Predictions using the OSMI Mental Health Database

Erica Junqueira
Peter Salveson
Kadi Thiessen

August 6, 2019

Contributions by Team Member:

- Discussions to determine data cleaning approach, validation and training set split, and questions to answer – All
- Data Cleaning Execution – Peter
- Data Training and Test Set – Peter
- Introduction – Kadi
- Data Explanation – Peter
- Report Compilation – Peter
- Exploratory Analysis:
    - Barplots of variable comparisons by country – Kadi
    - Tables comparing treatment vs. other variables – Kadi
    - Barplots of variables by country – Erica
    - Barplots of variables by wellness_program – Erica
    - Hierarchical Trees – Peter
- Classification:
    - Country Logistic Regression, LDA, and QDA – Kadi
    - Treatment Logistic Regression, LDA, and QDA – Kadi
    - Wellness Program Logistic Regression – Kadi
    - Country Ridge and Lasso Regression – Erica
    - Treatment Ridge and Lasso Regression – Erica
    - Wellness Program Ridge and Lasso Regression – Erica
    - Country Random Forest Classification – Peter
    - Treatment Random Forest Classification – Peter
    - Wellness Random Forest Classification – Peter
- Conclusion – All

Predictions using the OSMI Mental Health Database

Erica Junqueira
Peter Salveson
Kadi Thiessen

August 6, 2019

Contents

1 Introduction

In this analysis we explore the perception of mental health in the tech workplace by analyzing the responses to the 2014 Mental Health in Tech Survey. The survey asked questions regarding perceptions of Mental Health in the Tech Workplace, the extent of support for mental health conditions by employers, and consequences of mental health conditions in the tech industry. This survey highlights the vastly different levels of support and openness around Mental Health Conditions felt by individuals in the tech workplace, and also illustrates the confusion that exists around what type of Mental Health benefits may be available through an individual's employer.

Mental health is an important aspect of an individual's well-being, and may impact many aspects of an individual's life. The Healthy People initiative states that, "mental health is essential to a person's well-being, healthy family and interpersonal relationships, and the ability to live a full and productive life." It also goes on to say that, "mental health disorders also have a serious impact on physical health and are associated with the prevalence, progression, and outcome of some of today's most pressing chronic diseases, including diabetes, heart disease, and cancer. Mental health disorders can have harmful and long-lasting effects—including high psycho-social and economic costs—not only for people living with the disorder, but also for their families, schools, workplaces, and communities." [1] With the far-reaching impact of mental health conditions, it is imperative that individuals are able to receive mental health support to help combat the negative consequences associated with these conditions.

Through this exploration of the 2014 Mental Health in Tech Survey data, we aim to better understand the differences between mental health perceptions in US and Non-US countries, the factors that might influence an individual to seek treatment, and the factors that may be associated with more open workplace discussion of the topic. This is accomplished through exploratory analysis and regression modeling to identify possible trends.

2 Data

2.1 Description

The Mental Health in Tech Survey dataset [2] includes responses by 1259 individuals to a 2014 survey that measures attitudes towards mental health and frequency of mental health disorders in the tech workplace. See Table 2 and Table 3 in the Appendix for a description of the variables used in our analysis.

2.2 Data Cleaning

The following changes were made to the original dataset in preparation for analysis:

- The variables `State`, `Timestamp`, and `Comments` were removed.
- The variable `Country` was changed to reflect either residence within or without the USA.
- Missing values of the variable `work_interfere` were changed from missing to "No MHC" to reflect "no mental health condition".
- There were 47 unique responses to the question of `Gender`. The 47 responses were binned into one of three groups: Male, Female, and Non-binary.
- Observations with missing values were removed from analysis. This decreased the number of observations from 1259 to 1234.

2.3 Train and Test Sets

The training set is a sample of ~70% of the total observations (n=833). The data was stratified by `Country`, `treatment`, `Age`, and `wellness_program` prior to sampling. The remaining observations constitute our test set (n = 351)

3 EDA

3.1 Numeric Summaries

Summaries by variable:

Age	work_interfere	no_employees	leave
Min. :11	Never :207	1-5 :157	Don't know :552
1st Qu.:27	no-MHC :262	100-500 :172	Somewhat difficult:121
Median :31	Often :138	26-100 :283	Somewhat easy :262
Mean :32	Rarely :170	500-1000 : 59	Very difficult : 96
3rd Qu.:36	Sometimes:457	6-25 :283	Very easy :203
Max. :72		More than 1000:280	

Gender	benefits	care_options	wellness_program	seek_help
Female :246	Don't know:400	No :492	Don't know:182	Don't know:355
Male :978	No :368	Not sure:307	No :825	No :633
Non-binary: 10	Yes :466	Yes :435	Yes :227	Yes :246

anonymity	mental_health_consequence	phys_health_consequence
Don't know:802	Maybe:471	Maybe:272
No : 61	No :477	No :906
Yes :371	Yes :286	Yes : 56

coworkers	supervisor	mental_health_interview
No :256	No :385	Maybe:201
Some of them:764	Some of them:348	No :995
Yes :214	Yes :501	Yes : 38

phys_health_interview	mental_vs_physical	obs_consequence	tech_company
Maybe:548	Don't know:566	No :1055	No : 224
No :490	No :331	Yes: 179	Yes:1010
Yes :196	Yes :337		

Country	self_employed	family_history	treatment	remote_work
non-US:498	No :1091	No :752	No :611	No :868
US :736	Yes: 143	Yes:482	Yes:623	Yes:366

Number of respondents who sought treatment (`treatment`) by ease of taking leave (`leave`):

	Don't know	Somewhat difficult	Somewhat easy	Very difficult	Very easy
No	301	43	133	31	103
Yes	251	78	129	65	100

It appears that respondents for whom taking leave is somewhat or very difficult are more likely to seek treatment.

Number of respondents who sought treatment (`treatment`) by gender (`Gender`):

	Female	Male	Non-binary
No	78	532	1
Yes	168	446	9

It appears that non-binary and female individuals are more likely to seek treatment, compared to males.

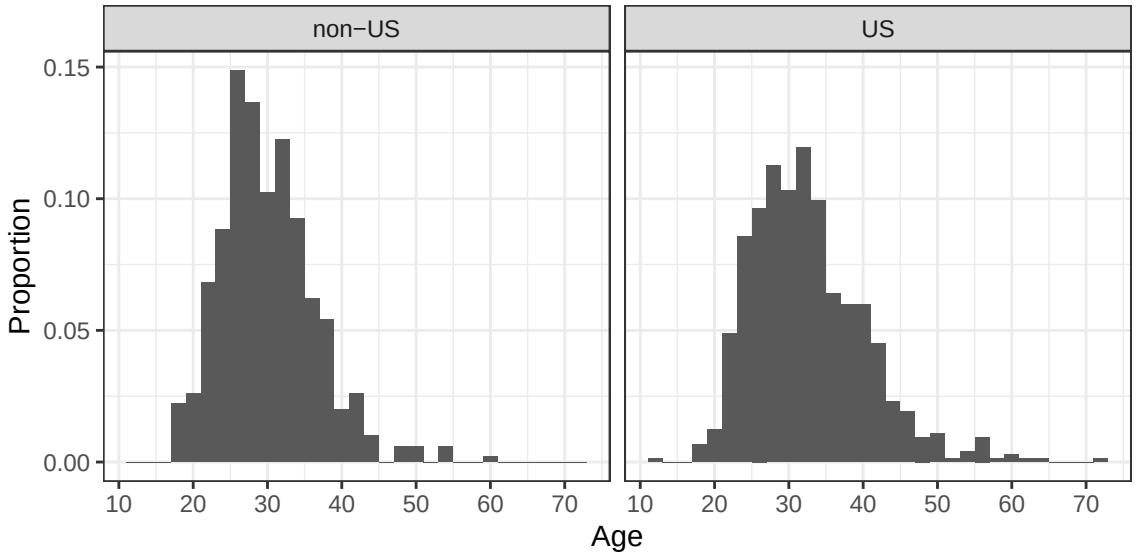
Number of respondents who sought treatment (`treatment`) by work interference (`work_interfere`):

	Never	no-MHC	Often	Rarely	Sometimes
No	177	258	21	49	106
Yes	30	4	117	121	351

It appears that respondents for whom a mental health condition interferes with their work are more likely to seek treatment compared to individuals who responded that they have no mental health condition or who's mental health condition never interferes with their work.

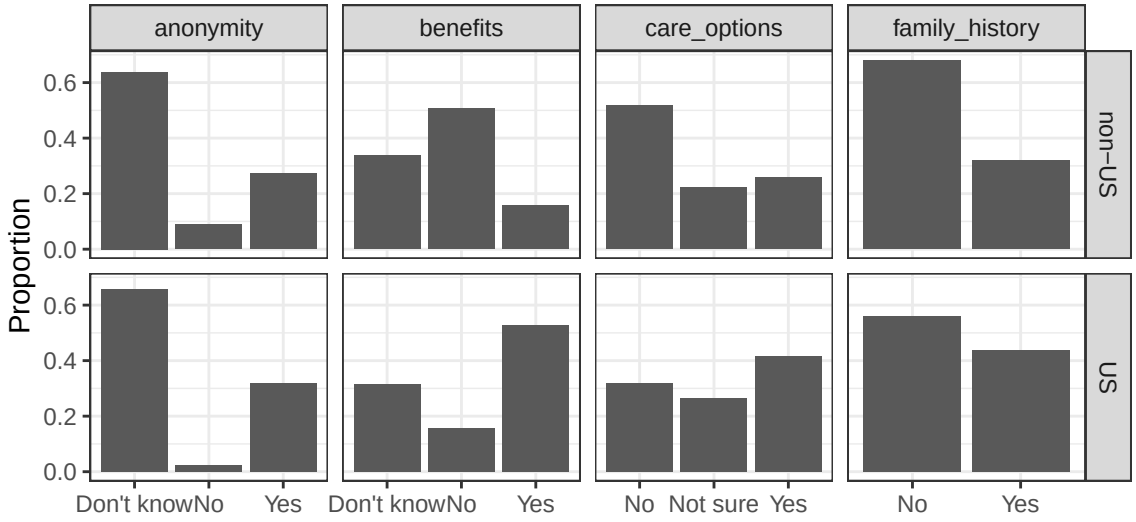
3.2 Plots

Age Distribution for US and Non-US Responses



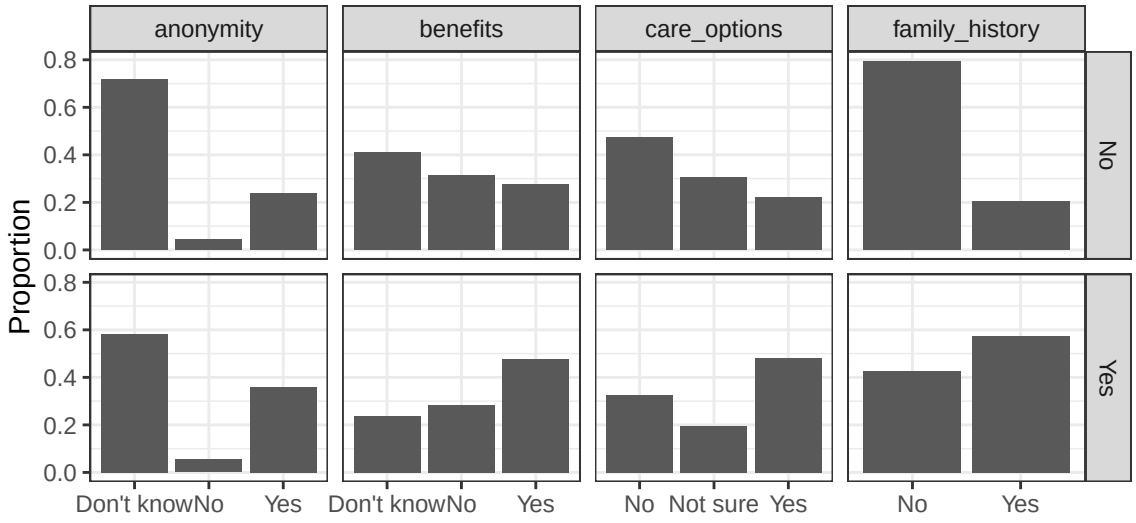
The US population of tech workers appears to skew older.

Responses to `anonymity`, `benefits`, `care_options`, and `family_history` partitioned by response to `Country`:



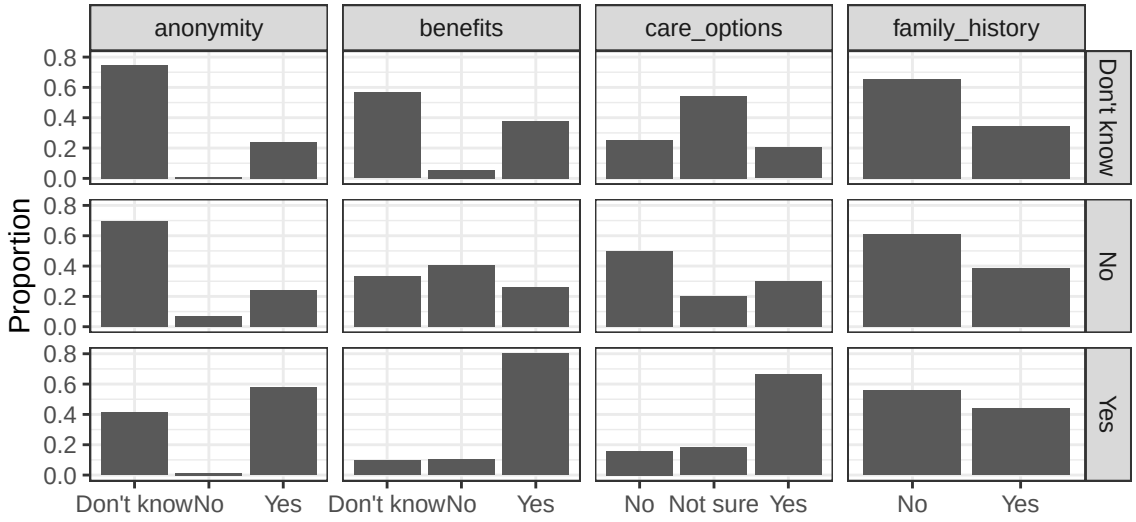
It appears that more US respondants had employer benefits, and were better educated about their care options.

Responses to `anonymity`, `benefits`, `care_options`, and `family_history` partitioned by response to `treatment`:



Respondants that sought treatment also appear to have more family history of mental conditions, more employer benefits, and more care options.

Responses to `anonymity`, `benefits`, `care_options`, and `family_history` partitioned by response to `wellness_program`:



Respondants who's employers have discussed mental health as part of a wellness program also have more awareness of care options, benefits, and greater anonymity.

3.3 Hierarchical Agglomerative Clustering

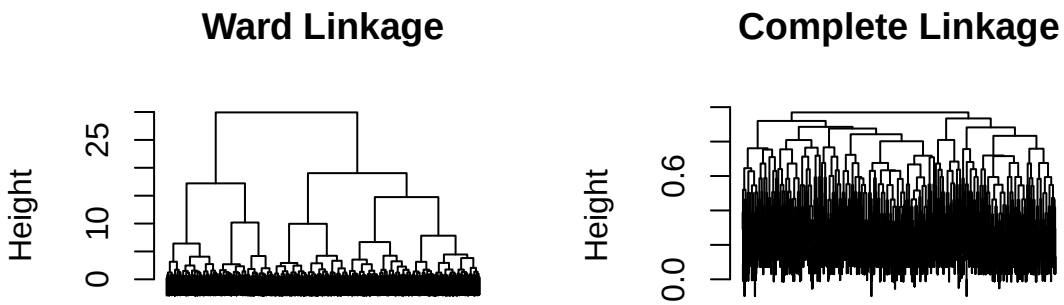


Figure 1: Dendrogram of HAC using Gower distance

Gower distance was used to create a dissimilarity matrix for all 24 variables. Gower distance was used instead of Euclidean distance due to the mixed nature of our dataset (continuous and categorical variables). Figure 1 shows the dendrograms for Ward's and Complete Linkage. Analysis of cuts of 2, 3, 4, and 5 for both linkage algorithms did not reveal any meaningful partitioning of the dataset.

4 Analysis

Please note that all of the following confusion matrices show the predicted test set classifications using models trained on the training set.

4.1 Predicting Respondent Country

4.1.1 Logistic Regression, Significant Predictors

Logistic regression models a binary response variable. Here we use logistic regression to model whether a respondent resides in the USA, or outside the USA. A predicted probability of more than .5 was assigned as the outcome in question (i.e., resides in the USA), while probabilities of less than .5 were assigned to the alternative outcome (i.e., resides outside the USA).

```
Confusion Matrix and Statistics

          Reference
Prediction non-US  US
    non-US     90  41
    US         55 165

               Accuracy : 0.726
                 95% CI : (0.677, 0.772)
    No Information Rate : 0.587
    P-Value [Acc > NIR] : 3.66e-08
```

Variables: Age, Gender, self_employed, family_history, benefits, care_options, seek_help, anonymity, leave, phys_health_consequence, obs_consequence

We are able to calculate the probability that an observation belongs to the reference class using the following:

$$P = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_i X_i}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_i X_i}}$$

β_i represents the coefficient for variable X_i. A positive coefficient increases the probability of the outcome, while a negative coefficient decreases the probability of the outcome.

Logistic regression model coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.67064	0.4883	-1.37342	1.70e-01
Age	0.04997	0.0118	4.23024	2.33e-05
GenderMale	-0.61026	0.2259	-2.70130	6.91e-03
GenderNon-binary	-2.07815	1.0983	-1.89219	5.85e-02
self_employedYes	-0.68750	0.2718	-2.52970	1.14e-02
family_historyYes	0.39187	0.1718	2.28141	2.25e-02
benefitsNo	-0.83564	0.2377	-3.51540	4.39e-04
benefitsYes	1.02253	0.2504	4.08404	4.43e-05
care_optionsNot sure	0.21052	0.2130	0.98832	3.23e-01
care_optionsYes	0.51409	0.2281	2.25428	2.42e-02
seek_helpNo	-0.52740	0.2122	-2.48564	1.29e-02
seek_helpYes	-0.41965	0.2897	-1.44836	1.48e-01
anonymityNo	-1.22698	0.4421	-2.77505	5.52e-03
anonymityYes	-0.14142	0.2178	-0.64938	5.16e-01
leaveSomewhat difficult	0.10114	0.3062	0.33034	7.41e-01
leaveSomewhat easy	-0.81099	0.2181	-3.71826	2.01e-04
leaveVery difficult	0.26334	0.3468	0.75935	4.48e-01
leaveVery easy	-0.95900	0.2459	-3.89959	9.64e-05
phys_health_consequenceNo	0.66943	0.2060	3.24942	1.16e-03
phys_health_consequenceYes	-0.00376	0.4311	-0.00872	9.93e-01
obs_consequenceYes	-0.70409	0.2487	-2.83150	4.63e-03

Each model coefficient can either increase and decrease the likelihood of an individual being from the US. Increases in age, having a family history of mental illness, having benefits, knowing their care options and not being sure about their care option, finding it very difficult to take leave, and believing that discussing a physical health issue with an employer would not cause a negative consequence all increased the likelihood that an individual was from the US. Not being female (i.e. being male or non-binary), being self-employed, not having benefits, not having employers that provided resources to seek help, not being confident that their anonymity would be respected, finding it somewhat easy or very easy to take leave for mental health issues, believing that discussing a physical health issue would have negative consequences with their employer, and having observed negative consequences for co-workers with mental illness all decreased the likelihood that the individual was from the US.

4.1.2 Logistic Regression with Ridge

Ridge regression is a method of shrinking regression coefficients to achieve better predictions. The regression coefficients and optimal value for the tuning parameter (λ) was found using ten-fold cross validation and the training set. That model was then used to predict the classification of the training set.

```
Confusion Matrix and Statistics

          Reference
Prediction non-US  US
    non-US     96  39
    US         49 167

               Accuracy : 0.749
                 95% CI : (0.701, 0.794)
    No Information Rate : 0.587
    P-Value [Acc > NIR] : 1.41e-10
```

$\lambda = .0291$

4.1.3 Logistic Regression with Lasso

Least absolute shrinkage and selection operator (LASSO) is similar to Ridge regression, and was performed with similar cross validation and lambda selection. Lasso replaces the L1 penalty from Ridge with an L2 penalty. This allows coefficients to be shrunk to zero, and thus exclude them from the model. Lasso has an advantage over ridge regression in terms of model interpretation. Lasso allows one to see a subset of predictor variables and of those, the most significant predictor variables will have a larger coefficient.

```
Confusion Matrix and Statistics

          Reference
Prediction non-US  US
    non-US     82  37
    US         63 169

               Accuracy : 0.715
                 95% CI : (0.665, 0.762)
    No Information Rate : 0.587
    P-Value [Acc > NIR] : 4.21e-07
```

$\lambda = .0291$

Non-zero coefficients: Age, Gender, self_employed, family_history, no_employees, remote_work, benefits, care_options, seek_help, anonymity, leave, phys_health_consequences, mental_health_interview, phys_health_interview

Lasso and logistic regression both found the following variables as non-zero / significant: Gender, self_employed, family_history, benefits, care_options, seek_help, anonymity, leave, and phys_health_consequence. Logistic regression also included obs_consequence. The Lasso included no_employees, remote_work, mental_health_interview, and phys_health_interview. Overall, the Lasso method included more variables than the Logistic regression method, with 10 of the variables overlapping between the two.

4.1.4 Linear Discriminant Analysis (LDA)

LDA produces similar results as logistic regression when the predictors are multivariate normal. However, LDA may be more stable than logistic regression in certain instances, such as if n is small, and also can be

applied to models that have >2 categorical responses.

```
Confusion Matrix and Statistics

          Reference
Prediction non-US  US
    non-US     91  40
    US         54 166

               Accuracy : 0.732
                 95% CI : (0.683, 0.778)
    No Information Rate : 0.587
    P-Value [Acc > NIR] : 9.93e-09
```

Variables: Age, Gender, self_employed, family_history, benefits, care_options, seek_help, anonymity, leave, phys_health_consequence, obs_consequence

4.1.5 Quadratic Discriminant Analysis (QDA)

QDA is similar to LDA, but it provides a quadratic decision boundary, which may provide a better fit to the true decision boundary. Additionally, there is a bias / variance trade off between LDA and QDA. LDA has lower variance and higher bias, whereas QDA increases variance, but may reduce bias.

```
Confusion Matrix and Statistics

          Reference
Prediction non-US  US
    non-US     75  42
    US         70 164

               Accuracy : 0.681
                 95% CI : (0.629, 0.729)
    No Information Rate : 0.587
    P-Value [Acc > NIR] : 0.000178
```

Variables: Age, Gender, self_employed, family_history, benefits, care_options, seek_help, anonymity, leave, phys_health_consequence, obs_consequence

4.1.6 Random Forest Classifier

The random forest classifier builds a number of classification trees (this report uses 1000 trees) from bootstrapped samples of our training data. For each split in each tree, a random sample of m predictors (where $m < p$, $m = 4$ in this report) are chosen as candidates for the split. This has the effect of "decorrelating" the classification trees and preventing a few strong predictors from making the classification trees too similar to one another.

```
Confusion Matrix and Statistics

          Reference
Prediction non-US  US
    non-US     93  37
    US         52 169

               Accuracy : 0.746
                 95% CI : (0.698, 0.791)
    No Information Rate : 0.587
    P-Value [Acc > NIR] : 2.97e-10
```

4.2 Predicting Treatment

4.2.1 Logistic Regression, Significant Predictors

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  132  18
       Yes  41 160

               Accuracy : 0.832
                 95% CI : (0.789, 0.87)
    No Information Rate : 0.507
    P-Value [Acc > NIR] : < 2e-16
```

Variables: Age, Gender, family_history, work_interfere, benefits, coworkers

Logistic regression model coefficients:

```
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              -3.3778     0.5796 -5.8280 5.61e-09
Age                       0.0345     0.0129  2.6805 7.35e-03
GenderMale               -0.6943     0.2727 -2.5456 1.09e-02
GenderNon-binary         13.8447   615.3679  0.0225 9.82e-01
family_historyYes         1.0142     0.2049  4.9489 7.46e-07
work_interfereno-MHC     -2.7134     0.7542 -3.5977 3.21e-04
work_interfereOften       3.4123     0.3926  8.6918 3.57e-18
work_interfereRarely      2.6067     0.3302  7.8935 2.94e-15
work_interfereSometimes   2.8064     0.2905  9.6604 4.44e-22
benefitsNo                0.2949     0.2467  1.1953 2.32e-01
benefitsYes               1.0090     0.2469  4.0877 4.36e-05
coworkersSome of them     0.1748     0.2393  0.7302 4.65e-01
coworkersYes              0.9479     0.3419  2.7725 5.56e-03
```

Two coefficients decrease the likelihood that an individual sought treatment: the individual being male, and the individual suggesting that they do not have a mental health condition in the question about work interference. There are other variables that increase the likelihood that an individual sought treatment for a mental health condition. The largest in magnitude is an individual identifying as a non- binary gender.

Additionally, having a family history of mental illness, reporting that their mental condition interfered with work sometimes, rarely, or often, having benefits, and being able to discuss mental health issues with coworkers all increased the likelihood that an individual sought treatment.

4.2.2 Logistic Regression with Ridge

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  134  22
       Yes  39 156

               Accuracy : 0.826
                 95% CI : (0.782, 0.864)
    No Information Rate : 0.507
    P-Value [Acc > NIR] : <2e-16
```

$\lambda = 0.0277$

4.2.3 Logistic Regression with Lasso

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  130   9
       Yes  43 169

               Accuracy : 0.852
                 95% CI : (0.81, 0.887)
    No Information Rate : 0.507
    P-Value [Acc > NIR] : < 2e-16
```

$\lambda = 0.0291$

Non-zero coefficients: Age, Gender, family_history, work_interfere, benefits, care_options, anonymity, coworkers, mental_health_interview, phys_health_interview

Logistic regression and Lasso Regression both found the variables of Age, Gender, family_history, work_interfere, benefits, and coworkers to be significant/nonzero. Lasso also included care_options, anonymity, mental_health_interview and phys_health_interview. Overall the Lasso method included more variables than the Logistic regression model, with 6 of the variables overlapping between the two.

4.2.4 LDA

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  122  10
       Yes  51 168

               Accuracy : 0.826
                 95% CI : (0.782, 0.864)
    No Information Rate : 0.507
    P-Value [Acc > NIR] : < 2e-16
```

Variables: Age, Gender, family_history, work_interfere, benefits, coworkers

4.2.5 QDA

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  123   9
       Yes  50 169

               Accuracy : 0.832
                 95% CI : (0.789, 0.87)
    No Information Rate : 0.507
    P-Value [Acc > NIR] : < 2e-16
```

Variables: Age, family_history, work_interfere, benefits, coworkers

4.2.6 Random Forest Classifier

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  136  17
       Yes  37 161

               Accuracy : 0.846
                 95% CI : (0.804, 0.882)
    No Information Rate : 0.507
    P-Value [Acc > NIR] : < 2e-16
```

4.3 Predicting Wellness Program

4.3.1 Multinomial Logistic Regression with Ridge

```
Confusion Matrix and Statistics

             Reference
 Prediction   Don't know  No Yes
   Don't know         10   2   2
   No                 33 224  22
   Yes                 7  13  38

 Overall Statistics

                Accuracy : 0.775
                  95% CI : (0.728, 0.818)
     No Information Rate : 0.681
     P-Value [Acc > NIR] : 6.39e-05
```

$\lambda = 0.154$

4.3.2 Multinomial Logistic Regression with Lasso

```
Confusion Matrix and Statistics

             Reference
 Prediction   Don't know  No Yes
   Don't know         15   1   2
   No                 27 223  20
   Yes                 8  15  40

 Overall Statistics

                Accuracy : 0.792
                  95% CI : (0.746, 0.833)
     No Information Rate : 0.681
     P-Value [Acc > NIR] : 2.42e-06
```

$\lambda = .01334$

Non-Zero Coefficients: Remote_work, tech_company, benefits, care_option, seek_help, mental_health_consequences, coworker, Gender, leave, mental_vs_physical, supervisor, mental_health_interview, self_employed, no_employees, anonymity, obs_consequences

4.3.3 LDA

```
Confusion Matrix and Statistics

             Reference
 Prediction   Don't know  No Yes
   Don't know         29  19   4
   No                 15 202  12
   Yes                 6  18  46

 Overall Statistics

                Accuracy : 0.789
                  95% CI : (0.743, 0.831)
     No Information Rate : 0.681
     P-Value [Acc > NIR] : 4.35e-06
```

Variables: benefits, care_options, seek_help, leave, mental_health_interview, obs_consequence

4.3.4 QDA

```
Confusion Matrix and Statistics

             Reference
 Prediction   Don't know  No Yes
   Don't know         36  33  12
   No                  9 191  18
   Yes                 5  15  32

 Overall Statistics

                Accuracy : 0.738
                  95% CI : (0.689, 0.783)
     No Information Rate : 0.681
     P-Value [Acc > NIR] : 0.011783
```

Variables: benefits, care_options, seek_help, leave, mental_health_interview, obs_consequence

4.3.5 Random Forest Classifier

```
Confusion Matrix and Statistics

             Reference
 Prediction   Don't know  No Yes
   Don't know          9   3   1
   No                 31 225  25
   Yes                10  11  36

 Overall Statistics

                Accuracy : 0.769
                  95% CI : (0.722, 0.812)
     No Information Rate : 0.681
     P-Value [Acc > NIR] : 0.000168
```

5 Test Accuracy Summary

Table 1: Summary of test accuracy by method and variable

	Country	treatment	wellness_program
Logistic Regression	0.726	0.832	-
Logistic Regression, Ridge	0.749	0.826	-
Logistic Regression, Lasso	0.715	0.852	-
LDA	0.732	0.826	0.789
QDA	0.681	0.832	0.738
Random Forest Classifier	0.746	0.846	0.769
Multinomial Log. Reg., Ridge	-	-	0.775
Multinomial Log. Reg., Lasso	-	-	0.792

6 Conclusion

We are able to predict country, treatment, and the employer wellness programs based on the survey response. The data shows perspectives on mental health in the tech industry varies whether the respondent resides in the USA, or in another country. These perspectives also impact whether an individuals seeks treatment for a mental health condition.

The differences between the USA versus other countries were first suggested during our exploratory data analysis and were confirmed by our analysis. These differences demonstrate areas where the USA excels, such as providing mental health benefits, as well as areas where the US can improve compared to other countries. One such area where the USA tech industry could improve is education of mental health benefits. USA tech employees were not sure if there would be difficulty taking leave for mental health issues. As a possible consequence, US employees are likely to say that a mental health condition interferes with their work sometimes or often.

We were able to determine key indicators that influence whether or not an individual has sought treatment for a mental health condition. Key areas identified that influence this were age, gender, family history of

mental illness, the extent their mental illness impacts their work, the benefits individuals have available to them, and an individual's willingness to discuss their mental health issue with a coworker or employer. In using these variables, we were able to predict if an individual has sought treatment for a mental health condition.

Lastly, we were also able to predict whether or not an employer has ever discussed mental health as part of a wellness program. Unsurprisingly, key predictors of this were related to other actions taken by the employers, such as if an individual had observed negative consequences for a coworker with a mental health condition and if the employer provides resources to learn more about mental illness and how to seek help. Additionally, whether or not the employer provides benefits for mental health and if the employees are aware of the care options available also influenced the prediction

The methods we employed to model `Country`, `treatment`, and `wellness_program` achieved similar test accuracies. See Table 1 for a summary of our results. Generally, the accuracies for Logistic Regression (with or without Ridge/Lasso), LDA, and Random Forest were very close. QDA did not perform as well as the other methods when predicting `Country` and `wellness_program`, but gave comparable results when predicting `treatment`.

Overall, valuable insights can be gained about mental health in the tech workplace based on the responses to the survey questions. Employers in every country can use these results to spark further discussion about mental health and help employees seek the treatment they need.

7 Appendix

7.1 Tables

Table 2: Variable Definitions

	Variable	Description of Variable	Values
1	Age	Respondent age.	Continuous
2	Gender	Respondent gender.	Female
2			Male
2			Non-binary
3	Country	Does the respondent live inside the USA?	US
3			Non-US
4	self_employed	Are you self-employed?	Yes
4			No
5	family_history	Do you have a family history of mental illness?	Yes
5			No
6	treatment	Have you sought treatment for a mental health condition?	Yes
6			No
7	work_interfere	If you have a mental health condition, do you feel that it	Often
7		interferes with your work?	Sometimes
7			Rarely
7			Never
7			No MHC
8	no_employees	How many employees does your company or organization have?	1-5
8			6-25
8			26-100
8			101-500
8			501-1000
8			>1000
9	remote_work	Do you work remotely (outside of an office) at least 50% of the time?	Yes
9			No
10	tech_company	Is your employer primarily a tech company/organization?	Yes
10			No
11	benefits	Does your employer provide mental health benefits?	Yes
11			No
11			Don't know
12	care_options	Do you know the options for mental health care your employer provides?	Yes
12			No
12			Don't know
13	wellness_program	Has your employer ever discussed mental health as part of an	Yes
13		employee wellness program?	No
13			Don't know
14	seek_help	Does your employer provide resources to learn more about mental health issues	Yes
14		and how to seek help?	No
14			Don't know

Table 3: Variable Definitions (cont.)

	Variable	Description of Variable	Values
15	anonymity	Is your anonymity protected if you choose to take advantage of mental health	Yes
15		or substance abuse treatment resources?	No
15			Don't know
16	leave	How easy is it for you to take medical leave for a mental health condition?	Very difficult
16			Somewhat difficult
16			Somewhat easy
16			Very easy
16			Don't know
17	mental_health_consequence	Do you think that discussing a mental health issue with your employer would	Yes
17		have negative consequences?	No
17			Maybe
18	phys_health_consequence	Do you think that discussing a physical health issue with your employer	Yes
18		would have negative consequences?	No
18			Maybe
19	coworkers	Would you be willing to discuss a mental health issue with your coworkers?	Yes
19			No
19			Some of them
20	supervisor	Would you be willing to discuss a mental health issue with your direct supervisor(s)?	Yes
20			No
20			Some of them
21	mental_health_interview	Would you bring up a mental health issue with a potential employer in an interview?	Yes
21			No
21			Maybe
22	phys_health_interview	Would you bring up a physical health issue with a potential employer in an interview?	Yes
22			No
22			Maybe
23	mental_vs_physical	Do you feel that your employer takes mental health as seriously as physical health?	Yes
23			No
23			Don't know
24	obs_consequence	Have you heard of or observed negative consequences for coworkers with	Yes
24		mental health conditions in your workplace?	No

References

[1] ODPHP. *Mental Health*. 2019. URL: https://www.healthypeople.gov/2020/leading-health-indicators/2020-lhi-topics/Mental-Health.

[2] OSMI. *Mental Health in Tech Survey*. 2014. URL: https://www.kaggle.com/osmi/mental-health-in-tech-survey.