Education Project

Jared Wilber, Shannon Chang, Nura Kawa, Manuel Horta

December 5, 2016

concordance=TRUE

Abstract

Our client is a non-profit NGO with an interest in minority success. Primarily, they'd like to identify schools that underserve minorities so that they can donate money to those schools in an effort to increase minority enrollment. They don't want their money to go to waste, so they'd like to target the best schools they can. We help in two ways. First, we provide a method for identifying those schools which underserve minorities. Second, we define a data-driven metric value metric that can be used to rank the schools.

1 Introduction

To help the NGO, we'll need data. We're utilizing data from [INSERT URL]. This dataset provides federally reported information about universities in the United States of America. We'll utilize this data to help the NGO achieve the following objective: determine whether a given university underserves minorities. By underserves, we mean has less than the US median percentage of minorities enrolled. We also create a data-driven metric that ranks the value of universities. Our analysis culminates in a user-friendly app that the NGO will use to identify whether or not a university serves minorities. The app also lists similar universities ranked in order of the value metric we created.

In this manner, the NGO can employ us in the following 2 ways:

1. Given multiple minority serving schools, determine which should receive funding.

2. Given a school, predict whether or not it will underserve or overserves minorities.

We realize the first goal with our created metric. We realize the second goal with our classifier, which takes in some features as input and outputs a binary label: whether or not a school underserves or adequately serves minorities.

1.1 Data

As stated previously, the data is freely available at collegescorecard.ed.gov. The data contains multiple datasets, each corresponding to a different year. Because of fluctuations regarding data completion (i.e. some datasets are more sparse than others), we opted to use the most recent dataset, as it was relatively dense. Furthermore, this dataset is more likely to reflect present day trends.

The dataset lives in very high-dimensions (roughly eighteen-hundred features), so our first order of business was to reduce dimensions. Data-reduction is important because it allows for more interpretable results,and it's crucial that our NGO understand our methods. We also took efforts to clean the data, such as imputing NA values and "PrivacySuppressed" values. We also removed columns which were over fifty percent sparse, as these are essentially useless. Following this, we used regular expressions to clean up university names and subset our data to four-year universities only. This resulted in a much sparser dataset, but one that still had a couple hundred of features (about five-hundred to be exact). Further dimensionality-reduction efforts are discussed later, with particular emphasis given to interpretability.

Because our goal is to identify minority serving schools, we need some feature to reflect minority enrollment. This metric was created as follows:

Using the data documentation we identified columns in our dataset that measured the percentage of students that identified as: Black, Asian, Hispanic, Pacific Islander, or two or more races. We simply added these columns to create a metric that measures the percentage of minorities that a school has.

To determine whether or not a school underserves or adequately serves minorities, we compared the above metric to the corresponding percantage of minorities in the US, via data we found online from the US Census.

We also create a metric by which to rank universities. This is discussed in more detail in the analysis section.

From a high-level, the entire data-munging is as follows:

1. Hand-select important features from data. This yielded about 500 variables.
2. Handle unruly data (e.g. NA values, NULL values, etc.)
3. Create our own variables: minority, ranked, and bestvalue
4. Subset data based on gradient boosted tree importance

Those methods in the above listed data-munging not already addressed will be discussed in the following analysis section.

2 Analysis

Our analysis consists of two primary objectives:

1. Define metric used to rank schools.

2. Define model used to predict minority serving.

Ranking schools

In order to determine the value of schools, we need some metric. Because the NGO wants a data-driven solution, we created this value in a data-driven manner. Creating this metric consisted of the following steps, discussed in turn:

1. Scrape US NEWS ranked universities and create ranked feature

2. Run LASSO with ranked feature on response

3. Run 'PCA' on obtained results, use loadings from first component as weights

4. Create a weighted metric that mirrors that which US NEWS uses, using the above PCA-weighted features.

Our first step was to create a column that detailed whether or not a given university was ranked or not. To achieve this, we scraped US NEWS rankings online from the Washington Post and merged them into our dataset. After the data was merged, we created a new binary feature that identified whether or not a university was ranked.

Once we had this ranked feature, we used it as a response in a LASSO regression model. We did this because LASSO will select features associated with the response; thus, this provided us with a much smaller set of features (HOW MANY) most associated with a university being ranked. They features made sense intuitively and included things such as average debt level and graduation rates.

Following this, we ran PCA on our smaller set of features. The scree plot for the PCA looked great, and is shown below:

We used the loadings from the first component of our PCA as weights for our LASSO-selected features, with the idea being that they'll help provide more weight to the features. I should note that this idea (using the PCA loadings as weights in this manner) was recommended by Professor Gaston Sanchez. This newly created metric is called QUALITY_INDEX.

Finally, after obtaining a new, much smaller set of weighted features related to ranking, we created our metric. In creating this metric, we wanted to mirror the ranking process by the 'holy grail' of school rankings, US News & Reports. The ranking is formulated as follows:

BV_SCORE = .60*(QUALITY_INDEX/DISCOUNTED_TOTAL_COST) + .25*PCTPELL + .15*(DISCOUNTED_TOTAL_COST/STICKER_PRICE)

This metric was then normalized to be between zero and one-half, for interpretability.

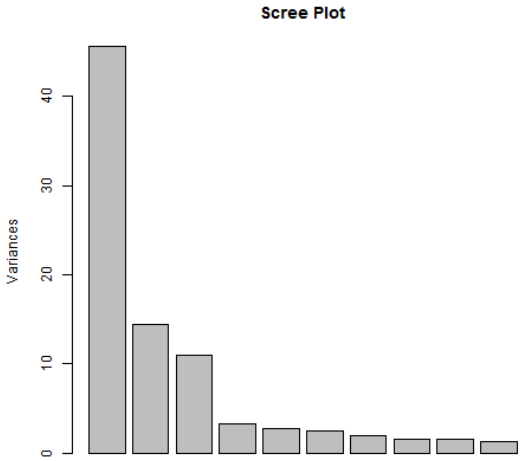


Figure 1: **Scree plot.**

Train XG-Boost model

To classify whether or not schools underserve or adequately serve minorities, we use a gradient boosted tree classification model. We use this model for three primary reasons: 1 - It's a very powerful classification technique with a high degree of accuracy. 2 - It handles correlated data without issue. 3 - It is interpretable in that it provides us with a means to rank features based on their importance.

At a high-level, a gradient boosted tree is just an ensemble of weak decision trees that performs well. Gradient boosted tree works as follows: First, we train a weak model on our data, with data drawn according to some weight distribution. Then, we keep track of how our model performed by increasing the weight of misclassified samples and decreasing the weight of correctly classified samples. Following this, we train another weak model on samples of data from our updated weight distribution. Thus, the algorithm iteratiely trains models on data that is 'difficult' to classify. This process results in an ensemble of models that are good at learning different parts of our dataset. Our boosted tree is this ensemble of models.

To test our model, we split our dataset into a training set and a testing set. We used a 70 percent split. We then used cross-validation to identify which hyper-parameters corresponded to the best model, and used them to fit a new and improved model. Finally, as our dataset is still too large for easy interpretability, we map out the feature importance of our variables with the respect to our classification model. The benefit of our boosted tree model is that we can easily identify feature importance. We simply check the amount

that each feature was used in the construction of the decision trees in the model. That is, for each tree we calculate the amount of performance improved from splitting on a given feature. We do this for all of the utilized features across all of the trees, and then average these results. In this way we have a ranking of values that are important to our model, as seen below:

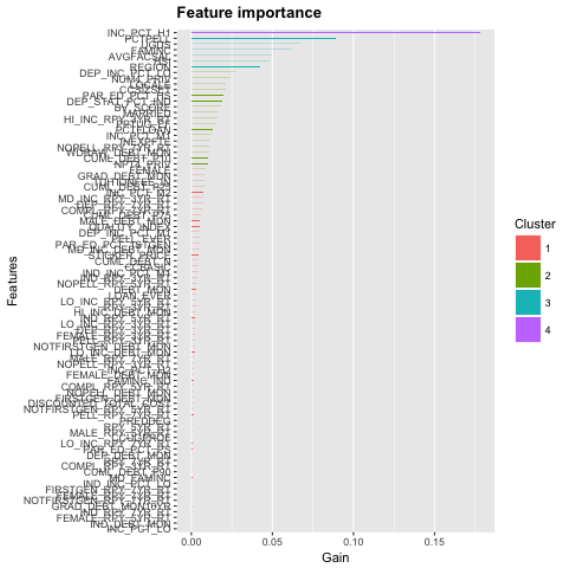


Figure 2: **Sorted significant variables.**

Because we want an easiliy useable product, we elect to choose the top ten features. We then fit another boosted-tree to this model, and repeat the training and cross-validation process. The feature importance for our new variables is displayed below:

Not only is this new model much more interpretable, but it is very accurate; with 95 percent accuracy we can predict whether or not a university underserves or adequately serves minorities.

Now, we have a much smaller, much more interpretable data set and a very accurate model. Thus, we've created a product that the NGO can easily use: they show us 10 components of a given school [INSERT COMPONENTS], and we'll tell them whether or not they should be targeted for minority-driven funding.

Step 1 of the above is self-explanatory and we selected the following variables (column names):

Variable names:

Step 2 of the above is more involved. To create BEST VALUE, we first scraped the wh_post (Washington Post best school rankings) for rankings. Once we had our ranked schools and appended them to our dataset, we used

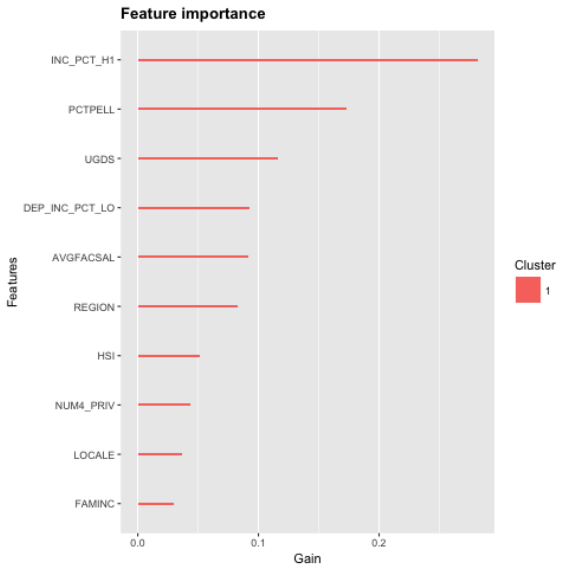


Figure 3: **Top 10 sorted significant variables.**

LASSO/XGBOOST to pick those features that were associated with a school being ranked (we treat this as a proxy for quality). We then used these variables to build a BEST-VALUE metric.

insert feature importance

For step 3, we computed another feature importance metric, this time with our response being the minority served classification. We get rid of variables that don't contribute to our model. Because we are predicting discrete categories of minority serving, we use xgboost for feature importance.

Step 4 is to check for correlated variables and to remove them as well.

Finally, we have our dataset. Now our dataset is much smaller **and** more relevant to our goals. Thus, our data is less computationally expensive, has higher prediction power, and is more interpretable.

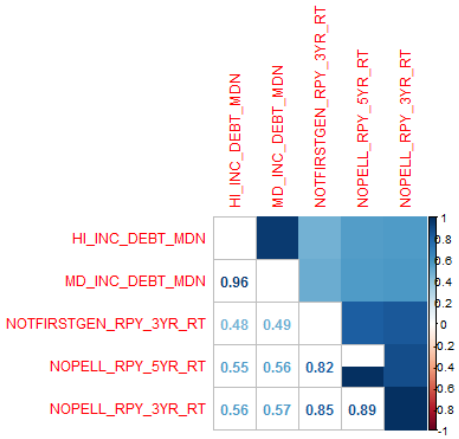
We end up with a data set consisting of 110 variables. These are then sorted by significance to minorities using xgboost.

2.1 Exploratory Data Analysis

After creating the dataset, variable relationships were analyzed with exploratory data analysis.

Correlation Plot

The above correlation plot is just a small sample of our features. However, the trend remains the same: our variables feature a relatviely high-degree of



correlation. Because we'll need this data for a classification task (predicting minority service), it's important to recognize this as it can effect our classification. For this reason, we use a gradient-boosted tree for our analysis, as it easily handles correlated data.

The PCA plot for our best value metric shows a pretty good degree of separation. We can see that the colors appear to disperse in groups, with similar scores clustering together.

PCA of our minority classification metric also shows some degree of separation. It appears that the primary source of separation between minority counts is across the y-axis.

The feature plot reveals the degree to which our minority metric varies by feature. For a lot of features, there appears to be little variation/effect. Although some features show different values from others, the majority show little fluctuation between minority serving amount. We'll investigate which features affect minority serving in more detail later.

Because the PCA results were unconclusive, we opted to use another technique to view our high-dimensional data. Whereas PCA will reduce the dimensions of the dataset (and thus mutate it), t-SNE (t-stochastic neighborhood embeddings) won't change our data. Instead, it computes a distance between our observations, then plots that distance in small dimensions. In this case, our data is plotted into two dimensions and color-coded by Best-Value.

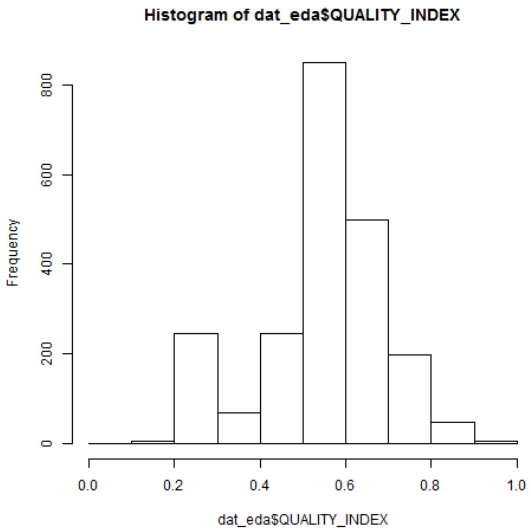


Figure 4: **Histogram of the Quality Index.**

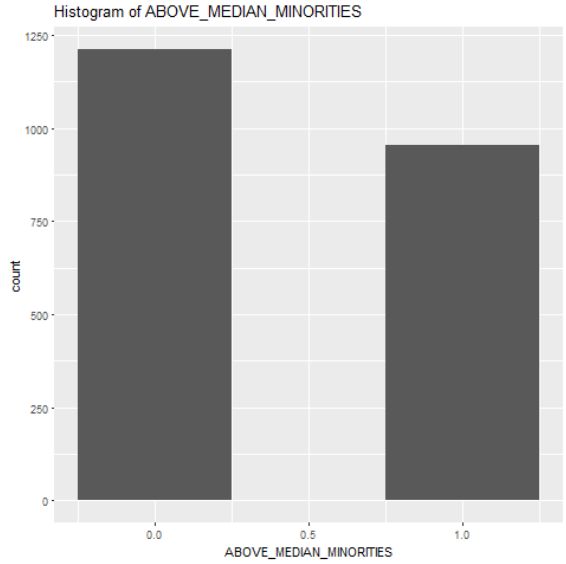


Figure 5: **Histogram of ABOVE_MEDIAN_MINORITIES.**

Figure 6: **Boxplots of School Value by Minority Count per Region.**

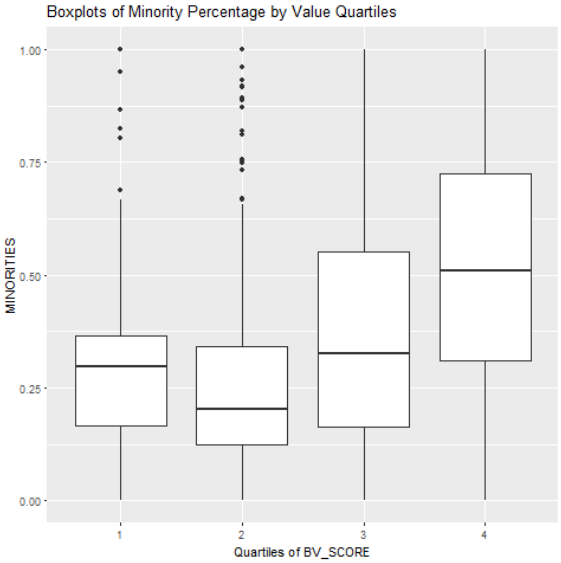


Figure 7: **Boxplots of Minority Percentage by Value Quartiles.**

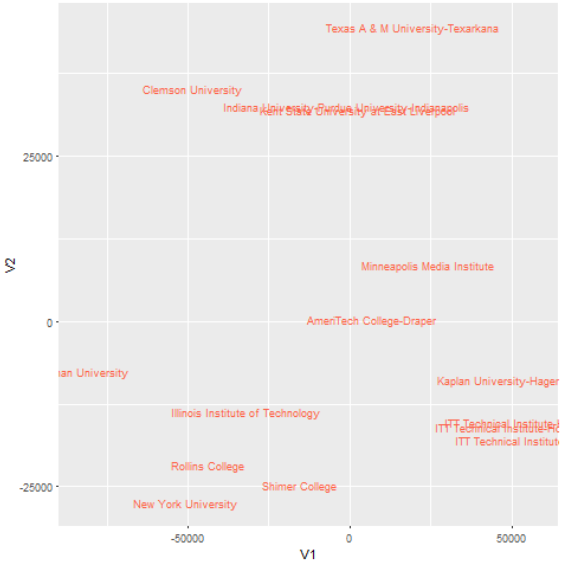


Figure 8: **IDK WHAT THIS IS.**

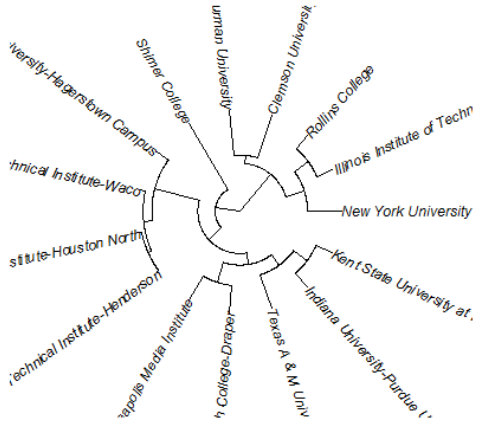


Figure 9: **Phylo Tree.**

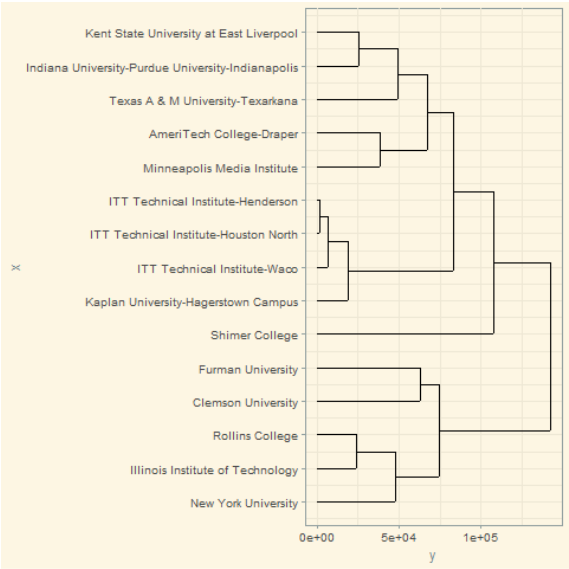


Figure 10: **HIERARCHICAL CLUSTERING.**

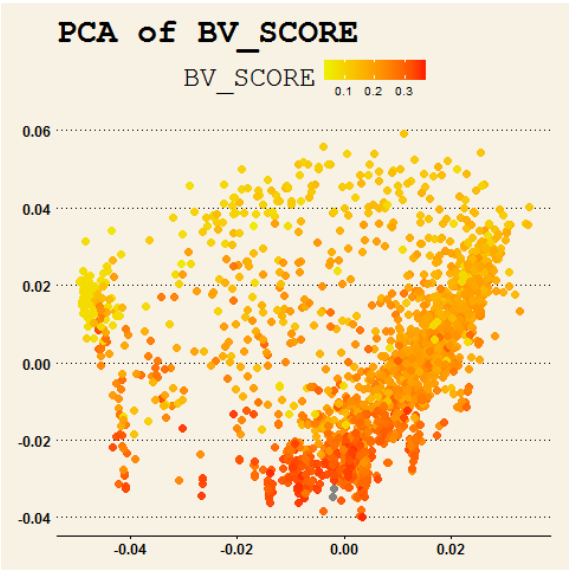


Figure 11: **PCA of QUALITY_INDEX**

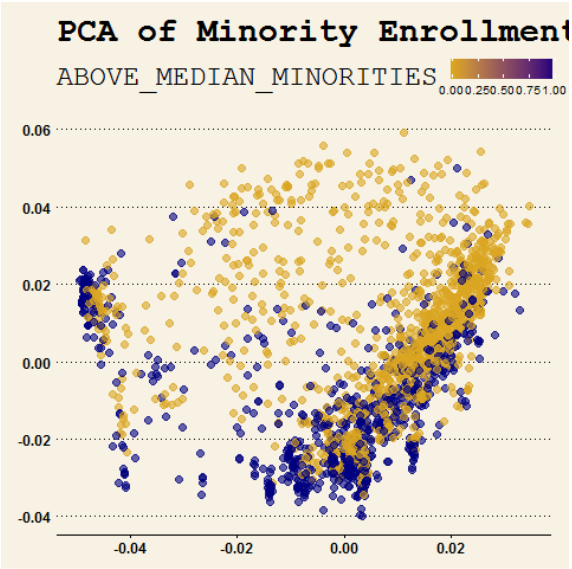


Figure 12: **PCA of Minority Enrollment Rate**

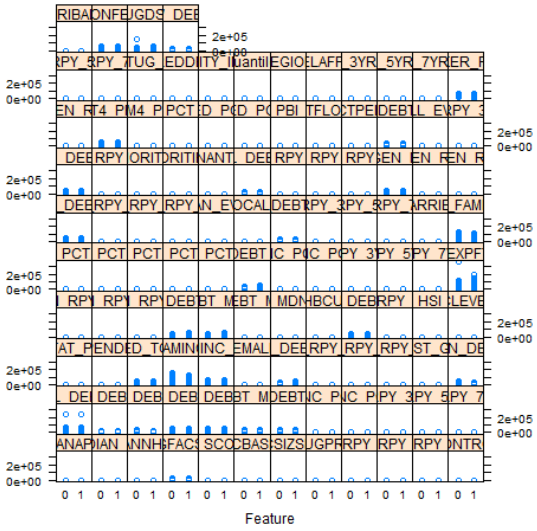


Figure 13: **Feature plot.**

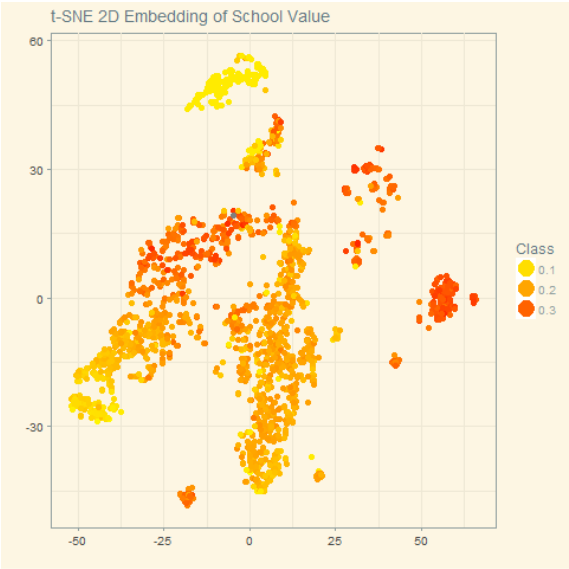


Figure 14: **t-SNE 2D Embedding of School Quality.**

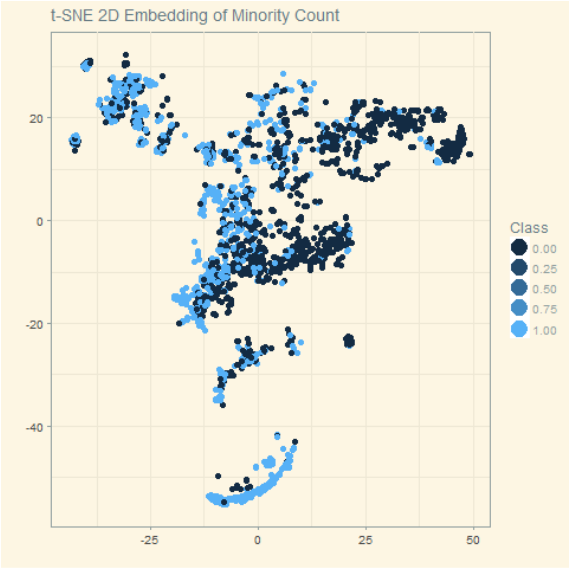


Figure 15: **t-SNE 2D Embedding of Minority Count.**

3 Results

The gradient boosting model reveals that the following predictors are most indicative:

1. Percentage of Families with Income of 75 to 100 thousand dollars anually.
2. Percentage of Pell-grant recipients
3. Number of Undergraduates
4. Family Income
5. Average Faculty Salary
6. Hispanic Serving Institution
7. Region
8. Percentage of Low-Income Dependent Students
9. Tuition, if Private Universtiy
10. Location Type (City, Town, etc.)

This diverse set of predictors spans a variety of features of universities - this covers cost, location, and student profiles. This small-dimensional profile makes it by far easier and more straight-forward for an "NGO" to identify whether or not a school serves minority students. Our initial gradient boosting model had an 85 percent accuracy; however, cross-validation lead to a 95 percent accuracy in our shiny app.

4 Conclusions

Our analysis has culminated into a user-friendly app that any "NGO" can use to target universities to invest in minority education.

Our final product is a shiny app that allows an "NGO" to see whether or not a new school serves minorities and finds the ten most similar universities whose profiles match that of the new school.

A "new school" is input by the user by adjusting the values of the top-ten features generated from fitting the gradient boosting model. The user selects from features such as "Region", "Percentage of Pell-Grant recipients", and "Average Family Income". Then, the shiny app uses the gradient boosting model to predict whether or not a school serves minorities. A distance matrix is then calculated, and the top ten most similar schools are shown.

We also created another app that uses Principal Component Analysis. This app allows the "NGO" to visualize the relationship between variables.