Mental Health in the Workplace – Analysis of the OSMI 2014 Data

Springboard Capstone Project
Elizabeth Matthews
April, 2017

Problem Statement & Project Aims

According to the National Institute of Mental Health, almost 18% of all adult Americans were diagnosed with mental illness in 2015 (NIH, n.d). Mental health disorders can present significant challenges to the workplace and often result in many direct and indirect costs to both individuals and their employers. Poor mental health among a company's employees can negatively impact the company's productivity in a number of ways, including: lost employee productivity, poor performance in the workplace, high rates of worker absenteeism or employee turnover, accidents on the job, as well as increasing costs related to insurance premiums and medications (Harvard Health, 2010).

However, recent research has found that when companies actively promote mental health wellness in the workplace, that there are significant financial benefits as well as improved employee moral (Harvard Health, 2010). As part of the efforts to better understand how to address work place mental health concerns, employers and their management can look to data science and the development of statistical models that can accurately predict the risk of mental health issues among employees. Developing these models can enable employers to better understand the factors that impact employees' mental health and then to work proactively to promote mental health well-being in their companies. These strategies could improve the workplace environment and help reduce costs associated with mental health illness.

The goals of this project were to:

- explore employee and workplace characteristics and their impact on treatment seeking and
- develop an accurate model to predict treatment seeking behavior

Data Set & Variables

Data for this project was obtained from the Open Source Mental Illness Project (OSMI). The data was collected via an online survey. The 2014 Mental Health in Tech Survey includes 1259 observations and 27 variables that include: demographic information of the respondents, workplace information and attitudes regarding mental health care and consequences of seeking mental health care. In addition to these variables, there is also an additional comments section that includes qualitative information on respondents' experiences with mental health issues in the workplace.

The original data set can be obtained on Kaggle @
https://www.kaggle.com/osmi/mental-health-in-tech-survey

Limitations

The OSMI contains a great deal of valuable information, however there are some limitations that affect the ability to build a predictive model to determine the likelihood of someone developing a mental health disorder. Firstly, there is no single variable assessing whether the respondent currently is experiencing a mental

health disorder, or the classification of the disorder. The two closest variables include the "seek help" and "treatment" variables. Both attempt to assess whether the individual has sought help or treatment for a mental health condition. The phrasing of the question posed to participants is such that neither variable may capture those individuals who have mental health concerns, but are not willing, or able to obtain treatment. Neither variable provides a time frame for the individual's treatment and whether treatment was sought recently or during the respondent's tenure with the current company.

Secondly, the demographic information is somewhat limited. Demographic variables in this data set are limited to age, gender, workplace type (tech/not tech), whether there is a family history of mental health issues, the location of the respondent and the size of the respondent's employer. It would be helpful to have other demographic information such as marital status, child(ren), education, and income.

Data Analysis Approach

Several approaches were utilized in analyzing the data and building a predictive model. First, the data set was manipulated and cleaned to allow for effective exploration and analysis. Then, descriptive and inferential statistics were used to better understand the variables, the demographic aspects of the respondents and their workplace settings as well as some of the relationships between various variables and treatment seeking behavior. This exploratory analysis was used to determine the selection of variables to be included in the regression analyses and model building.

Since the outcome variable's (i.e : treatment) levels were binary (yes/no), logistic regression was used to build the predictive model. The data was divided into a training and testing set. Once the model was developed with the training data, it was subsequently tested on the test data portion (or "unseen" data) to determine it's accuracy.

In addition to Logistic Regression, Recursive Partitioning was done to build a classification and regression tree to better understand the impact of the independent variables on treatment seeking. Similar to the Logistic Regression, both training and test data sets were used to evaluate the accuracy of the model.

As mentioned previously, a small number of qualitative comments were available in the data. Although the sample of comments was very small (only 163 comments), sentiment analysis was done in the final step since qualitative data can provide some rich context for understanding behavioral drives for treatment seeking.

The cleaned data set and full code for the data manipulation and analyses can be found at: https://github.com/belljar26/MentalHealthCapstone

Data Wrangling

Although I did not use all 27 variables in my analyses, I chose to clean the entire data set. Most of the data manipulation involved fixing spacing issues in the variable levels, correcting mistakes in the gender column, collapsing multiple levels of gender identification into one "other" level, addressing missing values and creating a new column for state's regions. In addition, a new variable was created with a binary 1, 0 levels to indicate whether or not the respondent made a comment in the survey.

For the sentiment analysis, the data was subset into a second data frame including only the treatment variable and comments variables. All rows where no comments were left, were removed.
The full documentation of the code used in the data manipulation can be found in a R Markdown document here:
https://github.com/belljar26/MentalHealthCapstone/blob/master/DataWranglingCapstone.Rmd

Exploratory Data Analyses

Descriptive and inferential statistics were utilized to better understand the data and to provide useful information necessary for developing an accurate predictive model.

Descriptive Analyses:

The mean age of the respondent is 32 years of age (SD = 7.23). Men represent 78% of respondents, 19% were female and 3 % identified as "other". Remote workers were not well represented in this data set (N=376).

The initial analysis found that 50% of respondents had sought treatment for a mental health disorder. Interestingly, 62% stated that they either don't know (if they do) or don't have access to mental health benefits their place of work.

Inferential Statistics

Chi Square Tests

As the majority of the variables in this data set are categorical, Chi Square tests were done on several different variables to better understand the behavioral health views and treatment seeking behavior of the respondents.

Neither working in the technology sector or working remotely greater than 50% of the time were associated with treatment seeking [(p=.29), (p=.39)].

In the initial exploratory analyses, some statistically significant relationships were found. As expected, family history of a mental health disorder does appear to be associated with seeking treatment (p<.01). Gender also was found to be associated with treatment seeking (p<.01) with women being more likely to seek treatment than men. Interestingly, perceptions and observations of consequences for disclosing mental health disorders appear to be associated with treatment seeking (p <.01).

Company size may have an impact on employees' willingness to disclose mental health disorders (p=.0018). There also seems to be a statistically significant relationship (p= .02) between country [where the respondent is located] and beliefs among respondents about consequences for disclosing mental health disorders. Understanding the relationships between company culture and respondents' fear of disclosure of mental health issues may provide some additional understanding of the factors that influence treatment seeking.

Qualitative Comments:

This data set also includes some qualitative comments made by respondents. 163 respondents provided comments about their experiences with mental health issues in the workplace. Leaving a comment was associated with seeking treatment (p< .0002).

Building a Predictive Model

Following the exploratory data analysis, promising independent variables were identified for use in building the logistic regression model. As noted previously, logistic regression was used because the outcome variable (treatment) is binary.

Logistic Regression

The logistic regression model was initially fit using the following independent variables:

- *Age*
- *Gender*
- *Family History*
- *Work Interference*
- *Region and*
- *Care Options*

The following output shows that Gender, Family History, Work Interference and Care Options were significant in predicting treatment seeking. Neither Region or Mental Health Consequence was found to be significant in this model.

Coefficients:

	Estimate	**Std.**	**Error**	**z value**	**Pr(>|z|)**
(Intercept)	-3.42149	0	.58378	-5.861	4.60e-09 ***
GenderMale	-1.01275	0	.27928	-3.626	0.000288 ***
GenderOther	-0.14486	0	.72537	-0.200	0.841711
Age	0.03803	0	.01391	2.734	0.006259 **
care_optionsUnsure	-0.31132	0	.26636	-1.169	0.242488
care_optionsYes	0.67307	0	.24122	2.790	0.005266 **
work_interfereOften	4.50708	0	.42010	10.729	< 2e-16 ***
work_interfereRarely	3.18151	0	.32761	9.711	< 2e-16 ***
work_interfSome	3.46857	0	.28282	12.264	< 2e-16 ***
regionMW	0.29859	0	.33057	0.903	0.366386
regionNE	-0.23320	0	.37514	-0.622	0.534192
regionS	-0.25133	0	.29524	-0.851	0.394606
regionW	0.55800	0	.30705	1.817	0.069173 .
ment_health_consNo	-0.03702	0.	23878	-0.155	0.876796
ment_health_consYe	-0.09316	0	.26282	-0.354	0.722989
family_historyYes	0.97772	0	.21232	4.605	4.13e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1133.87 on 817 degrees of freedom
Residual deviance: 618.34 on 802 degrees of freedom

AIC: 650.34

Number of Fisher Scoring iterations: 5

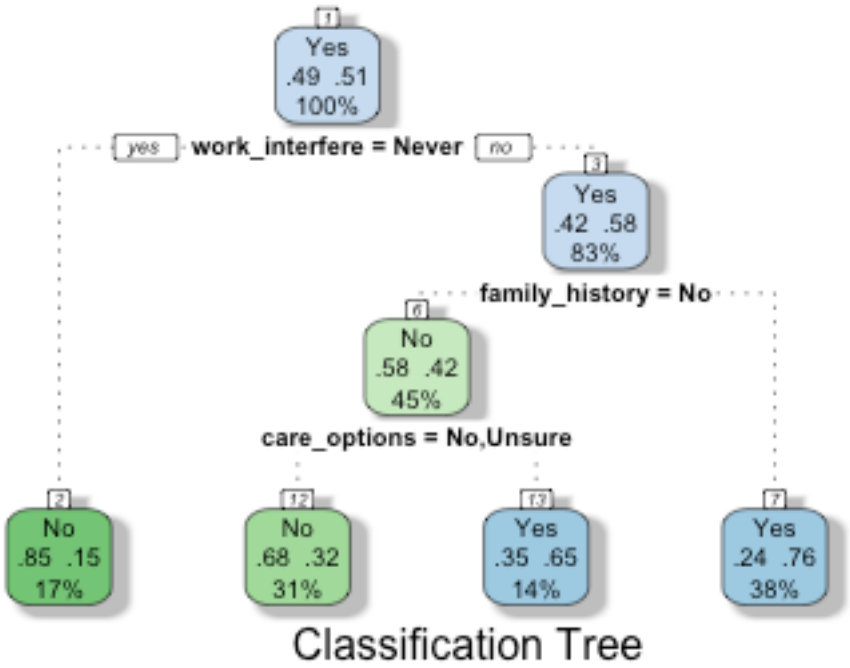
The following confusion matrix was created to determine the accuracy of the model on new data:

	False	**True**
No (Treatment)	165 (True -)	53 (False +)
Yes (Treatment)	24 (False -)	199 (True +)

This model correctly predicts treatment seeking 82.53% of the time and has a sensitivity of 89%. This model is significantly more accurate than the baseline of 50-50 (representing random guessing) or the baseline for treatment seeking represented in the data set (50% of respondents have sought treatment).

Recursive Partitioning

Recursive partitioning was used to produce a classification tree that provides a highly interpretable and hierarchical model for understanding the impact of the independent variables. Recursive partitioning is an unsupervised machine learning method, so all independent variables are entered into the model. The classification tree displayed below shows that beliefs about work interference are highly predictive for treatment seeking as well as family history and knowledge of care options.



Classification Tree

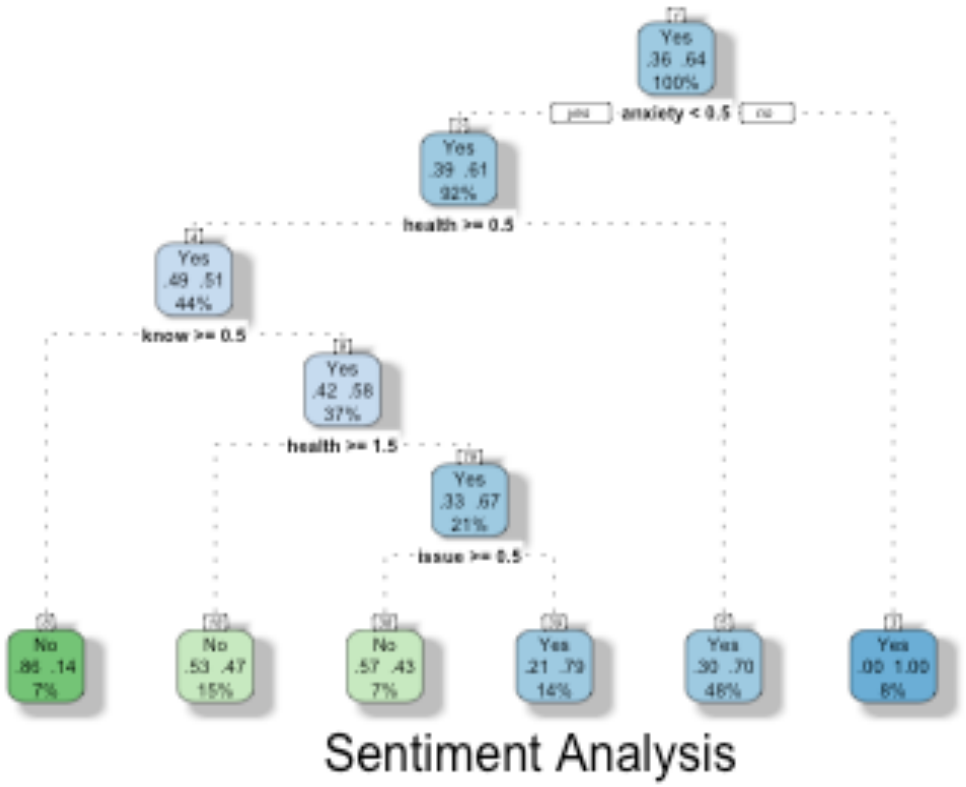
The confusion matrix demonstrates that there is a high degree of accuracy with this model with 75.6% (accuracy) predicting treatment seeking.

	False	True
No (Treatment)	146	41
Yes (Treatment)	51	140

Sentiment Analysis

Qualitative data can provide a rich context for understanding the underlying themes represented by quantitative analyses. In the OSMI, a small number of participants (n= 163) left comments. Despite the fact that this is an extremely small number of data, sentiment analysis was conducted to see if any additional insights could be gleaned.

The tm and Snowball C packages were used to pre process the data for analysis. Then, recursive partitioning was used to create a classification and regression tree to explore those words which are predictive of treatment seeking.



Based on this classification tree, the presence the words “anxiety and health” in the comments section are predictive of treatment seeking. While interesting and

fairly unsurprising, the overall predictive accuracy of this model is fairly low, at 63%. This may be largely due to the small sample size of comments.

Recommendations for the Workplace

The findings from this project provide some important insights into mental health and treatment seeking behaviors in the workplace. Key variables that appear to be associated with treatment seeking include gender, age, knowledge of care /benefit options and the employee's beliefs about the impact of mental health disorders on work performance. It also appears that the majority of respondents were unsure or did not know about the benefits that were available to them.

Importantly, it appears that respondents may sense a continuing stigma around mental health disorders. In fact, 61 % of respondents reported that there may be, or are consequences for disclosing a mental health disorder, although only a small minority of respondents (.09%) have reported actually witnessing a consequence for reporting a mental health disorder.

Accordingly, employers can leverage these findings to create a culture within their organizations where employees perceive support for their mental health well being and where concrete education about the impact of mental health and the available resources are provided.

Specifically, recommendations, based on the findings from this study include:

- **Provide Education on the Impact of Untreated Mental Health Disorders in the Work Place**. The findings from this study suggest that employees who believe that mental health disorders impact their work performance are more likely to seek treatment. This is an important finding, because employers can develop educational resources for employees to help them understand the implications of untreated mental health disorders on their wellbeing and on their work.

- **Ensure Employees are Aware of Mental Health Care Options/Benefits**. This analysis found that over 60% of employees do not know or unsure of the benefits and care options available to them. Employees' knowledge of care options appears to influence treatment seeking behavior. Accordingly, it is important that employers make this information available and ensure that employees understand the benefits and resources available to them

References

Harvard Health Publications(2010). Mental health problems in the workplace. Retrieved from: http://www.health.harvard.edu/newsletter_article/mental-health-problems-in-the-workplace

NIMH (n.d). Mental illness among adults. Retrieved from: https://www.nimh.nih.gov/health/statistics/prevalence/any-mental-illness-ami-among-us-adults.shtml