Is Work Making You Mentally Ill?

Identifying predictors of mental health issues in the workplace

Maximilian Rodrigues, Harsh Hareshkumar Shukla, Vivek Umeshkumar Bhavshar, Long Bao Nguyen, Ana Morrissey, Yash Pandya

Table of Contents

Abstract

In this analysis, we developed logistic regression models capable of predicting whether a person has sought treatment for a mental health condition among working professionals. Using a dataset created by Open Sourcing Mental Illness, LTD, each group member developed distinct models to derive predictive variables that included demographics, attitudes toward mental illness, and information about the individual's workplace. Models were developed on a random subset of the full dataset and tested on the complementary random subset. Probability cutoffs to determine who did or did not seek treatment were calculated and classification metrics such as specificity, sensitivity and accuracy were computed. The results show that the models can be used to predict whether a person has sought treatment for a mental health condition. These models can be of use to companies to help inform themselves if certain employees ought to seek treatment for a mental health condition.

Introduction

Mental health is a multi-faceted and complex subject that has an extensive literature in regards to causal factors of mental health conditions. The purpose of this project was to build an optimal predictive classification model to predict what factors may lead someone to seek treatment for a mental health condition.

The dataset includes demographics such as Gender, Location, Family History of Mental Illness as well as participant attitudes towards mental health and information related to the size and culture of their workplace. Research shows that genetic as well as environmental factors contribute to mental health illnesses Uher[1]. Past research has also shown that the age of onset of mental health conditions varies depending on the mental health disorder. Phobias and separation anxiety tend to have an early onset while generalized anxiety disorders, panic disorders and PTSD have a much later age of onset distribution (Kessler et al.[2]). Prevalence rates not only vary by age, but also by gender according to Eaton et al.[3] For example, women have higher rates of mood and anxiety disorders while men show higher rates of antisocial and substance abuse disorders. However, according to the World Health Organization[4], the overall prevalence of mental disorders does not appear to be different between the two sexes. However, according to Vessey & Howard[5], men are less likely to seek treatment for a mental health condition than women meaning that although overall prevalence rates are about equal between genders, there are differences in who seeks help and who doesn't. According to Corrigan, et al.[6], a factor that impedes care-seeking is the stigma associated with mental illness. In the same World Health Organization article that was previously mentioned, it is stated that rates of depression vary significantly between countries. This indicates that geographic factors may have an influence on mental health conditions. Research has also found that larger companies have higher rates of prevalence of psychological disorders than smaller companies. (Inoue et al.[7]).

We expect there to be biological and environmental factors that can be used to predict whether someone has sought treatment for a mental health illness. The dataset contains 1,239 observations and 25 meaningful variables, all of which are categorical except for Age.

The performance metrics for all of the models created were compared and the best model was selected. The following Methodology, Analysis, Results & Findings sections are from the selected Final Model. The team's individual reports can be found in Appendix – Individual Reports. The Appendix for the Final Model can found under Vivek Umeshkumar Bhavshar Individual report at the start of page 96.

Methodology

Data Preparation
we obtained the data from https://www.kaggle.com/osmi/mental-health-in-tech-survey.

The dataset was examined for missing values. The dataset contained 98% categorical variables and the remaining 2% were scalable or numerical data. Furthermore, all the possible dummy variables were created to perform logistic regression and to move one step further in the process of getting knowledge or common goal. However, the outliers were also removed to get the more accuracy on the prediction model and the interaction variables were also created.

Model Approach
In the same direction to move further, the different models were applied on the dataset in which first the data were partitioned in to training and testing sets with 70% in training and 30% in testing. Moreover, the forward selection, stepwise and backward elimination were performed to get the significant attribute information, in which we looked for predicted probability and used certain threshold. Furthermore, the decision tree was also applied on the same data with the hold out partitioning technique with 70% in training and 30% on testing to come up with the best final model.

Validation
The classification matrix or confusion matrix were used to check the multicollinearity also used some performance measures such as specificity, sensitivity and accuracy.

Analysis, Results and Findings

Analysis
This data set is about the mental health for humans at work environment. Moreover, to analyze the number or statistic that how many of them actual needed treatment, the frequencies were checked and observed that about 50% of them were needed the treatment or their mental health were not good. Furthermore, many other attributes were analyzed, and the dummy variables were created to preprocess the dataset. The treatment attribute was ***selected*** as a dependent attribute and the rest were selected as an independent attribute. The given dataset has two binary variables which are Age and Timestamp out of which we kept Age and formed a Boxplot with the dependent variable treatment. Figure 7 in appendix depicts the same.

Interaction variables
The interaction variables were formed by combining different important attributes such as Female, Work Interfere, Family History, Care Option and Benefit for the given data set and dummy variables were also created.

Collinearity
How collinear the variables are with each other were checked with the correlation matrix in the given data set and observed the variables collinear with our dependent variable Treatment. However, the maximum collinearity measure observed was 0.4.

Influential Points and Outlier
Performed action to find outlier and influential points in the dataset. Moreover, to improve the model accuracy the outliers and influential points were removed from the dataset.

MODEL 1

Before we ran our any of the applied model we divided the dataset into two parts Training and Testing with 70% in Training and 30% in Testing.

For variable selection we ran diverse selection technique, for model 1 Forward selection and Stepwise were used and the results for both the techniques were same as both the techniques had gave the equal number and same attributes in the table called Analysis of maximum likelihood estimates. Where the likelihood ratio was 566.89 and the P-values was observed <0.0001. Moreover, the classification table and the cut off or threshold (0.4) values with help of specificity and sensitivity were calculated for training set, we also computed predicted probability for testing set. In addition to that, confusion matrix for model was used to get the performance measures like accuracy, specificity, sensitivity, precision and recall. Below are the calculated statistics:

Accuracy: 85.94%, Sensitivity: 93.8%, Specificity: 76%, Precision: 83.12%

Predictor in Model 1

Model 1 has given total 10 significant variables out of all which are as below.

`Male Middle_Atlantic_USA Has_family_history work_ifr_Never work_ifr_Often work_ifr_Rarely work_ifr_Sometimes has_benefits know_care_options has_coworkers`

Figure 1 in appendix shows maximum likelihood estimates for Forward selection also Figure 8 shows the confusion matrix for model 1.

MODEL 2

In model 2 the backward elimination selection method was used in which the likelihood ration was 561.94 and the P-values was observed <0.0001. Even in model 2 we calculated the classification table. However, the cut off value calculated from specificity and sensitivity for this model was 0.45. In addition to that, confusion matrix for model was used to get the performance measures like accuracy, specificity, sensitivity, precision and recall. Below are the calculated statistics:

Accuracy: 86.73%, Sensitivity: 93.8%, Specificity: 74.84%, Precision: 85.65%

Predictor in Model 2

Model 2 has given total 9 significant variables out of all which are as below.

Female Has_family_history work_ifr_Never work_ifr_Often work_ifr_Rarely work_ifr_Sometimes has_benefits know_care_options has_coworkers

Figure 2 in appendix shows maximum likelihood estimates for Backward Elimination Also Figure 9 shows the confusion matrix for model 1.

Final Model Selection

As projected above, we ran two diverse selection techniques which gave different accuracy and different number of total predictor. Moreover, between two model the second model has been selected as such it gives higher accuracy and less number of predictors the likelihood ratio was 561.94 and the P-values was observed <0.0001. Moreover, in the final model we found influential points and outliers which were removed from the data set to further improve the accuracy and performance of the model. Furthermore, the possible outlier and influential points removed from the dataset are as below.

1257, 828, 651, 603.

Predictor in Final Model

Final Model has given total 9 significant variables out of all which are as below.

`Female Has_family_history work_ifr_Never work_ifr_Often work_ifr_Rarely work_ifr_Sometimes has_benefits know_care_options has_coworkers`

Figure 3 in appendix shows maximum likelihood estimates for Final Model. Also, the final regression model is as below.

Log((Need-treatment-=1)/(Need-treatment=0))= - 5.2866 + 0.6475* Female + 1.0760*(Has_Family_history)+2.4317*(work_ifr_Never)+5.8586*(work_ifr_Often)+4.6904*(work_ifr_Rarely)+5.1115*(work_ifr_Sometimes)+0.7596*(has_benefits)+0.8074*(know_care_options)+0.6837 *(has coworkers)

In addition to that the residual plots were also created to check the principle violations in the logistic regression. Figures 4, 5 & 6 in appendix shows I-plots for the same.

Different Approach of Binary Tree (not covered in class)

For the same dataset I applied decision tree with hold out partitioning technique with 70% in training and 30% in testing. However, I had applied different cases for parent and child to check for the different outcomes and accuracy. The optimal tree is marked in yellow line.

Index	Cases for N_P, Nc	Training Accuracy (70%)	Testing Accuracy (30%)	Complexity
1	N_P=100 N_c=50	74.9%	75%	3
2	N_P=35 N_c=17	84.1%	82.6%	15
3	N_P=28 N_c=14	84.7%	82.3%	16

The table above depict information on number of parent (NP) and child (Nc), training and testing percentage also complexity by the number of terminal nodes in each tree. Moreover, among all Models of regression analysis and decision tree for this dataset the regression analysis model performs overall good so it is preferred to go with model 2 when compared other applied model. Figure 10, 11 & 12 shows Final Tree, Classification Table & Important attributes Respectively.

Future Work

First of all, there are limitations on the dataset. It was collected from an open survey on the internet that anyone could access and take, hence the data gathered might not represent the population in the industry well. The bias can be shown by the distribution of the Treatment variable (49.4% answer No and 50.6% answer Yes), which indicates the rate of people having any mental illness in tech workplace is 50.6%, much higher than the 18.3% rate of US adults having any mental illness from the U.S. National Institute of Mental Health data.

The difference suggests that either the rate of mental illness in the tech industry are much higher than outside the tech industry or that the rate of mental illness in the tech industry are similar to outside the industry and the sample was incredibly biased in this respect. If the latter is true, we need a more representative sample before making any predictions about people outside our sample, more data would need to be collected in a more proper way.

Although the data might be biased, we still gathered some meaningful insights such as:

- There are certainly mental health resources already being provided by employer, this is indicated by 37.9% of responses confirming that employers offer mental health benefits or resources and 35.3% of the responses knowing the options for mental health care being provided by employer.
- However, communication about the provided resources could be better as 32.4% of the respondents indicated that they did not know what benefits or resources were available to them. Furthermore, 81.8% of respondents are not sure about or confirm that their employer has not discussed mental health as a part of an employee wellness program.
- Workers are open about mental illness, indicated by 79.3% of employees are willing to talk to their coworkers about their mental health conditions. This number is 68.8% when it comes to talking with their supervisors.

Secondly, we have gone through several variations in preprocessing data, creating different interaction variables, splitting training/testing data with different ratio, choosing varying cut-off values and experienced a wide range of model accuracies. Yet the variations have not run out and we could still explore even more variations to improve our final model accuracy.

Finally, for the prediction model, due to restriction of time, we only applied one classifier on the dataset to build the prediction model. That is Logistic Regression which learns to separate the classes based on a linear decision surface, the predicted log-odds is a linear function of the independent variables.

There is the chance that our classes are not very well linearly separable. We could try applying other classifiers including Decision tree, Random Forest or K-Nearest neighbor to compare the results and choose the best classifying method.

References

1. Uher, Rudolf. 2014. Gene-environment interactions in severe mental illness. *Frontiers in Psychiatry*. https://www.frontiersin.org/articles/10.3389/fpsyt.2014.00048/full Accessed 5 March 2018.

2. Kessler, R., Andermeyer, M., Anthony, J., De Graff, R., Demyttenaere, K., Gasquet, I., De Girolamo, G., Gluzman, S., Gureje, O., Haro, J.M., Kawakami, N., Karam, A., Levinson, D., Media Mora, M.E., Browne, M.A.O., Posada-Villa, J., Stein, D.J., Tsang, C., Aguilar-Gaxiola, S., Alonso, J., Lee, S., Heeringa, S., Pennell, B., Berglund, P., Gruber, M.J., Petukhova, M., Chatterji, S., Üstun, T.B. 2007. Lifetime prevalence and age-of-onset distributions of mental disorders in the World Health Organization's World Mental Health Survey Initiative. *World Psychiatry* 6(3): 168-176.

3. Eaton, NR., Keyes, KM., Krueger, RF., Balsis, S., Skodol, AE., Markon, KE., Grant, BF., Hasin, DS. 2012. An invariant dimensional liability model of gender differences in mental disorder prevalence: evidence from a national sample. *Journal of Abnormal Psychology.* 121(1): 282-288.

4. World Health Organization. (2003) Organization of Services for Mental Health. *Geneva: World Health Organization.* http://www.who.int/en/ Accessed 2 March 2018.

5. Vessey, J. T., & Howard, K. I. 1993. Who seeks psychotherapy? *Psychotherapy: Theory, Research, Practice, Training* 30(4), 546-553.

6. Corrigan PW., Druss, PG., Perlick, DA. 2014. The impact of mental illness stigma on seeking and participating in mental health care. *Psychological Science in the Public Interest.* 15(2): 37-70.

7. Inoue, A., Kawakami, N., Tsuchiya, M., Sakurai, K., Hashimoto, H. 2010. Association of Occupation, Employment Contract, and Company Size with Mental Health in a Representative Sample of Employees in Japan. *Journal of Occupational Health.* 52(4): 227-240.

Appendix – Individual Reports

Maximillian Rodrigues

Introduction

The goal of this analysis is the same as the group, refer to the main report for details.

Methodology

Obtaining Data

The data were obtained from www.kaggle.com[1].

Data Preparation

After downloading and importing the data into SAS, each variable was explored iteratively beginning with the dependent variable treatment to identify any patterns or data entry errors. Categorical variables were explored by ensuring the levels were all valid, and then the categorical variables were explored against the dependent variable treatment to identify any potential relationships. The one numeric variable (Age) was explored by examining its distribution and relation to treatment.

Variable Creation

Once all of the categorical variables were recoded as dummy variables and interaction variables were established, an initial regression model was created utilizing all of the predictor variables with the VIF option selected to see collinearity between numeric variables (Age and the interaction variables that included Age). A correlation matrix was also created for Age and interaction variables that included Age. No variable transformations were deemed necessary.

Model Approach

Two models were created with the first using a total of 59 predictors while the second used 51. For the two models, data was split randomly into training and testing sets (80% in the training set and 20% in the testing set). Three selection methods were used on each model (backward, forward, and stepwise) and the model with the fewest predictors as well as lowest AIC and SC values was used as a final model. A probability threshold for prediction was determined using an ROC curve where Youden's J statistic was maximized. A classification threshold table was also created to compare the maximized Youden's J statistic with the maximized sum of specificity and sensitivity. Using the determined threshold, the predicted values in the test set were compared with the actual values, and two observations were also predicted using made up data. The predicted probabilities and predictive intervals were then analyzed.

Validation Method

For the two models created, confusion matrices were created. Overall accuracy, sensitivity, specificity, positive predictive value and negative predictive value were calculated for each model.

Analysis, Results and Findings

Exploratory Analysis

For the Treatment variable, the frequency of Yes and No responses as to whether or not the individuals has sought treatment for a mental health condition was approximately 50-50 (see figure 1 in the appendix). In this sample the odds are about equal as to whether or not a person sought treatment for a mental health condition. It is important to note that not everyone in this sample actually has a mental health condition.

In order to address invalid entries in the Age variable, only observations within the age range of 18-90 were used. Once this range was selected, the histogram for Age (Appendix Figure 2) shows some minor positive

skew, indicating the population of the survey is rather young. This skew did not need to be addressed with a transformation. According to the Bureau of Labor Statistics[2], the median age of all workers in 2017 was 42.2 and the median age of all tech workers was 40.7 while this survey had a median age of 31, indicating there was possible selection bias when conducting this survey. Age was used as a numeric variable in the model but was also grouped into cohorts for exploratory purposes, since studies typically look at age differences in mental health by segmenting ages (Gatz & Hurwicz[3]). The graph in the appendix shows Age would likely not be a strong predictor if segmented into cohorts analogously to other studies since approximately 85% of participants were under 40 (figure 3 in the appendix).

Gender was a free response question in the survey which meant a high degree of data inconsistency. The Gender Literature is extensive, with researchers like Monro[4] suggesting Gender is more a spectrum than it is binary. For this reason, individuals were recoded as being either Male, Female, Non-Binary (e.g. androgynous, transgender, etc.), or "Other" for individuals who did not provide a valid gender. However, only three individuals declined to provide their gender so individuals who did not indicate gender were removed from further analysis. While the Non-Binary variable only has 11 observations, 10 of the 11 have sought treatment for a mental health condition. Figure 4 in the appendix also shows that more than twice as many females sought treatment than didn't seek treatment.

Since this was an international survey, there is reason to believe that social or other geographically related factors may have influence on the prevalence of mental health conditions. While ideally it would be best to use each country as a level in the variable "Country", the low rates of responses from certain countries makes it more sensible to join countries and states into regions. States were divided into regions based on regions determined by the U.S Census Bureau5 and most countries were placed into regions as determined by the United Nation's geoschemes[6]. Regions with fewer than 10 individuals were removed from analysis. The variations in frequencies of those who did or not seek treatment mostly looked like random variation, indicating this is likely a weak predictor.

All of the remaining variables (which were all categorical) were split by the "Treatment" variable using either bar charts or tables. Several variables showed clear relationships such as the more a person indicates mental health interferes with their work, the more likely the individual has sought treatment (figure 5 table in the appendix). Most of the remaining variables showed weak relationships (see SAS code to recreate tables and graphs).

Recoded and Interaction Variables

Prior to recoding the categorical variables as dummy variables, there were 23 predictors and after removing certain observations, the dataset had 1,212 observations. Given how much the literature stresses biological differences as predictors of mental health conditions, interaction variables were created using combinations of Age, Family History, and Gender. Dummy variables were created for all of the categorical variables and all variables were used in an initial model.

Collinearity

Since nearly all of the variables were dummy variables, the only variables necessary to check for multicollinearity were Age and the interaction variables comprised of Age. Correlations between each of these variables was below .5, so they were deemed suitable for model use.

Influential Points

Influence diagnostics for the full model showed that 1 observation (observation 1208) had a large amount of influence on the model. The Chi-Square Deletion Difference (that is, the change in the Chi-Square statistic

for deleting this observation) was 253.6 while the next highest value was under 50 (see figure 6 in the appendix). The observation also has a Pearson residual over 15 (measuring the individual contribution to the Pearson statistic) and a deviance deletion difference of about 12. This observation was deleted since the value is ill-fitted. After removing this observation and rerunning the full model, there were no further observations with extremely high diagnostic values.

Normal Probability and Residual Plots

Since the response distribution is not normal, the residuals for the model are not expected to be normal so there is no need to test for normality. Also, unlike the least squares estimation of normal-response models, the assumption of homoscedasticity does not apply to the maximum likelihood estimation of a logistic regression model.

Model 1

Overall Goodness of Fit for the Full Model 1

The Likelihood Ratio obtained from the full model was 834.03 with 60 degrees of freedom and a p-value <.0001 (see figure 7 in the appendix). This indicates that we reject the null hypothesis that all parameters are equal to 0.

Model 1 Selection

The testing and training sets were created by splitting the data randomly using 20% in the test set and 80% in the training set. For model selection, the three model selection methods of Forward, Backward and Stepwise were used on the training set and the model with the fewest significant predictors as well as lowest AIC and SC values were selected as the final model. The Stepwise and Forward methods provided identical outputs, and both methods had fewer predictors than the Backward method while having an equal AIC and SC value. The Forward/Stepwise selection method outputs were used to build the final model (see figure 8 for the model output). 10 Predictors reached significance, none of which were numeric or interaction variables.

Strongest Predictors for Model 1

The dummy variable "WorkIntNum3" (which corresponds to whether or not the participant "sometimes" feels a mental condition interferes with their work) is the strongest predictor since it had the highest standardized parameter estimate in figure 8a in the appendix. The work Interference dummy variables were among the strongest predictors, followed by the family history of mental health condition variable. Figure 9 in the appendix shows the odds ratios for each of the parameters used. The odds ratios can be interpreted as for every one unit change in the predictor, the odds ratio for the participant to have sought treatment for a mental health condition is expected to change by the value of the point estimate in the table, holding all other variables constant. Figure 10 is a table summarizing each of the converted odds ratios (since all of the variables used in model 1 are dummy variables, these values can be interpreted as the percentage increase in probability that the individual sought treatment when the dummy variable is equal to 1).

Model 1 Equation

Log(treatment=1/treatment=0) = -5.4086 – 0.6343*GenderNumM + .9929*fam_histNum + 2.9058*work_intNum + 6.1963*work_intNum1 + 5.2538*work_intNum2 + 5.8118*work_intNum3 + 0.7622*benefitsNum1 + 0.7217*careNum1 +.4702*coworkNum1 + 0.4489*phcNum

Determining a Probability Threshold for Model 1

According to Habibzadeh et al.[7], an appropriate way to select a cut-off probability for a logistic regression model is where Youden's J statistic is maximized. This is the point on the ROC curve with the highest vertical distance from the 45 degree diagonal line (Figure 11 shows the relevant ROC curve). Any observation with a probability above this threshold would be predicted as "sought treatment" and any observation below would be predicted as "did not seek treatment". The probability threshold obtained was .53245. To supplement this, a threshold classification table was also created, testing the classification metrics of the training set model at different probability thresholds (see figure 12). A probability level of .50 maximized Specificity + Sensitivity. The maximized Youden's J statistic was used as a threshold for the predictive model.

Classification Metrics Measuring Predictive Power for Model 1

Predicted values were compared to actual values in the test set (see figure 13 in the appendix for the confusion matrix). Below are the calculated metrics:

Accuracy: 82.64%
Sensitivity: 116/(116+13) = 89.92%
Specificity: 84/(84+29) = 74.33%
Positive Predictive Value: 116/(116+29) = 80.00%
Negative Predictive Value: 84/(84+13) = 86.60%

The model has a very high sensitivity rate, meaning it is very good at identifying participants who did seek treatment. Specificity is also high, but much lower than sensitivity.

Predicted Observations for Model 1

The ten variables were included in the final model which included dummy variables for gender, family history of mental illness, degree of work interference mental illness causes, whether the company offers mental health benefits, does the person have an understanding of the care options the company provides, whether the individual feels comfortable talking about mental illness with coworkers, and whether there would be negative consequences discussing physical health issues with an employer. Using these variables, two new observations were created using made up data. Figure 14 in the appendix shows what values were selected for each variable. The created values were then inputted to the final model to predict whether these made-up date would have sought treatment or not. Both observations were predicted to have sought treatment. The predicted probabilities for the created data were .70205 and .99970 with prediction intervals (.53613, .82769) and (.99854, .99994), respectively (see figure 15 in the appendix). With 95% certainty, we can say the true probabilities lie within the prediction intervals. Neither prediction interval contains the threshold value of .53245.

Model Improvement

During exploratory analysis, it was determined that participants who did not provide a response to the work interference variable do not actually have a mental health condition. Since the analysis is concerned with what information can be used to predict if a person sought treatment for a mental health condition, it makes sense to do an analysis exclusively among people who claim to have a mental health condition. Therefore, all observations that did not provide a response for the work interference variable were deleted so a model could be built for participants who do claim to have a mental health condition. The deletion of variables brought the dataset size down to 934 observations. Model 2 followed the same procedures as model 1 but on the reduced data set.

Model 2

Overall Goodness of Fit of Full Model 2

The Likelihood Ratio obtained from the full model was 421.99 with 51 degrees of freedom and a p-value <.0001. (see figure 16 in the appendix). This indicates that we reject the null hypothesis that all parameters are equal to 0.

Model 2 Selection

The same model selection process used for model 1 was used for model 2. Once again, the Stepwise and Forward selection methods provided identical outputs (both with fewer predictors than the Backward method with identical AIC and SC values). Eight variables reached significance which will all be used in the final model, which were once again all dummy variables (see figure 17 in the appendix).

Model 2 Equation

Log(treatment=1/treatment=0) = -1.6311 -0.856*GenderNumM + 1.0743*fam_histNum + 3.5958*work_intNEW + 2.5349*work_intNEW2 + 2.7031*work_intNEW3 + 0.8450*careNum1+ 0.6152*anonNum1 + -0.4906*coworkNum

Strongest Predictors for Model 2

The dummy variable "WorkIntNEW2" (which corresponds to whether or not the participant "sometimes" feels a mental condition interferes with their work) is the strongest predictor since it had the highest standardized parameter estimate in figure 17a in the appendix. Once again, the work Interference dummy variables were among the strongest predictors, followed by the family history of mental health condition variable. Figure 18 in the appendix shows the odds ratios for each of the parameters used. Figure 19 is a table summarizing each of the converted odds ratios.

Determining Probability Threshold for Model 2

The same method for determining a probability threshold in model 1 was used for model 2. The probability threshold obtained for Model 2 was .60178 (maximizing Youden's J).

Classification Metrics Measuring Predictive Power for Model 2

Predicted values were compared to actual values in the test set (see figure 20 in appendix). Below are the calculated metrics:

Accuracy: 74.19%
Sensitivity: 85/(85+24) = 77.98%
Specificity: 53/(53+24) = 68.83%
Positive Predictive Value: 85/(85+24) = 77.98%
Negative Predictive Value: 53/(53+24) = 68.83%

Compared to model 1, model 2 has much less predictive value. Both sensitivity and specificity are worse than model 1. This was expected, since one of the predictors in model one essentially indicated which participants did or did not have a mental health condition. So while predictive power is weaker in model 2, it is still a useful model when applied exclusively to individuals who have a mental health condition.

Predicted Observations for Model 2

Using the 8 variables from the final model, two new observations were created using made up data. The model created was then used to predict whether these made up observations would have sought treatment. Figure 21 in the appendix show what values were selected for each variable. The created values

were inputted to the final model to predict whether these made-up data observations would have sought treatment for a mental health condition. The predicted probabilities for the created data were .48203 and .95294 with 95% prediction intervals (.37232, .59351) and (.91382, .97479). Neither of these prediction intervals contained the threshold value of .60178 (see figure 22 in the appendix).

Findings

While the first model was much more accurate in prediction (higher sensitivity and specificity), the second model is more informative since it only applies to individuals who actually have a mental health condition. When attempting to predict who sought treatment regardless of whether they actually have a mental health condition, model 1 should be prescribed. For a sample that is comprised only of individuals who actually have a mental health condition, model 2 should be utilized.

References

1. Kaggle.com. 2018. Survey on Mental Health in the Tech Workplace in 2014. https://www.kaggle.com/osmi/mental-health-in-tech-survey. Accessed 3 March 2018

2. U.S. Department of Labor, Bureau of Labor Statistics. "Employed Persons by Detailed Occupation and Age" *Labor Force Statistics from the Current Population Survey,* U.S. Dept. of Labor, 19, Jan. 2018, https://www.bls.gov/cps/cpsaat11b.htm. Accessed 2 March 2018.

3. Gatz, M. and Hurwicz, M.-L. 1990. Are old people more depressed? Cross-sectional data on Center for Epidemiological Studies Depression Scale Factors. *Psychology and Aging* 5(2): 284-290.

4. Monro, S. 2005 Beyond Male and Female: Poststructuralism and the spectrum of gender. *International Journal of Transgenderism* 8(1), 3-22

5. United States Census Bureau Regions retrieved from https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf Accessed 2 March 2018

6. United Nations Geographic Regions retrieved from https://unstats.un.org/unsd/methodology/m49/ Accessed 3 March 2018

7. Habibzadeh, F., Habibzadeh, P., Mahboobeh, Y. 2016 On determining the most appropriate test cut-off value: the case of tests with continuous results. *Biochemia Medica* 26(3): 297-307

Appendix

Figure 1

Treatment		
treatment	Frequency	Percent
No	622	49.40
Yes	637	50.60

Figure 2

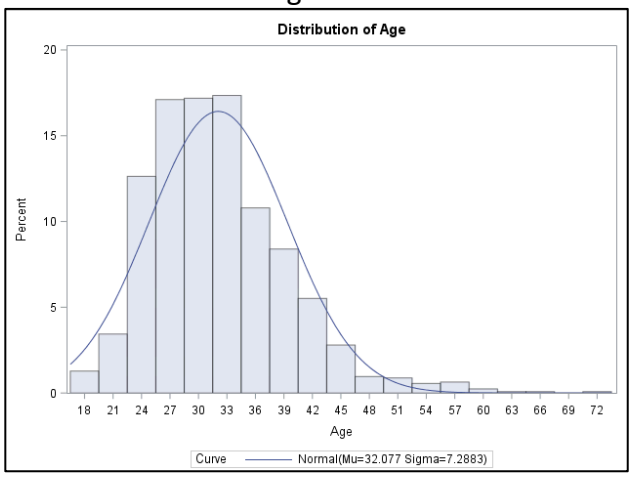


Figure 3

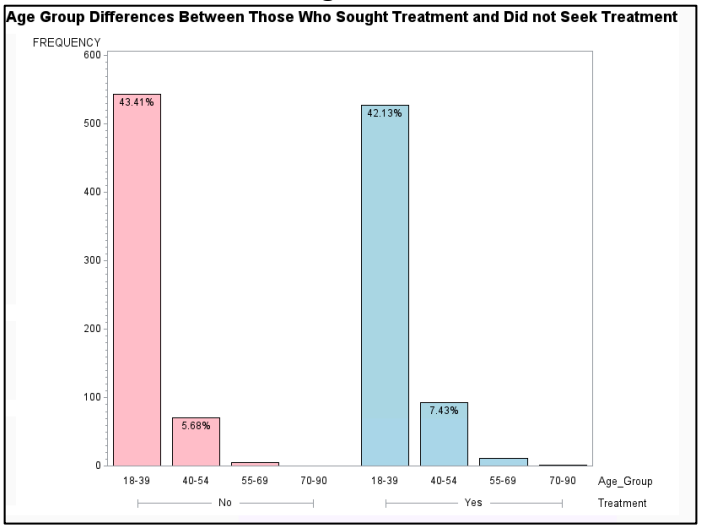


Figure 4

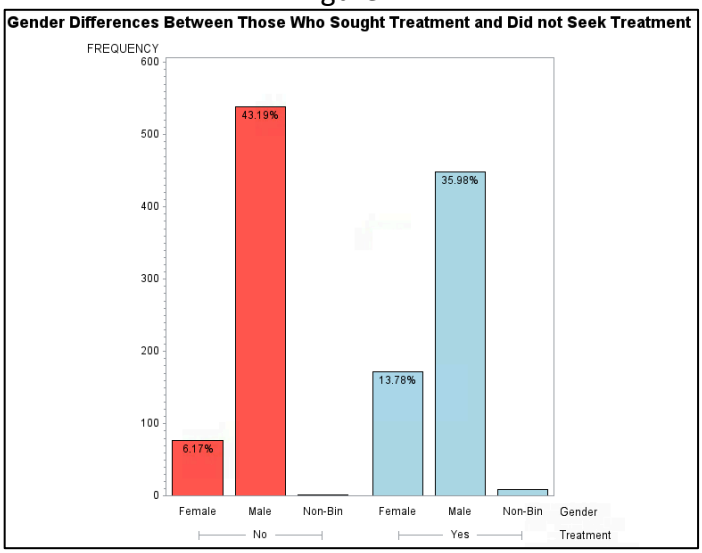


Figure 5

Table of treatment by work_interfere						
	work_interfere(Does Mental Health Condition Interfere With Work)					
treatment(Treatment)	Blank	Never	Often	Rarely	Sometimes	Total
No	260	183	21	51	107	622
Yes	4	30	123	122	358	637
Total	264	213	144	173	465	1259

Figure 6 a.

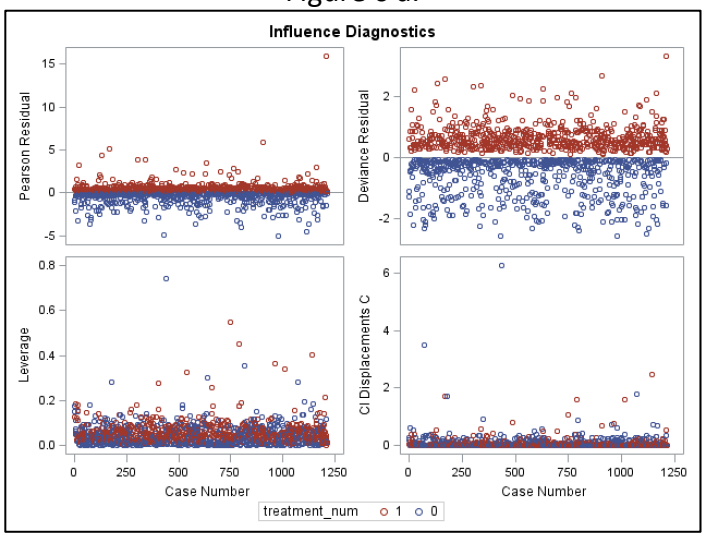


Figure 6 b.

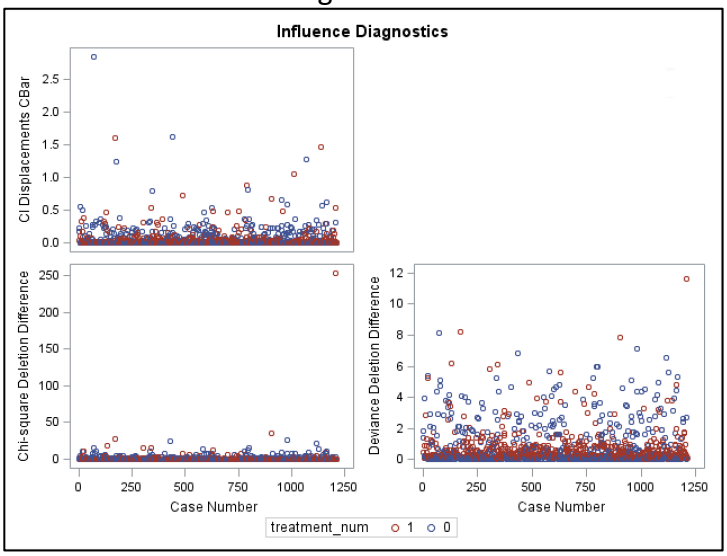


Figure 7

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	843.0240	60	<.0001
Score	662.1919	60	<.0001
Wald	280.3782	60	<.0001

Figure 8 a.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.4087	0.7769	48.4737	<.0001
GenderNumM	1	-0.6343	0.2473	6.5784	0.0103
fam_histNum	1	0.9929	0.2004	24.5558	<.0001
work_intNum	1	2.9058	0.7542	14.8440	0.0001
work_intNum1	1	6.1963	0.7676	65.1550	<.0001
work_intNum2	1	5.2538	0.7438	49.8861	<.0001
work_intNum3	1	5.8118	0.7310	63.2190	<.0001
benefitsNum1	1	0.7622	0.2237	11.6100	0.0007
careNum1	1	0.7217	0.2218	10.5889	0.0011
coworkNum1	1	0.4702	0.2671	3.1004	0.0783
phcNum	1	0.4489	0.2151	4.3561	0.0369

Figure 8 b

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1345.293	716.165
SC	1350.170	769.804
-2 Log L	1343.293	694.165

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	649.1279	10	<.0001
Score	518.5302	10	<.0001
Wald	229.8073	10	<.0001

Figure 9

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
GenderNumM	0.530	0.327	0.861
fam_histNum	2.699	1.823	3.997
work_intNum	18.279	4.169	80.157
work_intNum1	490.905	109.040	>999.999
work_intNum2	191.296	44.518	821.998
work_intNum3	334.227	79.773	>999.999
benefitsNum1	2.143	1.382	3.322
careNum1	2.058	1.332	3.178
coworkNum1	1.600	0.948	2.701
phcNum	1.567	1.028	2.388

Figure 10

GenderNumM	= (.530-1)*100 = -47.0%
Fam_histNum	= (2.699-1)*100 = 169.9%
work_intNum	= (18.279-1)*100 = 1,727.9%
work_intNum1	= (490.905-1)*100 = 48990.5%
work_intNum2	= (191.296-1)*100 = 19029.6%
work_intNum3	= (334.227-1)*100 = 33322.7%
benefitsNum1	= (2.143-1)*100 = 114.3%
careNum1	= (2.058 – 1) *100 = 105.8%
coworkNum1	= (1.600 -1) * 100 = 60%
phcNum	= (1.567 -1) * 100 = 56.7%

Figure 11

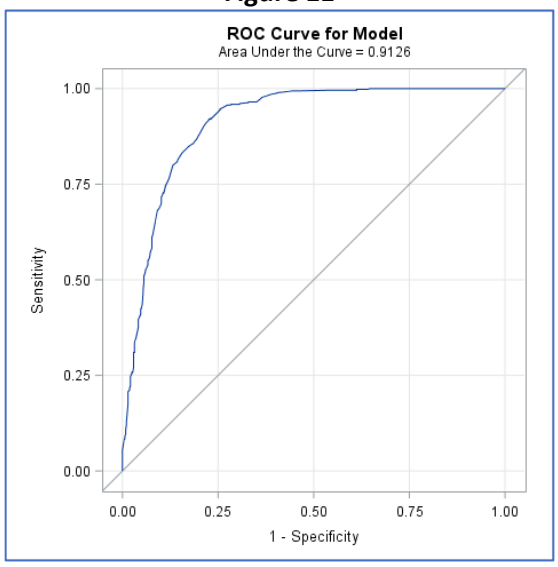


Figure 12

Classification Table

Prob Level	Correct		Incorrect		Percentages				
	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.400	466	350	132	21	84.2	95.7	72.6	22.1	5.7
0.450	448	370	112	39	84.4	92.0	76.8	20.0	9.5
0.500	448	372	110	39	84.6	92.0	77.2	19.7	9.5
0.550	422	374	108	65	82.1	86.7	77.6	20.4	14.8
0.600	412	390	92	75	82.8	84.6	80.9	18.3	16.1
0.650	397	407	75	90	83.0	81.5	84.4	15.9	18.1
0.700	363	422	60	124	81.0	74.5	87.6	14.2	22.7

Figure 13

treatment_num	pred_y		
	0	1	Total
0	84	29	113
1	13	116	129
Total	97	145	242

Figure 14

Male (1= Yes)	Family History of Mental Illness (1 = Yes)	Mental Illness Never Interferes with work (1 = Yes)	Mental Illness Often Interferes with work (1 = Yes)	Mental Illness Rarely Interferes with work (1 = Yes)	Mental Illness Sometimes Interferes with work (1 = Yes)	Company Offers Mental Health Benefits (1 = Yes)	Do You Know if Care Options are offered? (1 = Yes)	Would you discuss mental health with coworkers? (1 = Yes)	Are there negative consequences for talking about physical health at work? (1 = Yes)
1	0	0	0	1	0	0	1	1	0
1	1	0	1	0	1	1	0	0	1

Figure 15

phat	lcl	ucl	pred_y	threshold
0.70205	0.53613	0.82769	1	0.53245
0.99970	0.99854	0.99994	1	0.53245

Figure 16

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	421.9893	51	<.0001
Score	371.5902	51	<.0001
Wald	223.1833	51	<.0001

Figure 17 a

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-1.6311	0.3491	21.8244	<.0001	
GenderNumM	1	-0.8560	0.2628	10.6059	0.0011	-0.2023
fam_histNum	1	1.0743	0.2089	26.4446	<.0001	0.2958
work_intNEW	1	3.5958	0.4004	80.6559	<.0001	0.6946
work_intNEW1	1	2.5349	0.3257	60.5849	<.0001	0.5283
work_intNEW2	1	2.7031	0.2767	95.4076	<.0001	0.7447
careNum1	1	0.8450	0.2282	13.7073	0.0002	0.2287
anonNum1	1	0.6152	0.2442	6.3481	0.0118	0.1577
coworkNum	1	-0.4906	0.2392	4.2041	0.0403	-0.1064

Figure 17 b

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	959.228	654.638
SC	963.845	696.195
-2 Log L	957.228	636.638

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	320.5896	8	<.0001
Score	292.8906	8	<.0001
Wald	181.9674	8	<.0001

Figure 18

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
GenderNumM	0.425	0.254	0.711
fam_histNum	2.928	1.944	4.410
work_intNEW	36.445	16.628	79.882
work_intNEW1	12.615	6.663	23.883
work_intNEW2	14.925	8.677	25.673
careNum1	2.328	1.488	3.641
anonNum1	1.850	1.146	2.986
coworkNum	0.612	0.383	0.979

Figure 19

GenderNumM	= (.425-1)*100 = -57.5%
Fam_histNum	= (2.928-1)*100 = 192.8%
work_intNEW	= (36.445-1)*100 = 3544.5%
work_intNEW1	= (12.615-1)*100 = 1161.5%
work_intNEW2	= (14.925-1)*100 = 1392.5%
careNum1	= (2.328-1)*100 = 132.8%
anonNum1	= (1.850 – 1) *100 = 85.0%
coworkNum	= (0.612 -1) * 100 = -38.8%

Figure 20

	pred_y		
treatment_num	0	1	Total
0	53	24	77
1	24	85	109
Total	77	109	186

Figure 21

Male (1 = Yes)	Family History of Mental Health Conditions (1 = Yes)	Mental Health Often Interferes with Work (1 = Yes)	Mental Health Rarely Interferes with Work (1 = Yes)	Mental Health Sometimes Interferes with Work (1 = Yes)	Do You Know the Care Options Your Employer Provides? (1 = Yes)	Is Anonymity Protected if You Seek Help for a Mental Health Condition? (1 = Yes)	Are There Consequences for Discussing Physical Health with Your Employer? (1 = Yes)
1	0	0	0	1	0	0	1
0	1	0	1	0	1	1	0

Figure 22

phat	lcl	ucl	pred_y	threshold
0.48203	0.37232	0.59351	0	0.60178
0.95294	0.91382	0.97479	1	0.60178

Long Bao Nguyen

Introduction

The goal of this analysis is the same as the group, refer to the main report for details.

Methodology

Get data from Kaggle (https://www.kaggle.com/osmi/mental-health-in-tech-survey).

For data cleaning and exploratory analysis, the dependent variable Treatment and 24 independent variables will be explored iteratively beginning with Age, examining variables for data inconsistencies, data entry errors, and skewness that may be addresses with variable transformations.

Treatment is binary so we will apply Logistic regression model. The model uses log odds so we will not do y-variable transformation. We will not examine scatterplots of paired data or linearity of relationship because Y is binary, points are concentrated at 0 or 1.

Variables with complex levels will be recoded for simpler operation.

Interaction and dummy variables are created as necessary.

A full logistic regression model will be run using all of the explanatory variables.

The data will be split into a training and test set (66/34), I shall fit the model on training set, test the model performance on test set.

Before applying model selection, I check to make sure the model has little or no multicollinearity. I also run Pearson and Deviance residual diagnostics to detect and remove outliers close to or exceeding ±3 and Levee failure diagnostics to ***detect*** and remove influential points with |Dfbeta| > $2/\sqrt{n}$.

After testing the various model selection methods to build a model, I shall compare models based on selection criteria such as AIC, SC, the goodness of fit as Likelihood Ratio, significant values and standard errors of predictors to choose a final model for prediction on the test set.

Finally the model will be used to make predictions on made up data.

Data exploration

In this stage, I shall observe the distribution of data and detect, then remove any meaningless samples or outliers.

Treatment: this indicates whether the survey taker already sought treatment for their mental conditions, this will be my dependent variable. Figure 1 shows that there are 1259 observations in total, 49.4% of the observations answered No and 50.6% answered Yes, distribution is balanced.

Age: The data was collected to represent mental health in tech workplace so Age range should be in the working age, specifically 18 to 64.

According to above descriptive statistics, i see that there are 10 observations out of the range 18-64, among them, 5 are meaningless.

I could remove the 10 observations because it shows that the survey takers might be not serious, yet consider the fact that Age is a sensitive aspect for people to answer sincerely and this survey targeted people with possibly mental illness and those observations could affect the final result, I shall replace those 10 observations' values with 31, which is the median of the data population.

Figure 2 shows frequency of Age before and after cleaning.

Gender: The responses vary greatly with 44 different terms for gender. To simplify, I shall divide the observations into 3 groups based on definitions of gender and possible typos in the survey:

> ***Female:*** 'female', 'cis female', 'f', 'woman', 'femake', 'female ', 'cis-female/femme', 'female (cis)', 'femail'
>
> ***Male:*** 'm', 'male', 'male-ish', 'maile', 'cis male', 'mal', 'male (cis)', 'make', 'male ', 'man', 'msle', 'mail', 'malr', 'cis man'
>
> ***Others:*** 'queer/she/they', 'non-binary', 'nah', 'enby', 'fluid', 'genderqueer', 'androgyne', 'agender', 'guy (-ish) ^_^', 'male leaning androgynous', 'neuter', 'queer', 'ostensibly male, 'trans-female', 'trans woman', 'female (trans)', 'something kinda male?'

Figure 3 shows frequency of Gender before and after cleaning.

Country and State: the collected data was represented in 48 countries and 46 states in the US. I grouped the countries and US states based on United Nations Country Groupings and US regions and create the Region variable. Figure 4 shows the frequency of Country and State and the Region variables.

Figure 5 shows all the frequencies of the remaining variables which have 2 or 3 levels.

Self_employed: which asked the survey takers if they are self-employed. There are 18 observations with NA values. I shall put the NA values as Yes if the answer in No employees is 1-5, otherwise the NA values will be put as No.

Family history: which asked the survey takers if they have a family history of mental illness.

Work interfere: which asked the survey takers if they feel that their mental health conditions interfere with their works.
This is an important variable as it indicates whether the survey taker has a health condition or not. The observations with NA responses can be taken as that the person has no mental health condition. However, there are 4 NA responses from people who answered Yes in Treatment. I shall put the observations in 2 groups:

> No: NA which answers No in Treatment, Never
> Yes: NA which answers Yes in Treatment, Rarely, Sometimes, Often

No_employees: which asked for the number of employees in the company of the survey takers.
I shall group the Number of employees into 3 groups:

> Small: 1-5, 6-25, 26-100
> Medium: 100-500, 500-1000
> Large: More than 1000

Remote work: this asked the survey takers if they work remotely (outside of an office) at least 50% of the time.

Tech company: this asked the survey takers if their employer is primarily a tech company/organization.

Benefits: this asked the survey takers if their employer provides any mental health benefits.

Care options: this asked the survey takers if they know the options for mental health care their employer provides.

Wellness program: this asked the survey takers if their employer ever discussed mental health as part of an employee wellness program.

Seek help: this asked the survey takers if their employer provides resources to learn more about mental health issues and how to seek help.

Anonymity: this asked the survey takers if their anonymity is protected if they choose to take advantage of mental health or substance abuse treatment resources.

Leave: this asked the survey takers if it is easy for them to take medical leave for a mental health condition I shall put the observations into 3 groups:

Don't know: Don't know
Yes: Somewhat easy, Very easy
No: Somewhat difficult, Very difficult

Mental health consequence: this asked the survey takers if discussing a mental health issue with their employer would have negative consequences.

Physical health consequence: this asked the survey takers if discussing a physical health issue with their employer would have negative consequences.

Coworkers: this asked the survey takers if they are willing to discuss a mental health issue with their coworkers.

I shall group Some of them and Yes observations into same group: Yes.

Supervisor: this asked the survey takers if they are willing to discuss a mental health issue with their direct supervisor(s).

I shall group Some of them and Yes observations into same group: Yes.

Mental health interview: this asked the survey takers if they would bring up a mental health issue with a potential employer in an interview.

Physical health interview: this asked the survey takers if they would bring up a physical health issue with a potential employer in an interview.

Mental vs physical: this asked the survey takers if their employer takes mental health as seriously as physical health.

Obs consequence: this asked the survey takers if they ever heard of or observed negative consequences for coworkers with mental health conditions in their workplaces.

Interaction variables

Age and Family history: in my opinion, there is a positive effect from the combination of the two variables on the odds of Treatment.

If Age increases, the older the person is and the more vulnerable he/she is to mental illness. According to WHO fact sheet on Mental health of older adults: older people may experience life stressors common to all people, but also stressors that are more common in later life, like a significant ongoing loss in capacities and a decline in functional ability.

If there is a family history of mental illness, the person more likely inherits those illnesses. According to Lichtenstein et al. 2009: Schizophrenia has a heritability rate of 64 percent, and bipolar disorder has a heritability rate of 59 percent.

I also created the interaction variables between Age and Work interfere as well as Family history and Work interfere as Work interfere shows high positive correlation with Treatment. I am interested in observing the combined effects of them on the odds of Treatment.

Multicollinearity

The fact that most of the independent variables are qualitative and require dummy variables to be created for them makes results shown from multicollinearity matrix meaningless.

Except from Age and the interaction variables created from Age. However, according to Paul Allison from Statistical Horizons, those kinds of multicollinearity could be ignored.

Model selection method

Firstly, I applied the holdout method and split 66% of the data into training set and 34% of the data into testing set.

Then three selection methods were applied on the training set, Stepwise, Forward selection and Backward elimination and they all gave similar results with R-square of 0.4536, AIC = 1154 and SC = 1158.7. Figure 6 shows the results of the three selection methods.

The final model is:

LOG(TREATMENT=1/TREATMENT=0) = –3.4450 + 0.03*AGE – 0.85*GENDER3 +
1.1*NUMFAMILY_HISTORY + 3.43*NUMWORK_INTERFERE + 0.55*BENEFITS2 + 0.61*CARE_OPTIONS2

The cut-off value was determined as the value with highest sum of Sensitivity and Specificity, which is 0.52 in my case.

Predicted values were then compared with the observed values in the test set. Figure 7 shows the confusion matrix, from which statistics of the model were calculated as following:

Sensitivity	89.2%
Accuracy	84.1%
Precision	81.9%
Specificity	78.5%
F-metric	85.4%

The strongest predictor is Numwork_interfere with standardized estimate of 0.92, this variable represents survey takers who answered that their mental conditions do interfere their work. Figure 8 shows the table of standardized estimates.

Diagnostics were run on the training model to check for multicollinear and outliers, influential points. Figure shows there are no collinearity between the variables. Figure 9 shows there are 10 outliers (points higher than 3 or below -3 in Pearson residual) and figure 10 shows there are several influential points (points with |Dfbeta| > $2/\sqrt{831}$), however, after checking the data, all observations with outliers and influential points are within normal range so I decided to ignore the outliers and influential points.

Model assumptions are also not applied for this logistic regression model.

The final model is used to make two predictions: (see Figure 11 for details).

The probability that a male worker who is 25 years old, has no family history of mental illness and has no work interference from mental illness, whose employer provides mental health benefits and knows the options for mental health care being provided, is 9.5% with C.I of (5.8%, 15.3%).

The probability that a female worker who is 55 years old, has family history of mental illness and has work interference from mental illness, whose employer provides no mental health benefits but she knows the options for mental health care being provided is 160% with C.I of (148%, 166%).

References

1. Kaggle.com. 2018. Mental Health in Tech Survey. https://www.kaggle.com/osmi/mental-health-in-tech-survey. Accessed 18 March 2018.

2. World Health Organization. 2018. Mental health of older adults. http://www.who.int/mediacentre/factsheets/fs381/en/. Accessed 18 March 2018.

3. Lichtenstein et al. 2009. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. https://www.ncbi.nlm.nih.gov/pubmed/19150704. Accessed 18 March 2018.

4. Internet World Stats - ISO 3166. 2018. Country List by Geographical Regions. https://www.internetworldstats.com/list1.htm#EE. Accessed 18 March 2018.

5. U.S. Embassy. 2018. The Regions of the United States. https://usa.usembassy.de/travel-regions.htm. Accessed 18 March 2018.

6. Paul Allison. 2018. When can you safely ignore multicollinearity? https://statisticalhorizons.com/multicollinearity. Accessed 18 March 2018.

Appendix

Figure 1:

FREQUENCY-TREATMENT

The FREQ Procedure

treatment	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	622	49.40	622	49.40
1	637	50.60	1259	100.00

Figure 2:

FREQUENCY-AGE

The FREQ Procedure

Age	Frequency	Percent	Cumulative Frequency	Cumulative Percent
-1726	1	0.08	1	0.08
-29	1	0.08	2	0.16
-1	1	0.08	3	0.24
5	1	0.08	4	0.32
8	1	0.08	5	0.40
11	1	0.08	6	0.48
18	7	0.56	13	1.03
19	9	0.71	22	1.75
20	6	0.48	28	2.22
21	16	1.27	44	3.49
22	21	1.67	65	5.16
23	51	4.05	116	9.21
24	46	3.65	162	12.87
25	61	4.85	223	17.71
26	75	5.96	298	23.67
27	71	5.64	369	29.31
28	68	5.40	437	34.71
29	85	6.75	522	41.46
30	63	5.00	585	46.47

31	67	5.32	652	51.79
32	82	6.51	734	58.30
33	70	5.56	804	63.86
34	65	5.16	869	69.02
35	55	4.37	924	73.39
36	37	2.94	961	76.33
37	43	3.42	1004	79.75
38	39	3.10	1043	82.84
39	33	2.62	1076	85.46
40	33	2.62	1109	88.09
41	21	1.67	1130	89.75
42	20	1.59	1150	91.34
43	28	2.22	1178	93.57
44	11	0.87	1189	94.44
45	12	0.95	1201	95.39
46	12	0.95	1213	96.35
47	2	0.16	1215	96.51
48	6	0.48	1221	96.98
49	4	0.32	1225	97.30
50	6	0.48	1231	97.78
51	5	0.40	1236	98.17
53	1	0.08	1237	98.25
54	3	0.24	1240	98.49
55	3	0.24	1243	98.73
56	4	0.32	1247	99.05
57	3	0.24	1250	99.29
58	1	0.08	1251	99.36
60	2	0.16	1253	99.52
61	1	0.08	1254	99.60
62	1	0.08	1255	99.68
65	1	0.08	1256	99.76
72	1	0.08	1257	99.84
329	1	0.08	1258	99.92
99999999999	1	0.08	1259	100.00

FREQUENCY-AGE

The FREQ Procedure

Age	Frequency	Percent	Cumulative Frequency	Cumulative Percent
18	7	0.56	7	0.56
19	9	0.71	16	1.27
20	6	0.48	22	1.75
21	16	1.27	38	3.02
22	21	1.67	59	4.69
23	51	4.05	110	8.74
24	46	3.65	156	12.39
25	61	4.85	217	17.24
26	75	5.96	292	23.19
27	71	5.64	363	28.83
28	68	5.40	431	34.23
29	85	6.75	516	40.98
30	63	5.00	579	45.99
31	77	6.12	656	52.10
32	82	6.51	738	58.62
33	70	5.56	808	64.18
34	65	5.16	873	69.34
35	55	4.37	928	73.71
36	37	2.94	965	76.65
37	43	3.42	1008	80.06
38	39	3.10	1047	83.16
39	33	2.62	1080	85.78
40	33	2.62	1113	88.40
41	21	1.67	1134	90.07
42	20	1.59	1154	91.66
43	28	2.22	1182	93.88
44	11	0.87	1193	94.76
45	12	0.95	1205	95.71
46	12	0.95	1217	96.66
47	2	0.16	1219	96.82
48	6	0.48	1225	97.30
49	4	0.32	1229	97.62
50	6	0.48	1235	98.09
51	5	0.40	1240	98.49
53	1	0.08	1241	98.57
54	3	0.24	1244	98.81
55	3	0.24	1247	99.05
56	4	0.32	1251	99.36
57	3	0.24	1254	99.60
58	1	0.08	1255	99.68
60	2	0.16	1257	99.84
61	1	0.08	1258	99.92
62	1	0.08	1259	100.00

Figure 3:

FREQUENCY-GENDER

The FREQ Procedure

Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A litt	1	0.08	1	0.08
Agende	1	0.08	2	0.16
All	1	0.08	3	0.24
Androg	1	0.08	4	0.32
Cis Fe	1	0.08	5	0.40
Cis Ma	3	0.24	8	0.64
Enby	1	0.08	9	0.71
F	38	3.02	47	3.73
Femake	1	0.08	48	3.81
Female	126	10.01	174	13.82
Gender	1	0.08	175	13.90
Guy (-	1	0.08	176	13.98
M	116	9.21	292	23.19
Mail	1	0.08	293	23.27
Make	4	0.32	297	23.59
Mal	1	0.08	298	23.67
Male	618	49.09	916	72.76
Male (	1	0.08	917	72.84
Male-i	1	0.08	918	72.92
Malr	1	0.08	919	72.99
Man	2	0.16	921	73.15
Nah	1	0.08	922	73.23
Neuter	1	0.08	923	73.31
Trans	1	0.08	924	73.39
Trans-	1	0.08	925	73.47
Woman	3	0.24	928	73.71
cis ma	1	0.08	929	73.79
cis-fe	1	0.08	930	73.87
f	15	1.19	945	75.06
femail	1	0.08	946	75.14
female	62	4.92	1008	80.06
fluid	1	0.08	1009	80.14
m	34	2.70	1043	82.84
maile	1	0.08	1044	82.92
male	206	16.36	1250	99.29
male l	1	0.08	1251	99.36
msle	1	0.08	1252	99.44
non-bi	1	0.08	1253	99.52
ostens	1	0.08	1254	99.60
p	1	0.08	1255	99.68
queer	1	0.08	1256	99.76
queer/	1	0.08	1257	99.84
someth	1	0.08	1258	99.92
woman	1	0.08	1259	100.00

FREQUENCY-GENDER

The FREQ Procedure

Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	247	19.62	247	19.62
Male	991	78.71	1238	98.33
Others	21	1.67	1259	100.00

Figure 4:

FREQUENCY-COUNTRY

The FREQ Procedure

Country	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Australia	21	1.67	21	1.67
Austria	3	0.24	24	1.91
Bahamas, The	1	0.08	25	1.99
Belgium	6	0.48	31	2.46
Bosnia and Her	1	0.08	32	2.54
Brazil	6	0.48	38	3.02
Bulgaria	4	0.32	42	3.34
Canada	72	5.72	114	9.05
China	1	0.08	115	9.13
Colombia	2	0.16	117	9.29
Costa Rica	1	0.08	118	9.37
Croatia	2	0.16	120	9.53
Czech Republic	1	0.08	121	9.61
Denmark	2	0.16	123	9.77
Finland	3	0.24	126	10.01
France	13	1.03	139	11.04
Georgia	1	0.08	140	11.12
Germany	45	3.57	185	14.69
Greece	2	0.16	187	14.85
Hungary	1	0.08	188	14.93
India	10	0.79	198	15.73
Ireland	27	2.14	225	17.87
Israel	5	0.40	230	18.27
Italy	7	0.56	237	18.82

Country	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Japan	1	0.08	238	18.90
Latvia	1	0.08	239	18.98
Mexico	3	0.24	242	19.22
Moldova	1	0.08	243	19.30
Netherlands	27	2.14	270	21.45
New Zealand	8	0.64	278	22.08
Nigeria	1	0.08	279	22.16
Norway	1	0.08	280	22.24
Philippines	1	0.08	281	22.32
Poland	7	0.56	288	22.88
Portugal	2	0.16	290	23.03
Romania	1	0.08	291	23.11
Russia	3	0.24	294	23.35
Singapore	4	0.32	298	23.67
Slovenia	1	0.08	299	23.75
South Africa	6	0.48	305	24.23
Spain	1	0.08	306	24.31
Sweden	7	0.56	313	24.86
Switzerland	7	0.56	320	25.42
Thailand	1	0.08	321	25.50
United Kingdom	185	14.69	506	40.19
United States	751	59.65	1257	99.84
Uruguay	1	0.08	1258	99.92
Zimbabwe	1	0.08	1259	100.00

FREQUENCY-STATE

The FREQ Procedure

state	Frequency	Percent	Cumulative Frequency	Cumulative Percent
AL	8	0.64	8	0.64
AZ	7	0.56	15	1.19
CA	138	10.96	153	12.15
CO	9	0.71	162	12.87
CT	4	0.32	166	13.19
DC	4	0.32	170	13.50
FL	15	1.19	185	14.69
GA	12	0.95	197	15.65
IA	4	0.32	201	15.97
ID	1	0.08	202	16.04
IL	29	2.30	231	18.35
IN	27	2.14	258	20.49
KS	3	0.24	261	20.73
KY	5	0.40	266	21.13
LA	1	0.08	267	21.21
MA	20	1.59	287	22.80
MD	8	0.64	295	23.43
ME	1	0.08	296	23.51
MI	22	1.75	318	25.26
MN	21	1.67	339	26.93
MO	12	0.95	351	27.88
MS	1	0.08	352	27.96
NA	515	40.91	867	68.86
NC	14	1.11	881	69.98
NE	2	0.16	883	70.14
NH	3	0.24	886	70.37
NJ	6	0.48	892	70.85
NM	2	0.16	894	71.01
NV	3	0.24	897	71.25
NY	57	4.53	954	75.77
OH	30	2.38	984	78.16
OK	6	0.48	990	78.63
OR	29	2.30	1019	80.94
PA	29	2.30	1048	83.24
RI	1	0.08	1049	83.32
SC	5	0.40	1054	83.72
SD	3	0.24	1057	83.96
TN	45	3.57	1102	87.53
TX	44	3.49	1146	91.02
UT	11	0.87	1157	91.90
VA	14	1.11	1171	93.01
VT	3	0.24	1174	93.25
WA	70	5.56	1244	98.81
WI	12	0.95	1256	99.76
WV	1	0.08	1257	99.84
WY	2	0.16	1259	100.00

FREQUENCY-REGION

The FREQ Procedure

Region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
America	14	1.11	14	1.11
Asia	19	1.51	33	2.62
Canada	72	5.72	105	8.34
Europe	15	1.19	120	9.53
European Union	346	27.48	466	37.01
Oceania	29	2.30	495	39.32
Others	24	1.91	519	41.22
US Mid Atlantic	102	8.10	621	49.32
US Midwest	164	13.03	785	62.35
US New England	32	2.54	817	64.89
US South	121	9.61	938	74.50
US Southwest	59	4.69	997	79.19
US West	262	20.81	1259	100.00

Figure 5:

self_employed	Frequency	Percent	Cumulative Frequency	Cumulative Percent
NA	18	1.43	18	1.43
No	1095	86.97	1113	88.40
Yes	146	11.60	1259	100.00

FREQUENCY-SELF EMPLOYED

The FREQ Procedure

self_employed	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	1111	88.24	1111	88.24
Yes	148	11.76	1259	100.00

FREQUENCY-FAMILY HISTORY

The FREQ Procedure

family_history	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	767	60.92	767	60.92
Yes	492	39.08	1259	100.00

FREQUENCY-TREATMENT

The FREQ Procedure

treatment	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	622	49.40	622	49.40
Yes	637	50.60	1259	100.00

FREQUENCY-WORK INTERFERE

The FREQ Procedure

work_interfere	Frequency	Percent	Cumulative Frequency	Cumulative Percent
NA	264	20.97	264	20.97
Never	213	16.92	477	37.89
Often	144	11.44	621	49.32
Rarely	173	13.74	794	63.07
Sometimes	465	36.93	1259	100.00

FREQUENCY-WORK INTERFERE

The FREQ Procedure

work_interfere	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	477	37.89	477	37.89
Yes	782	62.11	1259	100.00

FREQUENCY-NO of EMPLOYEES

The F REQ Procedure

no_employees	Frequency	Percent	Cumulative Frequency	Cumulative Percent
100-500	176	13.98	176	13.98
26-100	289	22.95	465	36.93
43105	162	12.87	627	49.80
43276	290	23.03	917	72.84
500-1000	60	4.77	977	77.60
More than 1000	282	22.40	1259	100.00

FREQUENCY-NO EMPLOYEES

The FREQ Procedure

no_employees	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Large	282	22.40	282	22.40
Medium	236	18.75	518	41.14
Small	741	58.86	1259	100.00

The F REQ Procedure

remote_work	F requency	Percent	Cumulative Frequency	Cumulative Percent
No	883	70.14	883	70.14
Yes	376	29.86	1259	100.00

The F REQ Procedure

tech_company	F requency	Percent	Cumulative Frequency	Cumulative Percent
No	228	18.11	228	18.11
Yes	1031	81.89	1259	100.00

The F REQ Procedure

benefits	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Don't know	408	32.41	408	32.41
No	374	29.71	782	62.11
Yes	477	37.89	1259	100.00

The FREQ Procedure

care_options	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	501	39.79	501	39.79
Not sure	314	24.94	815	64.73
Yes	444	35.27	1259	100.00

The FREQ Procedure

wellness_program	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Don't know	188	14.93	188	14.93
No	842	66.88	1030	81.81
Yes	229	18.19	1259	100.00

The FREQ Procedure

seek_help	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Don't know	363	28.83	363	28.83
No	646	51.31	1009	80.14
Yes	250	19.86	1259	100.00

The FREQ Procedure

anonymity	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Don't know	819	65.05	819	65.05
No	65	5.16	884	70.21
Yes	375	29.79	1259	100.00

The FREQ Procedure

leave	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Don't know	563	44.72	563	44.72
Somewhat difficult	126	10.01	689	54.73
Somewhat easy	266	21.13	955	75.85
Very difficult	98	7.78	1053	83.64
Very easy	206	16.36	1259	100.00

FREQUENCY-LEAVE

The FREQ Procedure

leave	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Don't know	563	44.72	563	44.72
No	224	17.79	787	62.51
Yes	472	37.49	1259	100.00

The FREQ Procedure

mental_health_consequence	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Maybe	477	37.89	477	37.89
No	490	38.92	967	76.81
Yes	292	23.19	1259	100.00

The FREQ Procedure

phys_health_consequence	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Maybe	273	21.68	273	21.68
No	925	73.47	1198	95.15
Yes	61	4.85	1259	100.00

The FREQ Procedure

coworkers	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	260	20.65	260	20.65
Some of them	774	61.48	1034	82.13
Yes	225	17.87	1259	100.00

FREQUENCY-COWORKERS

The FREQ Procedure

coworkers	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	260	20.65	260	20.65
Yes	999	79.35	1259	100.00

The FREQ Procedure

supervisor	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	393	31.22	393	31.22
Some of them	350	27.80	743	59.02
Yes	516	40.98	1259	100.00

FREQUENCY-SUPERVISOR

The FREQ Procedure

supervisor	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	393	31.22	393	31.22
Yes	866	68.78	1259	100.00

The FREQ Procedure

mental_health_interview	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Maybe	207	16.44	207	16.44
No	1008	80.06	1215	96.51
Yes	44	3.49	1259	100.00

The FREQ Procedure

phys_health_interview	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Maybe	557	44.24	557	44.24
No	500	39.71	1057	83.96
Yes	202	16.04	1259	100.00

The FREQ Procedure

mental_vs_physical	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Don't know	576	45.75	576	45.75
No	340	27.01	916	72.76
Yes	343	27.24	1259	100.00

The FREQ Procedure

obs_consequence	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	1075	85.39	1075	85.39
Yes	184	14.61	1259	100.00

Figure 6:

Step 6. Effect Age entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1154.000	663.699
SC	1158.722	696.758
-2 Log L	1152.000	649.699

R-Square	0.4536	Max-rescaled R-Square	0.6048

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	502.3005	6	<.0001
Score	419.1478	6	<.0001
Wald	244.1919	6	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
30.0009	44	0.9469

Note: No effects for the model in Step 6 are removed.

Note: No (additional) effects met the 0.05 significance level for entry into the model.

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	numwork_interfere		1	1	374.2379		<.0001
2	numfamily_history		1	2	36.6057		<.0001
3	benefits2		1	3	20.5936		<.0001
4	gender3		1	4	9.1789		0.0024
5	care_options2		1	5	8.1639		0.0043
6	Age		1	6	4.5553		0.0328

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.4450	0.5408	40.5852	<.0001
Age	1	0.0300	0.0141	4.5211	0.0335
gender3	1	-0.8489	0.2627	10.4460	0.0012
numfamily_history	1	1.1011	0.2068	28.3448	<.0001
numwork_interfere	1	3.4262	0.2464	193.3555	<.0001
benefits2	1	0.5468	0.2318	5.5660	0.0183
care_options2	1	0.6071	0.2307	6.9251	0.0085

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Age	1.030	1.002	1.059

gender3	0.428	0.256	0.716
numfamily_history	3.007	2.005	4.511
numwork_interfere	30.759	18.978	49.855
benefits2	1.728	1.097	2.721
care_options2	1.835	1.168	2.884

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	89.9	Somers' D	0.800
Percent Discordant	9.9	Gamma	0.801
Percent Tied	0.2	Tau-a	0.401
Pairs	172638	c	0.900

Step 6. Effect Age entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1154.000	663.699
SC	1158.722	696.758
-2 Log L	1152.000	649.699

R-Square	0.4536	Max-rescaled R-Square	0.6048

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	502.3005	6	<.0001
Score	419.1478	6	<.0001
Wald	244.1919	6	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
30.0009	44	0.9469

Note: No (additional) effects met the 0.05 significance level for entry into the model.

Summary of Forward Selection					
Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	numwork_interfere	1	1	374.2379	<.0001
2	numfamily_history	1	2	36.6057	<.0001
3	benefits2	1	3	20.5936	<.0001
4	gender3	1	4	9.1789	0.0024
5	care_options2	1	5	8.1639	0.0043
6	Age	1	6	4.5553	0.0328

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.4450	0.5408	40.5852	<.0001
Age	1	0.0300	0.0141	4.5211	0.0335
gender3	1	-0.8489	0.2627	10.4460	0.0012
numfamily_history	1	1.1011	0.2068	28.3448	<.0001
numwork_interfere	1	3.4262	0.2464	193.3555	<.0001
benefits2	1	0.5468	0.2318	5.5660	0.0183
care_options2	1	0.6071	0.2307	6.9251	0.0085

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Age	1.030	1.002	1.059

gender3	0.428	0.256	0.716
numfamily_history	3.007	2.005	4.511
numwork_interfere	30.759	18.978	49.855
benefits2	1.728	1.097	2.721
care_options2	1.835	1.168	2.884

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	89.9	Somers' D	0.800
Percent Discordant	9.9	Gamma	0.801
Percent Tied	0.2	Tau-a	0.401
Pairs	172638	c	0.900

Step 44. Effect age_family_history is removed:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1154.000	663.699
SC	1158.722	696.758
-2 Log L	1152.000	649.699

R-Square	0.4536	Max-rescaled R-Square	0.6048

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	502.3005	6	<.0001
Score	419.1478	6	<.0001
Wald	244.1919	6	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
30.0009	44	0.9469

Note: No (additional) effects met the 0.05 significance level for removal from the model.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.4450	0.5408	40.5852	<.0001
Age	1	0.0300	0.0141	4.5211	0.0335
gender3	1	-0.8489	0.2627	10.4460	0.0012
numfamily_history	1	1.1011	0.2068	28.3448	<.0001
numwork_interfere	1	3.4262	0.2464	193.3555	<.0001
benefits2	1	0.5468	0.2318	5.5660	0.0183
care_options2	1	0.6071	0.2307	6.9251	0.0085

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Age	1.030	1.002	1.059
gender3	0.428	0.256	0.716
numfamily_history	3.007	2.005	4.511
numwork_interfere	30.759	18.978	49.855
benefits2	1.728	1.097	2.721
care_options2	1.835	1.168	2.884

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	89.9	Somers' D	0.800
Percent Discordant	9.9	Gamma	0.801
Percent Tied	0.2	Tau-a	0.401
Pairs	172638	c	0.900

Figure 7:

Fit logistic regression

The FREQ Procedure

Frequency	Table of treatment by pred_dis		
	pred_dis		
treatment	0	1	Total
0	161	44	205
1	24	199	223
Total	185	243	428

Figure 8:

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1154.000	663.699
SC	1158.722	696.758
-2 Log L	1152.000	649.699

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	502.3005	6	<.0001
Score	419.1478	6	<.0001
Wald	244.1919	6	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-3.4450	0.5408	40.5852	<.0001	
Age	1	0.0300	0.0141	4.5211	0.0335	0.1190
gender3	1	-0.8489	0.2627	10.4460	0.0012	-0.1905
numfamily_history	1	1.1011	0.2068	28.3448	<.0001	0.2961
numwork_interfere	1	3.4262	0.2464	193.3555	<.0001	0.9213
benefits2	1	0.5468	0.2318	5.5660	0.0183	0.1462
care_options2	1	0.6071	0.2307	6.9251	0.0085	0.1603

Figure 9:

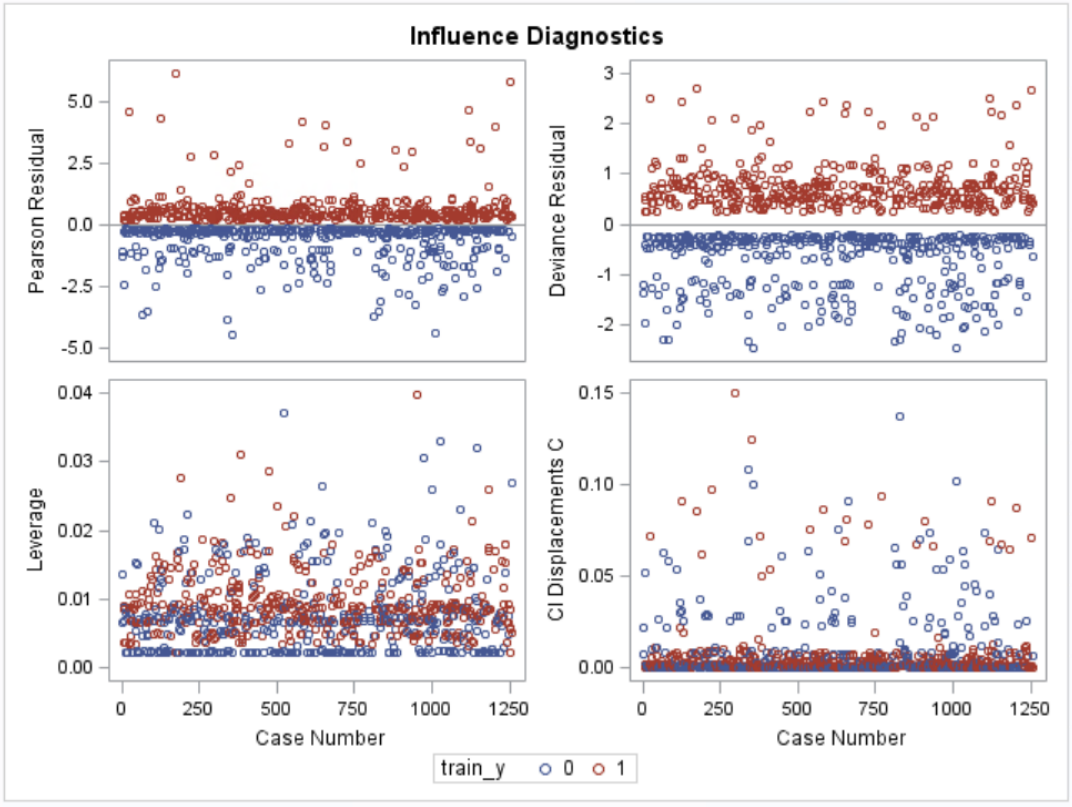


Figure 10:

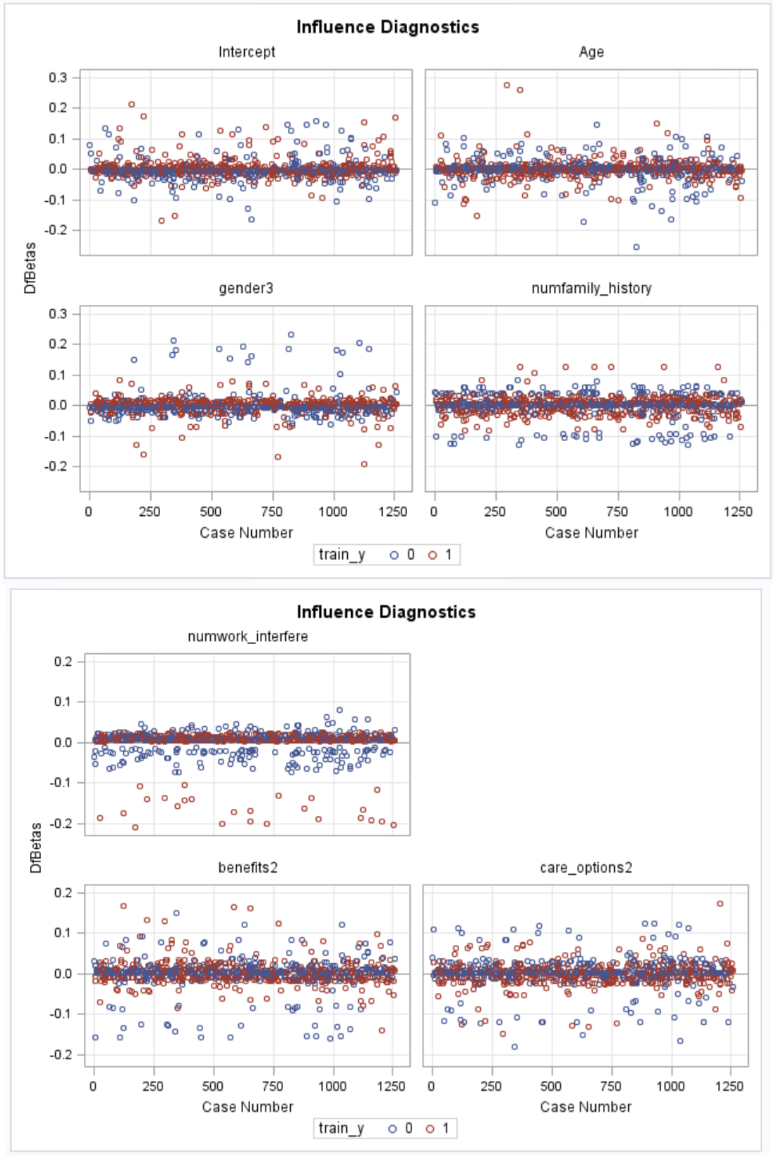


Figure 11:

Obs	age	gender3	numfamily_history	numwork_interfere	benefits2	care_options2	_LEVEL_	phat	lcl	ucl
1	25	1	0	0	1	1	1	0.09059	0.05648	0.14221
2	55	0	1	1	0	1	1	0.95573	0.90746	0.97939

Ana Morrissey

Introduction

Goal is the same as the group, refer to the introduction section in the main report for details.

Methodology

The data analyzed is from the Survey on Mental Health in the Tech Workplace conducted in 2014, downloaded from Kaggle [(1)].

Data

The original data contained 1,259 observations and 27 variables: 1 date/time; 2 location variables (country and state); 1 numerical (age); 1 ordinal (number of employees); 21 text variables, 19 with yes, no, don't know/not sure responses and 2 with varying levels. It was determined that timestamps and comments were not going to be included in the analysis and were removed from the dataset. Because the main objective of this project is to uncover the strongest indicators in determining whether an individual in a tech workplace will seek treatment for a mental illness, the ***treatment*** variable was selected as the response variable, leaving 24 potential variables to create the model.

Variable Creation

All variables were explorer for possible data entry errors and missing data. Additionally, the distribution and relationship of age with the treatment variable was examined; frequency rates were examined for all variables. Several variables were recoded for easier analysis, these were gender, country and state. All variables had varying levels of dummy variables created to facilitate the analysis. A breakdown of the dummy variables and interaction variables created for each variable are provided in Appendix A.

Once the exploratory analysis was completed a full logistic model was fitted to check for outliers, influential points and collinearity among variables. Pearson and Deviance residual diagnostics were examined to detect and remove outliers close to or exceeding ±3 and Levee failure diagnostics to detect and remove influential points with $|Dfbeta| > 2/\sqrt{n}$. A total of 34 observations were removed.

Due to the binary nature of the treatment variable logistic regression model was applied. The model uses log odds so the y-variable did not need to be transformed. Scatterplots of paired data and linearity of relationship were not examined because the y-variable is binary, points are concentrated at 0 or 1.

Model Approach

To evaluate the predictive power of the model, the data was split into training and test set (80/20), the model was fit using the training set and performance was tested on the test set. After testing various model selection methods (forward, backward and stepwise) selection methods, models were compared based on selection criteria the model with the fewest predictors as well as lowest AIC, SC values; Stepwise was selected.

Validation Methods

Sensitivity or Recall, Accuracy, Precision and Specificity value were calculated to measure the performance of the model.

Analysis, Results and Findings

Exploratory analysis

The dependent variable for this data is binary, therefore y-variable transformation is not required; linear relationships assumptions do not need to be met. The distributions, frequency rates and relationship with the treatment variable were examined. Logistic regression was applied to the dataset.

Figure 1 in Appendix B, shows the frequency of individuals who had sought treatment for a mental health condition were split *almost* 50/50. Additionally, the distribution of age was slightly skewed to the right with most of the participants ages falling under 32 years old. A possible outlier being the individual who is 72 years old (Appendix B, Figure 2).

The gender variable also seems to be unevenly distributed among the gender groups (Appendix, Figure 3). There is a larger number of males (979) than females (246) or non-binary individuals (17).

Interaction Variables

Two interaction variables were created. It was of interest to know if family history and gender as well as age and gender were significant predictors. A full model was fitted and it was determined that the interaction variables were not significant (Appendix B, Figure 4). They were removed from further analysis.

Collinearity

Most of the variables in this dataset are comprised of dummy variables; therefore, multicollinearity was tested on only two variables, age and the interaction variable gender*age. Figure 5 in Appendix B shows high correlation among the age and the gender*age interactions variables. Because the interactions were deemed non-significant no additional analysis was necessary.

Residuals and Influential points

Residual plots were examined solely as secondary method to detect outliers. A total of 34 observations were removed because they had Pearson and/or Deviance residuals that exceed +-3 (Appendix B, Figure 6). After the variables were removed a full model was fitted; the AIC and SC were slightly lower than the model containing the outliers (Appendix B, Figure 7).

Model

To evaluate the predictive power of the model, the data was split into training and test set (80/20), the model was fit using the training set and performance was tested on the test set. The model was built using the training set three different model selection methods were conducted and the predicted probability were calculated on the final model.

Model Selection

Stepwise, Backward and Forward Selection methods were applied to the data to determine the best set of predictive variables.

A comparison of the models can be found on Appendix B, Figure 8. Stepwise and Forward methods selected the same predictors (12 predictors) and had the same AIC, SC, Likelihood rations, and likelihood estimates. Backward method also had the same AIC, SC, Likelihood rations, and likelihood estimates but it contained a total 16 predictors, because of this the Stepwise method was selected.

Strongest Predictor

The strongest predictor is dwork_interfereS which is part of the was the work_interfere variable, which asked participants, “If you have a mental health condition, do you feel that it interferes with your work?” dwork_interfereS is the dummy variable for the response *Sometimes*. The dwork_interfereS variable has the highest standardized estimate of 1.79, followed by the other members of the work_interfere variables. Figure 10 ranks the variables by the strength level. The odds ratio are summarized in Figure 9, which provides evidence to work_interfere being the strongest predictor. If an individual feels that their mental health condition is interfering with work they are more likely to seek treatment.

Final Model

The final model had a likelihood ration of 758.92 with a p-value of <.0001. We can reject the null hypothesis and go with the alternative hypothesis, that there is at least one variable, in our case 12, that can predict if an individual will seek treatment for a mental health condition (Appendix B, Figure 9).

Final Model Equation

Log(treatment=1/treatment=0) = -7.11 + 0.04age – 1.10dgenderM + 0.59dstate_regionW + 1.55dfamily_history + 3.14dwork_interfereN + 7.54dwork_interfereO + 6.15dwork_interfereR + 6.75dwork_interfereS + 0.74dbenefitsY +0.85dcare_optionsY + 0.91dcoworkersY – 0.52dcoworkersN

Probability Threshold

A classification table was created to determine the probability threshold (Appendix B, Figure 11). Two levels were found to maximize Specificity and Sensitivity (0.55 and 0.6); therefore, a classification matrix was created for each (Appendix B, Figure 12 and 13). Threshold .55 proved to be a better threshold point because it had a greater sensitivity and accuracy percentage.

The calculated metrics are listed below:

- Sensitivity or recall: 92%
- Accuracy: 85%
- Precision: 79.4%
- Specificity: 78.9%

Calculations are available on Appendix B, Figure 11

Predictions

Two predictions were run against the final model (Appendix B, Figure 14). A description of the made-up observations can be found in Figure 14. The predicted probabilities for the created observations were:

Prediction #1 Predicted probability= 0.9702 with a 95% of the time, the predicted probability will fall within 0.94009, 0.98548

Prediction #2 Predicted probability= 0.7941 with a 95% of the time, the predicted probability will fall within 0.65802, 0.88554

The probabilities of each prediction lie within the prediction intervals.

Model Improvement

There were several things that could be improved in the model. One of them would be to obtain an even number of participants from the various gender groups, I believe that the reason why the other gender groups were not statistically significant was because there weren't enough data points. Additionally, it would be interesting to know if this model would apply to other industries not just technological companies. Also, being able to compared countries would have provided greater insights into how mental illness is viewed in different areas of the world.

Future Work

Refer to group's future work section.

Appendix A – Pre-processed Data

Below you will find the details on how the data was prepared for analysis, the creation of dummy variables and interaction variables.

Gender

Monro[2] argues that gender classification is non-binary and therefore the classification should not be solely male and female. As evidence to his argument, the gender variable had 41 unique responses, that range from female, male, queer, non-binary, trans, etc. Because of this the gender variable was recoded to 3 levels: female, male, and non-binary. Four (4) observations were deleted because of inadequate responses.

Location Variables

The survey was an international survey that contained responses from a total of 48 unique countries. While it would be ideal to determine differences among countries, there were not enough responses for each country to do so. So instead, countries were recoded into 5 regions, the regions were selected based on the United Nations Department of Economic and Social Affairs Statistics Division geographical regions [3]. The 48 countries were grouped into the following regions: Africa, Americas, Asia, Europe, and Oceania; the Americas was used as the base. Regions with fewer than 10 observations were removed from the analysis, Africa contained less than 10 observations so it was excluded from the analysis. The country variable was renamed to geo_region.

The State variable consisted of 46 states, these were recoded into 4 regions, Midwest, Northeast, South and West based on the classification of regions and divisions by the United States Census Bureau[4]. The state variable was renamed to state_region.

All other variables

The following dummy variables were created for the remaining variables:

Variable	Options	Dummy Level
Self_Employed	Yes, No, NA	2
Family_history	Yes, No	1
Treatment	Yes, No	1
Work_interfere	Never, Often, Rarely, Sometimes, NA	4
No_employees	1-5; 6-25; 26-100; 100-500; 500-1000; 1000+	5
Remote_work	Yes, No	1
Tech_Company	Yes, No	2
Benefits	Yes, No, Don't Know	2
Care_Options	Yes, No, Not Sure	2
Well_Programs	Yes, No, Don't Know	2
Seek_Help	Yes, No, Don't Know	2
Anonymity	Yes, No, Don't Know	2
Leave	Somewhat Difficult; Somewhat Easy; Very Difficult; Very Easy; Don't KNow	4
Mental_Health_Consequence	Yes, No, Maybe	2
Phys_Health_Consequence	Yes, No, Maybe	2
Coworkers	Yes, No, Some	2
Supervisor	Yes, No, Some	2
Mental_health_interview	Yes, No, Maybe	2
Phys_health_interview	Yes, No, Maybe	2
Obs_consequence	Yes, No	1

Appendix B -Relevant Output

Treatment (Dependent Variable)

Figure 1 – Treatment frequency

Q: Q: Have you sought treatment for a mental health condition?

Frequency treatment

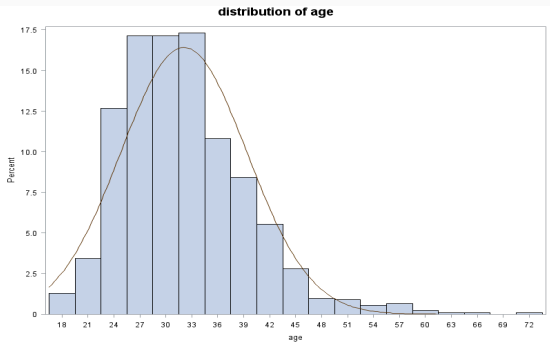
The FREQ Procedure

dtreatment	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	616	49.60	616	49.60
1	626	50.40	1242	100.00

1= Yes 0 = No

Independent Variables

Figure 2 - Age



Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
18	477	60	295
18	473	60	1221
18	285	62	515
18	280	65	555
18	119	72	1167

Basic Statistical Measures			
Location		Variability	
Mean	32.06602	Std Deviation	7.25669
Median	31.00000	Variance	52.65962
Mode	29.00000	Range	54.00000
		Interquartile Range	9.00000

There were 7 participants who were 18 years old and five (5) participants over the age of 60. The distribution of age is slightly skewed to the right with most of the participants ages falling under 32 years old. A possible outlier being the individual who is 72 years old.

Figure 3 - Gender

Frequency Gender

The FREQ Procedure

dgenderF	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	996	80.19	996	80.19
1	246	19.81	1242	100.00

dgenderM	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	263	21.18	263	21.18
1	979	78.82	1242	100.00

There seems to be a higher number of males (979) who responded to the survey compared to females (246) and non-binary (17). This could be largely to the fact that more males than any other gender group works in the tech industry.

Interaction Variables, Multicollinearity, Influential Points,

Figure 4 – Variables (Interaction) Significance

Analysis of Maximum Likelihood Estimates

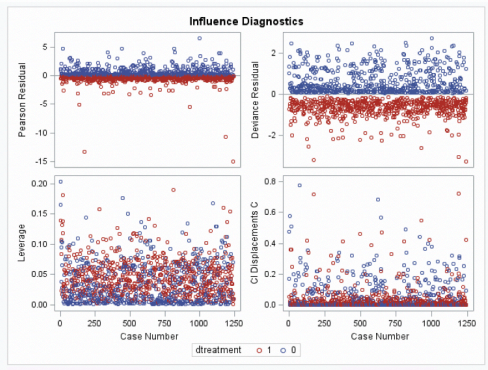
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.2180	2.9266	3.1789	0.0746
age	1	-0.0406	0.0915	0.1967	0.6574
dgenderF	1	-0.1460	2.8948	0.0025	0.9598
dgenderM	1	-1.9809	2.7677	0.5123	0.4742
dgeo_regionAs	1	-0.6473	0.6675	0.9404	0.3322
dgeo_regionEu	1	0.3682	0.3521	1.0932	0.2958
dgeo_regionOc	1	0.1123	0.6133	0.0335	0.8547
dstate_regionMW	1	0.3773	0.3902	0.9351	0.3335
dstate_regionNE	1	-0.1339	0.4294	0.0972	0.7552
dstate_regionS	1	-0.0459	0.3796	0.0146	0.9038
dstate_regionW	1	0.4781	0.3740	1.6342	0.2011
dself_employedY	1	0.4469	0.7458	0.3591	0.5490
dself_employedN	1	0.5958	0.6802	0.7672	0.3811
dfamily_history	1	2.1924	1.5822	1.9201	0.1658
dwork_interfereN	1	2.4586	0.5616	19.1660	<.0001
dwork_interfereO	1	5.9738	0.5936	101.2887	<.0001
dwork_interfereR	1	4.9201	0.5561	78.2642	<.0001
dwork_interfereS	1	5.3370	0.5405	97.5042	<.0001
dno_employees1	1	0.0806	0.3964	0.0413	0.8389
dno_employees6	1	0.2103	0.3110	0.4574	0.4988
dno_employees26	1	0.4075	0.2926	1.9392	0.1638
dno_employees100	1	0.3292	0.3157	1.0874	0.2971
dno_employees500	1	0.3681	0.4744	0.6022	0.4377
dremote_work	1	-0.0614	0.2087	0.0866	0.7685
dtech_company	1	-0.1538	0.2488	0.3823	0.5364
dbenefitsY	1	0.6914	0.2726	6.4341	0.0112
dbenefitsN	1	0.1094	0.2708	0.1634	0.6861
dcare_optionsY	1	0.9127	0.2765	10.8973	0.0010
dwellness_programY	1	-0.3010	0.3639	0.6841	0.4082
dwellness_programN	1	0.0135	0.3063	0.0019	0.9648
dseek_helpY	1	-0.7424	0.3253	5.2096	0.0225
dseek_helpN	1	-0.6252	0.2628	5.6585	0.0174
danonymityY	1	0.5061	0.2351	4.6335	0.0314
danonymityN	1	-0.00327	0.3991	0.0001	0.9935
dleaveSd	1	0.4211	0.3176	1.7580	0.1849
dleaveSe	1	-0.2908	0.2339	1.5450	0.2139
dleaveVd	1	0.1679	0.3493	0.2311	0.6307
dleaveVe	1	0.1480	0.2875	0.2649	0.6067
dmental_health_conse	1	-0.1212	0.2412	0.2525	0.6154
dmental_health_conse	1	-0.1966	0.2312	0.7236	0.3950
dphys_health_conseqY	0	0	.	.	.
dphys_health_conseqN	0	0	.	.	.
dcoworkersY	1	0.6248	0.2668	5.4845	0.0192
dcoworkersN	1	-0.3190	0.2293	1.9363	0.1641
dsupervisorY	0	0	.	.	.
dsupervisorN	0	0	.	.	.
dmental_health_intY	1	0.7102	0.6090	1.3602	0.2435
dmental_health_intN	1	0.3567	0.2844	1.5728	0.2098
dphys_health_intY	1	0.4740	0.2903	2.6663	0.1025
dphys_health_intN	1	0.1734	0.2021	0.7364	0.3908
dmentalphysY	1	-0.0510	0.2451	0.0434	0.8350
dmentalphysN	1	-0.0322	0.2278	0.0199	0.8877
dobs_conseqY	1	0.2435	0.2689	0.8200	0.3652
familyhistory_gender	1	-1.6626	1.6325	1.0372	0.3085
familyhistory_gender	1	-1.1805	1.5969	0.5465	0.4598
age_genderF	1	0.0405	0.0958	0.1784	0.6728
age_genderM	1	0.0681	0.0920	0.5487	0.4589

Figure 5 – Collinearity

Estimated Correlation Matrix

Parameter	Intercept	age	age_genderF	age_genderM
Intercept	1.0000	-0.4188	0.0427	0.0782
age	-0.4188	1.0000	-0.9045	-0.9341
age_genderF	0.0427	-0.9045	1.0000	0.9724
age_genderM	0.0782	-0.9341	0.9724	1.0000

Figure 6 – Influence Diagnostics



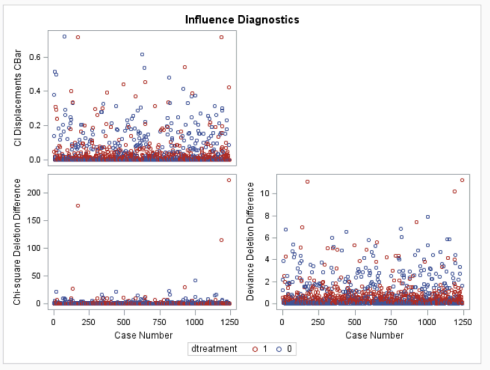


Figure 7 – Removal of Outliers/Influential points

Initial

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1723.697	993.868
SC	1728.822	1270.589
-2 Log L	1721.697	885.868

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	835.8295	53	<.0001
Score	662.9146	53	<.0001
Wald	310.5079	53	<.0001

After outliers/influential points were removed

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1676.312	762.947
SC	1681.409	1017.783
-2 Log L	1674.312	662.947

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1011.3654	49	<.0001
Score	745.9306	49	<.0001
Wald	252.9333	49	<.0001

Figure 8 Model Selection Method

Stepwise

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1341.280	606.354
SC	1346.154	669.719
-2 Log L	1339.280	580.354

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	758.9251	12	<.0001
Score	578.6857	12	<.0001
Wald	214.2439	12	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
50.1004	37	0.0737

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-7.1164	1.1504	38.2665	<.0001	
age	1	0.0461	0.0151	9.3396	0.0022	0.1849
dgenderM	1	-1.1030	0.3037	13.1879	0.0003	-0.2422
dstate_regionW	1	0.5943	0.2929	4.1178	0.0424	0.1335
dfamily_history	1	1.5515	0.2339	43.9937	<.0001	0.4140
dwork_interfereN	1	3.1477	1.0430	9.1071	0.0025	0.6403
dwork_interfereO	1	7.5429	1.0703	49.6654	<.0001	1.3159
dwork_interfereR	1	6.1565	1.0328	35.5350	<.0001	1.2055
dwork_interfereS	1	6.7530	1.0232	43.5610	<.0001	1.7934
dbenefitsY	1	0.7460	0.2688	7.7044	0.0055	0.1991
dcare_optionsY	1	0.8527	0.2526	11.3910	0.0007	0.2235
dcoworkersY	1	0.9194	0.3154	8.4995	0.0036	0.1965
dcoworkersN	1	-0.5260	0.2524	4.3410	0.0372	-0.1182

Forward

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1341.280	606.354
SC	1346.154	669.719
-2 Log L	1339.280	580.354

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	758.9251	12	<.0001
Score	578.6857	12	<.0001
Wald	214.2439	12	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
50.1004	37	0.0737

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-7.1164	1.1504	38.2665	<.0001	
age	1	0.0461	0.0151	9.3396	0.0022	0.1849
dgenderM	1	-1.1030	0.3037	13.1879	0.0003	-0.2422
dstate_regionW	1	0.5943	0.2929	4.1178	0.0424	0.1335
dfamily_history	1	1.5515	0.2339	43.9937	<.0001	0.4140
dwork_interfereN	1	3.1477	1.0430	9.1071	0.0025	0.6403
dwork_interfereO	1	7.5429	1.0703	49.6654	<.0001	1.3159
dwork_interfereR	1	6.1565	1.0328	35.5350	<.0001	1.2055
dwork_interfereS	1	6.7530	1.0232	43.5610	<.0001	1.7934
dbenefitsY	1	0.7460	0.2688	7.7044	0.0055	0.1991
dcare_optionsY	1	0.8527	0.2526	11.3910	0.0007	0.2235
dcoworkersY	1	0.9194	0.3154	8.4995	0.0036	0.1965
dcoworkersN	1	-0.5260	0.2524	4.3410	0.0372	-0.1182

Backward

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1341.280	591.377
SC	1346.154	674.238
-2 Log L	1339.280	557.377

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	781.9026	16	<.0001
Score	586.0808	16	<.0001
Wald	210.3031	16	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
29.2609	33	0.6539

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-7.5650	1.1987	39.8297	<.0001	
age	1	0.0550	0.0157	12.2755	0.0005	0.2208
dgenderM	1	-1.1719	0.3120	14.1040	0.0002	-0.2574
dgeo_regionEw	1	0.5955	0.2618	5.1730	0.0229	0.1499
dstate_regionW	1	0.8549	0.3163	7.3080	0.0069	0.1921
dfamily_history	1	1.6696	0.2429	46.6720	<.0001	0.4428
dwork_interfereN	1	3.1925	1.0477	9.2847	0.0023	0.6494
dwork_interfereO	1	8.0058	1.0884	54.1022	<.0001	1.3966
dwork_interfereR	1	6.3980	1.0417	37.7231	<.0001	1.2528
dwork_interfereS	1	7.0746	1.0348	46.7367	<.0001	1.8788
dbenefitsY	1	0.9725	0.3027	10.3233	0.0013	0.2595
dcare_optionsY	1	1.4941	0.3193	21.8940	<.0001	0.3916
dcare_optionsN	1	0.5965	0.2816	4.4877	0.0341	0.1608
dseek_helpY	1	-1.1962	0.3736	10.2534	0.0014	-0.2632
dseek_helpN	1	-1.0719	0.2899	13.6738	0.0002	-0.2956
dcoworkersY	1	0.8938	0.3204	7.7815	0.0053	0.1911
dcoworkersN	1	-0.6240	0.2585	5.8257	0.0158	-0.1402

Figure 9 – Final Model

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1341.280	606.354
SC	1346.154	669.719
-2 Log L	1339.280	580.354

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	758.9251	12	<.0001
Score	578.6857	12	<.0001
Wald	214.2439	12	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-7.1164	1.1504	38.2665	<.0001
age	1	0.0461	0.0151	9.3396	0.0022
dgenderM	1	-1.1030	0.3037	13.1879	0.0003
dstate_regionW	1	0.5943	0.2929	4.1178	0.0424
dfamily_history	1	1.5515	0.2339	43.9937	<.0001
dwork_interfereN	1	3.1477	1.0430	9.1071	0.0025
dwork_interfereO	1	7.5429	1.0703	49.6654	<.0001
dwork_interfereR	1	6.1565	1.0328	35.5350	<.0001
dwork_interfereS	1	6.7530	1.0232	43.5610	<.0001
dbenefitsY	1	0.7460	0.2688	7.7044	0.0055
dcare_optionsY	1	0.8527	0.2526	11.3910	0.0007
dcoworkersY	1	0.9194	0.3154	8.4995	0.0036
dcoworkersN	1	-0.5260	0.2524	4.3410	0.0372

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
age	1.047	1.017	1.079
dgenderM	0.332	0.183	0.602
dstate_regionW	1.812	1.020	3.217
dfamily_history	4.719	2.983	7.463
dwork_interfereN	23.282	3.014	179.829
dwork_interfereO	>999.999	231.634	>999.999
dwork_interfereR	471.777	62.321	>999.999
dwork_interfereS	856.645	115.312	>999.999
dbenefitsY	2.109	1.245	3.571
dcare_optionsY	2.346	1.430	3.849
dcoworkersY	2.508	1.352	4.653
dcoworkersN	0.591	0.360	0.969

Age → (1.047-1)100= 4.7%
dgenderM → (0.332 -1)100= -66.8%
dstate_regionW → (1.812 -1)100= 81.2%
dfamily_history → (4.719 -1)100= 371.9%
dwork_interfereN → (23.282-1)100= 2,228.2%
dwork_interfereO → (999.99 -1)100= 99,890
dwork_interfereR → (471.77-1)100= 47,077%
dwork_interferS →(856.62-1)100= 85,562%
dbenefitsY →(2.109-1)100= 110.9%
dcare_optionsY → (2.346-1)100= 134.6%
dcoworkersY →(2.50-1)100= 150%
dcoworkersN →(0.591-1)100= -40.9%

Figure 10 – Strongest Predictor

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-7.1164	1.1504	38.2665	<.0001	
age	1	0.0461	0.0151	9.3396	0.0022	0.1849
dgenderM	1	-1.1030	0.3037	13.1879	0.0003	-0.2422
dstate_regionW	1	0.5943	0.2929	4.1178	0.0424	0.1335
dfamily_history	1	1.5515	0.2339	43.9937	<.0001	0.4140
dwork_interfereN	1	3.1477	1.0430	9.1071	0.0025	0.6403
dwork_interfereO	1	7.5429	1.0703	49.6654	<.0001	1.3159
dwork_interfereR	1	6.1565	1.0328	35.5350	<.0001	1.2055
dwork_interfereS	1	6.7530	1.0232	43.5610	<.0001	1.7934
dbenefitsY	1	0.7460	0.2688	7.7044	0.0055	0.1991
dcare_optionsY	1	0.8527	0.2526	11.3910	0.0007	0.2235
dcoworkersY	1	0.9194	0.3154	8.4995	0.0036	0.1965
dcoworkersN	1	-0.5260	0.2524	4.3410	0.0372	-0.1182

Ranked by strength:

1. dwork_interfereS – 1.79
2. dwork_interfereO – 1.31
3. dwork_interfereR – 1.20
4. dwork_interfereN – 0.64
5. dfamily_history – 0.41
6. dgenderM – -0.24
7. dcare_optionsY – 0.22
8. dbenefitsY – 0.1991
9. dcoworkersY -0.1965
10. age – 0.18
11. dstate_regionW – 0.13
12. dcoworkersN – -0.11

Figure 11 – Classification Table

Classification Table

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.200	486	319	147	15	83.2	97.0	68.5	23.2	4.5
0.250	482	324	142	19	83.4	96.2	69.5	22.8	5.5
0.300	481	339	127	20	84.8	96.0	72.7	20.9	5.6
0.350	473	350	116	28	85.1	94.4	75.1	19.7	7.4
0.400	469	357	109	32	85.4	93.6	76.6	18.9	8.2
0.450	464	365	101	37	85.7	92.6	78.3	17.9	9.2
0.500	452	375	91	49	85.5	90.2	80.5	16.8	11.6
0.550	441	388	78	60	85.7	88.0	83.3	15.0	13.4
0.600	433	396	70	68	85.7	86.4	85.0	13.9	14.7

Figure 12 – Threshold .55

threshold .55

The FREQ Procedure

Frequency	Table of dtreatment by pred_y		
	pred_y		
dtreatment	0	1	Total
0	101	27	128
1	9	104	113
Total	110	131	241

		Predicted 1	Predicted 0
Actual	1	True Positive (TP) 104	False Negative (FN) 9
	0	False Positive (FP) 27	True Negative (TN) 101

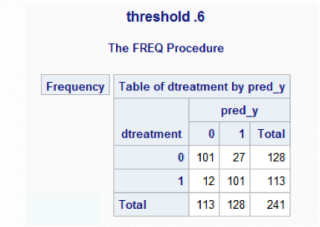
Sensitivity or Recall = TP/(TP+FN) = 104 / (104+9) = 104 / 113 = 92%
Proportion of correctly classified positives

Accuracy = (TP + TN) / (TP + TN + FP + FN) = (104+101) / (104 + 101 + 27 + 9) = 205 / 241 = 85%
Proportion of correctly classified positives and negatives

Precision = TP / (TP + FP) = 104 / (104 + 27) = 104 / 131 = 79.4%
Proportion of true positives among all predicted positives

Specificity = TN / (TN + FP)= 101 / (101 + 27) = 101 / 128 = 78.9%
Proportion of correctly classified negatives

Figure 13 – Threshold .6



threshold .6

The FREQ Procedure

Frequency	Table of dtreatment by pred_y		
	pred_y		
dtreatment	0	1	Total
0	101	27	128
1	12	101	113
Total	113	128	241

		Predicted	
		1	0
Actual	1	True Positive (TP) 101	False Negative (FN) 12
	0	False Positive (FP) 27	True Negative (TN) 101

Sensitivity or Recall = TP/(TP+FN) = 101 / (101+12) = 101 / 113 = 89.4%
Proportion of correctly classified positives

Accuracy = (TP + TN) / (TP + TN + FP + FN) = (101 + 101) / (101 + 101+ x27+ 12) = 202 / 241 **=** 83.8%
Proportion of correctly classified positives and negatives

Precision = TP / (TP + FP) = 101 / (101 + 27) = 101 / 128 = 78.9%
Proportion of true positives among all predicted positives

Specificity = TN / (TN + FP) = 101 / (101 + 27) = 101 / 128= 78.9%
Proportion of correctly classified negative

Figure 14 – Predictions

#1 A 30 year old female from the midwest with family history of mental illness, she states that the mental illness sometimes interferes with work but that her company provides mental health benefits but does not know the mental health care options and she doesn't feel comfortable speaking with her coworkers.

#2 A 28 year old male from California with family history of mental illness, states that the mental condition sometimes interferes with work but his company does provide mental health benefits but does not know the care options provided by his employer and is not willing to speak to his coworkers.

age= 30	Age = 28
dgenderM= 0	dgenderM = 1
dstate _regionW= 0	dstate_regionW = 1
dfamily_history= 1	dfamily_history = 1
dwork_interfereN= 0	dwork_interfereN= 0
dwork_interfereO= 0	dwork_interfereO= 0
dwork_interfereR=0	dwork_interfereR=0
dwork_interfereS= 1	dwork_interfereS= 1
dbenefitsY = 1	dbenefitsY = 1
dcare_optionsY= 0	dcare_optionsY= 0
dcoworkersY= 0	dcoworkersY= 0
dcoworkersN= 1	dcoworkersN= 1

phat	lcl	ucl	pred_y	threshold
0.97027	0.94009	0.98548	1	0.55
0.79417	0.65802	0.88554	1	0.55

Predictions

Obs	age	dgenderM	dstate_regionW	dfamily_history	dwork_interfereN	dwork_interfereO	dwork_interfereR	dwork_interfereS	dbenefitsY	dcare_optionsY	dcoworkersY	dcoworkersN
1	30	0	.	1	0	0	0	1	1	1	0	1
2	28	1	.	0	0	0	0	1	1	1	0	1
3	29	1	0	1	0	0	0	1	1	1	0	0

Prediction #1 Predicted probability= 0.9702 with a 95% of the time, the predicted probability will fall within 0.94009, 0.98548

- The corresponding 95% confidence limits for the odds ration are ([exp(0.94009)-1]*100) 156.02%, ([exp(0.98548)-1]*100) 167.90%

Prediction #2 Predicted probability= 0.7941 with a 95% of the time, the predicted probability will fall within 0.65802, 0.88554

- The corresponding 95% confidence limits for the odds ration are ([exp(0.65802)-1]*100) 93.09%, ([exp(0.88554)-1]*100) 142.43%

References

1. Kaggle.com. 2018. Survey on Mental Health in the Tech Workplace in 2014. https://www.kaggle.com/osmi/mental-health-in-tech-survey. Accessed 23 February 2018.

2. Monro, S. 2005. Beyond Male and Female: Poststructuralism and the spectrum of gender. *International Journal of Transgenderism* 8(1), 3-22

3. United Nations Department of Economic and Social Affairs Statistics Division. 2018. Standard Country or Area Codes for Statistical Use. https://unstats.un.org/unsd/methodology/m49/. Accessed 2 March 2018.

4. United States Census Bureau. 2013. Census Bureau Regions and Divisions with State FIPS Codes. https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf. Accessed 2 March 2018.

Harsh Hareshkumar Shukla

Introduction

Goal is the same as the group, refer to the introduction section in the main report for details.

Methodology

Data Collection: We have enriched our data set from Kaggle which can be obtained from the following link: " https://www.kaggle.com/osmi/mental-health-in-tech-survey "[1].

Data Preprocessed: As we did exploratory analysis to begin with we constructed age vs Treatment boxplot diagram to see if the age is an affecting factor in mental health. After doing that we realized there are certain noise in data which resulted the treatment boxplot in skewed output so initially we removed noise from the data and repeated the process to learn that it did not change much depending upon the age. On second exploratory stage we tried to figure out if the geographical location affects it in any order with bar graph. And we also tested a frequency report of each variables to gain much detailed domain knowledge.
Variable creation: As we learned more about the data we realized we have to create dummy variables for most parts since, almost all the variables in the dataset are categorical than computational.
Model Creation: we have divided the data into training and testing sets. Based upon the training set we have applied 3 different model selection method that are "Forward selection", "Backward selection", and "Stepwise selection". Using those methods, we are getting two different models among which forward and stepwise are yielding similar selection models while the backward is giving slightly different model. So, after that point we ran two different models and tested both the models on testing sets and calculated accuracy for both the models and based upon that we have selected our final model.
Other Approach (Out of class): The logistic regression is yielding efficient results, but to yield an advance result I have applied binary decision tree for which I have split the data using K-fold cross validation and then ran multiple models. Based on it I finalized a rule model depending on numerous factors.
Validation Method: For the created models we will also calculate specificity, Accuracy, Sensitivity and Precision.

Analysis, Results and Findings

Exploratory analysis: As we learned from our analysis, treatment will be our dependent variable and all of the variables in our dataset are categorical except age, so we ran a frequency report to check the overall number of cases in each attribute. Based upon the frequency reports we have learned the data had lot of noise in Gender variable and the dataset reflects that it was a manual entry so it had typing errors as well, so we derived all the different classes and based on it we divided them into 3 categories by creating dummy variables "Male", "Female" and "others". Once we were done with that procedure we decided to divide all the variables into dummy variables to get more accurate results. **We had two location based variables Country and States** For Country we have divided it into 11 different parts and for the states we took "NA" as base variable and divided the states of US into 9 different parts which are mentioned in **figure 1** .We have also run two bar graphs to see if the geographical location effects on the part affects on mental health, but as it can be observed from **Figure 5** we have almost equal number of observations for both yes and no categories from different countries so it can be possible that the results that were collected were mainly collected from few major countries so we cannot conclude that geographical location will affect in that part. For age we must consider only the observations which were from 5 to 100. We have removed rest of the observations to get optimal results then we have divided it into 5 distinct categories from childhood to mature adulthood. Our **Figure 2 and figure 3** represents that age is not as much affective in terms of mental health. Number of employees were divided into 5 different dummy variables depending upon the size of

their employer's employee strength. For the rest of the variables we have taken "NA" or "NO" as base line and created dummy variables to convert the variable range into [0,1].

Dimensionality Reduction: As we learned from our frequency table, Timestamp and Comment variables were not useful at all so we have not considered them in further model creations.

Collinearity: since all our variables were dummy variables including dependent variable so checking collinearity would not be affective at all.

Correlation: even though all of our variables were dummy variables we have taken Correlational report so that we can see which have a Pearson correlation coefficient to determine certain variables for interaction variable and model creation. Figure 4 represents it's model for yes_treatment vs all the other variables. Derived from figure4.

Interaction Variable and Interaction Model:

We have used following attributes as interaction variables female often_work_interfere rarely_work_interfere sometimes_work_interfere yes_obs_consequence Very_difficult_leave

From those points we have retrieved an Interaction model which is represented in FIG and from which we can determine certain variables that are:

Data split: once we have completed interaction model we decided to split the data using 'hold out partitioning technique' using 66% of the data into model creation while remaining 34% of the data for testing the accuracy, Sensitivity, Specificity of the model. I have learned from an article[6] which stated that why to use 66 and 34% ratio is better then other ratios yielding various model research and resulting that the accuracy difference between training and testing model is usually least in this ratio but ideally if you have more than 5000 observations then this ratio is considered much more advisable.

Model Selection: for model selection part we have used three diverse selection methods which are Forward, Backward and stepwise. Out of which two are giving us equivalent results which are yielding two different models as follows:

Model 1: The overall likelihood ratio obtained from model one was 536.28 and <.001 p-value which yields to reject null hypothesis and accept alternate hypothesis.

For model 1 we are using the attributes that are listed by forward and stepwise are as follows :

male yes_family_history never_work_interfere often_work_interfere rarely_work_interfere sometimes_work_interfere no_employees_3 yes_care_options Somewhat_easy_leave

We have taken 0.55 as threshold (based from figure 8) for misclassification on testing model and we ran testing on remaining 34% of the data which yields 82.39% accuracy

Model 2 : The overall likelihood ratio obtained from model one was 536.27 and <.001 p-value which yields to reject null hypothesis and accept alternate hypothesis

For Model 2 we are using the attributes that are listed by Backward selection which are as follows:

Female yes_family_history never_work_interfere often_work_interfere rarely_work_interfere sometimes_work_interfere no_employees_3 yes_care_options Somewhat_easy_leave

We have taken 0.5 as threshold (Based from Figure 9) value for misclassification on testing model and ran testing on remaining 34% of the data which yields 82.97% accuracy

Influential points: Based upon accuracy we have decided to go with the 2nd model, but to make it more accurate we have tried to see influential points in the model. And we have found 9 influential and 1 possible outlier in data observations which are as follows:
1250, 1204, 1178, 983, 923, 822, 652, 570, 379, 218
I have deleted all the influential models and yielded the final logistic regression model.
Residual plots: as we learned in exploratory analysis all our attributes are dummy variables and for that they will not have a normal distribution as on the residual plot will not yield normality can be observed in figure 16.

Final Model:

- **The final model yields that female gender who has a family history for mental health illness who does suffers any sort of work interface and does have care options are more likely to have the mental health treatment requirements.**

Attributes for final model: `Female yes_family_history never_work_interfere often_work_interfere rarely_work_interfere sometimes_work_interfere no_employees_3 yes_care_options Somewhat_easy_leave`

Final Model Equation: From Figure 15 we can derive following equation
Log ((Yes_treatment = 1)/(Yes_treatment = 0)) = -4.8473 + 0.8334*(Female) +1.0955*(Yes_Family_history)+2.5505*(never_work_interfere)+5.7053*(often_work_interfere)+4.5316*(Rarely_work_interfere)+4.9417*(Sometimes_work_interfere) +0.9501*(Yes_care_options)

Sensitivity	Accuracy	Precision	Specificity
77.82%	82.39%	93.86%	90.73%

Summarization for each predictors

Variable		
Female	(2.301-1)*100	130.1%
Yes_Family_history	(2.991-1)*100	199.1%
Never_work_interfere	(12.814-1)*100	1181.4%
often_work_interfere	(300.454-1)*100	29945.4%
rarely_work_interfere	(92.904-1)*100	9190.4%
Sometimes_work_interfere	(140.013-1)*100	13901.3%
Yes_care_options	(2.586-1)*100	158.6%

Testing:

Female	yes_family_history	never_work_interfere	often_work_interfere	rarely_work_interfere	sometimes_work_interfere	yes_care_options
0	1	0	1	1	0	1
1	0	1	1	0	1	1

- Figure 17 represents that the phat or predicted value for both the model will be around 1 based upon phat value which lies between ucl and lcl.

Different approach: Once, I have done the Regression approach I decided to apply more approach to data set to see if we can generate more optimal model which help [Oded Z, M., 2014]. For that purpose, I decided to go with a supervised learning technique which is named as binary decision tree. For that I have used K fold cross validation for split and testing where k will be 10 so, I have used following parent and child nodes. Out of which I have selected 80 and 40 which has 84% accuracy and 11 as complexity. Figure 17 represents the decision tree while figure 18 represents the misclassification matrix for selected variable. Rules for decision tree are mentioned in appendix section followed by importance table

Parent Node	Child Node	Complexity/ Terminal node	Total Node	Depth	Accuracy
100	50	10	19	4	83.1%
80	**40**	**11**	**21**	**5**	**84%**
60	30	13	25	6	84%
36	18	13	25	6	84%

Differently from Regression model, Binary Tree is a classification technique so rather than model creation we will have rules that will define it. But both of them does not requires to store the data.

Findings: once, we have come across all the models we have learned that almost all 3 of them provide equal accuracy but out of all the finalized logistic, regression makes most sense which is much more faster and accurate than decision tree.

References

1. Mental Health in Tech Survey | Kaggle. 2018. *Mental Health in Tech Survey | Kaggle*. https://www.kaggle.com/osmi/mental-health-in-tech-survey. Accessed 12 March 2018.

2. National Center for Biotechnology Information. 2018. *National Center for Biotechnology Information*. https://www.ncbi.nlm.nih.gov. Accessed 10 March 2018.

3. Oded Z, M., 2014. *Data Mining With Decision Trees: Theory And Applications*. 2nd ed. World Scientific: World Scientific.

4. The Official Blog of BigML.com. 2018. *Logistic Regression versus Decision Trees | The Official Blog of BigML.com*. https://blog.bigml.com/2016/09/28/logistic-regression-versus-decision-trees/. Accessed 14 March 2018.

5. census. 2000. *U.S. Census Bureau Regions and Divisions*. https://www.census.gov/geo/www/us_regdiv.pdf. Accessed 28 February 2018.

6. SpringerLink. 2018. *Investigating the influence of data splitting on the predictive ability of QSAR/QSPR models | SpringerLink*. https://link.springer.com/article/10.1007/s11224-011-9757-4. Accessed 8 March 2018.

Appendix

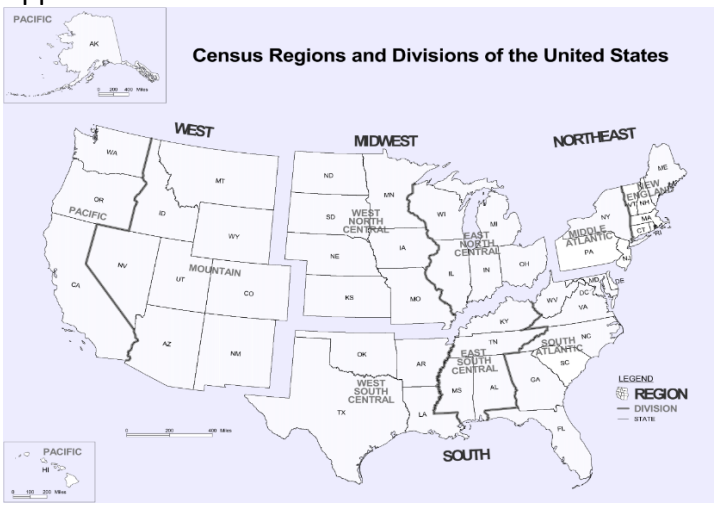


Figure 1 Reference for dummy variable creation for State variable reference[5]

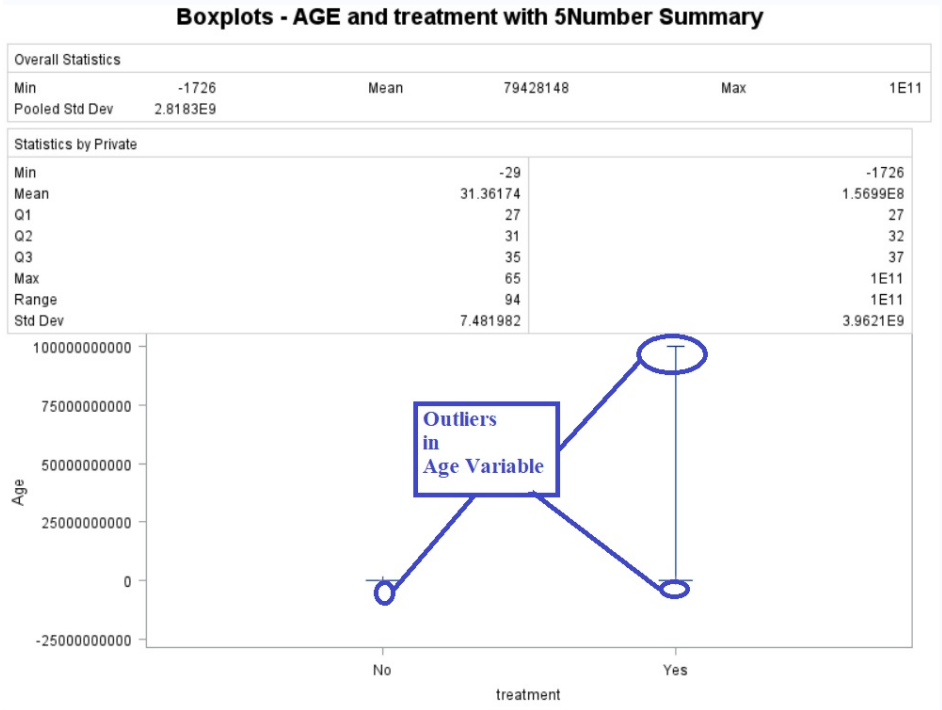


Figure 2 Exploratory Analysis for Age variable to see if the age affects in mental health

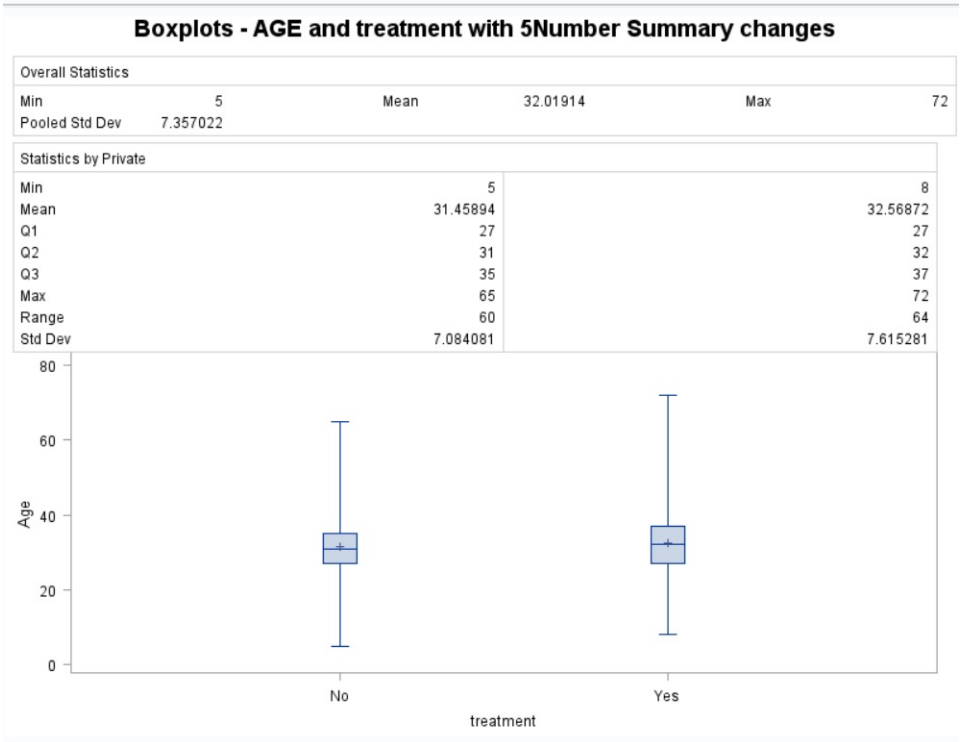


Figure 3 boxplot of age vs treatment after outlier removal from age variable

Figure 4 correlation matrix

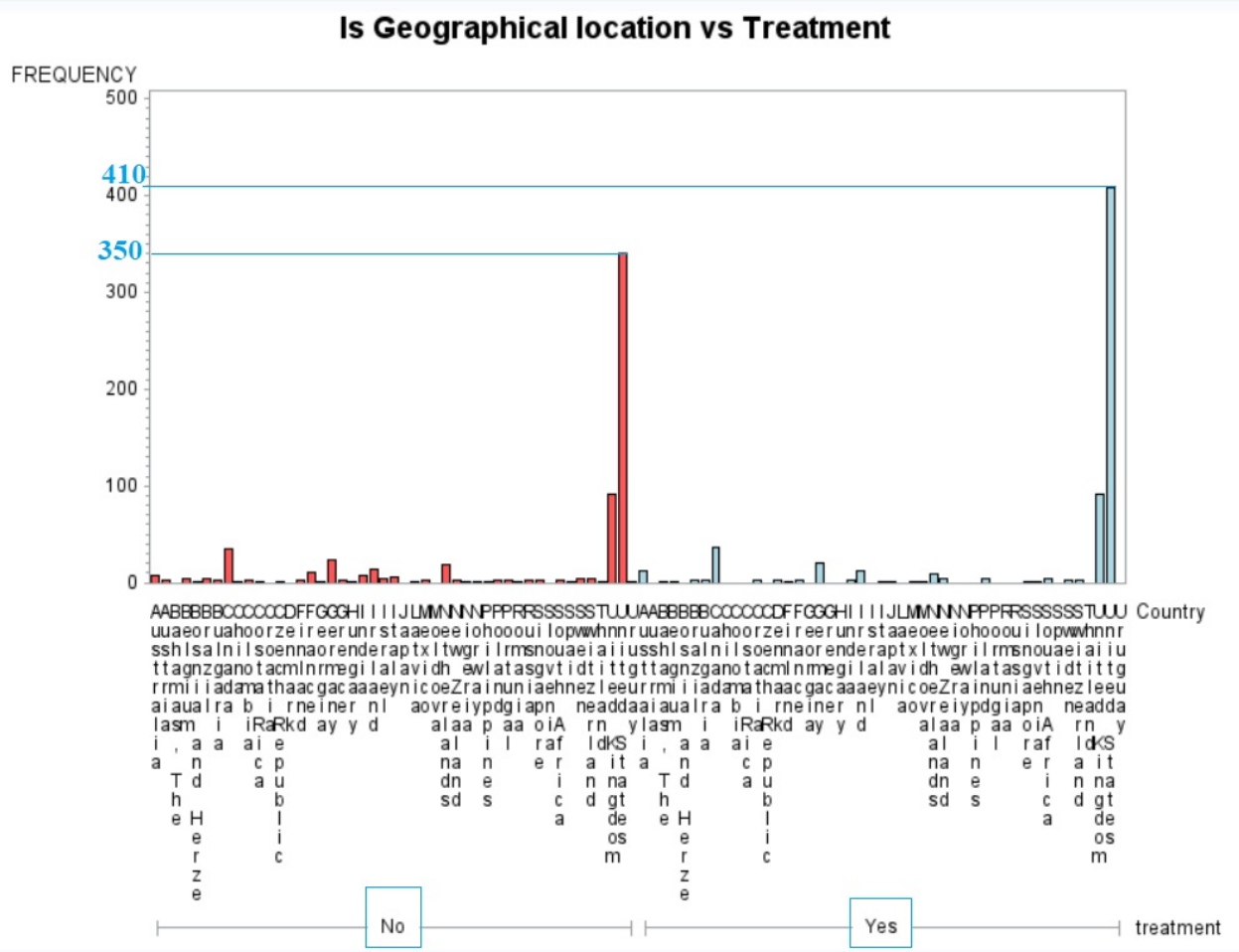


Figure 5 geographical location survey exploratory analysis

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.05425	0.01859	2.92	0.0036
Female	1	0.07434	0.05021	1.48	0.1390
often_work_interfere	1	0.74705	0.04436	16.84	<.0001
rarely_work_interfere	1	0.59300	0.03989	14.87	<.0001
sometimes_work_interfere	1	0.68499	0.02844	24.09	<.0001
yes_obs_consequence	1	0.07876	0.06546	1.20	0.2291
Very_difficult_leave	1	0.10778	0.10524	1.02	0.3060
int_var1	1	0.05487	0.09089	0.60	0.5462
int_var2	1	0.17435	0.08096	2.15	0.0315
int_var3	1	0.03033	0.06489	0.47	0.6403
int_var4	1	-0.02573	0.10021	-0.26	0.7974
int_var5	1	0.04755	0.08836	0.54	0.5906
int_var6	1	-0.07093	0.13445	-0.53	0.5979
int_var7	0	0	.	.	.
int_var8	1	-0.13825	0.10220	-1.35	0.1764
int_var9	1	-0.03447	0.15288	-0.23	0.8216
int_var10	1	-0.01836	0.08008	-0.23	0.8187
int_var11	1	-0.15230	0.11784	-1.29	0.1964
int_var12	1	-0.19978	0.10020	-1.99	0.0464
int_var13	0	0	.	.	.
int_var14	1	0.01741	0.07215	0.24	0.8094
int_var15	0	0	.	.	.

Figure 6 interractive variable selection for interective model

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.07143	0.01670	4.28	<.0001
often_work_interfere	1	0.77964	0.03493	22.32	<.0001
rarely_work_interfere	1	0.57198	0.03616	15.82	<.0001
sometimes_work_interfere	1	0.69797	0.02377	29.36	<.0001
int_var2	1	0.24295	0.06361	3.82	0.0001

Figure 7 Final Interaction Model

Step Wise Selection Model

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.6410	0.7577	37.5138	<.0001
Male	1	-0.5500	0.2617	4.4161	0.0356
yes_family_history	1	0.8719	0.2120	16.9157	<.0001
never_work_interfere	1	2.8513	0.7563	14.2139	0.0002
often_work_interfere	1	6.1995	0.7789	63.3442	<.0001
rarely_work_interfer	1	5.5411	0.7580	53.4334	<.0001
sometimes_work_inter	1	5.4613	0.7326	55.5664	<.0001
no_employees_3	1	1.1996	0.5579	4.6245	0.0315
yes_care_options	1	1.0203	0.2233	20.8835	<.0001
Somewhat_easy_leave	1	-0.4933	0.2463	4.0097	0.0452

Forward Selection Model

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.6410	0.7577	37.5138	<.0001
Male	1	-0.5500	0.2617	4.4161	0.0356
yes_family_history	1	0.8719	0.2120	16.9157	<.0001
never_work_interfere	1	2.8513	0.7563	14.2139	0.0002
often_work_interfere	1	6.1995	0.7789	63.3442	<.0001
rarely_work_interfer	1	5.5411	0.7580	53.4334	<.0001
sometimes_work_inter	1	5.4613	0.7326	55.5664	<.0001
no_employees_3	1	1.1996	0.5579	4.6245	0.0315
yes_care_options	1	1.0203	0.2233	20.8835	<.0001
Somewhat_easy_leave	1	-0.4933	0.2463	4.0097	0.0452

Figure 8 Forward and Stepwise Selection Model

Backward Selection Model

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.1938	0.7303	50.5832	<.0001
Female	1	0.5686	0.2715	4.3843	0.0363
yes_family_history	1	0.8776	0.2118	17.1683	<.0001
never_work_interfere	1	2.8626	0.7564	14.3244	0.0002
often_work_interfere	1	6.2026	0.7790	63.3938	<.0001
rarely_work_interfer	1	5.5482	0.7582	53.5466	<.0001
sometimes_work_inter	1	5.4660	0.7327	55.6452	<.0001
no_employees_3	1	1.1856	0.5582	4.5109	0.0337
yes_care_options	1	1.0253	0.2233	21.0901	<.0001
Somewhat_easy_leave	1	-0.4946	0.2465	4.0259	0.0448

Figure 9 Backward Selection Model

Classification Table										
Prob Level	Correct		Incorrect		Percentages					Sensitivity + Specificity
	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	FALSE POS	FALSE NEG	
0.1	397	265	158	8	80	98	62.6	28.5	2.9	160.6
0.15	394	275	148	11	80.8	97.3	65	27.3	3.8	162.3
0.2	388	280	143	17	80.7	95.8	66.2	26.9	5.7	162
0.25	386	297	126	19	82.5	95.3	70.2	24.6	6	165.5
0.3	386	301	122	19	83	95.3	71.2	24	5.9	166.5
0.35	385	304	119	20	83.2	95.1	71.9	23.6	6.2	167
0.4	383	306	117	22	83.2	94.6	72.3	23.4	6.7	166.9
0.45	376	307	116	29	82.5	92.8	72.6	23.6	8.6	165.4
0.5	376	317	106	29	83.7	92.8	74.9	22	8.4	167.7
0.55	375	317	106	30	83.6	92.6	74.9	22	8.6	167.5
0.6	333	352	71	72	82.7	82.2	83.2	17.6	17	165.4
0.65	327	355	68	78	82.4	80.7	83.9	17.2	18	164.6
0.7	305	361	62	100	80.4	75.3	85.3	16.9	21.7	160.6
0.75	282	376	47	123	79.5	69.6	88.9	14.3	24.6	158.5
0.8	188	397	26	217	70.7	46.4	93.9	12.1	35.3	140.3

Figure 10 Threshold for misclassification model selection in Model 2

Classification Table										Sensitivity + Specificity
Prob Level	Correct		Incorrect		Percentages					
	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	FALSE POS	FALSE NEG	
0.1	397	265	158	8	80	98	62.6	28.5	2.9	160.6
0.15	394	274	149	11	80.7	97.3	64.8	27.4	3.9	162.1
0.2	386	279	144	19	80.3	95.3	66	27.2	6.4	161.3
0.25	386	297	126	19	82.5	95.3	70.2	24.6	6	165.5
0.3	386	303	120	19	83.2	95.3	71.6	23.7	5.9	166.9
0.35	385	304	119	20	83.2	95.1	71.9	23.6	6.2	167
0.4	383	306	117	22	83.2	94.6	72.3	23.4	6.7	166.9
0.45	376	306	117	29	82.4	92.8	72.3	23.7	8.7	165.1
0.5	376	316	107	29	83.6	92.8	74.7	22.2	8.4	167.5
0.55	375	319	104	30	83.8	92.6	75.4	21.7	8.6	168
0.6	335	354	69	70	83.2	82.7	83.7	17.1	16.5	166.4
0.65	329	355	68	76	82.6	81.2	83.9	17.1	17.6	165.1
0.7	298	361	62	107	79.6	73.6	85.3	17.2	22.9	158.9
0.75	283	376	47	122	79.6	69.9	88.9	14.2	24.5	158.8
0.8	189	397	26	216	70.8	46.7	93.9	12.1	35.2	140.6

Figure 11 Threshold for misclassification model selection in Model 1

Confusion Matrix for Model 1

The FREQ Procedure

Frequency

ACCuracy = (137+214)/426 =82.39%

Table of yes_treatment by pred_y

yes_treatment	pred_y 0	pred_y 1	Total
0	137 (TN)	61	198
1	14	214 (TP)	228
Total	151	275	426

Figure 12 Accuracy for Model 1

COnfusion Matrix for model 2

The FREQ Procedure

Frequency

Accuracy =(137+214)/426 =82.97%

Table of yes_treatment by pred_y

yes_treatment	pred_y 0	pred_y 1	Total
0	137	61	198
1	14	214	228
Total	151	275	426

Figure 13 Accuracy for Model 2

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.8852	0.6131	63.4917	<.0001
Female	1	0.7871	0.2677	8.6423	0.0033
yes_family_history	1	1.0972	0.2087	27.6313	<.0001
never_work_interfere	1	2.5728	0.6349	16.4235	<.0001
often_work_interfere	1	5.8168	0.6802	73.1345	<.0001
rarely_work_interfer	1	4.5806	0.6345	52.1198	<.0001
sometimes_work_inter	1	5.0181	0.6111	67.4387	<.0001
no_employees_3	1	0.7751	0.5422	2.0439	0.1528
yes_care_options	1	0.9869	0.2151	21.0427	<.0001
Somewhat_easy_leave	1	-0.2772	0.2451	1.2786	0.2582

Figure 14 Regression model after removal of Influential points

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.8473	0.6036	64.4950	<.0001
Female	1	0.8334	0.2671	9.7358	0.0018
yes_family_history	1	1.0955	0.2082	27.6974	<.0001
never_work_interfere	1	2.5505	0.6325	16.2586	<.0001
often_work_interfere	1	5.7053	0.6727	71.9262	<.0001
rarely_work_interfer	1	4.5316	0.6291	51.8871	<.0001
sometimes_work_inter	1	4.9417	0.6045	66.8188	<.0001
yes_care_options	1	0.9501	0.2128	19.9364	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Female	2.301	1.363	3.884
yes_family_history	2.991	1.989	4.498
never_work_interfere	12.814	3.709	44.269
often_work_interfere	300.454	80.382	>999.999
rarely_work_interfer	92.904	27.073	318.803
sometimes_work_inter	140.013	42.813	457.888
yes_care_options	2.586	1.704	3.924

Figure 15 Final Regression Model

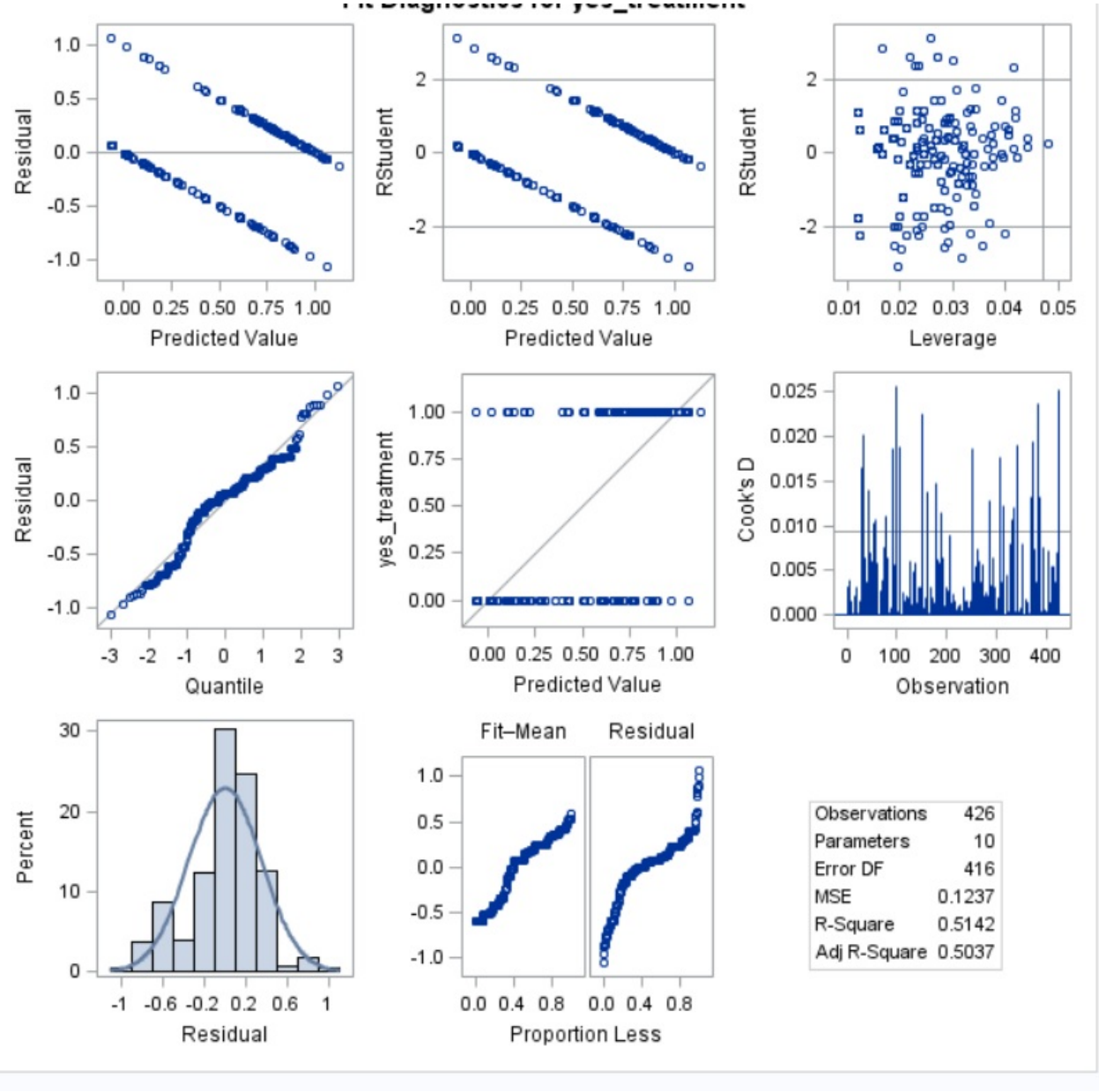


Figure 16 Residual Plots

phat	lcl	ucl
0.99970	0.99773	0.99996
0.99998	0.99938	1.00000

Figure 17 Phat and ucl and lcl values for tested datalines

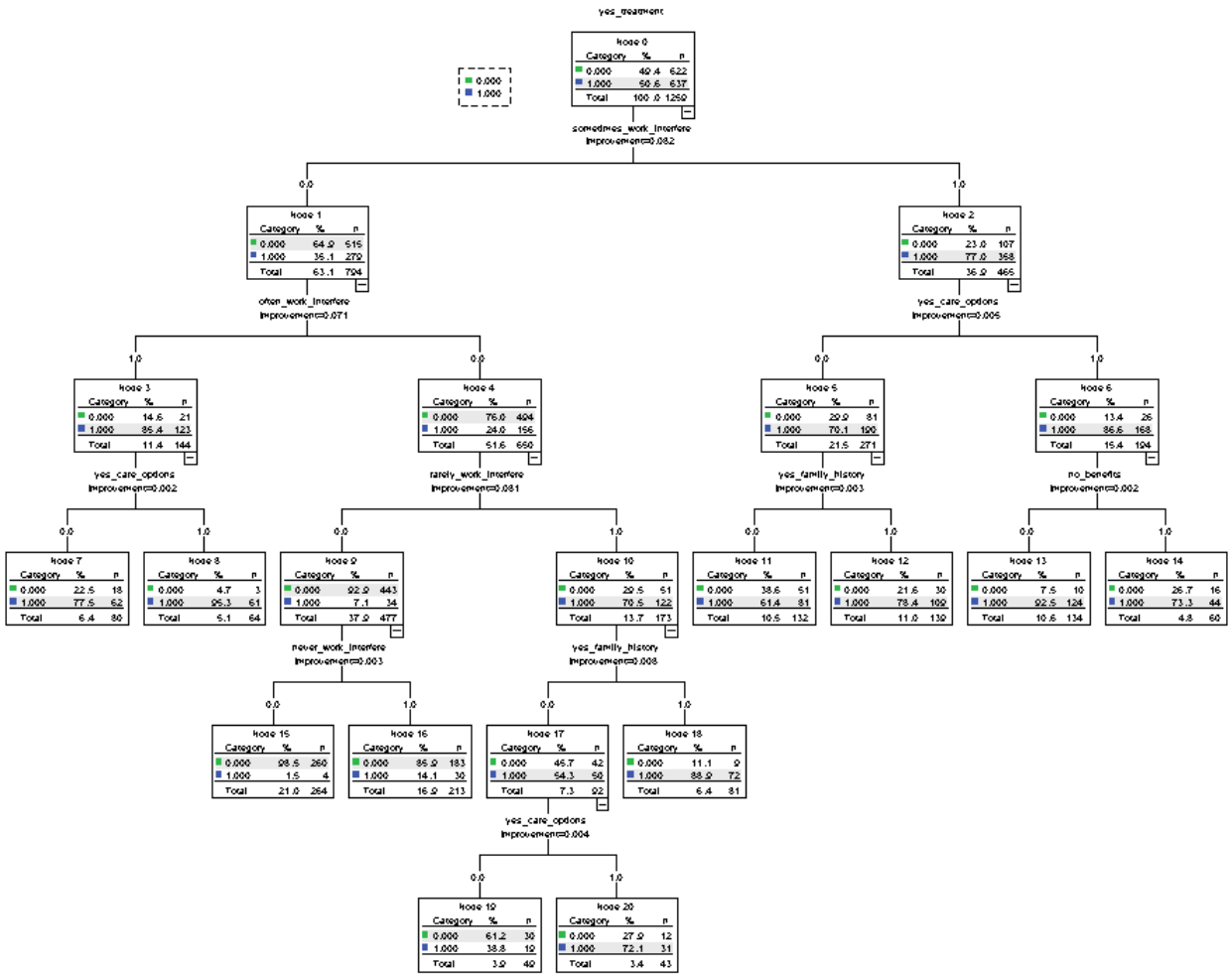


Figure 18 Tree model presentation

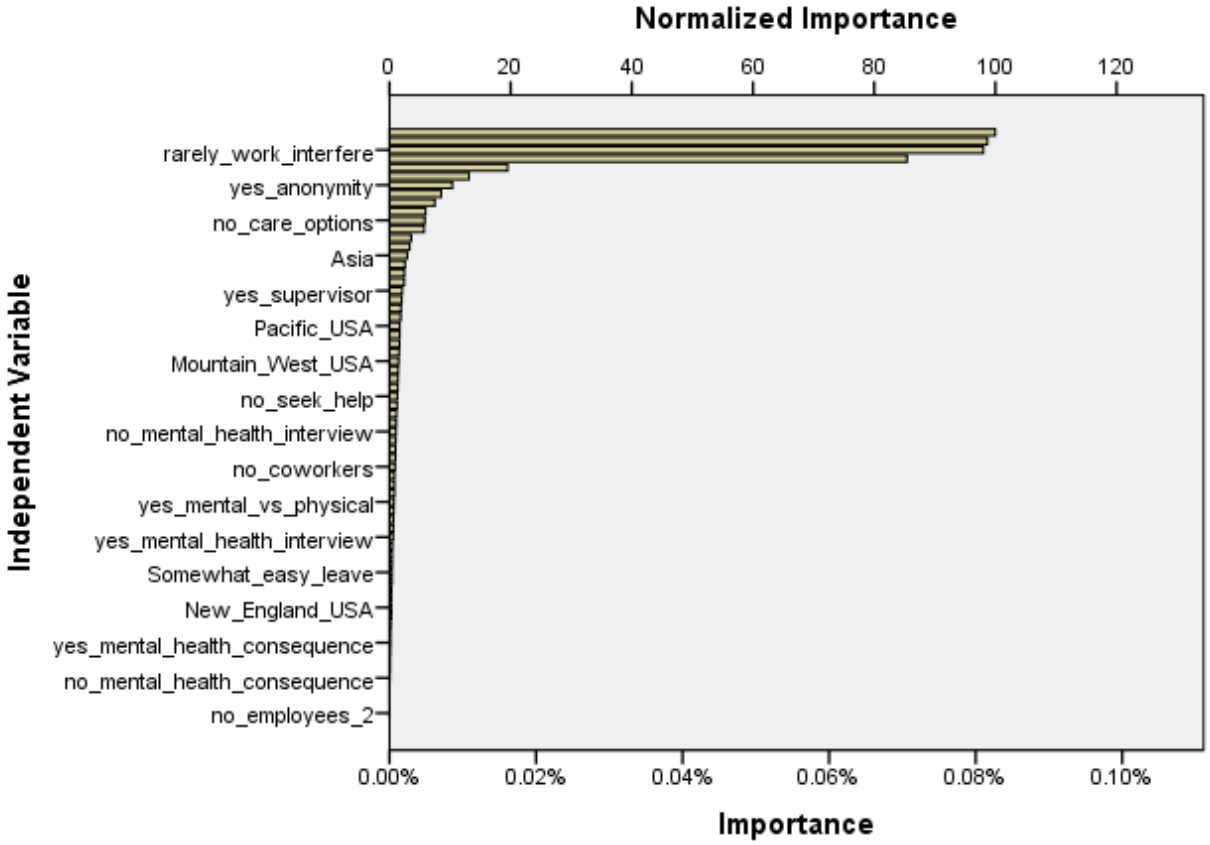
Classification

	Predicted		
Observed	0	1	Percent Correct
0	473	149	76.0%
1	53	584	91.7%
Overall Percentage	41.8%	58.2%	84.0%

Growing Method: CRT

Dependent Variable: yes_treatment

Figure 19 Misclassificcation matrix



⇨ Rules for The Decision tree

```
/* Node 7 */.
IF (((sometimes_work_interfere = 0) OR (sometimes_work_interfere != 1)  AND
((yes_family_history = 0) OR (yes_family_history != 1)  AND
(Very_difficult_leave != 1))))  AND  (((often_work_interfere = 1) OR
(often_work_interfere != 0)  AND  (Africa = 1)))  AND  (((yes_care_options =
0) OR (yes_care_options != 1)  AND  ((no_care_options = 1) OR
(no_care_options != 0)  AND  ((yes_anonymity = 0) OR (yes_anonymity != 1)
AND  ((yes_benefits = 0) OR (yes_benefits != 1)  AND  ((yes_wellness_program
= 0) OR (yes_wellness_program != 1)  AND  ((East_North_Centeral_USA = 0) OR
(East_North_Centeral_USA != 1)  AND  ((yes_seek_help = 0) OR (yes_seek_help
!= 1)  AND  ((no_wellness_program = 1) OR (no_wellness_program != 0)  AND
((Very_easy_leave = 0) OR (Very_easy_leave != 1)  AND  ((no_employees_1 = 0)
OR (no_employees_1 != 1)  AND  ((no_employees_4 = 1) OR (no_employees_4 != 0)
AND  ((yes_obs_consequence = 0) OR (yes_obs_consequence != 1)  AND
((yes_supervisor = 0) OR (yes_supervisor != 1)  AND  ((Middle_Atlantic_USA =
0) OR (Middle_Atlantic_USA != 1)  AND  ((New_England_USA = 0) OR
(New_England_USA != 1)  AND  ((yes_phys_health_consequence = 0) OR
(yes_phys_health_consequence != 1)  AND  ((yes_coworkers = 0) OR
(yes_coworkers != 1)  AND  ((not_self_employed = 1) OR (not_self_employed !=
0)  AND  ((yes_self_employed = 0) OR (yes_self_employed != 1)  AND
((yes_mental_health_interview = 0) OR (yes_mental_health_interview != 1)  AND
((Other = 0) OR (Other != 1)  AND  ((Africa = 0) OR (Africa != 1)  AND
```

```
((Very_difficult_leave = 0) OR (Very_difficult_leave != 1)  AND  ((Male = 1)
OR (Male != 0)  AND  ((Australia_and_New_Zealand = 0) OR
(Australia_and_New_Zealand != 1)  AND  ((Somewhat_easy_leave = 0) OR
(Somewhat_easy_leave != 1)  AND  ((Mountain_West_USA = 0) OR
(Mountain_West_USA != 1)  AND  ((Southeast_Europe = 0) OR (Southeast_Europe
!= 1)  AND  ((childhood = 0) OR (childhood != 1)  AND  (yes_tech_company !=
0))))))))))))))))))))))))))))))
```

THEN

Node = 7

Prediction = 1

Probability = 0.775000

```
/* Node 8 */.
IF (((sometimes_work_interfere = 0) OR (sometimes_work_interfere != 1)  AND
((yes_family_history = 0) OR (yes_family_history != 1)  AND
(Very_difficult_leave != 1))))  AND  (((often_work_interfere = 1) OR
(often_work_interfere != 0)  AND  (Africa = 1)))  AND  (((yes_care_options =
1) OR (yes_care_options != 0)  AND  ((no_care_options = 0) OR
(no_care_options != 1)  AND  ((yes_anonymity = 1) OR (yes_anonymity != 0)
AND  ((yes_benefits = 1) OR (yes_benefits != 0)  AND  ((yes_wellness_program
= 1) OR (yes_wellness_program != 0)  AND  ((East_North_Centeral_USA = 1) OR
(East_North_Centeral_USA != 0)  AND  ((yes_seek_help = 1) OR (yes_seek_help
!= 0)  AND  ((no_wellness_program = 0) OR (no_wellness_program != 1)  AND
((Very_easy_leave = 1) OR (Very_easy_leave != 0)  AND  ((no_employees_1 = 1)
OR (no_employees_1 != 0)  AND  ((no_employees_4 = 0) OR (no_employees_4 != 1)
AND  ((yes_obs_consequence = 1) OR (yes_obs_consequence != 0)  AND
((yes_supervisor = 1) OR (yes_supervisor != 0)  AND  ((Middle_Atlantic_USA =
1) OR (Middle_Atlantic_USA != 0)  AND  ((New_England_USA = 1) OR
(New_England_USA != 0)  AND  ((yes_phys_health_consequence = 1) OR
(yes_phys_health_consequence != 0)  AND  ((yes_coworkers = 1) OR
(yes_coworkers != 0)  AND  ((not_self_employed = 0) OR (not_self_employed !=
1)  AND  ((yes_self_employed = 1) OR (yes_self_employed != 0)  AND
((yes_mental_health_interview = 1) OR (yes_mental_health_interview != 0)  AND
((Other = 1) OR (Other != 0)  AND  ((Africa = 1) OR (Africa != 0)  AND
((Very_difficult_leave = 1) OR (Very_difficult_leave != 0)  AND  ((Male = 0)
OR (Male != 1)  AND  ((Australia_and_New_Zealand = 1) OR
(Australia_and_New_Zealand != 0)  AND  ((Somewhat_easy_leave = 1) OR
(Somewhat_easy_leave != 0)  AND  ((Mountain_West_USA = 1) OR
(Mountain_West_USA != 0)  AND  ((Southeast_Europe = 1) OR (Southeast_Europe
!= 0)  AND  ((childhood = 1) OR (childhood != 0)  AND  (yes_tech_company =
0))))))))))))))))))))))))))))))
```

THEN

Node = 8

Prediction = 1

Probability = 0.953125

```
/* Node 15 */.
IF (((sometimes_work_interfere = 0) OR (sometimes_work_interfere != 1)  AND
((yes_family_history = 0) OR (yes_family_history != 1)  AND
(Very_difficult_leave != 1))))  AND  (((often_work_interfere = 0) OR
(often_work_interfere != 1)  AND  (Africa != 1)))  AND
(((rarely_work_interfere = 0) OR (rarely_work_interfere != 1)  AND  (Other !=
1)))  AND  (((never_work_interfere = 0) OR (never_work_interfere != 1)  AND
((yes_phys_health_consequence = 0) OR (yes_phys_health_consequence != 1)  AND
```

```
((not_self_employed = 1) OR (not_self_employed != 0)  AND  ((Northern_America
= 0) OR (Northern_America != 1)  AND  ((yes_mental_health_consequence = 0) OR
(yes_mental_health_consequence != 1)  AND  ((mature_adulthood = 0) OR
(mature_adulthood != 1)  AND  ((yes_benefits = 0) OR (yes_benefits != 1)  AND
((yes_wellness_program = 0) OR (yes_wellness_program != 1)  AND
((yes_seek_help = 0) OR (yes_seek_help != 1)  AND  ((yes_self_employed = 0)
OR (yes_self_employed != 1)  AND  ((Australia_and_New_Zealand = 0) OR
(Australia_and_New_Zealand != 1)  AND  ((yes_care_options = 0) OR
(yes_care_options != 1)  AND  ((South_Atlantic_USA = 0) OR
(South_Atlantic_USA != 1)  AND  ((West_North_Centeral_USA = 0) OR
(West_North_Centeral_USA != 1)  AND  (Somewhat_difficult_leave !=
1))))))))))))))))
```

THEN

Node = 15

Prediction = 0

Probability = 0.984848

```
/* Node 16 */.
IF (((sometimes_work_interfere = 0) OR (sometimes_work_interfere != 1)  AND
((yes_family_history = 0) OR (yes_family_history != 1)  AND
(Very_difficult_leave != 1))))  AND  (((often_work_interfere = 0) OR
(often_work_interfere != 1)  AND  (Africa != 1)))  AND
(((rarely_work_interfere = 0) OR (rarely_work_interfere != 1)  AND  (Other !=
1)))  AND  (((never_work_interfere = 1) OR (never_work_interfere != 0)  AND
((yes_phys_health_consequence = 1) OR (yes_phys_health_consequence != 0)  AND
((not_self_employed = 0) OR (not_self_employed != 1)  AND  ((Northern_America
= 1) OR (Northern_America != 0)  AND  ((yes_mental_health_consequence = 1) OR
(yes_mental_health_consequence != 0)  AND  ((mature_adulthood = 1) OR
(mature_adulthood != 0)  AND  ((yes_benefits = 1) OR (yes_benefits != 0)  AND
((yes_wellness_program = 1) OR (yes_wellness_program != 0)  AND
((yes_seek_help = 1) OR (yes_seek_help != 0)  AND  ((yes_self_employed = 1)
OR (yes_self_employed != 0)  AND  ((Australia_and_New_Zealand = 1) OR
(Australia_and_New_Zealand != 0)  AND  ((yes_care_options = 1) OR
(yes_care_options != 0)  AND  ((South_Atlantic_USA = 1) OR
(South_Atlantic_USA != 0)  AND  ((West_North_Centeral_USA = 1) OR
(West_North_Centeral_USA != 0)  AND  (Somewhat_difficult_leave =
1))))))))))))))))
```

THEN

Node = 16

Prediction = 0

Probability = 0.859155

```
/* Node 19 */.
IF (((sometimes_work_interfere = 0) OR (sometimes_work_interfere != 1)  AND
((yes_family_history = 0) OR (yes_family_history != 1)  AND
(Very_difficult_leave != 1))))  AND  (((often_work_interfere = 0) OR
(often_work_interfere != 1)  AND  (Africa != 1)))  AND
(((rarely_work_interfere = 1) OR (rarely_work_interfere != 0)  AND  (Other =
1)))  AND  (((yes_family_history = 0) OR (yes_family_history != 1)  AND
((Male = 1) OR (Male != 0)  AND  ((Female = 0) OR (Female != 1)  AND
((no_phys_health_consequence = 1) OR (no_phys_health_consequence != 0)  AND
((yes_tech_company = 1) OR (yes_tech_company != 0)  AND
((no_phys_health_interview = 0) OR (no_phys_health_interview != 1)  AND
((yes_mental_health_consequence = 0) OR (yes_mental_health_consequence != 1)
```

```
AND  ((no_mental_health_consequence = 1) OR (no_mental_health_consequence !=
0)  AND  ((yes_anonymity = 0) OR (yes_anonymity != 1)  AND
((Very_difficult_leave = 0) OR (Very_difficult_leave != 1)  AND
((yes_benefits = 0) OR (yes_benefits != 1)  AND  ((yes_coworkers = 0) OR
(yes_coworkers != 1)  AND  ((yes_phys_health_consequence = 0) OR
(yes_phys_health_consequence != 1)  AND  ((West_North_Centeral_USA = 0) OR
(West_North_Centeral_USA != 1)  AND  ((East_North_Centeral_USA = 0) OR
(East_North_Centeral_USA != 1)  AND  ((yes_self_employed = 0) OR
(yes_self_employed != 1)  AND  ((Middle_Atlantic_USA = 0) OR
(Middle_Atlantic_USA != 1)  AND  ((yes_phys_health_interview = 0) OR
(yes_phys_health_interview != 1)  AND  ((not_self_employed = 1) OR
(not_self_employed != 0)  AND  ((Somewhat_difficult_leave = 0) OR
(Somewhat_difficult_leave != 1)  AND  ((no_employees_1 = 0) OR
(no_employees_1 != 1)  AND  ((no_mental_health_interview = 1) OR
(no_mental_health_interview != 0)  AND  ((yes_obs_consequence = 0) OR
(yes_obs_consequence != 1)  AND  ((mid_life = 0) OR (mid_life != 1)  AND
((no_anonymity = 0) OR (no_anonymity != 1)  AND  ((no_employees_3 = 0) OR
(no_employees_3 != 1)  AND  ((no_employees_2 = 0) OR (no_employees_2 != 1)
AND  (Other != 1))))))))))))))))))))))))))))))))  AND  (((yes_care_options = 0)
OR (yes_care_options != 1)  AND  ((no_care_options = 1) OR (no_care_options
!= 0)  AND  ((yes_seek_help = 0) OR (yes_seek_help != 1)  AND  ((yes_benefits
= 0) OR (yes_benefits != 1)  AND  ((yes_anonymity = 0) OR (yes_anonymity !=
1)  AND  ((yes_wellness_program = 0) OR (yes_wellness_program != 1)  AND
((yes_supervisor = 0) OR (yes_supervisor != 1)  AND  ((yes_coworkers = 0) OR
(yes_coworkers != 1)  AND  ((no_wellness_program = 1) OR (no_wellness_program
!= 0)  AND  ((no_seek_help = 1) OR (no_seek_help != 0)  AND
((Somewhat_easy_leave = 0) OR (Somewhat_easy_leave != 1)  AND
((no_employees_5 = 0) OR (no_employees_5 != 1)  AND  ((Male = 1) OR (Male !=
0)  AND  ((no_mental_health_interview = 1) OR (no_mental_health_interview !=
0)  AND  ((yes_mental_vs_physical = 0) OR (yes_mental_vs_physical != 1)  AND
((yes_phys_health_interview = 0) OR (yes_phys_health_interview != 1)  AND
((Female = 0) OR (Female != 1)  AND  ((Pacific_USA = 0) OR (Pacific_USA != 1)
AND  ((no_benefits = 0) OR (no_benefits != 1)  AND
((Australia_and_New_Zealand = 0) OR (Australia_and_New_Zealand != 1)  AND
((Very_easy_leave = 0) OR (Very_easy_leave != 1)  AND
((Somewhat_difficult_leave = 0) OR (Somewhat_difficult_leave != 1)  AND
((West_North_Centeral_USA = 0) OR (West_North_Centeral_USA != 1)  AND
((early_adulthood = 1) OR (early_adulthood != 0)  AND
((yes_mental_health_interview = 0) OR (yes_mental_health_interview != 1)  AND
((mature_adulthood = 0) OR (mature_adulthood != 1)  AND  ((no_supervisor = 1)
OR (no_supervisor != 0)  AND  ((East_North_Centeral_USA = 0) OR
(East_North_Centeral_USA != 1)  AND  ((Very_difficult_leave = 0) OR
(Very_difficult_leave != 1)  AND  ((no_phys_health_consequence = 0) OR
(no_phys_health_consequence != 1)  AND  ((Mountain_West_USA = 0) OR
(Mountain_West_USA != 1)  AND  ((no_anonymity = 0) OR (no_anonymity != 1)
AND  ((mid_life = 0) OR (mid_life != 1)  AND  ((Other = 0) OR (Other != 1)
AND  (South_America != 1))))))))))))))))))))))))))))))))))))))))))))))))))
```

THEN

Node = 19

Prediction = 0

Probability = 0.612245

```
/* Node 20 */.
IF (((sometimes_work_interfere = 0) OR (sometimes_work_interfere != 1)  AND
((yes_family_history = 0) OR (yes_family_history != 1)  AND
(Very_difficult_leave != 1))))  AND  (((often_work_interfere = 0) OR
```

```
(often_work_interfere != 1)  AND  (Africa != 1)))  AND
(((rarely_work_interfere = 1) OR (rarely_work_interfere != 0)  AND  (Other =
1)))  AND  (((yes_family_history = 0) OR (yes_family_history != 1)  AND
((Male = 1) OR (Male != 0)  AND  ((Female = 0) OR (Female != 1)  AND
((no_phys_health_consequence = 1) OR (no_phys_health_consequence != 0)  AND
((yes_tech_company = 1) OR (yes_tech_company != 0)  AND
((no_phys_health_interview = 0) OR (no_phys_health_interview != 1)  AND
((yes_mental_health_consequence = 0) OR (yes_mental_health_consequence != 1)
AND  ((no_mental_health_consequence = 1) OR (no_mental_health_consequence !=
0)  AND  ((yes_anonymity = 0) OR (yes_anonymity != 1)  AND
((Very_difficult_leave = 0) OR (Very_difficult_leave != 1)  AND
((yes_benefits = 0) OR (yes_benefits != 1)  AND  ((yes_coworkers = 0) OR
(yes_coworkers != 1)  AND  ((yes_phys_health_consequence = 0) OR
(yes_phys_health_consequence != 1)  AND  ((West_North_Centeral_USA = 0) OR
(West_North_Centeral_USA != 1)  AND  ((East_North_Centeral_USA = 0) OR
(East_North_Centeral_USA != 1)  AND  ((yes_self_employed = 0) OR
(yes_self_employed != 1)  AND  ((Middle_Atlantic_USA = 0) OR
(Middle_Atlantic_USA != 1)  AND  ((yes_phys_health_interview = 0) OR
(yes_phys_health_interview != 1)  AND  ((not_self_employed = 1) OR
(not_self_employed != 0)  AND  ((Somewhat_difficult_leave = 0) OR
(Somewhat_difficult_leave != 1)  AND  ((no_employees_1 = 0) OR
(no_employees_1 != 1)  AND  ((no_mental_health_interview = 1) OR
(no_mental_health_interview != 0)  AND  ((yes_obs_consequence = 0) OR
(yes_obs_consequence != 1)  AND  ((mid_life = 0) OR (mid_life != 1)  AND
((no_anonymity = 0) OR (no_anonymity != 1)  AND  ((no_employees_3 = 0) OR
(no_employees_3 != 1)  AND  ((no_employees_2 = 0) OR (no_employees_2 != 1)
AND  (Other != 1))))))))))))))))))))))))))))))))  AND  (((yes_care_options = 1)
OR (yes_care_options != 0)  AND  ((no_care_options = 0) OR (no_care_options
!= 1)  AND  ((yes_seek_help = 1) OR (yes_seek_help != 0)  AND  ((yes_benefits
= 1) OR (yes_benefits != 0)  AND  ((yes_anonymity = 1) OR (yes_anonymity !=
0)  AND  ((yes_wellness_program = 1) OR (yes_wellness_program != 0)  AND
((yes_supervisor = 1) OR (yes_supervisor != 0)  AND  ((yes_coworkers = 1) OR
(yes_coworkers != 0)  AND  ((no_wellness_program = 0) OR (no_wellness_program
!= 1)  AND  ((no_seek_help = 0) OR (no_seek_help != 1)  AND
((Somewhat_easy_leave = 1) OR (Somewhat_easy_leave != 0)  AND
((no_employees_5 = 1) OR (no_employees_5 != 0)  AND  ((Male = 0) OR (Male !=
1)  AND  ((no_mental_health_interview = 0) OR (no_mental_health_interview !=
1)  AND  ((yes_mental_vs_physical = 1) OR (yes_mental_vs_physical != 0)  AND
((yes_phys_health_interview = 1) OR (yes_phys_health_interview != 0)  AND
((Female = 1) OR (Female != 0)  AND  ((Pacific_USA = 1) OR (Pacific_USA != 0)
AND  ((no_benefits = 1) OR (no_benefits != 0)  AND
((Australia_and_New_Zealand = 1) OR (Australia_and_New_Zealand != 0)  AND
((Very_easy_leave = 1) OR (Very_easy_leave != 0)  AND
((Somewhat_difficult_leave = 1) OR (Somewhat_difficult_leave != 0)  AND
((West_North_Centeral_USA = 1) OR (West_North_Centeral_USA != 0)  AND
((early_adulthood = 0) OR (early_adulthood != 1)  AND
((yes_mental_health_interview = 1) OR (yes_mental_health_interview != 0)  AND
((mature_adulthood = 1) OR (mature_adulthood != 0)  AND  ((no_supervisor = 0)
OR (no_supervisor != 1)  AND  ((East_North_Centeral_USA = 1) OR
(East_North_Centeral_USA != 0)  AND  ((Very_difficult_leave = 1) OR
(Very_difficult_leave != 0)  AND  ((no_phys_health_consequence = 1) OR
(no_phys_health_consequence != 0)  AND  ((Mountain_West_USA = 1) OR
(Mountain_West_USA != 0)  AND  ((no_anonymity = 1) OR (no_anonymity != 0)
AND  ((mid_life = 1) OR (mid_life != 0)  AND  ((Other = 1) OR (Other != 0)
AND  (South_America = 1))))))))))))))))))))))))))))))))))))))))))))))))))))
```

THEN
Node = 20
Prediction = 1
Probability = 0.720930

```
/* Node 18 */.
IF (((sometimes_work_interfere = 0) OR (sometimes_work_interfere != 1)  AND
((yes_family_history = 0) OR (yes_family_history != 1)  AND
(Very_difficult_leave != 1))))  AND  (((often_work_interfere = 0) OR
(often_work_interfere != 1)  AND  (Africa != 1)))  AND
(((rarely_work_interfere = 1) OR (rarely_work_interfere != 0)  AND  (Other =
1)))  AND  (((yes_family_history = 1) OR (yes_family_history != 0)  AND
((Male = 0) OR (Male != 1)  AND  ((Female = 1) OR (Female != 0)  AND
((no_phys_health_consequence = 0) OR (no_phys_health_consequence != 1)  AND
((yes_tech_company = 0) OR (yes_tech_company != 1)  AND
((no_phys_health_interview = 1) OR (no_phys_health_interview != 0)  AND
((yes_mental_health_consequence = 1) OR (yes_mental_health_consequence != 0)
AND  ((no_mental_health_consequence = 0) OR (no_mental_health_consequence !=
1)  AND  ((yes_anonymity = 1) OR (yes_anonymity != 0)  AND
((Very_difficult_leave = 1) OR (Very_difficult_leave != 0)  AND
((yes_benefits = 1) OR (yes_benefits != 0)  AND  ((yes_coworkers = 1) OR
(yes_coworkers != 0)  AND  ((yes_phys_health_consequence = 1) OR
(yes_phys_health_consequence != 0)  AND  ((West_North_Centeral_USA = 1) OR
(West_North_Centeral_USA != 0)  AND  ((East_North_Centeral_USA = 1) OR
(East_North_Centeral_USA != 0)  AND  ((yes_self_employed = 1) OR
(yes_self_employed != 0)  AND  ((Middle_Atlantic_USA = 1) OR
(Middle_Atlantic_USA != 0)  AND  ((yes_phys_health_interview = 1) OR
(yes_phys_health_interview != 0)  AND  ((not_self_employed = 0) OR
(not_self_employed != 1)  AND  ((Somewhat_difficult_leave = 1) OR
(Somewhat_difficult_leave != 0)  AND  ((no_employees_1 = 1) OR
(no_employees_1 != 0)  AND  ((no_mental_health_interview = 0) OR
(no_mental_health_interview != 1)  AND  ((yes_obs_consequence = 1) OR
(yes_obs_consequence != 0)  AND  ((mid_life = 1) OR (mid_life != 0)  AND
((no_anonymity = 1) OR (no_anonymity != 0)  AND  ((no_employees_3 = 1) OR
(no_employees_3 != 0)  AND  ((no_employees_2 = 1) OR (no_employees_2 != 0)
AND  (Other = 1))))))))))))))))))))))))))))))))
```

THEN
Node = 18
Prediction = 1
Probability = 0.888889

```
/* Node 11 */.
IF (((sometimes_work_interfere = 1) OR (sometimes_work_interfere != 0)  AND
((yes_family_history = 1) OR (yes_family_history != 0)  AND
(Very_difficult_leave = 1))))  AND  (((yes_care_options = 0) OR
(yes_care_options != 1)  AND  ((no_care_options = 1) OR (no_care_options !=
0)  AND  ((yes_benefits = 0) OR (yes_benefits != 1)  AND  ((yes_anonymity =
0) OR (yes_anonymity != 1)  AND  ((yes_seek_help = 0) OR (yes_seek_help != 1)
AND  ((yes_wellness_program = 0) OR (yes_wellness_program != 1)  AND
((yes_obs_consequence = 0) OR (yes_obs_consequence != 1)  AND
((no_wellness_program = 1) OR (no_wellness_program != 0)  AND
((Middle_Atlantic_USA = 0) OR (Middle_Atlantic_USA != 1)  AND
((Northern_America = 0) OR (Northern_America != 1)  AND
((yes_mental_vs_physical = 0) OR (yes_mental_vs_physical != 1)  AND
```

```
((New_England_USA = 0) OR (New_England_USA != 1)  AND  ((no_employees_3 = 0)
OR (no_employees_3 != 1)  AND  ((Mountain_West_USA = 0) OR (Mountain_West_USA
!= 1)  AND  ((Australia_and_New_Zealand = 0) OR (Australia_and_New_Zealand !=
1)  AND  ((Female = 0) OR (Female != 1)  AND  ((Very_easy_leave = 0) OR
(Very_easy_leave != 1)  AND  ((Northern_Europe = 0) OR (Northern_Europe != 1)
AND  (no_anonymity != 1)))))))))))))))))))  AND  (((yes_family_history = 0)
OR (yes_family_history != 1)  AND  ((Western_Europe = 1) OR (Western_Europe
!= 0)  AND  ((no_benefits = 1) OR (no_benefits != 0)  AND  ((yes_benefits =
0) OR (yes_benefits != 1)  AND  ((Very_difficult_leave = 1) OR
(Very_difficult_leave != 0)  AND  ((yes_mental_health_consequence = 1) OR
(yes_mental_health_consequence != 0)  AND  ((no_employees_4 = 1) OR
(no_employees_4 != 0)  AND  ((yes_self_employed = 1) OR (yes_self_employed !=
0)  AND  ((not_self_employed = 0) OR (not_self_employed != 1)  AND
((West_South_Centeral_USA = 1) OR (West_South_Centeral_USA != 0)  AND  ((Asia
= 1) OR (Asia != 0)  AND  ((Very_easy_leave = 1) OR (Very_easy_leave != 0)
AND  ((yes_phys_health_consequence = 1) OR (yes_phys_health_consequence != 0)
AND  ((Adolescence = 1) OR (Adolescence != 0)  AND  ((no_supervisor = 1) OR
(no_supervisor != 0)  AND  ((yes_coworkers = 1) OR (yes_coworkers != 0)  AND
((no_care_options = 1) OR (no_care_options != 0)  AND  ((early_adulthood = 0)
OR (early_adulthood != 1)  AND  ((yes_seek_help = 1) OR (yes_seek_help != 0)
AND  ((yes_anonymity = 1) OR (yes_anonymity != 0)  AND  ((Mountain_West_USA =
1) OR (Mountain_West_USA != 0)  AND  ((mature_adulthood = 1) OR
(mature_adulthood != 0)  AND  ((no_seek_help = 1) OR (no_seek_help != 0)  AND
((no_mental_health_consequence = 1) OR (no_mental_health_consequence != 0)
AND  ((Australia_and_New_Zealand = 1) OR (Australia_and_New_Zealand != 0)
AND  ((no_coworkers = 0) OR (no_coworkers != 1)  AND
((East_South_Centeral_USA = 0) OR (East_South_Centeral_USA != 1)  AND
((no_mental_health_interview = 0) OR (no_mental_health_interview != 1)  AND
((yes_remote_work = 0) OR (yes_remote_work != 1)  AND
((yes_mental_vs_physical = 1) OR (yes_mental_vs_physical != 0)  AND  ((Other
= 1) OR (Other != 0)  AND  ((Southeast_Europe = 1) OR (Southeast_Europe != 0)
AND  ((Africa = 1) OR (Africa != 0)  AND  (no_employees_5 =
1)))))))))))))))))))))))))))))))))))))))
THEN
Node = 11
Prediction = 1
Probability = 0.613636

/* Node 12 */.
IF (((sometimes_work_interfere = 1) OR (sometimes_work_interfere != 0)  AND
((yes_family_history = 1) OR (yes_family_history != 0)  AND
(Very_difficult_leave = 1))))  AND  (((yes_care_options = 0) OR
(yes_care_options != 1)  AND  ((no_care_options = 1) OR (no_care_options !=
0)  AND  ((yes_benefits = 0) OR (yes_benefits != 1)  AND  ((yes_anonymity =
0) OR (yes_anonymity != 1)  AND  ((yes_seek_help = 0) OR (yes_seek_help != 1)
AND  ((yes_wellness_program = 0) OR (yes_wellness_program != 1)  AND
((yes_obs_consequence = 0) OR (yes_obs_consequence != 1)  AND
((no_wellness_program = 1) OR (no_wellness_program != 0)  AND
((Middle_Atlantic_USA = 0) OR (Middle_Atlantic_USA != 1)  AND
((Northern_America = 0) OR (Northern_America != 1)  AND
((yes_mental_vs_physical = 0) OR (yes_mental_vs_physical != 1)  AND
((New_England_USA = 0) OR (New_England_USA != 1)  AND  ((no_employees_3 = 0)
OR (no_employees_3 != 1)  AND  ((Mountain_West_USA = 0) OR (Mountain_West_USA
!= 1)  AND  ((Australia_and_New_Zealand = 0) OR (Australia_and_New_Zealand !=
1)  AND  ((Female = 0) OR (Female != 1)  AND  ((Very_easy_leave = 0) OR
(Very_easy_leave != 1)  AND  ((Northern_Europe = 0) OR (Northern_Europe != 1)
```

```
AND  (no_anonymity != 1))))))))))))))))))))  AND  (((yes_family_history = 1)
OR (yes_family_history != 0)  AND  ((Western_Europe = 0) OR (Western_Europe
!= 1)  AND  ((no_benefits = 0) OR (no_benefits != 1)  AND  ((yes_benefits =
1) OR (yes_benefits != 0)  AND  ((Very_difficult_leave = 0) OR
(Very_difficult_leave != 1)  AND  ((yes_mental_health_consequence = 0) OR
(yes_mental_health_consequence != 1)  AND  ((no_employees_4 = 0) OR
(no_employees_4 != 1)  AND  ((yes_self_employed = 0) OR (yes_self_employed !=
1)  AND  ((not_self_employed = 1) OR (not_self_employed != 0)  AND
((West_South_Centeral_USA = 0) OR (West_South_Centeral_USA != 1)  AND  ((Asia
= 0) OR (Asia != 1)  AND  ((Very_easy_leave = 0) OR (Very_easy_leave != 1)
AND  ((yes_phys_health_consequence = 0) OR (yes_phys_health_consequence != 1)
AND  ((Adolescence = 0) OR (Adolescence != 1)  AND  ((no_supervisor = 0) OR
(no_supervisor != 1)  AND  ((yes_coworkers = 0) OR (yes_coworkers != 1)  AND
((no_care_options = 0) OR (no_care_options != 1)  AND  ((early_adulthood = 1)
OR (early_adulthood != 0)  AND  ((yes_seek_help = 0) OR (yes_seek_help != 1)
AND  ((yes_anonymity = 0) OR (yes_anonymity != 1)  AND  ((Mountain_West_USA =
0) OR (Mountain_West_USA != 1)  AND  ((mature_adulthood = 0) OR
(mature_adulthood != 1)  AND  ((no_seek_help = 0) OR (no_seek_help != 1)  AND
((no_mental_health_consequence = 0) OR (no_mental_health_consequence != 1)
AND  ((Australia_and_New_Zealand = 0) OR (Australia_and_New_Zealand != 1)
AND  ((no_coworkers = 1) OR (no_coworkers != 0)  AND
((East_South_Centeral_USA = 1) OR (East_South_Centeral_USA != 0)  AND
((no_mental_health_interview = 1) OR (no_mental_health_interview != 0)  AND
((yes_remote_work = 1) OR (yes_remote_work != 0)  AND
((yes_mental_vs_physical = 0) OR (yes_mental_vs_physical != 1)  AND  ((Other
= 0) OR (Other != 1)  AND  ((Southeast_Europe = 0) OR (Southeast_Europe != 1)
AND  ((Africa = 0) OR (Africa != 1)  AND  (no_employees_5 !=
1))))))))))))))))))))))))))))))))))))))))))
THEN
Node = 12
Prediction = 1
Probability = 0.784173

/* Node 13 */.
IF (((sometimes_work_interfere = 1) OR (sometimes_work_interfere != 0)  AND
((yes_family_history = 1) OR (yes_family_history != 0)  AND
(Very_difficult_leave = 1))))  AND  (((yes_care_options = 1) OR
(yes_care_options != 0)  AND  ((no_care_options = 0) OR (no_care_options !=
1)  AND  ((yes_benefits = 1) OR (yes_benefits != 0)  AND  ((yes_anonymity =
1) OR (yes_anonymity != 0)  AND  ((yes_seek_help = 1) OR (yes_seek_help != 0)
AND  ((yes_wellness_program = 1) OR (yes_wellness_program != 0)  AND
((yes_obs_consequence = 1) OR (yes_obs_consequence != 0)  AND
((no_wellness_program = 0) OR (no_wellness_program != 1)  AND
((Middle_Atlantic_USA = 1) OR (Middle_Atlantic_USA != 0)  AND
((Northern_America = 1) OR (Northern_America != 0)  AND
((yes_mental_vs_physical = 1) OR (yes_mental_vs_physical != 0)  AND
((New_England_USA = 1) OR (New_England_USA != 0)  AND  ((no_employees_3 = 1)
OR (no_employees_3 != 0)  AND  ((Mountain_West_USA = 1) OR (Mountain_West_USA
!= 0)  AND  ((Australia_and_New_Zealand = 1) OR (Australia_and_New_Zealand !=
0)  AND  ((Female = 1) OR (Female != 0)  AND  ((Very_easy_leave = 1) OR
(Very_easy_leave != 0)  AND  ((Northern_Europe = 1) OR (Northern_Europe != 0)
AND  (no_anonymity = 1))))))))))))))))))))))  AND  (((no_benefits = 0) OR
(no_benefits != 1)  AND  ((yes_benefits = 1) OR (yes_benefits != 0)  AND
((yes_self_employed = 0) OR (yes_self_employed != 1)  AND
((not_self_employed = 1) OR (not_self_employed != 0)  AND  ((no_seek_help =
0) OR (no_seek_help != 1)  AND  ((no_anonymity = 0) OR (no_anonymity != 1)
```

```
AND  ((Very_difficult_leave = 0) OR (Very_difficult_leave != 1)  AND
((Western_Europe = 0) OR (Western_Europe != 1)  AND  ((mature_adulthood = 0)
OR (mature_adulthood != 1)  AND  ((Asia = 0) OR (Asia != 1)  AND
((Australia_and_New_Zealand = 0) OR (Australia_and_New_Zealand != 1)  AND
((West_South_Centeral_USA = 0) OR (West_South_Centeral_USA != 1)  AND
((South_America = 0) OR (South_America != 1)  AND  ((no_employees_4 = 0) OR
(no_employees_4 != 1)  AND  (Mountain_West_USA != 1))))))))))))))))
```

THEN

Node = 13

Prediction = 1

Probability = 0.925373

```
/* Node 14 */.
IF (((sometimes_work_interfere = 1) OR (sometimes_work_interfere != 0)  AND
((yes_family_history = 1) OR (yes_family_history != 0)  AND
(Very_difficult_leave = 1))))  AND  (((yes_care_options = 1) OR
(yes_care_options != 0)  AND  ((no_care_options = 0) OR (no_care_options !=
1)  AND  ((yes_benefits = 1) OR (yes_benefits != 0)  AND  ((yes_anonymity =
1) OR (yes_anonymity != 0)  AND  ((yes_seek_help = 1) OR (yes_seek_help != 0)
AND  ((yes_wellness_program = 1) OR (yes_wellness_program != 0)  AND
((yes_obs_consequence = 1) OR (yes_obs_consequence != 0)  AND
((no_wellness_program = 0) OR (no_wellness_program != 1)  AND
((Middle_Atlantic_USA = 1) OR (Middle_Atlantic_USA != 0)  AND
((Northern_America = 1) OR (Northern_America != 0)  AND
((yes_mental_vs_physical = 1) OR (yes_mental_vs_physical != 0)  AND
((New_England_USA = 1) OR (New_England_USA != 0)  AND  ((no_employees_3 = 1)
OR (no_employees_3 != 0)  AND  ((Mountain_West_USA = 1) OR (Mountain_West_USA
!= 0)  AND  ((Australia_and_New_Zealand = 1) OR (Australia_and_New_Zealand !=
0)  AND  ((Female = 1) OR (Female != 0)  AND  ((Very_easy_leave = 1) OR
(Very_easy_leave != 0)  AND  ((Northern_Europe = 1) OR (Northern_Europe != 0)
AND  (no_anonymity = 1))))))))))))))))))))  AND  (((no_benefits = 1) OR
(no_benefits != 0)  AND  ((yes_benefits = 0) OR (yes_benefits != 1)  AND
((yes_self_employed = 1) OR (yes_self_employed != 0)  AND
((not_self_employed = 0) OR (not_self_employed != 1)  AND  ((no_seek_help =
1) OR (no_seek_help != 0)  AND  ((no_anonymity = 1) OR (no_anonymity != 0)
AND  ((Very_difficult_leave = 1) OR (Very_difficult_leave != 0)  AND
((Western_Europe = 1) OR (Western_Europe != 0)  AND  ((mature_adulthood = 1)
OR (mature_adulthood != 0)  AND  ((Asia = 1) OR (Asia != 0)  AND
((Australia_and_New_Zealand = 1) OR (Australia_and_New_Zealand != 0)  AND
((West_South_Centeral_USA = 1) OR (West_South_Centeral_USA != 0)  AND
((South_America = 1) OR (South_America != 0)  AND  ((no_employees_4 = 1) OR
(no_employees_4 != 0)  AND  (Mountain_West_USA = 1))))))))))))))))
```

THEN

Node = 14

Prediction = 1

Probability = 0.733333

SAS Code

```
proc import datafile="survey.csv" out=survey replace;
delimiter=',';
getnames=yes;
run;
TITLE "Boxplots - AGE and treatment with 5Number Summary";
```

```
ods graphics off;
PROC SORT;
BY treatment;
RUN;
PROC BOXPLOT;
PLOT AGE*treatment ;
inset min mean max stddev/
header = 'Overall Statistics'
pos = tm;
insetgroup min mean Q1 Q2 Q3 max range stddev/
header = 'Statistics by Private';
RUN;
Title "The Dataset";
proc print data=survey(obs = 10);
run;
*To Check the Frequency of dataset and check all the requirements such as
preprocessing;
proc freq data=survey;
run;
*Removing Outliers from age variable;
data survey;
set survey;
if age>100 then delete;
else if age<4 then delete;
run;
TITLE "Boxplots - AGE and treatment with 5Number Summary changes";
ods graphics off;
PROC SORT;
BY treatment;
RUN;
PROC BOXPLOT;
PLOT AGE*treatment ;
inset min mean max stddev/
header = 'Overall Statistics'
pos = tm;
insetgroup min mean Q1 Q2 Q3 max range stddev/
header = 'Statistics by Private';
RUN;
*checking if Geographical location affects the mental health;
pattern1 color = lightred;
pattern2 color = lightblue;
proc gchart data = survey;
title 'Is Geographical location vs Treatment';
vbar country/ discrete width = 10 inside = percent group = treatment
patternID = group;
run;
*Creating Dummy Variables;
data survey;
set survey;
if (Gender = "Cis Fema") then Female = 1;
else if (Gender = "F") then Female = 1;
else if (Gender = "Femake") then Female = 1;
else if (Gender = "Female") then Female = 1;
else if (Gender = "Female (") then Female = 1;
else if (Gender = "Woman") then Female = 1;
else if (Gender = "f") then Female = 1;
else if (Gender = "cis-fema") then Female = 1;
```

```
else if (Gender = "femail") then Female = 1;
else if (Gender = "female") then Female = 1;
else if (Gender = "woman") then Female = 1;
else Female = 0;
if (Gender= "Cis Male") then Male = 1;
else if (Gender= "Cis Man") then Male = 1;
else if (Gender= "Guy (-is") then Male = 1;
else if (Gender= "M") then Male = 1;
else if (Gender= "Mail") then Male = 1;
else if (Gender= "Make") then Male = 1;
else if (Gender= "Mal") then Male = 1;
else if (Gender= "Male (CI") then Male = 1;
else if (Gender= "Male") then Male = 1;
else if (Gender= "Male-ish") then Male = 1;
else if (Gender= "Malr") then Male = 1;
else if (Gender= "Man") then Male = 1;
else if (Gender= "m") then Male = 1;
else if (Gender= "maile") then Male = 1;
else if (Gender= "male") then Male = 1;
else if (Gender= "male lea") then Male = 1;
else if (Gender= "msle") then Male = 1;
else Male =0;
if (Gender= "A little") then Other = 1;
else if (Gender= "Agender") then Other = 1;
else if (Gender= "All") then Other = 1;
else if (Gender= "Androgyn") then Other = 1;
else if (Gender= "Enby") then Other = 1;
else if (Gender= "Genderqu") then Other = 1;
else if (Gender= "Nah") then Other = 1;
else if (Gender= "Neuter") then Other = 1;
else if (Gender= "Trans wo") then Other = 1;
else if (Gender= "Trans-fe") then Other = 1;
else if (Gender= "fluid") then Other = 1;
else if (Gender= "non-bina") then Other = 1;
else if (Gender= "ostensib") then Other = 1;
else if (Gender= "p") then Other = 1;
else if (Gender= "queer") then Other = 1;
else if (Gender= "queer/sh") then Other = 1;
else if (Gender= "somethin") then Other = 1;
else Other = 0;
if (state= "AK") then Pacific_USA = 1;
else if (state= "WA") then Pacific_USA = 1;
else if (state= "OR") then Pacific_USA = 1;
else if (state= "CA") then Pacific_USA = 1;
else Pacific_USA = 0;
if (state= "MT") then Mountain_West_USA = 1;
else if (state= "ID") then Mountain_West_USA = 1;
else if (state= "WY") then Mountain_West_USA = 1;
else if (state= "NV") then Mountain_West_USA = 1;
else if (state= "UT") then Mountain_West_USA = 1;
else if (state= "CO") then Mountain_West_USA = 1;
else if (state= "AZ") then Mountain_West_USA = 1;
else if (state= "NM") then Mountain_West_USA = 1;
else Mountain_West_USA = 0;
if (state= "ND") then West_North_Centeral_USA = 1;
else if (state= "SD") then West_North_Centeral_USA = 1;
else if (state= "MN") then West_North_Centeral_USA = 1;
```

```
else if (state= "IA") then West_North_Centeral_USA = 1;
else if (state= "NE") then West_North_Centeral_USA = 1;
else if (state= "KS") then West_North_Centeral_USA = 1;
else if (state= "MO") then West_North_Centeral_USA = 1;
else West_North_Centeral_USA = 0;
if (state= "TX") then West_South_Centeral_USA = 1;
else if (state= "OK") then West_South_Centeral_USA = 1;
else if (state= "AR") then West_South_Centeral_USA = 1;
else if (state= "LA") then West_South_Centeral_USA = 1;
else West_South_Centeral_USA = 0;
if (state= "WI") then East_North_Centeral_USA = 1;
else if (state= "MI") then East_North_Centeral_USA = 1;
else if (state= "IN") then East_North_Centeral_USA = 1;
else if (state= "IL") then East_North_Centeral_USA = 1;
else if (state= "OH") then East_North_Centeral_USA = 1;
else  East_North_Centeral_USA =  0;
if (state= "KY") then East_South_Centeral_USA = 1;
else if (state= "TN") then East_South_Centeral_USA = 1;
else if (state= "MS") then East_South_Centeral_USA = 1;
else if (state= "AL") then East_South_Centeral_USA = 1;
else East_South_Centeral_USA = 0;
if (state= "NY") then Middle_Atlantic_USA = 1;
else if (state= "PA") then Middle_Atlantic_USA = 1;
else if (state= "NJ") then Middle_Atlantic_USA = 1;
else Middle_Atlantic_USA = 0;
if (state= "WV") then South_Atlantic_USA = 1;
else if (state= "MD") then South_Atlantic_USA = 1;
else if (state= "DC") then South_Atlantic_USA = 1;
else if (state= "DE") then South_Atlantic_USA = 1;
else if (state= "VA") then South_Atlantic_USA = 1;
else if (state= "NC") then South_Atlantic_USA = 1;
else if (state= "SC") then South_Atlantic_USA = 1;
else if (state= "GA") then South_Atlantic_USA = 1;
else if (state= "FL") then South_Atlantic_USA = 1;
else South_Atlantic_USA = 0;
if (state= "ME") then New_England_USA = 1;
else if (state= "VT") then New_England_USA = 1;
else if (state= "NH") then New_England_USA = 1;
else if (state= "MA") then New_England_USA = 1;
else if (state= "CT") then New_England_USA = 1;
else if (state= "RI") then New_England_USA = 1;
else New_England_USA = 0;
if (Country= "Austria") then Central_Europe = 1;
else if (Country= "Croatia") then Central_Europe = 1;
else if (Country= "Czech Republic") then Central_Europe = 1;
else if (Country= "Germany") then Central_Europe = 1;
else if (Country= "Hungary") then Central_Europe = 1;
else if (Country= "Poland") then Central_Europe = 1;
else if (Country= "Slovenia") then Central_Europe = 1;
else if (Country= "Switzerland") then Central_Europe = 1;
else Central_Europe = 0;
if (Country= "Bosnia and Herze") then Southeast_Europe = 1;
else if (Country= "Bulgaria") then Southeast_Europe = 1;
else if (Country= "Greece") then Southeast_Europe = 1;
else if (Country= "Romania") then Southeast_Europe = 1;
else if (Country= "Portugal") then Southeast_Europe = 1;
else if (Country= "Georgia") then Southeast_Europe = 1;
```

```
else if (Country= "Moldova") then Southeast_Europe = 1;
else Southeast_Europe = 0;
if (Country= "Belgium") then Western_Europe = 1;
else if (Country= "France") then Western_Europe = 1;
else if (Country= "Ireland") then Western_Europe = 1;
else if (Country= "Italy") then Western_Europe = 1;
else if (Country= "Netherlands") then Western_Europe = 1;
else if (Country= "Denmark") then Western_Europe = 1;
else if (Country= "United Kingdom") then Western_Europe = 1;
else Western_Europe = 0;
if (Country= "Finland") then Northern_Europe = 1;
else if (Country= "Sweden") then Northern_Europe = 1;
else if (Country= "Norway") then Northern_Europe = 1;
else if (Country= "Latvia") then Northern_Europe = 1;
else Northern_Europe = 0;
if (Country= "Portugal") then Southwestern_Europe = 1;
else if (Country= "Spain") then Southwestern_Europe = 1;
else Southwestern_Europe = 0;
if (Country= "Australia") then Australia_and_New_Zealand = 1;
else if (Country= "New Zealand") then Australia_and_New_Zealand = 1;
else Australia_and_New_Zealand = 0;
if (Country= "China") then Asia = 1;
else if (Country= "Thailand") then Asia = 1;
else if (Country= "Japan") then Asia = 1;
else if (Country= "India") then Asia = 1;
else if (Country= "Singapore") then Asia = 1;
else if (Country= "Philippines") then Asia = 1;
else if (Country= "Russia") then Asia = 1;
else if (Country= "Israel") then Asia = 1;
else Asia = 0;
if (Country= "Canada") then Northern_America = 1;
else if (Country= "Bahamas, The") then Northern_America = 1;
else Northern_America = 0;
if (Country= "Brazil") then South_America = 1;
else if (Country= "Colombia") then South_America = 1;
else if (Country= "Uruguay") then South_America = 1;
else South_America = 0;
if (Country= "Mexico") then Central_America = 1;
else if (Country= "Costa Rica") then Central_America = 1;
else Central_America = 0;
if (Country= "Nigeria") then Africa = 1;
else if (Country= "Zimbabwe") then Africa= 1;
else if (Country= "South Africa") then Africa= 1;
else Africa = 0;
if (self_employed="No") then not_self_employed=1;
else not_self_employed=0;
if(self_employed="Yes") then yes_self_employed=1;
else  yes_self_employed=0;
if (family_history="Yes") then yes_family_history=1;
else yes_family_history=0;
if (treatment="Yes") then yes_treatment=1;
else yes_treatment=0;
if (work_interfere="Never") then never_work_interfere=1;
else never_work_interfere=0;
if (work_interfere="Often") then often_work_interfere=1;
else often_work_interfere = 0;
if (work_interfere="Rarely") then rarely_work_interfere=1;
```

```
else rarely_work_interfere=0;
if (work_interfere="Sometimes") then sometimes_work_interfere=1;
else sometimes_work_interfere=0;
if (no_employees="100-500") then no_employees_1=1;
else no_employees_1=0;
if (no_employees="26-100")then no_employees_2=1;
else no_employees_2=0;
if (no_employees="500-1000")then no_employees_3=1;
else no_employees_3=0;
if(no_employees="6-25")then no_employees_4=1;
else no_employees_4=0;
if(no_employees="More than 1000") then no_employees_5=1;
else no_employees_5=0;
if(remote_work="Yes") then yes_remote_work=1;
else yes_remote_work=0;
if(tech_company="Yes") then yes_tech_company=1;
else yes_tech_company=0;
if(benefits="No") then no_benefits=1;
else no_benefits=0;
if(benefits="Yes") then yes_benefits=1;
else yes_benefits=0;
if(care_options="No") then no_care_options=1;
else no_care_options=0;
if(care_options="Yes") then yes_care_options=1;
else yes_care_options=0;
if (wellness_program="No") then no_wellness_program=1;
else no_wellness_program=0;
if (wellness_program="Yes") then yes_wellness_program=1;
else yes_wellness_program=0;
if (seek_help="No") then no_seek_help=1;
else no_seek_help=0;
if (seek_help="Yes") then yes_seek_help=1;
else yes_seek_help=0;
if (anonymity="No") then no_anonymity=1;
else no_anonymity=0;
if (anonymity="Yes") then yes_anonymity=1;
else yes_anonymity=0;
if (mental_health_consequence="No") then no_mental_health_consequence=1;
else no_mental_health_consequence=0;
if (mental_health_consequence="Yes") then yes_mental_health_consequence=1;
else yes_mental_health_consequence=0;
if (phys_health_consequence="No") then no_phys_health_consequence=1;
else no_phys_health_consequence=0;
if (phys_health_consequence="Yes") then yes_phys_health_consequence=1;
else yes_phys_health_consequence=0;
if (coworkers="No") then no_coworkers=1;
else no_coworkers=0;
if (coworkers="Yes") then yes_coworkers=1;
else yes_coworkers=0;
if (supervisor="No") then no_supervisor=1;
else no_supervisor=0;
if (supervisor="Yes") then yes_supervisor=1;
else yes_supervisor=0;
if (mental_health_interview="No") then no_mental_health_interview=1;
else no_mental_health_interview=0;
if (mental_health_interview="Yes") then yes_mental_health_interview=1;
else yes_mental_health_interview=0;
```

```
if (phys_health_interview="No") then no_phys_health_interview=1;
else no_phys_health_interview=0;
if (phys_health_interview="Yes") then yes_phys_health_interview=1;
else yes_phys_health_interview=0;
if (mental_vs_physical="No") then no_mental_vs_physical=1;
else no_mental_vs_physical=0;
if (mental_vs_physical="Yes") then yes_mental_vs_physical=1;
else yes_mental_vs_physical=0;
if (obs_consequence="Yes") then yes_obs_consequence=1;
else yes_obs_consequence=0;
if (leave="Somewhat difficult") then Somewhat_difficult_leave= 1;
else Somewhat_difficult_leave=0;
if (leave="Somewhat easy") then Somewhat_easy_leave= 1;
else Somewhat_easy_leave=0;
if (leave="Very difficult") then Very_difficult_leave= 1;
else Very_difficult_leave=0;
if (leave="Very easy") then Very_easy_leave= 1;
else Very_easy_leave=0;
if(age > 4 and age <12) then childhood = 1;
else childhood = 0;
if(age > 11 and age <21) then Adolescence = 1;
else adolescence = 0;
if(age>20 and age<35) then early_adulthood = 1;
else early_adulthood = 0;
if(age>34 and age<50) then mid_life = 1;
else mid_life = 0;
if(age>49 and age<80) then mature_adulthood=1;
else mature_adulthood=0;
run;
proc print data=survey (obs=10);
run;
*Correlation Matrix;
proc corr;
var yes_treatment Female Male Other Pacific_USA Mountain_West_USA
West_North_Centeral_USA West_South_Centeral_USA East_North_Centeral_USA
East_South_Centeral_USA Middle_Atlantic_USA South_Atlantic_USA
New_England_USA Central_Europe Southeast_Europe Western_Europe
Northern_Europe Southwestern_Europe Australia_and_New_Zealand Asia
Northern_America South_America Central_America Africa not_self_employed
yes_self_employed yes_family_history never_work_interfere
often_work_interfere rarely_work_interfere sometimes_work_interfere
no_employees_1 no_employees_2 no_employees_3 no_employees_4 no_employees_5
yes_remote_work yes_tech_company no_benefits yes_benefits no_care_options
yes_care_options no_wellness_program yes_wellness_program no_seek_help
yes_seek_help no_anonymity yes_anonymity no_mental_health_consequence
yes_mental_health_consequence no_phys_health_consequence
yes_phys_health_consequence no_coworkers yes_coworkers no_supervisor
yes_supervisor no_mental_health_interview yes_mental_health_interview
no_phys_health_interview yes_phys_health_interview no_mental_vs_physical
yes_mental_vs_physical yes_obs_consequence Somewhat_difficult_leave
Somewhat_easy_leave Very_difficult_leave Very_easy_leave childhood
Adolescence early_adulthood mid_life mature_adulthood;
run;
*interactive variables Creation;
data survey;
set survey;
int_var1 = female*often_work_interfere;
```

```
int_var2 = female*rarely_work_interfere;
int_var3 = female*sometimes_work_interfere;
int_var4 = often_work_interfere*yes_obs_consequence;
int_var5 = yes_obs_consequence*Very_difficult_leave;
int_var6 = often_work_interfere*Very_difficult_leave;
int_var7 = rarely_work_interfere*sometimes_work_interfere;
int_var8 = rarely_work_interfere*yes_obs_consequence;
int_var9 = rarely_work_interfere*Very_difficult_leave;
int_var10 = sometimes_work_interfere*yes_obs_consequence;
int_var11 = sometimes_work_interfere*Very_difficult_leave;
int_var12 = female*Very_difficult_leave;
int_var13 = often_work_interfere*sometimes_work_interfere;
int_var14 = female*yes_obs_consequence ;
int_var15 = often_work_interfere*rarely_work_interfere;
run;
*Interaction model with interactive variables;
proc reg data=survey;
model yes_treatment=female often_work_interfere rarely_work_interfere
sometimes_work_interfere yes_obs_consequence Very_difficult_leave int_var1
int_var2 int_var3 int_var4 int_var5 int_var6 int_var7 int_var8 int_var9
int_var10 int_var11 int_var12 int_var13 int_var14 int_var15;
run;
*final Interaction model with interactive variables;
title"Final interaction model";
proc reg data=survey;
model yes_treatment= often_work_interfere rarely_work_interfere
sometimes_work_interfere int_var2;
run;
* Split the data into training and test sets - 66/34;
proc surveyselect data=survey out=train_all seed=5277 samprate=0.66 outall;
proc print data=train_all (obs=10);
run;
*Write the sas code to create a new field called train_y;
data train_all;
set train_all;
if selected then train_y = yes_treatment;
proc print data=train_all (obs=10);
run;
*Model selection with traininf dataset with Forward;
proc logistic data = train_all;
model train_y(event='1')=Female Male Other Pacific_USA Mountain_West_USA
West_North_Centeral_USA West_South_Centeral_USA East_North_Centeral_USA
East_South_Centeral_USA Middle_Atlantic_USA South_Atlantic_USA
New_England_USA Central_Europe Southeast_Europe Western_Europe
Northern_Europe Southwestern_Europe Australia_and_New_Zealand Asia
Northern_America South_America Central_America Africa not_self_employed
yes_self_employed yes_family_history never_work_interfere
often_work_interfere rarely_work_interfere sometimes_work_interfere
no_employees_1 no_employees_2 no_employees_3 no_employees_4 no_employees_5
yes_remote_work yes_tech_company no_benefits yes_benefits no_care_options
yes_care_options no_wellness_program yes_wellness_program no_seek_help
yes_seek_help no_anonymity yes_anonymity no_mental_health_consequence
yes_mental_health_consequence no_phys_health_consequence
yes_phys_health_consequence no_coworkers yes_coworkers no_supervisor
yes_supervisor no_mental_health_interview yes_mental_health_interview
no_phys_health_interview yes_phys_health_interview no_mental_vs_physical
yes_mental_vs_physical yes_obs_consequence Somewhat_difficult_leave
```

```
Somewhat_easy_leave Very_difficult_leave Very_easy_leave childhood
Adolescence early_adulthood mid_life mature_adulthood/ selection=forward;
run;
*Model selection with traininf dataset with Backward;
proc logistic data = train_all;
model train_y(event='1')=Female Male Other Pacific_USA Mountain_West_USA
West_North_Centeral_USA West_South_Centeral_USA East_North_Centeral_USA
East_South_Centeral_USA Middle_Atlantic_USA South_Atlantic_USA
New_England_USA Central_Europe Southeast_Europe Western_Europe
Northern_Europe Southwestern_Europe Australia_and_New_Zealand Asia
Northern_America South_America Central_America Africa not_self_employed
yes_self_employed yes_family_history never_work_interfere
often_work_interfere rarely_work_interfere sometimes_work_interfere
no_employees_1 no_employees_2 no_employees_3 no_employees_4 no_employees_5
yes_remote_work yes_tech_company no_benefits yes_benefits no_care_options
yes_care_options no_wellness_program yes_wellness_program no_seek_help
yes_seek_help no_anonymity yes_anonymity no_mental_health_consequence
yes_mental_health_consequence no_phys_health_consequence
yes_phys_health_consequence no_coworkers yes_coworkers no_supervisor
yes_supervisor no_mental_health_interview yes_mental_health_interview
no_phys_health_interview yes_phys_health_interview no_mental_vs_physical
yes_mental_vs_physical yes_obs_consequence Somewhat_difficult_leave
Somewhat_easy_leave Very_difficult_leave Very_easy_leave childhood
Adolescence early_adulthood mid_life mature_adulthood/ selection=backward;
run;
*Model selection with traininf dataset with stepwise;
proc logistic data = train_all;
model train_y(event='1')= Female Male Other Pacific_USA Mountain_West_USA
West_North_Centeral_USA West_South_Centeral_USA East_North_Centeral_USA
East_South_Centeral_USA Middle_Atlantic_USA South_Atlantic_USA
New_England_USA Central_Europe Southeast_Europe Western_Europe
Northern_Europe Southwestern_Europe Australia_and_New_Zealand Asia
Northern_America South_America Central_America Africa not_self_employed
yes_self_employed yes_family_history never_work_interfere
often_work_interfere rarely_work_interfere sometimes_work_interfere
no_employees_1 no_employees_2 no_employees_3 no_employees_4 no_employees_5
yes_remote_work yes_tech_company no_benefits yes_benefits no_care_options
yes_care_options no_wellness_program yes_wellness_program no_seek_help
yes_seek_help no_anonymity yes_anonymity no_mental_health_consequence
yes_mental_health_consequence no_phys_health_consequence
yes_phys_health_consequence no_coworkers yes_coworkers no_supervisor
yes_supervisor no_mental_health_interview yes_mental_health_interview
no_phys_health_interview yes_phys_health_interview no_mental_vs_physical
yes_mental_vs_physical yes_obs_consequence Somewhat_difficult_leave
Somewhat_easy_leave Very_difficult_leave Very_easy_leave childhood
Adolescence early_adulthood mid_life mature_adulthood/ selection=stepwise;
run;
*Compute model-1 based upon Forward and Stepwise selection method;
Title "Model 1";
proc logistic data=train_all;
model train_y (event='1') = male yes_family_history never_work_interfere
often_work_interfere rarely_work_interfere sometimes_work_interfere
no_employees_3 yes_care_options Somewhat_easy_leave;
run;
*Compute model-2 based upon Backward selection method;
Title "Model 2";
proc logistic data=train_all;
```

```
model train_y (event='1') = Female yes_family_history never_work_interfere
often_work_interfere rarely_work_interfere sometimes_work_interfere
no_employees_3 yes_care_options Somewhat_easy_leave;
run;
*compute pred. probability for test set for model-1;
Title "Testing Model 1";
proc logistic data=train_all;
model train_y (event='1') = male yes_family_history never_work_interfere
often_work_interfere rarely_work_interfere sometimes_work_interfere
no_employees_3 yes_care_options Somewhat_easy_leave /ctable pprob = (0.1 to
0.8 by 0.05);
output out=pred(where =(train_y=.)) p=phat lower=lcl upper=ucl
predprob=(individual crossvalidate);
run;
proc print data = pred (obs=10);
run;
*Compute the predicted Y based on the cut-off value and;
*create confusion matrix TP,TN,..;
data probs;
set pred;
pred_y=0;
threshold=0.55;
*compute Predicted Y;
if phat>threshold then pred_y=1;
else pred_y=0;
run;
*confusion matrix;
title "Confusion Matrix for Model 1";
proc freq data = probs;
tables yes_treatment*pred_y/norow nocol nopercent;
run;
*compute pred. probability for test set for model-2;
Title "Testing Model 2";
proc logistic data=train_all;
model train_y (event='1') = Female yes_family_history never_work_interfere
often_work_interfere rarely_work_interfere sometimes_work_interfere
no_employees_3 yes_care_options Somewhat_easy_leave/ctable pprob = (0.1 to
0.8 by 0.05);
output out=pred(where =(train_y=.)) p=phat lower=lcl upper=ucl
predprob=(individual crossvalidate);
run;
proc print data = pred (obs=10);
run;
*Compute the predicted Y based on the cut-off value and;
*create confusion matrix TP,TN,..;
data probs;
set pred;
pred_y=0;
threshold=0.5;
*compute Predicted Y;
if phat>threshold then pred_y=1;
else pred_y=0;
run;
*confusion matrix;
title "COnfusion Matrix for model 2";
proc freq data = probs;
tables yes_treatment*pred_y/norow nocol nopercent;
```

```
run;
*Model-2 will be our final regression Model;
Title "Final Regression Model";
proc logistic data=train_all;
model train_y (event='1') = Female yes_family_history never_work_interfere
often_work_interfere rarely_work_interfere sometimes_work_interfere
no_employees_3 yes_care_options Somewhat_easy_leave;
run;
*Residual plots;
proc reg;
model yes_treatment= Female yes_family_history never_work_interfere
often_work_interfere rarely_work_interfere sometimes_work_interfere
no_employees_3 yes_care_options Somewhat_easy_leave ;
plot student.*(Female yes_family_history never_work_interfere
often_work_interfere rarely_work_interfere sometimes_work_interfere
no_employees_3 yes_care_options Somewhat_easy_leave);
plot student.*predicted.;
plot npp.*student.;
run;
*influential points;
PROC REG data=survey;
TITLE "Detecting - Outliers and Influential points";
MODEL yes_treatment = Female yes_family_history never_work_interfere
often_work_interfere rarely_work_interfere sometimes_work_interfere
no_employees_3 yes_care_options Somewhat_easy_leave /INFLUENCE R;
RUN;
*Deleting Influential points and outliers;
data pred;
set pred;
if _n_=1250 then delete;
if _n_=1204 then delete;
if _n_=1178 then delete;
if _n_=983 then delete;
if _n_=923 then delete;
if _n_=822 then delete;
if _n_=652 then delete;
if _n_=570 then delete;
if _n_=374 then delete;
if _n_=218 then delete;
run;
*final regression Model after influential and outlier removal;
Title "Final Regression Model with out influential points";
proc logistic data=train_all;
model train_y (event='1') = Female yes_family_history never_work_interfere
often_work_interfere rarely_work_interfere sometimes_work_interfere
no_employees_3 yes_care_options Somewhat_easy_leave;
run;
*final regression Model after influential and outlier removal;
Title "Final Regression Model with out influential points EXO";
proc logistic data=train_all;
model train_y (event='1') = Female yes_family_history never_work_interfere
often_work_interfere rarely_work_interfere sometimes_work_interfere
yes_care_options;
run;
data new;
```

```
input train_y Female yes_family_history never_work_interfere
often_work_interfere rarely_work_interfere sometimes_work_interfere
yes_care_options;
datalines;
1 0 1 0 1 1 0 1
1 1 0 1 1 0 1 1
;
run;
data pred;
set new train_all;
proc print data=pred (obs=10);
run;
proc logistic data=pred;
model train_y(event='1') = Female yes_family_history never_work_interfere
often_work_interfere rarely_work_interfere sometimes_work_interfere
no_employees_3 yes_care_options Somewhat_easy_leave;
output out=pred p=phat lower=lcl upper=ucl ;
run;
proc print data=pred (obs=2);
run;
```

SPSS CODE

```
* Encoding: UTF-8.
* Decision Tree.
TREE yes_treatment [n] BY Female [n] Male [n] Other [n] Pacific_USA [n] Mountain_West_USA [n]
    West_North_Centeral_USA [n] West_South_Centeral_USA [n] East_North_Centeral_USA [n]
    East_South_Centeral_USA [n] Middle_Atlantic_USA [n] South_Atlantic_USA [n] New_England_USA [n]
    Central_Europe [n] Southeast_Europe [n] Western_Europe [n] Northern_Europe [n] Southwestern_Europe
    [n] Australia_and_New_Zealand [n] Asia [n] Northern_America [n] South_America [n] Central_America
    [n] Africa [n] not_self_employed [n] yes_self_employed [n] yes_family_history [n]
    never_work_interfere [n] often_work_interfere [n] rarely_work_interfere [n]
    sometimes_work_interfere [n] no_employees_1 [n] no_employees_2 [n] no_employees_3 [n]
    no_employees_4 [n] no_employees_5 [n] yes_remote_work [n] yes_tech_company [n] no_benefits [n]
    yes_benefits [n] no_care_options [n] yes_care_options [n] no_wellness_program [n]
    yes_wellness_program [n] no_seek_help [n] yes_seek_help [n] no_anonymity [n] yes_anonymity [n]
    no_mental_health_consequence [n] yes_mental_health_consequence [n] no_phys_health_consequence [n]
    yes_phys_health_consequence [n] no_coworkers [n] yes_coworkers [n] no_supervisor [n] yes_supervisor
    [n] no_mental_health_interview [n] yes_mental_health_interview [n] no_phys_health_interview [n]
    yes_phys_health_interview [n] no_mental_vs_physical [n] yes_mental_vs_physical [n]
    yes_obs_consequence [n] Somewhat_difficult_leave [n] Somewhat_easy_leave [n] Very_difficult_leave
    [n] Very_easy_leave [n] childhood [n] Adolescence [n] early_adulthood [n] mid_life [n]
    mature_adulthood [n]
  /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES NODEDEFS=YES
SCALE=AUTO
  /DEPCATEGORIES USEVALUES=[VALID]
  /PRINT MODELSUMMARY CLASSIFICATION RISK
  /PLOT IMPORTANCE
  /RULES NODES=TERMINAL SYNTAX=INTERNAL TYPE=SCORING SURROGATES=INCLUDE
  /METHOD TYPE=CRT MAXSURROGATES=AUTO PRUNE=NONE
  /GROWTHLIMIT MAXDEPTH=AUTO MINPARENTSIZE=80 MINCHILDSIZE=40
  /VALIDATION TYPE=CROSSVALIDATION(10) OUTPUT=BOTHSAMPLES
  /CRT IMPURITY=GINI MINIMPROVEMENT=0.0001
  /COSTS EQUAL
  /PRIORS FROMDATA ADJUST=NO
  /MISSING NOMINALMISSING=MISSING.
```

Yash Pandya

This team member did not provide his individual report on time.

Vivek Umeshkumar Bhavshar – Final Model Selected

Introduction

Goal is the same as the group, refer to the introduction section in the main report for details.

Methodology

Data Preparation

we obtained the data from https://www.kaggle.com/osmi/mental-health-in-tech-survey. After getting the dataset looked for missing values and possibilities for adding dummy variable as the data set contained 98% categorical data and rest scalable or numerical data. Furthermore, all the possible dummy variables were created to perform logistic regression and to move one step further in the process of getting knowledge or common goal. However, the outliers were also removed to get the more accuracy on the prediction model and the interaction variables were also created.

Model Approach

In the same direction to move further, the different models were applied on the dataset in which first the data were partitioned in to training and testing sets with 70% in training and 30% in testing. Moreover, the forward selection, stepwise and backward elimination were performed to get the significant attribute information, in which we looked for predicted probability and used certain threshold. Furthermore, the decision tree was also applied on the same data with the hold out partitioning technique with 70% in training and 30% on testing to come up with the best final model.

Validation

The classification matrix or confusion matrix were used to check the multicollinearity also used some performance measures such as specificity, sensitivity and accuracy.

Analysis, Results & Findings

Analysis

This data set is about the mental health for humans at work environment. Moreover, to analyze the number or statistic that how many of them actual needed treatment, the frequencies were checked and observed that about 50% of them were needed the treatment or their mental health were not good. Furthermore, many other attributes were analyzed, and the dummy variables were created to preprocess the dataset. The treatment attribute was ***selected*** as a dependent attribute and the rest were selected as an independent attribute. The given dataset has two binary variables which are Age and Timestamp out of which we kept Age and formed a Boxplot with the dependent variable treatment. Figure 7 in appendix depicts the same.

Interaction variables

The interaction variables were formed by combining different important attributes such as Female, Work Interfere, Family History, Care Option and Benefit for the given data set and dummy variables were also created.

Collinearity

How colinear the variables are with each other were checked with the correlation matrix in the given data set and observed the variables collinear with our dependent variable Treatment. However, the maximum collinearity measure observed was 0.4.

Influential Points and Outlier

Performed action to find outlier and influential points in the dataset. Moreover, to improve the model accuracy the outliers and influential points were removed from the dataset.

MODEL 1

Before we ran our any of the applied model we divided the dataset into two parts Training and Testing with 70% in Training and 30% in Testing.

For variable selection we ran diverse selection technique, for model 1 Forward selection and Stepwise were used and the results for both the techniques were same as both the techniques had gave the equal number and same attributes in the table called Analysis of maximum likelihood estimates. Where the likelihood ratio was 566.89 and the P-values was observed <0.0001. Moreover, the classification table and the cut off or threshold (0.4) values with help of specificity and sensitivity were calculated for training set, we also computed predicted probability for testing set. In addition to that, confusion matrix for model was used to get the performance measures like accuracy, specificity, sensitivity, precision and recall. Below are the calculated statistics:

Accuracy: 85.94%, Sensitivity: 93.8%, Specificity: 76%, Precision: 83.12%

Predictor in Model 1

Model 1 has given total 10 significant variables out of all which are as below.

`Male Middle_Atlantic_USA Has_family_history work_ifr_Never work_ifr_Often work_ifr_Rarely work_ifr_Sometimes has_benefits know_care_options has_coworkers`

Figure 1 in appendix shows maximum likelihood estimates for Forward selection also Figure 8 shows the confusion matrix for model 1.

MODEL 2

In model 2 the backward elimination selection method was used in which the likelihood ration was 561.94 and the P-values was observed <0.0001. Even in model 2 we calculated the classification table. However, the cut off value calculated from specificity and sensitivity for this model was 0.45. In addition to that, confusion matrix for model was used to get the performance measures like accuracy, specificity, sensitivity, precision and recall. Below are the calculated statistics:

Accuracy: 86.73%, Sensitivity: 93.8%, Specificity: 74.84%, Precision: 85.65%

Predictor in Model 2

Model 2 has given total 9 significant variables out of all which are as below.

Female Has_family_history work_ifr_Never work_ifr_Often work_ifr_Rarely work_ifr_Sometimes has_benefits know_care_options has_coworkers

Figure 2 in appendix shows maximum likelihood estimates for Backward Elimination Also Figure 9 shows the confusion matrix for model 1.

Final Model Selection

As projected above, we ran two diverse selection techniques which gave different accuracy and different number of total predictor. Moreover, between two model the second model has been selected as such it gives higher accuracy and less number of predictors the likelihood ratio was 561.94 and the P-values was observed <0.0001. Moreover, in the final model we found influential points and outliers which were removed from the data set to further improve the accuracy and performance of the model. Furthermore, the possible outlier and influential points removed from the dataset are as below.

1257, 828, 651, 603.

Predictor in Final Model

Final Model has given total 9 significant variables out of all which are as below.

`Female Has_family_history work_ifr_Never work_ifr_Often work_ifr_Rarely work_ifr_Sometimes has_benefits know_care_options has_coworkers`

Figure 3 in appendix shows maximum likelihood estimates for Final Model. Also, the final regression model is as below.

Log((Need-treatment-=1)/(Need-treatment=0))= - 5.2866 + 0.6475* Female + 1.0760*(Has_Family_history)+2.4317*(work_ifr_Never)+5.8586*(work_ifr_Often)+4.6904*(work_ifr_Rarely)+5.1115*(work_ifr_Sometimes)+0.7596*(has_benefits)+0.8074*(know_care_options)+0.6837 *(has coworkers)

In addition to that the residual plots were also created to check the principle violations in the logistic regression. Figures 4, 5 & 6 in appendix shows I-plots for the same.

Different Approach of Binary Tree (not covered in class)

For the same dataset I applied decision tree with hold out partitioning technique with 70% in training and 30% in testing. However, I had applied different cases for parent and child to check for the different outcomes and accuracy. The optimal tree is marked in yellow line.

Index	Cases for N_P, Nc	Training Accuracy (70%)	Testing Accuracy (30%)	Complexity
1	N_P=100 N_c=50	74.9%	75%	3
2	N_P=35 N_c=17	84.1%	82.6%	15
3	N_P=28 N_c=14	84.7%	82.3%	16

The table above depict information on number of parent (NP) and child (Nc), training and testing percentage also complexity by the number of terminal nodes in each tree. Moreover, among all Models of regression analysis and decision tree for this dataset the regression analysis model performs overall good so it is preferred to go with model 2 when compared other applied model. Figure 10, 11 & 12 shows Final Tree, Classification Table & Important attributes Respectively.

Reference

1. SpringerLink. 2018. *Investigating the influence of data splitting on the predictive ability of QSAR/QSPR models | SpringerLink*. https://link.springer.com/article/10.1007/s11224-011-9757-4 Accessed 19 March 2018.

2. Suhrid Balakrishnan. 2018. *Decision Trees for Functional Variables*. https://pdfs.semanticscholar.org/adfc/d9a8f4e89d9854f7890bc202aa188e004851.pdf. Accessed 8 March 2018.

3. Statistics Solutions. (2016). Data analysis plan: Binary Logistic Regression. fhttp://www.statisticssolutions.com/membership-resources/member-profile/data-analysis-plan-templates/binary-logistic-regression/ Accessed 6 March 2018

Appendix

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.3602	0.5506	62.7175	<.0001
Male	1	-0.7380	0.2160	11.6740	0.0006
Middle_Atlantic_USA	1	-0.5872	0.2889	4.1306	0.0421
Has_family_history	1	0.9662	0.1728	31.2819	<.0001
work_ifr_Never	1	2.3527	0.5488	18.3766	<.0001
work_ifr_Often	1	5.7322	0.5673	102.1037	<.0001
work_ifr_Rarely	1	4.7686	0.5400	77.9854	<.0001
work_ifr_Sometimes	1	5.2031	0.5238	98.6746	<.0001
has_benefits	1	0.6490	0.1963	10.9324	0.0009
know_care_options	1	0.7089	0.1922	13.5989	0.0002
has_coworkers	1	0.6581	0.2328	7.9909	0.0047

Figure 20 Analysis of maximum Likelihood Estimates for Forward Selection.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.1046	0.5240	94.9026	<.0001
Female	1	0.7098	0.2225	10.1774	0.0014
Has_family_history	1	0.9617	0.1721	31.2239	<.0001
work_ifr_Never	1	2.3468	0.5477	18.3583	<.0001
work_ifr_Often	1	5.7339	0.5665	102.4591	<.0001
work_ifr_Rarely	1	4.7488	0.5386	77.7317	<.0001
work_ifr_Sometimes	1	5.1923	0.5227	98.6923	<.0001
has_benefits	1	0.6063	0.1943	9.7336	0.0018
know_care_options	1	0.7083	0.1917	13.6562	0.0002
has_coworkers	1	0.6411	0.2312	7.6894	0.0056

Figure 21 Analysis of Maximum Likelihood for backward Elimination.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.2866	0.6139	74.1655	<.0001
Female	1	0.6475	0.2674	5.8639	0.0155
Has_family_history	1	1.0760	0.2049	27.5666	<.0001
work_ifr_Never	1	2.4317	0.6401	14.4327	0.0001
work_ifr_Often	1	5.8586	0.6633	78.0174	<.0001
work_ifr_Rarely	1	4.6904	0.6268	56.0009	<.0001
work_ifr_Sometimes	1	5.1115	0.6081	70.6571	<.0001
has_benefits	1	0.7596	0.2327	10.6587	0.0011
know_care_options	1	0.8074	0.2295	12.3738	0.0004
has_coworkers	1	0.6837	0.2742	6.2178	0.0126

Figure 22 Analysis of Maximum Likelihood for FINAL MODEL.

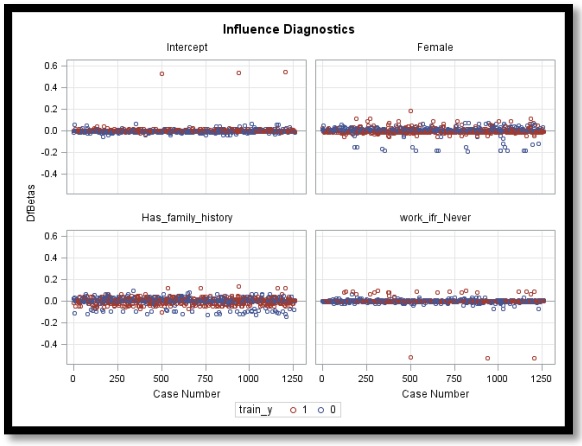


Figure 23 I-plots for final model Intercept, Female, Has Family history & work interfere never.

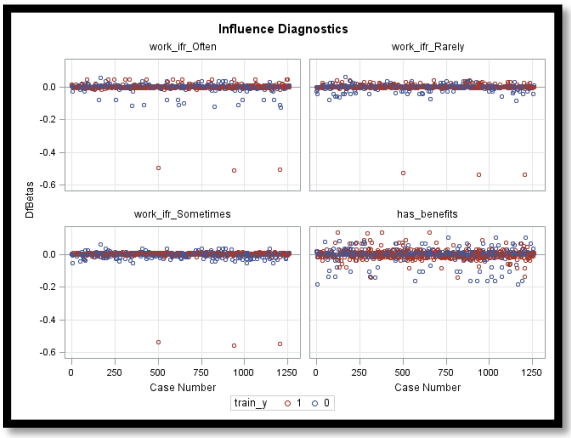


Figure 24 I-plots for final model work interfere often, work interfere rarely, work interfere sometimes & Has benefits.

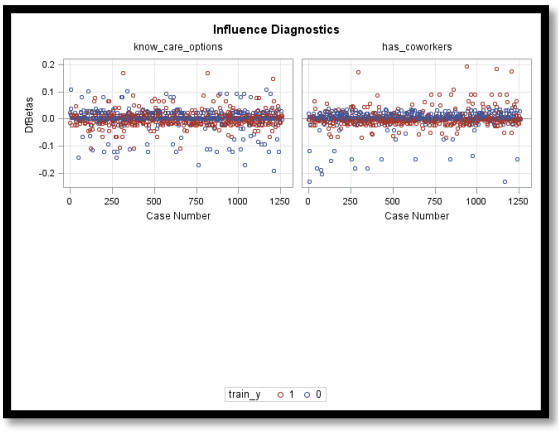


Figure 25 I-plots for final model know care option & has coworkers.

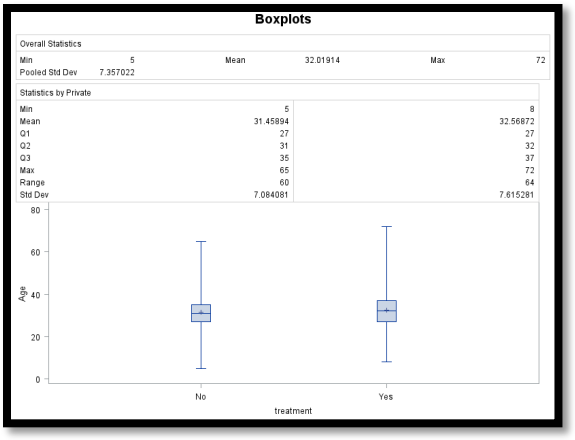
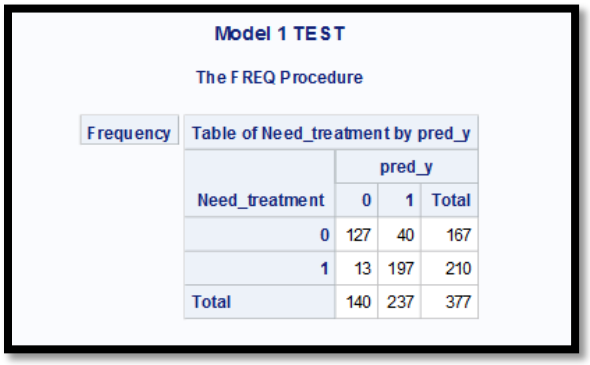


Figure 26 Boxplot for Age VS. Treatment

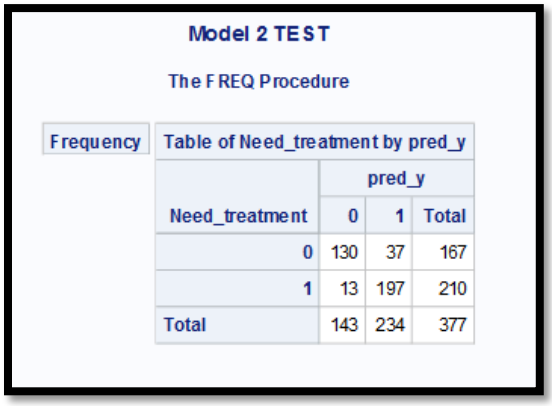


Model 1 TEST

The FREQ Procedure

Frequency	Table of Need_treatment by pred_y		
	pred_y		
Need_treatment	0	1	Total
0	127	40	167
1	13	197	210
Total	140	237	377

Figure 27 confusion matrix for model 1



Model 2 TEST

The FREQ Procedure

Frequency	Table of Need_treatment by pred_y		
	pred_y		
Need_treatment	0	1	Total
0	130	37	167
1	13	197	210
Total	143	234	377

Figure 28 confusion matrix for model 2.

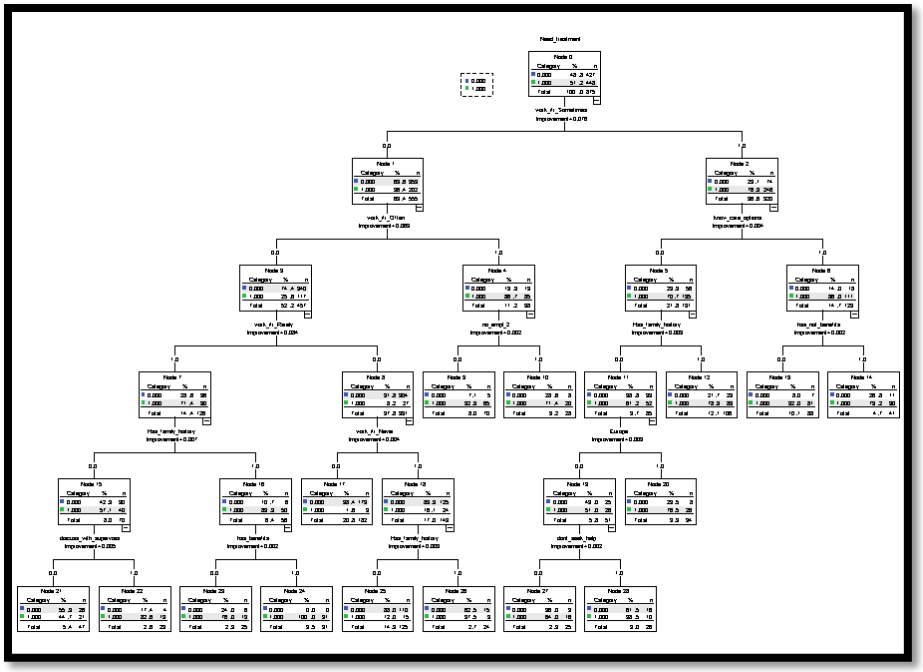


Figure 29 Final Tree Model.

Classification				
		Predicted		
Sample	Observed	0	1	Percent Correct
Training	0	346	81	81.0%
	1	58	390	87.1%
	Overall Percentage	46.2%	53.8%	84.1%
Test	0	149	46	76.4%
	1	21	168	88.9%
	Overall Percentage	44.3%	55.7%	82.6%
Growing Method: CRT Dependent Variable: Need treatment				

Figure 30 Classification Model for Final Tree.

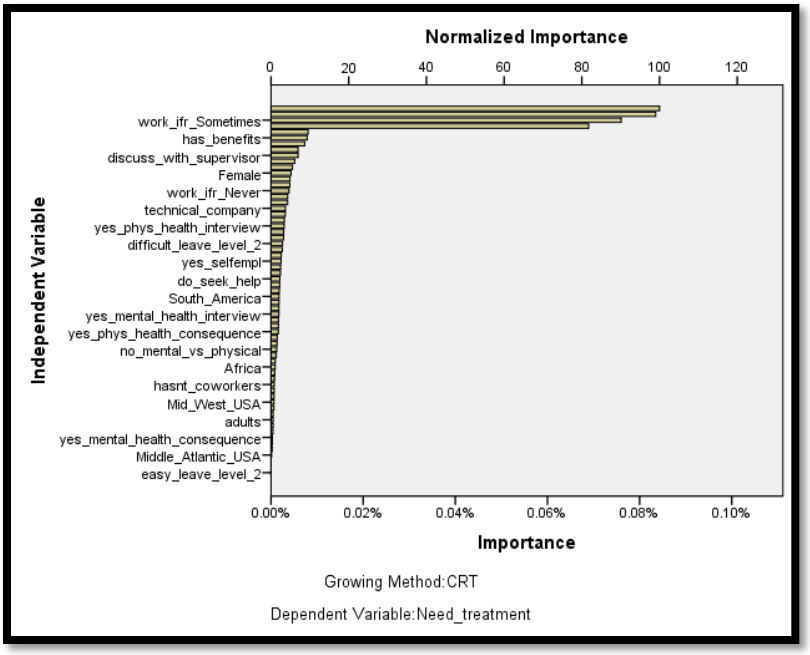


Figure 31 Final Tree attribute importance.

CODE

```
proc import datafile="survey.csv" out=survey replace;
delimiter=',';
getnames=yes;
run;
proc print;
run;
proc freq data=survey;
run;

*Creating Dummy Variables for each Attributes....;
data survey;
set survey;

*================================================================;
*Gender;

if (Gender= "male") then Male = 1;
else if (Gender= "Cis Man") then Male = 1;
else if (Gender= "Cis Male") then Male = 1;
else if (Gender= "male lea") then Male = 1;
else if (Gender= "msle") then Male = 1;
else if (Gender= "Make") then Male = 1;
else if (Gender= "Mal") then Male = 1;
else if (Gender= "Male (CI") then Male = 1;
else if (Gender= "Male") then Male = 1;
else if (Gender= "Male-ish") then Male = 1;
else if (Gender= "Malr") then Male = 1;
else if (Gender= "Man") then Male = 1;
else if (Gender= "m") then Male = 1;
else if (Gender= "maile") then Male = 1;
else if (Gender= "Guy (-is") then Male = 1;
```

```
else if (Gender= "M") then Male = 1;
else if (Gender= "Mail") then Male = 1;
else Male =0;

if (Gender = "Female") then Female = 1;
else if (Gender = "F") then Female = 1;
else if (Gender = "cis-fema") then Female = 1;
else if (Gender = "femail") then Female = 1;
else if (Gender = "Femake") then Female = 1;
else if (Gender = "f") then Female = 1;
else if (Gender = "female") then Female = 1;
else if (Gender = "Cis Fema") then Female = 1;
else if (Gender = "Female (") then Female = 1;
else if (Gender = "Woman") then Female = 1;
else if (Gender = "woman") then Female = 1;
else Female = 0;

if (Gender= "All") then Other = 1;
else if (Gender= "Agender") then Other = 1;
else if (Gender= "fluid") then Other = 1;
else if (Gender= "non-bina") then Other = 1;
else if (Gender= "Neuter") then Other = 1;
else if (Gender= "Trans wo") then Other = 1;
else if (Gender= "ostensib") then Other = 1;
else if (Gender= "Genderqu") then Other = 1;
else if (Gender= "p") then Other = 1;
else if (Gender= "A little") then Other = 1;
else if (Gender= "Androgyn") then Other = 1;
else if (Gender= "Trans-fe") then Other = 1;
else if (Gender= "queer") then Other = 1;
else if (Gender= "queer/sh") then Other = 1;
else if (Gender= "Enby") then Other = 1;
else if (Gender= "Nah") then Other = 1;
else if (Gender= "somethin") then Other = 1;
else Other = 0;

*=========================================================================;

*Age;

if(age > 0 and age <=14) then children = 1;
else children = 0;
if(age >=15 and age <=24) then youth = 1;
else youth = 0;
if(age>=25 and age<=64) then adults = 1;
else adults = 0;
if(age>=65) then senior = 1;
else senior = 0;

*=========================================================================
=;

*Countries by Continents;

*Europian Countries;
if (Country= "Austria") then Europe = 1;
else if (Country= "Croatia") then Europe = 1;
```

```
else if (Country= "Czech Republic") then Europe = 1;
else if (Country= "Germany") then Europe = 1;
else if (Country= "Hungary") then Europe = 1;
else if (Country= "Poland") then Europe = 1;
else if (Country= "Slovenia") then Europe = 1;
else if (Country= "Switzerland") then Europe = 1;
else if (Country= "Bosnia and Herze") then Europe = 1;
else if (Country= "Bulgaria") then Europe = 1;
else if (Country= "Greece") then Europe = 1;
else if (Country= "Romania") then Europe = 1;
else if (Country= "Portugal") then Europe = 1;
else if (Country= "Georgia") then Europe = 1;
else if (Country= "Moldova") then Europe = 1;
else if (Country= "Belgium") then Europe = 1;
else if (Country= "France") then Europe = 1;
else if (Country= "Ireland") then Europe = 1;
else if (Country= "Italy") then Europe = 1;
else if (Country= "Netherlands") then Europe = 1;
else if (Country= "Denmark") then Europe = 1;
else if (Country= "United Kingdom") then Europe = 1;
else if (Country= "Finland") then Europe = 1;
else if (Country= "Sweden") then Europe = 1;
else if (Country= "Norway") then Europe = 1;
else if (Country= "Latvia") then Europe = 1;
else if (Country= "Portugal") then Europe = 1;
else if (Country= "Spain") then Europe = 1;
else Europe = 0;

*Australia and Oceania;
if (Country= "Australia") then Australia_and_Oceania = 1;
else if (Country= "New Zealand") then Australia_and_Oceania = 1;
else Australia_and_Oceania = 0;

*Asian Countries;
if (Country= "India") then Asia = 1;
else if (Country= "Thailand") then Asia = 1;
else if (Country= "Japan") then Asia = 1;
else if (Country= "China") then Asia = 1;
else if (Country= "Singapore") then Asia = 1;
else if (Country= "Philippines") then Asia = 1;
else if (Country= "Russia") then Asia = 1;
else if (Country= "Israel") then Asia = 1;
else Asia = 0;

*North American Countries;
if (Country= "Canada") then North_America = 1;
else if (Country= "Bahamas, The") then North_America = 1;
else if (Country= "Costa Rica") then North_America = 1;
else if (Country= "Mexico") then North_America = 1;
else North_America = 0;

*South American Countries;
if (Country= "Brazil") then South_America = 1;
else if (Country= "Colombia") then South_America = 1;
else if (Country= "Uruguay") then South_America = 1;
else South_America = 0;
```

```
*African Countries;
if (Country= "Nigeria") then Africa = 1;
else if (Country= "Zimbabwe") then Africa= 1;
else if (Country= "South Africa") then Africa= 1;
else Africa = 0;

*================================================================================
========;
*United States by Region;

*West;
if (state= "CA") then West_USA = 1;
else if (state= "NV") then West_USA = 1;
else West_USA = 0;

*North West;
if (state= "MT") then North_West_USA = 1;
else if (state= "ID") then North_West_USA = 1;
else if (state= "WY") then North_West_USA = 1;
else if (state= "WA") then North_West_USA = 1;
else if (state= "OR") then North_West_USA = 1;
else if (state= "AK") then North_West_USA = 1;
else North_West_USA = 0;

*Mid-West;
if (state= "ND") then Mid_West_USA = 1;
else if (state= "SD") then Mid_West_USA = 1;
else if (state= "MN") then Mid_West_USA = 1;
else if (state= "IA") then Mid_West_USA = 1;
else if (state= "NE") then Mid_West_USA = 1;
else if (state= "KS") then Mid_West_USA = 1;
else if (state= "MO") then Mid_West_USA = 1;
else if (state= "MI") then Mid_West_USA = 1;
else if (state= "IN") then Mid_West_USA = 1;
else if (state= "IL") then Mid_West_USA = 1;
else if (state= "OH") then Mid_West_USA = 1;
else if (state= "WI") then Mid_West_USA = 1;
else if (state= "KY") then Mid_West_USA = 1;
else Mid_West_USA = 0;

*South West;
if (state= "TX") then South_West_USA = 1;
else if (state= "OK") then South_West_USA = 1;
else if (state= "UT") then South_West_USA = 1;
else if (state= "CO") then South_West_USA = 1;
else if (state= "AZ") then South_West_USA = 1;
else if (state= "NM") then South_West_USA = 1;
else South_West_USA = 0;

*Mid Atlantic;
if (state= "NY") then Middle_Atlantic_USA = 1;
else if (state= "PA") then Middle_Atlantic_USA = 1;
else if (state= "NJ") then Middle_Atlantic_USA = 1;
else if (state= "WV") then Middle_Atlantic_USA = 1;
else if (state= "MD") then Middle_Atlantic_USA = 1;
else if (state= "DC") then Middle_Atlantic_USA = 1;
else if (state= "DE") then Middle_Atlantic_USA = 1;
```

```
else if (state= "VA") then Middle_Atlantic_USA = 1;
else Middle_Atlantic_USA = 0;

*South East;
if (state= "NC") then South_East_USA = 1;
else if (state= "SC") then South_East_USA = 1;
else if (state= "GA") then South_East_USA = 1;
else if (state= "FL") then South_East_USA = 1;
else if (state= "TN") then South_East_USA = 1;
else if (state= "MS") then South_East_USA = 1;
else if (state= "AL") then South_East_USA = 1;
else if (state= "AR") then South_East_USA = 1;
else if (state= "LA") then South_East_USA = 1;
else South_East_USA = 0;

*North East;
if (state= "ME") then North_East_USA = 1;
else if (state= "VT") then North_East_USA = 1;
else if (state= "NH") then North_East_USA = 1;
else if (state= "MA") then North_East_USA = 1;
else if (state= "CT") then North_East_USA = 1;
else if (state= "RI") then North_East_USA = 1;
else North_East_USA = 0;

*=============================================================================
==;

*Self Employed;

if (self_employed="No") then not_selfempl=1;
else not_selfempl=0;
if(self_employed="Yes") then yes_selfempl=1;
else  yes_selfempl=0;

*=============================================================================
===;

*Family History;

if (family_history="Yes") then Has_family_history=1;
else Has_family_history=0;

*=============================================================================
=;

*Treatment;

if (treatment="Yes") then Need_treatment=1;
else Need_treatment=0;

*=============================================================================
==;

*Work Interfere;

if (work_interfere="Never") then work_ifr_Never=1;
else work_ifr_Never = 0;
```

```
if (work_interfere="Often") then work_ifr_Often=1;
else work_ifr_Often = 0;
if (work_interfere="Rarely") then work_ifr_Rarely=1;
else work_ifr_Rarely = 0;
if (work_interfere="Sometimes") then work_ifr_Sometimes=1;
else work_ifr_Sometimes=0;

*================================================================================
;

*Number of Employee;

if (no_employees="1-5") then no_empl_1=1;
else no_empl_1=0;
if(no_employees="6-25")then no_empl_2=1;
else no_empl_2=0;
if (no_employees="26-100")then no_empl_3=1;
else no_empl_3=0;
if (no_employees="500-1000")then no_empl_4=1;
else no_empl_4=0;
if(no_employees="More than 1000") then no_empl_5=1;
else no_empl_5=0;

*================================================================================
;

*Remote Work;
if(remote_work="Yes") then yes_working_remotely=1;
else yes_working_remotely=0;

*================================================================================
;

*Tech. Company;

if(tech_company="Yes") then technical_company=1;
else technical_company=0;

*================================================================================
;

*Benefits;

if(benefits="No") then has_not_benefits=1;
else has_not_benefits=0;
if(benefits="Yes") then has_benefits=1;
else has_benefits=0;

*================================================================================;

*Care Option;

if(care_options="No") then dont_know_care_options=1;
else dont_know_care_options=0;
if(care_options="Yes") then know_care_options=1;
else know_care_options=0;
```

```
*============================================================================;

*Wellness Program;

if (wellness_program="No") then not_discussed_wellness_program=1;
else not_discussed_wellness_program=0;
if (wellness_program="Yes") then Discussed_wellness_program=1;
else Discussed_wellness_program=0;

*============================================================================;
*Seek Help;

if (seek_help="No") then dont_seek_help=1;
else dont_seek_help=0;
if (seek_help="Yes") then do_seek_help=1;
else do_seek_help=0;

*============================================================================;
*Anonymity;

if (anonymity="No") then does_not_keep_anonymity=1;
else does_not_keep_anonymity=0;
if (anonymity="Yes") then keep_anonymity=1;
else keep_anonymity=0;

*============================================================================;

*Leave;

if (leave="Somewhat difficult") then difficult_leave_level_2= 1;
else difficult_leave_level_2=0;
if (leave="Somewhat easy") then easy_leave_level_2= 1;
else easy_leave_level_2=0;
if (leave="Very difficult") then difficult_leave_level_1= 1;
else difficult_leave_level_1=0;
if (leave="Very easy") then easy_leave_level_1= 1;
else easy_leave_level_1=0;

*============================================================================;

*Mental Health Consequence;

if (mental_health_consequence="No") then no_mental_health_consequence=1;
else no_mental_health_consequence=0;
if (mental_health_consequence="Yes") then yes_mental_health_consequence=1;
else yes_mental_health_consequence=0;

*============================================================================;

*Physical Health;

if (phys_health_consequence="No") then no_phys_health_consequence=1;
else no_phys_health_consequence=0;
if (phys_health_consequence="Yes") then yes_phys_health_consequence=1;
else yes_phys_health_consequence=0;

*============================================================================;
```

```
*CoWorker;

if (coworkers="No") then hasnt_coworkers=1;
else hasnt_coworkers=0;
if (coworkers="Yes") then has_coworkers=1;
else has_coworkers=0;

*================================================================================
;
*Supervisor;

if (supervisor="No") then not_discuss_with_supervisor=1;
else not_discuss_with_supervisor=0;
if (supervisor="Yes") then discuss_with_supervisor=1;
else discuss_with_supervisor=0;

*================================================================================;

*Mental Health Interview;

if (mental_health_interview="No") then no_mental_health_interview=1;
else no_mental_health_interview=0;
if (mental_health_interview="Yes") then yes_mental_health_interview=1;
else yes_mental_health_interview=0;

*================================================================================;

*Physical Health Interview;

if (phys_health_interview="No") then no_phys_health_interview=1;
else no_phys_health_interview=0;
if (phys_health_interview="Yes") then yes_phys_health_interview=1;
else yes_phys_health_interview=0;

*================================================================================;

*Mental vs. Physical;

if (mental_vs_physical="No") then no_mental_vs_physical=1;
else no_mental_vs_physical=0;
if (mental_vs_physical="Yes") then yes_mental_vs_physical=1;
else yes_mental_vs_physical=0;

*================================================================================;

*Observation Consequence;

if (obs_consequence="Yes") then observed_consequence=1;
else observed_consequence=0;
run;

*================================================================================;

*print the datset;
proc print;
run;
```

```
*create corelatino metrix;
proc corr;
var Need_treatment Male Female Other children youth adults senior Europe
Australia_and_Oceania Asia North_America South_America Africa West_USA
North_West_USA Mid_West_USA South_West_USA Middle_Atlantic_USA South_East_USA
North_East_USA not_selfempl yes_selfempl Has_family_history work_ifr_Never
work_ifr_Often work_ifr_Rarely work_ifr_Sometimes no_empl_1 no_empl_2
no_empl_3 no_empl_4 no_empl_5 yes_working_remotely technical_company
has_not_benefits has_benefits dont_know_care_options know_care_options
not_discussed_wellness_program Discussed_wellness_program dont_seek_help
do_seek_help does_not_keep_anonymity keep_anonymity difficult_leave_level_2
easy_leave_level_2 difficult_leave_level_1 easy_leave_level_1
no_mental_health_consequence yes_mental_health_consequence
no_phys_health_consequence yes_phys_health_consequence hasnt_coworkers
has_coworkers not_discuss_with_supervisor discuss_with_supervisor
no_mental_health_interview yes_mental_health_interview
no_phys_health_interview yes_phys_health_interview no_mental_vs_physical
yes_mental_vs_physical observed_consequence ;
run;

*interactive variables Creation;
data survey;
set survey;
interactive1 = know_care_options*has_benefits;
interactive2 = know_care_options*work_ifr_Sometimes;
interactive3 = know_care_options*work_ifr_Often;
interactive4 = know_care_options*Has_family_history;
interactive5 = know_care_options*Female;
interactive6 = has_benefits*work_ifr_Sometimes;
interactive7 = has_benefits*work_ifr_Often;
interactive8 = has_benefits*Has_family_history;
interactive9 = has_benefits*Female;
interactive10 = work_ifr_Sometimes*work_ifr_Often;
interactive11 = work_ifr_Sometimes*Has_family_history;
interactive12 = work_ifr_Sometimes*Female;
interactive13 = work_ifr_Often*Has_family_history;
interactive14 = work_ifr_Often*Female;
interactive15 = Has_family_history*Female;
run;

*Interaction model with interactive variables;
proc reg data=survey;
model Need_treatment=Female Has_family_history work_ifr_Often
work_ifr_Sometimes has_benefits know_care_options interactive1 interactive2
interactive3 interactive4 interactive5 interactive6 interactive7 interactive8
interactive9 interactive10 interactive11 interactive12 interactive13
interactive14 interactive15;
run;

*Create a Frequency table for variable Need_Treatment;
proc freq data=survey;
table Need_treatment;
run;

* Split the data into training and test sets - 70/30;
```

```
proc surveyselect data=survey  out= train_all seed= 27927 samprate = 0.70
outall;
proc print;
run;

* Write the SAS code to create a new field called train_y to assign the
existing value of survey for the training set;
data train_all;
set train_all;
if selected then train_y = Need_treatment;
proc print;
run;

* Model selection - Stepwise;
title'Model selection - Stepwise';
proc logistic;
model Need_treatment(event='1')= Male Female Other children youth adults
senior Europe Australia_and_Oceania Asia North_America South_America Africa
West_USA North_West_USA Mid_West_USA South_West_USA Middle_Atlantic_USA
South_East_USA North_East_USA not_selfempl yes_selfempl Has_family_history
work_ifr_Never work_ifr_Often work_ifr_Rarely work_ifr_Sometimes no_empl_1
no_empl_2 no_empl_3 no_empl_4 no_empl_5 yes_working_remotely
technical_company has_not_benefits has_benefits dont_know_care_options
know_care_options not_discussed_wellness_program Discussed_wellness_program
dont_seek_help do_seek_help does_not_keep_anonymity keep_anonymity
difficult_leave_level_2 easy_leave_level_2 difficult_leave_level_1
easy_leave_level_1 no_mental_health_consequence yes_mental_health_consequence
no_phys_health_consequence yes_phys_health_consequence hasnt_coworkers
has_coworkers not_discuss_with_supervisor discuss_with_supervisor
no_mental_health_interview yes_mental_health_interview
no_phys_health_interview yes_phys_health_interview no_mental_vs_physical
yes_mental_vs_physical observed_consequence / selection = stepwise ;
run;

* Model selection - Forward selection;
title'Model selection - Forward Selection';
proc logistic;
model Need_treatment(event='1')= Male Female Other children youth adults
senior Europe Australia_and_Oceania Asia North_America South_America Africa
West_USA North_West_USA Mid_West_USA South_West_USA Middle_Atlantic_USA
South_East_USA North_East_USA not_selfempl yes_selfempl Has_family_history
work_ifr_Never work_ifr_Often work_ifr_Rarely work_ifr_Sometimes no_empl_1
no_empl_2 no_empl_3 no_empl_4 no_empl_5 yes_working_remotely
technical_company has_not_benefits has_benefits dont_know_care_options
know_care_options not_discussed_wellness_program Discussed_wellness_program
dont_seek_help do_seek_help does_not_keep_anonymity keep_anonymity
difficult_leave_level_2 easy_leave_level_2 difficult_leave_level_1
easy_leave_level_1 no_mental_health_consequence yes_mental_health_consequence
no_phys_health_consequence yes_phys_health_consequence hasnt_coworkers
has_coworkers not_discuss_with_supervisor discuss_with_supervisor
no_mental_health_interview yes_mental_health_interview
no_phys_health_interview yes_phys_health_interview no_mental_vs_physical
yes_mental_vs_physical observed_consequence / selection = forward ;
run;

* Model selection - Backward selection;
title'Model selection - Backward Selection';
```

```
proc logistic;
model Need_treatment(event='1')= Male Female Other children youth adults
senior Europe Australia_and_Oceania Asia North_America South_America Africa
West_USA North_West_USA Mid_West_USA South_West_USA Middle_Atlantic_USA
South_East_USA North_East_USA not_selfempl yes_selfempl Has_family_history
work_ifr_Never work_ifr_Often work_ifr_Rarely work_ifr_Sometimes no_empl_1
no_empl_2 no_empl_3 no_empl_4 no_empl_5 yes_working_remotely
technical_company has_not_benefits has_benefits dont_know_care_options
know_care_options not_discussed_wellness_program Discussed_wellness_program
dont_seek_help do_seek_help does_not_keep_anonymity keep_anonymity
difficult_leave_level_2 easy_leave_level_2 difficult_leave_level_1
easy_leave_level_1 no_mental_health_consequence yes_mental_health_consequence
no_phys_health_consequence yes_phys_health_consequence hasnt_coworkers
has_coworkers not_discuss_with_supervisor discuss_with_supervisor
no_mental_health_interview yes_mental_health_interview
no_phys_health_interview yes_phys_health_interview no_mental_vs_physical
yes_mental_vs_physical observed_consequence/ selection = backward ;
run;

*====================*====================*==============================*
======================*=============*====================================;
*After running forward selection, stepwise and backward we are getting same
resuls for Forward Selection and Stepwise so that will be our MODEL 1;
*also the Backward selection will be MODEL 2;
*NOW, TO COME UP WITH THE BEST MODEL BETWEEN THIS TWO MODEL WE WILL CHECK THE
ACCURACY FOR BOTH THE MODELS;

*===============*=============*==============*=============*===========*====
==========================*======================*==================*=====;

*Model 1 selected from Forward selection and Stepwise (both gives the same
result);

                                    title'Model 1 TEST';
*generate the clasification table --> cut off value compute using training
set;
*compute predicted probability for test set;
proc logistic data= train_all;
model train_y (event='1') = Male Middle_Atlantic_USA Has_family_history
work_ifr_Never work_ifr_Often work_ifr_Rarely work_ifr_Sometimes has_benefits
know_care_options has_coworkers/ ctable pprob = (0.1 to 0.8 by 0.05);
*comute prd. probability for test set;
output out = pred(where=(train_y = .)) p= phat lower= lcl upper=ucl;
run;

*compute pred. y based on cut off value and create the confusion matrix;
data probs;
set pred;
*compute pred. y;
*calculated thresold from specificity and sensitivity values;
if phat>0.4 then pred_y = 1;
else pred_y = 0;
run;
proc print;
run;

*confusion matrix;
```

```
proc freq data = probs;
table Need_treatment*pred_y/ norow nocol nopercent;
run;

*=========================================*=============================
=========*====================================;

*Model 2 selected from Backward selection ;
                                        title'Model 2 TEST';
*generate the clasification table --> cut off value compute using training
set;
*compute predicted probability for test set;
proc logistic data= train_all;
model train_y (event='1') = Female Has_family_history work_ifr_Never
work_ifr_Often work_ifr_Rarely work_ifr_Sometimes has_benefits
know_care_options has_coworkers/ ctable pprob = (0.1 to 0.8 by 0.05);
*comute prd. probability for test set;
output out = pred(where=(train_y = .)) p= phat lower= lcl upper=ucl;
run;

*compute pred. y based on cut off value and create the confusion matrix;
data probs;
set pred;
*compute pred. y;
*calculated thresold from specificity and sensitivity values;
if phat>0.45 then pred_y = 1;
else pred_y = 0;
run;
proc print;
run;

*confusion matrix;
proc freq data = probs;
table Need_treatment*pred_y/ norow nocol nopercent;
run;

*=========================================================================
=====================================================;

*our final model will be Model 2 with the backward selection as it gives more
accuracy(86.73%) when compared to model 1(85.94%);
 title 'Final Regression Model';
proc logistic data= train_all;
model train_y (event='1') = Female Has_family_history work_ifr_Never
work_ifr_Often work_ifr_Rarely work_ifr_Sometimes has_benefits
know_care_options has_coworkers ;
run;
data new;
input train_y Female Has_family_history work_ifr_Never work_ifr_Often
work_ifr_Rarely work_ifr_Sometimes has_benefits know_care_options
has_coworkers ;
datalines;
1 0 1 0 1 1 1 0 1 0
1 1 0 1 1 0 1 1 1 0
;
run;
data pred;
```

```
set new train_all;
proc print data=pred;
run;
proc logistic data=pred;
model train_y (event='1') = Female Has_family_history work_ifr_Never
work_ifr_Often work_ifr_Rarely work_ifr_Sometimes has_benefits
know_care_options has_coworkers;
output out=pred p=phat lower=lcl upper=ucl ;
run;
proc print data=pred;
run;
*=====================*====================*==================*=======;
*influential points;
PROC REG;
TITLE "Outliers and Influential points";
MODEL Need_treatment = Female Has_family_history work_ifr_Never
work_ifr_Often work_ifr_Rarely work_ifr_Sometimes has_benefits
know_care_options has_coworkers /INFLUENCE R;
run;

*Now we will Remove the outliers and Influencial Points;
data pred;
set pred;
if _n_=1257 then delete;
if _n_=828 then delete;
if _n_=651 then delete;
if _n_=603 then delete;
run;

*Residual plots;
proc reg;
model Need_treatment = Female Has_family_history work_ifr_Never
work_ifr_Often work_ifr_Rarely work_ifr_Sometimes has_benefits
know_care_options has_coworkers;
plot student.*(Female Has_family_history work_ifr_Never work_ifr_Often
work_ifr_Rarely work_ifr_Sometimes has_benefits know_care_options
has_coworkers);
plot student.*predicted.;
plot npp.*student.;
run;

*Final Model Without Outliers and Influential Points;
title 'Final Regression Model';
proc logistic data= train_all;
model train_y (event='1') = Female Has_family_history work_ifr_Never
work_ifr_Often work_ifr_Rarely work_ifr_Sometimes has_benefits
know_care_options has_coworkers /stb corrb iplots;
run;

* Boxplot for AGE VS. Treatment Variable.;
data survey;
set survey;
if age>100 then delete;
else if age<4 then delete;
run;
TITLE "Boxplots";
ods graphics off;
```

```
PROC SORT;
BY treatment;
RUN;
PROC BOXPLOT;
PLOT AGE*treatment ;
inset min mean max stddev/
header = 'Overall Statistics'
pos = tm;
insetgroup min mean Q1 Q2 Q3 max range stddev/
header = 'Statistics by Private';
RUN;
```