## 

# Assignment (14.2) 21- Jan 2018

5. Problem Statement

1. Use the below given data set

DataSet

Problem- prediction of the number of comments in the upcoming 24 hours on

those blogs, The train data was generated from different base times that may

tempor ally overlap. Therefore, if you simply split the train into disjoint partitions,

the underlying time intervals may overlap. Therefore, the you should use the

provided, temporally disjoint train and test splits to ensure that the evaluation is

fair.

a. Read the dataset and identify the right features

b. Clean dataset, impute missing values and perform exploratory data analysis.

c. Visualize the dataset and make inferences from that

d. Perform any 3 hypothesis tests using columns of your choice, make conclusions

e. Create a linear regression model to predict the number of comments in the next 24 hours

(relative to basetime)

f. Fine tune the model and represent important features

g. Interpret the summary of the linear model

h. Report the test accuracy vs. the training accuracy

i. Interpret the final model coefficients

j. Plot the model result and compare it with assumptions of the model

Approximate Time to Complete Task

30 Minutes

#Sol :-

#a)

#reading the dataset and viewing

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

slr1<- slr

View(slr1)

#features

dim(slr1)

str(slr1)

library(psych)

describe(slr1)

summary(slr1)

#b) & #c)

#using dataset slr1

#using dataset airquality

airquality <- read.csv("D:/acadgild/airquality.csv")

View(airquality)

#visualization

hist(slr1$Advt ,xlab = "advt", ylab = "Frequency",main="Histogram of advt",col="red")

hist(slr1$Sales ,xlab = "sales", ylab = "Frequency",main="Histogram of sales",col="blue")

plot(slr1$Advt,slr1$Sales)

#missing values and imputing and cleaning

#with the help of summary function we can find which variable has how many NA value

#or check for missing values

summary(airquality)

#thus ozone and solar.r has missing values

#lets see the structure of airquality first

str(airquality)

#in another way say do for variable Solar.R in airquality dataset

newair =airquality

dim(newair)

str(newair)

summary(newair)

#before imputing

hist(newair$Solar.R ,xlab = "Solar.R", ylab = "frequency",main="histogram of Solar.R",col="red")

mean(newair$Solar.R)

mean(newair$Solar.R,na.rm = T)

#imputed my mean

newair$Solar.R[is.na(newair$Solar.R)]<- mean(newair$Solar.R,na.rm = T)

#check summary after done with imputing

summary(newair)

newair$Solar.R

#visualize after imputing the variable Solar.R with the mean

#lets visualize through histogram

#after imputing

hist(newair$Solar.R ,xlab = "Solar.R", ylab = "frequency",main="histogram of Solar.R",col="red")

#d)

#hypothesis tests

#we do paired test for continous variables

#using airquality dataset

#some of test are as follows

#define the null hypothesis

#Ho: Mean of first variable - Mean of 2 variable is equal to 0

#Ha: Mean of first variable - Mean of 2 variable is not equal to 0

t.test(x=airquality$Ozone, y=airquality$Solar.R ,alternative = "two.sided",mu=0 ,paired = TRUE)

t.test(x=airquality$Temp, y=airquality$Wind ,alternative = "two.sided",mu=0 ,paired = TRUE)

t.test(x=airquality$Ozone, y=airquality$Temp ,alternative = "two.sided",mu=0 ,paired = TRUE)

t.test(x=airquality$Day, y=airquality$Solar.R ,alternative = "two.sided",mu=0 ,paired = TRUE)

#as p value of this test is <0.05 we reject the null hypo

#and accept the alternative hypothesis which says there

#Mean of 1 variable - Mean of 2 variable is not equal to 0

#thus this are some test that we performed

#e),f),g)

#Answers

#using slr1 dataset

#linear regression model

model<- lm(slr1$Advt~slr1$Sales)

model

#Important features

#multiple r squared value

#p value of slope test

#F stats

#check and interpreting the summary

summary(model)

#\*\*NOTE\*\*

#Interpreting

#thus by multiple r squared value we see our model is good

#also our p value of slope test is <0.05 so good for our model

#adjusted r squared value is also good 0.891

#f stats value of 90.93 suggest our model is good and also its p value is <0.05

#our model accuracy is 0.9009 which is good

#predicting

Pred<- predict(lm(slr1$Sales~slr1$Advt))

Pred

pred<- predict(model,newdata= slr1Test,type = "response")

table(slr1$Advt,pred>= 0.5)

conf<- table(slr1$Advt,pred)

conf

#verfify residuals

error<- residuals(lm(slr1$Sales~slr1$Advt))

error

summary(error)

#h)

#test and training accuracy

#dataset slr1

set.seed(1)

split<- sample.split(slr1$Advt,SplitRatio = 0.70)

slr1Train <- subset(slr1,split == TRUE)

slr1Test<- subset(slr1, split == FALSE)

#train

model1<- lm(slr1Train$Advt~slr1Train$Sales)

model1

summary(model1)

#accuracy is 0.926

#test

model2<- lm(slr1Test$Advt~slr1Test$Sales)

model2

summary(model2)

#accuracy is 0.871

#i) & j)

#Answers

#plotting,comparing,interpreting

#result of all of our models

summary(model)

summary(model1)

summary(model2)

#model coefficients

model

model1

model2

slr1$coefficients<- NA

slr1$coefficients<- model$coefficients

slr1$coefficients

#Assumptions of the linear regression model and test

#normality

#linearity

#independence of error

#independence of error/constant error variance

predict(model)

Pred=predict(model)

slr1$predicted =NA

slr1$predicted =Pred

slr1$error =model$residuals

#name diff

slr1$error =model$error

hist(slr1$error,col = "red",horizontal =T)

boxplot(slr1$error,col = "blue",horizontal =T)

hist(slr1$Advt,col = "red",horizontal =T)

boxplot(slr1$Sales,col = "blue",horizontal =T)

plot(slr1$Observation.no,slr1$error,main="independence of error",col="green")

plot(slr1$Pred,slr1$error,main="independence of error",col="green")

#visualize

plot(slr1$Advt,slr1$Sales)

plot(slr1Test$Advt,slr1Test$Sales)

plot(slr1Train$Advt,slr1Train$Sales)