Austin James

5/2/2024

MGT-665

Machine Learning SAT Model

With the rise of technology and using data to solve world problems, education was also included to try and solve some of the most difficult challenges that the education system has faced. From how to improve education in low-income schools to the benefits of private school's vs public schools, using data and technology has helped us understand these challenges better and has brought us closer to our goals. My challenge was to find a problem within the school system and see if I could produce a solution to solve it using data and modeling technology. The challenge that I decided to work on was to make a model that would predict SAT scores. This model is to help us predict the future on how students will perform on scores and be able to see what might need to be the focus on which subjects in the future.

I was able to find data on past generations of SAT test taking from 1971 to 2016 with scores in all three aspects of critical reading, mathematics and writing that was later introduced into the test taking in 2006. This data was split between the boys and the girls for each section and would give a total mean score for that section for that year. So, with data already in my hands and the goal is to predict the future so I determined that the best model would be a supervised linear regression model since regression models focus more on predicting the future with supervised data that we already have created. The pictures show the coding that was done for each data set for math reading and writing. I decided to split the data 50/50 when creating the train and test data sets. A screenshot of a computer program

Description automatically generated

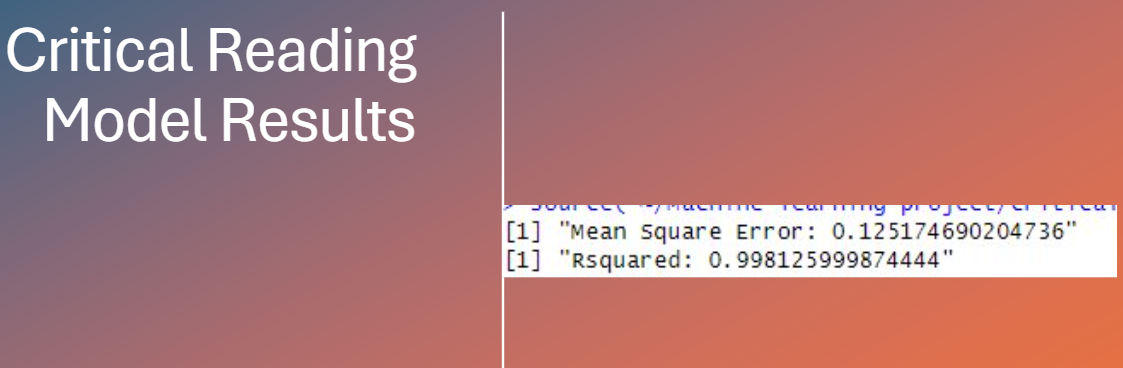
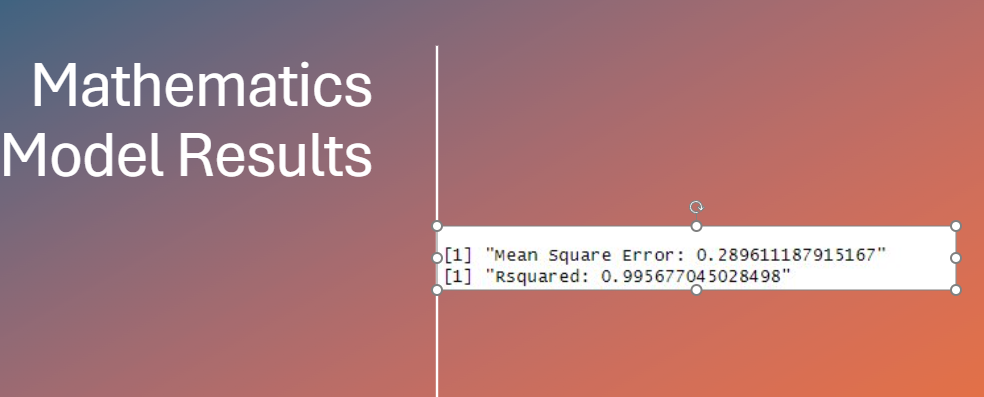
A screenshot of a computer program

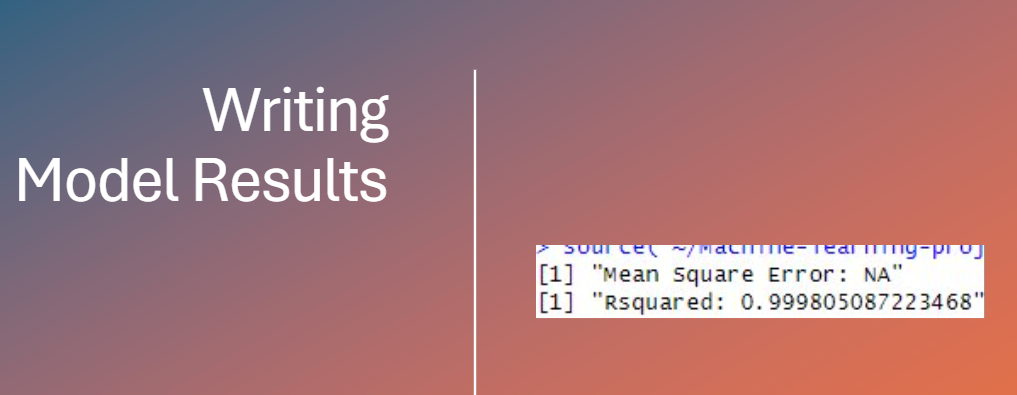
Description automatically generated

A screenshot of a computer program

Description automatically generated

Next was the important results that came of the model that was created. It was a shocking result that surprised me with all three models having extremely high accuracy with exceptionally low chance of error.





It was exciting to see that the model I created has such accuracy. However, with it so high I think it has flaws that impact the accuracy of the predation model. The first being that the data was only able to go as far back as 1970 so there was not a wide range of dataset to use, and this could have increased the accuracy due to the small sample size. Also, the variables were only from the boys scores and the girls scores. These were the only two variables used in the data set, and it could have been expanded to include other factors like state average scores, school funding in that state, average family income, and ethnic groups. The lack of variables could have also increased the model's accuracy, as there was less variation between them.

This model is a good starting point but can be expanded on with more time and data variation. This model in the future can be used to help solve the problem of how to increase the scores of SAT testing and help teachers understand what needs to be focused on more in the future.

Reference

Average SAT scores of college-bound seniors (1952. (n.d.). <https://www.erikthered.com/tutor/historical-average-SAT-scores.pdf>

Aldric, A. (n.d.). *Average SAT scores over time: 1972–2023*. Average SAT Scores Over Time: 1972–2023. <https://blog.prepscholar.com/average-sat-scores-over-time>