**A MACHINE LEARNING PROJECT:**

**PREDICTING PORTUGESE SECONDARY SCHOOL STUDENT PERFORMANCE**

**Annie Phan (Banner ID: B01309278)**

**DATA 1030 – Fall 2020, Brown University Data Science Initiative**

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

**Tertiary: https://github.com/annieptba/data1030\_fa20\_final\_project.git**

**1. Introduction**

- Follow the original research paper

- The dataset is interesting/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 (BI)/Data Mining (DM) to develop automated tools that can improve decision making and optimize success in education. For instance, some interesting questions for this domain that could be answered using BI/DM techniques: What type of courses can be offered to attract more students? Is it possible to predict student performance? What are the factors that affect student achievement? In my model, I seek to explore these similar questions but look further into demographic factors such as family support, romantic relationships, alcohol consumption, and internet access.

- This data contains 1,044 instances (students) including 395 Mathematics class students and 649 Portuguese language class students. There are 33 features. The dataset is already well-described and can be found here: <https://archive.ics.uci.edu/ml/datasets/student+performance>

- In the original research paper named “Using Data Mining To Predict Secondary School Student Performance” by Cortez and Silva in 2008, the two datasets were modeled under binary/five-level classification and regression tasks: “i) binary classification (pass/fail); ii) classification with five levels (from I very good or excellent to V - insufficient); and iii) regression, with a numeric output that ranges between zero (0%) and twenty (100%).” The results show that the students’ final grades can be predicted by the first and/or second school period grades and also other relevant features (e.g. number of absences, parent’s job and education, alcohol consumption). Another project that tackled this problem was the Michigan State University which also modeled the problem using three classification approaches: “binary: pass/fail; 3-level: low, middle, high; and 9-level: from 1 - lowest grade to 9 - highest score” (Minaei-Bidgoli et al. 2003). The best solution was obtained by a Naive Bayes method with an accuracy of 74%. It was also found that past school grades have a much higher impact than demographic variables (ibid, 2003). There are other researches that have touched on this problem, but I think that the two researches highlighted here are most important. I base my regression and classification problems on the approaches in these researches.

**2. Exploratory Data Analysis**

Chart, histogram

Description automatically generatedChart, bar chart, histogram

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, histogram

Description automatically generatedA picture containing chart

Description automatically generated

- I chose to show the distribution of student’s performance to have an idea for the classification problem. The figure demonstrates distribution of students across different final grades performance. Around 60% of the students have a fair performance, but 20% of them have a poor and 20% have good performance. This is a very high % of students have a poor performance and fit with the idea that many Portuguese students were underperforming.

-I chose to explore the relationship between study time and desire for higher education of children across gender to see study patterns and desires across gender. The figure demonstrates distribution of students' study time across different age, classified into whether the student wants to achieve higher education or not. It seems that the older the students get, the more they want to achieve higher education, but also overall, however, the distribution of students who don't want to achieve higher education is also quite close to those who want to, which raises some concerns about how the Portuguese school system and other social factors influence students' choice to achieve higher education. There is also a positive relationship between the students' study time and whether they want to achieve higher education. This makes sense because more motivated students will want to achieve higher education. However, this alerts that more measures should be taken to incentives students to study more and year higher education in a way that balance their well-being and happiness as well.

-I chose to explore the relationship between romantic status and students’ performance since the dataset covers students of adolescent ages, a group that tend to be “distracted” from school due to desires for romantic relationships. The figure demonstrates distribution of students' relationship status across different final grade performance. It seems that more students who don't have a relationship have better final grade performance and vice versa, though the disparity isn't significant. This signals that students should not be in a relationship if they want to do well in school, but since the difference is not big, intervention measures need to be sensible and cognizant of the students' social development.

-I chose to explore the relationship between alcohol consumption and students’ performance since the dataset covers students of adolescent ages, a group that tend to be more frivolous and excited to drink since they might just be approaching the drinking age. The figures demonstrate final grades across measures of distributions of students' weekend and weekday alcohol consumption, classified by gender. There are some expected observations here. Students who have less alcohol consumption tend to get higher grades and vice versa, and more male students consume alcohol than female students. Students drink significantly more on the weekend than weekday. This suggests that some intervention and educational efforts should be executed to teach students about appropriate drink habits and how it can affect their study.

**3. Methods**

- by creating an additional column called “pass\_fail” which categorizes the students into ‘pass’ performance if they score 10-20, and ‘fail’ if they score below 10. This variable is my target variable for the classification problem. Separate feature matrix and target variable. I drop the “G3” column because I don’t need it for classification and keeping it could skew my results.

- For my classification problem, I create an additional feature called “performance” which categorizes the students into ‘good’ performance if final score ia from 15 to 20, ‘fair’ if they score from 10 to 14, and ‘poor’ if they score below 9. My 3-level classification approach is a middle ground between the binary and 5-level classification approach in the original paper, which provides more insights into the students’ performance than a simple pass-fail problem but isn’t too computationally expensive to compute like the 5-level approach.

- Splitting strategy & 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 (“performance”) since this is a classification problem and the target variable needs to have class labels.

- For each dataset (Portuguese and Math scores), I have **3 sub-models: Model 1 consists of all midterm and final scores (G1, G2, G3), Model 2 consists of only the first midterm scores (G1) and final scores (G3), and Model 3 consists of no midterm and only final scores (G3).** This set up shows the predictive power of the models over the course of the semester to allow educators, parents, students, and even 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 altogether, they provide a good mix of pros and cons that 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 metric: I choose accuracy score because my datasets are both balanced (0.486 for Math and 0.644 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 allows me to see how many students my models predict correctly the performance of,

- To measure uncertainties due to splitting and non-deterministic ML models, I loop through 5 random states and calculate the mean and standard deviation of the test scores in each random state. Initially, I looped through 10 random states, but found that the computation time was taking too long and that there isn’t a huge difference between the mean and standard deviations of the test scores between these to numbers. Moreover, most of the iteration that gave me the best test score for each model lies within 5 random states.

- 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 (Lasso) and a saga solver and tune C. I chose Lasso because it shrinks the less important features’ coefficients, which works indirectly as a way of 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 undefit and overfit within the range for each of the parameter I chose). 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) and confusion matrix (CM)** analyses. For LR, I perform FT through 3 approaches: perturbation, linear coefficients (adding a StandardScaler to my 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 to have another method for evaluation but also learn a new approach. For RFC, I perform FT through 3 approaches: perturbation, native feature importance metrics of random forests, and SHAP (TreeExplainer). For SVC, I perform FT through 2 approaches: perturbation and SHAP (KernelExplainer). I wasn’t able to perform the coefficient FT approach for SVC because SVC with a linear kernel wasn’t able to run on my datasets, so I used the default RBF kernel.

**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

model 1 does best followed by 2 nd 3. 2 gives the most actianlbe insighs, fits into academic context, model 3 above baseline.

**-** Features importance vary by model, but generally, here are some commonly appeared. Midterm scores (G1 and G2) are definitely (and proven by the models) are most important features. Apart from that, romantic status appear the most frequent, followed by internet status, MJob and Fjob (servies, teacher), absences, School sup and fam sup are most important. Least important features are health, famsize, famrel, traveltime. Focus more on doing these things, improve romantic, internet, working with parents.

**-** Unexpected: Surprising – romantic status seems to be the most important – consistent with EDA but didn’t expect it to be so important. EDA shows alcohol consumption was quite important, didn’t see it in model. But seems like family background and support matters a lot. studytime

**5. Outlook**

- weakspot: used all features to predict, which meets the usual standard practice of data science, but for social sciences purposes, data has many categorical features, may not work best with these kind of feature importance

- to improve the model: may be better to use only the most important features, as suggested by FT analysis, for prediction. In the future, would like to do feature selection and use different combination of features to generate even more models. Feature engineering should also be done. Due to the scope of this project (building many-submodels), it was too computationally expensive for me to do feature engineering to avoid extra noise

- can also test with more robust ML techniques such as XGBoost, PCA, QDA LDA, Naïve Bayes to test, though my guess is that these models might be more complicated than I actually need for this data.

- additional info: homework scores, school attitude, more continuous or ordinal features that