Arun Agarwal

March 25th, 2021

Deep Analysis and Prediction of the Effects of Various Factors on the Performance of Students on Examinations

**I. Introduction:**

The future of every society, great and struggling, is in the hands of its young people. These future leaders and their store of knowledge, their abilities to reason, their worldviews and skills will determine the improvements and problems the rest of the century will bring. Education has been the path to better opportunity for generations of American strivers. By 1970, America had the world’s leading educational system, and until 1990, the gap between minority and white students was significantly closing. However, education gains in this country have plateaued since then, and educational inequality has become one of America’s most vexing problems.

According to Economist Rolan Fryer, in 2016, only 44 percent of American students are proficient in reading and math. Worse, the proficiency of African American students, many of them in underperforming schools, is even lower. And in a study by the Education Commission of the States, black and Hispanic students in kindergarten through 12th grade performed on par with the white students who sit in the lowest quartile of achievement. This gap in education starts as early as kindergarten, where black children are eight months behind their white peers in learning. As each year passes, the gap widens.

The issue extends beyond race, as other key factors causing this inequality—identified by numerous reputable agencies like the United States Department of Education—include poverty rates (which are three times higher for blacks than for whites), parental level of education, gender, diminished teacher and school quality, unsettled neighborhoods, ineffective parenting, personal trauma, and peer group influence, which only strengthens as children grow older.

Around the world, 59 million children of primary school age are being denied an education, and nearly 65 million for secondary school. Yet, even for students with the privilege of having access to a public education system with billions of tax dollars invested into it, many students still come out with poor performance scores across all subjects. This poor performance leads to a competitive atmosphere where students take drastic and, sometimes, counterintuitive measures (cheating) to fight for every percentage point in hopes on increasing their chances for a successful future. However, an understanding of the various causes of poor test scores can benefit both the learning of the student and teaching style of the educator.

The purpose of an examination is to understand the ability and learning of a student; thus, through careful analysis of patterns between parental level of education, gender, race, reading scores, writing scores, math scores, and other factors, one can understand which aspects have the largest impact on test outcomes as well as what are the best ways to improve student scores for the future. Furthermore, the improvement of test scores would benefit the economy at a national level by creating equality, leading to major GDP growth, and building a modern society. Correlated with improved test scores is a stronger education system, which for the learner, creates more employment opportunities, secures a higher income, develops problem-solving skills, provides a prosperous and happy life, and educates them to give back to the community.

Prediction of student academic performance in mathematics, reading, and writing based on various demographic and socioeconomic statistics can be performed through creation of various data science models such as linear regression, logistic regression, and k-NN. A model with good RMSE, MSE and MSAE scores signifies the model predicts student’s performance well, making the data useful in identifying methods to improve student performance for the future. The early detection of students who are vulnerable to suffering academic failure (through use of these models) can subsequently be used to design new teaching/mentoring strategies for a overall strengthening education system and society.

Motivation: Educational inequality has always been a large passion of mine due to the poor socioeconomic and demographic background I was brought up in that led me to learn in an impoverished education system. Furthermore, while I have always been able to push harder to reach higher levels of success in life, I realize that not all students can obtain such milestones due to the various external factors that prevent them from obtaining educational equality. After doing thorough research into developing a reimagined approach to education to better support educational equality, I have been motivated to utilize the skills I have gained in the data science sector to build models that would help one to understand which factors are the true causes of poor performance in school. After finding a Kaggle dataset, provided by Royce Kimmons, with student data on gender, race/ethnicity, parental level of education, and test scores, I have decided that I could perform analysis myself with the dataset to determine which factors have major impacts on student performance. It is my hope that others with more power over begetting change in the education system will examine my models and results to work to rectify the system.

Related Project: [Regression Models for Predicting Student Academic Performance in an Engineering Dynamics Course](8.%09https:/peer.asee.org/regression-models-for-predicting-student-academic-performance-in-an-engineering-dynamics-course.pdf), Created by Shaobo Huang and Ning Fang, staff members at Utah State University. By collecting, compiling, and processing data for 239 undergraduate students over three semesters, they were able to create multivariate linear regression models to predict student academic performance in Engineering Dynamics—a high enrollment, high-impact, and core engineering core that almost every mechanical or civil engineering student must take. The purpose of collecting and modeling this data was to help instructors develop a good understanding of how well or how poorly the students in their classes will perform so that instructors can take proactive measures to improve student learning. The inputs/independent variables of the model include the student’s cumulative GPA, gender, race, and other factors; the output/dependent variable of the models is a student’s final exam score in the Dynamics course stated above. Multiple criteria are utilized to evaluate and validate the predictive models, including R-square values, and average prediction accuracy. A good prediction was defined to have prediction error of plus or minor 10%. The results showed that the developed models had average prediction accuracy of 86.8% to 90.7%. This project is related to mine because it does a directly similar task of using models to determine student performance on exams.

**II. Proposed Work:**

Idea/Goal: The goal of this project is to use various supervised learning techniques including linear regression and logistic regression to ultimately predict student’s performance on examinations based on multiple socioeconomic and demographic statistics such as gender, race/ethnicity, parental level of education, and provided test scores. The dataset, provided by Royce Kimmons, has such data for 1000 students and includes both features—gender, race/ethnicity, parental level of education, test preparation course, lunch—and dependent variables/targets—scores in reading, writing, and mathematics. Unfortunately, not all essential and useful pieces of data, such as hours spent studying, are provided by this dataset, so I can only perform analysis of the effects these features have on performance. The model will be fit with a OLS Linear Regression model (because we are interested in the factors that influence performance). However, prediction of student performance will also be performed through creation of various types of graphs and other models such as Linear Least Squares Regression and Logistic Regression. Through analysis of these models, we will also be able to determine the effectiveness of test preparation courses and the effect food/lunch has on performance. For each model, I will use different hyperparameters and preprocessing techniques to optimize performance.

Procedure:

1. Source data from Kaggle: [Student Performance on Exams](https://www.kaggle.com/spscientist/students-performance-in-exams/tasks), data on various factors that can potentially affect student’s educational performance. The data is provided by Royce Kimmons, and the project is created on Kaggle by Jakki Seshaapanpu
2. Import Libraries:
   1. Basic imports: numpy, pandas
   2. Plotting imports: matplotlib, plotly, seaborn,
   3. Preprocessing imports: sklearn
   4. ML model imports: lightgym, xgboost, bayes\_opt
   5. Feature importance imports: shap
3. Preprocess data
   1. Separate data into two different categories: target (math score, reading score, writing score, and average score), and features (gender, race/ethnicity, parental level of education, lunch, and test preparation course)
   2. Read in data and perform preliminary analysis on it using methods/functions such as describe and info
   3. Create new column in data called average score, which is the average of math score, reading score, and writing score. Having one average value to use will be helpful for certain models
   4. Checking for missing values
   5. Determine which features are categorical and which are numerical
   6. Understand data by creating various that display data
      1. Pie charts for each category to understand size of categories in each feature
      2. Bar graphs for the target variables to demonstrate their continuous nature
      3. Other graphs I will create: heatmap, catplot, pairplot, and probably a few others
   7. Clean data
      1. Removing outliers (none removed due to dataset having only 1000 points)
      2. Label encoding for some categorical features
   8. Split the dataset into training/testing/validation sets
4. Numerous models will be used in this project to predict student performance on examinations, starting with support vector machine and linear regression
   1. One will use MSE, MAE, and RMSE from sklearn
   2. Above models will be created using above model imports
   3. Model metrics will be written, and scores for MSE, MAE, RMSE, and R-squared will be given for both the training set and test set
   4. Model performance will also be analyzed through creation of numerous graphs
5. A Logistic Regression model will also be created to perform the similar task as above
   1. Use k-fold cross-validation to train and test the model, as well as tune the hyperparameter for the regularization term
6. Evaluate the accuracy, precision, recall and F1 score of the classifications for each preprocessing technique
7. A Linear Least Squares Regression (type of linear regression model) will also be created to understand the strength of the relationship between various factors provided and scores
   1. Residuals—the deviation of the fitted values from the actual values—will be created for this and used to determine if it is good to predict the reading scores from one’s writing scores
   2. A coefficient of determination (R^2 value) and a plot of the two features graphed will also be included
8. Conduct a Principle Component Analysis (most likely the sklearn Neighborhood Component Analysis) to prepare for k-NN on preprocessed dataset using most effective preprocessing technique
9. Perform k-NN classification using sklearn.neighbors and plot the results
10. Compare results obtained from various models and make general conclusions about which factors are interrelated to each other and to performance on examinations.

**III. Timeline:**

By the end of Week 1 (March 18th-March 25th): Write the project proposal (this document), import the dataset, import required libraries, complete data preprocessing (including cleaning data, making graphs to understand data, label encoding, splitting the dataset, and more) and exploratory data analysis. Total: 5-7 hours

By the end of Week 2 (March 26th-April 4th): Start the creation of numerous models and complete work on support vector machine and OLS Linear Regression model. With completion of these models comes all corresponding pattern and model performance analysis. Also, progress report 1 will be completed. Total: 6-8 hours

By the end of Week 3 (April 5th-April 11th): Start and complete Linear Least Squares Regression Model. Also logistic regression model will be completed and accuracy, precision, recall and F1 score of the classifications will be evaluated. With completion of these models again comes all corresponding pattern and model performance analysis. Also, progress report 2 will be completed. Total: 6-8 hours

By the end of Week 4 (April 12th-April 20th): Principal Component Analysis (PCA), k-NN classification, and overall final plots/visualizations will be completed. Lightning talk will be prepared and completed as well. Total: 5-7 hours

By the end of Week 5 (April 21st-April 25th): Compare model performances, analyze results and relevance to student performance in real world, and add any final touch-ups to the coding section of the notebook (improve upon spacing, comments, descriptions, and layout of code and notebook). Also to be completed is the final report for the project. Total: 6-8 hours

**IV: References:**

1. <https://www.kaggle.com/spscientist/students-performance-in-exams/tasks>
2. <http://roycekimmons.com/tools/generated_data/exams>
3. <https://www.kaggle.com/ritikpnayak/regression-model-1-the-best-fit-line-with-r2-0-9>
4. <https://www.kaggle.com/josephchan524/studentperformanceregressor-rmse-12-26-r2-0-26>
5. <https://www.globalcitizen.org/en/content/9-facts-about-education/>
6. <https://www.habitatbroward.org/benefits-of-education/#:~:text=It%20helps%20people%20become%20better,rights%2C%20laws%2C%20and%20regulations>.
7. <https://files.eric.ed.gov/fulltext/EJ1151836.pdf>
8. <https://peer.asee.org/regression-models-for-predicting-student-academic-performance-in-an-engineering-dynamics-course.pdf>
9. <https://www.kdnuggets.com/2020/01/beginners-guide-nearest-neighbors-r.html>
10. <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.NeighborhoodComponentsAnalysis.html>
11. <https://towardsdatascience.com/logistic-regression-using-python-sklearn-numpy-mnist-handwriting-recognition-matplotlib-a6b31e2b166a>