Prepare support vector machines model for classifying the area under fire for forestfires data

##import the data set##

library(readr)

data <- read.csv(file.choose())

View(data)

str(data)

attach(forest)

##from the view data set we can see that data some character factor variables which will be not useful to your model bulding

## so we will remove two variable that is the day

forest <- data[,-c(1,2)]

View(forest)

##EDA

library(DataExplorer)

plot\_str(forest)

##to known the variables in the data set

plot\_missing(forest)

##knowning missing values in the data set

plot\_histogram(forest)

## knowning the continous data

plot\_density(forest)

##representing the data by plots

plot\_correlation(forest, type = 'continuous','area')

## from the plot you can see that the temperature is one variable which is effecting the forest area fire the rh is negatively correlated

plot\_bar(data)

##plotting the data of the categorical values

create\_report(forest)

##we should normalize all the variables

normalize <- function(x){

return((x-min(x))/(max(x)-min(x)))

}

forest$FFMC <- normalize(forest$FFMC)

forest$DMC <- normalize(forest$DMC)

forest$DC <- normalize(forest$DC)

forest$ISI <- normalize(forest$ISI)

forest$temp <- normalize(forest$temp)

forest$RH <- normalize(forest$RH)

forest$wind <- normalize(forest$wind)

forest$rain <- normalize(forest$rain)

## we will change the area variable area by into small and large large is >5 small is 5<

forest$size <- NULL

forest$size <- factor(ifelse(forest$area < 5, 1, 0),

labels = c("small", "large"))

##Splitting the data set into train and test

train <- sample(x = nrow(forest), size = 400, replace = FALSE)

##model building

library(kernlab)

library(caret)

##kernel = "polydot"

m.poly <- ksvm(size ~ temp + RH + wind + rain+FFMC+DMC+ISI+DC,

data = forest[train, ],

kernel = "polydot", C = 1)

m.poly##Training error : 0.28

##kernel = "vanilladot"

m.lin<- ksvm(size ~ temp + RH + wind + rain+FFMC+DMC+ISI+DC,

data = forest[train, ],

kernel = "vanilladot", C = 1)

m.lin##Training error : 0.28

## kernel = "rbfdot"

m.li<- ksvm(size ~ temp + RH + wind + rain+FFMC+DMC+ISI+DC,

data = forest[train, ],

kernel = "rbfdot", C = 1)

m.li##Training error : 0.2625

##kernel = "tanhdot"

m.l<- ksvm(size ~ temp + RH + wind + rain+FFMC+DMC+ISI+DC,

data = forest[train, ],

kernel = "tanhdot", C = 1)

m.l##Training error : 0.425

##Prediction of the model

pred <- predict(m.li, newdata = forest[-train, ], type = "response")

pred

table(forest[-train,"size"])

table(pred)

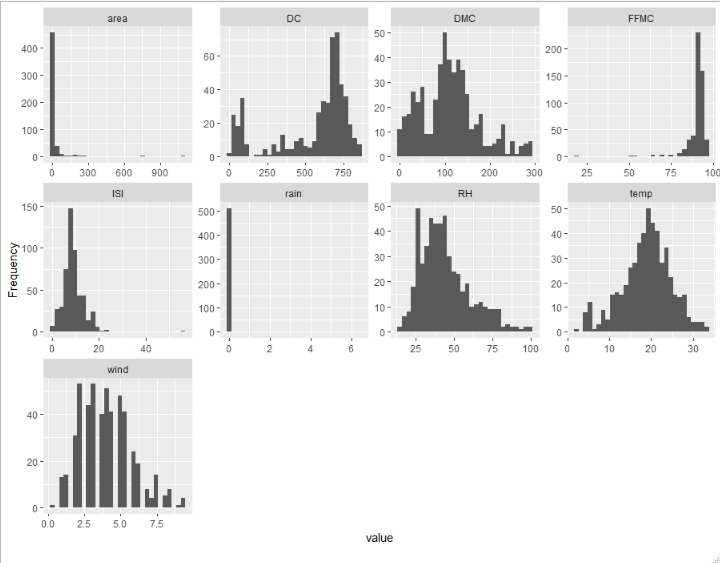
test <- forest[-train]

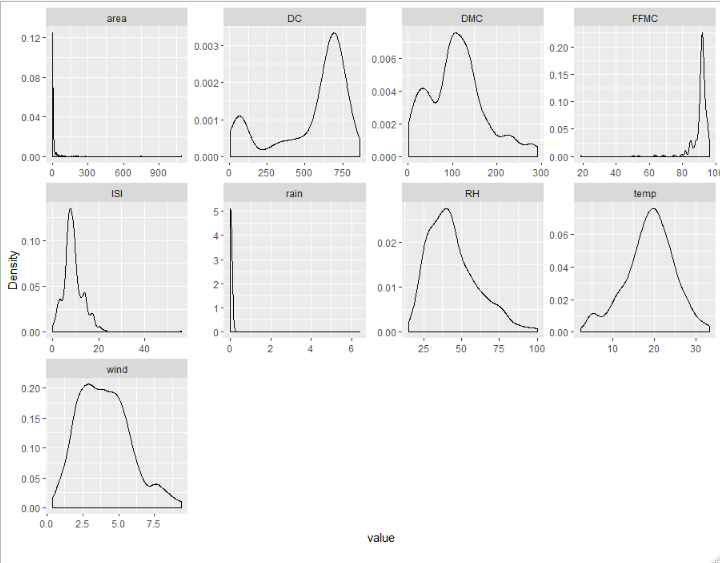
confusionMatrix(pred,test$)

confusionMatrix(table(pred, forest[-train, "size"]), positive = "small") # from the caret package, also need e1071

##Accuracy : 0.6752

EDA plots





SALARY DATA

##import the data set##

library(readr)

data <- read.csv(file.choose())

View(data)

## changing the factor variable into numeric variable for the independent variable

data$workclass <- as.numeric(factor(data$workclass))

data$education <- as.numeric(factor(data$education))

data$maritalstatus <- as.numeric(factor(data$maritalstatus))

data$occupation <- as.numeric(factor(data$occupation))

data$relationship <- as.numeric(factor(data$relationship))

data$race <- as.numeric(factor(data$race))

data$sex <- as.numeric(factor(data$sex))

data$native <- as.numeric(factor(data$native))

str(data)

attach(data)

##EDA

library(DataExplorer)

plot\_str(data)

##to known the variables in the data set

plot\_missing(data)

##knowning missing values in the data set

plot\_histogram(data)

## knowning the continous data

plot\_density(data)

##representing the data by plots

plot\_correlation(data, type = 'continuous',"Salary")

plot\_bar(data)

##plotting the data of the categorical values

create\_report(data)

##we should normalize all the variables

normalize <- function(x){

return((x-min(x))/(max(x)-min(x)))

}

nordata <- data[,-14]

nordata <- normalize(nordata)

##model building

library(kernlab)

library(caret)

##kernel = "polydot"

m.poly <- ksvm(Salary ~ .,

data = nordata,

kernel = "polydot", C = 1)

m.poly ##Training error : 0.189317

##kernel = "vanilladot"

m.lin <- ksvm(Salary ~ .,

data = nordata,

kernel = "vanilladot", C = 1)

m.lin ##Training error : 0.189284

## kernel = "rbfdot"

m.li <- ksvm(Salary ~ .,

data = nordata,

kernel = "rbfdot", C = 1)

m.li##Training error : 0.149

##Prediction of the model

test <- read.csv(file.choose())

attach(test)

test$workclass <- as.numeric(factor(test$workclass))

test$education <- as.numeric(factor(test$education))

test$maritalstatus <- as.numeric(factor(test$maritalstatus))

test$occupation <- as.numeric(factor(test$occupation))

test$relationship <- as.numeric(factor(test$relationship))

test$race <- as.numeric(factor(test$race))

test$sex <- as.numeric(factor(test$sex))

test$native <- as.numeric(factor(test$native))

##we should normalize all the variables

normalize <- function(x){return((x-min(x))/(max(x)-min(x)))

}

nordata1 <- test[,-14]

View(nordata1)

nordata1<- normalize(nordata1)

pred <- predict(m.li, newdata = nordata1, type = "response")

pred

confusionMatrix(pred,test$Salary)

##Accuracy : 0.8459