Feature Selection via Regulatization: School Proficiency Case

Andres Cambronero

Social scientists often rely on ordinary least squares regression and logistic regression to test their hypotheses and develop their theories. These methods provide coefficient estimates that are easy for researchers to interpret and for readers to understand. Consequently, these statistical tools are present in almost every recent social science article. Despite their popularity, most publications overlook a challenging aspect of these methods: variable selection. While it is common for researchers to develop models based on the findings of previous studies, readers are often unaware of the criteria that researchers use to develop their models.

This lack of transparency creates both ethical and statistical dilemmas. First, researchers might be tempted to control for specific variables to obtain their desired result. Such behavior is a prevalent pattern that Hindman (2015) called "pet variable problem." Second, models built exclusively on findings of previous research might overlook important associations between variables. As data becomes more granular, modern databases provide researchers with variables that were unavailable in previous decades, which heightens the risk of missing meaningful associations. Third, because issues of multicollinearity prevent researchers from including all variables in a regression, they are forced make some determinations, even if these are unfounded. Unfortunately, these issues often receive little attention in publications.

Regularization methods such as Ridge regression, LASSO and Elastic Net offers scientists tools to overcome these obstacles. When presented with a list of variables, these methods effectively shrink the coefficients of variables that are not strongly associated with the response variable to zero and allow predictors that are associated with it to have non-zero coefficients. Although these methods produce biased coefficients, they allow researchers to identify a few predictors that are strongly associated with the outcome, which can later be further studied. Ridge regression, LASSO and Elastic Net Regression provide researchers a way to efficiently and ethically identify predictors associated with their study's dependent variable.

Using data on elementary and middle schools' academic performance, this paper compares whether ridge regression, LASSO and Elastic Net identify similar factors as being associated with proficiency as does a logistic regression in which the independent variables were selected manually, as commonly done in social science. The results demonstrate that, with fewer variables, these methods develop models that are easy to interpret and achieve the same or better accuracy as the model built manually. This work shows that modern statistical methods can be used to overcome issues associated with variable selection and thereby bring greater transparency and reproducibility to social science.

Literature Review

When education researchers use a statistical methods, they often fail to explain the criteria used to determine the variables included in their regression. For example, Ainsworth-Darnell and Downey (1998) use OLS to test whether students' personal attitudes about school affect the racial gaps in educational achievement. In their work, researchers develop six hierarchical regression models that include somewhere between 2 to 17 predictors. Although the control variables in their model might be justified, the authors fail to explain why these variables were included. Apart from reviewing findings from previous research, the authors do not detail their process of variable selection.

Similarly, Logan et al. (2012) rely on OLS regression to examine the association between race and students' scores in mathematics and reading, but fail to clarify why certain variables were included in their final model. In line with findings listed in their literature review, the authors find that schools with a large share of minority and economically disadvantaged students tend to perform worse than schools that do not have these demographic characteristics. Beyond a side note detailing why the percentage white students was omitted from the analysis, the authors fail to explain the reasons for including other variables in the model. Again, the reader is given no explanation about the researchers' variable selection decisions.

Other studies often rely exclusively on the findings of previous research to develop their models, which risks ignoring potential associations between predictors and the outcome of interest. For example, Subedi and Howard (2013) developed a hierarchical generalized linear model to predict the drop-out and graduation rates based on the predictors that previous research deemed significant. In their literature review, the authors note that previous studies have found an association between drop out rates and socioeconomic status, behavioral issues, gender and race. Based on these findings, the authors develop a model that identifies a student's number of days suspended, a student's race, and a school's percent of English language learners as significant predictors of dropping out. The authors develop their model disregarding the fact that other variables could have been significant but might have been missed by previous research.

Methodology

Ridge Regression: According to Faraway (2014), ridge regression assumes that, among many predictors, only a few are associated with the outcome. Variables that do not predict the outcome should have coefficients close to zero and vice versa. In particular, ridge regression selects $\hat{\beta}$ such that:

$$\hat{\beta} = \text{argmin}_{\beta}(y - X\beta)^T(y - X\beta) + \lambda \sum_{j=1}^{p} \beta_j^2$$

where $\lambda \geq 0$ is a shrinkage parameter and $\sum_j \beta_j^2$ is a penalty term. As λ approaches infinity, β_j approaches 0, resulting in an intercept only model. As λ approaches 0, β_j approaches the least squares solution.

In other words, to minimize the expression, β_j must keep most coefficients close to 0 and allow a few powerfully predictive variables to have larger (in absolute terms) coefficients. While this method assigns most variables coefficients close to zero, it keeps all variables in the model. The coefficients obtained from ridge are biased towards zero. By introducing bias, this method provides less variable coefficients than OLS, in the presence of multicollinearity.

LASSO: LASSO solves a similar minimization problem as ridge regression. LASSO selects the $\hat{\beta}$ such that:

$$\hat{\beta} = \text{argmin}_{\beta}(y - X\beta)^T(y - X\beta) + \lambda ||\beta_j||_1$$

where $\lambda \geq 0$ is a shrinkage parameter and $||\beta_j||_1$ is a penalty term. Just as in ridge regression, as λ approaches infinity, β_j approaches 0, resulting in an intercept only model. As λ approaches 0, β_j approaches the least squares solution. As with ridge, the coefficients obtained from LASSO are biased toward zero. However, unlike with ridge regression, LASSO sets some coefficients exactly to zero and therefore performs variable selection. In the presence of highly correlated pair of predictors, LASSO keeps one of such pair in the model.

Elastic Net: This method uses two penalty terms to overcome some of the limitations of both ridge regression and LASSO. This method selects $\hat{\beta}$ such that:

$$\hat{\beta} = \text{argmin}_{\beta}(y - X\beta)^T(y - X\beta) + \lambda_2 ||\beta||^2 + \lambda_1 ||\beta||_1$$

According to Zou et al (2005), the l_1 penalty performance variable selection while the l_2 penalty removes the limitation on the number of selected variables and stabilizes the l_1 regularization path. Just as with LASSO, Elastic Net sets the coefficients of some predictors to zero and therefore performs variable selection.

Data Processing

First, the outcome variable "PROF_LEVEL" was created following guidelines of the Every Student Succeeds Act (ESSA) and New York's standard of proficiency. Under ESSA, a school is in need of improvement if English language learners, students receiving special education, racial minorities, or students in poverty consistently underperform. Using the Department of Education's 2015-16 3-8 Assessment Database, each subpopulation of interest within a school received a score of 1 (proficient) or -1 (not proficient) if it's assessment mean achieved the state's standards. Then, each school received a weighted score based on the school's demographics. If the final score was negative, the school was classified as "Not Proficient" and "Proficient" otherwise.

To have a completed dataset, values had to be artificially generated to supplement the information available on the New York State's Department of Education databases, the National Center for Educational Statistics and ACS. These databases contained the following incomplete variables: Suspensions, Average Class Size, Number of Teachers, Percent of Not Highly Qualified Teachers, Mean Teacher Score, Enrollment, Expenditure per student, Violent crimes and Property crimes. For these observations, random values were generated from a bounded normal distribution with the mean and variance of the observed values of a school's district. Any remaining missing observations were completed drawing numbers from a normal distribution with the mean and variance of the values observed in the column. With this process, each observation had complete information for each column.

The final dataset contained 26 variables. A full list of variables in the final dataset is provided in the index.

The final dataset contained 3327 observations. Out of that total, 277 (8%) observations were of class "Proficient" and the remaining were "Not Proficient." This dataset was randomly split into a training and test set using a 8:2 ratio. Both subsets retained the same proportions in proficiency level as the original data. Given that the variables were recorded on different scales, all variables except charter status were standardized before analysis.

Exploratory Analysis

This section of the paper explores the characteristics of the data. The goal is, first, to determine whether the data reflect some of the findings that previous research has found and, second, explore which variables could potentially be included in a logistic regression model.

Table 1 presents a summary the predictors in the training data. The table indicates that, on average, schools in New York State have a student body that is 50.3% White, 18.5% Black, 22% Hispanic and 6% Asian. Among the schools in the data, over half of students are categorized as economically disadvantaged and live in areas where 26% of individuals have a high school degree. The table demonstrate that the variables are measured in different scales and will be standardized before conducting formal analysis.

Table 1: Summary Statistics of Numeric Variables

	mean	sd	median	lower.25%	upper.75%
PER_FREE_LUNCH	46.708	27.263	44.000	23.667	71.333
PER_REDUCED_LUNCH	5.858	3.909	5.667	2.667	8.333
PER_LEP	7.103	10.087	3.000	0.333	9.667
PER_AM_IND	0.641	3.503	0.000	0.000	0.667
PER_BLACK	18.493	25.362	5.333	1.333	25.667
PER_HISP	21.951	23.260	12.667	4.333	31.667
PER_ASIAN	6.479	11.613	2.000	1.000	6.000
PER_WHITE	50.319	37.270	60.333	6.333	87.000
PER_Multi	2.043	2.404	1.333	0.333	3.000
PER_SWD	16.678	8.055	15.333	12.000	19.667
PER_FEMALE	48.812	5.045	48.667	47.333	50.333
PER_ECDIS	54.523	27.624	56.000	32.333	79.667
PER_SUSPENSIONS	3.267	6.323	1.333	0.333	3.667
MEAN_CLASS_SIZE	22.383	4.134	22.000	20.000	25.000
MEAN_ENROLL	472.462	272.857	405.667	295.667	575.667
PER_NOT_HQ	3.876	7.509	0.333	0.000	4.000
STUD_TEACH_RATIO	11.785	2.705	11.649	10.150	13.228
PER_HS_DEGREE	25.953	8.176	26.005	20.310	31.325
MEAN_INC	123552.701	47856.536	121649.687	90969.229	155960.598
EXP_PER_ST	19141169.345	39600632.038	264574.939	91714.595	25508199.738
MEAN_TEACH_SCORE	88.373	5.523	88.657	84.343	92.847
VIOL_CRIME_PER_CITIZEN	0.004	0.004	0.003	0.001	0.005
PROP_CRIME_PER_CITIZEN	0.018	0.012	0.015	0.013	0.018
TOTAL_DIST_POP	2693147.982	3808842.039	47578.000	15675.000	8461961.000

Figure 1 presents correlations between variables that previous literature has identified as significant predictors of school proficiency and students' academic performance. Clearly, some variables are highly correlated and therefore should not be included simultaneously in a regression. For example, "PER_ECDIS" is almost perfectly correlated (0.98) with "PER_FREE_LUNCH." The plot also reflects patterns that have been encountered in previous research. For instance, "PER_WHITE" shows a negative correlation with all indicators of economic hardship, suggesting that as a schools share of White students increases its share of students in economic need decreases. Conversely, "PER_BLACK" and "PER_HISPANIC" are positively correlated these economic hardship indicators. Additionally, schools with a higher percentage of White students have a lower percentage of teachers without proper qualifications. These results reflect those found in studies mentioned in the literature review.

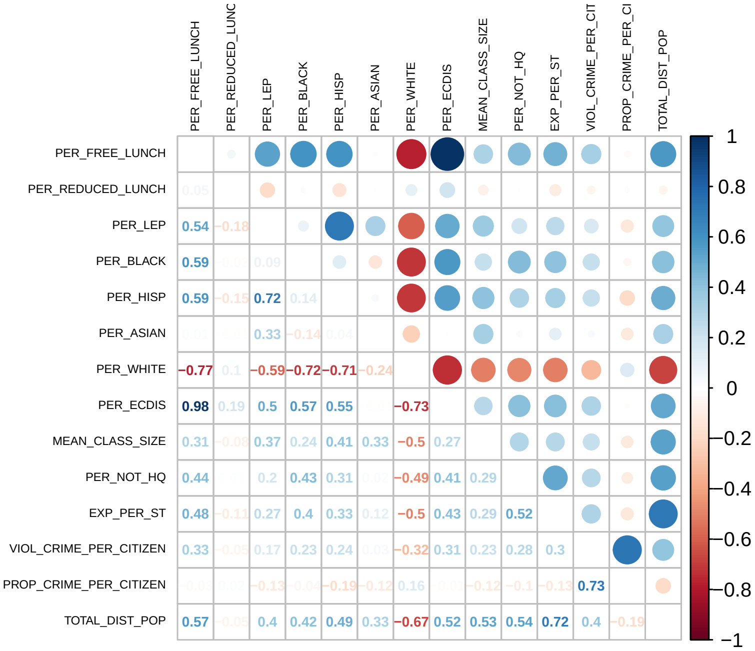

Figure 1: Correlation between Predictors

Figure 2 suggests that racial composition varies by a school's proficiency level. The figure presents the mean of "PER_BLACK," "PER_WHITE," "PER_HISP," and "PER_ASIAN" by proficiency level. The mean percentage of Black students in "Not Proficient" schools is 19.12% and 11.53% for "Proficient Schools." For both proficiency levels, there are clear outliers at the higher end of the distribution. The mean percentage of White students in "Not Proficient" schools is 50.58% and 47.45% for "Proficient Schools." The mean percentage of Hispanic students in "Not Proficient" schools is 22.4% and 16.94% for "Proficient Schools." Finally, the mean percentage of Asian students in "Not Proficient" schools is 5.09% and 21.75% for "Proficient Schools." These means highlight the existing racial disparities in educational opportunities identified in previous research.

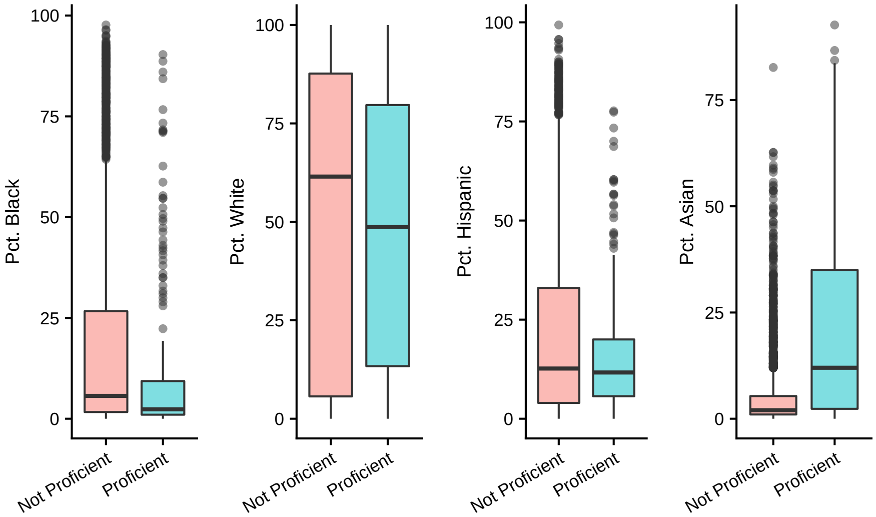

Figure 2: Racial Distribution between 'Proficient' and 'Not Proficient' Schools

Figure 3 reinforces the notion that racial composition differs between "Proficient" and "Not Proficient" schools.

The figure shows the distributions of the racial variables analyzed in Figure 2. First, most schools contain a low percentage of each of the racial minorities considered, which results in right skewed histograms. Second, the Pct.White plot's bimodal distribution suggests that schools suffer from racial segregation. School are either almost entirely White or schools lack almost entirely such students. Third, very few “Proficient” schools have high percentages of Black or Hispanic students, but a significant portion of these schools have almost all White student bodies. These exploratory findings suggest that race might be associated with a school's proficiency level.

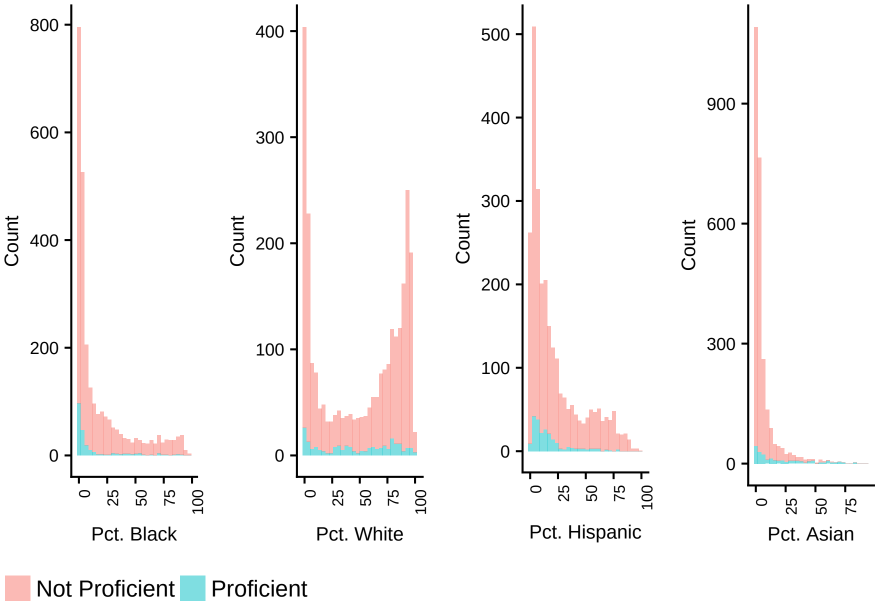

Figure 3: Racial Distribution between 'Proficient' and 'Not Proficient' Schools

In addition to racial differences between schools with different proficiency levels, these schools exhibit economic disparities. Figure 4 shows the distribution of students' economic status by schools' proficiency level. The boxplots suggest that the mean percentage of students that receive free lunch in “Not Proficient” schools is 48.07% compared to only 31.66% in “Proficient”" schools. Similarly, the mean percentage of economically disadvantaged students in “Not proficient” schools is 55.99% and 38.29% in “Proficient” schools. The histograms indicate that very few “Proficient” schools have a high percentage of students experiencing economic hardships. Conversely, many schools that have a high percentage of students in economic hardship are “Not Proficient.” This figure reinforces the connection that exist between educational opportunities and economic status and further highlight the correlation between the variables “PER__FREE__LUNCH” and “PER__ECDIS.”

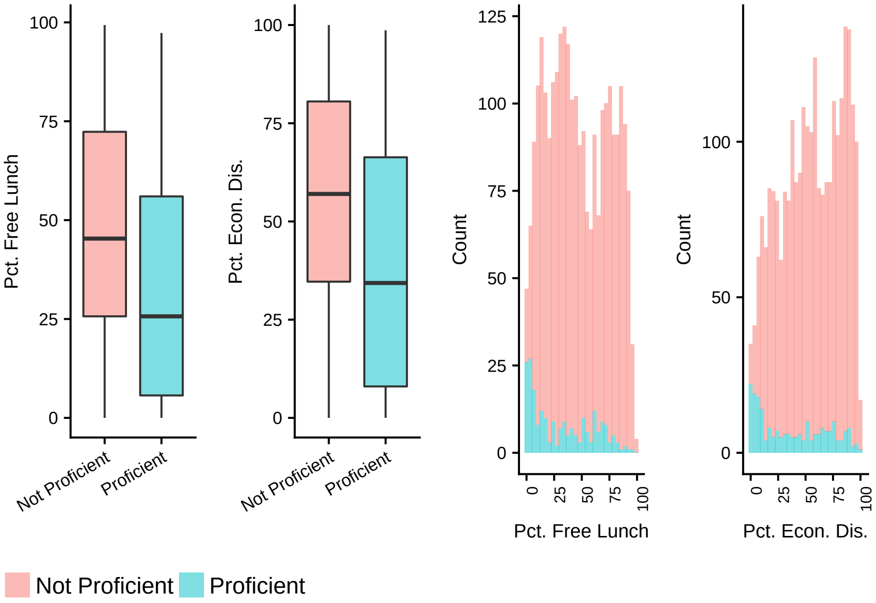

Figure 4: Comparison between 'Free Lunch' designation and 'Economically Disadvantaged' Status

Although charter education is controversial in educational settings, the training data suggest that the proportion of charter schools that are proficient is larger than the proportion of non-charter schools. Figure 5 shows the percentage of charter school compared to non-charter schools by proficiency level. Clearly, for both charter and non-charter schools, "Not Proficient" schools comprise the majority of observations. Nonetheless, a higher percent of charter schools are "Proficient"" compared to non-charter schools. Specifically, 19.8% of charter schools are considered "Proficient," while only 7.7% of non-charter schools are significant. This result reflects similar findings made in research listed in the literature review.

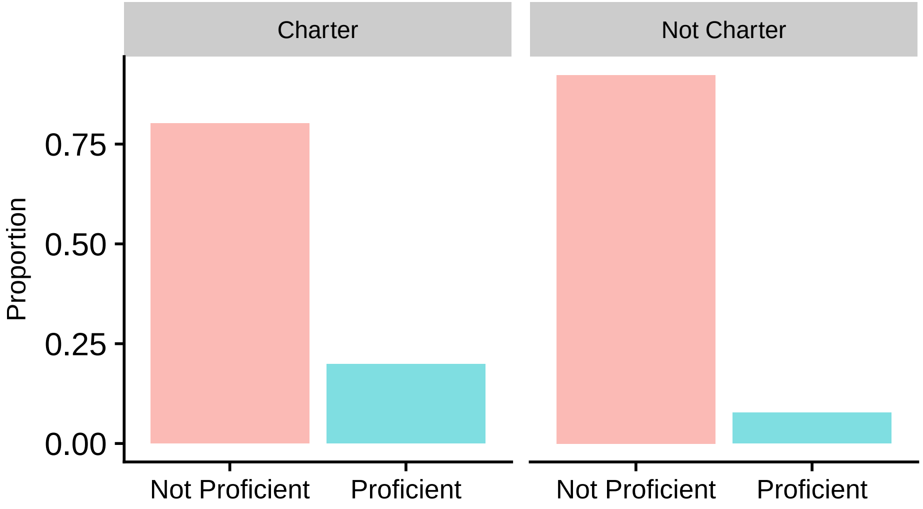

Figure 5: Proportion of 'Proficient' Schools within Charter School Designation

While research has found that students in areas of high crime tend to perform worse than students living in low crime areas, the data used for this project is inconclusive. Figure 6 shows the distribution of property crimes and violent crimes per citizen at the county level separated by schools' proficiency level. The histograms show that, clearly, the vast majority of schools are located in areas in which violent and property crimes per citizen are quite low. The box plots suggest that the median property crime per citizen is slightly higher in areas with “Not Proficient” schools than “Proficient” schools. Counterintuitively, the boxplot for violent crimes suggests that this sort of activity is higher in areas with “Proficient” Schools than “Not Proficient” schools. These patterns do not necessarily align with previous research.

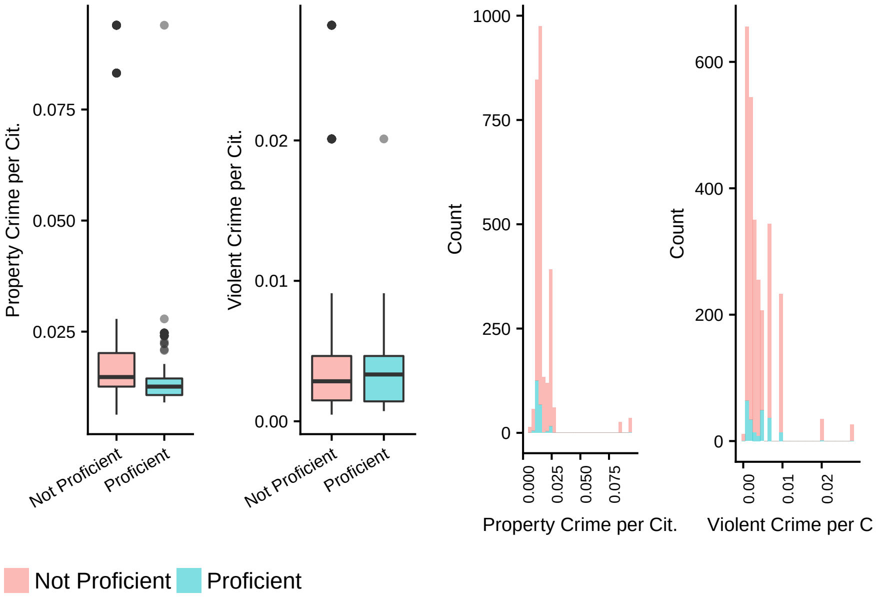

Figure 6: Criminal Activity in locations of ’Proficient’ and ’Not Proficient’ Schools

The exploratory exercise above indicates that a school’s proficiency status might be associated with several of the variables in the data. From this exercise, it seems that proficiency levels differ by racial composition, economic standing, charter status and level of criminal activity. The variables representing these factors will be included in a logistic regression model. In addition to these variables, the regression include variables that research has found to be predictive of student performance but did not show differences in the present dataset. These variables are: suspensions, income, expenditure per student and total population as a proxy for urban settings. A regression model with manually selected variables will be used to assess the association of these predictors and proficiency and to compare with the result of the regularization methods.

Analysis: Variable Selection Methods

This section compares the results obtained from a logistic regression model with variables selected based on previous research (selected variables model) with those obtained using regularization methods. Before doing so, there are important issues to be noted. First, given that social science does not usually concern itself with the effect of class imbalance of an outcome variable in coefficient estimation, this analysis does not address the effect that this characteristic of the data has on the prediction task. Second, a mixing parameter of $\alpha = 0.95$ was chosen via cross-validation and used throughout the analysis for the Elastic Net.

Based on the literature review and exploratory analysis, school proficiency levels vary across racial, economic, neighborhood-related school-related factors. In order to build a logistic regression model with manually selected variables, one variable representing each of these factors had to be selected. Given that these factors were captured by several variables in data and these variables were correlated to each other, only one variable for each concept was included in the logistic regression model. Between the highly correlated pairs PROP_CRIME_PER_CITIZEN and VIOL_CRIME_PER_CITIZEN and PER_ECDIS and PER_FREE_LUNCH, the formers were chosen arbitrarily, as commonly done in social science. Ultimately, the logistic model included the variables: PER_BLACK, PER_HISP, PER_ASIAN, PER_WHITE, PER_ECDIS, PER_SUSPENSIONS, MEAN_INC EXP_PER_ST, PROP_CRIME_PER_CITIZEN, TOTAL_DIST_POP and CHARTER.

The results for the selected variables logistic regression model are shown in Table 2. The results suggest that PER_ASIAN, PER_ECDIS, EXP_PER_ST, CHARTER are significant at the 0.05 level. In particular, a one standard deviation increase in a school's Asian student population is associated with a 1.422 increase the log odds of being proficient. Similarly, a 1 standard deviation increase in a school's economically disadvantaged students is associated with a 1.39 decrease in the log odds. Additionally, a one standard deviation increase in property crimes per citizen decreases the log odds by 0.324. Finally, charter schools are associated with a 3.586 increase in the log odds of being proficient. These results are roughly congruent with previous research.

Before continuing, it is important to note that diagnostic tests were conducted on this model and found that the logistic regression assumptions were reasonably met. The diagnostics revealed several outliers and influential points. However, because the observations were not a result of obvious data errors, the observations were included in the analysis. These diagnostics were omitted due to space limitations.

Table 2: Results of Regression with Manually Selected Predictors

	Dependent variable:
	PROF_LEVEL
PER_BLACK	1.052 (0.924)
PER_HISP	0.932 (0.838)
PER_ASIAN	1.422^{***} (0.430)
PER_WHITE	1.560 (1.337)
PER_ECDIS	-1.298^{***} (0.146)
PER_SUSPENSIONS	-0.082 (0.149)
MEAN_INC	-0.118 (0.082)
EXP_PER_ST	0.297^{**} (0.134)
PROP_CRIME_PER_CITIZEN	-0.324^{*} (0.192)
TOTAL_DIST_POP	0.141 (0.143)
CHARTER1	3.586^{***} (0.426)
Constant	-3.601^{***} (0.149)
Observations	2,661
Log Likelihood	-524.529
Akaike Inf. Crit.	1,073.058
Note:	*p<0.1; **p<0.05; ***p<0.01

In order to compare the results shown in Table 2 with the regression results using regularization methods, the size of the shrinkage parameter λ must be chosen. This parameter is often chosen via cross-validation and is often the value that minimizes misclassification error. This analysis uses the value of λ that is within one standard error of the λ that produced the minimum misclassification error in order to obtain sparser models. For each of the regularization methods, Figure 7 shows the misclassification error (y-axis) at different values of lambda (lower x-axis) and the number of variables with non-zero coefficients at that specific value of lambda (upper x-axis).

On the left, the plot for ridge regression shows that the minimum misclassification error is 0.0665, achieved at $log(\lambda) = -3.87$. At this point, ridge regression assigns non-zero coefficients to all 25 predictors. Additionally, the misclassification error for the λ that is one standard error away from the minimum is 0.071. The middle plot shows that the minimum error for Elastic Net is 0.065, achieved at $log(\lambda) = -6.79$. At this point, Elastic Net includes 20 variables. One standard error above the minimum, the misclassification error is 0.070 at $log(\lambda) = -4.00$. At this point, Elastic Net assigns non-zero coefficients to 5 predictors only. On the right, the minimum error for LASSO is 0.065 achieved at $log(\lambda) = -5.74$. At this point, LASSO includes 18 variables. One standard error above the minimum, the misclassification error is 0.0707 at $log(\lambda) = -4.168$. At this point, LASSO obtains only 5 predictors with non-zero coefficients. For Elastic Net and LASSO, these results suggest that, while including more predictors in the model reduces the misclassification error, increasing the number of predictors from 5 to 18 or 20 produces a marginal improvement. From this point on, the value of λ for each regularization method, is that which is one standard error above the λ that produced the minimum misclassification error. These values of λ are 0.0531 for ridge, 0.0182 for Elastic Net and 0.0154 for LASSO.

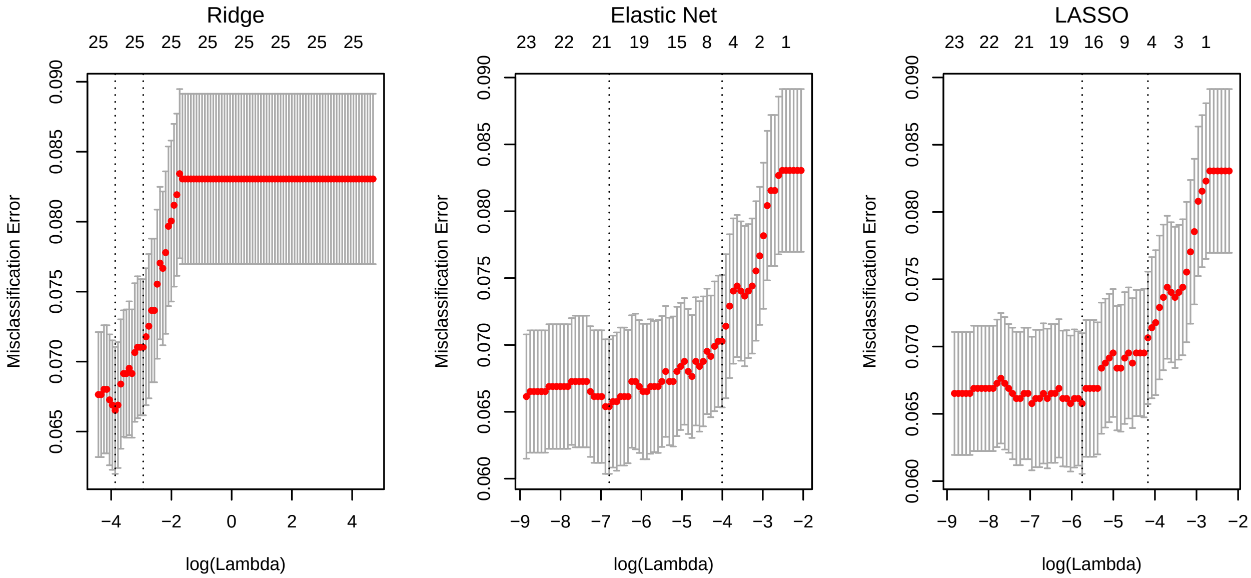

Figure 7: Misclassification at Different Values of Lambda

As mentioned in the methodology section, the value of λ determines the size of predictors' coefficients. Figure 8 shows how different values of lambda affect the coefficient size for the 10 predictors with largest coefficients in absolute terms. For all methods, when $\lambda = 1$, the method produces an intercept only model. Conversely, when $\lambda = 0$, the methods assign non-zero coefficients to all predictors, as in OLS regression. Across each method, the variables CHARTER, PER_ASIAN, VIO_CRIME and PER_ECDIS are the first predictors to achieve "large" non-zero coefficients. Using ridge regression, at the specified value of λ for this method, four predictors with the largest non-zero coefficients in absolute terms are CHARTER, PER_ASIAN, PER_ECDIS, PER_FREE_LUNCH. Using Elastic Net with the appropriate λ, the four predictors with largest non-zero coefficients are also CHATER, PER_ASIAN, PER_ECDIS, PER_FREE_LUNCH for Elastic Net. Similarly, for LASSO, those variables are CHATER, PER_ASIAN, PER_ECDIS, and MEAN_ENROLLMENT. These results suggest that these predictors are the most powerful when explaining the variation in schools' proficiency.

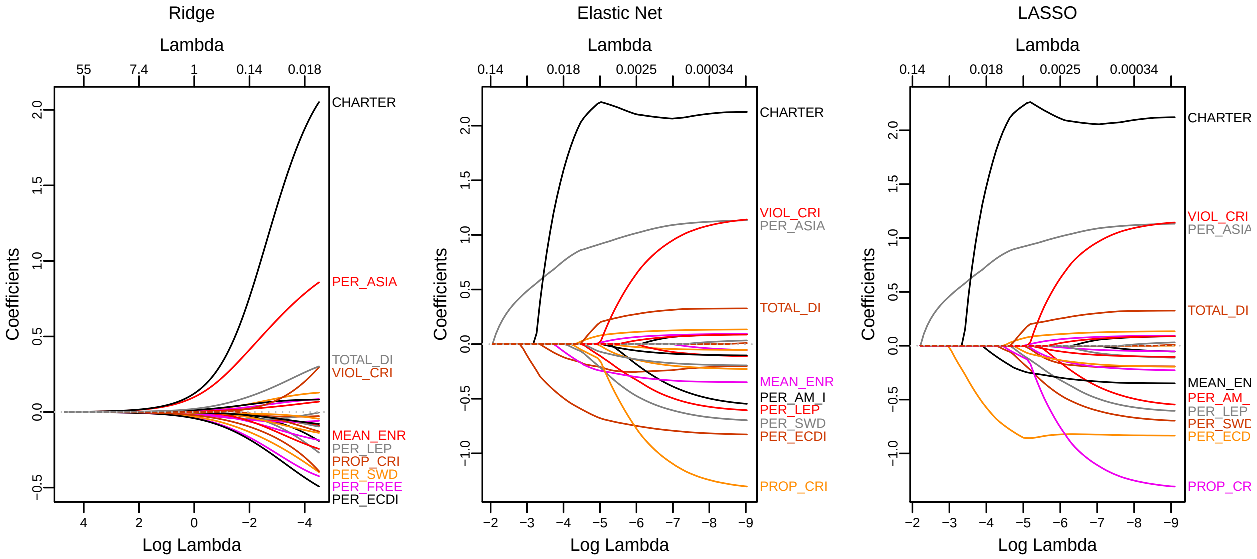

Figure 8: Coefficient Size at Different Values of Lambda

Table 3 shows the specific estimate obtained using ridge regression for the predictors with the 10 largest coefficients in absolute terms at the specified λ as well as a 95% bootstrap confidence interval on each estimate. Specifically, charter status are associated with a 1.32 increase in the log odds of being proficient. A one standard deviation increase in a school's Asian student population is associated with a 0.59 increase in log odds. A one standard deviation increase in PER_ECDIS and PER_FREE_LUNCH are both associated with roughly a 0.3 decrease in log odds of a school being proficient. Although other predictors were excluded from Table 3, they all received small but non-zero coefficient estimates.

Table 3: Coefficient Estimates and 95% Bootstrap CI on Ridge Regression

Predictor	Estimate	2.5%	97.5%
CHARTER	1.3248	0.9363	1.7262
PER_ASIAN	0.5981	0.5396	0.6650
PER_ECDIS	-0.2949	-0.3446	-0.2451
PER_FREE_LUNCH	-0.2682	-0.3127	-0.2240
PER_SWD	-0.2144	-0.4120	-0.0534
TOTAL_DIST_POP	0.1709	0.1121	0.2323
PROP_CRIME_PER_CITIZEN	-0.1689	-0.2291	-0.1014
MEAN_ENROLL	-0.1272	-0.2099	-0.0348
PER_WHITE	-0.1109	-0.1485	-0.0669
PER_FEMALE	0.0881	-0.0025	0.1686

Table 4 presents the specific estimate obtained using Elastic Net for the predictors with the 10 largest coefficients in absolute terms at the specified λ as well as a 95% bootstrap confidence interval on each estimate. Charter status is associated with a 1.6 increase in the log odds of being proficient. On average, a one standard deviation increase in a school's Asian student population results in a 0.76 increase in log odds. Additionally, a one standard deviation increase in PER_ECDIS is associated with roughly a 0.5 decrease in the log odds of a school being proficient. The cells with 0 as lower and upper bounds of the confidence interval correspond to variables that never achieved a non-zero coefficient in the 1000 bootstrap samples used to construct the 95% empirical confidence interval. Additionally, all variables excluded from the table obtained 0 as a coefficient estimate and confidence interval. This result indicate that Elastic Net highlights these few predictors as most substantial in explaining the variation in schools' proficiency.

Table 4: Coefficient Estimates and 95% Bootstrap CI on Elastic Net Regression

Predictor	Estimate	2.5%	97.5%
CHARTER	1.6003	0.7147	2.1894
PER_ASIAN	0.7617	0.6682	0.8614
PER_ECDIS	-0.4955	-0.6786	-0.2494
PER_FREE_LUNCH	-0.1206	-0.3196	-0.0074
MEAN_ENROLL	-0.0673	-0.2214	-0.0009
EXP_PER_ST	0.0000	0.0013	0.0336
MEAN_CLASS_SIZE	0.0000	-0.1667	-0.0005
MEAN_INC	0.0000	-0.0675	-0.0002
MEAN_TEACH_SCORE	0.0000	0.0012	0.0405
PER_AM_IND	0.0000	0.0000	0.0000

Table 5 presents the specific estimate obtained using LASSO for the predictors with the 10 largest coefficients in absolute terms at the specified λ as well as a 95% bootstrap confidence interval on each estimate. The results suggest that charter school is associated with a 1.7 increase in the log odds of being proficient. On average, a one standard deviation increase in a school's Asian student population is associated with a 0.78 increase in log odds. Similarly, a one standard deviation increase in PER_ECDIS results in roughly a 0.64 decrease in log odds of a school being proficient. Finally, schools that are one standard deviation above the mean see a 0.0756 decrease in their log odds of being proficient. The cells with 0 as lower and upper bounds of the confidence interval correspond to variables that never achieved a non-zero coefficient in the 1000 bootstrap samples used to construct the 95% confidence interval. This result indicate that LASSO highlights these few predictors as most substantial in explaining the variation in schools' proficiency.

Table 5: Coefficient Estimates and 95% Bootstrap CI on LASSO Regression

Predictor	Estimate	2.5%	97.5%
CHARTER	1.7012	0.7858	2.2992
PER_ASIAN	0.7840	0.6870	0.8909
PER_ECDIS	-0.6421	-0.7840	-0.2710
MEAN_ENROLL	-0.0756	-0.2393	-0.0014
EXP_PER_ST	0.0000	0.0016	0.0411
MEAN_CLASS_SIZE	0.0000	-0.1862	-0.0011
MEAN_INC	0.0000	-0.0707	-0.0003
MEAN_TEACH_SCORE	0.0000	0.0011	0.0352
PER_AM_IND	0.0000	0.0000	0.0000
PER_BLACK	0.0000	0.0000	0.0000

Given the limited size of the dataset, it is possible that the variables that the regularization methods selected would not be significant given a different sample from the same population. In order to investigate how frequently variables obtained non-zero coefficients in LASSO and Elastic Net, these methods are employed on 1000 bootstrap samples and the number of non-zero coefficients are presented in Table 6 and 7. Table 6 shows that in 1000 samples Elastic Net always gave PER_ASIAN and PER_ECDIS non-zero coefficients. Then, Charter, PER_FREELUNCH and MEAN_ENROLL were used in over 80 percent of regression models. In contrast, the variables shown in the table between PER_REDUCED_LUNCH and VIOL_CRIMES_PER_CITIZEN were never used. These results further indicate that the variables identified previously as important predictors of school proficiency are actually predictive of this outcome.

Table 6: Percentage of Bootstrap Samples Elastic Net Regression included a Predictor

Predictor	Percentage Used
(Intercept)	1.000
PER__ASIAN	1.000
PER__ECDIS	1.000
CHARTER	0.999
PER__FREE__LUNCH	0.831
MEAN__ENROLL	0.824
PER__SWD	0.531
PROP__CRIME__PER__CITIZEN	0.301
PER__FEMALE	0.245
STUD__TEACH__RATIO	0.173
MEAN__CLASS__SIZE	0.093
PER__LEP	0.074
TOTAL__DIST__POP	0.055
PER__HS__DEGREE	0.040
MEAN__INC	0.028
PER__SUSPENSIONS	0.005
MEAN__TEACH__SCORE	0.005
EXP__PER__ST	0.004
PER__NOT__HQ	0.001
PER__REDUCED__LUNCH	0.000
PER__AM__IND	0.000
PER__BLACK	0.000
PER__HISP	0.000
PER__WHITE	0.000
PER__Multi	0.000
VIOL__CRIME__PER__CITIZEN	0.000

Table 7 shows that in 1000 samples LASSO attributed non-zero coefficients to many of the same variables as did Elastic Net. PER__ASIAN and PER__ECDIS always obtained non-zero coefficients. Then, Charter, PER__ECDIS, MEAN__ENROLL were used in over 80 percent of the time. In contrast, the variables shown in the table between PER__REDUCED__LUNCH and VIOL__CRIMES__PER__CITIZEN were never used. Many of the variables that never obtained non-zero coefficients in LASSO are the same that did not obtain non-zero coefficients in Elastic Net. These results further indicate that the variables identified previously as significant predictors of school proficiency are actually predictive of this outcome.

Table 7: Percentage of Bootstrap Samples that produced non-zero Coefficient using LASSO

Predictor	Percentage Used
(Intercept)	1.000
PER_ASIAN	1.000
CHARTER	0.999
PER_ECDIS	0.995
MEAN_ENROLL	0.834
PER_SWD	0.513
PROP_CRIME_PER_CITIZEN	0.251
PER_FEMALE	0.248
PER_FREE_LUNCH	0.196
STUD_TEACH_RATIO	0.173
MEAN_CLASS_SIZE	0.094
PER_LEP	0.085
TOTAL_DIST_POP	0.048
PER_HS_DEGREE	0.039
MEAN_INC	0.028
PER_SUSPENSIONS	0.005
EXP_PER_ST	0.004
MEAN_TEACH_SCORE	0.003
PER_NOT_HQ	0.001
PER_REDUCED_LUNCH	0.000
PER_AM_IND	0.000
PER_BLACK	0.000
PER_HISP	0.000
PER_WHITE	0.000
PER_Multi	0.000
VIOL_CRIME_PER_CITIZEN	0.000

In conjunction Table 3, Table 4, Table 5, Table 6, and Table 7 demonstrate that regularization methods identified similar variables as significant predictors of school proficiency as did the model with manually selected predictors. Specifically, all four regressions highlight charter, PER_ASIAN, PER_ECDIS as significant predictors of school performance. These results reflect similar findings as those mentioned in the literature review, and they reinforce the association between race, economics and educational opportunities. Due to the biased introduced, the coefficients of these important variables obtained using regularization are closer to 0 than they are in the model with manually selected variables. However, more importantly, these results indicate that Elastic Net and LASSO can produce sparse models that identify factors associated with school proficiency with fewer predictors than the logistic regression model with manually selected variables.

These results also highlight that regularization methods are capable of appropriately avoiding problems of multicollinearity despite highly correlated variables in the dataset. For example, ridge regression and Elastic Net both include the highly correlated variables PER_ECDIS and PER_FREE_LUNCH in the model, but they assign one of the variables a much larger non-zero coefficient to one of the variables than the other. On the other hand, LASSO only included PER_ECDIS in the final model. In contrast, when manually selecting the predictors to include in the regression, the researcher must inspect the variable correlations and arbitrarily determining which variables to include. Employing regularization methods offer a more efficient process to select variables between correlated predictors.

While it is clear that regularization methods can produce sparse models, how well do they explain the variation in the outcome? Comparing Table 8 and the plots in Figure 9 suggests that decreasing the number of predictors in the model via LASSO and Elastic Net does not result in a sharp decrease in the amount of variation in the data explained by the models. Table 6 shows that the manually selected model explains about 31 percent of the variation. Figure 9 shows the fraction of Deviance explained by the models using regularization using different number of predictors. The plots show that with 25 predictors for ridge, 5 for Elastic Net and 5 for LASSO, the regularization methods explain about 25% of the variable in the data. Although these methods explain a smaller amount of variation, the decrease is not very substantial and results in sparser, more interpretable models.

Table 8: Overall Model Significant and Pseudo R-Squared using Selected Variables

Overall Sig.	1.18256540175288e-94
Pseudo R-Squared	0.3112

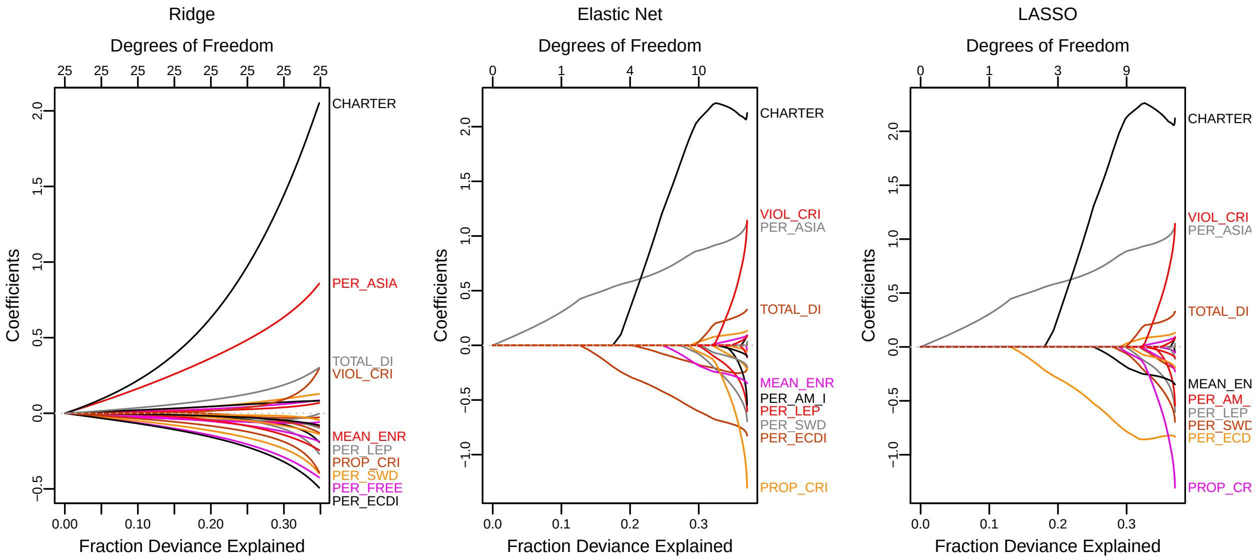

Figure 9: Coefficient Size at Different Values of Lambda

A regression model's ability to explain an outcome can also be assessed based on its predictive performance. Hindman (2015) argues that, although social scientists are primarily interested in obtaining accurate point estimates rather than prediction, regression models that accurately capture social phenomena should be able to predict out of sample observations relatively well. With that in mind, the tables below show the performance of each model (manually selected, ridge, LASSO, and Elastic) on the test set. The tables also present 95% confidence intervals on the overall test accuracy of models trained on 1000 bootstrap samples. Given the class imbalance mentioned earlier, the overall accuracy rate is high despite substantial misclassification of the minority class. However, since prediction was not the main goal of this project, methods achieve a better classification of the minority class was not investigated. Overall, examining the confusion tables along with these confidence intervals provides some insight into these models serve to explain social issues.

Table 9 shows the confusion matrix for the classification test results of the model with the variables selected manually and developed on the original dataset, and Table 10 shows this models performance over 1000 bootstrap samples. In Table 9, the rows represent the observed classes of the data points, while the columns represent predicted class. Out of 610 "Not Proficient" observations in the test data, the model misclassified only 4 observations. Out of 56 "Proficient" observations, the model only classified 10 correctly. Over 1000 bootstrap samples, 95 percent of the models obtained an overall test accuracy rate that ranged between 0.92 and 0.932.

Table 9: Confusion Matrix for Regression with Selected Predictors

	Not Proficient	Proficient
Not Proficient	606	4
Proficient	46	10

Table 10: 95% CI Accuracy Rate of Regression using Selected Variables

2.5% Bound	97.5% Bound
0.92	0.932

Table 11 shows the confusion matrix for the classification test results of the ridge regression and developed on the original dataset, and Table 12 shows this model's performance over 1000 bootstrap samples. In Table 11, the rows represent the observed classes of the data points, while the columns represent predicted class. Out of 610 "Not Proficient" observations in the test data, the model classified all observations in this level accurately. Out of 56 "Proficient" observations, the model only classified 8 correctly. In this respect, this model performs worse when accurately classifying observations of the minority class. Over 1000 bootstrap samples, 95 percent of the models obtained an overall test accuracy rate that ranged between 0.923 and 0.932. The overall accuracy in the bootstrap samples is the same for the ridge regression and the model with the manually selected variables.

Table 11: Confusion Matrix for Ridge Classifier

	Not Proficient	Proficient
Not Proficient	610	0
Proficient	48	8

Table 12: 95% CI Accuracy Rate of Classifier using Ridge

2.5% Bound	97.5% Bound
0.923	0.932

Table 13 shows the confusion matrix for the classification test results of the Elastic Net, and Table 14 shows this model's performance over 1000 bootstrap samples. In Table 13, the rows represent the observed classes of the data points, while the columns represent predicted class. Out of 610 "Not Proficient" observations in the test data, the model classified 603 observations accurately. Out of 56 "Proficient" observations, the model only classified 11 correctly. Compared to the model with manually selected variables and ridge, Elastic Net seems slightly better able to classify observations from the minority class. Over 1000 bootstrap samples, 95 percent of the models obtained an overall test accuracy rate that ranged between 0.923 and 0.931. The overall accuracy in the bootstrap samples is roughly equal for the Elastic Net as it was the manually selected variables. Together, these results suggest that, with a much more sparse model, Elastic Net is as capable of achieving similar prediction performance as the manually select model, which implies that only a few variables in the data truly represent factors associated with school proficiency.

Table 13: Confusion Matrix for Elastic Net Regression

	Not Proficient	Proficient
Not Proficient	607	3
Proficient	45	11

Table 14: 95% CI Accuracy Rate of Classifier using Elastic Net

2.5% Bound	97.5% Bound
0.923	0.931

Table 15 shows the confusion matrix for the classification test results of the LASSO, and Table 16 shows this model's performance over 1000 bootstrap samples. In Table 15, the rows represent the observed classes of the data points, while the columns represent predicted class. Out of 610 "Not Proficient" observations in the test data, the model classified 606 observations accurately. Out of 56 "Proficient" observations, the model only classified 11 correctly. Compared to the model with manually selected variables and ridge, LASSO slightly better able to classify observations from the minority class and exactly the same as the Elastic Net. Over 1000 bootstrap samples, 95 percent of the models obtained an overall test accuracy rate that ranged between 0.923 and 0.931. The overall accuracy in the bootstrap samples is roughly equal to that obtained by the previous models. Together, these results suggest that, with a much sparser model, LASSO is as capable of achieving similar prediction performance as other methods, which implies that only a few variables in the data truly represent factors associated with school proficiency.

Table 15: Confusion Matrix for LASSO Regression

	Not Proficient	Proficient
Not Proficient	606	4
Proficient	45	11

Table 16: 95% CI Accuracy Rate of Classifier using LASSO

2.5% Bound	97.5% Bound
0.923	0.931

Discussion

OLS and logistic regression play an important role in helping researchers develop social theories that are supported by quantitative data. Despite the importance of these methods in social science, few articles dedicate a significant portion of their work to explain the criteria used to select variables used in a regression model. Unfortunately, such practice leaves the reader unable to fully understand the researchers' decisions and opens the possibility for unethical behavior. To avoid such pitfalls, social science researchers could rely more heavily on regularization methods such as ridge regression, LASSO and Elastic Net.

This paper compared the results of logistic model with manually selected variables to those of ridge regression, LASSO and Elastic Net. Interestingly, all regularization methods identified the same factors as being associated with school proficiency, and these factors overlapped with the significant predictors in the model with manually selected variable. While regularization methods do not eliminate the need for thorough data exploration, these methods can help researchers identify important factors among thousands of possible predictors and avoid problems of multicollinearity. Although these methods return biased coefficient estimates, the direction of the association was the same for all three methods and the manually selected model. Such a result suggests that, once these variables have been identified as significant, researchers can explore methods to "de-bias" their coefficients and obtain more accurate point estimates. Overall, regularization methods can help social science researchers develop sparse models that reflect social patterns and interactions.

References

Ainsworth-Darnell, James, and Douglas Downey. “Assessing the Oppositional Culture Explanation for Racial/Ethnic Differences in School Performance.” American Sociological Review 63, no. 4 (1998).

Columbia University Mailman School of Public Health. “Ridge Regression.”

Faraway, Julian. “Linear Models with R.”" 2nd ed. Chapman and Hall/CRC.

Hindman, Matthew. “Building Better Models: Prediction, Replication and Machine Learning in Social Sciences.” Annal, AAPSS 659 (2015).

Logan, John, Elisabeta Minca, and Sinem Adar. “The Geography of Inequality: Why Separate Means Unequal in American Public Schools.” Sociology of Education 85, no. 3 (2012).

Qian, Junyang and Trevor Hastie. “Glmnet Vignette.” (2014).

Logan, John, Elisabeta Minca, and Sinem Adar. “The Geography of Inequality: Why Separate Means Unequal in American Public Schools.” Sociology of Education 85, no. 3 (2012).

Tibsharani, Robert. “Regularization: Ridge Regression and the LASSO.” Statistics 305. (2007).

Zou, Hui, and Trevor Hastie. “Regularization and Variable Selection via the Elastic Net.” Journal of the Royal Statistical Society 67, no. 2 (2005).

Index: List of Variables and Abreviations

The following is a list of variable abreviations:

```
-PROF_LEVEL: Proficiency level
-PER_FREE_LUNCH: Percent of student body in a school on free lunch
-PER_REDUCED_LUNCH: Percent of student body on reduced lunch
-PER_LEP: Percent of student body that is English Language Learner
-PER_AM_IND: Percent of student body that is Native American
-PER_BLACK: Percent of student body that is Black
-PER_HISP: Percent of student body that is Hispanic
-PER_ASIAN: Percent of student body that is Asian
-PER_WHITE:Percent of student body that White
-PER_Multi: Percent of student body that Multiracial
-PER_SWD: Percent of student body that has a disabiliy
-PER_FEMALE: Percent of student body is female
-PER_ECDIS: Percent of student body that is economically disadvantaged
-PER_SUSPENSIONS: Percent of student body that has received a suspension
-MEAN_CLASS_SIZE: Average class size in grades 3-8
-MEAN_ENROLL: School's average enrollment in 2014-2015
-PER_NOT_HQ: Percent of teachers that are not highly qualified
-STUD_TEACH_RATIO: Ratio of students to teacher
-PER_HS_DEGREE: Percent of people in school district that has a high school degree
-MEAN_INC: Mean income at the school district  level
-EXP_PER_ST: Expenditure per student
-MEAN_TEACH_SCORE: Average teacher assessment score
-CHARTER: Whether school has charter status
-VIOL_CRIME_PER_CITIZEN: Violent crimes per citizen at the county level
-PROP_CRIME_PER_CITIZEN: Property crimes per citizen at the county level
-TOTAL_DIST_POP: Total population at the school district level
```