LR-MODEL.R

User

2020-04-13

#title: "Logistic Regression Model on Loan Dataset"  
#author: "Usman Jibril"  
#date: "13/04/2020"  
#output: html\_document

#Using three algorithms to predict loan eligibility.  
#we used Logistic regression, Decision tree and Random forest algorithms  
#Using Logistic Regression   
loandata<-read.csv("C:\\Users\\User\\Documents\\Datasets\\loan\_data\_set.csv")  
View(loandata)  
str(loandata)

## 'data.frame': 614 obs. of 13 variables:  
## $ Loan\_ID : Factor w/ 614 levels "LP001002","LP001003",..: 1 2 3 4 5 6 7 8 9 10 ...  
## $ Gender : Factor w/ 3 levels "","Female","Male": 3 3 3 3 3 3 3 3 3 3 ...  
## $ Married : Factor w/ 3 levels "","No","Yes": 2 3 3 3 2 3 3 3 3 3 ...  
## $ Dependents : Factor w/ 5 levels "","0","1","2",..: 2 3 2 2 2 4 2 5 4 3 ...  
## $ Education : Factor w/ 2 levels "Graduate","Not Graduate": 1 1 1 2 1 1 2 1 1 1 ...  
## $ Self\_Employed : Factor w/ 3 levels "","No","Yes": 2 2 3 2 2 3 2 2 2 2 ...  
## $ ApplicantIncome : int 5849 4583 3000 2583 6000 5417 2333 3036 4006 12841 ...  
## $ CoapplicantIncome: num 0 1508 0 2358 0 ...  
## $ LoanAmount : int NA 128 66 120 141 267 95 158 168 349 ...  
## $ Loan\_Amount\_Term : int 360 360 360 360 360 360 360 360 360 360 ...  
## $ Credit\_History : int 1 1 1 1 1 1 1 0 1 1 ...  
## $ Property\_Area : Factor w/ 3 levels "Rural","Semiurban",..: 3 1 3 3 3 3 3 2 3 2 ...  
## $ Loan\_Status : Factor w/ 2 levels "N","Y": 2 1 2 2 2 2 2 1 2 1 ...

summary(loandata)

## Loan\_ID Gender Married Dependents Education   
## LP001002: 1 : 13 : 3 : 15 Graduate :480   
## LP001003: 1 Female:112 No :213 0 :345 Not Graduate:134   
## LP001005: 1 Male :489 Yes:398 1 :102   
## LP001006: 1 2 :101   
## LP001008: 1 3+: 51   
## LP001011: 1   
## (Other) :608   
## Self\_Employed ApplicantIncome CoapplicantIncome LoanAmount   
## : 32 Min. : 150 Min. : 0 Min. : 9.0   
## No :500 1st Qu.: 2878 1st Qu.: 0 1st Qu.:100.0   
## Yes: 82 Median : 3812 Median : 1188 Median :128.0   
## Mean : 5403 Mean : 1621 Mean :146.4   
## 3rd Qu.: 5795 3rd Qu.: 2297 3rd Qu.:168.0   
## Max. :81000 Max. :41667 Max. :700.0   
## NA's :22   
## Loan\_Amount\_Term Credit\_History Property\_Area Loan\_Status  
## Min. : 12 Min. :0.0000 Rural :179 N:192   
## 1st Qu.:360 1st Qu.:1.0000 Semiurban:233 Y:422   
## Median :360 Median :1.0000 Urban :202   
## Mean :342 Mean :0.8422   
## 3rd Qu.:360 3rd Qu.:1.0000   
## Max. :480 Max. :1.0000   
## NA's :14 NA's :50

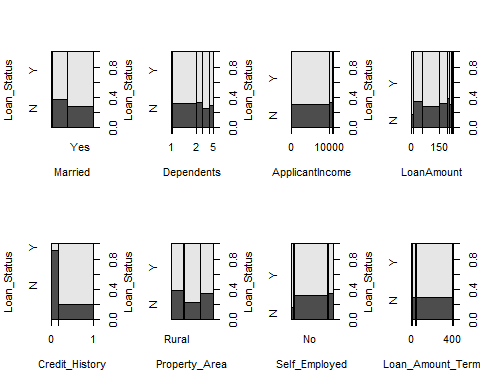
loandata<-na.omit(loandata) #if NAs are present we omit  
str(loandata)

## 'data.frame': 529 obs. of 13 variables:  
## $ Loan\_ID : Factor w/ 614 levels "LP001002","LP001003",..: 2 3 4 5 6 7 8 9 10 11 ...  
## $ Gender : Factor w/ 3 levels "","Female","Male": 3 3 3 3 3 3 3 3 3 3 ...  
## $ Married : Factor w/ 3 levels "","No","Yes": 3 3 3 2 3 3 3 3 3 3 ...  
## $ Dependents : Factor w/ 5 levels "","0","1","2",..: 3 2 2 2 4 2 5 4 3 4 ...  
## $ Education : Factor w/ 2 levels "Graduate","Not Graduate": 1 1 2 1 1 2 1 1 1 1 ...  
## $ Self\_Employed : Factor w/ 3 levels "","No","Yes": 2 3 2 2 3 2 2 2 2 2 ...  
## $ ApplicantIncome : int 4583 3000 2583 6000 5417 2333 3036 4006 12841 3200 ...  
## $ CoapplicantIncome: num 1508 0 2358 0 4196 ...  
## $ LoanAmount : int 128 66 120 141 267 95 158 168 349 70 ...  
## $ Loan\_Amount\_Term : int 360 360 360 360 360 360 360 360 360 360 ...  
## $ Credit\_History : int 1 1 1 1 1 1 0 1 1 1 ...  
## $ Property\_Area : Factor w/ 3 levels "Rural","Semiurban",..: 1 3 3 3 3 3 2 3 2 3 ...  
## $ Loan\_Status : Factor w/ 2 levels "N","Y": 1 2 2 2 2 2 1 2 1 2 ...  
## - attr(\*, "na.action")= 'omit' Named int 1 17 20 25 31 36 37 43 45 46 ...  
## ..- attr(\*, "names")= chr "1" "17" "20" "25" ...

#change dependents into numeric  
loandata$Dependents<-as.numeric(loandata$Dependents)  
#some NA's still showing in the dataset  
loandata<-na.omit(loandata)  
summary(loandata)

## Loan\_ID Gender Married Dependents Education   
## LP001003: 1 : 12 : 2 Min. :1.000 Graduate :421   
## LP001005: 1 Female: 95 No :188 1st Qu.:2.000 Not Graduate:108   
## LP001006: 1 Male :422 Yes:339 Median :2.000   
## LP001008: 1 Mean :2.741   
## LP001011: 1 3rd Qu.:4.000   
## LP001013: 1 Max. :5.000   
## (Other) :523   
## Self\_Employed ApplicantIncome CoapplicantIncome LoanAmount   
## : 25 Min. : 150 Min. : 0 Min. : 9.0   
## No :434 1st Qu.: 2900 1st Qu.: 0 1st Qu.:100.0   
## Yes: 70 Median : 3816 Median : 1086 Median :128.0   
## Mean : 5508 Mean : 1542 Mean :145.9   
## 3rd Qu.: 5815 3rd Qu.: 2232 3rd Qu.:167.0   
## Max. :81000 Max. :33837 Max. :700.0   
##   
## Loan\_Amount\_Term Credit\_History Property\_Area Loan\_Status  
## Min. : 36.0 Min. :0.0000 Rural :155 N:163   
## 1st Qu.:360.0 1st Qu.:1.0000 Semiurban:209 Y:366   
## Median :360.0 Median :1.0000 Urban :165   
## Mean :342.4 Mean :0.8507   
## 3rd Qu.:360.0 3rd Qu.:1.0000   
## Max. :480.0 Max. :1.0000   
##

par(mfrow=c(2,4))  
plot(Loan\_Status~Married, data=loandata)  
plot(Loan\_Status~Dependents, data=loandata)  
plot(Loan\_Status~ApplicantIncome, data=loandata)  
plot(Loan\_Status~LoanAmount, data=loandata)  
plot(Loan\_Status~Credit\_History, data=loandata)  
plot(Loan\_Status~Property\_Area, data=loandata)  
plot(Loan\_Status~Self\_Employed, data=loandata)  
plot(Loan\_Status~Loan\_Amount\_Term, data=loandata)



#plot boxplots of the variables against the target variables  
par(mfrow=c(2,3))  
boxplot(Dependents~Loan\_Status, ylab="Dependents", xlab="Loan\_Status", col="red", data=loandata)  
boxplot(ApplicantIncome~Loan\_Status, ylab="ApplicantIncome", xlab="Loan\_Status", col="yellow", data=loandata)  
boxplot(LoanAmount~Loan\_Status, ylab="LoanAmount", xlab="Loan\_Status", col="purple", data=loandata)  
boxplot(Credit\_History~Loan\_Status, ylab="Credit\_History", xlab="Loan\_Status", col="blue", data=loandata)  
boxplot(Loan\_Amount\_Term~Loan\_Status, ylab="Loan\_Amount\_Term", xlab="Loan\_Status", col="green", data=loandata)  
  
  
  
#Create Training and Test Samples  
data1<-loandata[which(loandata$Loan\_Status=="Y"),]  
data2<-loandata[which(loandata$Loan\_Status=="N"),]  
View(data1)  
View(data2)  
  
set.seed(150)  
  
#set traing dataset to 75% of the data  
trainingset1<-sample(1:nrow(data1), 0.75\*nrow(data1))  
trainingset2<-sample(1:nrow(data2), 0.75\*nrow(data2))  
View(trainingset1)  
View(trainingset2)  
  
training\_1<-data1[trainingset1,]  
training\_2<-data2[trainingset2,]  
  
trainingdata<-rbind(training\_1,training\_2)  
View(trainingdata)  
  
##create a Test Set of the remaining 25%  
test\_1<-data1[trainingset1, ]  
test\_2<-data2[-trainingset2, ]  
testdata<-rbind(test\_1,test\_2)  
View(testdata)  
  
#Get predictor values with two methods(Random Forest and Boruta methods)  
  
#confusion matrix is used to describe the performance of a classification model applied to classification problem, whereby you can   
  
  
#this shows the levels of importance of the variables in predicting our dependent variable.  
  
#Boruta is a feature selection algorithm.  
  
  
#boruta has shown the variables deemed important and unimportant.  
  
#our Mean Decrease GINI shows the measure of variable importance based on split in trees(used for random forest)  
  
#STEPS  
#tweak the algorithm (E.G change the parameters)  
#use a different machine learning algorithm  
  
  
  
#VISUALIZE THE VARIABLES AND THEIR IMPORTANCE  
  
#create a dataframe of the final result derived from boruta and store the result in boruta.df  
  
#we now build Logistics regression models and get predictions.All variables included  
#Generalized Linear Model(another name for Logistic regression) is an extention of  
#linear regression models that allow the dependent variable to be non-normal(i.e categorical)  
  
logisticsmodel1<-glm(Loan\_Status~Gender+Married+Dependents+Education+Self\_Employed+ApplicantIncome+CoapplicantIncome+LoanAmount+Loan\_Amount\_Term+Credit\_History+Property\_Area,  
 data=loandata,family=binomial(link="logit"))  
logisticsmodel1

##   
## Call: glm(formula = Loan\_Status ~ Gender + Married + Dependents + Education +   
## Self\_Employed + ApplicantIncome + CoapplicantIncome + LoanAmount +   
## Loan\_Amount\_Term + Credit\_History + Property\_Area, family = binomial(link = "logit"),   
## data = loandata)  
##   
## Coefficients:  
## (Intercept) GenderFemale GenderMale   
## 1.117e+01 2.089e-02 3.530e-01   
## MarriedNo MarriedYes Dependents   
## -1.348e+01 -1.302e+01 1.304e-01   
## EducationNot Graduate Self\_EmployedNo Self\_EmployedYes   
## -4.334e-01 -6.247e-01 -8.070e-01   
## ApplicantIncome CoapplicantIncome LoanAmount   
## 1.465e-05 -4.021e-05 -2.362e-03   
## Loan\_Amount\_Term Credit\_History Property\_AreaSemiurban   
## -1.357e-03 3.872e+00 1.033e+00   
## Property\_AreaUrban   
## 1.706e-01   
##   
## Degrees of Freedom: 528 Total (i.e. Null); 513 Residual  
## Null Deviance: 653.4   
## Residual Deviance: 470.1 AIC: 502.1

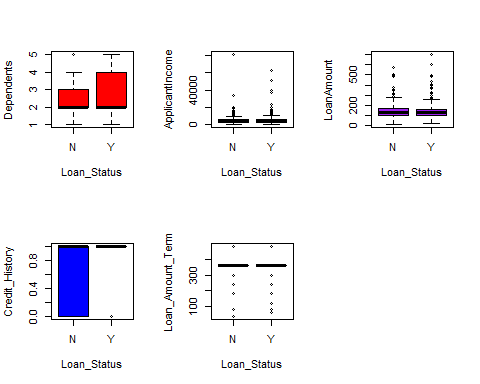
#we now test without the least important (5) variables according to boruta and random forest   
  
logisticsmodel2<-glm(Loan\_Status~Credit\_History+ApplicantIncome+CoapplicantIncome+LoanAmount+Property\_Area+Dependents+Loan\_Amount\_Term,  
 data=loandata,family=binomial(link="logit"))  
logisticsmodel2

##   
## Call: glm(formula = Loan\_Status ~ Credit\_History + ApplicantIncome +   
## CoapplicantIncome + LoanAmount + Property\_Area + Dependents +   
## Loan\_Amount\_Term, family = binomial(link = "logit"), data = loandata)  
##   
## Coefficients:  
## (Intercept) Credit\_History ApplicantIncome   
## -2.732e+00 3.872e+00 1.081e-05   
## CoapplicantIncome LoanAmount Property\_AreaSemiurban   
## -2.755e-05 -1.585e-03 1.048e+00   
## Property\_AreaUrban Dependents Loan\_Amount\_Term   
## 2.409e-01 2.012e-01 -1.508e-03   
##   
## Degrees of Freedom: 528 Total (i.e. Null); 520 Residual  
## Null Deviance: 653.4   
## Residual Deviance: 480 AIC: 498

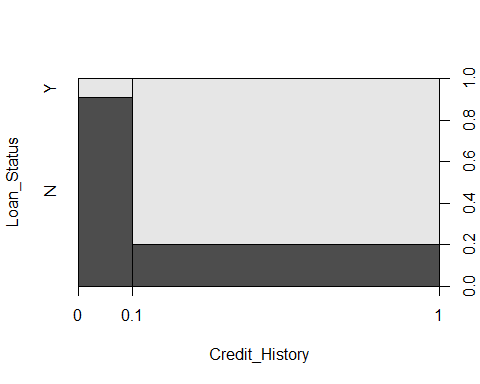
#predict scores and add it to the test data  
predicted<-predict(logisticsmodel2,testdata)  
View(predicted)  
  
#merge the prediction to the testdata  
testdata$predicted<-predicted  
View(testdata)  
  
##Analysis of logistics model and model fit  
modelsummary<-summary(logisticsmodel2)  
modelsummary$coefficients

## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -2.732222e+00 9.264597e-01 -2.9490993 3.187016e-03  
## Credit\_History 3.872217e+00 4.271196e-01 9.0658845 1.235973e-19  
## ApplicantIncome 1.081206e-05 2.468330e-05 0.4380314 6.613635e-01  
## CoapplicantIncome -2.754892e-05 4.197008e-05 -0.6563942 5.115705e-01  
## LoanAmount -1.585490e-03 1.735046e-03 -0.9138031 3.608203e-01  
## Property\_AreaSemiurban 1.048373e+00 2.875977e-01 3.6452762 2.671048e-04  
## Property\_AreaUrban 2.408702e-01 2.766388e-01 0.8707031 3.839163e-01  
## Dependents 2.012373e-01 1.173009e-01 1.7155651 8.624167e-02  
## Loan\_Amount\_Term -1.507929e-03 2.006944e-03 -0.7513558 4.524386e-01

#test and graph the model  
#plot of Loan Status Eligibility i.e NO or YES by Credit\_History  
  
#SAME IN BAR GRAPH  
par(mfrow=c(1,1))



plot(Loan\_Status~Credit\_History, data=loandata)



#Analysis of variance(ANOVA)  
#This backs up our earlier observations on significance of the variables  
anova(logisticsmodel1,test="Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: Loan\_Status  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 528 653.42   
## Gender 2 2.111 526 651.31 0.3480991   
## Married 2 4.840 524 646.46 0.0889023 .   
## Dependents 1 0.413 523 646.05 0.5203597   
## Education 1 3.484 522 642.57 0.0619616 .   
## Self\_Employed 2 3.939 520 638.63 0.1395291   
## ApplicantIncome 1 0.240 519 638.39 0.6239305   
## CoapplicantIncome 1 2.127 518 636.26 0.1447223   
## LoanAmount 1 1.625 517 634.64 0.2024412   
## Loan\_Amount\_Term 1 0.275 516 634.36 0.6000426   
## Credit\_History 1 149.282 515 485.08 < 2.2e-16 \*\*\*  
## Property\_Area 2 15.018 513 470.06 0.0005482 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(logisticsmodel1) #interprete the metrics

##   
## Call:  
## glm(formula = Loan\_Status ~ Gender + Married + Dependents + Education +   
## Self\_Employed + ApplicantIncome + CoapplicantIncome + LoanAmount +   
## Loan\_Amount\_Term + Credit\_History + Property\_Area, family = binomial(link = "logit"),   
## data = loandata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.3463 -0.3639 0.4986 0.7000 2.4848   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.117e+01 6.161e+02 0.018 0.985530   
## GenderFemale 2.089e-02 7.941e-01 0.026 0.979010   
## GenderMale 3.530e-01 7.495e-01 0.471 0.637700   
## MarriedNo -1.348e+01 6.161e+02 -0.022 0.982544   
## MarriedYes -1.302e+01 6.161e+02 -0.021 0.983135   
## Dependents 1.304e-01 1.267e-01 1.029 0.303692   
## EducationNot Graduate -4.334e-01 2.916e-01 -1.486 0.137176   
## Self\_EmployedNo -6.247e-01 5.919e-01 -1.055 0.291206   
## Self\_EmployedYes -8.070e-01 6.571e-01 -1.228 0.219401   
## ApplicantIncome 1.465e-05 2.500e-05 0.586 0.557750   
## CoapplicantIncome -4.021e-05 4.244e-05 -0.947 0.343397   
## LoanAmount -2.362e-03 1.756e-03 -1.345 0.178661   
## Loan\_Amount\_Term -1.357e-03 2.026e-03 -0.670 0.503047   
## Credit\_History 3.872e+00 4.337e-01 8.927 < 2e-16 \*\*\*  
## Property\_AreaSemiurban 1.033e+00 2.925e-01 3.533 0.000411 \*\*\*  
## Property\_AreaUrban 1.706e-01 2.839e-01 0.601 0.547867   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 653.42 on 528 degrees of freedom  
## Residual deviance: 470.06 on 513 degrees of freedom  
## AIC: 502.06  
##   
## Number of Fisher Scoring iterations: 13

#final step of predicting with the model and scoring back on the original dataset  
pred<-predict(logisticsmodel2,testdata,"link")  
head(pred)

## 493 250 468 197 459 127   
## 0.8651189 1.0658892 2.0540595 0.6899173 1.2335392 1.2740555

#create new class prediction column on the original data set so you now have  
#side by side comparison  
  
testdata$Class<-ifelse(pred>0,"Y","N")  
View(testdata)