Chris Shartrand and Toyya Pujol

ISyE 6416 Project Proposal

Can We Know How Much Money You Make Based on Where You Went to College?

Problem Statement:

Using the United States Government College Scorecard dataset, is it possible to use a combination of principal component analysis and k-means clustering to accurately predict the income from former college students, 6 and 10 years out, based on the college that they attended.

Data Discussion:

In 2013, the United States government released the College Score Card to help families make informed choices on college selection. The data consisted of information from every college in the US, and included over 1700 data elements and over 7400 colleges. Examples of the data on the colleges includes state, zip code, acceptance rates, percentage of male and female students, if the school is public or private, as well as percentage of the degrees awarded of each discipline. Data includes financial data such as the mean and median income of former college enrollees at years 6, 8 and 10 years, broken up by gender, first-generation college students, and other demographics of the students. Though, data includes college data from 2004 to 2015, there is financial data only from 2015. Hence, we can use the other years as test data.

Methodology:

Using topics we have learned in class, we will cluster the schools by low, medium and high income, based on 2015 data. Low income will be defined by median income of below $30,000, medium income between $30,000 and $75,000, and high income above $75,000. We have decided to use the median over the mean due to the median's characteristic of being robust relative to outliers. We also chose to compare the median income between 6 and 10 years to see if any schools switch categories and if our correct rate decreases. Due to the large size of the data, we have decided to delete any colleges that are missing more than 10% of data elements from consideration. After investigation of the missing data, we believe that EM was not a practical method. Most of the missing data was from non-traditional schools or trade schools (i.e. New Beginnings Schools of Cosmetology). Since EM uses other data to fill in the missing values, we didn't feel that a dataset of traditional schools would be similar enough to provide good estimates for the missing data.

We expect to have some issues with multicollinearity because of the data elements that are highly dependent. A good example, is the percentage of females in the college and the percentage of males. Clearly, the number of females is simply 1- the percentage of males. Our plan is to use PCA and stepwise regression for variable selection. We will take the best of the two methods (determined by R2), and use those predictors to implement K-means. PCA will give us the advantage of creating a combination of predictors, and hence ideally a better prediction for income. In most cases, the loss of meaningful interpretation of the regression coefficients makes PCA a worse choice. Since we are only looking to cluster the data, not having meaningful coefficients will not negatively affect the solution output.

Expected Results:

From the methodology described above, we expect to generate important observations from both the variable selection in PCA and the clustering in k-means. Foremost, we believe that the variables selected using PCA will generate a better fit and show improved predictability of our data over the possible variables selected through a stepwise regression. One drawback to variable selection using

stepwise regression is that its predictive capabilities are limited to the definitions created by the US College Scorecard study. Take for example three variables created by the US College Scorecard: graduation rate, region and ACT scores of incoming freshmen. It is our belief that it would be too difficult to predict income of former students 6 and 10 years out of college, based on such narrowly defined variables. Hence, we expect that PCA will be able to better capture the complexities involved in these income metrics.

Furthermore, as part of our methodology, we plan on producing a confusion matrix after running k-means clustering for the 6 and 10 year out income values. From these matrices, we would expect to see a decrease in the percentage of correct clustering within our three income brackets from 6 to 10 years out. This is due to our theory that it is less likely that the college a student attended significantly influences his or her income 10 years after the person last attended. We expect to get better results with the 6 years after the student last attended the college. From the colleges that are incorrectly classified, an analysis of their representative characteristics will be conducted. Within this analysis, we strongly feel that it will reveal two distinct groups of colleges. The first group would be small, specialized schools that produce qualified and competitive students. The second group would consist of colleges that do not produce students with graduate degrees. In timelines of 6 and 10 years out, these students could obtain graduate degrees from other institutions and increase their incomes. Hence their undergraduate institutions would be confounded with the graduate institutions and could possibly lead to misclassification.

Task Assignment

Chris

1. Proposal Writing
2. Final Report Writing
3. K-Means
4. Confusion Matrix
5. Analysis of Misclassified Colleges
6. Summary Statistics

Toyya

1. Proposal Writing
2. Final Report Writing
3. Cleaning Data
4. Variable Selection
    a. PCA
    b. Stepwise & Lasso Regression