Improving Student Academic Success

Executive Summary

Andrew Niewiarowski

Executive Summary:

The analysis used several different features, such as study time, to model the grade a student would receive at the end of the term. The purpose of this analysis was to identify which students may need extra help.

It was found that students that were behind at the beginning of the term stayed behind, and that students who had failed classes before were likely to fail again. This means that it is crucial that students get help early on. It was also found that very few people had extra school support, even those who were failing. For this reason, it is recommended that the school looks at its support programs and finds out why students are not taking advantage of them.

Problem Framing & Big Picture:

A large issue at the school is academic failure, which reflects poorly on the school's reputation and often results in students leaving. For this reason, a model was developed to predict a student's grade based on a variety of factors. This will be useful to identify students that could potentially be facing challenges academically either currently or in the future. Data can be plugged into the model, which will predict a student's final grade. Then, a decision could be made on whether that student will need extra help. Ideally, a student can receive support early on, so they do not get left behind.

Data Overview:

The dataset consists of 35 features relating to student backgrounds and performance in math. The target chosen to act as a proxy for academic success was the final grade a student received in math. This was the feature titled "G3".

The correlations of the features were compared to see which features best predicted the final grade. A heatmap of the correlations was created for all numerical features. The heatmap shows that past grades received in term 1 (G1) and term 2 (G2) are by far the best predictors of the final grade (G3). Past failures are the most negatively correlated with final grade.

The correlations of the student's final grade (G3) were also analyzed with ordinal features using the Spearman rank correlation. Final grades were most strongly correlated with the parents' education level, specifically the mothers. It was also correlated with commuting time and study time, although the correlation was weaker.

There were two models was created to predict a student's final grade: one with the G1/G2 columns and one without the columns. One reason for this is that the correlation between the G1/G2 column and G3 column was so much greater than with the other attributes, that it is likely that they overshadow those other attributes. This would hide valuable insights. The more important reason why the G1/G2 columns were removed, is that our analysis shows students need help early on. Finding out that a student needs extra help after he or she has already taken two out of three terms is unhelpful.

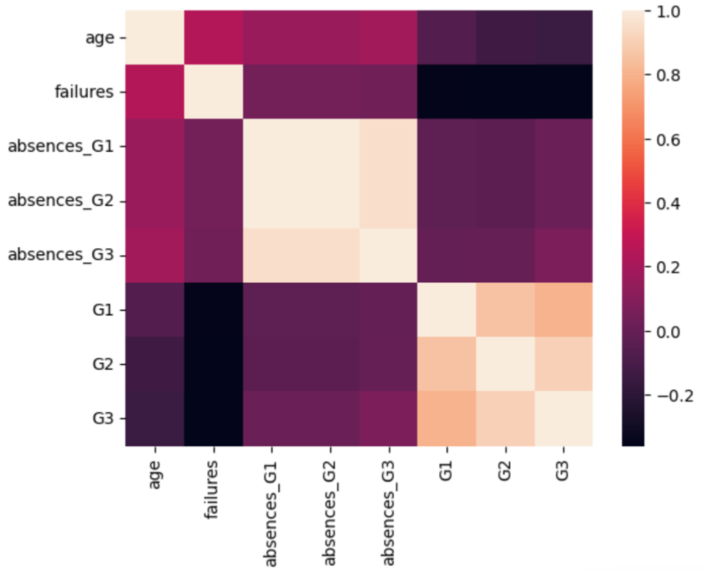


Figure 1: A heatmap showcasing the correlations between numerical features. G3 is the target and represents the final grade. G1 and G2 are the first and second term grades. The absences relate to the student's number of absences per term. Failures represent the number of times a student has failed a class. Age represents the age of the student

Analytical Insight:

The data was explored before building the model. This was done to find business insights and pick a final list of features to use in the model.

The line graphs below compare student past performance in the term to their final grade. One graph explores the relationship between the first term grade and final grade. The second graph looks at the relationship between the second term grade and final grade. It should be noted that they were almost perfectly correlated, besides the collection of dots on the bottom which represent students who dropped the class. This means that it is essential to get help for failing students early, as students who are behind in the first term tend to stay behind.

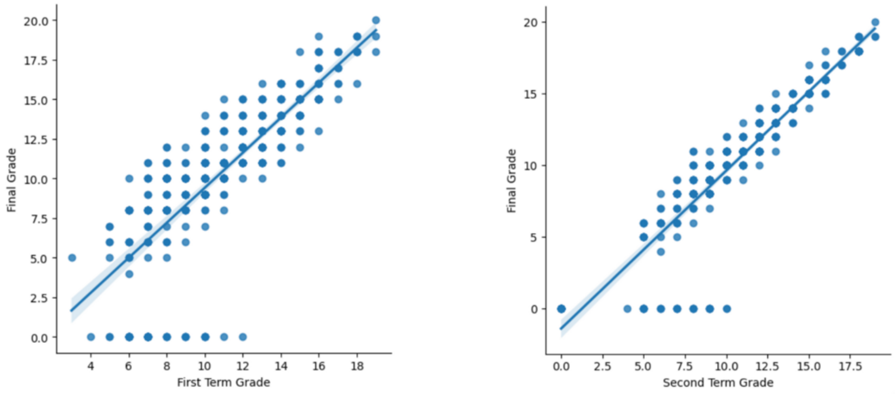


Figure 2: Line graph comparing the first and second term grades to the final grade

The bar chart below shows the correlations of ordinal features to their final grade. Ordinal data is related to a non-numerical scale. For example, the options for health are on the scale from “very bad” to “very good”. The Spearman rank correlation was used, since it can be used to compare ordinal data (Navarro 146). Looking at the chart, the parent’s education level has the greatest impact on the final grade. Party life has the most negative correlation on a student’s final grade. It is interesting that study time did not have a stronger correlation to the final grade. It would be expected that a student’s effort level would be the greatest factor in their success. However, it seems like a student’s environment is important in their academic success based on the correlations in the chart.

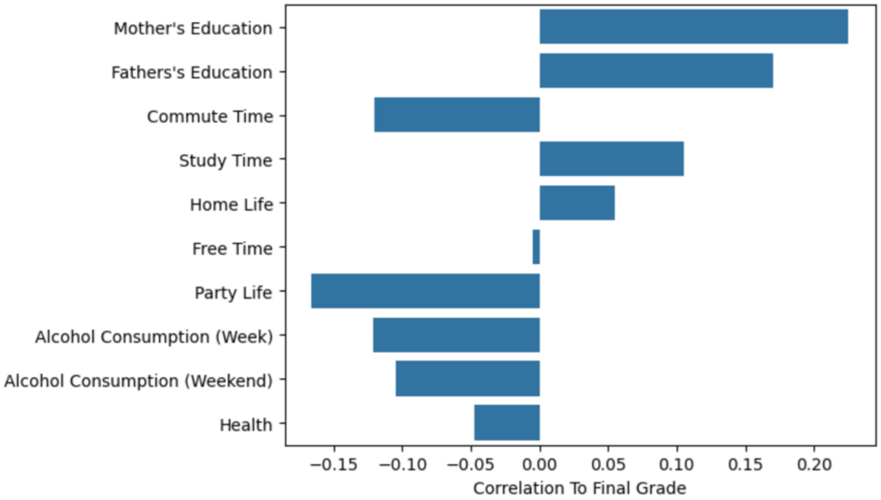


Figure 3: Bar chart showing the correlations of ordinal data to the final grade

The count plots below are count plots that display the total number of final grades for different attributes. The first plot compares the final grades of men and women. The second plot compares whether the student had access to a paid tutor. The first plot shows that there are more men who get the highest grades and more women who get lower grades. This means that the school may need to provide more resources for women. The second plot suggests that academic performance of those with paid tutors was average. It appears the paid tutors did not help as much as expected.

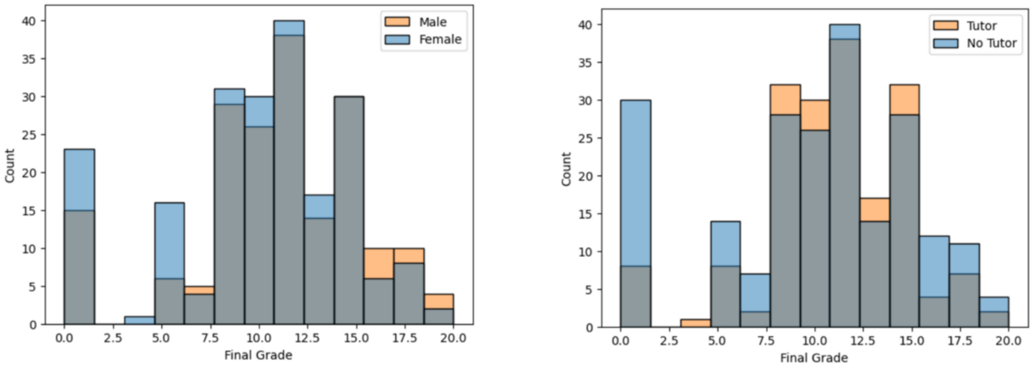


Figure 4: Count plots comparing the final grades between two categorical features: sex and private tutoring

The count plot below compares the final grades of students who had extra school support with those who did not. It should be noted that almost no students had any extra school support. This includes students who received a final grade under 10, which means the student failed the class. The school may need to see why so few students are receiving extra help. The students may not know that they can obtain extra support, or they may not see it as beneficial. Either way, this may be something to investigate.

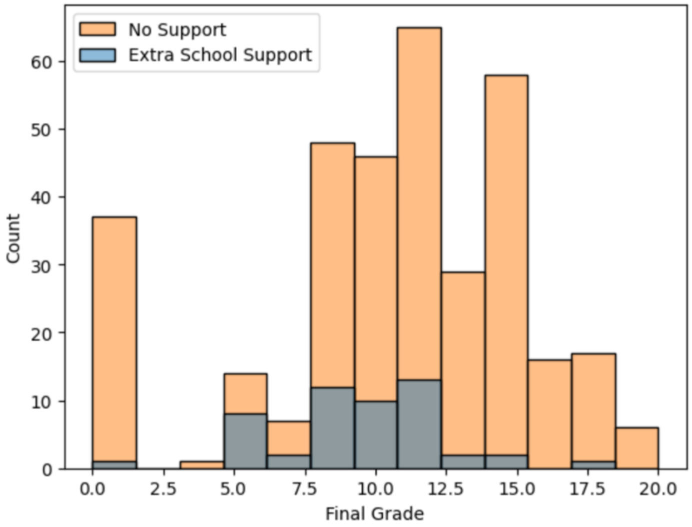


Figure 5: Count plot showing the final grades of students who received extra school support and students who did not receive extra support

The charts below were the most important insights gleamed from the data. Other variables, besides the ones mentioned above, simply did not have as much of a correlation with the student's final grade. Many other features, such as whether a student was involved in romantic relationships, had too small of an effect on the data.

Methodology:

Building the final model to predict student academic performance consisted of training three baseline models, tuning them, and picking the most accurate one as the final model. The models that were build were a linear regression, lasso regression, and support vector machine model.

The models were evaluated for accuracy using the root mean squared error metric. Each model generates predictions. These predictions are then compared to real values from the dataset in a cost function formula. If a prediction and real value are different, then the penalty generated by the cost function goes up. However, if the prediction and the real value are the same, then no penalty is applied. The model with the lowest penalty calculated by the cost function is the most accurate model. The cost function used for all the models in this project is root mean squared error (RMSE). It first subtracts the predictions from the real value. Then it adds all these numbers together for each prediction and real number. Finally, it takes the square root, so the number is always positive (Géron 43). A smaller RMSE means a more accurate model.

Equation 2-1. Root mean square error (RMSE)

$$\text{RMSE}(\mathbf{X}, h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(h\left(\mathbf{x}^{(i)}\right) - y^{(i)}\right)^2}$$

Equation 1: Root mean squared error cost function equation (Géron 43)

Linear regression models are the simplest. The model takes the equation of y = mx +b, except it adds more mx's to represent each feature. The b is represented by θ_0 and the mx are represented by θ_n. The algorithm used to solve this model is called ordinary least squares. It gets the θ_0 and θ_n values (Géron 132). While simple, linear regression is often very accurate. In this case, it was used as the final model.

Equation 4-1. Linear regression model prediction

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

Equation 2: Linear regression algorithm (Géron 132)

Lasso regressions are very similar to linear regression. The only difference is the cost function. The lasso regression adds a term to the cost function for regularization. Regularization is used to prevent overfitting and make the model more accurate. The lasso regression introduces an α hyperparameter. Setting the α larger will make the θ_n close to zero. It is set by default to 1. The lasso regression adds the absolute value of θ_n multiplied by α to the cost function. The cost function shown below is the lasso equation if mean squared error (MSE) is the cost function (Géron 158).

Equation 4-10. Lasso regression cost function

$$J(\boldsymbol{\theta}) = \text{MSE}(\boldsymbol{\theta}) + 2\alpha\Sigma_{i=1}^{n}\left|\theta_i\right|$$

Equation 3: Lasso regression algorithm (Géron 158)

The support vector machine model (svm) was the final model used. The algorithm works by segmenting the data between two lines, called support vectors. In a regression svm, it tries to maximize the number of points between the support vectors. The actual math behind the model is very complicated and will be skipped. However, the svm has two main hyperparameters that someone can choose: C and epsilon. Epsilon determines the width between the support vectors. C determines the number of points that will be inside the boundary of support vectors. The svm allows for more complex model building (Géron 184).

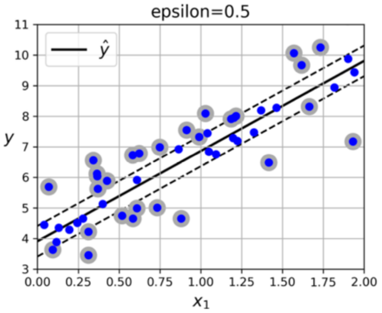
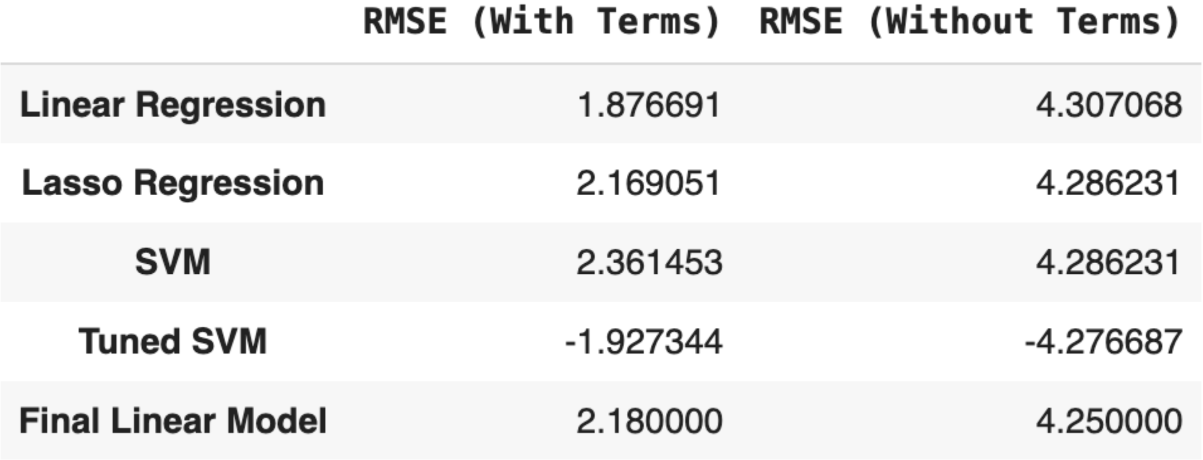


Figure 6: Graph of data points between two support vectors (Géron 158)

Key Results:

Each model was compared using RMSE. The values of RMSE are summarized below in a table. The “With terms” refers to the model trained with the G1/G2 columns and the without terms refers to the models trained without the G1/G2 columns. The G1/G2 columns were generally more accurate. The linear regression performed best on the training data, so it was used in the final model, final linear model. The model seemed to generalize well to the test data, but was not as accurate, demonstrating a little overfitting in the training set. It can be interpreted as predicting either 2.18 or 4.25 points off the correct final grade.

Predictions varying 2.18 off the final grade are acceptable, but 4.25 is too inaccurate. A final grade of 20 represents a perfect score. It was be recommended not to implement either model in production. The model with the G1/G2 terms is not useful, since students will have needed to complete two terms before it can identify if they need help or not. This is too long of a wait time, since students need help early on. The model without the terms is also not useful because it is too inaccurate. However, it has the potential to be useful. If it can accurately identify students at the beginning of the term who need help, then this would greatly improve student success.



	RMSE (With Terms)	RMSE (Without Terms)
Linear Regression	1.876691	4.307068
Lasso Regression	2.169051	4.286231
SVM	2.361453	4.286231
Tuned SVM	-1.927344	-4.276687
Final Linear Model	2.180000	4.250000

Table 1: Table comparing RMSE of different models. The final linear model is the final model trained on the test set.

Conclusion:

While the final model is not recommended to use in production, this analysis did unveil some key insights. First, is that students do not take advantage of extra school support, even if they are failing. The school may need to implement more policies to raise awareness, so that students can get the help they need. Second, the analysis exposed that females may not be given the support they need. The school should look about how it can better cater to females needs. Third, the model has potential. It just needs to be more accurate. There are many more algorithms to model data and many more combinations of features to try. If given more time, it is likely that a model that is accurate enough can be developed. These are some suggestions to improve academic success and reduce dropout rate here at the school.

Works Cited:

Géron, Aurélien. *Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow Concepts, Tools, and Techniques to Build Intelligent Systems Aurélien Géron Aut*. O'Reilly, 2020.

Navarro, Danielle. *Learning Statistics with R A Tutorial for Psychology Students and Other Beginners Danielle Navarro*. Danielle Navarro, 2018.