Exploring the Dynamic Between Mental Health and the Tech Industry

Brittany Stenekes
Cornell University
bss99@cornell.edu

William Xiao
Cornell University
wmx2@cornell.edu

Sami Smalling
Cornell University
ss2676@cornell.edu

Abstract

Mental health is a serious issue which is garnering more attention given the movement to virtual work and schooling due to the pandemic. In particular, the tech industry is known for heavily stigmatizing the subject. In this report, we discuss our plan to avoid over (and under-)fitting and to test the effectiveness of the models we develop. We also discuss the quality of our data, and display some results from preliminary results with their associated transformations. Finally, we explain what remains to be done, and how we plan to develop the project over the rest of the semester.

1 Descriptive Statistics

Data from surveys filled out in 2017-2019 were combined into a full dataset with 1,525 examples. The most pertinent questions were selected to go into the cleaned dataset, which now has 29 features. These questions were about the respondent's demographics, their mental health status, their comfortability discussing mental health, and how their workplace considers mental health concerns. Questions that were removed included repeated questions or ones relating to a previous employer. There were a few questions that were emphasized in bold in the survey, and these questions were in general the ones included in our cleaned dataset. We extracted basic statistics from the survey responses regarding employment type and environments from the combined survey results from 2017 to 2019. We first examined the age of survey participants, generating the following histograms. Since the distributions are very similar, we cannot conclude that age is a relevant metric to use in model fitting.

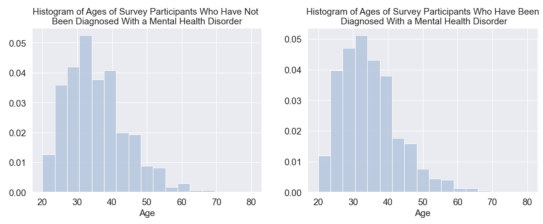

Figure 1: Age vs. Diagnosis

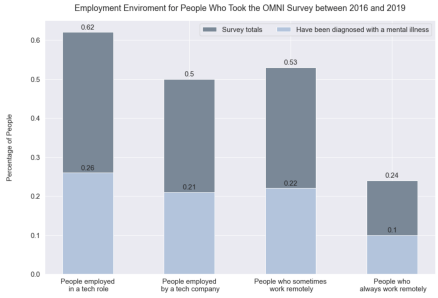

Figure 2: Respondent Employment Environment.

62% of the participants are employed in a tech role, with half of all participants working for a tech company. About half of the tech employees– illustrated by the light blue bars in Figure 2 have been diagnosed with a mental illness. Half of the participants sometimes work remotely, while 25% always work remotely. Whether or not remote work contributes to mental health illness in the workplace is a factor we will consider in our further model fitting and analysis because remote work has become so prevalent among companies in 2020, due to COVID-19.

Next, we separated the participants into two mutually exhaustive and exclusive groups: people who have been diagnosed with a mental illness and people who have not. As Figure 3 shows, people who have not been diagnosed with a mental illness are less comfortable talking about mental health issues compared to people who have been diagnosed with a disorder.

When analyzing data quality, we found that there are corrupted NaN values throughout the table, including in basic fields such as age, race, and gender. There are several optional survey questions that some did not choose to answer, including "Would you bring up mental health issues in an interview? Why or why not?", so the entries for those rows are missing. The most critical columns have <1% of the data missing. The average is around 14-50% missing, and the most is around 80%.

There were also plenty of messy values in the gender column, including misspellings that were reminiscent of uncleaned text-to-speech ('uhhhhhh genderqueer'),

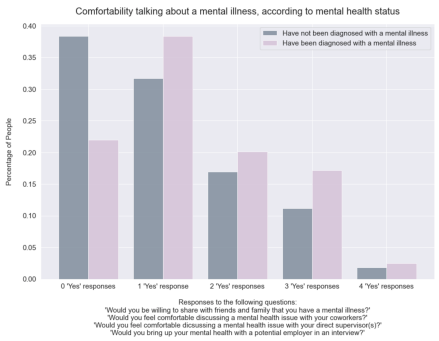

Figure 3: Comfortability of discussing mental health issues by diagnosis.

Gender	Coworkers	Supervisor(s)
Female	19.8%	31.8%
Male	20.8%	30.4%
Other	28.0%	40.2%

Table 1: Percentage of tech employees who feel comfortable talking to other groups about mental health, organized by gender.

making for tens of different options that had to be cleaned down to one of 'M', 'F', and 'Nonbinary'. There were few enough distinct values that this could be cleaned manually. The rest of the columns did not need to be cleaned too heavily.

2 Preliminary Analysis

We extracted demographic information about the survey participants and found that, of the respondents who work for a tech company, or in a tech role, about 28% of the respondents were female, 64% male, and the remaining 8% identified as another gender. 53% of the female tech employees have been diagnosed with a mental health disorder, while 38% of the men, and 7% of the remaining tech employees have been diagnosed with a mental health disorder. These basic statistics motivate us to explore the possibility that females in tech are more prone to mental health disorders than males. Although, it could also be the case that females are more comfortable seeking professional mental health diagnoses than males.

We also examined workplace comfortability according to gender.

These results are synthesized in table 1.

All percentages in the table above are 40% or less, suggesting most people who work in the tech industry do not feel comfortable talking about mental health. Both males and females feel especially uncomfortable talking to their coworkers about their mental health concerns.

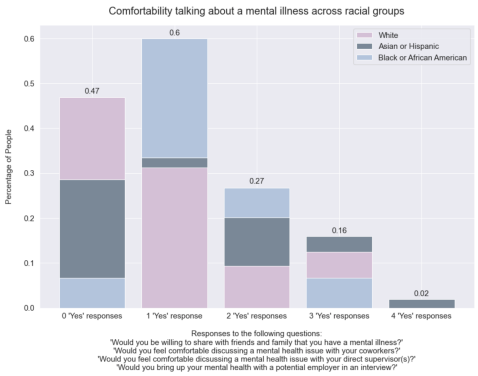

Figure 4: Comfortability of discussing mental health issues by race.

We then performed a similar analysis of comfortability discussing mental health for participants who identified as White or Caucasian, Asian, Hispanic, Black, or African American. 60% of people who identified as Black or African American, answered "Yes" to only 1 question asking about their comfort in discussing mental illness. About half of the people who identified as White answered "No" to all of the questions. There were mixed responses among people who identified as either Asian or Hispanic. About 3% of these people answered "No" to all questions, while 15% answered "Yes" at least 75% of the time. Based on these results illustrated in the plot below, it is clear that very few people are comfortable discussing mental health in their workplace.

In summary, our initial data explorations generated the following ideas which will be useful in moving forward:

- Mental illness seems common in the tech industry: About 1 in 2 tech employees who participated have been diagnosed with a mental health illness.
- Most people are uncomfortable discussing mental health issues during an interview or with their coworkers or supervisors.
- There is not a noticeable difference between the comfortability of males versus females. However, of the participants in the tech industry, 15% more females have been diagnosed with a mental health issue when compared to males.
- There are about the same proportions of participants with mental illnesses across age groups. Age does not appear to be an important factor to consider in model fitting.
- Few people who identified as White, Black, or African American are comfortable discussing mental health issues in any circumstances. Comparatively, people who identified as Asian or

Hispanic had more mixed opinions. Therefore, we think ethnicity could potentially be useful in model fitting, although we understand that survey different respondents have different past experiences that are not necessarily captured by survey results.

3 Initial Model

The feature transformations used mirrored those we used in the AirBnB homework. For real-valued fields (e.g. age), we kept the real values. For named values (e.g. gender, country), we decided to use one-hot encodings and many-hot encodings for set values. We have not yet decided how to handle ordinal data. We may use one-hot encodings as well. For text data, we may use one of the text embedding tools mentioned in class to extract the word embedding vectors for the text responses.

(Footnote): We realized while doing basic analysis that 863 out of 1,525 responses to the question "have you ever been diagnosed with a mental health disorder?" were NaN. Because questions following this question indicate that checkboxes were used for respondents to indicate whether or not they were diagnosed with a number of disorders, we assume that these NaN values correspond to respondents who have not been diagnosed with a mental health disorder, and categorize them as "no" values.

4 Next Steps

Besides linear regression, we also plan to try to minimize the error with other metrics, including logistic and hinge loss to reduce the effect of outliers. We will quantify the error in our model fits using these loss functions. We will choose to move forward with the loss function that results in a small and similar testing and training error.

Also, since there are questions in our dataset that are fairly similar, such as "Do you feel comfortable talking to a coworker about mental health?" and "Do you feel comfortable talking to a supervisor about mental health?", we plan to try to use a few different regularizers. As a result, we hope to make our model sparse, using fewer features and making the results easy to interpret. In order to help prevent over and underfitting, we will make sure to use features which we believe to have the highest amount of correlation with the target response, including gender, race, workplace factors, and support from friends and family. This way, we can cut down on the number of features we have while still making sure that we still preserve the most important information.

We will also attempt to prevent under and overfitting by using various train/test splitting techniques, including cross-validation by evaluating model error for multiple feature combinations and also bootstrapping. For the cross-validation procedure, we will begin by using the $\sim$10 features that we found to be relevant from our preliminary analysis. If our initial model training error is lower than our testing error, we will consider adding features about participants' previous experiences with mental illness to avoid underfitting. Alternatively, if our initial model training error is higher than our testing error, we will remove features from our model.