Predicting Student Performance:

An Application of Machine Learning

Minh Hoang Le

October 10, 2020

I. Introduction

Education is the passport to the future. My goal is to help educational institutions, and parents to improve the student academic performance. The problems that I want to solve from this project is to find out which attributes, including student grades, demographic, social, and school-related features, have an effect on high school student performance and from those effective attributes, predicting the student performance. Based on my analysis, schools and parents will know which factors and reasons affect student performance. With that knowledge, they can help the students to increase their performance, preparing them for the future.

II. Description of the dataset

The data that I am using was collected by using school reports and questionnaires. The data include two datasets (CSV files), one is the performance in mathematics, and the other is the performance in Portuguese. The data are acquired from the UCI Machine Learning Repository (Cortez).

The first step is to find unique values of each attribute (feature) for both datasets (Mathematics and Portuguese) to figure out whether it has missing values or not. There are 32 attributes in both datasets. Most of the features are categorical (nominal and ordinal) data, except absences (number of school absences), G1 (first-term grade), G2 (second-term grade), and G3 (final grade, also the output target).

```
In [5]: # Looking for missing data (Mathematics)
        for i in mat.columns:
            print(i,mat[i].unique())
```

```
school ['GP' 'MS']
sex ['F' 'M']
age [18 17 15 16 19 22 20 21]
address ['U' 'R']
famsize ['GT3' 'LE3']
Pstatus ['A' 'T']
Medu [4 1 3 2 0]
Fedu [4 1 2 3 0]
Mjob ['at_home' 'health' 'other' 'services' 'teacher']
Fjob ['teacher' 'other' 'services' 'health' 'at_home']
reason ['course' 'other' 'home' 'reputation']
guardian ['mother' 'father' 'other']
traveltime [2 1 3 4]
studytime [2 3 1 4]
failures [0 3 2 1]
schoolsup ['yes' 'no']
famsup ['no' 'yes']
paid ['no' 'yes']
activities ['no' 'yes']
nursery ['yes' 'no']
higher ['yes' 'no']
internet ['no' 'yes']
romantic ['no' 'yes']
famrel [4 5 3 1 2]
freetime [3 2 4 1 5]
goout [4 3 2 1 5]
Dalc [1 2 5 3 4]
Walc [1 3 2 4 5]
health [3 5 1 2 4]
absences [ 6  4 10  2  0 16 14  7  8 25 12 54 18 26 20 56 24 28  5 13 15 22  3 21
  1 75 30 19  9 11 38 40 23 17]
G1 [ 5  7 15  6 12 16 14 10 13  8 11  9 17 19 18  4  3]
G2 [ 6  5  8 14 10 15 12 18 16 13  9 11  7 19 17  4  0]
G3 [ 6 10 15 11 19  9 12 14 16  5  8 17 18 13 20  7  0  4]
```

Figure 1: Mathematics dataset

```
In [6]: # Looking for missing data (Portuguese)
        for i in por.columns:
            print(i,mat[i].unique())
```

```
school ['GP' 'MS']
sex ['F' 'M']
age [18 17 15 16 19 22 20 21]
address ['U' 'R']
famsize ['GT3' 'LE3']
Pstatus ['A' 'T']
Medu [4 1 3 2 0]
Fedu [4 1 2 3 0]
Mjob ['at_home' 'health' 'other' 'services' 'teacher']
Fjob ['teacher' 'other' 'services' 'health' 'at_home']
reason ['course' 'other' 'home' 'reputation']
guardian ['mother' 'father' 'other']
traveltime [2 1 3 4]
studytime [2 3 1 4]
failures [0 3 2 1]
schoolsup ['yes' 'no']
famsup ['no' 'yes']
paid ['no' 'yes']
activities ['no' 'yes']
nursery ['yes' 'no']
higher ['yes' 'no']
internet ['no' 'yes']
romantic ['no' 'yes']
famrel [4 5 3 1 2]
freetime [3 2 4 1 5]
goout [4 3 2 1 5]
Dalc [1 2 5 3 4]
Walc [1 3 2 4 5]
health [3 5 1 2 4]
absences [ 6  4 10  2  0 16 14  7  8 25 12 54 18 26 20 56 24 28  5 13 15 22  3 21
  1 75 30 19  9 11 38 40 23 17]
G1 [ 5  7 15  6 12 16 14 10 13  8 11  9 17 19 18  4  3]
G2 [ 6  5  8 14 10 15 12 18 16 13  9 11  7 19 17  4  0]
G3 [ 6 10 15 11 19  9 12 14 16  5  8 17 18 13 20  7  0  4]
```

Figure 2: Portuguese dataset

The next step is to find outlier values of the non-categorical features for both datasets using the z-score method. From the empirical rule, around 99.7% of the z-score will be within -3 and 3. The outlier will have the z-score value greater than 3 or less than -3. I calculate the z-score of each value in each feature and take the absolute value of it. If the value is greater than 3, that value is an outlier. Then I compare the outliers with the range of values of the non-categorical features provided in the dataset description (from the website) to check whether the outliers are due to incorrectly entered or measuring data or not. After close examination, there is no missing value in both datasets. However, there were outliers in both datasets. Based on the dataset description, since none of the outliers are due to incorrectly entered or measuring data, the outliers are kept in the datasets.

III. Initial findings from exploratory analysis

The goal of my project is to figure out which attributes, including student grades, demographic, social, and school-related features, have an effect on high school student performance and, from those attributes, predict the student performance (final grade). To see correlations between pairs of independent variables and between an independent and a dependent variable, I use correlation matrices using Pearson parametric correlation test. For both datasets, the correlation matrices show a moderate positive relationship between mother's education and father's education, Pearson correlation coefficients are around 0.6. Between workday alcohol consumption and weekend alcohol consumption of students, Pearson correlation coefficients are about 0.6. However, the correlation matrices also show that there is a strong positive linear relationship between first-term grade and second-term grade, and between the term grade (both terms) and final grade. For both datasets, the relationship between the second term grade and the

final grade is slightly stronger than between the first term grade and final grade (Figure 3 and Figure 4).

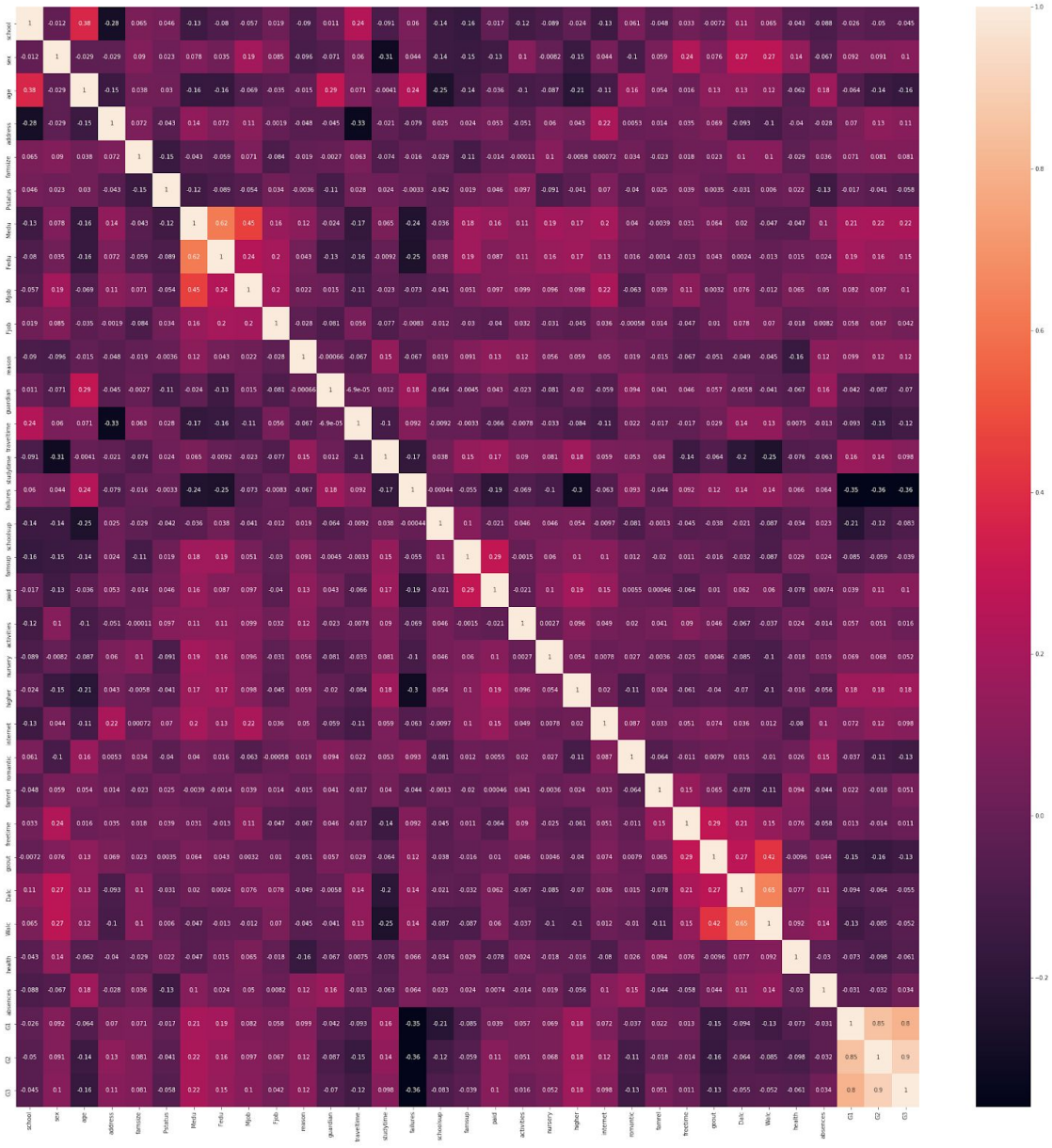

Figure 3: Correlation matrix heatmap - Mathematics

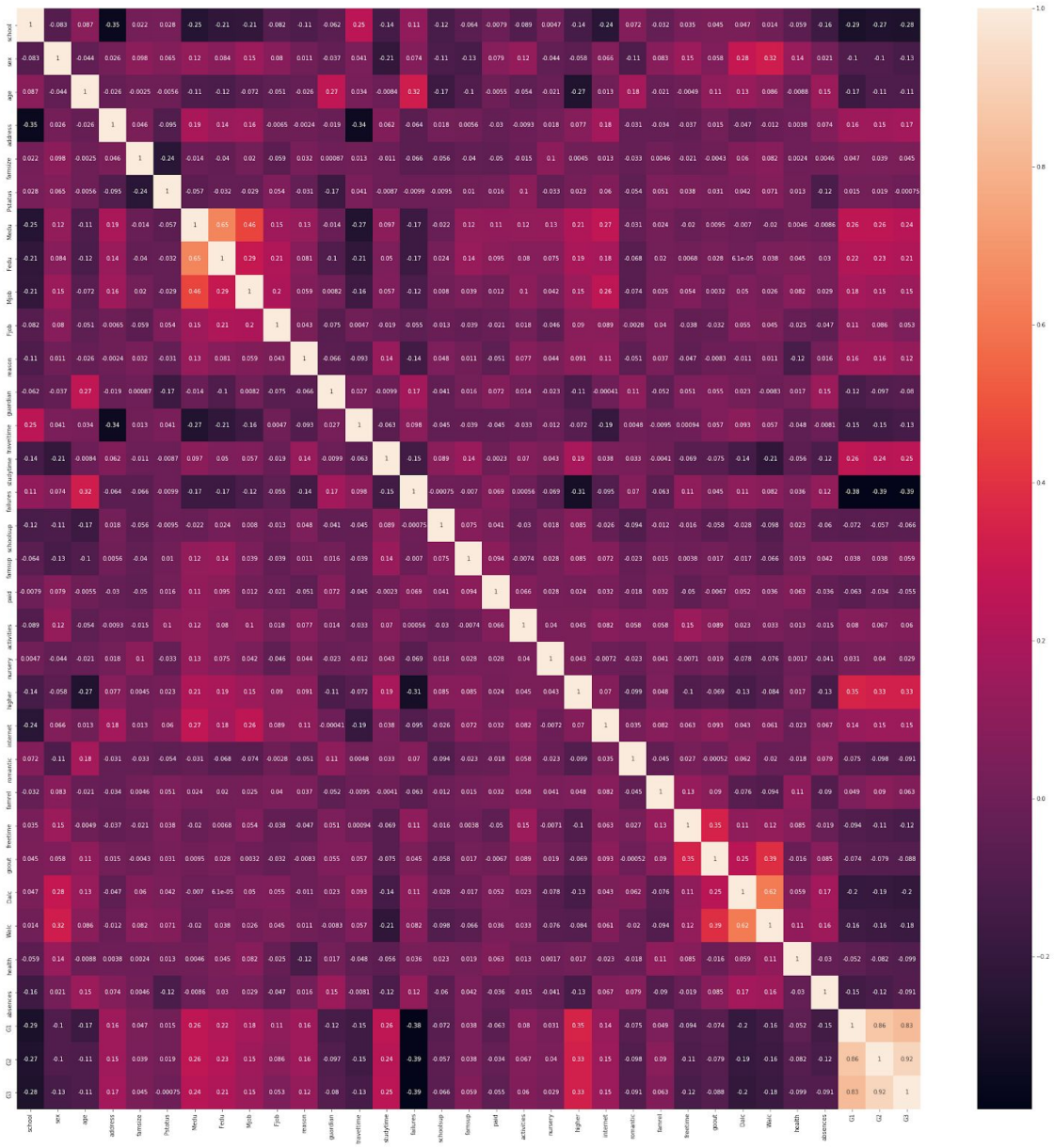

Figure 4: Correlation matrix heatmap - Mathematics

To see the relationship of the final grade between subgroups of each attribute, I compare the mean of the final grade of each subgroup in the same attribute to each other. To test the significant differences of the mean, I use one-tailed t-test for the means of two independent samples with the significance level (alpha) equals 0.5 and with the assumption that the population means from the two groups are not equal. For gender, the average mathematics final

grades of male students are higher than those of female students. In contrast, the average Portuguese final grades of female students are higher than those of male students (Figure 5).

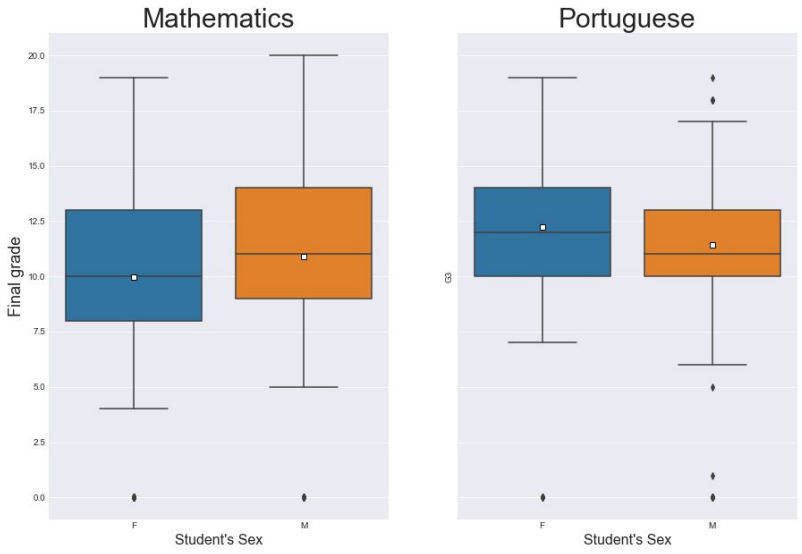

Figure 5: Student Gender and Final Grade

For wanting higher education, the average final grades in both mathematics and Portuguese of students who want higher education are higher than those of students who do not want a higher education (Figure 6).

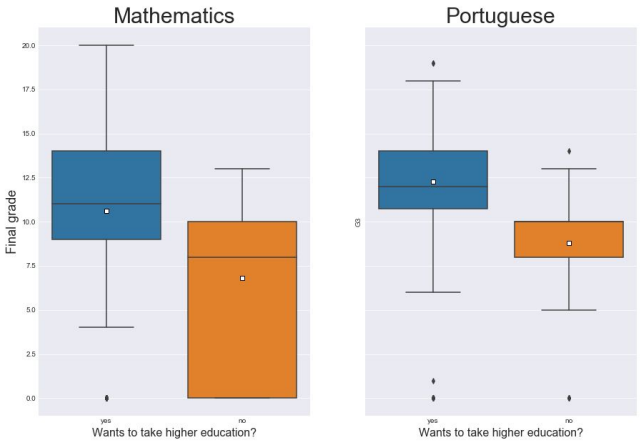

Figure 6: Wanting Higher Education and Final Grade

For extra educational support, the average final grades in both mathematics and Portuguese of students who have extra educational support are lower than those who do not have (Figure 7).

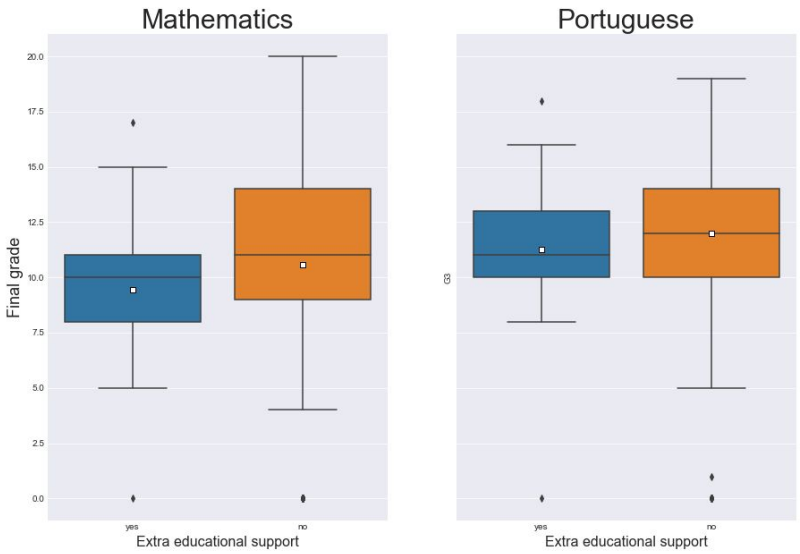

Figure 7: Extra Educational Support and Final Grade

For time spending on studying mathematics, surprisingly, there is no significant difference in the average final grade of students who spend a lot of time studying (more than 10 hours) and students who don't spend much (less than 2 hours). However, for Portuguese, students who spend more time studying (more than 10 hours) have higher average final grades than those who don't (less than 2 hours) (Figure 8).

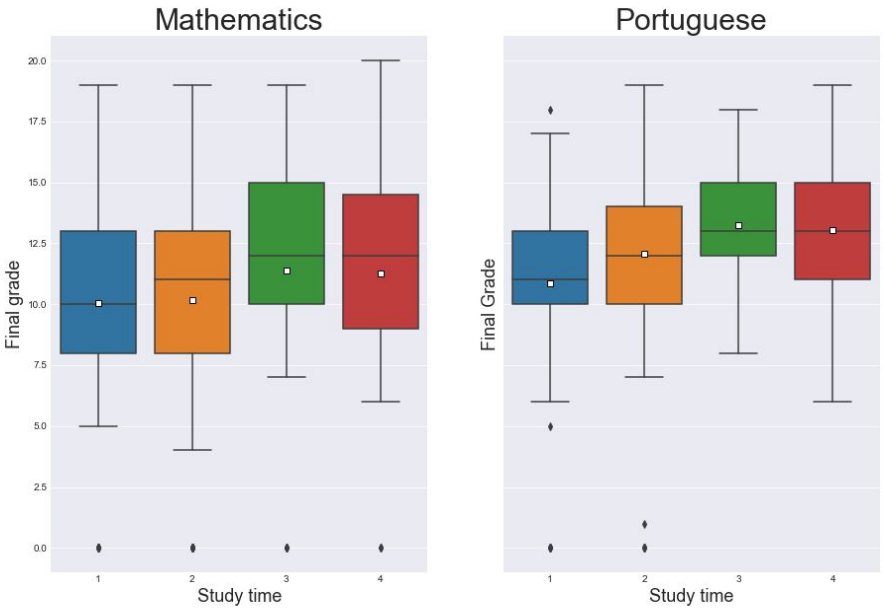

Figure 8: Study Time and Final Grade

For health, surprisingly, students who are the least healthy (1) have higher average final grade final grades in both subjects than students who are the healthiest (5). This can be because healthy students are more likely to do more extracurricular activities than those who are not healthy. Therefore, they will spend less time studying, which can affect their final grades (Figure 9).

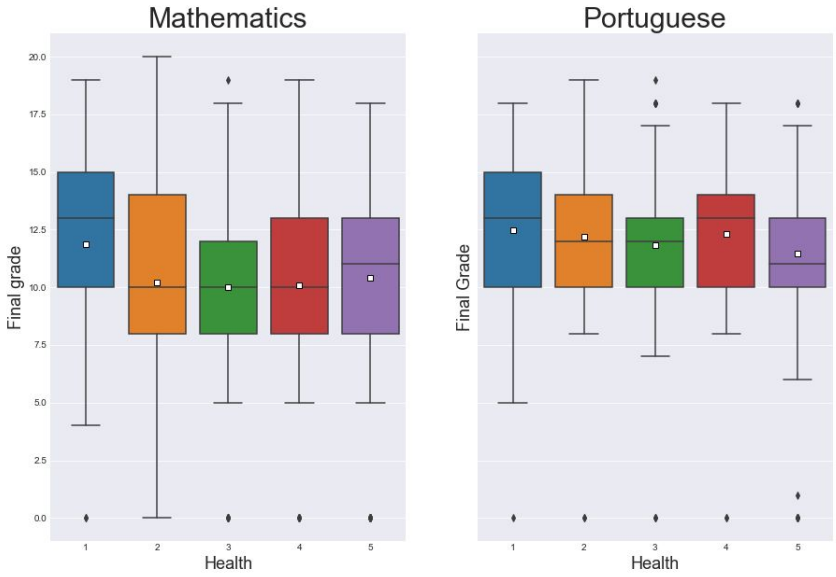

Figure 9: Health and Final Grade

IV. Machine learning

A. Pre-processing

Since both datasets contain nominal categorical features (dummy variable), the first step is to encode them. Next, the only way to know how well a model will generalize to a new case is to actually try it out on a new case. Therefore, before model selecting and training, the first step is randomly splitting data into two subsets: the test set and the training set. The reason behind it is to test how well a model will generalize to new cases. The ratio between training and test sets is 4:1. In other words, the training set is 80 percent of the original data, and the test set is 20

percent of the original data. The training data set is split into 10-fold and cross-validated with each set to avoid over-fitting.

B. Model Selection

For choosing the right types of machine learning, since there is a target variable, the final grade of students, supervised learning is preferred over unsupervised learning. Nine regression models are tested before selecting the best prediction model. They are linear regression, ridge regression, lasso regression, support vector machine-based regression, decision trees regression, random forest regression, gradient boosting regression, XGBoost, and Light Gradient Boosted Machine (LightGBM).

After trained on the training set, each model is evaluated once per validation set. Averaging out all the evaluations of a model gets a more accurate measure of its performance. Coefficient of determination (R-squared) and mean squared error (MSE) are used as evaluation metrics. R-squared is a direct indicator of how good our model is in terms of performance. The R-squared is the measure of the variance in the response variable, target variable, that can be predicted using predictor variables, or features. R-squared is the most common way to measure the strength of the model (Tripathi). The mean squared error tells how close a regression line is to a set of points by taking the distances from the points to the regression line and squaring them. The squaring removes any negative signs and gives more weight to larger differences, or errors (Glen). The top three or four models that have the highest average of R-squared scores and the lowest average of MSE scores are selected for hyperparameter tuning.

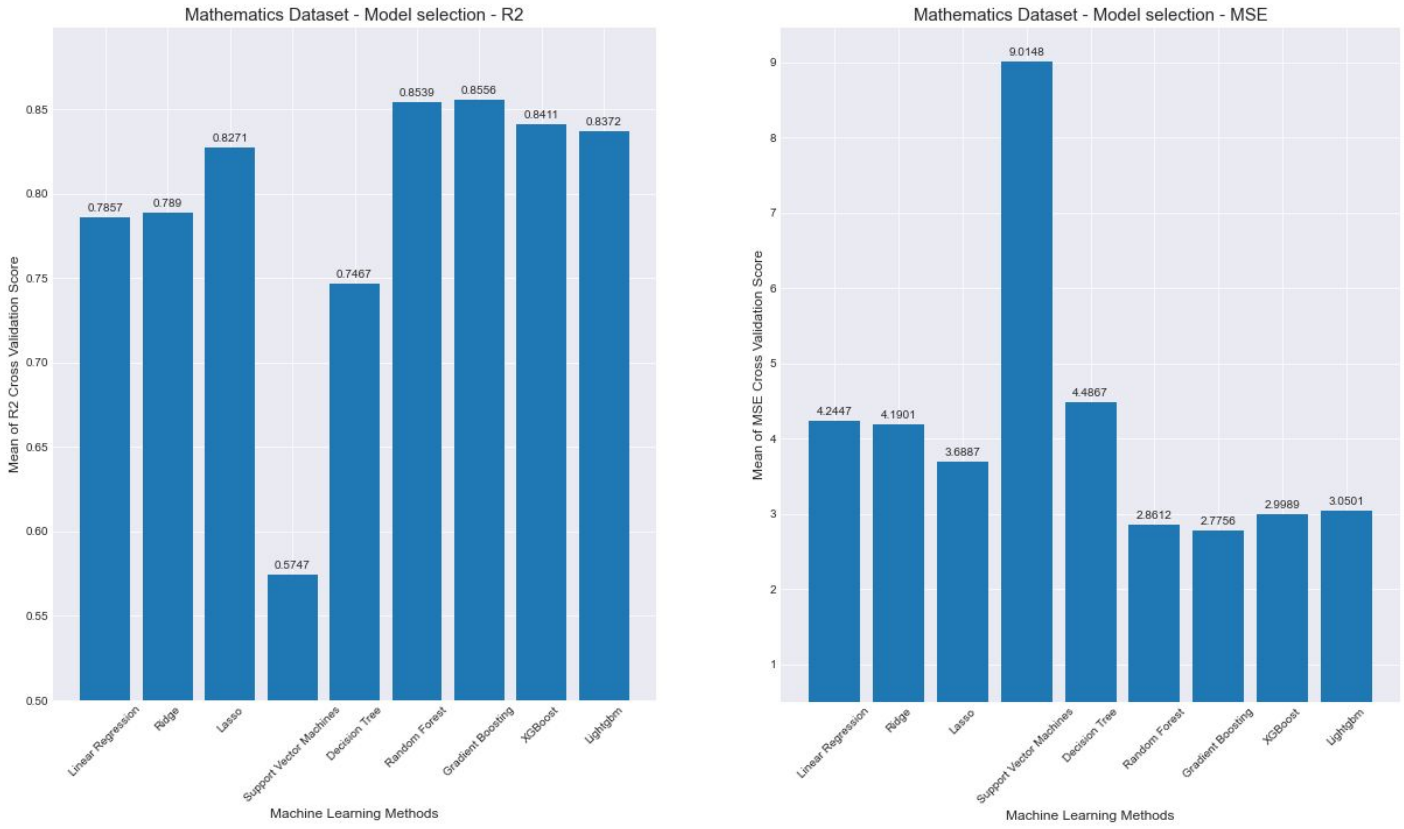

Figure 10: Average of Cross-Validation Score (R-squared and MSE) of Mathematics Dataset

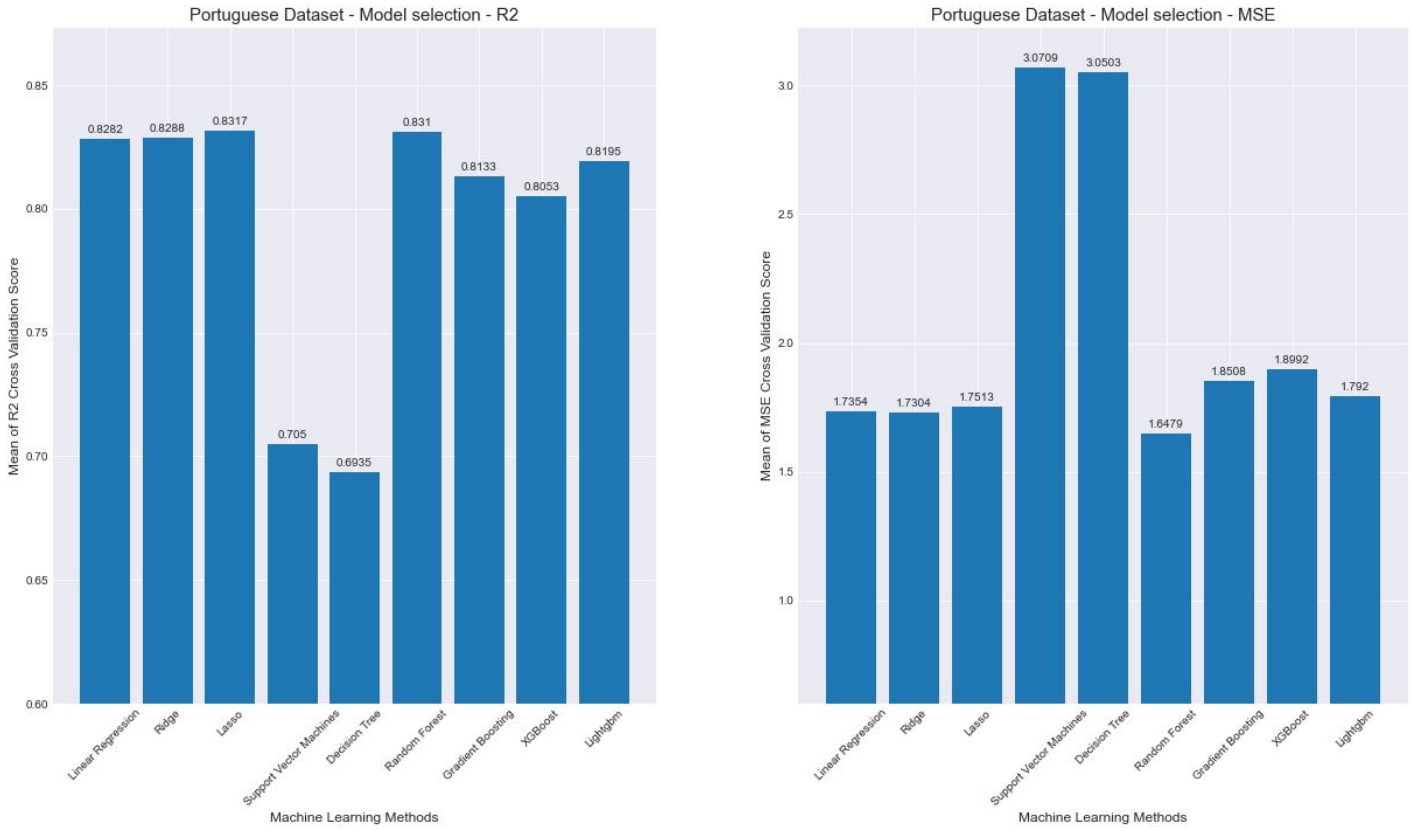

Figure 11: Average of Cross-Validation Score (R-squared and MSE) of Portuguese Dataset

For the mathematics dataset, Random forest, gradient boosting, and XGBoost have the best performances. For the Portuguese dataset, Linear regression, Ridge, Lasso, and random forest regression have the best performances. They all have the highest averages of R-squared cross-validation score and the lowest averages of MSE cross-validation score comparing to other models.

C. Hyperparameter Optimization

After model selection, grid-search cross-validation is used to tune the hyperparameters for the top models.

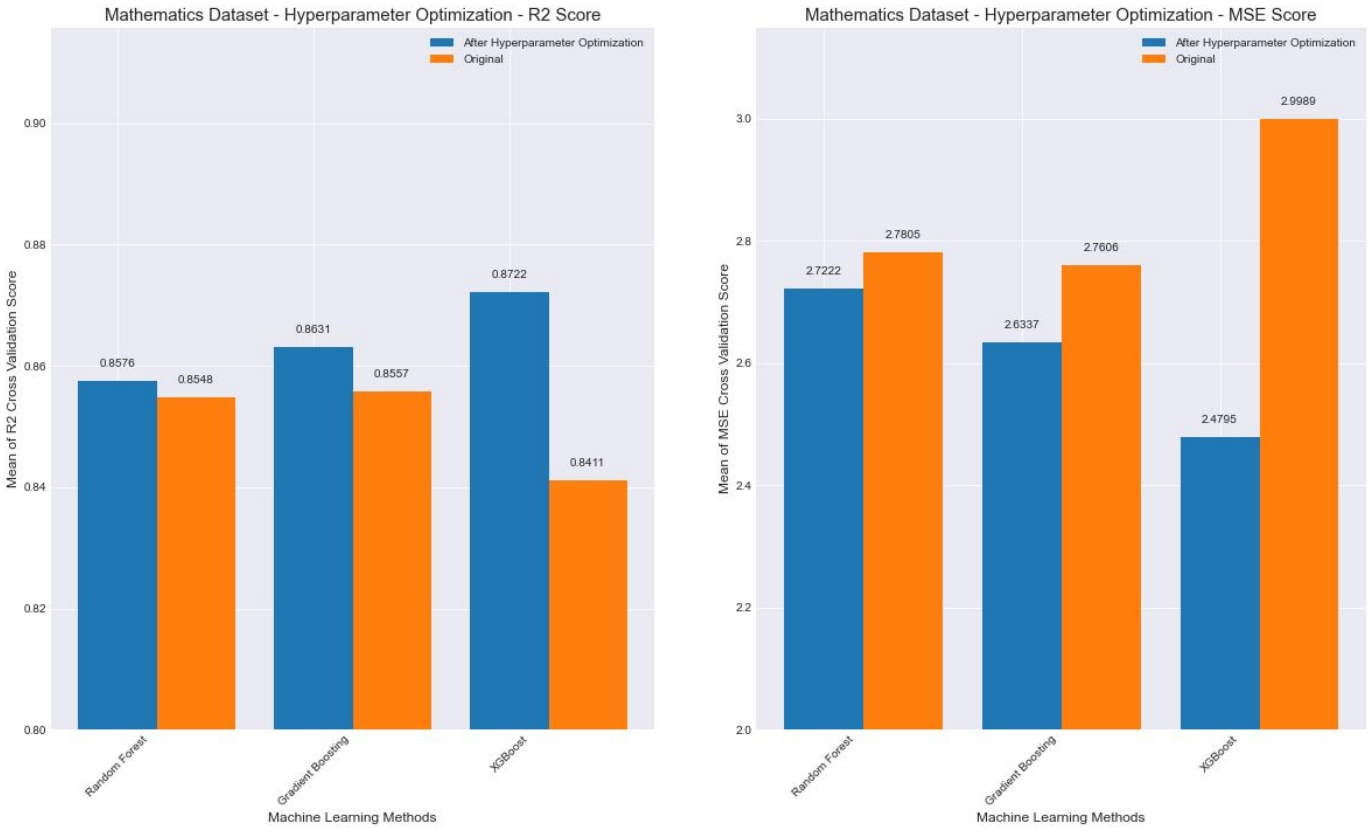

Figure 12: Average of Cross-Validation Score (R-squared and MSE) of Mathematics Dataset - After Hyperparameter Optimization

For mathematics datasets, the XGBoost regression model performs the best. It has the highest average of R-squared cross-validation score (0.8722) and the lowest average of MSE cross-validation score (2.4795). It also has the most improvement after hyperparameter tuning.

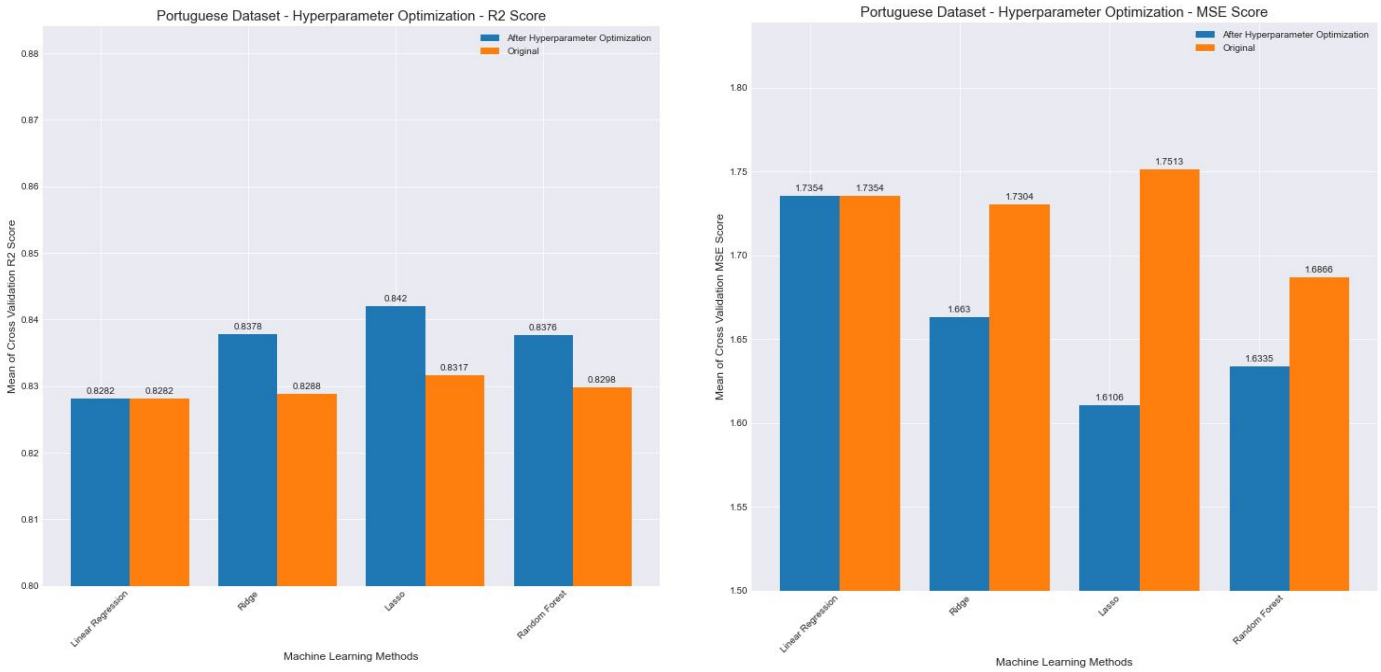

Figure 13: Average of Cross-Validation Score (R-squared and MSE) of Portuguese Dataset - After Hyperparameter Optimization

For the Portuguese dataset, the Lasso regression model performs the best. It has the highest average of R-squared cross-validation score (0.842) and the lowest average of MSE cross-validation score (1.6106).

D. Important Features

In the mathematics dataset, for the XGBoost model, the gain is used to determine the importance of the feature. Gain implies the relative contribution of the corresponding feature to the model calculated by taking each feature's contribution for each tree in the model. A higher value of this metric when compared to another feature implies it is more important for generating

a prediction. Second term grade (G2) is the most important feature, followed by absences, study time.

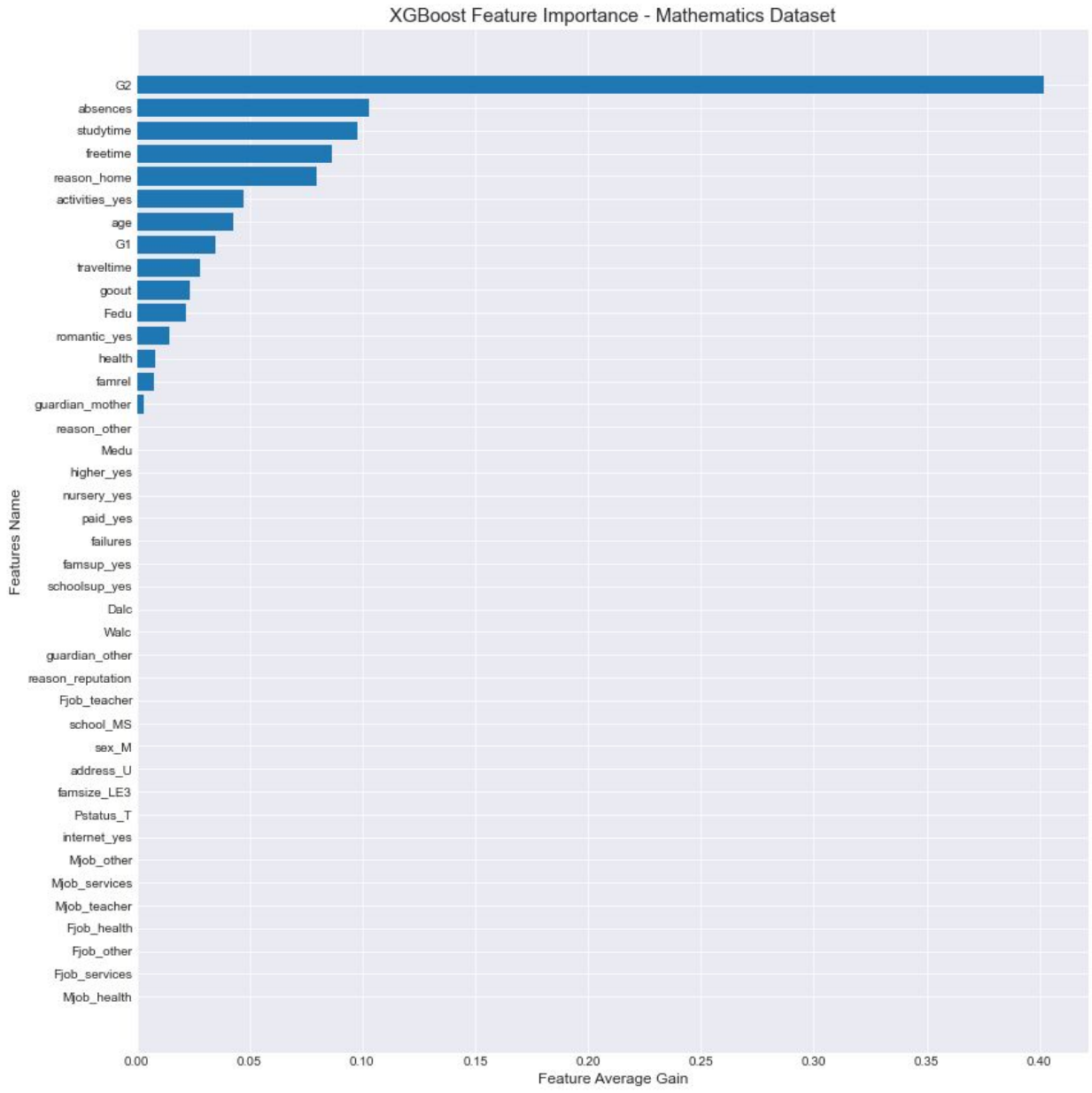

Figure 14: Feature Importance - Mathematics Dataset - XGBoost model

In the Portuguese datasets, for Lasso regression, the coefficient value of each feature signifies how much the mean of the target variable changes given a one-unit shift in the feature variable

while holding other variables in the model constant. The more important features will have higher coefficients. Most of the features do not have correlations with the final grade, having zero for the coefficients. Second term grade has the highest coefficient (0.85399). In other words, if the second term grade increase by 1 point, the average final grade will increase by 0.85399 points. Second term grade (G2) is the most important feature.

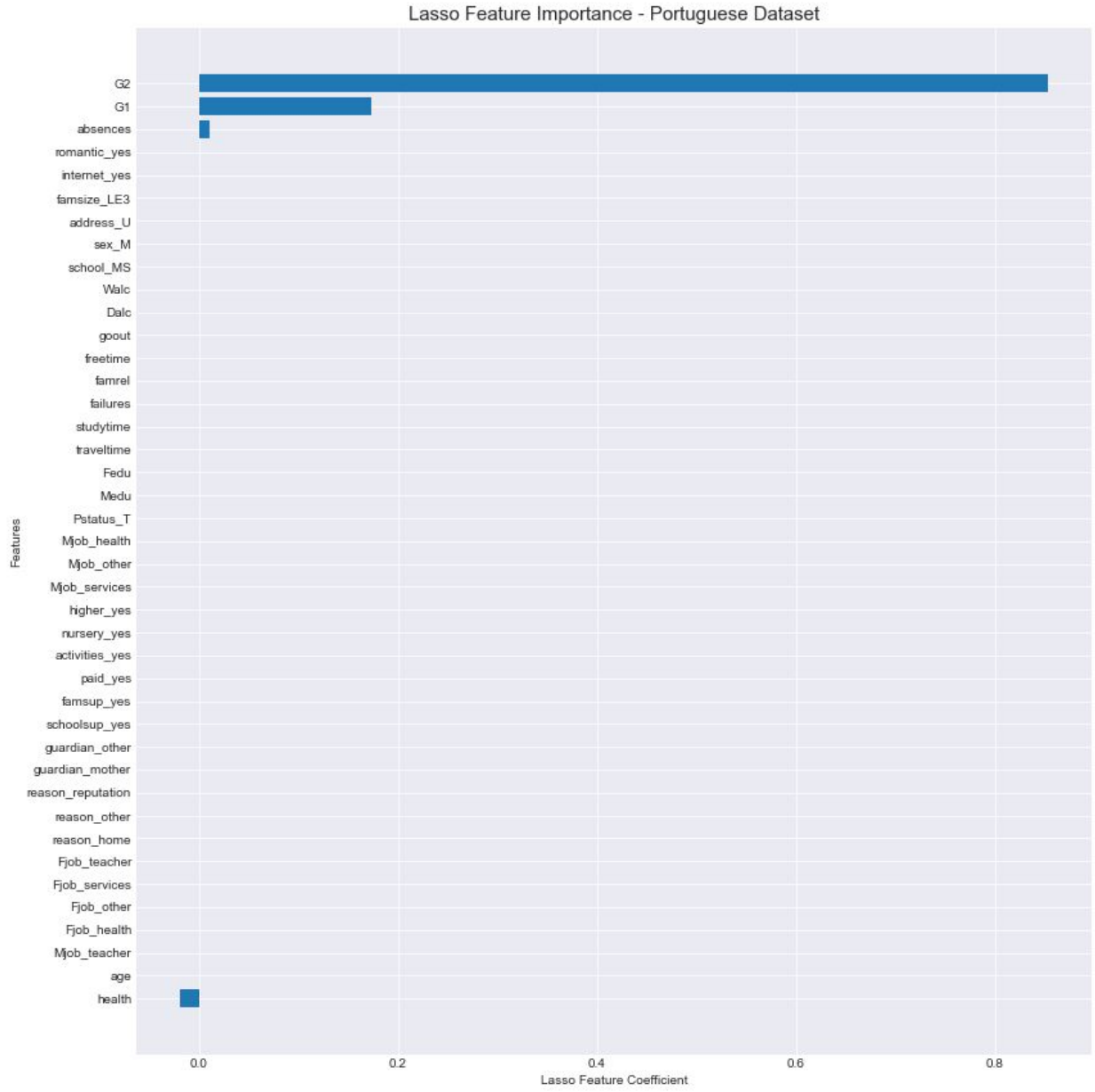

Figure 15: Feature Importance - Portuguese Dataset - Lasso Regression model

E. Performance on the test set

For the mathematics dataset, the XGBoost model performs very well on test data. The R-squared score is 0.8885, and the MSE is 2.2619. For the Portugueses dataset, the Lasso regression model also performs well on test data. The R-squared score is 0.8677, and the MSE is 1.6405.

F. Conclusion

The midterm grade and final grade have a strong correlation, so it is not surprising that midterm grade is the most important feature to predict student performance. Besides midterm grade, other features seem to be irrelevant. Through this project, one thing we can learn from is that instead of having one final test that determines student performance, having multiple mid-term is also a good predictor of student performance. Having multiple small exams throughout the year helps students to decrease stress and anxiety.

Reference

Cortez, P. (n.d.). Student Performance Data Set. Retrieved October 10, 2020, from https://archive.ics.uci.edu/ml/datasets/Student+Performance#

Glen, S. (n.d.). Mean Squared Error: Definition and Example. Retrieved October 10, 2020, from https://ashutoshtripathi.com/2019/01/22/what-is-the-coefficient-of-determination-r-square/

Tripathi, A. (n.d.). What is the Coefficient of Determination | R Square. Retrieved October 10, 2020,from https://ashutoshtripathi.com/2019/01/22/what-is-the-coefficient-of-determination-r-square/