D1.Definition and projects assignment.

Alícia Chimeno Sarabia

2023-09-18

1. Name of all group components by alphabetical order (sort by Family name)

Adrià Casanova Lloveras,Alicia Chimeno Sarabia, Victor-George Giurcoiu, Zhengoyng Ji, Victor Garcia Pizarro.

Data Source

2. Data source including the url or urls involved

https://www.kaggle.com/datasets/mishra5001/credit-card

```r
data <- read.csv("application_data.csv", sep = ",")
data
```

3. One paragraph explaining the process to get your data (basic download, more sophisticated processes when used). It is possible to enlarge your database with additional variables coming from other sources if you like, but do not invest too much time on that. Deliver dataset on time. If it is the case, provide all urls involved in your dataset.

Basic download, load the .csv.

Data Context

4. One paragraph explaining what data are about

Loan Application Data. This dataset contains social-economic information about clients who applied for a loan. The target variable is set to 1 if the client had a late payment and 0 if they did not. The data originates from a study conducted by IIIT Bangalore, the International Institute of Information Technology Bangalore, as mentioned by the author who updated the database on Kaggle. The columns_description.csv is a .csv file describing each variable (already done but we want to do owr own).

Data Description

5. Basic structure of data matrix: One paragraph with: a. nr of records (better if it is bigger than 2000, if you are working with countries in the world or other situations, this might be reconsidered) b. nr of variables c. nr of numerical variables (minimum of 7 numerical variables) d. nr of binary variables (minimum of 2 binary variables) e. nr of qualitative variables (minimum of 5 categorical variables) f. number and % of missing data per each variable g. % of missing data in the whole data matrix.

```r
glimpse(data)
summary(data)
names(data)
```

Number of records:

```
nrow(data)
```

```
## [1] 307511
```

Number of variables:

```
ncol(data)
```

```
## [1] 122
```

First, we convert the binary variables that are considered numeric into binary:

```
# Function to convert numeric columns with 2 distinct values to binary
convert_to_binary <- function(dataframe) {
  for (col in names(dataframe)) {
    if (is.numeric(dataframe[[col]]) && length(unique(dataframe[[col]])) == 2) {
      dataframe[[col]] <- as.factor(dataframe[[col]])
    }
  }
  return(dataframe)
}

# Apply the function to your data frame
data <- convert_to_binary(data)
```

We have the following types of variables: 16 qualitative variables, 33 binary variables, and 73 numerical variables.

number and % of missing data per each variable:

```
missing.values.df <- as.data.frame(skimr::skim(data))
missing.values.df <- missing.values.df[,2:3]
missing.values.df$percentage_missing <- missing.values.df$n_missing / nrow(data) * 100
missing.values.df
```

```
##                    skim_variable n_missing percentage_missing
## 1             NAME_CONTRACT_TYPE         0       0.000000e+00
## 2                    CODE_GENDER         0       0.000000e+00
## 3                   FLAG_OWN_CAR         0       0.000000e+00
## 4                FLAG_OWN_REALTY         0       0.000000e+00
## 5                NAME_TYPE_SUITE         0       0.000000e+00
## 6               NAME_INCOME_TYPE         0       0.000000e+00
## 7            NAME_EDUCATION_TYPE         0       0.000000e+00
## 8             NAME_FAMILY_STATUS         0       0.000000e+00
## 9              NAME_HOUSING_TYPE         0       0.000000e+00
## 10               OCCUPATION_TYPE         0       0.000000e+00
## 11    WEEKDAY_APPR_PROCESS_START         0       0.000000e+00
## 12             ORGANIZATION_TYPE         0       0.000000e+00
## 13            FONDKAPREMONT_MODE         0       0.000000e+00
## 14                HOUSETYPE_MODE         0       0.000000e+00
## 15            WALLSMATERIAL_MODE         0       0.000000e+00
## 16           EMERGENCYSTATE_MODE         0       0.000000e+00
## 17                        TARGET         0       0.000000e+00
## 18                    FLAG_MOBIL         0       0.000000e+00
## 19                FLAG_EMP_PHONE         0       0.000000e+00
## 20               FLAG_WORK_PHONE         0       0.000000e+00
## 21              FLAG_CONT_MOBILE         0       0.000000e+00
## 22                    FLAG_PHONE         0       0.000000e+00
## 23                    FLAG_EMAIL         0       0.000000e+00
## 24    REG_REGION_NOT_LIVE_REGION         0       0.000000e+00
## 25    REG_REGION_NOT_WORK_REGION         0       0.000000e+00
## 26   LIVE_REGION_NOT_WORK_REGION         0       0.000000e+00
## 27        REG_CITY_NOT_LIVE_CITY         0       0.000000e+00
## 28        REG_CITY_NOT_WORK_CITY         0       0.000000e+00
## 29       LIVE_CITY_NOT_WORK_CITY         0       0.000000e+00
## 30               FLAG_DOCUMENT_2         0       0.000000e+00
## 31               FLAG_DOCUMENT_3         0       0.000000e+00
## 32               FLAG_DOCUMENT_4         0       0.000000e+00
## 33               FLAG_DOCUMENT_5         0       0.000000e+00
## 34               FLAG_DOCUMENT_6         0       0.000000e+00
## 35               FLAG_DOCUMENT_7         0       0.000000e+00
## 36               FLAG_DOCUMENT_8         0       0.000000e+00
## 37               FLAG_DOCUMENT_9         0       0.000000e+00
## 38              FLAG_DOCUMENT_10         0       0.000000e+00
## 39              FLAG_DOCUMENT_11         0       0.000000e+00
## 40              FLAG_DOCUMENT_12         0       0.000000e+00
## 41              FLAG_DOCUMENT_13         0       0.000000e+00
## 42              FLAG_DOCUMENT_14         0       0.000000e+00
## 43              FLAG_DOCUMENT_15         0       0.000000e+00
## 44              FLAG_DOCUMENT_16         0       0.000000e+00
## 45              FLAG_DOCUMENT_17         0       0.000000e+00
## 46              FLAG_DOCUMENT_18         0       0.000000e+00
## 47              FLAG_DOCUMENT_19         0       0.000000e+00
## 48              FLAG_DOCUMENT_20         0       0.000000e+00
## 49              FLAG_DOCUMENT_21         0       0.000000e+00
## 50                    SK_ID_CURR         0       0.000000e+00
## 51                  CNT_CHILDREN         0       0.000000e+00
## 52              AMT_INCOME_TOTAL         0       0.000000e+00
## 53                    AMT_CREDIT         0       0.000000e+00
## 54                   AMT_ANNUITY        12       3.902299e-03
## 55               AMT_GOODS_PRICE       278       9.040327e-02
## 56    REGION_POPULATION_RELATIVE         0       0.000000e+00
## 57                    DAYS_BIRTH         0       0.000000e+00
## 58                 DAYS_EMPLOYED         0       0.000000e+00
## 59             DAYS_REGISTRATION         0       0.000000e+00
## 60               DAYS_ID_PUBLISH         0       0.000000e+00
## 61                   OWN_CAR_AGE    202929       6.599081e+01
## 62               CNT_FAM_MEMBERS         2       6.503832e-04
## 63          REGION_RATING_CLIENT         0       0.000000e+00
## 64   REGION_RATING_CLIENT_W_CITY         0       0.000000e+00
## 65       HOUR_APPR_PROCESS_START         0       0.000000e+00
## 66                  EXT_SOURCE_1    173378       5.638107e+01
## 67                  EXT_SOURCE_2       660       2.146265e-01
## 68                  EXT_SOURCE_3     60965       1.982531e+01
## 69                APARTMENTS_AVG    156061       5.074973e+01
## 70              BASEMENTAREA_AVG    179943       5.851596e+01
## 71   YEARS_BEGINEXPLUATATION_AVG    150007       4.878102e+01
## 72               YEARS_BUILD_AVG    204488       6.649778e+01
## 73                COMMONAREA_AVG    214865       6.987230e+01
## 74                 ELEVATORS_AVG    163891       5.329598e+01
## 75                 ENTRANCES_AVG    154828       5.034877e+01
## 76                 FLOORSMAX_AVG    153020       4.976082e+01
## 77                 FLOORSMIN_AVG    208642       6.784863e+01
## 78                  LANDAREA_AVG    182590       5.937674e+01
## 79          LIVINGAPARTMENTS_AVG    210199       6.835495e+01
## 80                LIVINGAREA_AVG    154350       5.019333e+01
## 81       NONLIVINGAPARTMENTS_AVG    213514       6.943296e+01
## 82             NONLIVINGAREA_AVG    169682       5.517916e+01
## 83               APARTMENTS_MODE    156061       5.074973e+01
## 84             BASEMENTAREA_MODE    179943       5.851596e+01
## 85  YEARS_BEGINEXPLUATATION_MODE    150007       4.878102e+01
## 86              YEARS_BUILD_MODE    204488       6.649778e+01
## 87               COMMONAREA_MODE    214865       6.987230e+01
## 88                ELEVATORS_MODE    163891       5.329598e+01
## 89                ENTRANCES_MODE    154828       5.034877e+01
## 90                FLOORSMAX_MODE    153020       4.976082e+01
## 91                FLOORSMIN_MODE    208642       6.784863e+01
## 92                 LANDAREA_MODE    182590       5.937674e+01
## 93         LIVINGAPARTMENTS_MODE    210199       6.835495e+01
## 94               LIVINGAREA_MODE    154350       5.019333e+01
## 95      NONLIVINGAPARTMENTS_MODE    213514       6.943296e+01
## 96            NONLIVINGAREA_MODE    169682       5.517916e+01
## 97               APARTMENTS_MEDI    156061       5.074973e+01
## 98             BASEMENTAREA_MEDI    179943       5.851596e+01
## 99  YEARS_BEGINEXPLUATATION_MEDI    150007       4.878102e+01
## 100             YEARS_BUILD_MEDI    204488       6.649778e+01
## 101              COMMONAREA_MEDI    214865       6.987230e+01
## 102               ELEVATORS_MEDI    163891       5.329598e+01
## 103               ENTRANCES_MEDI    154828       5.034877e+01
## 104               FLOORSMAX_MEDI    153020       4.976082e+01
## 105               FLOORSMIN_MEDI    208642       6.784863e+01
## 106                LANDAREA_MEDI    182590       5.937674e+01
## 107        LIVINGAPARTMENTS_MEDI    210199       6.835495e+01
## 108              LIVINGAREA_MEDI    154350       5.019333e+01
## 109     NONLIVINGAPARTMENTS_MEDI    213514       6.943296e+01
## 110           NONLIVINGAREA_MEDI    169682       5.517916e+01
## 111               TOTALAREA_MODE    148431       4.826852e+01
## 112     OBS_30_CNT_SOCIAL_CIRCLE      1021       3.320206e-01
## 113     DEF_30_CNT_SOCIAL_CIRCLE      1021       3.320206e-01
## 114     OBS_60_CNT_SOCIAL_CIRCLE      1021       3.320206e-01
## 115     DEF_60_CNT_SOCIAL_CIRCLE      1021       3.320206e-01
## 116       DAYS_LAST_PHONE_CHANGE         1       3.251916e-04
## 117   AMT_REQ_CREDIT_BUREAU_HOUR     41519       1.350163e+01
## 118    AMT_REQ_CREDIT_BUREAU_DAY     41519       1.350163e+01
## 119   AMT_REQ_CREDIT_BUREAU_WEEK     41519       1.350163e+01
## 120    AMT_REQ_CREDIT_BUREAU_MON     41519       1.350163e+01
## 121    AMT_REQ_CREDIT_BUREAU_QRT     41519       1.350163e+01
## 122   AMT_REQ_CREDIT_BUREAU_YEAR     41519       1.350163e+01
```

% of missing data in the whole data matrix:

```
sum(missing.values.df$n_missing) / (ncol(data) * nrow(data)) * 100
```

```
## [1] 22.35851
```