Data-Due-Diligence

Aamol Gote

6/6/2020

customers <- read\_csv("data/Customer\_Dataset\_File\_Original.csv")

sum(is.na(customers))

## [1] 124

sapply(customers,function(x)sum(is.na(x)))

## CustomerID Region TownSize   
## 0 0 0   
## Gender Age EducationYears   
## 33 0 0   
## JobCategory UnionMember EmploymentLength   
## 15 0 0   
## Retired HHIncome DebtToIncomeRatio   
## 0 0 0   
## CreditDebt OtherDebt LoanDefault   
## 0 0 0   
## MaritalStatus HouseholdSize NumberPets   
## 0 8 6   
## NumberCats NumberDogs NumberBirds   
## 7 8 34   
## HomeOwner CarsOwned CarOwnership   
## 13 0 0   
## CarBrand CarValue CommuteTime   
## 0 0 0   
## PoliticalPartyMem Votes CreditCard   
## 0 0 0   
## CardTenure CardItemsMonthly CardSpendMonth   
## 0 0 0   
## ActiveLifestyle PhoneCoTenure VoiceLastMonth   
## 0 0 0   
## VoiceOverTenure EquipmentRental EquipmentLastMonth   
## 0 0 0   
## EquipmentOverTenure CallingCard WirelessData   
## 0 0 0   
## DataLastMonth DataOverTenure Multiline   
## 0 0 0   
## VM Pager Internet   
## 0 0 0   
## CallerID CallWait CallForward   
## 0 0 0   
## ThreeWayCalling EBilling TVWatchingHours   
## 0 0 0   
## OwnsPC OwnsMobileDevice OwnsGameSystem   
## 0 0 0   
## OwnsFax NewsSubscriber   
## 0 0

# Feature engineering steps

# Data Imputation for customers.

#customers <- read\_csv("data/Customer\_Dataset\_File\_Original.csv")  
print("Number of rows with Gender = NA before imputation")

## [1] "Number of rows with Gender = NA before imputation"

sum(is.na(customers$Gender))

## [1] 33

numberOfFemales <- customers %>%   
 filter(Gender=="Female") %>%   
 nrow()  
print("Number of females before imputation")

## [1] "Number of females before imputation"

numberOfFemales

## [1] 2494

numberOfMales <- customers %>%   
 filter(Gender=="Male") %>%   
 nrow()  
count <- 0  
print("Number of males before imputation")

## [1] "Number of males before imputation"

numberOfMales

## [1] 2473

for(i in 1:nrow(customers)){  
 if(is.na(customers$Gender[i])){  
 #customers$Gender[i] <- "F"  
 count <- count + 1  
 if (count %% 2 == 0){  
 customers$Gender[i] <- "Female"  
 }  
 else{  
 customers$Gender[i] <- "Male"  
 }  
 }  
}  
  
print("Number of rows with Gender = NA after imputation")

## [1] "Number of rows with Gender = NA after imputation"

sum(is.na(customers$Gender))

## [1] 0

numberOfFemales <- customers %>%   
 filter(Gender=="Female") %>%   
 nrow()  
print("Number of females after imputation")

## [1] "Number of females after imputation"

numberOfFemales

## [1] 2510

numberOfMales <- customers %>%   
 filter(Gender=="Male") %>%   
 nrow()  
print("Number of males after imputation")

## [1] "Number of males after imputation"

numberOfMales

## [1] 2490

# Data Imputation for HouseholdSize.

#customers <- read\_csv("data/Customer\_Dataset\_File\_Original.csv")  
print("Number of rows with HouseholdSize = NA before imputation")

## [1] "Number of rows with HouseholdSize = NA before imputation"

sum(is.na(customers$HouseholdSize))

## [1] 8

summary(customers$HouseholdSize)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 1.000 1.000 2.000 2.202 3.000 9.000 8

customers$HouseholdSize[is.na(customers$HouseholdSize)] <- median(customers$HouseholdSize, na.rm = T)  
print("Check for missing values for HouseholdSize")

## [1] "Check for missing values for HouseholdSize"

sum(is.na(customers$HouseholdSize))

## [1] 0

# Data Imputation for HomeOwner.

#customers <- read\_csv("data/Customer\_Dataset\_File\_Original.csv")  
print("Number of rows with HomeOwner = NA before imputation")

## [1] "Number of rows with HomeOwner = NA before imputation"

sum(is.na(customers$HomeOwner))

## [1] 13

summary(customers$HomeOwner)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.0000 0.0000 1.0000 0.6296 1.0000 1.0000 13

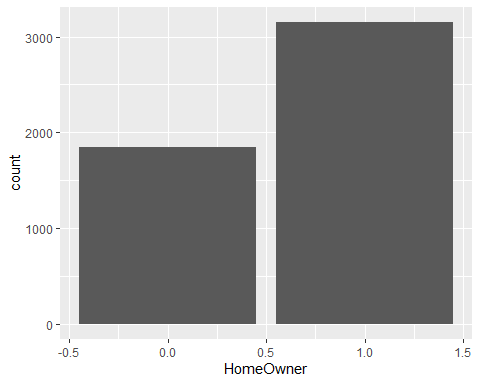
get\_mode <- function(x) {  
 unique\_x <- unique(x)  
 mode <- unique\_x[which.max(tabulate(match(x, unique\_x)))]  
 mode  
}  
mode\_value = get\_mode(customers$HomeOwner)  
customers$HomeOwner[is.na(customers$HomeOwner)] <- mode\_value  
print("Check for missing values for HomeOwner")

## [1] "Check for missing values for HomeOwner"

sum(is.na(customers$HomeOwner))

## [1] 0

ggplot(customers) +   
 geom\_bar(aes(x = HomeOwner))



# Data Imputation for JobCategory

#customers <- read\_csv("data/Customer\_Dataset\_File\_Original.csv")  
print("Number of rows with JobCategory = NA before imputation")

## [1] "Number of rows with JobCategory = NA before imputation"

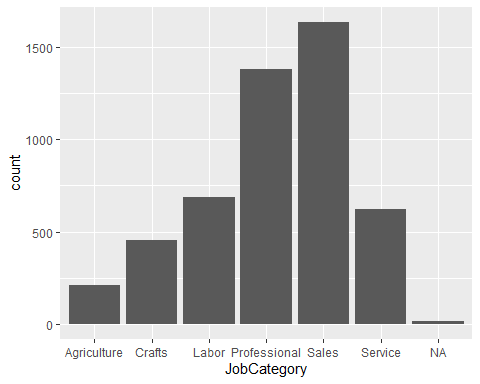
sum(is.na(customers$JobCategory))

## [1] 15

summary(customers$JobCategory)

## Length Class Mode   
## 5000 character character

ggplot(customers) +   
 geom\_bar(aes(x = JobCategory))



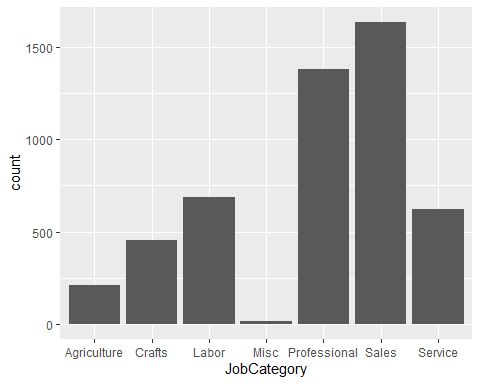
customers$JobCategory[is.na(customers$JobCategory)] <- "Misc"  
print("Check for missing values for HomeOwner")

## [1] "Check for missing values for HomeOwner"

sum(is.na(customers$JobCategory))

## [1] 0

ggplot(customers) +   
 geom\_bar(aes(x = JobCategory))



# Remove the dollar sign from household income

#customers <- read\_csv("data/Customer\_Dataset\_File\_Original.csv")  
customers <- customers %>%  
 replace(.=="#NULL!", NA) # replace with NA  
customers$HHIncome = as.numeric(gsub("\\,", "", gsub("\\$", "", gsub("\\-", "", customers$HHIncome))))   
customers$VoiceLastMonth = as.numeric(gsub("\\,", "", gsub("\\$", "", gsub("\\-", "", customers$VoiceLastMonth))))   
customers$VoiceOverTenure = as.numeric(gsub("\\,", "", gsub("\\$", "", gsub("\\-", "", customers$VoiceOverTenure))))   
customers$CardSpendMonth = as.numeric(gsub("\\,", "", gsub("\\$", "", gsub("\\-", "", customers$CardSpendMonth))))   
customers$EquipmentLastMonth = as.numeric(gsub("\\,", "", gsub("\\$", "", gsub("\\-", "", customers$EquipmentLastMonth))))   
customers$EquipmentOverTenure = as.numeric(gsub("\\,", "", gsub("\\$", "", gsub("\\-", "", customers$EquipmentOverTenure))))   
customers$DataLastMonth = as.numeric(gsub("\\,", "", gsub("\\$", "", gsub("\\-", "", customers$DataLastMonth))))   
customers$DataOverTenure = as.numeric(gsub("\\,", "", gsub("\\$", "", gsub("\\-", "", customers$DataOverTenure))))   
sapply(customers,function(x)sum(is.na(x)))

## CustomerID Region TownSize   
## 0 0 2   
## Gender Age EducationYears   
## 0 0 0   
## JobCategory UnionMember EmploymentLength   
## 0 0 0   
## Retired HHIncome DebtToIncomeRatio   
## 0 0 0   
## CreditDebt OtherDebt LoanDefault   
## 0 0 0   
## MaritalStatus HouseholdSize NumberPets   
## 0 0 6   
## NumberCats NumberDogs NumberBirds   
## 7 8 34   
## HomeOwner CarsOwned CarOwnership   
## 0 0 0   
## CarBrand CarValue CommuteTime   
## 0 0 2   
## PoliticalPartyMem Votes CreditCard   
## 0 0 0   
## CardTenure CardItemsMonthly CardSpendMonth   
## 0 0 7   
## ActiveLifestyle PhoneCoTenure VoiceLastMonth   
## 0 0 0   
## VoiceOverTenure EquipmentRental EquipmentLastMonth   
## 3 0 3296   
## EquipmentOverTenure CallingCard WirelessData   
## 3296 0 0   
## DataLastMonth DataOverTenure Multiline   
## 3656 3656 0   
## VM Pager Internet   
## 0 0 0   
## CallerID CallWait CallForward   
## 0 0 0   
## ThreeWayCalling EBilling TVWatchingHours   
## 0 0 0   
## OwnsPC OwnsMobileDevice OwnsGameSystem   
## 0 0 0   
## OwnsFax NewsSubscriber   
## 0 0

ncol(customers)

## [1] 59

# Adding of additional variables

customers$LastMonthTotalValue <- customers$VoiceLastMonth + customers$EquipmentLastMonth + customers$DataLastMonth  
customers$OverTenureTotalValue <- customers$VoiceOverTenure + customers$EquipmentOverTenure + customers$DataOverTenure  
customers$TotalDebt <- customers$CreditDebt + customers$OtherDebt

# Remove the dollar sign from CarValue

#customers <- read\_csv("data/Customer\_Dataset\_File\_Original.csv")  
customers$CarValue = gsub("\\ ", "", customers$CarValue)   
summary(customers$CarValue)

## Length Class Mode   
## 5000 character character

carValueLessThanZero <- customers %>%   
 filter(CarValue =="$(1,000.00)") %>%   
 nrow()  
  
customers$CarValue[customers$CarValue =="$(1,000.00)"] <- 0  
customers$CarValue = as.numeric(gsub("\\,", "", gsub("\\$", "", gsub("\\-", "", customers$CarValue))))   
customers$CarValue = as.numeric(customers$CarValue)  
  
carValueLessThanZero <- customers %>%   
 filter(is.na(customers$CarValue)) %>%  
 nrow()  
   
print("Number of records with CarValue = NA")

## [1] "Number of records with CarValue = NA"

carValueLessThanZero

## [1] 0

carOwnership <- customers %>%   
 filter(customers$CarOwnership == "-1") %>%  
 nrow()  
print("Number of records with CarOwnership == -1")

## [1] "Number of records with CarOwnership == -1"

carOwnership

## [1] 497

carBrand <- customers %>%   
 filter(customers$CarBrand == "-1") %>%  
 nrow()  
print("Number of records with CarBrand == -1")

## [1] "Number of records with CarBrand == -1"

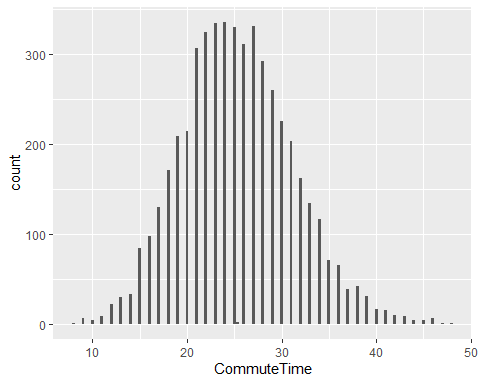
carBrand

## [1] 497

customers$CarOwnership[customers$CarOwnership =="-1"] <- NA  
customers$CarBrand[customers$CarBrand =="-1"] <- NA

# Impute Commute Time, there are 2 missing values

customers$CommuteTime = as.numeric(customers$CommuteTime)  
customers$CommuteTime[is.na(customers$CommuteTime)] <- mean(customers$CommuteTime, na.rm = T)  
ggplot(customers) +   
 geom\_bar(aes(x = CommuteTime))



summary(customers)

## CustomerID Region TownSize Gender   
## Length:5000 Min. :1.000 Length:5000 Length:5000   
## Class :character 1st Qu.:2.000 Class :character Class :character   
## Mode :character Median :3.000 Mode :character Mode :character   
## Mean :3.001   
## 3rd Qu.:4.000   
## Max. :5.000   
##   
## Age EducationYears JobCategory UnionMember   
## Min. :18.00 Min. : 6.00 Length:5000 Length:5000   
## 1st Qu.:31.00 1st Qu.:12.00 Class :character Class :character   
## Median :47.00 Median :14.00 Mode :character Mode :character   
## Mean :47.03 Mean :14.54   
## 3rd Qu.:62.00 3rd Qu.:17.00   
## Max. :79.00 Max. :23.00   
##   
## EmploymentLength Retired HHIncome DebtToIncomeRatio  
## Min. : 0.00 Length:5000 Min. : 9000 Min. : 0.000   
## 1st Qu.: 2.00 Class :character 1st Qu.: 24000 1st Qu.: 5.100   
## Median : 7.00 Mode :character Median : 38000 Median : 8.800   
## Mean : 9.73 Mean : 54760 Mean : 9.954   
## 3rd Qu.:15.00 3rd Qu.: 67000 3rd Qu.:13.600   
## Max. :52.00 Max. :1073000 Max. :43.100   
##   
## CreditDebt OtherDebt LoanDefault   
## Min. : 0.0000 Min. : 0.0000 Length:5000   
## 1st Qu.: 0.3855 1st Qu.: 0.9803 Class :character   
## Median : 0.9264 Median : 2.0985 Mode :character   
## Mean : 1.8573 Mean : 3.6545   
## 3rd Qu.: 2.0638 3rd Qu.: 4.3148   
## Max. :109.0726 Max. :141.4591   
##   
## MaritalStatus HouseholdSize NumberPets NumberCats   
## Length:5000 Min. :1.000 Min. : 0.000 Min. :0.0000   
## Class :character 1st Qu.:1.000 1st Qu.: 0.000 1st Qu.:0.0000   
## Mode :character Median :2.000 Median : 2.000 Median :0.0000   
## Mean :2.202 Mean : 3.067 Mean :0.5003   
## 3rd Qu.:3.000 3rd Qu.: 5.000 3rd Qu.:1.0000   
## Max. :9.000 Max. :21.000 Max. :6.0000   
## NA's :6 NA's :7   
## NumberDogs NumberBirds HomeOwner CarsOwned   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:1.000   
## Median :0.0000 Median :0.0000 Median :1.0000 Median :2.000   
## Mean :0.3928 Mean :0.1112 Mean :0.6306 Mean :2.131   
## 3rd Qu.:0.0000 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:3.000   
## Max. :7.0000 Max. :5.0000 Max. :1.0000 Max. :8.000   
## NA's :8 NA's :34   
## CarOwnership CarBrand CarValue CommuteTime   
## Length:5000 Length:5000 Min. : 0 Min. : 8.00   
## Class :character Class :character 1st Qu.: 9200 1st Qu.:21.00   
## Mode :character Mode :character Median :17000 Median :25.00   
## Mean :23332 Mean :25.35   
## 3rd Qu.:31100 3rd Qu.:29.00   
## Max. :99600 Max. :48.00   
##   
## PoliticalPartyMem Votes CreditCard CardTenure   
## Length:5000 Length:5000 Length:5000 Min. : 0.00   
## Class :character Class :character Class :character 1st Qu.: 6.00   
## Mode :character Mode :character Mode :character Median :14.00   
## Mean :16.66   
## 3rd Qu.:26.00   
## Max. :40.00   
##   
## CardItemsMonthly CardSpendMonth ActiveLifestyle PhoneCoTenure   
## Min. : 0.00 Min. : 69.7 Length:5000 Min. : 0.0   
## 1st Qu.: 8.00 1st Qu.: 1838.5 Class :character 1st Qu.:18.0   
## Median :10.00 Median : 2766.9 Mode :character Median :38.0   
## Mean :10.18 Mean : 3376.8 Mean :38.2   
## 3rd Qu.:12.00 3rd Qu.: 4187.5 3rd Qu.:59.0   
## Max. :23.00 Max. :39264.1 Max. :72.0   
## NA's :7   
## VoiceLastMonth VoiceOverTenure EquipmentRental EquipmentLastMonth  
## Min. : 2.70 Min. : 0.9 Length:5000 Min. : 17.00   
## 1st Qu.: 17.10 1st Qu.: 104.6 Class :character 1st Qu.: 30.35   
## Median : 28.65 Median : 350.0 Mode :character Median : 36.55   
## Mean : 40.41 Mean : 708.9 Mean : 38.12   
## 3rd Qu.: 49.65 3rd Qu.: 913.9 3rd Qu.: 44.26   
## Max. :539.55 Max. :13046.5 Max. :106.30   
## NA's :3 NA's :3296   
## EquipmentOverTenure CallingCard WirelessData   
## Min. : 12.05 Length:5000 Length:5000   
## 1st Qu.: 478.94 Class :character Class :character   
## Median :1153.50 Mode :character Mode :character   
## Mean :1379.63   
## 3rd Qu.:2100.30   
## Max. :6525.30   
## NA's :3296   
## DataLastMonth DataOverTenure Multiline VM   
## Min. : 12.70 Min. : 12.7 Length:5000 Length:5000   
## 1st Qu.: 27.95 1st Qu.: 472.5 Class :character Class :character   
## Median : 36.52 Median : 1270.5 Mode :character Mode :character   
## Mean : 39.81 Mean : 1569.9   
## 3rd Qu.: 47.71 3rd Qu.: 2334.1   
## Max. :186.25 Max. :12858.6   
## NA's :3656 NA's :3656   
## Pager Internet CallerID   
## Length:5000 Length:5000 Length:5000   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## CallWait CallForward ThreeWayCalling   
## Length:5000 Length:5000 Length:5000   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## EBilling TVWatchingHours OwnsPC OwnsMobileDevice   
## Length:5000 Min. : 0.00 Length:5000 Length:5000   
## Class :character 1st Qu.:17.00 Class :character Class :character   
## Mode :character Median :20.00 Mode :character Mode :character   
## Mean :19.64   
## 3rd Qu.:23.00   
## Max. :36.00   
##   
## OwnsGameSystem OwnsFax NewsSubscriber   
## Length:5000 Length:5000 Length:5000   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## LastMonthTotalValue OverTenureTotalValue TotalDebt   
## Min. : 41.50 Min. : 39.2 Min. : 0.000   
## 1st Qu.: 89.15 1st Qu.: 1174.6 1st Qu.: 1.620   
## Median :111.25 Median : 2990.4 Median : 3.279   
## Mean :121.82 Mean : 3781.8 Mean : 5.512   
## 3rd Qu.:143.45 3rd Qu.: 5681.7 3rd Qu.: 6.456   
## Max. :362.35 Max. :21057.0 Max. :211.381   
## NA's :4111 NA's :4111