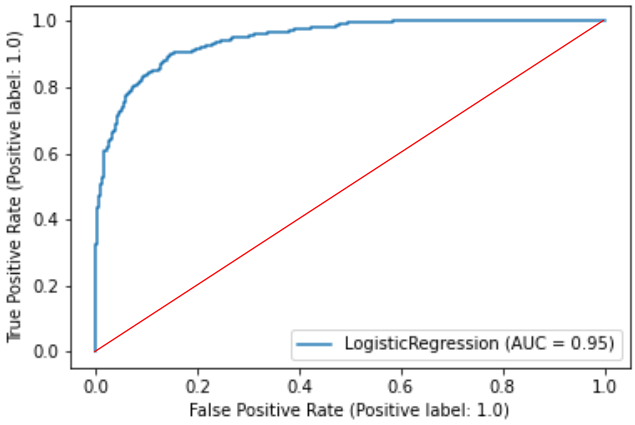


Figure 67: AUC Plot for the Correct/Incorrect Ratio for the Full Standardized-Transformed Dataset



I then performed the same method on the reduced version of this dataset using an L2 penalty, a value of 0.01 for the C parameter, and a maximum iteration of 200. This model achieved accuracies of 80.6%, 82.5%, and 80.7%, respectively for the training, testing, and validation datasets. It also produced AUCs of 89%, 90.4%, and 88.8%, again respectively for the training, testing, and validation datasets. Reducing the dataset to the 10 most independent variables did not aid this method in classifying the institutions. The number of correctly classified majority Pell institutions decreased to 500 while the number of incorrectly classified minority Pell institutions increased to 82. This poses an issue because we would miss a portion of the institutions that have a majority Pell population while taking in a set of institutions that do not.

Figure 68: Confusion Matrix for the Top Ten Most Independent Variables

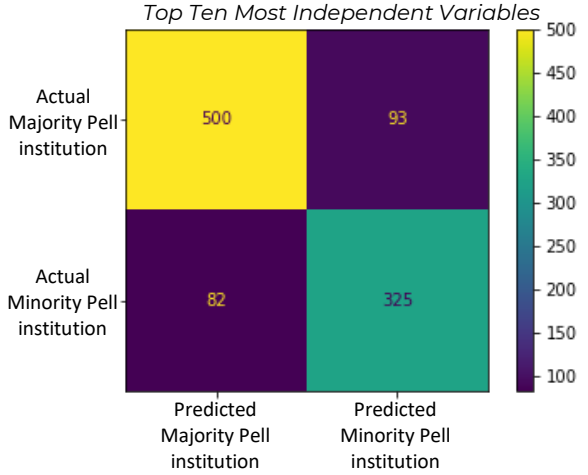
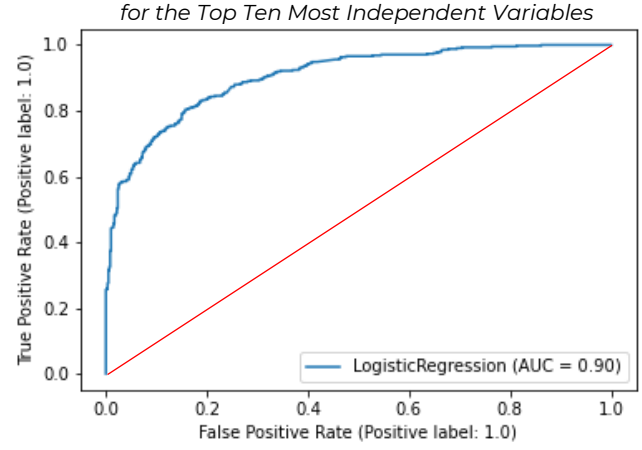


Figure 69: AUC Plot for the Correct/Incorrect Ratio for the Top Ten Most Independent Variables



K-Nearest Neighbors.

The next method I performed on the full standardized-transformed dataset was PCA with k-nearest neighbors using a Euclidean metric, a value of 55 for the number of neighbors, uniformed weights, and a leaf size of 35. This model achieved accuracies of 100%, 85.5%, and 86%, respectively for the training, testing, and validation datasets. It also produced AUCs of 100%, 92.2%, and 92.9%, again respectively for the training, testing, and validation datasets. This model performed very well and correctly classified 535 majority Pell institutions correctly while only misclassifying 62 majority Pell institutions as minority Pell. Even so, this model did not produce the highest number of correctly classified majority Pell institutions. While this method used the best model for the standardized-transformed data, it did not out-perform the overall best model for this project.

Figure 70: Confusion Matrix for the Full Standardized-Transformed Dataset

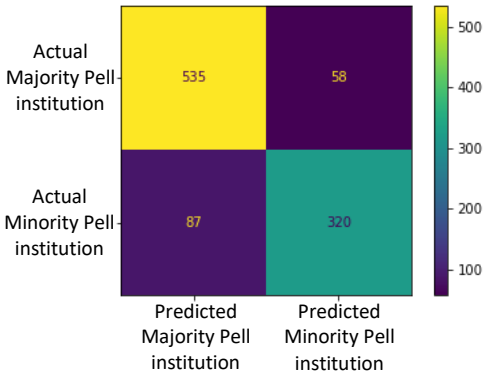
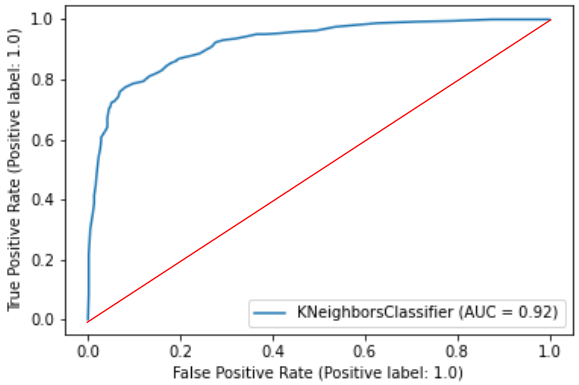


Figure 71: AUC Plot for the Correct/Incorrect Ratio for the Full Standardized-Transformed Dataset



I then performed the same method on the reduced version of this dataset using a Euclidean metric, a value of 55 for the number of neighbors, uniformed weights, and a leaf size of 35. This model achieved accuracies of 99.3%, 86.2%, and 87.6%, respectively for the training, testing, and validation datasets. It also produced AUCs of 99.99%, 93.5%, and 93.9%, again respectively for the training, testing, and validation datasets. Interestingly, this model produced the same number of correctly and incorrectly classified institutions as the method above, but the accuracy and AUC increased for the testing and validation datasets and decreased for the training data. Like the method above this model did not out-perform the overall best model for this project.

Figure 72: Confusion Matrix for the Top Ten Most Independent Variables

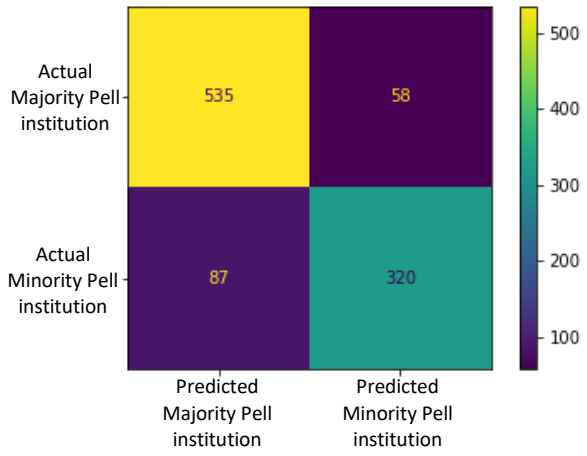
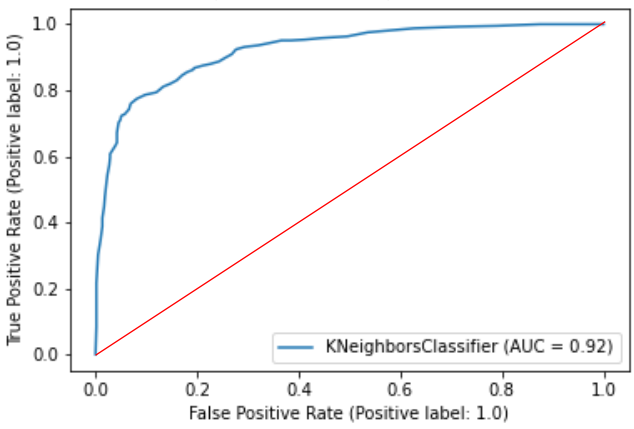


Figure 73: AUC Plot for the Correct/Incorrect Ratio for the Top Ten Most Independent Variables



XGBoost.

For the full standardized-transformed dataset, I performed the method XGBoost using a learning rate of 0.1, a max depth of 2, a gamma value of 0.5 and 400 estimators. This model achieved accuracies of 92.3%, 87%, and 89.3%, respectively for the training, testing, and validation datasets. It also produced AUCs of 98.8%, 96%, and 96.5%, again respectively for the training, testing, and validation datasets. Although this model did well at correctly classifying institutions, it still misclassified majority Pell institutions as minority Pell at a significant rate. For this reason, I concluded that this model was not the best to classify the standardized-transformed dataset.

Figure 74: Confusion Matrix for the Full Standardized-Transformed Dataset

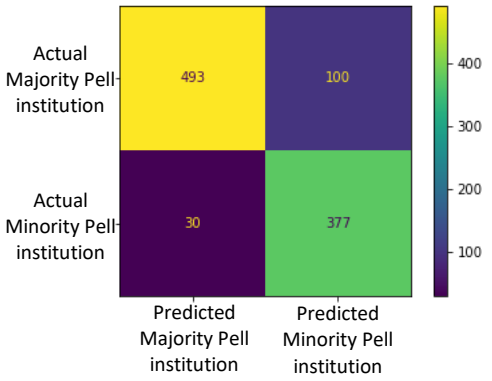
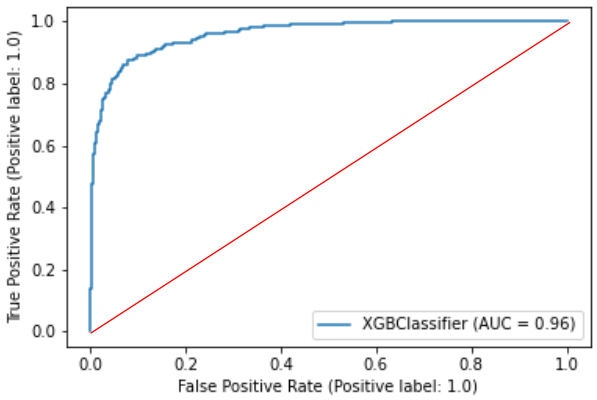


Figure 75: AUC Plot for the Correct/Incorrect Ratio for the Full Standardized-Transformed Dataset



I then performed this method on the reduced standardized dataset using a learning rate of 0.7, a max depth of 4, a gamma value of 0.01 and 1000 estimators. This model achieved accuracies of 94.1%, 80.8%, and 80.5%, respectively for the training, testing, and validation datasets. It also produced AUCs of 99.5%, 91.7%, and 92.4%, again respectively for the training, testing, and validation datasets. In comparison to the model above, this model misclassified a greater amount of minority Pell institutions while correctly classifying a negligibly greater amount of majority Pell institutions.

Figure 76: Confusion Matrix for the Top Ten Most Independent Variables

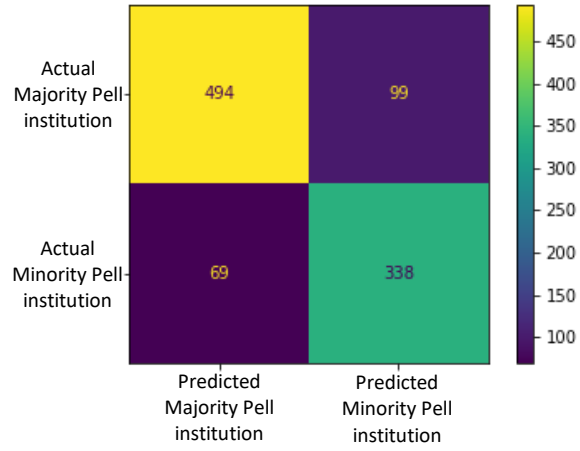
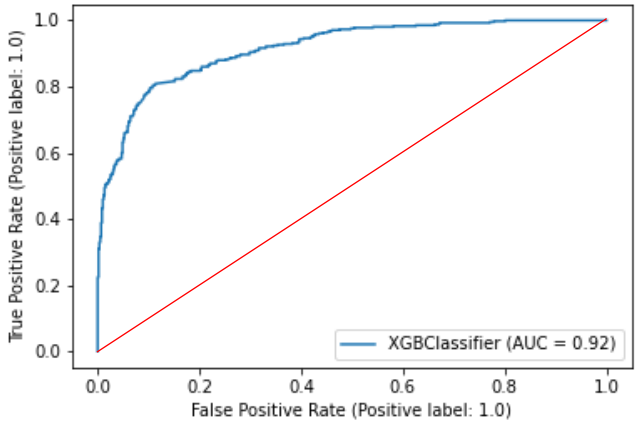


Figure 77: AUC Plot for the Correct/Incorrect Ratio for the Top Ten Most Independent Variables



Random Forest.

The next method I used to classify my data was random forest. For the full standardized dataset, I used the entropy criterion, a maximum depth of 6, a minimum leaf size of 6, and 700 estimators. This model achieved accuracies of 97.8%, 84.3%, and 83.9%, respectively for the training, testing, and validation datasets. It also produced AUCs of 99.9%, 93.4%, and 93.4%, again respectively for the training, testing, and validation datasets. This model did well at classifying minority Pell institutions (350 correct, 57 incorrect) but misclassified 95 majority Pell institutions as minority Pell. Since the goal is to classify majority Pell institutions, I did not consider this model to be better than the overall best model.

Figure 78: Confusion Matrix for the Full Standardized-Transformed Dataset

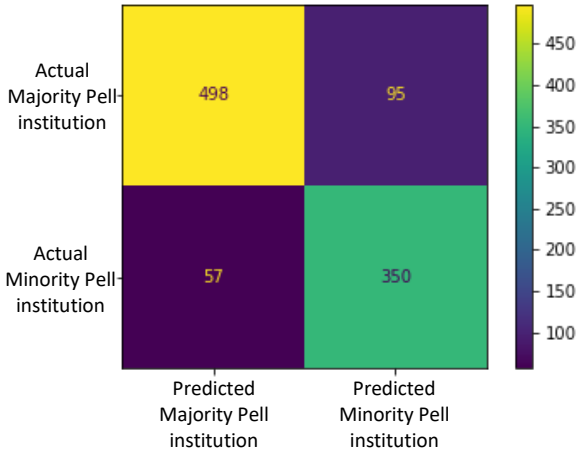
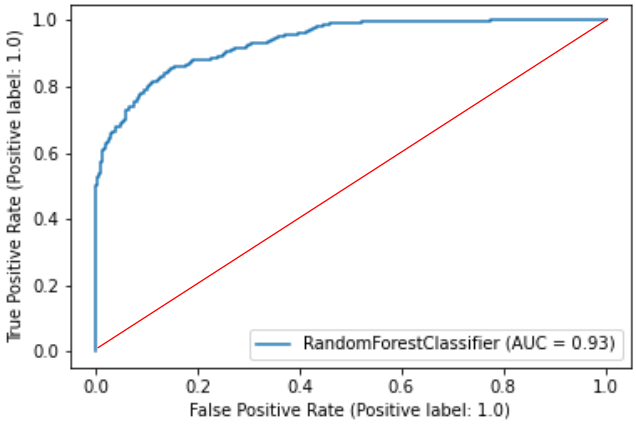


Figure 79: AUC Plot for the Correct/Incorrect Ratio for the Full Standardized-Transformed Dataset



For the reduced standardized dataset, I performed another random forest using the entropy criterion, a maximum depth of 6, a minimum leaf size of 6, and 700 estimators. This model achieved accuracies of 97.8%, 86.2%, and 88.1%, respectively for the training, testing, and validation datasets. It also produced AUCs of 99.9%, 93.5%, and 93.4%, again respectively for the training, testing, and validation datasets. This model did well at classifying majority Pell institutions (519 correct, 74 incorrect) but misclassified 90 minority Pell institutions as majority Pell. While this was one of the better models for this project it still did not out-perform the overall best model.

Figure 80: Confusion Matrix for the Top Ten Most Independent

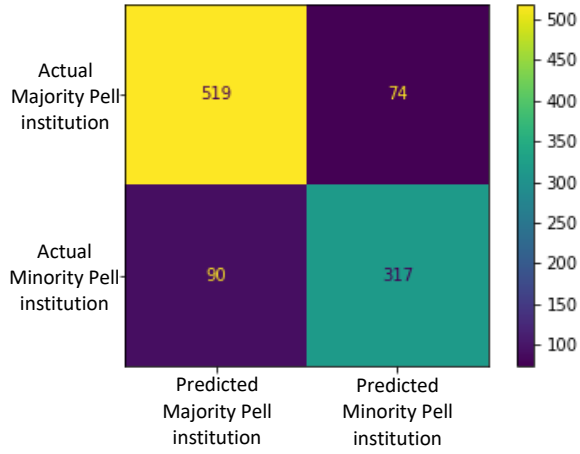
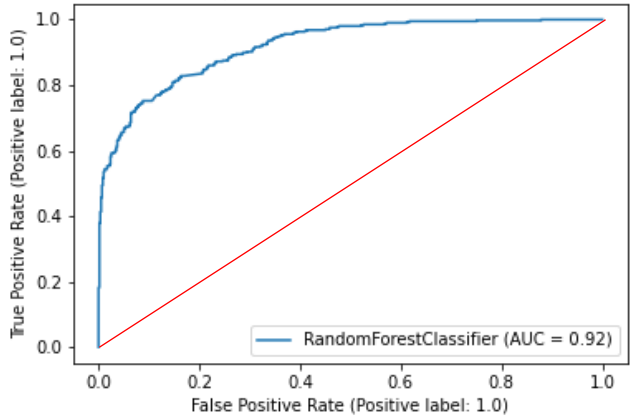


Figure 81: AUC Plot for the Correct/Incorrect Ratio for the Top Ten Most Independent Variables



Overall Conclusion:

I compared all 30 of these models against each other based on their accuracies, areas under the curve (AUC), and ratio of correct to incorrect classifications of majority and minority Pell institutions and found that the original data set with no alterations produced the least helpful models while standardization led to better classification.

Poorest Classification Methods

The worst models included the logistical regression and the logistical regression with principal component analysis (PCA) models for the original, full dataset. These models had an affinity for properly classifying only either majority Pell schools (Figure 16, pg. 9; original dataset with logistic regression) or minority Pell schools (Figure 20, pg. 10; original dataset with logistic regression with PCA). These models also exhibited low accuracies and AUCs in the training, testing, and validation datasets.

Best Classification Methods

The method using k-nearest neighbors on the standardized-transformed dataset produced the second-best model for classifying majority and minority Pell institutions. This model correctly classified 531 majority Pell institutions and 310 minority Pell institutions, while only making a total of 159 incorrect classifications across both groups. These metrics were important for the decision-making process because the accuracy and AUC of the training dataset reached 100% each while the testing and validation dataset produced respective accuracies of 85.5% and 86% and AUCs of 92.2% and 92.9%.

Using a random forest on the standardized dataset reduced to the top 10 most independent variables produced the number one model for classifying majority and minority Pell institutions. This model correctly classified 537 majority Pell institutions and 325 minority Pell institutions with only 138 incorrect classifications in total. It also produced the highest accuracies (97.8% training set, 86.2% testing set, 88.1% validation set) and AUCs (99.9% training set, 93.5% testing set, 93.4% validation set) of the top two models. With this information, I concluded that this model was sufficient in classifying the data and, in fact, the best model out of all 30 created.

Discussion

This project focused on exploring different methods and identifying which models worked best for the data in the CollegeScorecard dataset. The goal was to classify majority Pell institutions and explore the features that are most important to this classification of post-secondary institutions. I selected the variables for modeling using an imputation threshold, PCA, and categorical inference, rather than using methods such as forward/backward selection or Wald scores, because this research centered around the method selection process. Future research will focus on implementing variable selection processes in tandem with the method selection processes identified here to create even better models for classification of majority and minority Pell institutions.