DATA 621 - HW4

Andrew Bowen, Glen Davis, Shoshana Farber, Joshua Forster, Charles Ugiagbe

2023-10-30

Homework 4 - Binary Logistic Regression & Multiple Linear Regression

Data Exploration:

We load an auto insurance company dataset containing 8,161 records. Each record represents a customer, and each record has two response variables: `TARGET_FLAG` and `TARGET_AMT`. Below is a short description of all the variables of interest in the data set, including these response variables:

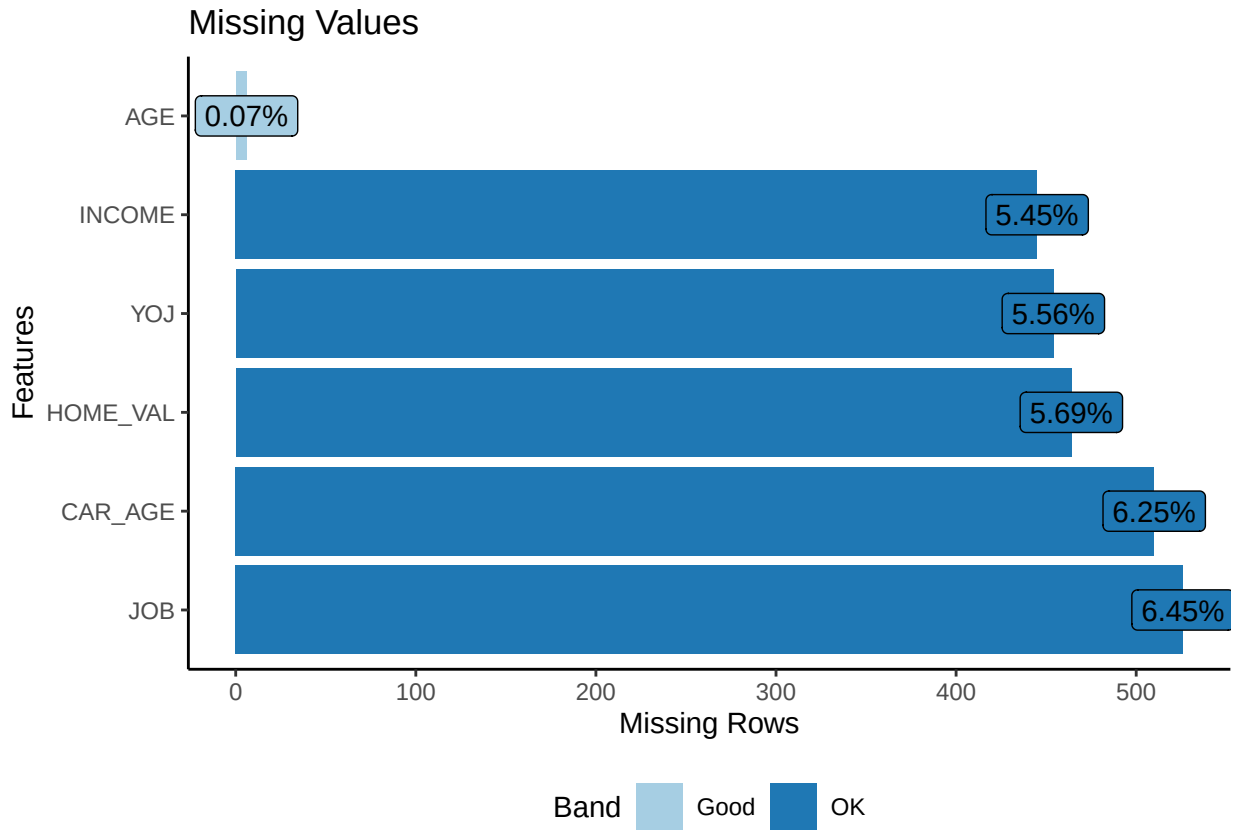
VARIABLE NAME	DEFINITION
`INDEX`	Identification Variable
`TARGET_FLAG`	Was Car in a crash? 1=YES 0=NO
`TARGET_AMT`	If car was in a crash, what was the cost
`AGE`	Age of Driver
`BLUEBOOK`	Value of Vehicle
`CAR_AGE`	Vehicle Age
`CAR_TYPE`	Type of Car
`CAR_USE`	Vehicle Use
`CLM_FREQ`	# Claims (Past 5 Years)
`EDUCATION`	Max Education Level
`HOMEKIDS`	# Children at Home
`HOME_VAL`	Home Value
`INCOME`	Income
`JOB`	Job Category
`KIDSDRIV`	# Driving Children
`MSTATUS`	Marital Status
`MVR_PTS`	Motor Vehicle Record Points
`OLDCLAIM`	Total Claims (Past 5 Years)
`PARENT1`	Single Parent
`RED_CAR`	A Red Car
`REVOKED`	License Revoked (Past 7 Years)
`SEX`	Gender
`TIF`	Time in Force
`TRAVTIME`	Distance to Work
`URBANICITY`	Home/Work Area
`YOJ`	Years on Job

Data Exploration

We remove the identification variable `INDEX` and take a look at a summary of the dataset's completeness.

rows	8161
columns	25
all_missing_columns	0
total_missing_values	2405
complete_rows	6045

None of our columns are completely devoid of data. There are 6,045 complete rows in the dataset, which is about 74% of our observations. There are 2,405 total missing values. We take a look at which variables contain these missing values and what the spread is.



A very small percentage of observations contain missing `AGE` values. The `INCOME`, `YOJ`, `HOME_VAL`, `CAR_AGE`, and `JOB` variables are each missing around 5.5 to 6.5 percent of values. There are no variables containing such extreme proportions of missing values that removal would be warranted on that basis alone.

Let's take a look at the summary statistics for each variable.

```
##   TARGET_FLAG       TARGET_AMT        KIDSDRIV          AGE
##  Min.   :0.0000   Min.   :     0   Min.   :0.0000   Min.   :16.00
##  1st Qu.:0.0000   1st Qu.:     0   1st Qu.:0.0000   1st Qu.:39.00
##  Median :0.0000   Median :     0   Median :0.0000   Median :45.00
##  Mean   :0.2638   Mean   :  1504   Mean   :0.1711   Mean   :44.79
##  3rd Qu.:1.0000   3rd Qu.:  1036   3rd Qu.:0.0000   3rd Qu.:51.00
##  Max.   :1.0000   Max.   :107586   Max.   :4.0000   Max.   :81.00
##                                                     NA's   :6
##     HOMEKIDS           YOJ          INCOME            PARENT1
```

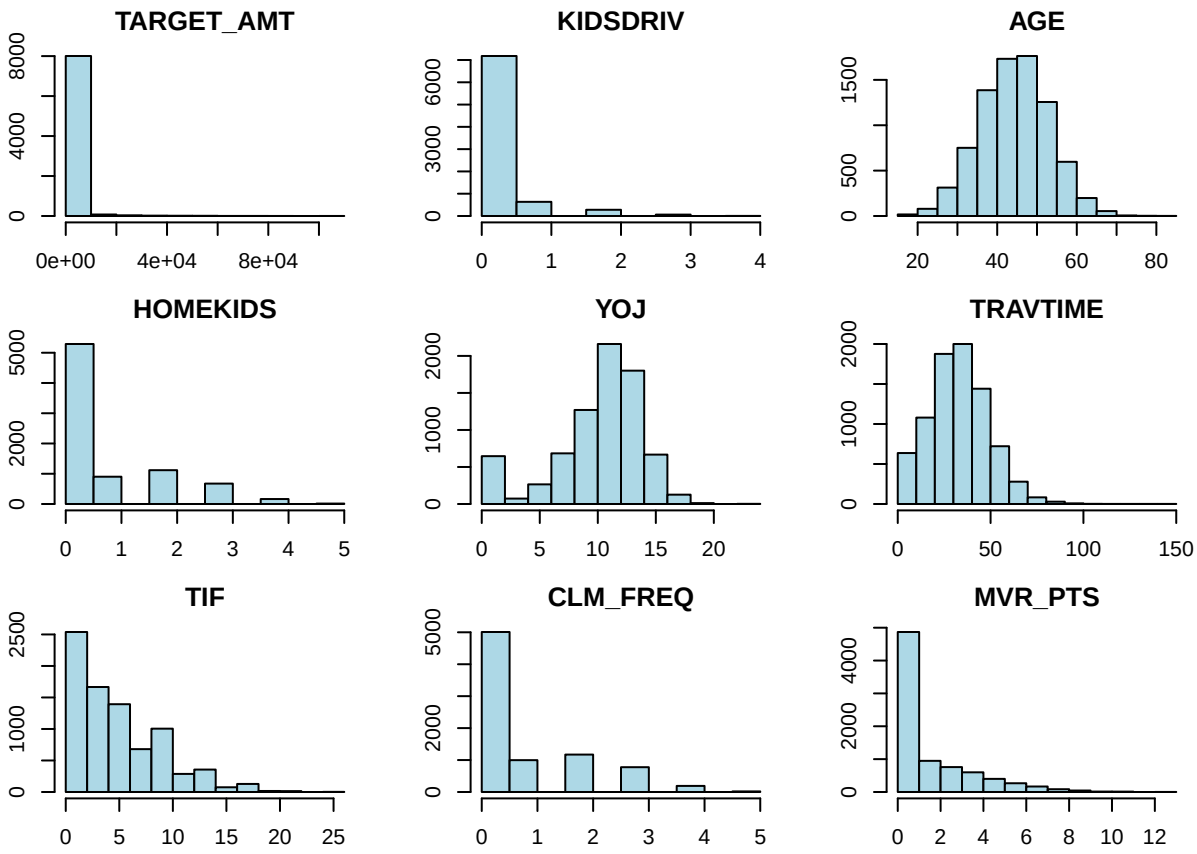
```
##  Min.   :0.0000   Min.   : 0.0   Length:8161        Length:8161
##  1st Qu.:0.0000   1st Qu.: 9.0   Class :character   Class :character
##  Median :0.0000   Median :11.0   Mode  :character   Mode  :character
##  Mean   :0.7212   Mean   :10.5
##  3rd Qu.:1.0000   3rd Qu.:13.0
##  Max.   :5.0000   Max.   :23.0
##                   NA's   :454
##    HOME_VAL           MSTATUS              SEX             EDUCATION
##  Length:8161        Length:8161        Length:8161        Length:8161
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##      JOB               TRAVTIME        CAR_USE            BLUEBOOK
##  Length:8161        Min.   :  5.00   Length:8161        Length:8161
##  Class :character   1st Qu.: 22.00   Class :character   Class :character
##  Mode  :character   Median : 33.00   Mode  :character   Mode  :character
##                     Mean   : 33.49
##                     3rd Qu.: 44.00
##                     Max.   :142.00
##
##       TIF            CAR_TYPE           RED_CAR            OLDCLAIM
##  Min.   : 1.000   Length:8161        Length:8161        Length:8161
##  1st Qu.: 1.000   Class :character   Class :character   Class :character
##  Median : 4.000   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 5.351
##  3rd Qu.: 7.000
##  Max.   :25.000
##
##     CLM_FREQ        REVOKED             MVR_PTS          CAR_AGE
##  Min.   :0.0000   Length:8161        Min.   : 0.000   Min.   :-3.000
##  1st Qu.:0.0000   Class :character   1st Qu.: 0.000   1st Qu.: 1.000
##  Median :0.0000   Mode  :character   Median : 1.000   Median : 8.000
##  Mean   :0.7986                      Mean   : 1.696   Mean   : 8.328
##  3rd Qu.:2.0000                      3rd Qu.: 3.000   3rd Qu.:12.000
##  Max.   :5.0000                      Max.   :13.000   Max.   :28.000
##                                                       NA's   :510
##   URBANICITY
##  Length:8161
##  Class :character
##  Mode  :character
##
##
##
##
```

There are 6 NAs in `AGE`, 454 in `YOJ`, and 510 in `CAR_AGE`.

The variables are a mixture of categorical and numerical.

`TARGET_FLAG` is a dummy variable for whether or not the car was involved in an accident. Let's make this variable a factor data type.

Let's take a look at the distributions of the variables.



Appendix: Report Code

Below is the code for this report to generate the models and charts above.

```
knitr::opts_chunk$set(echo = FALSE)
library(tidyverse)
library(DataExplorer)
library(knitr)

cur_theme <- theme_set(theme_classic())

my_url <- "https://raw.githubusercontent.com/andrewbowen19/businessAnalyticsDataMiningDATA621/main/data/
df <- read.csv(my_url, na.strings = "")

train_df <- df |>
    select(-INDEX)

remove <- c("discrete_columns", "continuous_columns", "total_observations",
            "memory_usage")

completeness <- introduce(train_df) |>
    select(-all_of(remove))
knitr::kable(t(completeness), format = "simple")

p1 <- plot_missing(train_df, missing_only = TRUE,
```

```
                    ggtheme = theme_classic(), title = "Missing Values")

p1 <- p1 +
    scale_fill_brewer(palette = "Paired")
p1

summary(train_df)

train_df$TARGET_FLAG <- as.factor(train_df$TARGET_FLAG)

# just numeric variables
numeric_train <- train_df[,sapply(train_df, is.numeric)]

par(mfrow=c(3,3))
par(mai=c(.3,.3,.3,.3))

variables <- names(numeric_train)

for (i in 1:(length(variables)-1)) {
  hist(numeric_train[[variables[i]]], main = variables[i], col = "lightblue")
}
```