**Achala Rao Shiravanthe**
**Urja Nadibail**
**Robert Bellinger**

**STAT 571**
**December 13, 2017**
**Final Project Paper**

## Abstract

Mental health issues are becoming increasingly prevalent and accounted as legitimate health issues in one's work environment. Using a single tree, neural nets, random forest, logistic regression and bagging, we have sought to identity factors correlated to employee's likelihood to seek treatment. The data used in the study was sourced from Kaggle. It is the 2016 Mental Health in Tech Survey from Open Source Mental Illness (OSMI), an Indiana based non-profit dedicated to raising awareness and resources to support mental wellness in the tech communities. Descriptive statistics made evident the amount of study participants who had sought treatment for mental illness. The survey included 1259 participants from 157 countries. The scope of data collection spans from workplace information concerning company size, coworkers, supervisors, and perceived consequences of mental illnesses. Moreover, basic demographic information and personal health data is gathered to provide a snapshot of the present state of affairs for tech employees managing mental illness. Analysis was conducted using logistic regression, neural network, random forest, bootstrap aggregation, and a single tree model.

## Introduction

A common issue for working professionals has been the management of stress, anxiety, and other personal health issues. These of course are compounded upon the tensions and issues from one’s separate home life, but the workplace is an excellent pool for sampling a cross-section of our own local community. Many of those close to us suffer from mental illness whether it is apparent or guarded by one’s professional countenance. Fortunately studies conducted by OSMI and other health professionals provide an insightful lens for being able to grasp the salience and prevalence of these very real personal health issues. In light of this, we hope that our research report can highlight factors which can be considered with greater care. Mental health is important.

According to the the National Institute of Mental Health (NIMH), roughly 43 million or 1 in 5 Americans suffer from mental health issues. Moreover, 10 million (or 1 in 25) experience such debilitating symptoms that their suffering induces serious functional impairment.[1]

Data developed by the Global Burden of Disease Study conducted by the World Health Organization reveal that mental illness, including suicide, accounts for over 15 percent of the burden of disease in established market economies, such as the United States. This is more than the disease burden caused by all cancers.[2]

18.1% of adults in the U.S. experienced an anxiety disorder such as posttraumatic stress disorder, obsessive-compulsive disorder and specific phobias.[3] Among the 20.2 million adults in the U.S. who experienced a substance use disorder, 50.5%—10.2 million adults—had a co-occurring mental illness.[4] As such, an understanding of the prevalence of mental health issues in the tech workspace can be useful for future researchers and managers. Our interest for entering the tech industry has also made this an interesting and relevant topic for analysis.

## Overview

Mental health issues are becoming increasingly prevalent and accounted as legitimate health issues in one’s work environment. Using a single tree, neural nets, random forest, logistic regression and bagging, we have sought to identity factors correlated to employee’s likelihood to seek treatment. The data used in the study was sourced from Kaggle. It is the

[1] https://www.nimh.nih.gov/about/directors/thomas-insel/blog/2015/mental-health-awareness-month-by-the-numbers.shtml

[2] http://www.who.int/topics/global_burden_of_disease/
The World Health Organization. The World Health Report 2004: Changing History, Annex Table 3: Burden of disease in DALYs by cause, sex, and mortality stratum in WHO regions, estimates for 2002. Geneva: WHO, 2004.

[3] Any Anxiety Disorder Among Adults. (n.d.). Retrieved January 16, 2015, from http://www.nimh.nih.gov/health/statistics/prevalence/any-anxiety-disorder-among-adults.shtml

[4] Substance Abuse and Mental Health Services Administration, *Results from the 2014 National Survey on Drug Use and Health: Mental Health Findings*, NSDUH Series H-50, HHS Publication No. (SMA) 15-4927. Rockville, MD: Substance Abuse and Mental Health Services Administration. (2015). Retrieved October 27, 2015 from http://www.samhsa.gov/data/sites/default/files/NSDUH-FRR1-2014/NSDUH-FRR1-2014.pdf

2014 Mental Health in Tech Survey from Open Source Mental Illness (OSMI). Descriptive statistics made evident the amount of study participants who had sought treatment for mental illness.

**Analysis Goals:**

This analysis aims to develop a model to predict the possibility of an individual seeking mental health treatment. Data was aggregated from various countries and states.

**Important Questions to Ask:**

- Does the availability of mental health benefits positively relate to the treatment of mental health issues?
- How does family history affect the treatment of mental illnesses?
- Does the stigma associated with mental health and it affecting one's career have an impact on an individual seeking treatment?

**Concerns and Limitations**

- The data contains a sampling bias from of the selection of survey respondents
- Participants may also be prone to voluntary response bias and may cause over representation of data due to their concerns of privacy or their openness, or unwillingness to divulge mental illness w/ or w/o family history
- The data is incomplete and contains several missing variables. Many of which needed to be cleaned and removed from the final analysis.
  - There contains significant amount of missing data pertaining to comments
- The data includes attributes which were excessively binary and did not provide a dimension of variance and differentiation for the sampled population.

Despite these limitations, the data represents an excellent starting point for developing a model which can identify the important factors  affecting mental health issues and treatment, especially in the high-technology world (Appendix A).

**Data Analysis for Model Selection**

Highlights: More males than females were included in the survey. The summary shows that there are 990 and 251 male and female participants, respectively. The majority of the parameters were categorical factors. This is an indication of the skewed gender ratio in a workplace.

Since the response variable, treatment, (exhibited in Appendix B) has relationships with numerous categorical variables, we applied five different models for classification. We applied a logistic regression model, random forest and neural net models to predict multinomial response. Additionally, we used bootstrap aggregating, or bagging, to improve classification while reducing variance and overfitting issues.

## Response Variable:

`treatment` Have you sought treatment for a mental health condition?

## Predictor Variables:

### Demographics, Mental Health Condition:

- `Timestamp`
- `Age`
- `Gender`
- `Country`
- `state`: If you live in the United States, which state or territory do you live in?
- `family_history`: Do you have a family history of mental illness?
- `work_interfere`: If you have a mental health condition, do you feel that it interferes with your work?

### Employment Background

- `self_employed`: Are you self-employed?
- `no_employees`: How many employees does your company or organization have?
- `remote_work`: Do you work remotely (outside of an office) at least 50% of the time?
- `tech_company`: Is your employer primarily a tech company/organization?

### Organizational Policies

- `benefits`: Does your employer provide mental health benefits?
- `care_options`: Do you know the options for mental health care your employer provides?
- `Wellness_program`: Has your employer ever discussed mental health as part of an employee wellness program?
- `seek_help`: Does your employer provide resources to learn more about mental health issues and how to seek help?
- `anonymity`: Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?
- `leave`: How easy is it for you to take medical leave for a mental health condition?

### Interpersonal Openness w/ Colleagues

- `mental_health_consequence`: Do you think that discussing a mental health issue with your employer would have negative consequences?
- `phys_health_consequence`: Do you think that discussing a physical health issue with your employer would have negative consequences?
- `coworkers`: Would you be willing to discuss a mental health issue with your coworkers?
- `supervisor`: Would you be willing to discuss a mental health issue with your direct supervisor(s)?
- `mental_health_interview`: Would you bring up a mental health issue with a potential employer in an interview?
- `phys_health_interview`: Would you bring up a physical health issue with a potential employer in an interview?
- `mental_vs_physical`: Do you feel that your employer takes mental health as seriously as physical health?
- `obs_consequence`: Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?
- `comments`: Any additional notes or comments

After data cleaning, the model was built using 1251 samples. 875(70%) samples were reserved for training and 376(30%) samples for testing.

**Data Cleaning**

One of first things we noticed when previewing the data was the extensive number of unique entries for `gender`. Since the data was collected as text boxes instead of providing gender options, we encountered different types of answers like “Guy - ish” and typos like “Maile” for *Male* and “f”, “fem”, “girl” and the likes indicating *Female*. To convert this into a categorical variable, we first had to normalize the `gender` data. After careful observation of the data, we recognized the commonly used words for both the genders and classified them into male and female buckets. For the values that we weren’t sure of the category, we created a separate non-M/F category.

Age was another variable that required some cleaning. We encountered ages that ranged from negative numbers to 120. Negative age values as well as outliers, as high as 120 were excluded from the data. Missing data for `work_interfere` was changed to “Never” to avoid eliminating additional variables. `Seek_help` and `anonymity` were not significant and removed. `Family_history`, `work_interfere`, `benefits`, and `care_options` were found to be significant at the 0.05 level. (See Anova Fit3 table in Appendix D).

After splitting the data, we performed a logistic regression and ran an ANOVA test to identify and remove insignificant factors. Since `state` and `self_employed` have NA values but are not significant at the 0.05 level, we could remove these columns from our data.

We removed predictors with substantial null values. For instance, the custom responses, or `comments` variable, was removed as their values were irrelevant to this analysis.

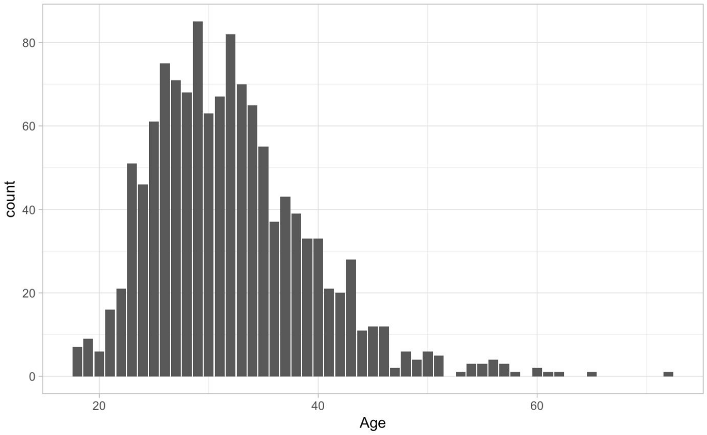
**Data Exploration**

In order to gain a general understanding, we first explored the data through univariate and bivariate analyses.

`Age` Distribution:
The tech industry is known for having a youthful work culture and unsurprisingly has a distribution skewed to the right. The median `Age` of 31 is less than the mean (32.077). SD is 7.29. The tech industry is primarily located on the US west coast, east coast, and also in highly developed countries. The demographics high cost of living, population density, and competitiveness of these atmosphere may contribute to the prevalence of mental illness as compared to the working conditions themselves. Interestingly, California accounted for 16%

of the survey data due to the high concentration of technology jobs based in the greater San Francisco Bay Area.

Following the data cleaning, we're left with a manageable age distribution. These values can be further transformed to categorize ages however as this is one of the few non-categorical parameters in this model, we maintain its function as a continuous variable.

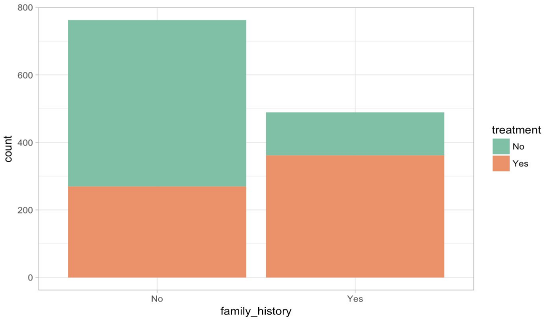


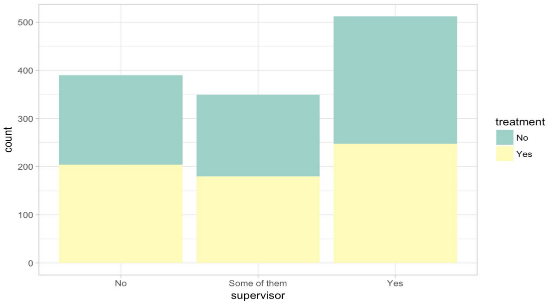
Willingness to Bring Up Health Issues in an Interview:

Males tend to be evenly split on their willingness to mention their physical health conditions whereas females expressed a slightly greater frequency of unwillingness. Few members of either gender were willing to bring up their mental health issues during job interviews. This may be attributed to the cultural stigma that such health issues may be impede job performance or that they may be overlooked for the very job offer they seek.

`phys_health_consequence` vs. `mental_health_consequence`:

While 74% of employees feel that they can be open about physical health issues with their employer, only 38% feel that way about personal mental health concerns. (Appendix E)



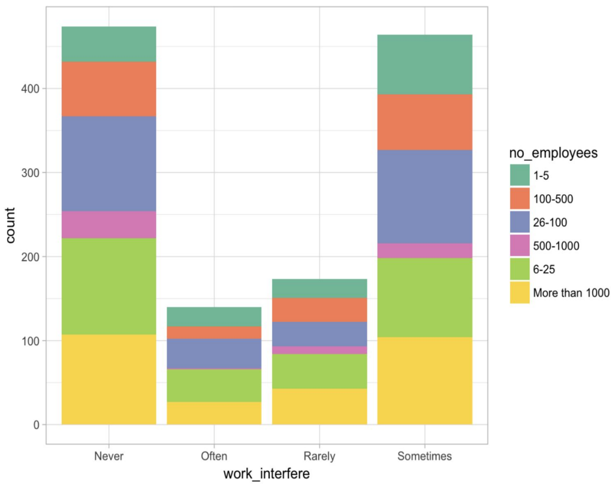


Above left, note that only 39.1% of respondents noted having a family history of mental illness. The chart apparent above right, is indicative of the trend for employees to shield and limit themselves from taking part in a dialogue with their supervisor for concern of mental health consequences. We found that the anonymity of workers was not a statistically significant factor for affecting individuals seeking treatment.

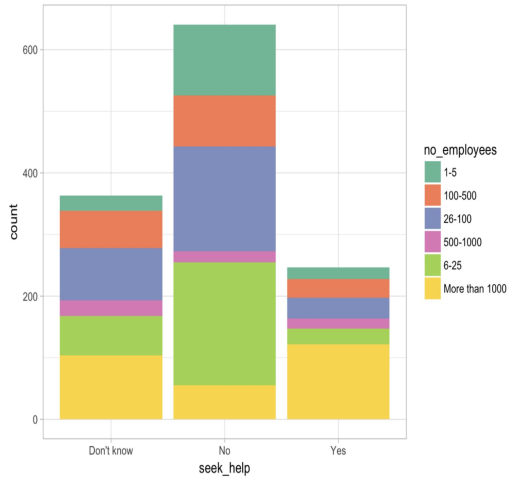
Employee Type:

The distribution of data with respect to tech and non-tech employees was clearly skewed in favor of tech companies with 81.9% of sampled respondents either working at a tech company or having a tech role in a non-tech company.  While a sizeable amount of people have mental health coverage via insurance benefits, this does not necessarily occur for all employees. Thus restricting and limiting the effectiveness and efficacy of the study relating to the extensiveness and distribution of such healthcare benefits.

We also observed that there were higher frequencies of employees in larger companies who felt that they could seek help. On the other hand, there's a stark difference with the amount of employees in small to mid-size companies (`no_employees` groups: 1-5, 5-25, and 25-100) who felt that they could seek help from their colleagues or supervisors. Midsize and large companies (`no_employees` groups 25-100 and 1000+) showed higher proportions of employees who experience interference with their work either `often` or `sometimes`.



Does the size of the company affect the awareness created about mental health in the organization?



An important limitation of this dataset as mentioned before is the bias which is presented by the actual respondent and the sampling bias based on the limitation of sampling to specific companies and geographies. Furthermore, the binary nature of yes and no questions limits the capacity to better qualify and the variation in mental health conditions.

Classification and prediction: We chose to use a limited number of variables from an original set of 27. While there are various techniques and methods for being able to complete this analysis we used classifiers which would be able to work best with this health related dataset with the intent of finding and making valuable and more accurate predictions. This for instance involves the weighing a sensitivity vs false positive analysis.

## MODEL SELECTION

### Logistic Regression

We used variable significance as a metric for model selection. In order to recognize and eliminate variables which were not significant at the 0.05 level, we did a p-test. The first model was built and we used chi-square test as a metric for testing the significance of each variable. Further, by doing a p-test, we followed this protocol with two subsequent fits. We decided that `anonymity`, `seek_help` were not very significant owing to their high p-values and hence removed them from our model. After a series of models, excluding the least significant predictor variable at every step, we arrived at the final model that includes the following variables:

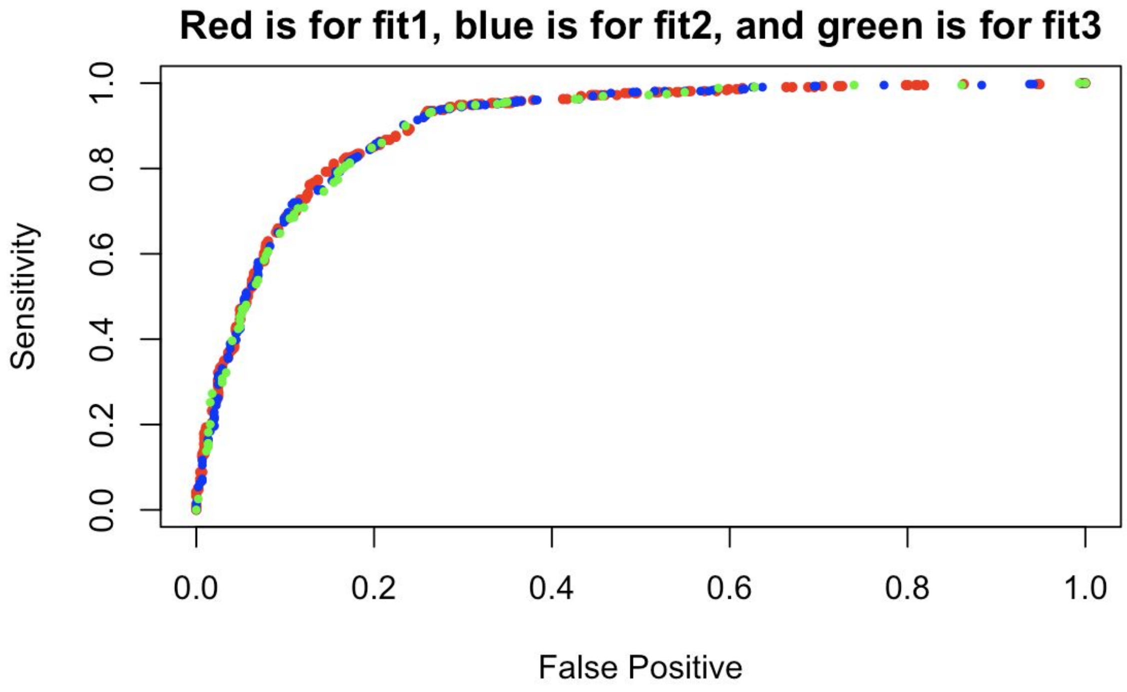
- `Family_history`
- `Work_interfere`
- `Benefits`
- `care_options`

The following is the final model that was built using logit:

```
treatment =   -3.3585 + 1.0321 * family_historyYes + 4.1556 * work_interfereOften
             + 3.1157 * work_interfereRarely + 3.6161*work_interfereSometimes
                -   0.1551 *benefitsNo + 0.8816 *benefitsYes
             - 0.2461*care_optionsNotsure +  0.5970 * care_optionsYes
```

Clearly `family_historyYes`, `work_interfereOften`, `work_interfereRarely`, `work_interfereSometimes`, and `care_optionsYes` correlate positively; whereas `benefitsNo` and `care_optionsNotsure` relate negatively with the final outcome which is to predict whether the worker would seek treatment.

The area under the ROC curve for the classifier was considered with a goal of minimizing the False Positive Ratio and to minimize the Positive ratio. The resultant misclassification error for the Logistic Regression model was 0.191. This seems like a good model with a decent prediction accuracy.



**Bagging**

In general, predictors from this type of model are useful for obtaining higher values of accuracy because the methodology is intrinsically designed for generating multiple versions of a predictor and using these to get an aggregated predictor.

We used 70% of the available data for training and the rest for testing as earlier. The underlying model here is a linear model. The resultant accuracy measured by our bagging predictor is 85.90% The highest level amongst all five of our models. This is likely because of the overlap in the training and the testing data.

| Observed Class | | |
|---|---|---|
| Class | No | Yes |
| No | 137 | 17 |
| Yes | 36 | 186 |

**Neural Network**

We first calculated the number of hidden layers/nodes and the decay parameters. After sizing our neural net model, tuning for an optimum number of intermediate hidden nodes, and avoiding overfitting, we tested it with 875 samples, 22 predictors, and the two classes of Yes/No. Our model.nn is a 90-1-1 network with 93 weights and decay of 0.1.

While developing the NN is rather simple, it is computationally intensive and requires more computing resources than many other models for classification. Since our data was very small with the number of observations being a little over 100, computational constraints were not our concerns. This model did yield a high accuracy rate of 82.44%. This is an improvement over recent logistic fittings which were noted at 0.8127.

Different kinds of errors like positive prediction, negative prediction and misclassification errors can be observed from the confusion matrix below.

| Confusion Matrix and Statistics | | |
|---|---|---|
| Prediction | No | Yes |
| No | 140 | 33 |
| Yes | 33 | 170 |

### Single Tree

We fitted a single tree model using an mtry of 2. This led to an OOB of 0.3824. Thus being our least accurate model measure at (1-OOB ) *100%= 61.76%

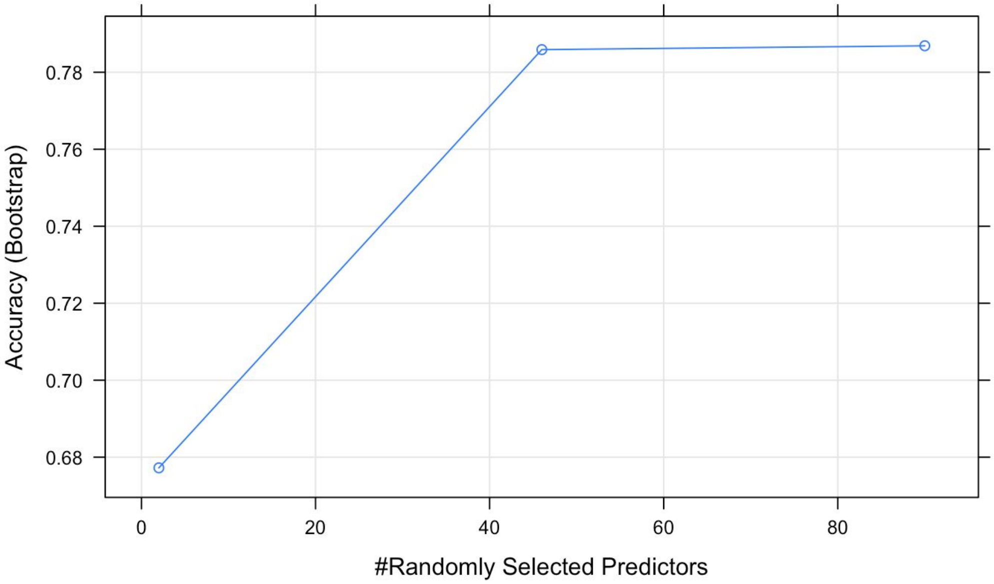
The advantages of using trees is that they are easy to explain, even more so than linear regression models. They tend to resemble or even mirror human decision making. Unfortunately, they don’t tend to have the same level of predictive accuracy as other regression and classification methods we used. It is no surprise that this has our lowest predictive performance level.

Confusion Matrix:

| | No | Yes | class.error |
|---|---|---|---|
| No | 151 | 11 | 0.06790123 |
| Yes | 111 | 46 | 0.70700637 |

### Random Forest

In order to improve the model built using a single tree, we used random forests. 22 Predictors were used in the Random Forest model. Further adjustment in our model has increased the accuracy level to 82.45% with a 95% confidence level for accuracy in the range of (0.7822, 0.8616). The optimal number of features is the sqrt of the number of predictors which in this case is between six to seven predictors.



Confusion Matrix:

| Prediction | No | Yes |
|---|---|---|
| No | 142 | 35 |
| Yes | 31 | 168 |

**Results : Learnings from the results and inferences**

In the data analysis phase of our research, we've gained insight into the factors affecting mental health at the workplace based on the results of the OSMI survey. The logistic regression analysis was based on several variables which had been transformed. Demographic variables did not exhibit statistically significant effects on the response variable as exhibited below in Appendix G. It is important note that into order to compare mental health statistics across groups, randomizations are necessary, While the preliminary EDA noted significantly higher frequency of treatment for female participants, in the overall model, gender did not stand out as a predictor. Rather our summary results showed that the most statistically significant factors were `Family_history, Work_interfere, Benefits, care_options.`

For an analysis of this nature, false positive results need to be weighed with the sensitivity of the factor analysis.

Some factors variables which would be interesting to include in the future would be data pertaining to salary and ethnicity. The former may provide insight into mental illness across socio-economic divisions, regardless of industry. The latter may also offer insight into mental health distresses affecting different cultural groups. These however could very well be statistically insignificant based on the performance of other demographic data.

**Conclusion:**

The objective of this project is to help tech managers and experts in the healthcare domain for decision making pertaining to the mental healthcare treatment. We can see from our predictive analysis that the frequency of mental illness varies by demographics by that this issue can be targeted and accurately classified based on core environmental factors such as the availability treatment benefits and of healthcare options provided by employers.

From this study we can infer that the availability of resources to tech employees and employer transparency regarding healthcare benefits highlight existing policies inclusive of mental health treatment. On the other hand, all three coefficients of the `work_interfere` factor levels (`Rarely`, `Sometimes`, and `often`) are of high significance. This is 3 times the factor of having a confirmed family history of mental illness and 4 and 6 times, respectively, for the availability of benefits and care options.

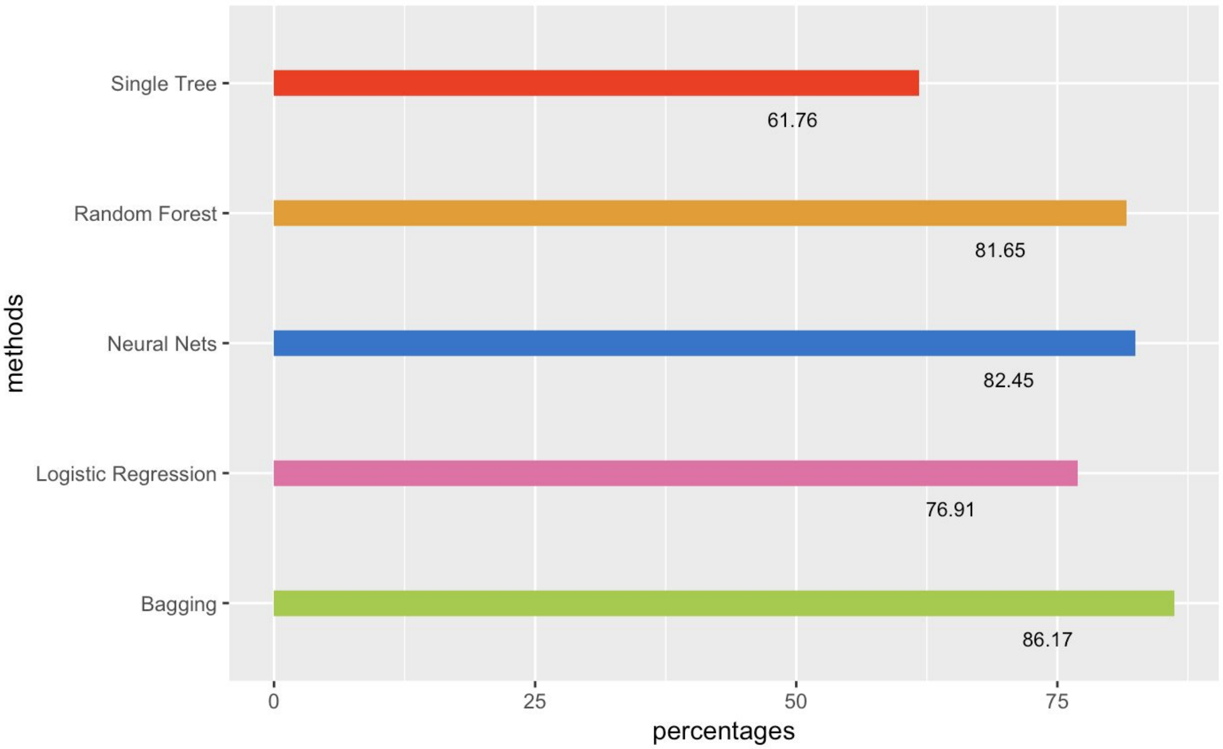
The significance of `work_interfere` shows that individuals are more likely to seek treatment if their mental health issues interfere with their day to day tasks at work. This clearly has more effect on an individual than them having a family history of mental illnesses. This encourages the need for further dialogue as these topics tend to surface only seasonally and following events of suicide.

Recommendations for future work:

- conduct a cost analysis for each predictor as this may be also weighed with a study pertaining to `work_interfere`
- Follow-up data collection on productivity levels may yield helpful insights into the value of treatment for company employees. Incorporating these models may also help treatment providers for sourcing observational data relevant to investing further infrastructure to support such treatment.
- Instead of binary inputs for issues like `work_interfere,` a categorical range, say from 1-5 with 1 being least affected by mental health at work to 5 being the max would yield better insights.
- Similarly, further data collection and privacy and anonymity fronts would give a deeper understanding of the stigma associated with mental health issues at workplace.
- Further follow up of the individuals that sought help with mental health issues and analyzing their career growth might give a good idea of the effectiveness of the current treatments. A positive outcome on this front would serve as a motivation to all those suffering from mental ailments and are hesitant to seek treatment.

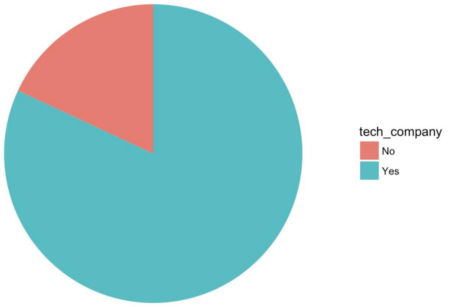
Success rates of predicting if an individual would seek treatment given the various factors:

| Models | Success rates |
|---|---|
| Logistic Regression | 76.91429 |
| Single Tree | 61.75549 |
| Random Forest | 81.64894 |
| Bagging | 86.17021 |
| Neural Nets | 82.44681 |

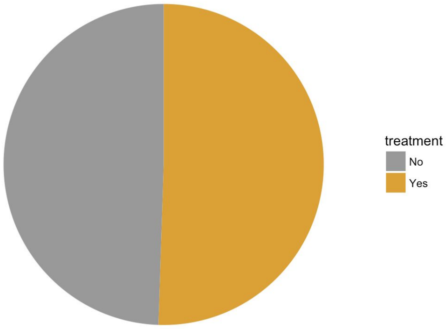


## Appendix:

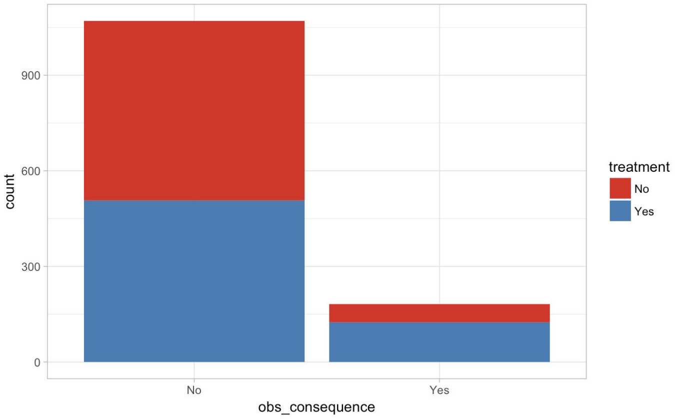
A:
Distribution of Data wrt Tech & Non-Tech Companies



## B. Seeking vs Not Seeking Mental Health Treatment



## C. Consequences of Workers Seeking Treatment



## D.

Significance Factors from Chi-Square tests

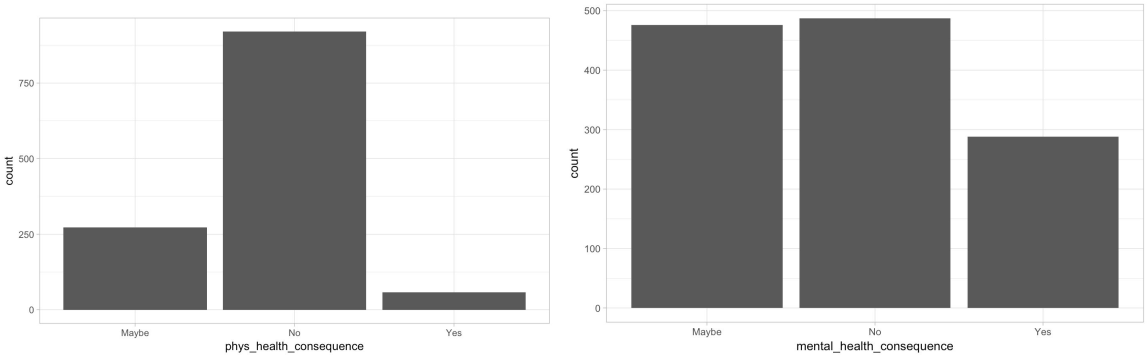
| | Df | Deviance | Resid. Df | Resid. Dev |
|---|---|---|---|---|
| NULL | | | 501 | 692.73 |
| Age | 1 | 0.154 | 500 | 692.58 |
| Country | 3 | 4.363 | 497 | 688.21 |
| state | 42 | 50.04 | 455 | 638.17 |
| self_employed | 1 | 0.226 | 454 | 637.95 |
| family_history | 1 | 61.552 | 453 | 576.39 |
| work_interfere | 3 | 224.572 | 450 | 351.82 |
| no_employees | 5 | 8.757 | 445 | 343.06 |
| remote_work | 1 | 0.128 | 444 | 342.94 |
| tech_company | 1 | 0.142 | 443 | 342.8 |
| benefits | 2 | 25.407 | 441 | 317.39 |
| care_options | 2 | 10.126 | 439 | 307.26 |
| care_options | 2 | 10.126 | 439 | 307.26 |
| wellness_program | 2 | 1.949 | 437 | 305.31 |
| seek_help | 2 | 6.374 | 435 | 298.94 |
| anonymity | 2 | 9.46 | 433 | 289.48 |
| leave | 4 | 1.178 | 429 | 288.3 |
| mental_health_consequence | 2 | 4.223 | 427 | 284.08 |
| phys_health_consequence | 2 | 1.625 | 425 | 282.45 |
| coworkers | 2 | 3.789 | 423 | 278.66 |
| supervisor | 2 | 1.24 | 421 | 277.42 |
| mental_health_interview | 2 | 1.628 | 419 | 275.8 |
| phys_health_interview | 2 | 1.622 | 417 | 274.17 |
| mental_vs_physical | 2 | 2.639 | 415 | 271.54 |
| obs_consequence | 1 | 0 | 414 | 271.53 |
| Gender | 2 | 3.813 | 412 | 267.72 |

ANOVA Fit3 for Logit Model

Analysis of Deviance Table

| | Df | Deviance Resid. | Df | Resid. Dev |
|---|---|---|---|---|
| family_history | 1 | 108.32 | 873 | 1104.43 |
| work_interfere | 3 | 343.65 | 870 | 760.79 |
| benefits | 2 | 29.86 | 868 | 730.93 |
| care_options | 2 | 13.41 | 866 | 717.51 |

E. Consequences of discussing Physical vs Mental Health with employer



F.

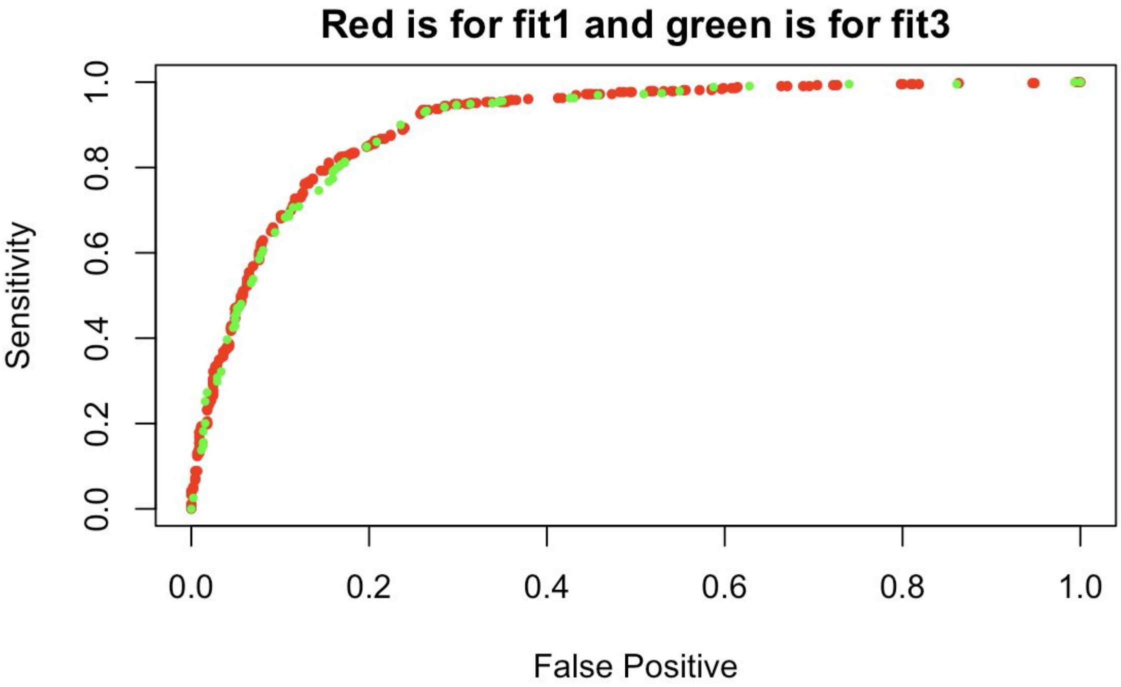
Resampling Results (for RF)

Across Tuning Parameters

| mtry | Accuracy | Kappa |
|---|---|---|
| 2 | 0.6772331 | 0.357714 |
| 46 | 0.7858835 | 0.5724892 |
| 90 | 0.7868806 | 0.5744342 |

H.

Fit 1 vs Fit3


Red is for fit1 and green is for fit3
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive