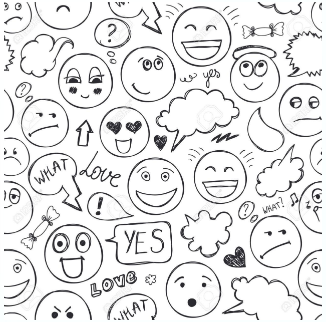
Mental Health in the Tech Sector

Linh-Chi Pham



"**Mental health is one of the least discussed subjects in the corporate world.** The tech industry adds another layer of complexity when taking into consideration the mental health discussion. The tech industry fosters a 'crunch' culture where demanding work must be completed in a short amount of time. The industry is known for high-stress: late nights, abnormal hours, and tight deadlines, all while being constantly available at any time of day."

-Naveen Bhateja, Chief People Officer of Medidata Solutions.

Background

- The CDC reports that **1 in 5 Americans** will experience a mental illness in a given year
- Mental health is an important component of overall health.
  + For example, depression increases the risk for many types of physical health problems, particularly long-lasting conditions like diabetes, heart disease, and stroke
- Mental health in the workplace is beginning to be more talked about
  + Mental wellbeing is an essential component of a healthy and effective workplace, particularly in fast-paced and high-growth sectors of the economy like tech

Our Research Question:

- What contributes most to the average mental health condition of tech workers?
- Using survey data, can we accurately predict if one has a current diagnosis for a mental illness?

Goals:

- Examine the prevalence of mental health disorders among tech workers
- From the final model built, propose suggestions on how to improve the workplace environment of the Tech/IT sector

Our Dataset

- Open Sourcing Mental Health (Formerly OSMI) is a campaign founded by Ed Finkler to change how we deal with mental health in the tech community.
- Online survey that aims to measure attitudes towards mental health in the tech workplace, and examine the frequency of mental health disorders among tech workers.
- 2016 survey has over 1400 responses

Data Wrangling & Transformation

Original **survey data**: 1433 observations, 64 variables/questions; Mostly **categorical**

1. **Clean + prepare data**
 - Filter out most relevant 20 predictors
 - Subset just US
 - Change variable names
 - Group similar unique answers into a few main categories (i.e. Gender, State)

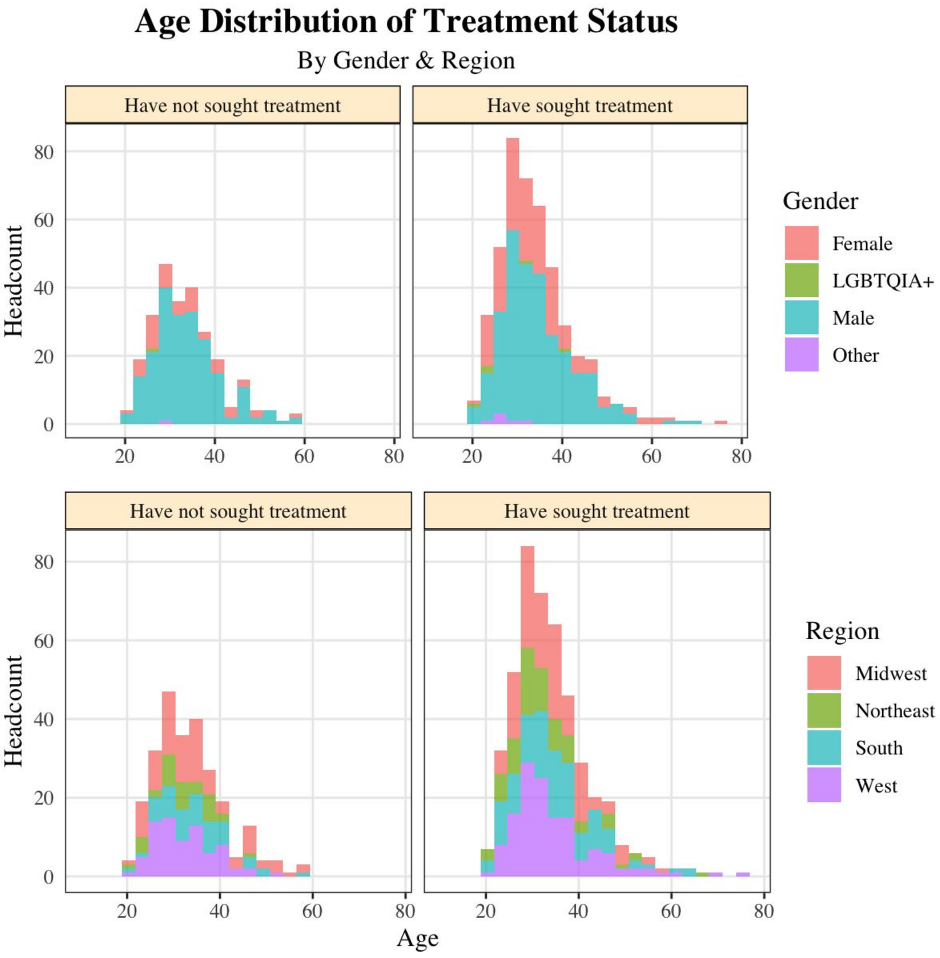
```
male <- c("male","male ","m","man","cis male","male.","male (cis)
          "mail","ml","male/genderqueer","cisdude","cis man")
female <- c("female", "f","i identify as female.","female ","cisg
            "female (props for making this a freeform field, thou
            "female assigned at birth ","woman","fm","cis female
lgbtqia <- c("non-binary","bigender","non-binary","transitioned,
             "other/transfeminine","female/woman","androgynous","
             "genderfluid","enby","mtf","queer", "agender","fluid
other <- c("n/a","other","none of your business","genderqueer","h
```

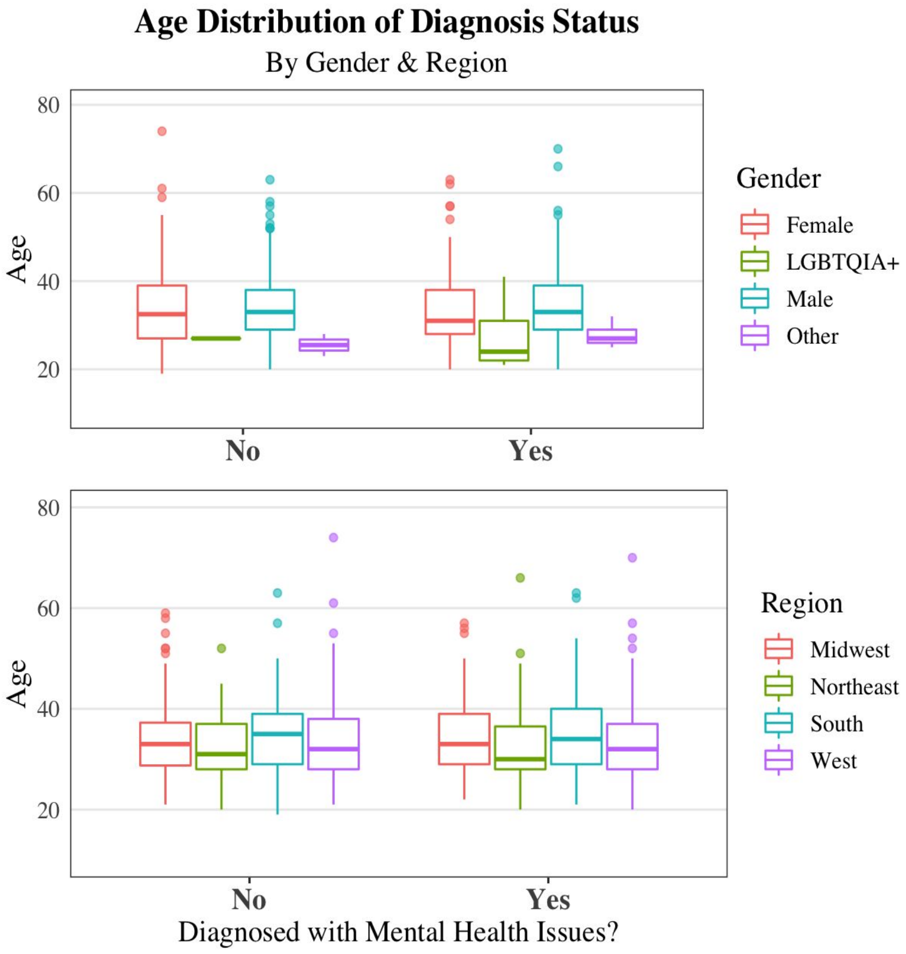
1. **One-hot encoding**
 - Convert categorical variables to numerical, binary variables that takes value of 1 and 0, which can be used for ML algorithms

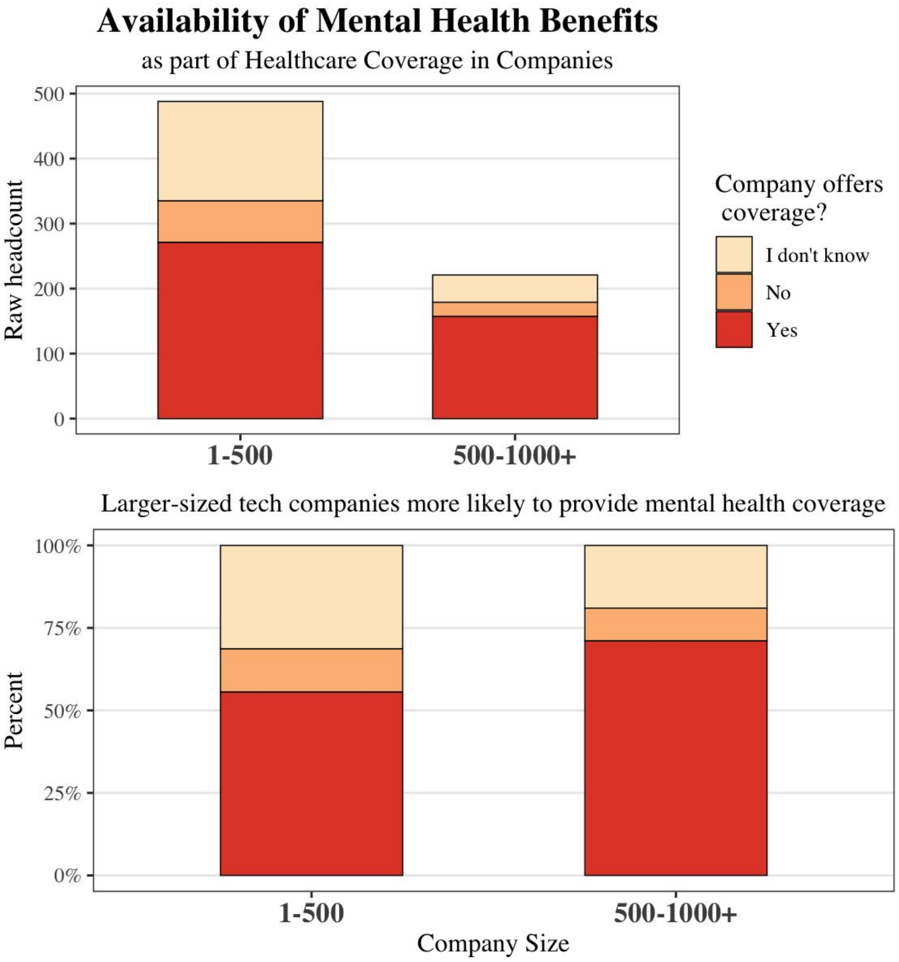
id	color
1	red
2	blue
3	green
4	blue

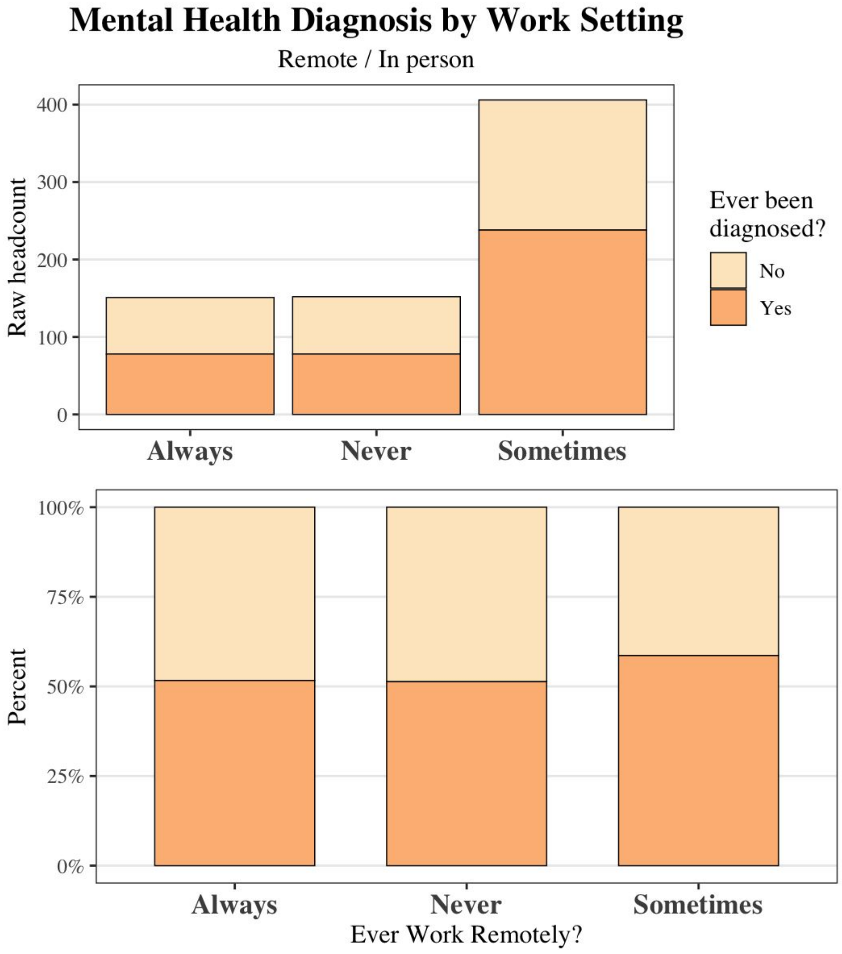
One Hot Encoding

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0


Age Distribution of Treatment Status
By Gender & Region
Have not sought treatment
Have sought treatment
Headcount
Gender
Female
LGBTQIA+
Male
Other
Have not sought treatment
Have sought treatment
Headcount
Region
Midwest
Northeast
South
West
Age


Age Distribution of Diagnosis Status
By Gender & Region
Age
No
Yes
Gender
Female
LGBTQIA+
Male
Other
Age
No
Yes
Diagnosed with Mental Health Issues?
Region
Midwest
Northeast
South
West


Availability of Mental Health Benefits
as part of Healthcare Coverage in Companies
Raw headcount
0
100
200
300
400
500
1-500
500-1000+
Company offers coverage?
I don't know
No
Yes
Larger-sized tech companies more likely to provide mental health coverage
Percent
0%
25%
50%
75%
100%
1-500
500-1000+
Company Size


Mental Health Diagnosis by Work Setting
Remote / In person
Raw headcount
0
100
200
300
400
Always
Never
Sometimes
Ever been diagnosed?
No
Yes
Percent
0%
25%
50%
75%
100%
Always
Never
Sometimes
Ever Work Remotely?

Statistical Modeling

- train:test ratio **80:20**
- 20 variables used to model:
- Response: current_diagnosis

```
> names(us.16.model)
 [1] "no_employees"          "tech_company"            "benefits"              "care_options"
 [5] "wellness_program"      "seek_help"               "anonyimity"            "leave"
 [9] "coworkers"             "supervisor"              "mental_vs_physical"    "obs_neg_consequence"
[13] "physhealth_interview"  "mentalhealth_interview"  "hurt_career"           "views"
[17] "willing_share"         "family_history"          "current_diagnosis"     "treatment"
```

1. Feature Selection → **2.** Logistic Regression → **3.** Neural Network (classification)

I. Feature Selection

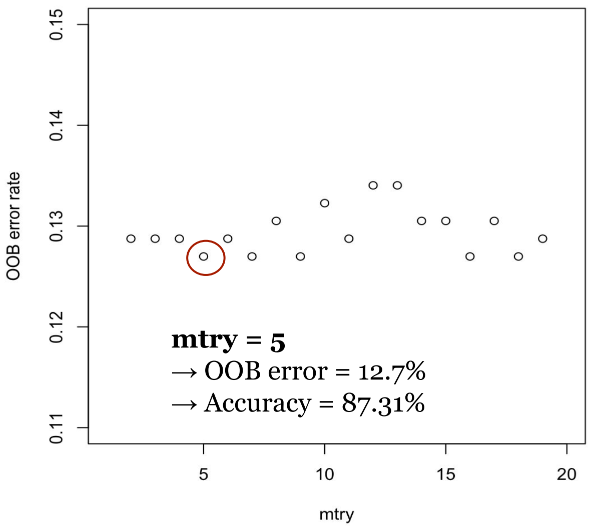
1. RANDOM FOREST

a. Default mtry = 4, ntree = 500

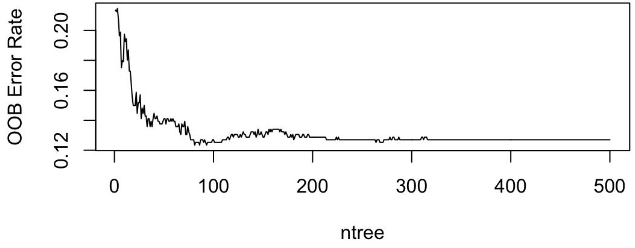
```
        OOB estimate of  error rate: 12.87%
Confusion matrix:
     No Yes class.error
No  191  60  0.23904382
Yes  13 303  0.04113924
```

b. Tune parameter **mtry**

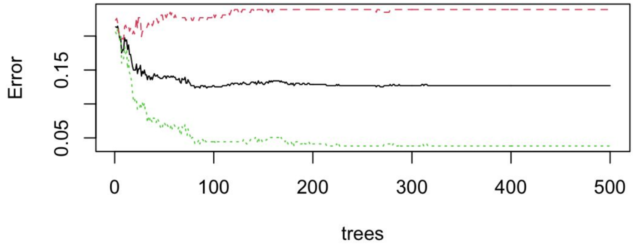
Choose best mtry by comparing OOB error rate



OOB error rate is stabilized at ntree = 500



Random forest model using ntree = 500, mtry = 5

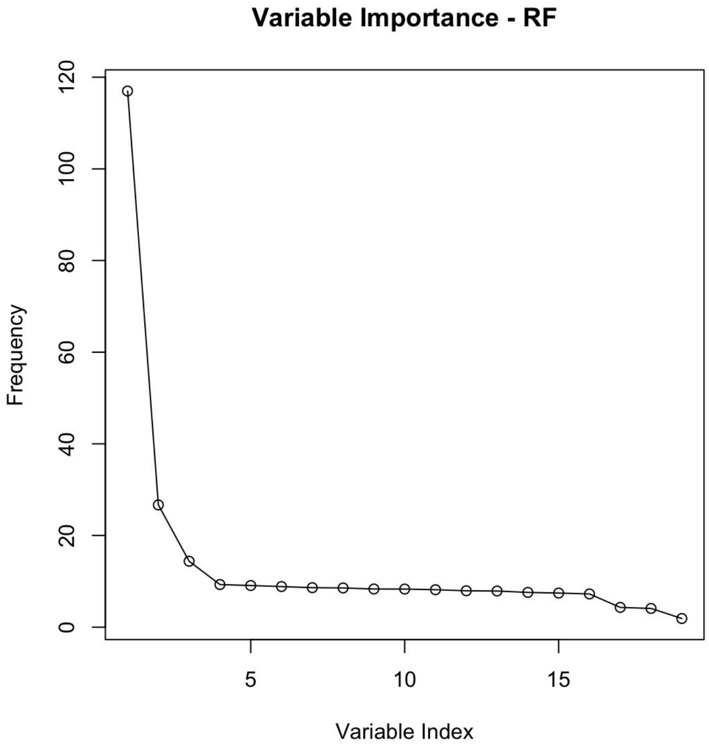



```
randomForest(formula = current_diagnosis ~ ., data = train, mtry = 5)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 5

        OOB estimate of  error rate: 12.7%
Confusion matrix:
     No Yes class.error
No  191  60  0.23904382
Yes  12 304  0.03797468
```

I. Feature Selection

Variable Importance (Gini Index)



	Predictor	Importance
1	treatment	116.989601
2	family_history	26.670507
3	care_options	14.377565
4	seek_help	9.299166
5	leave	9.081598
6	wellness_program	8.856187
7	physhealth_interview	8.610474
8	mental_vs_physical	8.552960
9	views	8.324414
10	supervisor	8.308691
11	willing_share	8.167052
12	coworkers	7.940869
13	anonyimity	7.883522
14	mentalhealth_interview	7.578577
15	benefits	7.437145
16	hurt_career	7.243354
17	no_employees	4.308412
18	tech_company	4.082824
19	obs_neg_consequence	1.883628

Model Rf2: 5 predictors

Model Rf3: 10 predictors

Model Rf1: All 19 predictors

I. Feature Selection

2. STEPWISE REGRESSION

a. Backward Selection: AIC 392.51

```
Call:
glm(formula = current_diagnosis ~ care_options + wellness_program +
    anonyimity + family_history + treatment, family = "binomial",
    data = train)
```

→ Model #4 back_mod

b. Forward Selection / Both directions: **AIC 391.87**

```
Call:
glm(formula = current_diagnosis ~ treatment + family_history +
    care_options + anonyimity + mental_vs_physical, family = "binomial",
    data = train)
```

→ Model #5 forward_both

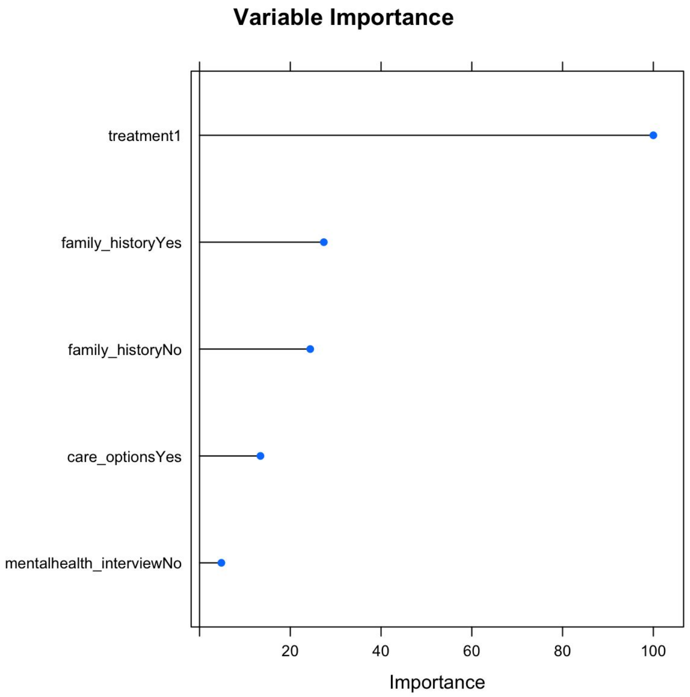
=> Both chose 5 predictors, AIC doesn’t differ too much

=> Try both models in logistic regression

I. Feature Selection

3. Recursive Partitioning

- Recursively partitioning the explanatory variables into the "purest groups" of two of the levels of the response variable
- Assessment of the purity of the resulting groups is decided by any number of metrics, but the most commonly used metric is the gini index.



	Overall
treatment1	100.000
family_historyYes	27.390
family_historyNo	24.400
care_optionsYes	13.425
mentalhealth_interviewNo	4.819

Model#6 Rpart

II. Logistic Regression

	Rf1	Rf2	**Rf3**	S1	**S2**	Rpart
Number of predictors	19	5	**10**	5	**5**	4
AIC	423.45	403.07	409.99	392.51	391.87	397.61
AUC	0.926	0.906	0.914	0.917	0.918	0.907
Test error rate	11.97%	13.38%	11.97%	12.68%	11.97%	11.97%
Accuracy	88.03%	86.62%	88.03%	87.32%	88.03%	88.03%

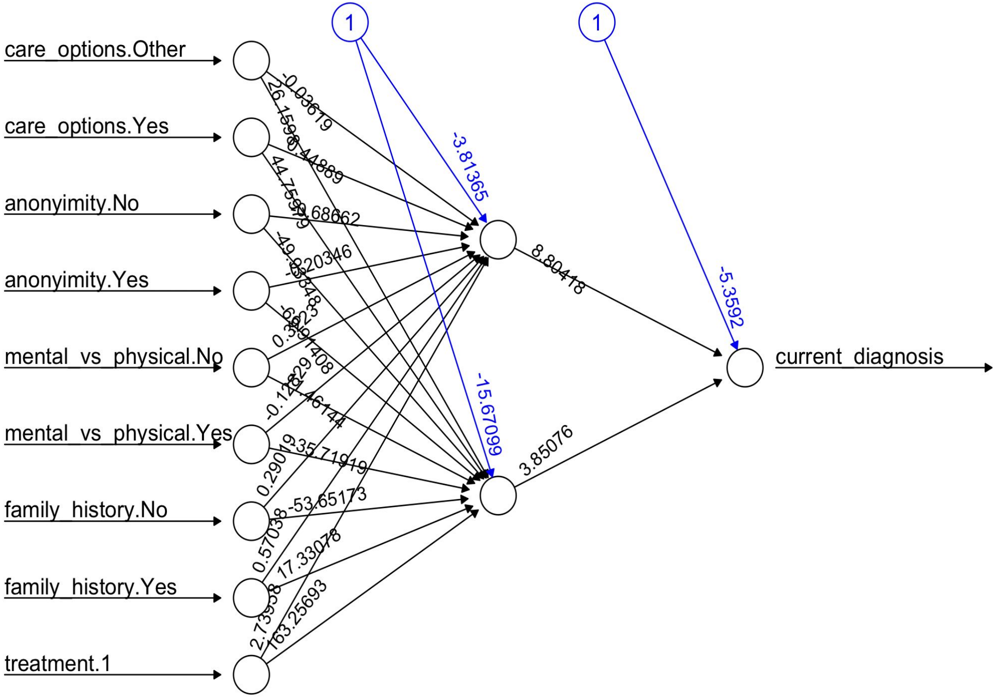
"Would a simple model with few predictors outperform a more complicated one with twice the number of predictors?"

Finalists

to be used for Neural Net

III. Classification Neural Network

Model S2: Forward Selection (5 variables)



- Hidden layers: 1
- Number of neurons per layer: 2

Reasoning:

- 2 hidden layers is more than enough. In this case adding 1 more hidden layer does not improve neuralnet performance → **1 layer**
- Number of neurons should usually be **2/3 of the input size** → **2 neurons** (3 performs worse)

Confusion Matrix:

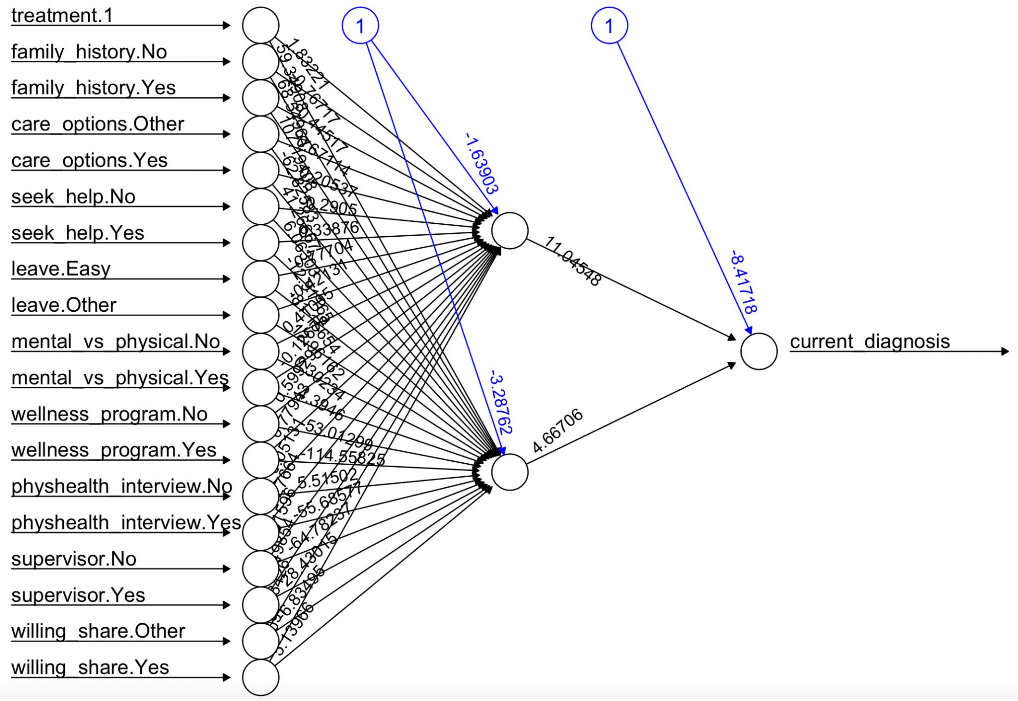
	FALSE	TRUE
FALSE	50	14
TRUE	4	74

→ Test error rate: **12.68%**

→ Accuracy: **87.32%**

III. Classification Neural Network

Model Rf3: Random Forest (10 variables)



- Hidden layers: 1
- Number of neurons per layer: 2

Reasoning:

- Adding 1 more hidden layer does not improve neuralnet performance → **1 layer**
- Number of neurons should usually be **2/3 of the input size** → **2 neurons** (3 performs worse)

Confusion Matrix:

	FALSE	TRUE
FALSE	51	13
TRUE	8	70

→ Test error rate: **14.79%**

→ Accuracy: **85.21%**

Final Result

	Model S2: Forward Selection	Model Rf3: Random Forest
Number of Predictors	**5**	10
Test error rate	**12.68%**	14.79%
Accuracy	**87.32%**	85.21%

→ **Model S2** with predictors chosen using forward selection is the best-performing model

→ **Final model** for predicting diagnosis status of Tech workers:

current_diagnosis ~ treatment + family_history + care_options + anonymity + mental_vs_physical

highest significance

Model Interpretation

Final model for predicting diagnosis status of Tech workers:

current_diagnosis ~ **treatment + family_history + care_options + anonymity + mental_vs_physical**

1. **treatment**: Have you sought treatment for a mental health condition?

   → only those with a current diagnosis would seek treatment but not necessarily all

2. **family_history:** Do you have a family history of mental illness?

   → Mental disorders are the result of both genetic and environmental factors. But family history can provide an increased risk for developing a mental illness.

3. **care_options**: Do you know the options for mental health care your employer provides?

   → those with current diagnoses may be more likely to know the options since they may be using them or plan to

4. **anonymity**: Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?
5. **mental_vs_physical:** Do you feel that your employer takes mental health as seriously as physical?

   → those with current diagnoses probably choose to stay at places where they feel comfortable and safe

Conclusions

- The results from the survey and our modeling highlights the prevalence of mental illness, especially in the tech sector.
- Tech companies should be making more effort to support mental health issues like they do with physical issues.
- We only had time to focus on a small portion of this survey, there is a lot more information to analyze.
- Future years should be analyzed and see if there are any trends
- Convenience sample may not be representative of the whole population

References

https://www.techtimes.com/articles/271446/20220204/mental-health-an-important-conversation-in-the-tech-industry.htm

https://www.cdc.gov/mentalhealth/learn/index.htm

Thank you! Questions?