A MACHINE LEARNING PROJECT: PREDICTING STUDENT PERFORMANCE

Annie Phan
DATA1030, Brown DSI

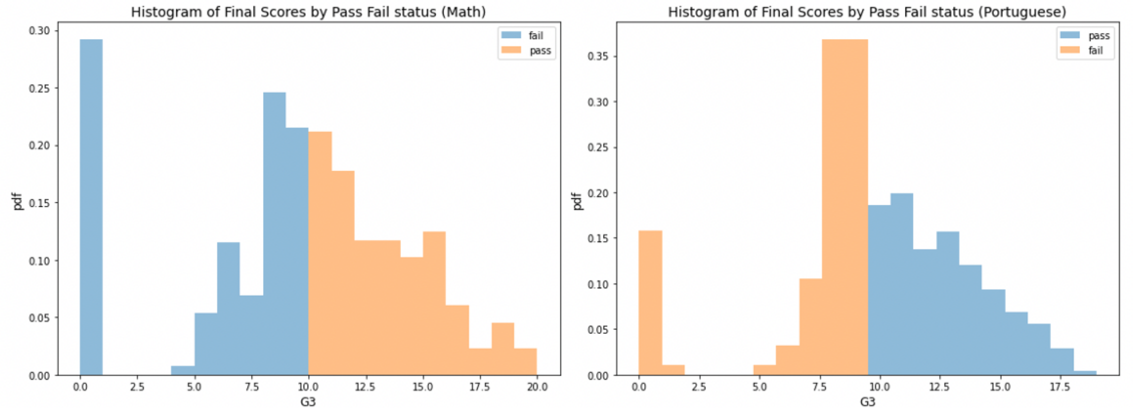
1. Introduction

- The project examines how social and educational factors impact students' performance to allow educators to intervene as necessary. I use ML to predict if students will pass or fail based on information such as family background, socioeconomic status, previous grades, etc. The students pass if their final score is 10 and fail otherwise.
- The analysis is important because although the educational level of the Portuguese population has increased in last decades, Portugal remains at Europe's tail end due to its high student failure rates. The data helps Portuguese educational institutions and government find attributes to best invest in to improve students' performance and Business Intelligence/Data Mining to develop automated tools that can improve decision making and optimize success in education.
- The data is obtained from the UCI ML Repository and my project is based on the paper "Using Data Mining to Predict Secondary School Student Performance" by Cortez and Silva in 2008. I work with 2 datasets for students Math and Portuguese scores. There are 395 students in Math and 649 students in Portuguese. Both datasets have 33 features, which include students' age, romantic status, internet access, parent's jobs, past classroom failures, etc. The measure of academic performance is final grades (0-20). There are 2 midterm grades and 1 final grade.

2. Exploratory Data Analysis

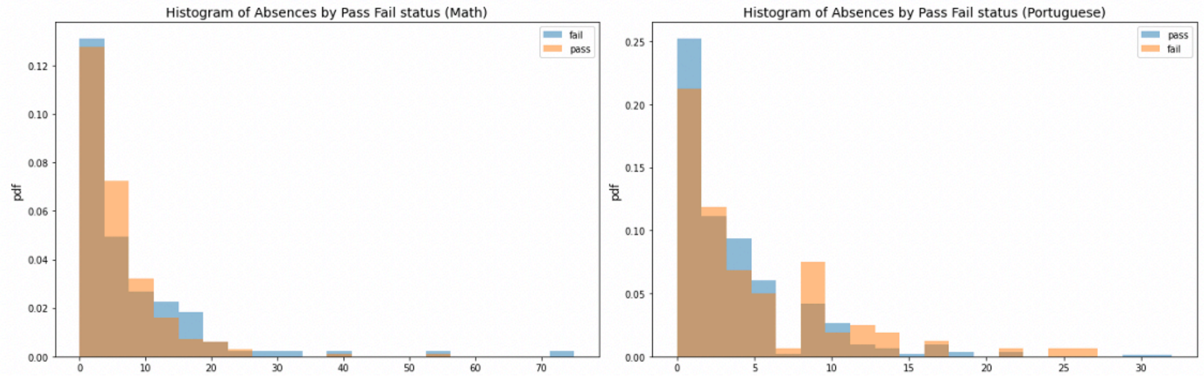
- I've updated my EDA to focus more on the target variable and most important features.

Figure 1:



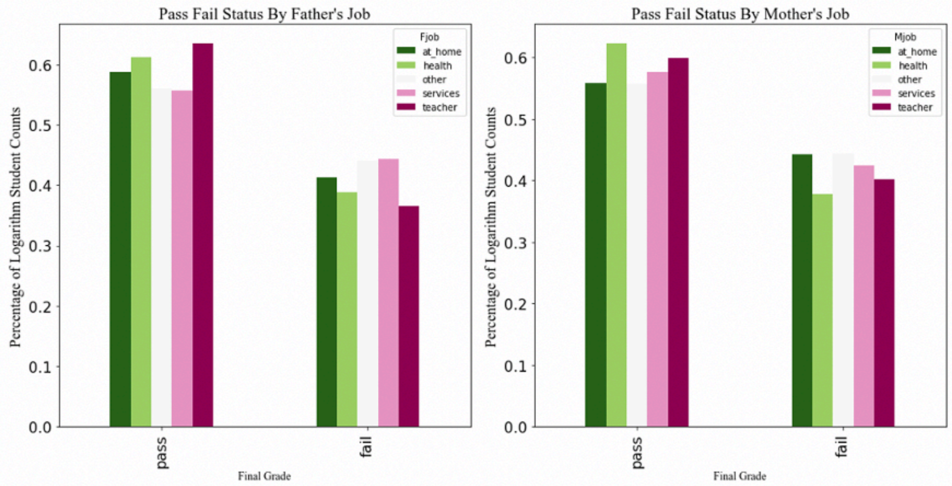
- Math score is more evenly distributed while the Portuguese score distribution is more right-ward skewed. This means students perform better in Portuguese than Math, suggesting Math is harder or Portuguese schools teach Math less well. The % of students who fail is quite high: 32.9% for Math and 15.4% for Portuguese. This means that the predictive modeling I'm doing is crucial for educators to intervene and improve students 'performance.

Figure 2:



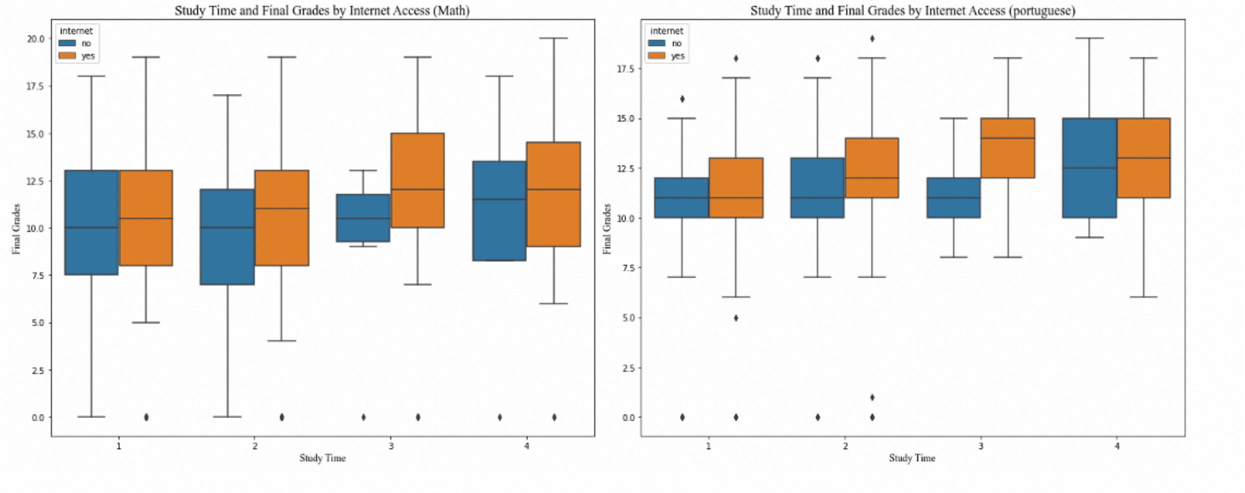
- Students have more absences in Math than Portuguese, which coupled with the lower performance in Math suggests that students are less interested in or struggle more in Math. More students who had higher absences were able to pass Portuguese than Math, suggesting that Math is harder than Portuguese. More research is needed to explore why students are absent more in Math, how or if they are struggling, and how to improve.

Figure 3:



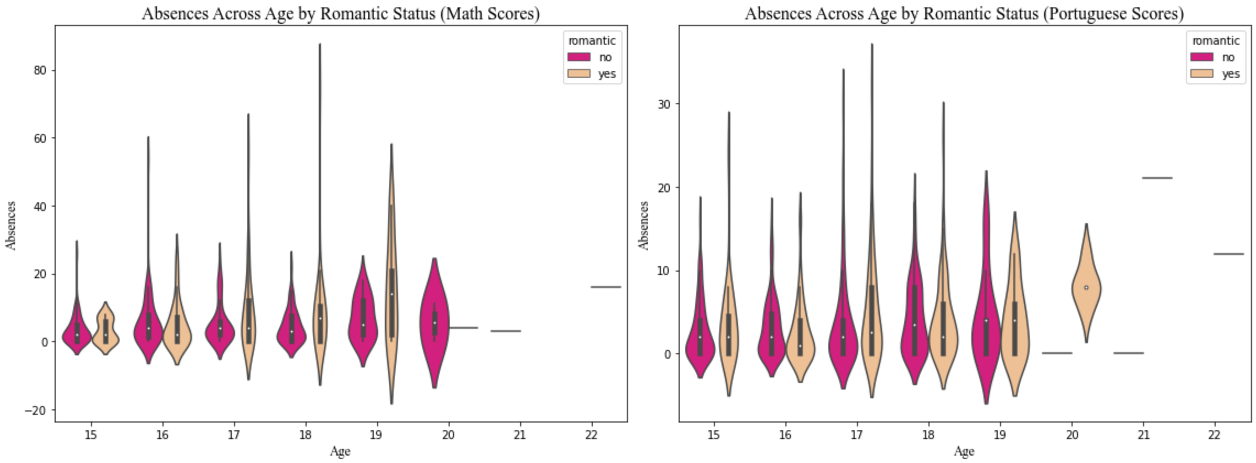
- There is correlation between the parents' job and students' performance. More students pass if their parents are in a "highly educated" field, such as health or teacher and vice versa. This makes sense, because the parents of these students probably have more earnings and knowledge to assist them, which points to socioeconomic inequal. There isn't a significant different between the number of students who pass and fail with respect to the parents' jobs, which suggests that both parents' jobs are equally important.

Figure 4:



-Students who have internet have higher grades than those who don't, which suggests that internet is important for studying. For students who study less in both subjects, there is a smaller difference in the final grades between those who don't have internet and those who do compared to students who study more. This suggests that strong students are more effective at using the internet for studying. There needs to be more research into what the students do on the Internet, how much time do use the internet for studying, and how schools can assist students in using the Internet for educational purposes.

Figure 5:



-Students who are in a relationship have more absences than those who don't, suggesting that relationships can distract students from school. In Math, students between 17 – 19 years old have the most absences and are in a relationship more, which confirms my assertion. In Portuguese, there are more students who are in a relationship than Math, and these students have more absences. This suggests that either students in Math are more "disciplined" or Math is harder and require students to study more.

3. Methods

- I **create a feature "pass_fail" which denotes that a student fail if their final score is below 10 and pass if otherwise.** This threshold is derived from the original paper. I make this my target variable and the rest of the features my feature matrix. I drop the final grade column because I want to predict the pass-fail status without knowing the final grade.
- **Splitting & preprocessing**: Both of my data sets are **IID, small, and don't have group structure, time series, and missing values.** Thus, I use a **basic train_test_split and a KFold split**. For preprocessing, since all the ordinal features have already been encoded, I only apply the **OneHotEncoder on the remaining non-bounded/ranked, categorical features and the MixMaxEncoder on the bounded continuous features**. I also applied a **LabelEncoder** to target variable (0=fail, 1=pass).
- For each dataset, I have **3 sub-models: Model 1 consists of all midterm and final scores, Model 2 consists of only the first midterm scores and final scores, and Model 3 consists of no midterm and only final scores.** This set up shows the predictive power of the models over the course of the semester to allow educators, parents, and policymakers to assist the students earlier in the semester.
- For each sub-model, I apply **3 ML algorithms: logistic regression (LR), random forest classifier (RFC), and support vector classifier (SVC)**. These algorithms work well with the nature of my datasets and complement one another: LR can capture non-linear relationships, provides smooth predictions, and is easy to interpret; RFC can provide higher accuracy through cross-validation, reduces overfitting through combining many trees, and handle higher dimensionality data; SVC can capture non-linear relationships and higher dimensionality data and is one of the most robust algorithms, though can be hard to interpret.
- **Evaluation metrics:** I choose accuracy score because my datasets are both balanced **(67.1% Class 1 for Math and 84.6% Class 1 for Portuguese)** and the accuracy score satisfies my goal to classify correctly the performance of as many students as possible. I also use confusion matrices and ROC curves.
- To measure uncertainties, I **loop through 10 random states** and **calculate the mean and standard deviation of the test scores** across the random states, to compare the performance of the model. For cross-validation, I also do a **confusion matrix and ROC curve** analysis.
- I **developed an ML pipeline using K-Fold Cross Validation (GridSearchCV),** which allows me to perform splitting, preprocessing, and hyperparameter tuning/ fitting the ML algorithm, and cross-validation efficiently.
- **Hyperparameter tuning**: For **LR, I use L1 regularization and a saga solver and tune C.** I chose Lasso because it shrinks the less important features' coefficients. I **tune C with 8 even values on a log space between 10^{-2} and 10^{2}.** For **RFC**, I use **n_estimators=100** because this is a large enough number of trees for accuracy but not too big that to slow the function. I tune **max depth with 9 values linearly spaced between 2 and 30 and max features with 5 values linearly spaced between 0.25 and 1.** I chose a max depth that is less than the number of features but still large enough to ensure accuracy and not to large that it would slow my function. For SVC, I use a

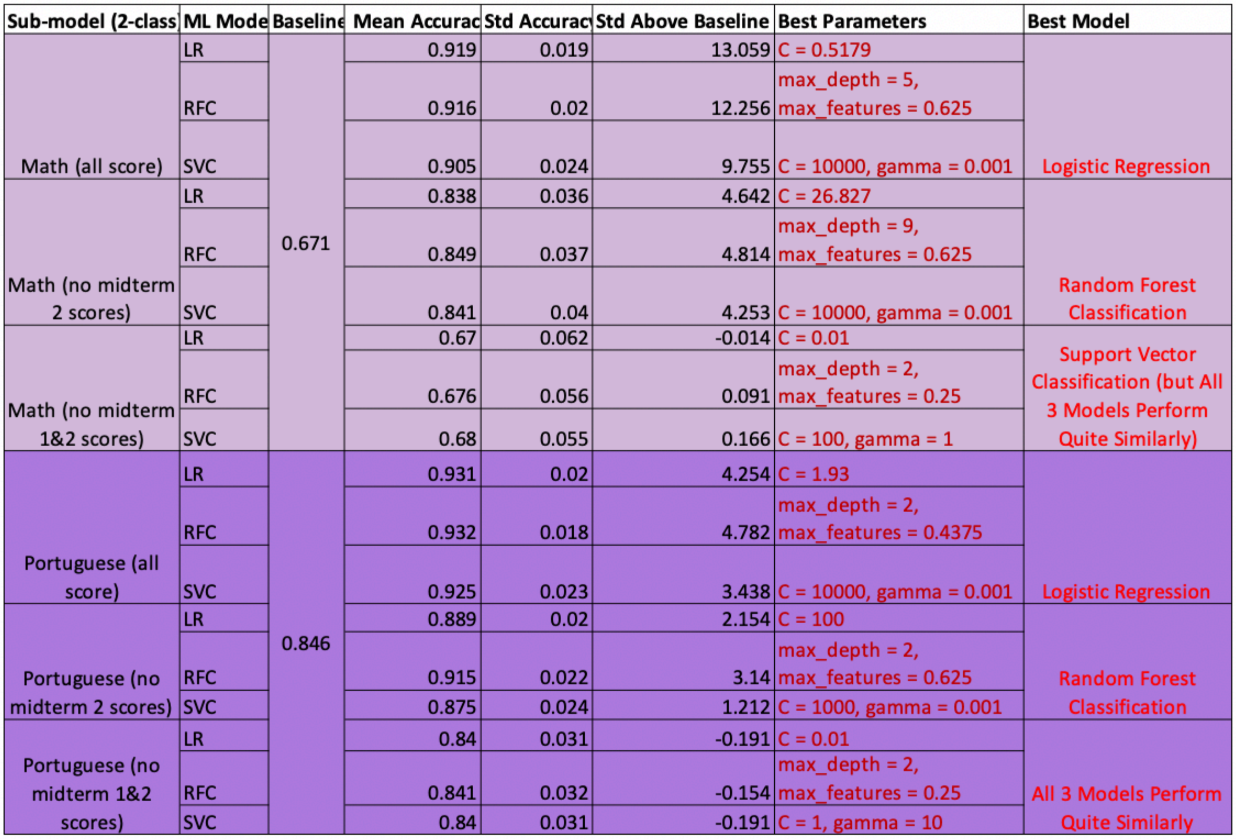
default RBF kernel and **tune C and gamma**, both with **8 even values on a log space between 10^{-3} and 10^4** to avoid edge cases.

- Prior to deciding on the exact range, **I've tried different ranges and use cv_results_** to print out the results to ensure that the range is wide enough (**I've seen both underfit and overfit**). I also adjust the parameters and range keeping in mind the computation time.
- Within each ML model, I also perform **feature importance** analysis. For LR, I perform **perturbation, linear coefficients, and SHAP(LinearExplainer).** Though I'm aware that SHAP isn't necessary for LR's feature importance, I still want to explore it for reference and learn a new approach. For RFC, I perform **perturbation, native feature importance metrics of random forests, and SHAP(TreeExplainer).** For SVC, I perform **perturbation and SHAP(KernelExplainer).** I wasn't able to gets coefficients because SVC with a linear kernel couldn't run on my datasets.

4. Results

- The table **summarizes the mean and standard deviation of the test accuracy scores and the number of standard deviations above the baseline for each ML algo.**

Figure 6: Summary table of results for the sub-models and ML algos

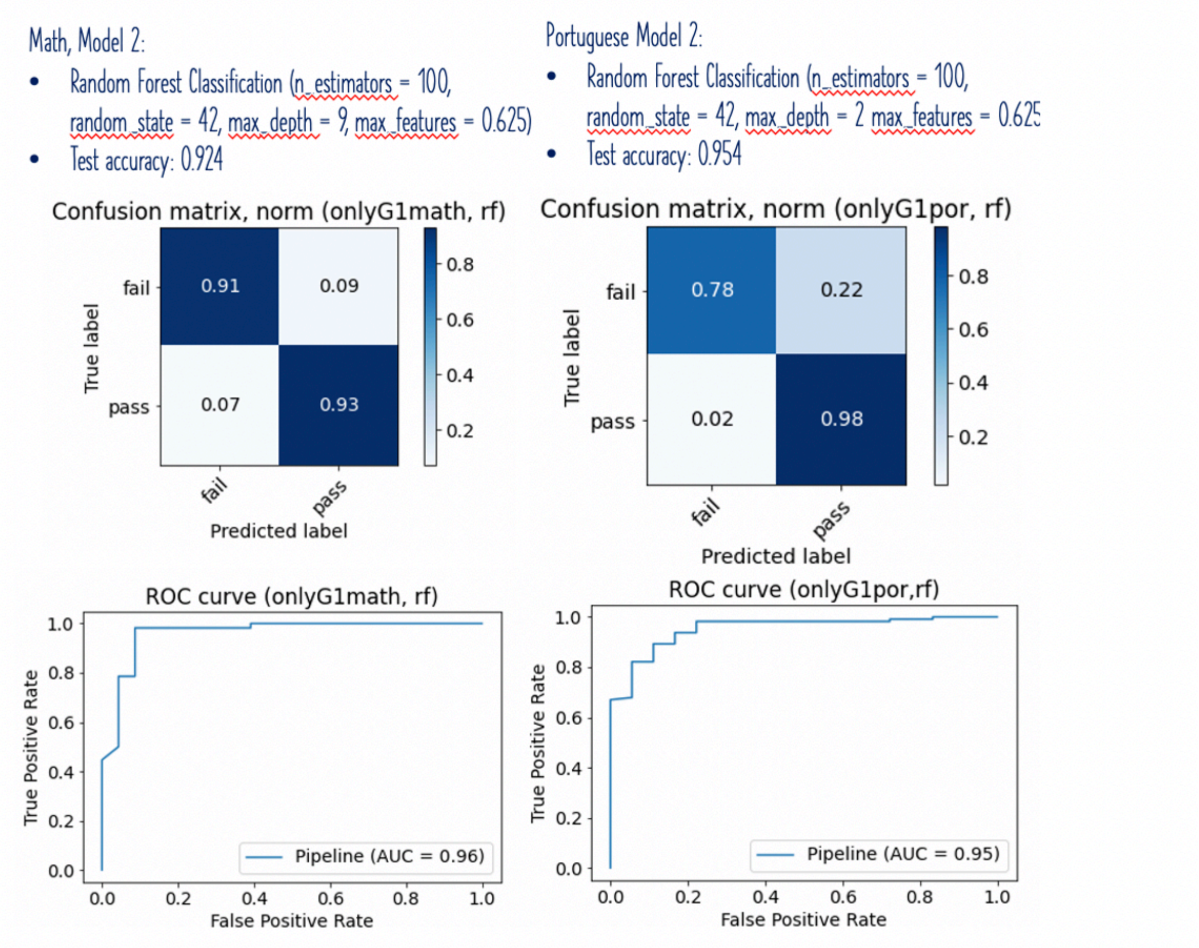


Sub-model (2-class	ML Mode	Baseline	Mean Accurac	Std Accurac	Std Above Baseline	Best Parameters	Best Model
Math (all score)	LR	0.671	0.919	0.019	13.059	C = 0.5179	Logistic Regression
	RFC		0.916	0.02	12.256	max_depth = 5, max_features = 0.625	
	SVC		0.905	0.024	9.755	C = 10000, gamma = 0.001	
Math (no midterm 2 scores)	LR		0.838	0.036	4.642	C = 26.827	Random Forest Classification
	RFC		0.849	0.037	4.814	max_depth = 9, max_features = 0.625	
	SVC		0.841	0.04	4.253	C = 10000, gamma = 0.001	
Math (no midterm 1&2 scores)	LR		0.67	0.062	-0.014	C = 0.01	Support Vector Classification (but All 3 Models Perform Quite Similarly)
	RFC		0.676	0.056	0.091	max_depth = 2, max_features = 0.25	
	SVC		0.68	0.055	0.166	C = 100, gamma = 1	
Portuguese (all score)	LR	0.846	0.931	0.02	4.254	C = 1.93	Logistic Regression
	RFC		0.932	0.018	4.782	max_depth = 2, max_features = 0.4375	
	SVC		0.925	0.023	3.438	C = 10000, gamma = 0.001	
Portuguese (no midterm 2 scores)	LR		0.889	0.02	2.154	C = 100	Random Forest Classification
	RFC		0.915	0.022	3.14	max_depth = 2, max_features = 0.625	
	SVC		0.875	0.024	1.212	C = 1000, gamma = 0.001	
Portuguese (no midterm 1&2 scores)	LR		0.84	0.031	-0.191	C = 0.01	All 3 Models Perform Quite Similarly
	RFC		0.841	0.032	-0.154	max_depth = 2, max_features = 0.25	
	SVC		0.84	0.031	-0.191	C = 1, gamma = 10	

- **I get the highest accuracy from Model 1 because the algo have more score information to predict the performance.** The accuracy score erodes over Model 2 and Model 3 because the algo have less score information to predict from. Model 3 performs just above the baseline. Though Model 1 performs the best, if teachers have to wait until the second midterm to know students' final performance, there will be limited actions and time to help the students. **Model 2 performs pretty well and allows schools and teachers to identify weak students early on to take necessary actions.**

- My CMs and ROC curves for Model 2 for both datasets confirms this, with high true positive and true negative rates.

Figure 7: Confusion Matrices and ROC Curves for Model 2 (no midterm 2 grades) of Math and Portuguese datasets



Math, Model 2:

- Random Forest Classification (n_estimators = 100, random_state = 42, max_depth = 9, max_features = 0.625)
- Test accuracy: 0.924

Portuguese Model 2:

- Random Forest Classification (n_estimators = 100, random_state = 42, max_depth = 2 max_features = 0.625
- Test accuracy: 0.954

- Features importance varies by model, but there is some commonality. **For Models 1 and 2, the midterm scores are the most important features.** This is quite obvious and isn't very informative for educators. Other important features beyond midterm scores that could help educators identify students' performance are **romantic status, internet access, the jobs of the parents, absences, and school supplement and fam supplement.** Besides that, some of the **least important features are health, family size & relationship, and travel time.** This means that to improve students' performance, educators should focus on helping students navigate adolescent relationships in a way that could positively impact their performance, work closely with parents of students with weak performance, and provide more supplement when necessary.

- I was surprised to see that romantic status is among one of the most important features. Coupled with low students' performance and my research that Portugal has high teenage pregnancies rates, I'm concerned that adolescent relationships can negatively affect students' performance. I'd research more on how Portugal can deliver better sex-ed or education about teenage relationship. I was also surprised that study time didn't regularly appear among the top 10 most important features. Regardless, I think that education on the students can use their time efficiently (especially internet time, per EDA) to study is vital. I was also surprised that the ML models perform poorly when no grades are provided. This suggests that apart from grades, other features in these data aren't great predictors of students' performance.

Figure 8: Some Global Feature Importance Graphics

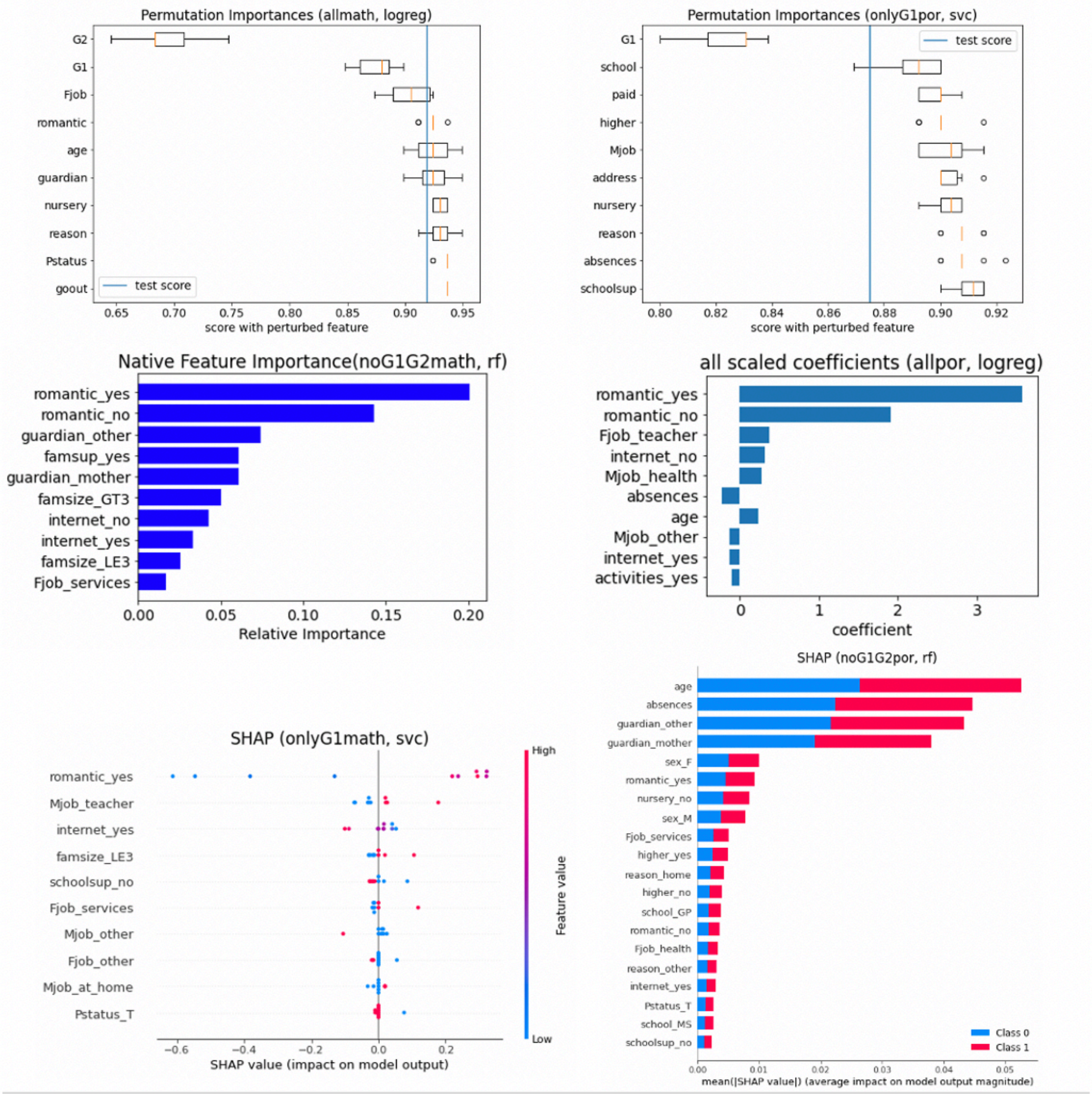
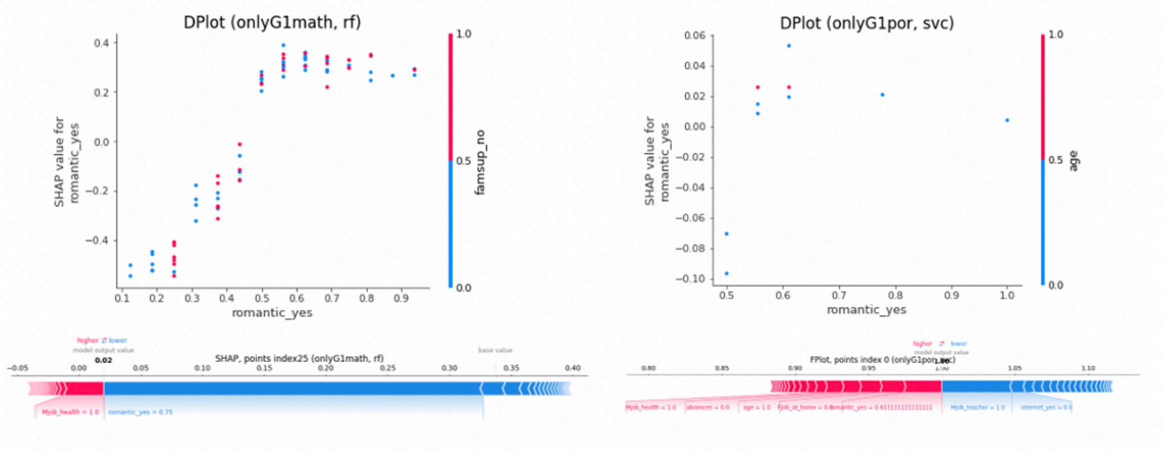


Figure 9: Some Local Feature Importance Graphics



5. Outlook

- **Weak spots:** I used all features for prediction, which while this meets the usual practice of data science, for social sciences purposes, it'd be better to use only some most important features to predict the students' performance. Due to having to build many-sub models, it was too computationally expensive for me to do so.
- **Improvement suggestions:** I'd do feature selection and use different combination of features to generate more specific models to see if I can get higher accuracies. I'd also do feature engineering based more detailed EDA to improve the model's performance. I'd also explore a multi-class classification which classifies the students into more specific levels of performance beyond just pass fail to help teachers assist the students better. I'd test more complex ML models such as XGBoost, PCA, KNN.
- **Additional info:** features such as homework scores, school attitude, prior exposureto/enjoyment of the subjects, might be great predictors of the students' performance that the data is missing. I'd also like to collect more continuous or ordinal features because the current data has a lot of categorical features. To expand the scope of this project beyond just Portuguese secondary students, I'd like to collect data and do analysis on students' performance in more countries, school levels, and subjects. I'd also use more holistic measures for academic performance, such as average grades across the school year or school attitude.

6. Reference

- https://archive.ics.uci.edu/ml/datasets/student+performance

- 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.

7. Github Repo: https://github.com/annieptba/data1030_project.git