Project

Categorical Analysis on Mental Health

January 2019

Categorical Analysis on Mental Health

January 2019

1 Introduction

Mental health cases are increasing with fast-paced industry growth. In order to figure out what kind of factors have big impacts on mental health, the categorical analysis was performed on the 2014 mental survey that measures attitudes towards mental health and frequency of mental health disorders in the tech workplace. Let Treatment (represents if people ever thought about mental health treat) as the response variable to perform the categorical analysis. It is important to know what the strongest predictors of mental health illness or certain attitudes towards mental health issue are. In addition, it would be great to predict the response variable individually by using the model we finally choose.

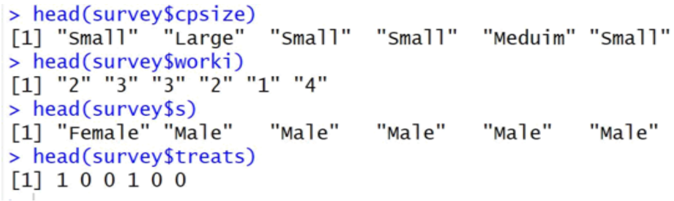
After reviewed the parts above we determined that it would be a good idea for us to go deeper with the data to get an idea of if a person have some probability of considering mental health treatment. The first major thing is that we should use this result to help him or her to deal with the mental health problem. Even if it is not definite if he or she definitely has mental health issue, but we can have some preventive measures by using the data providing the basic information, such as if he or she has family history of mental health issue, what kind of company he or she is working for. The mental health problem can be deal with putting efforts in that direction. So coming up some brand-new mental health problem prediction and caring strategies may lead to better working environment for staff and may lead to a little bit more success for a company in the long-term. The following sections will mention about the methodology and prediction models we used for this topic.

2 Methodology

The original dataset is from Open Sourcing Mental Illness surveyed by OMSIHELP. The data set we used is from a 2014 survey which is raw. So it is supposed to clean the data and assign value and groups in the beginning so that we can have a glance of what kind of data I will process in the next sections. To go deeper with the data set, it will be great to categorize the data into groups and do some visualization work to see if there are some differences on mental health illness between the groups of people. After getting the results of what kind of factors seemly have some impacts on the mental health, I built a logistic regression model and used the random forest method to figure out the importance of each factor.

2.1 Data Cleaning

I cleaned the data and assign value and groups in this section. For gender, I categorized them as female, male, undecided. And for treatment, I just put them into "1","0" as the response variable. When it comes to the work interfere, I also categorized them as 0-4 which represents thinking mental health interfere work to "NA, never, rarely, sometimes and often" degree respectively. For the company size, I put them into small,medium and large so as to do the data visualization more easily.



```
> head(survey$cpsize)
[1] "Small"  "Large"  "Small"  "Small"  "Meduim" "Small"
> head(survey$worki)
[1] "2" "3" "3" "2" "1" "4"
> head(survey$s)
[1] "Female" "Male"   "Male"   "Male"   "Male"   "Male"
> head(survey$treats)
[1] 1 0 0 1 0 0
```

Figure 1: A sample of what the data processed looks like

2.2 Categorical Analysis

Categorical data is data that classifies an observation as belonging to one or more categories. For example, I can divide data into different groups and do some boxplot or histogram plotting to have a better understanding of data set. Stat-graphics includes many procedures for dealing with such data, including modeling procedures contained in the sections on Analysis of Variance, Regression Analysis, and Statistical Process Control. In this part, I should go deeper with the data set and have a big picture of what kind of groups of people consider the mental health treatment more often.

- **Basic Data Analysis**

  In order to go deeper with the data set, I wanted to know some information about the people surveyed. I tried to figure out the age distribution, gender distribution and company size distribution. The age distribution is similar to a normal distribution. The middle-aged people in the data set accounts for the major population. The population of male has much more population in this survey than female. And majority of people in this survey are working for the companies of small size.

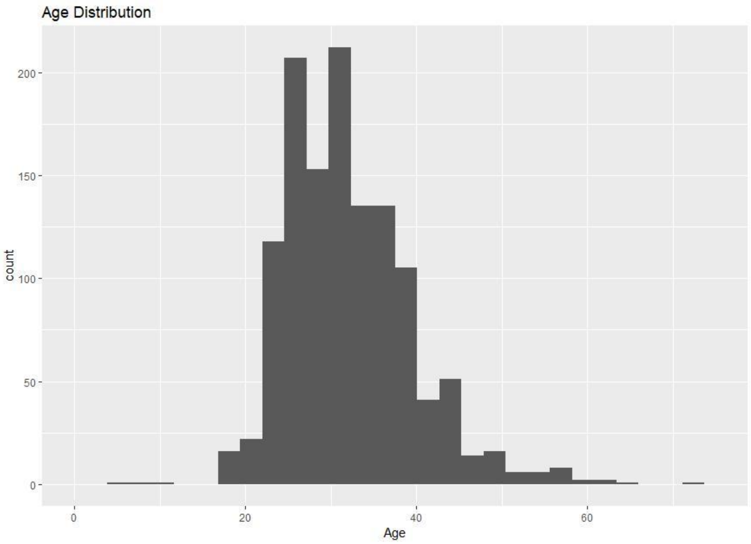


Figure 2: Basic Data Analysis (age,gender,company size) for people

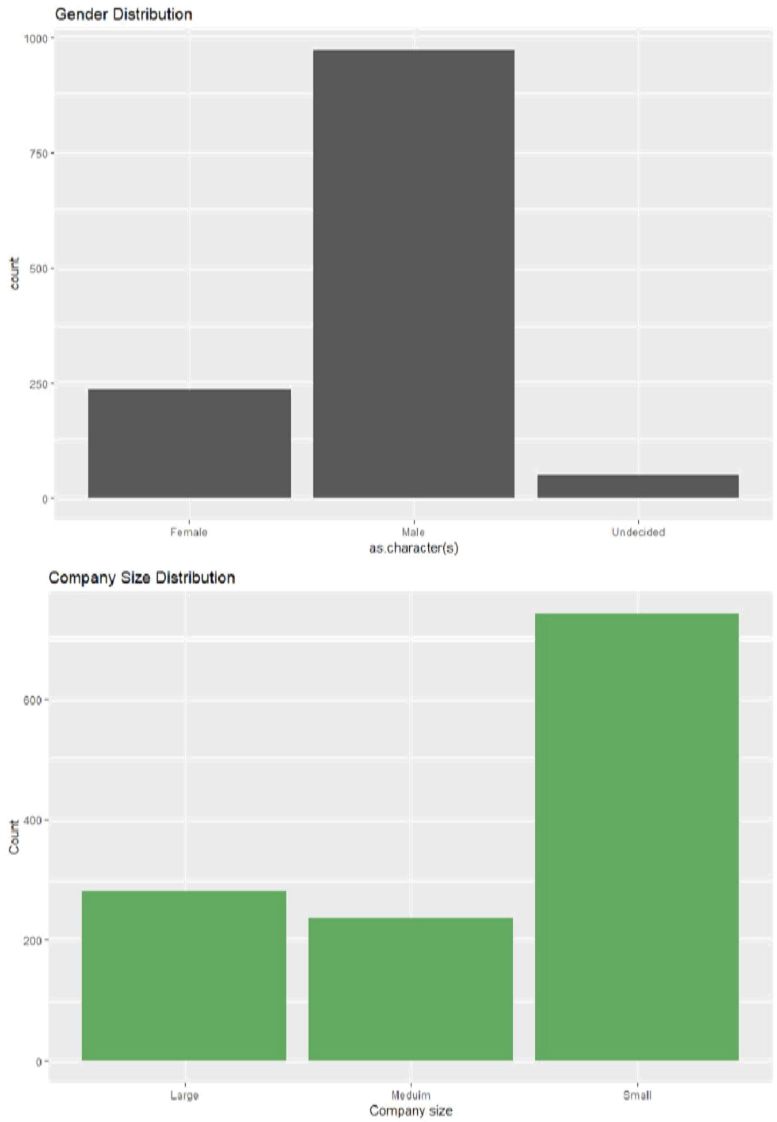


Figure 2: Basic Data Analysis (age,gender,company size) for people

- **Basic mental health analysis**

  In order to go deeper with the data set, I want to do some basic mental health information visualization about the people surveyed. I tried to figure out the probability of mental health illness by groups. There are no big differences among the people from different size of companies. People with family history of mental health problem have higher probability of considering mental health consideration. From the figure of benefits and care options group, people working in companies which offer care options have higher likelihood to think about the mental health problem.

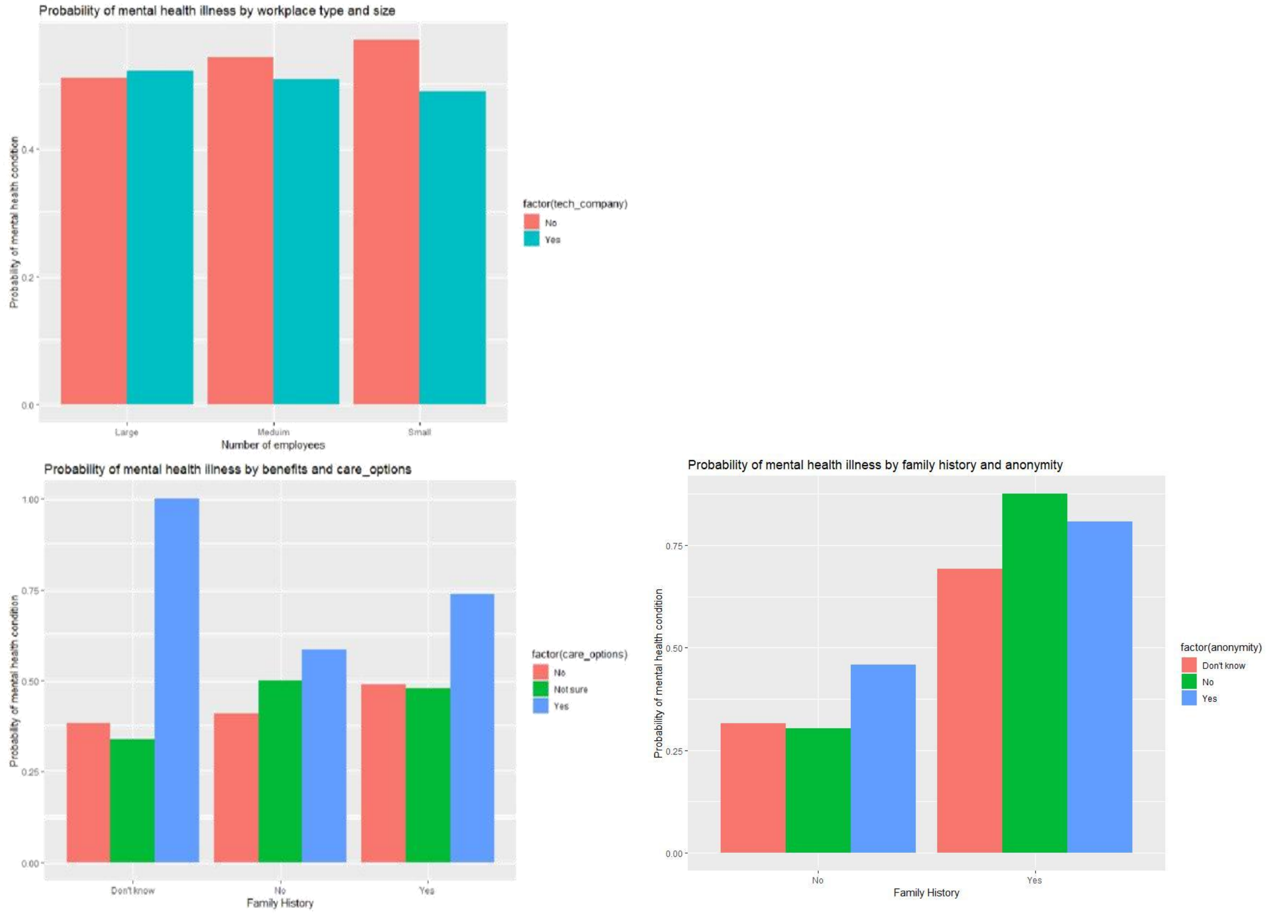


Figure 3: Probability of mental health illness by groups among people

- **Mental health treatment worldwide**
  In order to go have a big picture of the probability of considering mental health treatment among countries, I tried to calculate what is the percentage of people considering a mental health treatment worldwide. For example, we will try to figure out how many people are considering mental treatments inside USA. The graph shows the probability of every country mentioned in the data set con-sidering a mental health treatment. Different color represents different probability.

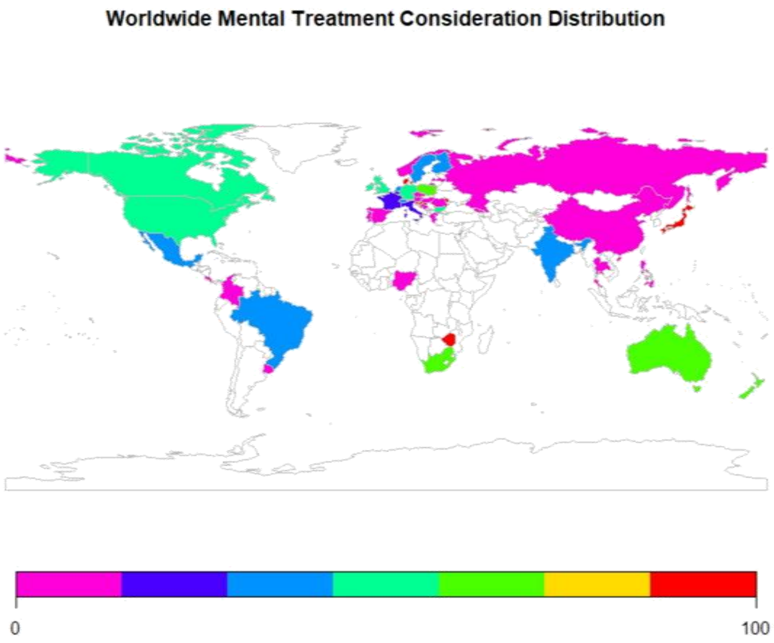


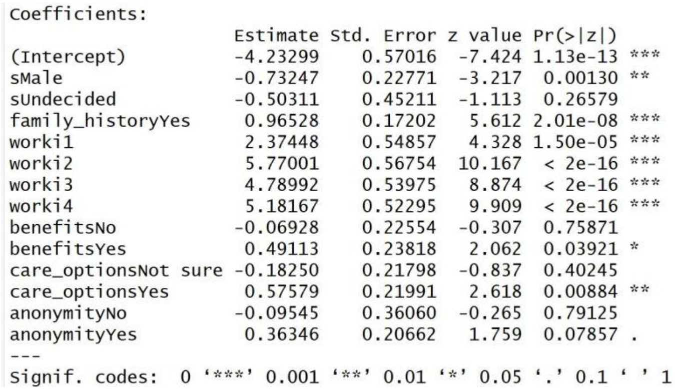
Figure 4: The worldwide mental treatment consideration distribution

2.3 Topic Modeling

Through the figures above, we can see that gender, family history, work interfere, benefits, care options and anonymity can have some impacts on the mental health treatment consideration. In this part, we will build two different models to check the assumption after dividing the data set into train and test set. One is logistic regression model and the other is neural network model. Questions worth exploring will be like: What are the strongest predictors of mental health illness or certain attitudes towards mental health? And which model performs better in prediction. For the first question, I just tried to use random forest model to plot the importance of predictors to have a check. The second one, I tried to plot ROC curve of each model to make the comparison.

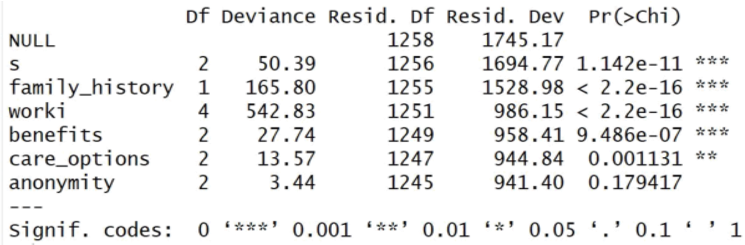
- **Check the variable importance**

  In order to go deeper with the variable importance, I built a regression model at first, using summary() and ANOVA() in R to have a initial check. Then we will use random forest method and varImpPlot() to have a double check, so that I can be sure what kind of factors are important to the prediction model.



```
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)          -4.23299    0.57016  -7.424 1.13e-13 ***
sMale                -0.73247    0.22771  -3.217  0.00130 ** 
sUndecided           -0.50311    0.45211  -1.113  0.26579    
family_historyYes     0.96528    0.17202   5.612 2.01e-08 ***
worki1                2.37448    0.54857   4.328 1.50e-05 ***
worki2                5.77001    0.56754  10.167  < 2e-16 ***
worki3                4.78992    0.53975   8.874  < 2e-16 ***
worki4                5.18167    0.52295   9.909  < 2e-16 ***
benefitsNo           -0.06928    0.22554  -0.307  0.75871    
benefitsYes           0.49113    0.23818   2.062  0.03921 *  
care_optionsNot sure -0.18250    0.21798  -0.837  0.40245    
care_optionsYes       0.57579    0.21991   2.618  0.00884 ** 
anonymityNo          -0.09545    0.36060  -0.265  0.79125    
anonymityYes          0.36346    0.20662   1.759  0.07857 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5: The summary of regression model



```
               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                            1258    1745.17              
s               2    50.39      1256    1694.77 1.142e-11 ***
family_history  1   165.80      1255    1528.98 < 2.2e-16 ***
worki           4   542.83      1251     986.15 < 2.2e-16 ***
benefits        2    27.74      1249     958.41 9.486e-07 ***
care_options    2    13.57      1247     944.84  0.001131 ** 
anonymity       2     3.44      1245     941.40  0.179417    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6: The ANOVA table of regression model

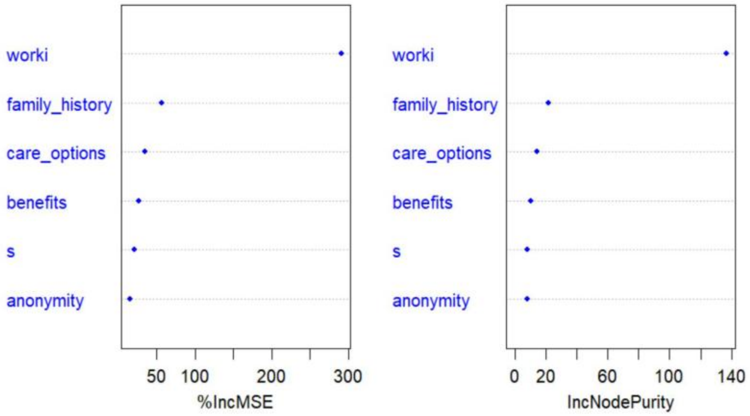


Figure 7: The important factors of regression model

- **Build logistic regression model**
  Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Logistic regression model is used very often in predicting binary response variable. In this part, we should divide the data set into training and testing groups so that we can compare their prediction models for deciding which model is better after fitting the data. The first around 66 percent of the data set is set to be training data and the rest is the test data. After that, we still try to figure out which one is better between the training data and testing data.

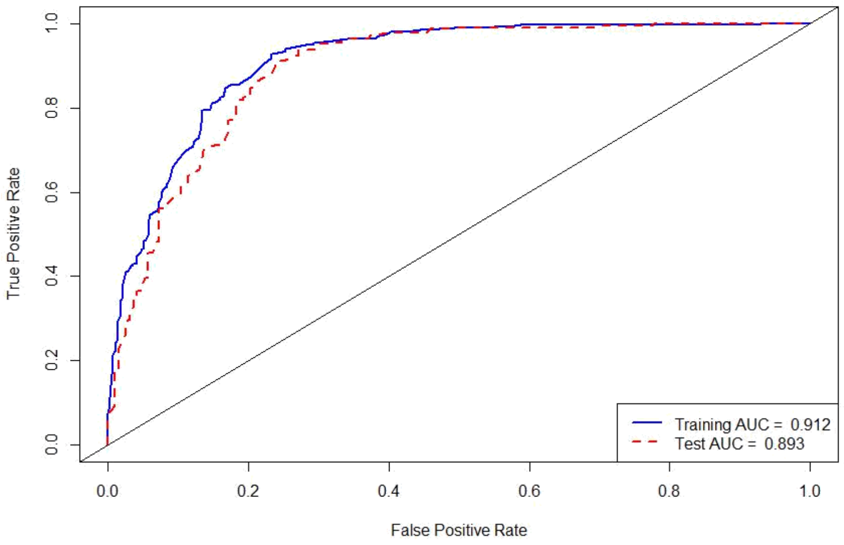


Figure 8: ROC for logistic regression classification

- **Build neural network model**

  Neural network is an interconnected group of nodes, similar to the vast network of neurons in a brain, which is used very often in binary response variable prediction. In this part, in order to compare which prediction method performs better, I tried to use the same data set with logistic regression. What I try to figure out is if neural network prediction model outperforms the logistic regression model. After building training and testing model, I ploted the ROC curves for each model. So that we can see which one is better for fitting the data set.

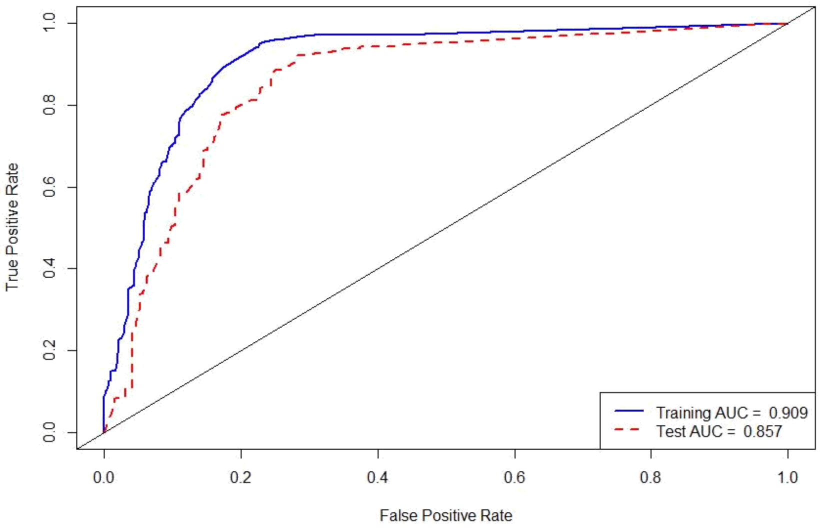
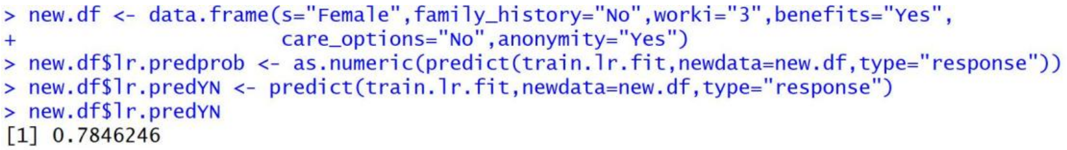


Figure 9: ROC for neural network classification

3 Findings

From the figures above, we can have a big picture of what kind of factors having more important impacts on mental health treatment consideration. The importance order for the predictors is work interfere, family history, care option, benefits and gender And after comparing the AUC value of both logistic regression and neural network prediction model, we can find out that logistic regression training model has higher AUC value which means this model fits the data better.

So I am supposed to use this model to the prediction. What I try to find out through the prediction model is: can we predict the respondent individually for the mental health treatment consideration after knowing the information about all the predictors. So here, I set up an example. There is a new respondent who is a woman, without family history considering mental illness interfere work sometimes, with benefits in workplace without care options in company, and anonymity is yes. So I tried to predict whether she considers the mental health treatment through the information given above. What we can find from the prediction is that the prediction accuracy is high and the result shows that the woman has 78.46 percent of probability to consider the mental health treatment.



```
> new.df <- data.frame(s="Female",family_history="No",worki="3",benefits="Yes",
+                      care_options="No",anonymity="Yes")
> new.df$lr.predprob <- as.numeric(predict(train.lr.fit,newdata=new.df,type="response"))
> new.df$lr.predYN <- predict(train.lr.fit,newdata=new.df,type="response")
> new.df$lr.predYN
[1] 0.7846246
```

Figure 10: Predict whether the woman considers mental health treatment

In addition, the figure of worldwide mental health treatment consideration shows us big diffierences among the countries. It shows that people Japan consider mental health treatment most, and Latin American countries follow behind Japan, other Asian countries seem not consider about mental health treatment.

4 Conclusion

The previous section, I ruled out some variables for our final model selection by visualization. We can see work interfering, family history, care options and benefits definitely matter, as well as bring the mental health issue to a potential employer. Because the person who thinks he/she needs mental health treatment may has mental health issue. So I can put this into consideration for checking if the staff has high probability of having some mental health issues. If he/she has, as leaders of a department or boss, we can have some conversations with the staff, caring about their mental health, which could bring warmness to the whole company and improve the relationship between staff members. In addition, from the figure of worldwide treatment consideration, we can see there are big differences among countries. Japan and Latin American countries have higher probability to consider about the treatment. We can see people from those countries may care about mental health more or may have more mental pressure than other countries.

What we can do for this mental health problem is, when we can know the information about a person when we are working or when we have the responsibility of keeping health and efficient working style for the company, we can have a forehead feeling about a person if he/she has some mental health consideration. For this kind of people, we can put more attention to them and give them more warm courage during the work. So we can make our contributions to the work place environment and world mental health as well.

Appendix

```
# load libraries
#install.packages("ggplot2")
#install.packages("dplyr")
library(ggplot2)
library(dplyr)
# read the data
survey <- read.csv("C:/STONY/Practice/R (No.10)/survey.csv",header = T)
# What does the dataset look like
head(survey)

#################
# Data Cleaning #
#################
# the survey data is a raw data set, so we have to clean it
# categorize gender as female, male, undecided
survey$s<- ifelse(survey$Gender %in% c("F" , "Female", "FEMALE" ,"female", "f"), "Female" ,
                ifelse(survey$Gender %in% c("Male", "male","M","m","MALE","maile"), "Male",
"Undecided"))

# categorize treatment score as 1(yes),0(no)
survey$treats <- ifelse(survey$treatment=="Yes", 1,0)

# categorize work interfere as 1(yes),0(no)
survey$worki <- ifelse(is.na(survey$work_interfere), "0", survey$work_interfere)



# categorize company size as small,meduim and large
survey$cpsize<- ifelse(survey$no_employees %in% c("1-5" , "6-25", "26-100"), "Small",
                    ifelse(survey$no_employees %in% c("100-500","500-1000"), "Meduim",
"Large"))

#####################
# Data Visulization #
```

```
#####################
## NO.1---the basic respondents information visualization ##

# plot the age distribution
age_plot <- ggplot(survey, aes(Age))+
  geom_histogram()+xlim(0,75)+labs(title="Age Distribution")
age_plot

# plot the gender distribution
gend_plot <- ggplot(survey, aes(x=as.character(s)))+
  geom_bar()+labs(title="Gender Distribution")
gend_plot

# plot the company size distribution
#Company size
cmp_plot <- ggplot(survey, aes(x=cpsize))+
  geom_bar(fill="#62AB61")+
  labs(x="Company size", y="Count",
       title="Company Size Distribution")+ theme(legend.position="none")
cmp_plot


## NO.2---the basic mental health information visualization ##

# plot the Probability of mental health illness by workplace type and size
ggplot(survey,aes(x=cpsize,y=treats, fill=factor(tech_company)), color=factor(vs)) +
  stat_summary(fun.y=mean,position=position_dodge(),geom="bar") +
  labs(x = "Number of employees", y = "Probability of mental health condition",
       title = "Probability of mental health illness by workplace type and size")

# plot the probablity of mental health illness by family history and anonymity
ggplot(survey,aes(x=family_history,y=treats, fill=factor(anonymity))) +
  stat_summary(fun.y=mean,position=position_dodge(),geom="bar") +
  labs(x = "Family History", y = "Probability of mental health condition",
       title = "Probability of mental health illness by family history and anonymity")
```

```
# plot the probablity of mental health illness by benefits and care_options
ggplot(survey,aes(x=benefits,y=treats, fill=factor(care_options))) +
  stat_summary(fun.y=mean,position=position_dodge(),geom="bar") +
  labs(x = "Family History", y = "Probability of mental health condition",
       title = "Probability of mental health illness by benefits and care_options")

# plot the worldwide mental treatment consideration distribution
# first calculate the percentage of each conutry poeple thinking about treat
install.packages("rworldmap")
library(rworldmap)
icountry <- group_by(survey, Country)
ic2 <- dplyr::summarise(icountry, add=sum(treats,na.rm=TRUE) , n=n())
ic2$treatpct <- ((ic2$add*100)/ (ic2$n))
ic2 <- arrange(ic2, desc(treatpct))

n <- joinCountryData2Map(ic2, joinCode="NAME", nameJoinColumn="Country")
mapCountryData(n, nameColumnToPlot="treatpct",
               mapTitle="Worldwide Mental Treatment Consideration Distribution",
               catMethod="fixedWidth", colourPalette = "rainbow")



#####################
# Prediction Model  #
#####################
# Preparing regression function for the use in other methods
regresion <- treats~
  s+
  family_history+
  worki+
  benefits+
  care_options+
  anonymity

# build the logistic regression for the model and check the variable importance
```

```
fit <- glm(regresion,family=binomial,data=survey)
summary(fit)
anova(fit,test="Chisq")

# build the random forest model and double check the variable importance
set.seed(1234)
data_fac=survey %>% mutate_if(is.character, as.factor)
rf.fit <- randomForest(regresion,data=data_fac,mtry=3,ntree=1000,importance=TRUE)
varImpPlot(rf.fit,color="blue",pch=20,cex=1.25,main="")



library(lattice)     # lattice plot
library(vcd)         # mosaic plots
library(nnet)        # neural networks
library(ROCR)        # ROC curve objects for binary classification

# user-defined function for plotting ROC curve using ROC objects from ROCR

plot.roc <- function(train.roc,train.auc,test.roc,test.auc) {
  plot(train.roc,col="blue",lty="solid",main="",lwd=2,
      xlab="False Positive Rate",ylab="True Positive Rate")
  plot(test.roc,col="red",lty="dashed",lwd=2,add=TRUE)
  abline(c(0,1))
  train.legend <- paste("Training AUC = ",round(train.auc,digits=3))
  test.legend <- paste("Test AUC = ",round(test.auc,digits=3))
  legend("bottomright",legend=c(train.legend,test.legend),
        lty=c("solid","dashed"),lwd=2,col=c("blue","red"))
}


# !! THIS PART HELPS WITH DEVIDING DATA INTO TRAIN AND TEST !!
set.seed(1234)
survey_origin = data_fac
partition <- sample(nrow(survey_origin),replace=FALSE)
```

```
survey_origin$group <- ifelse(partition<(2/3)*nrow(survey_origin),1,2)
survey_origin$group <- factor(survey_origin$group,levels=c(1,2),labels=c("TRAIN","TEST"))
train.df <- subset(survey_origin,subset=(group=="TRAIN"),
                   select=c("s","family_history","worki","benefits","care_options",
                           "anonymity","treats"))
test.df <-  subset(survey_origin,subset=(group=="TEST"),
                 select=c("s","family_history","worki","benefits","care_options",
                         "anonymity","treats"))
train.df <- na.omit(train.df)
test.df <- na.omit(test.df)
if(length(intersect(rownames(train.df),rownames(test.df)))!= 0) {
  print("\nProblem with partition")
}

##### [1] LOGISTIC REGRESSION #####

train.lr.fit <- glm(regresion,family=binomial,data=train.df)

# area under ROC curve for TRAINING data

train.df$lr.predprob <- predict(train.lr.fit,type="response")
train.lr.pred <- prediction(train.df$lr.predprob,train.df$treats)
train.lr.auc <- as.numeric(performance(train.lr.pred,"auc")@y.values)

# area under ROC curve for TEST data

test.df$lr.predprob <- as.numeric(predict(train.lr.fit,
                                        newdata=test.df,type="response"))
test.lr.pred <- prediction(test.df$lr.predprob,test.df$treats)
test.lr.auc <- as.numeric(performance(test.lr.pred,"auc")@y.values)

# ROC for logistic regression

train.lr.roc <- performance(train.lr.pred,"tpr","fpr")
test.lr.roc <- performance(test.lr.pred,"tpr","fpr")
```

```
plot.roc(train.roc=train.lr.roc,train.auc=train.lr.auc,
        test.roc=test.lr.roc,test.auc=test.lr.auc)

##### [2] NEURAL NETWORKS #####

set.seed(1234)
train.nnet.fit <- nnet(regresion,data=train.df,size=3,decay=0,
                       probability=TRUE,trace=FALSE)

# area under ROC curve for TRAINING data

train.df$nnet.predprob <- as.numeric(predict(train.nnet.fit,newdata=train.df))
train.nnet.prediction <- prediction(train.df$nnet.predprob,train.df$treats)
train.nnet.auc <- as.numeric(performance(train.nnet.prediction,"auc")@y.values)

# area under ROC curve for TEST data

test.df$nnet.predprob <- as.numeric(predict(train.nnet.fit,newdata=test.df))
test.nnet.prediction <- prediction(test.df$nnet.predprob,test.df$treats)
test.nnet.auc <- as.numeric(performance(test.nnet.prediction,"auc")@y.values)

# ROC for neural network classification

train.nnet.roc <- performance(train.nnet.prediction,"tpr","fpr")
test.nnet.roc <- performance(test.nnet.prediction,"tpr","fpr")
plot.roc(train.roc=train.nnet.roc,train.auc=train.nnet.auc,
        test.roc=test.nnet.roc,test.auc=test.nnet.auc)

### logistic regression model outperforms the neural network


################################
## PREDICTION OF NEW CUSTOMER ##
################################
```

```
# !! THIS PART HELPS WITH PREDICTING A NEW RESPONDENT WHEATHER HE CONSIDER
TREATS !!

# For example, there is a new respondent who is a woman,without family history
# considering mental illness interfere work sometimes, with benefits in workplace
# without care_options in company, and anonymity is yes. predict whether she considers
# mental health treats

new.df <- data.frame(s="Female",family_history="No",worki="3",benefits="Yes",
                     care_options="No",anonymity="Yes")
new.df$lr.predprob <- as.numeric(predict(train.lr.fit,newdata=new.df,type="response"))
new.df$lr.predYN <- predict(train.lr.fit,newdata=new.df,type="response")
new.df$lr.predYN
## the result is 1, so the woman will consider about mental health treat
```