Credit Risk Validation

Aneesh Avati

23 September 2018

# INTRODUCTION

## Credit Risk:

Credit risk is nothing but the default in payment of any loan by the borrower. In Banking sector this is an important factor to be considered before approving the loan of an applicant.

## Impact of Credit Risk:

The bank’s revenue pattern depends on the saving account and loans they offer. Once an account holder puts money in their saving account, they get certain interest on their account balance. The bank keeps some part of these balances in reserve to make payments in case of withdrawals. The rest of the money is given as loans to borrowers. The interest put on loan is the revenue earned by the banks. Therefore taking into consideration the factors that will decide whether a person will pay back the loan or not is crucial for banks. This is a common credit risk problem which is solved using analytics. While receiving application of a borrower, the bank asks certain details about the applicant. The general factors that help predict loan default are income, occupation, age, loan amount etc of the applicant.

# PROBLEM STATEMENT

## About Company:

Dream Housing Finance company deals in all home loans. They have presence across all urban, semi urban and rural areas. Customer first apply for home loan after that company validates the customer eligibility for loan.

## Problem:

Company wants to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amount so that they can specifically target these customers. Here they have provided a partial data set.

## Data:

library(readxl)  
variables <- read\_excel("C:/Users/anees/Desktop/variables.xlsx")  
print(variables)

## # A tibble: 13 x 2  
## Variable Description   
## <chr> <chr>   
## 1 Loan\_ID Unique Loan ID   
## 2 Gender Male/ Female   
## 3 Married Applicant married (Y/N)   
## 4 Dependents Number of dependents   
## 5 Education Applicant Education (Graduate/ Under Graduate)  
## 6 Self\_Employed Self employed (Y/N)   
## 7 ApplicantIncome Applicant income   
## 8 CoapplicantIncome Coapplicant income   
## 9 LoanAmount Loan amount in thousands   
## 10 Loan\_Amount\_Term Term of loan in months   
## 11 Credit\_History credit history meets guidelines   
## 12 Property\_Area Urban/ Semi Urban/ Rural   
## 13 Loan\_Status Loan approved (Y/N)

# Hypothesis:

1.Gender: Chances of loan approval to males can be more than chances of loan approval to females  
2.Marital Status: Chances of loan approval to person who is married is more  
3.Number of Dependents: As the number of dependents increases chances of loan approval decreases  
4.Education: Chances of loan aproval is higher to graduates  
5.Employment Status: chances of loan approval to a candidate who is presently employed is higher  
6.Loan Amount Term: as the loan amount term increases chances of approval is higher  
7.Property Area:loan approval chances to people residing in rural area is less  
8.Applicant Income: person with decent income has more chances of loan approval  
9.Co applicant Income : Person with Co-applicant income has more chances of loan approval

# Data Exploration

## Data Cleaning

## Training data set

crp<-read.csv("D:/imarticus/R/logistic regression/Logistic regression and SVM\_Case study/R\_Module\_Day\_10.2\_Credit\_Risk\_Train\_data (1).csv",na.strings=c(""," ","NA"))  
summary(crp)

## Loan\_ID Gender Married Dependents Education   
## LP001002: 1 Female:112 No :213 0 :345 Graduate :480   
## LP001003: 1 Male :489 Yes :398 1 :102 Not Graduate:134   
## LP001005: 1 NA's : 13 NA's: 3 2 :101   
## LP001006: 1 3+ : 51   
## LP001008: 1 NA's: 15   
## LP001011: 1   
## (Other) :608   
## Self\_Employed ApplicantIncome CoapplicantIncome LoanAmount   
## No :500 Min. : 150 Min. : 0 Min. : 9.0   
## Yes : 82 1st Qu.: 2878 1st Qu.: 0 1st Qu.:100.0   
## NA's: 32 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

Checking for no of missing Values in Training data set

colSums(is.na(crp))

## Loan\_ID Gender Married Dependents   
## 0 13 3 15   
## Education Self\_Employed ApplicantIncome CoapplicantIncome   
## 0 32 0 0   
## LoanAmount Loan\_Amount\_Term Credit\_History Property\_Area   
## 22 14 50 0   
## Loan\_Status   
## 0

## Some observations:

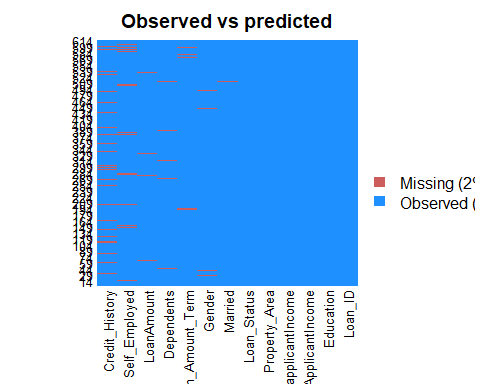
1.The mean of Credit\_history variable is 0.8422. That’s weird knowing that this variable has value of 1 for customers who have credit history and 0 otherwise.  
2.There are blank fields in Gender, Married, Dependents and Self\_Employed.  
3.There are NAs in LoanAmount, Loan\_Amount\_term and Credit\_History.  
4.Loan\_ID is unique key for each person. Hence its of no use to build model.

library(Amelia)

## Loading required package: Rcpp

## ##   
## ## Amelia II: Multiple Imputation  
## ## (Version 1.7.5, built: 2018-05-07)  
## ## Copyright (C) 2005-2018 James Honaker, Gary King and Matthew Blackwell  
## ## Refer to http://gking.harvard.edu/amelia/ for more information  
## ##

missmap(crp, main="Observed vs predicted")

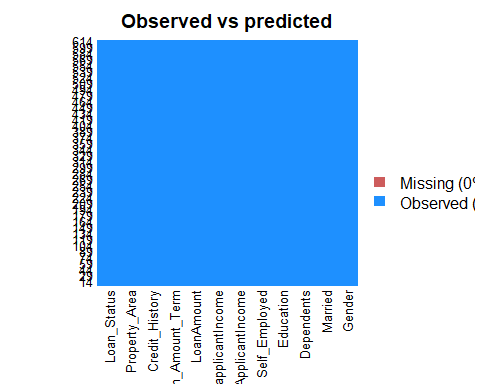


## Missing values imputation on training data set

crp<-crp[,-1]  
crp$Credit\_History=as.factor(crp$Credit\_History)  
crp$Gender[is.na(crp$Gender)]<- 'Male'  
crp$Married[is.na(crp$Married)]<- "Yes"  
crp$Dependents[is.na(crp$Dependents)]<- 0  
crp$Self\_Employed[is.na(crp$Self\_Employed)]<-"No"  
crp$LoanAmount[is.na(crp$LoanAmount)]<-median(crp$LoanAmount,na.rm=T)  
crp$Loan\_Amount\_Term[is.na(crp$Loan\_Amount\_Term)]<-median(crp$Loan\_Amount\_Term,na.rm=T)  
crp$Credit\_History[is.na(crp$Credit\_History)]<- 1  
colSums(is.na(crp))

## Gender Married Dependents Education   
## 0 0 0 0   
## Self\_Employed ApplicantIncome CoapplicantIncome LoanAmount   
## 0 0 0 0   
## Loan\_Amount\_Term Credit\_History Property\_Area Loan\_Status   
## 0 0 0 0

library(Amelia)  
missmap(crp, main="Observed vs predicted")



## Validation Data

crp\_v<-read.csv("D:/imarticus/R/logistic regression/Logistic regression and SVM\_Case study/R\_Module\_Day\_8.2\_Credit\_Risk\_Validate\_data.csv",na.strings=c(""," ","NA"))  
summary(crp\_v)

## Loan\_ID Gender Married Dependents Education   
## LP001015: 1 Female: 70 No :134 0 :200 Graduate :283   
## LP001022: 1 Male :286 Yes:233 1 : 58 Not Graduate: 84   
## LP001031: 1 NA's : 11 2 : 59   
## LP001035: 1 3+ : 40   
## LP001051: 1 NA's: 10   
## LP001054: 1   
## (Other) :361   
## Self\_Employed ApplicantIncome CoapplicantIncome LoanAmount   
## No :307 Min. : 0 Min. : 0 Min. : 28.0   
## Yes : 37 1st Qu.: 2864 1st Qu.: 0 1st Qu.:100.2   
## NA's: 23 Median : 3786 Median : 1025 Median :125.0   
## Mean : 4806 Mean : 1570 Mean :136.1   
## 3rd Qu.: 5060 3rd Qu.: 2430 3rd Qu.:158.0   
## Max. :72529 Max. :24000 Max. :550.0   
## NA's :5   
## Loan\_Amount\_Term Credit\_History Property\_Area outcome  
## Min. : 6.0 Min. :0.0000 Rural :111 N: 77   
## 1st Qu.:360.0 1st Qu.:1.0000 Semiurban:116 Y:290   
## Median :360.0 Median :1.0000 Urban :140   
## Mean :342.5 Mean :0.8254   
## 3rd Qu.:360.0 3rd Qu.:1.0000   
## Max. :480.0 Max. :1.0000   
## NA's :6 NA's :29

crp\_v<-crp\_v[,-1]  
colSums(is.na(crp\_v))

## Gender Married Dependents Education   
## 11 0 10 0   
## Self\_Employed ApplicantIncome CoapplicantIncome LoanAmount   
## 23 0 0 5   
## Loan\_Amount\_Term Credit\_History Property\_Area outcome   
## 6 29 0 0

## Missing values imputation on Validation data set

crp\_v$Credit\_History<-as.factor(crp\_v$Credit\_History)  
crp\_v$LoanAmount[is.na(crp\_v$LoanAmount)] <- median(crp\_v$LoanAmount,na.rm=T)  
crp\_v$Loan\_Amount\_Term[is.na(crp\_v$Loan\_Amount\_Term)] <- median(crp\_v$Loan\_Amount\_Term,na.rm=T)  
crp\_v$Credit\_History[is.na(crp\_v$Credit\_History)]<-1  
crp\_v$Self\_Employed[is.na(crp\_v$Self\_Employed)]<-"No"  
crp\_v$Dependents[is.na(crp\_v$Dependents)]<-0  
crp\_v$Gender[is.na(crp\_v$Gender)]<-"Male"  
summary(crp\_v)

## Gender Married Dependents Education Self\_Employed  
## Female: 70 No :134 0 :210 Graduate :283 No :330   
## Male :297 Yes:233 1 : 58 Not Graduate: 84 Yes: 37   
## 2 : 59   
## 3+: 40   
##   
##   
## ApplicantIncome CoapplicantIncome LoanAmount Loan\_Amount\_Term  
## Min. : 0 Min. : 0 Min. : 28.0 Min. : 6.0   
## 1st Qu.: 2864 1st Qu.: 0 1st Qu.:101.0 1st Qu.:360.0   
## Median : 3786 Median : 1025 Median :125.0 Median :360.0   
## Mean : 4806 Mean : 1570 Mean :136.0 Mean :342.8   
## 3rd Qu.: 5060 3rd Qu.: 2430 3rd Qu.:157.5 3rd Qu.:360.0   
## Max. :72529 Max. :24000 Max. :550.0 Max. :480.0   
## Credit\_History Property\_Area outcome  
## 0: 59 Rural :111 N: 77   
## 1:308 Semiurban:116 Y:290   
## Urban :140   
##   
##   
##

## Test Set

crp\_test<-read.csv("D:/imarticus/R/logistic regression/Logistic regression and SVM\_Case study/R\_Module\_Day\_10.3\_Credit\_Risk\_Test\_data.csv",na.strings=c(""," ","NA"))  
summary(crp\_test)

## Loan\_ID Gender Married Dependents Education   
## LP001015: 1 Female: 70 No :134 0 :200 Graduate :283   
## LP001022: 1 Male :286 Yes:233 1 : 58 Not Graduate: 84   
## LP001031: 1 NA's : 11 2 : 59   
## LP001035: 1 3+ : 40   
## LP001051: 1 NA's: 10   
## LP001054: 1   
## (Other) :361   
## Self\_Employed ApplicantIncome CoapplicantIncome LoanAmount   
## No :307 Min. : 0 Min. : 0 Min. : 28.0   
## Yes : 37 1st Qu.: 2864 1st Qu.: 0 1st Qu.:100.2   
## NA's: 23 Median : 3786 Median : 1025 Median :125.0   
## Mean : 4806 Mean : 1570 Mean :136.1   
## 3rd Qu.: 5060 3rd Qu.: 2430 3rd Qu.:158.0   
## Max. :72529 Max. :24000 Max. :550.0   
## NA's :5   
## Loan\_Amount\_Term Credit\_History Property\_Area  
## Min. : 6.0 Min. :0.0000 Rural :111   
## 1st Qu.:360.0 1st Qu.:1.0000 Semiurban:116   
## Median :360.0 Median :1.0000 Urban :140   
## Mean :342.5 Mean :0.8254   
## 3rd Qu.:360.0 3rd Qu.:1.0000   
## Max. :480.0 Max. :1.0000   
## NA's :6 NA's :29

colSums(is.na(crp\_test))

## Loan\_ID Gender Married Dependents   
## 0 11 0 10   
## Education Self\_Employed ApplicantIncome CoapplicantIncome   
## 0 23 0 0   
## LoanAmount Loan\_Amount\_Term Credit\_History Property\_Area   
## 5 6 29 0

## Missing values imputation on Test data set

crp\_test$Credit\_History<-as.factor(crp\_test$Credit\_History)  
crp\_test$LoanAmount[is.na(crp\_test$LoanAmount)] <- median(crp\_test$LoanAmount,na.rm=T)  
crp\_test$Loan\_Amount\_Term[is.na(crp\_test$Loan\_Amount\_Term)] <- median(crp\_test$Loan\_Amount\_Term,na.rm=T)  
crp\_test$Credit\_History[is.na(crp\_test$Credit\_History)]<-1  
crp\_test$Self\_Employed[is.na(crp\_test$Self\_Employed)]<-"No"  
crp\_test$Dependents[is.na(crp\_test$Dependents)]<-0  
crp\_test$Gender[is.na(crp\_test$Gender)]<-"Male"  
summary(crp\_test)

## Loan\_ID Gender Married Dependents Education   
## LP001015: 1 Female: 70 No :134 0 :210 Graduate :283   
## LP001022: 1 Male :297 Yes:233 1 : 58 Not Graduate: 84   
## LP001031: 1 2 : 59   
## LP001035: 1 3+: 40   
## LP001051: 1   
## LP001054: 1   
## (Other) :361   
## Self\_Employed ApplicantIncome CoapplicantIncome LoanAmount   
## No :330 Min. : 0 Min. : 0 Min. : 28.0   
## Yes: 37 1st Qu.: 2864 1st Qu.: 0 1st Qu.:101.0   
## Median : 3786 Median : 1025 Median :125.0   
## Mean : 4806 Mean : 1570 Mean :136.0   
## 3rd Qu.: 5060 3rd Qu.: 2430 3rd Qu.:157.5   
## Max. :72529 Max. :24000 Max. :550.0   
##   
## Loan\_Amount\_Term Credit\_History Property\_Area  
## Min. : 6.0 0: 59 Rural :111   
## 1st Qu.:360.0 1:308 Semiurban:116   
## Median :360.0 Urban :140   
## Mean :342.8   
## 3rd Qu.:360.0   
## Max. :480.0   
##

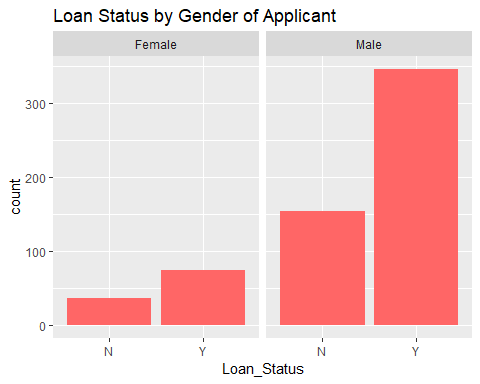
colSums(is.na(crp\_test))

## Loan\_ID Gender Married Dependents   
## 0 0 0 0   
## Education Self\_Employed ApplicantIncome CoapplicantIncome   
## 0 0 0 0   
## LoanAmount Loan\_Amount\_Term Credit\_History Property\_Area   
## 0 0 0 0

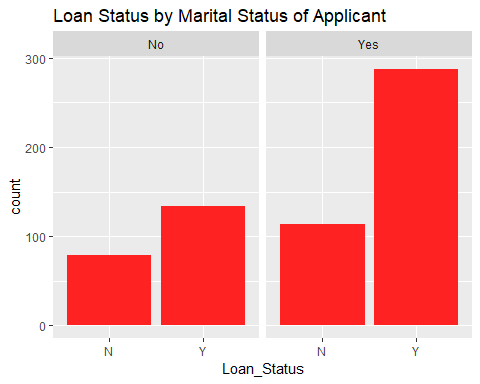
## Data Visualistaion

### understanding Train data set

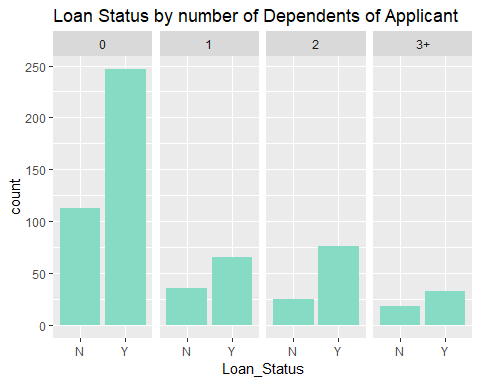
library(ggplot2)  
ggplot(crp, aes(x=Loan\_Status)) + geom\_bar(fill="#FF6666") +   
 facet\_grid(.~Gender) + ggtitle("Loan Status by Gender of Applicant")



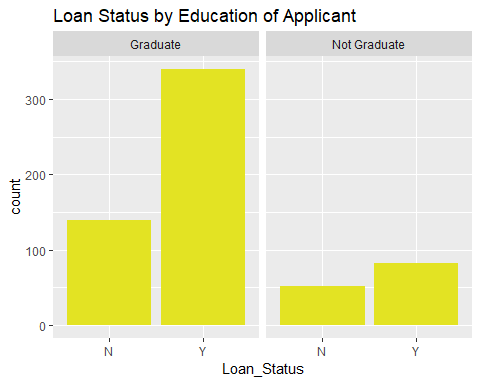
ggplot(crp, aes(x=Loan\_Status))+ geom\_bar(fill="#FF2222") +   
 facet\_grid(.~Married) + ggtitle("Loan Status by Marital Status of Applicant")



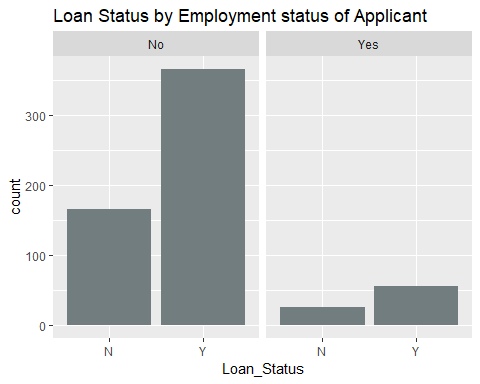
ggplot(crp, aes(x=Loan\_Status)) + geom\_bar(fill="#86DBC4") +  
 facet\_grid(.~Dependents) + ggtitle("Loan Status by number of Dependents of Applicant")



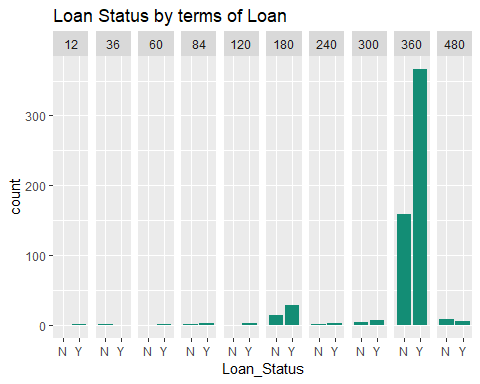
ggplot(crp, aes(x=Loan\_Status)) + geom\_bar(fill="#E3E323") +  
 facet\_grid(.~Education) + ggtitle("Loan Status by Education of Applicant")



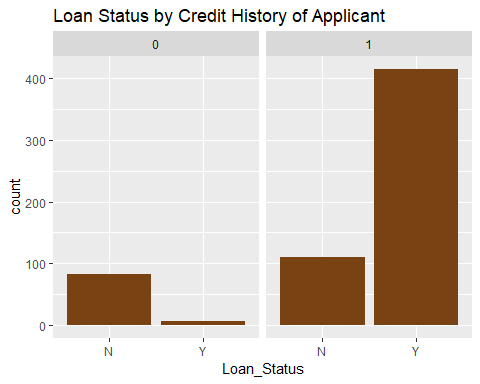
ggplot(crp, aes(x=Loan\_Status)) + geom\_bar(fill="#717d7e") +  
 facet\_grid(.~Self\_Employed) + ggtitle("Loan Status by Employment status of Applicant")



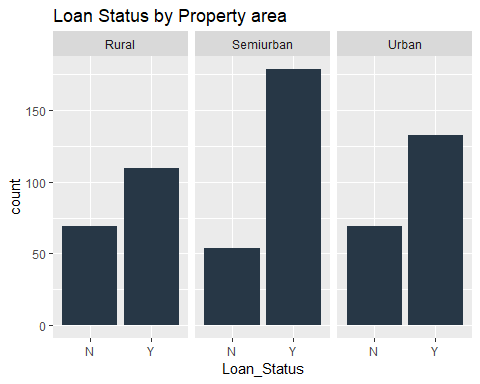
ggplot(crp, aes(x=Loan\_Status)) + geom\_bar(fill="#138d75") +  
 facet\_grid(.~Loan\_Amount\_Term) + ggtitle("Loan Status by terms of Loan")



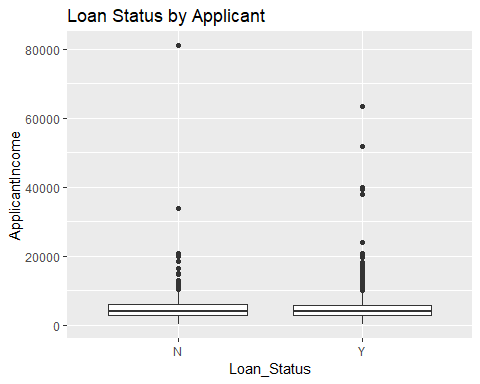
ggplot(crp, aes(x=Loan\_Status)) + geom\_bar(fill="#784212") +  
 facet\_grid(.~Credit\_History) + ggtitle("Loan Status by Credit History of Applicant")



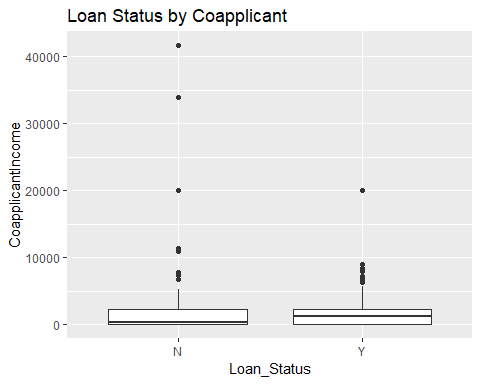
ggplot(crp, aes(x=Loan\_Status)) + geom\_bar(fill="#273746") +  
 facet\_grid(.~Property\_Area) + ggtitle("Loan Status by Property area")



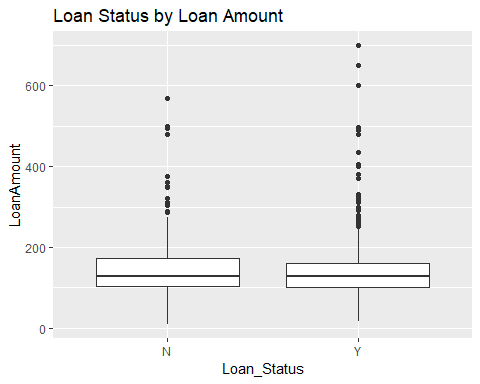
ggplot(crp, aes(x=Loan\_Status, y=ApplicantIncome)) + geom\_boxplot() + ggtitle("Loan Status by Applicant")



ggplot(crp, aes(x=Loan\_Status, y=CoapplicantIncome)) + geom\_boxplot() + ggtitle("Loan Status by Coapplicant")



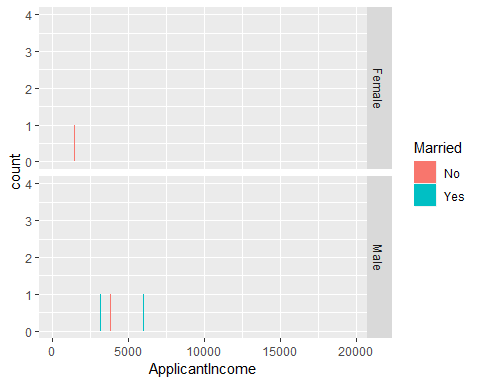
ggplot(crp, aes(x=Loan\_Status, y=LoanAmount)) + geom\_boxplot() + ggtitle("Loan Status by Loan Amount")



### Visualization on Applicant Income and Coapplicant Income:

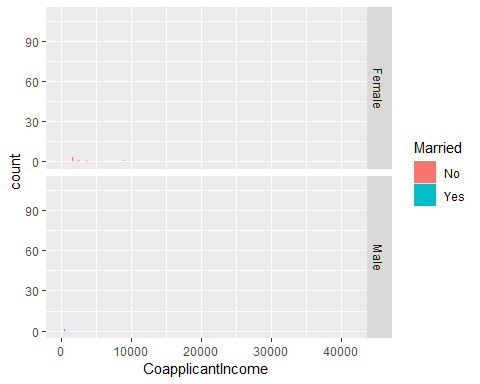
The first variables I will deal with are Applicant Income and Coapplicant Income. Some of the applicants are males, so, presumably, the coapplicants are female and vice versa.

print(ggplot(data=crp[crp$ApplicantIncome<20000,],aes(ApplicantIncome,fill=Married))+geom\_bar(position="dodge")+facet\_grid(Gender~.))



print(ggplot(data=crp[crp$ApplicantIncome<20000,],aes(CoapplicantIncome,fill=Married))+geom\_bar(position="dodge")+facet\_grid(Gender~.))

## Warning: position\_dodge requires non-overlapping x intervals

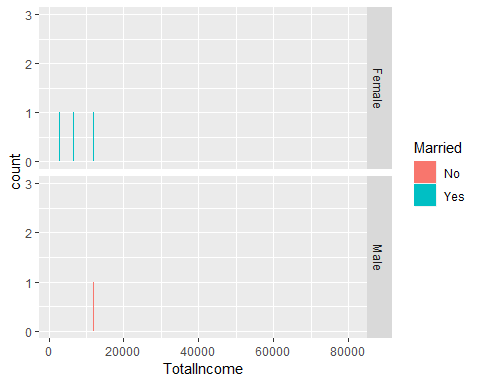


### Adding both Applicant and Coapplicant Income:

I think it might be more intuitive to look at total income of the family, rather than coapplicant income, though for a machine learning algorithm this is unlikely to make any difference.

library(plyr)  
crpFE<-mutate(crp,TotalIncome=ApplicantIncome+CoapplicantIncome)  
print(ggplot(data=crpFE,aes(TotalIncome,fill=Married))+geom\_bar(position="dodge")+facet\_grid(Gender~.))

## Warning: position\_dodge requires non-overlapping x intervals



# Model Building

# logistic Model:

### model

modelp2<-glm(Loan\_Status~Married+Credit\_History+Property\_Area,family=binomial(link="logit"),data=crp)  
summary(modelp2)

##   
## Call:  
## glm(formula = Loan\_Status ~ Married + Credit\_History + Property\_Area,   
## family = binomial(link = "logit"), data = crp)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.0859 -0.3857 0.4910 0.7432 2.5740   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.2756 0.4614 -7.100 1.25e-12 \*\*\*  
## MarriedYes 0.5332 0.2151 2.479 0.01318 \*   
## Credit\_History1 3.8879 0.4155 9.358 < 2e-16 \*\*\*  
## Property\_AreaSemiurban 0.9096 0.2647 3.436 0.00059 \*\*\*  
## Property\_AreaUrban 0.2390 0.2508 0.953 0.34064   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 762.89 on 613 degrees of freedom  
## Residual deviance: 568.62 on 609 degrees of freedom  
## AIC: 578.62  
##   
## Number of Fisher Scoring iterations: 5

### Validation of our model on validation set crp\_v

fitted.resultsp1 <- predict(modelp2,newdata=crp\_v[,-12],type='response')  
fitted.resultsp1 <- ifelse(fitted.resultsp1 > 0.5,1,0)

### Plotting auc curve

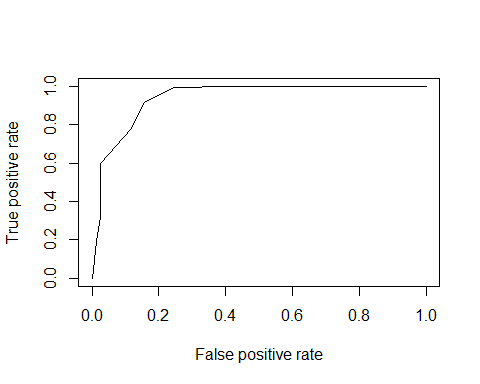
library(ROCR)

## Loading required package: gplots

##   
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':  
##   
## lowess

p2 <- predict(modelp2, newdata=crp\_v[,-12], type="response")  
pr2 <- prediction(p2, crp\_v[,12])  
prf2 <- performance(pr2, measure = "tpr", x.measure = "fpr")  
plot(prf2)



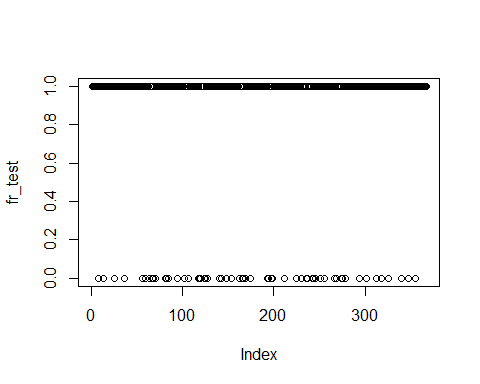
aucp <- performance(pr2, measure = "auc")  
aucp1 <- aucp@y.values[[1]]  
print(aucp1)

## [1] 0.9402821

### Logistic Regression model Accuracy: 0.9402

## Predicting on Test set

fr\_test <- predict(modelp2,newdata=crp\_test,type='response')  
fr\_test <- ifelse(fr\_test > 0.5,1,0)  
plot(fr\_test)



crp\_test<-crp\_test[,-13]  
outcome\_test<-fr\_test  
crp\_test=data.frame(crp\_test,outcome\_test)  
View(crp\_test)  
print(head(crp\_test,4))

## Loan\_ID Gender Married Dependents Education Self\_Employed  
## 1 LP001015 Male Yes 0 Graduate No  
## 2 LP001022 Male Yes 1 Graduate No  
## 3 LP001031 Male Yes 2 Graduate No  
## 4 LP001035 Male Yes 2 Graduate No  
## ApplicantIncome CoapplicantIncome LoanAmount Loan\_Amount\_Term  
## 1 5720 0 110 360  
## 2 3076 1500 126 360  
## 3 5000 1800 208 360  
## 4 2340 2546 100 360  
## Credit\_History Property\_Area outcome\_test  
## 1 1 Urban 1  
## 2 1 Urban 1  
## 3 1 Urban 1  
## 4 1 Urban 1

# SVM model:

## model building

### SVM based based on grid scearch

library(e1071)  
tunesvmcrp=tune(svm,Loan\_Status~.,  
 data=crp,  
 ranges = list(gamma=2^(-1:1),cost=2^(2:9)))  
summary(tunesvmcrp)

##   
## Parameter tuning of 'svm':  
##   
## - sampling method: 10-fold cross validation   
##   
## - best parameters:  
## gamma cost  
## 0.5 4  
##   
## - best performance: 0.2282919   
##   
## - Detailed performance results:  
## gamma cost error dispersion  
## 1 0.5 4 0.2282919 0.07008191  
## 2 1.0 4 0.2331571 0.06052101  
## 3 2.0 4 0.2965891 0.05654963  
## 4 0.5 8 0.2511105 0.08844315  
## 5 1.0 8 0.2542570 0.05427326  
## 6 2.0 8 0.3031465 0.05551376  
## 7 0.5 16 0.2641988 0.07296777  
## 8 1.0 16 0.2689582 0.05566406  
## 9 2.0 16 0.3097039 0.06447675  
## 10 0.5 32 0.2656795 0.07305565  
## 11 1.0 32 0.2771814 0.06573552  
## 12 2.0 32 0.3096774 0.06157373  
## 13 0.5 64 0.2884717 0.07849886  
## 14 1.0 64 0.2853517 0.07204586  
## 15 2.0 64 0.3063723 0.05726084  
## 16 0.5 128 0.3080645 0.06943159  
## 17 1.0 128 0.2884981 0.06362163  
## 18 2.0 128 0.3112639 0.06060605  
## 19 0.5 256 0.3097039 0.07543480  
## 20 1.0 256 0.2884188 0.06997115  
## 21 2.0 256 0.3128768 0.05126305  
## 22 0.5 512 0.3179799 0.08724805  
## 23 1.0 512 0.2950026 0.06740431  
## 24 2.0 512 0.3063458 0.04295748

### gamma=2 , cost=4

library(e1071)  
classifier2<-svm(formula =Loan\_Status~.,data=crp,type = 'C-classification',kernel="linear",gamma=2,cost=4)  
summary(classifier2)

##   
## Call:  
## svm(formula = Loan\_Status ~ ., data = crp, type = "C-classification",   
## kernel = "linear", gamma = 2, cost = 4)  
##   
##   
## Parameters:  
## SVM-Type: C-classification   
## SVM-Kernel: linear   
## cost: 4   
## gamma: 2   
##   
## Number of Support Vectors: 330  
##   
## ( 194 136 )  
##   
##   
## Number of Classes: 2   
##   
## Levels:   
## N Y

# validation of our model using validation set

### if type = response is not mentioned it will take log(odd(probability)), its for backtransforming it to categorical variable

fitted.resultssvmcrp2 <- predict(classifier2,newdata=crp\_v[,-12])

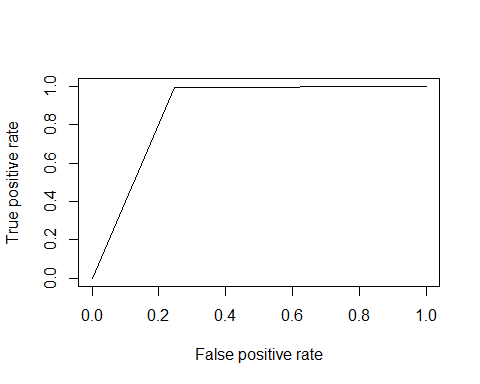
#Confusion matrix  
svmcf2<-table(fitted.resultssvmcrp2 , crp\_v[,12])  
  
  
#function for accuracy for logistic linear  
acc<-function(svmcf1){  
 Totp<-svmcf2[2,1]+svmcf2[2,2]  
 TP<-svmcf2[2,2]  
 c<-TP/Totp  
 c  
}  
acc(svmcf2)

## [1] 0.9383117

## SVM model confusion matrix accuracy : 0.9383

## plotting auc curve for linear i.e classifier2

svmp <- predict(classifier2, newdata=crp\_v[,-12])  
svmp <- as.numeric(svmp)  
crp\_v$outcome <-as.numeric(crp\_v$outcome)  
pr2 <- prediction(svmp, crp\_v[,12])  
prf2 <- performance(pr2, measure = "tpr", x.measure = "fpr")  
plot(prf2)



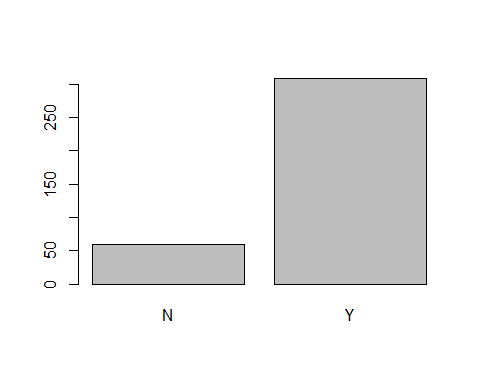
aucsvmp <- performance(pr2, measure = "auc")  
aucsvmp1 <- aucsvmp@y.values[[1]]  
aucsvmp1

## [1] 0.8748992

## SVM model accuracy = 0.8748

## predict for test set

svmfr\_test <- predict(classifier2,newdata=crp\_test)  
fr\_test <- ifelse(fr\_test > 0.5,1,0)  
plot(svmfr\_test)



crp\_test<-crp\_test[,-13]  
outcomesvm\_test<-svmfr\_test  
crp\_test=data.frame(crp\_test,outcomesvm\_test)  
print(head(crp\_test,4))

## Loan\_ID Gender Married Dependents Education Self\_Employed  
## 1 LP001015 Male Yes 0 Graduate No  
## 2 LP001022 Male Yes 1 Graduate No  
## 3 LP001031 Male Yes 2 Graduate No  
## 4 LP001035 Male Yes 2 Graduate No  
## ApplicantIncome CoapplicantIncome LoanAmount Loan\_Amount\_Term  
## 1 5720 0 110 360  
## 2 3076 1500 126 360  
## 3 5000 1800 208 360  
## 4 2340 2546 100 360  
## Credit\_History Property\_Area outcomesvm\_test  
## 1 1 Urban Y  
## 2 1 Urban Y  
## 3 1 Urban Y  
## 4 1 Urban Y

# Naive Bayes

## model building on train data

library(e1071)  
crpnaivem1 <- naiveBayes(Loan\_Status~., data=crp)  
summary(crpnaivem1)

## Length Class Mode   
## apriori 2 table numeric   
## tables 11 -none- list   
## levels 2 -none- character  
## call 4 -none- call

## validation data

crp\_v<-read.csv("D:/imarticus/R/logistic regression/Logistic regression and SVM\_Case study/R\_Module\_Day\_8.2\_Credit\_Risk\_Validate\_data.csv",na.strings=c(""," ","NA"))  
crp\_v<-crp\_v[,-1]  
crp\_v$Credit\_History<-as.factor(crp\_v$Credit\_History)  
crp\_v$LoanAmount[is.na(crp\_v$LoanAmount)] <- median(crp\_v$LoanAmount,na.rm=T)  
crp\_v$Loan\_Amount\_Term[is.na(crp\_v$Loan\_Amount\_Term)] <- median(crp\_v$Loan\_Amount\_Term,na.rm=T)  
crp\_v$Credit\_History[is.na(crp\_v$Credit\_History)]<-1  
crp\_v$Self\_Employed[is.na(crp\_v$Self\_Employed)]<-"No"  
crp\_v$Dependents[is.na(crp\_v$Dependents)]<-0  
crp\_v$Gender[is.na(crp\_v$Gender)]<-"Male"  
  
CRPNaive\_pred = predict(crpnaivem1, newdata = crp\_v)  
cmCRPNaive = table(CRPNaive\_pred, crp\_v$outcome)  
library(caret)

## Loading required package: lattice

cfNaiveCRP1<-confusionMatrix(CRPNaive\_pred,crp\_v$outcome)  
cfNaiveCRP1

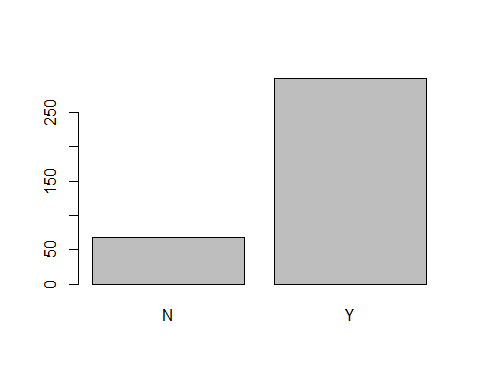
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction N Y  
## N 58 10  
## Y 19 280  
##   
## Accuracy : 0.921   
## 95% CI : (0.8885, 0.9464)  
## No Information Rate : 0.7902   
## P-Value [Acc > NIR] : 7.364e-12   
##   
## Kappa : 0.751   
## Mcnemar's Test P-Value : 0.1374   
##   
## Sensitivity : 0.7532   
## Specificity : 0.9655   
## Pos Pred Value : 0.8529   
## Neg Pred Value : 0.9365   
## Prevalence : 0.2098   
## Detection Rate : 0.1580   
## Detection Prevalence : 0.1853   
## Balanced Accuracy : 0.8594   
##   
## 'Positive' Class : N   
##

#0.921

## Naive Bayes model accuracy : 0.921

## predict for test set

naivefr\_test <- predict(crpnaivem1,newdata=crp\_test)  
plot(naivefr\_test)



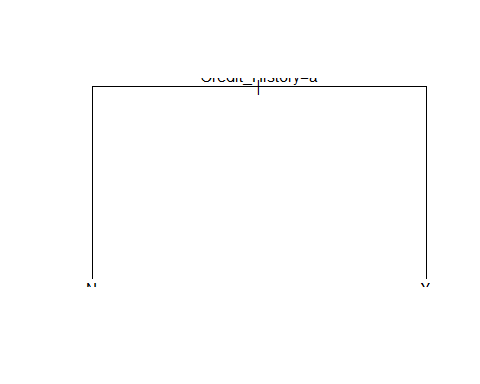
crp\_test<-crp\_test[,-13]  
naiveoutcome\_test<-naivefr\_test  
crp\_test=data.frame(crp\_test,naiveoutcome\_test)  
print(head(crp\_test,4))

## Loan\_ID Gender Married Dependents Education Self\_Employed  
## 1 LP001015 Male Yes 0 Graduate No  
## 2 LP001022 Male Yes 1 Graduate No  
## 3 LP001031 Male Yes 2 Graduate No  
## 4 LP001035 Male Yes 2 Graduate No  
## ApplicantIncome CoapplicantIncome LoanAmount Loan\_Amount\_Term  
## 1 5720 0 110 360  
## 2 3076 1500 126 360  
## 3 5000 1800 208 360  
## 4 2340 2546 100 360  
## Credit\_History Property\_Area naiveoutcome\_test  
## 1 1 Urban Y  
## 2 1 Urban Y  
## 3 1 Urban Y  
## 4 1 Urban Y

# DECESION TREES

## Model Building

library(rpart)  
modelCRPDT1 <- rpart(formula = Loan\_Status ~., data=crp)  
plot(modelCRPDT1)  
text(modelCRPDT1)



## validation

CRPDT\_pred = predict(modelCRPDT1, newdata = crp\_v, type = 'class')  
  
# confusion matrix  
cmCRPDT = table(CRPDT\_pred, crp\_v$outcome)  
cmCRPDT

##   
## CRPDT\_pred N Y  
## N 58 1  
## Y 19 289

library(caret)  
cfCRPDT1<-confusionMatrix(CRPDT\_pred,crp\_v$outcome)  
cfCRPDT1

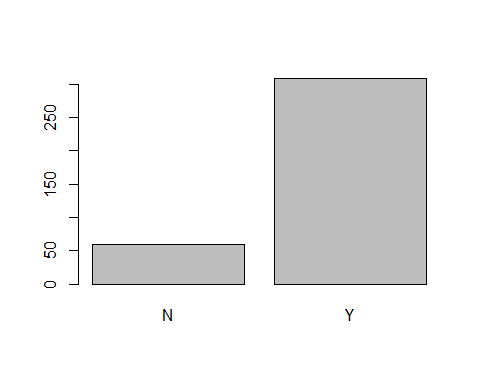
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction N Y  
## N 58 1  
## Y 19 289  
##   
## Accuracy : 0.9455   
## 95% CI : (0.9171, 0.9664)  
## No Information Rate : 0.7902   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.8202   
## Mcnemar's Test P-Value : 0.0001439   
##   
## Sensitivity : 0.7532   
## Specificity : 0.9966   
## Pos Pred Value : 0.9831   
## Neg Pred Value : 0.9383   
## Prevalence : 0.2098   
## Detection Rate : 0.1580   
## Detection Prevalence : 0.1608   
## Balanced Accuracy : 0.8749   
##   
## 'Positive' Class : N   
##

#0.9455

## Decesion Tree accuracy : 0.9455

## predict for test set

DTCRPfr\_test <- predict(modelCRPDT1,newdata=crp\_test,type='class')  
plot(DTCRPfr\_test)



View(DTCRPfr\_test)

crp\_test<-crp\_test[,-13]  
DTCRPoutcome\_test<-DTCRPfr\_test  
crp\_test=data.frame(crp\_test,DTCRPoutcome\_test)  
print(head(crp\_test,4))

## Loan\_ID Gender Married Dependents Education Self\_Employed  
## 1 LP001015 Male Yes 0 Graduate No  
## 2 LP001022 Male Yes 1 Graduate No  
## 3 LP001031 Male Yes 2 Graduate No  
## 4 LP001035 Male Yes 2 Graduate No  
## ApplicantIncome CoapplicantIncome LoanAmount Loan\_Amount\_Term  
## 1 5720 0 110 360  
## 2 3076 1500 126 360  
## 3 5000 1800 208 360  
## 4 2340 2546 100 360  
## Credit\_History Property\_Area DTCRPoutcome\_test  
## 1 1 Urban Y  
## 2 1 Urban Y  
## 3 1 Urban Y  
## 4 1 Urban Y

# RANDOM FOREST

## fitting random forest classification to the training set

library(randomForest)

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':  
##   
## margin

randomforestCRP = randomForest(x = crp[-12],y = crp$Loan\_Status, ntree = 50)  
print(randomforestCRP)

##   
## Call:  
## randomForest(x = crp[-12], y = crp$Loan\_Status, ntree = 50)   
## Type of random forest: classification  
## Number of trees: 50  
## No. of variables tried at each split: 3  
##   
## OOB estimate of error rate: 19.87%  
## Confusion matrix:  
## N Y class.error  
## N 97 95 0.49479167  
## Y 27 395 0.06398104

## predicting the test set results

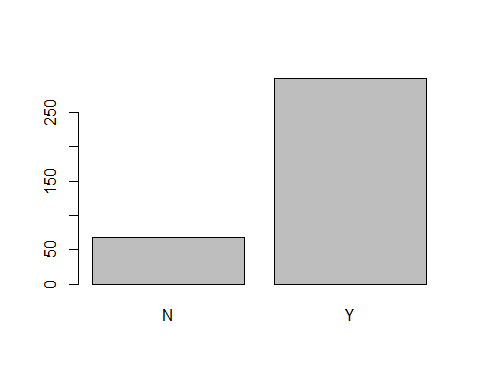
crp\_v<-read.csv("D:/imarticus/R/logistic regression/Logistic regression and SVM\_Case study/R\_Module\_Day\_8.2\_Credit\_Risk\_Validate\_data.csv",na.strings=c(""," ","NA"))  
crp\_v<-crp\_v[,-1]  
crp\_v$Credit\_History<-as.factor(crp\_v$Credit\_History)  
crp\_v$LoanAmount[is.na(crp\_v$LoanAmount)] <- median(crp\_v$LoanAmount,na.rm=T)  
crp\_v$Loan\_Amount\_Term[is.na(crp\_v$Loan\_Amount\_Term)] <- median(crp\_v$Loan\_Amount\_Term,na.rm=T)  
crp\_v$Credit\_History[is.na(crp\_v$Credit\_History)]<-1  
crp\_v$Self\_Employed[is.na(crp\_v$Self\_Employed)]<-"No"  
crp\_v$Dependents[is.na(crp\_v$Dependents)]<-0  
crp\_v$Gender[is.na(crp\_v$Gender)]<-"Male"  
randomCRP\_pred = predict(randomforestCRP,newdata = crp\_v[,-12])  
#making the confucion matrix  
cmCRPrandom = table(crp\_v$outcome,randomCRP\_pred)  
acc(cm)

## [1] 0.9383117

## Random forest Accuracy : 0.9166

## predict for test set

randomCRPfr\_test <- predict(randomforestCRP,newdata=crp\_test)  
plot(randomCRPfr\_test)



crp\_test<-crp\_test[,-13]  
randomCRPoutcome\_test<-randomCRPfr\_test  
crp\_test=data.frame(crp\_test,randomCRPoutcome\_test)  
print(head(crp\_test))

## Loan\_ID Gender Married Dependents Education Self\_Employed  
## 1 LP001015 Male Yes 0 Graduate No  
## 2 LP001022 Male Yes 1 Graduate No  
## 3 LP001031 Male Yes 2 Graduate No  
## 4 LP001035 Male Yes 2 Graduate No  
## 5 LP001051 Male No 0 Not Graduate No  
## 6 LP001054 Male Yes 0 Not Graduate Yes  
## ApplicantIncome CoapplicantIncome LoanAmount Loan\_Amount\_Term  
## 1 5720 0 110 360  
## 2 3076 1500 126 360  
## 3 5000 1800 208 360  
## 4 2340 2546 100 360  
## 5 3276 0 78 360  
## 6 2165 3422 152 360  
## Credit\_History Property\_Area randomCRPoutcome\_test  
## 1 1 Urban Y  
## 2 1 Urban Y  
## 3 1 Urban Y  
## 4 1 Urban Y  
## 5 1 Urban Y  
## 6 1 Urban Y

# Conclusion:

## Model Accuracy:

1.Logistic Regression model Accuracy: 0.9402  
2.SVM model confusion matrix accuracy : 0.9383  
3.Naive Bayes model accuracy : 0.921  
4.Decesion Tree accuracy : 0.9455  
5.Random forest Accuracy : 0.9166

## CHAMPION MODEL: DECESION TREE AND LOGISTIC REGRESSION

From Logistic regression we can conclude that Significant Varibales to predict if a person is eligible for loan or not are **Martial Status**, **Credit History**, **Property Area**.