Project Check-in Outline

Problem Statement

Description: Predicting student exam scores based on various academic, environmental, and personal factors. The research question is "What factors most significantly predict student exam performance, and how accurately can we predict exam scores using suitable machine learning models?" The machine learning tasks I have chosen are Regression and supervised learning, because this would help universities or educational institutions in general to identify students who potentially could be affected by certain factors and aid in personal intervention strategies.

Significance

- Understanding factors that affect academic achievement.
- Obtaining early indicators of student underperformance.
- Improving resource allocation for the affected students.
- Obtaining personalized learning pathways for certain students.

Dataset Description

Source: Pulled from Kaggle.com

Number of Observations: 6,608

Features (19 predictors, 1 response variable)

Quantitative Features

- Hours_Studied
- Attendance
- Sleep_Hours
- Previous_Scores
- Tutoring_Sessions
- Physical_Activity

Qualitative Features

- Parental_Involvement
- Access_to_Resources
- Extracurricular_Activities
- Motivation_Level
- Internet_Access
- Family_Income
- Teacher_Quality
- School_Type
- Peer_Influence
- Learning_Disabilities

- Parental_Education_Level
- Distance_From_Home
- Gender

Response Variable

- Exam_score

Justifying the suitability of this dataset

- Appropriate sample size.
- Response variable is continuous, appropriate for regression analysis.
- Relevant features align with educational research on student performance.

Data Exploration

Summary statistics and Visualizations: Mean, SD (for numerical variables). Frequencies for categorical variables.

Relationships and patterns: Correlation analysis between quantitative predictors and Exam_Score, distribution of Exam_Score, relationship between Hours_Studied and Exam_Score, Relationship between Previous_Scores and Exam_Score.

Handling of Missing/Imbalanced Data

- I will be reporting missing NA values in this dataset, if any

Data Visualizations: Histogram of Exam_Score, boxplot of Exam_Score by categorical variable, scatterplot of Hours_Studied vs. Exam_Score, scatterplot of Previous_Scores vs. Exam_Score, scatterplot of Attendance vs. Exam_Score, Correlation heatmap between quantitative predictors, Residual plots.

Interpretations: I will be identifying important trends and discussing any potential outliers.

Model Selection, Application, and Evaluation

Models chosen

- OLS Linear Regression
- Ridge Regression
- Lasso Regression

The linear models are suitable because the response variable is continuous and since I am predicting a score, I utilize linear regression. Additionally, with many predictors, some may be correlated. Multicollinearity makes it hard to tell which predictor matters, but regularization helps us rectify that problem. It is also important to use Lasso and Ridge since there are a lot of predictors, which would increase even more after encoding. There is also a risk of overfitting with OLS.

Method Applications

- Loading and cleaning of data.
- Train-test-split.
- Standardization.
- Implementing cross-validation for Lasso and Ridge respectively.
- Explaining cross validation and the selection of the best lambda value.

Model Evaluation

Evaluation metrics include:

- MSE
- RMSE
- Residual analysis

I will also be sure to create a table of comparisons that analyzes metrics across different models, while also discussion bias-variance tradeoff.

Results Presentation

- Tables
- Visualizations
- Key statistics
- Tying it back to my original statement.
- Making sure that I am able to interpret the results I obtain.
- Discuss different limitations.
- Being prepared to explain how this project has prepared me to work in my preferred industry (software and data engineering).

Code Implementation and Documentation

- Making sure code runs with no errors.
- Ensuring proper documentation through commenting and detailed pushes to GitHub.
- Using the README.md file to give instruction to run code.
- Making sure my code is well organized and easy to interpret.