HW # 1 - EDA and Machine Learning Analysis of Student Performance Data

Noori Selina, Haig Bedros, Julia Ferris, Matthew Roland

In this project, we analyzed two datasets focused on student performance and used machine learning algorithms to predict exam outcomes. Dataset 1, the larger of the two, includes a broad range of variables that capture different aspects of student behavior and their academic environment. These variables cover factors such as hours studied, attendance, parental involvement, access to resources, extracurricular activities, sleep hours, and previous scores. It also includes details such as motivation level, tutoring sessions, family income, and teacher quality, along with categorical factors like school type, peer influence, and parental education level.

Dataset 2 is more straightforward, focusing primarily on academic outcomes, with variables such as math, reading, writing, and placement scores. While it includes fewer variables, it offers a clear overview of students' academic performance. The goal of our analysis was to identify the factors that most influence student exam scores and assess how well machine learning models such as Random Forest and XGBoost can predict these outcomes using the available data.

Are the Columns of Your Data Correlated?

The correlation analysis in our exploratory data analysis revealed significant insights into the relationships between variables in Dataset 1. The correlation matrix showed that hours studied and attendance had moderate positive correlations with exam scores, with correlation values of 0.45 and 0.58, respectively. This suggests that students who study more and attend school regularly tend to achieve better results. Other variables, such as sleep hours and physical activity, exhibited weak correlations with exam scores, indicating little influence on performance.

In Dataset 2, correlations between variables were much weaker. The strongest correlation found was between math and writing scores, but even this was low at 0.13. This lack of correlation made it challenging to predict placement scores based on other exam results, as the variables appeared largely independent of each other.

Are There Labels in Your Data? Did That Impact Your Choice of Algorithm?

Both datasets had clearly defined target variables, or labels: the exam score for Dataset 1 and the placement score for Dataset 2. These labels influenced our choice of machine learning algorithms. Because these scores were continuous but discrete (i.e., they had no decimal values), we considered algorithms that could handle both categorical and numeric data while capturing non-linear relationships. While linear regression was initially considered, it was discarded due to the discrete nature of the data. We ultimately selected Random Forest Regression and XGBoost, which are better suited for managing mixed data types and non-linear interactions.

What Are the Pros and Cons of Each Algorithm You Selected?

Random Forest Regression is particularly useful when working with a combination of categorical and numeric data, and is good at identifying non-linear relationships. One of its strengths is that it's relatively resistant to overfitting, which makes it reliable when handling datasets such as Dataset 1, since it includes diverse variables. However, a downside of Random Forest is that it can be computationally heavy, especially as the size of the data grows, and it doesn't offer the same transparency as simpler models like linear regression. This lack of interpretability can be a drawback when it comes to explaining exactly how the model arrives at its predictions.

XGBoost, on the other hand, is known for its speed and effectiveness, particularly with larger, and more complex datasets. Its regularization feature helps prevent overfitting, which is especially valuable when the dataset includes many features. But XGBoost can be more sensitive to tuning, meaning that achieving the best performance often requires careful adjustment of its parameters. While powerful, it shares a similar challenge with Random Forest in that both models can be difficult to interpret—this becomes especially problematic when the data itself, like in Dataset 2, shows weak relationships, making it harder for even sophisticated models to deliver accurate predictions.

How Does Your Choice of Algorithm Relate to the Datasets?

Our choice of algorithms was influenced by the structure of the datasets. Since Dataset 1 had a mix of both categorical and numeric variables, we decided to run Random Forest and XGBoost. These algorithms are well-suited for handling non-linear relationships and complex interactions between factors like parental involvement and teacher quality and how they impact exam scores.

On the other hand, Dataset 2 posed a bigger challenge. The weak correlations between the variables made it hard for even advanced algorithms like Random Forest and XGBoost to make accurate predictions. This highlighted that, no matter how sophisticated the algorithm is, it can still perform poorly if the dataset doesn't have strong, meaningful relationships between the features and the target variable.

Which Result Would You Trust if You Need to Make a Business Decision?

If we had to make a business decision based on this analysis, I would rely on the results from the Random Forest model applied to Dataset 1. With an R-squared value of 0.54 and an MAE of 1.32, the Random Forest model provided a reasonable level of accuracy in predicting exam scores. It explained 54% of the variance in student performance, making it a useful tool for identifying which factors most influence outcomes. In contrast, the models used on Dataset 2,

including XGBoost, performed poorly, with negative R-squared values and high error rates. This suggests that Dataset 2 is not reliable for predictive modeling..

Do You Think Analysis Could Be Prone to Errors When Using Too Much or Too Little Data?

Analysis can definitely be prone to errors when working with too much or too little data. For example, Dataset 2, had a relatively small number of features and weak correlations between variables which led to poor model performance. Without strong relationships between the input variables and the target variable, machine learning models struggle to make accurate predictions. On the other hand, using too many irrelevant features in a dataset can create complexity and potentially lead to overfitting. In Dataset 1, the balance between numeric and categorical variables allowed Random Forest to perform well, but including too many unrelated features would have likely reduced the model's accuracy.

How Does the Analysis Between Datasets Compare?

The comparison between Dataset 1 and Dataset 2 highlights the importance of data quality and structure when applying machine learning models. Dataset 1 was much more suitable for predictive modeling because it contained a mix of variables that were correlated, which allowed algorithms like Random Forest and XGBoost to identify important patterns in the data. Both algorithms performed reasonably well on Dataset 1, with Random Forest achieving an R-squared value of 0.54 and a low MAE. However, there is still room for improvement, particularly in exploring variables like *motivation* and their relationship to performance, which could potentially enhance the model's accuracy. In contrast, Dataset 2's weak correlations and limited number of features made it difficult for either algorithm to produce meaningful predictions. The negative R-squared values and high error rates demonstrated that Dataset 2 is better suited for exploratory data analysis rather than predictive modeling. Future analyses could focus on deepening these investigations to improve predictive outcomes.

Conclusion

Overall, the choice of Random Forest and XGBoost for Dataset 1 was appropriate given the mix of categorical and numeric variables and the presence of non-linear relationships. By using these models, we were able to identify important patterns in the data, though there is room for improvement. Dataset 2, on the other hand, posed significant challenges for predictive modeling due to the weak correlations between variables. This project highlights how crucial it is to choose the right machine learning models based on the structure and quality of the dataset. It also emphasizes the importance of conducting thorough exploratory data analysis before attempting any predictive modeling, as it can significantly improve the accuracy and effectiveness of the modeling process.