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:

Minimal missing values were observed, primarily in qualitative variables such as EthnicGroup, ParentEduc, TestPrep, and others. Mode imputation was used for all except TransportMeans, with samples still showing missing values after imputation excluded.

Model Assessment:

Model Selection:

Although correlations among variables were not evident, the inclusion of interaction terms is justified based on theoretical and practical grounds. Full models incorporating all 11 predictors and their pairwise interactions were developed. Subsequent analysis will focus on choosing a simpler variable subset to reduce overfitting risks.

Initially, we combined automated methods and criteria-based selection, using both AIC and BIC for model choice. The AIC criterion, yielding fewer variables, was preferred for backward elimination.

Due to retaining too many variables, we applied LASSO for penalization, using cross-validation to find the optimal lambda. Interaction terms with shrinkage coefficients below 0.5 were removed, resulting in three more efficient, nested models.

Cook's Distance (Residual vs Leverage Plot)

Cook's Distance, combined with the Residual vs. Leverage Plot, identifies influential observations in regression models. It measures each observation's impact on the fitted values to highlight potential outliers.

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

VIF (Variance Inflation Factor)

Adjusted Generalized Variance Inflation Factor (GVIF) was used instead of traditional VIF to assess multicollinearity, especially suitable for categorical variables and models with interaction terms.

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

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.

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

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%).

By using a backward elimination approach based on the AIC criterion on the full model with interaction terms, followed by penalization with the LASSO method, we derived three succinct and optimal models. Diagnostic plots for these models indicate no issues.

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.

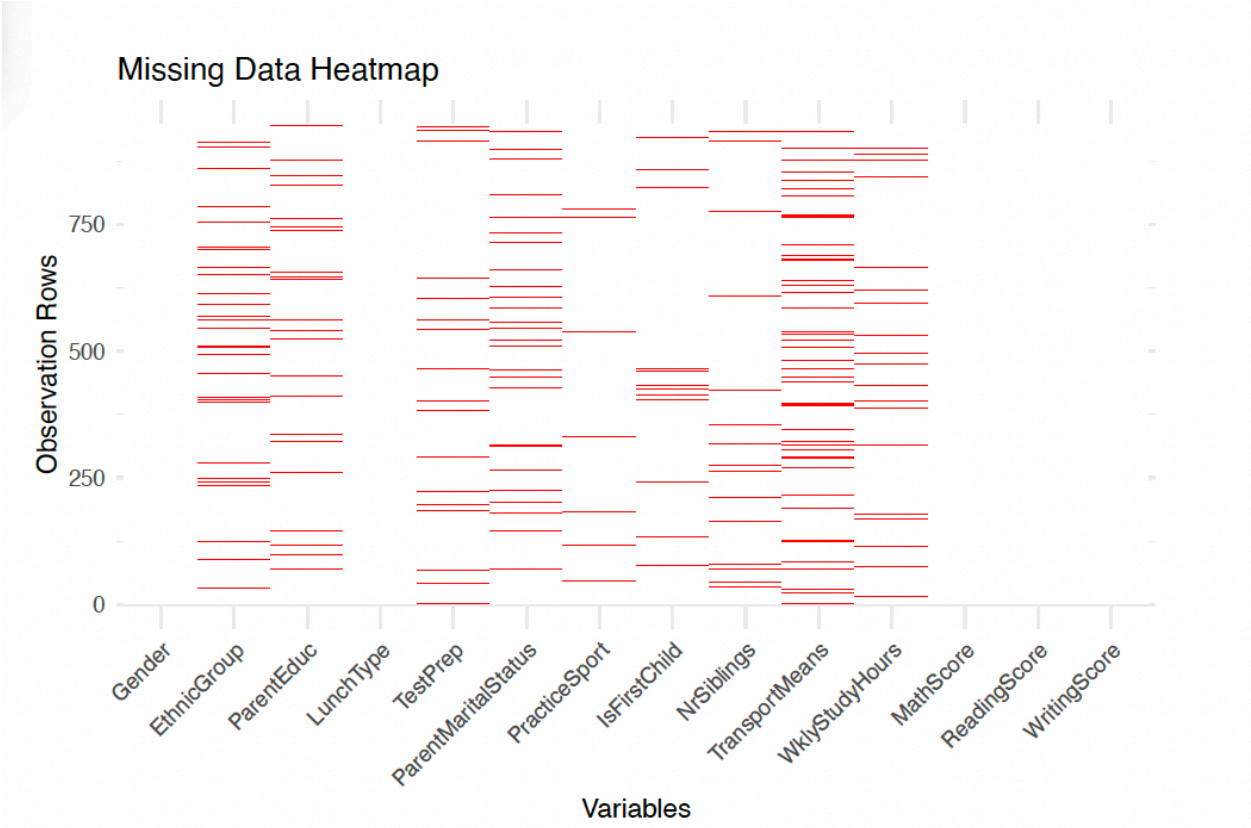
Conclusions/Discussion

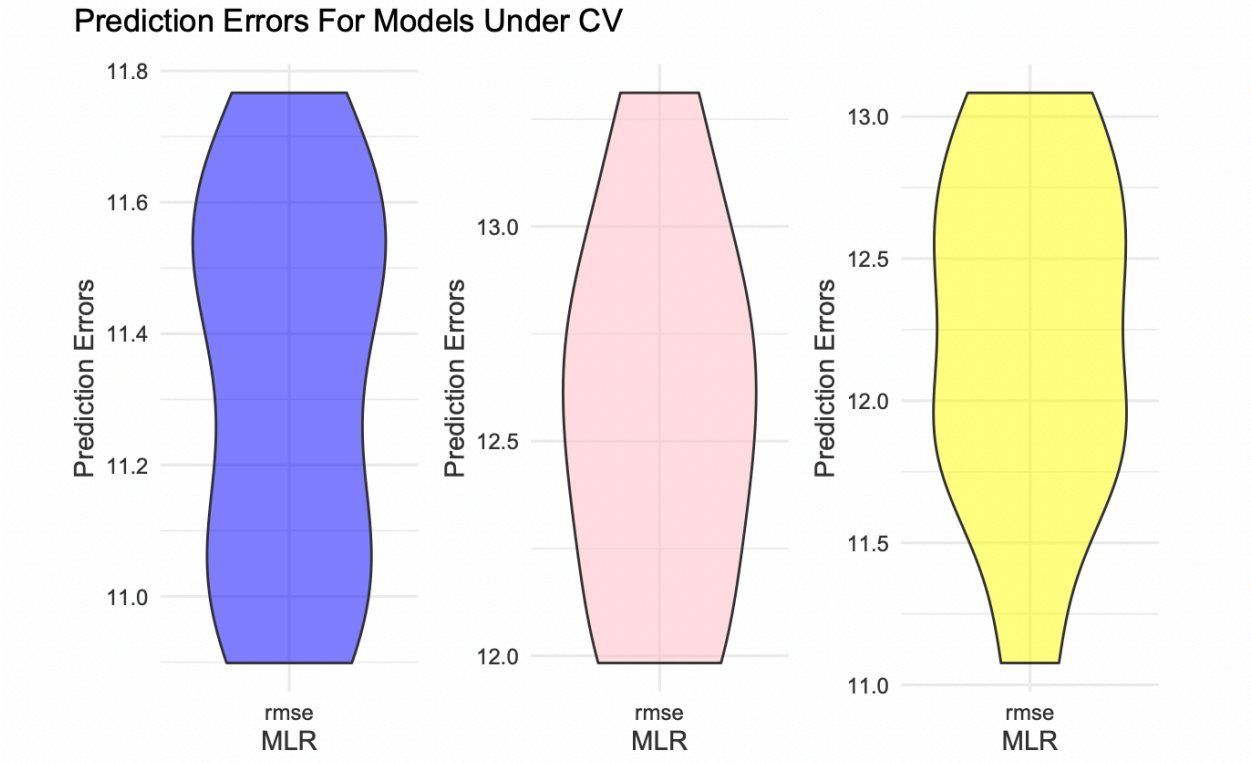
Your content here.

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.