University of Michigan STATS 503

Final Project

The Value of Higher Education: Classifying Future Income for Students Across U.S. Universities

Taylor Spooner
- LDA/QDA, Clustering and unsupervised learning, data collection, exploratory analysis

Elliott Evans
- Random forests, SVM, multinomial logistic regression, introduction, PCA

November 4, 2017

1 Introduction

There exists a continuing debate on what the true value of a college education is. For some, it is measured by the opportunities available to students post-graduation. These opportunities could be quantified by the sheer number of jobs for which a student is qualified, but also by salary. With this in mind, we consider median student incomes six years out of college.

If we can construct an effective model to classify the value of a student's education in terms of income six years after graduating, then we open the doors to more complex problems like forecasting the value of new community colleges, or the value of universities that have seen shifts in their demographics or tuition structures.

We tackle this problem by using government provided information such as tuition, average SAT scores, diversity of student body, etc, to classify discretized versions of median future student earnings, here broken down by quintiles, for a variety of U.S. colleges. The baseline comparison for all of our models will be one that assigns one of five income quintiles randomly to each college, i.e. a classifier that assigns a target "randomly". Such a baseline model would produce an expected error rate of 80% (since each guess has a one-fifth chance of being right).

Our goal is to beat this baseline method and create a set of models that can more accurately determine the median future income quintiles for students in colleges across the country.

2 Data Collection

Our data set comes from the United States Department of Education 2014-2015 College Scorecard Data. The original dataset contains roughly 2,000 columns of information on all undergraduate degree-granting universities across the United States, including academic and admissions statistics, size of school, demographics about the student body, cost of attendance, financial aid, and student earnings after completion[1]. We decided to look at a subset of 88 (78 quantitative, 10 categorical) of these columns to perform different supervised and unsupervised learning techniques to classify and cluster the data. Our response of interest is the median earnings of students working and not enrolled 6 years after entry to the university. We partitioned the median future incomes for each university into five quantiles, where all units are in US Dollars. The five quantiles are [8.3e+03,1.86e+04], (1.86e+04,2.29e+04], (2.29e+04,2.74e+04], (2.74e+04,3.32e+04], (3.32e+04,1.2e+05].

The covariates used to classify the data can be split into four groups. The first group is basic information about the school. This includes information such as the region of the United States in which the campus is located, whether or not the school is in an urban city or rural town, and average faculty salary. The second group can be thought of as academic information about the school, which includes the percentage of degrees awarded in different subjects. Another major group of covariates can be thought of as admissions information. We looked at the midpoint in SAT and ACT scores for incoming freshmen in the different schools, as well as admissions rates. The final group is information about the students who attend the university, captured in the median household income of the students and the total share of enrollment of undergraduates for different races and genders.

Before any analysis could be done, data cleaning techniques had to be utilized. We first

noticed that about 1,800 of the universities did not have values for our response variable, thus we had to remove them, leaving us with 5,900 universities to analyze.

The next step that had to be taken before we could start our analysis was to handle any missing data. For some columns, mainly SAT and ACT score, we saw that many colleges had missing values. We saw no evidence that the missingness in these columns depended on values of the missing data, thus we feel safe assuming that the data is missing at random (MAR) instead of missing not at random (MNAR). This assumption is important to proceed with imputation on the missing data. All of the data was imputed using the R library *mice* and the predictive mean matching method.

We ran the majority of our analysis on Principal Components of the quantitative variables while also including ten categorical variables. The categorical variables are as follows [1]:

- A flag if the campus is the main university campus or not.
- The predominant degree awarded: Certificate, associate's, bachelor's, or graduate.
- The highest degree awarded at the university.
- The region in which the university is located.
- The locale of the institution.
- The Carnegie classification score[3].
- The Carnegie classification size and setting.
- The Carnegie classification undergraduate profile.
- A flag if the university is still open.
- The level of the institution, i.e. four-year, two-year, less than two-year.

3 Models and Methods

3.1 Unsupervised Learning: Clustering

In order to better understand the underlying structure of our data, we ran different clustering algorithms on the scaled data without considering our response variable. We started by running a Gaussian mixture model on our data to cluster. Using the BIC criteria, we picked the best three models to determine the number of clusters, k, to use while running other clustering algorithms on the data. Using these values of k, we clustered our data using a Gaussian mixture model, the K-Means method, and Hierarchical Clustering using single, average, and complete linkage, and Ward's Method. For the K-Means algorithm, since the initial values are randomly selected, we implemented a k++ algorithm to better choose initial values. After picking the number of clusters, we ran k-means clustering using the Hartigan-Wong algorithm 100 times and obtained the average silhouette width. For each clustering algorithm, we observed the silhouette plots for each of the three values of k and chose the k with the largest average silhouette width. We then compared the results of the clusterings for the best performing models for each algorithm.

3.2 Principal Component Analysis

We used Principal Component Analysis (PCA) to reduce the dimensionality of our dataset. We ran PCA on the 78 quantitative variables to allow us to visualize some exploratory analysis on our data and to run different classification algorithms.

3.3 Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA)

We began our classification analysis by performing LDA and QDA. For LDA and QDA, we performed 5-fold cross validation (CV) to obtain training error rates and CV error rates for each dimension-reduced data set that included one principal component to the maximum number of principal components, which is 78. The optimal number of principal components was selected for each model based on the CV error rate. We also looked at the training and CV error rate for each of the five classes.

3.4 Support Vector Machine

We proceeded to use support vector machines with three different kernel types: linear, radial, and 2nd-degree polynomial. For each method, we performed 5-fold cross validation (CV) to obtain training error rates and CV error rates. These error rates were obtained for each model trained on the dimension-reduced data set that included one principal component to the maximum 78 (since there are 78 quantitative variables in our data set), along with ten relevant categorical variables. The optimal number of principal components was selected for each model (selected based on the CV error rate and on a preference toward simpler models) with a fixed cost parameter of 1 (the penalty for slackness).

Based on the results of the SVMs with differing kernel types, we selected the kernel that performed the best via cross validation and subsequently tuned the cost parameter (also with 5-fold cross validation).

3.5 Multinomial Logistic Regression

We proceeded to construct a multinomial logistic regression on the data set using $[8.3e+03, 1.86e+04]$ as the reference income quantile. Similar to the SVM construction, we chose the number of principal components on which to build the final model by observing the CV error rate across 1 through 78 principal components, along with ten relevant categorical variables.

3.6 Random Forests

For our random forest models, we fixed the number of principal components used at 66, due to its success in the models above. We also kept fixed the number of trees generated at 500. The number of variables considered at each split of each tree was considered the parameter on which to tune. We considered random forests with subsets of 1 through 76 variables, by multiples of five.

4 Analysis and Results

4.1 Exploratory Data Analysis

In Figures 1 and 2 we view some basic exploratory data analysis to try and better understand the structure and layout of our data. First, in Figure 1 we look at the relationship between the location of the university and the earnings bracket of a subset (only universities in the continental United States) of the universities in our dataset. This plot gives some slight evidence of class clustering by the region of the U.S. in which the university is located. We see that in the Northeast, the San Fransisco Bay area, and in Los Angeles, there are many schools in the top two earnings brackets. However, in the Southeastern and Midwestern parts of the U.S. we see more schools in the lower two earnings brackets. In the left plot of Figure 2 we observe the data projected onto two columns: median SAT math score and median household income. Examining the coloring based on our response variable, median earnings, we notice a distinction between the highest earning bracket and the lowest. However, for the middle brackets, it is harder to see distinct groups. Finally, on the right plot of Figure 2, we see the data projected onto the first two MDS directions. We once again see a distinction between the highest and lowest earning classes but also considerable overlap between the other classes. The results from our exploratory analysis show that while there is some noticeable separation between the lowest and highest classes, it might be difficult to classify and cluster the three middle income quantiles.

Figure 1: Universities in continental U.S. colored by class

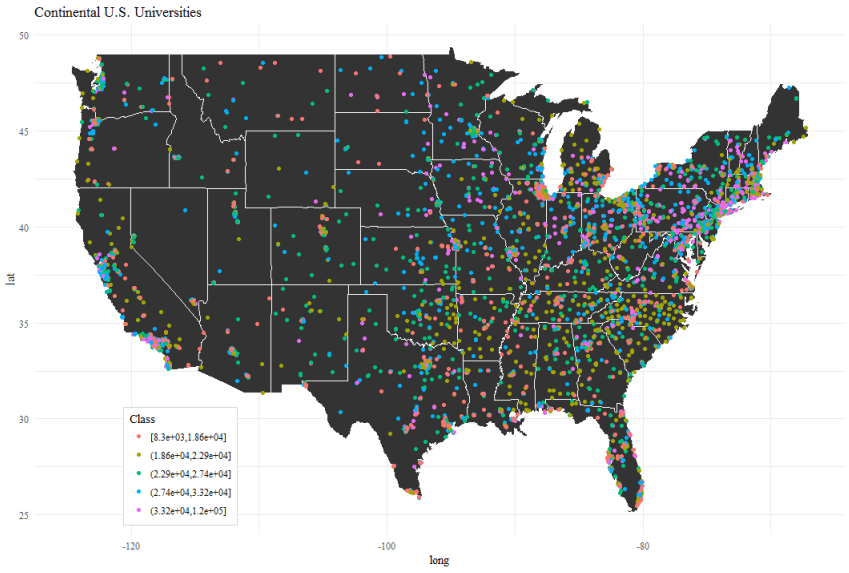
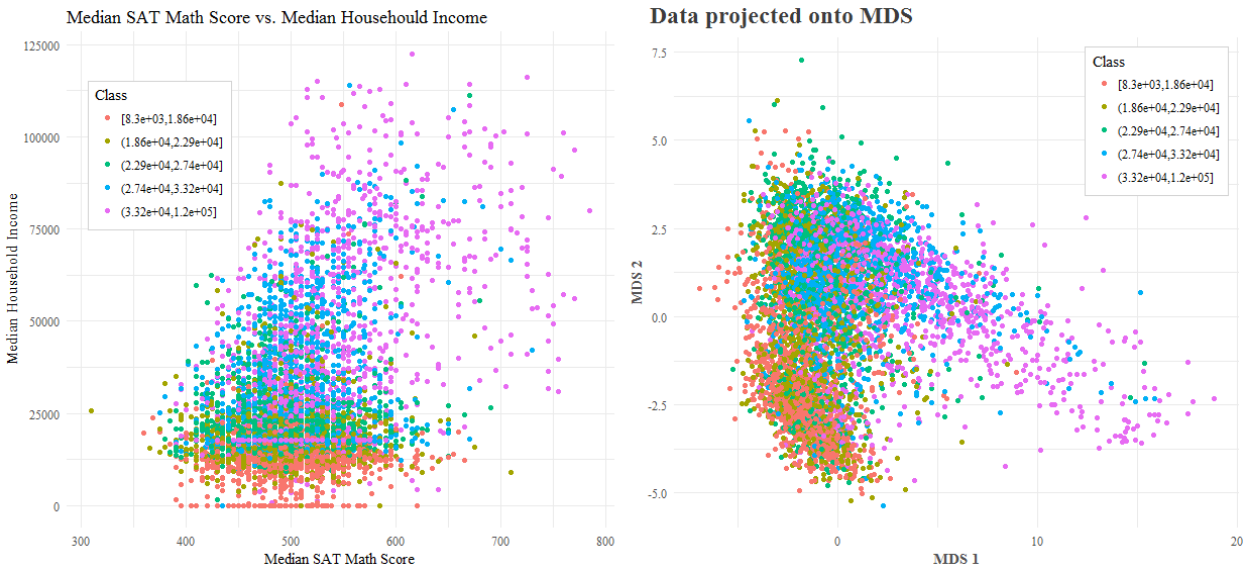


Figure 2: Median SAT Math Score vs. Median Household Income (left) and the data projected onto MDS directions (right)



4.2 PCA

In the case of multinomial logistic regression and support vector machines, we found that utilizing 66 principal components produced the ideal models in terms of cross-validation error. In the scree plot from Figure 3, we see that the marginal decrease in component standard deviation from 60 to 66 components is slight when compared to, say, the marginal decrease from 5 to 10 PCs. However, since we are more concerned with classification error rather than dimension reduction, we ultimately found that it sufficed to choose a greater number of principal components when building our models.

For instance, when we consider Figure 3, a histogram of bootstrap samples of the percentage of variance explained by 66 principal components shows a distribution densely concentrated at 99.283%. Furthermore, a 95% confidence interval using these bootstrap samples is $(99.253, 99.314)$. Thus, we an be quite confident that we have lost an incredibly small percentage of variation in our data set by disregarding the final 12 principal components, while still retaining accuracy in our classification models.

4.3 Clustering

Based on the BIC criteria, the top three models for clustering with a Gaussian Mixture Model are the ellipsoidal, equal shape model (VEV) with 7 clusters, VEV model with 9 clusters, and then VEV model with 8 clusters.Using the number of clusters $k = 7, 8, 9$, we obtained average silhouette widths for the three clustering algorithms; the results can be seen in Table 1.

As we see in Table 1, the clustering techniques were not very successful, giving us evidence that our data does not fall into separable clusters. Looking at our results we see that the clustering algorithm that performs the best is Ward's Method. Here, all three clusterings

Figure 3: Principal component Analysis Results

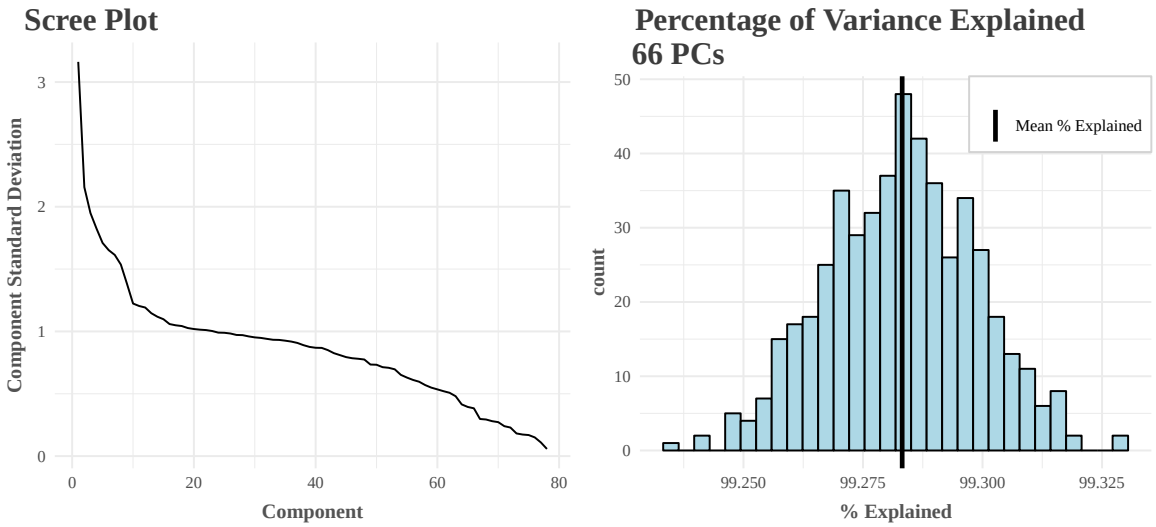


Table 1: Average silhouette width for all clustering algorithms for $k = 7, 8, 9$

Algorithm	k		
	7	8	9
Mixture Model	-0.06	-0.07	-0.09
K-Means	-0.05	-0.07	-0.08
Single Linkage	-0.43	-0.49	-0.52
Complete Linkage	-0.43	-0.43	-0.47
Average Linkage	-0.44	-0.47	-0.51
Ward's Method	0.05	0.05	0.04

produced positive silhouette widths. The next best models are K-Means with 7 clusters and the Gaussian mixture model for 7 clusters. In Figure 4, we see the silhouette plot for each of these algorithms. From these plots we can further see that the clustering was not performed very successfully. We note that for the K-means plot, the silhouette plot represents one run of the K-means algorithm while the average silhouette width in Table 1 represents the average across 100 runs. Looking at Figure 5 it is interesting to see how the different algorithms clustered the same data. The K-Means algorithm seems to have four very separable clusters and three very small clusters while the mixture model and Ward's method produce seven equal-sized clusters that are less separable.

Figure 4: Silhouette plots for the three clustering algorithms for $k = 7$: K-Means (far left), Mixture Model (middle), and Ward's Method (right)

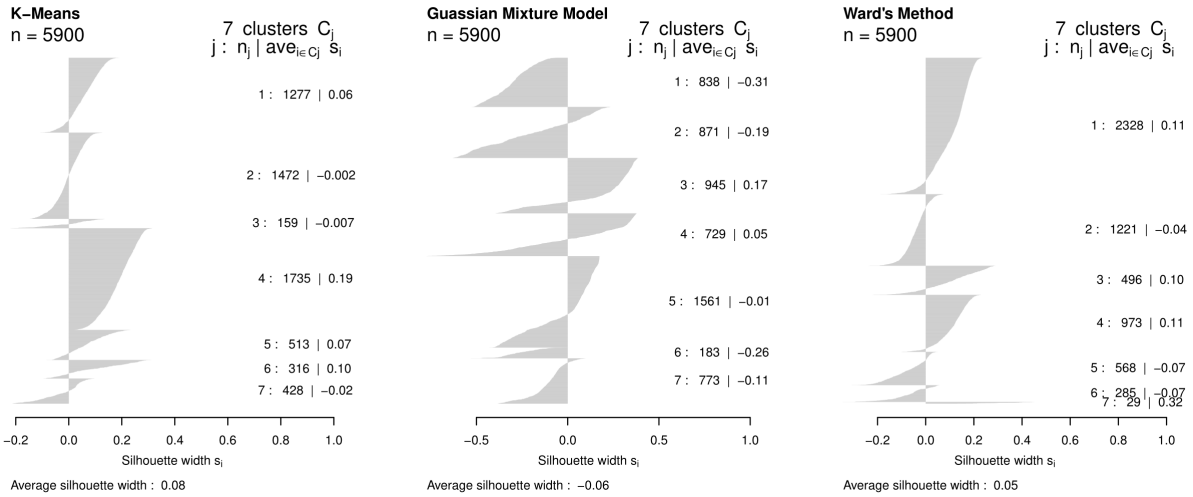
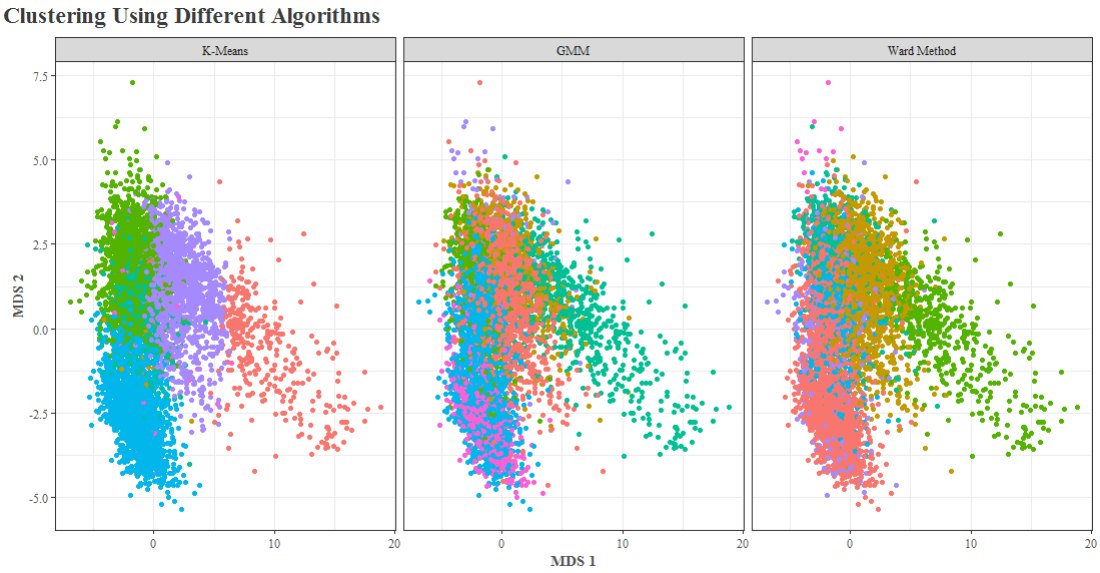


Figure 5: Clustering results for different algorithms using 7 clusters



4.4 LDA and QDA

The cross validation results for LDA and QDA can be found in Figure 6. We notice for LDA, as we increase the number of principal components, our CV error decreases steadily until about 63 principal components, where it starts to level off around a CV error rate of 45%. LDA performed the best when using 69 principal components with a CV error rate of 44.3%. Looking at Table 2 we can see that the LDA model performed the best when classifying the lowest income quantile and the highest income quantile with CV error rates under 40% but performed worse on the middle three quantiles.

Looking at the QDA classification results, also shown in Figure 6, we see interesting and unexpected results. The QDA model reaches its optimal performance using only 12 principal components with a CV error rate of 50.19%. After this point, the model begins to perform worse in both the CV and training error. Further research could be done to provide insight on why the training error would increase when including more principal components for the QDA model.

Figure 6: LDA and QDA Classification Results

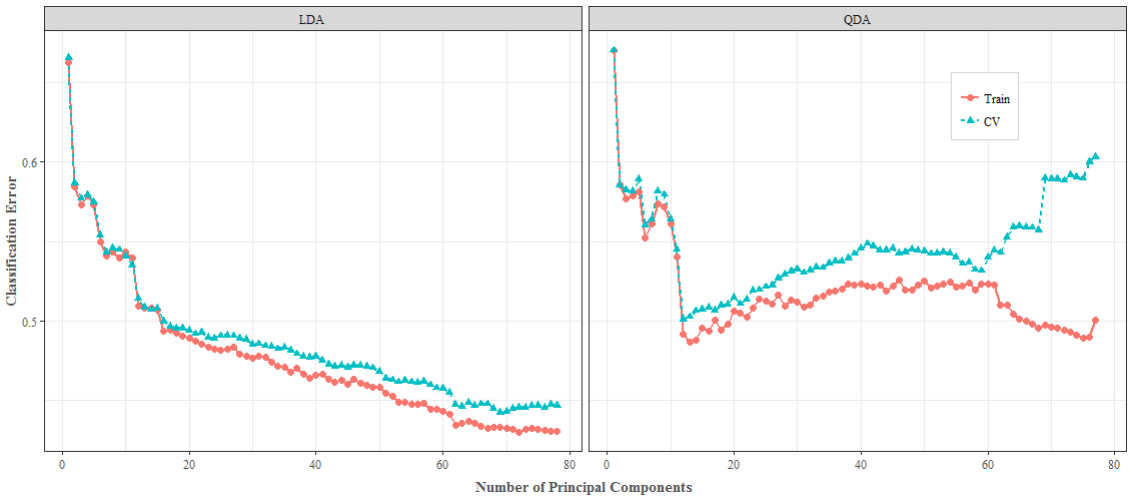


Table 2: Final error rates for LDA model with 69 PCs

Income Quantile	CV	Training
[8.3e+03,1.86e+04]	0.3270	0.3202
(1.86e+04,2.29e+04]	0.4819	0.4726
(2.29e+04,2.74e+04]	0.5045	0.4898
(2.74e+04,3.32e+04]	0.5238	0.5169
(3.32e+04,1.2e+05]	0.3799	0.3705
Total	**0.4425**	**0.4332**

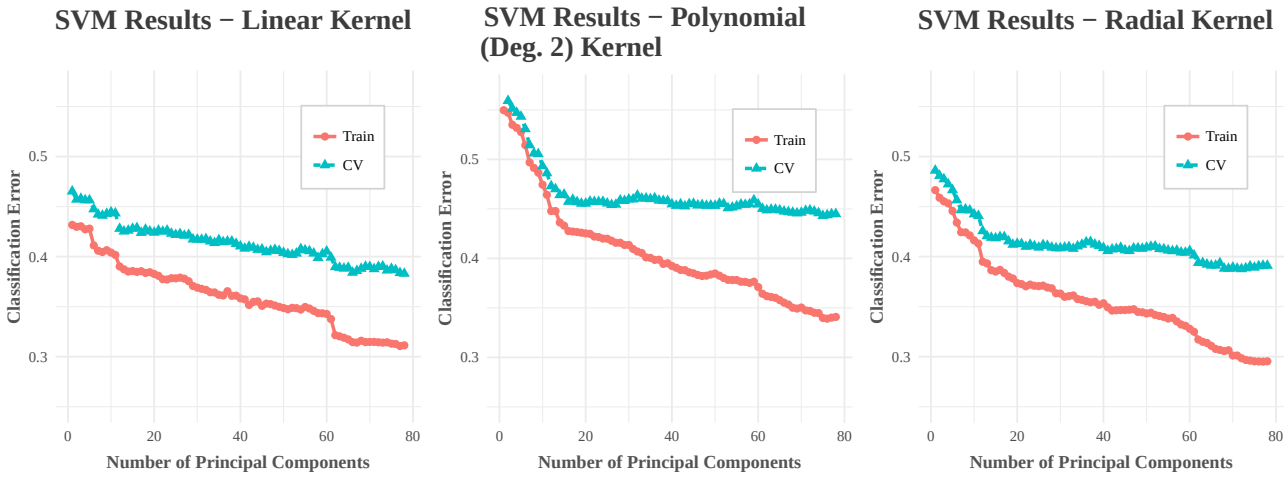
Table 3: Final error rates for QDA model with 12 PCs

Income Quantile	CV	Training
[8.3e+03,1.86e+04]	0.3308	0.3268
(1.86e+04,2.29e+04]	0.5669	0.5535
(2.29e+04,2.74e+04]	0.6138	0.6015
(2.74e+04,3.32e+04]	0.6847	0.6929
(3.32e+04,1.2e+05]	0.3105	0.2890
Total	**0.5010**	**0.4918**

4.5 Support Vector Machine

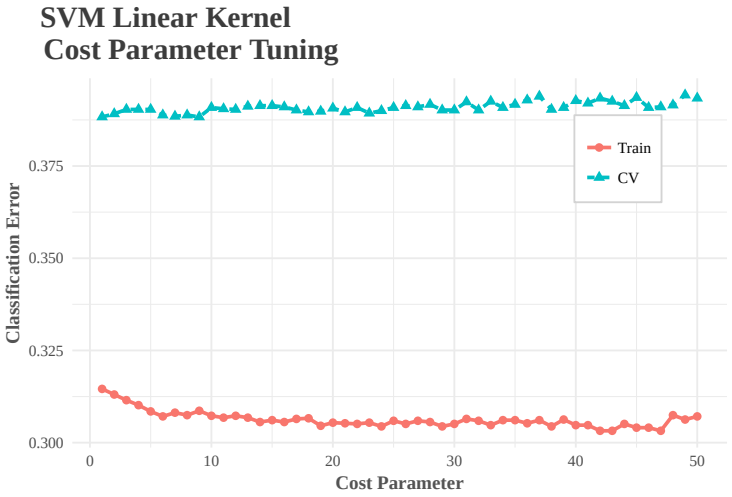
The cross validation results for the support vector machines with cost parameter valued at 1 can be found in Figure 7. We notice that the SVM second degree polynomial model performed the worst, with its best CV error rate of 44.3% (75 principal components). The radial model performed better, with its best CV error rate of 38.8% (72 PCs). The SVM with a linear kernel performed the best in cross validation, however, with a best CV error rate of 38.3% (using all 78 PCs). We notice that the SVM linear kernel model performed almost just as well using the first 66 principal components with a CV error rate of 38.4%. Since we do have a preference for simpler models, we proceeded to choose the SVM linear kernel model with 66 PCs as our best model choice, and subsequently tuned the cost parameter based on this decision.

Figure 7: Tuning of number of principal components for different support vector machines



As we can see from Figure 8, the SVM linear kernel model utilizing 66 principal components and all 10 categorical variables was not particularly sensitive to the tuning parameter, but performed the best using a cost parameter valued at 9. This model gave a CV error rate of 38.8%. Note that this error rate was greater than the original CV error rate of 38.4% when finding the best number of principal components to use in the final model, but this variability is expected due to the randomness in cross-validation folds at each subsequent step of the model selection process.

Figure 8: Results for tuning the cost parameter for the SVM linear kernel model using 66 principal components and ten categorical variables



The final error rates for the best SVM model are given in Table 4. The support vector machine with 66 PCs, all 10 categorical variables, and a cost parameter of 9 gives a total CV error rate of about 39%, which is much better than the expected baseline error rate of 80%, determined by guessing income bracket randomly. We also observe that the income bracket most easily predicted in cross validation is $(3.32e+04, 1.2e+05]$, the highest income bracket, which produced a CV error rate of 25.8%. The income bracket most difficult to predict in cross validation was $(2.29e+04, 2.74e+04]$, which produced a CV error rate of 51.3%.

Table 4: Final error rates for the best SVM model: linear kernel with 66 PCs, 10 categorical variables, and a cost parameter of 9

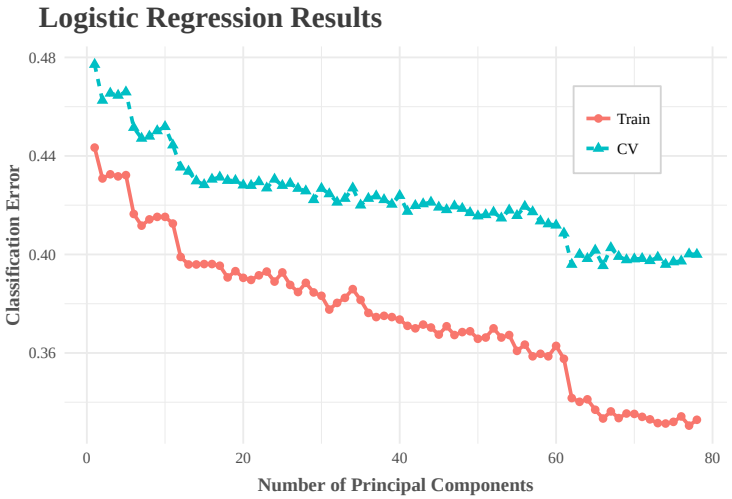
Income Quantile	CV	Training
[8.3e+03,1.86e+04]	0.2752	0.2260
(1.86e+04,2.29e+04]	0.4505	0.3429
(2.29e+04,2.74e+04]	0.5127	0.4112
(2.74e+04,3.32e+04]	0.4576	0.3773
(3.32e+04,1.2e+05]	0.2579	0.1878
Total	**0.3903**	**0.3086**

4.6 Multinomial Logistic Regression

After running multinomial logistic regression on our dimension reduced data sets (obtained via PCA), we produced the 5-fold cross validation error rates in Figure 9. We find that the number of principal components that minimizes the CV error rate is sixty-six, which gives us a model with an error rate of 39.5%. Thus, we use this number of components to include in the final logistic regression model.

We observe the final error rates for our best model in Table 5. The final logistic model produces a CV error rate of approximately 39.4%. These error rates are extremely similar

Figure 9: Cross validation (5-fold) error rates for multinomial logistic regression models using varying numbers of principal components



in magnitude and pattern to the error rates for the final SVM model in Table 4. The only noticeable difference is that the category $[8.3e+03, 1.86e+04]$ was only slightly easier to predict in the logistic regression model, based on the CV error rate.

Table 5: Final error rates for the best multinomial logistic regression model: 66 PCs and 10 categorical variables

Income Quantile	CV	Training
[8.3e+03,1.86e+04]	0.2512	0.2154
(1.86e+04,2.29e+04]	0.4735	0.4195
(2.29e+04,2.74e+04]	0.5272	0.4340
(2.74e+04,3.32e+04]	0.4505	0.4129
(3.32e+04,1.2e+05]	0.2698	0.2170
Total	**0.3939**	**0.3388**

While inference and interpretation can be difficult follwing dimension reduction via PCA, we can understand some dependencies on the original predictors by observing the model estimates. In Table 6, we find that the estimated effects on the log-odds of being in each future median income quantile are negative. We may also note that the first principal component is dominated by variables based on SAT scores, ACT scores, tuition fees, and the median family income for each student. The loadings for these variables in the first principal direction are also negative. One could reason, then, that increasing test scores, tuition fees, and median family incomes are associated with increasingly negative PC1 scores, and therefore increasingly positive effects in the multinomial logistic model, relative to the lower income reference level. Thus, it may be the case that higher values for these variables are associated with greater log-odds of universities having greater median future earnings for their students.

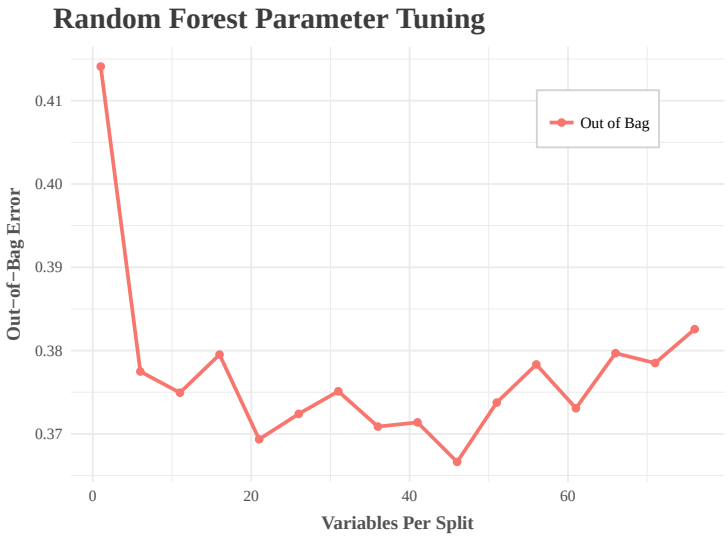
Table 6: Estimated coefficients for the first principal component in our multinomial logistic regression model. The lowest income quantile was excluded since it was the reference level.

Income Quantile	PC1 Estimated Logit Effects
(1.86e+04,2.29e+04]	-0.6764
(2.29e+04,2.74e+04]	-0.9344
(2.74e+04,3.32e+04]	-1.3341
(3.32e+04,1.2e+05]	-1.5436

4.7 Random Forests

After tuning the model to the number of variables to be considered at each split in each tree, we obtain the out-of-bag error rates in Figure 10. The training error rates were all 0.0%, an occurrence that is not rare for random forests, since the trees are allowed to grow to maximum size and overfit to the training set. We found that the model using a random subset of 46 variables at each split produced the best out-of-bag error rate of 36.7%. Hence, we choose this as our final random forest model.

Figure 10: Out-of-Bag error rates for random forest models with varying numbers of variables considered at each split. Each model used the first 66 principal components of the data set, along with ten categorical variables.



In Table 7, we observe the out-of-bag error rates by class. Our random forest model outperformed the multinomial logistic regression model in every class with the exception of the quantile $(2.74e+04, 3.32e+04]$, where the out-of-bag error rate was about 1.8% higher than the CV error rate for multinomial logistic regression. In addition, the random forest model outperformed the final support vector machine in three of the five classes, and outperformed both the LDA and QDA models in five-of-five classes.

Table 7: Results of the final random forest model: 500 trees and random subsets of 46 splitting variables.

Income Quantile	Out-of-Bag Error Rate
[8.3e+03,1.86e+04]	0.2408
(1.86e+04,2.29e+04]	0.4630
(2.29e+04,2.74e+04]	0.4796
(2.74e+04,3.32e+04]	0.4687
(3.32e+04,1.2e+05]	0.2092
Total	**0.3714**

5 Conclusion

The final cross validation error rates in predicting future median income of students for colleges six years post-graduation were generally near 39%. In Table 8, we see that our best SVM and multinomial logistic regression models produced CV error rates of approximately 39%, about half the 80% baseline error rate. The best performing model was the Random Forest model with 46 splitting variables with an out of bag error rate of 37.14%. The random forest model outperformed every other model in a majority of the five income brackets that were classified.

We also found that the data set essentially contained approximately twelve dimensions that were extraneous based on their total percentage of variation explained for the data set, i.e. sixty-six principal components captured about 99.3% of the variation in the quantitative variables. In general, this number of principal components was the point in the tuning process where CV error rate leveled off.

One possible limitation in our work is the amount of missing data in our data set. As mentioned in Section 2, we used the R library *mice* and the imputation method of predictive mean matching to impute the missing values. However, not having this missingness or imputing the data using a different method could have changed our results. A further extension of this project could entail performing this analysis under different imputation methods and observing the changes in results.

Table 8: Results of the final classifiers on the data set in predicting future median income of students from each college post-graduation

Final Models	CV/OOB Error Rates
SVM Linear Kernel	0.3903
Multinomial Logistic Regression	0.3939
LDA	0.4425
QDA	0.5010
Random Forest	0.3714

Acknowledgements

We would like to acknowledge Long Nguyen, Roger Fan and Jesús Arroyo for providing lecture notes and example code used to complete this project.

References

[1] U.S. Department of Education, *Data Documentation* (2016), available at `https://collegescorecard.ed.gov/assets/FullDataDocumentation.pdf`

[2] Analysis of 2015 from U.S. Department of Education, "College Scorecard Data"

[3] The Carnegie Classification of Institutions of Higher Education, `http://carnegieclassifications.iu.edu`