Mid-term Project Code Document

Student Name: Arpita Kaushik

CWID: A20336287

Contents

[**1.** **R Code** 2](#_Toc444523736)

# **R Code**

#----------------------------------------------------------------------------

# Extracting Data & Data Modification

#----------------------------------------------------------------------------

# Extracting the CSV file (originally downloaded from Kaggle) and import it #in R Studio environment

insurance<-read.csv(file="train.csv", head=TRUE, sep=",");

#View the metadata of the input file

# Number of Fields, Type of Fields, Values in fields etc.

str(insurance)

#Change all the Categorical fields as Factors in the dataset

#The reason for doing this step is to change all the categorical data as #factors. Therefore while building the model, categorical values are not #treated as numeric.

insurance$Medical\_History\_1 <- as.factor(insurance$Medical\_History\_1)

insurance$Medical\_History\_2 <- as.factor(insurance$Medical\_History\_2)

insurance$Medical\_History\_3 <- as.factor(insurance$Medical\_History\_3)

insurance$Medical\_History\_4 <- as.factor(insurance$Medical\_History\_4)

insurance$Medical\_History\_5 <- as.factor(insurance$Medical\_History\_5)

insurance$Medical\_History\_6 <- as.factor(insurance$Medical\_History\_6)

insurance$Medical\_History\_7 <- as.factor(insurance$Medical\_History\_7)

insurance$Medical\_History\_8 <- as.factor(insurance$Medical\_History\_8)

insurance$Medical\_History\_9 <- as.factor(insurance$Medical\_History\_9)

insurance$Medical\_History\_10 <- as.factor(insurance$Medical\_History\_10)

insurance$Medical\_History\_11 <- as.factor(insurance$Medical\_History\_11)

insurance$Medical\_History\_12 <- as.factor(insurance$Medical\_History\_12)

insurance$Medical\_History\_13 <- as.factor(insurance$Medical\_History\_13)

insurance$Medical\_History\_14 <- as.factor(insurance$Medical\_History\_14)

insurance$Medical\_History\_15 <- as.factor(insurance$Medical\_History\_15)

insurance$Medical\_History\_16 <- as.factor(insurance$Medical\_History\_16)

insurance$Medical\_History\_17 <- as.factor(insurance$Medical\_History\_17)

insurance$Medical\_History\_18 <- as.factor(insurance$Medical\_History\_18)

insurance$Medical\_History\_19 <- as.factor(insurance$Medical\_History\_19)

insurance$Medical\_History\_20 <- as.factor(insurance$Medical\_History\_20)

insurance$Medical\_History\_21 <- as.factor(insurance$Medical\_History\_21)

insurance$Medical\_History\_22 <- as.factor(insurance$Medical\_History\_22)

insurance$Medical\_History\_23 <- as.factor(insurance$Medical\_History\_23)

insurance$Medical\_History\_24 <- as.factor(insurance$Medical\_History\_24)

insurance$Medical\_History\_25 <- as.factor(insurance$Medical\_History\_25)

insurance$Medical\_History\_26 <- as.factor(insurance$Medical\_History\_26)

insurance$Medical\_History\_27 <- as.factor(insurance$Medical\_History\_27)

insurance$Medical\_History\_28 <- as.factor(insurance$Medical\_History\_28)

insurance$Medical\_History\_29 <- as.factor(insurance$Medical\_History\_29)

insurance$Medical\_History\_30 <- as.factor(insurance$Medical\_History\_30)

insurance$Medical\_History\_31 <- as.factor(insurance$Medical\_History\_31)

insurance$Medical\_History\_32 <- as.factor(insurance$Medical\_History\_32)

insurance$Medical\_History\_33 <- as.factor(insurance$Medical\_History\_33)

insurance$Medical\_History\_34 <- as.factor(insurance$Medical\_History\_34)

insurance$Medical\_History\_35 <- as.factor(insurance$Medical\_History\_35)

insurance$Medical\_History\_36 <- as.factor(insurance$Medical\_History\_36)

insurance$Medical\_History\_37 <- as.factor(insurance$Medical\_History\_37)

insurance$Medical\_History\_38 <- as.factor(insurance$Medical\_History\_38)

insurance$Medical\_History\_39 <- as.factor(insurance$Medical\_History\_39)

insurance$Medical\_History\_40 <- as.factor(insurance$Medical\_History\_40)

insurance$Medical\_History\_41 <- as.factor(insurance$Medical\_History\_41)

insurance$Family\_Hist\_1 <- as.factor(insurance$Family\_Hist\_1)

insurance$InsuredInfo\_1 <- as.factor(insurance$InsuredInfo\_1)

insurance$InsuredInfo\_2 <- as.factor(insurance$InsuredInfo\_2)

insurance$InsuredInfo\_3 <- as.factor(insurance$InsuredInfo\_3)

insurance$InsuredInfo\_4 <- as.factor(insurance$InsuredInfo\_4)

insurance$InsuredInfo\_5 <- as.factor(insurance$InsuredInfo\_5)

insurance$InsuredInfo\_6 <- as.factor(insurance$InsuredInfo\_6)

insurance$InsuredInfo\_7 <- as.factor(insurance$InsuredInfo\_7)

insurance$Insurance\_History\_1 <- as.factor(insurance$Insurance\_History\_1)

insurance$Insurance\_History\_2 <- as.factor(insurance$Insurance\_History\_2)

insurance$Insurance\_History\_3 <- as.factor(insurance$Insurance\_History\_3)

insurance$Insurance\_History\_4 <- as.factor(insurance$Insurance\_History\_4)

insurance$Insurance\_History\_7 <- as.factor(insurance$Insurance\_History\_7)

insurance$Insurance\_History\_8 <- as.factor(insurance$Insurance\_History\_8)

insurance$Insurance\_History\_9 <- as.factor(insurance$Insurance\_History\_9)

insurance$Employment\_Info\_2 <- as.factor(insurance$Employment\_Info\_2)

insurance$Employment\_Info\_3 <- as.factor(insurance$Employment\_Info\_3)

insurance$Employment\_Info\_5 <- as.factor(insurance$Employment\_Info\_5)

insurance$Product\_Info\_1 <- as.factor(insurance$Product\_Info\_1)

insurance$Product\_Info\_2 <- as.factor(insurance$Product\_Info\_2)

insurance$Product\_Info\_3 <- as.factor(insurance$Product\_Info\_3)

insurance$Product\_Info\_5 <- as.factor(insurance$Product\_Info\_5)

insurance$Product\_Info\_6 <- as.factor(insurance$Product\_Info\_6)

insurance$Product\_Info\_7 <- as.factor(insurance$Product\_Info\_7)

insurance$Medical\_Keyword\_1 <- as.factor(insurance$Medical\_Keyword\_1)

insurance$Medical\_Keyword\_2 <- as.factor(insurance$Medical\_Keyword\_2)

insurance$Medical\_Keyword\_3 <- as.factor(insurance$Medical\_Keyword\_3)

insurance$Medical\_Keyword\_4 <- as.factor(insurance$Medical\_Keyword\_4)

insurance$Medical\_Keyword\_5 <- as.factor(insurance$Medical\_Keyword\_5)

insurance$Medical\_Keyword\_6 <- as.factor(insurance$Medical\_Keyword\_6)

insurance$Medical\_Keyword\_7 <- as.factor(insurance$Medical\_Keyword\_7)

insurance$Medical\_Keyword\_8 <- as.factor(insurance$Medical\_Keyword\_8)

insurance$Medical\_Keyword\_9 <- as.factor(insurance$Medical\_Keyword\_9)

insurance$Medical\_Keyword\_10 <- as.factor(insurance$Medical\_Keyword\_10)

insurance$Medical\_Keyword\_11 <- as.factor(insurance$Medical\_Keyword\_11)

insurance$Medical\_Keyword\_12 <- as.factor(insurance$Medical\_Keyword\_12)

insurance$Medical\_Keyword\_13 <- as.factor(insurance$Medical\_Keyword\_13)

insurance$Medical\_Keyword\_14 <- as.factor(insurance$Medical\_Keyword\_14)

insurance$Medical\_Keyword\_15 <- as.factor(insurance$Medical\_Keyword\_15)

insurance$Medical\_Keyword\_16 <- as.factor(insurance$Medical\_Keyword\_16)

insurance$Medical\_Keyword\_17 <- as.factor(insurance$Medical\_Keyword\_17)

insurance$Medical\_Keyword\_18 <- as.factor(insurance$Medical\_Keyword\_18)

insurance$Medical\_Keyword\_19 <- as.factor(insurance$Medical\_Keyword\_19)

insurance$Medical\_Keyword\_20 <- as.factor(insurance$Medical\_Keyword\_20)

#Adding new levels for variables.

#The ‘test’ dataset (the one for which we need to find out the risk level) contains certain records having additional levels for the following variables:

# a) Employment\_Info\_2

# b) Medical\_History\_3

# c) Medical\_History\_33

# d) Medical\_History\_38

# e) Product\_Info\_3

#In order to build a model which fits all, we need to add these new levels to #the ‘train’ dataset as well

insurance$Product\_Info\_3<-factor(insurance$Product\_Info\_3, levels=c(levels(insurance$Product\_Info\_3), "14", "25", "35"))

insurance$Employment\_Info\_2<-factor(insurance$Employment\_Info\_2, levels=c(levels(insurance$Employment\_Info\_2), "8"))

insurance$Medical\_History\_3<-factor(insurance$Medical\_History\_3, levels=c(levels(insurance$Medical\_History\_3), "3"))

insurance$Medical\_History\_33<-factor(insurance$Medical\_History\_33, levels=c(levels(insurance$Medical\_History\_33), "2"))

insurance$Medical\_History\_38<-factor(insurance$Medical\_History\_38, levels=c(levels(insurance$Medical\_History\_38), "3"))

insurance$Product\_Info\_3<-factor(insurance$Product\_Info\_3, levels=c(levels(insurance$Product\_Info\_3), "7", "14", "36", "37"))

levels(insurance$Product\_Info\_3)

insurance$Employment\_Info\_2<-factor(insurance$Employment\_Info\_2, levels=c(levels(insurance$Employment\_Info\_2), "8", "24"))

insurance$Employment\_Info\_2<-factor(insurance$Employment\_Info\_2, levels=c(levels(insurance$Employment\_Info\_2), "8"))

insurance$Response <- as.factor(insurance$Response)

#----------------------------------------------------------------------------

# Exploratory Analysis

#----------------------------------------------------------------------------

# Identifying Predictor Variables

# 1. For Continues Variables

# Barplots: barplots are used to identify if there is an association between # the categorical variable(Response) and continues variable (possible #predictor)

# Following are the variables for which there was a trend noticed

# a) BMI

# b) Employment\_Info\_6

boxplot(BMI~Response, data=insurance, xlab="Response", ylab="BMI")

boxplot(Employment\_Info\_6~Response, data=insurance, xlab="Response", ylab="Employment Info6")

# 2. For Categorical Variables

# Frequency table – a frequency table determines association between two #categorical variables. In this scenario we have 43 variables which denoted #an association with the response variable

# Generate Relative Frequency table for the Categorical variable

# which will be included in Multinomial Regression

prop.table(table(insurance$Response, insurance$Product\_Info\_3),2)

prop.table(table(insurance$Response, insurance$Employment\_Info\_2),2)

prop.table(table(insurance$Response, insurance$InsuredInfo\_2),2)

prop.table(table(insurance$Response, insurance$InsuredInfo\_5),2)

prop.table(table(insurance$Response, insurance$InsuredInfo\_7),2)

prop.table(table(insurance$Response, insurance$Insurance\_History\_2),2)

prop.table(table(insurance$Response, insurance$Medical\_History\_2),2)

prop.table(table(insurance$Response, insurance$Medical\_History\_4),2)

prop.table(table(insurance$Response, insurance$Medical\_History\_5),2)

prop.table(table(insurance$Response, insurance$Medical\_History\_6),2)

prop.table(table(insurance$Response, insurance$Medical\_History\_16),2)

prop.table(table(insurance$Response, insurance$Medical\_History\_17),2)

prop.table(table(insurance$Response, insurance$Medical\_History\_20),2)

prop.table(table(insurance$Response, insurance$Medical\_History\_23),2)

prop.table(table(insurance$Response, insurance$Medical\_History\_29),2)

prop.table(table(insurance$Response, insurance$Medical\_History\_30),2)

prop.table(table(insurance$Response, insurance$Medical\_History\_31),2)

prop.table(table(insurance$Response, insurance$Medical\_History\_33),2)

prop.table(table(insurance$Response, insurance$Medical\_History\_38),2)

prop.table(table(insurance$Response, insurance$Medical\_History\_39),2)

prop.table(table(insurance$Response, insurance$Medical\_History\_40),2)

prop.table(table(insurance$Response, insurance$Medical\_Keyword\_1),2)

prop.table(table(insurance$Response, insurance$Medical\_Keyword\_3),2)

prop.table(table(insurance$Response, insurance$Medical\_Keyword\_4),2)

prop.table(table(insurance$Response, insurance$Medical\_Keyword\_9),2)

prop.table(table(insurance$Response, insurance$Medical\_Keyword\_10),2)

prop.table(table(insurance$Response, insurance$Medical\_Keyword\_13),2)

prop.table(table(insurance$Response, insurance$Medical\_Keyword\_14),2)

prop.table(table(insurance$Response, insurance$Medical\_Keyword\_15),2)

prop.table(table(insurance$Response, insurance$Medical\_Keyword\_16),2)

prop.table(table(insurance$Response, insurance$Medical\_Keyword\_19),2)

prop.table(table(insurance$Response, insurance$Medical\_Keyword\_22),2)

prop.table(table(insurance$Response, insurance$Medical\_Keyword\_23),2)

prop.table(table(insurance$Response, insurance$Medical\_Keyword\_24),2)

prop.table(table(insurance$Response, insurance$Medical\_Keyword\_25),2)

prop.table(table(insurance$Response, insurance$Medical\_Keyword\_27),2)

prop.table(table(insurance$Response, insurance$Medical\_Keyword\_31),2)

prop.table(table(insurance$Response, insurance$Medical\_Keyword\_33),2)

prop.table(table(insurance$Response, insurance$Medical\_Keyword\_35),2)

prop.table(table(insurance$Response, insurance$Medical\_Keyword\_36),2)

prop.table(table(insurance$Response, insurance$Medical\_Keyword\_38),2)

prop.table(table(insurance$Response, insurance$Medical\_Keyword\_43),2)

prop.table(table(insurance$Response, insurance$Medical\_Keyword\_46),2)

prop.table(table(insurance$Response, insurance$Medical\_Keyword\_48),2)

# 3. Running Chi-square test to understand association

# Chi-square test helps to identify if there is a significant level of #relation between the two variable or the null hypothesis can be rejected

# if the value of Chi-square is lower than 0.05 then we can reject the null #hypothesis

chisq.test(table(insurance$Response, insurance$InsuredInfo\_2))

chisq.test(table(insurance$Response, insurance$InsuredInfo\_5))

chisq.test(table(insurance$Response, insurance$InsuredInfo\_7))

chisq.test(table(insurance$Response, insurance$Medical\_History\_4))

chisq.test(table(insurance$Response, insurance$Insurance\_History\_2))

chisq.test(table(insurance$Response, insurance$Medical\_History\_5))

chisq.test(table(insurance$Response, insurance$Medical\_History\_23))

chisq.test(table(insurance$Response, insurance$Medical\_History\_31))

chisq.test(table(insurance$Response, insurance$Medical\_History\_38))

# Treating missing values

ins.family.hist2.miss <-insurance[insurance$Family\_Hist\_2 %in% c(NA), ]

ins.family.hist3.miss <-insurance[insurance$Family\_Hist\_3 %in% c(NA), ]

ins.family.hist4.miss <-insurance[insurance$Family\_Hist\_4 %in% c(NA), ]

ins.family.hist5.miss <-insurance[insurance$Family\_Hist\_5 %in% c(NA), ]

ins.Employment.Info6.miss <-insurance[insurance$Employment\_Info\_6 %in% c(NA), ]

ins.family.hist2.prsnt <-insurance[!(insurance$Family\_Hist\_2 %in% c(NA)), ]

ins.family.hist3.prsnt <-insurance[!(insurance$Family\_Hist\_3 %in% c(NA)), ]

ins.family.hist4.prsnt <-insurance[!(insurance$Family\_Hist\_4 %in% c(NA)), ]

ins.family.hist5.prsnt <-insurance[!(insurance$Family\_Hist\_5 %in% c(NA)), ]

ins.Employment.Info6.prsnt <-insurance[!(insurance$Employment\_Info\_6 %in% c(NA)), ]

mean.family2<-mean(ins.family.hist2.prsnt$Family\_Hist\_2)

mean.family3<-mean(ins.family.hist3.prsnt$Family\_Hist\_3)

mean.family4<-mean(ins.family.hist4.prsnt$Family\_Hist\_4)

mean.family5<-mean(ins.family.hist5.prsnt$Family\_Hist\_5)

mean.Emp.Info6<-mean(ins.Employment.Info6.prsnt$Employment\_Info\_6)

insurance$Family\_Hist\_2[is.na(insurance$Family\_Hist\_2)]<- mean.family2

insurance$Family\_Hist\_3[is.na(insurance$Family\_Hist\_3)]<- mean.family3

insurance$Family\_Hist\_4[is.na(insurance$Family\_Hist\_4)]<- mean.family4

insurance$Family\_Hist\_5[is.na(insurance$Family\_Hist\_5)]<- mean.family5

insurance$Employment\_Info\_6[is.na(insurance$Employment\_Info\_6)]<- mean.Emp.Info6

#----------------------------------------------------------------------------

# Multinomial Logistic Regression

#----------------------------------------------------------------------------

# Build a multinomial model for the insurance data

library(nnet)

test <- multinom(Response ~ Employment\_Info\_6+BMI+Product\_Info\_3+Employment\_Info\_2+InsuredInfo\_2+InsuredInfo\_5+InsuredInfo\_7+Insurance\_History\_2+Medical\_History\_4+Medical\_History\_5+Medical\_History\_6+Medical\_History\_16+Medical\_History\_17+Medical\_History\_20+Medical\_History\_23+Medical\_History\_29+Medical\_History\_30+Medical\_History\_31+Medical\_History\_33+Medical\_History\_38+Medical\_History\_39+Medical\_History\_40+Medical\_Keyword\_1+Medical\_Keyword\_3+Medical\_Keyword\_4+Medical\_Keyword\_9+Medical\_Keyword\_10+Medical\_Keyword\_13+Medical\_Keyword\_14+Medical\_Keyword\_15+Medical\_Keyword\_16+Medical\_Keyword\_19+Medical\_Keyword\_22+Medical\_Keyword\_23+Medical\_Keyword\_24+Medical\_Keyword\_25+Medical\_Keyword\_27+Medical\_Keyword\_31+Medical\_Keyword\_33+Medical\_Keyword\_35+Medical\_Keyword\_36+Medical\_Keyword\_38+Medical\_Keyword\_43+Medical\_Keyword\_46+Medical\_Keyword\_48,

data = insurance, MaxNWts = 8000, maxit=900)

summary()

#----------------------------------------------------------------------------

# Predicting values in the Test Data

#----------------------------------------------------------------------------

#Extracting the test file and import it in R Studio environment

test.data<-read.csv(file="test1.csv", head=TRUE, sep=",");

#View the metadata of the input file

str(test.data)

#Change all the Categorical fields as Factors in the dataset

# as the model treats the following as categorical. The following variables #needs to be converted into Categorical for ‘Test’ data as well

test.data$Medical\_History\_1 <- as.factor(test.data$Medical\_History\_1)

test.data$Medical\_History\_2 <- as.factor(test.data$Medical\_History\_2)

test.data$Medical\_History\_3 <- as.factor(test.data$Medical\_History\_3)

test.data$Medical\_History\_4 <- as.factor(test.data$Medical\_History\_4)

test.data$Medical\_History\_5 <- as.factor(test.data$Medical\_History\_5)

test.data$Medical\_History\_6 <- as.factor(test.data$Medical\_History\_6)

test.data$Medical\_History\_7 <- as.factor(test.data$Medical\_History\_7)

test.data$Medical\_History\_8 <- as.factor(test.data$Medical\_History\_8)

test.data$Medical\_History\_9 <- as.factor(test.data$Medical\_History\_9)

test.data$Medical\_History\_10 <- as.factor(test.data$Medical\_History\_10)

test.data$Medical\_History\_11 <- as.factor(test.data$Medical\_History\_11)

test.data$Medical\_History\_12 <- as.factor(test.data$Medical\_History\_12)

test.data$Medical\_History\_13 <- as.factor(test.data$Medical\_History\_13)

test.data$Medical\_History\_14 <- as.factor(test.data$Medical\_History\_14)

test.data$Medical\_History\_15 <- as.factor(test.data$Medical\_History\_15)

test.data$Medical\_History\_16 <- as.factor(test.data$Medical\_History\_16)

test.data$Medical\_History\_17 <- as.factor(test.data$Medical\_History\_17)

test.data$Medical\_History\_18 <- as.factor(test.data$Medical\_History\_18)

test.data$Medical\_History\_19 <- as.factor(test.data$Medical\_History\_19)

test.data$Medical\_History\_20 <- as.factor(test.data$Medical\_History\_20)

test.data$Medical\_History\_21 <- as.factor(test.data$Medical\_History\_21)

test.data$Medical\_History\_22 <- as.factor(test.data$Medical\_History\_22)

test.data$Medical\_History\_23 <- as.factor(test.data$Medical\_History\_23)

test.data$Medical\_History\_24 <- as.factor(test.data$Medical\_History\_24)

test.data$Medical\_History\_25 <- as.factor(test.data$Medical\_History\_25)

test.data$Medical\_History\_26 <- as.factor(test.data$Medical\_History\_26)

test.data$Medical\_History\_27 <- as.factor(test.data$Medical\_History\_27)

test.data$Medical\_History\_28 <- as.factor(test.data$Medical\_History\_28)

test.data$Medical\_History\_29 <- as.factor(test.data$Medical\_History\_29)

test.data$Medical\_History\_30 <- as.factor(test.data$Medical\_History\_30)

test.data$Medical\_History\_31 <- as.factor(test.data$Medical\_History\_31)

test.data$Medical\_History\_32 <- as.factor(test.data$Medical\_History\_32)

test.data$Medical\_History\_33 <- as.factor(test.data$Medical\_History\_33)

test.data$Medical\_History\_34 <- as.factor(test.data$Medical\_History\_34)

test.data$Medical\_History\_35 <- as.factor(test.data$Medical\_History\_35)

test.data$Medical\_History\_36 <- as.factor(test.data$Medical\_History\_36)

test.data$Medical\_History\_37 <- as.factor(test.data$Medical\_History\_37)

test.data$Medical\_History\_38 <- as.factor(test.data$Medical\_History\_38)

test.data$Medical\_History\_39 <- as.factor(test.data$Medical\_History\_39)

test.data$Medical\_History\_40 <- as.factor(test.data$Medical\_History\_40)

test.data$Medical\_History\_41 <- as.factor(test.data$Medical\_History\_41)

test.data$Family\_Hist\_1 <- as.factor(test.data$Family\_Hist\_1)

test.data$InsuredInfo\_1 <- as.factor(test.data$InsuredInfo\_1)

test.data$InsuredInfo\_2 <- as.factor(test.data$InsuredInfo\_2)

test.data$InsuredInfo\_3 <- as.factor(test.data$InsuredInfo\_3)

test.data$InsuredInfo\_4 <- as.factor(test.data$InsuredInfo\_4)

test.data$InsuredInfo\_5 <- as.factor(test.data$InsuredInfo\_5)

test.data$InsuredInfo\_6 <- as.factor(test.data$InsuredInfo\_6)

test.data$InsuredInfo\_7 <- as.factor(test.data$InsuredInfo\_7)

test.data$Test.data\_History\_1 <- as.factor(test.data$Test.data\_History\_1)

test.data$Test.data\_History\_2 <- as.factor(test.data$Test.data\_History\_2)

test.data$Test.data\_History\_3 <- as.factor(test.data$Test.data\_History\_3)

test.data$Test.data\_History\_4 <- as.factor(test.data$Test.data\_History\_4)

test.data$Test.data\_History\_7 <- as.factor(test.data$Test.data\_History\_7)

test.data$Test.data\_History\_8 <- as.factor(test.data$Test.data\_History\_8)

test.data$Test.data\_History\_9 <- as.factor(test.data$Test.data\_History\_9)

test.data$Employment\_Info\_2 <- as.factor(test.data$Employment\_Info\_2)

test.data$Employment\_Info\_3 <- as.factor(test.data$Employment\_Info\_3)

test.data$Employment\_Info\_5 <- as.factor(test.data$Employment\_Info\_5)

test.data$Product\_Info\_1 <- as.factor(test.data$Product\_Info\_1)

test.data$Product\_Info\_2 <- as.factor(test.data$Product\_Info\_2)

test.data$Product\_Info\_3 <- as.factor(test.data$Product\_Info\_3)

test.data$Product\_Info\_5 <- as.factor(test.data$Product\_Info\_5)

test.data$Product\_Info\_6 <- as.factor(test.data$Product\_Info\_6)

test.data$Product\_Info\_7 <- as.factor(test.data$Product\_Info\_7)

test.data$Medical\_Keyword\_1 <- as.factor(test.data$Medical\_Keyword\_1)

test.data$Medical\_Keyword\_2 <- as.factor(test.data$Medical\_Keyword\_2)

test.data$Medical\_Keyword\_3 <- as.factor(test.data$Medical\_Keyword\_3)

test.data$Medical\_Keyword\_4 <- as.factor(test.data$Medical\_Keyword\_4)

test.data$Medical\_Keyword\_5 <- as.factor(test.data$Medical\_Keyword\_5)

test.data$Medical\_Keyword\_6 <- as.factor(test.data$Medical\_Keyword\_6)

test.data$Medical\_Keyword\_7 <- as.factor(test.data$Medical\_Keyword\_7)

test.data$Medical\_Keyword\_8 <- as.factor(test.data$Medical\_Keyword\_8)

test.data$Medical\_Keyword\_9 <- as.factor(test.data$Medical\_Keyword\_9)

test.data$Medical\_Keyword\_10 <- as.factor(test.data$Medical\_Keyword\_10)

test.data$Medical\_Keyword\_11 <- as.factor(test.data$Medical\_Keyword\_11)

test.data$Medical\_Keyword\_12 <- as.factor(test.data$Medical\_Keyword\_12)

test.data$Medical\_Keyword\_13 <- as.factor(test.data$Medical\_Keyword\_13)

test.data$Medical\_Keyword\_14 <- as.factor(test.data$Medical\_Keyword\_14)

test.data$Medical\_Keyword\_15 <- as.factor(test.data$Medical\_Keyword\_15)

test.data$Medical\_Keyword\_16 <- as.factor(test.data$Medical\_Keyword\_16)

test.data$Medical\_Keyword\_17 <- as.factor(test.data$Medical\_Keyword\_17)

test.data$Medical\_Keyword\_18 <- as.factor(test.data$Medical\_Keyword\_18)

test.data$Medical\_Keyword\_19 <- as.factor(test.data$Medical\_Keyword\_19)

test.data$Medical\_Keyword\_20 <- as.factor(test.data$Medical\_Keyword\_20)

test.data$Insurance\_History\_2 <- as.factor(test.data$Insurance\_History\_2)

test.data$Insurance\_History\_3 <- as.factor(test.data$Insurance\_History\_3)

test.data$Insurance\_History\_4 <- as.factor(test.data$Insurance\_History\_4)

test.data$Insurance\_History\_7 <- as.factor(test.data$Insurance\_History\_7)

test.data$Insurance\_History\_8 <- as.factor(test.data$Insurance\_History\_8)

test.data$Insurance\_History\_9 <- as.factor(test.data$Insurance\_History\_9)

# treating missing values for ‘Test’ data set

test.Employment.Info6.miss <-test.data[test.data$Employment\_Info\_6 %in% c(NA), ]

test.Employment.Info6.prsnt <-test.data[!(test.data$Employment\_Info\_6 %in% c(NA)), ]

mean.tst.Emp.Info6<-mean(test.Employment.Info6.prsnt$Employment\_Info\_6)

test.data$Employment\_Info\_6[is.na(test.data$Employment\_Info\_6)]<- mean.tst.Emp.Info6

# Predicting the probability for ‘Test’ dataset

# ‘predict’ function is used to predict values for the test data set using #the model that was built on ‘Train’ data set.

# the predicted values are rounded off to get a more readable value for each risk level

test.response<-predict (test, newdata=test.data, "probs")

test.resp.round<-round(test.response, digit=2)

test.resp.round<-as.data.frame(test.resp.round)

str(test.resp.round)

# The max value, among the 8 level of risk, for each of record type is #calculated and stored as predicted risk level for that record

for (i in 1:nrow(test.resp.round))

{

test.resp.round$maximum[i] <- max(test.resp.round[i,1:8])

test.resp.round$response[i]<-which.max(test.resp.round[i,1:8])

}

test.data$Response<-test.resp.round$response

# The Test data with predicted ‘Risk’ level is exported to an excel using #read.xlsx comman

library(xlsx)

write.xlsx(test.data, "Test\_response.xlsx")