Analysis of Student Performance in Portuguese Schools

Introduction and Motivation:

As a previous high school teacher, I was always looking to find ways improve my student’s performance. High schools have different ways of measuring student success, like graduation rates, Ivy league admissions, or AP scores. My previous high school’s goal was to promote college acceptance by looking at metrics such as homework, quiz, and test results that would ultimately lead to better final grades. However, from my experience in the classroom, I found that other factors such as developing relationships with students, parent involvement, and one-on-one time (via office hours) made students more eager to improve their high school grades and ultimately excited to attend college. Through machine learning and data analysis, I will determine what features other than assessment scores are important in predicting a student’s final grade and at what accuracy do these features predict a passing or failing grade.

Problem definition and Solution:

This project uses a random forests classifier to predict whether a student passes or fails the final assessment of the year. It is implemented using python with sci-kit learn, numpy, matplotlib, and panda packages on a Jupyter notebook. The input will consist of 28 features consisting of student habits, family life, extracurricular activities, free time, and more. The output will be a binary number, pass and fail, of the final 3rd period grade.

Data:

The dataset was taken from “Using Data Mining To Predict Secondary School Student Performance” from Paulo Cortez and Alice Silva at the University of Minho, Portugal. It consists of 395 students from 2 different Portuguese schools. Note that United States student data is usually protected and difficult to come by. The original dataset included two features, weekly and daily alcohol consumption, that were removed since alcohol consumption is not relevant to the motivation of my project. Also the social norms of alcohol consumption in Portugal is not the same in America. Since the motivation is to find which factors, such as student habits and family life, are important in student success, previous grades were removed. Also the first two period grades, G1 and G2, are obvious indicators of how well the student would perform on G3. All features that are not already binary or numeric are converted by means of categorical coding. The output feature is the final assessment of the student on a scale of 0 to 20. According to the Portuguese scale, anything above a 10 is considered passing.

Machine Learning Design and Implementation:

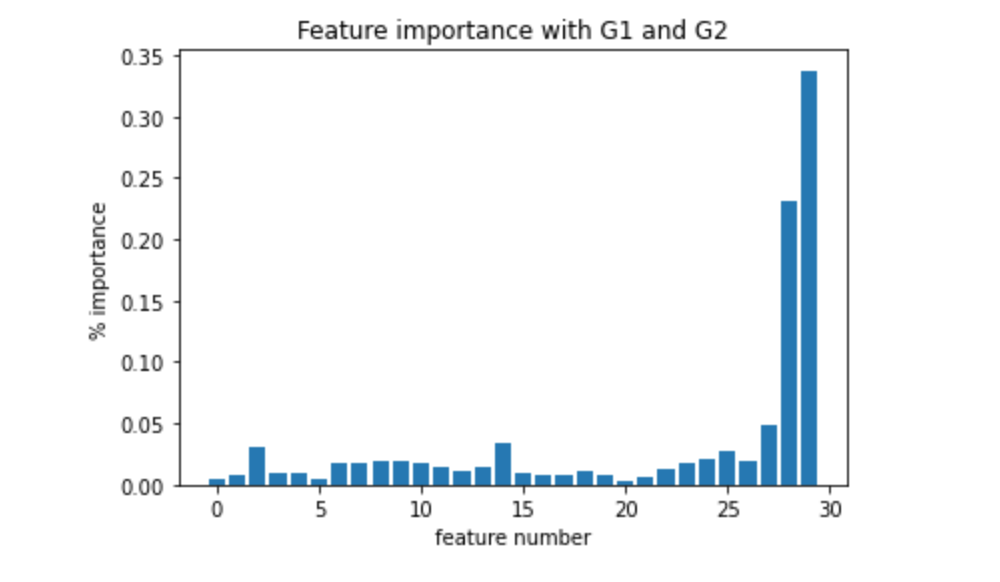
Random forest classification is an ensemble learning algorithm that randomly chooses a division of the training data and performs a system of decision trees. It totals the decision tree outputs to choose whether, in this project, the student passed or failed. The advantage of a random forest classifier is that it models very quickly, usually has high prediction accuracy, implements easily, and allows for feature importance analysis.

To optimize the dataset, any categorical/nominal data needs to be converted to numerical data. For instance, a feature in this dataset is the reason the school was chosen, where there are 4 categorical options to choose from. Through categorical coding, each of the options is represented as a value from 1 to 4. Binary data that was categorical, for instance whether the student was male or female, was changed to either 0 or 1.

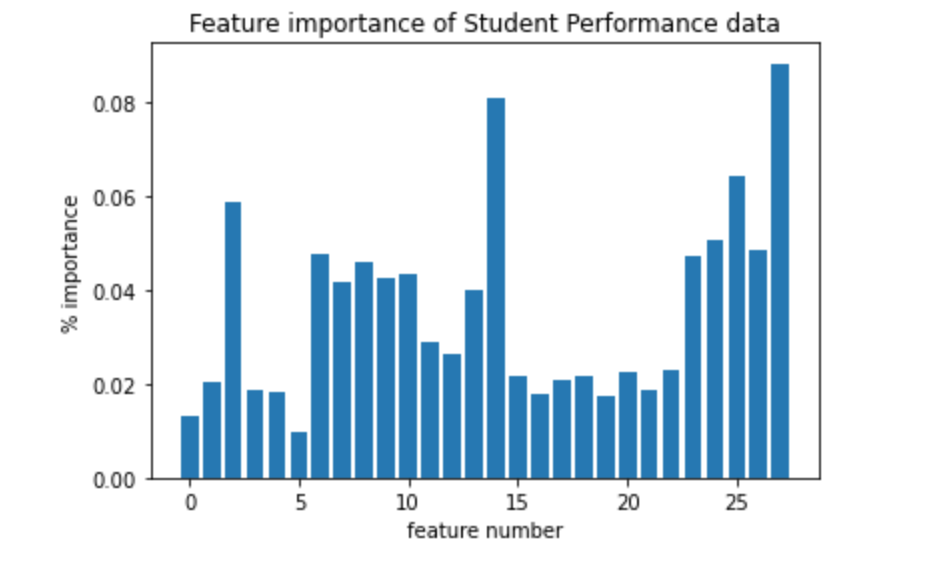
I used the Random Forest Classifier package from sklearn. I split the data into 80% training and 20% testing, and I specified 128 trees to be used. I adjusted this value starting at 50 and going up to 200, and 128 trees produced the best result. Using the accuracy score function from skelarn metrics, I received an accuracy score of 72%.

Results:

To first show that the student scores, G1 and G2, are heavily biased in predicting the G3 grade, a random forest classifier and feature importance analysis was performed. The final test prediction accuracy is 92%, and the feature importance percentage of G1 and G2 (shown in graph 1) are 23% and 34% respectively. After removing the two test grades, the accuracy dropped to 76%. After performing another feature importance analysis, the variables with most importance are (shown in graph 2): absences (feature 27, 8.8%), failure (feature 14, 8.1%), how frequently they go out with friends (feature 25, 6.4%) age (feature 2, 5.9%), family relationship (feature 23, 4.7%), free time (feature 24, 5.1%), Mothers job (feature 8, 4.6%), reason they chose the school (feature 10, 4.4%) and father and mothers education (feature 6 and 7, 4.3%). Low performing features include: sex (feature 1, 2%), parents cohabitation status (feature 5, 0.09%), wanting to take higher education (feature 20, 2.3%), and do they have internet (feature 21, 1.8%).



Graph 1: Feature importance on dataset including test grades 1 and 2 (ylabel is in decimal format)



Graph 2: Feature importance on dataset without test grades 1 and 2 (y label is in decimal format)

Conclusions:

With the G1 and G2 test scores included in the analysis, I can understand why rigorous course material, quiz results, and test performance are important to the high school I taught at. However, the second round of feature importance analysis is significant because it shows where we can improve as either teachers or parents. Family relationship was in the middle of the pack in importance, but it can affect many of the other features including absences. Absences were the highest feature at 8.8%. Knowing how important school attendance is, we as teachers need to motivate our students by developing more positive relationships. The frequency of past failed classes affected student performance by 8.1%. Failure continues to propagate through grade levels, and so students need to be engaged and start deep learning at an earlier age. As a past chemistry teacher, I had so many students not do well in dimensional analysis because they didn’t learn basic algebra in their formative years. It is also important to give our students free time after school. I would like to see further research be done about how much homework should be given to our students. With this information, more testing can be done by gathering more student data with a focus on how early on students start to fail, cause of absences, how to give students more free time, and more.

Some of the feature that I expected to be higher in importance ended up being some of the lowest performing features, such as wanting higher education or having internet or cohabitation status. It needs to be taken into account that Portugal has many cultural differences, and the country may have access to more public internet. However, knowing this information, teachers and parents should push all students to work hard and study regardless of the student’s higher education expectation.

To further the research of student performance, I would like to zoom in on G1 and G2 performance. More data can be gathered to analyze what features led to those successes, such as quizzes, homework completion, and student engagement in class. In conclusion, while homework and test scores are important, we can look beyond them to help further student performance.

Sources:

<https://archive.ics.uci.edu/ml/datasets/Student+Performance>

<https://builtin.com/data-science/random-forest-algorithm>