DSC 630 Spring 2020 Term Course Project

David G. Kinney

Bellevue University

**Executive Summary**

It is generally accepted that “eating badly” results in poor health. However, over the course of my lifetime, what constitutes eating healthy has changed, sometimes drastically (such as consuming fat, dairy, etc.). For this project I wanted to see if I could identify any eating habits that would directly correlate to a person being classified as obese (i.e., individuals with a Body Mass Index – BMI – greater than or equal to 30).

I leveraged the American Time Use Survey (ATUS) Eating and Health (EH) Module which contains information related to eating, meal preparation and health in the United States. [1] This overview is provided by the United States Department of Agriculture (USDA) Economic Research Service website. There is quite a bit of data offered by this module, and through my analysis, I was able to eliminate the majority of it for purposes of this study, In the end, I settled on the following factors that I felt would directly affect obesity: whether or not a respondent consumed soda on a regular basis, whether or not they engaged in exercise on a regular basis, whether they consumed fast food on a regular basis, and how they self-described their overall health.

As it turns out, based on the data analyzed, I was not able to find strong correlation between these factors and obesity. At best, there was about a 68% correlation, which is not terribly convincing (I would have liked to see at least 80%). However, I am not convinced there is no correlation. It is worth noting that this data is based on a *survey*. I personally believe that when humans are asked questions that potentially put them in a bad light, they tend to provide some version of the truth. Therefore, perhaps 68% plus the “truth” factor at least might be enough to declare a correlation.

**Technical Report**

# Introduction

The American Time Use Survey (ATUS) Eating and Health (EH) Module was fielded from 2006 and 2008, and again in 2014 and 2016. The EH Module data files contain information related to eating, meal preparation and health in the United States. [1] This overview is provided by the United States Department of Agriculture (USDA) Economic Research Service website.

Individual decisions about how to use the 24 hours in a day have short- and long-term implications for income and earnings, health, and other aspects of well-being. Understanding time use patterns can provide insight into economic behaviors associated with eating patterns as well as the diet and health status of individuals. Knowing more about eating patterns, grocery shopping, and meal preparation, as well as understanding whether participants in food and nutrition assistance programs face different time constraints than nonparticipants can inform the design of food assistance and nutrition policies and programs. [2].

# Methods

The CRISP-DM method was followed for purposes of this project, making use of two Jupyter notebooks. For Data Understanding and Data Preparation, the **Python** programming language was used. For Modeling and Evaluation, the **R** programming language was used.

***Data Understanding***

There are three datasets available within the EH Module:

1. The **EH Respondent** file, which contains information about EH respondents, including general health and body mass index. There are 37 variables.
2. The **EH Activity** file, which contains information such as the activity number, whether secondary eating occurred during the activity, and the duration of secondary eating. There are 5 variables.
3. The **EH Replicate weights** file, which contains miscellaneous EH weights. There are 161 variables.

I eliminated the last two datasets. EH Activity related only to secondary information, which in the interest of brevity, I did not include in the analysis. The Weights file was also not used, although if this was a more extensive study may have added something tangible to the results. The **EH Respondent** file was quite clean; there was no missing data in any of the variables, and all of the data was numeric.

***Data Preparation***

After reviewing the definitions of the original 37 variables, I created a new dataset containing just these:

|  |  |
| --- | --- |
| eusoda | Does the respondent consume soda |
| eufastfd | Has the respondent consumed fast food in the last 7 days |
| eufastfdfrq | How many times has the respondent consumed fast food in the last 7 days |
| eugenhth | How does the respondent rate their overall health on a scale of 1 – 5 |
| euexercise | Has the respondent exercised in the last 7 days |
| euexfreq | How often has the respondent exercised in the last 7 days |
| euhgt | Respondent’s height, in inches |
| euwgt | Respondent’s weight, in pounds |

In addition, I performed some additional data preparation:

* An additional variable was added—**BMI**—calculated as *euwgt / euhgt2 \* 703.* (See [Figure 1](#Figure1), [Figure 2](#Figure2), and [Figure 3](#Figure3))
* An additional categorical feature was added—**obese**--, set as true if BMI > 29.
* Dropped rows where weight was less than 98 pounds or greater than 340 pounds.
* Dropped rows where height was less than 56 inches or greater than 77 inches.

(See [Table 1](#Table1) for a sample of the prepared data.)

For the remainder of my analysis I used the **R** programming language. Using the dataset prepared with **Python**, I first checked for near zero variance. Near zero-variance predictors can cause numerical problems during resampling for some models, such as linear regression. [3] None of the variables posed a problem in this regard ([Table 2](#Table2)).

Lastly, the *euwgt* (weight) and *euhgt* (height) variables were removed. I did not want my models to attempt to predict obesity predicated on an individual’ s height and weight, but rather solely on their eating and exercise habits.

# Results

The EH Respondent dataset was split into Train and Test datasets; each dataset had roughly the same distribution of Obese = True and Obese = False:

set.seed(42)

health\_train <- health[1:8510,]

health\_test <- health[8501:10637,]

print(prop.table(table(health\_train$obese)))

print(prop.table(table(health\_test$obese)))

False True

0.6520564 0.3479436

False True

0.6588676 0.3411324

Various Predictive Models were trained, yielding the results shown in [Table 3](#Table3). The initial model examined was the **C5.0 Classification Tree.** Variable importance was ranked (using the *varImp* function (see [Table 3](#Table3) and [Figure 4](#Figure4)). The *xgbTree* (an eXtreme Grading Boosting model) model returned the best results, with an accuracy of 0.677 and Kappa equal to 0.206. It should be noted that “best” was quite subjective. Initially (and somewhat naively) I planned on using a sum of *accuracy* and *Kappa.*  The Kappa statistic is a measure of concordance for categorical data that measures agreement relative to what would be expected by chance. Values of 1 indicate perfect agreement, while a value of zero would indicate a lack of agreement. However I ended up relying almost entirely on the accuracy measurement.

XGBoost belongs to a family of boosting algorithms that convert weak learners into strong learners. A weak learner is one which is slightly better than random guessing. [4] Refer to Table 4 for a summary of all models evaluated, as well as the accompanying Jupyter R notebook. All models were trained and evaluated using Max Kuhn’s *caret* library.

# Discussion and Conclusion

The American Time Use Survey (ATUS) Eating and Health (EH) Module dataset is fairly robust and conducive to deeper analysis. While I did not achieve the accuracy of predicting obesity based on survey respondents’ eating and exercise habits I had hoped for, given more time I am confident I could have achieved better results. Firstly, I did not apply the weighting file at all. Second, a more experienced and knowledgeable Data Scientist would likely reap better results. Last, as this is survey data, the stoic in me feels compelled to note that humans are not always completely truthful when you are asking them about their (potentially) bad habits.

My conclusion therefore is, 68% predictive accuracy is not great, but given the possible reasons why it was not higher it compels me to say, it’s not terrible either. One other approach would be to also drop the respondent’s self-evaluation of their overall health. This is likely a highly subjective question to ask someone; I feel it is fair to say, most of us probably delude ourselves into thinking we’re in better shape than we actually are. Further, eliminating this variable would truly leave only “lifestyle” variables based on eating and exercise.

# Acknowledgements

I would like to acknowledge my fellow classmates who helped me make it through this course. They were all very supportive and quick to provide advice as well as moral support when I needed it most. And, as always, my wonderful wife Annette, who has put up with me being holed up in my office nights and weekends as I make my way through this Masters program.

**References**

[1] https://www.kaggle.com/bls/eating-health-module-dataset

[2] <https://www.ers.usda.gov/data-products/eating-and-health-module-atus/>.

[3] “Building Predictive Models in R” M. Kuhn, Journal of Statistics Software Vol 28, Issue 5, November 2008 American Statistical Association

[4] “Beginners Tutorial on XGBoost and Parameter Tuning in R” Manish Saraswat n.d. <https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/beginners-tutorial-on-xgboost-parameter-tuning-r/tutorial/>

**Tables**

Table 1

Sample of Prepared Data

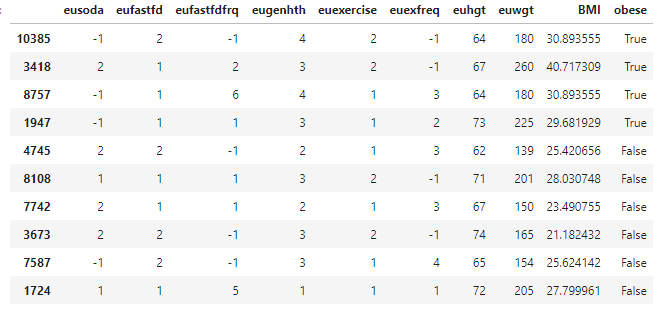


Table 2

Near-Zero Variance

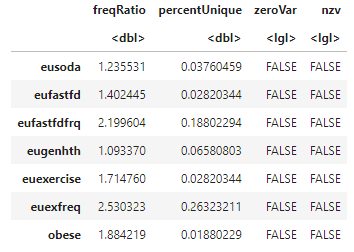


Table 3

C5.0 Variable Importance

|  |  |
| --- | --- |
| eugenhth | 100 |
| eusoda | 71.66 |
| eufastfd | 64.19 |
| euexfreq | 42.19 |
| eufastfdfrq | 3.12 |
| euexercise | 0 |

Table 4

Predictive Models with Accuracy and Kappa Scores

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **function** | **Type** | **Accuracy** | **Kappa** | **A+K** |
| xgbTree | eXtreme Gradient Boosting | 0.677 | 0.206 | 0.883 |
| xgbDART | eXtreme Gradient Boosting | 0.676 | 0.201 | 0.877 |
| deepboost | Classification | 0.675 | 0.204 | 0.879 |
| gbm | Gradient Boosting Machine | 0.673 | 0.186 | 0.859 |
| nnet | Neural Network | 0.672 | 0.191 | 0.863 |
| ada | Boosted Classification Tree | 0.671 | 0.196 | 0.867 |
| rf | Random Forest | 0.67 | 0.176 | 0.846 |
| C5.0 | Classification Tree | 0.669 | 0.213 | 0.882 |
| Jrip | Rule-based Classifier | 0.666 | 0.22 | 0.886 |
| OneR | Single Rule Classifier | 0.665 | 0.155 | 0.82 |
| adaboost | Classification Tree | 0.626 | 0.218 | 0.844 |

**Figures**

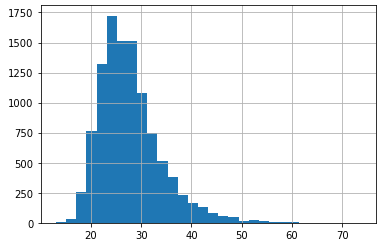
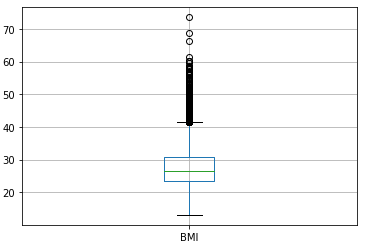
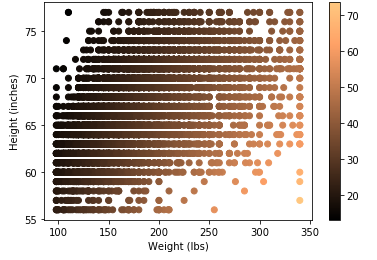


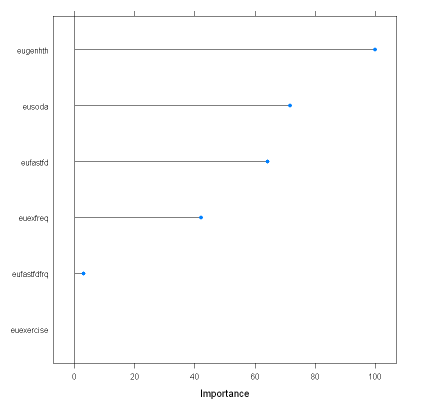
Figure 1. Histogram of calculated BMI



*Figure 2.* Boxplot of BMI



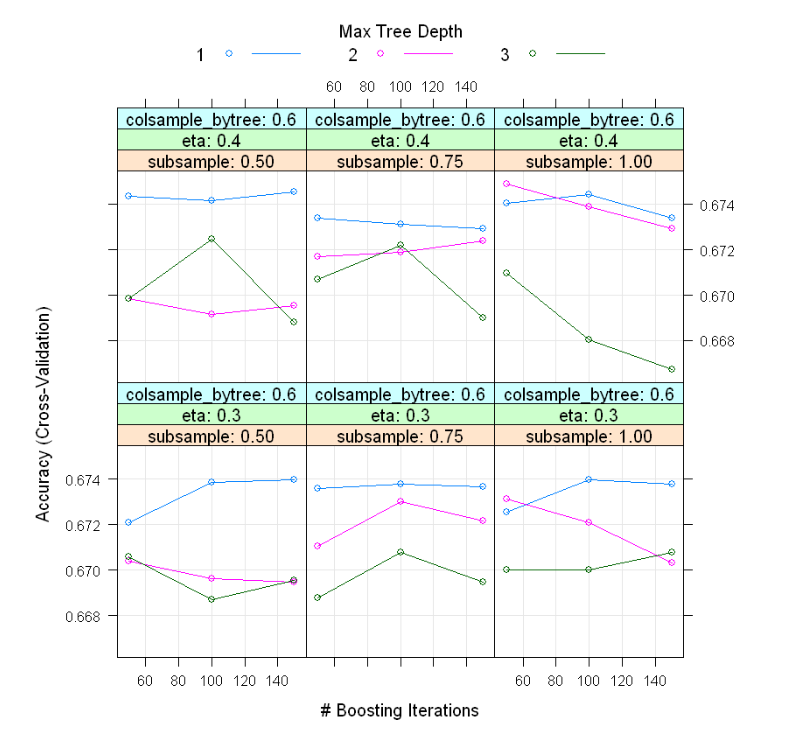
*Figure 3.* Relationship of Height, Weight, and BMI

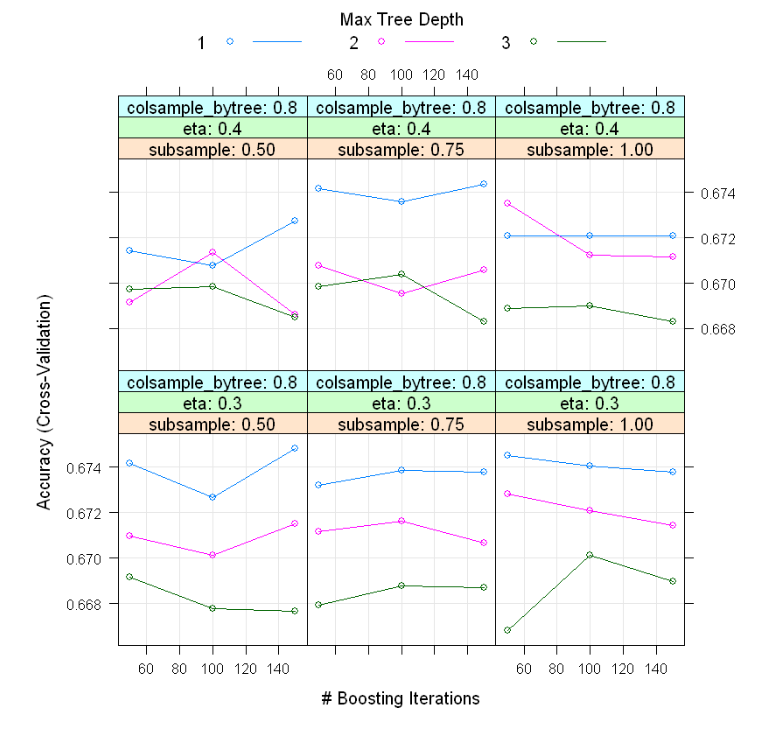


*Figure 4.* C5.0 Feature Ranking

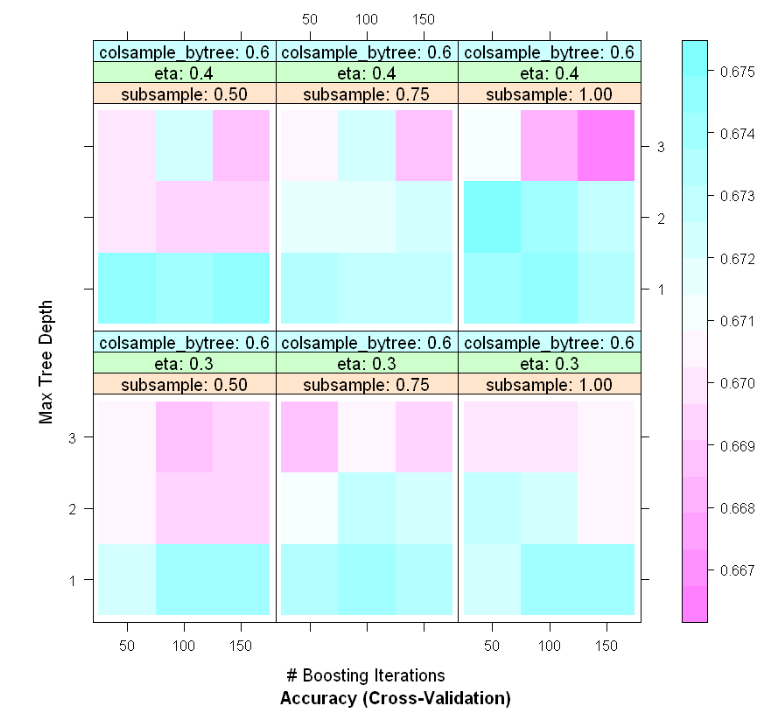
**xgbTree Predictive Model Visualizations**

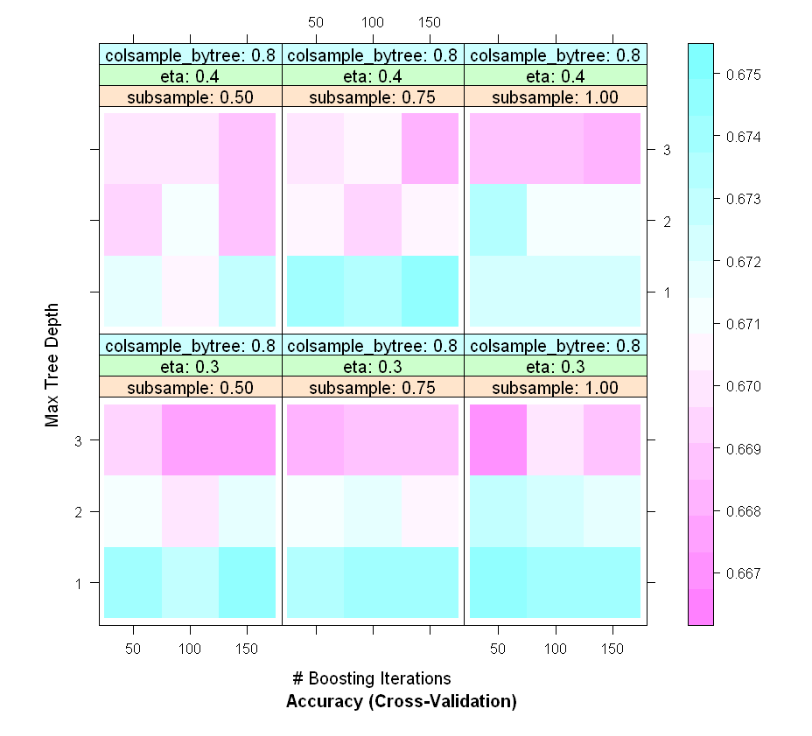
**Classification accuracy versus the tuning factors**





**A level plot of the accuracy values**





**Density plots of the bootstrap estimates of accuracy and Kappa for the final model**

