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

temporally 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.

1. 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)

> install.packages("Boruta")  
> library(Boruta)

> setwd("../Data/Loan\_Prediction")  
> traindata <- read.csv("train.csv", header = T, stringsAsFactors = F)

> str(traindata)  
> names(traindata) <- gsub("\_", "", names(traindata))

> summary(traindata)

> traindata[traindata == ""] <- NA

> traindata <- traindata[complete.cases(traindata),]

> convert <- c(2:6, 11:13)  
> traindata[,convert] <- data.frame(apply(traindata[convert], 2, as.factor))

> set.seed(123)  
> boruta.train <- Boruta(LoanStatus~.-LoanID, data = traindata, doTrace = 2)  
> print(boruta.train)

Boruta performed 99 iterations in 18.80749 secs.  
5 attributes confirmed important: ApplicantIncome, CoapplicantIncome,  
CreditHistory, LoanAmount, LoanAmountTerm.  
4 attributes confirmed unimportant: Dependents, Education, Gender, SelfEmployed.  
2 tentative attributes left: Married, PropertyArea.

> plot(boruta.train, xlab = "", xaxt = "n")  
> lz<-lapply(1:ncol(boruta.train$ImpHistory),function(i)  
boruta.train$ImpHistory[is.finite(boruta.train$ImpHistory[,i]),i])  
> names(lz) <- colnames(boruta.train$ImpHistory)  
> Labels <- sort(sapply(lz,median))  
> axis(side = 1,las=2,labels = names(Labels),  
at = 1:ncol(boruta.train$ImpHistory), cex.axis = 0.7)

> final.boruta <- TentativeRoughFix(boruta.train)  
> print(final.boruta)

Boruta performed 99 iterations in 18.399 secs.  
Tentatives roughfixed over the last 99 iterations.  
6 attributes confirmed important: ApplicantIncome, CoapplicantIncome,  
CreditHistory, LoanAmount, LoanAmountTerm and 1 more.  
5 attributes confirmed unimportant: Dependents, Education, Gender, PropertyArea,  
SelfEmployed.

> getSelectedAttributes(final.boruta, withTentative = F)  
[1] "Married"           "ApplicantIncome"   "CoapplicantIncome" "LoanAmount"  
[5] "LoanAmountTerm"    "CreditHistory"

> boruta.df <- attStats(final.boruta)  
> class(boruta.df)  
[1] "data.frame"  
> print(boruta.df)

                     meanImp   medianImp   minImp    maxImp   normHits    decision

Gender             1.04104738  0.9181620 -1.9472672  3.767040 0.01010101  Rejected

Married            2.76873080  2.7843600 -1.5971215  6.685000 0.56565657  Confirmed

Dependents         1.15900910  1.0383850 -0.7643617  3.399701 0.01010101  Rejected

Education          0.64114702  0.4747312 -1.0773928  3.745441 0.03030303  Rejected

SelfEmployed      -0.02442418 -0.1511711 -0.9536783  1.495992 0.00000000  Rejected

ApplicantIncome    6.05487791  6.0311639  2.9801751  9.197305 0.94949495  Confirmed

CoapplicantIncome  5.76704389  5.7920332  1.9322989 10.184245 0.97979798  Confirmed

LoanAmount         5.19167613  5.3606935  1.7489061  8.855464 0.88888889  Confirmed

LoanAmountTerm     5.50553498  5.3938036  2.0361781  9.025020 0.90909091  Confirmed

CreditHistory     59.57931404 60.2352549 51.7297906 69.721650 1.00000000  Confirmed

PropertyArea       2.77155525  2.4715892 -1.2486696  8.719109 0.54545455  Rejected