Jeffery Tse
CS-UY 4563
Professor Sellie
December 8, 2021

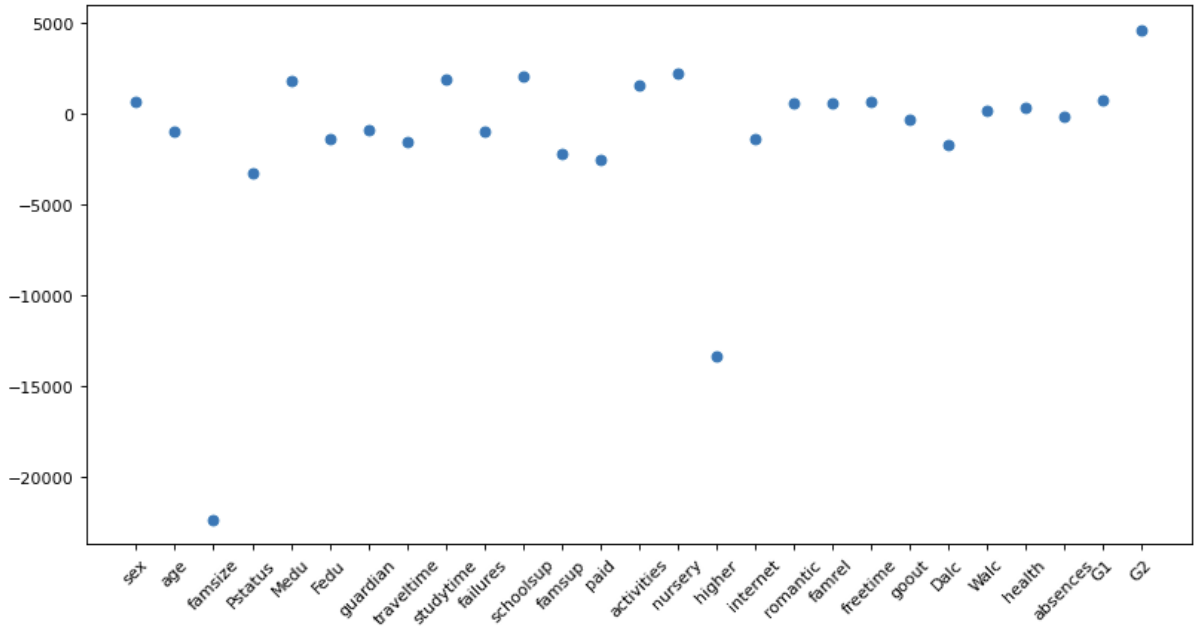
Student Grades Prediction

For this final project, I wanted to find a dataset that was interesting to me. Since this was towards the end of the semester, grades were on top of my mind. Originally, I had a partner that picked a dataset related to stock predictions. However, the partner has since withdrawn from the class so I was left to do the project by myself. I had a lot of assignments backed up and decided to look for a different dataset because I was left with an incomplete mess of code. I decided to look through kaggle.com for some datasets and ran across school related datasets. One of them was the prediction of students from a particular highschool getting into the UC schools. However, that set was slightly complex and I did not see the prediction as that interesting. Instead, I found this separate dataset that was on the grade prediction of some students in Portugal. It has many interesting features that explain the background of the specific student including parental status, study time, extracurricular activities, romantic relationships, and even alcohol consumption. There were a total of 30 features and 395 students. Using these features, I would predict if the students were going to pass or fail.

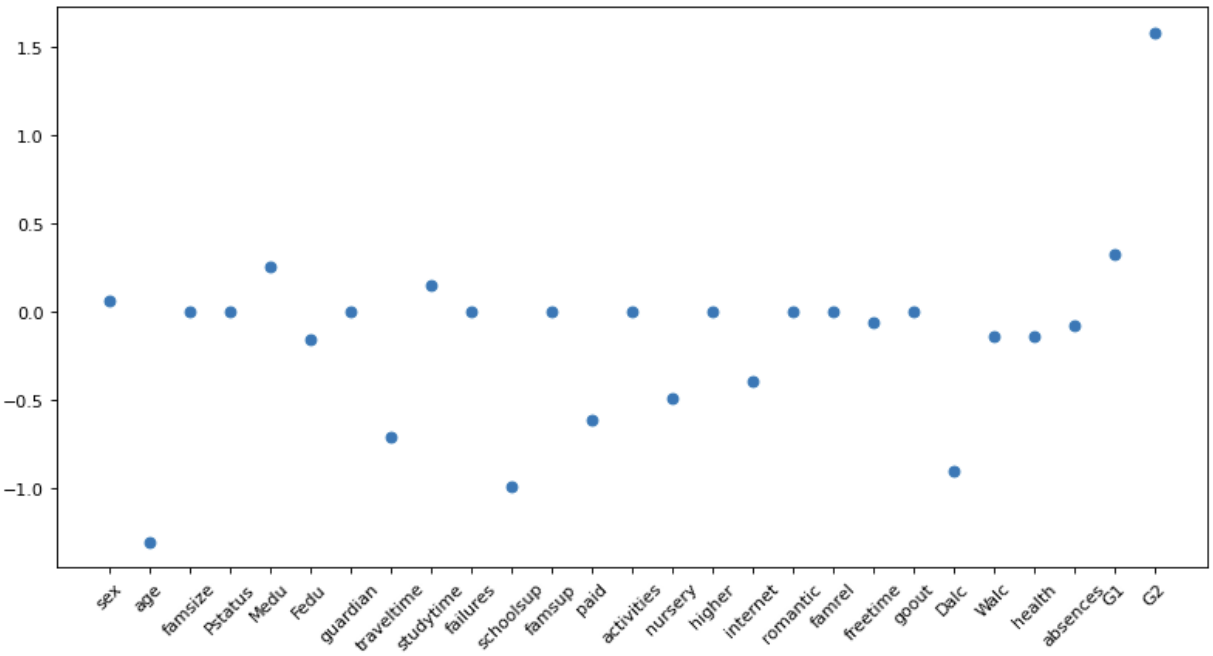
For the preprocessing of the data, I just converted the non-numerical data to numeric data points. Features such as if the student went to a nursery were in yes or no form. I would convert that to just ones and zeros. Other features such as guardian, there were three options which included mother, father, and other. For this, I just made “other” as number two. I also removed a few features from the provided data because they did not seem to have much of an impact, at least from my initial thought process. I removed school because I thought that out of the two schools, if there were a difference, it would not be a big difference. Even if there were, it would be a universal application to the whole student population. I also ignored the addresses of the students as that should have little to no impact on their performance in school. The mother and father’s job did seem important, but they were categorized in a way that did not really make sense to me. And finally, there was the reason for choosing the specific school. I thought that this feature did not matter much because there were only two schools and the reasons seemed more to do with parental reasoning. The final grade that I planned on predicting was out of 20 as are the other two grades for the first period and second period of the school year. To make it a bit easier to predict, I decided to make the prediction categorized. Thirteen out of twenty would be 65% and I made any grade above that passing. This meant that I would only need the models to predict if the students’ final grades were over 13.

The supervised analysis required three different models. I chose to use logistic regression, support vector machines, and neural networks. Each one of these models had their own transformations. I was able to use the sklearn library to assist in all the configurations of the different models and feature transformations.

The first model I implemented was logistic regression. I started off with the L1 regularization.

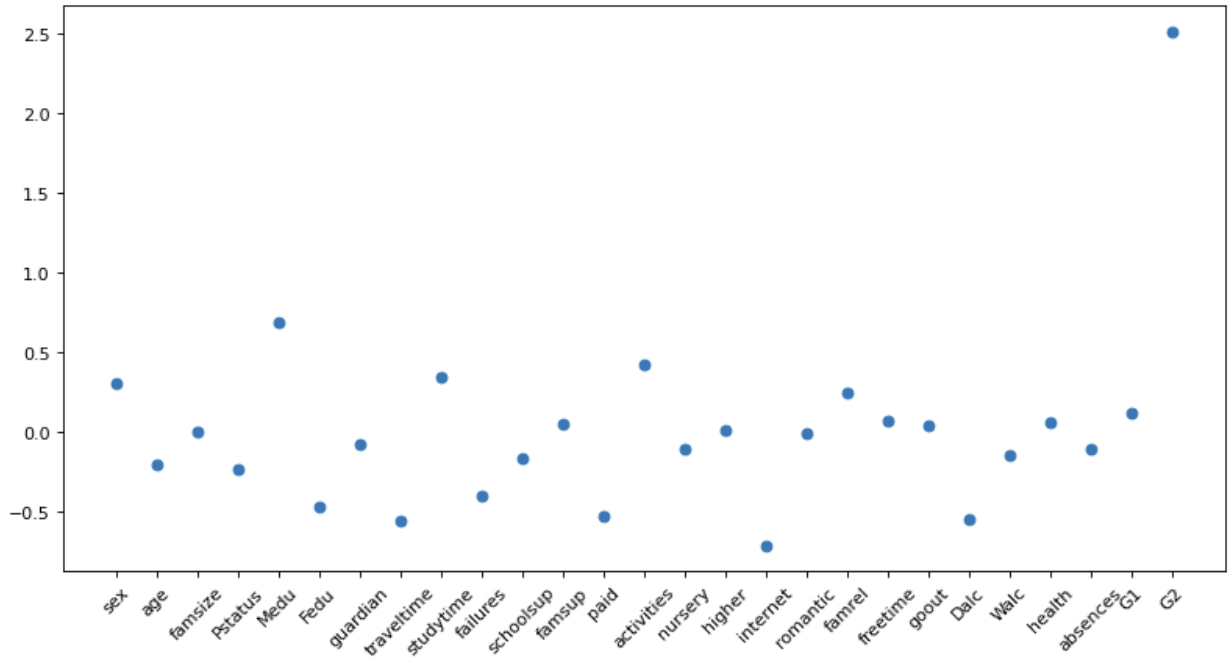


This was the graph that I got for the weights for this model. It seemed a bit weird as the weights were way too large for some reason. For one of the runs, the training accuracy turned out to be 100% while the testing accuracy was 92%. This seemed slightly under-fitted as the testing accuracy was slightly lower than the training accuracy. I then moved on to the L1 regularization.



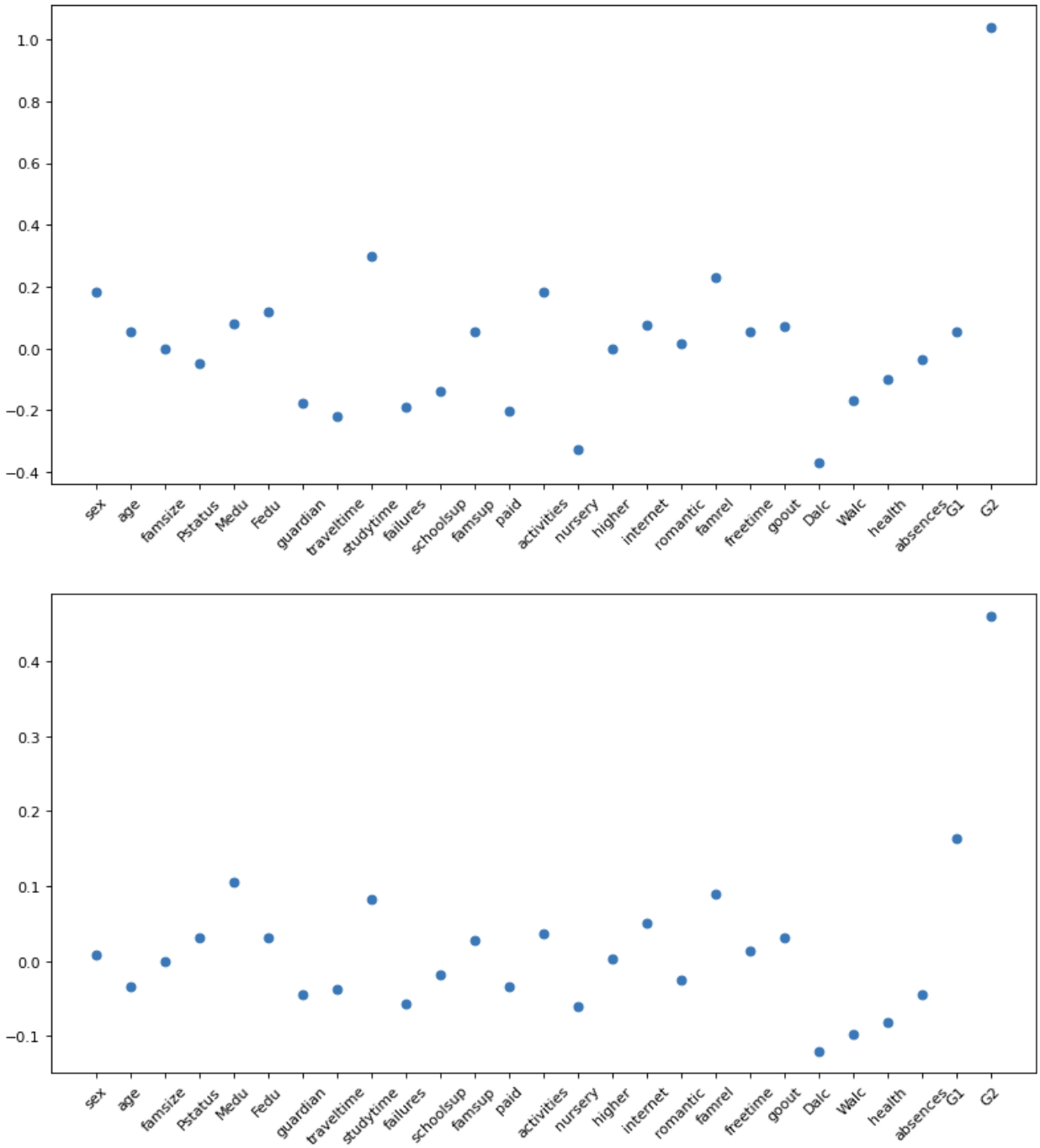
The weights here can be seen as a lot more normal. It looks like the G2 was the most impactful for the final grade. This run resulted in a training accuracy of 96% and a testing accuracy of

94%. This was slightly more reasonable than the previous attempt and had a slightly higher testing accuracy. The final configuration I did for the logistic regression is the L2 regularization.



This model seemed to weigh the G2 results even further. Since all three of them had a higher weight than everything else, it would seem that G2 is an important factor in determining the final grade of the student. This would make sense because the G2 feature basically provides a number representing how the student is doing most closely to the final grade.

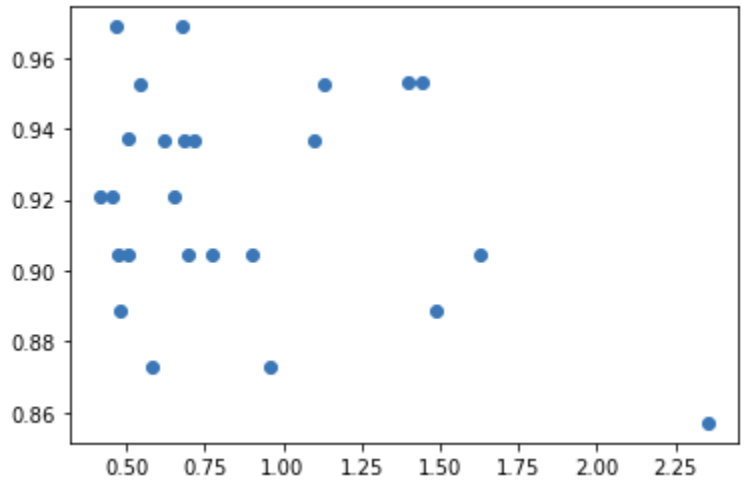
The second method I used was the support vector machines. I used two different kernels for the SVMs and also tested different c values. One of the more accurate results included the polynomial kernel with a c value of 100. This resulted in a training accuracy of 99% and a testing accuracy of 95%. I was not able to get the weights of the features because the built-in method for the coefficients of the features was not available for the polynomial kernel. I did perform some tests with the linear kernel. Two of the better configurations included a c value of 0.1 and a c value of 0.01. The first one had a training accuracy of 98% and a testing accuracy of 97%. The second test had a training accuracy of 97% and a testing accuracy of 97%. The weights for the features for these two are as follows.



It can be seen again that the G2 seemed to have the most impact on the final grade prediction. On the second one in particular, G1 also had slightly more impact than every other feature, which also makes sense as it is the grade of the first period. On the first test, there were also other interesting and higher weighted features that include the Dalc which is the student's alcohol consumption. This would obviously have a negative weight as more alcohol would be worse for the student. Study time is also one that makes sense as more studying hopefully results in a better grade. One feature that was surprising is the nursery which previously did not have much impact.

For this attempt however, it showed that going to the nursery had a significant negative effect on the grade of the student. I do not think that this is entirely accurate so maybe more data would help with figuring out this case. For the SVMs, the most accurate run was with the polynomial kernel with a c value of 1. I am not sure if this is a fluke because the training accuracy and the testing accuracy came out to be exactly the same percentage of 97.46835%.

The third model that I used is the neural networks. I was not really sure on how to display the weights of the runs compared to the features because the built in weights provided the weights in the amount of layers that were provided. When running the neural networks, I decided to test for the effect of the learning rate as well as the architecture. I first tested a few different learning rates with a constant architecture of three layers of eight nodes each. With the four learning rates of 1, 0.1, 0.01, and 0.01, it seemed that 0.1 performed the best with a training accuracy of 97% and a testing accuracy of 97%. Interestingly, when going smaller by a magnitude of 10, it got really inaccurate past the 0.01 learning rate. At 0.001, the training accuracy fell to 73% and the testing accuracy to 82%. This is easily seen as overfitted as the training accuracy is much lower than the testing accuracy. I stuck with the 0.1 learning rate and tested a few different layer configurations. The three cases that stood out the most were the three layers of eight nodes, the two layers with eight nodes, and the one layer with 8 nodes. The latter two both gave 100% training accuracy and 94% testing accuracy. It turns out that the three layers were still the best architecture from the ones that I have tested. I also had an interesting graph with the fitting times plotted against the accuracies.



It shows that there was not a big difference in how long the neural network took. From watching the presentations of other students, I have understood that this should have been the most difficult to do and the longest to run. However, for me, this was almost as quick as the other models. I take that due to my limited dataset of just 395 rows and a reduced 25 features, it did not make a difference in the length of time it took to run the neural networks.

In conclusion, it was interesting to see that the features about the students were able to provide enough information to accurately predict the final grade of the students. All three models

were resulting in pretty accurate results. This can be seen highlighted in the chart below with all the runs that I have done.

model	configuration	training accuracy	testing accuracy
Logistic Regression	No regularization	100%	92%
	L1 regularization	96%	94%
	L2 regularization	98%	96%
SVM	poly, c = 1	97%	97%
	poly, c = 10	98%	96%
	poly, c = 100	99%	95%
	linear, c = 0.1	98%	97%
	linear, c = 0.01	97%	97%
	linear, c = 0.001	94%	95%
Neural Networks	(8,8,8), learning rate = 0.001	73%	82%
	(8,8,8), learning rate = 0.01	96%	95%
	(8,8,8), learning rate = 0.1	97%	97%
	(8,8,8), learning rate = 1	100%	92%
	(8), learning rate = 0.1	100%	94%
	(2), learning rate = 0.1	98%	89%
	(8,8), learning rate = 0.1	100%	94%
	(64,64), learning rate = 0.1	100%	92%
	(64,64,8), learning rate = 0.1	100%	91%

The method that was the best for this dataset seemed to be the SVM with the polynomial kernel and the c value of 1. Although it seems to have the same results as the best contender in neural networks, the accuracy in the SVM proved to be better fitted as the results were so similar. If I had more time to work on this dataset, I would be curious on how the models would perform in predicting the success of a student without any concrete features on the students' previous grades. Possibly removing the G1 and G2 would be interesting because the models would be

purely judging the student on their background with family and other factors unrelated to academics. Also, a larger dataset would definitely help with the accuracy and better fits as this is not a lot of students. As for now, the project showed that a lot of factors go into the success of a student including a lot of conditions that are totally out of their control. Environmental factors such as the parents' education, marriage status, as well as if they have gone to a nursery school. Personally, this might allow myself to blame my own lack of resolve to factors I cannot control, but it is also apparent that the features such as study time are effective. Overall, this was an interesting project.

Dataset taken from this link:
https://www.kaggle.com/dipam7/student-grade-prediction
P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.