Linear Regression Analysis

1 Introduction

This report documents the process of analyzing student performance using linear regression models. The dataset was obtained from Kaggle, and features such as socioeconomic score, study hours, sleep hours, and attendance percentage were analyzed to predict student grades.

2 Data Exploration and Visualization

The initial exploration involved visualizing the data through scatter plots to understand the relationships between features and the label (grades). The following figure shows the scatter plots for the features against grades:

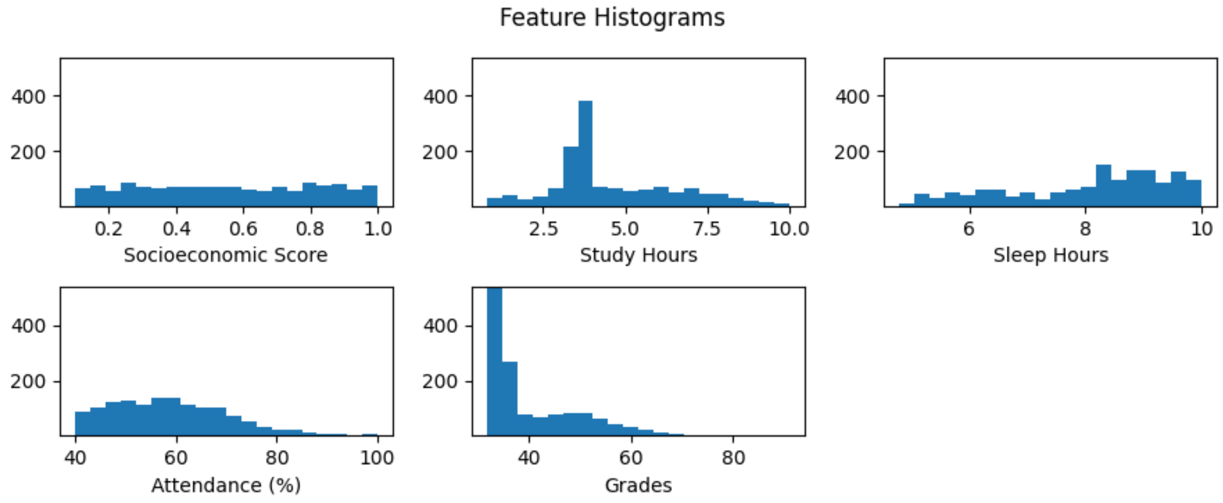

Figure 1: Histogram plots of features vs. grades

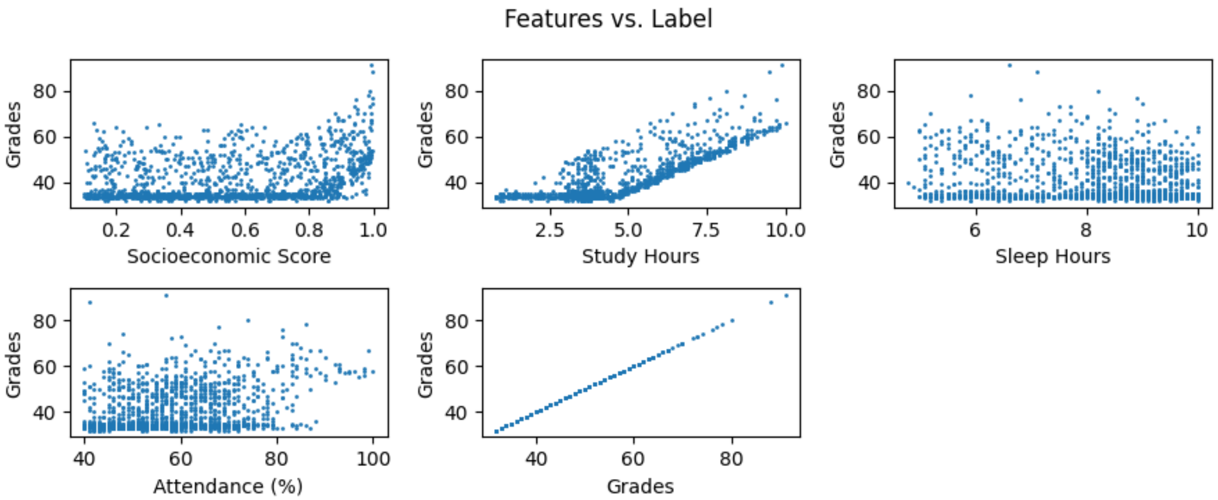

Figure 2: Scatter plots of features vs. grades

Observations:

- Study hours and socioeconomic score show some linear trends with grades. But I can already tell that a linear function won't be enough.
- Sleep hours and attendance percentage have weaker or unclear relationships with grades.

3 Model Training and Results

Several linear regression models were trained with different preprocessing and parameter configurations. The key results are summarized below:

3.1 First Try: No Preprocessing

- **R^2:** -2.52e+18
- **MSE:** 2.17e+20
- **MAE:** 1.36e+10

This attempt gave very poor results, there is a clear need for data preprocessing.

3.2 Second Try: Standard Scaler

- **$\mathbf{R^2}$:** 0.766
- **MSE:** 20.20
- **MAE:** 3.45

The introduction of standard scaling significantly improved the model's performance.

3.3 Third Try: Parameter Tuning

- **$\mathbf{R^2}$:** 0.740
- **MSE:** 22.44
- **MAE:** 3.34

This approach did not improve compared to the second try, suggesting that standard scaling alone was enough.

3.4 Fourth Try: Bayesian Regression with Polynomial Features

- **$\mathbf{R^2}$:** 0.926
- **MSE:** 6.38
- **MAE:** 1.96

This model provided the best results, showcasing the importance of adapting the model and preprocessing techniques to the data.

4 Model Visualizations

The following figures illustrate the fit of each model:

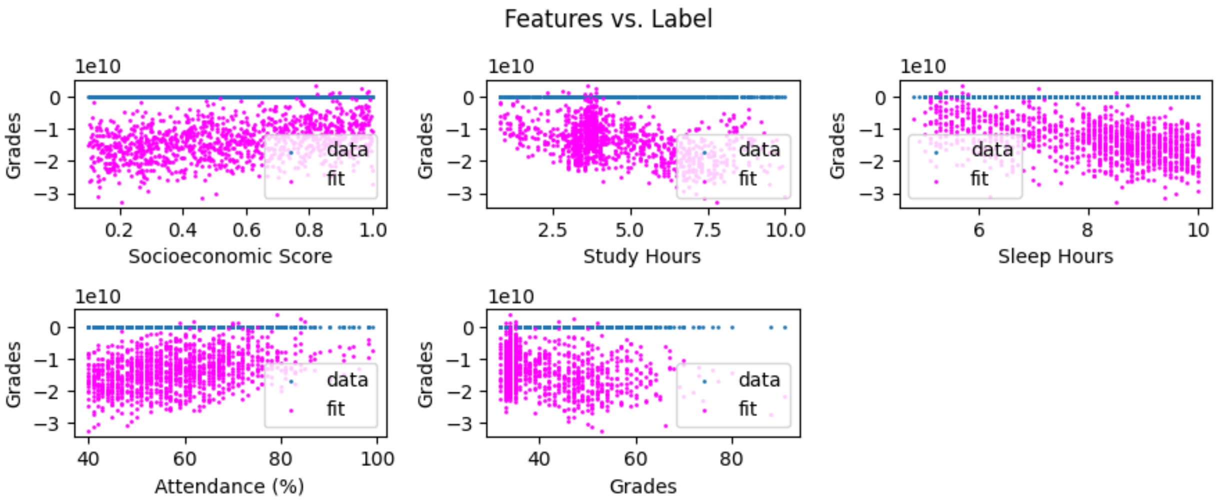

Figure 3: Model 1: No Preprocessing

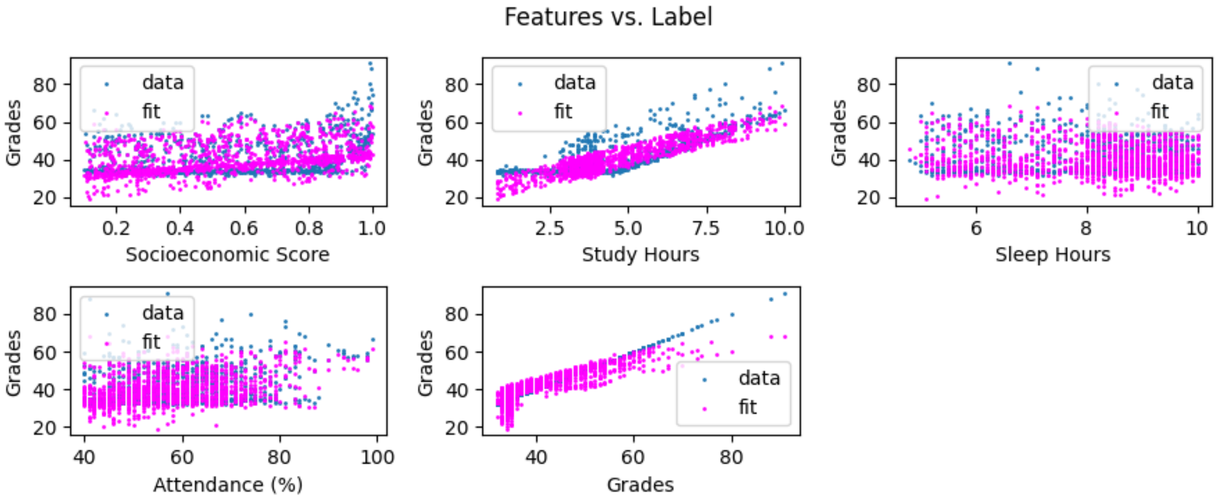

Figure 4: Model 2: Standard Scaler

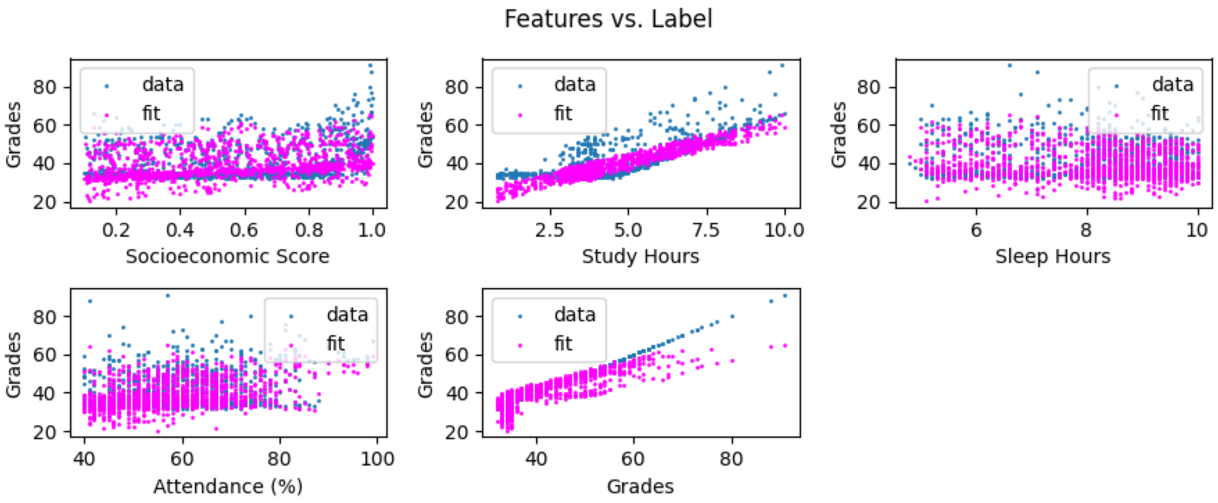

Figure 5: Model 3: Standard Scaling with Parameter Tuning

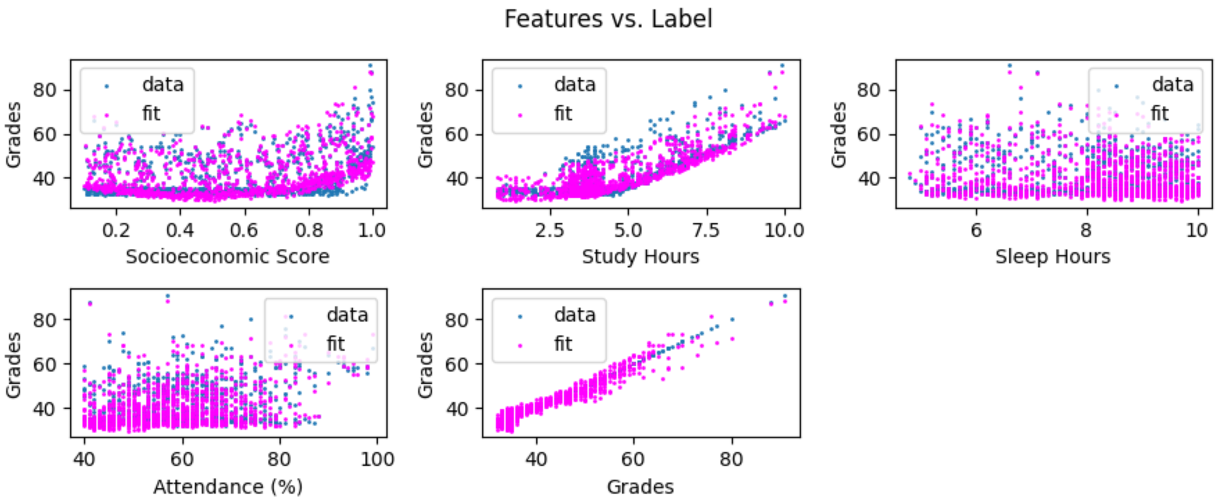

Figure 6: Model 4: Bayesian Regression with Polynomial Features

5 Conclusion

With this I understood how crucial it is to match the training process to the characteristics of the data itself. The initial results were less than ideal, but by changing the approach—like incorporating scaling or choosing a better fitted regression method—the model performance improved a lot. This shows how important it is to really "mold" the training process to fit the data. In particular, the Bayesian regression with polynomial features was able to capture the underlying trends effectively, achieving an R^2 score of 0.926. This demonstrates that understanding the data and experimenting with preprocessing can make a huge difference in predictive modeling.