Assignment1_Kwok

James Kwok

2024-01-08

```
# Set file path for the data
INFILE <- '/Users/jck/Documents/MSDS 422/Unit 1/Assignment 1/HMEQ_Loss.csv'

# Read in the data file
df <- read_csv(INFILE, show_col_types = FALSE)

# Define your target and predictor variables
TARGET_L <- "TARGET_LOSS_AMT"
cols_with_missing <- c('MORTDUE', 'VALUE', 'YOJ', 'DEROG', 'DELINQ', 'CLAGE', 'NINQ', 'C
LNO', 'DEBTINC')
categorical_cols_with_missing <- c('REASON','JOB')

# Store initial missing values count
initial_missing <- sapply(df[cols_with_missing], function(x) sum(is.na(x)))

# Fill in missing values for TARGET_LOSS_AMT
df[[TARGET_L]] <- ifelse(is.na(df[[TARGET_L]]), 0, df[[TARGET_L]])

# Handle numeric columns with missing values
for (col in cols_with_missing) {
  # Identify and remove outliers using IQR method
  Q1 <- quantile(df[[col]], 0.25, na.rm = TRUE)
  Q3 <- quantile(df[[col]], 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR
  temp_col <- df[[col]]
  temp_col[temp_col < lower_bound | temp_col > upper_bound] <- NA

  # Impute missing values with median (from non-outlier data)
  median_val <- median(temp_col, na.rm = TRUE)
  df[[paste0('IMP_', col)]] <- ifelse(is.na(df[[col]]), median_val, df[[col]])
}

# Handle categorical columns with missing values
for (col in categorical_cols_with_missing) {
  # Fill missing values with 'Unknown'
  df[[col]] <- ifelse(is.na(df[[col]]), 'Unknown', df[[col]])

  # One-hot encode using pivot_wider
  df <- df %>%
    mutate("{col}" := as.character(.[[col]])) %>%
    pivot_wider(
      names_from = col,
      values_from = col,
      values_fill = list(col = 0),
      names_prefix = paste0("OHE_", col, "_")
    )
}
```

```
## Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
## ℹ Please use `all_of()` or `any_of()` instead.
##   # Was:
##   data %>% select(col)
##
##   # Now:
##   data %>% select(all_of(col))
##
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
# Store final missing values count after imputation
final_missing_imp <- sapply(df[grep("^IMP_", names(df))], function(x) sum(is.na(x)))

# Display the initial and final missing values count
missing_values_summary <- data.frame(
  Initial = initial_missing,
  Final_IMP = final_missing_imp
)
print(missing_values_summary)
```

```
##         Initial Final_IMP
## MORTDUE     518         0
## VALUE       112         0
## YOJ         515         0
## DEROG       708         0
## DELINQ      580         0
## CLAGE       308         0
## NINQ        510         0
## CLNO        222         0
## DEBTINC    1267         0
```

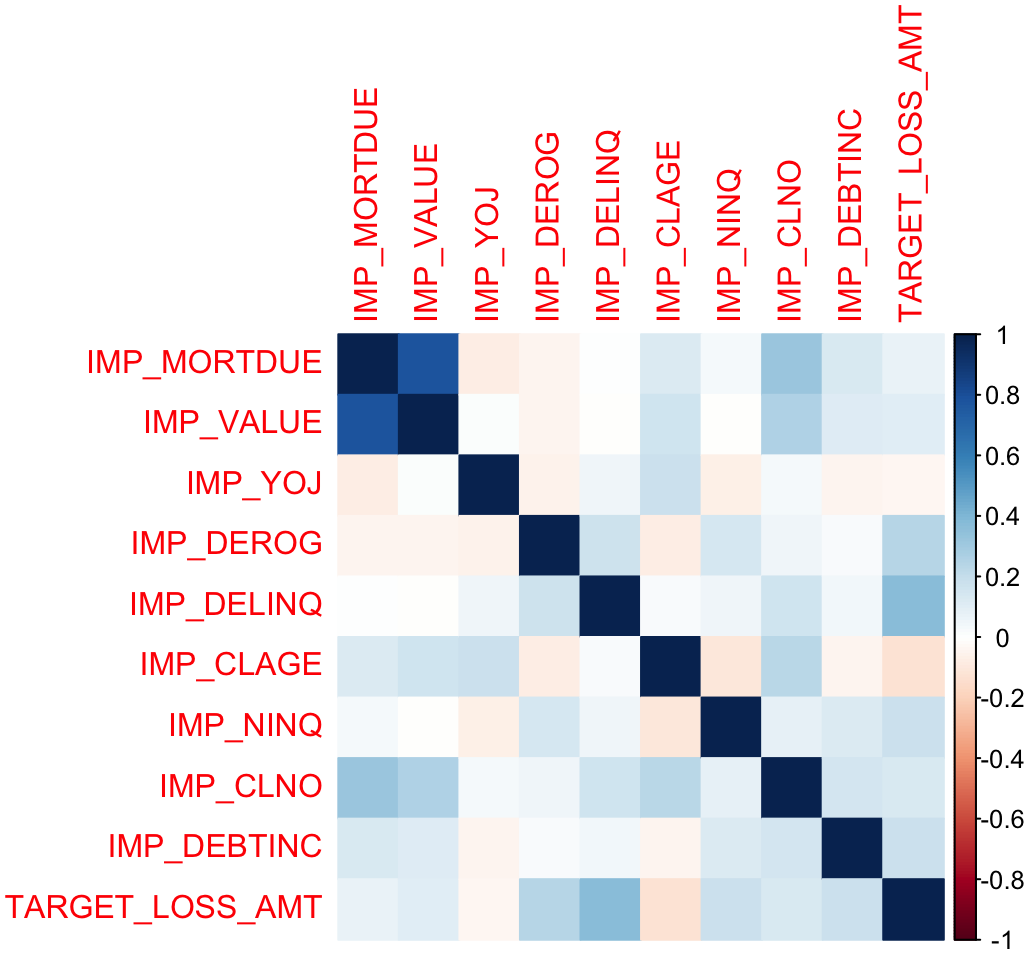
```
# Display the dataframe after imputation
print(head(df, 5))
```

```
## # A tibble: 5 × 31
##   TARGET_BAD_FLAG TARGET_LOSS_AMT  LOAN MORTDUE  VALUE   YOJ DEROG DELINQ CLAGE
##             <dbl>           <dbl> <dbl>   <dbl>  <dbl> <dbl> <dbl>  <dbl> <dbl>
## 1               1             641  1100   25860  39025  10.5     0      0  94.4
## 2               1            1109  1300   70053  68400   7       0      2 122.
## 3               1             767  1500   13500  16700   4       0      0 149.
## 4               1            1425  1500      NA     NA  NA      NA     NA  NA
## 5               0               0  1700   97800 112000   3       0      0  93.3
## # ℹ 22 more variables: NINQ <dbl>, CLNO <dbl>, DEBTINC <dbl>,
## #   IMP_MORTDUE <dbl>, IMP_VALUE <dbl>, IMP_YOJ <dbl>, IMP_DEROG <dbl>,
## #   IMP_DELINQ <dbl>, IMP_CLAGE <dbl>, IMP_NINQ <dbl>, IMP_CLNO <dbl>,
## #   IMP_DEBTINC <dbl>, OHE_REASON_HomeImp <chr>, OHE_REASON_Unknown <chr>,
## #   OHE_REASON_DebtCon <chr>, OHE_JOB_Other <chr>, OHE_JOB_Unknown <chr>,
## #   OHE_JOB_Office <chr>, OHE_JOB_Sales <chr>, OHE_JOB_Mgr <chr>,
## #   OHE_JOB_ProfExe <chr>, OHE_JOB_Self <chr>
```

```
# Heatmap for Correlation Matrix
correlation_matrix <- cor(df %>% select(starts_with("IMP_"), TARGET_L), use="complete.ob
s")
```

```
## Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
## ℹ Please use `all_of()` or `any_of()` instead.
##   # Was:
##   data %>% select(TARGET_L)
##
##   # Now:
##   data %>% select(all_of(TARGET_L))
##
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
corrplot(correlation_matrix, method = "color")
```


IMP_MORTDUE
IMP_VALUE
IMP_YOJ
IMP_DEROG
IMP_DELINQ
IMP_CLAGE
IMP_NINQ
IMP_CLNO
IMP_DEBTINC
TARGET_LOSS_AMT
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1