Richard Tran

3/22/2019

STAT 404

Student Performance on Exams based on Socioeconomic Factors

According to the Office of Management and Budget, $1.1 trillion dollars of tax payer money was allocated towards discretionary spending in 2015. The biggest category for this portion of the budget spending is the Pentagon and related military program at 53.71% of discretionary spending, followed not so closely by government (6.5%) and education (6.28%). From working at large outreach programs to private tutoring prep school students, I've noticed a huge gap in knowledge across students with different backgrounds. Every child has to go through the school system, yet, just a fraction of the overall tax budget. The business question for this project was simple: Where is the best allocation of tax payer money be into education?

The data set I used analyze this business question was found on Kaggle. The generated sample of 10,000 students had 8 different variables. The categorical variables included gender (M / F), race/ethnicity (Groups A – E), parental level of education (some high school – Master's degree), lunch status (Standard vs Free/Reduced), and test preparation course (Yes / No). The numerical variables were how well the students placed on Math / Reading / Writing exams on a scale from 1-100.

	gender	race	degree	lunch	prep	math	reading	writing
0	male	group D	bachelor's degree	standard	none	87	81	84
1	male	group A	high school	free/reduced	completed	88	86	81
2	male	group D	associate's degree	standard	none	70	58	63

Figure 1: Example of the data set

All of the other variables were relatively normally distributed with no major outliers in the data. Of course, some of the race/ethnicity and parental education levels were uneven; however, there were enough samples in each category to assume it's normal.

Since the test scores ended up being normal, I decided to standardize them using the normal distribution. I created a new variable 'pass' that determined whether the student would pass or fail school based on the variables describing their socioeconomic status. According to the National Center for Education Statistics, the graduation rate for public high school students was 84 percent in the 2015-2016 school year. This is roughly equal to the area above 1 standard deviation below the mean in the normal distribution. I deemed anyone with an average test score in this range to pass. This equated to roughly a 53 percent average across the three tests.

After assigning all the students a binary pass (1) and no pass (0) value, I ran my model across multiple types of regressions. I originally thought about doing a linear to model the data set. This would give me a rough estimate as to how influential each variable was; however, I found that the regression results did not properly reflect the data sets. I only wanted to use the socioeconomic factors as predictors; however, the variables themselves did not explain the score very well. The lack of any numerical variables carried over so there was very little accuracy in the linear regression.

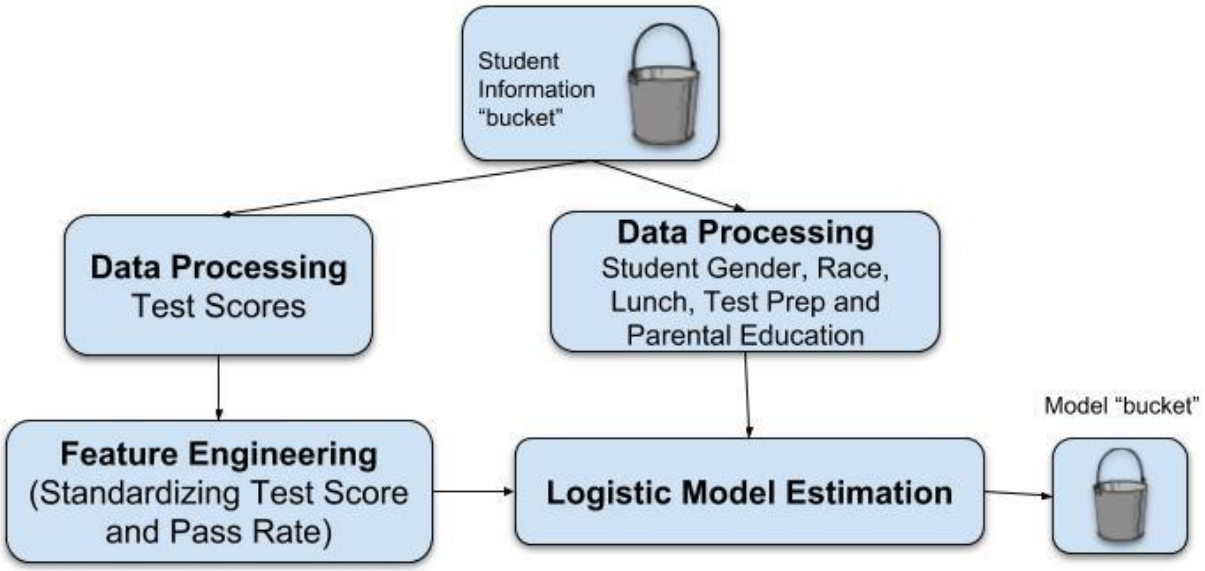

Figure 2: Architecture Diagram for Training

Figure 2 is an example of the architecture diagram for training my data set. Student information is processed in two different ways. The test scores were standardized in order to create the pass variable. The socioeconomic factors were turned into binary values for the logistic model estimation. The model uses this pass rate and socioeconomic factors to output a formula that determined if the student should pass or not.

In this project, I was mainly interested in whether a student would be able to pass high school given their socioeconomic status. Since this was just a binary classification question, I decided that the best test would be the logistic regression. The logistic regression output gives estimated probabilities using the logistic function. I based the effectiveness of the predictor variables on these probabilities. If the value of this logistic regression formula was greater than .5, the student was likely to pass high school given his/her factors.

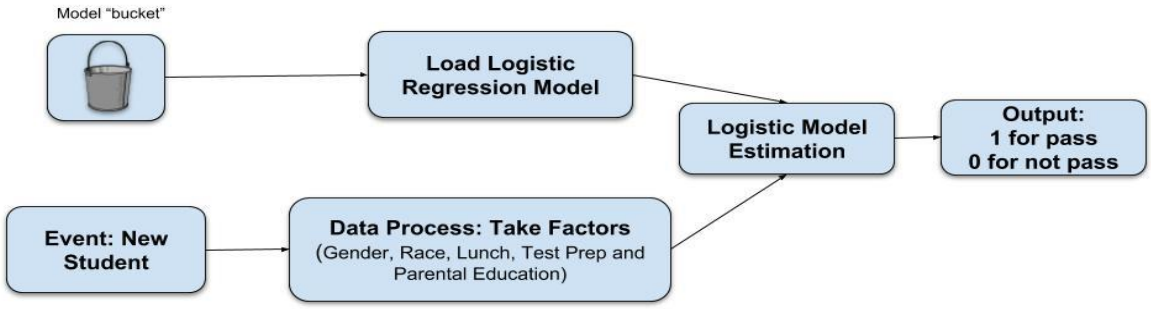

Figure 3: Architecture Diagram for Scoring

Figure 3 is an example of the architecture diagram for scoring. After the logistic model is computed, any new student information can be inputted into the formula to output whether the student passed or not.

	Features	Coefficient	odds
	race_group A	-0.060547	0.941250
	lunch_free/reduced	-0.081301	0.921916
	race_group B	-0.109264	0.896494
	degree_high school	-0.190592	0.826470
	degree_some high school	-0.263665	0.768231
14	lunch_standard	0.680471	1.974808
15	prep_completed	0.561419	1.753159
10	degree_master's degree	0.551538	1.735920
6	race_group E	0.504315	1.655851
0	gender_female	0.441778	1.555470

Figure 4: Head and tail of the Logistic Regression output

As seen in Figure 4, the most influential socioeconomic traits seem to be the type of lunch and the parental level of education. Students who had parents that only attended high school ended up having the lowest pass rates. Following parental education level, whether the student had free/reduced lunch or if they were in race group A or B had negative coefficient pass rates. On the other side, we see that the most influential factor for passing was having standard lunch. Usually students with free/reduced lunch are well below the median income so it's very apparent that income levels have a huge impact on student performance.

By using this model, I can pinpoint certain socioeconomic factors that the government could lend a helping hand. Since having parents that only attended high school or certain race groups had the lowest odds ratio, grants/scholarship/outreach programs should be implemented for first generation students and these minority groups. The government funding should go directly to benefit teachers and students so that they may do their research on what's best for the next generation. Any further steps to expand on this model would be to follow the student's progress as these programs are implemented. If we were to see any increase in these students, it'll be more beneficial for the United States than any dollar spent on the military.

References

https://www.nationalpriorities.org/budget-basics/federal-budget-101/spending/

https://nces.ed.gov/programs/coe/indicator_coi.asp

https://www.kaggle.com/spscientist/students-performance-in-exams