**Predicting Student Exam Performance**

Dustin Casey

Department of Data Science, Bellevue University

DSC 680: Applied Data Science

Professor Catie Williams

October 16th, 2022

**Abstract**

Using student demographic data to predict which students are more prone to struggle with exam scores and in the classroom.

**Business Problem & Background/History**

The problem I am attempting to solve is to identify the students that perform poorly on exams versus those that perform well and see if their backgrounds have any impact on their performance. I believe looking at a variety of personal, social, and economic factors that I can help identify those students and schools for school districts, who can redistribute and allocate their resources accordingly. Using predictive analytics and machine learning models, I can identify and predict students who may be more likely to struggle with passing these tests. From there, I can allocate resources to students and schools who may need additional help.

It is a common theory that a child’s background affects how they perform in school, which could affect their options for college and thus impact the rest of their lives. If this theory is correct, a way that could help combat the trajectory of a child’s life could be the resources they receive during the school day. The problem I am attempting to solve is to identify the students that perform poorly on exams versus those that perform well and see if their backgrounds have any impact on their performance. I can help identify those students and schools for school districts, who can redistribute and allocate their resources accordingly.

By looking at several variables including a student’s race/ethnicity, their parent’s education levels, their gender, whether they prepared for the exam, or whether they could afford lunch, thus giving me an indication of their family’s financial status, I hope to draw some conclusions that can help me to predict the children that will perform poorly. By predicting which children will perform poorly, I can identify which students could potentially need additional resources that can help them to be more successful not only throughout school, but to help them succeed in life. On the other hand, I could also predict which students perform better and do not need additional resources to help them, which would allow schools to save money on resources or distribute those resources to students who need them more. Finally, I could also see which students should be performing better than they are and maybe that could point me in the direction of identifying a learning disability that has yet to be diagnosed.

**The Data**

The first step of any good project process is to identify the data you will be working with, then to go through and understand what the different fields and variables mean and establish any assumptions that have been made about the data. My dataset is fictional, came from Royce Kimmons, and was found on Kaggle. The website is listed at the end of this document. The dataset in totality has 1000 rows and 8 columns, which consist of math, reading, and writing scores, as well as a variety of personal, social, and economic factors that could influence the scores. The only assumption I made about the data is that all the students had the same equipment and testing environment. While there are sometimes distractions while taking exams, I have no idea what kind of distractions would bother one student over another, and I do not have any data on that. So, for my purposes, I am assuming that all students had access to the same equipment and environment.

In the dataset, the 1000 rows represent 1000 different students. There are eight columns, three of which are math, reading, and writing scores. The other five columns are ‘gender’, ‘race/ethnicity’, ‘parental level of education’, ‘lunch’, and ‘test preparation course’. For the ‘gender variable’, there are two options, ‘male’ or ‘female’. For ‘race/ethnicity’, there are groups A, B, C, D, and E. I do not know which race/ethnicity is represented by each group since this is a fictional dataset, but I believe that information could be made available if this model were to actually be used by a school district The ‘parental level of education’ variable consists of ‘some high school’, ‘high school’, ‘some college’, ‘associate’s degree’, ‘bachelor’s degree’, and ‘master’s degree’. For the lunch variable, there is either ‘standard’ or ‘free/reduced’. For the test preparation course, the two options are ‘completed’ or ‘none’. The parent’s level of education could have a personal impact on the student’s learning. If one student’s parent has a master’s degree, I might expect them to perform better than a student’s parent who has only completed some high school. For lunch and test preparation course, these variables could have a social and economic impact on the student. Perhaps the student does not come from a financially stable background, and they cannot afford the standard lunch or the fee of a preparation course. Or maybe a student can afford the standard lunch but had a social conflict on the day the test preparation course was offered, and they could not make it.

**Methods/Analysis**

Given that this was a relatively small and fictional dataset, there was not much data cleansing that I had to do. However, I did do a few things in order to try and avoid any potential pitfalls or problems that I might have encountered farther down the line. One thing I did was remove all punctuation from the various variables. For example, in “bachelor’s degree” there is an apostrophe in the word “bachelor’s”. I removed the apostrophes in that field and every other field, so that when dealing with those values down the road I could avoid any issues. Another step was to replace any blank space with an underscore in the headers. For example, the “Test Prep” variable has a space between the words ‘Test’ and ‘Prep’. I replaced that blank space with an underscore. After that, I looked at some correlation charts between math and reading scores, math and writing scores, and reading and writing scores. There was a strong, positive correlation between all these scores, especially between the reading and writing scores. This led me to believe that students who performed well on one test, most likely did well on the other tests too. Having discovered that, I decided to create a total score variable, which combined all the scores from the math, reading, and writing tests. Finally, I conducted some initial exploratory data analysis, so that I could see the distribution of my different column variables.

Chart, bar chart

Description automatically generated

This EDA allowed me to discover several things about my data relatively easy. I discovered that more students did not complete a test prep course before the exam. Then I found out that slightly more females were tested than males. I was also able to identify that ethnicity Group C had the most participants. Additionally, I saw that more students had the standard lunch than the free/reduced version.

My goal was to predict student performance on the three different test types (reading, writing, and math). The two methods employed to evaluate this was to create predictive models to predict their combined score (liner regression model) and to predict whether they would pass or fail (logistic regression model).

For linear regression models, I tested three models to see their effectiveness, accuracy, and model fit. I observed Simple Linear Regression (SLR), Random Forest Regression (RFR), and Extra Trees Regression (ETR). When it came to model evaluation, the linear regression models did not perform very well. The best fitting model from an R2 standpoint was the SLR model with a value of 0.216, which is still very poor. The RFR and ETR models had negative R2 values, meaning they have a worse fit than if you just took the mean score and drew a line at that point. I also attempted to scale my input and target features to see if that would have any affect on the model accuracy and fit. I used sklearn’s StandardScaler which helped standardize features by removing the mean and scaling to unit variance. In the end, even with scaling, the results were the same (my models still had poor fit). Overall, I was disappointed that the regression models were not doing a great job of predicting scores, so I changed my tactics to change my target variable from an exact score prediction to be whether a student would be likely to pass or fail the tests.

This led me to employ a logistic regression model where the target variable was pass or fail (failure being a 1 if the student scored less than 60%). The logistic regression model worked a lot better and was much more accurate. My unweighted model had an accuracy of 0.787, precision of 0.706, recall of 0.308, and F1 score of 0.429. This is a much better fit than my linear regression models, all while still achieving my goal of identifying students at risk of poorer performances. I also ran a weighted version of my logistic regression model to place a higher emphasis on recall (limiting false negatives in favor of false positives) to help ensure that I am not missing as many students who need additional help. The weighted version was less accurate (0.693) but had a significantly higher recall (0.769).

Graphical user interface, application

Description automatically generated

**Conclusion**

After formulating a plan to identify and assist students who many need additional help to improve their testing performance, I explored what factors played the biggest impact on test scores as well as created predictive analytics models to predict which new students may be at risk. Whether they ate the standard lunch or had the free/reduced version was one of the highest correlated and significant features in test results. Outside of that, there were not many highly impactful or correlated features. As for predicting test scores, my linear regression models were unsuccessful in predicting an exact score for students and yielded poor R2 scores as a result. Moving forward, I will use my weighted logistic regression model to help predict students who are at a higher risk of failing the tests. This will assist me in identifying students who may need more assistance in preparation for these tests. I can also allocate more resources to those schools with students that need the most assistance.

**Assumptions/Limitations/Challenges**

Because this dataset is fictional, it is somewhat limited in size with only 1,000 records. I also make the assumption that all students have access to the same equipment and environment. Because of these limitations, there were challenges at times to accurately predict which students would be more prone to struggles.

**Future Uses/Recommendations/Implementation Plan**

For how this should be implemented, I would start by using the results to see which schools and districts have more students who are prone to struggle and allocate resources to those districts. I could also use this data to start after school programs to give students more options or opportunities to receive help.

**Ethical Assessment**

Obviously, there are ethical concerns that arise when dealing with a topic like this. Just because a student’s background might indicate they would be a poor performing student, that is not always the case. Just like a student’s background might suggest they would perform well, does not mean they will. I cannot individually pick and choose the students that receive help based off of their backgrounds, just like I cannot force a student to use these additional resources. At the end of the day, it is up to the student to receive help if they want it. All I can do is help school districts allocate which schools need additional resources more than others.

**References**

Kimmons, R. (n.d.). Exam scores. Retrieved October 12th, 2022, from http://roycekimmons.com/tools/generated\_data/exams

Seshapanpu, J. (2018, November 09). Students performance in exams. Retrieved October 12th, 2022, from https://www.kaggle.com/spscientist/students-performance-in-exams

**Appendix**

Dataset Codebook and Field Descriptions:

Graphical user interface, text, application, email

Description automatically generated