Telecomunication EDA

Nzambuli Daniel

2023-11-09

# Task

Perform Exploratory data analysis

1. data preparation
   * data with missing observations
   * explore
   * selecting the variables
2. assumptions for linear regression
   * normality
   * linearity
   * homoscedasticity
   * multi-colinearity
3. linear regression analysis - continuous
   * handling anomalies
4. GLMs
   * logostic regression - uses data that is binary, multinomial, ordinal
   * poisson regression - count data
   * negative binomial
   * quasi binomial
5. Non Parametric tests
   * sign rank test = one-sample t-test, one-sample z-test
   * wilcoxcon sign rank test = paired t-test
   * mann u test = independent t-test
   * Kruskal-wallis H-test = one way ANOVA
   * Pearson Correlation = spearman rank correlation
6. Handling correlated data/ multivariate data
7. Exploratory
   1. PCA, Factor Analysis
   2. Fixed effect and random effect models
   3. handling missing data
   4. assessing missingness types:
      * MAR
      * CMAR
      * MNAR
      * GEE regression
8. Non parametric

## 1. The Data Set

library(readxl)  
Telecommunication\_Data <- read\_excel("Telecommunication Data.xlsx")  
View(Telecommunication\_Data)

**See the data inline**

head(Telecommunication\_Data)

## # A tibble: 6 × 22  
## region tenure age marital address income employ retire gender reside  
## <chr> <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <chr> <chr> <dbl>  
## 1 Zone 2 13 44 Married 9 64 5 No Male 2  
## 2 <NA> 11 33 Married 7 136 5 No Male 6  
## 3 Zone 3 68 52 Married 24 NA 29 No Female 2  
## 4 Zone 2 33 33 Unmarried 12 33 0 No Female NA  
## 5 Zone 2 23 30 Married 9 30 2 No Male 4  
## 6 Zone 2 41 39 Unmarried 17 78 16 No Female 1  
## # ℹ 12 more variables: tollfree <chr>, tollten <dbl>, equipten <dbl>,  
## # cardten <dbl>, wireten <dbl>, loglong <dbl>, logtoll <dbl>, logequi <dbl>,  
## # logcard <dbl>, logwire <dbl>, custcat <chr>, churn <chr>

### Data Exploration

**Column names of the data**

colnames(Telecommunication\_Data)

## [1] "region" "tenure" "age" "marital" "address" "income"   
## [7] "employ" "retire" "gender" "reside" "tollfree" "tollten"   
## [13] "equipten" "cardten" "wireten" "loglong" "logtoll" "logequi"   
## [19] "logcard" "logwire" "custcat" "churn"

## 2 Data Preparation

**Finding the missing values in each column**

missing\_values = data.frame(columns = colnames(Telecommunication\_Data),  
 vals\_missing = colSums(is.na(Telecommunication\_Data)),  
 perc\_missing = colMeans(is.na(Telecommunication\_Data)) \* 100)  
  
head(missing\_values)

## columns vals\_missing perc\_missing  
## region region 7 0.7  
## tenure tenure 8 0.8  
## age age 5 0.5  
## marital marital 9 0.9  
## address address 8 0.8  
## income income 8 0.8

### Dealing with extreme missing values

# columns with more than 50% missing values need to be removed   
Telecommunication\_Data = Telecommunication\_Data[rowMeans(is.na(Telecommunication\_Data))< 0.5, colMeans(is.na(Telecommunication\_Data)) < 0.5]  
colnames(Telecommunication\_Data)

## [1] "region" "tenure" "age" "marital" "address" "income"   
## [7] "employ" "retire" "gender" "reside" "tollfree" "tollten"   
## [13] "equipten" "cardten" "wireten" "loglong" "logcard" "custcat"   
## [19] "churn"

### Split the data into Train and Test

set.seed(123)  
  
# split witha 70/30 ration  
ind = sample(1:nrow(Telecommunication\_Data), 0.7 \* nrow(Telecommunication\_Data))  
  
train = Telecommunication\_Data[ind,]  
test = Telecommunication\_Data[-ind,]  
  
nrow(train)

## [1] 700

nrow(test)

## [1] 300

## 3. Select The Variables

### Linear Regression

linear\_regress\_data = train

#### Na values in Str columns

library(dplyr, warn.conflicts = FALSE)  
  
char\_columns = colnames(linear\_regress\_data %>% select\_if(is.character))  
char\_columns

## [1] "region" "marital" "retire" "gender" "tollfree" "custcat" "churn"

#### Spread in the character columns

library(ggplot2)  
for(column in char\_columns){  
 uniq\_values = linear\_regress\_data[,column]  
 col\_vals = data.frame(table(uniq\_values))  
 col\_vals$percent = col\_vals$Freq / sum(col\_vals$Freq) \* 100  
 out1 = paste("sample statistics of the column", column)  
 out2 = paste(out1, "are: ")  
 ggplot(col\_vals,   
 aes(x = "", y = Freq, fill = uniq\_values))+  
 geom\_bar(stat = "identity", width = 1)+  
 coord\_polar("y")+  
 geom\_text(aes(label = paste0(round(percent), "%")), position = position\_stack(vjust = 0.5))+  
 ggtitle("Distribution of char\_column with Percentages")+  
 theme\_minimal()  
}