Predicting Reading Scores with ML Algorithm

Introduction

The Program for International Student Assessment (PISA) is a triennial assessment that measures the abilities of 15-year-old students worldwide in reading, mathematics, science and other 21st century skills. This assessment serves as a tool for benchmarking student performance globally. In this analytic report, we aim to forecast the reading scores of U.S. students who participated in the 2009 PISA examination using a subset of the variables measured.

Data Description and Exploration

Overview

The data used originates from the 2009 PISA Public-Use Data Files, distributed by the United States National Center for Education Statistics (NCES). This data subset contains information about American participants of the exam. It is important to note that while the data does not contain direct identifiers of the students, users of this data are legally obligated to adhere to the NCES data use agreement. This agreement strictly forbids any efforts to identify individual students from the data.

The original PISA data has many features, and this US sample has been filtered to only include 24 variables, which include demographic information and potential predictors related to reading skill. The sample is split as 70% for training and 30% for testing: the training contains 3663 observations, and the testing set has 1570. Each row represents an individual student who took part in the PISA exam in that year.

The dataset includes both numerical and categorical variables. Key variables include grade, binary gender, various demographic indicators like race/ethnicity, educational history and plan, educational background of parents, and several others related to the student's learning environment and habits. The outcome variable “readingScore” is a numerical variable representing the student’s score on a 1000-point scale.

For specific variable information, please refer to table 1 below.

Table 1: Structured overview of the variables in the PISA 2009 dataset

Variable Name	Description	Type	Range/Options
grade	Grade in school of the student	Integer	8-12, but most are 10th grade for 15-year-olds
male	Whether the student is male	Binary	1 (Male), 0 (Not Male)
raceeth	Race/ethnicity composite of the student	Categorical	Various Categories
preschool	Whether the student attended preschool	Binary	1 (Yes), 0 (No)
expectBachelors	Whether the student expects to obtain a bachelor's degree	Binary	1 (Yes), 0 (No)
motherHS	Whether the student's mother completed high school	Binary	1 (Yes), 0 (No)
motherBachelors	Whether the student's mother obtained a bachelor's degree	Binary	1 (Yes), 0 (No)
motherWork	Whether the student's mother has part-time/full-time work	Binary	1 (Yes), 0 (No)
fatherHS	Whether the student's father completed high school	Binary	1 (Yes), 0 (No)

Variable Name	Description	Type	Range/Options
fatherBachelors	Whether the student's father obtained a bachelor's degree	Binary	1 (Yes), 0 (No)
fatherWork	Whether the student's father has part-time/full-time work	Binary	1 (Yes), 0 (No)
selfBornUS	Whether the student was born in the United States of America	Binary	1 (Yes), 0 (No)
motherBornUS	Whether the student's mother was born in the USA	Binary	1 (Yes), 0 (No)
fatherBornUS	Whether the student's father was born in the USA	Binary	1 (Yes), 0 (No)
englishAtHome	Whether the student speaks English at home	Binary	1 (Yes), 0 (No)
computerForSchoolwork	Whether the student has access to a computer for schoolwork	Binary	1 (Yes), 0 (No)
read30MinsADay	Whether the student reads for pleasure for 30 minutes/day	Binary	1 (Yes), 0 (No)
minutesPerWeekEnglish	Number of minutes per week the student spends in English class	Numeric	0-2400 with a mean of 265.7
studentsInEnglish	Number of students in the student's English class at school	Numeric	1-90
schoolHasLibrary	Whether the student's school has a library	Binary	1 (Yes), 0 (No)
publicSchool	Whether the student attends a public school	Binary	1 (Yes), 0 (No)
urban	Whether the student's school is in an urban area	Binary	1 (Yes), 0 (No)

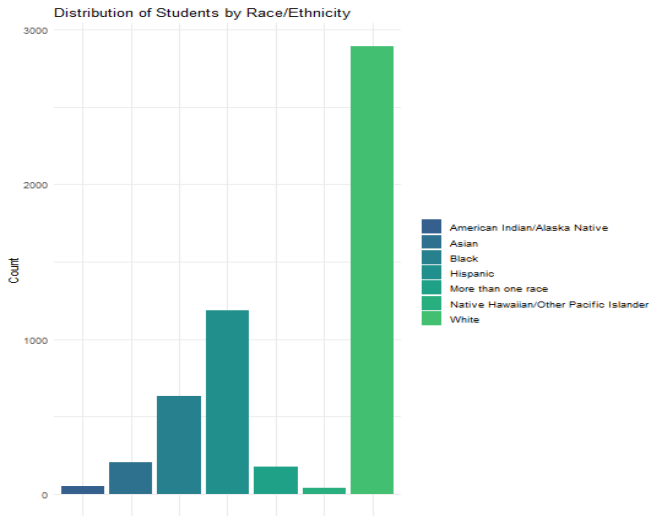
Variable Name	Description	Type	Range/Options
schoolSize	Number of students in the student's school	Numeric	100-6694
readingScore	The student's reading score	Numeric	1000-point scale

Table 2 provides a summary of the counts for each race/ethnicity category, and graph 1 provides the distribution of students by race/ethnicity. White is the dominating student group, and Hispanic constitutes the second largest student group.

Table 2: Race/Ethnicity Category

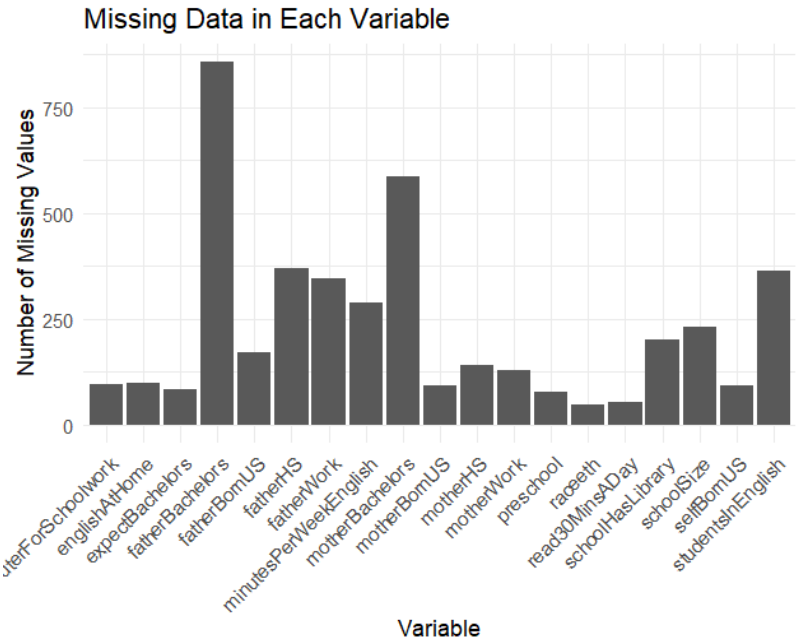
Race/Ethnicity Category	Count
American Indian/Alaska Native	51
Asian	204
Black	635
Hispanic	1184
More than one race	177
Native Hawaiian/Other Pacific Islander	40
White	2894

Graph 1:



Missing Data

Graph 2 demonstrates the counts of missing values in each feature. If the observations with missing values are to be deleted, then only approximately 62% of the observations would be retained. Thus simply removing those is not an viable option.



A missing data imputation was conducted, for the numerical variables, more specifically, the number of minutes per week the student spend in English class, the number of students in this student's English class at school, and the number of students in this student's school, I used the mean imputation.

Also, by evaluating the cross-tabulation, and then the results of chi-square tests of parental educational history and work. They are found to be significantly associated. Based on the observation, we assume the miss data pattern is Missing Not at Random, which means missingness is related to the unobserved data itself. The lack of information about the parent (missing data) is directly related to the reason why the data is missing (the student not knowing much about the parent). As a result, if the variables indicating the parent finished college, which is 1, replace the missing high school variable with “1”. For other missing variables, “Missing” are filled in.

Methods and Results

In our study, we partitioned the dataset into two subsets: 70% of the data was allocated for training the models, while the remaining 30% was used for testing purposes. We employed a variety of machine learning algorithms to model our data, including linear regression, LASSO regression, Classification and Regression Trees (CART), Random Forest (RF), and Support Vector Regression (SVR). To fine-tune our models and to ensure their robustness, we applied 10-fold cross-validation (CV) solely on the training data. This approach not only helps in model selection but also in hyperparameter tuning for LASSO, CART, RF, and SVR, optimizing each model's performance based on a decade-spanning CV framework.

Once the models were adequately trained and tuned, we evaluated their performance using the testing data. The metrics employed for this evaluation included Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). These metrics provide a comprehensive view of the models' predictive accuracy, with lower values indicating more precise predictions.

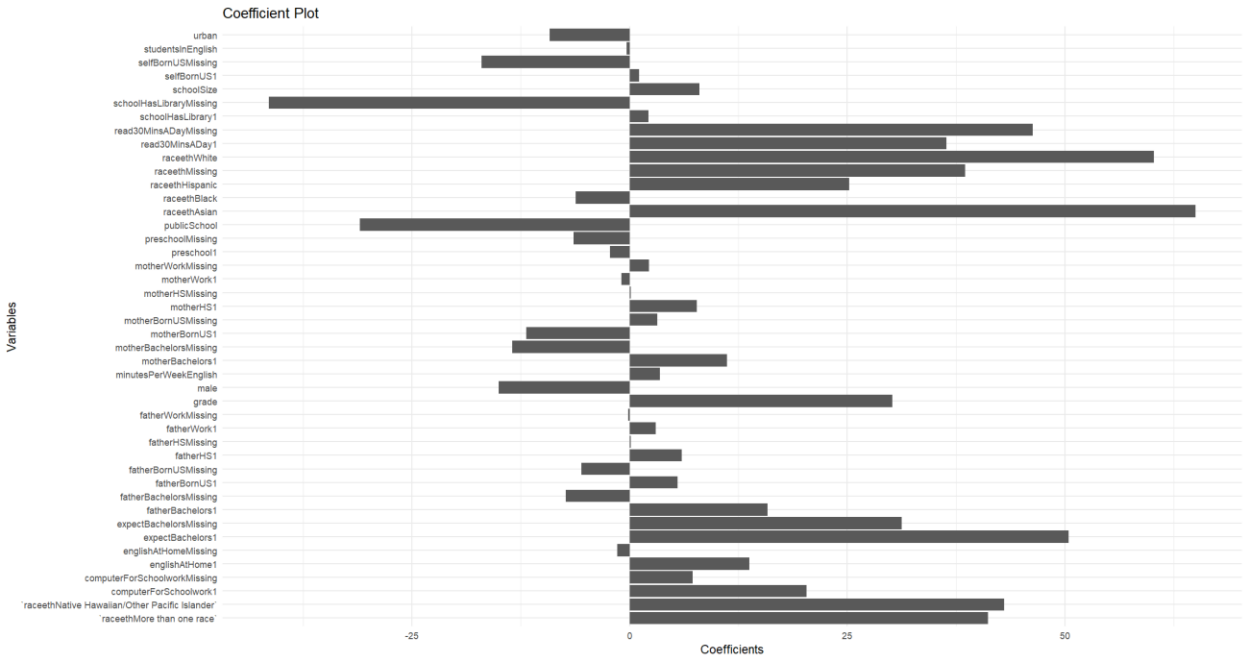
Moreover, for all models except SVR, we identified the most influential predictors. We selected the top four features that exhibited the highest coefficients in magnitude for linear regression and LASSO, as well as those with the greatest importance scores in CART and RF. These features are pivotal as they significantly contribute to the predictive power of the models, offering deeper insights into the underlying patterns within the data. The exception for SVR is due to its nature of modeling in a high-dimensional space, where feature importance is not as straightforwardly determined as in other models.

Table 3 is the performance metrics for linear regression and graph 3 illustrates the coefficients of the model. The most import predictors with estimates are the student's race is Asian (64.99), the student's race is White (60.24), the student expects to have a bachelor degree (50.43), the student reads 30 mins for pleasure a day (46.29).

Table 3: Linear Regression Metrics

Performance Metrics	Values
MSE	5838.98
RMSE	76.41
MAE	60.98

Graph 3: Linear Regression Coefficients Plot



LASSO regression was then fitted to the data, and based on CV, the lambda value that gives minimum mean cross-validated error is 0.51. Table 4 is the performance metrics for Lasso regression. The most import predictors with estimates are the student expects to have a bachelor degree (20.18), the student's grade (16.93), the student's race is White (16.72), the student reads 30 mins for pleasure a day (14.99).

Table 4: Performance metrics for Lasso regression

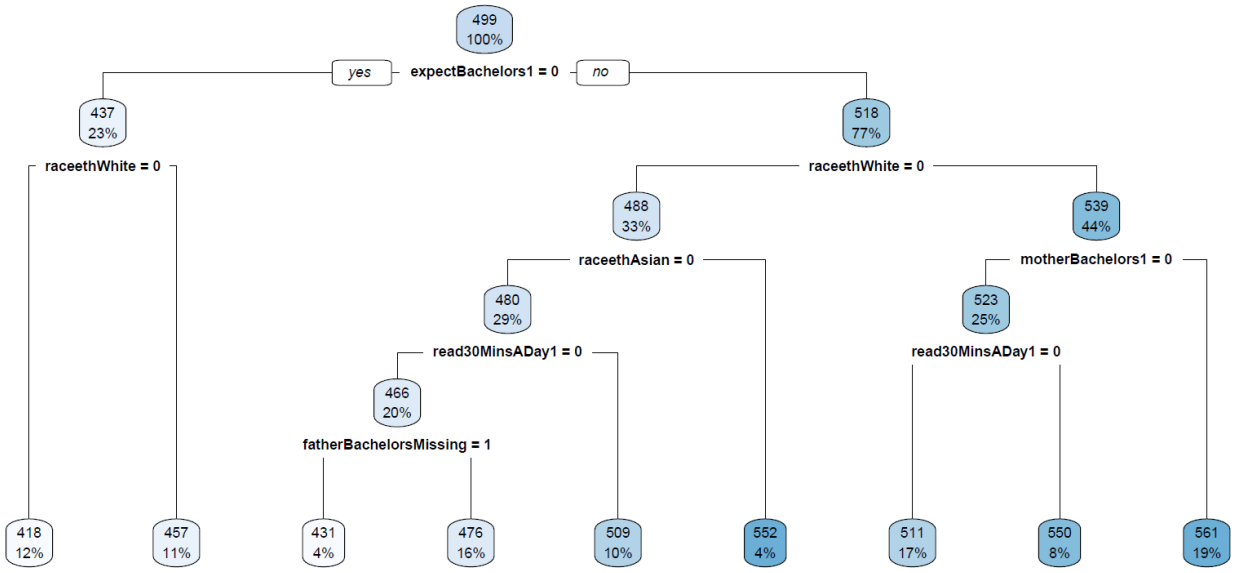
Performance Metrics	Values
MSE	5940.72
RMSE	77.08
MAE	61.60

CART was then employed, and table 5 gives the performance measures. The CP value is around 0.0065. The most import predictors are as followed: the student's grade (0.43), the student reads 30 mins for pleasure a day (0.36), the student's father has a bachelor's degree (0.31), the student is African American (0.27). These scores in the round brackets indicate how much each predictor contributes to the model's ability to make accurate predictions. A higher score means the variable is more important. Graph 4 illustrates the tree diagram.

Table 5: Performance metrics for CART

Performance Metrics	Values
MSE	7029.64
RMSE	83.84
MAE	67.53

Graph 4: Tree Diagram for CART

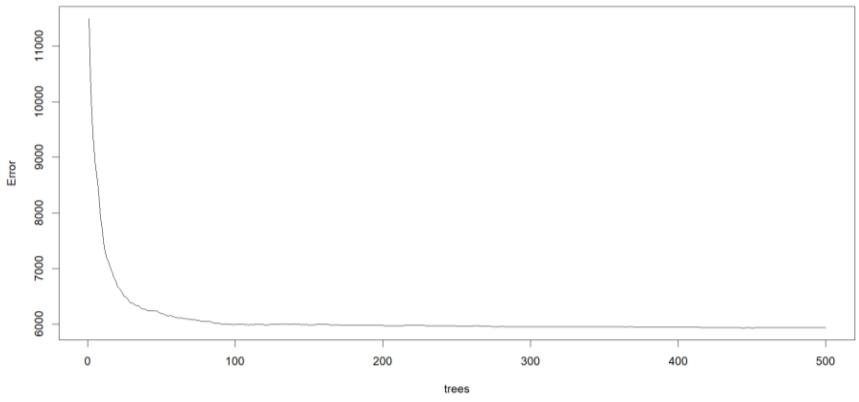


Next algorithm applied is random forest. Table 6 gives the performance metrics. Graph 5 gives the out of bag error rate. The number of variables randomly sampled as candidates at each split is 12, and the number of trees is 500. The most import predictors in a descending order are school size, the number of minutes per week the student spends in English class, the student expects to have a bachelor's degree, the number of students in the student's English class at school, as shown in graph 6.

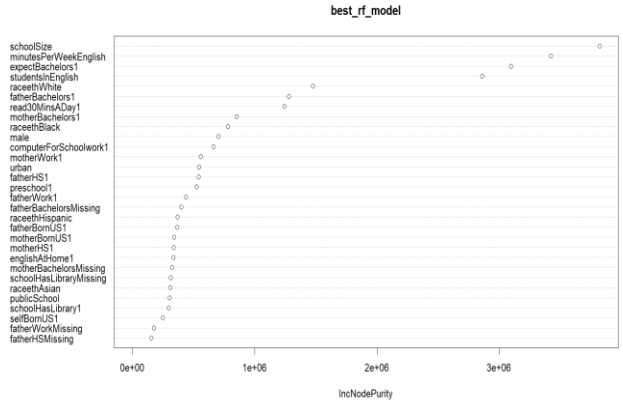
Table 6: Performance metrics for Random Forest

Performance Metrics	Values
MSE	13376.82
RMSE	115.66
MAE	92.28

Graph 5: OOB Error Rate



Graph 6: Important Predictors Graph for Random Forests



The last model used is support vector regression. Radial basis function (RBF) kernel was used, since a popular choice for SVM due to its locality and finite response across the entire range of the real x-axis. We tried five different sets of tuning parameters. One thing to notes is we converted sigma to gamma since the svm function in R uses the gamma parameter rather than sigma. The optimal sigma is 0.025 and C is 0.25

Table 7: Performance metrics for Support Vector Regression

Performance Metrics	Values
MSE	9606.82
RMSE	98.01
MAE	79.52

Table 8: Summary of all model metrics

Method	MSE	RMSE	MAE
Linear Regression	5838.98	76.41	60.98
LASSO Regression	5940.72	77.08	61.6
Support Vector Regression	9606.82	98.01	79.52
Random Forest	13376.82	115.66	92.28
CART	7029.64	83.84	67.53

Conclusion and Discussion

Among the models, linear regression shows the best performance in terms of the lowest MSE, RMSE, and MAE. The student's race and whether they read for pleasure are consistently found to be important across different models. The expectation to have a bachelor's degree also appears as a significant predictor in multiple models.