EDA & PreProcessing Credit Scoring

Carmine Santone

Variabili del dataset

ID: numero identificativo del cliente

CODE_GENDER: sesso del cliente

FLAGOWNCAR: indicatore del possesso di un'automobile

FLAGOWNREALTY: indicatore del possesso di una casa

CNT_CHILDREN: numero di figli

AMTINCOMETOTAL: reddito annuale

NAMEINCOMETYPE: tipo di reddito

NAMEEDUCATIONTYPE: livello di educazione

NAMEFAMILYSTATUS: stato civile

NAMEHOUSINGTYPE: tipo di abitazione

DAYS_BIRTH: numero di giorni trascorsi dalla nascita

DAYS_EMPLOYED: numero di giorni trascorsi dalla data di assunzione (se positivo, indica il numero di giorni da quando è disoccupato)

FLAG_MOBIL: indicatore della presenza di un numero di cellulare

FLAGWORKPHONE: indicatore della presenza di un numero di telefono di lavoro

FLAG_PHONE: indicatore della presenza di un numero di telefono

FLAG_EMAIL: indicatore della presenza di un indirizzo email

OCCUPATION_TYPE: tipo di occupazione

CNTFAMMEMBERS: numero di familiari

TARGET: variabile che vale 1 se il cliente ha una elevata affidabilità creditizia (pagamento costante delle rate), 0 altrimenti

SUMMARY

```
df <- df %>%
  mutate_if(is.character, factor)
summary(df)
```

```
##        ID          CODE_GENDER FLAG_OWN_CAR FLAG_OWN_REALTY  CNT_CHILDREN
##  Min.   :5008804   F:227916    N:213196     N:107120        Min.   : 0.0000
##  1st Qu.:5439602   M:110511    Y:125231     Y:231307        1st Qu.: 0.0000
##  Median :5878907                                          Median : 0.0000
##  Mean   :5821200                                          Mean   : 0.4289
##  3rd Qu.:6140206                                          3rd Qu.: 1.0000
##  Max.   :6841875                                          Max.   :19.0000
##
##  AMT_INCOME_TOTAL              NAME_INCOME_TYPE
##  Min.   :  26100   Commercial associate: 78090
##  1st Qu.: 121500   Pensioner           : 57841
##  Median : 162000   State servant       : 28113
##  Mean   : 187654   Student             :    17
##  3rd Qu.: 225000   Working             :174366
##  Max.   :6750000
##
##                       NAME_EDUCATION_TYPE            NAME_FAMILY_STATUS
##  Academic degree              :   232                 :     1
##  Higher education             : 91062   Civil marriage      : 28516
##  Incomplete higher            : 11387   Married             :231494
##  Lower secondary              :  3177   Separated           : 20809
##  Secondary / se               :     1   Single / not married: 42509
##  Secondary / secondary special:232568   Widow               : 15098
##
##            NAME_HOUSING_TYPE    DAYS_BIRTH     DAYS_EMPLOYED      FLAG_MOBIL
##                     :     1   Min.   :-25201   Min.   :-17531   Min.   :1
##  Co-op apartment    :  1162   1st Qu.:-19482   1st Qu.: -3116   1st Qu.:1
##  House / apartment  :304410   Median :-15622   Median : -1485   Median :1
##  Municipal apartment: 10819   Mean   :-15998   Mean   : 60239   Mean   :1
##  Office apartment   :  2968   3rd Qu.:-12524   3rd Qu.:  -380   3rd Qu.:1
##  Rented apartment   :  4442   Max.   : -7489   Max.   :365243   Max.   :1
##  With parents       : 14625   NA's   :1        NA's   :1        NA's   :1
##  FLAG_WORK_PHONE    FLAG_PHONE       FLAG_EMAIL        OCCUPATION_TYPE
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000              :103342
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   Laborers   : 60146
##  Median :0.0000   Median :0.0000   Median :0.0000   Core staff : 33527
##  Mean   :0.2114   Mean   :0.2933   Mean   :0.1052   Sales staff: 31652
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   Managers   : 27384
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Drivers    : 20020
##  NA's   :1        NA's   :1        NA's   :1        (Other)    : 62356
##  CNT_FAM_MEMBERS      TARGET
##  Min.   : 1.000   Min.   :0.00000
##  1st Qu.: 2.000   1st Qu.:0.00000
##  Median : 2.000   Median :0.00000
##  Mean   : 2.197   Mean   :0.08782
##  3rd Qu.: 3.000   3rd Qu.:0.00000
##  Max.   :20.000   Max.   :1.00000
##  NA's   :1
```

Si procede all'analisi delle singole variabili che presentano valori anomali

DAYS_EMPOLYED

Il valore massimo della variabile è:

```
max_days_emp <- max(df$DAYS_EMPLOYED, na.rm = T)
print(max_days_emp)

## [1] 365243
```

che equivale a 1.000 anni di lavoro, impossibile.

```
table(df$DAYS_EMPLOYED == max_days_emp, df$NAME_INCOME_TYPE)

##
##         Commercial associate Pensioner State servant Student Working
##   FALSE                78090         0         28113      17  174365
##   TRUE                     0     57841             0       0       0
```

Quindi, per i pensionati, la variabile **DAYS_EMPLOYED** assume valore **365243**. C'è bisogno di un encoding della variabile, si propone la creazione di una variabile che la sostituisca. La variabile **DAYS_EMPLOYED**, codificata in questo modo, non ha senso perchè un manager appena assunto avrebbe un valore minore rispetto ad un lavoratore alla prima esperienza. Si propone la sostituizione con una variabile **STATUS_EMPLOYED** categoriale, con 3 modalità, per occupati, disoccupati e pensionati. Innanzitutto controllo se ci sono disoccupati:

```
sum(df$DAYS_EMPLOYED >= 0 & df$DAYS_EMPLOYED != max_days_emp, na.rm = T)

## [1] 0
```

Non essendoci disoccupati, **STATUS_EMPLOYED** sarà una dummy, con modalità **Employed** e **Pensioner**.

```
df$STATUS_EMPLOYED <- factor(ifelse(df$DAYS_EMPLOYED < 1, "Employed",
"Pensioner"))
df$DAYS_EMPLOYED <- NULL
```

OCCUPATION_TYPE

Stringa vuota come modalità:

```
table(df$OCCUPATION_TYPE)

##
##                              Accountants        Cleaning staff
##                103342               12281                 4594
##         Cooking staff            Core staff              Drivers
##                  6248               33527                20020
## High skill tech staff              HR staff           IT staff
##                 13399                 567                 436
##              Laborers    Low-skill Laborers             Managers
##                 60146                1714                27384
##        Medicine staff Private service staff        Realty agents
##                 10438                2787                 852
```

```
##           Sales staff           Secretaries        Security staff
##                 31652                  1577                  6218
##  Waiters/barmen staff
##                  1245
```

Controllo se il valore mancante sia dovuto al fatto di essere pensionato, in quanto non è presente come modalità in **OCCUPATION_TYPE**:

```
table(df$STATUS_EMPLOYED == "Pensioner", df$OCCUPATION_TYPE)

##
##               Accountants Cleaning staff Cooking staff Core staff Drivers
##   FALSE 45500       12281           4594          6248      33527   20020
##   TRUE  57841           0              0             0          0       0
##
##         High skill tech staff HR staff IT staff Laborers Low-skill Laborers
##   FALSE                 13399      567      436    60146                1714
##   TRUE                      0        0        0        0                   0
##
##         Managers Medicine staff Private service staff Realty agents Sales staff
##   FALSE    27384          10438                  2787           852       31652
##   TRUE         0              0                     0             0           0
##
##         Secretaries Security staff Waiters/barmen staff
##   FALSE        1577           6218                 1245
##   TRUE            0              0                    0
```

Su **103342** valori mancanti, **57841** sono da imputare come **Pensioner** e imputo gli altri valori mancanti con una nuova modalità, **Other**:

```
levels(df$OCCUPATION_TYPE) <- c(levels(df$OCCUPATION_TYPE), "Pensioner", "Other")
df$OCCUPATION_TYPE[df$OCCUPATION_TYPE == "" & df$STATUS_EMPLOYED == "Pensioner"]
<- "Pensioner"
df$OCCUPATION_TYPE[df$OCCUPATION_TYPE == ""] <- "Other"
df$OCCUPATION_TYPE <- droplevels(df$OCCUPATION_TYPE)
table(df$OCCUPATION_TYPE)

##
##           Accountants        Cleaning staff         Cooking staff
##                 12281                  4594                  6248
##            Core staff               Drivers High skill tech staff
##                 33527                 20020                 13399
##              HR staff            IT staff              Laborers
##                   567                   436                 60146
##    Low-skill Laborers              Managers        Medicine staff
##                  1714                 27384                 10438
## Private service staff         Realty agents           Sales staff
##                  2787                   852                 31652
##           Secretaries        Security staff  Waiters/barmen staff
##                 1577                  6218                  1245
##             Pensioner                 Other
##                 57841                 45501
```

DAYS_BIRTH

Trasformo la variabile **DAYS_BIRTH** in **AGE**

```
df$AGE <- -df$DAYS_BIRTH
df$AGE <- floor(df$AGE/365)

df$DAYS_BIRTH <- NULL
```

ALTRI VALORI MANCANTI

In totale ci sono **7** valori mancanti…

```
sum(is.na(df))

## [1] 7
```

… che appartengono all'osservazione **338427**

```
na_sum <- rowSums(is.na(df))
na_obs <- which(na_sum != 0)
print(na_obs)

## [1] 338427
```

Trattandosi di una sola osservazione, si procede all'eliminazione dal dataset

```
df <- df[-na_obs,]
```

DUPLICATI

La matrice dei dati è composta da **268652** osservazioni duplicate

```
sum(duplicated(df[,-1]))

## [1] 268652
```

Prendo in considerazione solo le righe uniche

```
df <- unique(df[,-1])
```

ESPORTAZIONE DEL DATASET PULITO

Esporto il dataset pulito per utilizzarlo in ambiente Python per la parte di applicazione di modelli di Machine Learning

```
write.csv(df, "credit_scoring_preprocessed.csv", row.names = FALSE)
```