DTSC 670 Executive Summary: Tyler Stark, 9/24/24

Project Summary

The goal of this project is to predict student performance based upon the data provided from the large Portuguese school system for which I work. I have analyzed the data and created a Machine Learning model to help predict future data in hopes of identifying students who might need additional assistance and interventions to improve their grades.

In summary, I found that the most important predictors of student performance are those which we already have guessed: a student's record of absence, paired with their grades from previous trimesters. However, there are a few other features that appear to be correlated with a student's final grade, listed as follows:

1. Parent's level of education
2. Interest in pursuing higher education
3. Relationship status
4. Age
5. Failures in previous courses

Although there is correlation between these features and a student's final grade, I did not find that they are significant enough indicators of a student's success - or lack thereof - based upon the data provided. The most reliable predictors continue to be absences and previous term grades, and I propose building a model based off of these predictors until further study may be accomplished.

My recommendation: tutors and study groups should be made available throughout the G2 and G3 terms based upon a student's performance in the following metrics: absences and previous term grades. Within these metrics, more resources should be allocated to the following subgroups:

1. Older students in their final two years of school
2. Students who have previously failed 1 or more classes
3. Students who would be first-generation college students if they pursue higher education

Problem Framing & Big Picture

Without properly utilizing this data, we are shooting in the dark when it comes to helping our students succeed. Final G3 grades will continue to trend low unless we implement purposeful interventions to help these specific subsets of students. Not only could implementing this model help secure a brighter future for our students, but it could also aid in securing future funding and grants for our institution. We have the data, and we can begin to show the improvement immediately upon implementation.

For example, an older student who did not struggle with their grades at 15 and 16 may be surprised to find that they are struggling in their final years of school. They also may be afraid to ask for help because they have never needed it before, and the embarrassment that may come with that. By creating built-in guard rails for older students, they will not have to traverse asking for help because it will already be provided. The same goes for future first-generation college students. They may want help, but just have no idea how to get it. By providing these tools up front to the students that need help based on our data, we save precious time. Waiting to see how a student performs in their G1/G2 terms could be the fine line between success and failure for some of these students.

Data Overview

This dataset is composed of each student's past performance in terms G1, G2, and G3, in tandem with a record of absences and failures, along with categorical/ordinal data describing various aspects of the student's life (i.e how they spend their free time, their overall health, etc.).

The target data of our model is the student's final G3 grade. The G3 data is key in forming our predictive model: we can see how each student performed and possibly connect this performance with other features from the data.

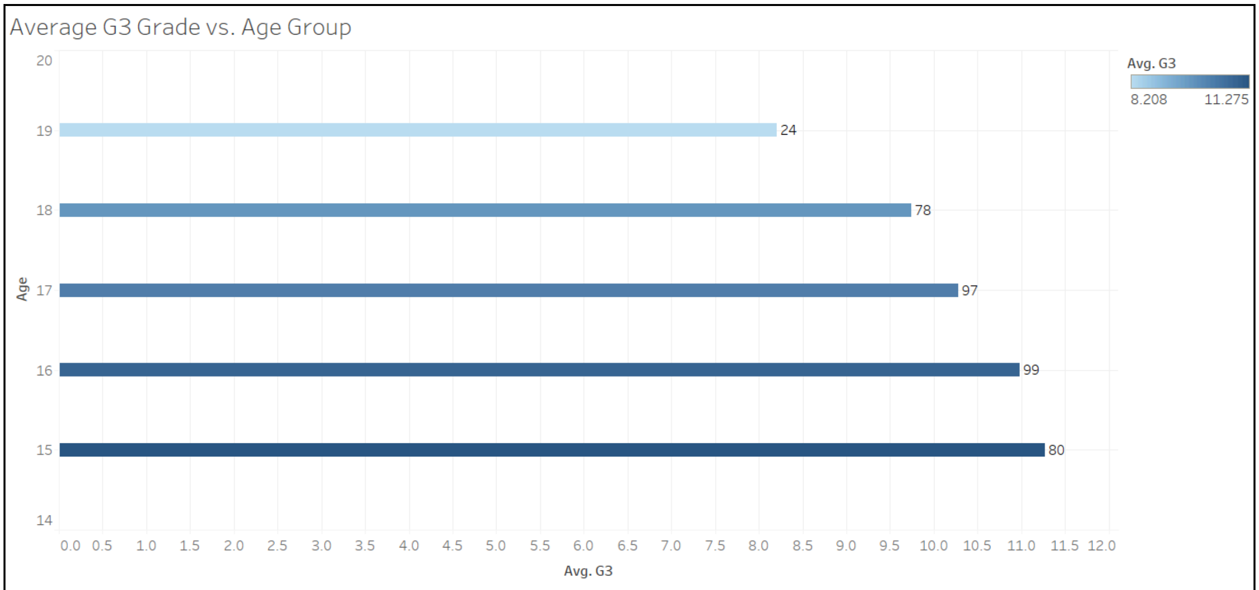
As listed above, a correlation was found between the following features and a student's final G3 grade: parent's level of education (positive correlation), interest in pursuing higher education (positive correlation), relationship status (higher average grade if not in a relationship), age (higher average grade for younger students), and failures in previous courses (negative correlation).

Two models were created for this project: one that includes a student's G1/G2 term grades and one that does not. The G1/G2 inclusive model performs better in terms of lower Root-Mean-Squared-Error and higher R-Squared scores, but there are significant drawbacks to relying on this model. The most pressing is that intervention may come too late if we wait to see how a student performs in their G1 and G2 terms. The non-inclusive G1/G2 model performs far worse in terms of these metrics, but may offer hope in providing significant early intervention for struggling students.

Analytical Insight

Although the non-inclusive G1/G2 model's performance is lacking, I have included visuals below to show why I think this model should encourage further study. With more data and a better tuned model, we could possibly predict which students need intervention far earlier than before.

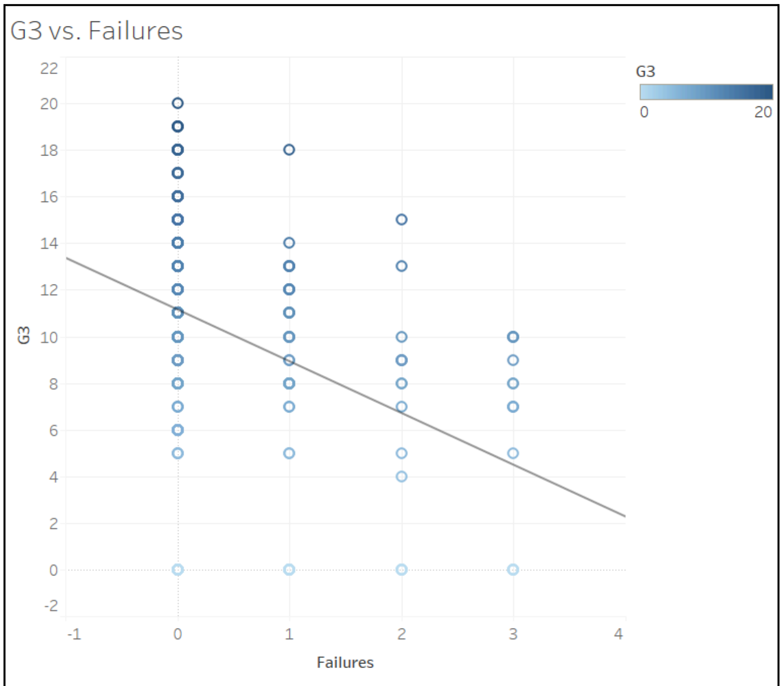
Figure 1



The plot of average of G3 for Age. Color shows average of G3. The marks are labeled by count of Age. The data is filtered on Age, which excludes 20, 21 and 22. I have excluded these ages due to the small size of each group.

Figure 1 displays a horizontal bar chart, comparing the average G3 grade for each age group. A significant decrease in average grade is noticed as students age, almost a 28% decrease between ages 15 and 19. Resources may be needed for older students, regardless of their grade status, strictly because the coursework becomes more difficult as they age. This truism would only complicate matters for an already-struggling student.

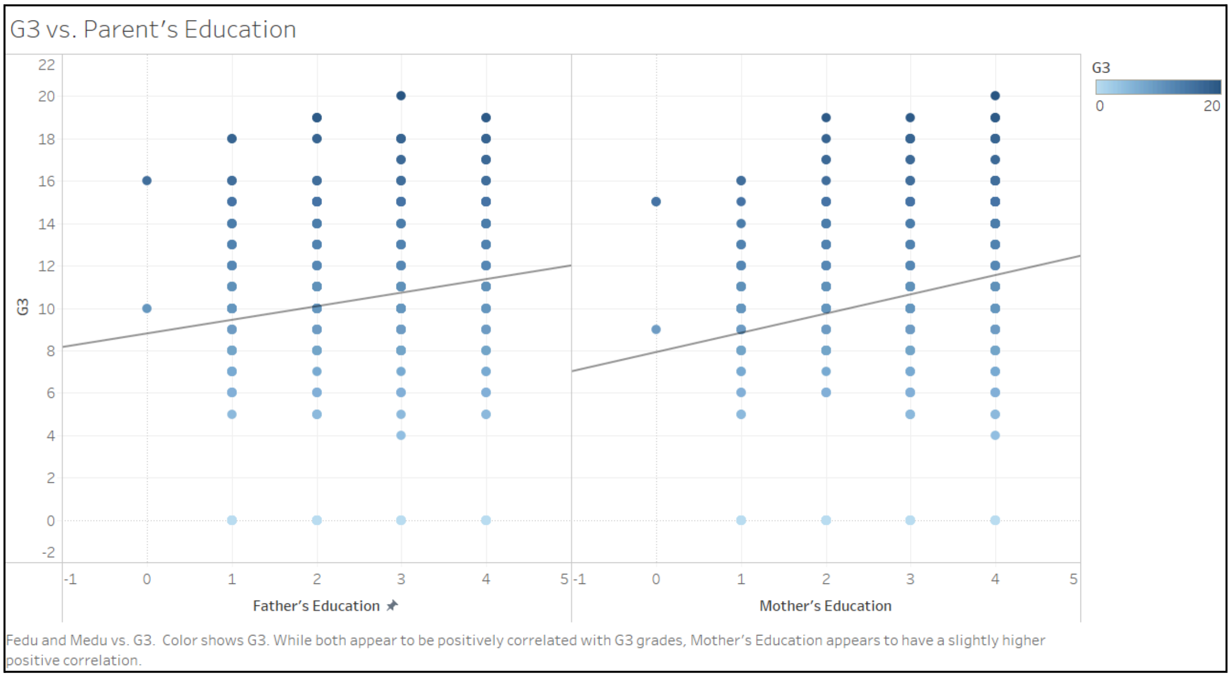
Figure 2



Failures vs. G3. Color shows G3. Whether or not a student has failed a class in the past seems to have a correlation with the student's G3 grade.

Figure 2 displays the correlation between a student’s final G3 grade and their past class failures. The more failures a student has, the lower their final G3 grade. This could generally be assumed without the data, but the trendline makes it very clear. It is significant to notice that students with 3 failures did not achieve a grade above 10. It would be a simple intervention to provide tutors or study groups to students who have failed at least one class.

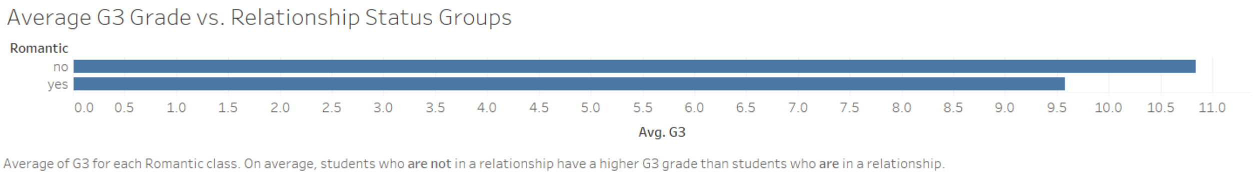
Figure 3


G3 vs. Parent’s Education

Fedu and Medu vs. G3. Color shows G3. While both appear to be positively correlated with G3 grades, Mother’s Education appears to have a slightly higher positive correlation.

Figure 3 indicates a positive correlation between a student’s final G3 grade and the highest level of education obtained by their parent(s). One thing to note, the trend line for “Mother’s Education” is slightly steeper than “Father’s Education”. It is common for post-secondary institutions to provide resources to “first-generation” students (students whose parents lack a college degree). I could be advisable to learn from institutions who do this well.

Figure 4


Average G3 Grade vs. Relationship Status Groups

Average of G3 for each Romantic class. On average, students who **are not** in a relationship have a higher G3 grade than students who **are** in a relationship.

Figure 4 is a simple bar chart showing the difference in average grade between students who are in a relationship (“yes”) and students who are not (“no”). I was actually shocked by this finding, because I assumed there would be little to no difference. The challenge will be figuring out how to help the students in the “no” category without displaying some sort of relationship bias.

I have chosen to leave out my visuals for study time and travel time compared to G3. There is a positive correlation for study time, and a negative correlation for travel time. However, the trend lines were not significant enough to merit inclusion in the model.

I have also excluded my visualizations for the relationship between absences, G1 and G2 with G3. I have done this because the correlation was already assumed, and the data confirmed this assumption.

Methodology

I utilized three separate machine learning models in this project: Linear Regression, Support-Vector-Machine Regression (SVM) and Lasso Regression. The final two models are based on SVM Regression.

Linear Regression attempts to predict the future data based on the linear trend found in the given data. The trend line used is the model that has the least amount of error between the line and the given data. This is the statistical equivalent of "an object in motion stays in motion".

SVM Regression is similar to Linear Regression, but instead of using a thin trend line to predict future data, we could visualize more of a "trend avenue".[1] SVM Regression attempts to fit as much data on the avenue as it can, while limiting the amount of data that does not fall on the avenue.

Lasso Regression is also similar to Linear Regression, but it adds more importance to the weighter features, while penalizing the less interesting, outlying features.[2] This model could be helpful when attempting to find trends in a noisy data set.

Two metrics were used to determine the performance of these models:
(1) Root-Mean-Squared-Error (RMSE) and (2) R-Squared.

For RMSE, the lower the better. This number describes how much distance lies between actual and predicted values in the data. A lower number means that these values are more similar. R-Squared is a number ranging from 0 to 1, and it tells us how well the predictor features determine the predicted value. For example, an R-Squared of .9 means that the predictors account for 90% of the variance in the predicted value. Usually, a good model has an R-Squared of .8 or higher.

Key Results

With these definitions in mind, I have included the metrics of the two models below:

RMSE for the model **with** G1/G2 grades: 2.11
R-Squared for the model **with** G1/G2 grades: 0.78

RMSE for the model **without** G1/G2 grades: 4.32
R-Squared for the model **without** G1/G2 grades: 0.09

The model **with** G1/G2 grade performed significantly better, as could be guessed. The G1/G2 grades account for a significant amount of the variability in a student's G3 grade, while our other features (i.e age, relationship, etc.) only account for roughly 10% of the variability.

At this point in the research, I would suggest the following: (1) begin implementing the model **with** G1/G2 grades immediately, and (2) collect more data and fine tune the model **without** the G1/G2 grades for future implementation.

Neither model is optimal, but I think it is a start. It is safe to say that reactive tutoring and study groups can begin for students who have struggled throughout the G1 and G2 term, for the correlation is obvious. Furthermore, proactive resources can be developed for the students who fall into the other categories I have explored to be implemented at a future date. The correlation is not as significant, but it is still present and should not be ignored.

Conclusion

Before the project, I believed the biggest struggle would be creating the pipelines and custom transformers for the dataset, while the easiest part would be interpreting the results of the models. However, I found the opposite to be true. I was surprised that neither model performed as well as I had hoped. It is eye opening to see that you can put a lot of work into a model, but that doesn't guarantee its success. Also, personally, I don't think I would implement either of the models just yet. I think a model that doesn't include the G1/G2 grades might do better as a decision tree or random forest model because of the categorical nature of the data. I think the correlation is there, but the regression models we chose were not effective. For next steps, I would explore those models and go from there.

---

1. Géron, Aurélien. *Hands-on Machine Learning with Scikit-Learn, Keras and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc, 2023, pp. 184-186.
2. Ibid. pp. 158-159