Validating The Association Between
The Tech Workplace & Mental Illness

I. ABSTRACT

The Tech sector has been and will continue to be one of the most relevant and disruptive industries in the world; and a Tech career is oftentimes marketed with “high pay” and “great career prospects” as the main keywords. However, with its high-stress nature and tight deadlines, more often than not, working in Tech is associated with increased likelihood of developing mental health issues. This paper strives to validate this association by (a) examining the prevalence of mental health disorders among Tech workers, (b) identifying the most significant contributors to the average mental health conditions in the Tech workplace through Feature Engineering (Stepwise Regression, Random Forest, Recursive Partitioning), and (c) predicting if a Tech worker is currently diagnosed with a mental illness with classification models (Logistic Regression, Neural Network).

Keywords: Tech Industry, Mental Health, Sustainable Productivity, Feature Engineering, Classification, Statistical Machine Learning

II. BACKGROUND & SIGNIFICANCE

At Amazon Web Services, the “on-call” rotational procedure expects software developers to get paged and be available any time during the day, even 2am, to resolve troubled tickets in only 15 minutes. As reported by employees on Blind, taking on this “on-call” responsibility is not compensated by extra pay, which further adds on to the existing stress and exhaustion. Stories like these have led many to naturally associate working in Tech with higher chances of developing mental illness.

On the other hand, there are well-known perks to working in Tech. According to *mthree*, “Most (Tech) companies won’t usually keep you tied to a strict 9-to-6 schedule like most other employees. Tech industry has a stimulating culture that isn’t just based on work, with great opportunities to show off your skills and socialize with like-minded people.”

Recognizing these differentiating opinions, we are interested in assessing the connection (if any) between Tech workplace and likelihood of developing mental illness. Through our final machine learning model, we aim to use statistically backed-up results to answer the following questions:

- How prevalent are mental health disorders in the Tech workplace?
- Does working in the Tech industry increase one’s likelihood of developing mental health illness?
- Using survey data, can we accurately predict if one has a current diagnosis for a mental illness?

III. METHODOLOGY

Dataset & Source

To answer our research questions, we rely on 2016 data from Open Sourcing Mental Health (formerly OSMI), a campaign founded by Ed Finkler to change how we deal with mental health in the Tech community. This campaign collects data in the form of an open-answer survey, whose aim is to measure attitudes towards mental health in the Tech workplace, and examine the frequency of mental health disorders among Tech workers. The original dataset has around 1500 observations and 64 variables (mostly categorical in the form of long survey questions).

Variable Creation

To prepare the data, we first filtered out the most relevant 20 variables as predictors, subsetted to just the US population (as other countries have very little observations with many missing values) and changed the variable names. Since the survey allows for free-form answers, there is a great amount of variation in responses for specific categories such as gender, region, work setting, etc. Our optimal solution was to look at all unique values of the original survey question and recategorize them into more concise and comprehensible subcategories (female/male/lgbtqia+). We paid special attention to keeping the number of subcategories of each factor variable as low as possible yet still informative enough to avoid over-complication in our next data wrangling method - one-hot encoding. Later on in the modeling process, we utilized the neuralnet package in R to build a classification neural network, however the challenge presented at hand is that this algorithm only takes in numerical inputs. To satisfy this requirement, we followed the one-hot encoding process and converted our old categorical variables into numerical/binary variables that take values of 1 and 0. For example, our original family_history (Do you have a family history of mental illness) variable with 3 factorized subcategories Yes/No/Other was split into 3 binary variables: family_history.Yes, family_history.No and family_history.Other.
Our response variable is current_diagnosis (Have you been diagnosed with a mental health condition by a medical professional?) with 2 factor levels: Yes/No. The final clean dataset used for modeling has a total

of 19 predictors, all of which are categorical and have no more than 3 factor levels. The full list and explanation of our final dataset variables can be found in *(Appendix: Table 1)*.

Modeling

Our statistical analysis process can be summarized as follows: Feature Engineering -> Logistics Regression -> (Classification) Neural Network. Each of the different models used the exact same training set and testing set, with 80% of the Tech workers placed in the train set and the remaining 20% placed in the test set.

In the first step, we used 3 different types of algorithms to perform Feature Selection. The first one is Permutation Feature Importance, which is used to determine the effects of the variables in the Random Forest model. We first built a Random Forest model *(Appendix: Graph 3)* with default parameters ntree and mtry, then proceeded to tune these parameters using out-of-bag error rate as the selection criteria. With the tuned Random Forest model, we used the *varImp* function *(Appendix: Graph 4)* in the randomForest package to calculate variable importance, which by default scales the measures of importance up to 100. This method calculates the increase in the prediction error (MSE) after permuting the feature values. If the permuting doesn't change the model error, the related feature is considered unimportant. The second set of algorithms we used to select features were from Stepwise Regression, including Backward, Forward and Both directions Selection. The last Feature Engineering method we used was Recursive Partitioning, which places 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 ranking the Gini index.

In our second step, we built Logistic Regression models using different sets of predictors we got from Feature Selection to assess how each performs in classifying current_diagnosis *(Appendix: Table 2)*. Given the fact that our sets of predictors are not too drastically different from each other, oftentimes only differing by 1-2 predictors, we used 3 main selection criteria: Test error rate/Accuracy (for preliminary elimination), AUC (for added precision - how likely the model is to make accurate classification), and finally, the number or predictors (interpretability of the models).

For the final step, we built Classification Neural Networks *(Appendix: Graph 7, 8)* from the best–performing sets of predictors based on Logistic Regression assessment metrics. As mentioned above, to run this neuralnet algorithm, we transformed all categorical predictors into binary variables (one-hot encoding). The splitting process caused the number of predictors to increase exponentially, as for each level there would be a new variable. To avoid over-complicating the inputs of our Neural Networks, we used p-value of predictors from Feature Selection as the elimination criteria, and removed insignificant predictors (with high p-values/ > 0.05 significance level). We chose 1 hidden layer and 2 neurons (per layer) as the tuning parameters.

The final predictor set is chosen by comparing the test error rate (computed from confusion matrix) of the Neural Network models *(Appendix: Table 3)*.

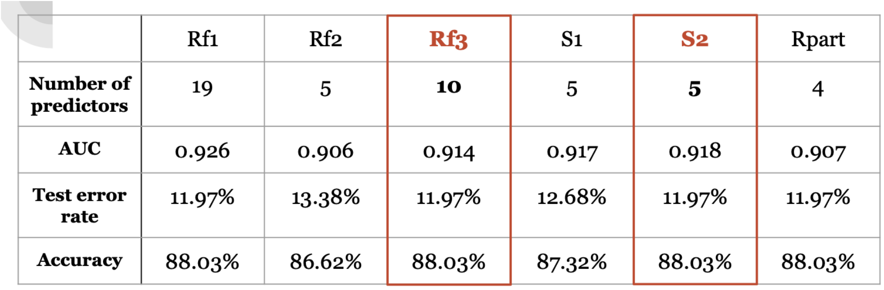
IV. RESULTS

1. *Feature Selection*

From Feature Importance in Random Forest, we were able to rank predictors on a 100 scale and extracted 3 sets of predictors: Rf1 (all 19 predictors), Rf2 (5 most important predictors) and Rf3 (10 most important predictors). From Stepwise Regression, we had 2 sets of predictors: S1 from Backward Selection, S2 from

Forward Selection & Both directions (both yielded the same result). From Recursive Partitioning, we had 1 last set of predictors Rpart. Notably, *treatment*, *family_history* and *care_options* consistently rank as the most important variables from all 3 Feature Selection methods discussed above.

2. *Logistic Regression*



	Rf1	Rf2	Rf3	S1	S2	Rpart
Number of predictors	19	5	**10**	5	**5**	4
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%

This table *(Appendix: Table 2)* summarizes the performance of all 6 sets of predictors from Feature Selection. We compared the models by group (Rf/S/Rpart), then placed the best-performing models of 3 groups against each other. Specifically, Rf3 was chosen to represent Random Forest group, as it has significantly less variables than Rf1 while having higher AUC and accuracy then Rf2. S2 outperforms S2 both in terms of AUC and accuracy/test error rate. Rpart does not perform as well as Rf3 and S2. Thus, we have our finalists sets of predictors to be used for Neural Networks: Rf3 and S2.

3. *Classification Neural Networks*

The tuning parameters we chose for our neural networks are 2 neurons in 1 hidden layer. According to Jeff Heaton, **one hidden layer** allows a neural network to approximate any function involving “a continuous mapping from one finite space to another.” With 2 hidden layers, the network is able to “represent an arbitrary decision boundary to arbitrary accuracy.” The most commonly used neuron:input size ratio is ⅔, and we originally have 5 variables (before one-hot encoding). Our strategy was to try building neural networks with 2 and 3 neurons for 1 hidden layer, and did the same for 2 hidden layers. The simplest neural network with 1 hidden layer and 2 neurons performs the best for both Rf3 and S2.

	Rf3	**S2**
Test Error Rate	14.79%	12.68%
Accuracy	85.21%	87.32%

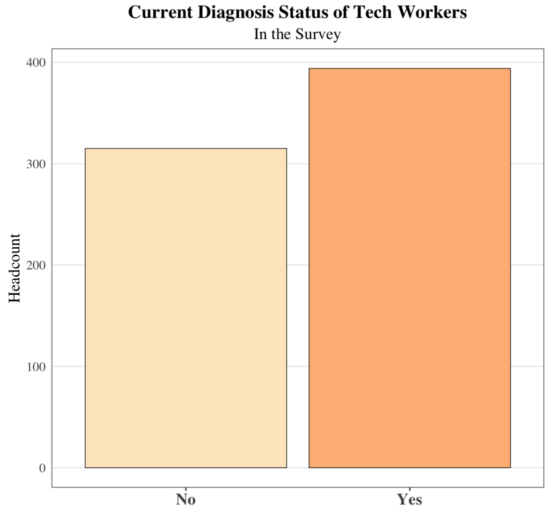
This table *(Appendix: Table 3)* compares the performance of 2 neural networks built from 2 sets of predictors Rf3 and S2. S2 outperforms Rf3 with lower test error rate and thus higher accuracy.

After 2 elimination rounds, models built with Forward Selection (S2) predictors perform the best. Our final set of predictors, ranked in descending order of significance, is as follows: treatment, family_history, care_options, anonymity, mental_vs_physical.

V. DISCUSSION & CONCLUSION

In this section, we address each of our research questions using the results from our exploratory data analysis and statistical modeling processes.

1. How prevalent are mental health disorders in the Tech workplace?



There are more Tech workers diagnosed with a mental health illness in the survey than those who are not diagnosed. *(Appendix: Graph 1)* looks more closely at the age distribution of Tech workers by gender and region, finding that most of the survey participants are in their 20s-40s and work in the Midwest region. Overall, women tend to seek treatment for mental health issues more so than men; however in terms of diagnosis status, men and women are equally likely, given little to no difference between the Female and Male boxplots. We excluded gender, age and region from our model to avoid bias/discrimination, and these results support our decision.

2. Does working in the Tech industry increase one's likelihood of developing a mental health illness?
Our final set of predictor variables include: treatment, family_history, care_options, anonymity and mental_vs_physical. However, treatment, family_history, and care_options consistently rank as the 3 most significant predictors from all Feature Selection methods.
To further explain our 3 main predictors: (a) In terms of treatment, only those with a current diagnosis would seek treatment but not necessarily all. (b) Mental disorders are the result of both genetic and environmental factors. But family history can provide an increased risk for developing a mental illness. (c) Those with current diagnoses may be more likely to know the options since they may be using them or plan to.
Overall, these 3 predictors lean more towards the background, history and awareness of Tech workers themselves, rather than the characteristics of their company. This means that working in the Tech industry/ the Tech workplace is not necessarily correlated with higher/lower chances of being diagnosed with a mental health issue. From our analysis, we thus do not have strong evidence to support the assumption that the high-stress nature of the Tech industry itself is solely to blame for increased likelihood of developing mental illness.

3. Using survey data, can we accurately predict if one has a current diagnosis for a mental illness?
For Logistic Regression, no model yields lower than 87% prediction accuracy. Area under the curve (AUC) of all models are all higher than 0.9, indicating strong probability of these models to produce accurate predictions. For Neural Networks, both 2 models have higher than 85% accuracy, with the better-performing one yielding approximately 87% accuracy, which is a solid statistic. Our final predictors (treatment, family_history, and care_options) are also consistent with the findings of this study, which lends credibility to our work.

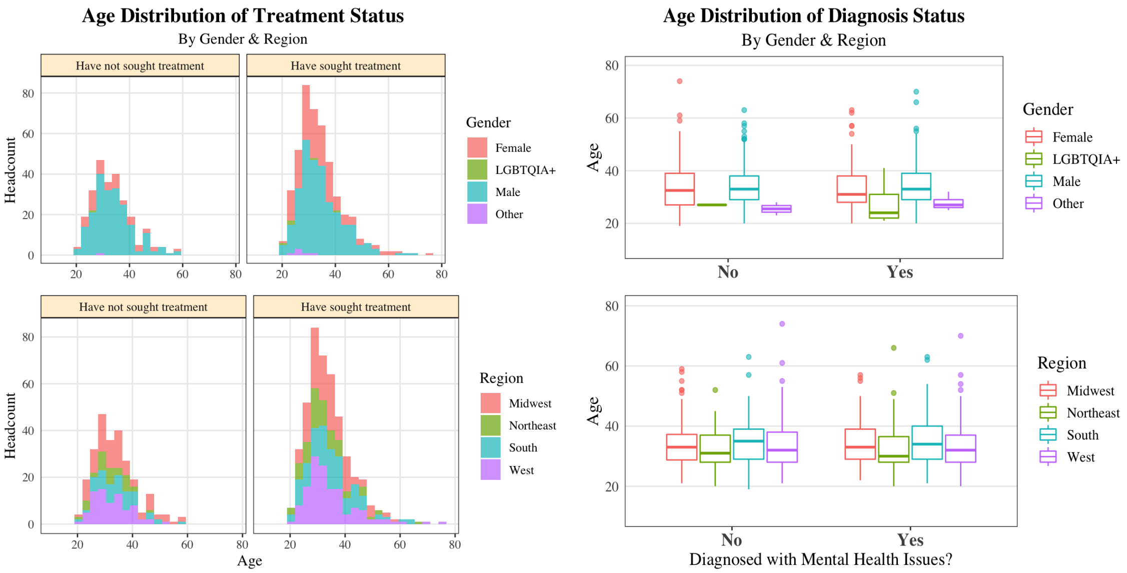
APPENDIX

Table 1: Meaning and levels of 20 variables used in this study

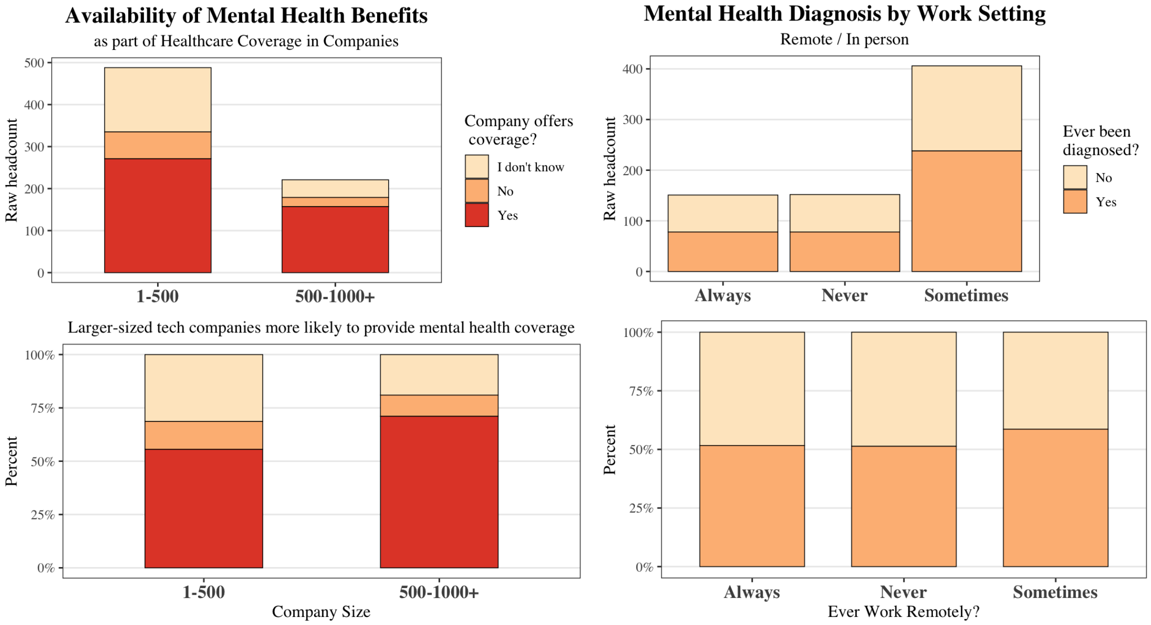
Variable	Meaning	Levels
current_diagnosis	Have you been diagnosed with a mental health condition by a medical professional?	Yes; No
no_employees	How many employees does your company or organization have?	1-500; 500-1000+
tech_company	Is your employer primarily a tech company/organization?	0; 1
benefits	Does your employer provide mental health benefits as part of healthcare coverage?	Yes; No; I don't know
care_options	Do you know the options for mental health care available under your employer-provided coverage?	Yes; No; Other
wellness_program	Has your employer ever formally discussed mental health (for example, as part of a wellness campaign or other official communication)?	Yes; No; I don't know
seek_help	Does your employer offer resources to learn more about mental health concerns and options for seeking help?	Yes; No; I don't know
anonymity	Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources provided by your employer?	Yes; No; I don't know
leave	If a mental health issue prompted you to request a medical leave from work, asking for that leave would be:	Difficult; Easy; Other
coworkers	Would you feel comfortable discussing a mental health disorder with your coworkers?	Yes; No; Maybe
supervisor	Would you feel comfortable discussing a mental health disorder with your direct supervisor(s)?	Yes; No; Maybe
mental_vs_physical	Do you feel that your employer takes mental health as seriously as physical health?	Yes; No; I don't know
obs_neg_consequence	Have you heard of or observed negative consequences for co-workers who have been open about mental health issues in your workplace?	Yes; No
physhealth_interview	Would you be willing to bring up a physical health issue with a potential employer in an interview?	Yes; No; Maybe
mentalhealth_interview	Would you bring up a mental health issue with a potential employer in an interview?	Yes; No; Maybe
hurt_career	Do you feel that being identified as a person with a mental health issue would hurt your career?	Yes; No; Maybe

views	Do you think that team members or coworkers would view you more negatively if they knew you suffered from a mental health issue?	Yes; No; Maybe
willing_share	How willing would you be to share with friends and family that you have a mental illness?	Yes; No; Other
family_history	Do you have a family history of mental illness?	Yes; No; I don't know
treatment	Have you sought treatment for a mental health condition?	0; 1

Graph 1: Age distribution of treatment & diagnosis status, categorized by gender & region

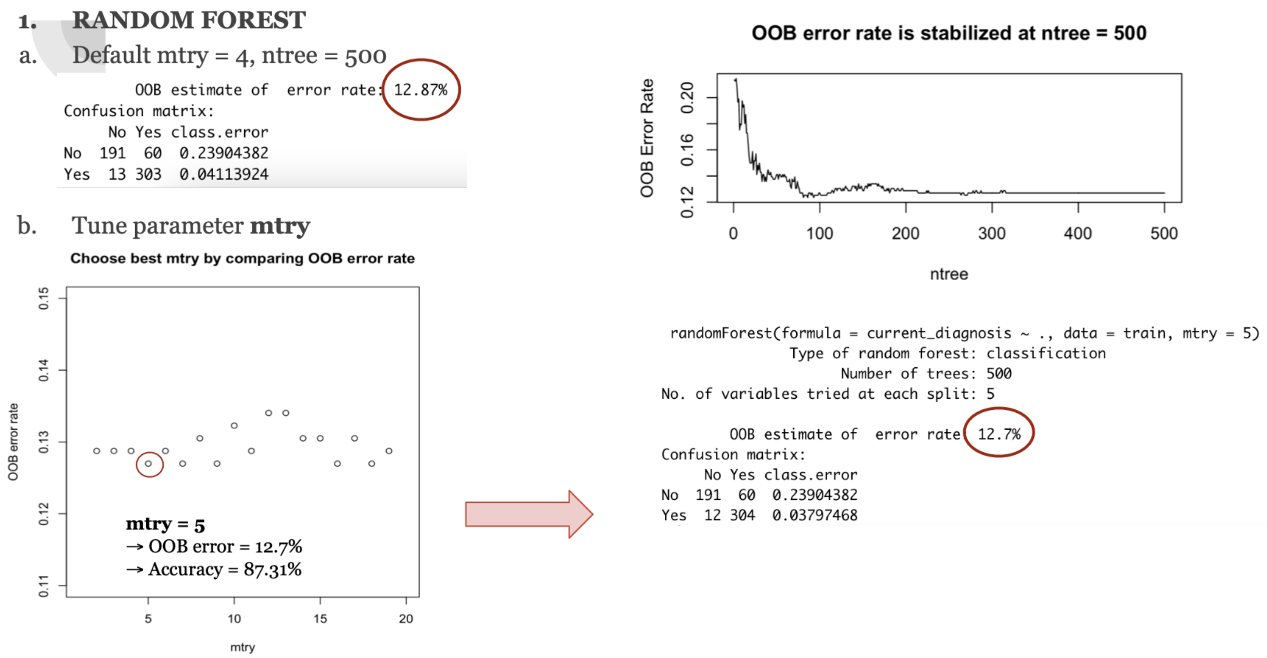


Graph 2: Company mental health coverage & work setting



In the survey, there were twice as many workers at the 1-500+ company as those at the 500-1000+ range. Larger-sized Tech companies are more likely to provide mental health benefits as part of the standard *healthcare coverage. In terms of work setting, most of the survey respondents work in the hybrid format (in-person and sometimes remote). However, there is roughly an equal likelihood of being diagnosed with a mental health illness.*

Graph 3: Building & tuning random forest



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
```

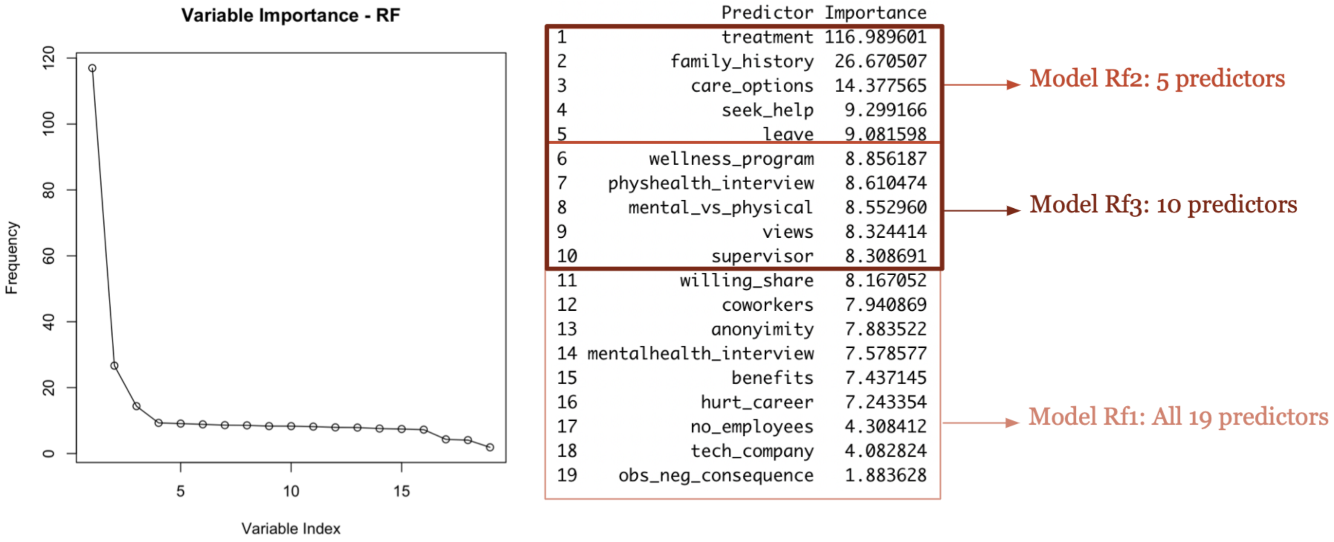
b. Tune parameter **mtry**

mtry = 5
→ OOB error = 12.7%
→ Accuracy = 87.31%

```
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
```

Graph 4: Variable Importance ranking from tuned Random Forest model



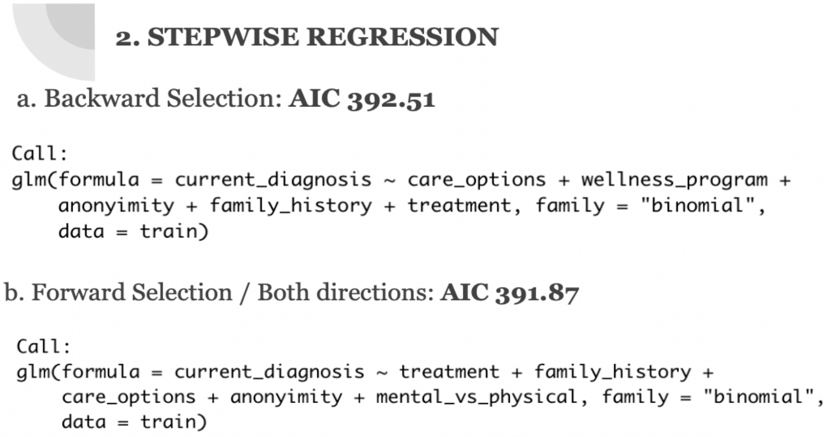
```
              Predictor  Importance
1             treatment 116.989601
2        family_history  26.670507
3          care_options  14.377565
4             seek_help   9.299166
5                 leave   9.081598
6      wellness_program   8.856187
7   physhealth_interview   8.610474
8     mental_vs_physical   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

Graph 5: 2 sets of predictors yielded from Stepwise Regression



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 **S1**

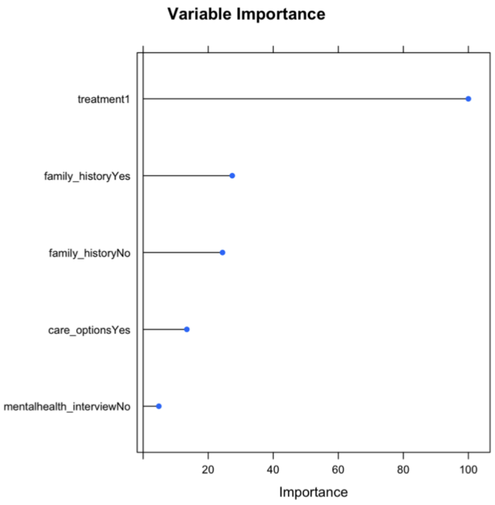
b. Forward Selection / Both directions: **AIC 391.87**

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

→ Model #5 **S2**

Graph 6: Recursive Partitioning

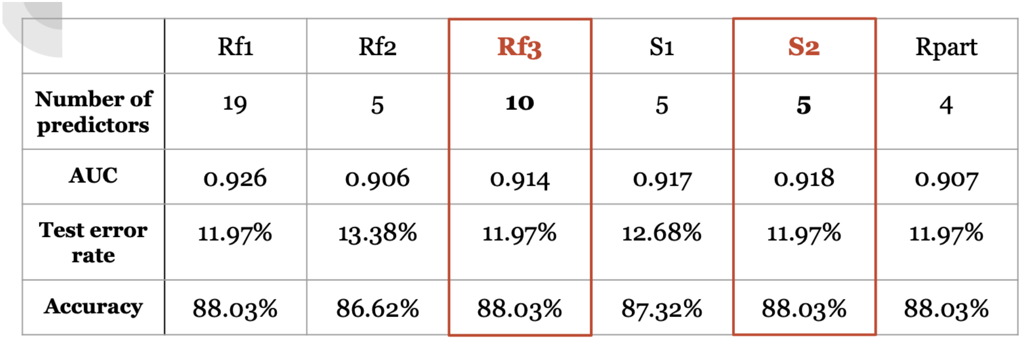
3. Recursive Partitioning



```
                            Overall
treatment1                  100.000
family_historyYes            27.390
family_historyNo             24.400
care_optionsYes              13.425
mentalhealth_interviewNo      4.819
```

Model#6 Rpart

Table 2: Logistic Regression Model Performance - Summary & Comparison



	Rf1	Rf2	Rf3	S1	S2	Rpart
Number of predictors	19	5	**10**	5	**5**	4
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%

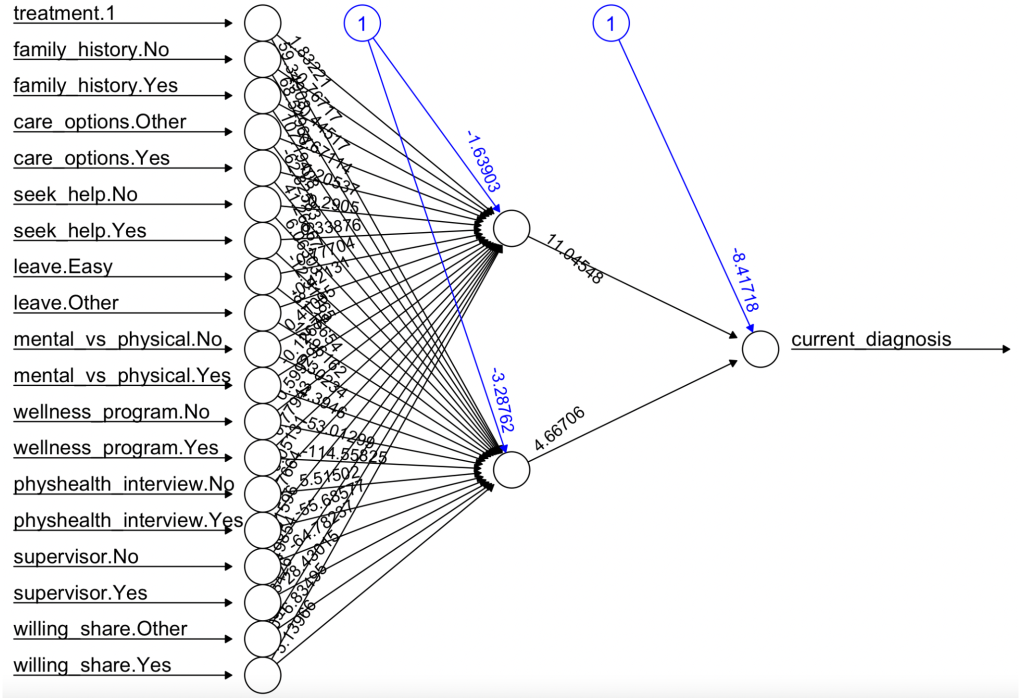
Finalists

to be used for Neural Net

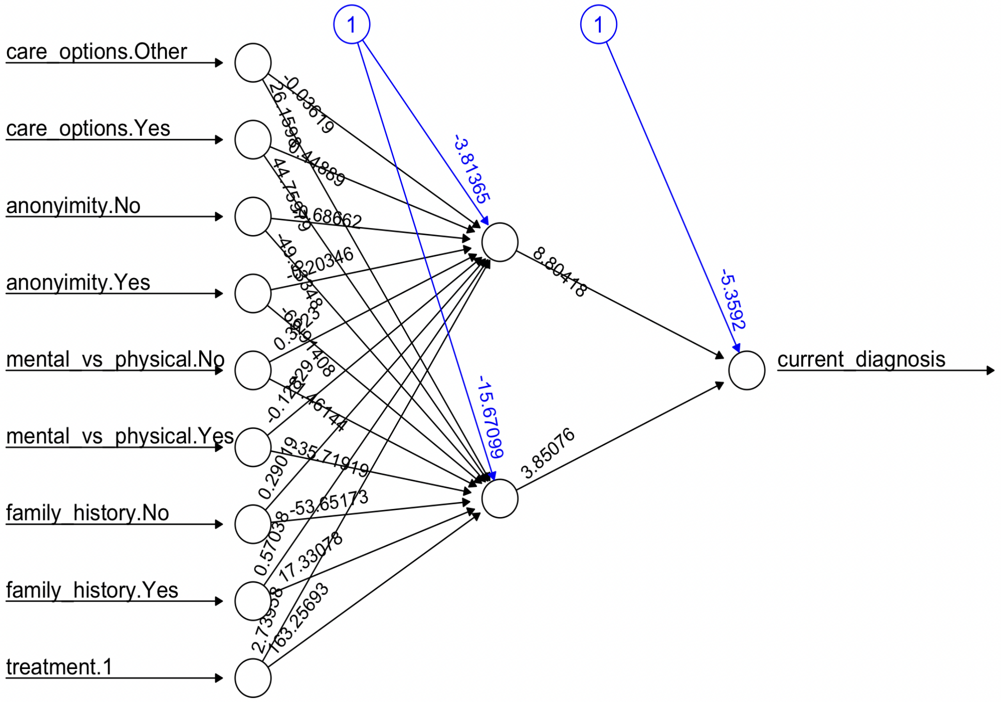
Table 3: Neural Network Performance - Summary & Comparison

	Rf3	**S2**
Test Error Rate	14.79%	12.68%
Accuracy	85.21%	87.32%

Graph 7: Neural Network using Rf3 (Random Forest) predictors



Graph 8: Neural Network using S2 (Forward Selection) predictors


care_options.Other
care_options.Yes
anonyimity.No
anonyimity.Yes
mental_vs_physical.No
mental_vs_physical.Yes
family_history.No
family_history.Yes
treatment.1
1
1
-0.03619
26.1598
0.44889
44.75949
0.68662
-49.95848
0.20346
-6.23
0.391408
-0.12829
1.46144
0.29019
35.71919
0.57038
-53.65173
17.33078
2.73958
163.25693
-3.81365
-15.67099
8.80418
3.85076
-5.3592
current_diagnosis