Can You Predict Health Perceptions Based on GPA?

A deeper understanding of factors that determine how a student perceives their health

Catherine Groh
Industrial Engineering
UW - Madison

Abstract

Our society has become increasingly focused on health. Just look at the expansive industries dedicated to answering: what people should eat, how they should exercise, how they should manage stress, and the the countless other minutiae surrounding "healthy lifestyles." Less has sprouted up around and less work seems to have been done on quantifying the benefits of being healthy. It is a goal that is nearly universally sought, but few rarely stop to measure the tangible benefits from achieving this state. Bearing this in mind, I thought it worthwhile to pursue the question of how health translates to academic achievement and vice versa. . Many people would naturally assume that a more positive outlook on life would lead to a higher GPA. But the actual relationship may be more complex. This work attempts to extract correlations between health and academic performance (and the influence of a few other factors). It aims to be beneficial in helping people understand how to focus their energy when trying to improve either their GPA or overall health.

This paper discusses two models that predict health perception based on GPA. Health perceptions are based on survey responses to the statements "I feel healthy" and "I live a rewarding life" which were asked to a college population and the results later distributed on Kaggle in a data set called "Food Choices".

Preliminary findings were that it is possible to predict with some measure of confidence a person's health perception based on their GPA. Different R packages were used to validate and create these models. The Caret package was widely used in the analysis that follows. The work here leaves the door open for further refinement of the models and a better understanding of exactly which factors are most influential in predicting outcomes.

1 Background and Purpose

Young adults are largely assumed to be healthy and for that reason healthcare takes a backseat to competing priorities amongst the college population. This analysis contends that health should be a bigger topic amongst students, parents, and faculty. The data set presented and its conclusions are a matter of interest to most of the people in class, considering they fall within the target group represented by this data. The models described in this paper were chosen to effectively evaluate the correlation between a student's GPA and how healthy they are (or perceive themselves being). To draw firm conclusions, these would need to show a clear cut correlation between high GPAs and

increased life satisfaction and overall health. These models were built over a dataset from Kaggle called “Food Choices”[1]. That data set consisted of responses to questions focused on capturing discrete information around students’ lifestyles, how rewarding they feel their lives are, and their GPA.

The obvious benefit a college student might see from these models is the realization that additional attention to how they are feeling might produce a bump in their their GPA, or vice versa. I was unable to find substantial academic studies asking the same question that this analysis was attempting to answer. A brief literature survey produced very little to go off. It appears there have been a few tangential studies but their focus was largely limited to specific diseases and their impact on GPA as seen in the article *GPA-MDS: a visualization approach to investigate genetic architecture among phenotypes using GWAS results* [2]. The goal for these models is to help students understand there is at least a soft link between their health perception and their GPA and to possibly incentivise students to live healthier lifestyles.

This analysis could go wrong if the data shows that people with worse diets (poorer health) actually outperform their counterparts. This might lead people to believe that health is a secondary factor and can be ignored.

A stretch goal of this analysis is for the information presented within to help redesign how students study and cook. With increased emphasis on the positive outcomes potentially leading to healthier snack and meal options at school along with more time focused on exercise and other healthy behaviors.

There are two models constructed from this data both of which were posted on the site because each shows an interesting prediction. One pertains to how students’ answers to “I feel healthy” correlates with their GPA. The other was built off of the answers to “I live a rewarding life” and its relationship to GPA. Within the full data set, there were other factors or variables that that had the potential to be beneficial in the analysis., However, I felt that many were too narrowly focused and mostly added noise when trying to draw a proper conclusion. So the decision was made to proceed with these core questions i as they best encompassed overall health rather than one specific part of health.

2 **Statement of Data & Transformations & Abstractions**

The data used for these models consisted of 125 responses from college students on 65 unique questions about their GPA, health choices, education, and weight along with other health related prompts.

Data cleaning was necessary to remove some of the missing responses, as well as some abstraction around missing numeric data. Exploratory data analysis [3] was used to comb the dataset in R and get

a better sense for where the value laid. The response type for each of the questions and how useful those responses would lend themselves in an algorithm to describe the set was carefully considered. Because there was such an abundance of questions and many related over nuanced facets of health, the higher level questions of “how healthy do you feel?” and “do you live a rewarding life?” were chosen as a proxy for overall student health. The analysis greatly benefited from a simple, high level understanding of health rather than getting lost amongst a sea of variables too narrow in scope.

At the completion of the exploratory data analysis, the mutate command was used to ensure that the GPA and weight values were numeric. This helped create clearly defined categories for the visualization. The “as.numeric” function was useful for this purpose. After this transformation, any rows with missing values were filtered out from both columns because they are by far the most substantive drivers of the models and responses missing these values could not contribute to meaningful analysis.

The data set wasn’t sparse, but a few nice to have elements were missing. If geographic data for where these students went to school, numeric data for how old they were, or data on their coursework/degree existed, it would have nicely bolstered the data that was available. These would have provided for further contextualization and complemented some of the already calculated metrics.

For the models, GPA was chosen as the independent variable and the answers to the survey prompts as the dependent variables. The data was split into “good” and “bad” ranges based on the answer to each prompt: “I feel healthy” and “I live a rewarding life.” Since it was on a scale from 1 (strongly disagree) to 10 (strongly agree), greater than 5 was good and less than or equal to 5 was bad. As a next step, I plan on fitting the model to predict at n a more granular scale because 1-5 is fairly large. Figures 1 and 2 below illustrate the split between the “good” and “bad” responses for both health questions modeled in this analysis.

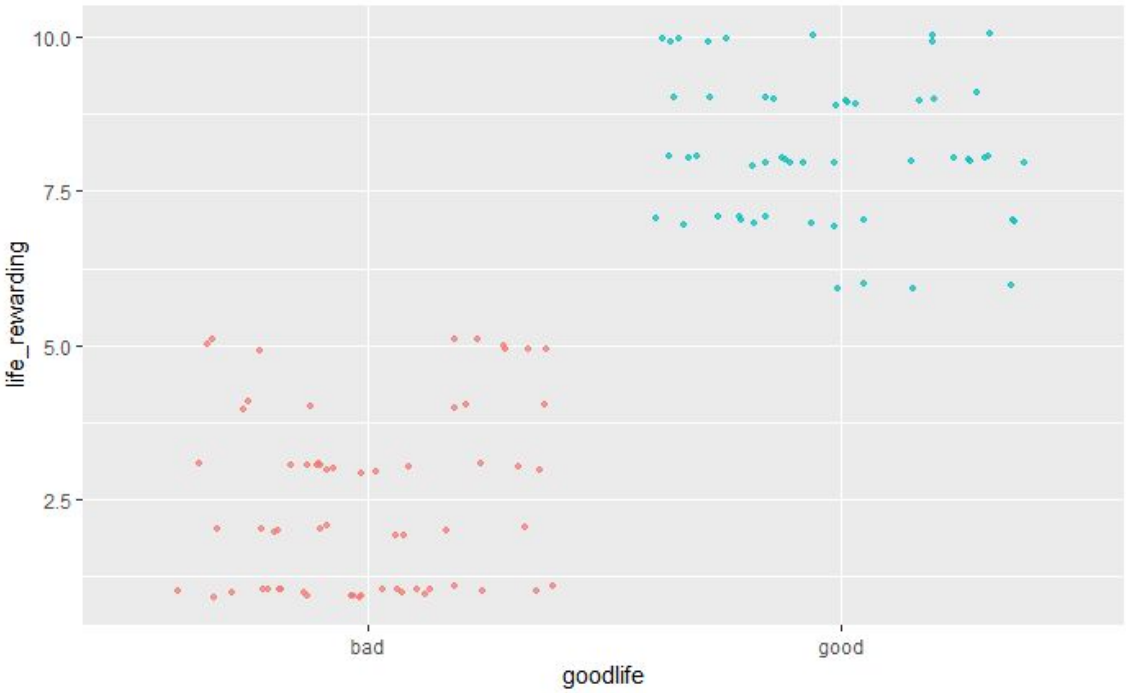

Figure 1. Split between “good” and “bad” for the statement “I live a rewarding life”

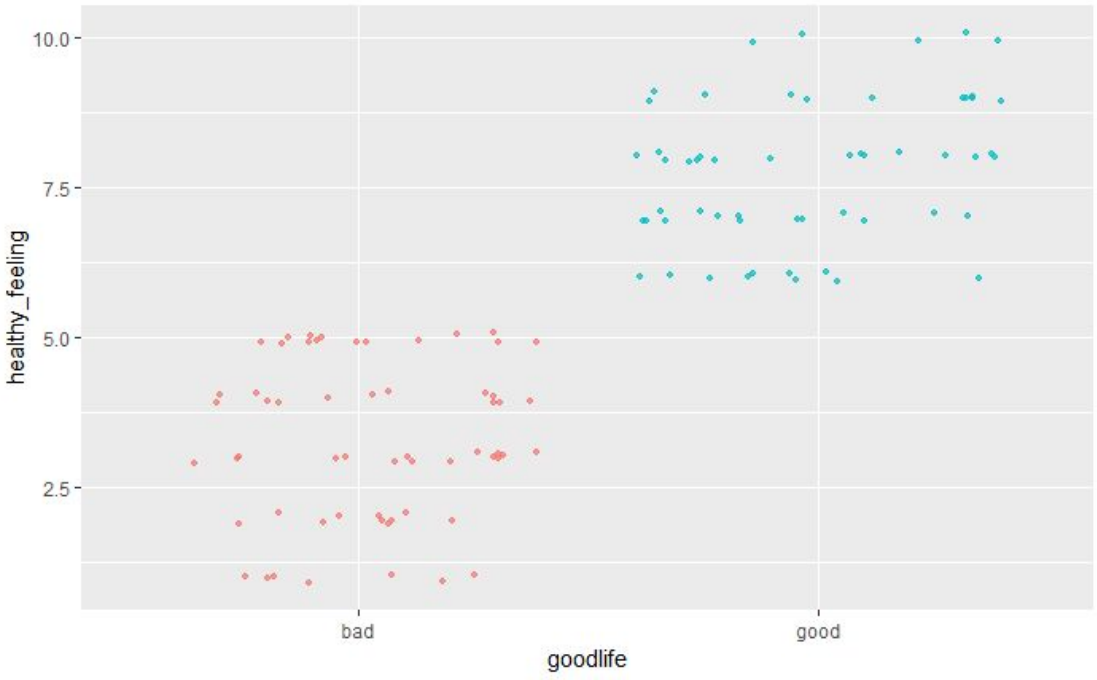

Figure 2. Split between “good” and “bad” for the statement “I feel healthy”

3 **Model Results**

The figures below represent the accuracy, sensitivity, and specificity for each of the methods applied on the models . Figure 3 shows that the extreme gradient boost method was consistently the most accurate across both models.

Figure 3. Measurements for each method used to understand validity of the model

		“I feel healthy”	“I live a rewarding life”
glm	Accuracy	0.7857	0.8214
	Sensitivity	0.7143	0.8667
	Specificity	0.8571	0.7692
svm	Accuracy	0.7857	0.8214
	Sensitivity	0.8214	0.9333
	Specificity	0.8571	0.6923
xgb	Accuracy	0.9286	0.8571
	Sensitivity	0.9286	0.9333
	Specificity	0.9286	0.7692

To better understand model accuracy, figures 4 and 5 below show the confidence intervals based on cross validation for the different methods applied to each model.

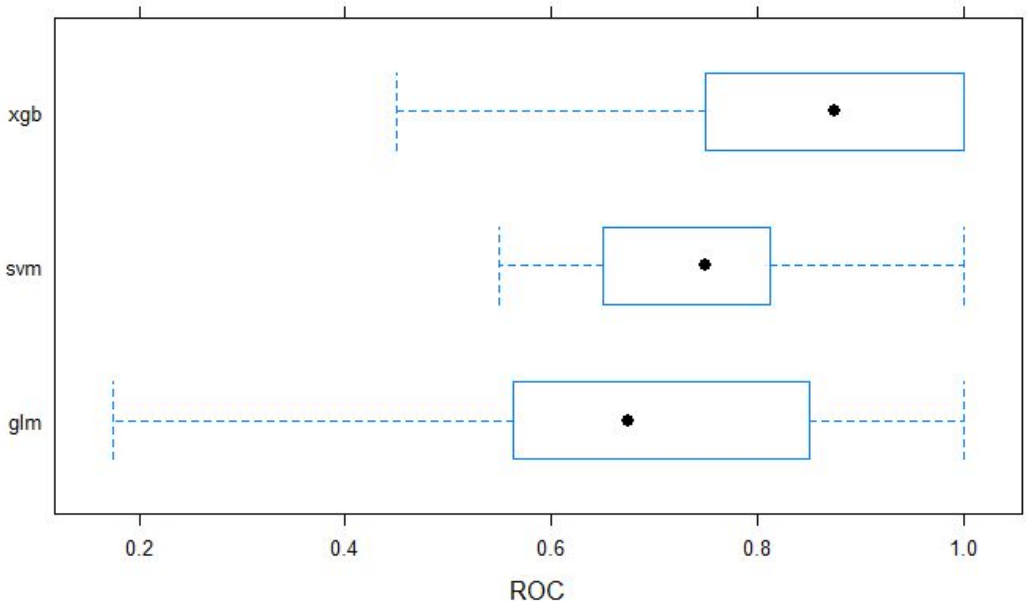

Figure 4. Confidence intervals for cross validation of each method for the prompt“I feel healthy”

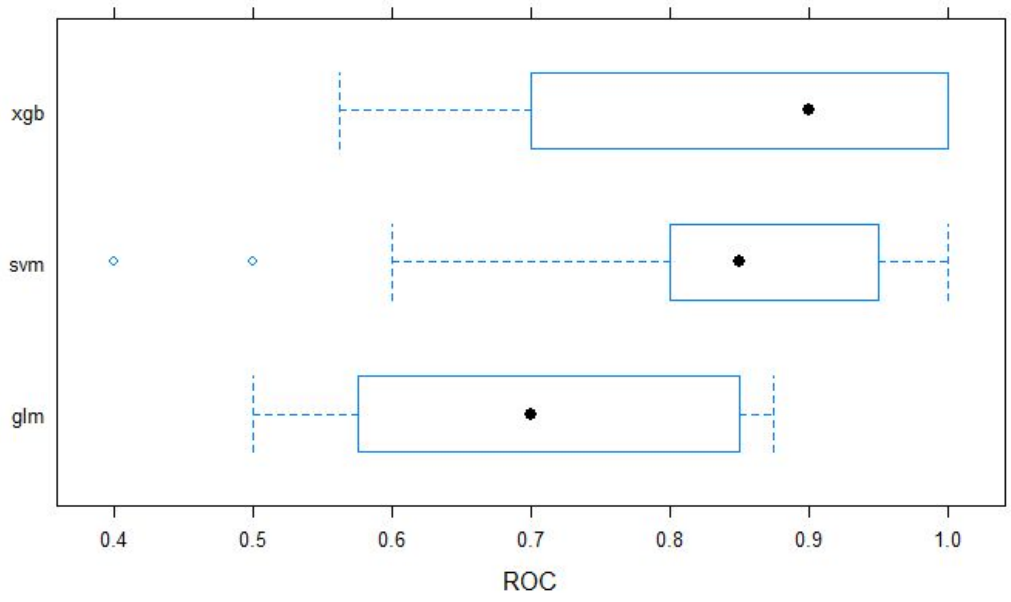

Figure 5. Confidence intervals for cross validation of each method for the prompt “I live a rewarding life”

The graphs in figures 6 and 7 show the predictions of the models compared to the actual values. This allows for a sense of how accurate and valid the model is overall.

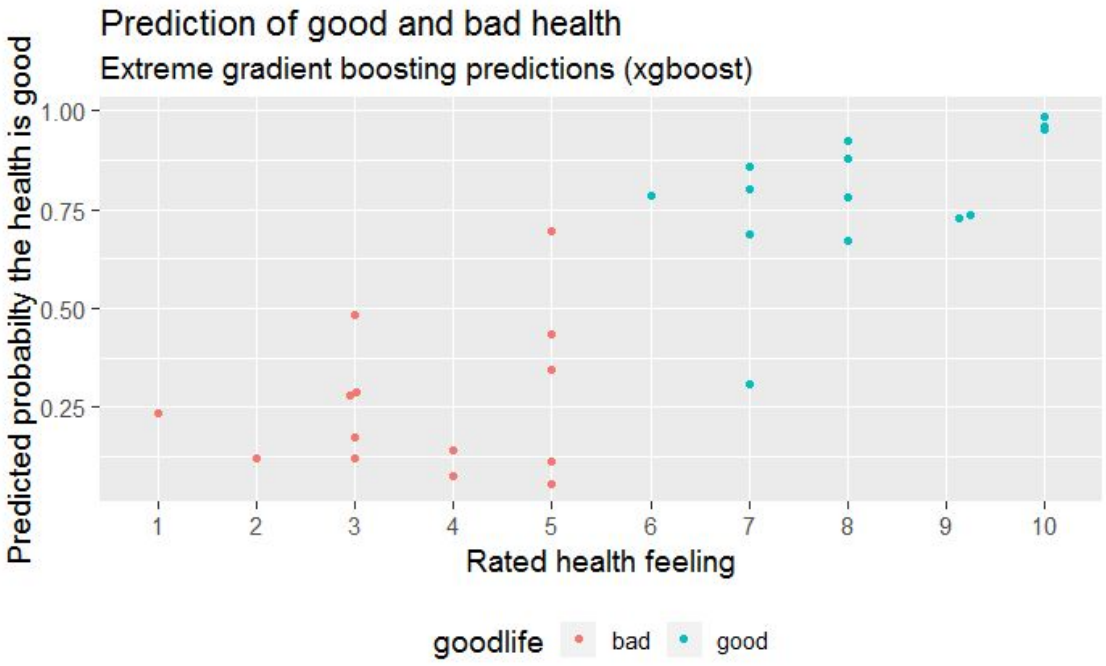

Figure 6. Extreme Gradient Boosting Predictions for the prompt “I feel healthy”

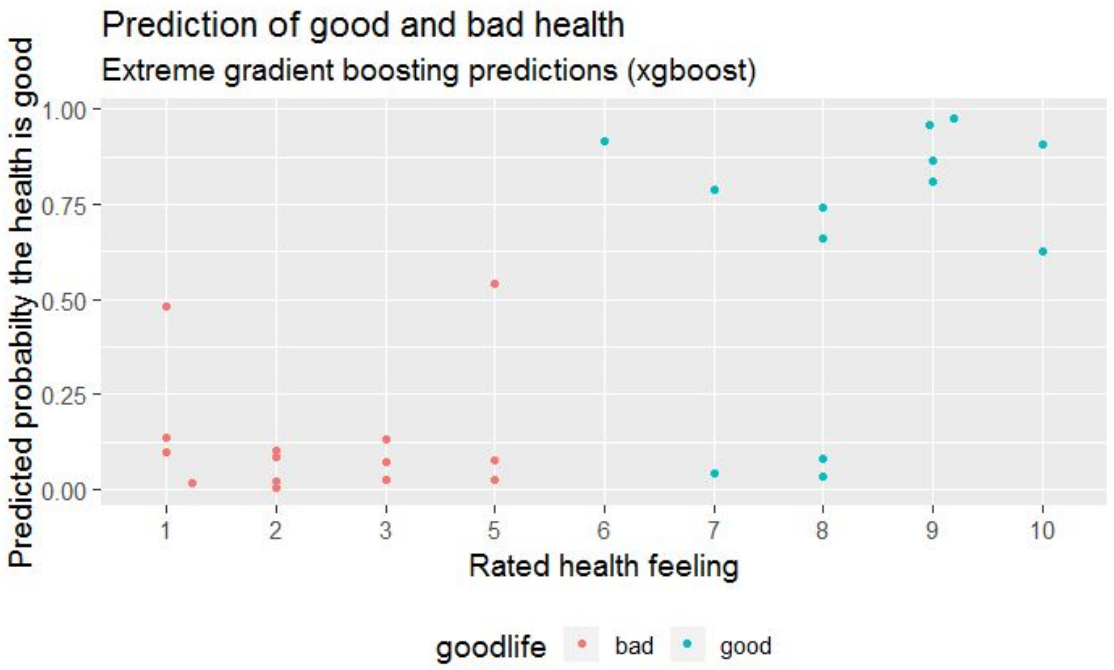

Figure 7. Extreme Gradient Boosting Predictions for the prompt “I live a rewarding life”

In the variable importance graphs shown below, it is observed that the answers to both of these questions are the most important factors in determining each of the models. There exists the possibility that these findings might be further refined by combining unused variables with this information for an even more accurate model.

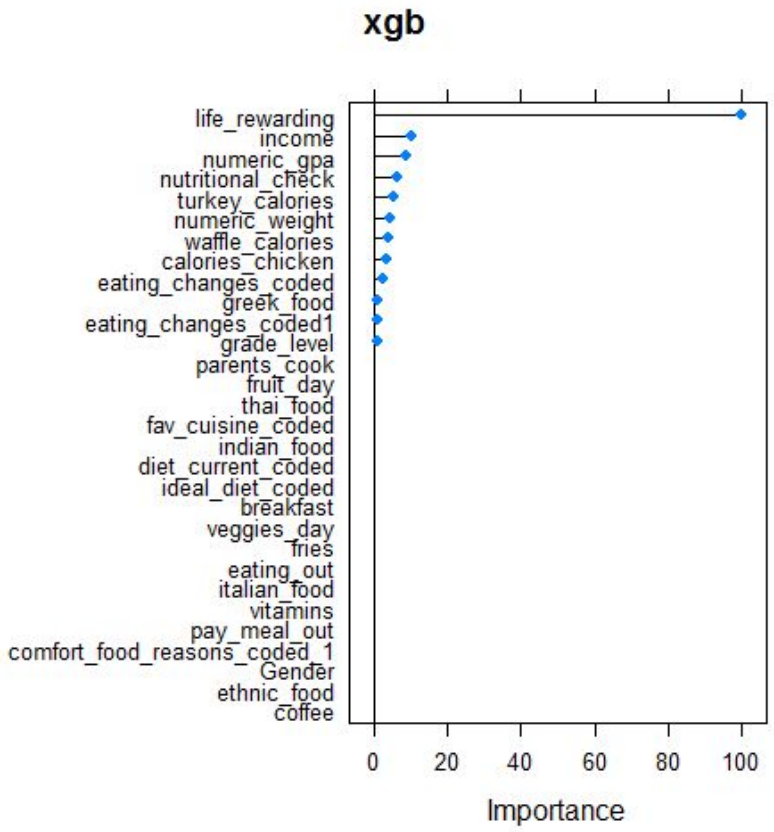

Figure 8. Variable Importance Plot for statement “I feel healthy”

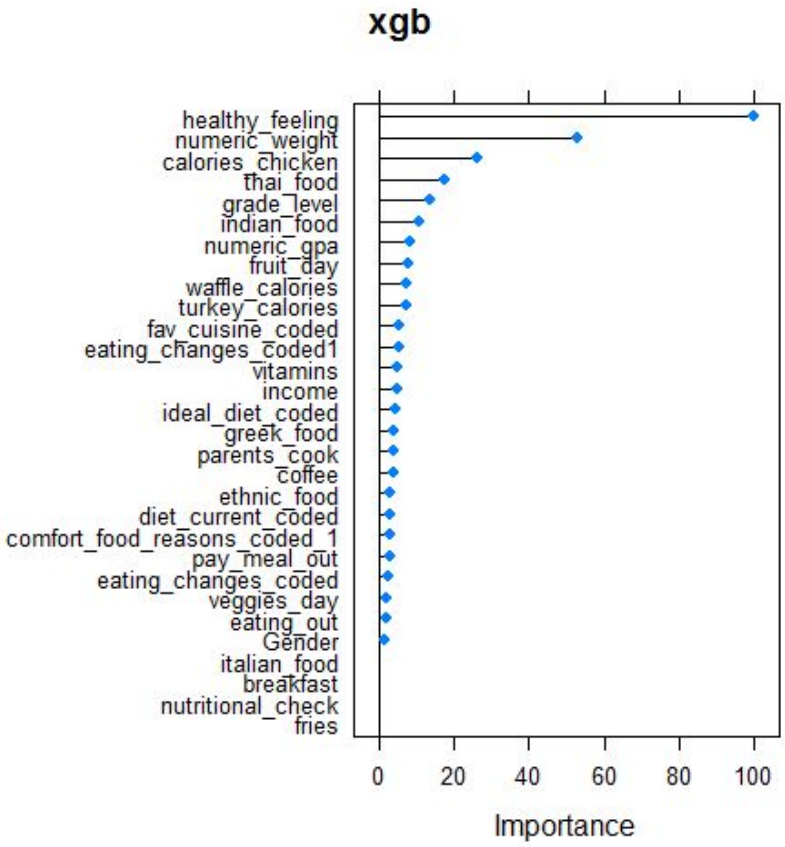

Figure 9. Variable Importance Plot for statement “I live a rewarding life”

To parse the data even further, an attempt was made at using the Lime package in R to increase understanding of how each factor affected the models. As another follow up, I plan to work around R not printing the plots correctly when using this package. Possibly after an update to the Lime package, I will be able to delve into the true affect each factor has on the model. That will significantly strengthen the hunch around the which variables are important in determining whether students reply that they live a rewarding and healthy life and provide for a more quantitative argument.

Figure 10. Lime output example from prompt “I live a rewarding life”

5 **Validity & Generalizability**

I was able to perform holdout and cross-validation on each model. This was done by separating the data into a training and testing data. The model was trained with about 75% of the data and that trained model was tested using the remaining 25%. The results of these tests can be seen above in the Model Results section.

Throughout the creation of the model, the construct validity was strongly considered.This was accomplished through choosing the responses to the statements of “I feel healthy” and “I live a rewarding life”. By focusing on questions that people self reported it is harder to create unfair outcomes. It also leads to less ambiguity around how to interpret the results. There is less medical knowledge required and room for error in taking a person’s health to be close to what they assess it to be than there is to try and calculate it from raw inputs. This might seem overly simplistic, but take, as a counter example, a complex relationship like that between the calories in a steak a student had the night before an exam and the score they received on an exam. The simple approach removes much of the guesswork and avoids the need for any unfounded assumptions and calculations

There was not much control over the external validity or generalizability for this model given that the data set was not collected by me. The utos, or sample, for the experiment was somewhat limited with only 115 student’s responses. Given the power to re-conduct the survey and prove that this model was

more generalizable, I would ensure a larger sample size with a larger variance in student demographics for answering surveyees.

4 **Ethical Concerns**

As discussed during class, there are a number of ethical concerns to consider when creating an algorithm for use by the general public. This model could be unintentionally biased and conclude that a person will have a healthier or more rewarding life if they are female versus male or similar problems drawn down non-contributing characteristics (race is another). It is unclear if this concern is valid, but with further research I will be able to suss these out and prevent them in the model.

Another concern might be that students could potentially see this model and think that their academic standing is not significant because it does not show a high correlation with health of life satisfaction. When presenting on this model it would be important to explain any of these anomalies and to make it clear that these are not the only factors in play and that reason and good judgement should guide above all else.

5 **Conclusions**

Although GPA was selected to drive these models initially, as seen in the variable importance plots, it is not as significant a factor in predicting whether people live a rewarding life or feel healthy as one might think. The most important factor in predicting if a student finds their life rewarding is to know if they feel healthy and vice versa. Several methods were used to come to this conclusion and the extreme gradient boost was consistently the most reliable.

Future will be done on this result set using the Lime package to dig deeper into the complete set of factors and gain a better understanding of how each variable truly affects the model and its outcomes.

References

[1] Panasiuk, R. (n.d.). Retrieved November 21, 2018, from https://www.kaggle.com/rafalpanasiuk/food-choices-data-exploration-analysis/comments

[2] Wei, W., Ramos, P. S., Hunt, K. J., Wolf, B. J., Hardiman, G., & Chung, D. (2016). GPA-MDS: a visualization approach to investigate genetic architecture among phenotypes using GWAS results. *International journal of genomics*, *2016*.

[3] Wickham, H., & Grolemund, G. (2016). *R for data science: import, tidy, transform, visualize, and model data*. " O'Reilly Media, Inc.".

Appendix

Link to Github: https://github.com/katiegroh/859Project

Link to Lime question for future work:
https://stackoverflow.com/questions/53708321/plot-features-is-printing-on-itself