A Statistical Analysis of Happiness

December 14, 2018

Alexa Bren, Barbara Sudol, Lindsay Finn and Maxwell Wulff

Final Project for ORIE 4740 - Statistical Data Mining

Abstract:

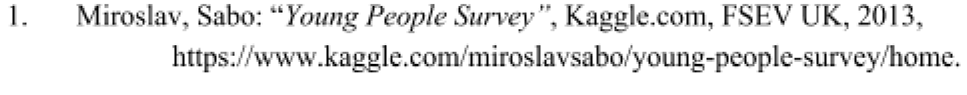
The purpose of this project is to use a survey for the young-age demographic and statistical data mining methods to build a model predicting overall happiness based on 100+ variables, which range from music preferences to lifestyle habits. This analysis could be used, for example, by universities and schools to better allocate funds towards overall well-being. After pre-processing the data to remove individuals who likely answered the survey dishonestly, we applied subset selection, boosting trees, and GAM methods to build and compare models. Furthermore, we distributed the same survey from which the data originated to our peers at Cornell University, and generated real test data to utilize in validating our model. We were able to observe several lifestyle variables that were significant in predicting happiness across all methods, including loneliness, energy levels, a desire to change the past, healthy eating and time spent with friends.

Table of Contents:

1. Introduction

For this project, we were interested in quantifying and analyzing different humanistic measures such as personality traits, habits and feelings. Using the database Kaggle, we found a dataset from a study exploring the preferences, interests, habits, opinions and fears of young people in Slovakia[1]. One thousand and ten Slovakian young adults took this survey ranging from ages 15-30. The survey asked questions in a range of topics, falling into the main categories of: Music & movie preferences, Hobbies & interests, Phobias, Personality traits, Views on life, & opinions, Spending habits and Demographics. For all questions (excluding height and weight), responses were self-reported and measured on a 1-5 linear scale with 1 being low/strongly disagree and 5 being high/strongly agree. Sample questions include: “How afraid of the dark are you?”, “I find it difficult to get up in the morning.”, or “I believe all my personality traits are positive.” Out of all factors, we found the most relevant one to be happiness and decided to analyze how the other factors influence or predict happiness. To measure happiness, we used the (1-5) responses for the question: “I am 100% happy with my life.” The main objective of our project was to determine the strongest predictors of this variable.

We used three popular classification methods to answer our question: best subset selection, boosting trees, and generalized additive models. Our first approach was to use forward, backward and best subset selection to obtain a subset of predictors that most strongly predicted happiness with a linear relationship. Our second approach was to use a tree-based method to stratify the predictor space into sample regions using recursive binary splitting. We decided to use the boosting tree method as this is known to be one of the most powerful tree based models.



[1] Miroslav, Sabo: “*Young People Survey*”, Kaggle.com, FSEV UK, 2013, https://www.kaggle.com/miroslavsabo/young-people-survey/home.

Our third approach was to use generalized additive models to analyze a series of non-linear models based on the subsets of predictors we found from best subset selection and boosting trees.

Lastly, in order to measure how successful our models were, we collected our own sample data by creating a similar survey to the original but with only the top predictors/questions we found from best subset selection and boosting trees. This data made a new test set which we analyzed on our models to check their accuracy.

2. Pre-processing the Data

2.1 Using Averages to Fill Missing Data Points

When preprocessing the data, we noticed that there were a number of missing elements spread out evenly throughout our data set. When removing all observations with any missing data, we lost about 30% of our dataset. This is because the missing data was not localized to a few variables or predictors. Since our data set is of a relatively small size, we decided that a better approach was to replace missing non - categorical data points with the average of that column, with the main purpose to keep as much data as possible. Our data has a small variance, so adding averages does not heavily affect the overall statistics. Furthermore, the missing data is spread over the predictors, so not one single variable had too many average values filled in.

2.2 Detecting Dishonesty

Due to the fact that our data is self-reported responses, we rely on honesty in completing the form in creating an accurate model. Therefore, it is natural to search for individuals who may have answered the survey dishonestly and remove their responses from the set. It will be impossible to know for certain if answers were truly dishonest or not, however for some, we can say with high confidence that this is in fact the case. For the analysis, we chose to look at two different types of dishonesty: answering the survey randomly and answering with the same response repeatedly. With the procedure described in the following sections, we selected 6 people we believed answered randomly, and marked 2 people who answered 90% of the survey with the same response. For the future sections, we removed these individuals from the data set. The indices of the these responses are: 97, 261, 342, 413, 459, 513, 687, 967.

2.2.1 Answering the Survey Randomly

It was difficult in this section to define what could be considered a random answer, so we chose to corroborate two different measures of randomness by looking for outliers in the data set: total distance from the mean answer and Local Outlier Factor (LOF).

To measure the total distance from the mean answer, we first computed the mean response of every predictor. Then for each answer, we added the absolute distance each response was from the mean of each predictor for a total sum of the distance that each answer was from the means of the data. A high distance value will signify individuals who were consistently far away from the mean response on each predictor. We then marked individuals as potentially answering the survey randomly if their distance value was three standard deviations above the mean distance value. We found 8 sets of answers like this.

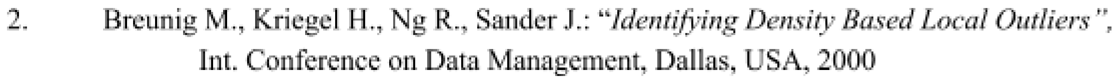
As corroboration and a further analysis of we looked at Local Outlier Factor (LOF). LOF is a method proposed by Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng and Jörg Sander (2000). It is a robust way to analyze and mark potential outliers in a dataset because it takes into consideration density of near clusters, so it is less likely to mark a point as an outlier if it is in a sparsely populated area, but will mark a point if it is a smaller distance away from a densely populated cluster[2]. We performed LOF across a range of 100 to 150 in neighborhood size selection, and chose each highest LOF value from the neighborhood selection. The higher the LOF value, the more likely an answer is to be an outlier. When comparing these points to our 8 marked points from above: 6 of the points above were also in the top 1% of LOF values. By this conclusion we know that these answers where large outliers, and likely answered the survey randomly. We choose not to include these answers in our analysis.

2.2.2 Answering the Survey With the Same Answer Repeatedly

Detecting individuals who answered the survey repeatedly was simple- we used a function that calculated the mode of each response. If the mode of a response was over 90% of the total number of questions, we marked the response as potentially dishonest. There were 2 responses marked. Upon visual inspection of these responses we could see that the individuals had answered with the same response. We chose not include these responses in the model.

3. Approach 0: Baseline

In this approach, we use our intuition to choose predictors of happiness for a baseline model to compare with our later models. We initially reasoned that the most significant predictor variables could be sportiness, BMI, loneliness, smoking, internet usage, and healthy eating. We fit a linear model to these variables and compared the fit (in terms of prediction MSE and R^2 value) to the later models. See Figure 3.1 for the R output for this linear model. We also



[2] Breunig M., Kriegel H., Ng R., Sander J.: "*Identifying Density Based Local Outliers*", Int. Conference on Data Management, Dallas, USA, 2000

predicted happiness against the mean of the happiness column. The results of this baseline are summarized in Figure 3.2.

4. Approach 1: Subset Selection

Since our dataset has a large amount of predictors, we wanted to determine which variables would be most significant in predicting happiness, since we suspected not all of them would be relevant (e.g. participants' music tastes). Our first approach was subset selection. We performed forward, backward, and best subset selection for 5 variables, as well as forward and backward on 10 variables (best subset selection at 10 variables was too heavy computationally). All of the methods for the 5 variable subset selections returned the same set: Energy levels, Loneliness, Changing the past, Personality, and Dreams. We performed a linear regression with these variables as predictors and happiness as the response to see their levels of relevance and if they positively or negatively impacted happiness. As seen in figure 4.1, in this analysis we found that Loneliness and Changing the Past are inversely related to Happiness, and the other predictors are directly related, with Energy levels, Loneliness, Changing the past being most significant.

For the 10 variable selections, both methods also selected the same set of Energy levels, Loneliness, Changing the Past, Personality, Dreams, Fun with Friends, Parents Advice, Reliability, Achievements, and Internet Usage, although some variables were selected in a slightly different order. We performed another linear regression, with results summarized in figure 4.2. We can see that with the inclusion of more variables, some variables from the previous model reduced in significance. The variables that remained the most significant, again, were Energy Levels, Loneliness, and Changing the Past. The variables on Internet usage are by far the least significant.

While this analysis gave us a sense of what may be some of the more important predictors, we wanted to try to fit different models since a linear model may not be the true model of the data.

5. Approach 2: Boosting Trees

Along with subset selection, we also wanted to try a tree based approach to determine the most significant predictors of happiness. To do so we used the boosting tree approach. Boosting involves fitting thousands of trees where each tree is grown using information from the previous, such that the model improves over time. There are a few tuning parameters including: B, number of trees; λ, shrinkage parameter; and d, number of splits in each tree. In order to determine how different predictors affect happiness with a boosting tree, we had to convert the happiness data from a 1-5 scale to a binary scale with 0 being not happy and 1 being happy. To do so, we assigned all Happiness values greater than the mean (3.7) to be Happy (1) and all values less than the mean to be Unhappy (0).

To execute the boosting model, we start by setting the tree function to be zero and all the residuals to be equal to the values in the training set. Repeating these next steps B times, we fit a new tree with d splits. We then update our original tree model by adding a shrunken version of the new tree. Lastly, we update the residuals of the model by subtracting a shrunken version of the new trees from the original residuals. Using the gbm library in R, we executed the boosting tree method and found the results shown in Figure 5.1.

Looking at the summary from our boosting method, we can see that the most influential variables for predicting happiness are: Energy Levels, Loneliness, BMI, Wishing to Change the Past, Height, Number of Friends, Healthy Eating, and Shopping. We tried a number of different parameters for the boosting method and found the lowest classification error rate (~.32) to be when B=2000, λ=.001 (default), and d=4. Greater values of B tended to overfit the data and lesser depths were not complex enough for the number of predictors we were analyzing. Each time we ran the boosting tree algorithm we got slightly different results, but most of the top predictors (Energy Levels, BMI, Loneliness) commonly occurred across all trials. The least important predictor was consistently gender with relative influence of approximately .0005.

6. Approach 3: Generalized Additive Models & ANOVA

Using ANOVA, we were able to compare multiple generalized additive models featuring the output variables of subset selection, variables with the highest relative influence in the boosting trees and a mix of variables from both. We compared the prediction accuracy of each best model to directly compare them. By fitting GAMs with various combinations of splines, 2nd degree polynomials and linear predictor variables, we narrowed down relationships between single predictors and the response. More polynomials and splines were applied to predictors that were proved to have nonlinear relationships with our response variable. The degrees of freedom were determined by visual inspection of data as well as cross validation (K-Fold). In total, we fit GAMs to 6 variables found to be significant in subset selection and used a total of 11 basis functions.

6.1 GAMs Using Top Predictors Predicted by Best Subset Selection

The most influential predictors according to best subset selection were: Energy levels, Loneliness, Wishing to Change the Past, Personality, Dreams, Fun with Friends, Reliability, etc. After performing K-fold cross validation (K = 20) on each variable for polynomials of degree 1 up to degree 4, we found the lowest error for: Degree 1: Loneliness, Wishing to change the past, Dreams, Fun with friends; Degree 2: Personality and Energy Levels.

We defined 6 gams using these variables:

1. Completely Linear
2. Completely linear except: degree 2 polynomial of personality
3. Completely linear except: degree 2 polynomial of energy levels, 2 DF spline of personality, and 2 DF spline of energy levels

4. Completely linear except: degree 2 polynomials of energy levels and personality,
5. Completely linear except: 2 DF splines of energy levels and personality.

Results: The only model with a high F value (low p value) was the GAM with two splines for personality and energy levels (Model 4). This indicates that having a spline with two degrees of freedom for both of these variables, while keeping the other variables linear, is a better model than a simple linear function of the best subset selection variables.

6.2 GAMs Using Top Predictors Determined by Boosting Trees

The most influential predictors according to their relative significance levels and the classification error are: Energy Levels, Loneliness, BMI, Wishing to Change the Past, Number of Friends, Healthy Eating and Mood Swings.

After performing K-fold cross validation (K = 20) on each variable for polynomials of degree 1 up to degree 4, we found the lowest error for: Degree 1: Loneliness, Wishing to Change the Past, BMI, Healthy Eating, Mood Swings; Degree 2: Energy Levels and Number of Friends.

We defined GAMs of the variables similarly to in the analysis above: combinations of degree 2 polynomials and splines against the linear model of the variables. The results for this ANOVA concluded that no model was significantly better than the linear model, so we inspected the variables' relationship with the response visually in order to hypothesize new functions and add them to new GAMs (See section below: *6.2.1 Visualizing Non-Linearity)*.

The new GAMs we tested were:

1. Completely linear
2. Completely linear except: degree 2 polynomial of loneliness, mood swings and BMI
3. Completely linear except: degree 2 polynomial of changing the past, 2 DF spline of energy
4. Completely linear except: degree 2 polynomial of energy, loneliness and mood swings
5. Completely linear except: spline of energy, degree 2 polynomial of loneliness, changing the past and healthy eating
6. Completely linear except: splines of energy and loneliness, degree 2 polynomial of changing the past

Results: Multiple models had low p values relative to the linear model, but two in particular were significant to the 0.001 magnitude: the third and sixth model listed:

> Model 3: Happiness.in.life ~ s(Energy.levels, 2) + Loneliness + poly(Changing.the.past, 2) + Number.of.friends + Mood.swings + BMI + Healthy.eating
> p = 2.765e-08 ***, RSS = 335.55
>
> Model 6: Happiness.in.life ~ s(Energy.levels, 2) + s(Loneliness, 3) +poly(Changing.the.past, 2) + Number.of.friends + Mood.swings + BMI + Healthy.eating
> p = 0.006609 **, RSS = 335.08

6.2.1 Visualizing Non-Linearity

To assist with building our GAMs, we plotted certain predictors against our response to see if we could visually approximate any potential non-linearity. Due to the discrete nature of our data we had to use ggplots with density-colored data points to look for trends. Fortunately, this was a valuable exercise and we saw predictors with clear non-linearity. Examples can be found in Figure 6.1.

6.3 Comparing GAMs using top predictors from both methods

Given our best models from the ANOVA above, we now test them on a subset of our data and compare prediction accuracy. The data was randomly split into ¾ training data and ¼ testing data, and these two partitions of data were used to check accuracy across all four models. Using the predict() function in R, we made predictions of happiness based on the linear model with best subset variables, the best model from the ANOVA in section 6.1, as well as the two best models from the ANOVA in section 6.2. The mean squared errors for each GAM are:

1. Linear model with best subset: 0.4629476
2. Model 4 from Section 6.1: 0.4440928
3. Model 3 from Section 6.2: 0.45379
4. Model 6 from Section 6.2: 0.45799

In summary, the MSE values are similar across all four functions, so they are relatively equally good predictors of happiness, especially because the MSE for all are well below 1 while the response variable is on a 1-5 scale. This suggests that all models predict within 1 point on the happiness scale. To compare to our baseline, R^2 increased by about 0.15 between the baseline and the linear model with best subset variables. The baseline MSE was 0.57430, fairly higher than the MSE of the above models.

7. Testing Our Models With New Test Data

Using the most significant variables discussed above, we compiled a new survey with the corresponding questions and generated new data (from the same age demographic, but within the Cornell University population). The data is in the same scale, with 1-5 responses to each question except for BMI, and there was no need to clean the data as there were no missing data points. We performed the same analysis on this data using our GAMs trained on the original data, and the predict() function on the new data in order to calculate MSE.

Our results:

1. Baseline Model - Predicting Against the Mean: 1.648721
2. Linear Model with Best Subset: 0.7338698
3. Model 4 from Section 6.1: 0.9959549
4. Model 3 from Section 6.2: 1.010603
5. Model 6 from Section 6.2: 0.7189165

The MSE is of similar magnitude to those computed using the real data, however on average it is about 0.3 higher. We conclude that in the context of this University (and by hypothesis, the greater demographic of American University students), the best models to use are model #1 and model #4 in order to make predictions and decisions based on lifestyle data. These models are still within one point of accuracy in terms of the 1-5 response scale and showed significant improvement over our baseline model, further reinforcing that they are robust.

8. Conclusion

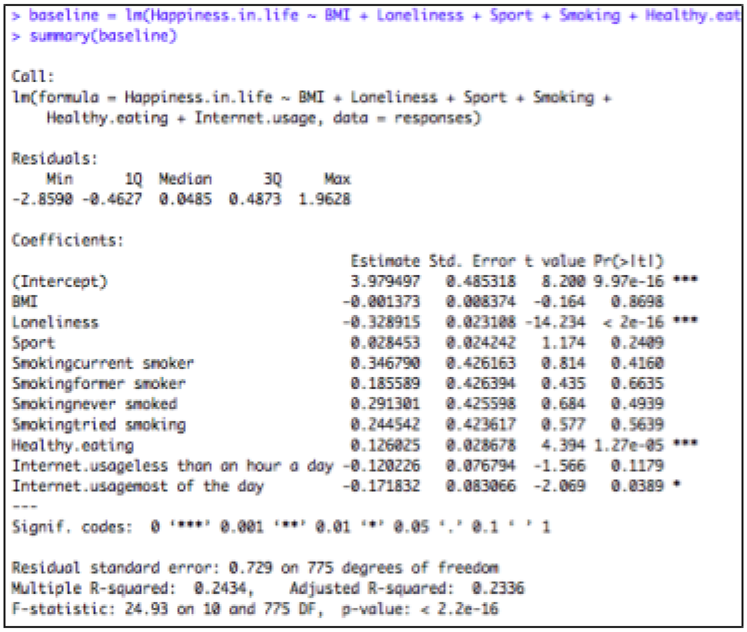
A brief summary of the results from our various approaches. Approach 1: When applying best subset selection on our predictors we were able to see a Residual Standard Error of 0.729, an improvement from the RSE of 0.822 when modeling against the mean value in Approach 0. Approach 2: We used Boosting Trees to determine which variables were the strongest predictors to use in GAM analysis. Approach 3: Amassing variables found in best subset selection and boosting trees analysis we built various GAM models.

Based on the models we created, we can confidently conclude that Loneliness, Energy Levels, BMI, Number of Friends, and Changing the Past in polynomial two are all strong predictors for happiness within the provided dataset. The strong performance of our model on fitting the test data we gathered from Cornell students further validated our model as a reliable predictor for happiness beyond just the scope of the original study.

It is important to note that our definition of happiness is defined as participants' level of agreement with the phrase "I am 100% happy with my life." We realize that this may be a presumptuous definition. Additionally, we recognize that there are an infinite amount of possible predictors that could relate to overall well being, and although we examined many, there may be significant ones that our data does not include.

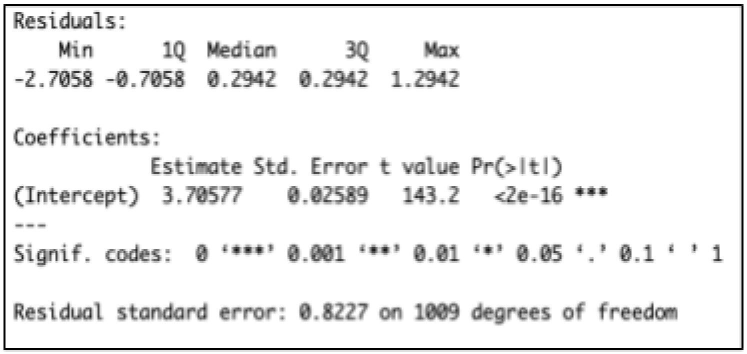
Overall, we hope to use our results to help inform future decisions on health, happiness and well being. For example, this could include sharing these findings to universities nationwide to help assess the satisfaction of their student body, evaluate the effectiveness of their current mental health support systems, and inform future spending initiatives.

9. Appendix: Tables and Figures



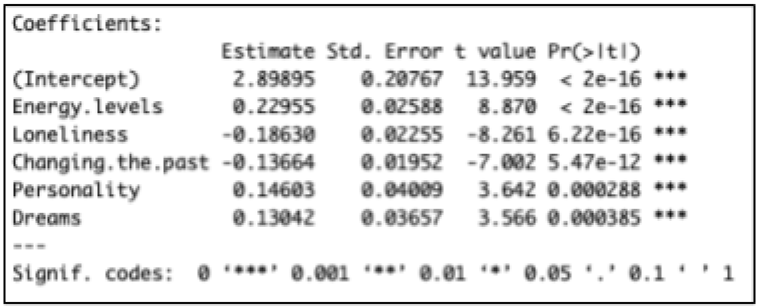
```
> baseline = lm(Happiness.in.life ~ BMI + Loneliness + Sport + Smoking + Healthy.eat
> summary(baseline)

Call:
lm(formula = Happiness.in.life ~ BMI + Loneliness + Sport + Smoking +
    Healthy.eating + Internet.usage, data = responses)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8590 -0.4627  0.0485  0.4873  1.9628

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         3.979497   0.485318   8.200 9.97e-16 ***
BMI                                -0.001373   0.008374  -0.164   0.8698
Loneliness                         -0.328915   0.023108 -14.234  < 2e-16 ***
Sport                               0.028453   0.024242   1.174   0.2409
Smokingcurrent smoker               0.346790   0.426163   0.814   0.4160
Smokingformer smoker                0.185589   0.426394   0.435   0.6635
Smokingnever smoked                 0.291301   0.425598   0.684   0.4939
Smokingtried smoking                0.244542   0.423617   0.577   0.5639
Healthy.eating                      0.126025   0.028678   4.394 1.27e-05 ***
Internet.usageless than an hour a day -0.120226   0.076794  -1.566   0.1179
Internet.usagemost of the day      -0.171832   0.083066  -2.069   0.0389 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.729 on 775 degrees of freedom
Multiple R-squared:  0.2434,    Adjusted R-squared:  0.2336
F-statistic: 24.93 on 10 and 775 DF,  p-value: < 2.2e-16
```

Figure 3.1



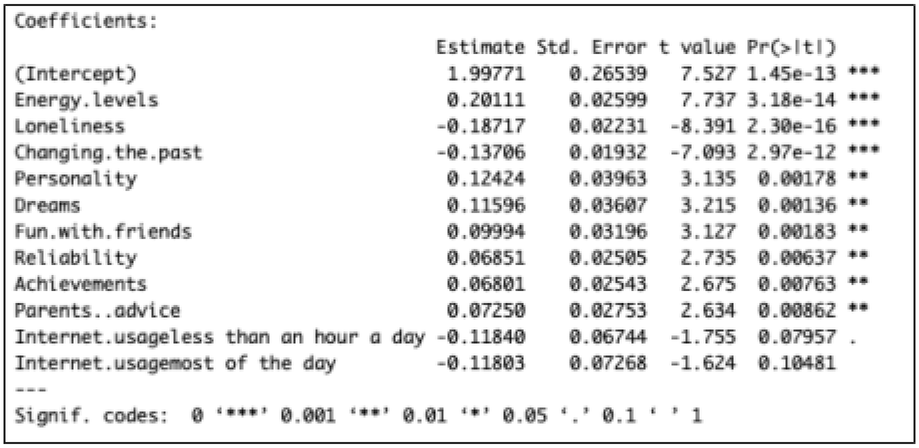
```
Residuals:
    Min      1Q  Median      3Q     Max
-2.7058 -0.7058  0.2942  0.2942  1.2942

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.70577    0.02589   143.2   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8227 on 1009 degrees of freedom
```

Figure 3.2



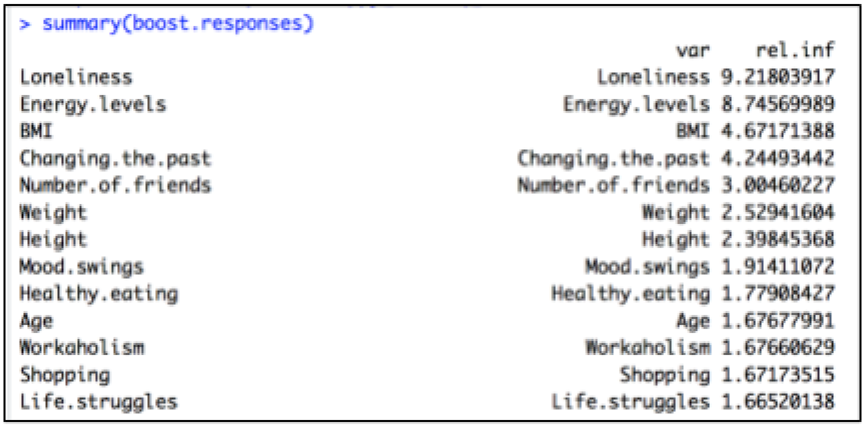
```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.89895    0.20767  13.959  < 2e-16 ***
Energy.levels       0.22955    0.02588   8.870  < 2e-16 ***
Loneliness         -0.18630    0.02255  -8.261 6.22e-16 ***
Changing.the.past  -0.13664    0.01952  -7.002 5.47e-12 ***
Personality         0.14603    0.04009   3.642 0.000288 ***
Dreams              0.13042    0.03657   3.566 0.000385 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4.1



```
Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)                          1.99771    0.26539   7.527 1.45e-13 ***
Energy.levels                        0.20111    0.02599   7.737 3.18e-14 ***
Loneliness                          -0.18717    0.02231  -8.391 2.30e-16 ***
Changing.the.past                   -0.13706    0.01932  -7.093 2.97e-12 ***
Personality                          0.12424    0.03963   3.135  0.00178 ** 
Dreams                               0.11596    0.03607   3.215  0.00136 ** 
Fun.with.friends                     0.09994    0.03196   3.127  0.00183 ** 
Reliability                          0.06851    0.02505   2.735  0.00637 ** 
Achievements                         0.06801    0.02543   2.675  0.00763 ** 
Parents..advice                      0.07250    0.02753   2.634  0.00862 ** 
Internet.usageless than an hour a day -0.11840    0.06744  -1.755  0.07957 .  
Internet.usagemost of the day       -0.11803    0.07268  -1.624  0.10481    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4.2



```
> summary(boost.responses)
                                    var    rel.inf
Loneliness                   Loneliness 9.21803917
Energy.levels             Energy.levels 8.74569989
BMI                                 BMI 4.67171388
Changing.the.past     Changing.the.past 4.24493442
Number.of.friends     Number.of.friends 3.00460227
Weight                           Weight 2.52941604
Height                           Height 2.39845368
Mood.swings                 Mood.swings 1.91411072
Healthy.eating           Healthy.eating 1.77908427
Age                                 Age 1.67677991
Workaholism                 Workaholism 1.67660629
Shopping                       Shopping 1.67173515
Life.struggles           Life.struggles 1.66520138
```

Figure 5.1

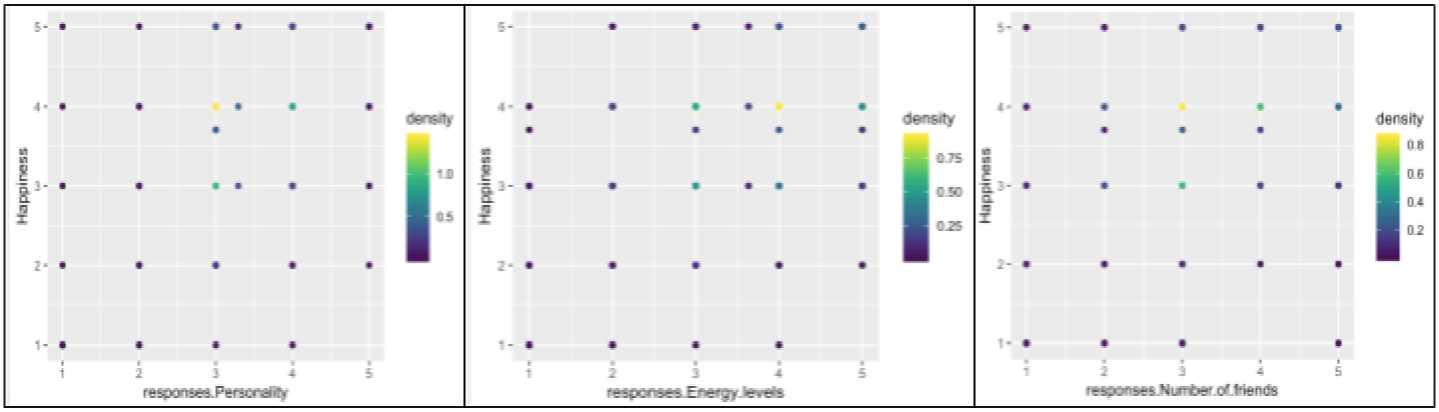


Figure 6.1 These ggplots displaying density of each discrete point illustrate potential non-linearity between a predictor and happiness. These non-linearities were taken into consideration when building our GAM models.