Behavioral Data Analytics

Understanding college student's health and GPA

Catherine Groh
Industrial Engineering
UW - Madison

1 Background and Purpose

The purpose of this model is to study the effect of diet and exercise on college student's GPAs. Machine learning would be helpful here because there is a lot of different data that has already been collected and using different supervised learning techniques, it will be clear to see what the data can predict.The dataset used can be found on Kaggle called "Food Choices". It is comprised of survey data from several college students with both numeric and text entries.

The ultimate purpose of applying the algorithm is to help determine if better diets and exercise lead to higher GPAs. Getting a better understanding of how health correlates to intelligence. This could go wrong if the data shows that people with worse diets actually tend to do better in school. This could lead people to believe that their health is not as important as maybe we think it would be. This could help redesign how students study and cook. Maybe leading to healthier snack and meal options at school along with more time focused on working out and leading a healthier lifestyle.

There are two models created and posted on this site because they both show interesting predictions. One is how students' answers to "I live a healthy life" correlate to their GPA. The other is showing the answers to "I live a rewarding life" correlating to GPA as well.

2 Statement of Data & Transformations & Abstractions

I used just one data source for my visualization, the food choices dataset from Kaggle. This data consisted of 115 responses from college students on 65 different questions about their GPA, health choices, education, and weight along with other health related questions.

This data had a bit of missing responses, as well as some numeric data that needed to be cleaned up in order to create an accurate data model. I was able to look at the data using exploratory data analysis with the dataset in R. I went through to see what types of responses

there were for each of the questions and how I could use those responses to create a useful algorithm.

After the exploratory data analysis, I started by using the mutate command to make sure that the GPA and weight values were numeric. This would be helpful to create clear categories later for the visualization. I did this by using the as.numeric function. I then made sure to filter out any of the NAs that were included in either of those columns because I felt that they were the most important values to look at and if those responses were missing it would not work for the data that I wanted to analyze.

Had I had the data for where these students went to school, or how old they were, or their coursework/degree, I would have used that to combine the data. I think it could have been useful to compare with some of those metrics as well.

I had the GPA as the independent variable and the answers as the dependent. The data was split into “good” and “bad” based on the answer. Since it was 1 - strongly disagree to 10-strongly agree, I said that >5 was good and <= 5 was bad.

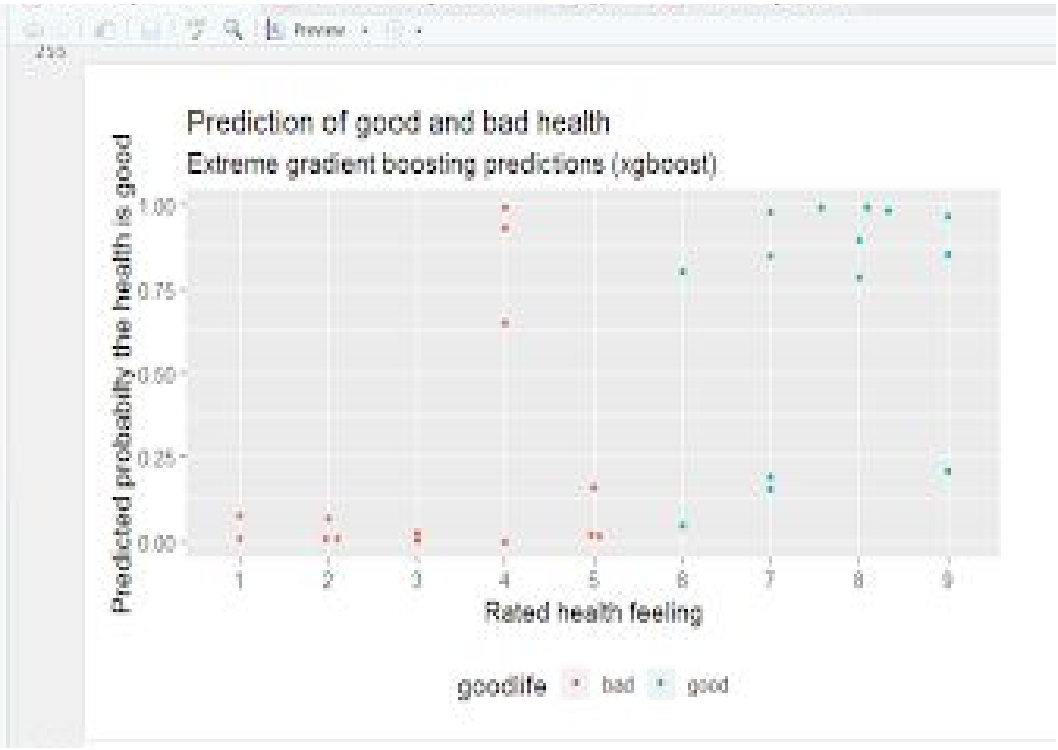
3 **Validity & Generalizability**

As you can see in the “I live a healthy life” data model, the validity is not very high. With more data, this could be avoided potentially. However, it may be that these questions aren’t generalizable and do not have any meaningful predictions to make. The “I live a rewarding life” actually has a pretty high validity or accuracy and you can see both of these values in the table below.

		“I live a healthy life”	“I live a rewarding life”
glm	Accuracy	0.6071	0.8929
	Sensitivity	0.5714	0.8000
	Specificity	0.6429	1.000
svm	Accuracy	0.7143	0.8571
	Sensitivity	0.7857	1.000
	Specificity	0.6429	0.6923
xgb	Accuracy	0.7500	0.9286
	Sensitivity	0.7857	0.8667
	Specificity	0.7143	1.000

4 Graphical Depiction

“I live a healthy life”



“I live a rewarding life”