AN ANALYSIS OF THE ATTITUDES TOWARDS MENTAL HEALTH IN THE TECH INDUSTRY BY MEANS OF CLASSIFICATION TREES, REGRESSION TREES AND TRADITIONAL METHODS

William Pritchard

Austin Peay State University

August 2019

Contents

1 Abstract

This paper explores the attitudes regarding mental health in the tech industry and seeks to create a predictive model to improve the mental health of their employees. Our data was gathered via web survey hosted by Open Source Mental Illness (OSMI) and contains over one-thousand participants of various demographics. Using random forest models, our results suggest that by creating an open environment open to mental health needs, we can increase the likelihood that at-risk employees will seek treatment.

2 Introduction

In our current job landscape, nine out of ten startups will fail (Patel, 2015) and established tech companies have been reported to be the most stressful places to work (Simoes, 2013). We can see how a large mental burden may be placed on tech industry employees. For this paper, we will be using the data provided by Open Sourcing Mental Illness (OSMI): Mental Health in Tech Survey 2017-2018 to explore the attitudes towards mental health in the tech industry. Our goal is to create a predictive model for the common predictors of mental illness and to explore prevalent attitudes towards mental health in the tech industry.

According to Forbes, only 25% of IT jobs are held by women (Davis, 2018). In addition, the Journal of Applied Social Psychology published an article by Rubin, Mark, et al. establishing that sex is an important factor when analyzing mental health in the workplace. This information will motivate further exploration in our dataset.

This paper will be divided into several sections: data cleaning, exploratory data analysis (EDA), traditional methods for analyzing survey data, and model building with regression and classification trees. All methods will be explained in detail before they are implemented. For regression and classification trees, we will make heavy use of software packages in R. All R code may be found in the appendix.

3 Data Cleaning

The data we are working with in this paper is a combination of two datasets that were posted on Kaggle.com: “OSMI Mental Health in Tech Survey 2017.csv” and “OSMI Mental Health in Tech Survey 2018.csv”. In addition, extensive changes have been made to allow for easier analysis. Most of these changes were made in Excel prior to being imported into R. Since several field names were lengthy (some over 200 characters long), they have been condensed or reformatted. Several fields have been removed entirely.

This dataset was gathered using a web form. As is traditional with surveys, conditional logic was implemented. As a result, certain fields may have numerous NA values. For example, a question such as “Have you been diagnosed with a mental disorder?” could exclude further questioning for those that answer with “No”. These NA’s may not be a large issue for traditional survey methods, but that will be for decision trees and random forest methods. To prevent this from skewing our analysis, we have excluded certain fields with an excess number of NAs or that contained values that we deemed unreliable.

Since our survey did not ask the respondents their sex, we have parsed the gender data to create this variable. This process was performed in Excel. For our purposes, we have defined “sex” as the biological and physiological characteristics that define men and women (Mills, 2011). Only responses that answered male or female, or anything similar (e.g. “m”, “f”, “cis male”, “cis female”, etc.), to the “gender” question have been included. All others were categorized as “Other”.

Our cleaned data still contains several different variables; some are of limited use to our analysis. The data provided in comment fields will be analyzed using word cloud. These comment fields will then be excluded from the dataset once we build our decision trees.

4 Exploratory Analysis and Traditional Survey Analysis Methods

4.1 Exploratory Analysis

Effective exploratory analysis can be seen to serve two major roles: familiarize the researcher with the data and present questions that may warrant further analysis. As we can see in figure 1, our data is mostly men in the US. And, our age distribution is about as expected with a majority being around their 30's.

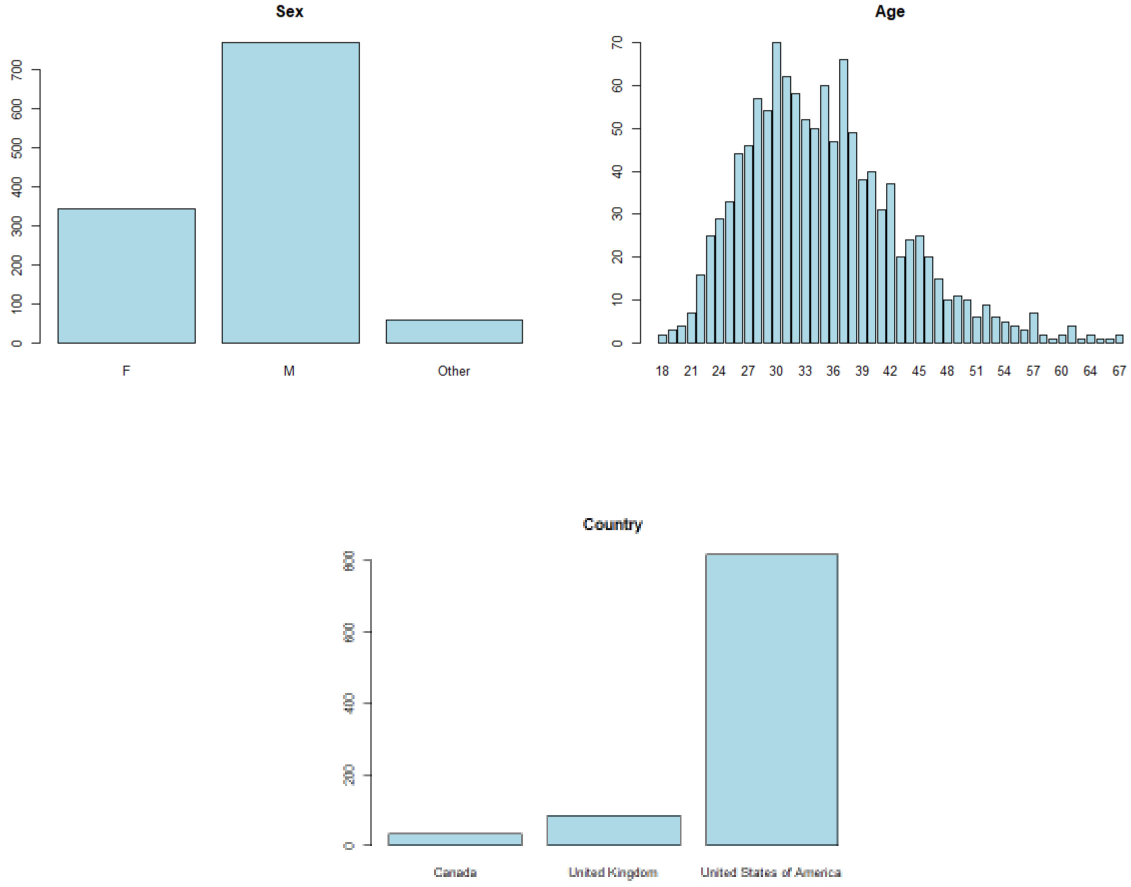


Figure 1: Sex, age, and top 3 countries for the OSMI survey respondents.

Our initial exploration will regard the sex of the respondent. We will begin our inquiry with a word cloud analysis for our comment data. For brevity, we will only be focusing on a few questions asked in this survey: describe the circumstances of a badly handled response to a mental health situation and describe what you think the industry as a whole could do to improve the mental health support of its employees.

Describe the circumstances of a well-handled response to mental health issue.

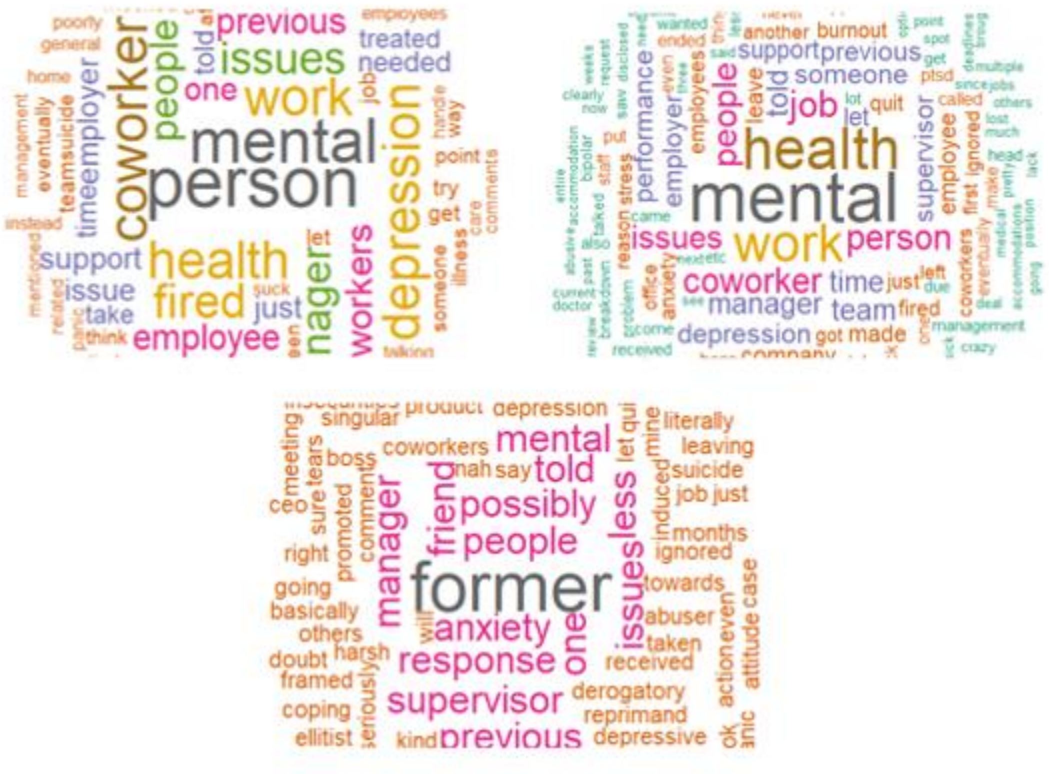


Figure 2: From left to right, a word cloud for the male, female, and “Other”(bottom figure) responses.

Words clouds can be a useful when analyzing unstructured data. When looking at badly handled responses to a mental health situation at work (figure 2), there doesn’t appear to be much difference between the male and female responses. Although, the word “former” for the those categorize as “Other” is interesting. When looking back at the specific responses that contain the word “former”, there appears to be anxiety that mental health issues may permanently damage their career.

Our word clouds for the improving mental health conditions is far less illuminating (figure 3). Given the nature of our survey, the words “mental” and “health” are expected to be common. We did create a word cloud for all unstructured text data (not displayed here). The outcome was similar to the word clouds below.

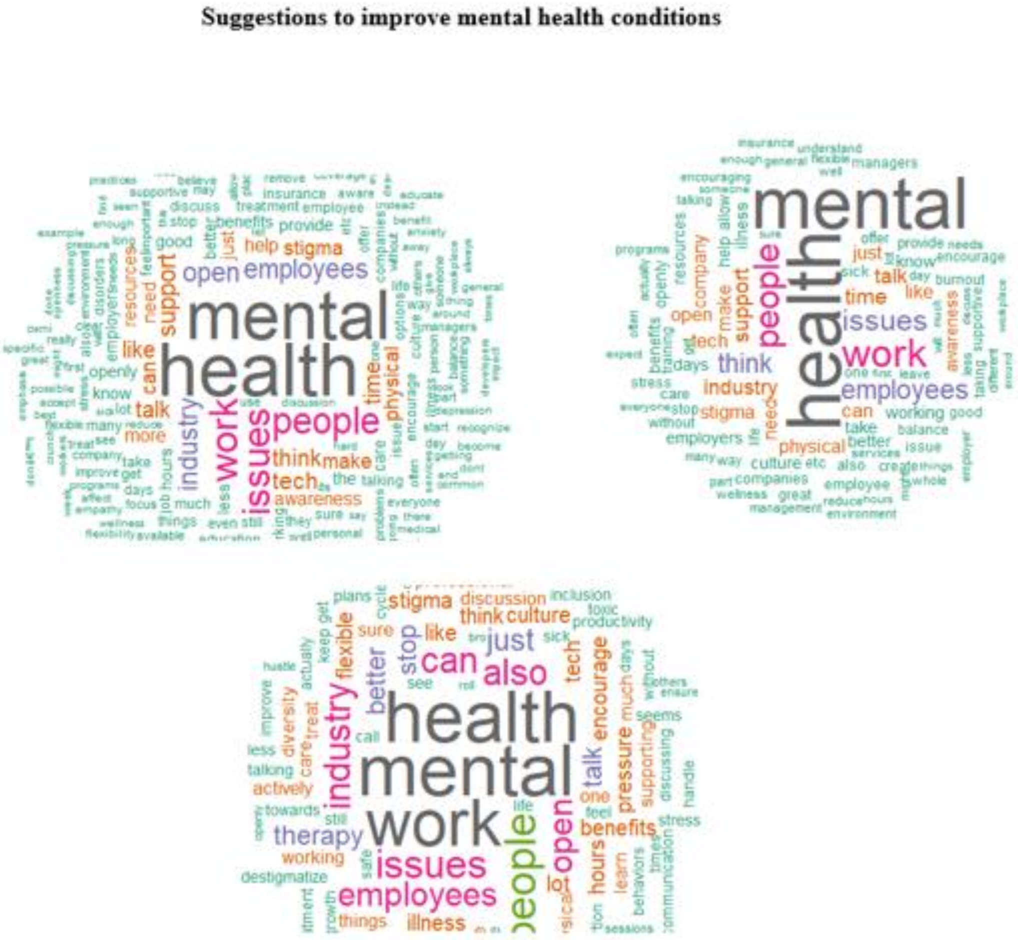

Suggestions to improve mental health conditions

Figure 3: From left to right, a word cloud for the male, female, and “Other”(bottom figure) responses.

4.2 Traditional Methods

It is common practice in survey analysis to utilize chi-square tests to assess independence among variables. Although ANOVA is not uncommon, survey data tends to contain a lot of categorical variables. This lends it well to cross tabulation and chi-square tests. Later in this section, we'll discuss the dimension reduction via principle component analysis (PCA) and multiple correspondence analysis (MCA).

We'll be starting by analyzing the awareness of mental health options on insurance coverage. Our data suggests that the sex of the respondent plays a major role in the awareness of mental health options. Specifically, it is shown in figure 4 that the observed values were lower than expected for males and higher for females. That is, males were not as aware of mental health options under employer's insurance coverage.

Aware of insurance coverage for mental health issues

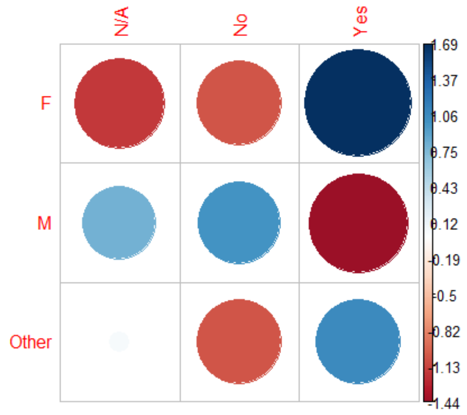


Figure 4: Residual values for the chi-square test.

Next, we will explore the sex basis for openly identifying as having a mental health issue. As figure 5 suggests, we observed more females answering "Yes" than expected and less for males. Those in the "Other" category, twice as many answered "Yes" than expected.

Openly identified as having a mental health issue

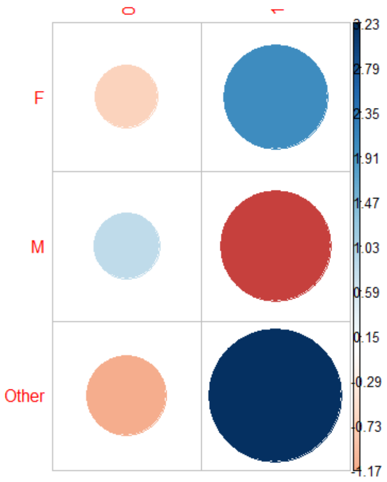


Figure 5: Residual values for the chi-square test. Here, 0 is "No" and 1 is "Yes".

It will be important for us to look at whether the sex of our respondent plays a role in their seeking treatment. Our data suggest there is a strong dependence here. In figure 6, we can see that more females sought treatment than was expected. Conversely, far fewer males sought treatment than would be expected.

Sought Treatment

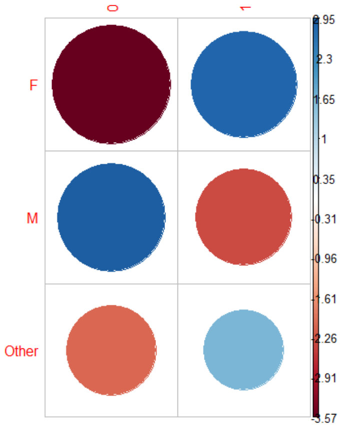


Figure 6: Residual values for the chi-square test. Here, 0 is "No" and 1 is "Yes".

Of course, more time could be spent analyzing confounding and other interactions between our variables. For the next sections, we will be using decision trees to create predictive models for our target interests: employees seeking treatment.

These tests are interesting but exhausting since our survey contains over 50 different variables. Techniques such as principle component analysis (PCA) and multiple correspondence analysis (MCA) are most often utilized. Since our data is recorded as largely categorical information, MCA will work better. Figure 7 shows our initial scree plot.

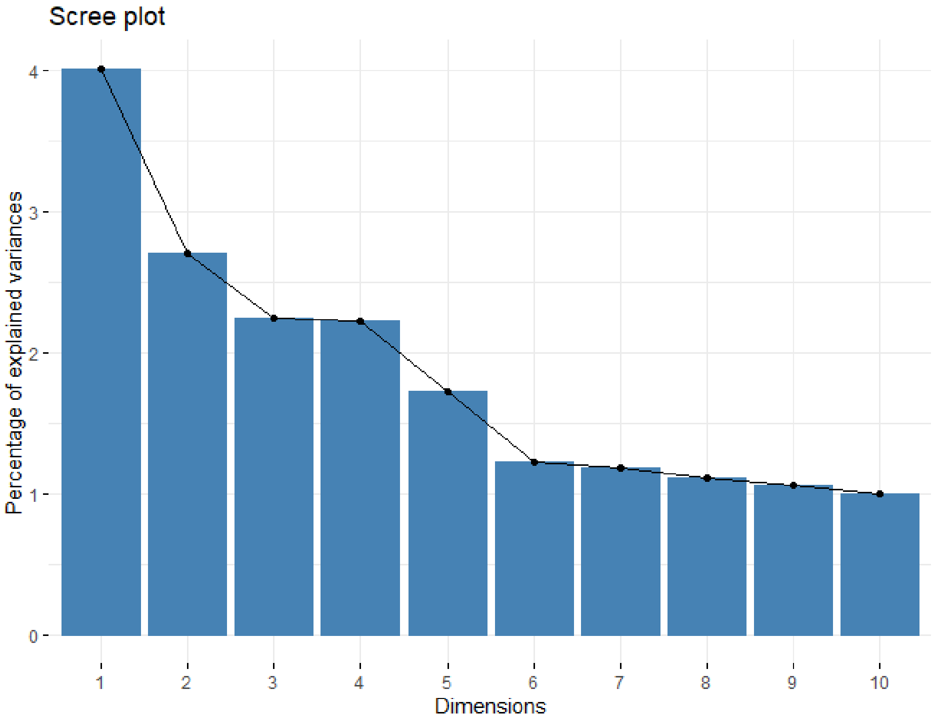


Figure 7: Scree plot for MCA on survey data

Our scree plots show us that a handful of components contribute a small amount to our overall variance. Looking at our contribution to dimensions 1 and 2 may be more insightful.

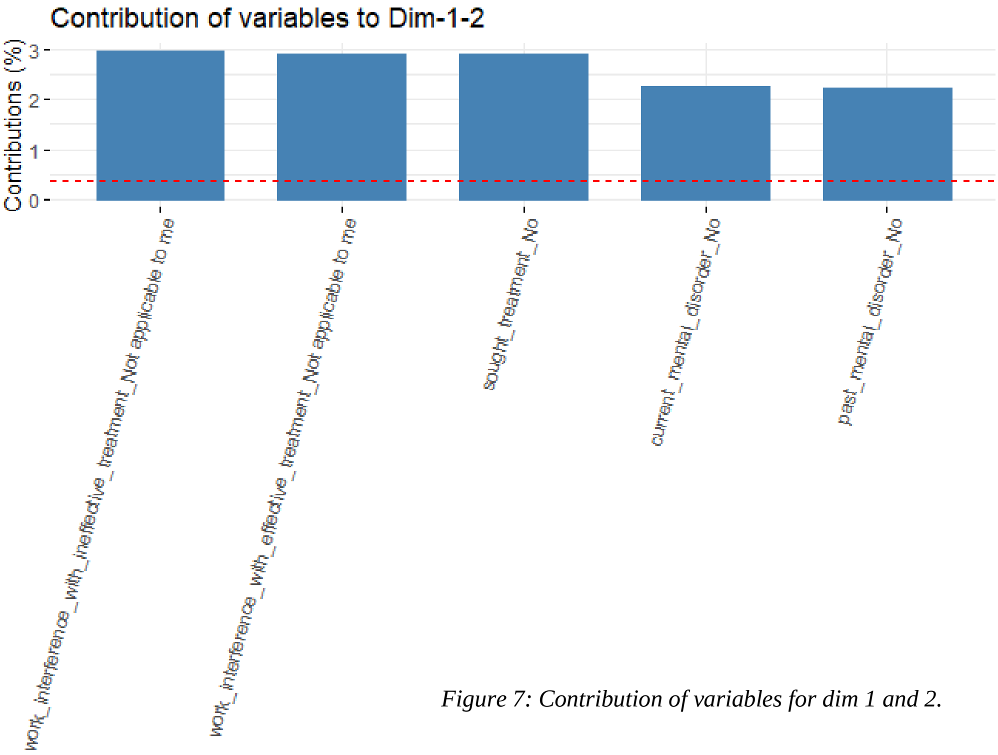
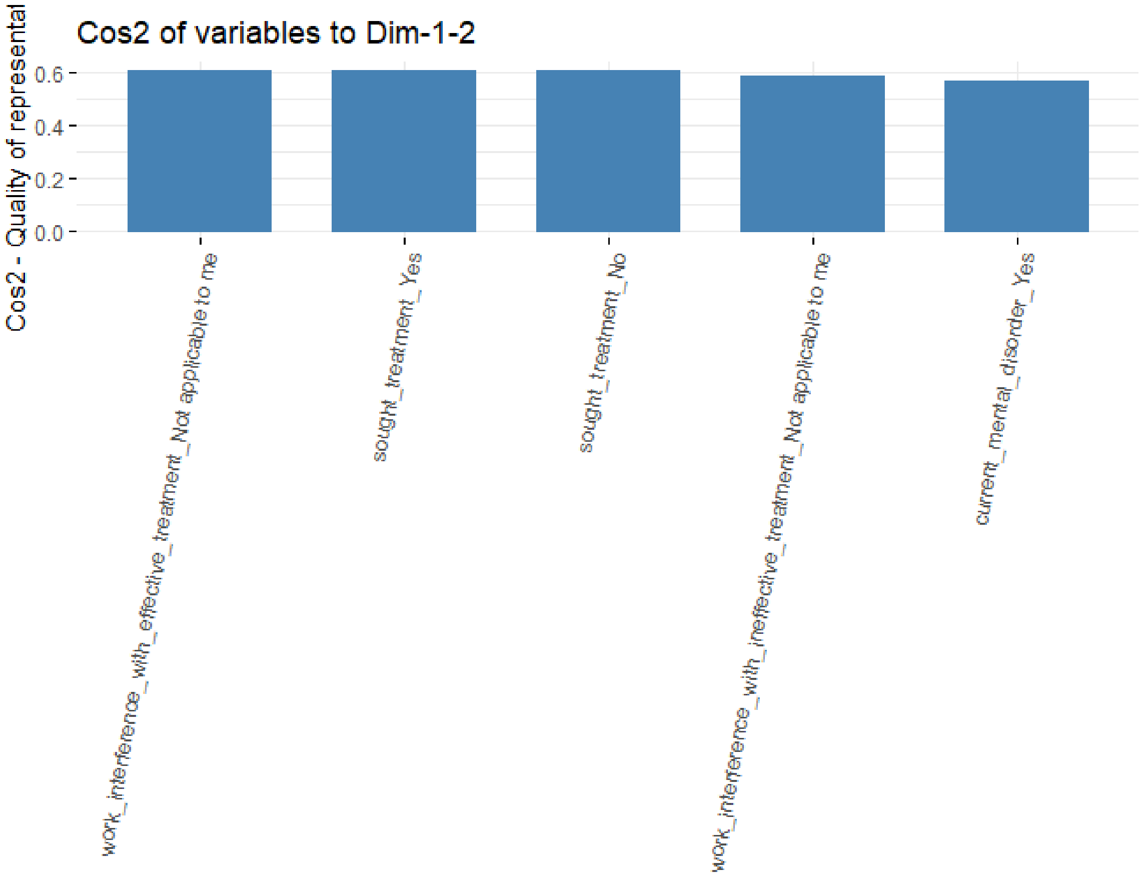


Figure 7: Contribution of variables for dim 1 and 2.

From our graph above, the primary contributions to our first two dimensions are as follows: those that answered “not applicable” to whether work interfered with ineffective or effective treatment, those that answered “No” to seeking treatment, those answering “No” to currently having a mental disorder, and final those answering “No” to having a past mental disorder. Looking at the Cos2 values,



5 Decision Tree and Random Forest Methods

Decision trees are based on the stratification of our feature space. For example, linear regression is a technique that divides our feature space into two parts using a line. This is an excellent tool if the data is truly linear. On the other hand, decision trees subset the data into multiple feature spaces. This method does not assume that our data is linear.

In "An Introduction to Statistical Learning" (ISLR), the process of feature stratification is summarized using 2 steps:

1. Divide the set of all possible predictor values X_1, X_2, … , X_p into J distinct overlapping regions, R_1, R_2, … , R_j.
2. If an observation falls within a region R_j, our prediction will be the mean value of the responses for the training observation in R_j.

The goal here is to produce a set of regions which minimizes the residual sum of squares (RSS), given by

$$\sum_{j=1}^{J}\sum_{i\in R_j}(y_i - \hat{y}_{R_j})^2 \tag{1}$$

This line of reasoning leads us to a "greedy" approach called "recursive binary splitting". This is "greedy" since we are maximizing the RSS at each step of the process. This can lead to overfitting on our training set. To correct this, as detailed in ISLR, growing a large tree then pruning it down to a subtree can reduce this overfitting. ISLR explains the following algorithm for building such regression trees:

1. Use recursive binary splitting to grow the tree on training data. Only stop the once a node has less than a chosen number of observations.
2. Apply cost complexity pruning to obtain a sequence of subtrees, as a function of α.
3. Divide the training set into K folds. For each k = 1, …, K:
    a. Repeat Steps 1 and 2 on all but the k^{th} fold.
    b. Evaluate the mean squared prediction error on the left-out k^{th} fold as a function of α.

    Average each value of α, and pick the α that minimizes the average error.

4. Return the subtree from Step 2 that corresponds to the chosen α.

As is mentioned in ISLR, classification trees work in a similar way. Instead of quantitative data and averages, we have qualitative data and frequencies. Instead of minimizing the RSS we minimize classification error rate. This error rate can be calculated as the fraction of observations in a region that do not belong to the most common class. As ISLR explains, this method is not good for tree-growing. Instead, there are two different methods for measuring classification error: Gini index and Gini entropy. Each equation is defined as follows, where $\hat{p}_{mk}$ is the proportion of training observations in the mth region from the kth class.

Gini index

$$G = \sum_{k=1}^{K} \hat{p}_{mk} \left(1 - \hat{p}_{mk}\right) \tag{2}$$

Gini entropy

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \, log \hat{p}_{mk} \tag{3}$$

Both of these equations assess node purity. That is, the closer we are to 1, the more our observations match the most common class.

Now that we have a solid idea of how to create decision trees, we should address the major limitations. The core advantage of decision trees is that they are easy to interpret and they can handle categorical variables easily. Unfortunately, decision trees tend to underperform when compared to other methods such as regression. In addition, as mentioned in ISLR, a small change in our dataset could result in a large overall change in our model. For our core analysis will be using random forests to overcome many of the limitations previously mentioned.

Random forest is a bagged decision tree technique that splits each node on a subset of variables. “Bagging” is an aggregated bootstrapping method. That is, different training sets are created by sampling, with replacement, from the dataset (i.e. “bootstrapping”) and averaging all predictions. Most of this heavy lifting will be carried out by the R package randomForest().

6 Results

Our initial decision tree was used to create an actionable plan for addressing mental health issues in the workplace. To this end, we've chosen to focus as a few predictors: whether they sought treatment and their awareness of mental health options.

We began with a simple decision tree to predict whether an employee will seek treatment. This variable is binary with 0 being "No" and 1 being "Yes". Our resulting regression tree shows us that those that have discussed mental health issues with coworkers, and are open with friends and family, are more likely to seek treatment.

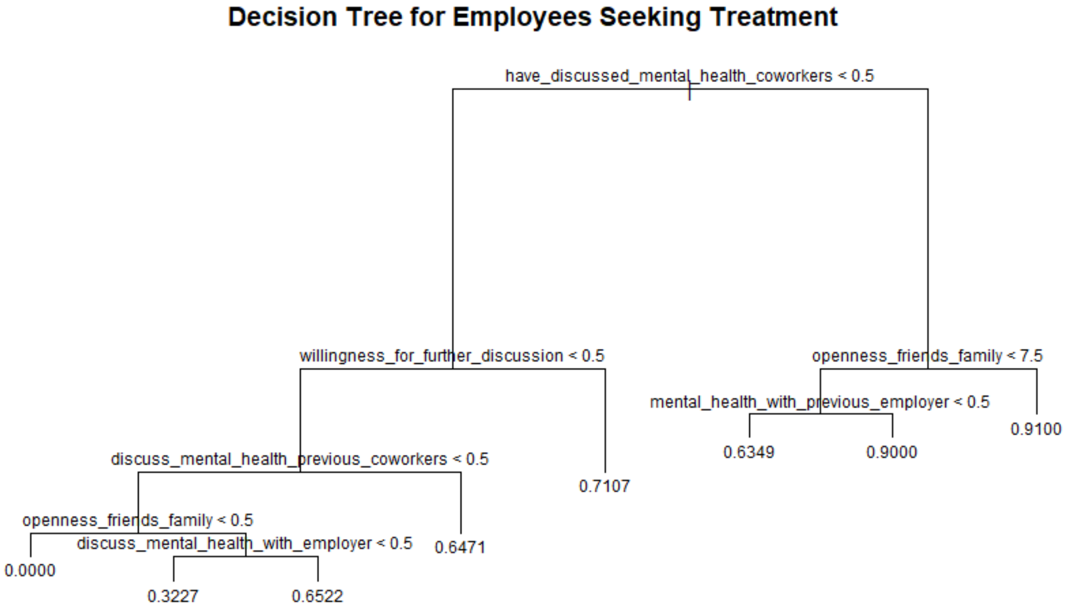

Decision Tree for Employees Seeking Treatment
have_discussed_mental_health_coworkers < 0.5
willingness_for_further_discussion < 0.5
openness_friends_family < 7.5
mental_health_with_previous_employer < 0.5
0.9100
0.6349
0.9000
discuss_mental_health_previous_coworkers < 0.5
0.7107
openness_friends_family < 0.5
discuss_mental_health_with_employer < 0.5
0.6471
0.0000
0.3227
0.6522

Figure 8: Naïve decision tree for employees seeking treatment

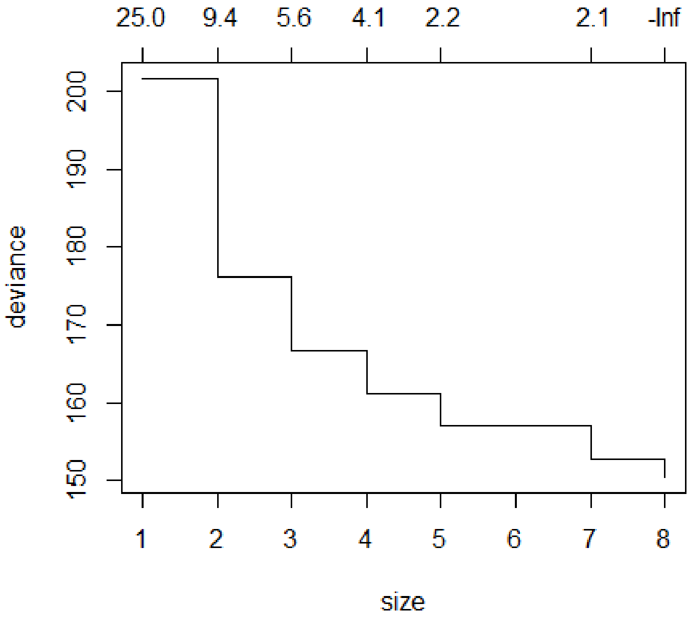


Figure 9: Size of our naïve model versus the sum of squared errors.

The trend here is clear. Those that keep silent about concerning mental health tend to not seek treatment. Although it is known that these models are far less reliable that, this model will motivate stronger random forest models.

As we previously mentioned, random forest models are known to be a more robust method for creating classification and regression trees. Our first model is concerned with the respondents seeking treatment. The initial random forest model produced an out-of-bag error rate of 11.48%. Figure 10 lists the variables by importance.

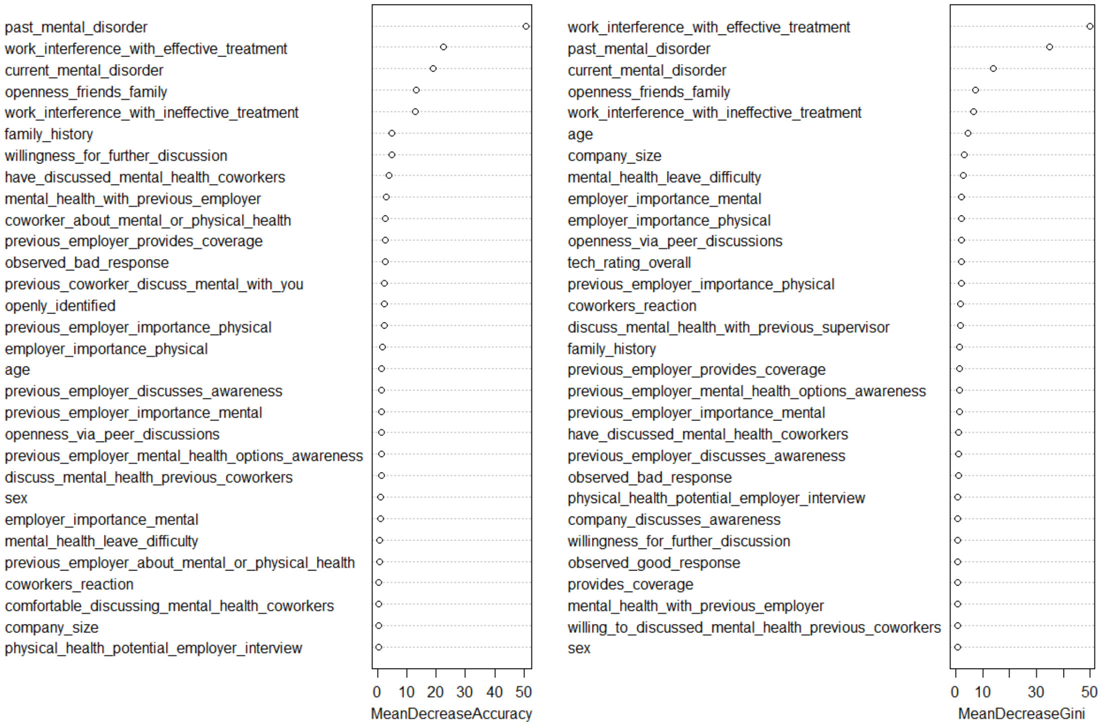


Figure 10: Variable importance for the random forest model predicting whether the respondent chose to seek treatment.

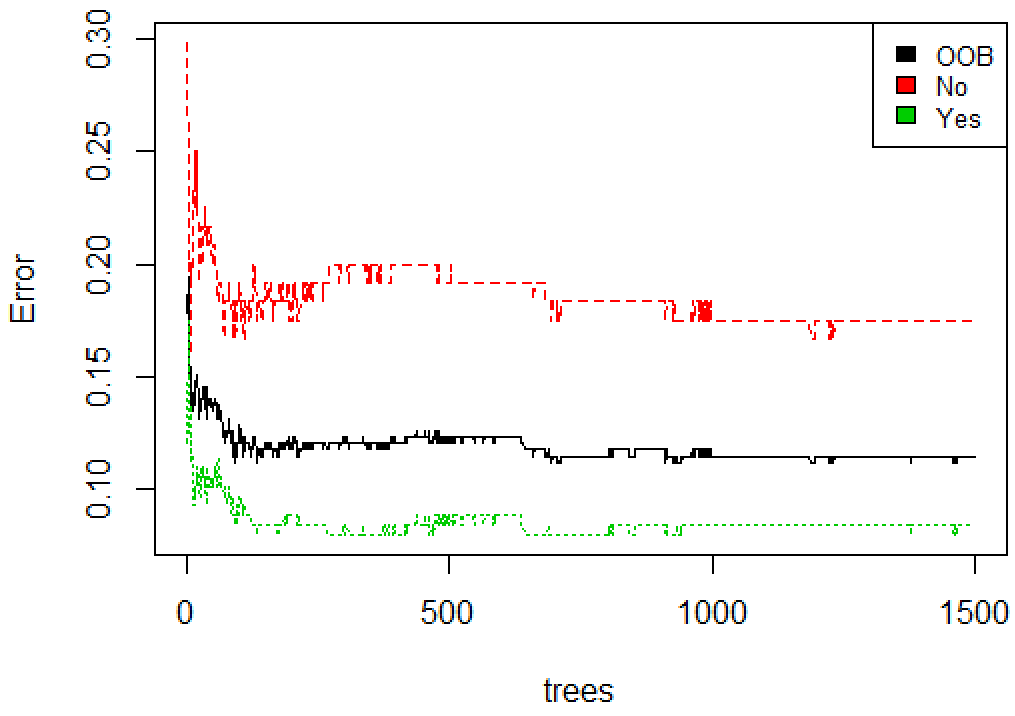


Figure 11: Error rates and tree sizes for the random forest model predicting whether the respondent chose to seek treatment.

The second variable of interest in the awareness of health insurance mental health options. The initial model we created had and error rate of 28.08%. After adjusting the number of trees to create and the number of variables that will be tried at each node split, we were only able to achieve a 26.89% error rate. We can see from figure 10 that the N/A values could not be correctly classified.

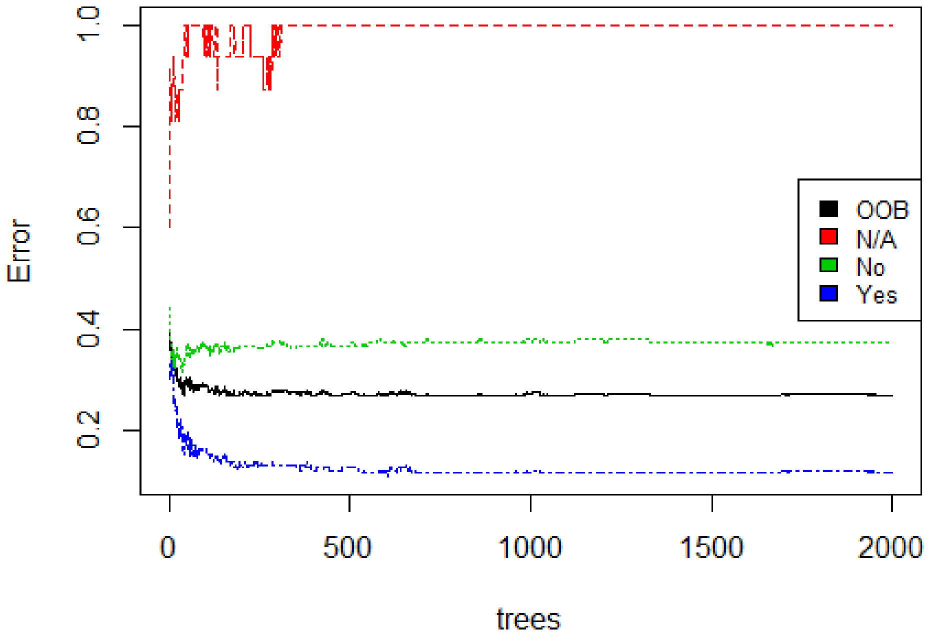


Figure 12: Error rates and tree sizes for the random forest model predicting the awareness of mental health options on employee insurance.

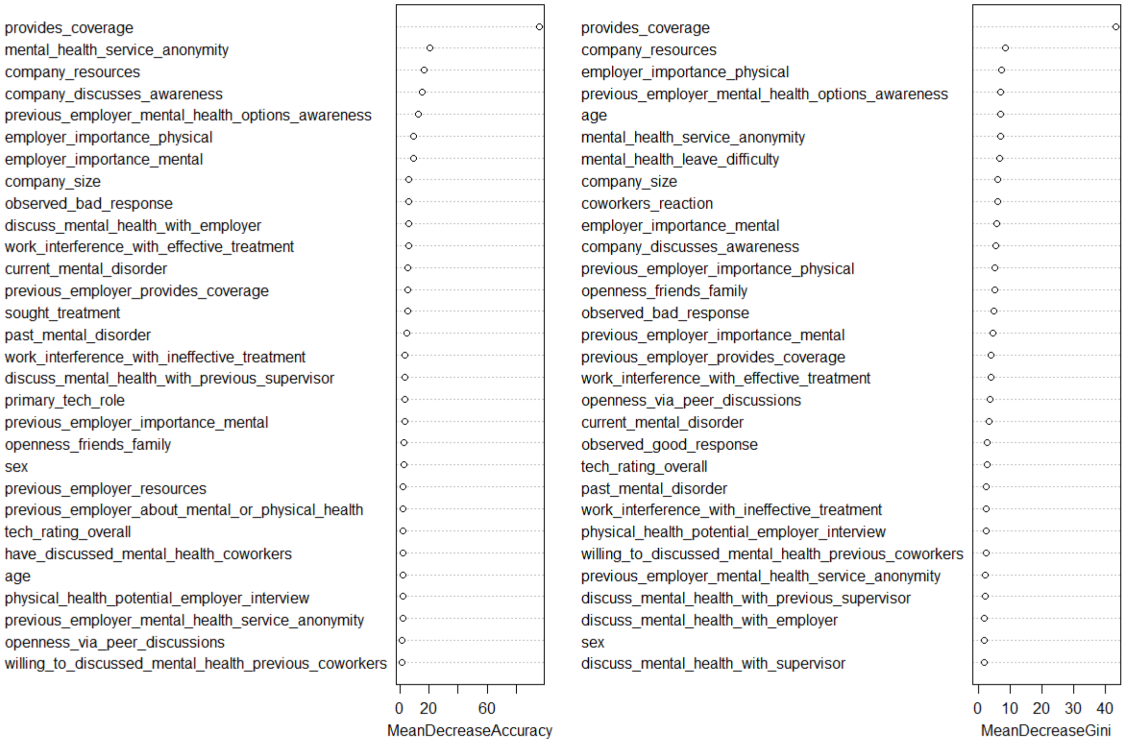


Figure 13: Variable importance for the random forest model predicting awareness of mental health option on employer's insurance.

7 Conclusion

In this paper we have explored the attitudes regarding mental health in the tech industry and developed actionable predictive models. Using random forest models, our results suggest that by creating an open environment where employee feel free to talk about mental health, we can increase the likelihood that an employee will seek treatment. As technology professions become more prevalent, increasing mental health awareness in the workplace will become more important.

8 Literature Cited

Davis, Bob. “Gender Equality: A Trend The Tech Sector Needs To Get Behind.” Forbes, Forbes Magazine, 27 June 2018, www.forbes.com/sites/forbestechcouncil/2018/06/27/gender-equality-a-trend-the-tech-sector-needs-to-get-behind/#4bc8c38d717b. Accessed June 20019.

Grolemund, Garrett, and Hadley Wickham. “R For Data Science.” 7 Exploratory Data Analysis, r4ds.had.co.nz/exploratory-data-analysis.html. Accessed June 20019.

James, Gareth, et al. An Introduction to Statistical Learning: with Applications in R (p. 303-331). Springer, 2017.

Kasbergen, Nara. “Supporting Mental Health in the Tech Workplace.” InfoQ, InfoQ, 4 May 2019, www.infoq.com/articles/mental-health-tech-workplace/. Accessed June 20019.

Marcelo C Anjos, Ding. “MCA - Multiple Correspondence Analysis in R: Essentials.” STHDA, 24 Sept. 2017, www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/114-mca-multiple-correspondence-analysis-in-r-essentials/. Accessed June 2019.

Mills, Michael. “Sex Difference vs. Gender Difference? Oh, I'm So Confused!” Psychology Today, Sussex Publishers, Oct 20. 2011, www.psychologytoday.com/us/blog/the-how-and-why-sex-differences/201110/sex-difference-vs-gender-difference-oh-im-so-confused. Accessed June 2019.

Osmi. “OSMI Mental Health in Tech Survey 2018.” Kaggle, 31 Dec. 2018, www.kaggle.com/osmihelp/osmi-mental-health-in-tech-survey-2018. Accessed June 2019.

Osmi. “OSMI Mental Health in Tech Survey 2017.” Kaggle, 23 May 2018, www.kaggle.com/osmihelp/osmi-mental-health-in-tech-survey-2017. Accessed June 2019.

Patel, Neil. “90% Of Startups Fail: Here's What You Need To Know About The 10%.” Forbes, Forbes Magazine, 2 Sept. 2015, www.forbes.com/sites/neilpatel/2015/01/16/90-of-startups-will-fail-heres-what-you-need-to-know-about-the-10/#181204d16679. Accessed June 2019.

Rubin, Mark, et al. “A Confirmatory Study of the Relations between Workplace Sexism, Sense of Belonging, Mental Health, and Job Satisfaction among Women in Male-dominated Industries.” Journal of Applied Social Psychology, Feb. 2019. EBSCOhost, doi:10.1111/jasp.12577. Accessed June 2019.

Simoes, Mariana. “Don't Work For A Tech Company If You Want A Stress-Free Job.” Business Insider, Business Insider, 11 Feb. 2013, www.businessinsider.com/tech-companies-stressful-places-to-work-2013-1. Accessed June 2019.

Snobar, Abdullah. “Getting Honest About Mental Health In The World Of Tech Startups.” Forbes, Forbes Magazine, 8 Aug. 2018, www.forbes.com/sites/forbestechcouncil/2018/08/08/getting-honest-about-mental-health-in-the-world-of-tech-startups/#1188c5af641a. Accessed June 2019.

“World Mental Health Day: the State of the Tech Industry.” *EM360*, 10 Oct. 2018, www.em360tech.com/tech-news/world-mental-health-day-tech-industry/. Accessed June 2019.

9 Appendix

9.1 Data Cleaning and Basic Exploratory Analysis

```
#Create a data frame that excludes comment fields and several fields that
will not be used in the analysis.
survey <- data.frame(Survey_2017_2018[,2:14], Survey_2017_2018[,16:17],
Survey_2017_2018[,19], Survey_2017_2018[,21:40], Survey_2017_2018[,42:43],
Survey_2017_2018[,45], Survey_2017_2018[,47:50], Survey_2017_2018[,90:97],
Survey_2017_2018[,99],Survey_2017_2018[,101:105], Survey_2017_2018[,107],
Survey_2017_2018[,109], Survey_2017_2018[,112:113],
Survey_2017_2018[,115:117], Survey_2017_2018[,119:120],
Survey_2017_2018[,124])


library(dplyr)

#Too many values for country to make a clear tree. Common values were used.
#The dataset here will be used for our tre models.
survey_subset1 <- subset(survey_subset, survey_subset$country=="United States
of America" | survey_subset$country=="Canada" | survey_subset$country=="Unite
d Kingdom")
survey_subset2 <- subset(survey_subset1, survey_subset1$work_country=="United
States of America" | survey_subset1$work_country=="Canada" | survey_subset1$w
ork_country=="United Kingdom")

cleaned_survey <- survey_subset2 %>% mutate_if(is.character, as.factor)
cleaned_survey[cleaned_survey$sought_treatment==0,]$sought_treatment <- "No"
cleaned_survey[cleaned_survey$sought_treatment==1,]$sought_treatment <- "Yes"

#Dataset below will be used for MCA
factor_survey <- cleaned_survey %>% mutate_if(is.numeric, as.factor)

#Basic barplots

country <- data.frame(sort(table(survey$work_country)))
nrow(country)
[1] 58
country3 <- country[56:58,]
par(mfrow = c(2,2))
barplot(table(survey$sex), main = "Sex" ,col = "light blue")
barplot(table(survey$age), main = "Age", col = "light blue")
barplot(country3$Freq, main = "Country" , names.arg = country3$Var1, col =
"light blue")



#Chi-Square analysis plots
library(corrplot)
t1 <- table(survey$sex, survey$mental_health_options_awareness)
chisq1 <- chisq.test(t1)
corrplot(chisq1$residuals, is.cor = FALSE)


table(survey$sex, survey$mental_health_options_awareness)

        N/A  No Yes
  F      25 130 151
  M      74 321 256
  Other   5  17  25
```

```
t1 <- table(survey$sex, survey$mental_health_options_awareness)
chisq.test(t1)

        Pearson's Chi-squared test

data:  t1
X-squared = 11.335, df = 4, p-value = 0.02305
chisq1 <- chisq.test(t1)
corrplot(chisq1$residuals, is.cor = FALSE)

t2 <- table(survey$sex, survey$openly_identified)
t2

          0   1
  F     291  53
  M     701  69
  Other  42  15

chisq.test(t2)

        Pearson's Chi-squared test

data:  t2
X-squared = 21.954, df = 2, p-value = 1.709e-05

chisq2 <- chisq.test(t2)
corrplot(chisq2$residuals, is.cor = FALSE)

t3 <- table(survey$sex, survey$sought_treatment)
chisq3 <- chisq.test(t3)
chisq3

        Pearson's Chi-squared test

data:  t3
X-squared = 41.878, df = 2, p-value = 8.059e-10

corrplot(chisq3$residuals, is.cor = FALSE)
```

9.2 Word Cloud Analysis

```
library("tm")
library("SnowballC")
library("wordcloud")
library("RColorBrewer")

#Initial subsetting of our data
F <- subset(Survey_2017_2018, Survey_2017_2018$sex == "F")
M <- subset(Survey_2017_2018, Survey_2017_2018$sex == "M")
Other <- subset(Survey_2017_2018, Survey_2017_2018$sex == "Other")

Q1_badly_handled_F <- as.matrix(F[,106])
Q1_badly_handled_M <- as.matrix(M[,106])
Q1_badly_handled_Other <- as.matrix(Other[,106])

Q2_improve_F <- as.matrix(F[,110])
Q2_improve_M <- as.matrix(M[,110])
```

```
Q2_improve_Other <- as.matrix(Other[,110])

toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))

###############Q1 Female response####################
Q1F_text <- Corpus(VectorSource(Q1_badly_handled_F))
Q1F_text <- tm_map(Q1F_text, toSpace, "/")
Q1F_text <- tm_map(Q1F_text, toSpace, "@")
Q1F_text <- tm_map(Q1F_text, toSpace, "\\|")

Q1F_text <- tm_map(Q1F_text, content_transformer(tolower))
Q1F_text <- tm_map(Q1F_text, removeNumbers)
Q1F_text <- tm_map(Q1F_text, removeWords, stopwords("english"))

Q1F_text <- tm_map(Q1F_text, removePunctuation)
Q1F_text <- tm_map(Q1F_text, stripWhitespace)

Q1F_dtm <- TermDocumentMatrix(Q1F_text)
Q1F_m <- as.matrix(Q1F_dtm)
Q1F_v <- sort(rowSums(Q1F_m),decreasing=TRUE)
Q1F_d <- data.frame(word = names(Q1F_v),freq=Q1F_v)

set.seed(1234)
wordcloud(words = Q1F_d$word, freq = Q1F_d$freq, min.freq = 3, max.words=200,
random.order=FALSE, rot.per=0.35, colors=brewer.pal(8, "Dark2"))

title("Circumstances of badly handled response (F)")

###############Q2 Female response####################
Q2F_text <- Corpus(VectorSource(Q2_improve_F))
Q2F_text <- tm_map(Q2F_text, toSpace, "/")
Q2F_text <- tm_map(Q2F_text, toSpace, "@")
Q2F_text <- tm_map(Q2F_text, toSpace, "\\|")

Q2F_text <- tm_map(Q2F_text, content_transformer(tolower))
Q2F_text <- tm_map(Q2F_text, removeNumbers)
Q2F_text <- tm_map(Q2F_text, removeWords, stopwords("english"))

Q2F_text <- tm_map(Q2F_text, removePunctuation)
Q2F_text <- tm_map(Q2F_text, stripWhitespace)

Q2F_dtm <- TermDocumentMatrix(Q2F_text)
Q2F_m <- as.matrix(Q2F_dtm)
Q2F_v <- sort(rowSums(Q2F_m),decreasing=TRUE)
Q2F_d <- data.frame(word = names(Q2F_v),freq=Q2F_v)

set.seed(1234)
wordcloud(words = Q2F_d$word, freq = Q2F_d$freq, min.freq = 3, max.words=200,
random.order=FALSE, rot.per=0.35, colors=brewer.pal(8, "Dark2"))

title("Suggestions to improve mental health conditions (F)")


###############Q1 Male response####################
Q1M_text <- Corpus(VectorSource(Q1_badly_handled_M))
Q1M_text <- tm_map(Q1M_text, toSpace, "/")
Q1M_text <- tm_map(Q1M_text, toSpace, "@")
Q1M_text <- tm_map(Q1M_text, toSpace, "\\|")
```

```
Q1M_text <- tm_map(Q1M_text, content_transformer(tolower))
Q1M_text <- tm_map(Q1M_text, removeNumbers)
Q1M_text <- tm_map(Q1M_text, removeWords, stopwords("english"))

Q1M_text <- tm_map(Q1M_text, removePunctuation)
Q1M_text <- tm_map(Q1M_text, stripWhitespace)

Q1M_dtm <- TermDocumentMatrix(Q1M_text)
Q1M_m <- as.matrix(Q1M_dtm)
Q1M_v <- sort(rowSums(Q1M_m),decreasing=TRUE)
Q1M_d <- data.frame(word = names(Q1M_v),freq=Q1M_v)

set.seed(1234)
wordcloud(words = Q1M_d$word, freq = Q1M_d$freq, min.freq = 4, max.words=200,
random.order=FALSE, rot.per=0.35, colors=brewer.pal(8, "Dark2"))
title("Circumstances of badly handled response (M)")

###############Q2 Male response###################
Q2M_text <- Corpus(VectorSource(Q2_improve_M))
Q2M_text <- tm_map(Q2M_text, toSpace, "/")
Q2M_text <- tm_map(Q2M_text, toSpace, "@")
Q2M_text <- tm_map(Q2M_text, toSpace, "\\|")

Q2M_text <- tm_map(Q2M_text, content_transMormer(tolower))
Q2M_text <- tm_map(Q2M_text, removeNumbers)
Q2M_text <- tm_map(Q2M_text, removeWords, stopwords("english"))

Q2M_text <- tm_map(Q2M_text, removePunctuation)
Q2M_text <- tm_map(Q2M_text, stripWhitespace)

Q2M_dtm <- TermDocumentMatrix(Q2M_text)
Q2M_m <- as.matrix(Q2M_dtm)
Q2M_v <- sort(rowSums(Q2M_m),decreasing=TRUE)
Q2M_d <- data.frame(word = names(Q2M_v),freq=Q2M_v)

set.seed(1234)
wordcloud(words = Q2M_d$word, freq = Q2M_d$freq, min.freq = 3, max.words=200,
random.order=FALSE, rot.per=0.35, colors=brewer.pal(8, "Dark2"))

title("Suggestions to improve mental health conditions (M)")

###############Q1 Other response###################
Q1Other_text <- Corpus(VectorSource(Q1_badly_handled_Other))
Q1Other_text <- tm_map(Q1Other_text, toSpace, "/")
Q1Other_text <- tm_map(Q1Other_text, toSpace, "@")
Q1Other_text <- tm_map(Q1Other_text, toSpace, "\\|")

Q1Other_text <- tm_map(Q1Other_text, content_transformer(tolower))
Q1Other_text <- tm_map(Q1Other_text, removeNumbers)
Q1Other_text <- tm_map(Q1Other_text, removeWords, stopwords("english"))

Q1Other_text <- tm_map(Q1Other_text, removePunctuation)
Q1Other_text <- tm_map(Q1Other_text, stripWhitespace)

Q1Other_dtm <- TermDocumentMatrix(Q1Other_text)
Q1Other_m <- as.matrix(Q1Other_dtm)
Q1Other_v <- sort(rowSums(Q1Other_m),decreasing=TRUE)
Q1Other_d <- data.frame(word = names(Q1Other_v),freq=Q1Other_v)
```

```
set.seed(1234)
wordcloud(words = Q1Other_d$word, freq = Q1Other_d$freq, min.freq = 1, max.wo
rds=200, random.order=FALSE, rot.per=0.35, colors=brewer.pal(8, "Dark2"))
title("Circumstances of badly handled response (Other)")

##############Q2 Other response###################
Q2Other_text <- Corpus(VectorSource(Q2_improve_Other))
Q2Other_text <- tm_map(Q2Other_text, toSpace, "/")
Q2Other_text <- tm_map(Q2Other_text, toSpace, "@")
Q2Other_text <- tm_map(Q2Other_text, toSpace, "\\|")

Q2Other_text <- tm_map(Q2Other_text, content_transformer(tolower))
Q2Other_text <- tm_map(Q2Other_text, removeNumbers)
Q2Other_text <- tm_map(Q2Other_text, removeWords, stopwords("english"))

Q2Other_text <- tm_map(Q2Other_text, removePunctuation)
Q2Other_text <- tm_map(Q2Other_text, stripWhitespace)

Q2Other_dtm <- TermDocumentMatrix(Q2Other_text)
Q2Other_m <- as.matrix(Q2Other_dtm)
Q2Other_v <- sort(rowSums(Q2Other_m),decreasing=TRUE)
Q2Other_d <- data.frame(word = names(Q2Other_v),freq=Q2Other_v)

set.seed(123)
wordcloud(words = Q2Other_d$word, freq = Q2Other_d$freq, min.freq = 1, max.wo
rds=100, random.order=FALSE, rot.per=0.35, colors=brewer.pal(8, "Dark2"))
title("Suggestions to improve mental health conditions (Other)")


#MCA
library("FactoMineR")
library("factoextra")
#NA values will cause a lot of problems with our MCA the following code was
#used to deal with how NAs are typically handled with MCA
library("missMDA")
complete <- imputeMCA(factor_survey, ncp = 5)
survey_mca <- MCA(factor_survey, tab.disj = complete$tab.disj)

print(survey_mca)
**Results of the Multiple Correspondence Analysis (MCA)**
The analysis was performed on 927 individuals, described by 53 variables
*The results are available in the following objects:

   name               description
1  "$eig"             "eigenvalues"
2  "$var"             "results for the variables"
3  "$var$coord"       "coord. of the categories"
4  "$var$cos2"        "cos2 for the categories"
5  "$var$contrib"     "contributions of the categories"
6  "$var$v.test"      "v-test for the categories"
7  "$ind"             "results for the individuals"
8  "$ind$coord"       "coord. for the individuals"
9  "$ind$cos2"        "cos2 for the individuals"
10 "$ind$contrib"     "contributions of the individuals"
11 "$call"            "intermediate results"
12 "$call$marge.col" "weights of columns"
13 "$call$marge.li"  "weights of rows"

> eigenvalues <- get_eigenvalue(survey_mca)
> head(round(eigenvalues, 2))
      eigenvalue variance.percent cumulative.variance.percent
Dim.1       0.16             4.01                        4.01
Dim.2       0.11             2.70                        6.72
Dim.3       0.09             2.25                        8.97
Dim.4       0.09             2.22                       11.19
Dim.5       0.07             1.72                       12.91
Dim.6       0.05             1.22                       14.13

fviz_screeplot(survey_mca, addlabels = TRUE, ylim = c(0, 10))
title("Scree Plot")
fviz_contrib(survey_mca, choice = "var", axes = 1:2, top = 5, xtickslab.rt =
75)
fviz_cos2(survey_mca, choice = "var", axes = 1:2, top = 5, xtickslab.rt = 80)
```

9.3 Decision Tree and Random Forest Analysis

```
#Remove columns with a large # of NAs
naCols <- vector(length=ncol(survey))
for (i in 1:ncol(survey)){ naCols[i]<- sum(is.na(survey[,i]))}
survey_subset <- survey[,which(naCols < 200)]

#Create the initial tree excluding country
library(tree)
survey.tree <- tree(survey_subset$sought_treatment~.-survey_subset$country, d
ata=survey_subset)

summary(survey.tree)

Regression tree:
tree(formula = survey_subset$sought_treatment ~ . - survey_subset$country, 
    data = survey_subset)
Variables actually used in tree construction:
[1] "have_discussed_mental_health_coworkers"   "willingness_for_further_discu
ssion"
[3] "discuss_mental_health_previous_coworkers" "openness_friends_family"
[5] "discuss_mental_health_with_employer"      "mental_health_with_previous_e
mployer"
Number of terminal nodes:  8 
Residual mean deviance:  0.1772 = 150.6 / 850 
Distribution of residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
-0.91000 -0.32270  0.09005  0.00000  0.34780  0.67730 

plot(survey.tree)
text(survey.tree, cex=.75, pretty =0)
title("Decision Tree for Employees Seeking Treatment")
plot(prune.tree(survey.tree))


library(ISLR)
library(randomForest)
cleaned_survey$sought_treatment <- as.factor(cleaned_survey$sought_treatment)

train <- sample(1: nrow(cleaned_survey), nrow(cleaned_survey)/2)

#Model to predict respondents' that sought treatment

set.seed(123)
```

```
rf_sought_treatment <- randomForest(sought_treatment~.,data=cleaned_survey, s
ubset=train, mtry=52, ntree=1500, importance=TRUE, na.action = na.exclude, pr
oximity=T)
rf_sought_treatment

Call:
 randomForest(formula = sought_treatment ~ ., data = cleaned_survey,      mtr
y = 52, ntree = 1500, importance = TRUE, proximity = T,      subset = train,
na.action = na.exclude)
               Type of random forest: classification
                     Number of trees: 1500
No. of variables tried at each split: 52

        OOB estimate of  error rate: 11.48%
Confusion matrix:
    No Yes class.error
No  99  21  0.17500000
Yes 20 217  0.08438819

plot(rf_sought_treatment)
legend("topright", colnames(rf_sought_treatment$err.rate),col=1:4,cex=0.8,fil
l=1:4)
varImpPlot(rf_sought_treatment)



#Model to predict respondents’ awareness of mental health options

rf_awareness <- randomForest(mental_health_options_awareness~.,data=cleaned_s
urvey, subset=train, mtry=52, importance=TRUE, na.action = na.exclude
rf_awareness

Call:
 randomForest(formula = mental_health_options_awareness ~ ., data = cleaned_s
urvey,      mtry = 52, importance = TRUE, subset = train, na.action = na.excl
ude)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 52

        OOB estimate of  error rate: 28.08%
Confusion matrix:
    N/A No Yes class.error
N/A   6 14   1   0.7142857
No    6 89  55   0.4066667
Yes   1 21 156   0.1235955


rf_awareness <- randomForest(mental_health_options_awareness~.,data=cleaned_s
urvey, subset=train, mtry = 17, ntree=2000, importance=TRUE, na.action = na.e
xclude, proximity=T)
rf_awareness

Call:
 randomForest(formula = mental_health_options_awareness ~ ., data = cleaned_s
urvey,      mtry = 17, ntree = 2000, importance = TRUE, proximity = T,      s
ubset = train, na.action = na.exclude)
               Type of random forest: classification
                     Number of trees: 2000
No. of variables tried at each split: 17

        OOB estimate of  error rate: 26.89%
Confusion matrix:
    N/A No Yes class.error
```

```
N/A    0 12   4   1.0000000
No     1 99  58   0.3734177
Yes    0 21 162   0.1147541
```

```
plot(rf_awareness)
legend("right", colnames(rf_awareness$err.rate),col=1:4,cex=0.8,fill=1:4)
varImpPlot(rf_awareness)
```

9.4 Dataset

Survey
2017-2018.xlsx