let's talk

MENTAL HEALTH

You're not alone

You're not the first to go through it

You're not going to be the last to go through it

Mental Health at Work: Data Analysis

Purple Team 9

PRESENTERS

Harkirat Kaur

Ashima Dogra

Grace Zhou

Saswati Prusty

Jagan

Weiming Luo

WHY ASSESSING MENTAL HEALTH AT WORK IMPORTANT?

Understanding the human element is key to overcoming workplace mental health hurdles.

McKinsey & Company

Source: Association of Small & Medium Enterprises, Singapore; McKinsey & Company

72%
OF EMPLOYEES
want their employer to champion mental health and well-being[1]

DEPRESSION ALONE COSTS
$200 million
in lost workdays[2]

81%
BELIEVE ON-THE-JOB STRESS
affects relationships with friends and family[3]

1 in 5
ADULTS
experience mental health issues, and 60 percent aren't seeking treatment[4]

[1] Survey of office workers, Peldon Rose, 2018
[2] CDC.gov
[3] Mental Health America
[4] NAMI.org

DATA ANALYSIS OBJECTIVE

As a Mental Health Consulting firm, our objective is to analyse factors that contribute to the need of mental health treatment at work through the results of a survey.

Based on the identified predictors, we will perform predictions and exploratory analysis for the Tech Industry employees who would need treatment.

We can accordingly modify our prediction models to work on different employee demographics and extend these models to other industries.

This analysis will benefit potential employers who wish to assess mental health of their employees and offer them treatment benefits.

AGENDA

About the dataset

Defining the key elements of the dataset.

Preprocessing of the dataset

Details of preprocessing and cleaning of dataset for analysis

Initial extrapolatory Analysis

Discuss data visualization and exploratory analysis to draw initial results.

Models & Insights

Discuss on different predictive model and their business insights

Summary and Questions

Summarise results and take questions if any.

ABOUT THE DATASET

1 Origin

A survey dataset from kaggle.com, collected for OSMI research, that measures attitudes towards mental health and frequency of mental health disorders in the tech workplace.
Kaggle - https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey

OSMI - https://osmihelp.org/research

2 Key elements

The observations across 27 variables, such as age, gender, state, country, family history with mental health, treatment received, employment details etc.

3 Observations

Out of the predictors, what came out to be stronger for usage in designing our models were - work_interference, age, gender, care_options and observed negative consequences amongst coworkers

4 Regions

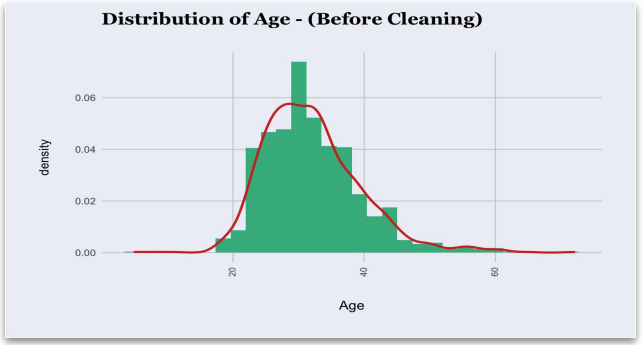
The dataset has observations made in regions such as United States, Croatia, UK, France, Switzerland, Italy, South Africa, India, Thailand, to name a few and provides a diverse view of the variables across countries.

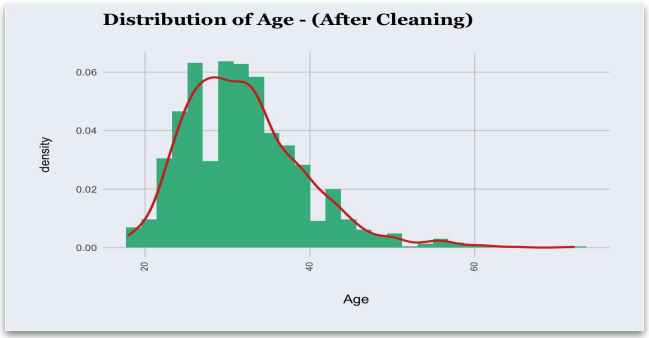
PRE-PROCESSING OF DATASET

Pre-processing & Cleaning

1. Remove any outliers in age
2. Categorized gender into three categories - spelling errors
3. Deleted any missing values/rows in self-employed column.
4. No of employees within organization - created bins.

Example - Age distribution before and after cleaning





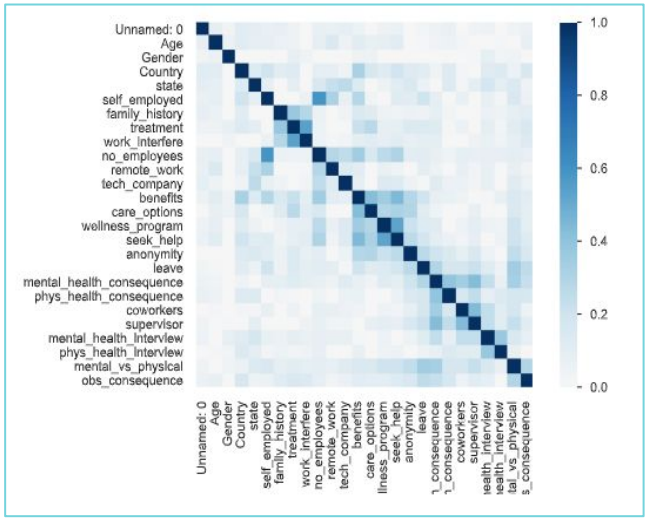
INITIAL EXPLORATORY ANALYSIS

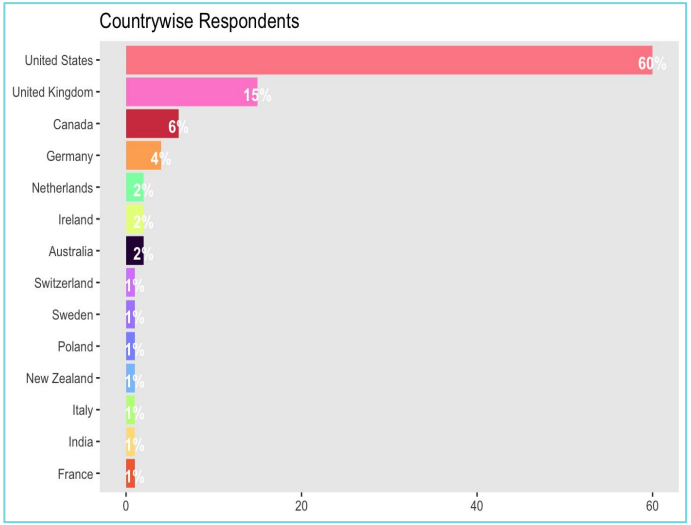
Observations from initial investigations -

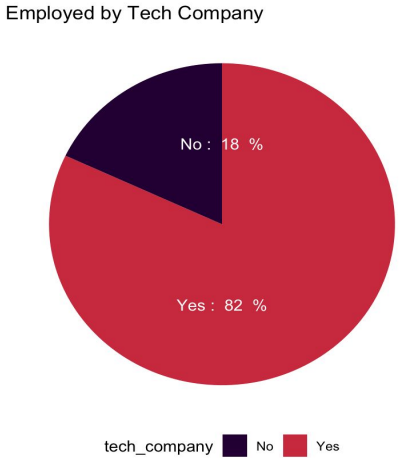
Correlation Chart

Country distribution

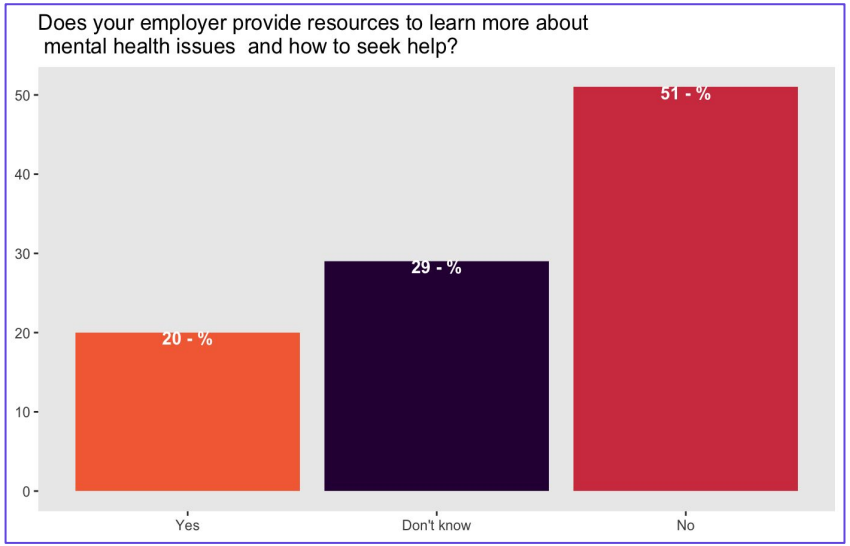
Industry Distribution

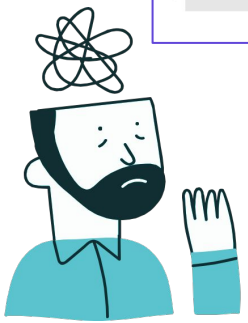




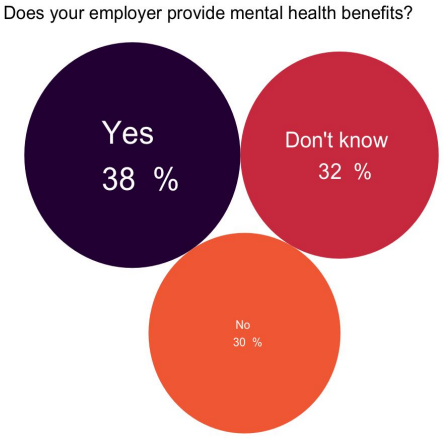


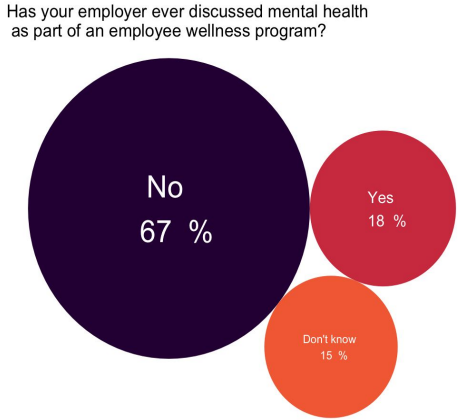
INITIAL EXPLORATORY ANALYSIS



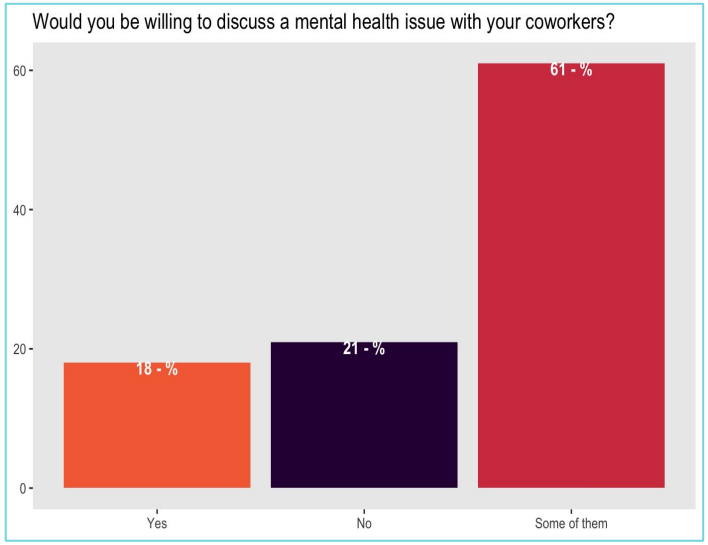


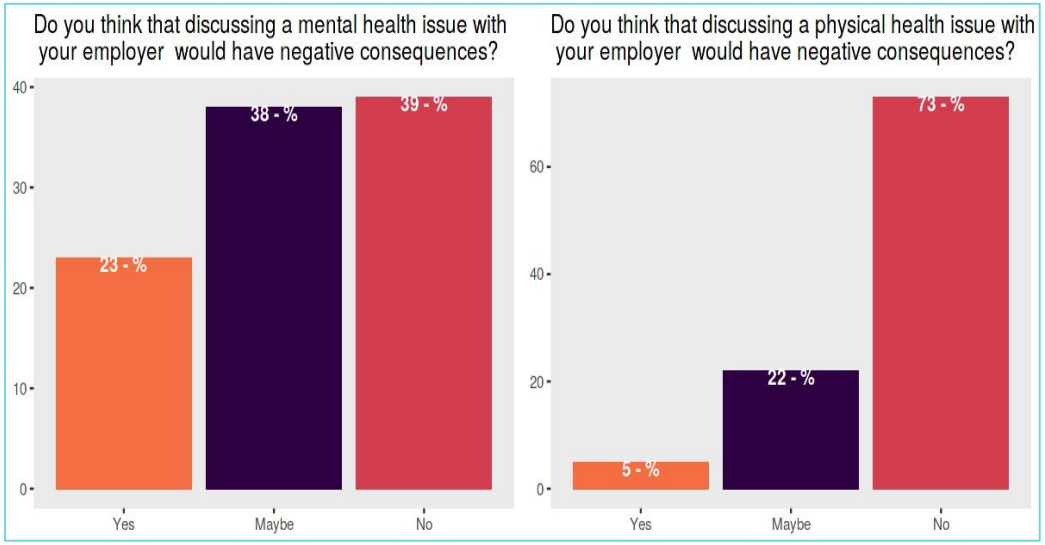
Employer Readiness





INITIAL EXPLORATORY ANALYSIS

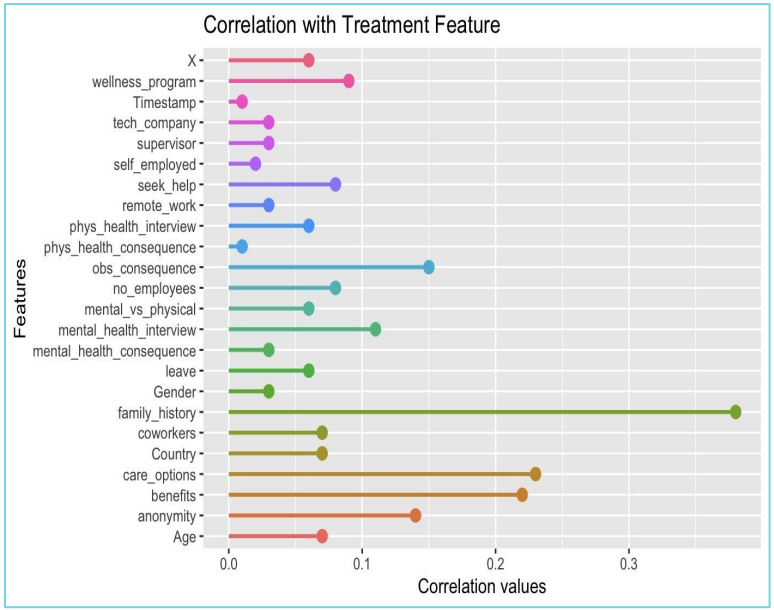


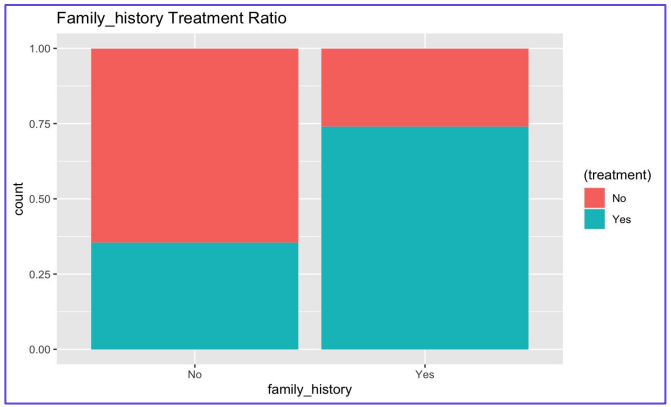


Attitude difference
- mental vs physical health

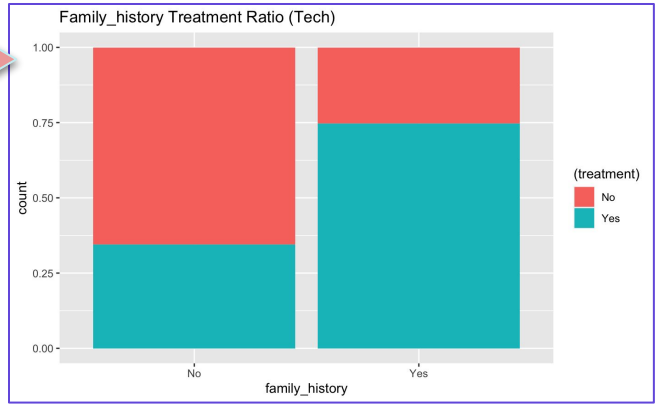
INITIAL EXPLORATORY ANALYSIS FOR TREATMENT

Correlation with treatment feature



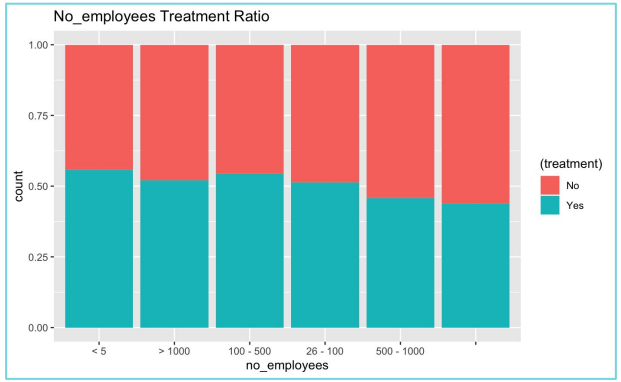


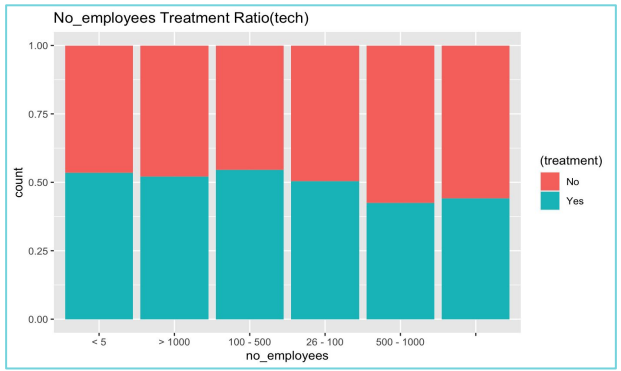
Family History



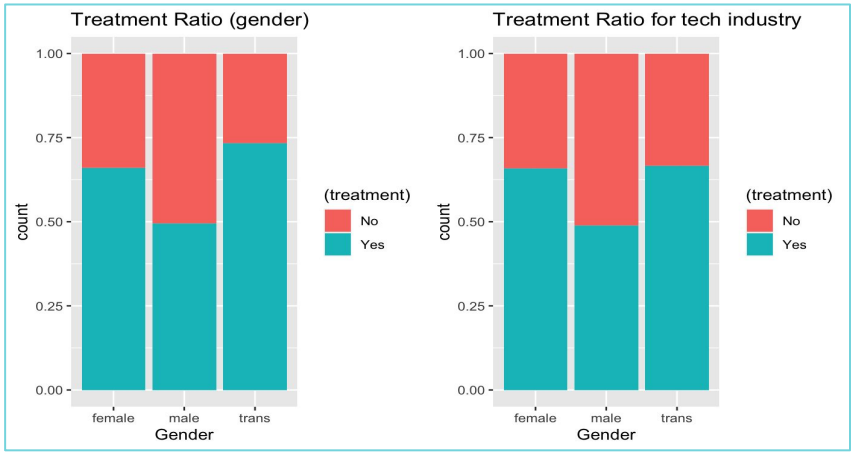
INITIAL EXPLORATORY ANALYSIS FOR TREATMENT

Treatment ratio - Number of Employees

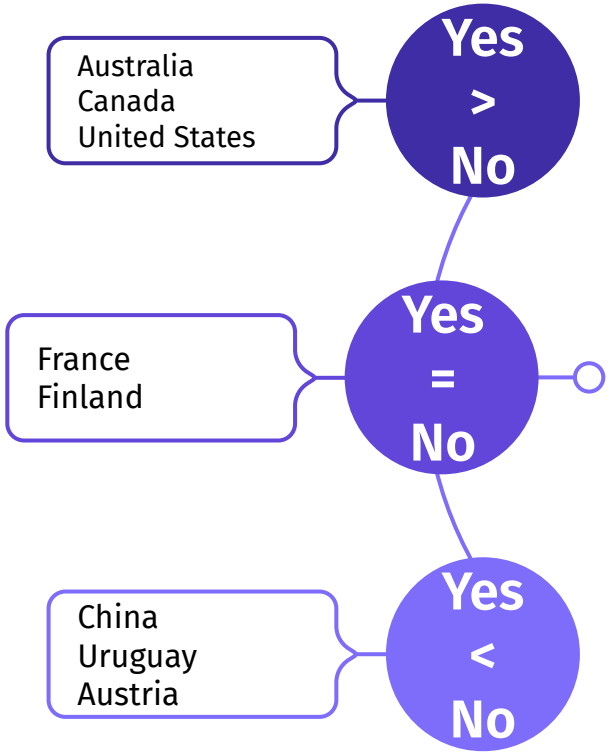


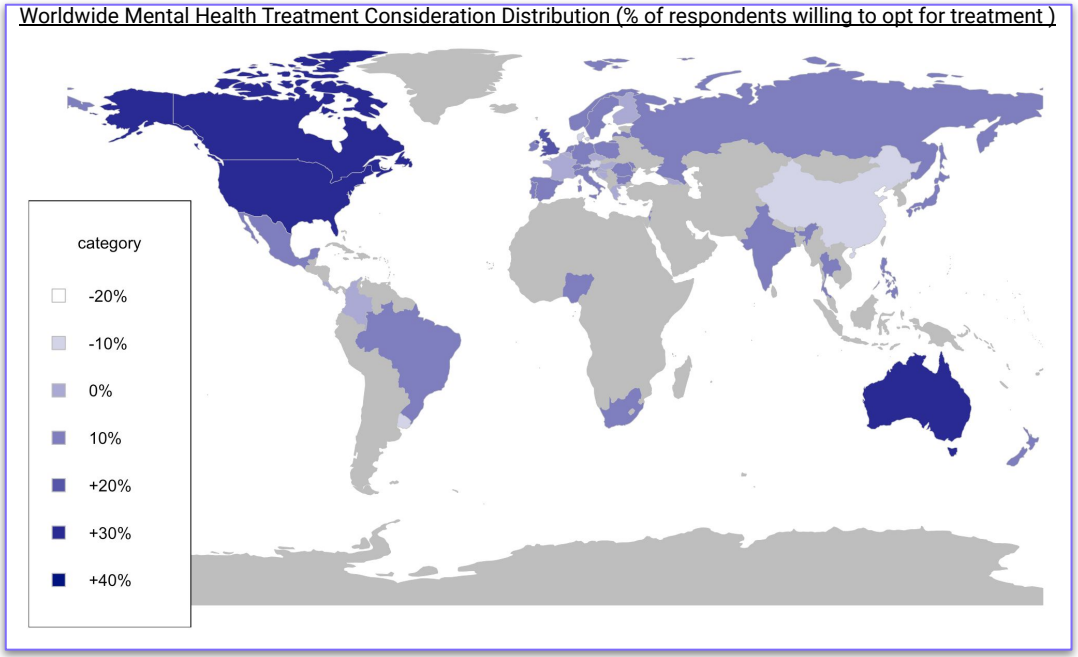


Treatment ratio -Gender



INITIAL EXPLORATORY ANALYSIS



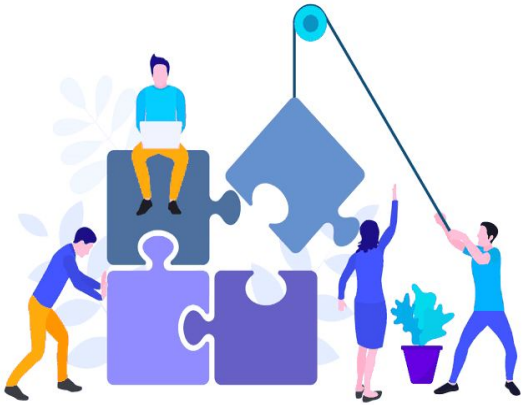


LOGISTIC REGRESSION MODEL - INTRODUCTION

Target Variable - Treatment

Predictors - Age, gender, no_employees, work_interference, family_history, self_employed, benefits, care_options, mental_health_interview, seek_help+anonymity

Goal - Predict Treatment requirement based on the impact of each predictor variable



Pre-processing -

- Binning of age into age groups
- Dropped state and comments - very high cardinality
- Dropped country - high cardinality and major data for United States.

LOGISTIC REGRESSION- MODEL RESULTS

```
Call:
glm(formula = treatment ~ Age + Gender + no_employees + work_interfere +
    family_history + self_employed + benefits + care_options +
    mental_health_interview + seek_help + anonymity, family = "binomial",
    data = train.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4039  -0.4551   0.3933   0.6693   2.6934

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -0.11639    1.06032  -0.110   0.9126
Age<20                     0.03639    0.67751   0.054   0.9572
Age>60                     2.07238    1.39249   1.488   0.1367
Age31-40                  -0.01010    0.21831  -0.046   0.9631
Age41-50                   0.34607    0.34703   0.997   0.3187
Age51-60                   1.53555    0.87884   1.747   0.0806 .
Gendermale                -0.08137    0.44976  -0.181   0.8564
Gendertrans               -0.06789    0.97471  -0.070   0.9445
no_employees> 1000         0.20376    0.41582   0.490   0.6241
no_employees100 - 500      0.59518    0.43425   1.371   0.1705
no_employees26 - 100       0.30967    0.38077   0.813   0.4161
no_employees500 - 1000     0.67876    0.63106   1.076   0.2821
no_employees6-25           0.13459    0.36083   0.373   0.7091
work_interfereOften        3.82608    0.40079   9.546  < 2e-16 ***
work_interfereRarely       2.58873    0.33179   7.802 6.08e-15 ***
work_interfereSometimes    3.19837    0.30335  10.543  < 2e-16 ***
family_historyNo          -1.10565    0.20618  -5.363 8.20e-08 ***
```

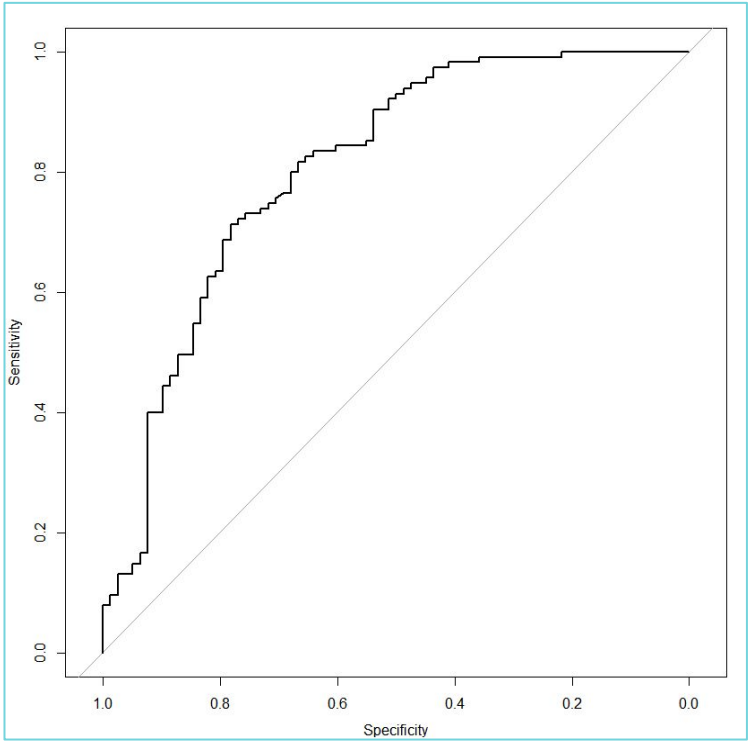
```
self_employedYes           0.12812    0.34532   0.371   0.7106
benefitsDon't know        -0.20225    0.28903  -0.700   0.4841
benefitsYes                0.56620    0.31545   1.795   0.0727 .
care_optionsNot sure      -0.30032    0.26345  -1.140   0.2543
care_optionsYes            0.61681    0.27068   2.279   0.0227 *
mental_health_interviewMaybe -1.86445  0.82445  -2.261   0.0237 *
mental_health_interviewNo  -1.71342    0.78993  -2.169   0.0301 *
seek_helpDon't know        0.83376    0.32989   2.527   0.0115 *
seek_helpNo                0.36528    0.30991   1.179   0.2385
anonymityDon't know       -0.49895    0.25529  -1.954   0.0506 .
anonymityNo               -0.56642    0.47721  -1.187   0.2353
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1008.65  on 776  degrees of freedom
Residual deviance:  671.99  on 749  degrees of freedom
  (223 observations deleted due to missingness)
AIC: 727.99

Number of Fisher Scoring iterations: 5
```

$$\log it = \log(Odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q$$

CONFUSION MATRIX & ROC CURVE

p = 0.5		Actual class	
		0	1
Predicted class	0	41	11
	1	37	104

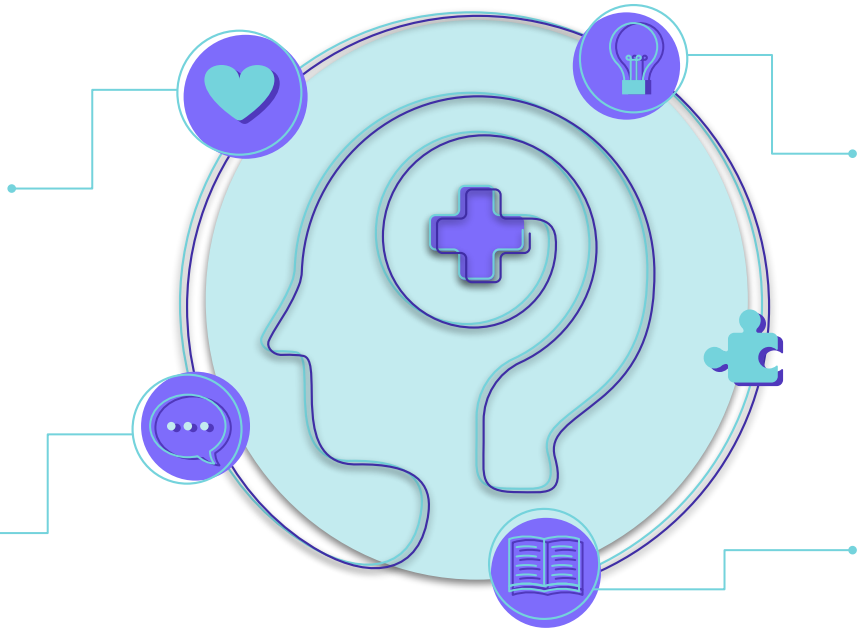
Threshold	Specificity	Sensitivity
0.9	0.92	0.33
0.747	0.78	0.71
0.5	0.52	0.9



LOGISTIC REGRESSION - BUSINESS INSIGHTS

Males are less likely to have sought treatment. Behaviour/mindset ? Trans are less likely, than females. Is this really the case ? Trans - higher targets ? Could be due to low data availability.

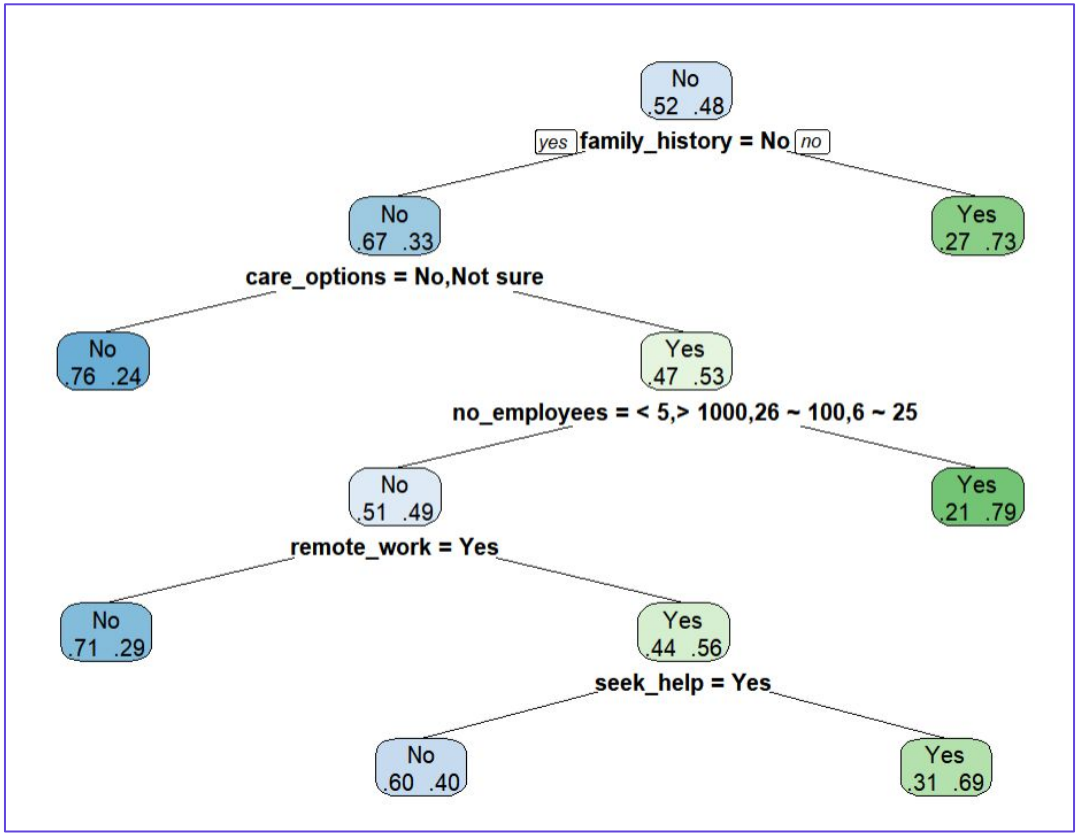
Employees with family history of mental health are **5 times more likely** to receive mental health treatment. Family history matters a lot!! Or maybe they are more aware.



For employees having interference in work due to mental health conditions, they are more likely to pursue mental health treatment. **Positive sign!!**

In general, if employers provide **resources, benefits, care options and employees anonymity** result in seeking more mental health care…

CART MODEL - First Attempt



Target Variable :
Treatment (Yes or No)

Predictors :
All variables excluding country, state, comments

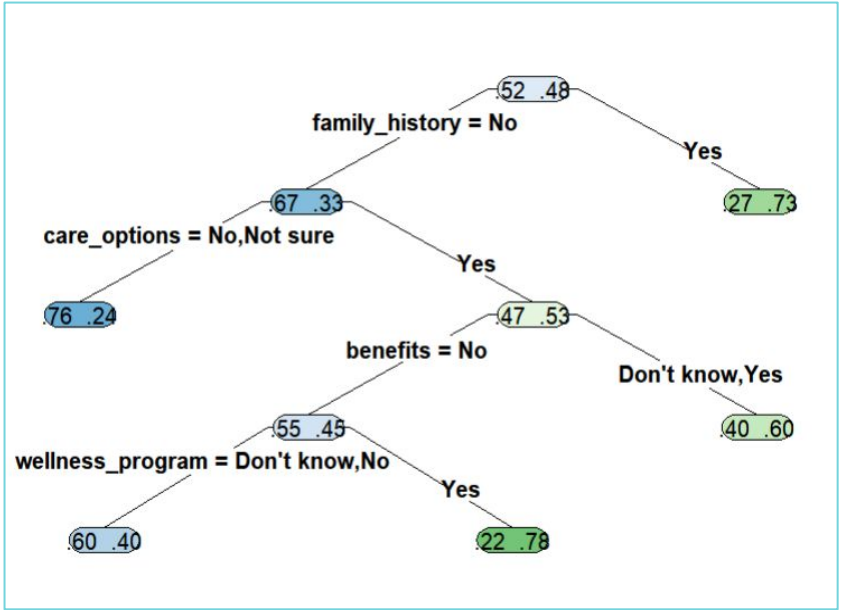
Pre-processing :

- Dropped country, state and comments
- For improved decision tree, dropped physhealthinterview, mentalhealthinterview, mentalvsphysical

Accuracy	Sensitivity	Specificity
68%	0.6807	0.7031

FINAL CART MODEL

Improving Model: remove "mental_interview", "physical_interview" and "mental vs physical" data



Accuracy	Sensitivity	Specificity
69%	0.6807	0.7031

Decisions:

IF family_history = Yes THEN treatment = Yes

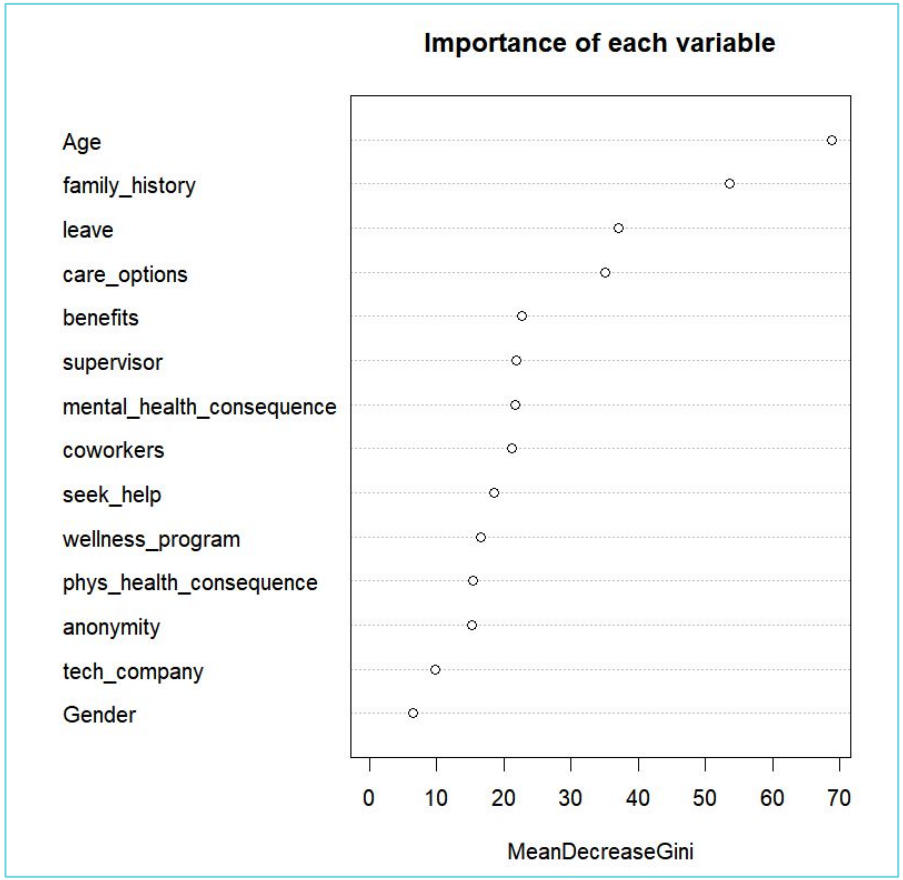
IF family_history = No AND care_options = Yes AND benefits = Don’t know/Yes THEN treatment = Yes

IF family_history = No AND care_options = Yes AND benefits = No AND wellness_program = Yes THEN treatment = Yes

IF family_history = No AND care_options = No/Not sure THEN treatment = 0

IF family_history = No AND care_options = Yes AND benefits = No AND wellness_program = Don’t know/No THEN treatment = No

RANDOM FOREST MODEL



Accuracy	Sensitivity	Specificity
71%	0.6891	0.7344

CART & RANDOM FOREST - BUSINESS INSIGHTS

Among all factors, the following could be main reasons whether it is likely for an employee to seek treatment:

- whether an employee has a family history of mental illness
- whether the employer provide mental health benefits
- Whether the employee knows the options for mental health care

The possible reason for employees with a family history is that they would be more aware of potential mental illness, thus seeking treatment actively.

Employees in companies with mental health benefits and care options could have an increased awareness of mental health among themselves and their co-workers, which could create beneficial intervention that increases engagement and creates an environment of inclusion and support.

COMPARATIVE VIEW OF TRAINING MODELS

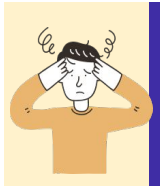
Metric vs Model	Logistic Regression	CART	Random Forest
Accuracy	75%	69%	71%
Sensitivity	0.71	0.68	0.69
Specificity	0.78	0.70	0.73

Logistic Regression is the best suited model for Mental Health dataset analysis.

CHALLENGES

Participation Bias

At least 78.98% of the responses claimed to have a mental illness. This may lead to the inference that there might be biased responses to the survey.



Age and Gender distribution

The data is not representative of general population - the Gender breakdown has significantly more men; and the Age breakdown centers on those aged 20-45, with a big drop-off after that

Regional disparity in responses

The respondents mainly belong to the US. Responses from other regions are very minimal.

OSMI Survey

We are all human.
And we all struggle.

Don't suffer in silence.

Don't feel embarrassed to
ask for help.