Significant Predictors for Final Grades

Vincent Johnson, Joseph Kang, Catherine Sung

April 2025

Project Overview

Vincent Johnson, Joseph Kang, Catherine Sung

- **Objective:** Predicting final grades (G3) for Math and Portuguese using statistical and machine learning methods.
- **Data:** Student performance data from two Portuguese secondary schools.
- **Methods:** A variety of statistical and machine learning methods including regression, classification, and ensemble methods.
- **Goal:** To identify key academic, personal, and social factors that predict student performance.

Data Overview

Vincent Johnson, Joseph Kang, Catherine Sung

- **Data Source:** UCI Machine Learning Repository, Cortez, P. (2008).
- **Variables:** 33 variables, including demographic, academic, and social factors.
- **Merging Data:** Merging datasets for Math and Portuguese, removing duplicates and redundant columns.
- **Key Variables:** G1, G2, and G3 (Grades), family background, study habits, and extracurricular activities.

Exploratory Data Analysis

Vincent Johnson, Joseph Kang, Catherine Sung

- **Visualizations:** Graphs and plots to explore relationships between variables.
- **Summary Statistics:** Descriptive statistics to understand the distribution of key variables (e.g., G1, G2, G3).
- **Insights:** Identify correlations, outliers, and missing values.

Methodology Overview

- **Regression Models:** Linear regression, logistic regression (for classification).
- **Ensemble Methods:** Random forest, bagging, boosting.
- **Dimensionality Reduction:** PCA, PLS.
- **Variable Selection:** Best subset selection, stepwise regression, lasso, ridge.
- **Hyperparameter Tuning:** Cross-validation for tuning model parameters.

Method Implementation

- **Model Selection:** Selection of models based on exploratory data analysis and research objectives.
- **Tuning:** Use of grid search and cross-validation to find optimal parameters.
- **Model Fitting:** Estimations, significance tests, and fitting results.
- **Interpretation:** Interpretation of model results and how they address the research questions.

Preparing the Data

Vincent Johnson, Joseph Kang, Catherine Sung

- **Bring in Data:** We began by bringing in the two necessary data sets, then combining them.
- **Check the Data for NA's:** There were no empty or blank values.
- **Training and Test Data:** Use consistent training and test datasets for comparison using seed 38520251.
- **Recommendations:**We will use this data moving forward.

Math EDA

Vincent Johnson, Joseph Kang, Catherine Sung

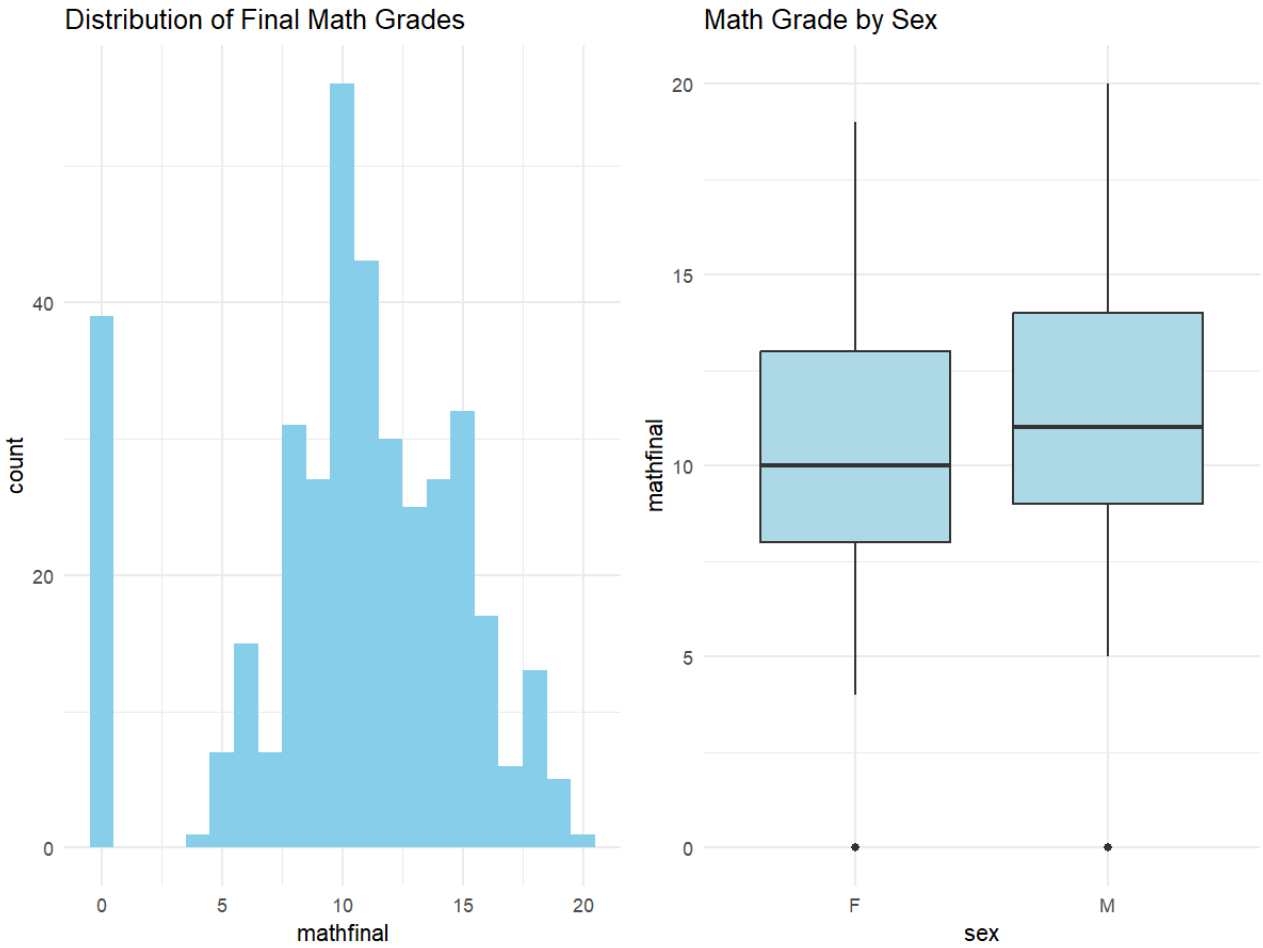

Figure: Histogram of Math Final Grades

Math EDA

Vincent Johnson, Joseph Kang, Catherine Sung

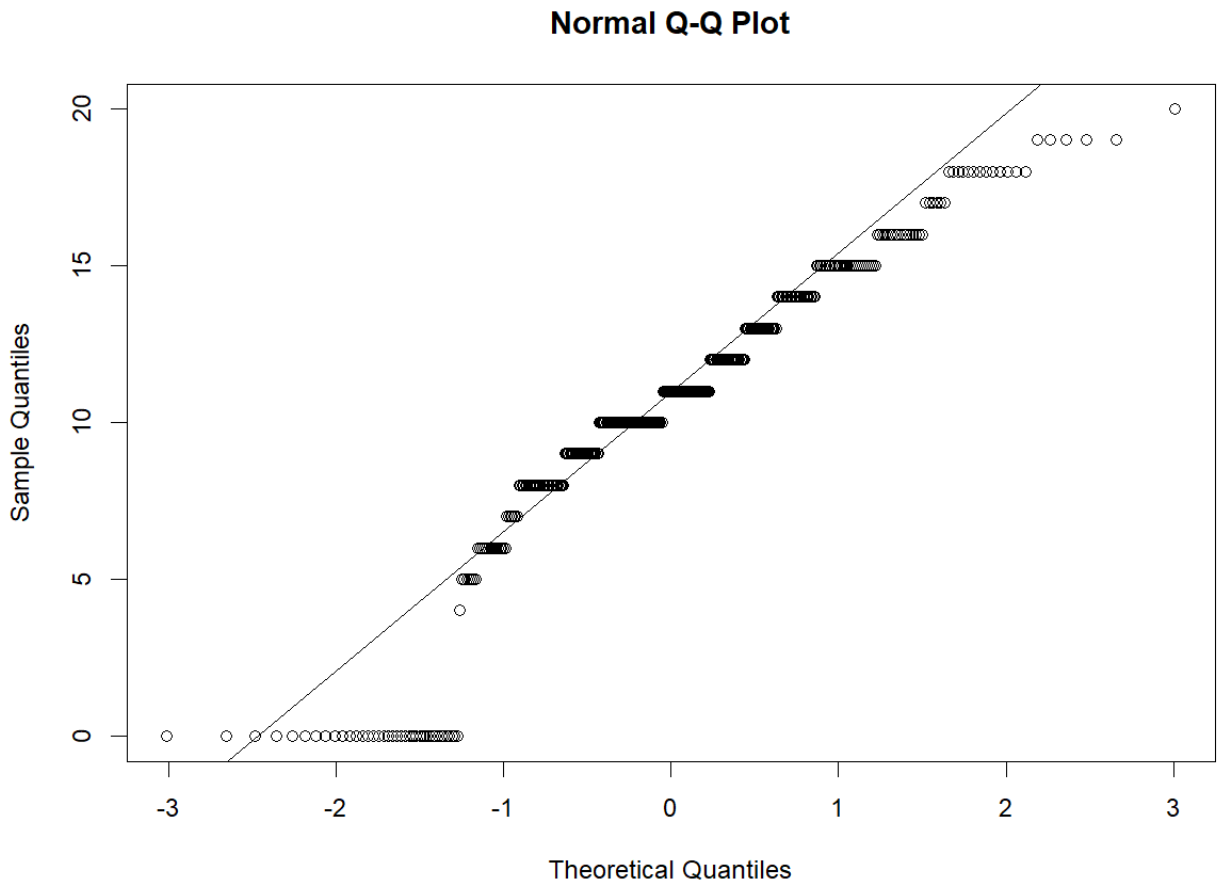

Figure: Linearity Check

Math EDA

Vincent Johnson, Joseph Kang, Catherine Sung

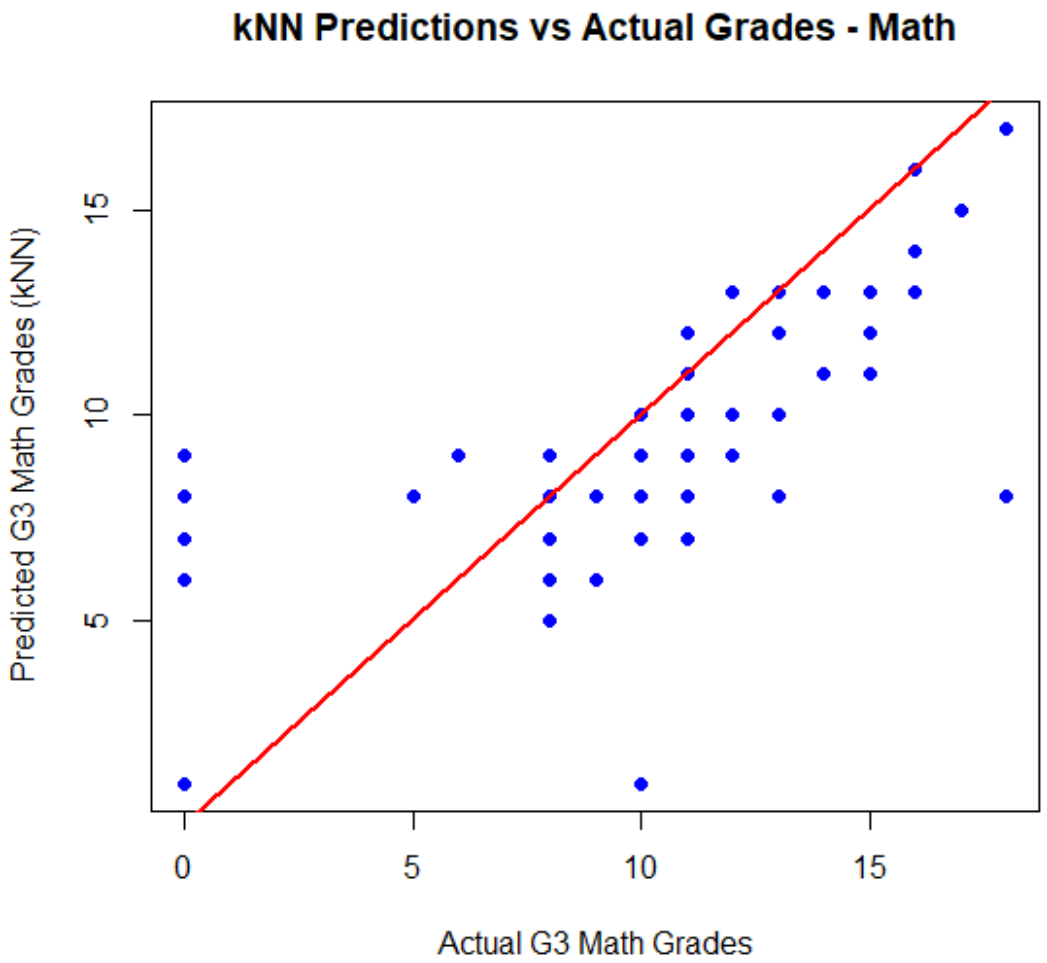

Figure: kNN Predictions vs Actual Grades

Math EDA

Vincent Johnson, Joseph Kang, Catherine Sung

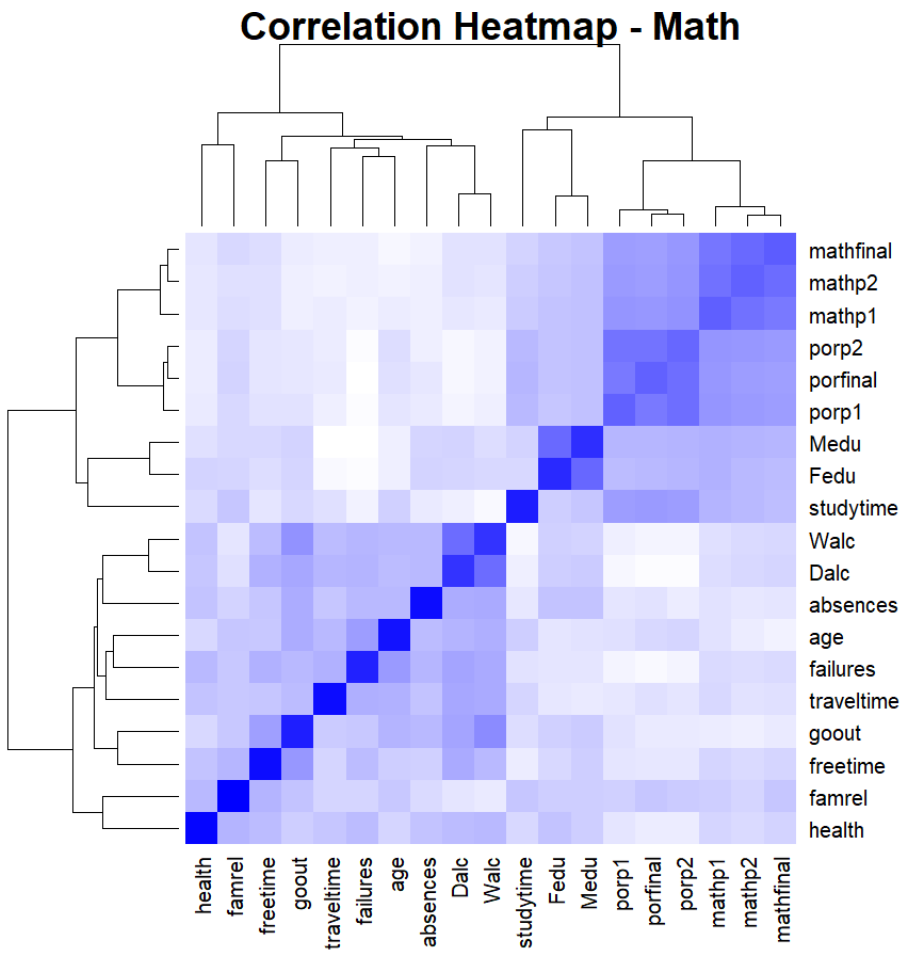

Figure: Heat Map for Correlation

Math EDA

Vincent Johnson, Joseph Kang, Catherine Sung

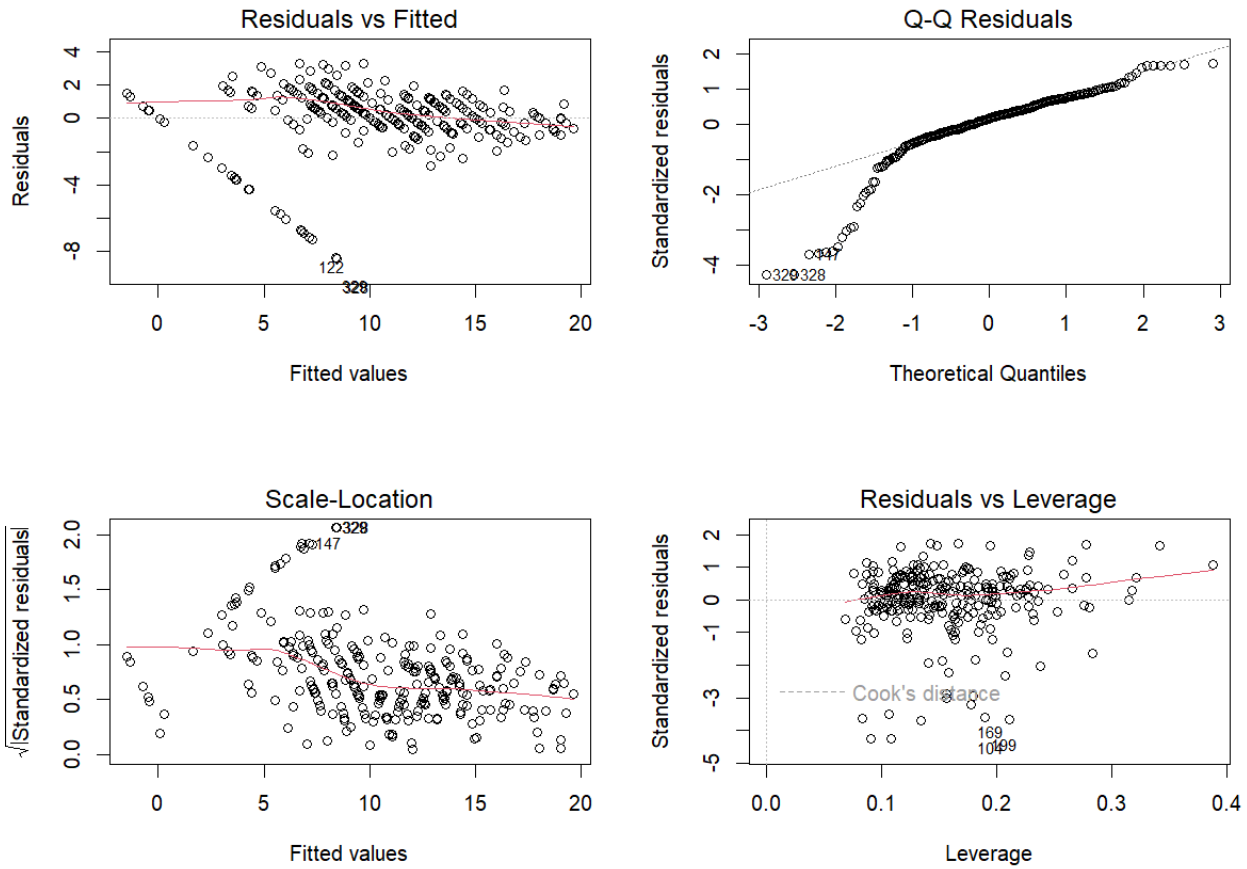

Figure: Assumptions check

Math EDA

Vincent Johnson, Joseph Kang, Catherine Sung

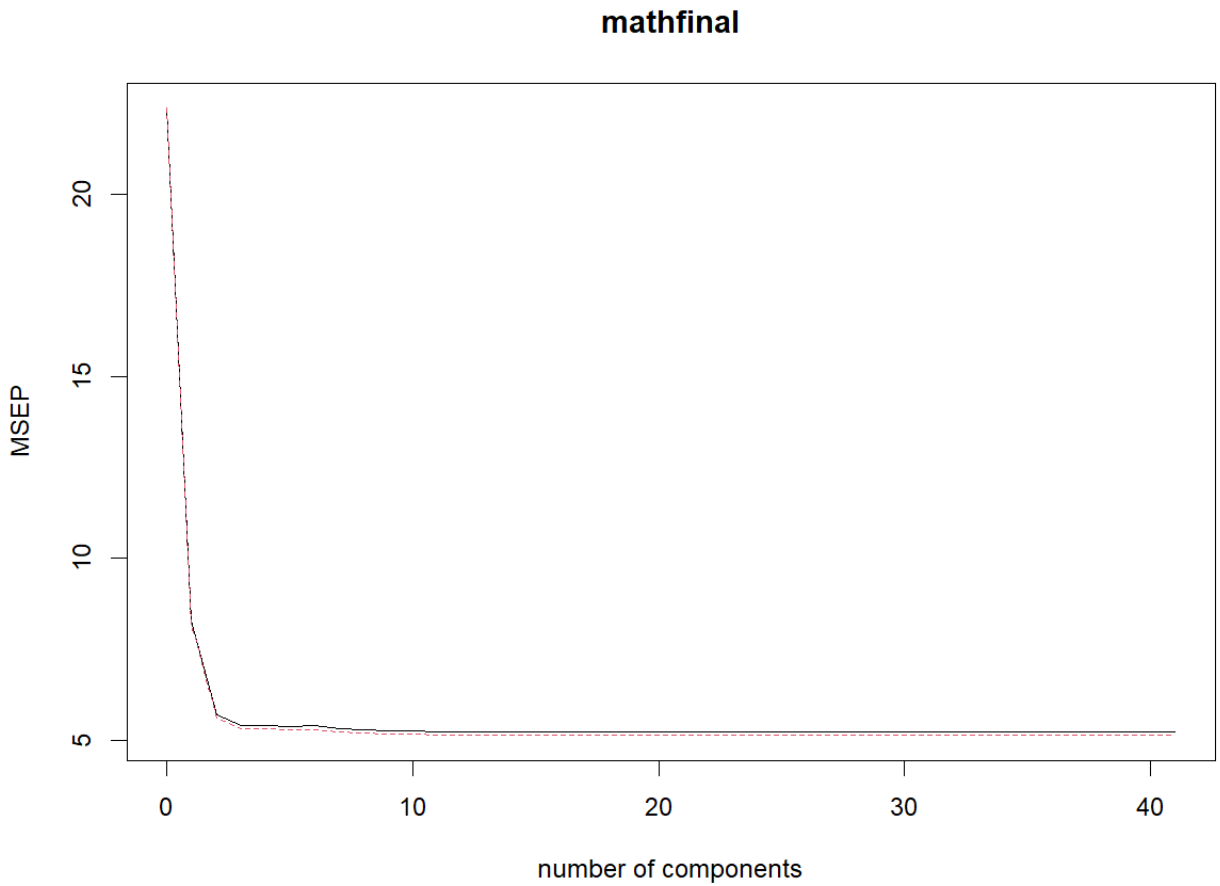

Figure: Partial least Squares Regression

Math EDA

Vincent Johnson, Joseph Kang, Catherine Sung

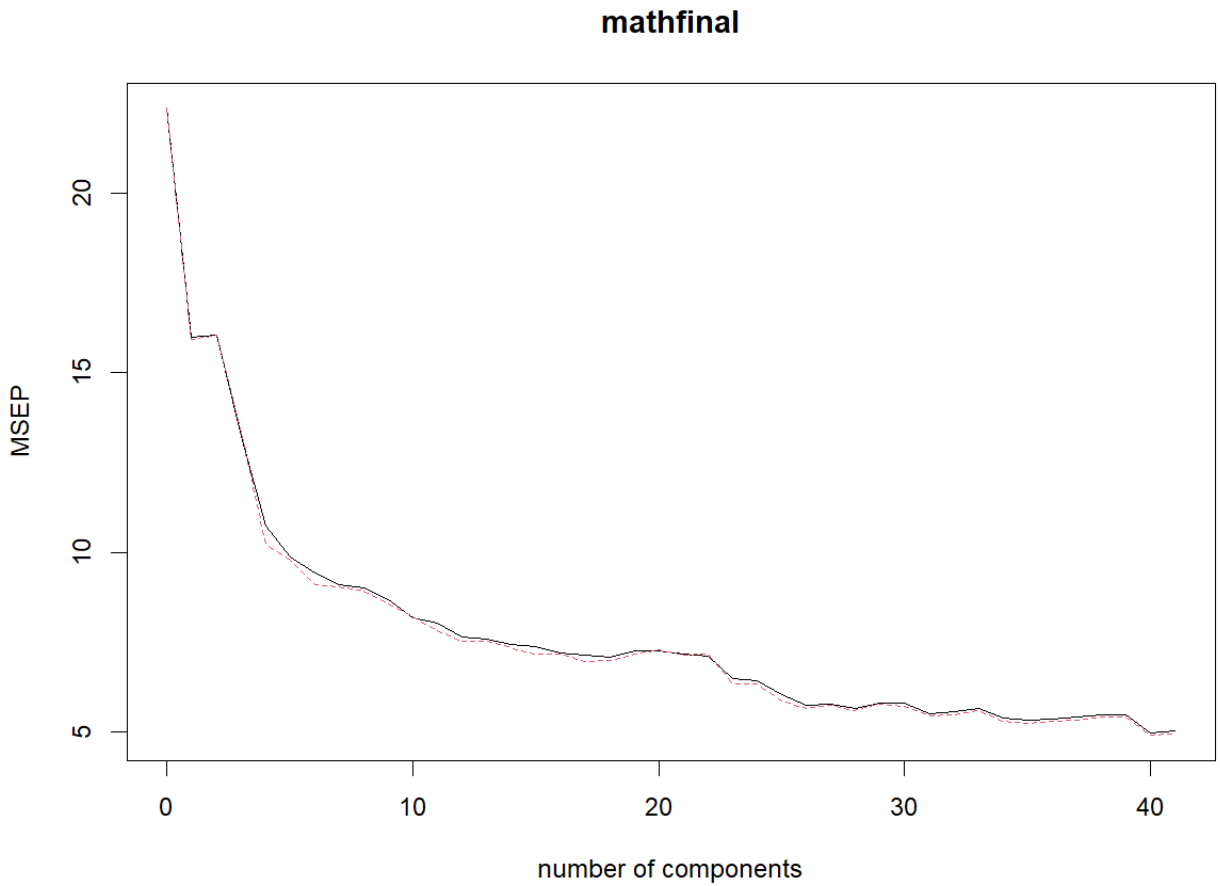

Figure: Partial Component Regression

Math EDA

Vincent Johnson, Joseph Kang, Catherine Sung

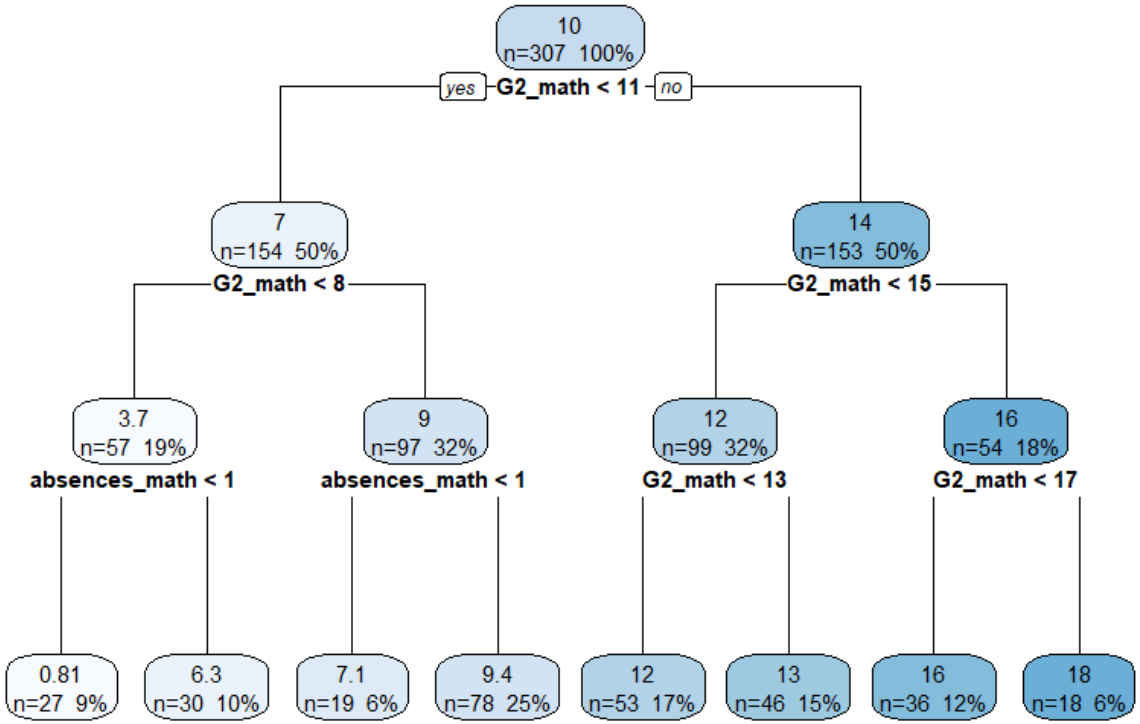

Figure: Decision Tree

Math EDA

Significant Predictors for Final Grades

Vincent Johnson, Joseph Kang, Catherine Sung

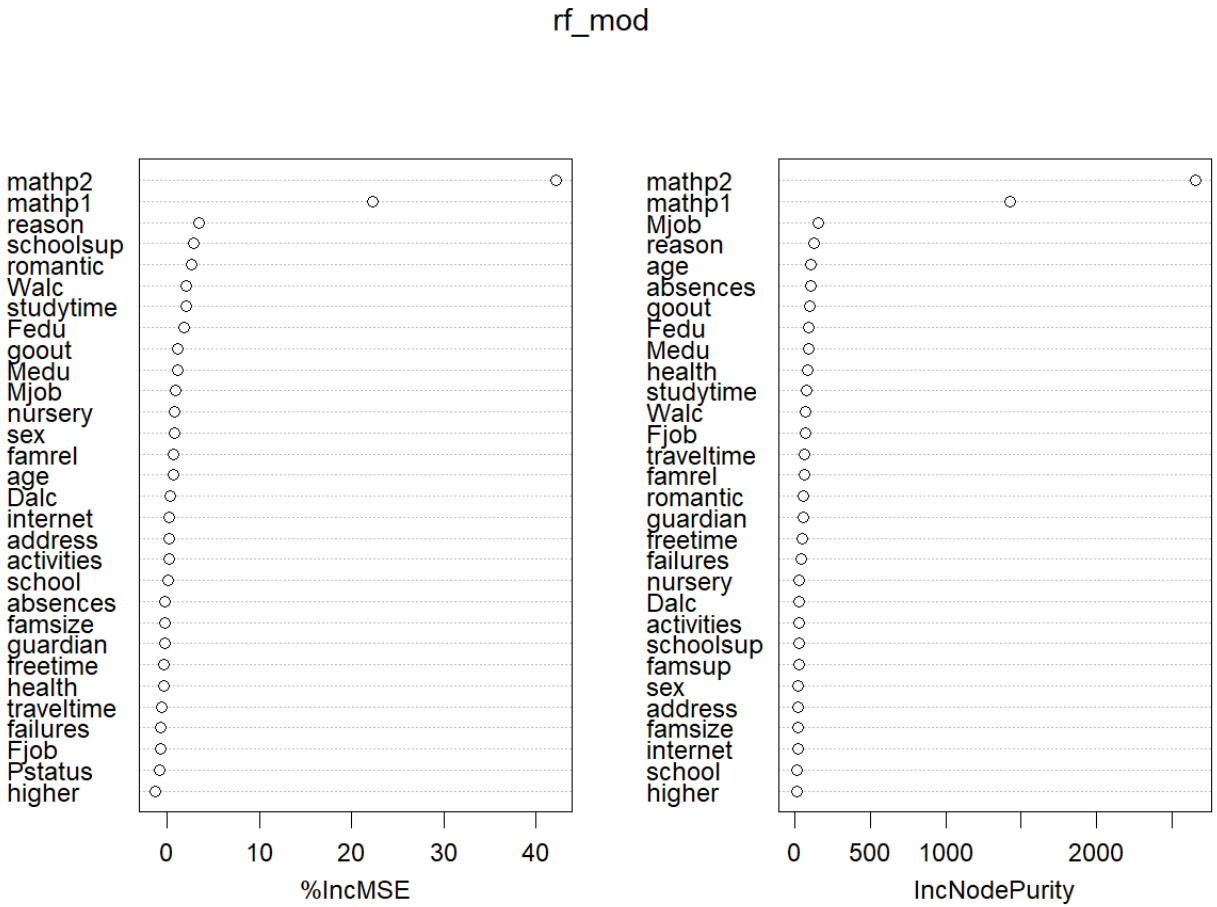

Figure: Variable Importance

Math Model Results

Vincent Johnson, Joseph Kang, Catherine Sung

Model	**RMSE**
Linear Model	1.892848
Stepwise AIC	1.836098
Stepwise BIC	1.867317
Forward Selection	1.836098
Backward Selection	1.84349
Lasso	1.846767
PCR	2.986807
PLS	1.946433
Bagging	1.883409
Boosting	2.20185
kNN Predictions	3.647830
Random Forest	2.097384
SVM	2.19725

Table: Math Model Performance: RMSE for Different Models

Portuguese EDA

Vincent Johnson, Joseph Kang, Catherine Sung

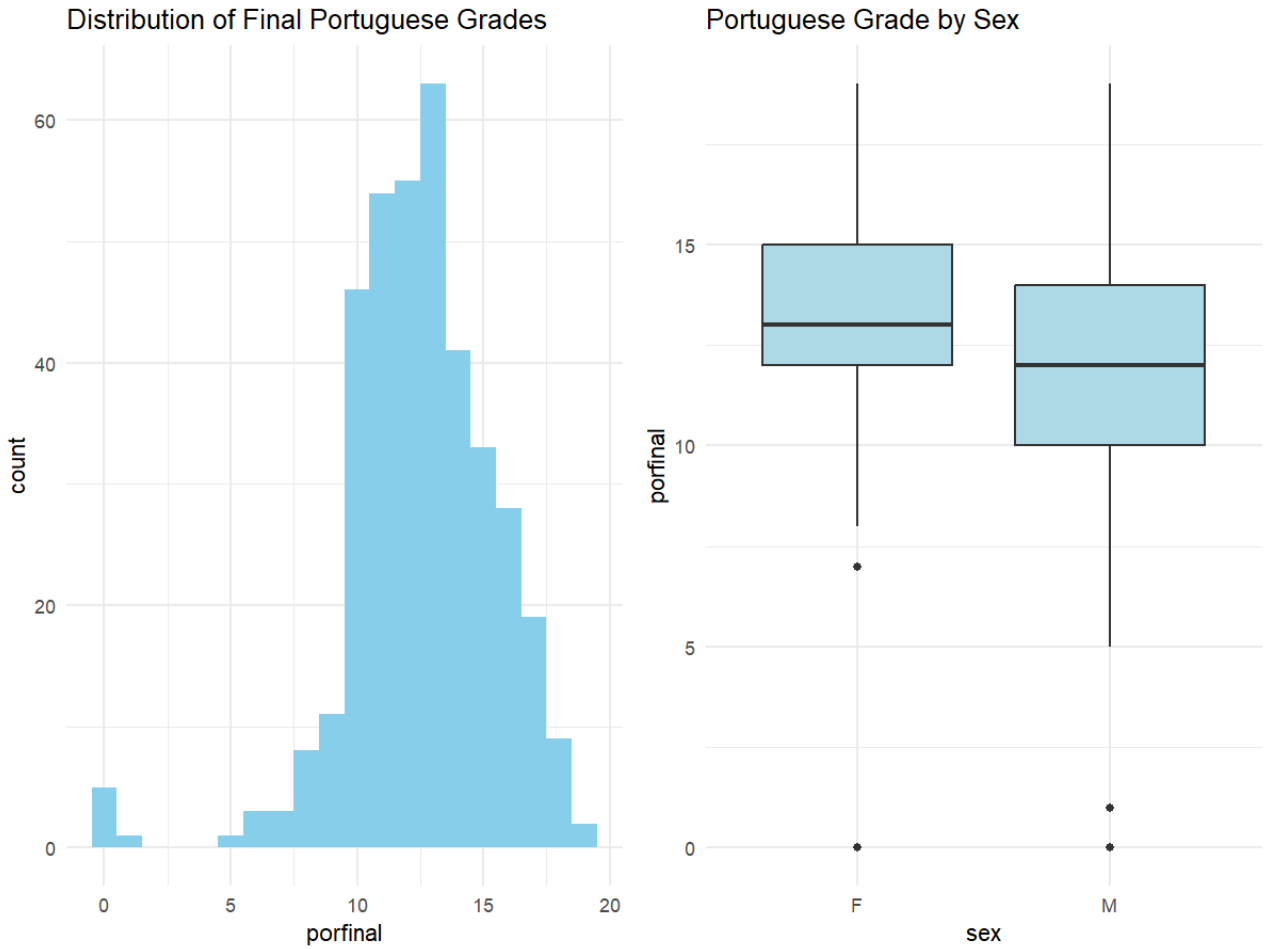

Figure: Histogram of Portuguese Final Grades

Portuguese EDA

Vincent Johnson, Joseph Kang, Catherine Sung

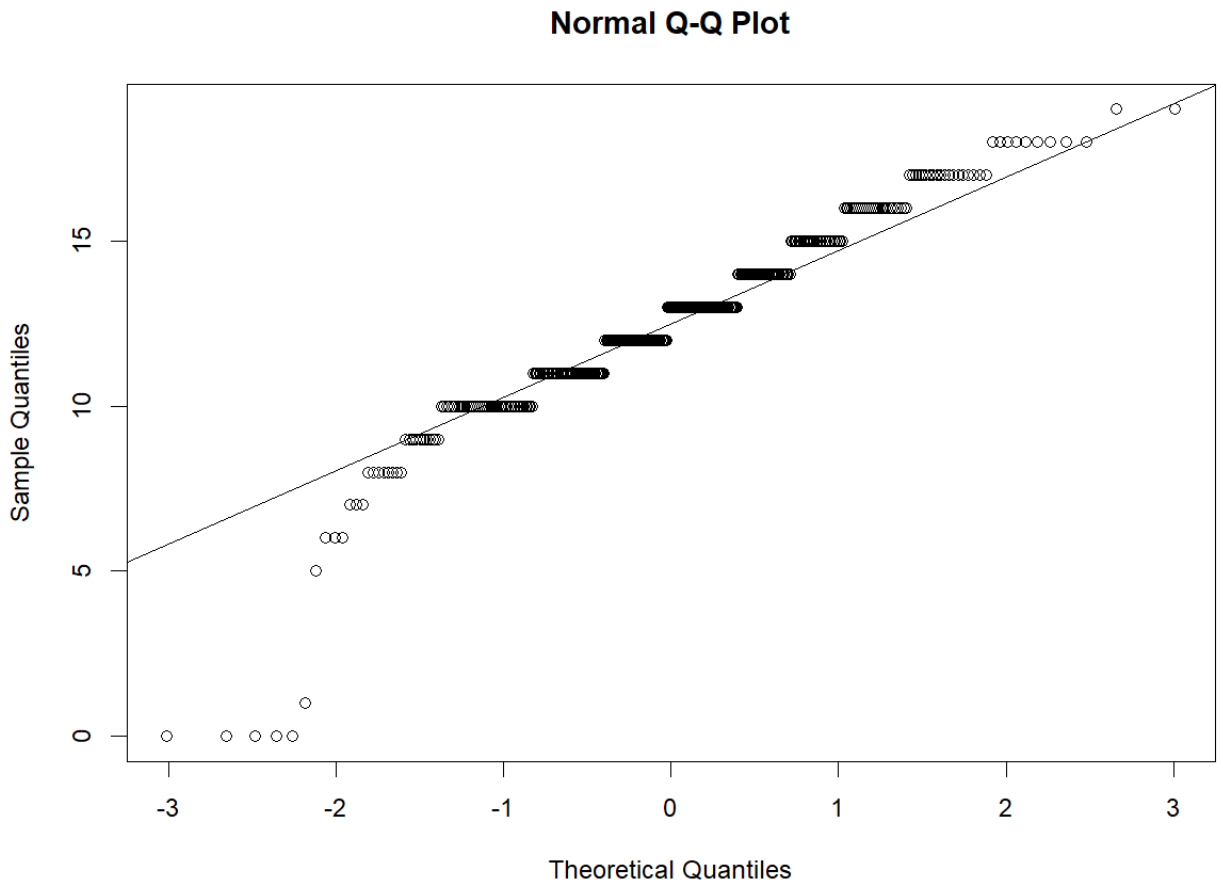

Figure: Portuguese assumption Check Final Grades

Portuguese EDA

Vincent Johnson, Joseph Kang, Catherine Sung

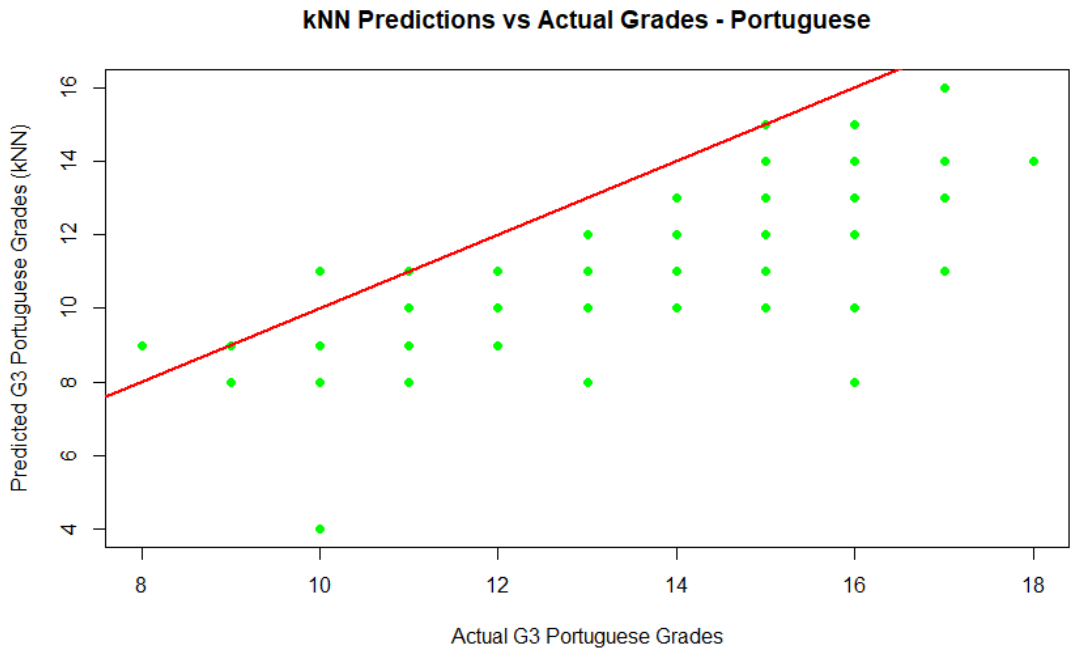

Figure: kNN Predictions vs Actual Grades

Portuguese EDA

Vincent Johnson, Joseph Kang, Catherine Sung

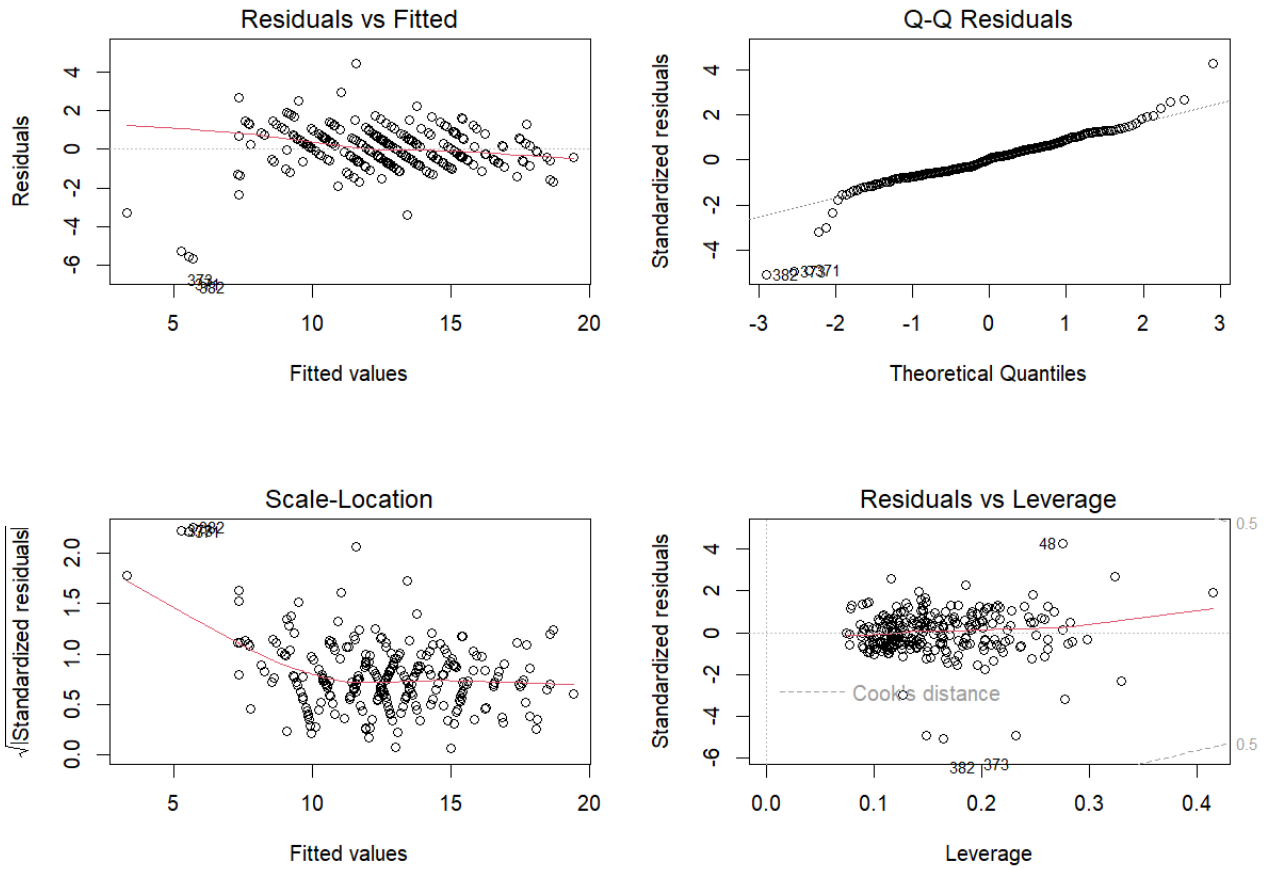

Figure: Portuguese Linearity Assumption

Portuguese EDA

Vincent Johnson, Joseph Kang, Catherine Sung

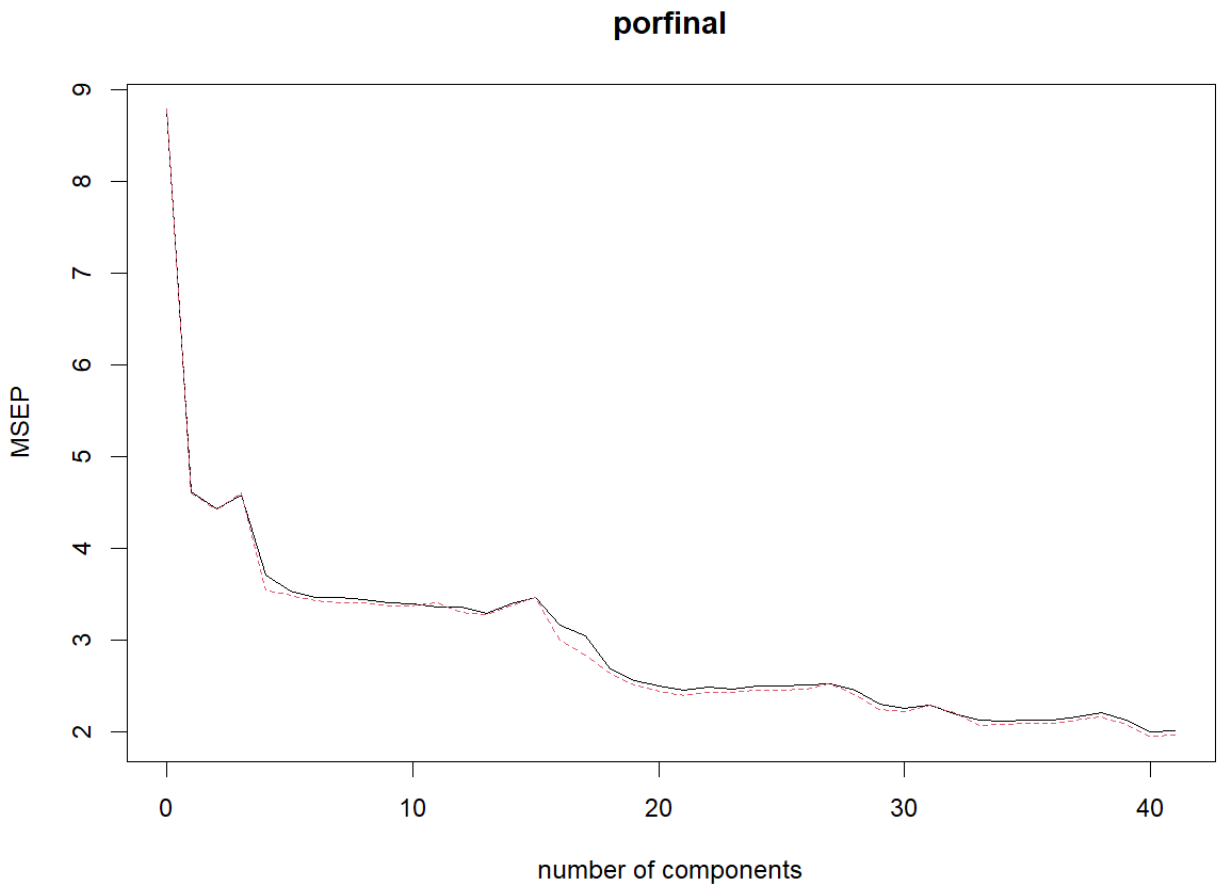

Figure: Partial Component Regression

Portuguese EDA

Vincent Johnson, Joseph Kang, Catherine Sung

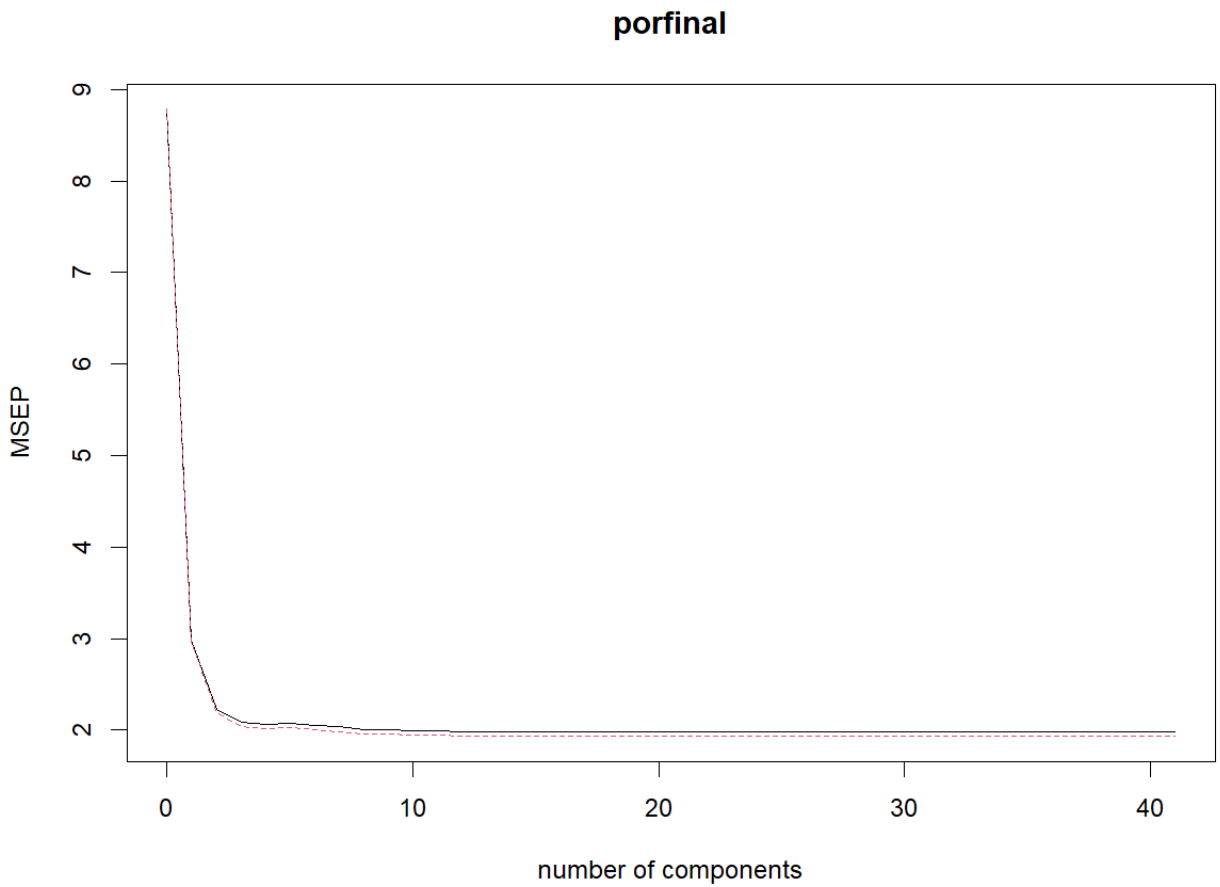

Figure: Partial Least Squares Regression

Portuguese EDA

Vincent Johnson, Joseph Kang, Catherine Sung

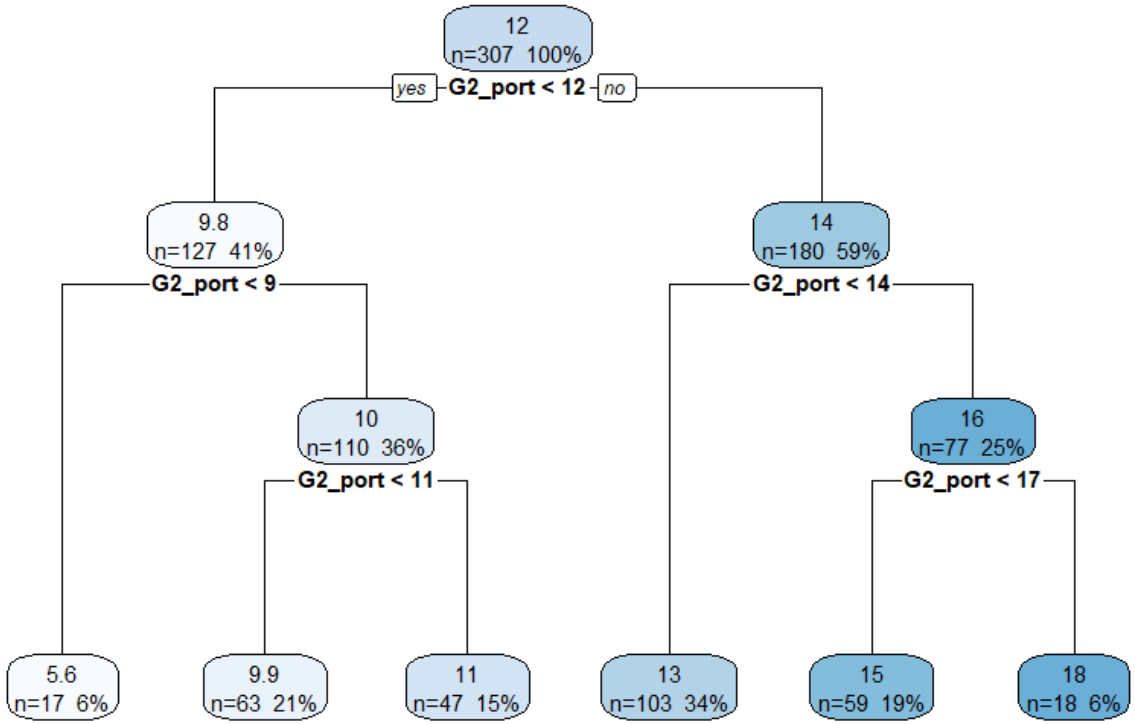

Figure: Decision Tree

Portuguese EDA

Vincent Johnson, Joseph Kang, Catherine Sung

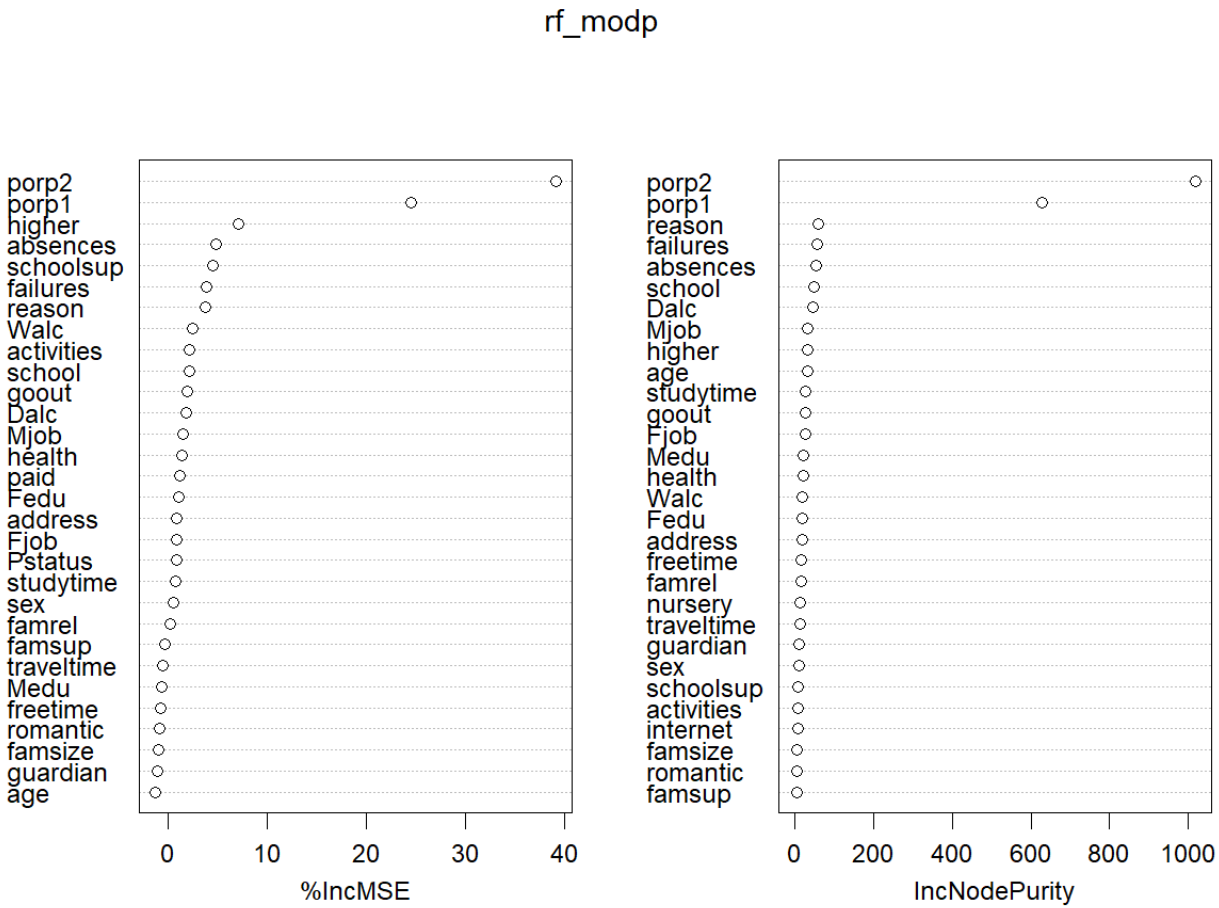

Figure: Variable Importance

Portuguese Model Results

Vincent Johnson, Joseph Kang, Catherine Sung

Model	**RMSE**
Linear Model	1.537861
Stepwise AIC	1.538056
Stepwise BIC	1.525514
Forward Selection	1.538056
Backward Selection	1.538056
Lasso	1.503412
PCR	1.721132
PLS	1.575459
Bagging	1.570195
Boosting	1.811127
kNN Predictions	3.0243457
Random Forest	1.519134
SVM	1.528807

Table: Portuguese Model Performance: RMSE for Different Models

Conclusion

Vincent Johnson, Joseph Kang, Catherine Sung

Math

Model	RMSE
Linear Model	1.892848
Stepwise AIC	1.836098
Stepwise BIC	1.867317
Forward Selection	1.836098
Backward Selection	1.84349
Lasso	1.846767
PCR	2.986807
PLS	1.946433
Bagging	1.883409
Boosting	2.20185
kNN Predictions	3.647830
Random Forest	2.097384
SVM	2.19725

Portuguese

Model	RMSE
Linear Model	1.537861
Stepwise AIC	1.538056
Stepwise BIC	1.525514
Forward Selection	1.538056
Backward Selection	1.538056
Lasso	1.503412
PCR	1.721132
PLS	1.575459
Bagging	1.570195
Boosting	1.811127
kNN Predictions	3.0243457
Random Forest	1.519134
SVM	1.528807

Dataset	Best Model	Best RMSE
Math	Stepwise AIC / Forward Selection	1.836098
Portuguese	Lasso	1.503412

Questions and Discussion

- Does anyone have any questions they would like to ask?