Prediction in Mental Health Attitudes in the Tech Industry

Christine Yao
Biomedical Engineering
10/25/24
[Github Link]

School of Engineering
Brown University

Introduction / Problem

Mental Health is a growing problem in the workplace

- Large portion of our time at work, forming connections and earning a living
    - Work affects our mental and physical health and an employee’s productivity and performance.
    - In the United States, healthcare tends to be part of the employer's business. There is a growing need for better mental health resources.
        - Goal: predict→ if people have sought treatment or might need treatment (binary classification problem)

School *of* Engineering
BROWN UNIVERSITY

Data Source

OSMI- Open Sourcing Mental Illness

- Aims to raise awareness, provide support and create an open dialogue about mental health issues in the tech industry by conducting research through surveys
- OSMI encourages open conversations + fights stigma
- I dug into Kaggle and found a dataset on mental health
    - Mental Health in Tech Survey data was collected from the 2014 mental health in tech survey from OSMI
        - Survey answers are usually Yes/No questions, with some open ended questions

Raw Dataset

Rows: 1259 Columns: 27

- Survey translated into CSV format
- Overwhelmingly Yes/No multiple choice questions
  - categorical

	Timestamp	Age	Gender	Country	state	self_employed	family_history	treatment	work_interfere	no_employees	...	leave	mental_health_consequence	phys_health_consequence	coworkers
0	2014-08-27 11:29:31	37	Female	United States	IL	NaN	No	Yes	Often	6-25	...	Somewhat easy	No	No	Some of them
1	2014-08-27 11:29:37	44	M	United States	IN	NaN	No	No	Rarely	More than 1000	...	Don't know	Maybe	No	No
2	2014-08-27 11:29:44	32	Male	Canada	NaN	NaN	No	No	Rarely	6-25	...	Somewhat difficult	No	No	Yes
3	2014-08-27 11:29:46	31	Male	United Kingdom	NaN	NaN	Yes	Yes	Often	26-100	...	Somewhat difficult	Yes	Yes	Some of them
4	2014-08-27 11:30:22	31	Male	United States	TX	NaN	No	No	Never	100-500	...	Don't know	No	No	Some of them

Raw Dataset

```
#datatypes
df.dtypes
```

```
df.isnull().sum()
```

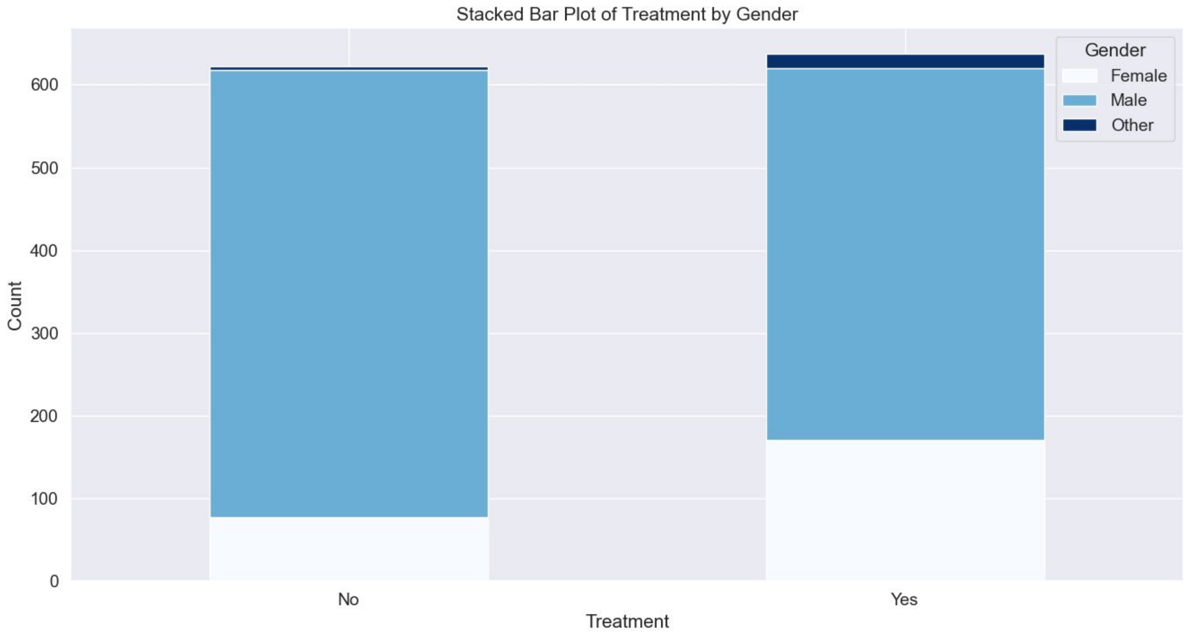
```
Timestamp                     object
Age                            int64
Gender                        object
Country                       object
state                         object
self_employed                 object
family_history                object
treatment                     object
work_interfere                object
no_employees                  object
remote_work                   object
tech_company                  object
benefits                      object
care_options                  object
wellness_program              object
seek_help                     object
anonymity                     object
leave                         object
mental_health_consequence     object
phys_health_consequence       object
coworkers                     object
supervisor                    object
mental_health_interview       object
phys_health_interview         object
mental_vs_physical            object
obs_consequence               object
comments                      object
dtype: object
```

```
Timestamp                       0
Age                             0
Gender                          0
Country                         0
state                         515
self_employed                  18
family_history                  0
treatment                       0
work_interfere                264
no_employees                    0
remote_work                     0
tech_company                    0
benefits                        0
care_options                    0
wellness_program                0
seek_help                       0
anonymity                       0
leave                           0
mental_health_consequence       0
phys_health_consequence         0
coworkers                       0
supervisor                      0
mental_health_interview         0
phys_health_interview           0
mental_vs_physical              0
obs_consequence                 0
comments                     1095
dtype: int64
```

School *of* Engineering
BROWN UNIVERSITY

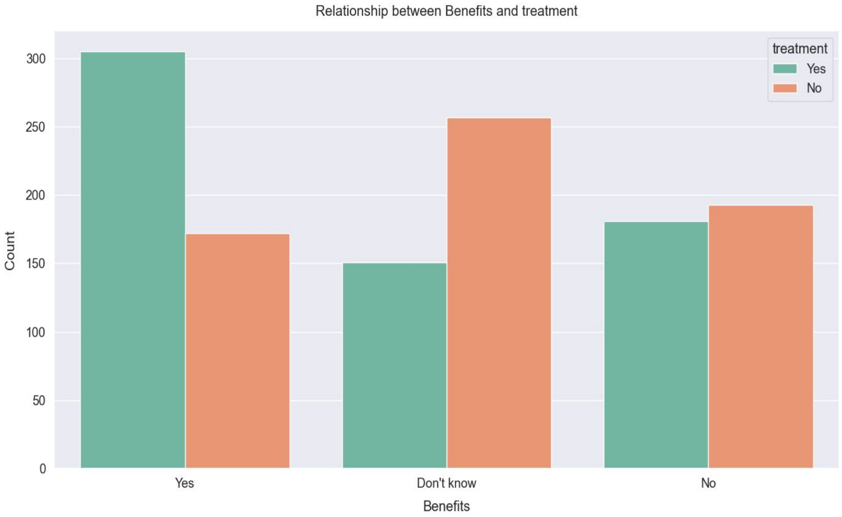
EDA - Target Variable

- Balanced target Variable

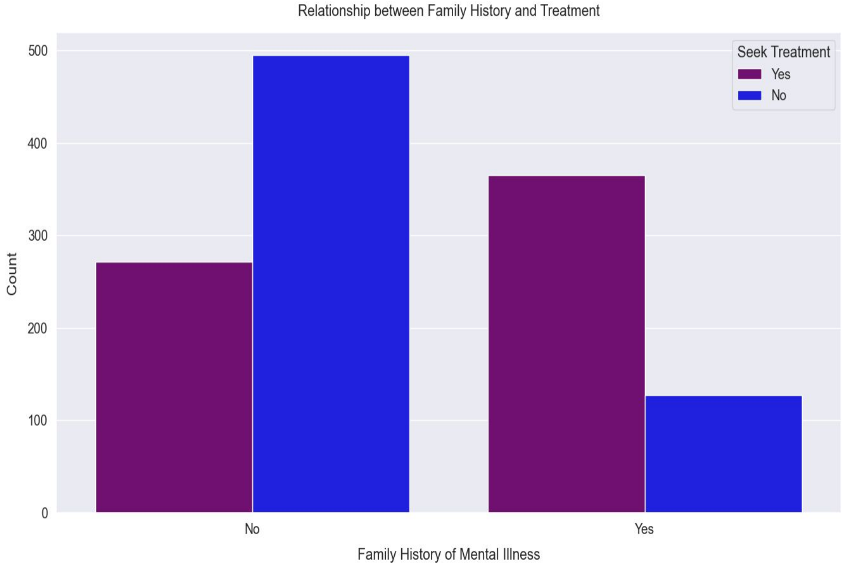


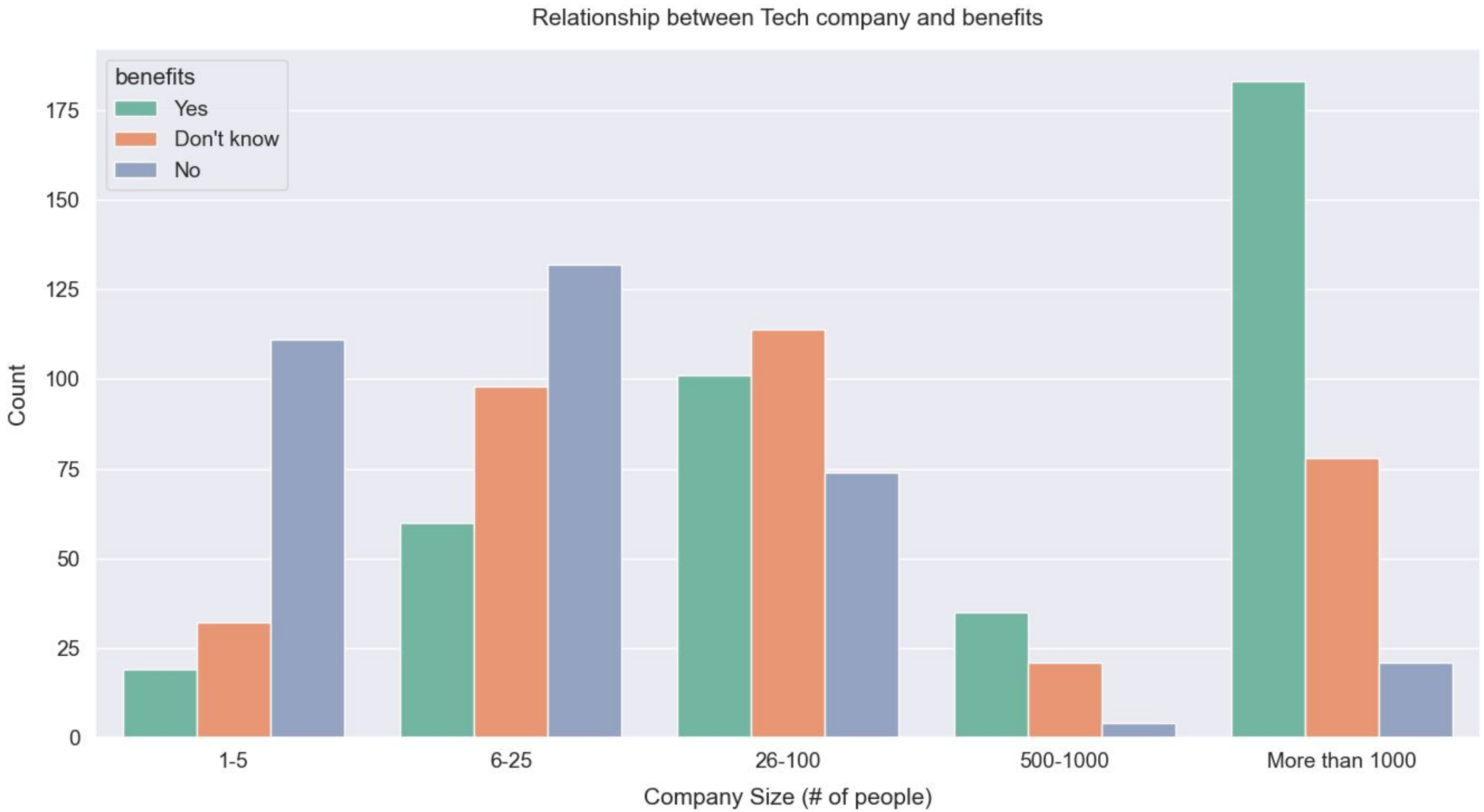
EDA- Interesting Figures

Treatment vs. Benefits



Treatment vs. Family History




Relationship between Tech company and benefits
benefits
Yes
Don't know
No
Count
0
25
50
75
100
125
150
175
1-5
6-25
26-100
500-1000
More than 1000
Company Size (# of people)

Cleaning

Columns Dropped:

1. **Comments, State, Timestamp**

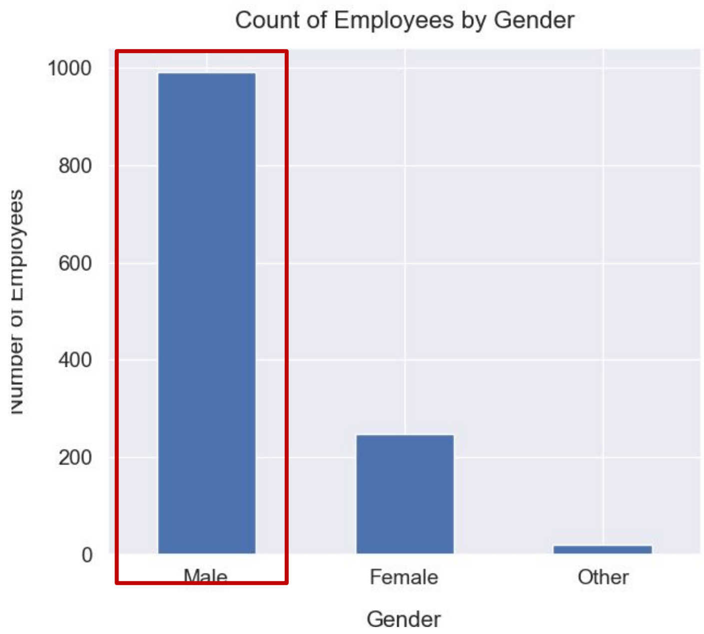
Rows Dropped:

- **Outliers in Age**

Shape of new dataframe: (1251, 24)

Gender - Cleaning

After sorting into Male, Female or Other:

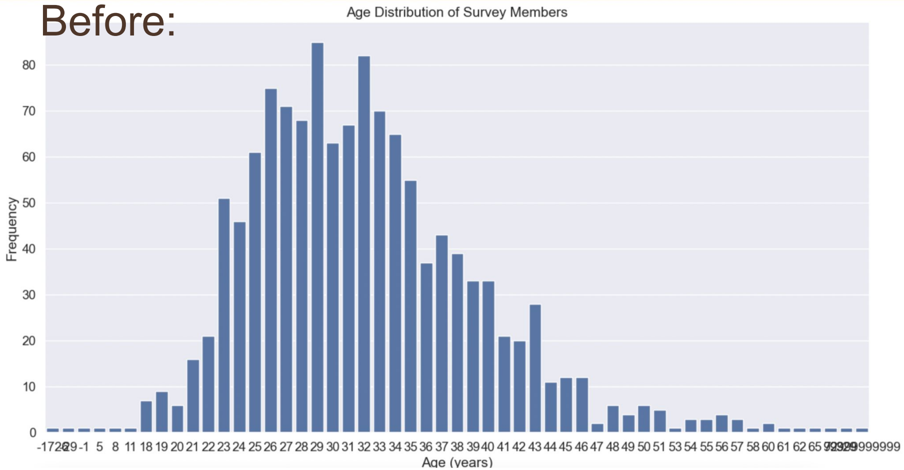


Age- Cleaning

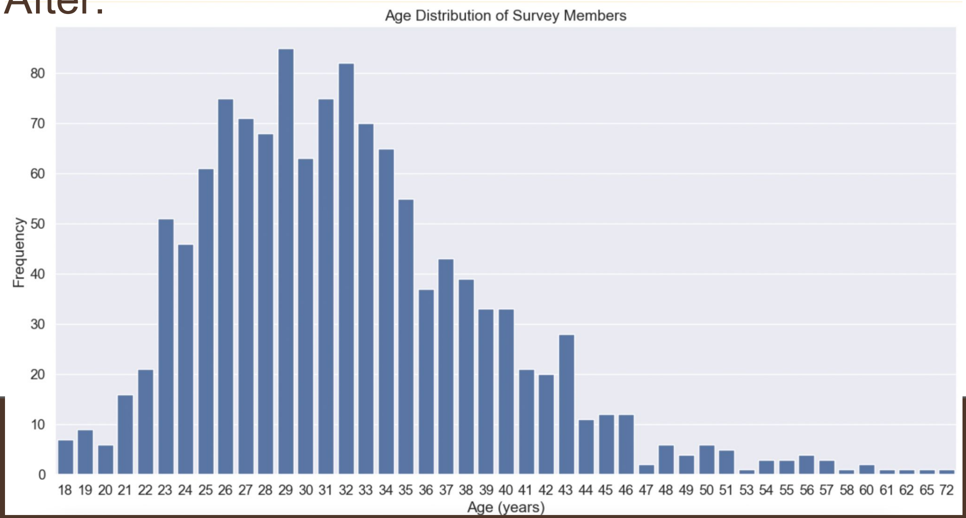
- drop rows >100 and <18
- Most respondents
    - 20-30 range

array([37, 44, 32, 31, 33,
35, 39, 42, 23, 29,
36, 27, 46, 41, 34,
30, 40, 38, 50, 24,
18, 28, 26, 22, 19,
25, 45, 21, -29, 43,
56, 60, 54, 329, 55,
99999999999, 48, 20, 57, 58,
47, 62, 51, 65, 49,
-1726, 5, 53, 61, 8,
11, -1, 72])

Before:



After:



Missing Values

Missing values

- The percentage of missing values in self_employed column is 1.43%
- The percentage of missing values in work_interfere column is 20.97%

- Work_interfere is a feature that only applies to those who identify/ know they have mental illnesses. If they do not, it was a question they were allowed to skip. That is why the missing values is a ⅕ of the dataset. I fixed this through adding a new category called “Unknown”.
- Self employed can be mode imputed

Split Data

- Predict the outcome of ‘treatment’.
  - Treatment is a categorical variable
- Small dataset (< 2000 rows)
- Balanced target variable allowed me to use the basic splitting technique to dataset
  - 80% in train
  - 20% in test
- Shape of split data: Training set: (1000, 23) , Test set: (251, 23) , Validation Set

Preprocessing

One Hot Encoding on categorical features:

```
['Gender','Country', 'self_employed', 'family_history','remote_work',
 'tech_company','benefits' ,'care_options', 'wellness_program','seek_help',
 'anonymity','mental_health_consequence', 'phys_health_consequence',
 'coworkers','supervisor','mental_health_interview','phys_health_interview',
 'Mental_vs_physical','obs_consequence']
```

Ordinal Encoding on ordered categorical features:

```
ordinal_ftrs = ['work_interfere','no_employees','leave']
```

MinMaxScaler on Age

```
Minmax_ftrs = [‘Age’]
```

New Shape after Preprocessing: (1000, 97)

Thank you!

Questions?

School *of* Engineering
BROWN UNIVERSITY