An Analysis on Student Performance Factors

Huaxing Zeng
DSI, Brown University
GitHub

1. Introduction

Purpose

Accurately identifying the factors that influence student performance is essential in improving educational strategies, shaping effective learning environments, and ultimately improving academic success. Analyzing various attributes related to student habits, parental involvement, and resource availability provides a comprehensive understanding of the elements that contribute to enhanced learning outcomes. Employing machine learning methods to determine which factors play the most significant roles in academic achievement can allow educators and policy makers to tailor interventions that address specific needs and challenges faced by students [1].

Student Performance Factors Dataset

The dataset utilized in this study is a synthetic dataset generated to simulate realistic scenarios especially for analysis, and it presents a holistic overview of multiple aspects that potentially impact student examination results. In this dataset, the target variable is the final exam score. The features include study patterns, attendance rates, levels of parental engagement, access to learning materials, participation in extracurricular activities, sleep habits, previous scores, motivational levels, and family income status. Other features, such as internet accessibility, frequency of tutoring sessions, teacher quality, and school characteristics, are also represented. Each student record comprises details on these attributes and a final exam score. There is a total of 6607 datapoints and 20 feature columns, and 229 out of 6607 data points are missing values (78 in Teacher_Quality, 90 in Parental_Education_Level, 67 in Distance_from_Home, where some datapoints were missing in more than one feature).

Previous Related Research

Previous research in educational analytics has found that university students with exceptionally high life satisfaction performed better academically than those with average or low satisfaction, demonstrating greater engagement, higher self-efficacy, more approach-oriented goals, and lower stress, as well as earning higher GPAs [2]. Also, other research projects have studied how various academic and social motivational factors—such as beliefs about ability, efficacy, control, values, and goals—interact to shape students' academic outcomes. Additionally, it highlights the role of contextual influences, including teachers' instructional methods and relationships with students, in fostering motivation and improving academic performance [3].

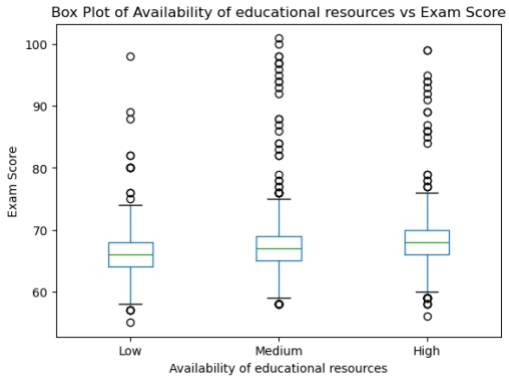
2. Feature Analysis

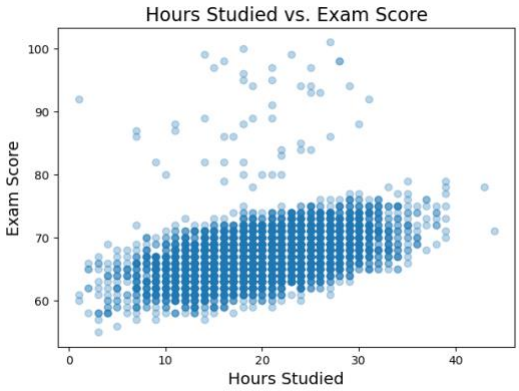
EDA Graphs

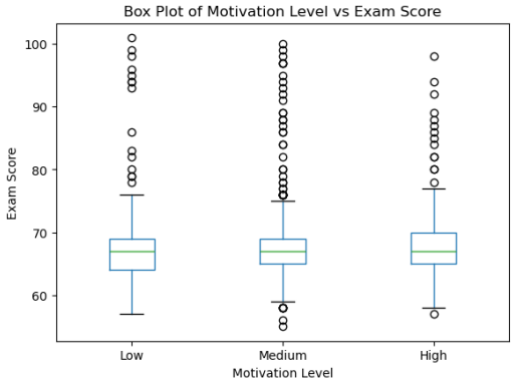
To gain a deeper understanding of the dataset, descriptive statistics and exploratory visualizations were performed. Histograms and distribution plots were generated to visualize the range and frequency of key continuous features, offering a clearer picture of their underlying distributions.

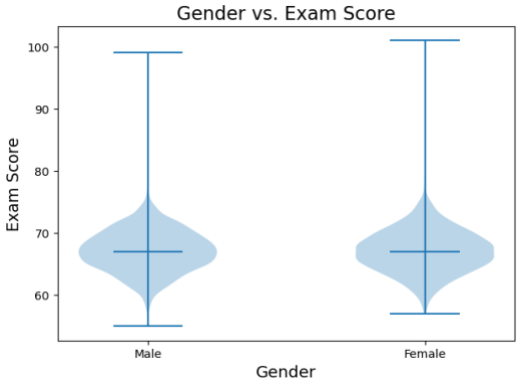
Several plots were produced to explore relationships between predictors and the target variable (exam score). For example, the box plot of educational resource availability versus exam score suggests that students with higher access to resources tend to achieve better results, though a range of scores is still present at all resource levels. The violin plot comparing gender to exam score visualizes the score distributions for male and female students, showing that the mean and median scores are quite similar for both groups, indicating no significant difference in exam performance between male and female students. The scatter plot of hours studied versus exam score reveals a generally positive relationship, indicating that increased study time often correlates with improved outcomes. Lastly, examining the box plot of motivation level versus exam score shows that higher motivation levels align with higher median scores, again underscoring the importance of such qualitative factors.

These initial findings serve as a useful guide for feature selection and model development, helping identify which variables may hold the greatest predictive power. By combining summary statistics with a variety of plots, we can gain insights regarding how both quantitative and qualitative factors contribute to student academic performance.

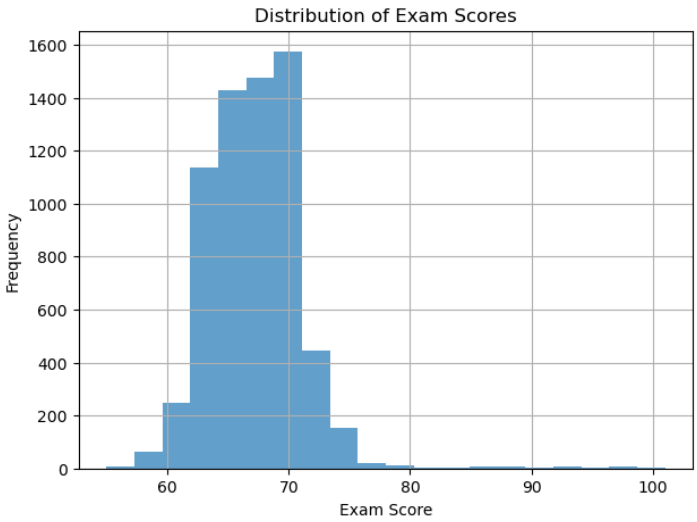








The Target Variable
The target variable in this dataset is the final exam score, a continuous numerical value representing each student's performance. The distribution plot reveals that the majority of the scores cluster within a relatively narrow band below the mid-70s, indicating that reaching or surpassing the 73-point threshold—commonly considered a passing 'C' grade—is not frequently achieved. This distribution suggests that many students may be struggling to attain what might be considered a satisfactory level of performance. The skewed nature of the scores implies that, while a small number of students excel well above this benchmark, the overall tendency leans toward lower exam results. Understanding this distribution is critical, as it provides a useful threshold as we move along the machine learning steps into splitting the data.



3. Methods

Data Preprocessing
The dataset initially contained a mix of ordinal, nominal, and numerical variables, with some missing values present in the categorical and ordinal features. To handle this, missing entries in categorical data were replaced with a new “NA” category, and specifically, missing entries in ordinal data were ordered to be below the lowest rank. Ordinal features such as Parental_Involvement and Motivation_Level were encoded using an OrdinalEncoder with defined category orders. Nominal features like Gender and School_Type were transformed via OneHotEncoder to avoid imposing any artificial hierarchy. Continuous variables, including Hours_Studied and Previous_Scores, were standardized using StandardScalar to ensure all features contributed more equitably to the models. After preprocessing, categorical data expanded due to the mechanisms of the OneHotEncoder, resulting in 24 total features.

Data Splitting Strategy
To evaluate and tune the machine learning models, the dataset was split into training, validation, and testing sets. Given the continuous and skewed nature of the target variable exam score as discussed before, a stratification approach was employed using a binary threshold at 73 points—

a passing C grade—creating two strata: students with scores below 73 and those at or above. Stratified K-Fold cross-validation with 5 folds was applied to the training and validation data.

ML Pipeline
The pipeline was built using scikit-learn's Pipeline and ColumnTransformer. The pipeline first applied all preprocessing steps, then passed the transformed data to the chosen model.

Model Selection and Hyperparameter Tuning
Six models were used here:

Model	Parameters
Lasso	Alpha: [0.01, 0.1, 1, 10, 100]
Ridge	Alpha: [0.01, 0.1, 1, 10, 100]
Elastic Net	Alpha: [0.01, 0.1, 1, 10, 100] L1_ratio': [0.1, 0.5, 0.7, 0.9]
Random Forest	max_depth: [none, 10, 20, 30] min_samples_split: [2, 5]
SVR	C: [0.1, 1, 10, 100] kernel: ['linear', 'rbf'] Epsilon: [0.01, 0.1, 1, 10]
XGBoost	learning_rate: [0.01, 0.1, 0.2] max_depth: [1, 3, 5, 7] subsample: [0.6, 0.8, 1.0]

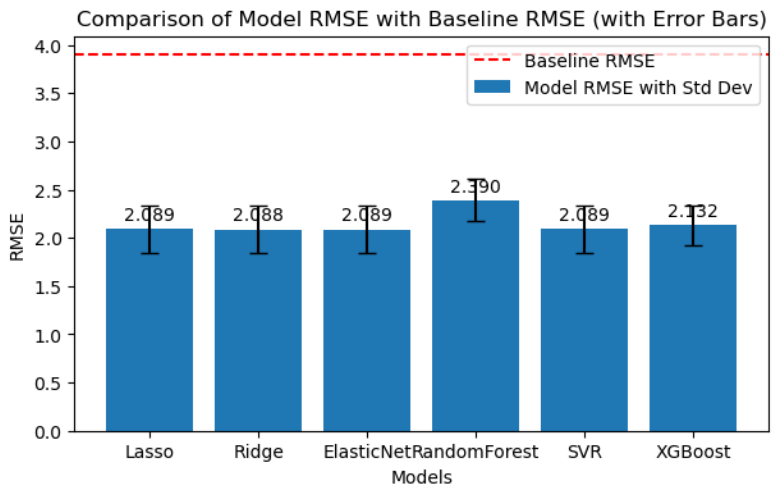
Each model underwent hyperparameter tuning using GridSearchCV. The best parameters for each model are highlighted.

Evaluation Metric and Uncertainty Estimation
The primary evaluation metric was the Root Mean Squared Error (RMSE), where minimizing RMSE meant producing more reliable predictions. To avoid the influence of one single random state being too strong, 5 random states were used during model fitting.

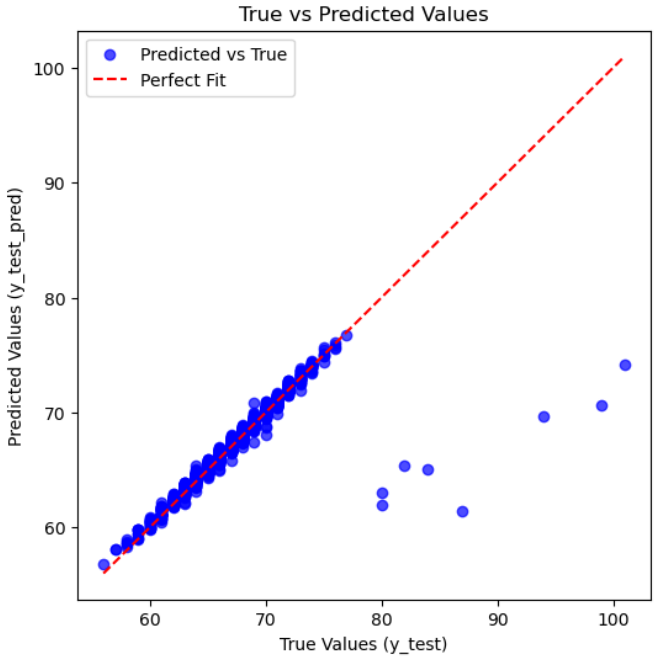
4. Results

Here are the resulting test scores and the corresponding standard deviation of each model after running the ML pipeline:

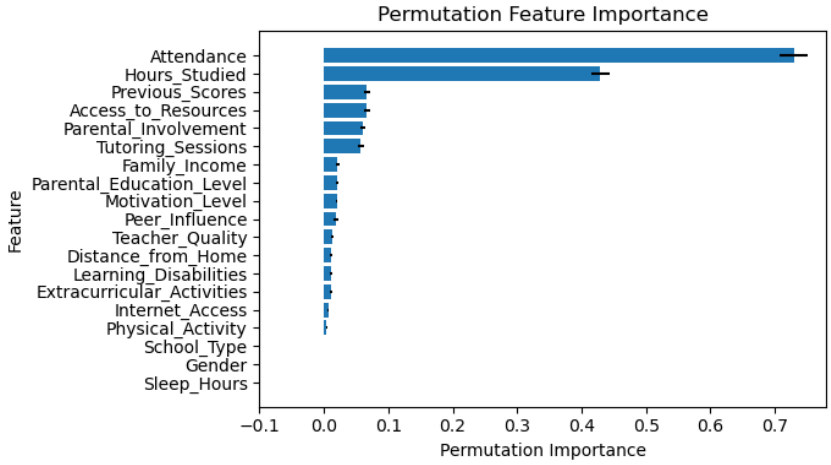
Model	Mean Test Score	Standard Deviation	Standard Deviations Above Baseline
Lasso	2.089	0.246294	7.341
Ridge	2.088	0.245694	7.363
Elastic Net	2.089	0.246657	7.330
Random Forest	2.394	0.214074	7.021
SVR	2.089	0.242938	7.442
XGBoost	2.132	0.206515	8.547

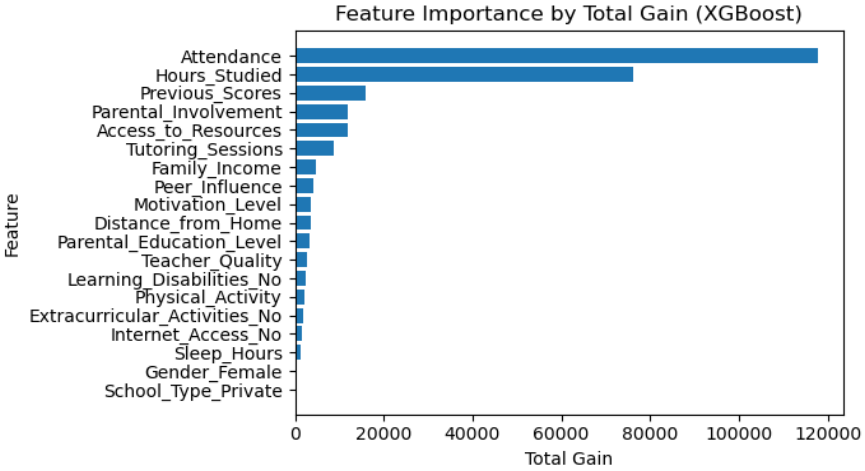


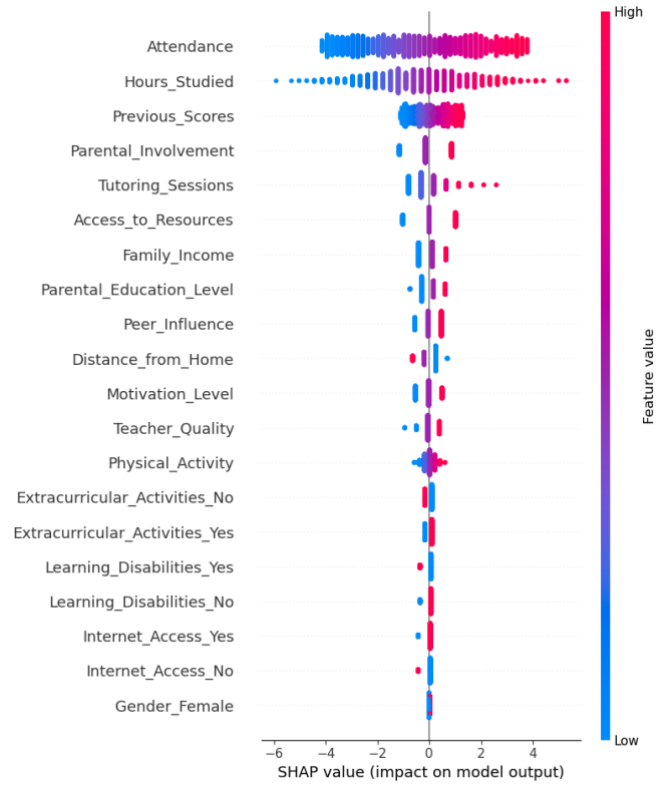
The baseline RMSE here is 3.897, calculated using a baseline model that predicts everything as the mean of the target variables. All models showed improvement comparing to the baseline model as they show smaller RMSE compared with the baseline. The best model is Ridge with alpha = 10, with an RMSE of approximately 2.088.



The plot of predicted values vs. true values using the best model (Ridge with alpha =10) shows that the model does a relatively good job as most of the points line around the diagonal line (where y_pred = y_true), with only a few outliers.

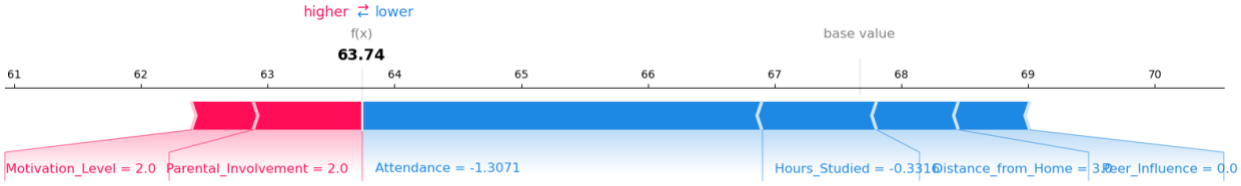

Permutation Feature Importance
Attendance
Hours_Studied
Previous_Scores
Access_to_Resources
Parental_Involvement
Tutoring_Sessions
Family_Income
Parental_Education_Level
Motivation_Level
Peer_Influence
Teacher_Quality
Distance_from_Home
Learning_Disabilities
Extracurricular_Activities
Internet_Access
Physical_Activity
School_Type
Gender
Sleep_Hours
Feature
−0.1 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
Permutation Importance


Feature Importance by Total Gain (XGBoost)
Attendance
Hours_Studied
Previous_Scores
Parental_Involvement
Access_to_Resources
Tutoring_Sessions
Family_Income
Peer_Influence
Motivation_Level
Distance_from_Home
Parental_Education_Level
Teacher_Quality
Learning_Disabilities_No
Physical_Activity
Extracurricular_Activities_No
Internet_Access_No
Sleep_Hours
Gender_Female
School_Type_Private
Feature
0 20000 40000 60000 80000 100000 120000
Total Gain

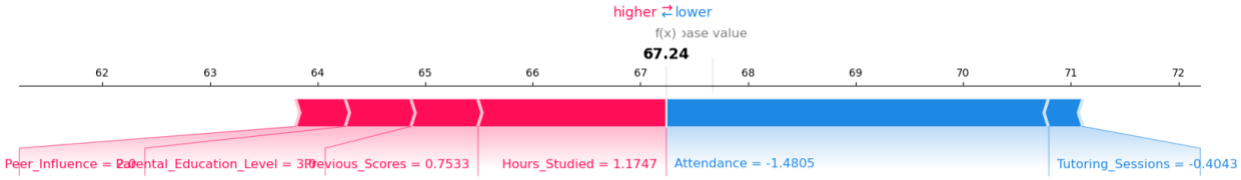

Attendance
Hours_Studied
Previous_Scores
Parental_Involvement
Tutoring_Sessions
Access_to_Resources
Family_Income
Parental_Education_Level
Peer_Influence
Distance_from_Home
Motivation_Level
Teacher_Quality
Physical_Activity
Extracurricular_Activities_No
Extracurricular_Activities_Yes
Learning_Disabilities_Yes
Learning_Disabilities_No
Internet_Access_Yes
Internet_Access_No
Gender_Female
−6 −4 −2 0 2 4
SHAP value (impact on model output)
High
Feature value
Low

Permutation importance, XGBoost total gain, and SHAP values are used to determine which feature contributes the most to the prediction. Across all three importance measures—permutation importance, XGBoost total gain, and SHAP values—Attendance**,** Hours_Studied, and Previous_Scores consistently emerged as the top predictors of exam performance. Variables like Parental_Involvement and Access_to_Resources also played meaningful roles. This strong alignment across multiple methods suggests that improving class attendance, increasing study time, and building upon previous academic successes are key levers for enhancing exam scores.

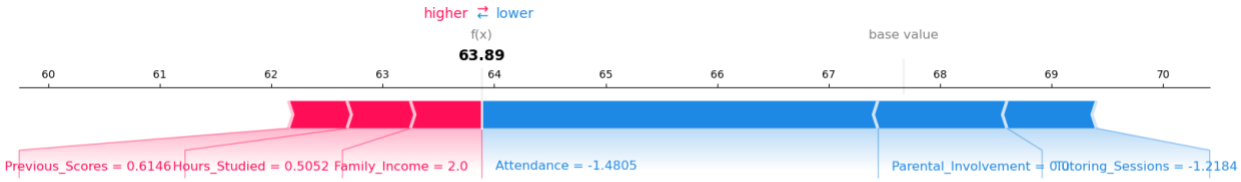
Index 0



Index 100



Index 200



The SHAP force plots for students at indices 0, 100, and 200 show how individual features push their predicted exam scores above or below the model's baseline. Red bars indicate features that increase the prediction, while blue bars represent factors pulling it down. Common influential drivers, such as attendance and previous scores, typically shift predictions upward, whereas lower motivation or parental involvement can reduce the final estimate. This individualized perspective allows educators or policymakers to pinpoint the most critical features for each student and tailor supportive interventions accordingly.

5. Outlook

Moving forward, the predictive power of the model could be enhanced through both technical and contextual refinements. Feature engineering, such as incorporating interaction terms or non-linear transformations, may help capture subtler patterns in student performance. Collecting additional data would also further strengthen the foundation of the model. For instance, richer behavioral data including test-taking anxiety, study habits, and detailed sleep patterns and more nuanced socioeconomic information such as granular family income levels or resource availability could further our understanding of external influences on student performance. Moreover, integrating indicators of instructor quality or the classroom environment could highlight previously overlooked educational dynamics. Taken together, these steps would not only improve model accuracy but also enable more targeted, effective interventions for student success.

6. References

[1] Hearn, J. C. (2006). *Student success: What research suggests for policy and practice.* National Postsecondary Education Cooperative, U.S. Department of Education. https://nces.ed.gov/npec/pdf/synth_hearn.pdf

[2] Antaramian, S. (2017). The importance of very high life satisfaction for students' academic success. *Cogent Education, 4*(1), 1307622. https://doi.org/10.1080/2331186X.2017.1307622

[3] Wentzel, K. R., & Wigfield, A. (1998). Academic and social motivational influences on students' academic performance. *Educational Psychology Review*, *10*(2), 155–175.

7. GitHub

https://github.com/Huaxing-Zeng/DATA1030_final_project