Final Project

Aiying Huang Mia Yu Eunice Wang

Abstract

a brief introduction, brief description of methods, and main results into a one-paragraph summary

Introduction

brief context and background of the problem

Methods (data description and statistical methods)

Missing Value Treatment:

The dataset exhibited minimal missing values, predominantly in qualitative variables such as EthnicGroup, ParentEduc, TestPrep, ParentMaritalStatus, PracticeSport, IsFirstChild, NrSiblings, and TransportMeans. Most variables had a low incidence of missing data (<10%), with the exception of TransportMeans. Given the qualitative nature of these variables, mode imputation was applied for all except TransportMeans. Samples still exhibiting missing values post-imputation were directly excluded.

Model Assessment:

Cook's Distance (Residual vs Leverage Plot)

Cook's Distance and the Residual vs. Leverage Plot are tools used in regression analysis to identify influential observations that might unduly affect the model's parameters. Cook's Distance measures the influence of each observation on the fitted values, highlighting points that might be outliers or have a significant effect on the model's overall fit.

$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{pMSE}.$

where: D_i represents the Cook's Distance for the i^{th} observation. $\hat{Y}_j$ is the predicted value for the j^{th}observation, based on the model fitted with all data points. $\hat{Y}_{j(i)}$ is the predicted value for the$ j^{th}$ observation, based on the model fitted without the $ i^{th} $observation. p is the number of parameters in the model.

MSE is the Mean Squared Error, which is the average of the squared residuals of the model. The Residual vs. Leverage Plot helps to visualize these influential points, displaying residuals against leverage values. Points with high leverage and large residuals are often considered influential.

VIF (Variance Inflation Factor)

Due to the categorical nature of the dataset variables and the presence of numerous interaction terms in our models, the traditional VIF was not adequate for assessing multicollinearity. Instead, we utilized the adjusted Generalized Variance Inflation Factor (GVIF).

$GVIF^{(1/(2\times Df))} = \left(\frac{1}{1-R_j^2}\right)^{(1/(2\times Df_j))}.$

where: R_j^2 is the coefficient of determination for the explanatory variable j (which may be a categorical variable) in a regression model. df_j is the degrees of freedom for the variable j. For categorical variables, the degrees of freedom are usually the number of categories minus one.

An adjusted GVIF value ranging from 1 to 3 indicates no significant issues, while values exceeding 5 or 10 suggest notable multicollinearity.

Model Validation:

10-fold Cross Validation

A 10-fold cross-validation approach was employed. This method involves partitioning the data into 10 subsets, using 9 for training and 1 for validation in a rotating fashion. This process is repeated until each subset has been used for validation.

The Mean Squared Prediction Errors (MSPE)

The Mean Squared Prediction Error (MSPE) is a statistical measure used to evaluate the accuracy of a predictive model. It calculates the average of the squares of the prediction errors, where a prediction error is the difference between the observed values and the predicted values made by the model. MSPE is particularly useful in regression analysis and time series forecasting, as it provides a clear indication of how well a model predicts new data.

$MSPE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

where:

MSPE represents the Mean Squared Prediction Error. $\frac{1}{n}$ is the fraction representing the average over n observations. $\sum_{i=1}^{n}$ is the summation notation, summing over all n observations. $(y_i - \hat{y}_i)^2$

represents the squared difference between the actual value y_i and the predicted value$\widehat{y}_i$ for each observation i.

Results

□□□□□□□□□□+After missing value treatment, we left 846 samples for further analysis.

□□□□□□□□

To facilitate internal validity assessment in subsequent modeling, the data was split into a training set (80%) and a test set (20%).+□□□□□□□

Examination of the residual vs leverage plots for the three models revealed a few outlier observations, notably in samples 181 and 268. However, closer inspection showed that their leverage did not exceed 0.5, and Cook's distances were below 0.1. Additionally, no data entry anomalies were identified in these samples. Removing these samples and reconstructing the models showed negligible differences from the original models. Therefore, no adjustments were made, and the original data was used for modeling. As for the multicollinearity check, we found no substantial multicollinearity concerns in any of the model variables.

The cross-validation results indicated that the RMSE for the Math model was concentrated around 11, while for the Reading and Writing models, RMSE values were centered around 12.5. To test for potential overfitting, the initially separated test data was employed to evaluate the predictive performance of the models. The Mean Squared Prediction Errors (MSPE) for the Math, Reading, and Writing models were 256.9196, 134.7220, and 127.4526, respectively. Contrary to the RMSE results, the Math model exhibited a lower performance compared to the Reading and Writing models. Given that the Math model incorporated the most predictors, it suggests a potential issue with overfitting.

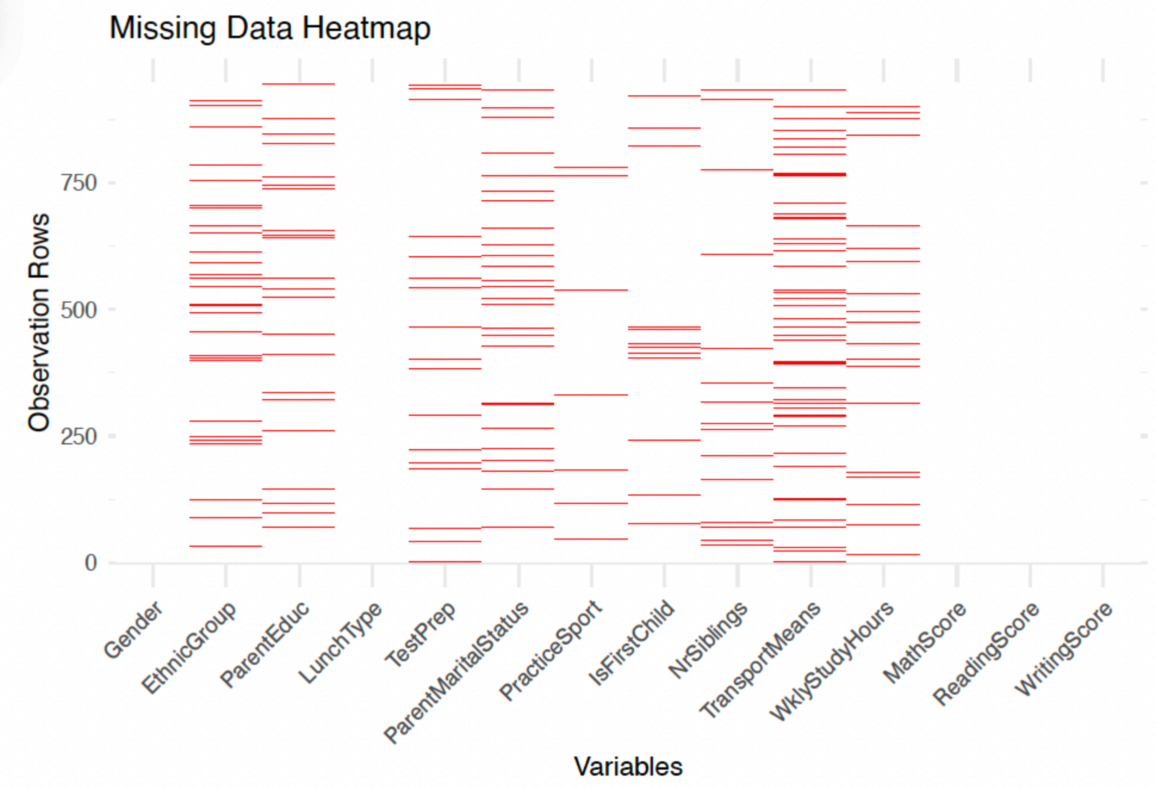
Conclusions/Discussion

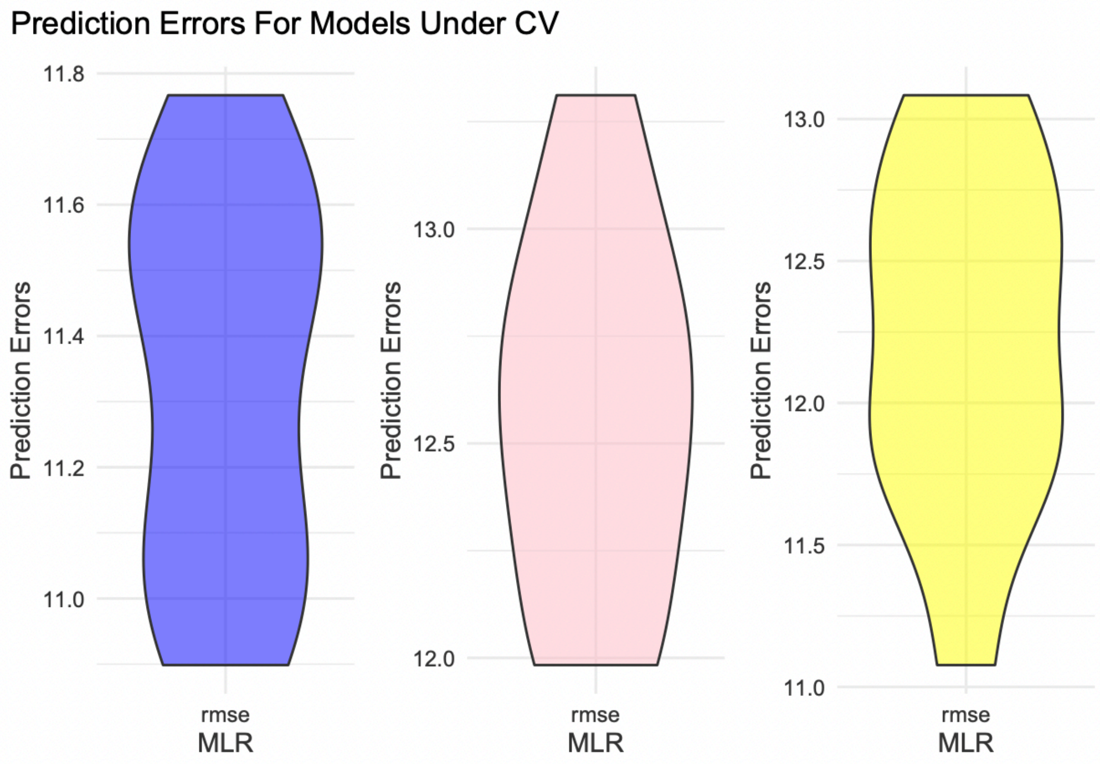
Your content here.

A brief summary on each group member’s contribution

Your content here.

Figures and Tables





References□□□□APA□

Bollinger, G. (1981). Book Review: Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Journal of Marketing Research, 18(3), 392-393. https://doi.org/10.1177/002224378101800318

Fox, J., & Monette, G. (1992). Generalized Collinearity Diagnostics. Journal of the American Statistical Association, 87(417), 178-183. https://doi.org/10.2307/2290467

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice (2nd ed.). OTexts.

Appendix

Appendix content goes here.