1/15/26 - Regression_Case_Study_Student_Performance

Name: Koume Matsutori
UW ID: 2263445

1. Best Model Selection (Test Performance) - From your results table, which model (**Linear / Poly-2 / Poly-3 / Poly-4**) gives the **best Test R²** in **Scenario A** (dropping Extracurricular Activities)? ✅ Report the **Test R² value**, and briefly explain what it means.

Linear	Poly-2	Poly-3	Poly-4
0.988701	0.988681	0.988693	0.988628

Linear regression gives the best Test R^2 value, as shown in the table above. The R^2 value, 0.9888701, means that the linear model explains ~98.87% of the variance in student performance on tests.

2. Effect of Polynomial Degree - Does increasing the polynomial degree always improve **test performance**? ✅ Support your answer using at least **one metric trend** (example: Test MSE or Test R²) from your results.

No. As discussed in question 1, the linear model had the best R^2 value, and the R^2 value slightly decreased as the polynomial degree increased.

3. Did “Extracurricular Activities” Help?

Compare:

- **Scenario A:** Drop Extracurricular Activities
- **Scenario C:** One-hot encode Extracurricular Activities

Did including extracurricular information improve the model’s **test performance**?
✅ Justify your answer using **Test MSE / Test MAE / Test R²**.

	Test MSE (linear)	Test MAE (linear)	Test R^2 (linear)
Scenario A	4.181380	1.630177	0.988701
Scenario C	4.066564	1.609044	0.989011

The table above compares the linear regression results of Test MSE, Test MAE, and Test R^2 values for Scenario A and C. Using this table, it is clear that extracurricular activities improved the test performance, as MSE and MAE decreased, whereas R^2 increased. MSE and MAE

lowering signifies lower error – higher accuracy, and higher R^2 means Scenario C explains more variance in student performance.

4. Separate Models vs One Combined Model

Compare:

- **Scenario B:** Separate models for `Extracurricular Activities = Yes` and No
- **Scenario C:** One combined model with one-hot encoded extracurricular features

Which approach appears to generalize better on the **test set**, and why?
✅ Use test metrics to support your claim.

Average Test MSE Scenario B: 4.146977
Test MSE Scenario C: 4.066564

Average Test MAE Scenario B: 1.610302
Test MAE Scenario C: 1.609044

Average Test R^2 Scenario B: 0.9888865
Test R^2 Scenario C: 0.989011

I took the average of the Test MSE, Test MAE, and R^2 by averaging the “YES” and “NO” values of each linear regression model. As shown, MSE and MAE values for Scenario C are lower than the average MSE and MAE values from Scenario B, and the Scenario C’s R^2 value is greater than Scenario B. Therefore, I can conclude that Scenario C appears to generalize better on the test set.

5. Encoding a Categorical Feature: 0/1 vs One-Hot Vectors - The feature **Extracurricular Activities** is categorical with two values: **Yes** and **No**.

Two possible encodings are:

- **Binary encoding:** No → `0`, Yes → 1
- **One-hot encoding:** Yes → `[1, 0]`, No → `[0, 1]`

Which encoding is more appropriate for regression models and why?
✅ In your answer, comment on whether binary encoding introduces an *artificial ordering* or *distance* between categories, and why one-hot encoding can avoid that.

One-hot encoding is more appropriate. If No is assigned to 0 and Yes is assigned to 1 (binary encoding), it assigns Yes to a higher numerical value, which introduces artificial ordering and hierarchy. On the other hand, one-hot encoding completely avoids this issue as each category is represented as a separate indicator variable.