Integrating ML/AI with Health and Wellness

Team M2S2 (Jason McCoy, Jeff Moore, David Shaw, Sudha Solayappan)

1. Introduction with Context

1.1. Survey Description

We are a health and wellness data and analytics solutions company. Our business revolves around being an independent third-party surveyor and benchmark provider of health indicators of employees in a company.

As an independent third-party vendor, we ask employees through a confidential survey about their eating, sleeping, exercise and behavioral habits. No individual data will be shared with the client; data privacy is of utmost importance.

Clients would use this data to determine how effectively they could promote healthy alternatives to their employees to reduce sick days while improving productivity without violating the Health Insurance Portability and Accountability Act (HIPAA) or any other applicable labor laws.

1.2. Research Question

How to build a benchmark of healthy behavior habits adopted in the United States for US-based companies to compare against? The context is to feed the health behavioral habits into the aggregation to be compared against benchmarks for the United States.

Clients can then use our data and analysis to design their personalized programs to incentivize employees to adopt better eating, exercising, sleeping and reduce addictive habits.

2. Sample Plan with Justification

2.1. Population Determination

The survey will be conducted one company at a time. When the survey is launched for a company, the population would be all the employees in the company including all the levels regardless of title and role. It is assumed that the population makeup is different for each company. Depending on the population size of the client we will use one of two methods:

First, if the number of employees in an organization is less than 500, then a census survey would be done. No sampling would be done and hence no inference required. The target response rate would be a minimum of 85% since this is a census survey. This target would be achieved by (a) partnering with our client's management to announce the commencement of the survey and convey the importance that the management places on all selected employees completing the survey. Issue weekly reminders from individual line supervisors reiterating the survey's importance and encouraging employees to complete the questionnaire truthfully. These periodic reminders can be delivered through established communications channels such as email and weekly team meetings, and (b) creating awareness among the employees about the importance of the survey.

Second, if the number of employees exceeds 500, then a stratified sampling would be done. The sample determination process is described in the Section 2.2.

2.2. Sample Determination

If the number of employees in a company exceeds 500, then stratified sampling would be conducted to make inferences about the population. This is done to save time and cost associated with surveying large number of employees and possibly incur biases because of lower response

rates. Therefore, our method is to use a stratified sample to identify target employees in the client organization.

In order to obtain the stratified sample, we would follow the steps outlined by the University of California at Davis:

1. Name the target population.
2. Name the categories (stratum) in the population.
3. Figure out what sample size is needed.
4. List all the cases within each stratum.
5. Make a decision rule to select cases (for example, we might select the items using the largest set of random numbers).
6. Assign a random number to each case.
7. Sort each case by random number.
8. Follow the decision rule (#5 above) to choose the participants.

(Statistics How to, 2018)

The characteristics that are critical to our research and define the strata are:

(a) Gender;

(b) Age group;

(c) Job Type – office-based vs. field roles, sales vs. other field-based roles, research vs. admin roles, managerial vs. non-managerial roles, etc.;

(d) Ethnicity; and

(e) Income levels - Low, Medium and High;

Based on the employee data from the company's personnel system, we would obtain data on gender, age, income levels, ethnicity, etc. This data would then be used to make strata for out stratified sampling. From each stratum, employees would be:

(a) chosen randomly, and

(b) the number of employees chosen would be proportional to the total number of employees in that strata.

The survey would be conducted over a two-week period. The idea behind taking a stratified sample is that the sample should be a good representation of the population. If we do simple random sampling, it is possible that no employee is selected from certain strata or categories, leading to sampling bias.

To ensure complete survey data integrity each participant who is selected for the survey will be required to submit a fully executed questionnaire. We provide survey assistance to our clients to ensure participation targets are met by designing and implementing an incentive-based program that is built on past models. This might include token gift cards offering small value cash back or percent discount on health-based commodities and/or services to those employees who complete the survey as desired. To protect the privacy of the participants the survey would be conducted anonymously via a secure website. The survey questionnaire would be sent to all the employees in case the company has less than 500 employees. For companies with more than 500 employees, the survey questionnaire will be sent only to those selected from our sampling process. The sample size would be large enough to ensure sufficient precision. The minimum employees selected from any company would be 500. To determine the sample size, we would refer to Table 1. For our survey purpose, 95% confidence level is selected with 2.5% margin of error. The chosen sample size would be max of 500 and the sample size obtained from Table 1. Thus, for a company

with less than 500 employees, all of them would be chosen (census). For a company with 1000 employees, 606 of them would be chosen in the sample and then put into different stratum.

Table 1: Determining Sample Size

	Confidence level = 95%			Confidence level = 99%		
	Margin of error			Margin of error		
Population size	5%	2,5%	1%	5%	2,5%	1%
100	80	94	99	87	96	99
500	217	377	475	285	421	485
1.000	278	606	906	399	727	943
10.000	370	1.332	4.899	622	2.098	6.239
100.000	383	1.513	8.762	659	2.585	14.227
500.000	384	1.532	9.423	663	2.640	16.055
1.000.000	384	1.534	9.512	663	2.647	16.317

Table 1 is obtained using Cochran's formula. The Cochran formula allows to calculate an ideal sample size given a desired level of precision, desired confidence level and population size (Statistics How to., 2012). The Cochran formula is:

$$n_0 = \frac{Z^2 pq}{e^2}$$

Where:

(a) e is the desired level of precision (i.e. the margin of error)

(b) p is the (estimated) proportion of the population which has the attribute in question, q is 1 – p. For maximum variability, p = 0.5

(c) Z is based on the confidence level chosen. Z = 1.96 for 95% confidence level.

The sample size calculated is then modified in the above formula by using the equation:

$$n = \frac{n_0}{1 + \frac{(n_0 - 1)}{N}}$$

Here n_0 is Cochran's sample size recommendation, N is the population size, and n is the new, adjusted sample size.

The next section talks about our plan to analyze the data that we collect from our survey procedures and how to sell them to the client.

3. Data Analysis Plan

Before getting into the specifics of the data analysis plan the survey team decided that only completed surveys would be accepted. Employees are free to enter and exit the survey as many times as they would like (saving each time before exiting the survey) but the survey can only be submitted when all of the fields have a valid answer. This precautionary step will lessen the likelihood of incomplete data/surveys. On the other hand, introducing this step is likely to reduce the total number of surveys received. This risk has been factored into the sample size requirements.

3.1. Data Visualization

Since the survey will open and close to all participants per the same timeline, there is no need to phase/stage data collection and analysis efforts. The current exploratory data analysis plan includes 4 steps. The first step is to import or load the data into a data mining or visualization tool so that preliminary graphs and correlations can be performed. Line charts and/or histograms will be built to show basic data like number of men vs. women, age groups, ethnicity groups, number of responses by income level, and a breakdown by job type or role.

3.2. Correlation and Basic Statistics

The second step in the data analysis process will be to build a correlation matrix as well as perform basic statistical calculations. The matrix will allow the survey team to graphically show all of the variables compared with each other. Looking for tightly correlated features will be an important input for the next step in the data analysis plan as these features or variables will be

studied further. In addition, basic statistics calculations will show means, modes, mins, maxs and the quartiles while box and whisker diagrams will allow the team to see outliers. The outliers will serve as another input into the next phase of the analysis plan. Any numerical variable that is statistically higher or lower than the baseline data will also need to be further analyzed.

3.3. Detailed Data Analysis

Step 3 in the process is complete the detailed analysis on the correlated variables, the outliers and the variables that are statistically outside of the baseline data for a given category like gender or ethnicity. The step is all about determining or understanding why the results are as they are. Understanding the data and the why's will serve as input into the final phase of the data analysis plan which is the report out including action plan and/or recommendations.

3.4. Report Out

Making recommendations based on insights discovered or developed is perhaps the most important step in the process. Clients/companies want to know what their specific data shows as well as actions can be taken in effort to address significant issues. While making the recommendations the survey team will be careful to highlight any important assumptions made in order to complete the analysis. Lastly, the survey team is always interested in improving so feedback will be captured in order to make future survey and data analysis efforts better.

3.5. Managing Bias

The survey was designed and randomized samples were chosen in an effort to address four different types of bias. The survey design team discussed proper survey population definition, randomization (including the samples selected and normalizing effects of geography on survey results), blinding (important that survey responses are anonymous and cannot be traced) and publication (both good and bad survey results will be made known). Once all of the data analysis

steps are completed, the team will perform one additional bias assessment before releasing the results to the client.

4. Survey Questionnaire with Routing Logic

The survey questionnaire that would be sent to the employees in the sample is described below. There would be a test survey before the final survey to identify the limitations of the survey.

1. Rate yourself on a scale of 1 to 10 on your perception of your fitness. (1 means least fit and 10 means very fit). __

   Comments, if any: _____

2. Stack rank the items from most important to least important when it comes to your health (1=Most important, 5=Least Important):
   a. Working out __ (Ex: Working out for at least 30 minutes twice or thrice a week)
   b. Eating right __ (Ex: Including 2-3 servings of fruits or vegetables in your diet)
   c. Sleeping well __ (Ex: Getting 8 hours of uninterrupted sleep)
   d. Letting go of addictions __ (Ex: Not feeling compelled to smoke or drink alcohol)
3. In your mind what is the biggest gap when it comes to your health (choose one):
   a. Working out
   b. Eating right
   c. Sleeping well
   d. Letting go of addictions
   e. None
4. What is your height in inches? _____ (1 feet = 12 inches)

   (A calculator for conversion to be included)
5. What is your weight in lbs? _____ (1 kg = 2.2 lb)

(A calculator for conversion to be included)

6. Looking back at the last week, how many filling meals did you typically eat each day? (A filling meal is anything that is not a snack)
    - <3 __
    - 3 __
    - >3 __
7. On a typical day, during which meal do you eat the most?
    - Breakfast __
    - Lunch __
    - Dinner __
8. Do you go to the gym? Yes__ No__

Branching Logic: If answer to Question 8 was Yes, then show Question 9

9. How often do you go to the gym?
    - 1-2 days a week __
    - 3-4 days a week __
    - >4 days a week __
10. Does your organization have a gym? Yes __ No__

Branching Logic: If answer to Question 10 was No, then show Question 11

11. Would you go to the gym if there is a free/subsidized membership gym in your organization? Yes __ No__
12. Do you exercise outside of the gym? Yes __ No__

Branching Logic: If answer to Question 12 was Yes, then show Question 13

13. How often do you exercise outside of the gym?
    - 1-2 hours a week __

- 3-7 hours a week __
- >7 hours a week __

14. In a typical week, how many times do you eat in a restaurant?
    - 0-1 meal __
    - 2-5 meals __
    - >5 meals __
15. On an average, how many hours do you sleep in a day?
    - <6 hours __
    - 6-8 hours __
    - >8 hours __
16. How often do you go for a regular check-up in a year?
    - Never __
    - Once __
    - 2 or more times __
17. Do you smoke? Yes__ No__

Branching Logic: If answer to Question 17 was Yes, then show Question 18

18. How many cigarettes do you smoke in a day?
    - <3 __
    - 3-10 __
    - >10 __
19. Do you drink alcoholic beverages? Yes__ No__

Branching Logic: If answer to Question 19 was Yes, then show Question 20

20. How many days in a typical month do you drink one or more alcoholic beverages?
    - <3 days__

- 3-8 days __
- >8 days__

5. References

Davenport, T. H., & Harris, J. G. (2017). Competing on Analytics The New Science of Winning. Boston, MA: Harvard Business School Publishing Corporation.

Niles, Robert. (2006). "Robert Niles' Journalism Help: Statistics Every Writer Should Know," RobertNiles.com. Retrieved from http://www.robertniles.com/stats/.

Statistics How to. (2012). Sample size in statistics (how to find it): Excel, Cochran's formula, general tips. Retrieved from http:// www.statisticshowto.com/probability-and-statistics/find-sample-size/.