**A MACHINE LEARNING PROJECT: PREDICTING STUDENT PERFORMANCE**

**Annie Phan**

**DATA 1030, Brown DSI**

[**https://github.com/annieptba/data1030\_project.git**](https://github.com/annieptba/data1030_project.git)

**1. Introduction**

- The project examines how social and education factors impact students’ performance to allow educators to intervene as necessary. The question I explore with my ML model is whether I can predict if students will pass or fail based on information such as family background, socioeconomic status, previous grades, etc.

- 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 not only helps Portuguese educational institutions and government find attributes to best invest in to improve students’ performance or identify students that need assistance, but also provides data for 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 Machine Learning Repository and my project is based on the paper named “Using Data Mining to Predict Secondary School Student Performance” by Cortez and Silva in 2008. I work with two datasets for students Math and Portuguese scores. There are 395 in Math and 649 students in Portuguese. Both datasets have 33 features, covering information such as the students’ age, romantic status, internet access, parent’s jobs, past classroom failures. The measures of academic performance are grades achieved in the classroom on a scale of 0 to 20 points. There are 2 midterm grades and 1 final grade.

- My model is a classification problem on the pass-fail status of the students, where the students pass if they score 10 and above and fail otherwise.

**2. Exploratory Data Analysis**

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

**Figure 1:**

Chart, histogram

Description automatically generated

- The math score distribution is relatively even on two sides while the Portuguese score distribution is more right-ward skewed. This means these students have better performance in Portuguese than math. Though in both subjects there are more students who pass, the % of students who fail is still quite high: 32.9% for Math and 15.4% for Portuguese. This means that the kind of predictive modeling I’m doing is crucial for educators to intervene and improve students ‘performance.

**Figure 2:**

Chart, histogram

Description automatically generated

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

**Figure 3:**

Chart, bar chart, histogram

Description automatically generated

- There seems to be 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. There isn’t a significant different between the number of students who pass and fail with respect to the mother or father’s job, which suggests that both parents’ jobs are equally important. This suggests some degree of socioeconomic inequality in education.

**Figure 4:**

Chart, box and whisker chart

Description automatically generated

-Overall, students who have internet have higher grades than those who don’t, which suggests that internet is quite 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 they dedicate their internet use to studying, and how schools can assist students in using the Internet for more education purposes.

**Figure 5:**

A picture containing chart

Description automatically generated

-Overall, students who are in a relationship have more absences than those who don’t. This suggests 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 also 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**

- For classification, I **create an additional feature called “pass\_fail” which denotes that a student fail if their score us below 10 and pass if otherwise.** This threshold is derived from the original paper, which reflects the true threshold for pass and fail in the Portuguese secondary school the data was collected from. I make this feature 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, relatively small, and don’t have group structure, time series, and missing values.** Thus, I use a **basic train\_test\_split and a KFold split** to estimate uncertainty due to random splitting of the train and validation sets. For preprocessing, since all the ordinal features have already been encoded, I only need to 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 because it needs class labels.

- For each dataset (Portuguese and Math scores), 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 intervene in the students’ performance earlier on in the semester.

- For each sub-model, I apply the following **3 ML algorithms: logistic regression (LR), random forest classifier (RFC), and support vector classifier (SVC)**. I chose these algorithms because they work well with the nature of my datasets and they complement each other: LR can capture non-linear relationships but is simple and quick to implement, 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 and accurate algorithm, though can be hard to interpret.

- **Evaluation metrics:** I choose accuracy score because my datasets are both balanced (0.671 Class 1 for Math and 0.846 Class 1 for Portuguese) and the accuracy score satisfies the goal of my project, which is to classify correctly the performance of as many students as possible. The accuracy score also works well with the confusion matrix, which shows me how many students my models predict correctly the performance of.

- To measure uncertainties due to splitting and non-deterministic ML models, I **loop through 10 random states** and **calculate the mean and standard deviation of the test scores** across the random states, which allows me to compare the performance of the model. For additional cross-validation, I also do a **confusion matrix (CM)** analysis.

- For the splitting, preprocessing, and hyperparameter tuning/ fitting the ML algorithm, I **developed an ML pipeline using K-Fold Cross Validation (GridSearchCV),** which allows me train learners using one set of data and testing them using another set and perform both cross-validation and hyperparameter tuning 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, which indirectly performs feature selection. I **tune C with 8 values spaced evenly on a log space between 10-2 and 102.** For **RFC**, I use **n\_estimators = 100** because this is a good balance between a large enough number of trees to increase accuracy but not a too big number that could slow down the function significantly. 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. This because after encoding, I have around 40 – 50 features, so I need 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 down my function. For SVC, I use a default RBF kernel and **tune C and gamma**, both with **8** **values evenly spaced on a log space between 10-3 and 104** to avoid edge cases. Prior to deciding on the exact range of these parameters, I have tried with different range of parameters and use cv\_results\_ to print out the results to ensure that the range I choose is wide enough (I have seen both underfit and overfit). I also adjust which parameters to tune and range to set keeping in mind the computation time.

- Within each ML model, I also perform **global and local feature importance (FT)** analyses. For LR, I perform FT through: perturbation, linear coefficients (adding a StandardScaler to ML pipeline), and SHAP (LinearExplainer). Though I’m aware that SHAP isn’t necessary for LR’s feature importance, I still want to explore this method for reference and learn a new approach. For RFC, I perform FT through: perturbation, native feature importance metrics of random forests, and SHAP (TreeExplainer). For SVC, I perform FT through perturbation and SHAP (KernelExplainer). I wasn’t able to gets coefficients for SVC because SVC with a linear kernel wasn’t able to run on my datasets.

**4. Results**

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



**-** As expected, I get the highest accuracy from Model 1 because the algo have more score information to leverage from 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 has very low accuracy scores, with the models performing just right at the baseline. Though Model 1 performs the best because the midterm scores are good indicators of the students’ final performance, if the teaches have to wait until the time of the second midterm to know students’ final performance, there will be limited actions and time available for the students to help the students. Model 2 performs pretty well, and its results fit into an academic context because it allows schools and teachers to identify weak students early on to take necessary actions.

- This is confirmed by my CMs and ROC curves for Model II for both the Math and Portuguese datasets, as we see that the model has quite high true positive and true negative rates.

**-** Features importance varies by model, but generally, there are 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 who have 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 relatively low students’ performance in Portuguese and from my research that Portugal has one of the highest teenage pregnancies in Europe, I’m concerned that adolescent relationships might be negatively affecting the students’ performance. I’d like to do more research into how Portugal can deliver better sex-ed or education about teenage relationship. I was also surprised to see that study time didn’t regularly appear among the top 10 most important features. Regardless, I still think that education on the students can use their time efficiently (and especially internet use time, as EDA suggests) to study is vital I was also surprised to see that the ML models perform very poorly when no grades are provided. This suggests that apart from grades, the features presented in these data may not be the best predictors of the students’ performance.

**5. Outlook**

**- Weak spots:** I used all features for prediction in my models, which while this meets the usual standard practice of data science, for social sciences purposes, it’d be better to use only some selected most important features to predict the students’ performance. Due to the scope of this project (building many-sub models), it was too computationally expensive for me to do feature. Moreover, since my data has many categorical features, some FT analyses

- **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 analysis to improve the model’s performance. I’d also want to 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. For these questions, I’d test more complex ML models such as XGBoost, PCA, KNN, etc.

- **Additional info:** features such as homework scores, school attitude, prior exposure to the subject, enjoyment of the subjects, might be great predictor 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, across different school levels, and more subjects. I also want to revise the method of using only final grades as a measure of academic performance, because education goes beyond grades, and more holistic measures such as average grades across the school year or school attitude should be considered.

**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.