# import relevant packages and headers

source("/Users/sam/All-Program/App/IIT-Projects/Insurance-Defaulters/Data\_Cleaner.R")

library(caret)

library(dummy)

################# Data File Location ##################

# Load the Data

datam = read.csv2("/Users/sam/All-Program/App-DataSet/IIT-Projects/Insurance-Defaulters/LoanStats3a.csv", header=TRUE, sep=",", skip=1)

varm = read.csv2("/Users/sam/All-Program/App-DataSet/IIT-Projects/Insurance-Defaulters/Variables-Sheet1.csv", header=TRUE, sep=",")

#clean\_data = read.csv2("/Users/sam/All-Program/App-DataSet/IIT-Projects/Insurance-Defaulters/LoanStats3a.csv", header=TRUE, sep=",", skip=1)

################# Data File Location ##################

# Remove all the rows with response as "current"

unique(datam$loan\_status)

dat <- subset(datam, loan\_status == c("Charged Off","Fully Paid"))

# Check How many are null or NA

summary(dat)

# Data on Status:

feature\_set1<-subset(varm, Status=="1")

feature\_set1$LoanStatNew

feature\_set0<-subset(varm, Status=="0")

feature\_set0$LoanStatNew

feature\_set2<-subset(varm, Status=="2")

feature\_set2$LoanStatNew

# Get the data for the given feature set.

intrsct<-intersect(feature\_set1$LoanStatNew, colnames(dat))

# Column names that are present in varm but not a part of datam column\_names

setdiff(feature\_set1$LoanStatNew,intrsct)

# Finding the subset of datam with features decided as initial analysis

dat\_1<-subset(dat, select = intrsct)

head(dat\_1)

############ Clean columns 15-30 ##########

# Remove Columns:

col\_names <- c("id")

x\_clean <- clean(dat\_1)

x\_clean <- drop\_columns(x\_clean,col\_names)

dim(datam)

dim(dat)

dim(x\_clean)

# Select sample random data from the Dataframe

sample\_x\_clean <- x\_clean[sample(nrow(x\_clean), 100), ]

######### creating Dummy Data ##########

sample\_x\_clean <- var\_cleaner(sample\_x\_clean)

grade <- predict(dummyVars(~grade, data = sample\_x\_clean), newdata = sample\_x\_clean)

application\_type <- predict(dummyVars(~application\_type, data = sample\_x\_clean), newdata = sample\_x\_clean)

home\_ownership <- predict(dummyVars(~home\_ownership, data = sample\_x\_clean), newdata = sample\_x\_clean)

pymnt\_plan <- predict(dummyVars(~pymnt\_plan, data = sample\_x\_clean), newdata = sample\_x\_clean)

grade <- predict(dummyVars(~grade, data = sample\_x\_clean), newdata = sample\_x\_clean)

######### Check by fiiting a simple model #############

loan\_status1 <- predict(dummyVars(~loan\_status, data = sample\_x\_clean), newdata = sample\_x\_clean)

# Fit a simple linear Model

fit0<-lm(sample\_x\_clean$loan\_status.new~sample\_x\_clean$annual\_inc,data=sample\_x\_clean)

# Fit a simple logit model

glm.out = glm(sample\_x\_clean$loan\_status.new~sample\_x\_clean$annual\_inc, family=binomial(logit), data=sample\_x\_clean)

glm.out = glm(sample\_x\_clean$loan\_status.new~sample\_x\_clean$dti, family=binomial(logit), data=sample\_x\_clean)

summary(glm.out)