UNDERSTANDING THE HEMINGWAY MEASURE OF ADULT CONNECTEDNESS SURVEY BY UTILIZING DATA ANALYSIS

Ping Ye*, Gildardo Bautista-Maya
University of North Georgia, USA

Contact Email*: ping.ye@ung.edu

Abstract

This paper analyzes the dataset collected from students participating in the Boy With A Ball (BWAB) program, a faith-based community outreach group, through the Hemingway Measure of Adult Connectedness©, a questionnaire measuring the social connectedness of adolescents. This paper first approaches the data in the conventional method provided by the Hemingway website. Then it identifies which questions are strong determiners in deciding whether a student has completed the BWAB program or not. With the goal of utilizing the logistic regression, the set of questions to those only identified as significant in other methods is reduced. These methods include linear regression, decision tree, and random forest. The results are explained from a psychological perspective of social adolescent development.

Keywords: Hemingway scoring method, student characteristics**,** connectedness

1. Introduction

Boy With A Ball (BWAB) is a non-profit organization that works to make cities better places by reaching young people and equipping them to be leaders capable of transforming their communities. The BWAB program works in multiple locations, both across the nation and globally, intending to develop troubled youth and thus develop communities. These solutions include mentoring, faith-based camps, scholarships, and community development. BWAB relocated its global headquarter from St. Antonio Texas to Atlanta Georgia in July 2013. The authors have been worked with BWAB since 2017 to help analyze data and evaluate its mentoring program under the support of the MAA PIC Math Grant and the UNG LEAP into Action Grant. Given the record of the BWAB program, which includes increased academic performance and graduation rates for students who are part of the program, the program believes that getting these kids connected is working positively in satisfying Maslow's needs (Maslow, 1943) and overall improving the student

participants' future potential. The Hemingway Measure of Adolescent Connectedness survey is the first research-based measure of adolescent connectedness. The Hemingway was developed in response to the need for an effective way to evaluate the impact of a high school mentoring program. Utilizing data gathered through the Hemingway *Measure of Adult Connectedness*© questionnaire, administered by BWAB during 2010-2013, the authors analyze question importance through linear regression, decision tree, random forest, and logistic regression. Furthermore, the authors use the Hemingway scoring method to compare participants who have completed the program to those who have not to see which aspects of social connectedness separate the two.

The dataset contains the question answers for 220 ninth-grade students, now are called participants, who were referred through their schools to participate in the BWAB program. 38 of the 220 have participated in BWAB in the past, assigned a “Program” value of 1. 182 of the 220 have not completed the BWAB program, assigned a “Program” value of 0. For future reference, the implication of which group a participant is relevant in the sense that someone who has not completed the program still needs it, while those who have do not. Furthermore, the group assigned with a value of 0 included those who were new to BWAB as well as those who did not complete the program in its entirety. Those who have been assigned a “Program” value of 1 completed the questionnaire as a post-survey. In contrast, those assigned a value of 0 completed the questionnaire finished it as either a pre-survey.

Each survey question was answered as one of the following five categories: “Not at all true”, “Not really true”, “Sort of true”, “True”, “Very true”, and “Unclear”. Generally, “Not at all true” was assigned a score of 1, “Not really true” was assigned a score of 2, “True” was assigned a score of 4, while “Very true” was assigned a score of 5; however, if the question is worded in such a way as to be reverse scored, “Not at all true” was assigned a score of 5, “Not really true” was assigned a score of 4, “True” was assigned a score of 2, while “Very true” was assigned a score of 1. In both cases, “Sort of true” and “Unclear” were assigned a value of 3, whether it was graded reversed or not.

2. Experimental Section

2.1 Comparison of Category Means

By comparing the resulting means, the categories in which the groups coincide or differ can be visualized. Furthermore, the categories in which each group scores low or high on

connectedness can be observed. As far as the Hemingway *Measure of Adult Connectedness*© questionnaire, the scope of the study will supersede past its intended purpose.

The authors begin by using the Hemingway scoring method, which consists of scoring the participants based on question categories. The 57 questions are broken up into ten different aspects of social connectedness: Neighborhood, Friends, Present-Self, Parents, Siblings, School, Peers, Teachers, Future-Self, and Reading. Upon finding the average score for each category, the authors determine whether the difference between those who have completed the program (1) to those who have not (0), and whether the measured difference is significant.

Question values were set to the Hemingway standard from 1 to 5 and the values of reverse-graded questions inverted. Thus, the mean score for each question was taken from each group. Questions were then grouped according to their categories and given a cumulative mean. Without loss of generality, question numbers ending in 1 were from into the category of “Neighborhood”, 2 from “Friends”, 3 from “Present-Self”, 4 from “Parents”, 5 from “Siblings”, 6 from “School”, 7 from “Peers”, 8 from “Teachers”, 9 from “Future-Self”, and 0 from “Reading”.

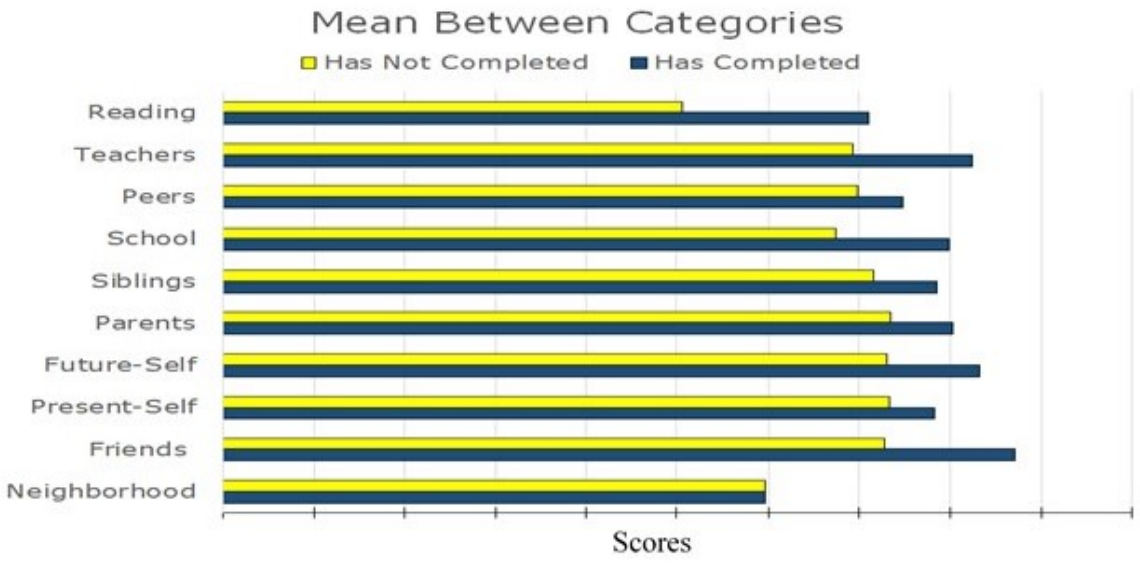

Figure1: Mean Score between Categories

Above Figure1 shows the visualized means between those who have completed the program to those who have not. As stated by the Hemingway manual, scores measuring below a score of 3.5 denote “low connectedness”, while scores at or above 3.5 denote “high connectedness”.

The following table1 summarizes the category means of each group with “high connectedness” being denoted in blue and “low connectedness” being denoted in gold. As can be seen, only one category is marked as denoting low connectedness, which is “Neighborhood”. Meanwhile, the group that has not completed the program has four categories denoting low connectedness, which are “Neighborhood”, “School”, “Peers”, and “Teachers”.

	Neighborhood	Friends	Present-Self	Future-Self	Parents
Has Completed	2.982	4.351	3.912	4.158	4.013
Has Not Completed	2.981	3.639	3.667	3.653	3.672
	Siblings	School	Peers	Teachers	Reading
Has Completed	3.926	3.991	3.737	4.114	3.553
Has Not Completed	3.579	3.372	3.493	3.467	2.522

Table1: Category Means

After computing the mean of each category, the significance of each was determined. Although not every individual question is significant to the 5% confidence interval, the combination of multiple questions results in most differences in the categories being significant.

Even if a category has a noticeable difference, it does not necessarily imply that the value of the questions was significant in determining which group a participant was in. Although the 'Reading' category denotes a large and significant difference between those who have completed the program and those who have not; yet, no question from that category was found to be ultimately relevant in predicting whether a participant was in the program or not.

The results are particularly useful in terms of observing the dataset from a psychological aspect. We note that there is reason to believe in a difference between the social connectedness of those who have completed the program and those who have not; however, we must go beyond the Hemingway's given categories and isolate which individual questions matter most in determining whether a student completes the program or not.

2.2 Identifying Significant Questions

By developing models to select which questions are the best predictors of which group a participant is in, we can narrow the full range of data to a few critical questions. Furthermore, we

can observe the category in which these questions originate, thus finding an aspect of social connectedness in which the program can focus its efforts.

One of the issues of using linear regression is that the model may "overshoot" and predict values that are above the maximum value or below the minimum value. As seen in the linear regression model, over half the entries were assigned a value below 0, even though as a categorical variable, it would never be anything less than 0.

2.2.1 Logistic Regression Model

The ideal model when predicting a dichotomous categorical variable such as the participant needs the program or not would be the logistic regression. Here, the logistic regression model is used to reduce the number of independent variables by removing questions that are not impactful in deciding which participant needs the program.

For all models, the same training set and testing set are used. The training set and testing set are split in the way that the training set contains 70% of the dataset and has a size of 154 while the testing set contains the remaining 30% and has a size of 66.

After applying the logistic regression with significance level 0.05, questions 34, 39, 45, 50, and 56 are statistically significant. The confusion matrix of the testing set is computed, and it shows the accuracy of the logistic regression model is 86.364%.

2.2.2 Decision Tree Model

The authors want to compare different classification models to eliminate the independent variables. The second model used is the decision tree. In a decision tree, the end goal is to assign the entry a probability of whether a participant needs the BWAB program or not. After applying the decision tree model, the end of each branch assigns a predicted probability based on the Hemingway survey responses to a few questions. The decision tree results are as the graph.

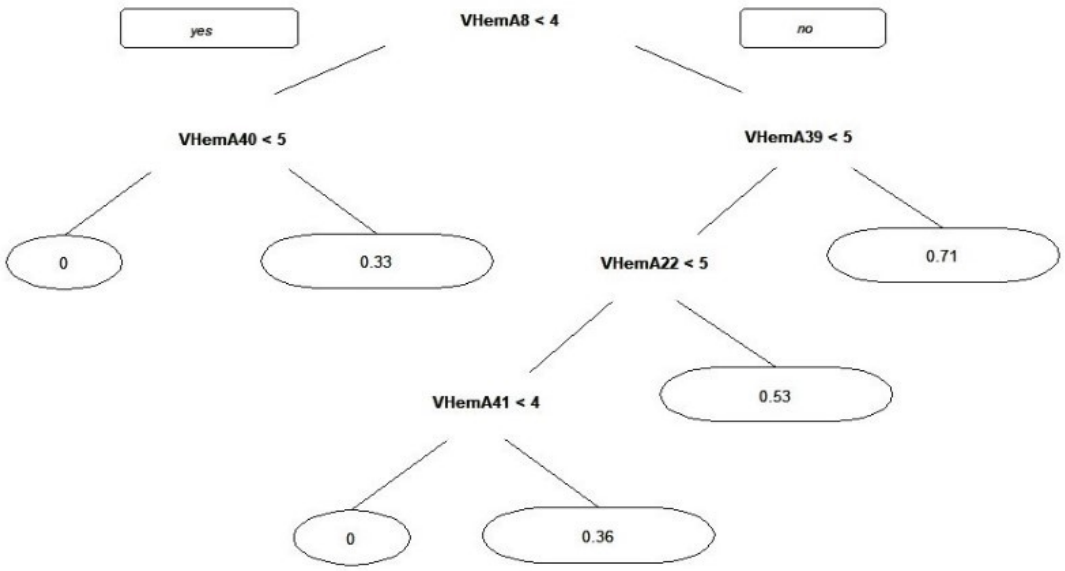

Figure2: Decision Tree Model

The questions that affect the model the most are shown below. The questions with a "variable importance" of 5 or larger are used. Thus, questions 6, 8, 22, 24, 39, 47, 48, and 56 will be taken for using the decision tree model. Finally, the confusion matrix shows the accuracy of the decision tree model is 81.818%.

```
# Variable importance
#  VHemA8 VHemA39 VHemA22 VHemA48 VHemA6  VHemA47 VHemA24 VHemA56 VHemA21 VHemA19 VHemA41 VHemA40
#      16      12       8       6      6       5       5       5       4       4       4       4
# VHemA27 VHemA17 VHemA29 VHemA11 VHemA10 VHemA20 VHemA38 VHemA33 VHemA30 VHemA35 VHemA16  VHemA3
#       3       3       3       2       2       2       2       1       1       1       1       1
```

Figure3: Variable Importance Chart

It is worth noting that while a question may not be present on the decision tree visual, it does not imply that the question is an irrelevant predictor. Furthermore, just because a question is currently on the decision tree does not mean that the question is a good enough predictor. For example, take the branches that split for a participant's response to question 41. It does not matter which answer they may or may not have marked, if the participant reached that part of the decision tree, they would have been assigned a value of 0 through utilizing our testing set.

2.2.3 Random Forest Model

The third model used is the random forest, which is essentially creating multiple decision trees were using the aggregate to find the best predictors. The authors created a random forest of 500 trees to minimize the error. By using the random forest to predict the values on the testing set, it shows that the model predicted results with an accuracy of 90.909%.

As shown in the graph below, the 'purity' of each question is a measure of how influential a variable being to the model. While there is no explicit cutoff to say which questions are more telling, only the variables with purity higher than one are used here. Thus, questions 8, 20, 39, and 40 will be taken from the random forest model.

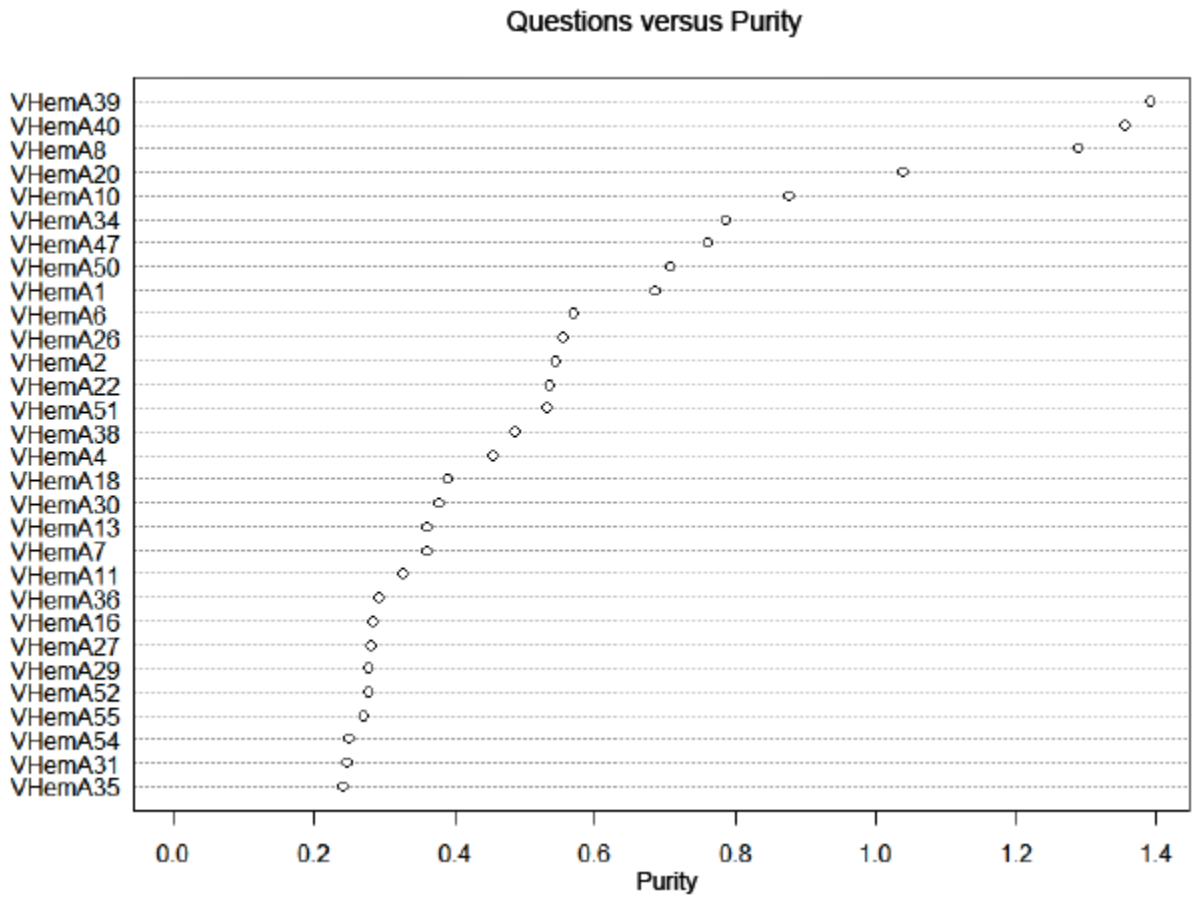

Figure4: Questions versus Purity

3. Results

Through these three preliminary models, the best predictors from each model are taken to use for a refined logistic regression model. By doing so, the number of independent variables is

reduced from 57 to 13. The following table contains the questions marked significant from each model, thus showing that some questions are marked significant in all three models.

Question	Logistic Regression	Decision Tree	Random Forest
6		•	
8		•	•
20			•
22		•	
24		•	
34	•		
39	•	•	•
40			•
45	•		
47		•	
48		•	
50	•		
56	•		

Table2: Significant Questions among Three Models

After applying the new logistic regression model for the above 13 survey questions, it shows that six questions out of the 13 candidates are statistically significant in determining whether a participant needs the BWAB program or not. Furthermore, the authors see whether the correlation between the question and a participant's "Program" value.

With the new logistic modelling complete, the significant questions in order of lowest to highest p-value are shown as the following: Question 8 (0.236%), Question 50 (0.262%), Question 22 (0.468%), Question 39 (0.806%), Question 24 (1.207%), and Question 56 (3.106%). The categories in which these questions came out of are "Teachers", "Friends", "Parents", "Future-Self", "School". Two of those, "School" and "Teachers", are categories in which those who have not finished the program received a mean score considered "low connectedness", which those who have finished the program did not.

4. Discussions

Concerning the Hemingway *Measure of Adult Connectedness*© questionnaire, each group's cumulative mean for a question category is measured and interpreted as "low connectedness" or "high connectedness". For review, a category is interpreted as a sign of "low connectedness" if it receives a score being less than 3.5; conversely, a category is interpreted as a sign of "high connectedness" if it gets a score of 3.5 or higher.

From the previous mean comparison, the group of participants who have not completed the program obtained a lower measure of connectedness in all categories when compared to those who have completed the program. Notably, the scores received in the categories "School", "Peers", "Teachers" were marked as "low connectedness" in the group of those who have not completed the program. In contrast, those who have completed the program were marked as "high connectedness". These categories have a real-world connection in the sense that these categories are shaped by what the participant experiences at school. Additionally, the highest-scoring category for any mean is the score for 'Friends' from the group that finished the program, who had a mean score of 4.351. It indicates that friends may be more critical than siblings in the adolescence period, as the focus is on expanding relationships beyond the family.

From the refined logistic regression result, it shows how disproportionally represented the "Teacher" and "School" question categories are when compared to any other category on the survey. Out of the six questions found significant, three were from the categories "Teacher" and "School". Questions such as question 8 ("I care what my teachers think of me") and question 50 ("I usually like my teachers") stress the importance of teacher interactions in a student's life, with question 8 questioning how much the participant values their teacher's opinion and question 50 asking how the participant feels about their teachers. Relating to the mean comparison, question 22 ("Spending time with my friends is a big part of my life") highlights the importance of spending time with friends. Such friends may surround the participant at school, in the program, or in their neighborhood.

An interesting result from the logistic regression model is that questions 24 ("I enjoy spending time with my friends is a big part of my life") and 56 ("Doing well in school is important to me") have a negative correlation associated with them. This is to say, the more 'True' the statement was to the participant, the more likely they were to be considered part of the group which has not completed the program, thus implying that they still needed the program. Yet, if the

participant's response is accurate, perhaps they were well socially connected such that they did not need the program to begin with.

A notable observation is the accuracy of our decision tree model and the random forest model. The accuracy obtained from performing the modeling on the testing set denotes that this accuracy is not a sign of overfitting, but rather, that there exist questions whose responses are strong determiners of which group the participant is in, thus reinforcing our motive of isolating these questions for future study.

While we focused on what was significant, consider the categories which were not significant, “Neighborhood” and “Siblings”. Both groups scored almost identically in the category of “Neighborhood” with those who have completed the program scored a 2.982 while those who have not scored a 2.981. Furthermore, these scores are considered a sign of “low connectedness” in Hemingway. While the striking similarity between the groups is puzzling, it is worth considering that the data was collected during 2010-2013, a time whether technology does not require participants to be physically associated in terms of the questions asked in the questionnaire. Finally, it’s worth noting that questions that reside within the “Sibling” category were the ones that were not filled out and had to be provided a substitute. As a result, many students have an unbiased score of 3 for questions in this category, helping explain the lack of significance in this category.

Without a doubt, there is more than meets the eye in any data analysis. Adolescents’ level of “connectedness” to family, school, friends, and self has been found to contribute to academic performance but also predict violence and substance use. Fortunately, the school environment directly influences students’ levels of connectedness such that connectedness appears malleable to school-based interventions. Though the group consisting of those who completed the program scored higher than those who have not completed the program, the authors cannot establish that participation in the BWAB program definitively caused this change; however, what the authors can say is that those students with low scores in certain categories could benefit from the program, i.e., a pre-survey could be used to identify the students who need the BWAB program. The authors would like to utilize the above data analysis tools to the BWAB Atlanta data of the year 2014-2020 for future study to discover the importance of connectedness for teaching and learning in multilingual, multiracial and multicultural Atlanta school environments.

References

[1] The Hemingway Measure of Adult Connectedness©
http://adolescentconnectedness.com/media/Hem5.5short.pdf

[2] Analyzing Student Data as a Measurement of Success for Boy With A Ball by *Ye, P. & Allen, D., etc., Mathematics Teaching Research Journal VOL 10, N1, 2018.*

[3] An Introduction to Statistical Learning, with Applications in R *by James, Witten, Hastie and Tibshirani: Springer, 2013.*

[4] Data Science For Business-What You Need to Know About Data Mining and Data-Analytic Thinking. *Foster Provost and Tom Fawcett.*

[5] Fundamentals of Machine Learning for Predictive Data Analytics Algorithms, Wored Examples, and Case Studies. *John D Kelleher, Brian Mac Namee, and Aoife D'arcy.*

[6] Maslow and Your Self Esteem Needs. *Karl Perera. More-selfesteem.com. Better Internet Bureau, n.d., 12 Jan. 2015. Web. 25 Mar. 2017*

[7] Maslow's Hierarchy of Needs. Simply Psychology. *Saul McLeod. Creative Commons, 16 Sept. 2016. Web. 25 Mar. 2017.*

[8] Probability and Statistics. *MIchael J Evans and Je_rey S Rosenthal.*

[9] Statistical Design and Analysis of Experiments. *Wiley, NY. 1989.*