**Data Product 2: Supervised Machine Learning**

Data Selection and Preprocessing:

Students’ annual assessment scores by school and by district in the State of Ohio. School year 2021-2022

(The Ohio Department of Education)

<https://reportcard.education.ohio.gov/download>

Description of Data:

Student’s annual assessments for 2021-2022

606 districts (rows)

29 (columns)



Average Income (Independent Variable) – Range $29,716 to $481,096 - Mean = $65,335

Math Scores\* (Dependent Variable) - 13.5% to 92.0% - Mean = 63.52%

\*Percentage of the district's students who scored proficient or better

Model Description:

Regression using the Average income of the area (IV) to predict the district Math Scores (DV)

Train/Split = 80/20 484/122

5-fold cross-validation to optimize the hyperparameters

The dataset was split into five subsets, and the model was trained on four folds while being validated on the remaining one, repeating this process five times.

Results:

Presenting the key performance metrics:

Fold 1: RMSE = 11.785544544686529

Fold 2: RMSE = 12.652971300061818

Fold 3: RMSE = 14.661560745752126

Fold 4: RMSE = 10.545333073946992

Fold 5: RMSE = 12.52522219062801

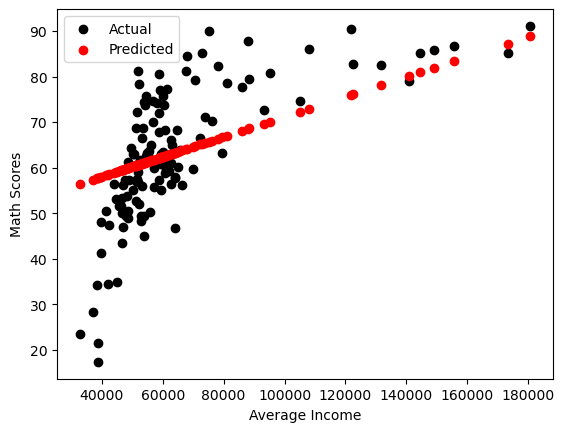
Average RMSE across all folds: 12.434126371015093

Linear Regression Coefficients: [0.00023929]

Linear Regression Intercept: 47.884174976821896

Mean Squared Error: 138.8990602147904

R-squared: 0.37054595223700526



Results Interpretation

The MSE is 138.90 and suggests a low level of accuracy.

R-squared of 0.37 or 37.05% of the DV is explained by the IV. This could be higher in a more responsive model.

Linear Regression Coefficients: [0.00023929] also suggest that the variable portion of the equation has very little weight on the overall model.

The average RMSE across all folds is approximately 12.43.

The graph (visually) supports these interpretations. The model has a low level of accuracy.

Conclusion

The model’s low level of accuracy would preclude it from being useful as a tool. Maybe a more complex model (Polynomial Regression), or the use of a more predictive variable or variables.

Ervin Matthews

BUS-759-901: Machine Learning and AI

Dr. Josh Lambert

December 12, 2023