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:

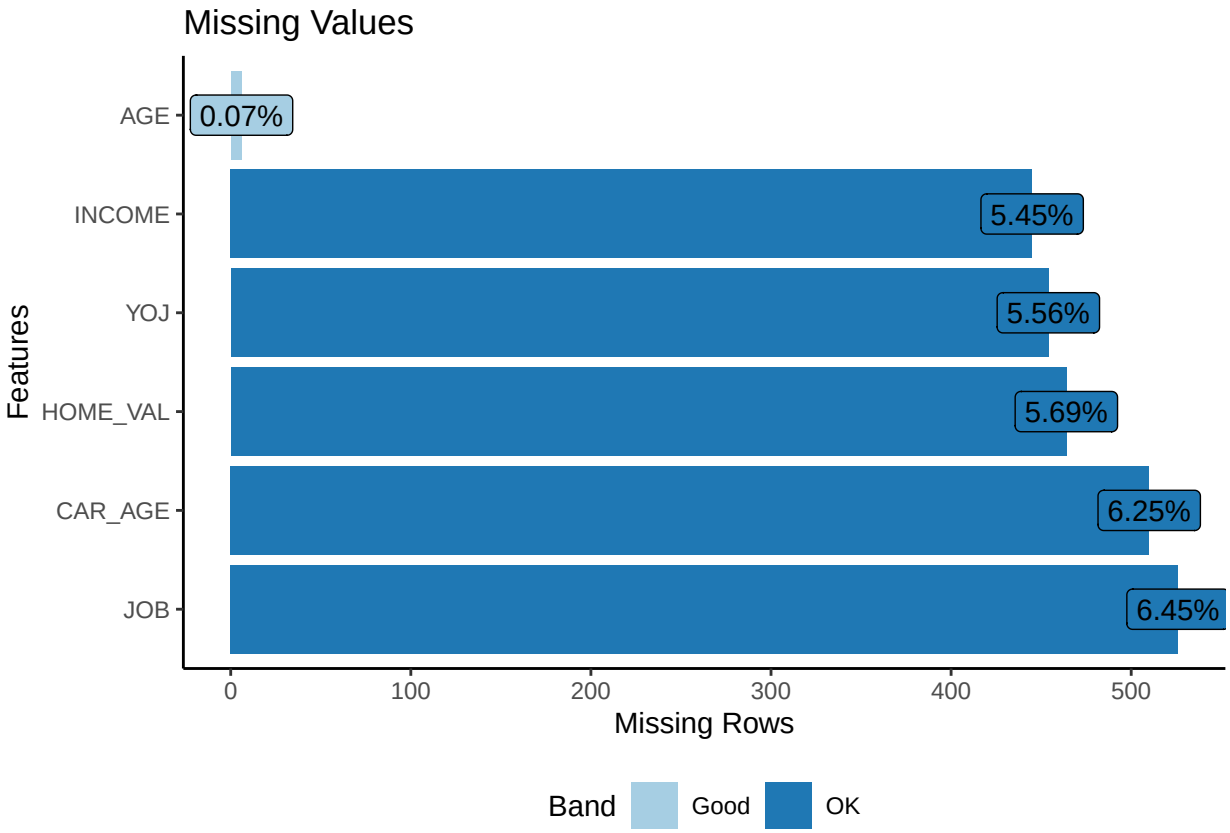
VARIABLE NAME	DEFINITION
`INDEX`	Identification Variable
`TARGET_FLAG`	Was Car in a crash? 1=YES 0=NO
`TARGET_AMT`	If car was in a crash, what was the cost
`AGE`	Age of Driver
`BLUEBOOK`	Value of Vehicle
`CAR_AGE`	Vehicle Age
`CAR_TYPE`	Type of Car
`CAR_USE`	Vehicle Use
`CLM_FREQ`	# Claims (Past 5 Years)
`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

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.

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/
```

```r
train_df <- read.csv(my_url, na.strings = "")

train_df <- train_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
```