Predicting Student Performance

Tatiana Ediger, Michael Lamontagne

Abstract

In this study, we predict secondary school students' performance in Math and Language classes based on various demographic, social, and academic aspects of the student's experience. We apply machine learning techniques involving both regression and classification to examine this question.

Introduction

We are planning to address the problem of student's performance in select secondary education classes given demographic, social and academic factors. In particular, we look at data from Portugese and Math classes in Portugal. This problem is relevant, because in order to understand how students learn, it is important to examine the different factors that affect their academic performance. Possible applications could be advice to parents and teachers on how they could improve student performance, how to make education fairer for all regardless of their circumstances, what advice should be given to students if they're looking to improve academic performance, and predicting what students will need extra support.

Education has always been an important issue to investigate. In the increasingly fractured modern world, it is important to examine how different aspects of a student's socioeconomic status can impact their academic success. It is also important in an increasingly virtual world to examine how social aspects of a student's experience can influence their academic success.

Technical Approach

We trained machine learning models to predict student performance in secondary education based on various demographic, school, and social-related features. We used both regression and classification to solve the problem.

We used both Ridge and Lasso regression algorithms to model student performance as a continuous output. We used ridge regression to avoid any overfitting of the data and we used lasso to see what, if any features, were more or less relevant to student performance.

We also used several classification algorithms. First, to define usable classes, we binned the output grade into four classes. Because the dataset is limited, it is impossible to accurately perform classification for the original twenty-one classes. Binning the target variable provides fewer target classes and less information from the results, but it does allow the model to accurately predict student performance.

To perform classification, we used three models. The first was Support Vector Machine, which allowed us to separate the data and make predictions on it. We also used Decision Trees. Our data had a large number of categorical features, so we took advantage of a Decision Tree's unique ability to handle these sorts of features. We also used Neural Networks as another method to predict student performance.

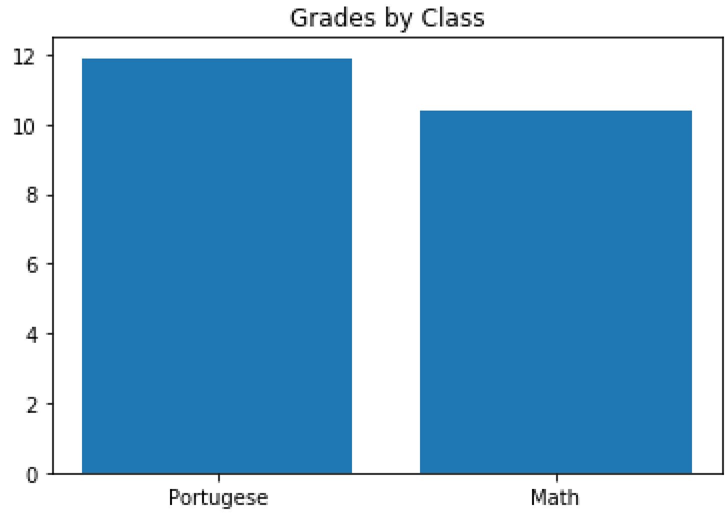
Experimental Results

The data we used to predict student performance was the Student Performance Dataset from the University of California Irvine's Machine Learning Repository. This dataset includes information on students' grades in secondary education in two Portugese schools. There were 21 classes in our dataset, representing final grades in the range of 0 to 20. After we binned the data, there were 230 in class 0, 153 in class 1, 367 in class 2, and 294 in class 3. Before binning, the samples per class were: [53,1, 0, 0, 1, 8, 18, 19, 67, 63, 153, 151, 103, 113, 90, 82, 52, 35, 27, 7, 1].

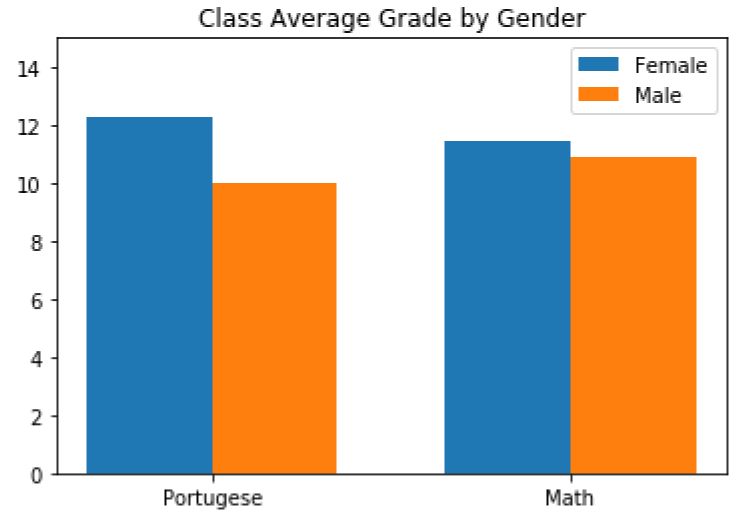
We used two different techniques to preprocess the data. The first was binning the targets to allow us to accurately predict performance. Many features in this dataset were not numerical, though. These features were categorical. To address the categorical nature of these features, and to be able to use them in our numerical models, such as regression and support vector machine, we needed to apply one-hot encoding to them. This split each of these categorical features into several features that reflected, with a binary choice, whether a particular category defined in this feature applied to that sample.

Originally our dataset was composed of two different datasets which all contained the same features and outputs, but one was for Math class and one was for Portugese class. We combined these two together, and added another feature representing whether it was

Portugese or Math class. There were 31 different features in our dataset. These different features were: student’s school, sex, age, home address, family size, parental cohabitation status, mother’s education, father’s education, mother’s job, father’s job, reason to choose school, student’s guardian, home to school travel time, weekly study time, number of past class failures, extra educational support, extra family support, extra paid classes within the subject, involvement in extra-curricular activities, attended nursery school, wants to take higher education, internet access at home, with a romantic relationship, quality of family relationships, free time after school, going out with friends, workday alcohol consumption, weekend alcohol consumption, current health status, number of school absences, and whether they’re in Math or Portugese class.

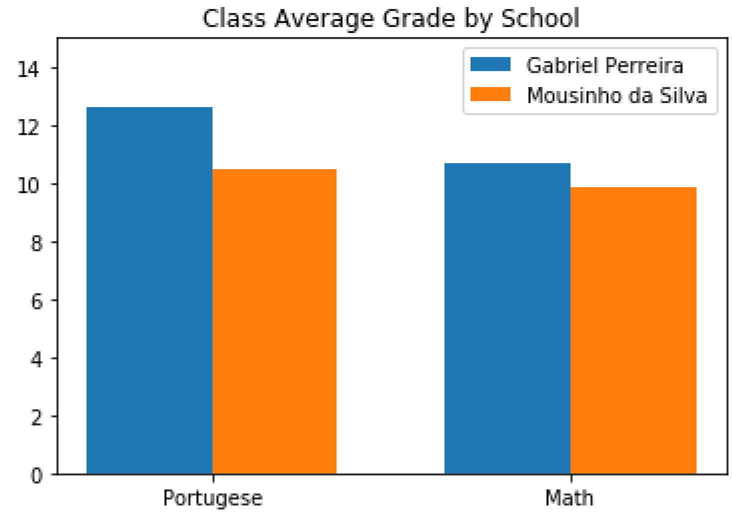


Some features were important to defining the dataset. It is important to note that students tended to perform much better in Portugese class than their math classes. This is notable, as it indicates that on the whole, students performed well in their language classes and poorer in their math classes, which can have an impact on the results.

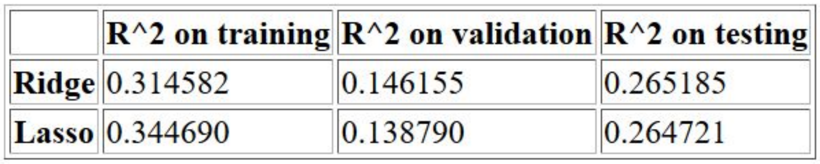


Another one of these features was gender. As we can see, gender went a long way to explaining certain disparities in the grades based on demographic circumstances. In language class, it is clear that females outperform males by a significant margin of more than two grade categories using the pre-binned classes. In math class, female students still outperform their male counterparts, but by a much smaller amount.

Another interesting revelation is that female students tend to do better in language class than they do in math class, and male students tend to perform better in math than in language classes. This addresses a very topical issue about gender inequality in education.

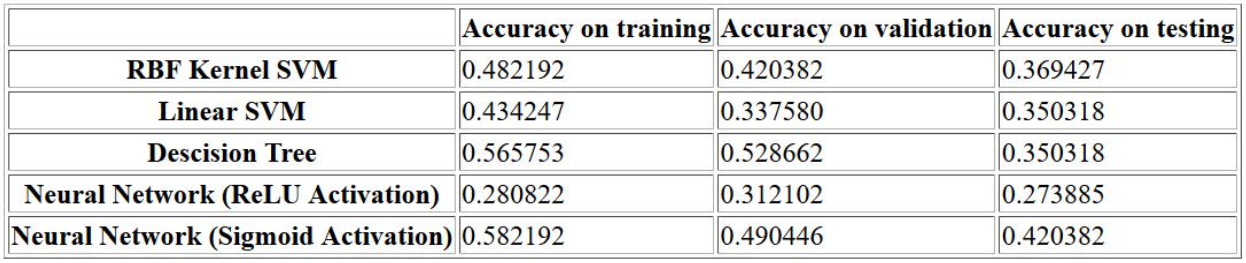


Another important feature is the school. Across the board, students at Gabriel Perreira outperform their peers at Mousinho da Silva. This indicates that students who are at Gabriel Perreira have a higher chance of succeeding than their counterparts. This highlights an important disparity between these two schools.



	R^2 on training	R^2 on validation	R^2 on testing
Ridge	0.314582	0.146155	0.265185
Lasso	0.344690	0.138790	0.264721

For our regression algorithms, we split up our data into training/validation/testing sets, with a 70%-15%-15% split, setting our target variable to be the final grade (G3 in our dataset). We then performed hyperparameter tuning for alpha for both Ridge and Lasso regression. We created a list of parameters: [0.001, 0.01, 0.1, 1, 10, 100], and trained models with these different alpha values on our validation set. We found the best alpha value to use by finding the best r^2 value on our validation set. For Ridge, we found that the best alpha value was 100, and for Lasso it was 0.01. We then used this best alpha value with our training and testing datasets, and compared their r^2 values. We found that when we applied our model to the testing dataset, we got comparable r^2 values for both Ridge and Lasso, at around 0.26. On the training datasets, the r^2 value for Ridge was 0.315, compared to Lasso, which had an r^2 value of 0.345. This indicates that there was slight overfitting when we used Lasso. Still, these are promising results for the effectiveness of the algorithm.



	Accuracy on training	Accuracy on validation	Accuracy on testing
RBF Kernel SVM	0.482192	0.420382	0.369427
Linear SVM	0.434247	0.337580	0.350318
Descision Tree	0.565753	0.528662	0.350318
Neural Network (ReLU Activation)	0.280822	0.312102	0.273885
Neural Network (Sigmoid Activation)	0.582192	0.490446	0.420382

For our classification algorithms, we used the same training/validation/testing split, and used four bins of the grades as our classes. For the Support Vector Machine using RBF kernel, we began by using the training and validation sets to perform hyperparameter tuning. We tested various values for gamma - 0.001, 0.01, 0.1, 1, 10, and 100 - in the rbf, and found that the best value for sigma was 0.001. This yielded a maximum accuracy on our validation set of 42%. For the Support Vector Machine using a linear kernel, we did not perform any hyperparameter tuning. However, the validation accuracy is shown as another set of test data. Both Support Vector Machine algorithms perform better on the training data than on testing, indicating that while overfitting may be lessened by hyperparameter tuning, it has not completely disappeared.

We also used a Decision Tree algorithm. In theory, this algorithm should perform very well, due to the presence of a high amount of categorical data. We performed hyperparameter tuning on the maximum depth of this algorithm in order to reduce overfitting. For validation, we tested maximum tree depths of 1, 5, 10, 25, and 50, and found that the best maximum depth was 25. Again, this yielded some overfitting, but it still performed very well on the testing data.

Finally, we used Neural Networks to model these data. We used two shallow, feed-forward networks, one with a sigmoid activation function and one with a rectified linear unit (ReLU) activation function. Because our dataset was not large enough, a deeper network would lead to overfitting. Our hidden layer had 48 nodes. Again, we did not perform hyperparameter tuning, so the performance on the validation set is included as an additional test. The network that used ReLU did not perform as well as any other classification method, while the network using a sigmoid function performed better than the others. Again, these algorithms were fairly effective at predicting student performance.

Participants Contribution

Ediger, Tatiana - Pair programmed all visualizations, machine learning algorithms, and data processing; wrote about the importance of the project, features, and described the dataset; wrote about regression algorithms
Lamontagne, Michael - Pair programmed all visualizations, machine learning algorithms, and data processing; wrote about the importance of the project, the technical approach and described important features; wrote about classification algorithms