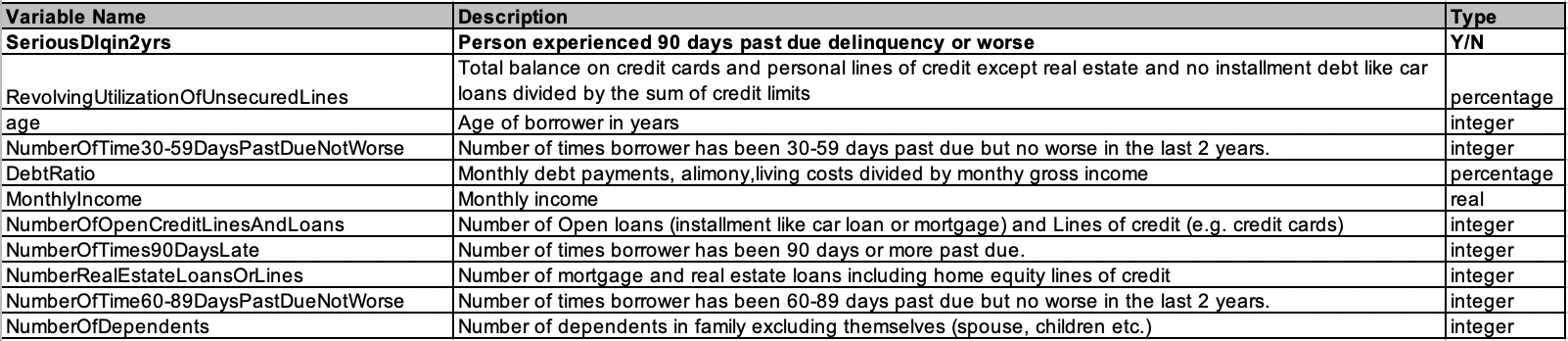
3. **Assumption and data preprocessing**

There are in total ten features in this data set. Below is the table that shows the variable names and description.



It can be seen that the SeriousDlqin2yrs is the dependent variable, and all other variables are independent variables. The independent variables measure the past arrears and basic information including the financial situations and marriage situations.

According to life experience, a person's default status of bank debt can be inferred from the following aspects: age, marital status, income and default status, default history and so on. As it can be seen, this dataset contains most of the above information. Of course, it does not take into account the deeper level of information, such as the macro information during sample extraction, such as the overnight lending rate of Banks, CPI, etc., but this data has played a good role in the preliminary study of bank credit default, so other variables are not considered.

For the features including ‘RevolvingUtilizationOfUnsecuredLines’, ‘DebtRatio’ and ‘MonthlyIncome’, it is discovered from the density plots that they have the problem of outliers. So the values outside the quantile one minus three times of the differences between quantile three minus quantile one and the quantile one plus three times of the differences between quantile three minus quantile one are replaced as NA. For the feature ‘age’, it is found that the minimum value of age is zero, which is unreasonable, so they are transformed as NA. For features relating to past default records, it is discovered that they have many unreasonable values which are larger than 90, so these values are transformed to NA. After these transformations, the NAs are replaced with the mean value of each predictors.