Jaehwan Han

Machine Learning Project Report

Predicting secondary school student performance

1. Introduction

This study applies various machine learning models to predict student achievement in secondary education using various education-related predictors. The data was collected from two Portuguese secondary schools with a number of demographics, social, and school-related variables and also the student achievement in Portuguese. The outcome of interest (i.e., Portuguese achievement) was converted into three forms—continuous, binary, and multi-nary, and used for regression, binary classification, and multi-nary classification tasks respectively. The best predictive model was selected for each task and the important features for prediction were discovered and discussed.

2. Method

2. 1. Data

The dataset was obtained from 649 Portuguese students from two schools, and had total 33 variables. Three of them were their grades: G1 – first period grade, G2 – second period grade, ad G3 – final grade. All of them were in numeric scale from 0 to 20. Since G1 and G2 could trivially predict G3, and out interest was in finding education-related predictors for achievement, G1 and G2 were removed from the analysis, and only applied at the final step by adding them to the final predictive model to check their effects. The outcome variable G3 was converted into 3 types: 1. No conversion with range from 0 to 20 for regression, 2. Binary classification with Pass / Fail (0-9 Fail, 10 – 20 Pass), and 3. Multi-nary classification with five levels (0-9 fail (V), 10-11 sufficient (IV), 12-13 satisfactory (III), 14-15 good (II), 16-20 excellent(I)).

The other 30 variables were used as predictors. They include demographics such as sex and age, social variables such as mother's education and job, father's education and job, and family relationship, school-related variables such as number of past failed classes and extra paid classes, and other education-related variables such as study-time and whether they want to take higher education.

All of the predictors except age and number of absences were categorical variables even though they were encoded to numeric in the raw data. One-hot-encoding was applied to dummy-code all of them, and the total number of predictors was 97 after that transformation.

The dataset was split into two sets—training (80%) and test (20%) sets. For classification tasks in particular, it was split using stratified-split based on the number of classes. The test set was left untouched in the training process, and tested only for the final models in regression and classification tasks to validate their effectiveness.

2.2. Machine learning on regression

For the regression task, six machine learning (ML) models—Ridge regression, Lasso, Random forest, Gradient boosting, Support vector regression, and Multilayer perceptron—were considered, and incorporated into analytical pipelines of step (1) removing zero or near-zero variance (NZV) predictors, (2) scaling using standard or minmax scaler, and (3) regression using one of those ML models. For the step (1), NZV predictors may make the estimated model unstable because they may be included in the training set but not in the validation set, or vice versa, and thus needs to be removed. The cut-off variance was either 0.01 or 0.05. For step (2), various ML models are sensitive to scaling because the different variances of predictors may affect the model differently and lower its predictive performance. The selected method was either standard scaling or min-max scaling, but no scaling was provided for tree-based ensemble models.

The performance of various hyper-parameter combinations of those models was estimated and compared using 2-times repeated 5-fold cross-validation (CV) strategy in grid search algorithm of Python Scikit-learn library. For Ridge and Lasso, regularization parameter alpha was one of (0.001, 0.01, 0.1, 1, 10, 100, 100). For Random forest (RF), number of trees was (100, 200, 400, 800), and max features was (5, 11, 20, 30). For Gradient Boosting (GB), number of trees was the same with RF, learning rate was (0.001, 0.01, 0.1, 1), and max depth was (1, 2, 4). For Support vector regression (SVR), kernel was either linear or radial-basis (RBF), C was (0.001, 0.01, 0.1, 1, 10, 100, 1000), epsilon was (0.01, 0.1, 0.5, 1, 5) and gamma for RBF was (0.001, 0.01, 0.1, 1, 10, 100, 1000). For multilayer perceptron (MLP), alpha was (0.001, 0.01, 0.1, 1, 10, 100), learning rate was (0.01, 0.1, 1), and hidden layer was one with hidden units from (1, 3, 5, …, 21) or two with hidden units (1, 5, 9, 13, 17, 21) x (1, 5, 9, 13, 17, 21).

The performance metric was R-square, and the best CV score model was selected from each ML model, and the final model was chosen among them with the highest CV R-square score. Finally, the training and test R-square score of the final model was provided.

2.3. Machine learning on classification (binary and multi-nary)

We used TSNE to visualize the possibility of classification tasks. With T-SNE nonlinearly transforming distance and plotting points on the two-dimensional embedded space, we could somehow tell, although not guaranteed, whether the dataset can be classified into two or five levels.

For both binary and multi-nary classification tasks, the number of outcome classes was imbalanced. Thus, we used three scenarios— (1) no resampling and use the imbalanced classes as is, (2) weighting inversely according to class frequency using 'balanced' option, and (3) SMOTE (Synthetic Minority Over-Sampling Technique). SMOTE is an oversampling algorithm to reduce overfitting from simply reproducing more instance in minority class. It generates synthesized new minority instances based on feature space similarity by using a nearest neighbor's algorithms. For multi-nary classification, additionally, we performed (4) multi-label classification (using label_binarizer) applying one vs. rest classifiers to use AUC (area under the curve) metric since those three methods above does not provide AUC (more details below).

The same 3-step pipelines as in regression were applied as well with the last step as classification using one of those ML models. Only when the scenario (3) SMOTE was used, the pipeline was 4-steps with SMOTE at the very first step. Six machine learning (ML) models—Ridge logistic regression, Lasso logistic regression, Random forest, Gradient boosting, Support vector machine, and Multilayer perceptron—were used as a classifier. The performance of various hyper-parameter combinations of those models was estimated and compared using 2-times repeated 5-fold stratified cross-validation (CV) strategy. The parameter combinations were the same as in regression except that C instead of alpha was used for Ridge and Lasso logistic regression and SVM did not have a parameter epsilon.

For the binary classification task, the performance metric was AUC (roc_auc), which was robust that accuracy for imbalanced classes. For the multi-nary classification task, the performance metric for the first three methods was accuracy, which is the same as micro f1-score in multi-nary classification with only one prediction for each, but was AUC for the multi-label classification scenario.

For the binary classification, the best CV AUC model was selected in each scenario, and the final model was chosen among them with the highest CV AUC score. For the multi-nary classification, the best CV accuracy model was selected from scenario (1) to (3), and compared with the best CV AUC model from scenario (4) using micro f1-score, and the better model was chosen as the final model.

3. Result

3.1. Descriptive analysis

For a preliminary visual exploration of the data, two figures of strip-plots were drawn with regard to regression and classification tasks respectively. Figure 1 consists of 10 strip-plots with the continuous outcome variable G3 at the y-axis and each of the potentially important predictors such as school, age, study time, and failures at the x-axis. Figure 2 also consists of 10 strip-plots each of which has two of those potentially important predictors at the x- and y-axis with the binary outcome variable G3_bin distinguishing the data points in two different colors.

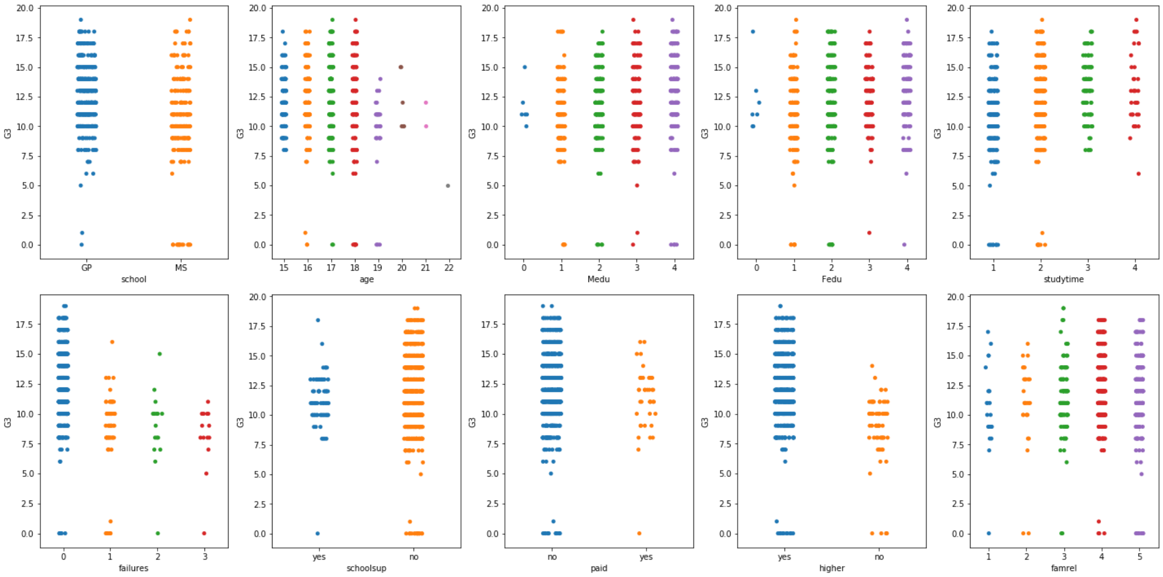

Figure 1. Strip-plots between predictors and the continuous outcome

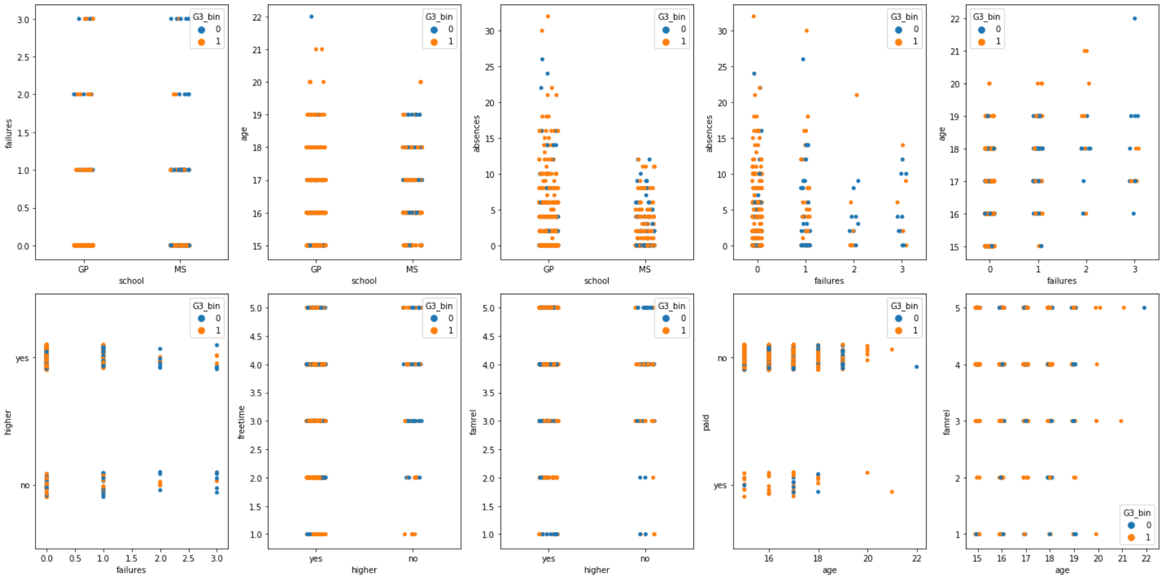

Figure 2. Strip-plots between predictors with the binary outcome in two colors

3.2. Regression results

The best linear model (including Ridge and Lasso regression) with the best CV R-square (0.295) score was a Ridge regression with alpha 100, with an NZV threshold 0.01 and a min-max scaler. The best ensemble model (including Random forest and Gradient boosting) with the best CV R-square (0.335) was a Random forest with max features 20 and number of estimators 800, with an NZV threshold 0.01 and no scaler. The best SVR model with the best CV R-square (0.315) was one with RBF kernel, C 10, epsilon 0.5, and gamma 0.001, with an NZV threshold 0.01 and a standard scaler. The best MLP model with the best CV R-square (0.306) was one with alpha 100, learning rate 0.01, 1 hidden layer with 19 units, with an NZV threshold 0.05 and a standard scaler.

The final model with the best CV score among them was the RF model (R-square 0.335). Its training and test R-square were 0.910 and 0.245 respectively, and test RMSE (root mean squared error) was 2.625. The top 10 important features were selected and shown in Figure 3.

Those CV scores and final scores are presented in table 1.

Table 1. Regression results. Training and test scores were obtained only for the final model.

	Linear	**Random Forest**	SVR	MLP
CV R-square	0.295	**0.335**	0.315	0.306
Training R-square		0.910		
Test R-square		0.245		
Test RMSE		2.625		

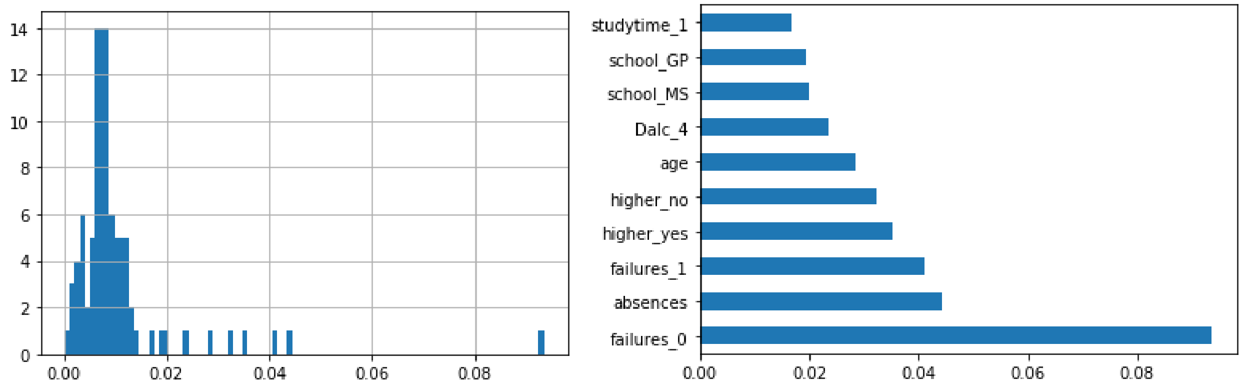

Figure 3. (Left) distribution of all feature importance. (Right) top 10 feature importance.

3.3. Binary classification results

The number of data points for each class was 100 for class 0 (fail) and 549 for class 1 (pass).

T-SNEs with perplexity 50 for the two classes is shown in Figure 4. The left panel is displayed with non-weighted classes, and the right panel is after SMOTE.

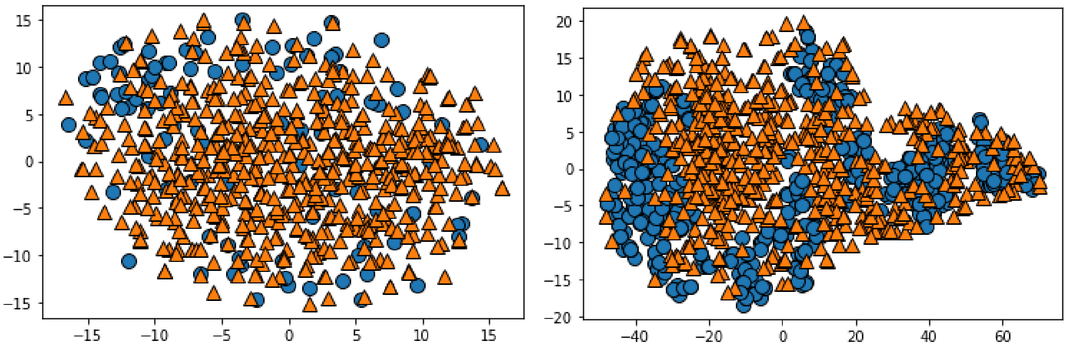

Figure 4. T-SNE for binary classification task. (Left) no weighting. (Right) after SMOTE.

For the scenarios (1) no-weighting and (2) weighting (i.e. balanced), the best linear model (including Ridge and Lasso logistic regression) with the best CV AUC (0.834) score was a Lasso logistic regression with C 0.1 and balanced classes, with an NZV threshold 0.01 and a min-max scaler. The best ensemble model (including Random forest and Gradient boosting) with the best CV AUC (0.848) was a Random forest with max features 20, number of estimators 100, and no-weighting, with an NZV threshold 0.01 and no scaler. The best SVM model with the best CV AUC (0.840) was one with RBF kernel, C 0.001, gamma 0.01, and balanced classes, with an NZV threshold 0.01 and a standard scaler. The best MLP model with the best CV AUC (0.829) was one with alpha 100, learning rate 0.1, 1 hidden layer with 21 units and no-weighting, with an NZV threshold 0.01 and a standard scaler.

For (1) and (2), the best model with the best score among them was the RF model (AUC 0.848).

For the scenario (3) SMOTE, the best linear model with the best CV AUC (0.824) score was a Ridge logistic regression with C 0.01, with an NZV threshold 0.05 and a min-max scaler. The best ensemble model with the best CV AUC (0.841) was a Random forest with max features 20, number of estimators 400, with an NZV threshold 0.05 and no scaler. The best SVM model with the best CV AUC (0.825) was one with RBF kernel, C 1 and gamma 0.001, with an NZV threshold 0.01 and a min-max scaler. The best MLP model with the best CV AUC (0.818) was one with alpha 10, learning rate 0.01, 2 hidden layers with (17, 21) units, with an NZV threshold 0.05 and a min-max scaler.

For (3), the best model with the best CV score among them was the RF model (AUC 0.841).

Although the CV AUC of the best model from (1) and (2) was slightly higher (0.07) than that from (3), the classifier based on SMOTE may provide more stable predictions in the future (Chawla et al. 2002), and both models are RF models with equivalent interpretability. Thus, the Random Forest model using SMOTE was chosen as the final model. Its training and test AUC were 1.0 and 0.831 respectively, and test accuracy was 0.885. The top 11 important features were selected and shown in Figure 5.

Those CV scores and final scores are presented in table 2.

Table 2. Binary classification results. Training and test scores were obtained only for the final model.

Non-SMOTE	Linear	Random Forest	SVM	MLP
CV AUC	0.834	0.848	0.840	0.829

SMOTE	Linear	**Random Forest**	SVM	MLP
CV AUC	0.824	**0.841**	0.825	0.818
Training AUC		1.0		
Test AUC		0.831		
Test Accuracy		0.885		

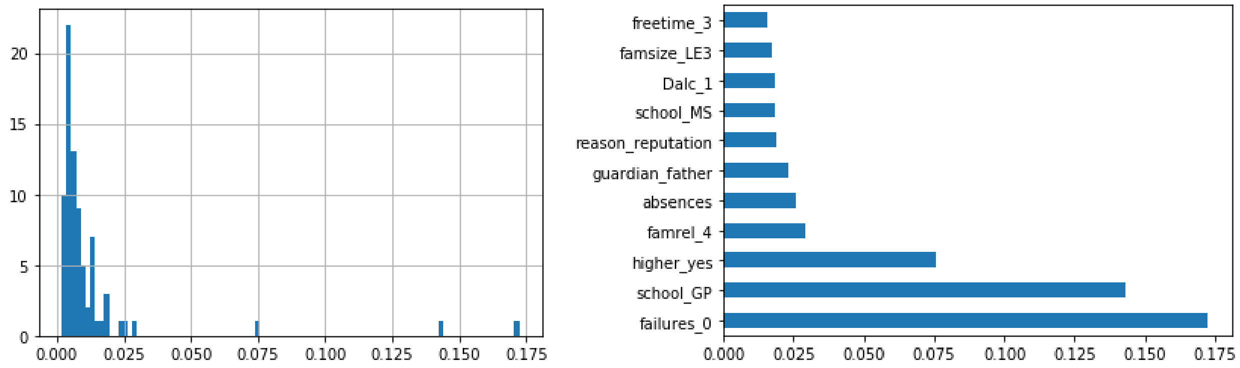

Figure 5. (Left) distribution of all feature importance. (Right) top 11 feature importance.

3.4. Multi-nary classification results

The number of data points for each class was 100 for class 0 (V), 201 for class 1 (IV), 154 for class 2 (III), 112 for class 3 (II), and 82 for class 4 (I).

T-SNEs with perplexity 50 for the two classes is shown in Figure 6. The left panel is displayed with non-weighted classes, and the right panel is after SMOTE.

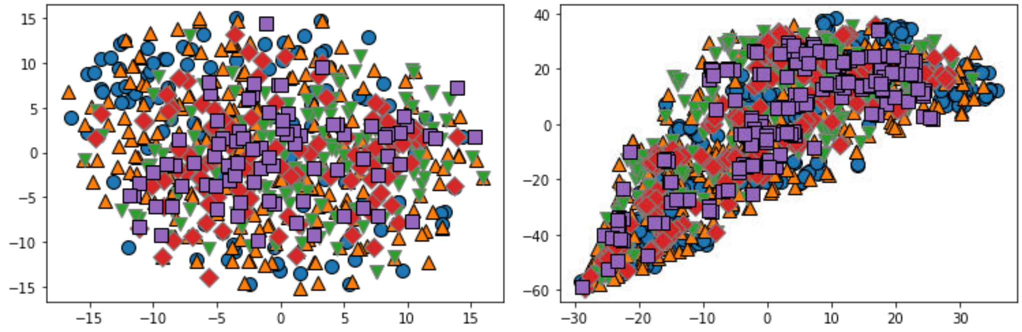

Figure 6. T-SNE for multi-nary classification task. (Left) no weighting. (Right) after SMOTE.

For the scenarios (1) no-weighting and (2) weighting (i.e. balanced), the best linear model with the best CV accuracy (0.368) score was a Ridge multinomial logistic regression with C 0.01 and no-weighting, with an NZV threshold 0.05 and a min-max scaler. The best ensemble model with the best CV accuracy (0.384) was a Random forest with max features 5, number of estimators 800, and no-weighting, with an NZV threshold 0.05 and no scaler. The best SVM model with the best CV accuracy (0.373) was one with RBF kernel, C 10, gamma 0.1, and no-weighting, with an NZV threshold 0.05 and a min-max scaler. The best MLP model with the best CV accuracy (0.373) was one with alpha 10, learning rate 1, 1 hidden layer with 1 unit, and no-weighting with an NZV threshold 0.05 and a standard scaler.

For (1) and (2), the best model with the best CV score among them was the RF model (accuracy 0.384)—final model candidate 1.

For the scenario (3) SMOTE, the best linear model with the best CV accuracy (0.361) score was a Lasso logistic regression with C 1, with an NZV threshold 0.05 and a min-max scaler. The best ensemble model with the best CV accuracy (0.382) was a Random forest with max features 11, number of estimators 200, with an NZV threshold 0.05 and no scaler. The best SVM model with the best CV accuracy (0.376) was one with RBF kernel, C 1 and gamma 0.1, with an NZV threshold 0.05 and a min-max scaler. The best MLP model with the best CV accuracy (0.383) was one with alpha 10, learning rate 0.1, 2 hidden layers with (13, 21) units, with an NZV threshold 0.05 and a standard scaler.

For (3), the best model with the best CV score among them was the MLP model (accuracy 0.383)—final model candidate 2.

For the scenario (4) multi-label classification with one vs. rest classifiers, the best linear model with the best CV AUC (0.700) score was a Ridge logistic regression with C 0.001 and balanced classes, with an NZV threshold 0.05 and a standard scaler. The best ensemble model with the best CV AUC (0.694) was a Random forest with max features 5, number of estimators 400, and balanced classes, with an NZV threshold 0.05 and no scaler. The best SVM model with the best CV AUC (0.702) was one with RBF kernel, C 0.1 and gamma 0.01, and balanced classes, with an NZV threshold 0.05 and a standard scaler. The best MLP model with the best CV AUC (0.676) was one with alpha 10, learning rate 1, 1 hidden layer with 21 units, with an NZV threshold 0.05 and a standard scaler.

For (4), the best model with the best CV AUC among them was the SVM model (AUC 0.702), and its CV micro f1-score was calculated to be 0.369—final model candidate 3.

For the final model candidates 1 and 2, accuracy is the same as micro f1-score because the multi-class prediction was achieved with only one predicted class for each observation, and the candidate 1 is better than the candidate 2 because it has a slightly higher accuracy and better interpretability. The micro f1-score for the candidate 3 is calculated slightly differently because it allows multiple predictions for each observation, so not directly comparable with the above, but still this value was lower (0.15) than that of the candidate 1. Thus, the Random forest model with accuracy 0.384 (candidate 1) was chosen as our final model. Its top 15 important features were selected and shown in Figure 7.

Those CV scores and final scores are presented in table 3.

Table 3. Multi-nary classification results. Training and test scores were obtained only for the final model.

Non-SMOTE	Linear	**Random Forest**	SVM	MLP
CV Accuracy	0.368	**0.384**	0.373	0.373
Training accuracy		**1.0**		
Test Accuracy		**0.323**		

SMOTE	Linear	Random Forest	SVM	MLP
CV Accuracy	0.361	0.382	0.376	0.383

Multi-label	Linear	Random Forest	SVM	MLP
CV AUC	0.700	0.694	0.702	0.676
CV micro f1			0.369	

3.5. Final models including G1 and G2

The final regression and classification models were re-fit with G1 and G2 as additional predictors. The final Random forest regression model with G1 and G2 showed CV R-square of 0.795, test R-square of 0.809, and test RMSE of 1.32. The final Random forest binary classification model (with SMOTE) with G1 and G2 showed CV AUC of 0.963, test AUC of 0.974, and test accuracy of 0.923. The final Random forest multi-nary classification model with G1 and G2 showed CV accuracy of 0.651 and test accuracy of 0.70. The performance of all the three models was quite higher than those without G1 and G2, since G1 and G2 are likely to be trivial predictors of G3.

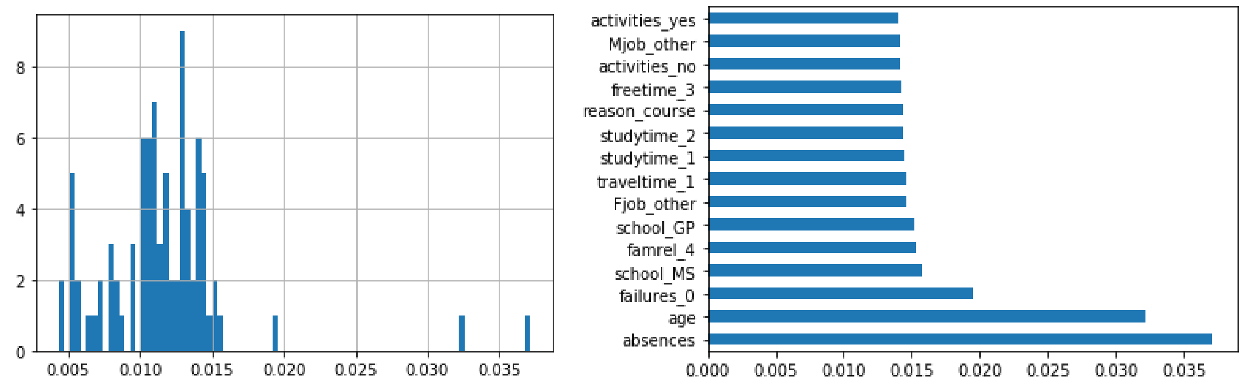

Figure 7. (Left) distribution of all feature importance. (Right) top 15 feature importance.

4. Discussion

The objective of this study was to find out the relationship between social environmental attributes and student academic achievement. According to the Figures 1 and 2, there are some notable expectations as related to the overall purpose of this study:

1. Most of the students attended Gabriel Pereira (GP) school generally have higher final grades than the students attended Mousinho da Silveria (MS) and much less students in GP school are located in lower score. However, when it comes to past class failure and number of school absences per each GP and MS school doesn't show a huge difference. It means, even if both school students have similar amount of past class failure and school absences and even there are a few students from GP who missed a lot classes but passed the exam, students from GP is more likely to pass the class. So we could expect that GP school might make the test relatively easier than MS school. 2. Regardless of the school, increasing study time and the students who want to take higher education have influenced positively on getting high scores. 3. Surprisingly, when we look at the plot which represents relationship between higher (wants to take higher education) and famerl (quality of family relationship), even if students have passion and expectation about studying, the students couldn't pass the exam under very bad quality of family relationship because they can hardly focus on their study. 4. Obviously, students who missed class shows more fail their classes. Also, most high scorer is located in none of past class failure. It indicates that past class failure may well be to do with final grade. Moreover, we could expect that our binary classification might be able to yield pretty acceptable accuracy through T-SNE technique in Figure 4. It does not guarantee but possibly imply that two classes might be well distinguished.

After making expectations based on the two preliminary visual exploration of data, we implement ML model and extract (1) top 10 important features from the Random forest regression model, (2) top 11 important features from the Random forest binary classification model, and (3) top 15 important features from the Random forest multi-nary classification model, that are highly related to final grade from our final model in Figure 3, 5, 7, respectively. In regression, past class failure, number of school absence, desire to take higher education, age were the most important features, followed by frequently alcohol consumption, school (MS, GP), and study time. In binary classification, past class failure, school (GP, MS), desire to take higher education were the most important features, followed by quality of family relationship, number of school absence, student's guardian, reason to choose this school, alcohol consumption, family size, free time after school. In multi-nary classification, absences, age, failures were the most important features, followed by school, family relationship, father's job, study time, reason to choose the school.

When taking into consideration that high feature-importance attributes appear in those ML model, we could simply describe that past class failure, number of school absence, desire to take higher education and school they attend. These features are exactly same that we expected above, even before implementing machine learning. There is also farmel_4 features which is quality of family relationship. This has a large feature importance in binary classification model and it is in concordance with our prediction as well. Therefore, even if the features that are selected seems like an obvious, studying steadily, no absent from school and offering students a better environment in which to study have been demonstrated again through this research.

As for performance of binary classification, most of ML models yield high AUC score (Best CV score 0.841 and test AUC 0.831) and it describes two classes are well distinguishable as implied by the t-SNE method (Figure 4). For multi-nary classification, the AUC of multi-label classification was rather good with the highest CV score 0.702 and test score 0.740 using one vs. rest classifiers, but the test accuracy from aggregating them all was not very good with 0.438 due to the not-clearly-distinguishable nature of the multiple classes that was implied by t-SNE transformations (Figure 6)..

Through model implementation and analyzing of each figures and characteristics of each features, we realized that the importance of data exploration before developing machine learning algorithms. It is quite surprising the similarity of assumption and result of ML model. People who don't even work on this research might be able to predict that past failure record, school absence and aspiration for academic achievement will be the influential attributes to student's academic ability. However, when things are more complicated other than simple student final grade and the topic can be accompanied by interaction between social determinants, people don't want decision only depend on their speculation.

In that case, you could roughly great estimate the outcome just "by exploring and visualizing dataset in detail" and it is an essential initial step of data analysis and should not be treated "less important" part. ML model is more of confirming your prediction and further deep analyzing technique.

References

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.