Understanding Mental Health in Tech Workplaces

Aditi Jayashankar, Eddie Kong, Sahana Vijaya Prasad

STAT 571: Modern Data Mining
The Wharton School
Dr. Linda Zhao

Table of Contents

Executive Summary

According to the National Institute of Mental Health, one in four adults in America experiences mental illness. These can include anxiety, schizophrenia, depression, bipolar disorder, and others. Yet, approximately 60% of adults do not seek treatment. Mental illness affects our work and personal lives. Serious mental illness costs America almost $200,000,000,000 in lost earnings per year. With the internet boom, and the growing tech world, we are interested in gauging how mental health is viewed within the tech/IT workplace, and the prevalence of certain mental health disorders within the tech industry. As Computer Engineers who will be joining this tech world workforce next year, our goal was to answer the following questions:

1. What workplace factors contribute to getting diagnosed with mental health problems?
2. What is the best model to predict whether someone will be diagnosed with mental health problems?

Using the dataset from the Open Source Mental Illness (OSMI) Mental Health in Tech survey, conducted in 2016, we cleaned and engineered features we felt were relevant to our study. A thorough exploratory data analysis revealed that 56% of the responders have been diagnosed with a mental health issue. Building a LASSO model with Logistic Regression, we were able to determine the responsible work factors that affect an employees probability to get diagnosed with a mental health issue were related to options for mental health care provided by the employer and willingness to share mental health issues openly without being judged.

For choosing the best model to predict whether someone will be diagnosed with mental health problems, we tried Logistic Regression from the LASSO model, Random Forest classifiers, and 2-Layer Neural Networks. Our final model is the Logistic Regression model that has an AUC of 0.75 and MCE of 0.31. Other models like Support Vector Machines and Boosting could be explored but are not used in this study.

The findings of this study can be used as motivation for tech companies to promote awareness of mental health, include more options in the healthcare packages, and encourage employees to be more open to each other about any issues. The more we acknowledge the role of mental health in our industry, the more we can help people get access to better healthcare, retain and support underrepresented employees, and build an authentic culture of inclusion.

About the Data

Description

This data was obtained from the Open Source Mental Illness (OSMI) survey conducted in 2016 which primarily aims to examine the prevalence and attitudes towards mental health among tech workers. The data is freely available from Kaggle [1] and contains the response of 1,433 participants to a 63 question survey. The survey responses include participant demographics, work background, attitudes towards mental health issues, types of diagnoses, and whether the patient was diagnosed with a mental health problem by a medical professional. Our goal is to use the results from the survey and identify which response variables are key in predicting a medically diagnosed mental health condition.

Data Cleaning

Our process for pre-processing and cleaning the data was extensive. We decided focus our investigation on only the respondents who work and live in the U.S. (roughly over 60% of all participants). In addition, several questions were either open ended that over 90% of respondents chose not to answer and left blank (why or why not questions). This is irrelevant for the purposes of this study (such as country or foreign province) or for legal reasons could be answered only by a small subset of participants and left blank (details about working with previous employers). These were excluded from our investigation.

We also restructured some messy variables. All variables were renamed for the purpose of clarity. Gender contained various ambiguous or missing responses such as “Unicorn”, “none of your business” or blanks and were binned into “Other” category. For job roles, participants had the option of selecting various occupations so we only chose their primary role in our analysis. Similarly, self-reported diagnosis with more than one response, the primary one was taken into consideration. Rows with invalid responses (eg. 323 years for Age) were removed. Finally, our binary classification variable “medical_diagnosis” was encoded as 0 for No and 1 for Yes. This indicates whether the person was medically diagnosed with a mental health disorder, or not.

Final Dataset

These are the 35 final variables used in our analysis, and the question associated with each of the variables:

- **self_employed:** Are you self-employed?
- **no_employees:** How many employees does your company or organization have?
- **benefits**: Does your employer provide mental health benefits as part of healthcare coverage?
- **care_options**: Do you know the options for mental health care available under your employer-provided coverage?
- **wellness_program**: Has your employer ever formally discussed mental health (for example, as part of a wellness campaign or other official communication)?
- **seek_help**: Does your employer offer resources to learn more about mental health concerns and options for seeking help?
- **anonymity**: Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources provided by your employer?
- **leave**: If a mental health issue prompted you to request a medical leave from work, asking for that leave would be:
- **mental_health_consequence**: Do you think that discussing a mental health disorder with your employer would have negative consequences?
- **physical_health_consequence**: Do you think that discussing a physical health issue with your employer would have negative consequences?
- **coworkers**: Would you feel comfortable discussing a mental health disorder with your coworkers?
- **supervisor**: Would you feel comfortable discussing a mental health disorder with your direct supervisor(s)?
- **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 co-workers who have been open about mental health issues in your workplace?
- **previous_employers**: Do you have previous employers?
- **physical_health_interview**: Would you be willing to bring up a physical health issue with a potential employer in an interview?
- **mental_health_interview**: Would you bring up a mental health issue with a potential employer in an interview?
- **hurt_career**: Do you feel that being identified as a person with a mental health issue would hurt your career?
- **negative_coworker**: Do you think that team members/co-workers would view you more negatively if they knew you suffered from a mental health issue?

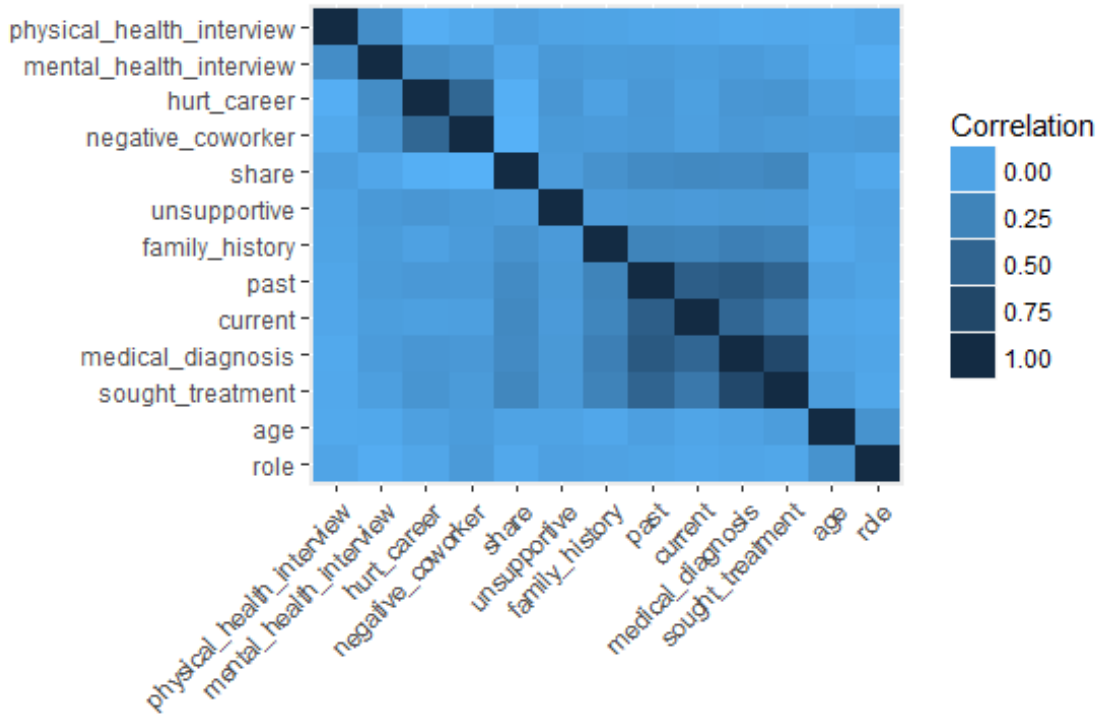
- **share**: How willing would you be to share with friends and family that you have a mental illness?
- **unsupportive**: Have you observed or experienced an unsupportive or badly handled response to a mental health issue in your current or previous workplace?
- **observation**: Have your observations of how another individual who discussed a mental health disorder made you less likely to reveal a mental health issue yourself in your current workplace?
- **family_history**: Do you have a family history of mental illness?
- **past**: Have you had a mental health disorder in the past?
- **current**: Do you currently have a mental health disorder?
- **yes_diagnosis**: If yes, what condition(s) have you been diagnosed with?
- **maybe_diagnosis**: If maybe, what condition(s) do you believe you have?
- **diagnosis_result**: If so, what condition(s) were you diagnosed with?
- **sought_treatment**: Have you ever sought treatment for a mental health issue from a mental health professional?
- **age**: What is your age?
- **gender**: What is your gender?
- **state**: What US state or territory do you work in?
- **role**: Which of the following best describes your work position?
- **remote**: Do you work remotely?
- **medical_diagnosis**: Have you been diagnosed with a mental health condition by a medical professional?

Exploratory Data Analysis

In order to understand the variables, we performed exploratory analysis of the data.

Correlation among Variables

If there is correlation between the variables, the model developed would not be of much importance. Hence, we first plot the correlation heatmap to determine the correlation between the variables. We choose a subset of variables for which correlation made sense. The heatmap is as shown.

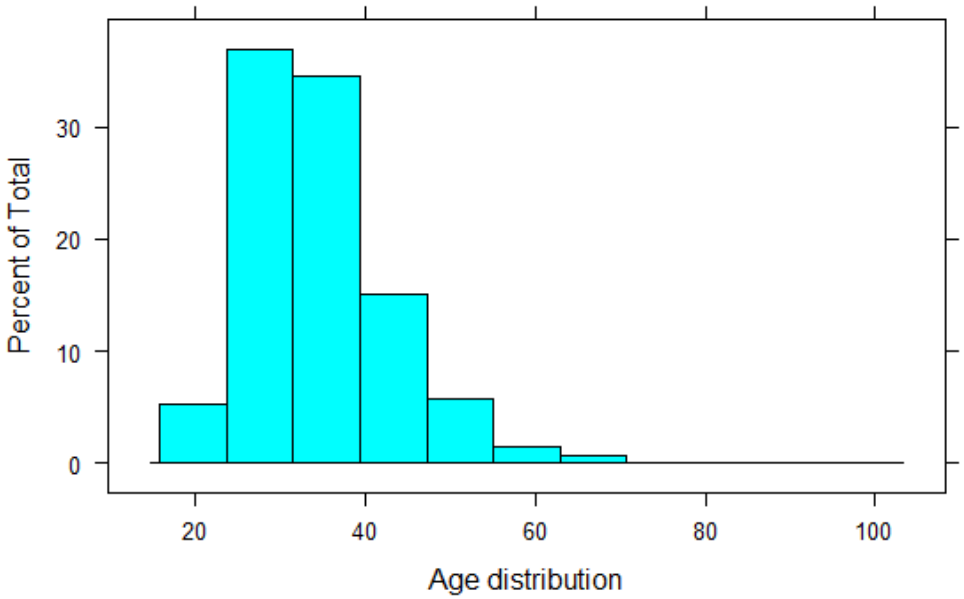


We make the following observations from the HeatMap:

1. People who feel that being identified as a person with a mental health issue would hurt your career also think that team members/co-workers would view you more negatively if they knew you suffered from a mental health issue.
2. People who had a mental health disorder in the past have a high correlation of having a mental health disorder in the present.
3. Being diagnosed with a mental health condition is more correlated to having a mental health disorder in the past than current.
4. There is high correlation between being diagnosed with a mental health condition and seeking treatment, which is trivial! We remove this variable from all our models.

Age Distribution

We then determine the age distribution of the sample.

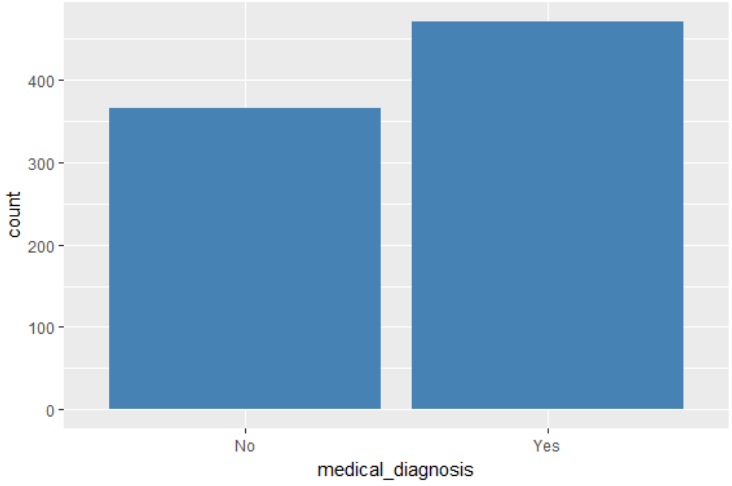


Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
19.0	29.0	33.0	34.6	39.0	99.0

Most of our participants fall within the 25 - 40 yrs of age group, having a mean of 34.6 years. This is typical of the current tech world [2].

Medically Diagnosed

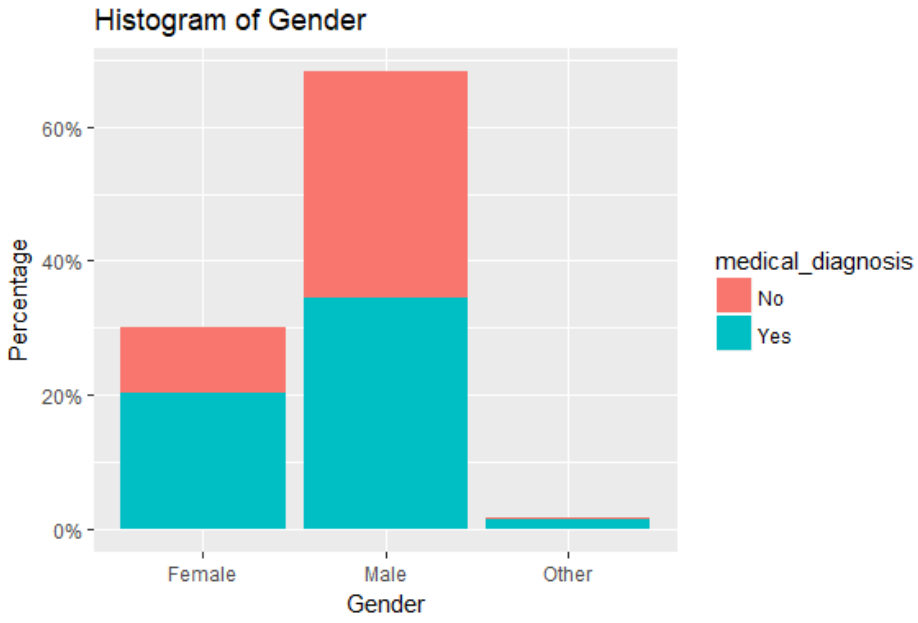
Out of the 836 number of participants, 471 number of them are medically diagnosed as having a mental problem.



It is extremely alarming to know that nearly 56% of people who work in tech have been medically diagnosed as having a mental health problem. Hence, it is worth exploring if the current climate of work culture in tech companies could have something to do with that.

Gender Distribution

We examine the gender distribution, and plot it against percentage who are medically diagnosed as having mental health disorder.

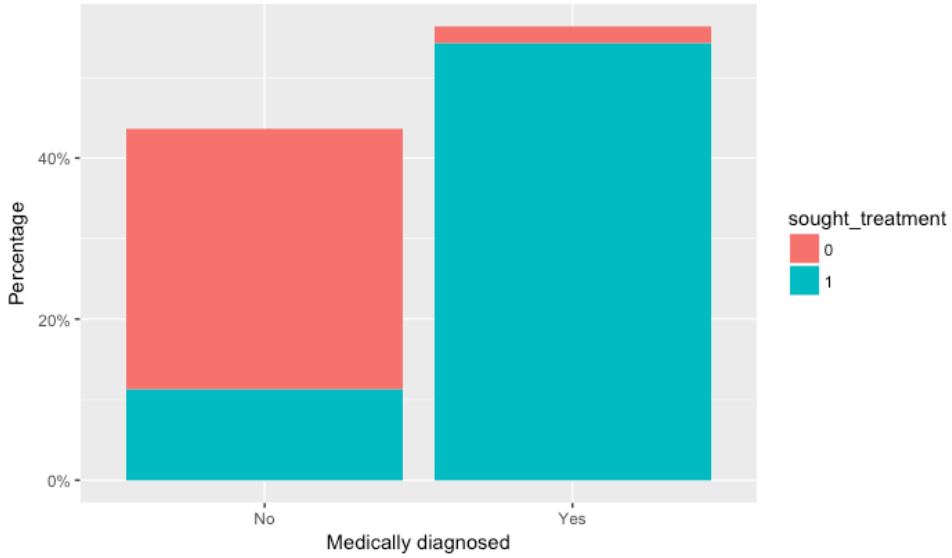


From the graph, we can make the following observations:

1. In our dataset, Males form the majority, around 68%, followed by Females at 30% and Others at 2%. This is extremely indicative of the present working condition of Tech in US. [3]
2. People categorized as "Other" have the highest percentage of people who are medically diagnosed as having mental health disorder.
3. Females have a higher percentage of people who are medically diagnosed with mental health problems, than Males.

Distribution of people who sought Medical treatment

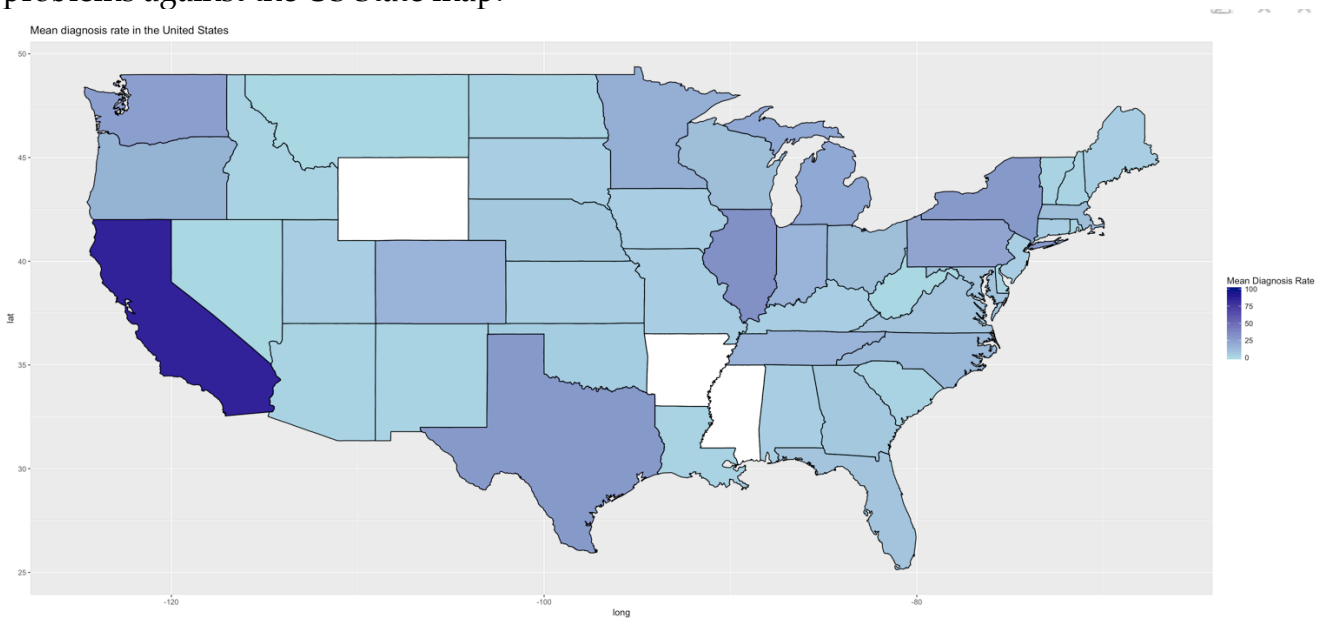
We analyzed the distribution of people who sought medical treatment, and plotted it against the percentages of being medically diagnosed as having mental health problems.



From the graph, we can observe that not all those who were medically diagnosed as having mental health problem sought treatment. Some people sought treatment even though they were not medically diagnosed as having a mental problem.

Distribution according to State

Finally, we plot the people who have been medically diagnosed as having mental health problems against the US State map.



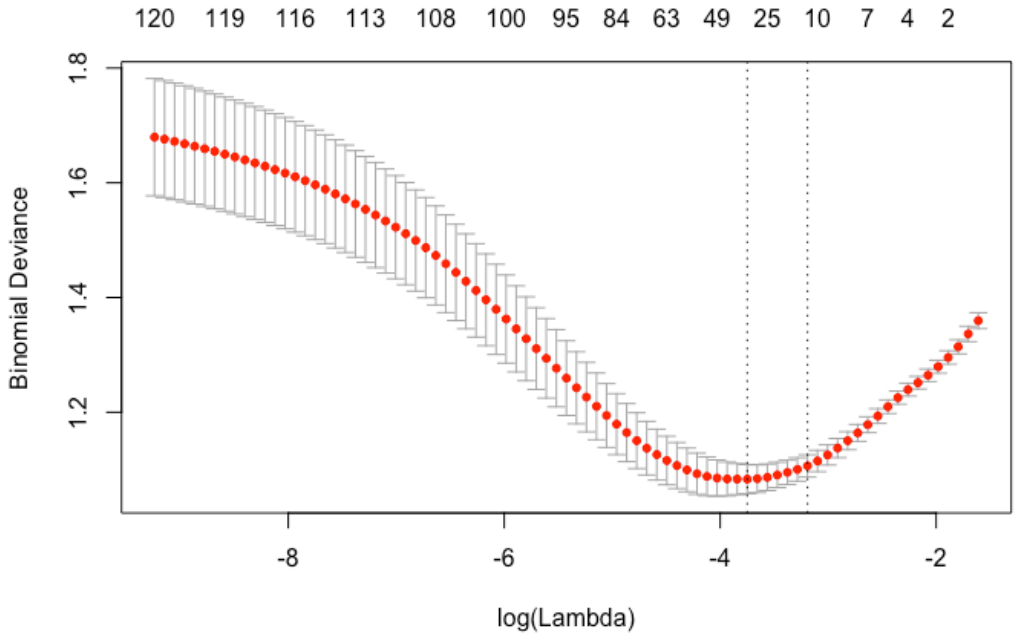
From the map, we can see that the State of California has the highest number. This is not surprising given that most of the tech companies are situated in the "Silicon Valley".

Final Model: Logistic Regression with LASSO

Since our goal is predicting a diagnosis for a mental health problem, we took out variables from the cleaned data that were about different types of diagnoses, since correlation would affect the model. We were left with 29 predictors. The dataset was split in train and test subsets by the ratio 2:1 respectively. The test set was only used for prediction and calculating errors.

LASSO

Since 29 predictors are a lot, we use LASSO to identify the contributing predictors in a binary logistic regression classification model. Setting alpha = 0.99, we ran the LASSO model with cross-validation.



Choosing lambda.1se = 0.0411128 as our lambda value, the non-zero variables are:

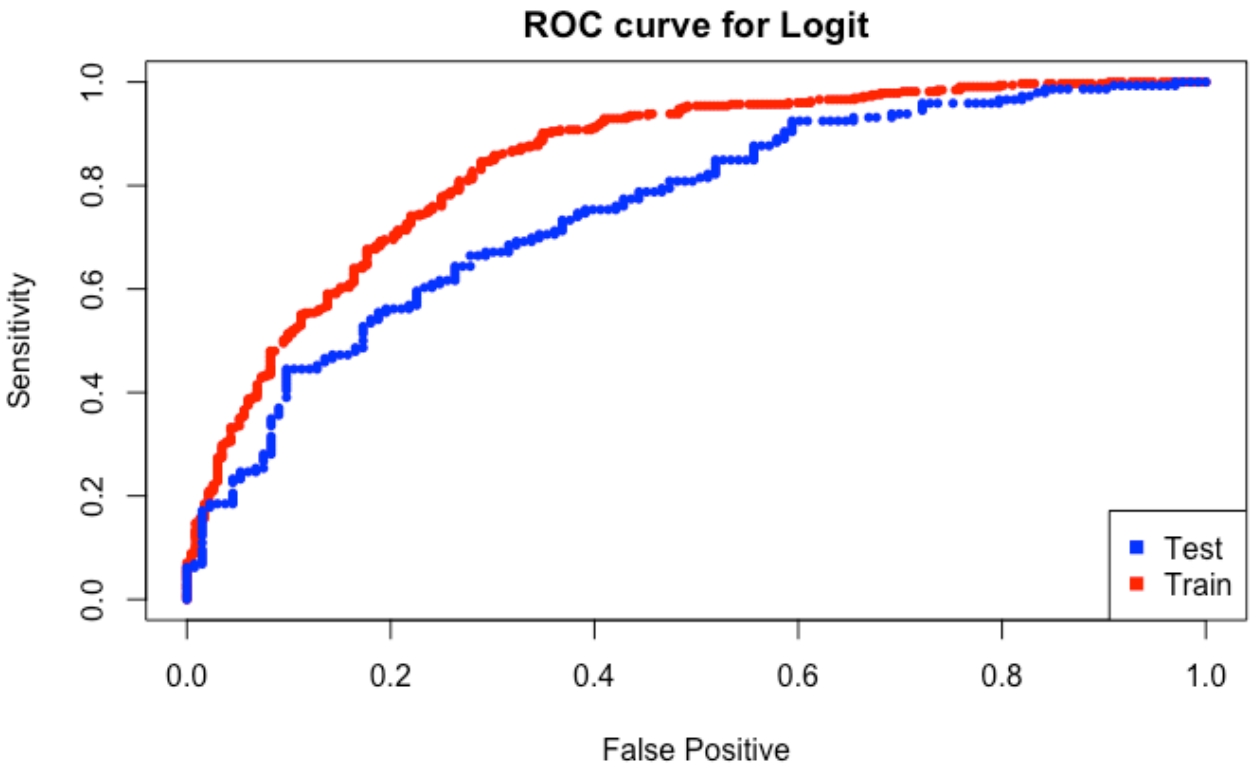
```
 [1] "(Intercept)"
 [2] "care_optionsNo"
 [3] "care_optionsYes"
 [4] "mental_health_interviewNo"
 [5] "negative_coworkerNo, they do not"
 [6] "shareNot applicable to me (I do not have a mental illness)"
 [7] "shareVery open"
 [8] "unsupportiveYes, I experienced"
 [9] "family_historyNo"
[10] "family_historyYes"
[11] "genderMale"
```

Logistic Regression

From the results of the LASSO model, the predictors used in the logit model are:

- care_options
- mental_health_interview
- negative_coworker
- share
- unsupportive
- family_history
- gender

The decision boundary for classification was 0.5. With the train sample, the AUC was 0.8433 and MCE was 0.213. With the test sample, the AUC was 0.7534 and MCE was 0.318. The ROC is shown below.

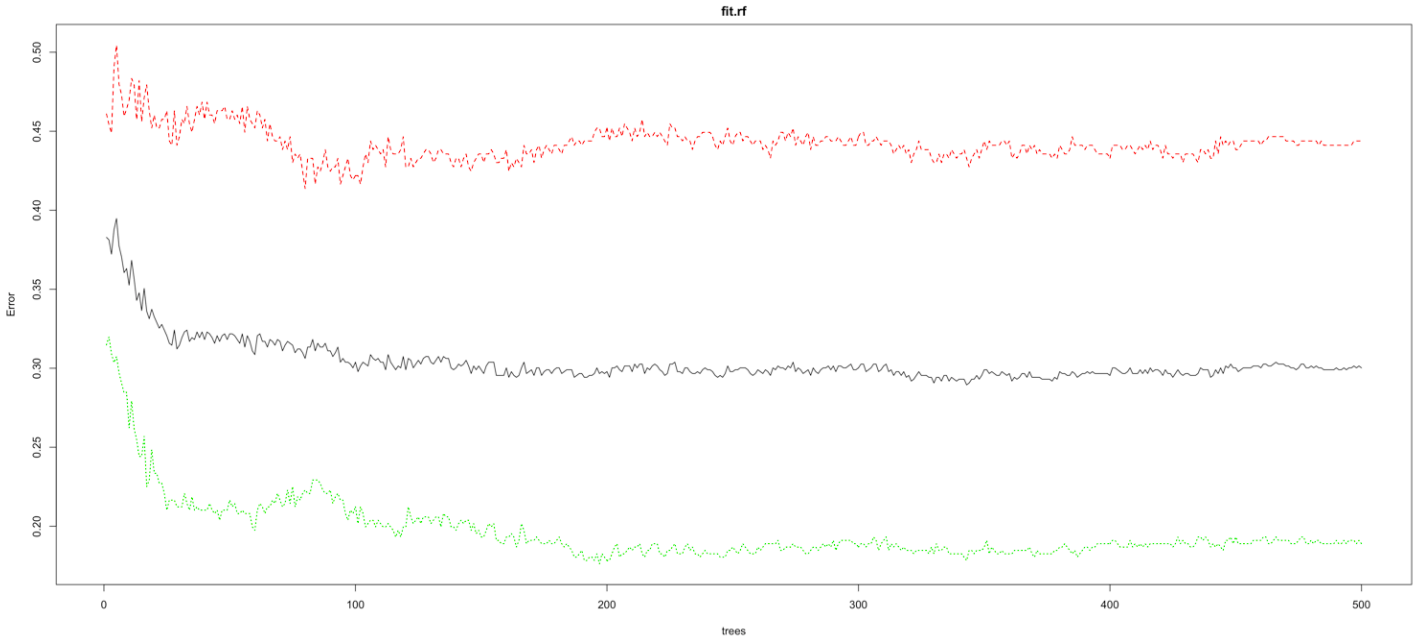


Reasons for choosing this model as our final one are the lower MCE and almost similar AUC when compared to the Random Forest model. Alternative methods of our study are in the subsequent sections of this report.

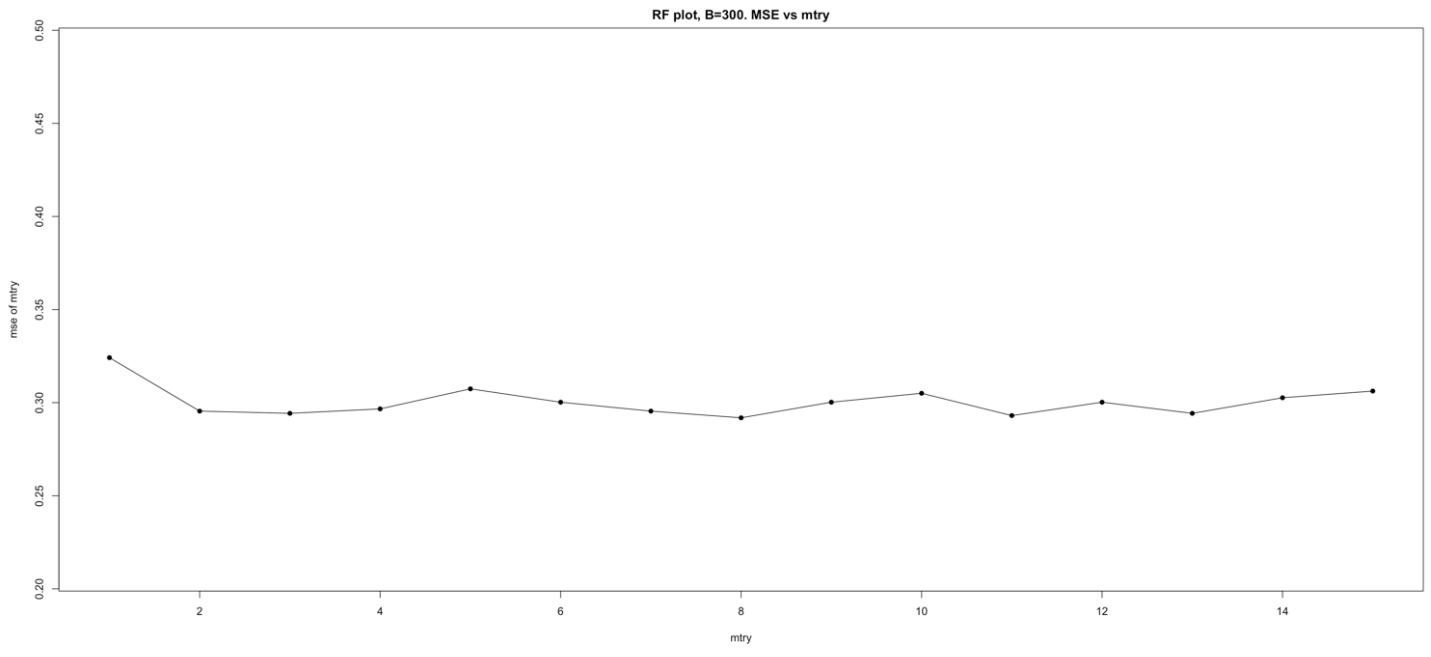
Alternative Methods

Random Forest Classifier

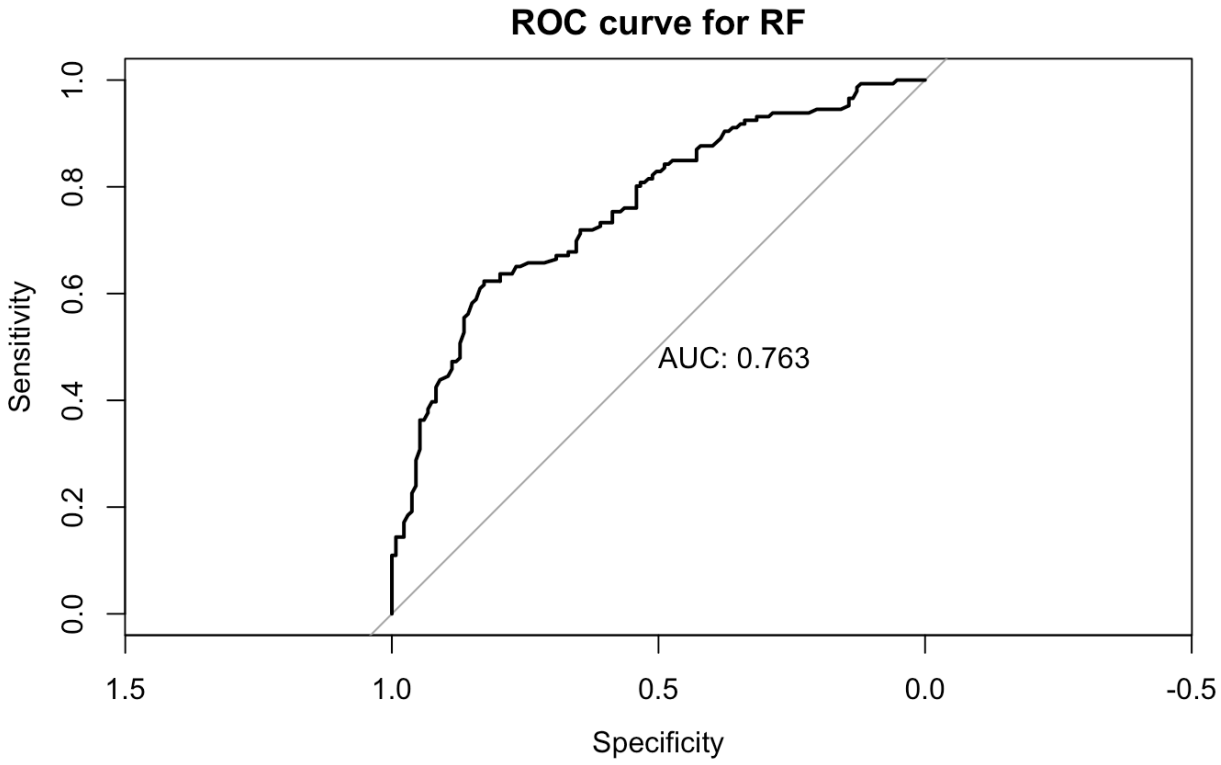
We decided to explore using Random Forests as a model. Decision trees are a very popular model in the medical field because they are easy to interpret and are generally quite useful in diagnosis decision support systems. We first proceed to tune the number of bootstrap samples B. Using a fixed mtry (the number of random variables we sample from when building a tree), we plot the error as a function of B and determine we require about 300 trees to settle out-of-bag testing errors.



We then tune our mtry, using B = 300 and plot the error vs mtry. We determine that the mtry that minimizes the error is 8.



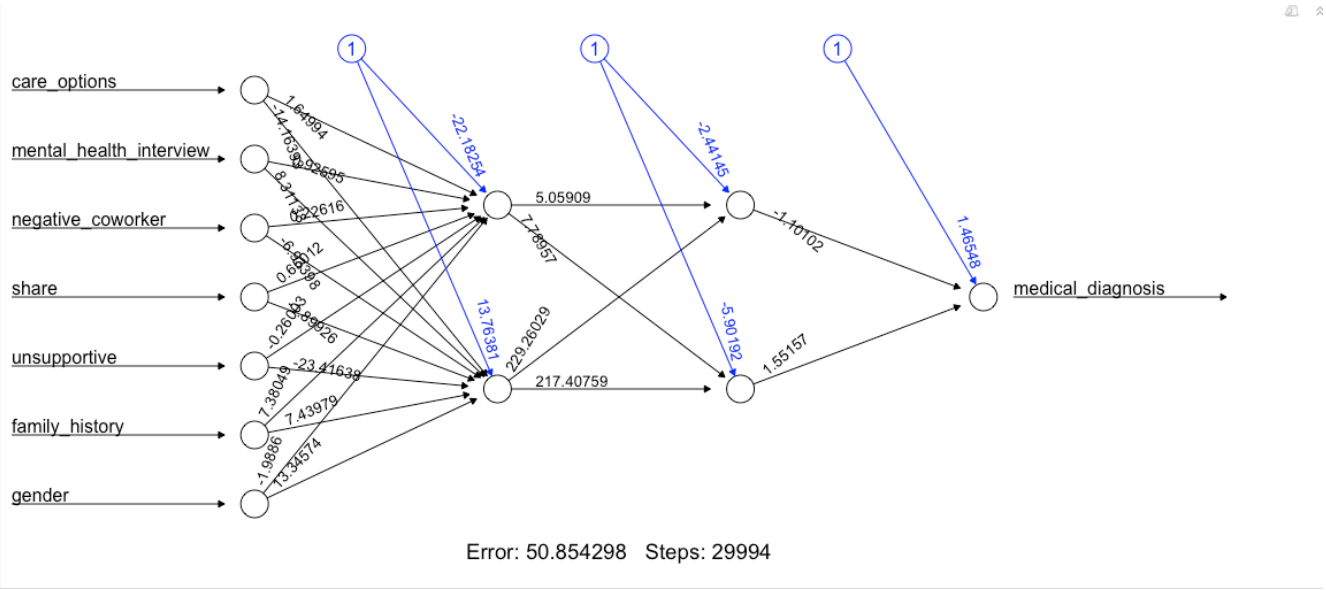
The MCE on testing data for this model is 0.3297491. The AUC for the ROC curve is 0.7633.



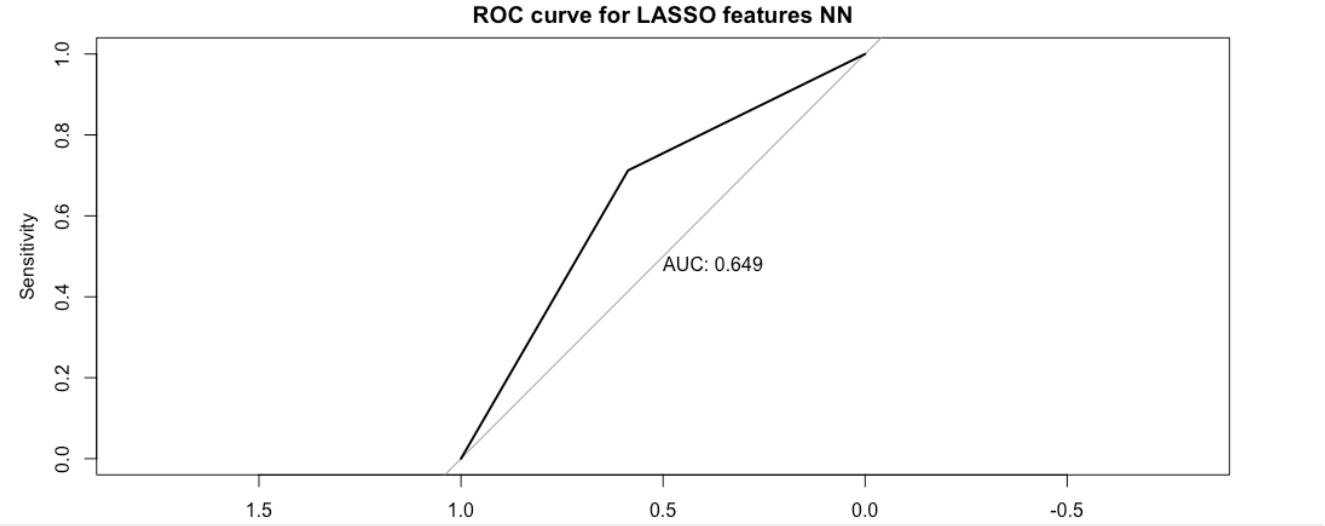
Neural Networks

For the purposes of using neural networks, all factors were turned into numerical inputs. We tested two different neural net architectures, 2 layers with 5 neurons each and 10 layers with 5 neurons each. Using all variables, we get a net that converges with highly predictive power. We found that the MCE for the 5 neuron layers was lower. For the 5 neuron model, the MCE was 0.010752688 and the AUC for the roc curve was 0.9890565.

However, if we remove the variables that are highly correlated to the response such as yes_diagnosis or maybe_diagnosis then we had difficulty achieving a neural net architecture that converges. We therefore used the features we selected from LASSO above (care_options + mental_health_interview + negative_coworker + share + unsupportive + family_history + gender). We found that a simpler architecture of 2 layers with 2 nodes was the only one that was both computationally efficient and could still converge.



The MCE for this model was 0.3476702 and the AUC for the ROC curve was 0.6493975.



Conclusion

Results

From our final LASSO/Logit model, the key predictors in leading to mental health issues while working in the tech industry are:

- The options for mental health care available under your employer-provided coverage
- Bringing up a mental health issue with a potential employer
- Thinking that team members/co-workers would view you more negatively if they knew you suffered from a mental health issue
- Willingness to share with friends and family that you have a mental illness
- Having observed or experienced an unsupportive or badly handled response to a mental health issue in your current or previous workplace
- Family history of mental illness
- Gender

This shows that there is room for improvement on how tech companies treat their employees. There is scope for improvement in awareness and care of mental health issues in the workplace.

Furthermore, based on these features, we can use the model to predict whether an employee is susceptible to a mental health diagnosis.

Suggestions & Limitations

The dataset and the survey questions are centric towards the work culture. However, it would be helpful to our analysis if there was some supporting data available. For example, demographic information like race and annual income could be used. Work habits would also be useful.

As for model building, other methods like Support Vector Machines and Boosting could be tested out. Another interesting question to ask from this dataset is, having being diagnosed with a disorder, what is the likelihood that the employee would seek out treatment? This could be a deeper dive into what companies can do to assure the good mental health of their employees.

References

[1] https://www.kaggle.com/osmi/mental-health-in-tech-2016

[2] http://www.businessinsider.com/median-tech-employee-age-chart-2017-8

[3] https://www.forbes.com/sites/christinawallace/2016/10/20/girls-in-coding-the-problem-is-getting-worse/