**Project 1: Factors that Contribute to Student Performance on Exams**

Christina Vosnak

Data Science, Bellevue University

DSC 680

Professor Iranitalab

April 6th, 2025

**Business Problem**

Student success is of the utmost importance in any educational institution. Students are evaluated by a multitude of factors including projects, overall grades, and most conventionally, exam scores. This project aims to identify the most significant factors affecting exam performance and hopes to provide insights to help educators implement targeted change to enhance student learning outcomes.

**Background/History**

Multiple stakeholders across the educational landscape are deeply invested in understanding what drives student performance. School administrators seek this insight to meet performance benchmarks, parents are eager to identify the factors that can support their children's academic growth and long-term success and students themselves are increasingly interested in recognizing the habits, environments, and support systems that can help them achieve their full potential.

While environmental, behavioral, and socioeconomic factors have been analyzed individually in the past, taking a holistic approach to looking at these factors through a modern analytical lens may help uncover key insights and correlations that influence exam scores and overall student performance.

**Data Explanation (Data Prep/Data Dictionary/etc)**

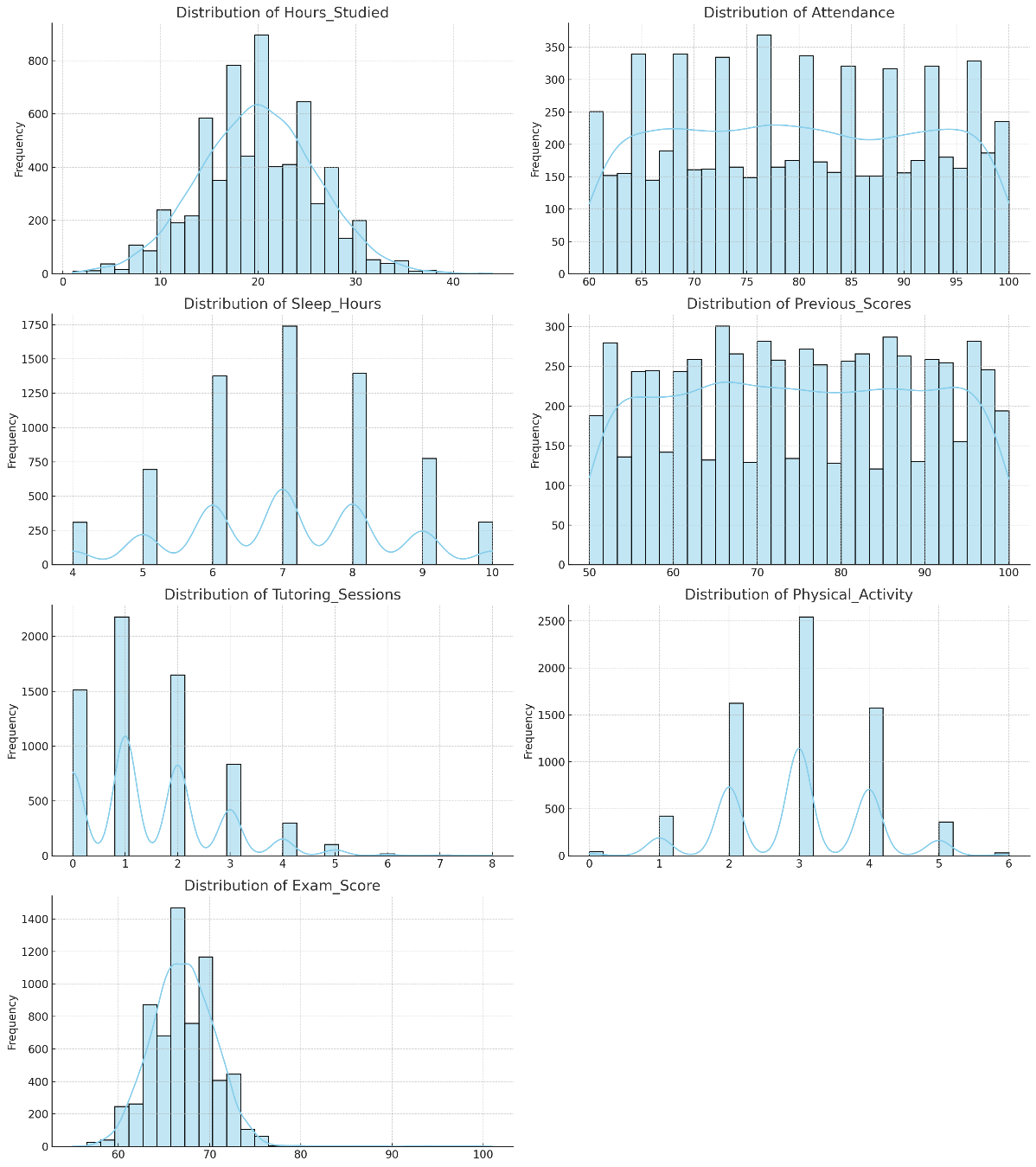
To begin looking at this topic, I sourced a preliminary dataset from Kaggle which contains 20 performance related features over 6000+ student records. The data contained inside this dataset is synthetic, however, using this information as a baseline to build a predictive model can be useful in the future to looking at actual data from the department of education or within a specific school system.

From this data set, I conducted an exploratory data analysis and cleaned the dataset to prepare it for modeling. I analyzed each categorical and quantitative feature to identify if any missing values or outliers existed in the data set and fixed these components accordingly. I found that three factors: Teacher Quality, Parental Education Level, and Distance From Home all were columns that had a significant amount of missing values so rather than trying to fill I the missing values, I removed the columns from the dataset entirely to focus on accurate results.. I then created box plots for all numerical columns to identify outliers. The chart below shows that Hours Studied, Tutoring Sessions and Exam Scores, have a significant amount of outliers.

A group of blue boxes with white text

AI-generated content may be incorrect.

However, when you look at the distribution charts, the so-called outliers in the data are really just representations of skewness. Therefore, I did not remove any of the recommended outliers, and kept them in the dataset for modeling purposes.



Finally, the data has quite a few categorical variables. To include them in a model, I used one-hot encoding to transform the information into dummy variables and then created a correlation heatmap to understand how each variable may relate to one another, especially the future target variable: Exam Scores.

**Methods**

A screenshot of a graph

AI-generated content may be incorrect.For this analysis, I will first be looking at correlations within each variable. I want to show how each variable correlates to one another using a correlation matrix to understand if there are any duplicate variables within the dataset and to get a preliminary understanding of which factors may most contribute to exam scores. From the preliminary chart below, we can see that attendance and hours studied factor mostly into exam scores.

Additionally, I plan to build 3 predictive models to analyze the data. I will use a linear regression model to look at linear relationships between the variables as well as a random forest model to evaluate relationships that are non-linear. I will also use a Gradient Boosting model to capture any complex non-linear relationship that might not be captured with a Random Forest. I am including it to compare results with the Random Forest regressor specifically. I will evaluate the accuracy of these models using evaluation metrics including MAE, R^2, and RMSE. MAE will provide a clear and interpretable measure of average error in the same units as the target, helping me understand typical prediction accuracy. R^2 score will explain the variations within the target variable. RMSE combines the benefits of both, offering an interpretable metric that still penalizes larger errors, making it a balanced and widely used evaluation tool that may be more sensitive to the outliers still included in the dataset. Finally, I will also include feature importance within each model to understand which factors most contribute to exam scores to provide recommendations on how to improve scores overall. I will also create a bar chart of the top 3 features of the best model to understand their importance and how strongly each feature contributes to exam scores.

**Analysis**

After building three predictive models, the information shows that a linear regression performed best for this dataset with lowest RMSE and MAE, and highest R^2 score.

A screenshot of a graph

AI-generated content may be incorrect.

The top 3 features for each model are listed below.

A screenshot of a computer

AI-generated content may be incorrect.

This shows that in the Linear Regression model, parental involvement regardless of level, contribute most to exam scores while looking at the other two models, attendance, hours studied and previous scores contribute most, aligning more with the results of the correlation matrix.

**Conclusion**

A red and white bar graph

AI-generated content may be incorrect.Overall, the linear regression model, which performed best, strongly suggests that parental involvement contributes most to student’s exam scores. The chart below shows the importance coefficients of these factors in relation to exam score.

Medium Involvement contributes most with a correlation coefficient of almost 3.5 followed by high involvement at 3, and low involvement at 2.7. This shows that schools should encourage parents and guardians to participate in their child’s academic curriculum to help improve exams scores.

**Assumptions**

This data is synthetic, and assumes that the information provided can be obtainable information in the real world. It is only a small sample of the data that actually exists. In using actual data and a significantly larger amount of data, different variables may emerge as the strongest contributing factors.

**Limitations**

The few columns that had missing data could affect the overall analysis. Categorical data can be subject to interpretation and this data only accounts for one subset of an overall population. This data doesn’t account for any sort of difference in curriculum that may exist between various school systems.

**Challenges**

Data across multiple institutions can vary and curriculums can vary. So, one institution that has a vast curriculum may out perform an institution that has a limited curriculum regardless of exam preparation or attendance.

**Future Uses/Additional Applications**

This data can be used to help students who are struggling in their courses to improve exam results. It can also be used by institutions to understand what predictors most contribute to exam performance. If a different school has internet access for example as a significant predictor in exam scores, schools can look into funding to provide families with access to the internet to help achieve academic success.

**Recommendations**

Based on the results per school, I’d recommend action plans to help lower scoring students improve their exam results. Because parental involvement is a significant factor in the linear regression, I’d suggest parents becoming more involved in their child’s academic curriculum. As a result of this finding, I dove deeper into a professional journal that studied the effect of parental involvement amongst students by William Jeynes. In his meta-analysis, Jeynes (2005) found that "parental involvement, as a whole, was associated with all the academic variables by about 0.7 to 0.75 of a standard deviation unit," indicating a significant positive relationship between parental involvement and academic achievement. However, looking at the other two models and the correlation matrix, I’d also suggest rewarding great attendance and developing a program to improve study habits. Looking further into real-world research studies, one study found that class attendance had a significantly positive impact on academic performance, especially among lower-performing students (Chen & Lin, 2023). Another study concluded that students who practiced structured and consistent study habits achieved higher academic outcomes, suggesting that teaching effective study techniques can lead to measurable improvements in performance (Shahrakipoor et al., 2021).

**Implementation Plan**

This model needs a significant amount of more data before it can be implemented within actual school systems. I suggest collecting more information and fine tuning the model prior to implementation.

**Ethical Assessment**

Since this data correlates to individual child performance, it is important to exclude any private information such as student name and exact address to mitigate any privacy concerns within the data. It is also important to address potential bias in the data such as socioeconomic disparities or gender-based differences.

**Frequently Asked Questions**

1. **Why did you choose to include Gradient Boosting alongside Random Forest?**
   1. **Answer:** I included Gradient Boosting to capture complex non-linear relationships that Random Forest might miss.
2. **Why did you remove features with missing data instead of inputting them?**
   1. **Answer:** The three columns that had missing data had a significant amount of missing data. Additionally, the data missing was categorical so to preserve model integrity and avoid introducing bias through imputation, I chose to remove them.
3. **Why did Linear Regression outperform the more complex models?**
   1. **Answer:** The relationships in the data appear to be primarily linear, particularly between parental involvement. This allowed Linear Regression to perform well compared to the other models, which don’t focus on linear relationships.
4. **How did you handle outliers in the data?**
   1. **Answer:** I created boxplots and distribution charts to visualize the existence of outliers. While outliers were present, they mostly reflected skewness, not data errors. I decided to retain them to avoid losing valuable variance.
5. **How reliable are the results since the data is synthetic?**
   1. **Answer:** While the data is synthetic, it simulates real-world patterns. The findings provide a strong foundation, but results may vary with actual data from educational institutions.
6. **How do you ensure your model is not overfitting?**
   1. **Answer:** I used a train-test split and evaluated all models on the test set using the evaluation metrics. I also chose simpler models and avoided excessive tuning to prevent overfitting.
7. **How can this model be applied in real schools?**
   1. **Answer:** Schools can use insights to prioritize the importance of parental involvement towards academic success. Schools can also emphasize attendance initiatives and structure study time. With real data, the model can guide targeted support strategies for struggling students. While this model shows parental involvement, attendance and study hours as strong correlations to test scores, different institutions may see different results.
8. **Were there any surprising features that didn't affect performance as expected or perform better than expected?**
   1. **Answer:** I found it surprising that in the linear regression model, that parental involvement served as the top 3 important features. This shows the difference between a correlation matrix and building a predictive model to uncover meaningful relationships that you may be unable to see otherwise.
9. **What would you improve or do next with this project?**
   1. **Answer:** I would collect real-world data across different schools, include more granular curriculum details, and try cross-validation or tuning to improve model robustness.
10. **Is this model ready for deployment?**
    1. **Answer:** Not yet. These models should be further evaluated with real world data to understand how they perform on a larger capacity scale. I think it is interesting how the linear regression model find significant in different factors than non-linear models and I’d like to fine tune each approach to confidently select a model to deploy in the real world.

**References**

Chen, Y., & Lin, H. (2023). The effects of class attendance on academic performance. *Journal of*

*Economic Behavior & Organization*. <https://doi.org/10.1016/j.jebo.2023.08.017>

Jeynes, W. H. (2005). A meta-analysis of the relation of parental involvement to urban

elementary school student academic achievement. Urban Education, 40(3), 237–269. <https://doi.org/10.1177/0042085905274540>

Shahrakipoor, M., Ghaffari, S., & Aghaei, M. (2021). Relationship between study habits and

academic achievement in students of medical sciences. *Advances in Medical Education and Practice, 12*, 103–109. https://doi.org/10.2147/AMEP.S291036