Nathaniel Jones Stat4310: Data Mining

Classification Project  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:**

**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.

*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 was 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 groupings which include: ID variables, variables that related to a Classification of Instructional Program (CIP) code, variables related to institutional demographics, variables related to an institutions attending-student demographics, variables specifying the institutions geolocation, variables specifying the academic outcomes at each university, and variables describing an institutions typical loan outcome proportion, 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%.

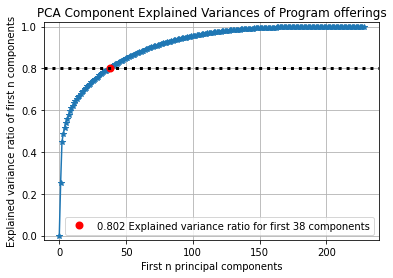
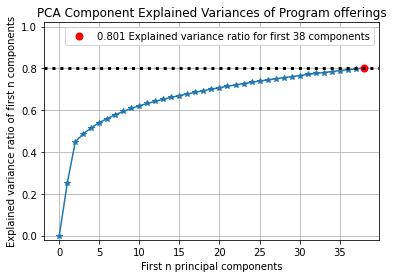
This method resulted in poor model performance and training speed. I replanned my method and backtracked to before One-hot encoding the categorical features. A new plan of action was formed, I decided to perform Principal Component Analysis on individual subgroups listed above to reduce the number of 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 that, 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.

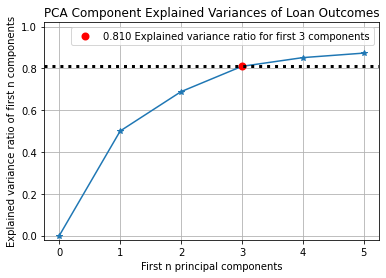
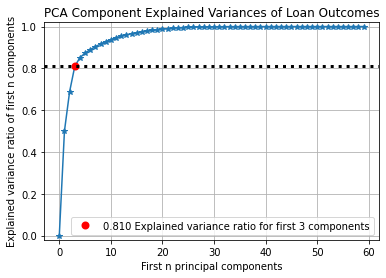


From the plot above, we can see that the first 38 Principal Components reached an explained variance ratio of 0.802. It is interesting that 38 Principal Components accounted for 80.1% of the explained variance ratio because my dataset included 38 different CIP code programs. With just the top five components, 54.1% of the explained variance ratio is 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*

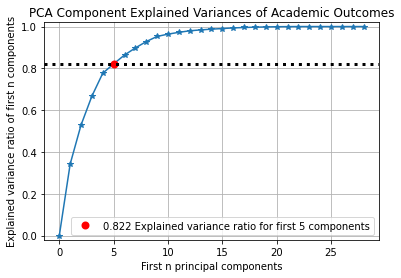
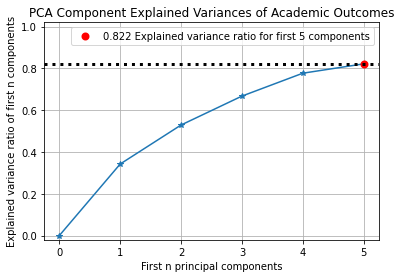
There were 64 variables that described different median loan outcomes or adjusted cohort counts for the loan outcomes of students after they exit their institution. These variables are collected for the 1, 2, 4-, 5-, 10-, or 20-years cohorts. It was found that three components accounted for 81% of the explained variance. 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 second component was best represented by the loan repayment rate of undergraduates who received an award 4-years after graduating.
* The third component was best represented by the number of parents in the 1-year loan outcome cohort who received a PLUS loan.



*Variables related to Academic Outcomes:*

There were 28 variables related to academic outcomes a student may exit their institution with in either 6 years (150% completion time) and 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, still enrolled at the same institution after 8 years without receiving an award. For each of these outcome groups, four variables are recorded for a combination 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.



From the model, 82% of the explained variance was gained from the first five components. 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).

I think it is very interesting that the proportion of full-time students’ who is still enrolled without receiving an award after 8 years was an important feature. This is because 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:*

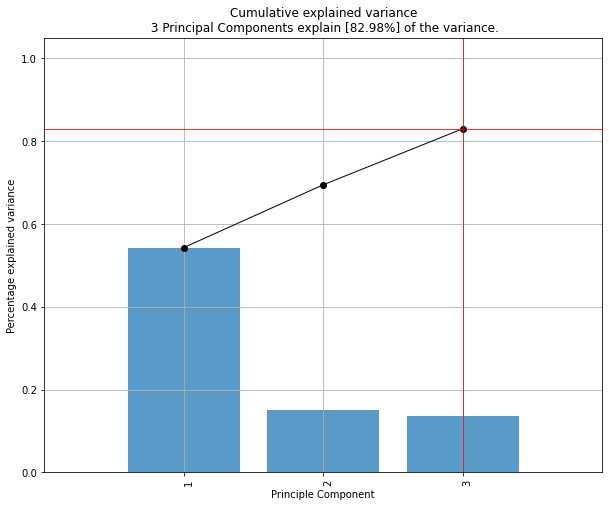
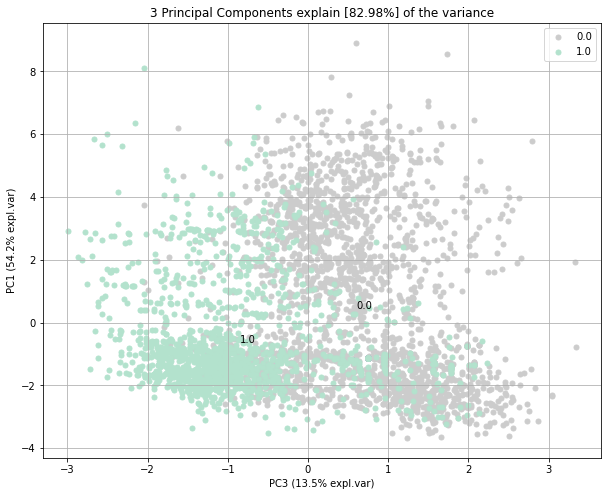
I created models for the original dataset, a standardized version of the original dataset, and a standardized version of the original dataset with various variable transformations. For each dataset I ran a Principal Component Analysis, Logistic Regression, XGBoost algorithm, K-nearest neighbors, and 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 variable selections will return values equal to zero. I decided that 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.**

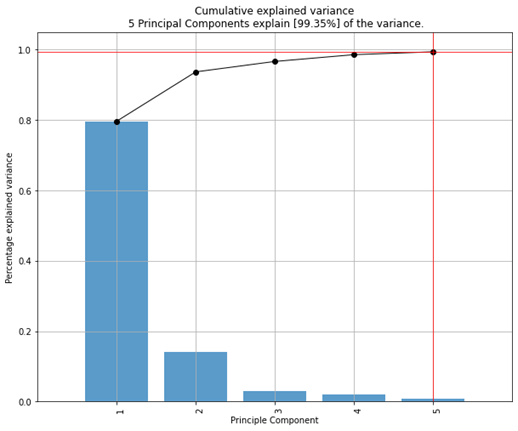
The first analysis considered all variables from the original data set and found that the first 3 components accounted for 83% of the explained variance. Figure 2 below visually displays the difference in the explained variance for each of the three components. The first principal component accounted for 54.2% of the explained variance while the other two components explained less than 15% each. Figure 3 displays a scatter plot 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 clustering of majority Pell schools (in grey) in the top right quadrant of the plot.

**Figure 3: Variance of Principal Component 3 by Principal Component 1**

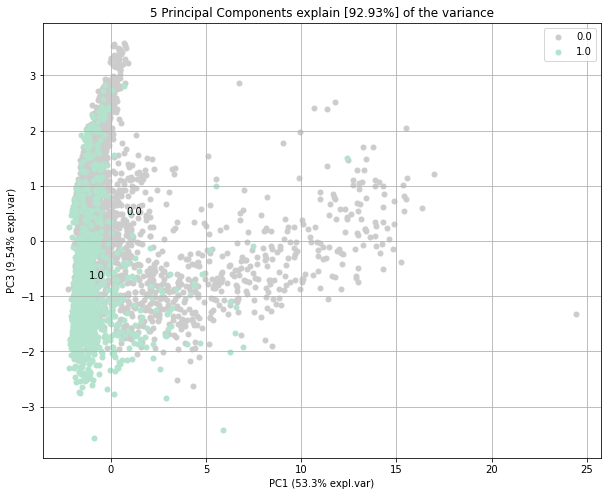


**Figure 2: Cumulative Explained Variance by the 5 Principal Components**

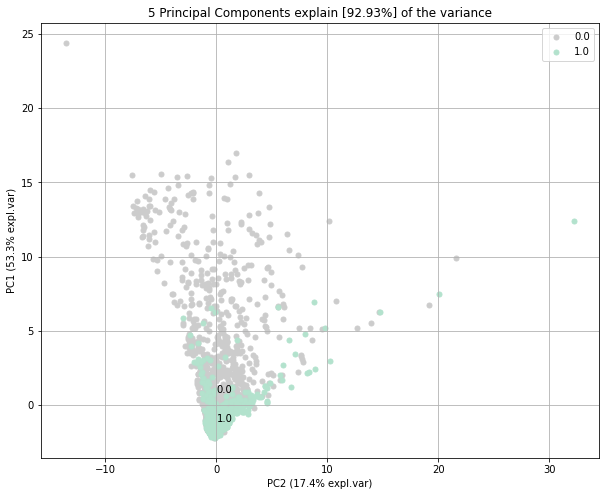
After selecting 10 variables from the original dataset and performing the analysis again, I found that the first five components in the model accumulated 99.3% of the explained variance. Figure 4 shows that the first component made up nearly 80% of the explained variance which was greater than four times 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 5 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 6 on the next page 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 4: Cumulative Explained Variance by the 5 Principal Components**



**Figure 5: Variance of Principal Component 3 by Principal Component 1**



**Figure 6: Variance of Principal Component 1 by Principal Component 2**

**Logistic regression.**

Start typing here.

FULL

REDUC

**Logistic regression with Principal Component Analysis.**

Start typing here.

FULL

REDUC

**k-Nearest Neighbors.**

Start typing here.

FULL

REDUC

**XGBoost.**

Start typing here.

FULL

REDUC

**Random Forest.**

Start typing here.

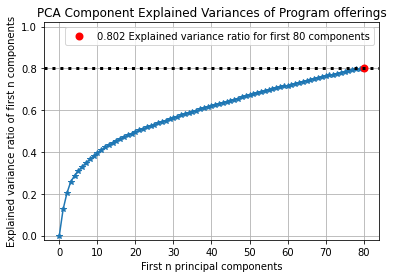
FULL

REDUC

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 10). 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 12 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 12 begins to show group differences.



**Figure 10: PCA Component Explained Variances of Program Offerings**

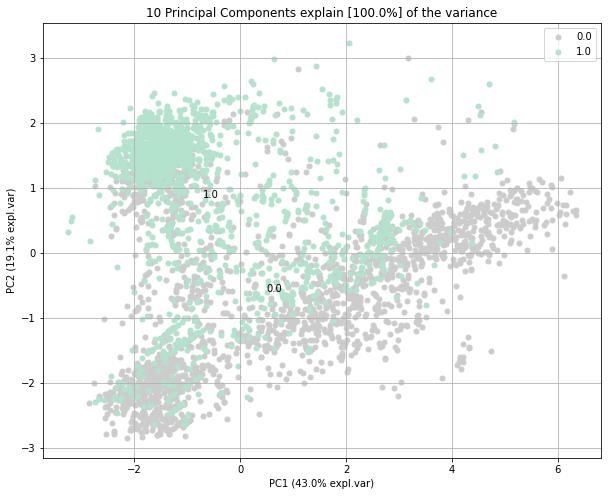
**Figure 12: Variance of Principal Component 7 by Principal Component 3**



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 13). 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 14 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 13: PCA Component Explained Variances of Program Offerings**



**Figure 14: Variance of Principal Component 2 by Principal Component 1**

**Logistic regression.**

Start typing here.

FULL

REDUC

**Logistic regression with Principal Component Analysis.**

Start typing here.

FULL

REDUC

**k-Nearest Neighbors.**

Start typing here.

FULL

REDUC

**XGBoost.**

Start typing here.

FULL

REDUC

**Random Forest.**

Start typing here.

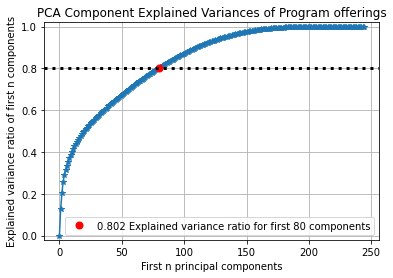
FULL

REDUC

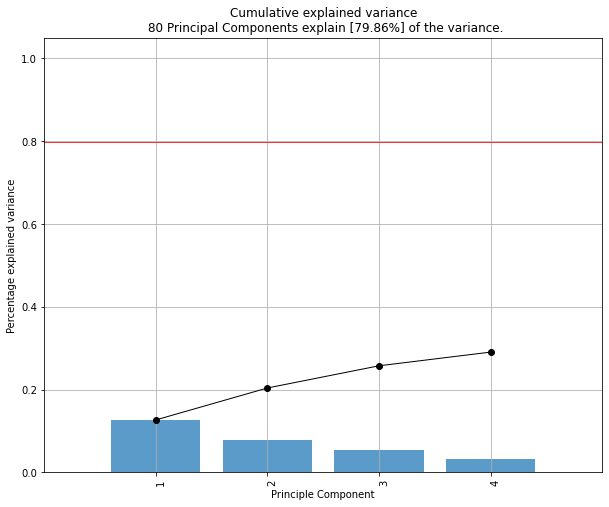
Analyses on the Standardized Dataset with Variable Transformations

**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 18). Of the 80 components, the first four explains 29.05% of the total explained variance and the most important features to these components are the institutions 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 institutions are for-profit private school or not (PC3), and the percent of Full-time students that received an award within 8 years (PC4). Figure 20 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.

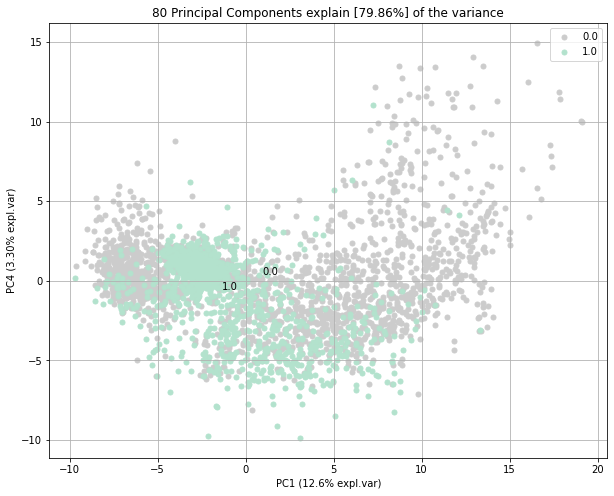


**Figure 18: PCA Component Explained Variances of Program Offerings**

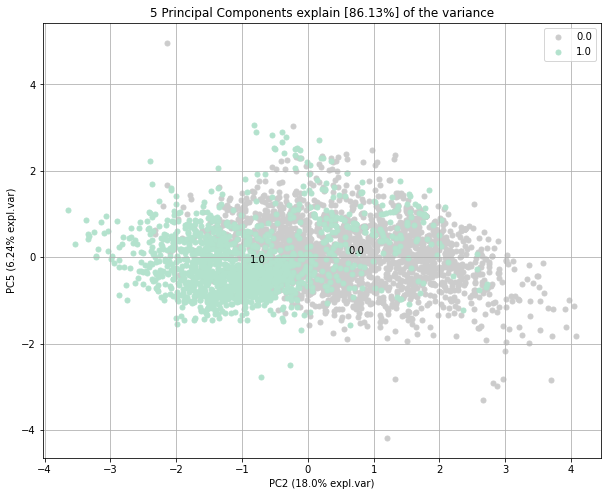


**Figure 19: PCA Component Explained Variances of Program Offerings**

**Figure 20: Explained Variance of Principal Component 2 by Principal Component 4**

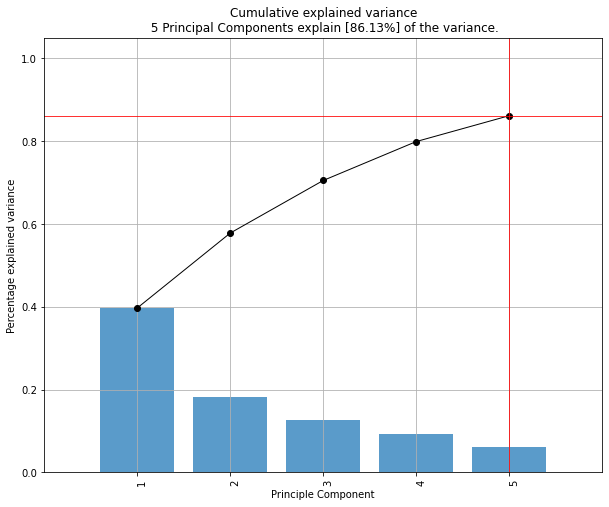


For the final principal component analysis, I reduced the standardized dataset with the transformed variables by finding the 10 most independent variables. Of the explained variance, the first three components held 83.6% of the explained variance ratio. These components are the out-of-state tuition and fees for an institution (PC1), the percentage of federal loan borrows (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 19 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.

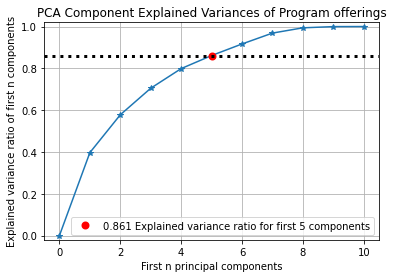


**Figure 19: Explained Variance of Principal Component 2 by Principal Component 5**

**Figure 20: Cumulative Explained Variance by the Top 5 Principal Components**



**Figure 18: PCA Component Explained Variances of Program Offerings**



**Logistic regression.**

Start typing here.

FULL

REDUC

**Logistic regression with Principal Component Analysis.**

Start typing here.

FULL

REDUC

**k-Nearest Neighbors.**

Start typing here.

FULL

REDUC

**XGBoost.**

Start typing here.

FULL

REDUC

**Random Forest.**

Start typing here.

FULL

REDUC

*Overall Conclusion:*