Analyzing Student Academic Performance Factors: Single Linear, Multiple Linear, and Logistic Regression Techniques

Hayley Hawkins

November 23, 2024

I. Introduction

This dataset describes the various factors that affect student performance in exams. In includes information on attendance, study habits, parental involvement and influence, as well as other factors that might influence academic success. This dataset was downloaded from Kaggle, which is an online data science platform with thousands of ready-to-use datasets.

The data set for this study is titled "Student Performance Factors" and it contains 6,607 observations and 20 variables. The objective of my study is to analyze how some of these features related to student life impact performance on an exam. I am interested in comparing how certain factors improve student scores. As a student myself, it will be interesting to gain some insight on the study habits that help students get better exam scores. There are three main hypotheses for my study, and I will be using three different regression techniques to examine these hypotheses.

A. Hypotheses

1. **The number of study hours per week is positively correlated with a student's final exam score.**

For my first hypothesis, I will be using simple linear regression to examine if a linear relationship exists between study hours and final exam scores. I will only be utilizing two variables: Hours Studied (independent variable) and Exam Score (dependent variable). I believe that study hours is one of the biggest indicators of success on an exam, so I am expecting there to be a strong positive correlation.

2. **A combination of parental education level, family income, and parental involvement is predictive of a student's overall performance.**

To examine my second hypothesis, I will utilize multiple linear regression with multiple categorical predictors (Parental Education Level, Family Income, Parental Involvement) to model the final exam score (Exam Score). I am interested to see how the characteristics of a student's parents impact their performance on an exam.

3. **Students with higher attendance are more likely to pass the final exam.**

My third hypothesis will examine if attendance is related to passing the final exam. I will use logistic regression analysis to model the relationship, using a binary outcome (pass/fail) based on exam scores above 70 percent. The two variables in this hypothesis are Attendance and Exam Score. Similar to study hours, I expect a higher attendance makes a bigger difference in a student passing an exam than other variables, so I am particularly interested to see the relationship between these variables.

B. Data Processing

For this study, a number of different data processing techniques were implemented to clean the dataset. The data was subsetted to include only the variables of interest, and all nulls and duplicates were dropped. Variable names were simplified, and the categorical data was checked for input or typing errors. Additionally, the data was checked for outliers. My dependent variable, Exam Score, had numerous scores above 100, all of which were removed.

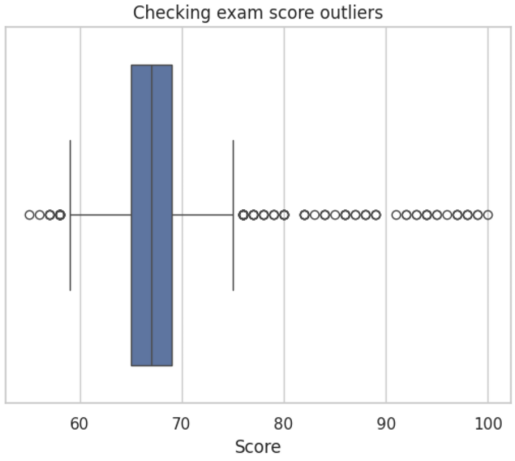

Fig. 1. Boxplot of Exam Outliers

C. Exploratory Data Analysis

To explore my selected variables, I performed a number of different analyses, including finding summary statistics, univariate analysis, bivariate analysis, multivariate analysis, and generating data visualizations.

TABLE I
Descriptive Statistics

Variable	Mean	Min	Max	Stdev
Exam Score	67.23	55	100	3.91
Attendance	75.66	60.0	100.0	3.91
Hours Studied	67.23	1	44	3.91

I generated distribution graphs for each of my quantitative variables and frequency counts for my categorical variables.

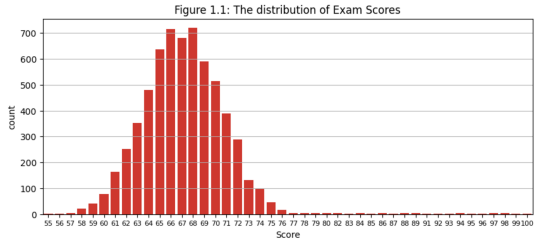

Fig. 2. Distribution of Exam Scores

I also completed bivariate analysis, which compared the relationships between variables such as Attendance and Exam Score, Study Hours and Exam Score, and Attendance and Study Hours. For these correlation studies, I created scatter plots with a line of best fit. To explore the relationships between each of my categorical variables, I created box plots.

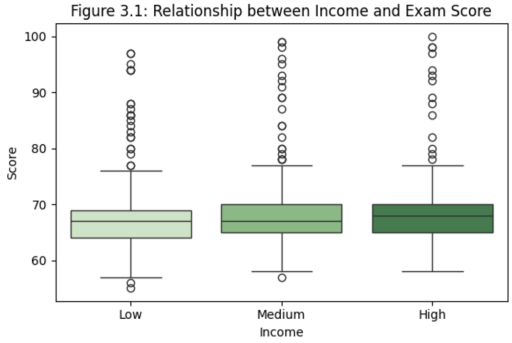

Fig. 3. Boxplot of Income and Exam Score

Multivariate analysis was also completed in the form of heatmaps. I compared the relationships between Study Hours, Attendance, and Score, and also the relationships between Parental Income, Attendance, Parental Education, and Exam Score.

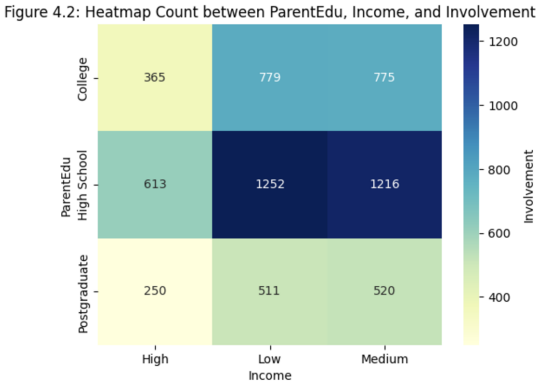

Fig. 4. Heatmap of Relationship between Parental Income, Attendance, Parental Education, and Exam Score

II. Methodology

A. Simple Linear Regression

Simple Linear Regression is a regression technique that examines the relationships between two continuous variables. Simple Linear Regression can be described as a straight line through a set of points that minimizes the vertical distances between the data points and the fitted line. In other words, it minimizes the sum of squared residuals. The equation for simple linear regression is the following:

$$y = b_0 + b_1 * x + \epsilon \quad (1)$$

I have selected Simple Linear Regression to test my Hypothesis 1, which states: "The number of study hours per week is positively correlated with a student's final exam score." I have chosen Simple Linear Regression because I only have two variables in this hypothesis, and I am curious to know the correlation between the two. Simple Linear Regression will clearly and efficiently show me the relationship between the number of hours a student spends studying and their resulting exam score. In this study, the Hours Studied variable will be my predictor/independent variable. My second variable will be Exam Score, which is my response/dependent variable.

B. Multiple Linear Regression

Multiple Linear Regression is another regression technique, similar to Simple Linear Regression, but it looks at the relationships between two or more variables. Specifically, Multiple Linear Regression looks at multiple explanatory/independent variables, and one dependent variable. The equation for Multiple Linear Regression is:

$$Y = b_0 + b_1 * X_1 + b_2 * X_2 + \cdots + b_n * X_n + e \quad (2)$$

I will use Multiple Linear Regression to test Hypothesis 2, which states: "A combination of parental education level, family income, and parental involvement is predictive of a student's overall performance." This hypothesis addresses more than two variables, therefore Simple Linear Regression would not be appropriate. By using Multiple Linear Regression, I will be able to find the relationship between all four variables simultaneously. For this model, I will have Parental Education Level, Family Income, and Parental Involvement be my three explanatory variables. Exam Score will be my dependent variable.

To evaluate my model, I will check certain assumptions, which are described below: 1. Linearity: The relationship between predictors and outcome variable should be linear. 2. Independence: The residuals (prediction errors) should be independent of each other. 3. Homoscedasticity: The variance of residuals should be constant across all levels of predictors. 4. Normality: The residuals are normally distributed. 5. Multicollinearity: Ensuring that predictors are not perfectly correlated with each other.

C. Logistic Regression

Logistic Regression is a statistical regression technique that measures the relationship between a categorical dependent variable and one or more independent variables using a logistic function. It predicts the probability of an event happening and is used specifically for binary or multi-class problems.

For Hypothesis 3, I am interested in seeing how student attendance affects the likelihood of passing an exam. Hypothesis 3 states: "Students with higher attendance are more likely to pass the final exam." Since passing an exam is a binary classifier (pass/fail), Logistic Regression is appropriate for this hypothesis.

In this problem, I will have to adjust my Exam Score variable to reflect a binary pass/fail criterion. To do this, I will create a condition in which scores greater than or equal to 70 are defined as "pass", and scores less than 70 are defined as "fail".

Once these adjustments have been made, the binary Exam Score variable will be my categorical dependent variable. Attendance will be my independent variable.

When I found initial evaluation metrics, I noticed that my recall was quite low. In order to mitigate this issue, I implemented SMOTE (Synthetic Minority Over-sample Technique). SMOTE synthesizes new minority instances from existing ones, helping to handle imbalanced data sets.

III. Results

A. Simple Linear Regression

TABLE II
SIMPLE LINEAR REGRESSION METRICS

Intercept	Slope	MSE	R-Squared
61.5014	0.2871	12.2490	0.1972

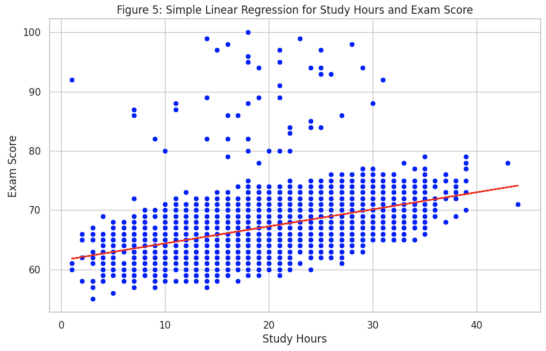

Fig. 5. Simple Linear Regression for Study Hours and Exam Score

With an intercept of approximately 61.50, this model predicts that with zero hours of study, students are expected to achieve an average exam score close to this baseline. The slope of 0.287 suggests that for each additional hour a student spends studying, there is a predicted increase of about 0.29 points in the final exam score.

The slope also indicates a positive but relatively weak relationship between study hours and exam scores. This implies that while studying more does have a positive impact on performance, there are likely other factors that may affect exam performance more.

The R squared value of 0.197 demonstrates that study hours explain around 19.7% of the variance in exam scores. Therefore, while study hours have some predictive value, other factors will likely contribute to student performance as well. The Mean Squared Error value of 12.25 reflects the average squared deviation between the actual score and predicted score, which points to a moderate level of error in the predictions.

B. Multiple Linear Regression

TABLE III
MULTIPLE LINEAR REGRESSION METRICS

MSE	R-Squared
13.02412	0.0648

In this multiple linear regression analysis, I explored whether parental education level, family income, and parental involvement could predict students' final exam scores.

My model's Mean Square Error of 13.02 suggests a moderate amount of prediction error while the R-squared value of 0.065 indicates that only about 6.5% of the variance in examples scores is explained by my three independent variables. This means that these factors are probably not strong predictors of exam performance in this dataset.

After completing my model, I went on to check my model assumptions. First, my linearity check showed a relatively random scatter, suggesting that the assumption is met.

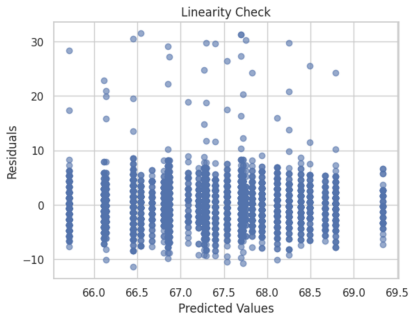

Fig. 6. Linearity Check

The next assumption, homoscedasticity, checks that the variance of residuals is constant across all levels of predictors. Based on the random scatter on my scatterplot, it seems like my model passes this assumption.

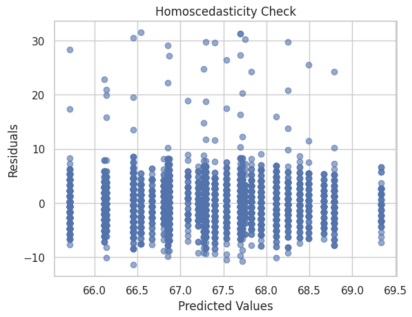

Fig. 7. Homoscedasticity Check

The Durbin-Watson statistic, which is used to check independence, is approximately 2.0, which suggests minimal autocorrelation in the residuals.

Normality is checked by plotting a Q-Q Plot to see if the residuals are normally distributed. My model shows a normal distribution up until right after the 2nd quantile.

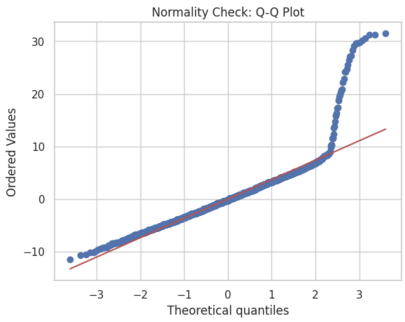

Fig. 8. Q-Q Plot

Finally, the Variance Inflation Factor (VIF) values are all below 1.01, indicating a low level of multicollinearity among the predictor variables. This suggests that the predictors are relatively independent of each other, which supports the model's stability and interpretability.

In summary, while parental education, family income, and parental involvement show low multicollinearity and minimal autocorrelation, they are relatively weak predictors of final exam scores in this model, explaining only a small fraction of the variance in student performance.

C. Logistic Regression

TABLE IV
PERFORMANCE METRICS

Data	Accuracy	Precision	Recall	F1	AUC
No SMOTE	79.00%	60.70%	48.90%	54.17%	69.07%
SMOTE	75.66%	50.39%	83.87%	62.95%	85.05%

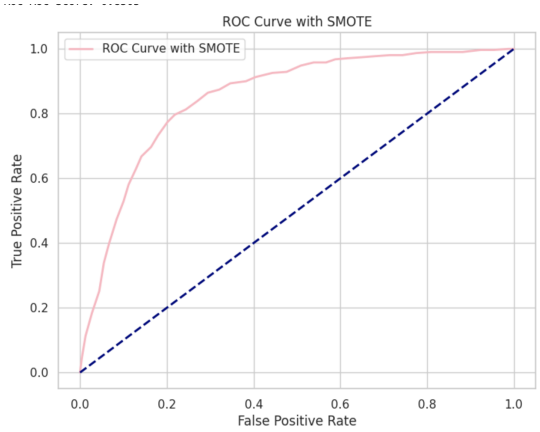

Fig. 9. ROC Curve with SMOTE

In my logistic regression analysis, I explored whether students with higher attendance are more likely to pass a final exam. I found initial evaluation metrics, but noticed that my recall was quite low. In order to mitigate this issue, I implemented SMOTE (Synthetic Minority Over-sample Technique). SMOTE synthesizes new minority instances from existing ones, helping to handle imbalanced data sets. Below is an interpretation of my final evaluation metrics:

My final accuracy is slightly lower compared to the previous accuracy (0.7900). This decrease is expected because with the introduction of SMOTE, the model now better identifies the minority class, which can slightly reduce overall accuracy.

Precision also dropped compared to the original model (0.6070). This indicates that the model now produces more false positives, which is because SMOTE generates synthetic data to balance the classes (pass/fail). The model is overestimating the likelihood of passing for some students. This indicates that there may be other influential factors.

There was a big improvement from the previous recall (0.4890). This means the model is now much better at identifying students who actually pass. This supports the idea that attendance plays a role in passing rates.

The F1 score improved compared to the previous value (0.5417), reflecting a better balance between precision and recall. This indicates the model is now more effective in handling the pass/fail classification.

The jump in ROC AUC score (from 69.1% to 85.0%) indicates the model is much better at ranking students by their likelihood of passing. This reinforces the idea that higher attendance has a strong correlation with passing.

IV. CONCLUSION

This study analyzed a number of different factors that influence student performance on a final exam. In this project, I utilized three different regression techniques to provide insights into the factors that influence exam success.

Using simple linear regression, I found that study hours are positively correlated with exam scores, with an R squared value of 0.197. A multiple linear regression model demonstrated that parental education, family income, and parental

involvement are somewhat weaker predictors of exam scores. My findings from this study suggest that they do not strongly influence student performance in this particular dataset. When using logistic regression, I found that attendance significantly impacts the likelihood of passing an exam. I incorporated SMOTE to improve the model's recall and ROC AUC score, effectively addressing class imbalance in my binary classification study. My model ended with an accuracy of 75.66%, and I found that higher attendance was strongly associated with passing rates.

There are many ways in which the findings from this study could be applied to educational practices and future research. My findings reveal that factors such as attendance and studying have large impacts on the likelihood of student success. While factors related to parent or family situations were less influential, it is still important that these factors be considered when trying to help students succeed. Furthermore, this study demonstrates how applying regression techniques can help educational institutions make data-driven decisions.

REFERENCES

STAT650 Lecture17 Descriptive Statistics
STAT650 Lecture08 Overview EDA by Python
STAT650-Lecture27-Regression01-Simple Linear
STAT650-Lecture28-Regression02-Multiple Linear
STAT650-Lecture33-Regression07-Logistic Regression

- Link to Google Colab Code
- Link to Dataset