Credit scoring

context

Credit score cards are a common risk control method in the financial industry. It uses personal information and data submitted by credit card applicants to predict the probability of future defaults and credit card borrowings. The bank is able to decide whether to issue a credit card to the applicant. Credit scores can objectively quantify the magnitude of risk.

Generally speaking, credit score cards are based on historical data. Once encountering large economic fluctuations. Past models may lose their original predictive power. Logistic model is a common method for credit scoring. Because Logistic is suitable for binary classification tasks and can calculate the coefficients of each feature. In order to facilitate understanding and operation, the score card will multiply the logistic regression coefficient by a certain value (such as 100) and round it.

At present, with the development of machine learning algorithms. More predictive methods such as Boosting, Random Forest, and Support Vector Machines have been introduced into credit card scoring. However, these methods often do not have good transparency. It may be difficult to provide customers and regulators with a reason for rejection or acceptance.

Task

Build a machine learning model to predict if an applicant is 'good' or 'bad' client, different from other tasks, the definition of 'good' or 'bad' is not given. You should use some techique, such as vintage analysis to construct you label

Data cleaning process:

Missing values

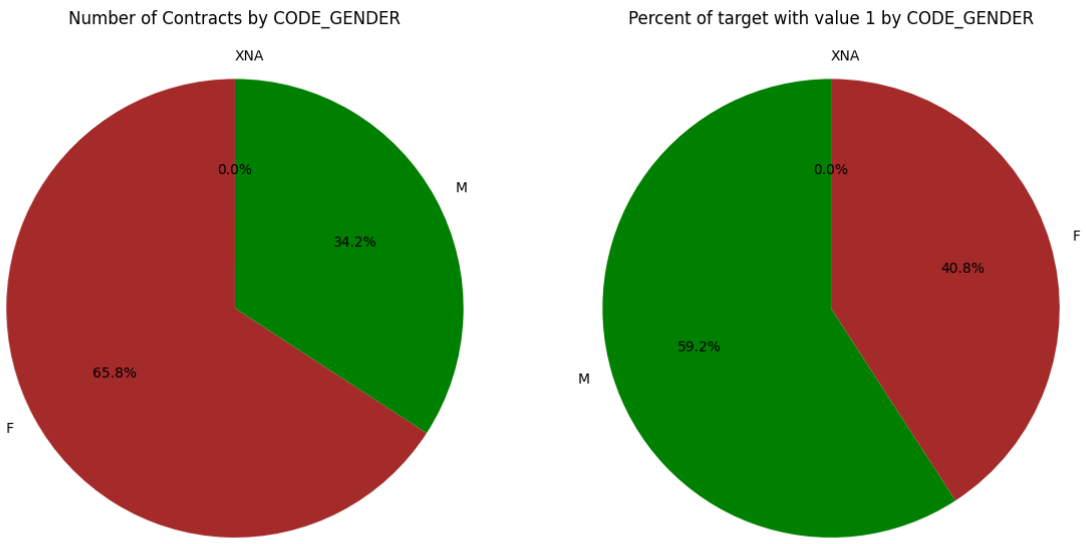
finding missing values in each dataset

drop rows that have maximum missing values exceed the threshhold=0.3

output csv files(dropped)

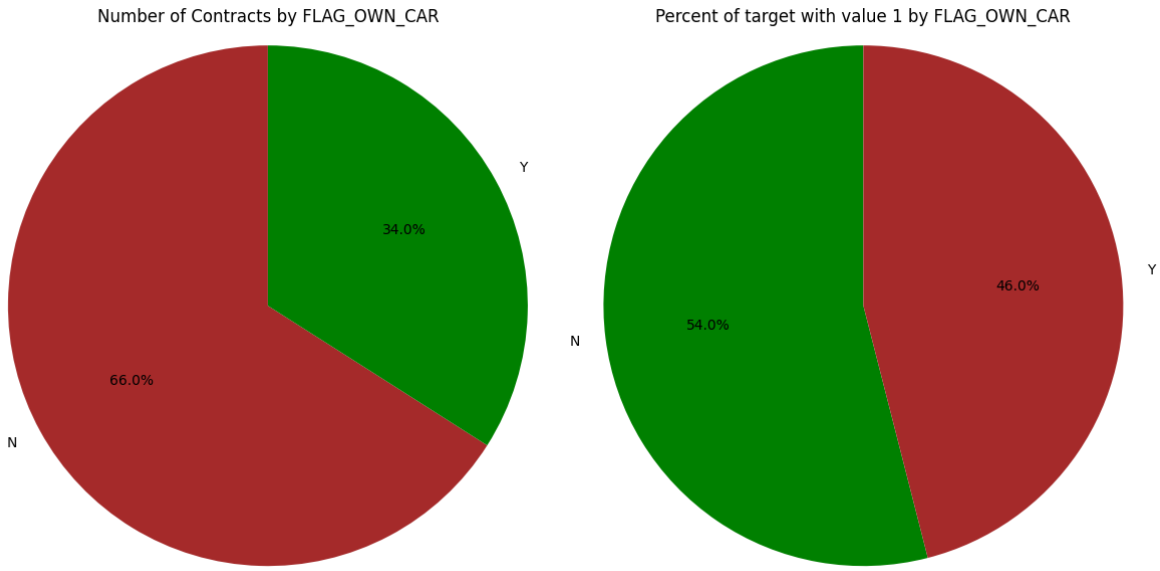
Getting some insights from categorical features :

develop a function plot_stats_pie that generate two pie charts . The first pie chart shows the number of contracts for each gender and the second one displays the percentage of loans with a target value of 1(not returning loans) for each gender



The number of female clients is almost double the number of male clients. Looking to the percent of defaulted credits, males have a higher chance of not returning their loans ~60% comparing with women ~40%

we can consider female are better client than male.



The clients that owns a car are almost a half of the ones that doesn't own one. The clients that owns a car are less likely to not repay a car that the ones that own.

Data filtering:

identifying categorical variables and numerical variables,

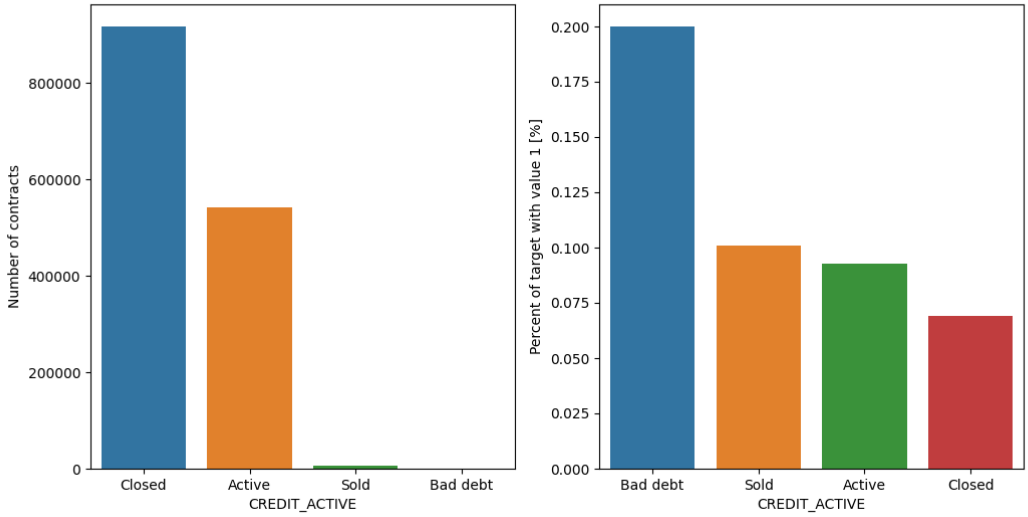
categorical variables: 16
numerical variables: 106

merging *application_train* with *bureau*.

analize the *application_bureau_train* data.

Credit status

Let's see the credit status distribution. We show first the number of credits per category (could be *Closed*, *Active*, *Sold* and *Bad debt*).



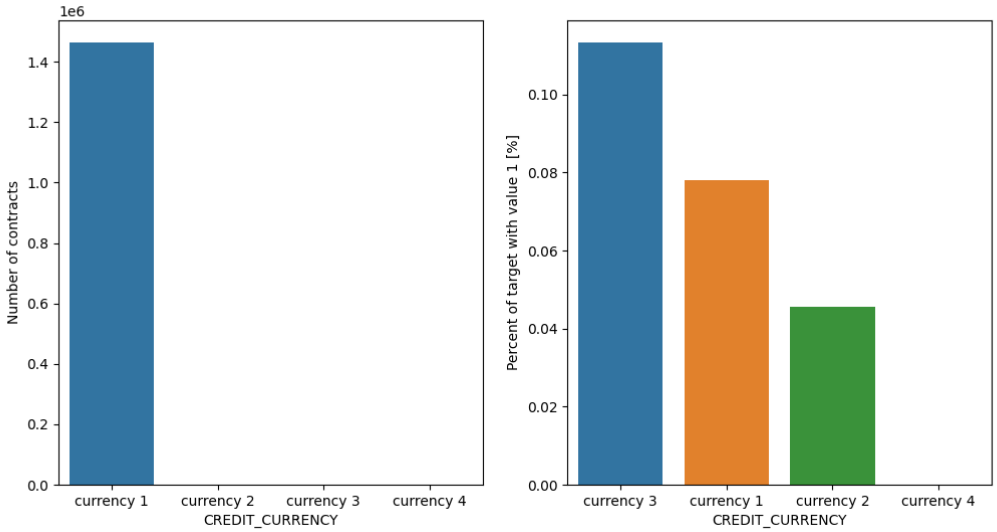
Most of the credits registered at the Credit Bureau are in the status *Closed* (~900K). On the second place are the *Active* credits (a bit under 600K). *Sold* and *Bad debt* are just a few.

In the same time, as percent having **TARGET = 1** from total number per category, clients with credits registered to the Credit Bureau with *Bad debt* have 20% default on the currrent applications.

Clients with credits *Sold*, *Active* and *Closed* have percent of **TARGET == 1** (default credit) equal or less than 10% (10% being the rate overall). The smallest rate of default credit have the clients with credits registered at the Credit Bureau with *Closed* credits.

That means the former registered credit history (as registered at Credit Bureau) is a strong predictor for the dafault credit, since the percent of applications defaulting with a history of *Bad debt* is twice as large as for *Sold* or *Active* and almost three times larger as for *Closed*.

Credit currency

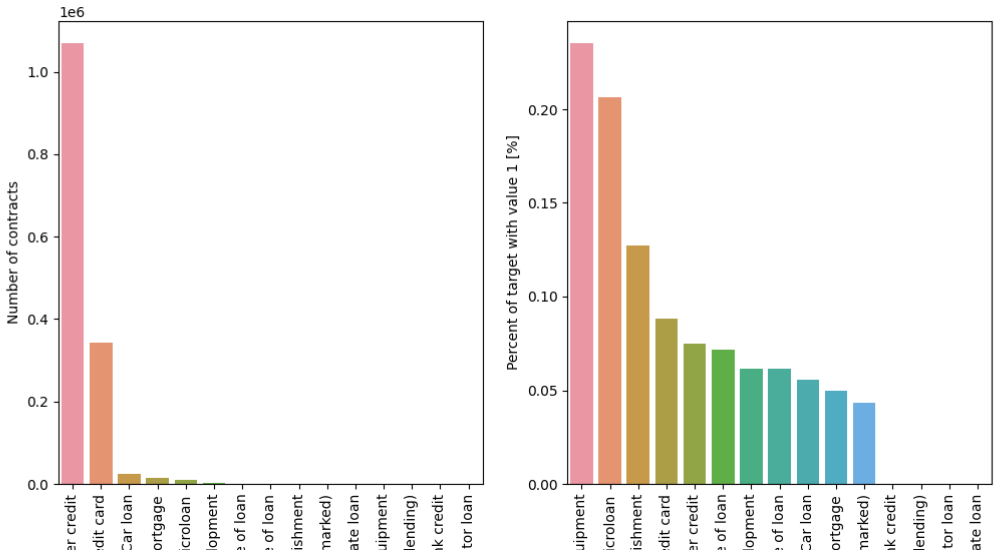


Credits are mostly in *currency_1*.

Depending on the currency, the percent of clients defaulting is quite different. Starting with *currency_3*, then *currency_1* and *currency_2*, the percent of clients defaulting is 11%, 8% and 5%. Percent of defaulting applications for clients that have credits registered with *currency_4* is close to 0.

Credit type

Let's check now the credit types for credits registered at the Credit Bureau.



Majority of historical credits registered at the Credit Bureau are *Consumer credit* and *Credit card*. Smaller number of credits are *Car loan*, *Mortgage* and *Microloan*.

Looking now to the types of historical credits registered at the Credit Bureau, there are few types with a high percent of current credit defaults, as following:

- *Loan for the purchase of equipment* - with over 20% current credits defaults;
- *Microloan* - with over 20% current credits defaults;
- *Loan for working capital replenishement* - with over 12% current credits defaults.