CSE 6242 Project Final Report

Jared Babcock, Sanjeed Hakim, Grayson Niehaus,
Nathaniel Rich, Matthew Riddle, Fred Sackfield

12/3/19

1 Introduction

There have been many studies on factors that contribute to educational success, however, many of these studies focus only on a limited amount of data and are often difficult to interpret. Additionally, the data is most commonly presented in a static, difficult to understand format. Our project aims to bring together historical state-level data to determine social, economic, and policy-related factors that are important to educational success, and to visualize that data over time. The hope is that bringing a wide variety of data into an easy to understand interactive map will empower decision-makers to understand important contributors to education and how those change over time to better inform policy. In addition to providing a canvas for exploration, we also hope to provide users the ability to work directly with the project data sources to build custom predictive models. Our unique approach to data collection, modeling, and visualization provides decision makers with an intuitive way to explore and understand how socioeconomic and policy decisions impact educational success.

2 Problem Definition

There are two main components to our project: explanation and prediction. We first aim to identify and explain which variables are the most important for success. To this end, we aim to create a map showing state-by-state education data that cycles over time to show longer-term changes. Tool-tips

over every state help visualize the most important factors contributing to the education level of that state at that time. For prediction, we plan to allow the user to input a variety of factors, such as state, funding, and year, and our models will produce an estimated education level under those parameters. The prediction feature adds a simulation component where those policy-makers can examine projections about the impact that certain decisions may make.

3 Literature Survey

As identified in a study by Ashenfelter and Kreuger [11], education is one of the key components of future economic success. In said study, a small population of separated twins were perfect controls for ability while education was the only variable. Likewise, Clark and Martorell [5] study the economic success of those that barely passed compared to those that barely failed final graduation exams in the 1990s. They identify the importance of a high-school diploma in signaling worker potential to employers. Similarly, Chevalier et al. [4] use a 1970s dataset to find limited support that education level is not just a signal of ability, but also enhances worker productivity. Furthermore, Schoeni and Blank [15] determine that education is a strong predictor of income and impacts the effects of welfare.

Previous academic work didn't acknowledge the difference between social science data and other datasets. In fact, many papers will exclusively use tests scores/grades to analyze students. Ferreira suggests that this is a naive approach and a researcher must take into account other demographic information, social information, etc. [8]. Liebowitz and Kelly [12] discuss the bias in state education rankings, based solely on standardized tests, and analyze new metrics that include ethnicity and state funding while Figueroa [9] identifies a correlation between state school funding and the economy. Similarly, York et al. provides a recommended framework, which can be used as a starting point [18].

In current research on modeling education data, linear regression is widely used, but GLM may be better when dealing with categorical data [17]. Shin also suggests that a Hierarchical Linear Model works well for multi-level data such as repeated measures of student performance. [16]. Other researchers

have found that decision trees and random forests perform well when evaluating the wide range of factors because those models can filter out irrelevant predictors [6]. However, some studies have explored the limits of regression-based models in social science, and find that simpler metrics may be more robust [7].

Many policies can be considered when adapting/understanding a prescriptive data approach. Brown and Conley [2] argue that legislators should consider requiring alignment of public state university curriculum with K-12 to support transition to higher education. Brock and Leblanc [1] recommend building supportive college systems by analyzing the systems of successful states, and outline how those systems are increasing educational attainment.

A briefing by Loeb and Plank [14] describes the need for robust information systems to drive education policy decision making. In a separate briefing, Loeb et al. [13] outlines requirements for an educational data system in California. We believe the first step towards building an effective system is to integrate many relevant factors into a database. Additionally, many states have seen success by implementing a SUR system to track their students and analyze the effects of local policy decisions. This should be generalized to the state level [10].

4 Proposed Method

4.1 Intuition

Our project will be better than state-of-the-art because of four main innovations:

1. We are leveraging data from a variety of sources, which can better capture the disparate factors contributing to education level

2. We are also emphasizing usability in our project; similar projects do not come with accessible visualization and modeling tools (both across states and over time) and thus are not as easy to make decisions on.

3. We are implementing a prediction feature that allows decision-makers to simulate the policy effects and thus to make more informed decisions.

4. The majority of academic papers are difficult to access and understand. However, we break the status quo by cloud hosting all of our results in an intuitive mobile-optimized website with a basic click or drag and drop interface.

4.2 Data Collection and Integration

Our methodology begins by identifying data points related to educational success that will be helpful in modeling and visualization. Because a key innovation is aggregating disparate data sources related to state education, we sought to identify a wide range of potential predictors. Many of these come from an article by Liebowitz, which we have combined with other predictors such as: crime rate, average teacher salary, teacher to student ratio, and education spending. Additionally, because our plan includes using the dataset for modeling purposes as well as visualization, we desire a clean and complete dataset that contains no missing values across all columns and for all years.

After scanning a multitude of public sources for accessible data sets, we were able to import more than 70 columns that spanned 12 years of state-level data. To help with the importing and integration process, we wrote web-scraping programs in Python, and several wrangling programs in Python, SQL, and R to re-structure the data, to assess the scope and availability of the data, and to impute missing values.

We have merged each data file into a master csv for analysis. Because of the large amount of features in our dataset, we will need to monitor multicollinearity, which is identified in our literature [9]. To clearly describe what each feature measures, we created a data dictionary outlining each variable.

4.3 Modeling and Analysis

Our dataset contains average scores in reading and math for both fourth graders and eight graders. The values appear to occur every two years from 2003 to 2015. So, for each state we have seven data points. First, we combined the average scores into an average of averages response variable. Additionally, we recognized that our monetary figures needed to be adjusted for inflation, so we pegged all of them to 2019 dollars. Then to address the normality in data assumption we needed to perform a natural log transformation of nearly all the possible predictors to address the obvious fact that

larger states would have larger revenues, larger student populations, and such. Next data was split into a training and validation set based on a 75-25 split and a stepwise process to maximize the validation R^2 was performed. The final model found that ln(Local Revenue) and ln(Median Income) had a statistically positive impact on average scores, while ln(Federal Revenue) and Student-Teacher ratio had a statistically negative impact, if all else is the same. Diagnostic plots showed that were had not violated the normality in errors assumption, our data did not have any outliers in the predictors, and our Q-Q plot showed no underlying non-linear structure. However, our validation R^2 is only about 33.8%, so we are not able to explain even half of the variability in average test scores

4.3.1 OLS: Making Statistically valid inferences

OLS is excellent for making statistically valid inferences, however the poor fit of our model means our predictions may not be as accurate as possible. One solution to this is to use machine learning methods such as Random Forests or Boosted Trees to capture subtle and complex non-linear relationships. However, these ML methods can quickly become very complex, with hundreds of layers, making inferences difficult. With our ML approaches we may be able to improve on predictions, but we cannot control for issues such as multicollinearity, which at best may increase bias, but at worst can totally invalidate our results due to issues of hetroscedasticity or more. Therefore, we would like to avoid making inferences and instead focus solely on improving predictive power with our ML methods. Issues of non-normality or multicollinearity are not as destructive to prediction as they are when making inferences. Some possible problems are that our parameters may have larger confidence intervals and we may experience some leptokurtosis in the distribution of our predictions.

4.3.2 Boosted Trees and Random Forests

A key issue is that each state relatively few times in our data, so based on the random splitting of our data, it is possible we may exclude states with high leverage in the training or validation process. Furthermore, the boosted ML methods we would like to use benefit tremendously from a slow learning rate, however, with our limited data the model may stop before it can converge on the optimal solution. Sure enough, when we ran our boosted tree and

forest models, we noticed that the fit would vary wildly based on validation and training split. One possible fix for this is to borrow a common technique used often in political science, known as bootstrapping, where data from a sample is resampled repeatedly with replacement to have more to work with. As long as the structure of the bootstrapped data is similar to the sampled data and we assume that the sampled data represents the true population, the inferences from bootstrapped data are of similar quality as those from a large sample (sample of the bootstrapped size). Therefore, we resampled a sample of size 50 out of 350, since there are fifty states, 100 times. The distribution of our data set with 5000 bootstrapped rows is nearly identical to our original sample. In fact, a simple OLS shows that the validation R^2 is around 35%, so we practically see no improvement in fit even though the bootstrap sample is more than 14 times the size. However, the Random Forest method benefited monumentally with the bootstrap sample. A variety of tuning adjustments finally honed into a model with 58 trees per forest and 5 terms sampled per split. It has a validation R^2 of around 99.8%, which may be overfit. We will continue refining our parameters and incorporate a test set as well to ensure we are not overfitting. Likewise, a Boosted Tree model converges with 500 layers, 3 splits per tree, and a learning rate of 0.11, with a validation R^2 of around 96.6%. This model essentially stops increasing validation R^2 at 300 layers, so it may be possible to achieve a more parsimonious model

4.4 Visualization

Analyzing educational differences between states over time fits well into a choropleth map similar to the one that we made in this course. For this map, we are charting our "academic success" variable in each state and dynamically updating the year and corresponding data after a certain interval. For now, we are using a placeholder standardized score for academic success, just to show viability with our data and selected key values. See Figure 1. We have implemented UI controls that enable the end user to step through each year by selecting "Next" or "Previous" buttons; additionally, by selecting the "Play" button, the end user can animate the choropleth and see how standardized scores in each state vary over time.

In addition to the animation features, we enabled further interactivity and exploration with zoom-in capability. As shown in Figure 2 below, upon clicking a state, the map zooms in and opens a tool tip that displays quan-

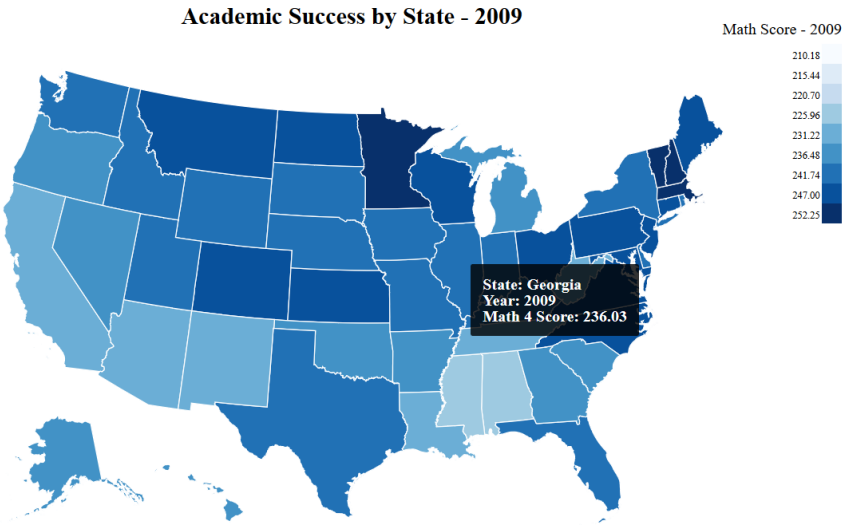

Figure 1: 2009 snapshot of plotted academic success

ties for the most important predictor variables.

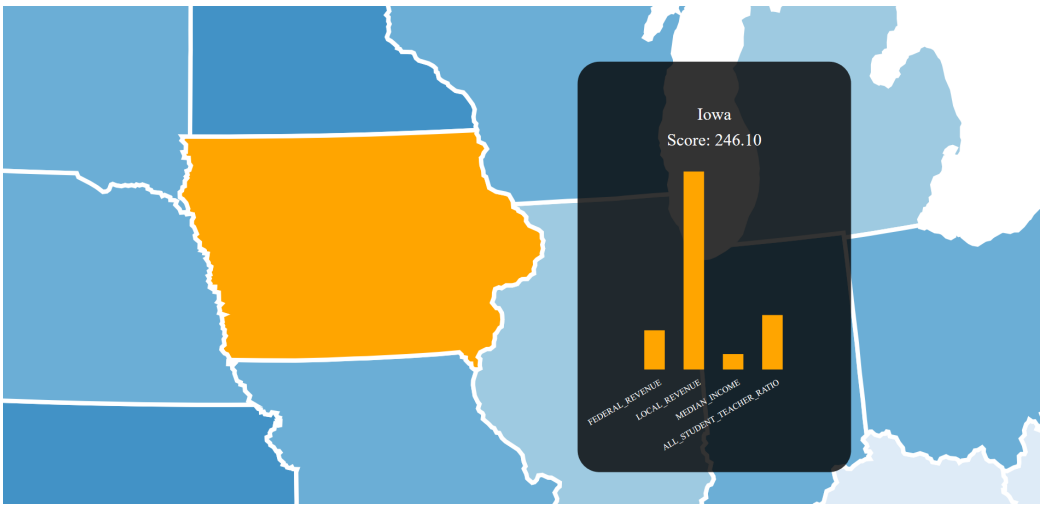

Figure 2: 2003, State of Iowa tool tip snapshot of key predictor quantities

We have also incorporated an interactive heat map visualization. To provide more insight into the many other socioeconomic factors that were collected and analyzed with respect to educational success. This heat map displays cells for all 50 states over all 7 years of data. The heat map is particularly helpful for visualizing the variability in different factors across states. For example, Figure 3 below shows the variability in poverty rate across all 50 states.

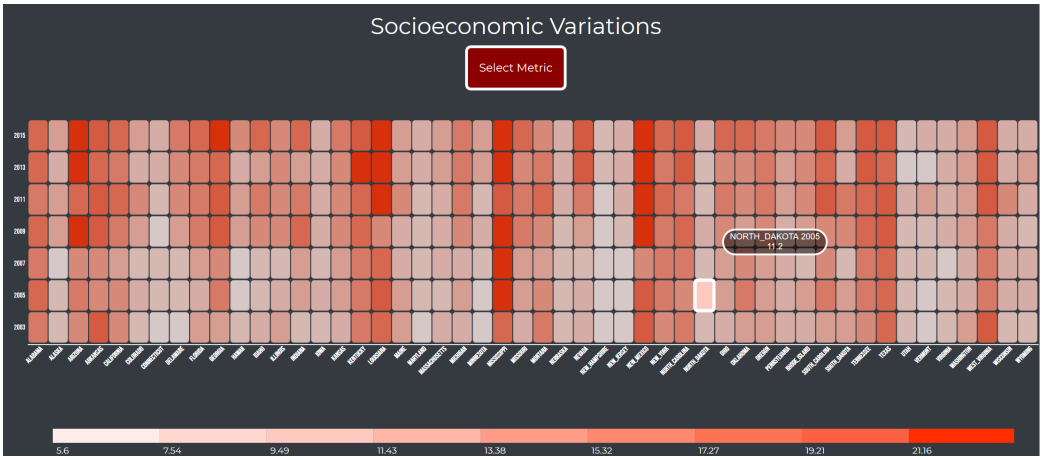

Figure 3: Heat Map showing Poverty Rate across all 50 states from 2003-2015

The ultimate goal of our visualizations is to provide the end-user with easy access to a wide range of factors that relate to educational success. The choropleth and heat map work together to provide a holistic view of education over the past two decades, both at the state-level, and nationally.

5 Experiment/Evaluation

Through our experimentation, we are trying to find the features that most contribute to students' success. We experimented with different evaluation metrics (training R^2, RMSE, validation R^2) to find which provides the most explanatory power. Based on the metric chosen, we are training and comparing different modeling methods to see which method has the most predictive power for our response. To ensure our methodology and experimentation was rigorous, we tested multiple different modeling techniques, with a wide range of predictors collected from various sources. These techniques include

stepwise regression, boosted trees, boosted forest, and neural network. After comparing results from many different models and subsets of explanatory features, we found that the best models, or those with highest explanatory power in terms of validation R^2, are boosted tree, boosted forest, and then neural network in that order.

However, because we want to empower the end users and decision makers with our interactive tools, a key feature of our integrated product is the predictor app, which gives the end user the ability to experiment with and evaluate different predicting variables and modeling techniques to analyze predictions. We believe that the trial and error process of model building and validation is critical to understanding the importance of the different factors as they relate to educational success. One consideration for future improvement of this tool would be to allow the end user to define and incorporate interaction effects into the custom models.

6 Conclusions and Discussion

Our project's data searching and collection efforts were successful in identifying a wide range of different variables to include for exploration and modeling; however, we were somewhat limited in the scope and availability of our newfound data sources. Despite only having seven full years of data, our modeling efforts seemed to work quite well for predictive purposes, especially when using the advanced boosting and random forest techniques. On a related note, the presence of so many new predicting variables made it difficult at first to rationalize and determine which combinations or subsets would be most useful for explanatory and predictive power. Stepwise regression was helpful for model selection, but more advanced feature selection techniques, as well as different interaction effects could also be considered to enhance ease of interpretation and model reliability.

By emphasizing customization and interactivity, our choropleth visualization stands out as a unique tool in the realm of data and analytics for education. Additionally, by integrating our predictor app into the same hosted site, we enable much broader exploration and modeling capabilities with side-by-side tool integration. The hope is that our integrated tool provides the end user with convenience as well as flexibility for data exploration and analysis.

In the future, we believe improvements can be made by focusing efforts on quantifying specific impacts of policy-making and different legislative acts on educational success. We could leverage Natural Language Processing techniques to scan bill summaries and determine which states have the most effective policy makers and determine which categories (e.g. Teacher Quality, School Safety, etc.) should receive more attention moving forward. With growing populations and increasing challenges within state-level education, being able to understand the relationships between academic success and socioeconomic factors is key for anyone who is a stakeholder or decision maker in public education. Our project not only provides this insight interactively, but it also provides the ability to predict and forecast the effects of various policy decisions or financial downturns. While many improvements can be made in the future, we believe our project is a proof of concept tool from which educators and policy makers alike can gain insight.

7 Work Distribution

7.1 Old Plan of Activities

Activity	Jared	Matthew	Nathan	Fred	Sanjeed	Grayson
Proposal Ideas	✓(2hrs)			✓(2hrs)		
Modeling	✓(4hrs)			✓(4hrs)	✓(8hrs)	✓(6hrs)
Visualization	✓(8hrs)		✓(12hrs)		✓(4hrs)	
Hypothesizing		✓(2hrs)				
Data Acquisition		✓(4hrs)	✓(8hrs)	✓(5hrs)		✓(4hrs)
Data Wrangling				✓(5hrs)		✓(6hrs)
Statistical Testing					✓(4hrs)	

All team members have contributed a similar amount of effort.

7.2 New Plan of Activities

Activity	Jared	Matthew	Nathan	Fred	Sanjeed	Grayson
Interpreting Results	✓(4hrs)					
Modeling	✓(4hrs)		✓(4hrs)	✓(4hrs)	✓(14hrs)	✓(2hrs)
Visualization		✓(5hrs)	✓(8hrs)			
Web App Dev			✓(8hrs)			
Data Acquisition		✓(5hrs)		✓(8hrs)		✓(6hrs)
Data Wrangling	✓(8hrs)			✓(4hrs)	✓(4hrs)	
Database Creation						✓(8hrs)
Data Tracking Tool		✓(10hrs)				

All team members have contributed a similar amount of effort.

References

[1] Brock, Thomas; Leblanc, Allen. (2005). Promoting Student Success in Community College and Beyond. MDRC.
https://files.eric.ed.gov/fulltext/ED485507.pdf.

[2] Brown, Richard S. Conley, David T. (2007) Comparing State High School Assessments to Standards for Success in Entry-Level University Courses, Lawrence Erlbaum Associates, Inc., 12:2, 137-160, DOI: 10.1080/10627190701232811

[3] Callan, Patrick M.; Finney, Joni E.; Kirst, Michael W.; Usdan, Michael D.; Venezia, Andrea. (2006). "State Policymaking for Improving College Readiness and Success. The National Center for Public Policy and Higher Education. https://www.tandfonline.com/doi/pdf/10.1080/10627190701232811?needAccess=true

[4] Chevalier, A., Harmon, C., Walker, I., Zhu, Y. (2004). Does Education Raise Productivity, or Just Reflect it? The Economic Journal, 114(499). doi: 10.1111/j.1468-0297.2004.00256.x

[5] Clark, D., Martorell, P. (2014). The Signaling Value of a High School Diploma. Journal of Political Economy, 122(2), 282–318. doi: 10.1086/675238

[6] Cortez, P., Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008), 5-12.
https://repositorium.sdum.uminho.pt/bitstream/1822/8024/1/student.pdf

[7] Dana, J., Dawes, R. (2004). The Superiority of Simple Alternatives to Regression for Social Science Predictions. Journal of Educational and Behavioral Statistics, 29(3), 317-331.
https://journals.sagepub.com/doi/pdf/10.3102/10769986029003317

[8] Ferreira, F. H. G., Gignoux Jeremie. (2011). The Measurement of Educational Inequality Achievement and Opportunity [Discussion Paper No. 6161]. Retrieved from
http://ftp.iza.org/dp6161.pdf

[9] Figueroa, Eric; Leachman, Michael; Masterson, Kathleen. (2017). A Punishing Decade for School Funding. Center on Budget and Policy Priorities.
`https://www.cbpp.org/research/state-budget-and-tax/a-punishing-decade-for-school-funding`

[10] Hearn, James McLendon, Michael Mokher, Christine. (2008). Accounting for Student Success: An Empirical Analysis of the Origins and Spread of State Student Unit-record Systems. Research in Higher Education. 49. 665-683. 10.1007/s11162-008-9101-z.

[11] Krueger, A., Ashenfelter, O. (1992). Estimates of the Economic Return to Schooling from a New Sample of Twins. American Economic Association. doi: 10.3386/w4143

[12] Liebowitz, S. J., Kelly, M. (2018). Fixing the Currently Biased State K-12 Education Rankings. SSRN Electronic Journal. doi: 10.2139/ssrn.3185152

[13] Loeb, S., Beteille, T., Perez, M. (2008). Building an Information System to Support Continuous Improvement in California Public Schools. Policy Brief 08-2. Policy Analysis for California Education, PACE (NJ3).

[14] Loeb, S., Plank, D. N. (2008). Learning What Works: Continuous Improvement in California's Education System. Policy Brief 08-4. Policy Analysis for California Education, PACE (NJ1).

[15] Schoeni, R., Blank, R. (2000). What has Welfare Reform Accomplished? Impacts on Welfare Participation, Employment, Income, Poverty, and Family Structure. National Bureau of Economic Research doi: 10.3386/w7627

[16] Shin, Jongho, et. al (2004). Use of Hierarchical Linear Modeling and Curriculum-Based Measurement for Assessing Academic Growth and Instructional Factors for Students with Learning Difficulties. Education Research Institute. Asia Pacific Education Review: Vol. 5, No. 2

[17] Theobald, E., Aikens, M., Eddy, S., Jordt, H. (2019). Beyond linear regression: A reference for analyzing common data types in discipline based education research. Physical Review Physics Education Research,

15.
https://journals.aps.org/prper/pdf/10.1103/
PhysRevPhysEducRes.15.020110

[18] York, T. T., Gibson, C., Rankin, S. (2015). Defining and Measuring Academic Success. Practical Assessment, Research Evaluation, 20.