**Appendix**

# load libraries

#install.packages("ggplot2")

#install.packages("dplyr")

library(ggplot2)

library(dplyr)

读取数据（数据是从Kaggle上下载下来的，是2014年的OSMI 的mental health in tech survey）

# read the data

survey <- read.csv("C:/STONY/Practice/R (No.10)/survey.csv",header = T)

查看数据长什么样子，有哪些变量

# What does the dataset look like

head(survey)

数据清理环节

#################

# Data Cleaning #

#################

# the survey data is a raw data set, so we have to clean it

因为原始数据是很杂乱的，为了方便做数据分析，将性别，是否考虑过心理治疗，是否认为心理健康对工作影响，所在公司规模大小根据实际情况，进行一个再分类。例如，我们根据公司的规模数量，分成了small,medium,large

# 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"))

数据可视化

#####################

# Data Visulization #

#####################

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

首先我们要查看参与调差的群体的年龄，性别以及所在公司的规模的人员分布情况

# 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

然后，做一与是否考虑过接受心理治疗有关的可视化数据分析

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

首先，我们查看群体中，不同公司类型以及公司大小的人们的考虑过心理治疗的概率分别是多少

# 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")

我们查看群体中，是否有家族心理问题历史以及匿名是否受到保护的考虑过心理治疗的概率分别是多少

# 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")

我们查看群体中，是否有公司福利以及公司是否提供心理健康关怀的考虑过心理治疗的概率分别是多少

# 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")

其次，做一个世界范围内，各个国家考虑过心理治疗的概率大小的分布图，可以从中也分析出一些跟culture相关的考虑心理治疗的原因

# 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")

建立预测模型。初步分析了数据之后，我们了解到从图种可看出的较为重要的影响变量是什么，所以我们将这些变量提取出来，建立预测模型，从而当我们得到一个新人的一系列数据后，我们可以初步预测他是否会考虑心理治疗，从而判断其心理健康状态，从而作出战略性的沟通上的情感关怀。

#####################

# Prediction Model #

#####################

# Preparing regression function for the use in other methods

regresion <- treats~

s+

family\_history+

worki+

benefits+

care\_options+

anonymity

首先，我们建立逻辑回归模型。将是否考虑心理治疗作为结果变量，其他的作为原因变量

查看基本的逻辑回国模型中，我们选取的原因变量中，较为重要的有哪些

# 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")

建立随机森林模型，通过Varlmppot()这个function查看原因变量的重要性排序，横坐标值越大，证明其越重要

# 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

在正式建立预测模型之前，我们需要定义一个画 ROC曲线的公式。ROC曲线的面积越大，证明模型效果越好，该公式可以用于比较后续的预测模型

# 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"))

}

建立正式的预测模型之前，我们需要将数据集分成训练集和测试集，然后根据ROC曲线的面积，来判断训练模型和测试模型的表现

# !! 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")

}

建立逻辑回归的预测模型

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

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

建立出逻辑回归中的训练模型

# 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

建立逻辑回归中的测试模型

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)

建立神经网络的预测模型

##### [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

建立神经网络中的训练模型

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

建立神经网络中的测试模型

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

根据每个模型中的AUC值，以及ROC曲线的面积大小，我们可以看到逻辑回归中的训练模型的表现是最好的。

预测新人是否会考虑心理治疗

################################

## PREDICTION OF NEW CUSTOMER ##

################################

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

这里举了一个例子，例如现在我们有一个没有家族心理疾病，认为心理健康影响工作，公司有福利，不提供心理健康的关心服务，匿名受到保护的女职员，我们要根据以上信息预测她是否会考虑心理治疗，从而判断是否她有心理疾病的倾向

# 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

预测结果显示为0.7 它大于0.5，证明我们预测她会考虑心理治疗，侧面说明她可能有心理疾病

## the result is 1, so the woman will consider about mental health treat