Predicting Student Performance Using Binary Classification - James Ro, Brown University

December 7, 2021

Github Repository: https://github.com/jamesjooyoung/Data-1030-Project.git

1 Introduction

While there have been many studies on building models to predict student performance, the predictive accuracy of these models are complicated by numerous factors that affect student performance. For example, factors ranging from previous academic performance to a student's socioeconomic background must be considered. The features of such models and the model of this project are variables that describe a student's academic background, demographics, among others, that influence how a student performs. The target variable is the student's academic performance in the classroom, which is measured with the final grade a student receives. As with most prediction models, there is no guarantee of perfect predictive accuracy. However, there is considerable value in building even a moderately accurate predictive tool as it could benefit education professionals in their efforts to identify struggling students, which would help education professionals better allocate school resources and improve education quality. This is hugely important in current times, as education professionals fight against growing inequalities and falling standards in education in certain countries.

This project attempts to build a highly accurate predictive tool that utilizes a binary classification model in machine learning to classify a student's performance as pass or fail. The dataset used for this project came from the UCI Machine Learning Repository and was published by the University of Minho in Portugal. The data is derived from school reports and consists of 32 total attributes across 395 students that measures student grades specifically in mathematics, and describes the demographics and the social and school features of students. While the dataset utilizes features 'G1', 'G2', and target variable 'G3' to measure student performance in the first, second, and final period respectively on a 0 (low) - 20 (high) integer scale, a binary scale for 'G1', 'G2', and 'G3' will be used to build the binary classification model in this project, where 0 represents a failing score ranging from 0-9 on the original integer scale and 1 represents a passing score ranging from 10-20 on the original integer scale.

In Cortez's original publication of the database, Cortez used a combination of three data mining goals, which included classification and regression techniques, and four data mining methods (Decision Trees, Random Forests, Neural Networks, Support Vector Machines) to study whether it would be possible to achieve a high predictive accuracy of student performance. One model implemented by Cortez recorded a 93.0% accuracy, proving that it was indeed possible to achieve a high predictive accuracy [1]. In another study by Satyanarayana and Nuckowski, the authors used the dataset to implement multiple classifiers (Decision Trees, Naive Bayes, Random Forest) to see

whether predictive accuracy can be further improved if an ensemble filtering technique is utilized to improve the data quality. In fact, they found that predictive accuracy did increase with ensemble filtering, with 95% accuracy [2]. However, what the predictive accuracies of these two studies show is that there is still room to improve on these models to achieve the highest predictive accuracy possible.

2 Exploratory Data Analysis

The following figures in this section were created during exploratory data analysis.

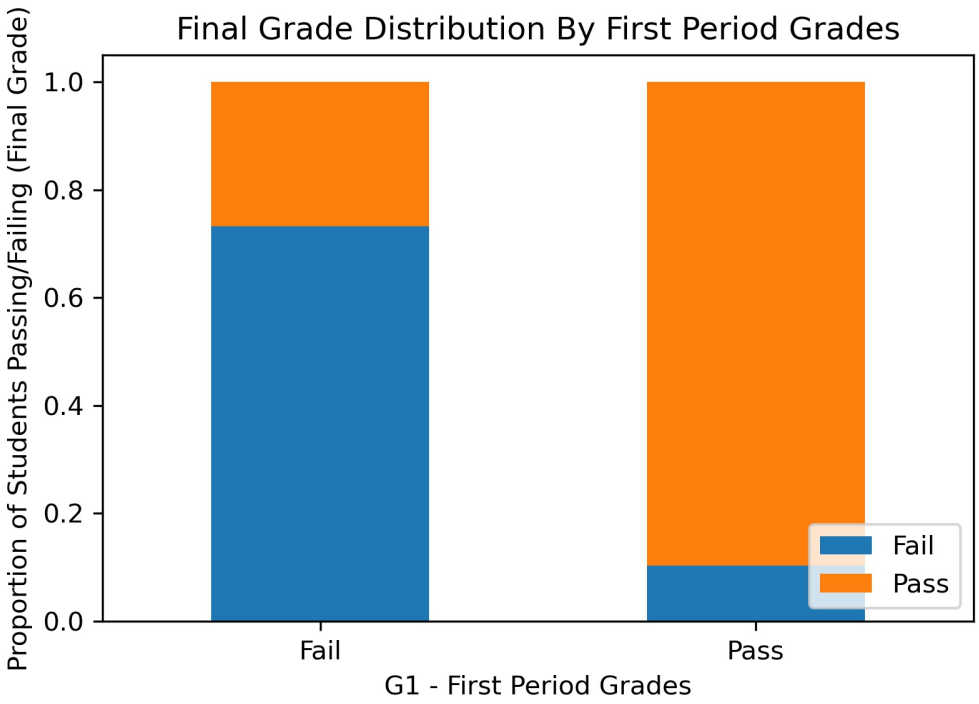

Figure 1 This figure depicts a stacked bar plot showing the proportion of students who receive a final grade of pass or fail in mathematics according to the first period grades students received. For the most part, students who received a passing grade in the first period received a passing final grade, and students who received a failing grade in the first period received a failing final grade. This suggests that the 'G1' feature may be significant and may be highly indicative of student performance.

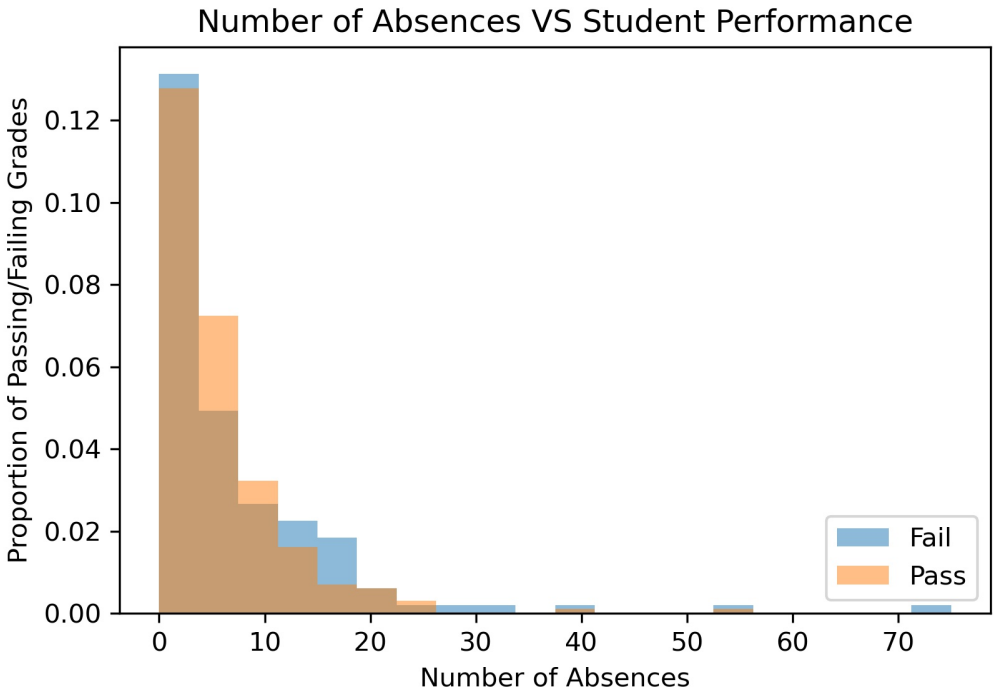

Figure 2 This figure displays a category-specific histogram showing the proportion of students passing or failing in mathematics according to the number of days a student missed class. From the histogram, it is evident that there is a slightly larger proportion of students who achieve a passing final grade than those who achieve a failing final grade when the number absences are low. As the number of absences increases, there is a slight increase in the proportion of students who receive a failing final grade, which suggests absences may have a small negative influence on student performance.

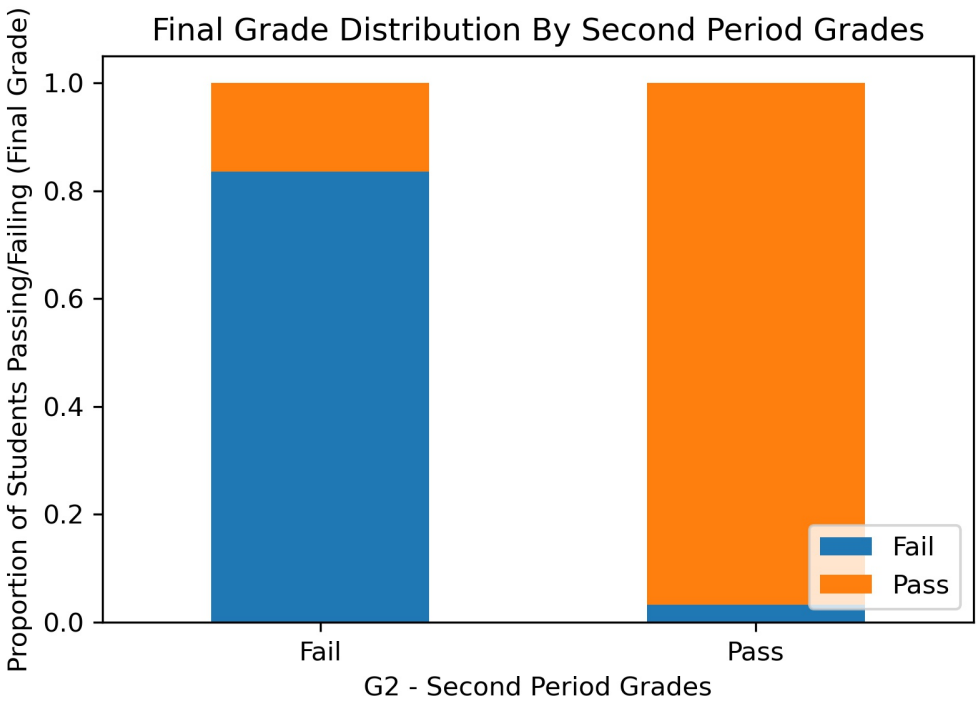

Figure 3 This figure depicts a stacked bar plot showing the proportion of students who receive a final grade of pass or fail in mathematics according to the second period grades students received. Similar to what Figure 1 shows, second period grades are more indicative than first period grades of the final grades that students receive. The fact that second period grades are more closely aligned to final grades suggests that the 'G2' feature is plays a more significant role in predicting final student performance than 'G1'.

3 Methods

3.1 Data Splitting and Preprocessing

Each row, or observation, of the data accounted for one individual student, which meant that the data was assumed to be independent and identically distributed with no time-series or group structure. To account for the small number of observations in the data and to address the need for more cases to test and validate the models given the small dataset, 20% of observations were initially split into testing using train_test_split and the other 80% of observations were allocated to 5-fold cross-validation. This splitting approach was trained on the machine learning model as it accounted for variability in random splits that could occur with a small dataset. The preprocessor then fit and transformed training folds before fitting and transforming the testing and validation sets in each iteration of cross-validation. Some categorical features such as 'Medu' were already ordered and converted into integer values in the data, so these features did not need to be encoded using the ordinal encoder. However, the preprocessor applied the StandardScaler on these features for the purpose of converting them to have mean = 0 and standard deviation = 1. The preprocessor applied one-hot encoder on the remaining categorical features such as 'school', as they were all unordered and could not be clearly ranked. Finally, the preprocessor applied the MinMaxEncoder

on the remaining continuous features such as 'age', as they were all bounded by reasonable ranges. The target variable 'G3' was not encoded in the preprocessor.

From the exploratory data analysis, it was concluded that 'G1' and 'G2' features would be highly influential in predicting student performance. Thus, three input configurations were tested on machine learning models: 1) configuration with all features except target variable 'G3', 2) similar to 1) but with 'G2' removed, 3) similar to 1) but with 'G1' and 'G2' removed. There were 32 total features in the final preprocessed data for the first configuration, 31 total features for the second configuration, and 30 total features for the third configuration.

3.2 Training Models

Six machine learning models were trained and compared to find the best algorithm for the dataset: 1) Logistic Regression with L1 regularization, 2) Logistic Regression with L2 regularization, 3) Logistic Regression Model with ElasticNet regularization, 4) Random Forest Classifier, 5) Support Vector Machine Classifier, 6) K-Nearest Neighbors Classifier. All models were trained using the previously mentioned splitting and preprocessing methods and were hyperparameter tuned using a ML pipeline. For the L1 and L2 models, the parameter C was tuned, with penalties set to 'l1' and 'l2' respectively and solver set to 'saga'. In addition to parameter C, the parameter l1 ratio was tuned in the ElasticNet model, with penalty set to 'elasticnet' and solver set to 'saga'. The RF Classifier tuned the max_depth and max_features parameters, the SVC model tuned the gamma and C parameters, and the KNN Classifier tuned the n_neighbors and weights parameters. The ML pipeline took in preprocessor, machine learning algorithm, and parameter grid as parameters and utilized GridSearchCV to perform cross validation to tune the models across 10 different random states for 10 different splits. Furthermore, accuracy score served as the evaluation metric for our models in our binary classification problem since the dataset was quite balanced, given that 67.1% and 32.9% of points were found to belong to class 1 and class 0 respectively. The ML pipeline also returned the best test scores, model parameters, and the grid.

4 Results

4.1 Model Comparisons

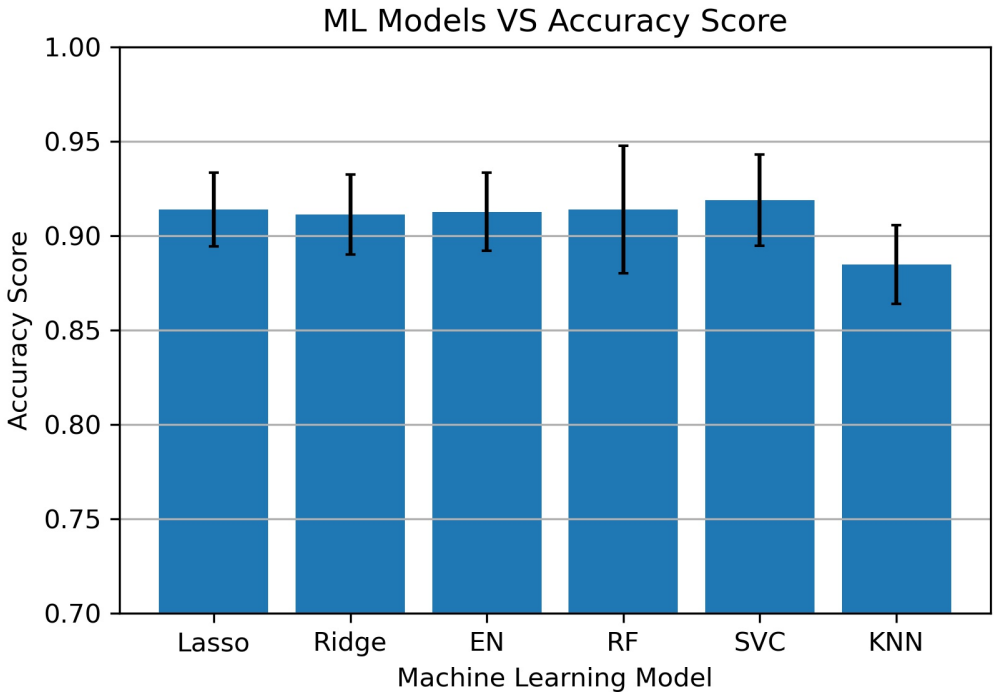

Figure 4 Average accuracy scores for each machine learning model across 10 different random states, with 'G1' and 'G2' included

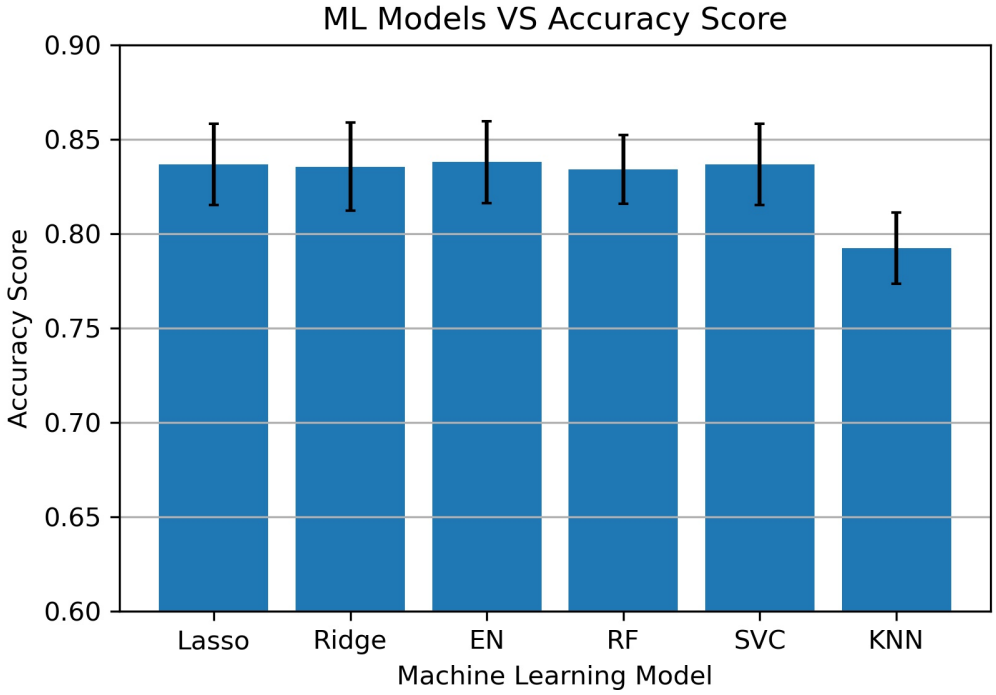

Figure 5 Average accuracy scores for each machine learning model across 10 different random states, with 'G2' removed

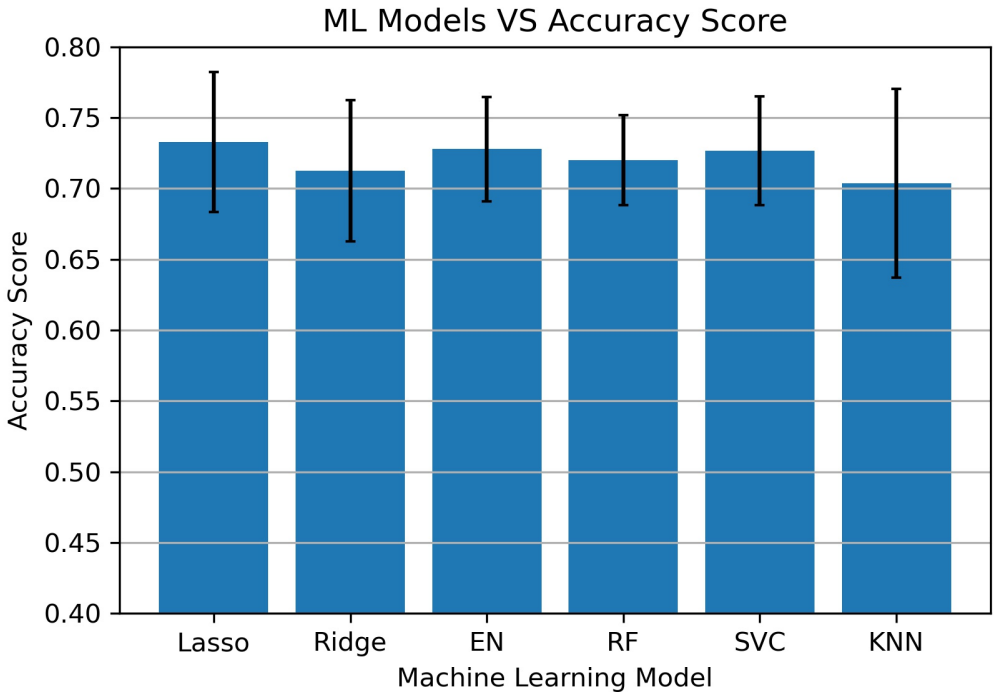

Figure 6 Average accuracy scores for each machine learning model across 10 different random states, with 'G1' and 'G2' removed

In the first setup where 'G1' and 'G2' were not removed as seen in Figure 4, the SVC model barely edged out the RF Classifier as the most predictive model as it had the best overall test score of 0.962 and had the higher mean test score of 0.919, which was 10.3 standard deviations above the baseline accuracy score of 0.671 or the proportion of points in class 1. In the second setup where 'G2' was removed as seen in Figure 5, the performance of all models except KNN Classifier was nearly identical where L2 model had the best overall test score of 0.886 and a mean test score of 0.835, which was 7.05 standard deviations above the baseline accuracy score of 0.671. In the third setup where 'G1' and 'G2' were removed as seen in Figure 6, L1 model was the most predictive model as it had the best overall test score of 0.785 and had a mean test score of 0.733, which was 1.25 standard deviations above the baseline accuracy score of 0.671. As previous period grades were dropped, there were noticeable drops in model performance and in accuracy scores evaluated by the models.

4.2 Interpretations

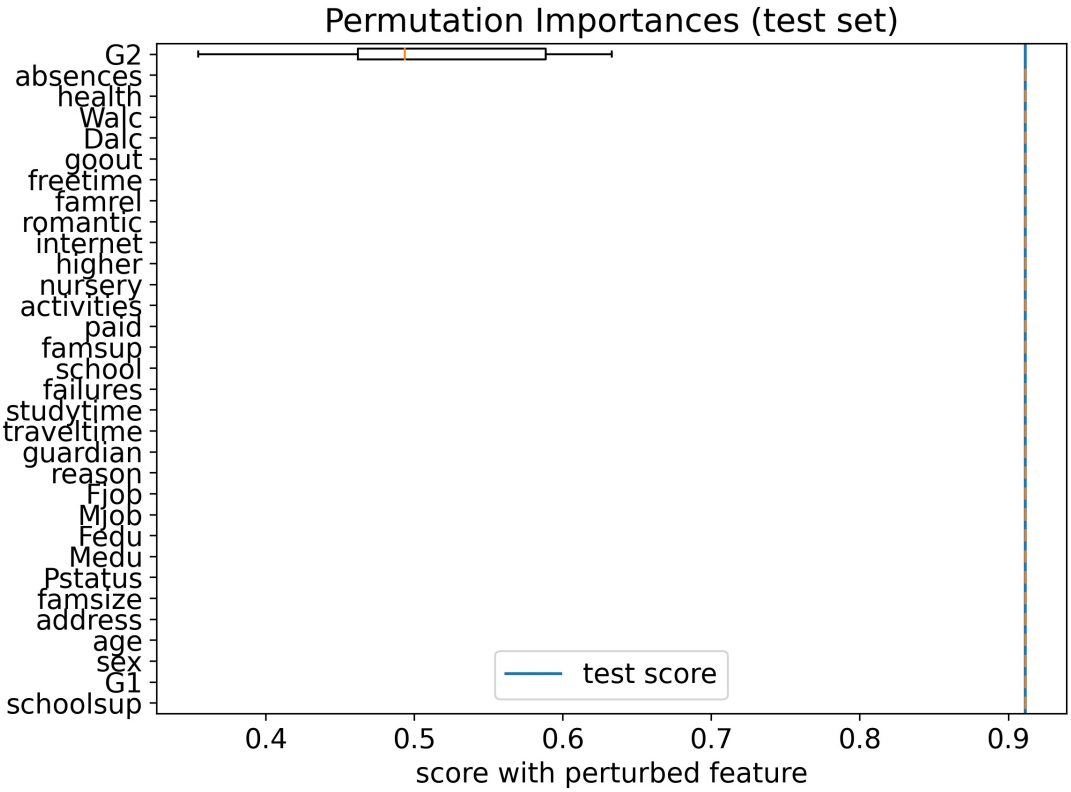

Figure 7 Permutation importances of features, including 'G1' and 'G2'

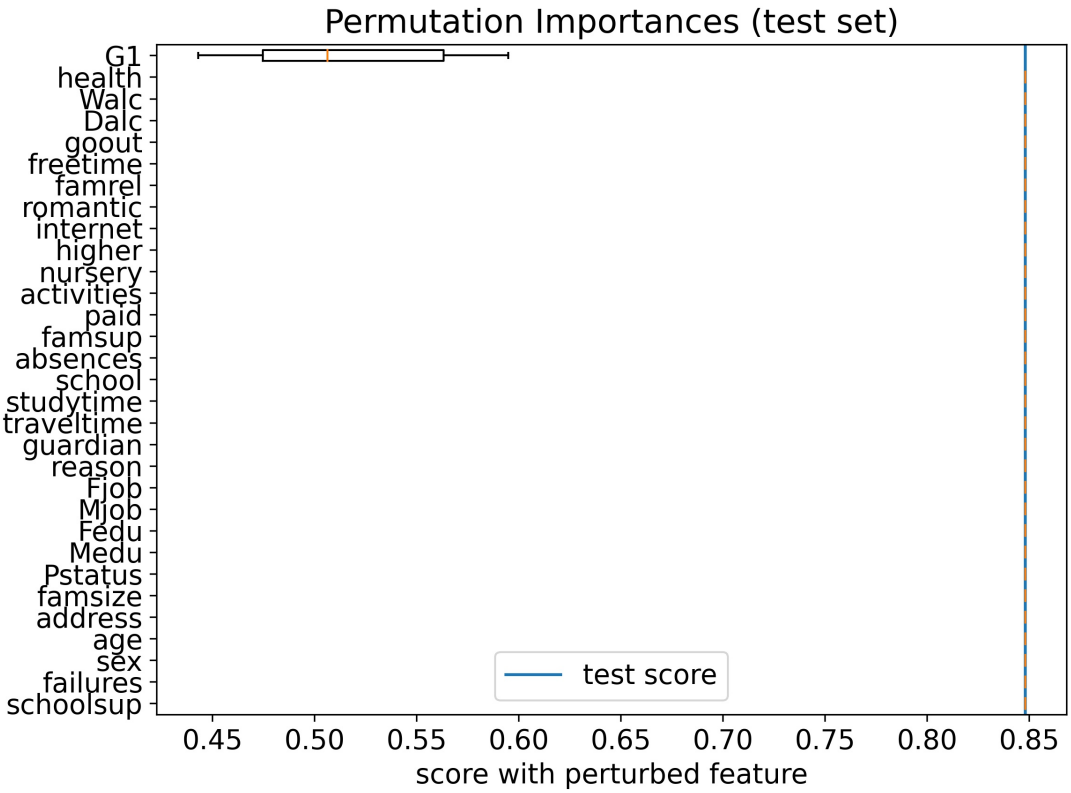

Figure 8 Permutation importances of features, with 'G2' removed

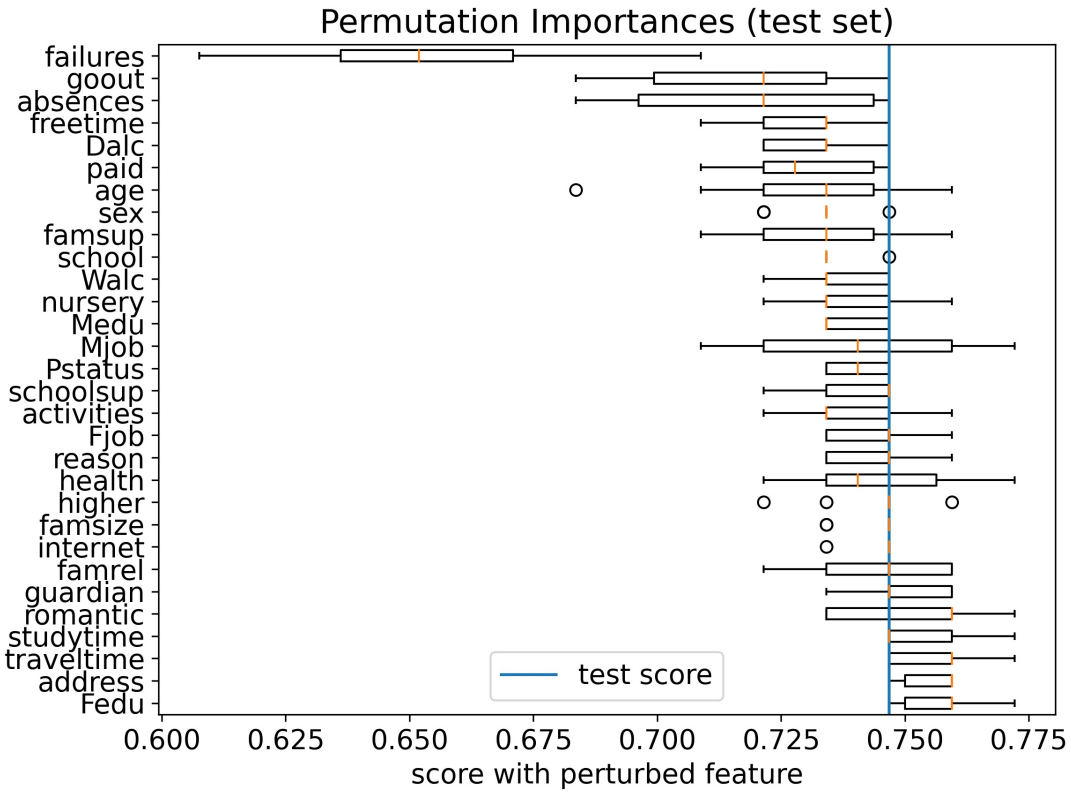

Figure 9 Permutation importances of features, with 'G1' and 'G2' removed

To visualize feature importances in our dataset, a global permutation method was adopted and the RF Classifier was selected as the model to be implemented because it was concluded to be very predictive in the first setup where all previous grades were included and because it was most efficient to implement when all grades are factored into the models. Figure 7-9 showed that the second period grade was the most influential feature in predicting the final student performance. When the second period grade was dropped, the first period grade was the biggest indicator of final student performance. When all previous period grades were dropped, other features such as number of school absences, past classes failed, and time spent outside with friends emerged as the important features. Features like a student's sex and the school a student goes to were found to be the least important. While this was somewhat expected that previous period grades play a huge role in predicting student performance, it was surprising that the importances of features other than previous period grades only started to become more evident when previous period grades were removed from our models.

5 Outlook

There are a few ways that the models built in this project can be further improved. First, the dataset can be enlarged so that it includes more schools and more school years. This would enrich and diversify data collected on students, and open possibilities of exploring new features in the models such as grades from previous school years. Second, more rigorous feature selection should be considered so that less important features are filtered out. While the models built in this project

were specifically trained to account for highly influential features such as 'G1' and 'G2', they did not account for features that had minimal importance. By implementing a more rigorous feature selection that targets less important features, the models may have higher predictive accuracies and may see an improvement in its performance.

6 References

[1] P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-' 9077381-39-7.

[2] A. Satyanarayana and M. Nuckowski, "Data Mining using Ensemble Classifiers for Improved Prediction of Student Academic Performance," Middle Atl. Sect. Spring 2016 Conf. (ASEE 2016), no. April.2016.