Analysis of Mental Health in the Tech Industry

Evan Downs, Madhumitha Ganesan, Sivapriya Vellaichamy, and Ashwin Nallan

04/11/21

Introduction

Project Motivation

- Mental health is an increasingly important concern in today’s world.
- 16.2 million adults in the United States, or 6.7 percent of American adults, have had at least one major depressive episode each year [1]
- Depression tends to affect people in their prime working years and is one of the top 3 workplace problems.

[1] https://www.healthline.com/health/depression/facts-statistics-infographic

Project Description

- The purpose of our project is to generate a profile for adults most prone to depression by analysing data on mental health disorders in the IT workplace
- This analysis will help understand the risk factors that contribute to mental health disorders.
- It can also help guide company policies regarding mental health resources to protect their employees in the workplace.

Scientific Research Questions

- Investigate the impact of geographical location on the amount or the probability of seeking treatment for mental illnesses.
- Determine the effect of a company's mental health coverage policy on the amount of employees seeking treatment
- Investigate the effect of negative workplace sentiment on the number of people seeking treatment

Scientific Research Questions

- Determine whether a person with a history of mental illness is more prone to facing another mental health issue.
- Determine if an individual’s past employer’s views on mental health is correlated with an individual’s present employer’s positions.
- Determine if an employee’s gender influences whether they feel that mental health concerns can be brought up in their workplace.
- Determine factors that are significantly determine the probability of people seeking treatment for mental illness.

Data Collection

Dataset

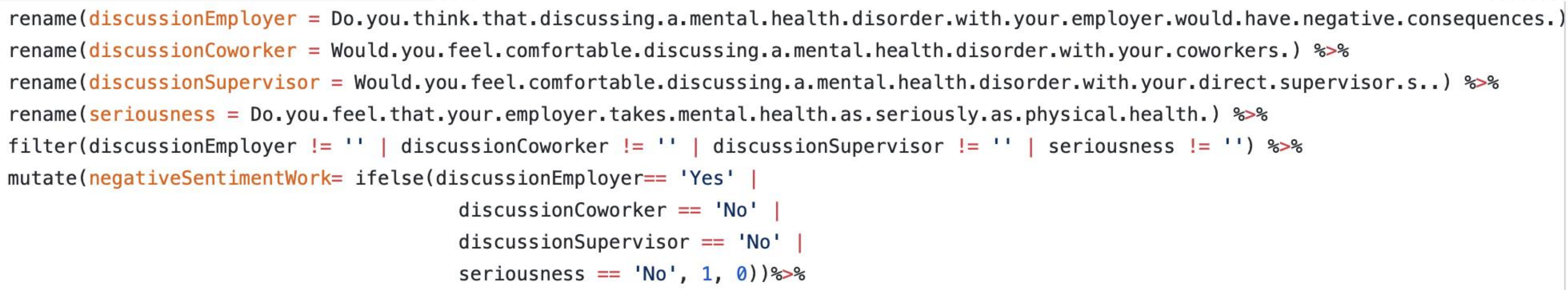
- The dataset for our study has been obtained from https://www.kaggle.com/osmi/mental-health-in-tech-2016.

- This dataset has 1433 responses from a survey to measure attitudes towards mental health in the IT/tech workplace and examine the frequency of mental health disorders among tech workers. The dataset contains 63 columns, including categorical and text-based columns.

- Data cleaning and feature engineering has been done to select specific columns which are relevant and largely contain non-empty values

Data Cleaning – Negative Sentiment

- The data pre-processing techniques vary between which research question is addressed
- Engineered feature – negative workplace sentiment
  - To reduce the complexity of workplace views on mental health 4 columns are reduced to 1
  - Ensuring that this feature does not have NA values requires omitting 287 rows of data

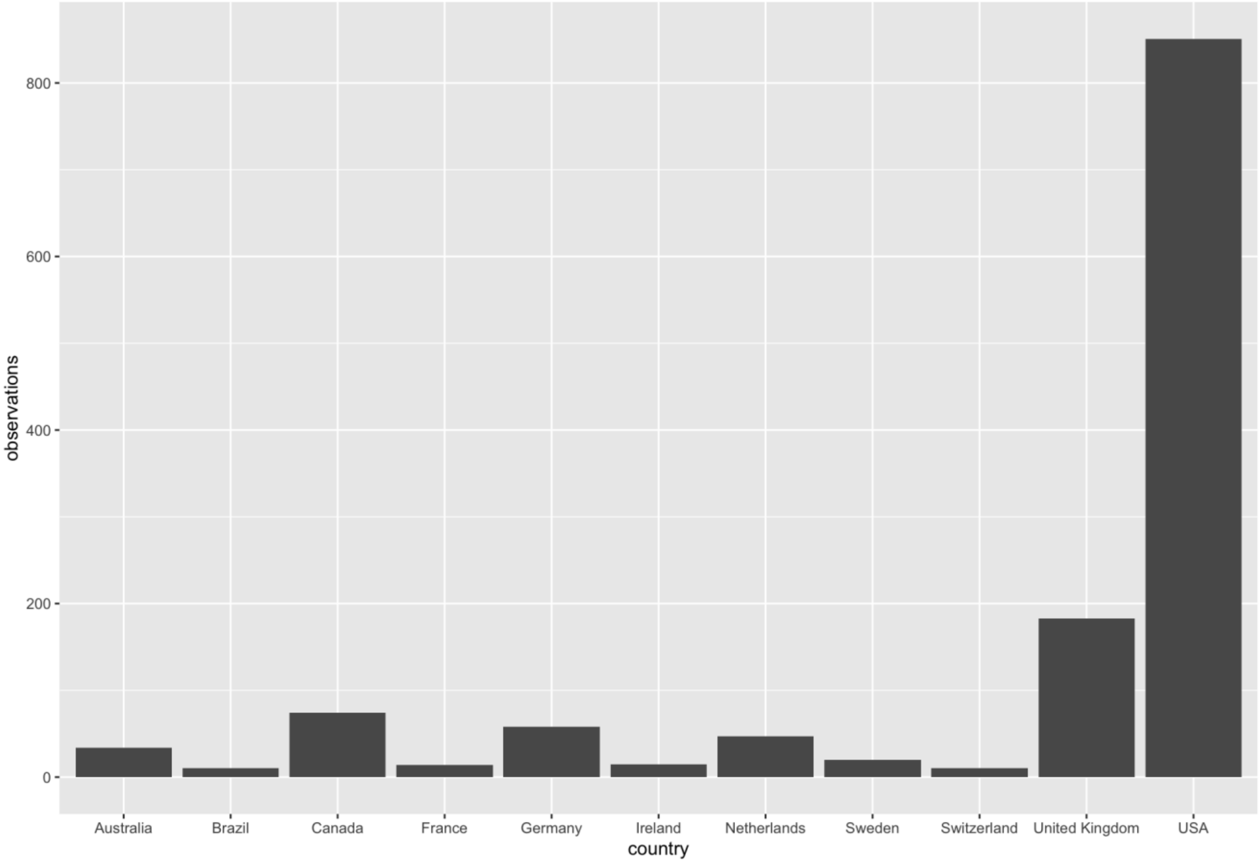


```
rename(discussionEmployer = Do.you.think.that.discussing.a.mental.health.disorder.with.your.employer.would.have.negative.consequences.)
rename(discussionCoworker = Would.you.feel.comfortable.discussing.a.mental.health.disorder.with.your.coworkers.) %>%
rename(discussionSupervisor = Would.you.feel.comfortable.discussing.a.mental.health.disorder.with.your.direct.supervisor.s..) %>%
rename(seriousness = Do.you.feel.that.your.employer.takes.mental.health.as.seriously.as.physical.health.) %>%
filter(discussionEmployer != '' | discussionCoworker != '' | discussionSupervisor != '' | seriousness != '') %>%
mutate(negativeSentimentWork= ifelse(discussionEmployer== 'Yes' |
                                       discussionCoworker == 'No' |
                                       discussionSupervisor == 'No' |
                                       seriousness == 'No', 1, 0))%>%
```

R code snippet shown
for implementation

Data Cleaning - Country

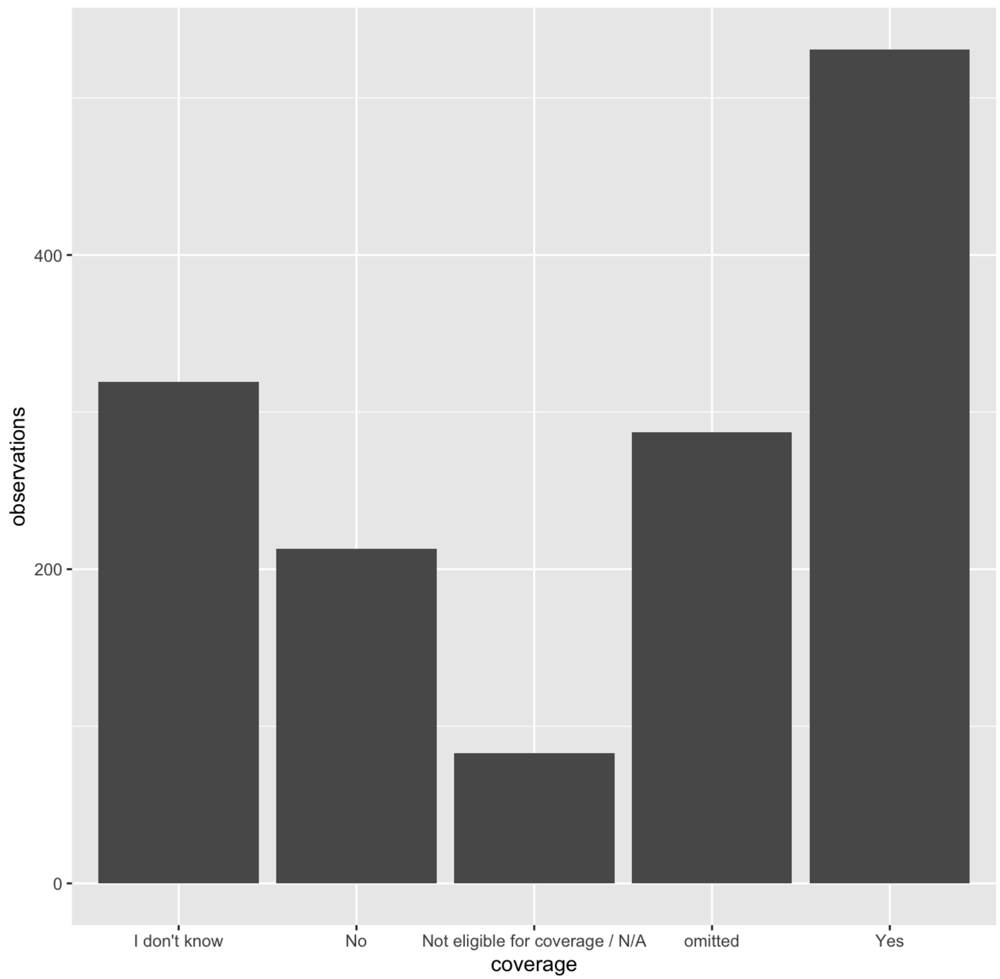
- 53 countries are represented but the observations are not evenly divided among them
- Will only control for the top 5 countries creating indicator variables 1 or 0 if observations are from them



Bar plot displays total observations for the top 11 countries

Data Cleaning – Coverage Policy

- The variable for mental health coverage policy takes on 4 kinds of values
- To control for coverage only need Yes and No
- Requires omitting 689 rows



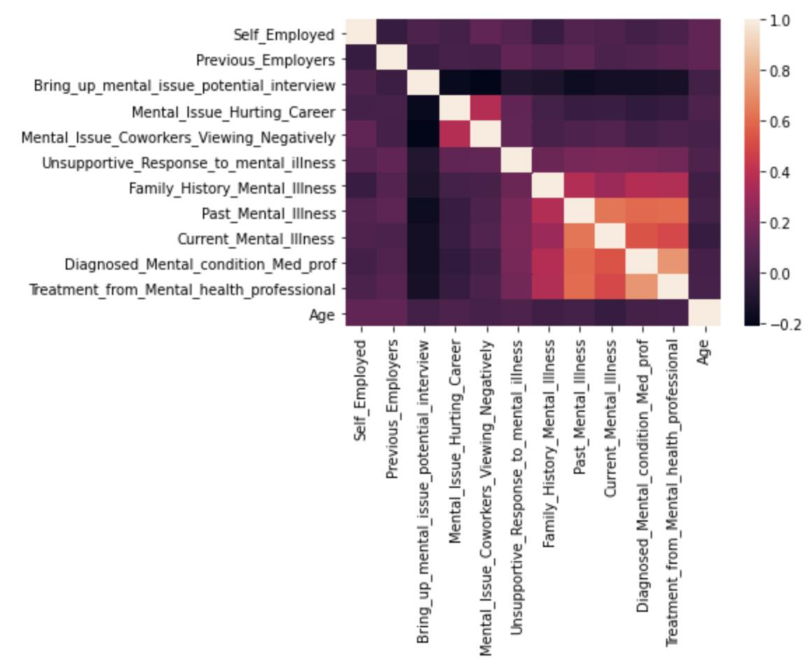
Data Cleaning – EDA and Classification

- Reducing multiple categories to closely-related fewer categories for simpler analysis and interpretation.
    1. Categories 'Somewhat Easy', 'Very Easy' merged as 'Easy', and Categories 'Somewhat difficult', 'Very difficult' merged as 'Difficult'
    2. Merging "Not Eligible" to a "No" Category depending on the context.
    3. Merging text-based responses like 'No, I don't think it would', 'No, it has not' to a broader category 'No'.
- Cleaning input responses for gender like "Male", "Female", "Woman", "Non-Binary", "Transgender" to three categories : "M", "F", and "Other.
- Imputing missing values for predictors. This includes risk of errors due to large amount of categorical data, but a greater emphasis has been given to avoiding dropping of valuable data points.

EDA (Exploratory Data Analysis)

- Correlation Matrix -

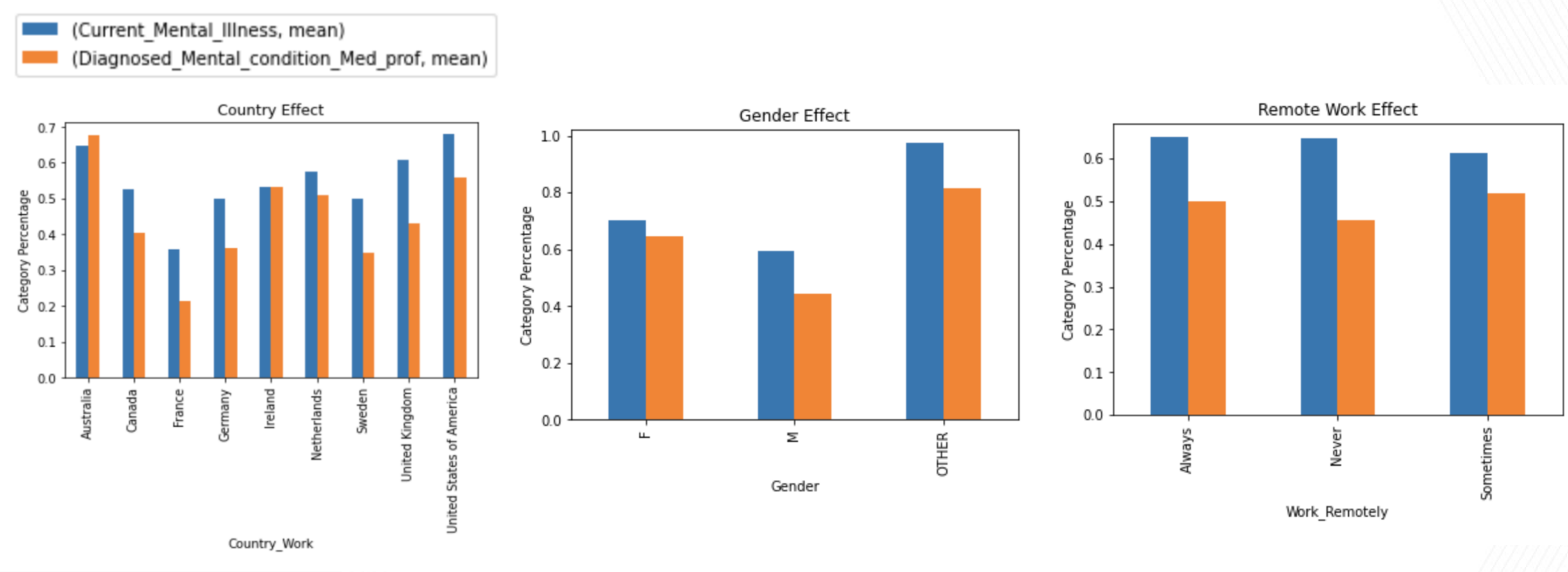
An initial analysis to observe the correlation between the various columns of the data gave the following matrix



We can see that “Diagnosed Medical condition Med Prof” is highly correlated to “Current Mental Illness”, “Past Mental Illness” and “Family History Mental Illness”. Other categorical variables like Gender, Country and Work style are not included because a numerical correlation cannot be calculated.

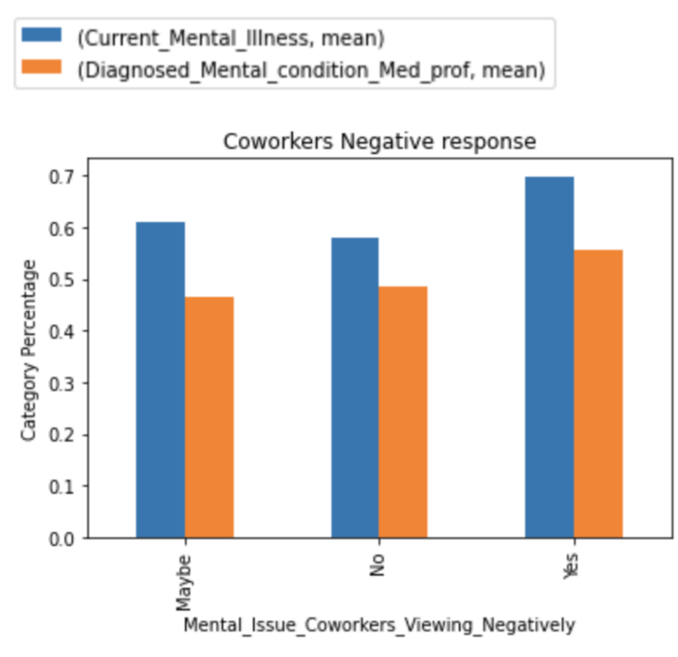
EDA

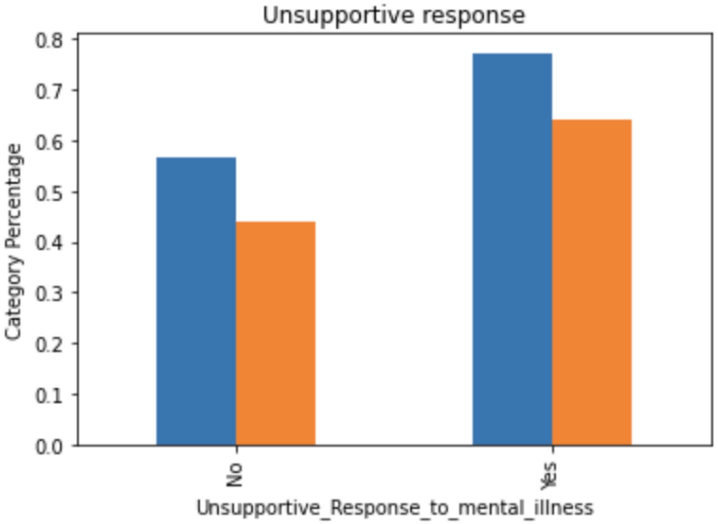
- To observe the effect between Current Mental Illness, Diagnosed Mental Illness and other categorical columns, bar graphs were generated as follows -



- USA, UK and Australia are one of the top few countries with highest percentage of employees suffering from mental illness or seeking medical help. (Australia does however have very few data points and the analysis may not be very representative of the truth)
- There seems to be a larger percentage of employees categorized under the "Other" gender category. It could mean that the LGBTQ community faces a greater risk for suffering from mental illness at a tech workplace.
- There seems to be no significant effect of the work style (remote or not) on mental condition.

EDA





- Some other interesting graphs show the effect of society's response to mental health discussions
- It is seen clearly that when there is a negative response from coworkers or unsupportive response from the company, employees are more prone to having a mental illness

Methods

Poisson Regression

- A Poisson regression model can be useful in determining the effect of certain factors on a response count variable.
- The dataset does not inherently contain a count response variable
- Aggregating a variable such as treatment or illness, the dataset can be transformed into count data format
- The Poisson model will be used to see the effect of geographic location, negative workplace sentiment, and providing coverage on the number of people seeking treatment and the number of diagnosed mental illnesses
- Will be used for descriptive interpretation of the coefficients and the F-test will be used for testing coefficients' significance, so a dispersion calculation is not needed

Logistic Regression

- To answer our research questions, 2 Logistic regression models will be built :
    1. To predict probability of employee's current mental illness
    2. To predict the probability of him/her seeking treatment for mental health issues.
- Missing data among predictors have been imputed to avoid dropping data points that would have other significant predictors.
- Relevant factors(like Past mental illness, Employer's take on mental illness, Previous Employers, Age, Gender, Family History etc.) have been considered for the models.
- p-value of the factors will be used to determine their statistical significance.
- Model results will be used to descriptively interpret the coefficients against the response variable.

Random Forest

- **Algorithm:** Decision Trees and their extension Random Forests are robust and easy-to-interpret machine learning algorithms for both classification and regression tasks.
- **Analysis:**

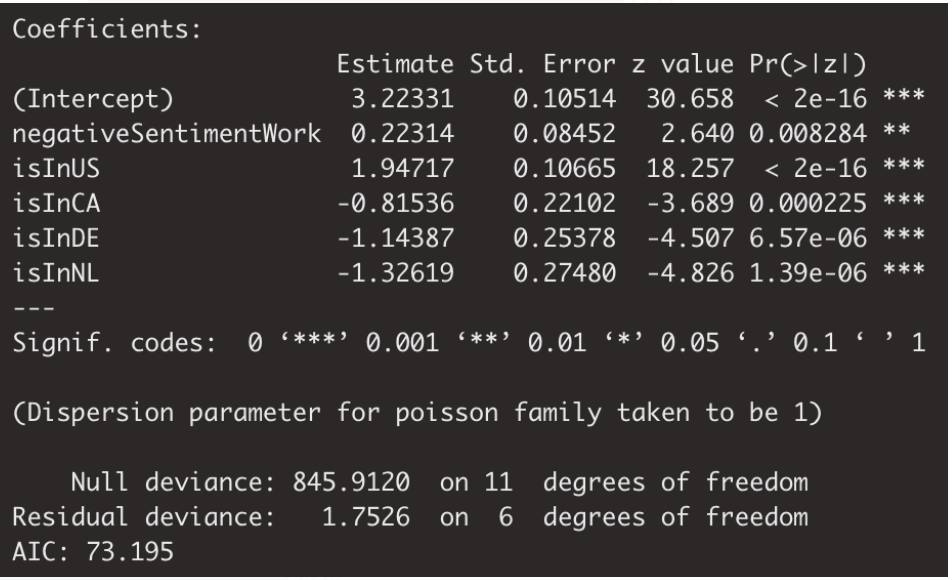
- **Variable Importance:** The chart shows us a list of variables in decreasing order of their predictive power towards the built Randomforest model

- **Evaluation:** The Receiver Operating Characteristic (ROC) is a plot that can be used to determine the performance and robustness of a binary or multi-class classifier. The x-axis is the false positive rate (FPR) and the y-axis is the true positive rate (TPR). This ensures we are taking into account the accuracies of both classes while classifying.

Results and Discussion

Poisson Regression – Sentiment/Country

- Poisson model fit initially with the negative sentiment work feature and indicator variables for the top 5 countries as explanatory variables and the count of mental illness as the response
- F-test indicates that the UK does not have a significant coefficient with a p-value of 0.646 so it is dropped from the model
- Full model shown below with all variables having significant coefficients through an F-test
- Key takeaways:
  - The expected number of mental illnesses in the U.S. is higher than the expected number of individuals in other countries outside of the top 5 countries with observations (intercept)
  - The expected number of mental illnesses in Canada, Denmark, and the Netherlands are each lower than the expected number of countries outside of the top 5 countries with observations (intercept)
  - The expected number of mental illnesses is higher when employees have a negative experience concerning mental health at their work

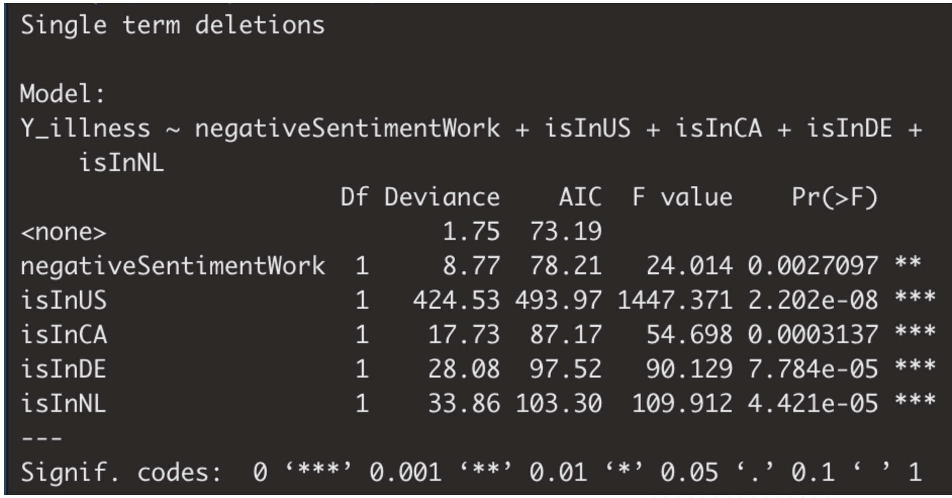


```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)            3.22331    0.10514  30.658  < 2e-16 ***
negativeSentimentWork  0.22314    0.08452   2.640 0.008284 ** 
isInUS                 1.94717    0.10665  18.257  < 2e-16 ***
isInCA                -0.81536    0.22102  -3.689 0.000225 ***
isInDE                -1.14387    0.25378  -4.507 6.57e-06 ***
isInNL                -1.32619    0.27480  -4.826 1.39e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 845.9120  on 11  degrees of freedom
Residual deviance:   1.7526  on  6  degrees of freedom
AIC: 73.195
```

Results from the final Poisson Regression model chosen



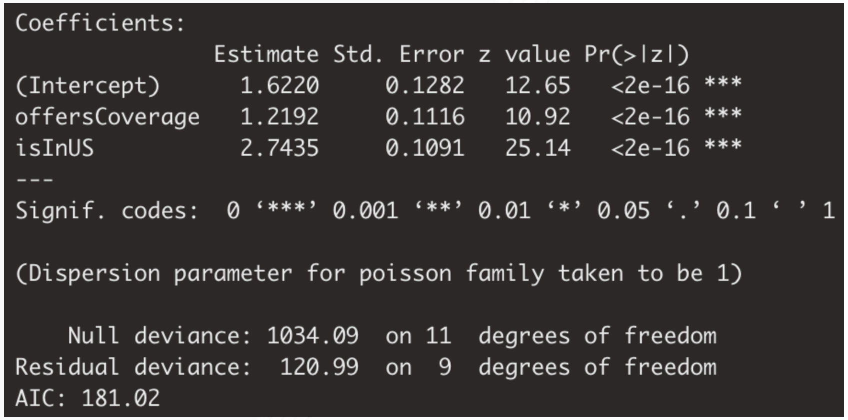
```
Single term deletions

Model:
Y_illness ~ negativeSentimentWork + isInUS + isInCA + isInDE + 
    isInNL
                      Df Deviance    AIC  F value     Pr(>F)    
<none>                       1.75  73.19                        
negativeSentimentWork  1     8.77  78.21   24.014 0.0027097 ** 
isInUS                 1   424.53 493.97 1447.371 2.202e-08 ***
isInCA                 1    17.73  87.17   54.698 0.0003137 ***
isInDE                 1    28.08  97.52   90.129 7.784e-05 ***
isInNL                 1    33.86 103.30  109.912 4.421e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

F-test results from the drop1 function for the final model chosen

Poisson Regression – Coverage/Country

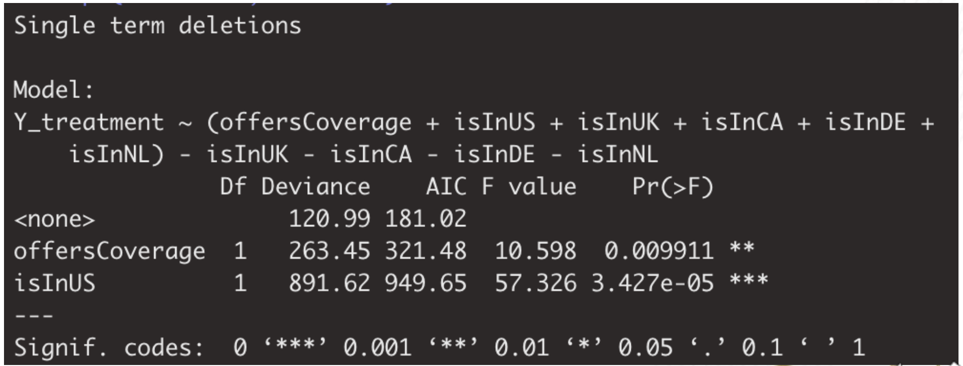
- Poisson model fit initially with explanatory variables for offering coverage and all 5 top countries and response variable the count of individuals seeking treatment
- Repeated F-tests with drop1 showed that the only variables with statistically significant coefficients were for offering coverage and is located in the US
- Key takeaways:
  - The expected number of people seeking treatment is greater in the United States than in other countries outside of the top 5 countries with observations
  - The expected number of those seeking treatment is greater when employer's provide coverage for mental health visits



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.6220     0.1282   12.65   <2e-16 ***
offersCoverage  1.2192     0.1116   10.92   <2e-16 ***
isInUS          2.7435     0.1091   25.14   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1034.09  on 11  degrees of freedom
Residual deviance:  120.99  on  9  degrees of freedom
AIC: 181.02
```

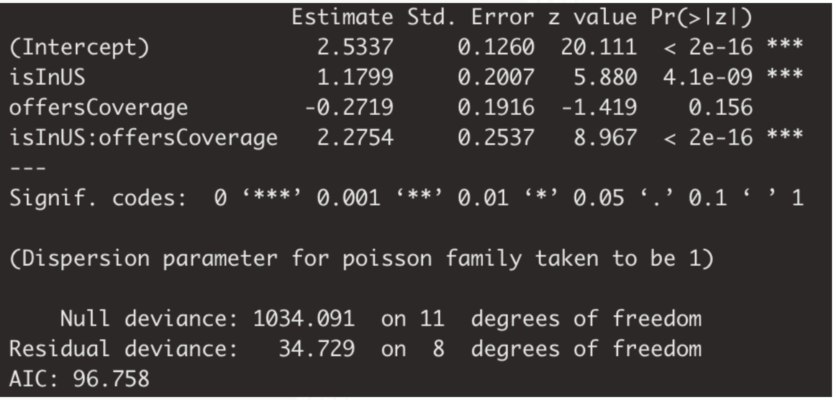


```
Single term deletions

Model:
Y_treatment ~ (offersCoverage + isInUS + isInUK + isInCA + isInDE + 
    isInNL) - isInUK - isInCA - isInDE - isInNL
               Df Deviance    AIC F value     Pr(>F)    
<none>              120.99 181.02                       
offersCoverage  1   263.45 321.48  10.598   0.009911 ** 
isInUS          1   891.62 949.65  57.326 3.427e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Poisson Regression –Coverage/Country Interaction

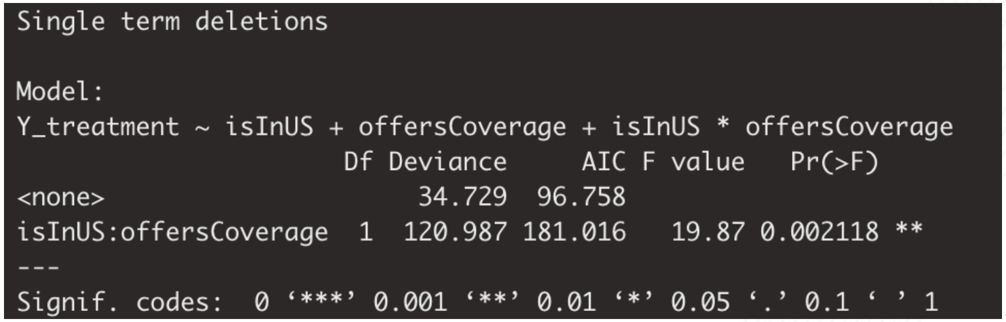
- Interaction effects were noticed to be statistically significant between location in the US and if the employer offers coverage
- Interpretation:
  - 2.275 is the amount to add to the coverage coefficient to get the coefficient for individuals living in the U.S.
  - Therefore, employer's providing coverage in the U.S. will have a multiplied effect on the number of people seeking treatment in the U.S.



```
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)             2.5337     0.1260  20.111  < 2e-16 ***
isInUS                  1.1799     0.2007   5.880  4.1e-09 ***
offersCoverage         -0.2719     0.1916  -1.419    0.156
isInUS:offersCoverage   2.2754     0.2537   8.967  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1034.091  on 11  degrees of freedom
Residual deviance:   34.729  on  8  degrees of freedom
AIC: 96.758
```

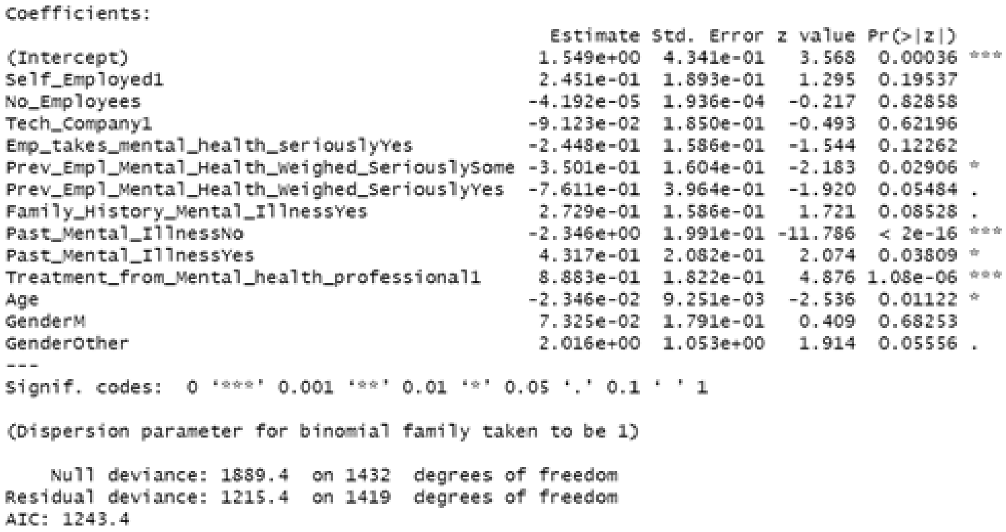


```
Single term deletions

Model:
Y_treatment ~ isInUS + offersCoverage + isInUS * offersCoverage
                      Df Deviance     AIC F value    Pr(>F)
<none>                     34.729  96.758
isInUS:offersCoverage  1  120.987 181.016   19.87 0.002118 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Logistic Regression Model 1

"Current Mental Illness" Vs Predictors



```
Coefficients:
                                             Estimate Std. Error z value Pr(>|z|)    
(Intercept)                                 1.549e+00  4.341e-01   3.568  0.00036 ***
Self_Employed1                              2.451e-01  1.893e-01   1.295  0.19537    
No_Employees                               -4.192e-05  1.936e-04  -0.217  0.82858    
Tech_Company1                              -9.123e-02  1.850e-01  -0.493  0.62196    
Emp_takes_mental_health_seriouslyYes       -2.448e-01  1.586e-01  -1.544  0.12262    
Prev_Empl_Mental_Health_Weighed_SeriouslySome -3.501e-01  1.604e-01  -2.183  0.02906 *  
Prev_Empl_Mental_Health_Weighed_SeriouslyYes  -7.611e-01  3.964e-01  -1.920  0.05484 .  
Family_History_Mental_IllnessYes            2.729e-01  1.586e-01   1.721  0.08528 .  
Past_Mental_IllnessNo                      -2.346e+00  1.991e-01 -11.786  < 2e-16 ***
Past_Mental_IllnessYes                      4.317e-01  2.082e-01   2.074  0.03809 *  
Treatment_from_Mental_health_professional1  8.883e-01  1.822e-01   4.876 1.08e-06 ***
Age                                        -2.346e-02  9.251e-03  -2.536  0.01122 *  
GenderM                                     7.325e-02  1.791e-01   0.409  0.68253    
GenderOther                                 2.016e+00  1.053e+00   1.914  0.05556 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1889.4  on 1432  degrees of freedom
Residual deviance: 1215.4  on 1419  degrees of freedom
AIC: 1243.4
```

Interpretation:

- ➢ **Previous employers that weighed mental health seriously reduced the odds of mental illness among employees as compared to employers that do not take mental health seriously.**
  - Somewhat seriously weighed concerns : Reduced odds of illness by 0.350
  - Very seriously weighed concerns : Reduced odds of illness by 0.76

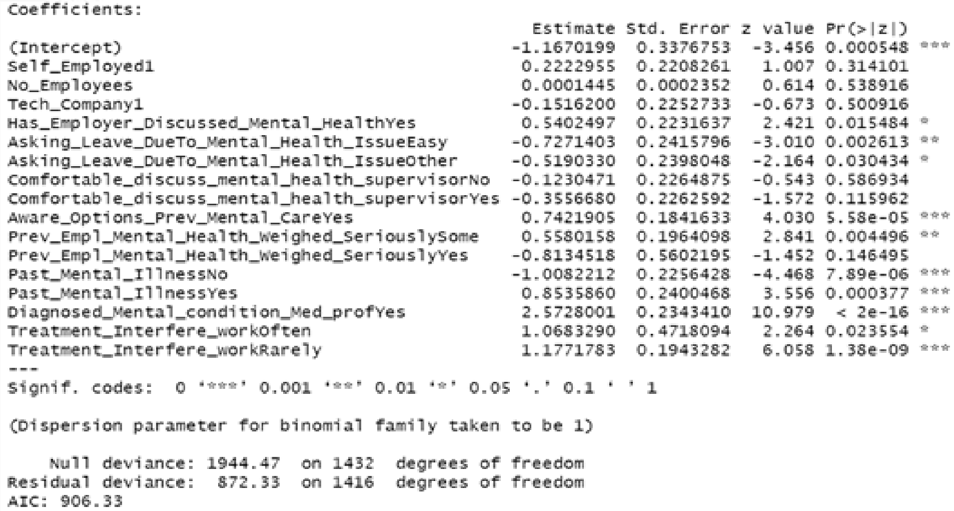
- ➢ **Past Mental illness is significant in determining a person's current mental illness.**
  - Employee without Past mental illness : Reduced odds of current illness by 0.24

- ➢ **Age contributes to a person's mental illness.**
  - Person with higher age : Reduced odds of illness by 0.023

- ➢ **Gender or family history has no effect on a person's mental illness.**

Logistic Regression Model 2

"Employees Seeking Treatment" Vs Predictors



```
Coefficients:
                                               Estimate Std. Error z value Pr(>|z|)    
(Intercept)                                  -1.1670199  0.3376753  -3.456 0.000548 ***
Self_Employed1                                0.2222955  0.2208261   1.007 0.314101    
No_Employees                                  0.0001445  0.0002352   0.614 0.538916    
Tech_Company1                                -0.1516200  0.2252733  -0.673 0.500916    
Has_Employer_Discussed_Mental_HealthYes       0.5402497  0.2231637   2.421 0.015484 *  
Asking_Leave_DueTo_Mental_Health_IssueEasy   -0.7271403  0.2415796  -3.010 0.002613 ** 
Asking_Leave_DueTo_Mental_Health_IssueOther  -0.5190330  0.2398048  -2.164 0.030434 *  
Comfortable_discuss_mental_health_supervisorNo -0.1230471  0.2264875  -0.543 0.586934    
Comfortable_discuss_mental_health_supervisorYes -0.3556680  0.2262592  -1.572 0.115962    
Aware_Options_Prev_Mental_CareYes             0.7421905  0.1841633   4.030 5.58e-05 ***
Prev_Empl_Mental_Health_Weighed_SeriouslySome  0.5580158  0.1964098   2.841 0.004496 ** 
Prev_Empl_Mental_Health_Weighed_SeriouslyYes  -0.8134518  0.5602195  -1.452 0.146495    
Past_Mental_IllnessNo                        -1.0082212  0.2256428  -4.468 7.89e-06 ***
Past_Mental_IllnessYes                        0.8535860  0.2400468   3.556 0.000377 ***
Diagnosed_Mental_condition_Med_profYes        2.5728001  0.2343410  10.979  < 2e-16 ***
Treatment_Interfere_workOften                 1.0683290  0.4718094   2.264 0.023554 *  
Treatment_Interfere_workRarely                1.1771783  0.1943282   6.058 1.38e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1944.47  on 1432  degrees of freedom
Residual deviance:  872.33  on 1416  degrees of freedom
AIC: 906.33
```

Interpretation:

- **Employers that discuss mental illness with their employees increase the chances of employees seeking treatment for depression.**
  - Employer who discussed concerns : Increased odds of employee seeking treatment by 0.540

- **Employees awareness on Mental health benefits increases chances of seeking treatment.**
  - Employee aware of health benefits : Increased odds of seeking treatment by 0.742

- **Past mental illness significantly impacts employees seeking treatment for depression.**
  - Person with prior mental illness : Increases odds of seeking treatment by 0.853

Random Forest Classifier for Prediction

Problem Statement: To identify factors that increase the risk of Mental illness

Dependent Variable – Respondents who currently having a mental illness or are diagnosed for a mental illness by a mental health professional.
Predictor variables – We see different categories of predictors in the dataset:

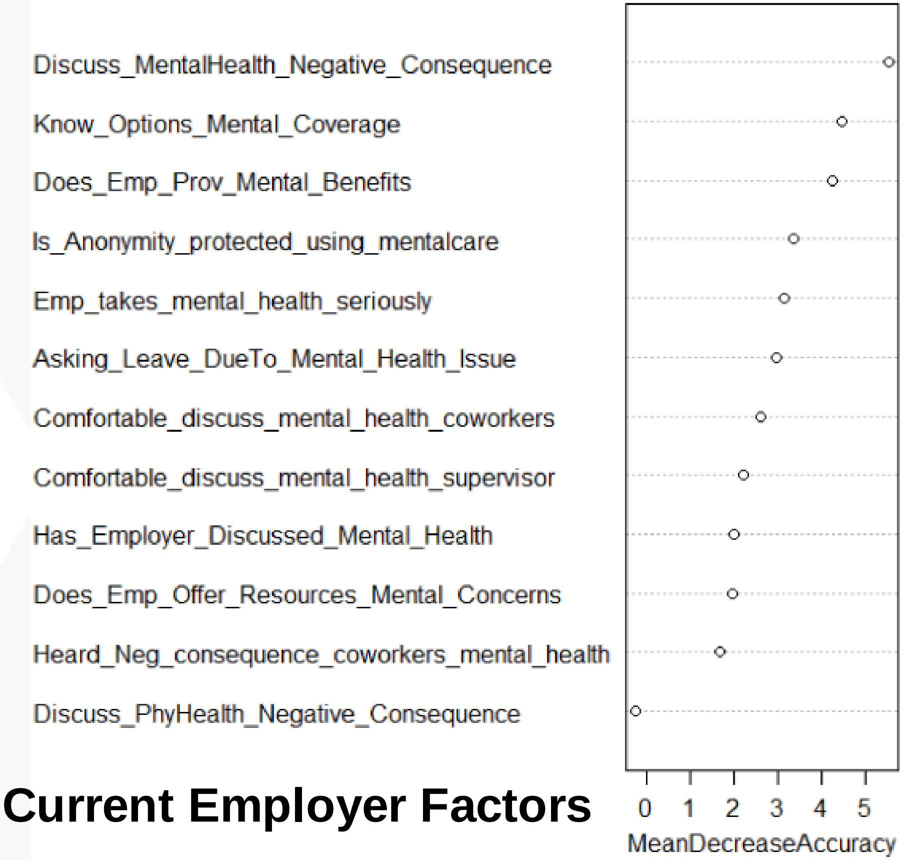
- Variables pertaining to current employers

- Variables pertaining to previous employers

- Variables pertaining to personal characteristics

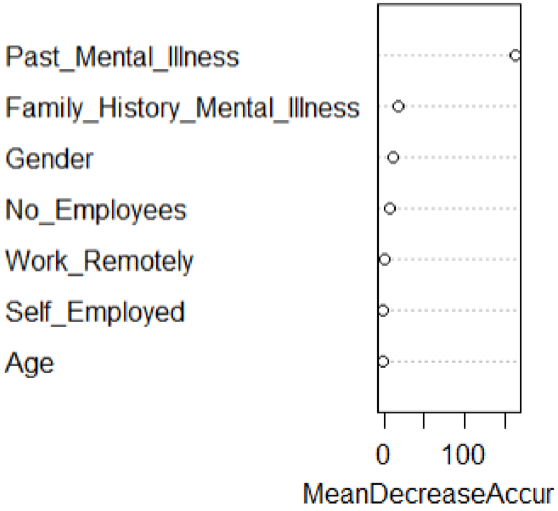
We initially build individual Randomforest models for each category letting us have a closer look within the category.

Random Forest – Variable Importance for predicting the risk of mental Illness



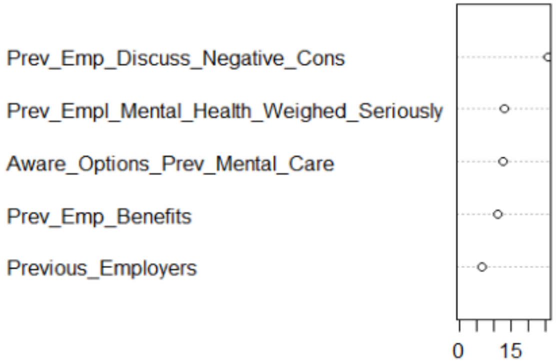
Current Employer Factors

In both cases of previous and current employers, employees who perceive that a discussion around mental health could causes negative consequence on Mental Health are more likely to be affected



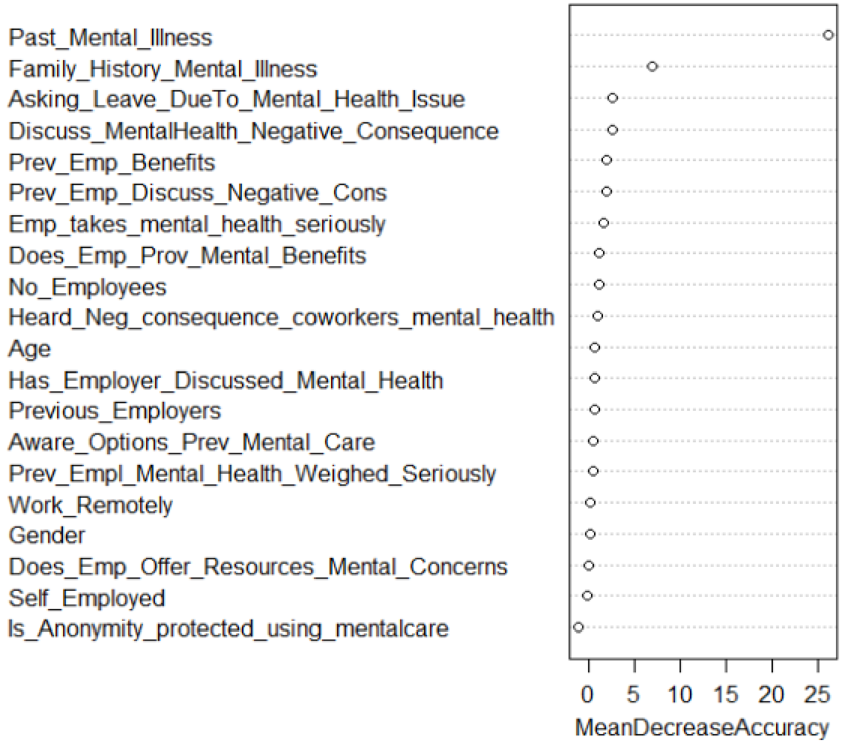
Personal Factors

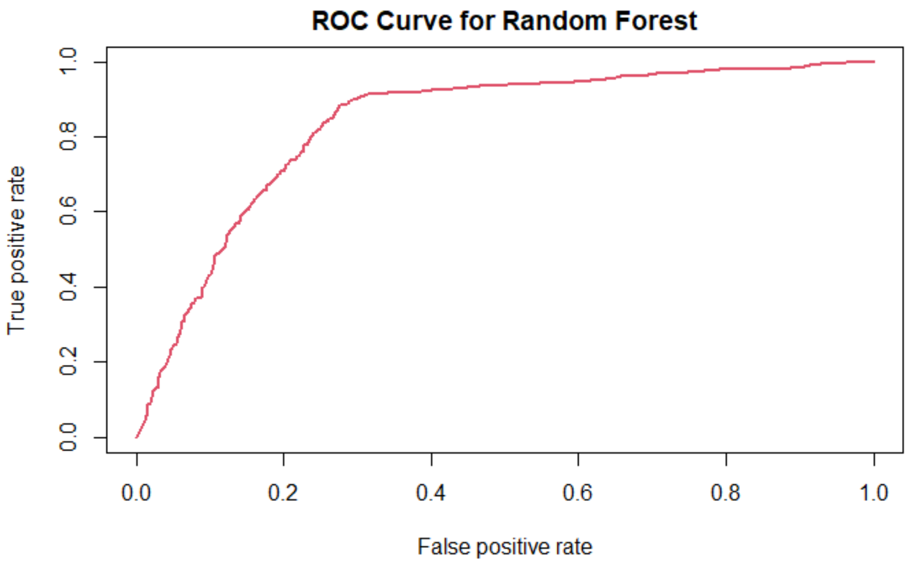
Employees with past mental illnesses are more susceptible to have a mental illness

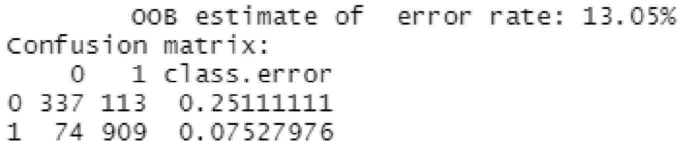


Previous Employer Factors

Random Forest – Overall Variable Importance for predicting risk of mental Illness







```
        OOB estimate of  error rate: 13.05%
Confusion matrix:
    0   1 class.error
0 337 113  0.25111111
1  74 909  0.07527976
```

- We see that of all the factors, the 'Past Mental Illness' seems to be a very important factor – meaning, this variable highly predicts the risk of an individual.

- The top subsequent factors are variables belonging to a mix of different categories and hence highlights the importance of all categories.

The predictive model with the variables listed has an accuracy of ~87%

Conclusion

Summary

- The expected number of mental illnesses is statistically greater in the U.S. over other countries
- Employers offering benefits packages including mental health resources increases the number of employees seeking treatment
- An employee with a past mental illness has a significantly higher chance of a recurring illness.
- Employers weighing mental health seriously have reduced the chances of depression among employees as compared to employers that do not take mental health seriously.
- Employee's awareness on mental health benefits increases their chances of seeking treatment.

Recommendations to Tech Companies

- Employers discussion of mental health in the workplace and promoting a positive attitude about discussing mental health is very important, and its absence can lead to a greater presence of mental illnesses
- Employers are highly recommended to offer their employees coverage for mental health visits, especially if the company is in the United States and if their employee has had a history of mental illness in the past

Georgia Tech
CREATING THE NEXT

End

Georgia Tech

CREATING THE NEXT