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). 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 provides 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.

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 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.

---

[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

Analysis Goals:

This analysis aims to develop a model to predict mental health treatment. Data was aggregated from various countries and states.

Predict the prevalence of employees seeking mental health treatment.

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?

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 represent 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). 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.

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.

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.We also use a random forest model 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:

- Timestamp
- Age
- Gender
- Country
- state: If you live in the United States, which state or territory do you live in?

Mental Health Condition:

- self_employed: Are you self-employed?
- 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?
- 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?
- 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?
- 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

There were multiple types of 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". To convert this into a categorical variable, we first had to normalize the `gender` data. For this we converted whatever was intended to be Male as Male and likewise for Female. For the values that we weren't sure of the category, we created a separate non-M/F category.

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) #Needs more explanation**

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.

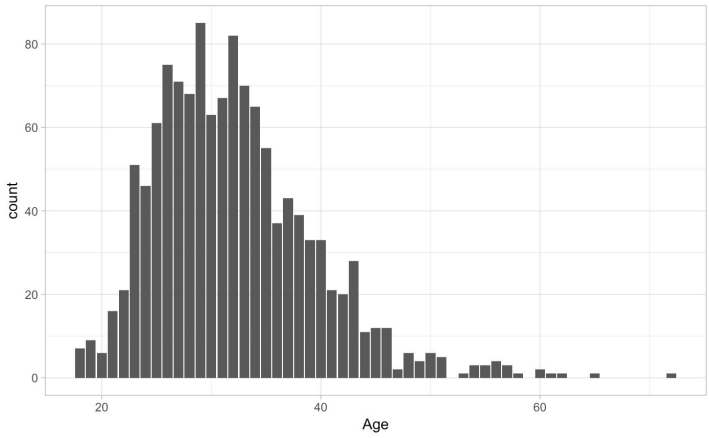
irrelevant columns etc

Data Exploration

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

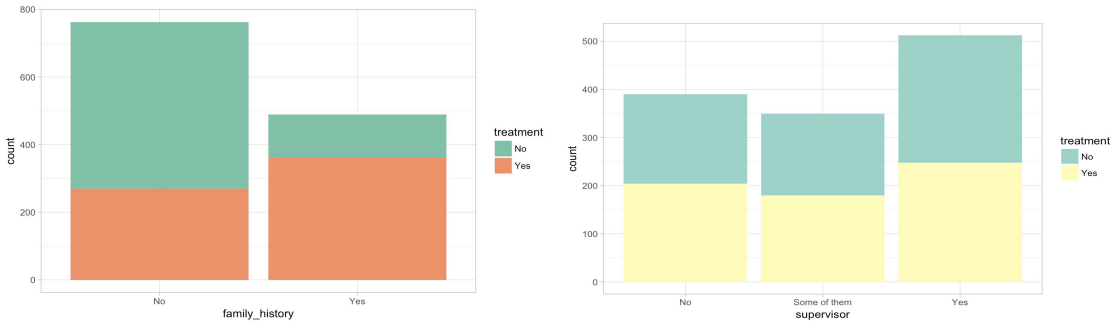
Age Distribution:
The tech industry is known for having a youthful work culture and unsurprisingly has a distribution skewed to the right. The Median age of 31 is less than the mean Mean is 32.07674. sd= 7.288272

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.



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. 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. Furthermore, the binary nature of yes and no questions limits the capacity to better qualify and the variation in mental health conditions. The chart apparent above 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.

The distribution of data with respect to tech and non-tech companies was clearly skewed in favor of tech companies. Only 39.1% of respondents noted having a family history of mental illness.



Above right, note the willingness of workers to discuss their mental health issues with their supervisor. We found that the anonymity of workers was not a statistically significant factor for affecting individuals seeking treatment.

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

This first model was built and used as a metric for testing the significance of each variable by doing a p-test. We decided that `anonymity`, `seek_help`, `state`, and `self_employed` were not very significant due to their p-values and hence decided not to use it in 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:

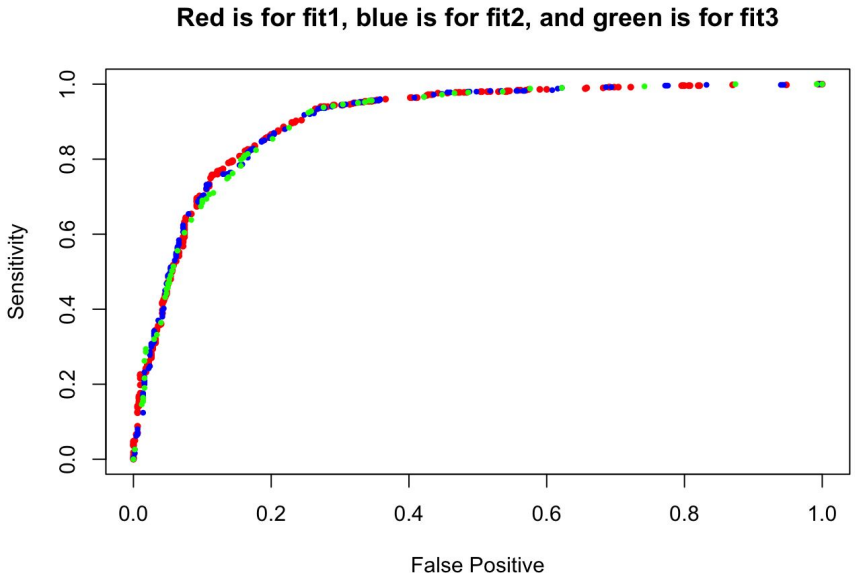
```
treatment =  -3.3585 + 1.0321 * family_historyYes + 4.1556 * work_interfereOften
             + 3.1157 * work_interfereRarely + 3.6161*work_interfereSometimes
             + -0.1551 *benefitsNo + 0.8816 *benefitsYes - 0.2461 *care_optionsNotsure
             +  0.5970 * care_optionsYes
```

Clearly `Family_history, Work_interfere, Benefits, care_options` correlate positively and

{note to remember vars to insert here }relate negatively with the final outcome which is to predict whether the worker would seek treatment.

The large number of categorical variables supported the use of a logistic reg

A classifier based on the ROC curve was built with the 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.



Bagging

In general, predictor for bootstrap aggregators (bagging) 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.

After fitting out bag model to our response variable, treatment, we trained the model to our training set. While also creating a matrix describing for each observation, the number of trees that assigned it to each class, which in this case is health.test. The resultant accuracy measured by our baggin predictor is 83.66% The highest level amongst all five of our models.

Neural Network

We first calculated the number of hidden layers/nodes and the decay parameters. By sizing our neural net model tuned an optimum number of intermediate hidden nodes and avoiding overfitting, out model contains 875 samples with 22 predictors and the two classes of Yes/N=o. Our model.nn is a 90-1-1 network with 93 weights and with an entropy fitting 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. This can be less than ideal, however, this model did yield a high accuracy rate of 82.44%. This is an improvement over recent NN fittings which were noted at 0.8127.

We can ascertain a variety of conclusions based on our models, for example the MCE, false positive, and true positive rates.

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:

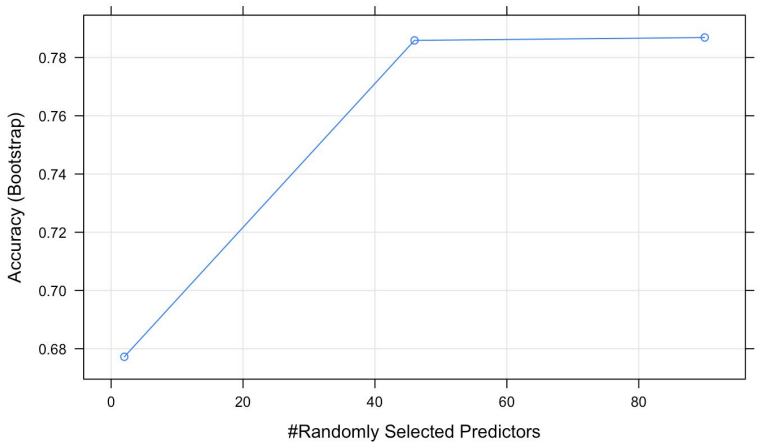
	No	Yes	class.error
No	151	11	0.06790123
Yes	111	46	0.70700637

Random Forest

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).

Confusion Matrix:

Prediction	No	Yes
No	142	35
Yes	31	168



Results : Learnings from the results and inferences

While adjustments to our models has yielded a higher accuracy rate for our models, namely the neural net, RF, and bagging models, this did come at a decrease of nearly 4%.

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 noted below in Appendix (D - {don't forget}). It is important note that into order to compare mental health statistics across groups, randomized is necessary, While a preliminary EDA notes a effect by gender, namely female, inclusion in the model overall notes 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

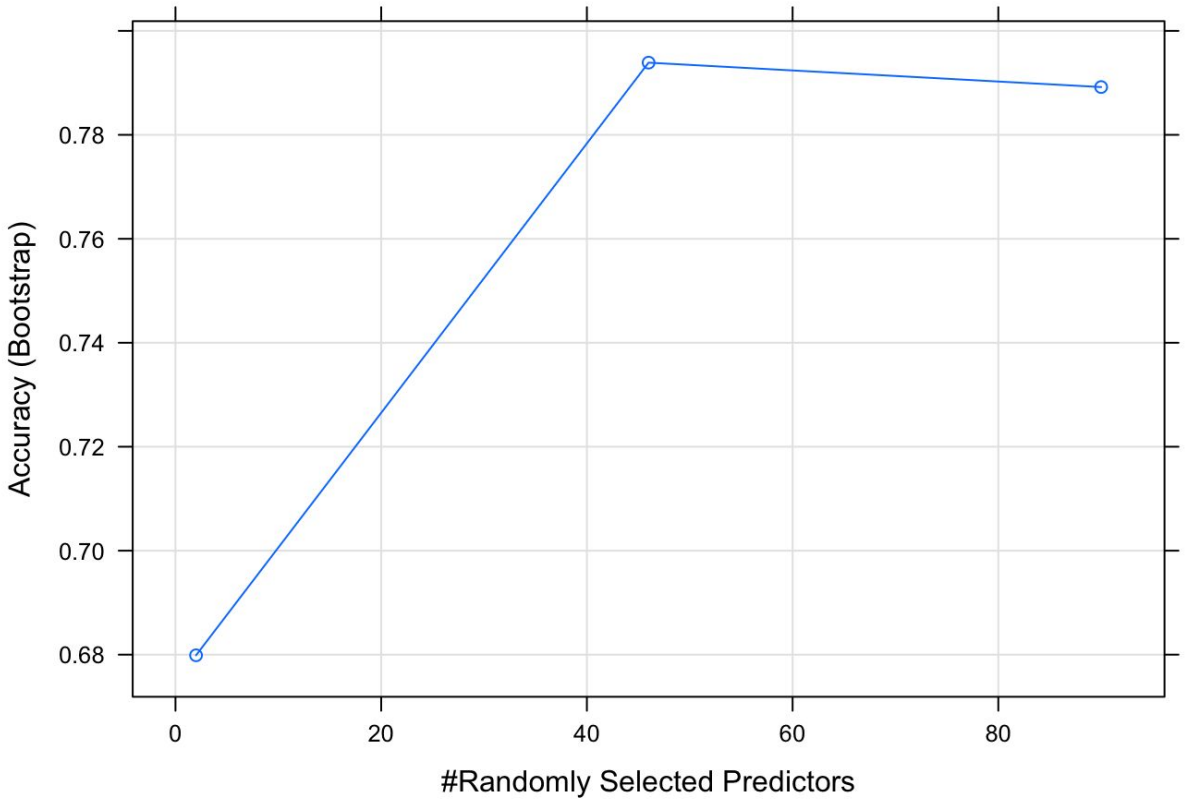
health distresses affecting different cultural groups. These however could very well be statistically insignificant based on the performance of other demographic data.

Compare test set MSE for RF and bagging.
Rank var importance.

EDA: Describe the graphs plotted, the inferences from them

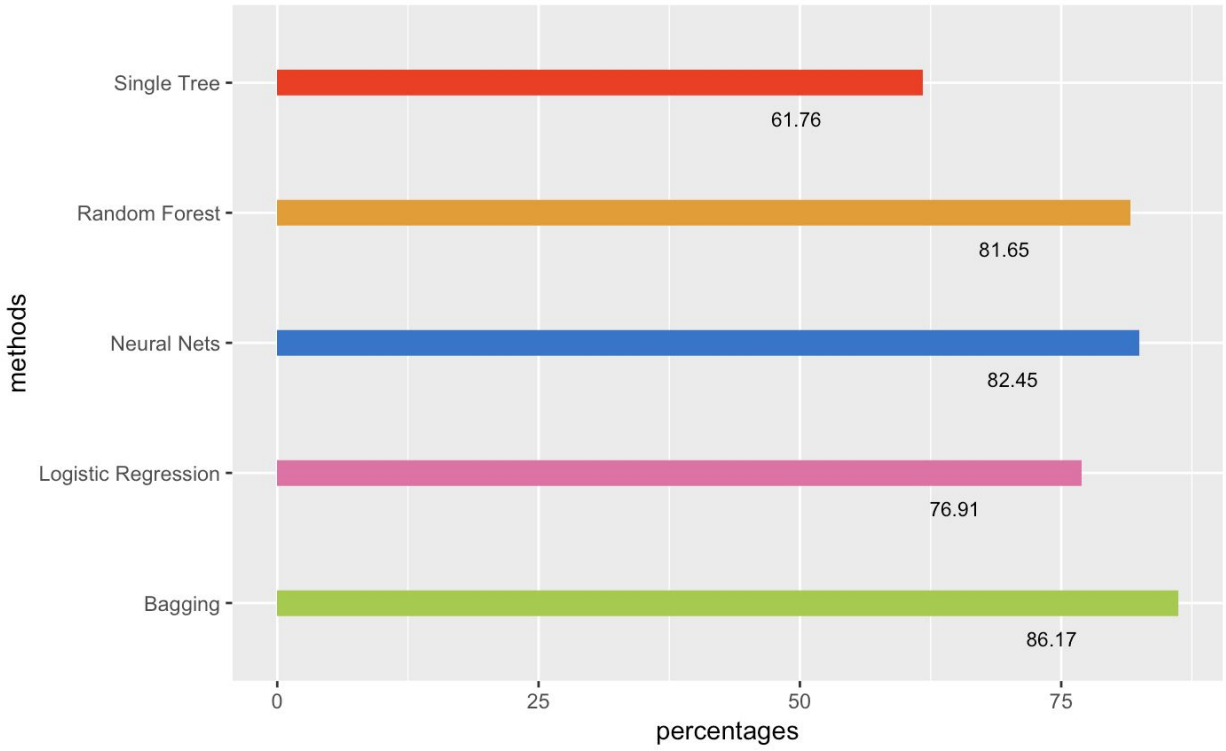
1. Model exploration: Test the significance based on p-values and decide to include or not the predictor

2. Model selection: describe random forests, decision trees, boosting, logistic regression and neural nets - describe the parameters in each of them



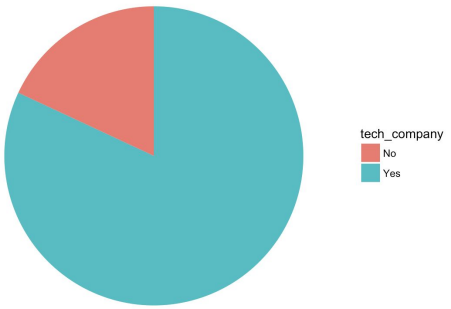
Success rates for different methods:

methods <fctr>	percentages <dbl>
Logistic Regression	76.91429
Single Tree	61.75549
Random Forest	81.64894
Bagging	86.17021
Neural Nets	82.44681

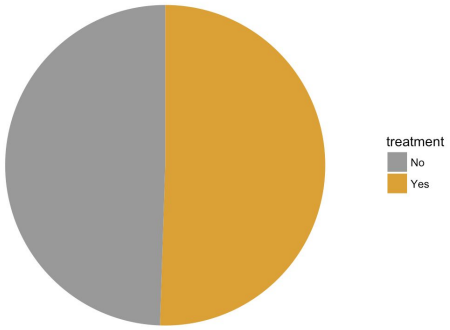


3. Future work - what other types of data could we collect that would lead us to better understanding of the mental health scenario in the tech industry

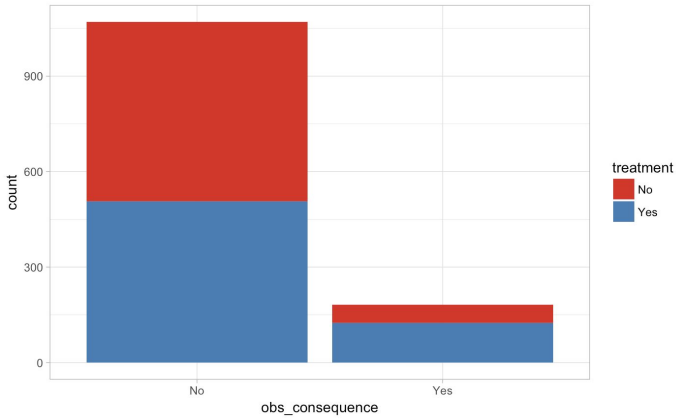
Appendix:
A:
Distribution of Data wrt Tech & Non-Tech Companies



B. Seeking vs Not Seeking Mental Health Treatment



C. Consequences of Workers Seeking Treatment



D. Analysis of Deviance Table

E. Random Forest

mtry	Accuracy	Kappa

2	0.6772331	0.357714
46	0.7858835	0.5724892
90	0.7868806	0.5744342