Analysis of College Scorecard

Sheryan Resutov
The Cooper Union for the
Advancement of Science and Art
New York, New York 10003
Email: resutov.sheryan@gmail.com

Eugene Sokolov
The Cooper Union for the
Advancement of Science and Art
New York, New York 10003
Email: eugsokolov@gmail.com

Harrison Zhao
The Cooper Union for the
Advancement of Science and Art
New York, New York 10003
Email: harrisonzhao.cooper@gmail.com

***Abstract*—We attempted to classify and predict higher educations institutions in the United States using the College Scorecard data set provided by Kaggle. We predicted student information such as mean earnings, loan default rates for different income brackets and majors, and completion rate. We used a combination of lasso regression, logistic regression, support vector machines, and random forests to predict the targets. The classification accuracy varied in the range of 40-78%. We also deduced which features are significant in predicting the aforementioned targets.**

I. Introduction

There are over 5,000 higher educational institutions in the United States, colleges and universities, that provide a post secondary education. These institutions have drastically changed since the first college, Harvard University, was founded in 1636. The basis of these changes stem from religious affiliations, academic and athletic guidance, faculty focus, and administrative power. We attempt to analyze present day institutions based on various features including administrative, faculty and student statistics.

Data was taken from an open Kaggle competition that started on 25 September 2015 and ends on 14 January 2016. The competition has no rewards. Kaggle is hosting a College Scorecard data set that draws information from the US Department of Education and from student financial aid systems, which are based on federal tax returns. There are a number of features provided including admissions, academics, cost, financial aid information, and several other interesting features. The data set covers 7,804 universities and colleges in the United States for each year from 1996 to 2013 [1].

II. Related Works

There has been some previous work done regarding higher education in the United States. An article by Washington Monthly in September 2011 states that inflation adjusted tuition at public universities has tripled since 1980. Total spending by higher education institutions has tripled since then. In addition, faculty to student ratios have remained constant while administrative to student ratios have increased by 36 percent [2]. Another Huffington post article stated that in 2014 the number of non-academic administrative employees at U.S. colleges and universities has more than doubled in the last 25 years [3]. Research done by the American Institutes for Research titled an article "Think Again: Administrators Ate My Tuition! Really?". In the article, the authors discuss the administrative bloat and change in higher education from 1990 to 2012. It is well known that higher education has become a business in the United States and that students are being cheated out of tens of thousands of dollars for their four years of higher education.

The College and University Professional Association for Human Resources titled a research article "Administrators in Higher Education Salary Survey for the 2013-14 Academic Year". The goal of their research was to collect surveys from various institutions and create summaries of higher education in the United States. Their data covers 1,247 higher education institutions. The report shows expenses by various features such as institution type, Carnegie classification, and several other metrics [4].

In addition, work has been done at the Cooper Union to classify schools based on a set of various features. The research focused on creating several indexes to show predictive trends between different features across higher education institutions in the United States. Their analysis shows that high endowment at a university is the single most important feature in showing higher entrance SAT scores, higher starting and mid-career student salaries, higher administrative salaries, and higher faculty salaries [5].

III. Our Approach and Methods

We encountered two main problems in our analysis, sparsity of the data and classifying continuous output classes. Although the Kaggle dataset contains approximately 1,700 features for over 7,000 higher educational institutions, many of the attributes were null. In addition, a large number of attributes were "Privacy Suppressed" because, due to certain privacy laws, Kaggle was not legally allowed to show the value. After removing these features and selecting all rows of data that contained valid values for our desired features, we were left with approximately 1,000 to 1,500 records, depending on the feature set used.

The second problem we dealt with was predicting salaries, SAT scores, and other continuous values. Predicting salaries to an exact value would not give good results, given that salary ranges from approximately 20,000 to 120,000. Thus, we bucketed values into discrete ranges to solve this problem. Given a feature set, we were able to divide the set into discrete bins, where every feature in the set would be mapped to a bin number. The size of each bin was determined by subtracting

the minimum value in the feature set from the maximum value and dividing the result by the number of bins. By mapping continuous output classes to discrete bins, we were able to achieve much higher accuracies.

Our analysis focused on using logistic regression, SVM, random forests and lasso regression. Logistic regression is a special case of a generalized linear model and analogous to linear regression although based on different assumptions. It models the posterior probabilities of K classes via linear functions in x, while at the same time ensuring that they sum to one to remain in [0,1]. The model has the form

$$\begin{aligned}
\log \frac{\Pr(G=1|X=x)}{\Pr(G=K|X=x)} &= \beta_{10} + \beta_1^T x \\
\log \frac{\Pr(G=2|X=x)}{\Pr(G=K|X=x)} &= \beta_{20} + \beta_2^T x \\
&\vdots \\
\log \frac{\Pr(G=K-1|X=x)}{\Pr(G=K|X=x)} &= \beta_{(K-1)0} + \beta_{K-1}^T x.
\end{aligned}$$

This ratio is commonly referred to as the logit and creates a continuous criterion as a transformed version of the dependent variable. The logistic function, sigmoid σ is defined as follows

$$\sigma(x) = \frac{e^x}{e^x + 1} \tag{1}$$

Support vector machines are a form of sparse kernel machines which utilize maximum margin classifiers. SVMs approach the classification problem through the concept of the margin, which is defined to be the smallest distance between the decision boundary and the support vectors. The support vectors are determined by a subset of data points, which are points closest to the opposing class. As seen in Figure 1, the support vectors are circled and the margin is maximized. SVM attempts to create a linear hyperplane to separate two classes, within the maximized margin [7].

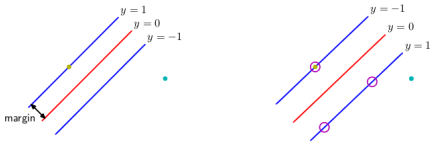

Fig. 1: Simple implementation of SVM [7]

Random Forests is an ensemble classifier, which is a modification of bagging, that builds a large collection of decorrelated trees and then averages them. It is implemented in three steps. The first step is bootstrapping. This builds a collection, or forest, of B decision trees. The next step is to split each level of the decision tree randomly on a subset m of p variables. After the forest is grown, the random forest predictor is

$$\hat{f_{rf}^B}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x;\Theta_b) \tag{2}$$

Random forests are robust against overfitting and are unbiased because the model is created through dense randomness [6]. It works well with large data sets due to its scalability and robustness.

Lasso regression is a form of linear regression with an added regularization parameter. The lasso estimate is defined by

$$\begin{aligned}
\hat{\beta}^{lasso} &= argmin \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \\
subjectto &\sum_{j=1}^{p} x_{ij}|\beta_j| \leq t.
\end{aligned} \tag{3}$$

This is similar to ridge regression, except the penalization term is replaced by the L_1 lasso penalty. This constraint makes the solutions nonlinear in the y_i, and computation is now a quadratic programming problem [6]. Due to the nature of the constraint, making t sufficiently small will cause the coefficients of the less significant variables to be zero. Using lasso regression as a feature reduction technique, we applied our reduced feature set to logistic regression, SVM and random forest models.

Our initial input feature set consisted of admission rate, average cost of attendance, net tuition, out-of- state tuition, in-state tuition, faculty salary, total undergraduate students, percent of undergraduate white are white, percent of undergraduates are black, percent of undergraduates above the age of 25, percent of undergraduates working part time, percent of undergraduate first generation college students, percent of students receiving federal loans, percent of students receiving Pell Grants, percent of students whose family income is over $110,000 (wealthy), percent of students whose family income is under $40,000 (poor), median debt loan principal, reading, writing and math SAT scores in the 75th percentile.

IV. Results

For each of the following results, we show our prediction accuracy for logistic regression, SVMs and random forest after performing feature reduction with lasso regression. We also show the bin size of the feature, along with the minimum and maximum value. All training/validation was done with 5 fold cross validation in order to ensure that each instance of our data got to be in the training and testing sets and in order to prevent overfitting. In addition, each model was trained through the use of grid search to sweep the parameters and find the best ones. For random forest, the parameters considered were the number of trees in the forest and the number of features to consider when looking for the best split. For SVM, the parameters considered were the C value and the gamma, since we just used the rbf kernel. For logistic regression, the parameter considered was the C value.

First, we attempted to predict the mean earnings of students 10 years post graduation. As seen in Figure 2, the reduced feature set includes tuition costs and percentage of wealthy students.

BinSize	26.7k
Min	13.1k
Max	227k
LogReg	0.71
SVM	0.59
RF	0.73

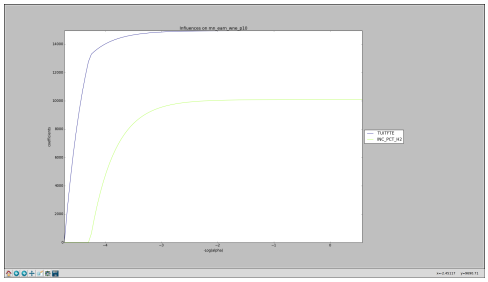

Fig. 2: Reduced Lasso plot for mean student earnings 10 years after graduation

Next, we attempted to predict the four year completion rate. As seen in Figure 3, the reduced feature set includes verbal SAT scores, admission rate, out-of-state tuition and tuition cost. Verbal SAT scores were most prevalent. Admissions rate and tuition cost were inversely related.

BinSize	0.18
Min	0
Max	1
LogReg	0.41
SVM	0.38
RF	0.62

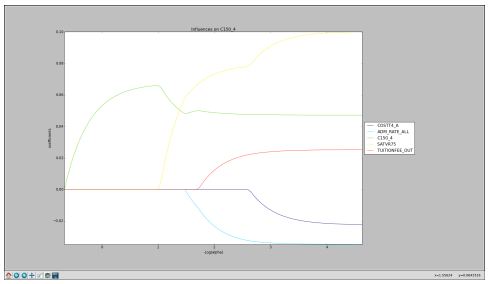

Fig. 3: Reduced Lasso plot for completion rate analysis

Next, we attempted to predict the default rate after 3 years of graduation based on students' family income. As seen in Figure 4, we were not able to reduce any of the features and are left all four income brackets which include $30k-48k, 48k-75k, 75k-110k, 110k+. This makes sense intuitively because the income bracket that your family is in should say a lot about whether you would be able to pay back your loans.

BinSize	0.1
Min	0
Max	1
LogReg	0.53
SVM	0.50
RF	0.75

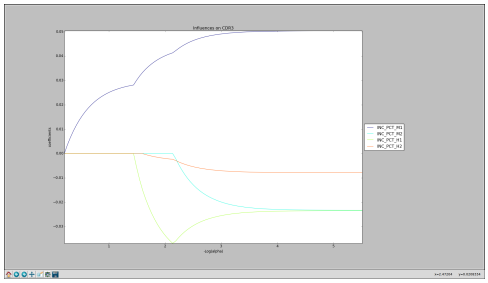

Fig. 4: Reduced Lasso plot for loan default rate after 3 years based on family income

Similarly, we attempted to predict the loan default rate after 3 years of graduation based on students' major. As seen in Figure 5, we see that students who major in Art are more likely to default on their students loans after 3 years. Following that are engineers and lastly, architects.

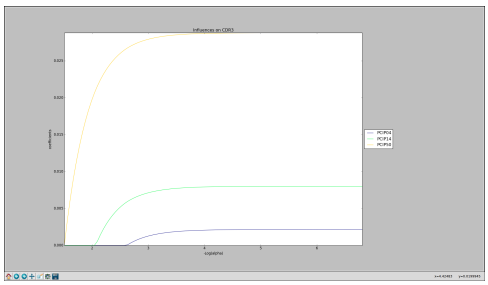

Fig. 5: Reduced Lasso plot for loan default rate after 3 years based on major

V. Conclusion

In conclusion, we have successfully analyzed the College Scorecard data set that was provided by Kaggle. Our accuracies were consistently in the 60-75% range, which is reasonable result. The accuracies could have been easily increased by making the bin sizes larger. However, the results would have been less helpful in that case because the bin sizes would have been too large to be meaningful. For example, a bin size of 20,000 for the expected mean salary makes a lot more sense and is much more helpful that a bin size of 50,000. Unfortunately, not as many features were present in the data set as we hoped, and certain interesting features such as "student earnings" and other tax information was "Privacy Suppressed". These features might have led to some extremely interesting analysis. Nonetheless, we provided some interesting analysis based on the features available. We show that student mean earnings are depending heavily on the tuition costs of an institution and the percentage of wealthy students in that institution. It would be possible to further pursue this data set and try different classifiers and continue reducing features.

Acknowledgment

The authors would like to thank Sam Keene for his guidance in Machine Learning and Data Science. In addition, we thank Kaggle for providing the data set, although it turned out to be

shit and leave out half the features we wanted to analyze to begin with.

REFERENCES

[1] Kaggle. US Dept of Education: College Scorecard. 2016.

[2] Ginsberg, Benjamin. Administrators Ate My Tuition. Washington Monthly. 2011.

[3] Marcus, Jon. New Analysis Shows Problematic Boom in Higher Ed Administrators. Huffington Post. 2014.

[4] Kirshstein, Rita. American Institutes for Research. 1000 Thomas Jefferson Street, NW Washington, DC 20007.

[5] Koe, Andrew and Resutov, Sheryan and Sokolov, Eugene. Administrative and Educational Analysis of Higher Education in the United States. The Cooper Union for the Advancement of Science and Art. May 2015. unpublished.

[6] Hastie, Trevor and Robert Tibshirani, and J. H. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer, 2009. Print.

[7] Bishop, Christopher M. Pattern Recognition and Machine Learning. New York: Springer, 2006. Print.