Investigating 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 what kinds of factors make one more likely to need or seek treatment for mental health. We show our methods for missing value imputation, as well as the models we used to try and fit the data. We also reflect on whether these models did well, our confidence in using them in a production setting, and also whether the results of this project might be used as a weapon of math destruction.

1 Problem Description

Mental health has become an ever-present issue in modern society. With Americans working harder than ever, with constant stressors looming over, it is easy for those with such burdens to become overwhelmed, and not know how to proceed. As such, the mental health of those suffering may often tend to decrease. Mental health has become even more important these past few years, with mental health organizations and efforts on the rise. Given the pandemic, people being isolated and working from home only compounds this issue of potentially deteriorating mental state. In this project, we attempt to discover what kinds of factors contribute to one's involvement with mental health, and whether one seeks treatment for issues or not.

2 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.

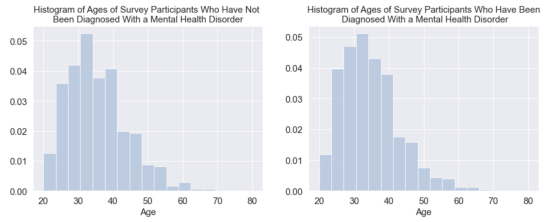

Figure 1: Age vs. Diagnosis

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

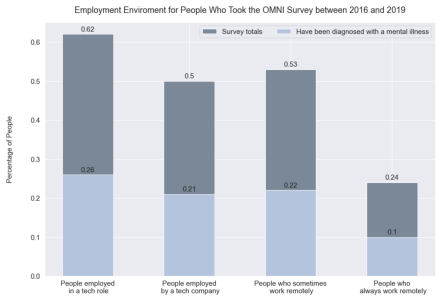

Figure 2: Respondent Employment Environment.

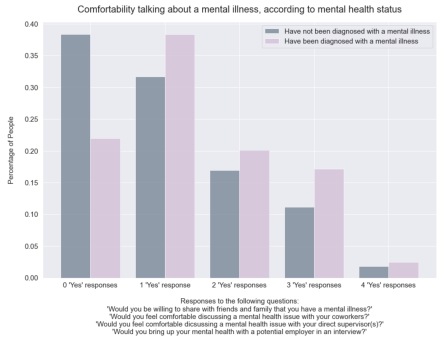

Figure 3: Comfortability of discussing mental health issues by diagnosis.

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'), 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.

3 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

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.

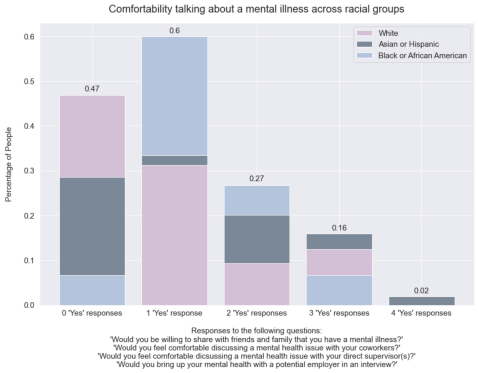

Figure 4: Comfortability of discussing mental health issues by race.

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.

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.

4 Techniques Used

4.1 Missing Value Imputation

Because this project involves working with survey data that had roughly 20% missing values, imputation was essential before moving forward with constructing models. Two different imputed datasets were created using different imputation techniques which were compared while evaluating model performance. The features were broken down into five distinct groups: real-valued features, boolean features, categorical features, ordinal features, and those with values that were not imputed (such as long-form text data, the country and state where the participant lives and works, and their gender and race). The unimputed features were chosen because there are categories within those features that could potentially not be within the observed data. For example, there could be participants from countries not listed within the countries observed in the dataset.

In both datasets, boolean features were cleaned by converting to binaries and ordinal features were encoded as factors, with NaN values intact.

In all three datasets, ordinal, real-valued, and boolean data were all imputed using KNN nearest neighbors imputation, using sklearn's `KNNImpute` [1]. The KNN nearest neighbors algorithm imputes missing values by averaging values from the n nearest neighbors. Two samples are designated to be close if the values of other features from that row that are not missing are close, measured with Euclidean distance. In all three datasets, the number of nearest neighbors, n, was chosen to be 3. It is recommended that the n chosen is an odd number to avoid ties in choosing the nearest neighbor.

After the imputation was performed, values were rounded so that encoded values could be converted back to categoricals and ordinals. For ordinal values, this was equivalent to rounding to the closest ordinal value. For the one-hot encoded categorical data, this was equivalent to choosing the largest value within the set of categories for each feature. If there were ties (two categories had values of 0.5), then all values for that row were set to 0 and the conversion chose the first category as the imputed value. After imputation was complete on those features that required it, the data which was not imputed was concatenated to the imputed data. Evaluation of the imputed dataset showed 100% accuracy in imputing the observed (not missing) values in both attempts at imputation.

4.2 Model 1: Logistic Regression

The first model we used to predict an individual's relation to mental health issues was logistic regression. There were two main variables we were interested in classifying:

- Whether a person sought treatment
- Whether a person had ever been diagnosed with a mental illness

The first was encoded in our datasets in a column called "sought_treatment", while the second was encoded as "ever_diagnosed". Both of these are boolean variables - where 1 represents True and 0 represents False.

We decided to use `sklearn`'s functionality to help drive this logistic regression. In particular, we decided to use the `DictVectorizer` located in `sklearn.feature_extraction`, and `LogisticRegression` from `sklearn.linear_model` whose loss function is

$$l_{logistic} = \sum -y log(y') + (1-y) log(1-y')$$

.

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html`

Dataset	Ordinal / Boolean/ Real values	Categorical Values
Set 1	KNN imputation (n=3), rounded	Most common element used to fill in missing data in each column
Set 2	KNN imputation (n=3), rounded	Converted to a one-hot encoded dataset, KNN imputation (n=3), then rounded and converted back separately from other data

Table 2: Breakdown of imputation treatment for each dataset created.

In addition, to validate our results, we decided to run K-Fold cross-validation. We wanted to see whether the results we got from a simple one-time training-validation split (using `train_test_split` from `sklearn.model_selection`) were more or less accurate and representative for the entire dataset. Since our dataset was not too large, we were also able to use Leave-One-Out cross-validation for maximum confirmation of accuracy.

The features we attempted to fit on were as follows:

1. Age
2. Country
3. State (if applicable)
4. Gender
5. Race
6. Family History
7. Whether one would talk about mental health with their friends and family.
8. Whether one works primarily in the tech industry or not
9. Whether one knows resources for mental health.
10. Why one would (or would not) bring up mental health in an interview.

Encoding was done as follows:

- Features 2 through 6 were one-hot encoded using `DictVectorizer` from `sklearn`
- Gender was encoded with three categories: male, female, and non-binary using one-hot
- Features 1, 7, and 9 were ordinal encoded
- Age was already a discrete ordinal variable and was not transformed
- Comfortability with talking to friends and family was on a scale from 0-10 and not transformed Knowledge about resources was encoded as follows: 0 = "No, I don't know any", 1 = "I know some", and 2 = "Yes, I know several"
- Feature 10 was converted from text to continuous values using `CountVectorizer`

We also did some data analysis / preprocessing using NLP tools. Using Google's Universal Sentence Encoder, we decided to try and visually see if there were a lot of overlaps in similarity / meaning between different people's responses for why they might bring up mental health in an interview, and why they might not, as that was the most prominent text-based field in our dataset, and we wanted to see if there could be similarities drawn between all of these responses, should they be helpful in classifying people into one category or another.

4.3 Model 2: Hinge Loss

We made a second attempt at fitting a model to classify whether or not a person has sought treatment for mental health, this time using Hinge Loss with Quadratic Regularization. We denote our input space, $X \in \mathbb{R}^{1525\times n}$, where each row corresponds to a survey participant, and each column corresponds to one of the n features we are using to make predictions. And our output space, $Y = \{1, 0\}$, where

$$y_i = \begin{cases} 1 & \text{person has sought treatment} \\ 0 & \text{otherwise} \end{cases}$$

Hinge loss enables us to determine a decision boundary and margins on the support vector to separate our predicted population of people who we expect to have sought treatment for mental health from those who have not. The loss function is specified as $\ell_{hinge}(x, y; w) = (1 - yw^T x)_+$. Notice that ℓ_{hinge} is 0 when $yw^T x \geq 1$, and $\ell_{hinge} > 0$ otherwise. Therefore, hinge loss does not penalize misclassified points that lie within the margin of the classification boundary.

We leveraged SVMs (support vector machines) in `sklearn` to implement the regression analysis. We performed a random 80-20 split of the data to construct a training set, and a testing set. Below in the results section, you can find accuracy and F1 scores for models with various features chosen. Here we ignore the respondent's demographic information to get a sense of how important a workplace's organizational structure is in relating to someone's mental health status.

Our initial choice of X includes the following 15 features:

1. Country they work in
2. Willingness to mention mental health in interview
3. Awareness of resources

Model Type	Train Acc	Test Acc.	F1 Score
80/20 Split	0.701	0.652	0.745
5-fold CV	0.681	0.660	0.736

Table 3: Performance of Hinge Loss on different datasets.

4. Awareness of employee resources
5. Awareness of mental health benefits in contract
6. Willingness to discuss with coworkers
7. Willingness to discuss with supervisor
8. Whether it is easier to talk about Physical Health
9. Revealing mental health disorder to clients
10. Comfortability sharing information with family/friends
11. Number of employees
12. Witnessing negative consequences from discussing Mental Health
13. Has medical coverage
14. Works for a tech company
15. Works in a tech role

- Feature 1 was encoded according to whether or not the survey participant works in the United States. Therefore, if the participant answered "United States of America", the feature is transformed into a 1, and 0 otherwise.
- Features 2 through 10, and 12 through 15 were encoded using a one-hot boolean encoding such that 1 denotes the feature statement is true, and 0 to denote the feature statement is false. For example, if someone responded 'Yes" to being aware of mental health resources, the feature was transformed to a 1. "No" responses were transformed to a 0.
- Feature 11 was transformed into ordinal values:
  - 0 = small sized company with 1-5 employees or 6-25 employees
  - 1 = medium sized company with 100-500 employees or 500-1000 employees
  - 2 = large company with more than 1000 employees

We used k-cross validation again to confirm that our initial model results were reasonable. Notice in table 3, the average train and test accuracies front the cross validation models are close to the performance of the initial model split. We now have confidence that the initial model split performance was not a coincidence.

5 Results

5.1 Model 1

In assessing our models, we decided to first use the field for "sought_treatment" as the basis for what we would do our classification on, where 1 represents that the individual sought treatment, and 0 represents that they did not.

Num	Model Features	Test Acc.	F1 Score
1.1	Baseline: Age, Country, Gender	0.726	0.841
1.2	Age, Country, Gender, Race	0.721	0.838
1.3	Age, Country, Gender, Family History	0.695	0.799
1.4	Age, Country, Gender, Race, Family History	0.695	0.799
1.5	Age, Country, Gender, Race, Family History, Discussing MH in interview	0.732	0.819
1.6	Age, Country, Gender, Race, Family History, Discussing MH in interview / Share with Friends and Family	0.742	0.824
1.7	Age, Country, Gender, Race, Family History, Discussing MH in interview / Share with Friends and Family, Primarily Tech	0.742	0.824

Table 4: Performance of Logistic Regression with different feature sets.

We used two main metrics: exact match accuracy and F1. Exact match accuracy is the percentage of the predicted y-values that match the actual y-values. F1 is an average of precision and recall.

When splitting our data, we used an 80/20 train/validation split.

The results can be found in table 4

Next, we decided to run the data using Leave-One-Out cross validation on our most complicated model to see the effects of running our model on different datasets to see if we could further validate our model and see if we could achieve better performance with a different dataset. We found that by doing so, our mean exact match score over all different models trained was 0.749 and our mean F1 score was 0.616. The exact match improving after using LOO was to be expected, as it allowed us to try all sorts of datasets, some of which might produce models with parameters that better represented the underlying data. However, the mean F1 score dropping significantly was somewhat surprising. We chalk this up to calculating F1 over the validation set. With the validation set size being 1, this means the precision and recall that make up the F1 score are going to be very binary (either 0 or 1), so this makes maintaining an accurate representation of the actual F1 score over many trials difficult, hence why the value decreased significantly.

Next, we decided to run k-Fold Cross validation with k=50 different splits. When we ran this, we found that our mean exact match score was 0.748, similar to the previous step, and our mean F1 score was 0.842,

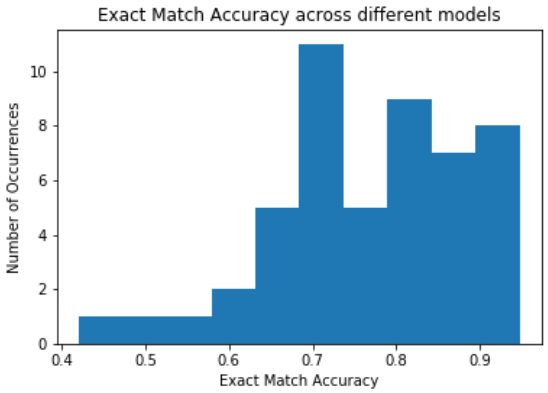

Figure 5: Exact match histogram across different cross validation datasets.

much better than the previous LOO model. This is not surprising. Like the LOO model, K-fold cross validation can find the best model over many different options with underlying datasets, making it easier to find the best model for the data. The exact match score was marginally lower than previously, since it had fewer datasets to run its trials over. However, the F1 score was much better and in line with our thoughts.

When we plotted a histogram of our exact match accuracies after each trial, we ended up with figure 5, and we can see that around 0.7 seems to be the most common accuracy, and our histogram is somewhat left-skewed.

5.1.1 Side note: ever_diagnosed

We also tried running our model to predict whether someone was ever diagnosed with a mental illness. However, when we did so, we realized that this would be a difficult task, as around 99% of the entries are "yes" and the rest are "no". Although it is somewhat discouraging that our data is so incredibly skewed in one direction, it makes sense - people who have been diagnosed with mental illness are more likely to care about mental illness and the issues that surround it, so they would be more likely to fill out a survey directly pertaining to mental health issues.

When we ran this fitting, we found that we received a mean accuracy score of 0.986 and a mean F1 score of 0.993 over the validation set. However, we cannot rejoice at this seemingly incredibly high accuracy classification. The fact that the data was incredibly skewed means that it is incredibly easy for a model to simply output all "yes"s for every single entry and still get an incredibly high accuracy score.

We decided to include analysis on the short-answer

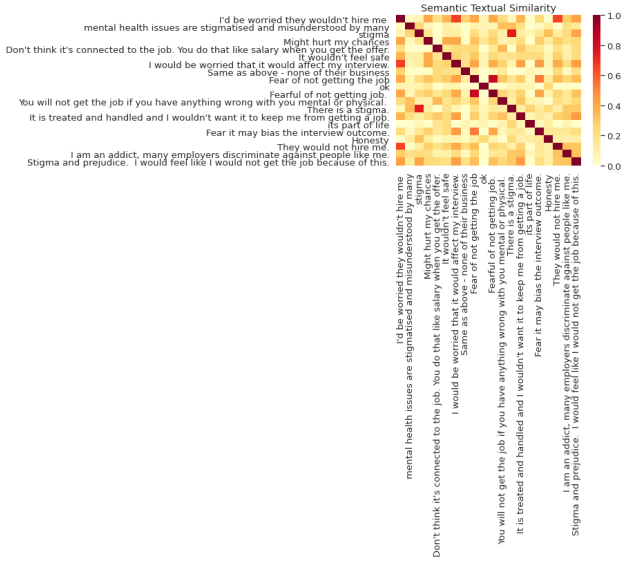

Figure 6: Heatmap of similarities between sample sentences in the dataset.

responses for why someone might (or might not) mention mental health in an interview. We did this to include the overall pattern of responses as a feature in the model. A heatmap of a subset of the entries is shown in Figure 6.

From here, we can see that most of the similarities are mid-range, even though they all end up with a similar theme. This makes sense, as it is somewhat taboo of a subject to talk about when you are trying to portray yourself in the best possible light. A histogram of similarities between all entries removing similarities of a response with itself is shown in Figure 7.

Figure 7 has a right-skewed distribution, with the majority of similarities seem to be sitting around 0.2. When calculating statistics, we found that the mean similarity was 0.183, the max similarity was 0.837, and the standard deviation of similarity was 0.156.

5.2 Model 2

The Hinge Loss model was tested with five combinations of features, denoted in the table above. The feature combinations were selected to reflect themes including:

- Someone's willingness to discuss mental health
- Their accessibility to mental health resources
- Their company type and size

So for example, the "Willingness to discuss Mental Health" feature set includes data about comfortability discussing mental health with friends, family, co-workers, supervisors, and in an interview. The combination of accessibility and willingness features

Num	Model Features	Train Acc.	Test Acc.	F1 Score
2.1	Features 1 through 15	0.665	0.685	0.768
2.2	Willingness to discuss Mental Health (Features 2, and 6 through 10)	0.695	0.694	0.778
2.3	Accessibility: Awareness of resources and access to mental health coverage (Features 3, 4, 5, and 13)	0.619	0.622	0.766
2.4	Company Type: country, size, tech or non-tech (Features 1, 11, 14, 15)	0.642	0.635	0.717
2.5	Model 2 + 3 (Features 2, 4-10, 13)	0.699	0.698	0.779

Table 5: Predicting whether someone sought treatment (Hinge Loss With Quadratic Regularization)

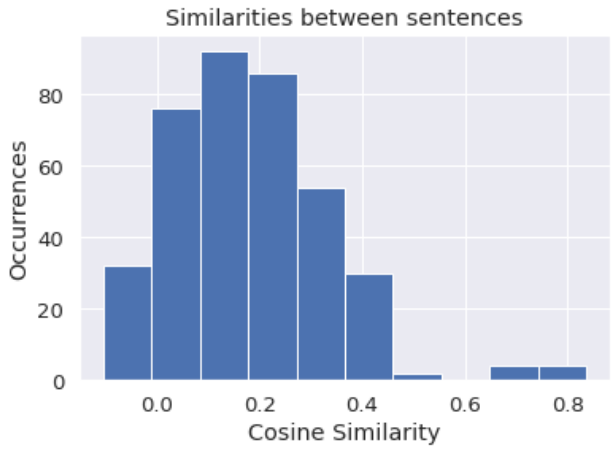

Figure 7: Histogram of similarities between sentences.

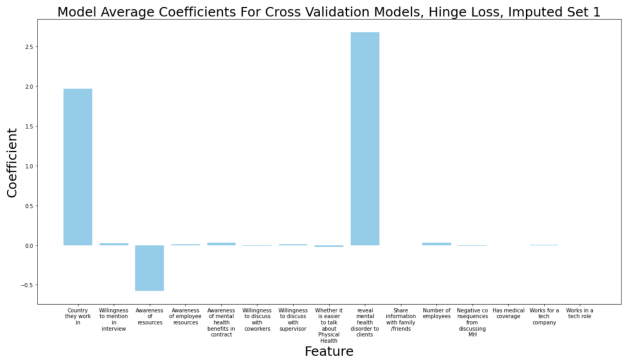

Figure 8: Hinge Loss coefficients on the first imputed dataset.

proved to be the most effective model, yielding approximately 70% classification accuracy. This is unsurprising that being able to talk about mental health without worrying about negative repercussions, and being able to access services to support mental health are good predictors for knowing whether or not someone is suffering from a mental health concern.

6 Discussion

6.1 Model 1

6.1.1 Logistic Regression

In the logistic regression model, we found that the first few models actually underfit the data, as there was not enough information to make meaningful predictions. The last model that utilized all the features

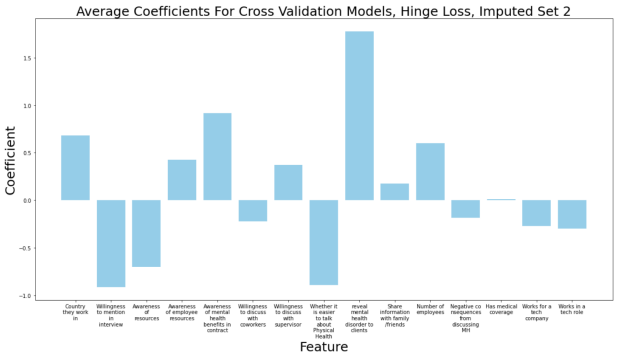

Figure 9: Hinge Loss coefficients on the second imputed dataset.

resulted in decent overfitting of the data since the training accuracy was about 15% higher than the testing accuracy. Playing around with the transformations led to a middle model (containing age, country, race, gender, and family history) that only slightly overfit the training data by about 2%. Whether the full model's overfitting is an acceptable margin of overfitting depends on the context in which this data will be used. We suspect the use of `CountVectorizer` made it easier to overfit the training data, since the data was sparse and the responses were each fairly short. This made it difficult to generalize the text field given the other features included.

6.2 Model 2

In the hinge loss model, quadratic regularization was applied to add a penalty for respondents who were outliers. Using all 15 features resulted in slight overfitting of the data since the training accuracy was about 2% higher than the testing accuracy. We experimented with different feature configurations, using just four features for Model 2.3 resulted in slight underfitting since the training accuracy is slightly lower than the testing accuracy. Using between 5 and 9 features reduced the overfitting and underfitting and yielded nearly equivalent accuracies for the training and testing sets (see Model 2.2 and 2.5). Model 2.5, which yielded the highest average prediction accuracy, had a 0.1% difference between training and testing accuracy. On another note, since the dataset only included about 1500 participants, we suspect the small amount of data to cause such an ideal margin. In the end, the best test accuracy score achieved with reasonable avoidance of over and underfitting was 0.698.

6.3 Synthesis

Overall, our models and the data suggest that the culture of a workplace has an impact on workers' propensity to reach out for professional help. The features addressing employees' **willingness to speak** about their mental health and their **awareness of mental health resources** had larger coefficients compared to others, and models which included these features outperformed other models. Although this outperformance is slight, the take-home message remains relevant as something for employers to consider. More data would be necessary to get more distinct proof, and we hope that this survey continues to be distributed yearly to continue work in this important field.

We are somewhat confident in our results, but the number of samples in our dataset is not large enough that we would be confident in using this model in a production setting. A good amount of the data was missing and had to be imputed, and there were not enough rows to begin with to build a model that could generalize well over all possible cases. We expect people who filled out this survey to be people who were more passionate about mental health causes, so the survey responses reflected this. Responses tended to be more skewed towards people who had gone out and sought treatment or had already been diagnosed with mental illnesses. If we had a much larger dataset with tens of thousands of examples that was more accurately sampled over the entire population, rather than having a small sample of those who are more invested in the causes fill it out on their own volition, we could build a model that we would be more confident in deploying to production and use in the real world. For the time being, however, we are not willing to make any claims of such.

Although this model does not have the ability to force someone to get treatment, in the hands of certain employers this model could have a negative impact on potential employees, suggesting that this model could theoretically be used as a weapon of math destruction. For example, if a potential employee is being interviewed for a new position, and the model indicates that the employee is predicted to seek treatment for mental health issues, the employer may decide against hiring them for fear that they may have mental health issues. Even with a more accurate model, having potential or current employers be unwilling to speak to their employees about mental health could have negative consequences for those suffering with these illnesses.

7 Conclusion

Predicting the propensity of workers in the tech industry towards reaching out for help with mental illnesses could help reach individuals who are quietly suffering to lead them to professional help. Using survey data from OSMI which was imputed using KNN nearest neighbors imputation, we utilized natural language processing, logistic regression, and Hinge regression to find the best prediction which reached close to 75% accurate. A lack of data, including many missing fields in the surveys, was a limitation of this study. In the future, for companies looking to improve their culture and take care of their employees, it could be useful to de-stigmatize mental health through group discussions, explicit advertisement of resources, and leadership from a supervisor to encourage these changes.