DataCleaning\_and\_Merging.R

Admin

Thu Aug 22 13:22:03 2019

###############################  
#Data Cleaning and Preparation#  
###############################  
  
library(openxlsx)  
library(rio)  
library(tidyr)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

#Reading the Profit Loss data  
branch\_pl <- import("F:/Vivek/Branch Pofitability wo allocationsFY18-19 (003).xls", skip=2)  
branch\_pl <- branch\_pl[-c(56:58),]  
branch\_pl <- branch\_pl %>% separate("Branch",c("BranchCode","branch"), remove = TRUE) #To split the data on hyphen, space or other character  
str(branch\_pl)

## 'data.frame': 55 obs. of 17 variables:  
## $ BranchCode : chr "001" "002" "003" "004" ...  
## $ branch : chr "DUTTAPULIA" "HANSKHALI" "MATIKUMRA" "LAUHATI" ...  
## $ Disbursement : num 48385000 50067800 46877800 54197000 61663000 ...  
## $ Outstanding : num 30689334 31743931 32385961 28679093 34690797 ...  
## $ Interest Collection: num 5323078 5472744 6722031 6404707 6663294 ...  
## $ Reverse : num 256071 234283 328214 26664 34410 ...  
## $ Booked : num 213939 298481 26215 132217 144133 ...  
## $ Utkarsh BC : num 0 0 2357464 0 0 ...  
## $ Mas/Assignment : num 1187421 1228225 0 1109642 1342245 ...  
## $ Own Income : num 4094990 4310182 4064033 5402084 5432237 ...  
## $ P.Fees : num 455040 470677 0 425233 514371 ...  
## $ Utkarsh Commission : num 0 0 1098541 0 0 ...  
## $ MAS Commission : num 357937 370237 0 334491 404607 ...  
## $ Other Income : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Total Income : num 4907966 5151095 5162574 6161808 6351215 ...  
## $ Expenditure : num 4649638 4370939 2039140 4438315 5342270 ...  
## $ Profit/(Loss) : num 258328 780157 3123434 1723493 1008945 ...

branch\_pl <- branch\_pl[,-c(3:14)]  
names(branch\_pl)[5] <- "finalvalue"  
colnames(branch\_pl) <- tolower(make.names(colnames(branch\_pl)))  
branch\_pl$result <- ifelse(branch\_pl$finalvalue > 0,"Profit","Loss")  
  
#Reading demographic data  
demographic <- read.xlsx("Masterdata.xlsx",sheet = "Demographic")  
str(demographic)

## 'data.frame': 138 obs. of 7 variables:  
## $ district : chr "North Twenty Four Parganas" "South Twenty Four Parganas" "Barddhaman" "Murshidabad" ...  
## $ population: chr "10,009,781" "8,161,961" "7,717,563" "7,103,807" ...  
## $ increase : chr "12.04 %" "18.17 %" "11.92 %" "21.09 %" ...  
## $ sex.ratio : chr "955" "956" "945" "958" ...  
## $ literacy : chr "84.06 %" "77.51 %" "76.21 %" "66.59 %" ...  
## $ density : chr "2445" "819" "1099" "1334" ...  
## $ state : chr "westBengal" "westBengal" "westBengal" "westBengal" ...

demographic$increase <- as.numeric(gsub("\\%","",demographic$increase)) #removing % sign  
demographic$literacy <- as.numeric(gsub("\\%","",demographic$literacy)) #removing % sign  
demographic$population <- as.numeric(gsub("\\,","",demographic$population)) #removing comma in population data  
demographic <- transform(demographic, sex.ratio = as.numeric(sex.ratio), density = as.numeric(density))  
demographic[demographic=="westBengal"]<- "West Bengal"  
demographic[demographic=="bihar"]<- "Bihar"  
demographic[demographic=="Odissa"]<- "Odisha"  
demographic[demographic=="jharkhand"]<- "Jharkhand"  
  
#Reading Unemployment Data  
unemp <-read.xlsx("Masterdata.xlsx", sheet = "UnempData")  
str(unemp)

## 'data.frame': 28 obs. of 2 variables:  
## $ Unemployment.Rate.(%): chr "States (India)" "Andhra Pradesh" "Assam" "Bihar" ...  
## $ Unemployment.Rate.(%): chr "Jun 2019" "3.6" "5.5" "10.1" ...

names(unemp)[1]<- "state"  
names(unemp)[2]<- "rate"  
unemp <- unemp[-c(1,29),]  
unemp$rate <- as.numeric(unemp$rate)

## Warning: NAs introduced by coercion

#Reading MFI Growth Data  
mfigrowth <-read.xlsx("Masterdata.xlsx", sheet = "MFIGrowth")  
str(mfigrowth)

## 'data.frame': 33 obs. of 3 variables:  
## $ Region.State : chr "A & N Islands (UT)" "Andhra Pradesh" "Arunachal Pradesh" "Assam" ...  
## $ Total.No..of.SHGs: num 4998 884508 3351 292071 224469 ...  
## $ Savings.Amount : num 125 262950 256 9943 29667 ...

names(mfigrowth)[1]<- "state"  
names(mfigrowth)[2]<- "count.shg"  
names(mfigrowth)[3]<- "saving.amount.lakh.rs"  
  
#Reading GSDP Data  
gsdp <- read.xlsx("Masterdata.xlsx",sheet = "GSDPPCData")  
str(gsdp)

## 'data.frame': 35 obs. of 5 variables:  
## $ State : chr "State" "Goa" "Delhi" "Sikkim" ...  
## $ GSDP.per.capita..INR.at.Current.prices. : chr "18-19" "-" "402,172" "-" ...  
## $ GSDP.per.capita..INR.at.Current.prices..1: chr "17-18" "-" "362,790" "340,705" ...  
## $ GSDP.per.capita..INR.at.Current.prices..2: chr "16-17" "413,334" "329,836" "309,428" ...  
## $ GDP.per.capita...billion. : chr "2016" "6,045" "4,824" "4,526" ...

gsdp <- gsdp[-c(1,36),-c(2:3)]  
names(gsdp)[1] <- "state"  
names(gsdp)[2] <- "percapita16-17"  
names(gsdp)[3] <- "billion.percapita"  
gsdp$`percapita16-17` <- as.numeric(gsub("\\,","",gsdp$`percapita16-17`))  
gsdp$billion.percapita <- as.numeric(gsub("\\,","",gsdp$billion.percapita))  
  
  
#Reading State wise branchList  
branch <- import("State and District wise branch list.xlsx")  
str(branch)

## 'data.frame': 59 obs. of 5 variables:  
## $ Sl No. : num 1 2 3 4 5 6 7 8 9 10 ...  
## $ Branch Code : chr "000" "001" "002" "003" ...  
## $ Branch Name : chr "HEAD OFFICE" "DUTTAPULIA" "HANSKHALI" "MATIKUMRA" ...  
## $ District Name: chr "NORTH 24 PARGANAS" "NADIA" "NADIA" "NADIA" ...  
## $ State : chr "West Bengal" "West Bengal" "West Bengal" "West Bengal" ...

colnames(branch)=tolower(make.names(colnames(branch)))  
names(branch)[4] <- "district"  
unique(demographic$district)

## [1] "North Twenty Four Parganas" "South Twenty Four Parganas"  
## [3] "Barddhaman" "Murshidabad"   
## [5] "Paschim Medinipur" "Hugli"   
## [7] "Nadia" "Purba Medinipur"   
## [9] "Haora" "Kolkata"   
## [11] "Maldah" "Jalpaiguri"   
## [13] "Bankura" "Birbhum"   
## [15] "Uttar Dinajpur" "Puruliya"   
## [17] "Koch Bihar" "Darjiling"   
## [19] "Dakshin Dinajpur" "Ranchi"   
## [21] "Dhanbad" "Giridih"   
## [23] "Purbi Singhbhum" "Bokaro"   
## [25] "Palamu" "Hazaribagh"   
## [27] "Pashchimi Singhbhum" "Deoghar"   
## [29] "Garhwa" "Dumka"   
## [31] "Godda" "Sahibganj"   
## [33] "Saraikela Kharsawan" "Chatra"   
## [35] "Gumla" "Ramgarh"   
## [37] "Pakur" "Jamtara"   
## [39] "Latehar" "Kodarma"   
## [41] "Simdega" "Khunti"   
## [43] "Lohardaga" "Ganjam"   
## [45] "Cuttack" "Mayurbhanj"   
## [47] "Baleshwar" "Khordha"   
## [49] "Sundargarh" "Jajapur"   
## [51] "Kendujhar" "Puri"   
## [53] "Balangir" "Kalahandi"   
## [55] "Bhadrak" "Bargarh"   
## [57] "Kendrapara" "Koraput"   
## [59] "Anugul" "Nabarangapur"   
## [61] "Dhenkanal" "Jagatsinghapur"   
## [63] "Sambalpur" "Rayagada"   
## [65] "Nayagarh" "Kandhamal"   
## [67] "Malkangiri" "Nuapada"   
## [69] "Subarnapur" "Jharsuguda"   
## [71] "Gajapati" "Baudh"   
## [73] "Debagarh" "Araria"   
## [75] "Arwal" "Aurangabad"   
## [77] "Banka" "Begusarai"   
## [79] "Bhagalpur" "Bhojpur"   
## [81] "Buxar" "Darbhanga"   
## [83] "Gaya" "Gopalganj"   
## [85] "Jamui" "Jehanabad"   
## [87] "Kaimur" "Katihar"   
## [89] "Khagaria" "Kishanganj"   
## [91] "Lakhisarai" "Madhepura"   
## [93] "Madhubani" "Munger"   
## [95] "Muzaffarpur" "Nalanda"   
## [97] "Nawada" "Pashchim Champaran"   
## [99] "Patna" "Purbi Champaran"   
## [101] "Purnia" "Rohtas"   
## [103] "Saharsa" "Samastipur"   
## [105] "Saran" "Sheikhpura"   
## [107] "Sheohar" "Sitamarhi"   
## [109] "Siwan" "Supaul"   
## [111] "Vaishali" "Nagaon"   
## [113] "Dhubri" "Sonitpur"   
## [115] "Cachar" "Barpeta"   
## [117] "Kamrup" "Tinsukia"   
## [119] "Dibrugarh" "Kamrup Metropolitan"   
## [121] "Karimganj" "Sivasagar"   
## [123] "Jorhat" "Golaghat"   
## [125] "Lakhimpur" "Goalpara"   
## [127] "Morigaon" "Karbi Anglong"   
## [129] "Baksa" "Darrang"   
## [131] "Kokrajhar" "Udalguri"   
## [133] "Nalbari" "Bongaigaon"   
## [135] "Dhemaji" "Hailakandi"   
## [137] "Chirang" "Dima Hasao"

unique(branch$district)

## [1] "NORTH 24 PARGANAS" "NADIA" "NORTH DINAJPUR"   
## [4] "MURSHIDABAD" "SOUTH DINAJPUR" "KISHANGANJ"   
## [7] "PURNIA" "PURI" "ARARIA"   
## [10] "JAMTARA" "CUTTACK" "SOUTH 24 PARGANAS"  
## [13] "KATIHAR" "BHADRAK" "DHANBAD"

demographic$district <- toupper(x = demographic$district)  
demographic$district <- ifelse(demographic$district == "NORTH TWENTY FOUR PARGANAS","NORTH 24 PARGANAS",demographic$district)  
demographic$district <- ifelse(demographic$district == "SOUTH TWENTY FOUR PARGANAS","SOUTH 24 PARGANAS",demographic$district)  
  
#Merging to data frames  
d1 <- left\_join(branch,demographic, by = "district")  
  
#Imputing missing values  
  
summary(d1)

## sl.no. branch.code branch.name district   
## Min. : 1.0 Length:59 Length:59 Length:59   
## 1st Qu.:15.5 Class :character Class :character Class :character   
## Median :30.0 Mode :character Mode :character Mode :character   
## Mean :30.0   
## 3rd Qu.:44.5   
## Max. :59.0   
##   
## state.x population increase sex.ratio   
## Length:59 Min. : 791042 Min. :11.99 Min. :909.0   
## Class :character 1st Qu.: 1698730 1st Qu.:12.22 1st Qu.:947.0   
## Mode :character Median : 5167600 Median :13.05 Median :950.0   
## Mean : 4602471 Mean :18.84 Mean :947.6   
## 3rd Qu.: 7103807 3rd Qu.:28.33 3rd Qu.:956.0   
## Max. :10009781 Max. :30.40 Max. :981.0   
## NA's :6 NA's :6 NA's :6   
## literacy density state.y   
## Min. :51.08 Min. : 437 Length:59   
## 1st Qu.:55.46 1st Qu.: 897 Class :character   
## Median :74.97 Median :1316 Mode :character   
## Mean :69.71 Mean :1203   
## 3rd Qu.:77.51 3rd Qu.:1316   
## Max. :85.50 Max. :2445   
## NA's :6 NA's :6

for(i in 1:ncol(d1)){  
 if (is.numeric(d1[,i]) == TRUE){  
 d1[is.na(d1[,i]),i] <- median(d1[,i], na.rm = TRUE)  
 }  
 }  
  
datafinal <- d1[-c(1),-c(1,11)]  
str(datafinal)

## 'data.frame': 58 obs. of 9 variables:  
## $ branch.code: chr "001" "002" "003" "004" ...  
## $ branch.name: chr "DUTTAPULIA" "HANSKHALI" "MATIKUMRA" "LAUHATI" ...  
## $ district : chr "NADIA" "NADIA" "NADIA" "NORTH 24 PARGANAS" ...  
## $ state.x : chr "West Bengal" "West Bengal" "West Bengal" "West Bengal" ...  
## $ population : num 5167600 5167600 5167600 10009781 5167600 ...  
## $ increase : num 12.2 12.2 12.2 12 13.1 ...  
## $ sex.ratio : num 947 947 947 955 950 947 947 958 958 955 ...  
## $ literacy : num 75 75 75 84.1 75 ...  
## $ density : num 1316 1316 1316 2445 1316 ...

names(datafinal)[2] <- "branch"   
names(datafinal)[4] <- "state"  
head(datafinal)

## branch.code branch district state population increase  
## 2 001 DUTTAPULIA NADIA West Bengal 5167600 12.22  
## 3 002 HANSKHALI NADIA West Bengal 5167600 12.22  
## 4 003 MATIKUMRA NADIA West Bengal 5167600 12.22  
## 5 004 LAUHATI NORTH 24 PARGANAS West Bengal 10009781 12.04  
## 6 005 KARNAJORA NORTH DINAJPUR West Bengal 5167600 13.05  
## 7 006 ARANGHATA NADIA West Bengal 5167600 12.22  
## sex.ratio literacy density  
## 2 947 74.97 1316  
## 3 947 74.97 1316  
## 4 947 74.97 1316  
## 5 955 84.06 2445  
## 6 950 74.97 1316  
## 7 947 74.97 1316

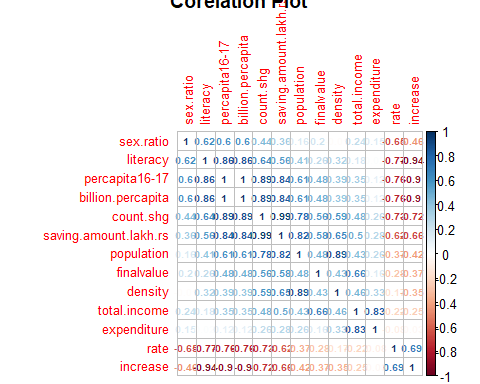
#Combining demographic data and Branch level profit data  
  
finaldata <- datafinal %>%  
 left\_join(branch\_pl, by = "branch")%>%  
 left\_join(unemp,by = "state")%>%  
 left\_join(mfigrowth,by = "state")%>%  
 left\_join(gsdp,by = "state")  
  
head(finaldata)

## branch.code branch district state population increase  
## 1 001 DUTTAPULIA NADIA West Bengal 5167600 12.22  
## 2 002 HANSKHALI NADIA West Bengal 5167600 12.22  
## 3 003 MATIKUMRA NADIA West Bengal 5167600 12.22  
## 4 004 LAUHATI NORTH 24 PARGANAS West Bengal 10009781 12.04  
## 5 005 KARNAJORA NORTH DINAJPUR West Bengal 5167600 13.05  
## 6 006 ARANGHATA NADIA West Bengal 5167600 12.22  
## sex.ratio literacy density branchcode total.income expenditure  
## 1 947 74.97 1316 001 4907966 4649638  
## 2 947 74.97 1316 002 5151095 4370939  
## 3 947 74.97 1316 003 5162574 2039140  
## 4 955 84.06 2445 004 6161808 4438315  
## 5 950 74.97 1316 005 6351215 5342270  
## 6 947 74.97 1316 006 5641977 4542549  
## finalvalue result rate count.shg saving.amount.lakh.rs percapita16-17  
## 1 258328.5 Profit 6.8 760941 127348 91114  
## 2 780156.6 Profit 6.8 760941 127348 91114  
## 3 3123434.1 Profit 6.8 760941 127348 91114  
## 4 1723492.9 Profit 6.8 760941 127348 91114  
## 5 1008944.6 Profit 6.8 760941 127348 91114  
## 6 1099427.5 Profit 6.8 760941 127348 91114  
## billion.percapita  
## 1 1333  
## 2 1333  
## 3 1333  
## 4 1333  
## 5 1333  
## 6 1333

finaldata <- finaldata[,-c(1,10)]  
str(finaldata)

## 'data.frame': 58 obs. of 17 variables:  
## $ branch : chr "DUTTAPULIA" "HANSKHALI" "MATIKUMRA" "LAUHATI" ...  
## $ district : chr "NADIA" "NADIA" "NADIA" "NORTH 24 PARGANAS" ...  
## $ state : chr "West Bengal" "West Bengal" "West Bengal" "West Bengal" ...  
## $ population : num 5167600 5167600 5167600 10009781 5167600 ...  
## $ increase : num 12.2 12.2 12.2 12 13.1 ...  
## $ sex.ratio : num 947 947 947 955 950 947 947 958 958 955 ...  
## $ literacy : num 75 75 75 84.1 75 ...  
## $ density : num 1316 1316 1316 2445 1316 ...  
## $ total.income : num 4907966 5151095 5162574 6161808 6351215 ...  
## $ expenditure : num 4649638 4370939 2039140 4438315 5342270 ...  
## $ finalvalue : num 258328 780157 3123434 1723493 1008945 ...  
## $ result : chr "Profit" "Profit" "Profit" "Profit" ...  
## $ rate : num 6.8 6.8 6.8 6.8 6.8 6.8 6.8 6.8 6.8 6.8 ...  
## $ count.shg : num 760941 760941 760941 760941 760941 ...  
## $ saving.amount.lakh.rs: num 127348 127348 127348 127348 127348 ...  
## $ percapita16-17 : num 91114 91114 91114 91114 91114 ...  
## $ billion.percapita : num 1333 1333 1333 1333 1333 ...

#Original copy of final data  
original <- finaldata  
  
#Imputing missing data for branches not mentioned in previous list  
finaldata[sapply(finaldata, is.numeric)] <- lapply(finaldata[sapply(finaldata, is.numeric)], function(x) ifelse(is.na(x), median(x, na.rm = TRUE), x))  
finaldata$result <- as.factor(ifelse(finaldata$finalvalue > 0,"Profit","Loss"))  
finaldata <- finaldata%>% select(-result,result) #Moving Result column to the end !  
#write.csv(finaldata, file = "Finaldataforanalysis1.csv")   
  
  
#Correlation Plot to check multicollinearity  
finaldata <- finaldata  
corrplot::corrplot(cor(finaldata[,unlist(lapply(finaldata, is.numeric))]),method = "number",  
 diag = TRUE,title = "Corelation Plot",order = "AOE",tl.cex = 0.8,number.cex = 0.6)



#Building the models  
  
m1 = glm(result~ density+sex.ratio+literacy,family=binomial(link = "logit"), data = finaldata)  
m2 = glm(result~ density+sex.ratio+literacy+count.shg,family=binomial(link = "logit"), data = finaldata)  
  
m3 = glm(result~ density+sex.ratio+rate,family=binomial(link = "logit"), data = finaldata)  
  
summary(m3)

##   
## Call:  
## glm(formula = result ~ density + sex.ratio + rate, family = binomial(link = "logit"),   
## data = finaldata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.3683 -0.4605 0.2240 0.4076 1.9962   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.027e+02 4.280e+01 -2.399 0.0164 \*   
## density 7.925e-03 2.031e-03 3.902 9.52e-05 \*\*\*  
## sex.ratio 9.852e-02 4.190e-02 2.352 0.0187 \*   
## rate 2.081e-01 2.784e-01 0.747 0.4548   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 75.934 on 57 degrees of freedom  
## Residual deviance: 36.214 on 54 degrees of freedom  
## AIC: 44.214  
##   
## Number of Fisher Scoring iterations: 6

library(stargazer)

##   
## Please cite as:

## Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

stargazer(m1,m2,m3,type = "text")

##   
## ===================================================  
## Dependent variable:   
## ---------------------------------  
## result   
## (1) (2) (3)   
## ---------------------------------------------------  
## density 0.008\*\*\* 0.013\*\* 0.008\*\*\*   
## (0.002) (0.005) (0.002)   
##   
## sex.ratio 0.078\*\* 0.126\*\* 0.099\*\*   
## (0.033) (0.051) (0.042)   
##   
## literacy -0.006 0.072   
## (0.046) (0.105)   
##   
## count.shg -0.00001   
## (0.00001)   
##   
## rate 0.208   
## (0.278)   
##   
## Constant -80.619\*\*\* -133.487\*\*\* -102.699\*\*  
## (30.417) (50.856) (42.801)   
##   
## ---------------------------------------------------  
## Observations 58 58 58   
## Log Likelihood -18.384 -16.843 -18.107   
## Akaike Inf. Crit. 44.769 43.687 44.214   
## ===================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

#Creating Test Data Set  
  
test\_data <- subset(demographic, district %in% c("BHAGALPUR","SAHARSA","KHAGARIA","BADARPUR","KARIMGANJ","CACHAR")) #to be extracted from demographic file as the branch is not present in Finaldata  
test\_data <- test\_data %>%   
 left\_join(unemp, by ="state")%>%  
 left\_join(gsdp, by = "state")%>%  
 left\_join(mfigrowth, by="state")  
  
  
#m1 Predict  
  
m1.predictedvalues <- as.factor(if\_else(fitted.values(m1) >0.5,"Profit","Loss"))  
confusionMatrix(m1.predictedvalues,finaldata$result)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Loss Profit  
## Loss 19 7  
## Profit 2 30  
##   
## Accuracy : 0.8448   
## 95% CI : (0.7258, 0.9265)  
## No Information Rate : 0.6379   
## P-Value [Acc > NIR] : 0.0004424   
##   
## Kappa : 0.6805   
## Mcnemar's Test P-Value : 0.1824224   
##   
## Sensitivity : 0.9048   
## Specificity : 0.8108   
## Pos Pred Value : 0.7308   
## Neg Pred Value : 0.9375   
## Prevalence : 0.3621   
## Detection Rate : 0.3276   
## Detection Prevalence : 0.4483   
## Balanced Accuracy : 0.8578   
##   
## 'Positive' Class : Loss   
##

m1.predict <- predict(m1,test\_data,type = "response");m1.predict

## 1 2 3 4 5   
## 0.02712272 0.02840296 0.12903717 0.04962254 0.27246825

m1.predict <- ifelse(m1.predict >0.5,"Profit","Loss");m1.predict

## 1 2 3 4 5   
## "Loss" "Loss" "Loss" "Loss" "Loss"

#m2 Predict  
  
m2.predictedvalues <- as.factor(if\_else(fitted.values(m2) >0.5,"Profit","Loss"))  
confusionMatrix(m2.predictedvalues,finaldata$result)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Loss Profit  
## Loss 15 1  
## Profit 6 36  
##   
## Accuracy : 0.8793   
## 95% CI : (0.767, 0.9501)  
## No Information Rate : 0.6379   
## P-Value [Acc > NIR] : 3.506e-05   
##   
## Kappa : 0.7246   
## Mcnemar's Test P-Value : 0.1306   
##   
## Sensitivity : 0.7143   
## Specificity : 0.9730   
## Pos Pred Value : 0.9375   
## Neg Pred Value : 0.8571   
## Prevalence : 0.3621   
## Detection Rate : 0.2586   
## Detection Prevalence : 0.2759   
## Balanced Accuracy : 0.8436   
##   
## 'Positive' Class : Loss   
##

m2.predict <- predict(m2,test\_data,type = "response");m2.predict

## 1 2 3 4 5   
## 0.014001881 0.009260216 0.081254848 0.034000209 0.502052383

m2.predict <- ifelse(m2.predict >0.5,"Profit","Loss");m2.predict

## 1 2 3 4 5   
## "Loss" "Loss" "Loss" "Loss" "Profit"

#m3 Predict  
  
m3.predictedvalues <- as.factor(if\_else(fitted.values(m3) >0.5,"Profit","Loss"));m3.predictedvalues

## [1] Profit Profit Profit Profit Profit Profit Profit Profit Profit Profit  
## [11] Profit Profit Profit Profit Profit Profit Profit Profit Profit Profit  
## [21] Profit Profit Profit Profit Profit Profit Profit Profit Profit Profit  
## [31] Profit Profit Loss Profit Loss Profit Loss Profit Loss Profit  
## [41] Loss Loss Loss Loss Loss Loss Loss Loss Profit Loss   
## [51] Loss Loss Loss Loss Profit Loss Loss Profit  
## Levels: Loss Profit

finaldata$modelresult <- as.factor(if\_else(fitted.values(m3) >0.5,"Profit","Loss"))  
confusionMatrix(m3.predictedvalues,finaldata$result)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Loss Profit  
## Loss 16 3  
## Profit 5 34  
##   
## Accuracy : 0.8621   
## 95% CI : (0.7462, 0.9385)  
## No Information Rate : 0.6379   
## P-Value [Acc > NIR] : 0.0001331   
##   
## Kappa : 0.6951   
## Mcnemar's Test P-Value : 0.7236736   
##   
## Sensitivity : 0.7619   
## Specificity : 0.9189   
## Pos Pred Value : 0.8421   
## Neg Pred Value : 0.8718   
## Prevalence : 0.3621   
## Detection Rate : 0.2759   
## Detection Prevalence : 0.3276   
## Balanced Accuracy : 0.8404   
##   
## 'Positive' Class : Loss   
##

summary(finaldata$result)

## Loss Profit   
## 21 37

m3.predict <- predict(m3,test\_data,type = "response");m3.predict

## 1 2 3 4 5   
## 0.01066627 0.01195792 0.08284793 0.03125115 0.21478874

m3.predictvalues <- ifelse(m3.predict >0.5,"Profit","Loss");m3.predictvalues

## 1 2 3 4 5   
## "Loss" "Loss" "Loss" "Loss" "Loss"

test\_data$modelresult = as.factor(if\_else(m3.predict > 0.5,"Profit","Loss"));test\_data

## district population increase sex.ratio literacy density state rate  
## 1 BHAGALPUR 3037766 25.36 880 63.14 1182 Bihar 10.1  
## 2 KHAGARIA 1666886 30.19 886 57.92 1122 Bihar 10.1  
## 3 SAHARSA 1900661 26.02 906 53.20 1127 Bihar 10.1  
## 4 CACHAR 1736617 20.19 959 79.34 459 Assam 5.5  
## 5 KARIMGANJ 1228686 21.90 963 78.22 679 Assam 5.5  
## percapita16-17 billion.percapita count.shg saving.amount.lakh.rs  
## 1 37478 548 224469 29667  
## 2 37478 548 224469 29667  
## 3 37478 548 224469 29667  
## 4 75857 1109 292071 9943  
## 5 75857 1109 292071 9943  
## modelresult  
## 1 Loss  
## 2 Loss  
## 3 Loss  
## 4 Loss  
## 5 Loss

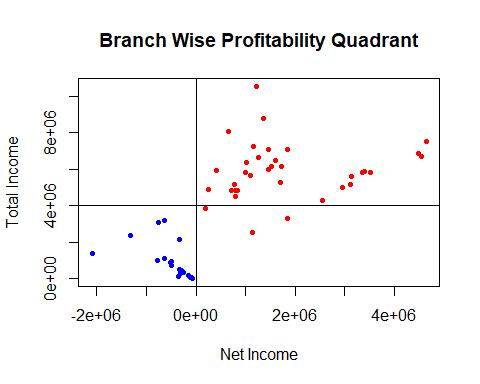
test\_data[,c(1,13)]

## district modelresult  
## 1 BHAGALPUR Loss  
## 2 KHAGARIA Loss  
## 3 SAHARSA Loss  
## 4 CACHAR Loss  
## 5 KARIMGANJ Loss

finaldata$truevalue <- 0  
finaldata$truevalue <- if\_else((finaldata$result == finaldata$modelresult),'True','False')  
head(finaldata[,c(1,17:19)],n = 10)

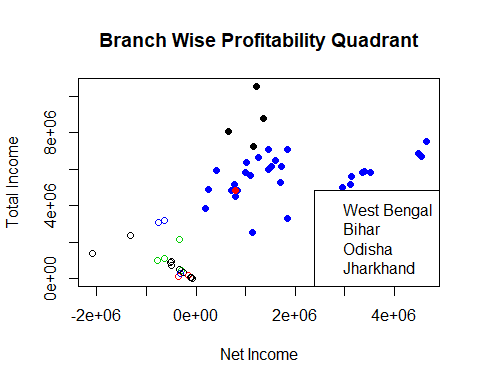
## branch result modelresult truevalue  
## 1 DUTTAPULIA Profit Profit True  
## 2 HANSKHALI Profit Profit True  
## 3 MATIKUMRA Profit Profit True  
## 4 LAUHATI Profit Profit True  
## 5 KARNAJORA Profit Profit True  
## 6 ARANGHATA Profit Profit True  
## 7 KRISHNAGANJ Profit Profit True  
## 8 SAGARPARA Profit Profit True  
## 9 SEKHPARA Profit Profit True  
## 10 SINDRANI Profit Profit True

#Result Plots  
  
# #Average Profit per State  
# profit.state = finaldata %>%  
# group\_by(state)%>%  
# summary(finalvalue)  
  
  
#aggregate(finaldata$finalvalue, list(finaldata$state), mean)  
plot(finaldata$finalvalue,finaldata$total.income,col = c("Blue", "red")[finaldata$result],pch = 20, main = "Branch Wise Profitability Quadrant", xlab = "Net Income", ylab = "Total Income")  
abline(a=0,b=4000000,h = 4000000)



#Main Graph with 2 variations for profit loss and State  
#plot(finaldata$finalvalue,finaldata$total.income,col = as.numeric(factor(finaldata$state)),pch = c(1,20)[finaldata$result], main = "Branch Wise Profitability Quadrant", xlab = "Net Income", ylab = "Total Income")  
  
plot(finaldata$finalvalue,finaldata$total.income,col = as.numeric(factor(finaldata$state)),pch = c(1,19)[finaldata$result], main = "Branch Wise Profitability Quadrant", xlab = "Net Income", ylab = "Total Income")  
legend(x="bottomright", legend=unique(finaldata$state), col=as.numeric(finaldata$state), pch=1)

## Warning in legend(x = "bottomright", legend = unique(finaldata$state), col  
## = as.numeric(finaldata$state), : NAs introduced by coercion



#Predicted Profitability Graph  
# plot(finaldata$finalvalue,finaldata$total.income,col = as.numeric(factor(finaldata$state)),pch = c(1,20)[finaldata$modelresult], main = "Branch Wise Predicted Profitability Quadrant", xlab = "Net Income", ylab = "Total Income")  
# legend(x="bottomright", legend=unique(finaldata$state), col=c('blue','black','green','red'), pch=19, cex = 0.8, bg = )  
#   
#   
# plot(finaldata$finalvalue,finaldata$total.income,col = as.numeric(factor(finaldata$state)),pch = 19, main = "Branch Wise Profitability Quadrant", xlab = "Net Income", ylab = "Total Income")  
# legend(x="bottomright", legend=unique(finaldata$state), col=c('blue','black','green','red'), pch=19, cex = 0.8, bg = "transparent")  
#   
# plot(finaldata$finalvalue,finaldata$total.income,col = as.numeric(factor(finaldata$state))[finaldata$state)[finaldata$result],pch = c(1,3,15,20)[finaldata$state], main = "Branch Wise Profitability Quadrant", xlab = "Net Income", ylab = "Total Income")  
#   
#   
# scatterplot(data = finaldata, finalvalue ~ total.income,pch = 20, main = "Branch Wise Profitability Quadrant", xlab = "Net Income", ylab = "Total Income")  
#   
#   
# #Computing State Wise Profit-Loss  
# aggprofit <- (finaldata %>% group\_by(result,state) %>% summarise(aggPL = mean(finalvalue)))  
# aggprofit1 <- (finaldata %>% group\_by(state) %>% summarise(aggPL = mean(finalvalue)))  
#   
# library(ggplot2)  
#   
#   
# #Average State wise Profitability  
# ggplot(aggprofit1,aes(factor(state),aggPL)) +  
# geom\_bar(stat = "identity", position = "dodge", fill = 'steelblue') +  
# scale\_fill\_brewer(palette = "Paired")+ ggtitle("State-wise Average Profitability") + ylab("Average Net Income (INR)") + xlab("States") +  
# geom\_text(aes(label = round(aggPL,digits = 0)),cex = 3.5, vjust = -0.2)+  
# theme\_minimal()  
#   
# #State Wise Average Profit and Loss  
# ggplot(aggprofit,aes(factor(state),aggPL, fill = result)) +  
# geom\_bar(stat = "identity", position = "dodge") +  
# scale\_fill\_brewer(palette = "Paired")+ ggtitle("State-wise Average Profit - Loss") + ylab("Average Net Income (INR)") + xlab("States") +  
# geom\_text(aes(label = round(aggPL,digits = 0)),cex = 3.5, vjust = -0.3, position = 'nudge')+  
# theme\_minimal()  
#   
#   
# install.packages("plotly")  
# library(plotly)  
#   
# Sys.setenv("plotly\_username"="rushikesh12")  
# Sys.setenv("plotly\_api\_key"="NlhrrEvudQc2O4W6sMyN")  
#   
# #original Plot  
# #plot\_ly(finaldata,x = ~total.income, y = ~expenditure , hovertext = ~branch, type = "scatter", mode = 'markers', size = ~finalvalue, color = ~result, colors = 'Paired', marker = list(opacity = 0.8, sizemode = 'area'))%>%  
# # layout(title = "Branch Profitability - Predicted", xaxis = list(showgrid = TRUE), yaxis = list(showgrid = TRUE), showlegend = TRUE)  
#   
# p1 <-plot\_ly(finaldata,x = ~total.income, y = ~expenditure , hoverinfo = 'text',  
# text = ~paste('</br>Branch: ',branch,'</br>NetProfit: ', finalvalue),  
# type = "scatter", mode = 'markers', size = ~finalvalue, color = ~result, colors = 'Paired',  
# marker = list(opacity = 0.8, sizemode = 'area'))%>%  
# layout(title = "Branch Profitability - Actual", xaxis = list(showgrid = TRUE, title = 'Total Income'),  
# yaxis = list(showgrid = TRUE, title = 'Expenditure'), showlegend = TRUE)  
#   
# api\_create(p1, filename = "Branch Profitability - Actual")  
#   
#   
# p2 <- plot\_ly(finaldata,x = ~total.income, y = ~expenditure , hoverinfo = 'text',  
# text = ~paste('</br>Branch: ',branch,'</br>NetProfit: ', finalvalue),  
# type = "scatter", mode = 'markers', size = ~finalvalue, color = ~modelresult, colors = 'Paired',  
# marker = list(opacity = 0.8, sizemode = 'area'))%>%  
# layout(title = "Branch Profitability - Predicted", xaxis = list(showgrid = TRUE, title = 'Total Income'),  
# yaxis = list(showgrid = TRUE, title = 'Expenditure'), showlegend = TRUE)  
#   
# api\_create(p2, filename = "Branch Profitability - Predicted")  
#   
#   
# p3 <- plot\_ly(finaldata,x = ~total.income, y = ~expenditure , hoverinfo = 'text',  
# text = ~paste('</br>Branch: ',branch,'</br>Net Profit: ',finalvalue),  
# type = "scatter", mode = 'markers', size = ~finalvalue, color = ~factor(truevalue), colors = 'RdYlGn',  
# marker = list(opacity = 0.8, sizemode = 'area'))%>%  
# layout(title = "Branch Profitability - Comparison", xaxis = list(showgrid = TRUE, title = 'Total Income'),  
# yaxis = list(showgrid = TRUE, title = 'Expenditure'), showlegend = TRUE)  
#   
# api\_create(p3, filename = "Branch Profitability - Comparison")  
#   
# #plot\_ly (x =c(1,2, 3 ),y =c(5,6, 7 ),type = 'scatter',mode = 'markers',size =c(1,5,10),  
# # marker = list(color =c('red','blue','green')))  
#