Analysis of the dataset
Credit Card Default

Gaspare Ferraro (520549)
Javad Khalili (546677)
Mario Matovic (583449)

January 3, 2019

University of Pisa
Exam: Data Mining
Year: 2018/2019
Instructors: Dino Pedreschi, Anna Monreale, Riccardo Guidotti

Contents

1. Introduction

This research aimed at the case of customers default payments in Taiwan and compares the predictive accuracy of probability of default among six data mining methods. From the perspective of risk management the binary result of classification will valuable for identifying credible or not credible clients.

TEST TEST TEST TEST TEST TEST TEST TEST TEST TEST TEST TEST TEST TEST TEST TEST TEST TEST TEST TEST TEST TEST TEST TEST TEST TEST TEST TEST TEST TEST TEST TEST TEST TEST TEST

2. Data Understanding

The dataset is composed by 10000 records. Each record represents a customer, described by 24 different attributes.

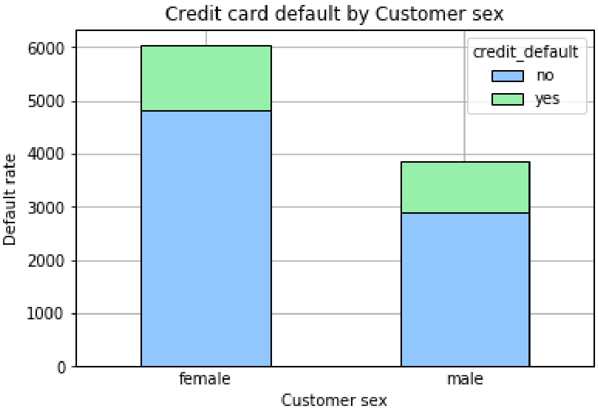
2.1 Data semantics, distribution and statistics

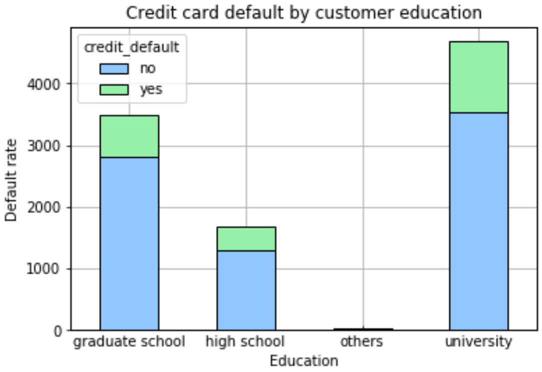
In this section we will analyze, for each attribute, its semantic and we will show interesting statistic and plot. We have used.

We have discretized the continuous attribute by the natural binning method. For these attributes, mode has also been reported.

Sex

Gender of the customer.
A categorical attribute that can assume the values of *male* (3868 of 10000) or *female* (6032 of 10000). Both of the gender values have a similar default rate (25% for males and 20% for females).





Education

Qualification of the customer.

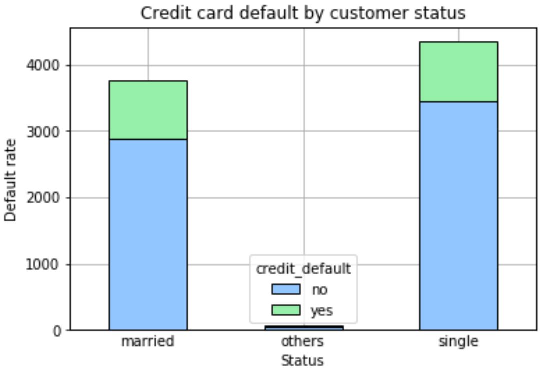
A categorical attribute that can assume the values of *university* (4685 of 10000), *high school* (1672 of 10000), *graduate school* (3480 of 10000) or *others* (36 of 10000). The default rate is again very similar for all the qualifications (around the 20%), except for the *others* which is equal to 5%, but its number of records is very low to make any assumptions.

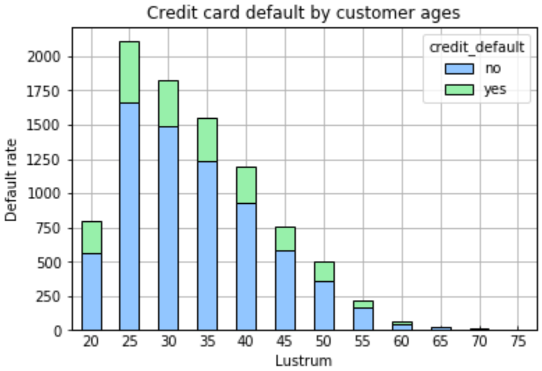
Status

Marital status of the customer.

A categorical attribute that can assume the values of *married* (4685 of 10000), *single* (3757 of 10000) or *other status* (75 of 10000).

The default rate is very similar for all the status (around the 25%).





Age

Age of the customer.

An attribute that can assume integer values in $[21, 75]$ (this is due that in Taiwan the age of majority is 20) and seems to be arranged according to a Gaussian distribution.

The average is 35.5 and the standard deviation is 9.22, 50% of the ages lie in $[28, 41]$. The bin with most elements is 25.

Again the default rate is similar for all the bins (around 25%).

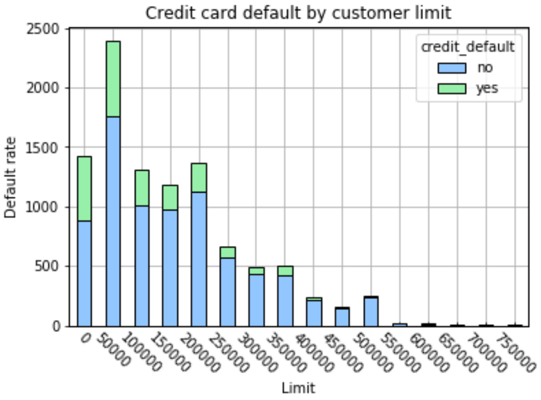
Limit

Limit of the credit card (expressed in NT dollar).

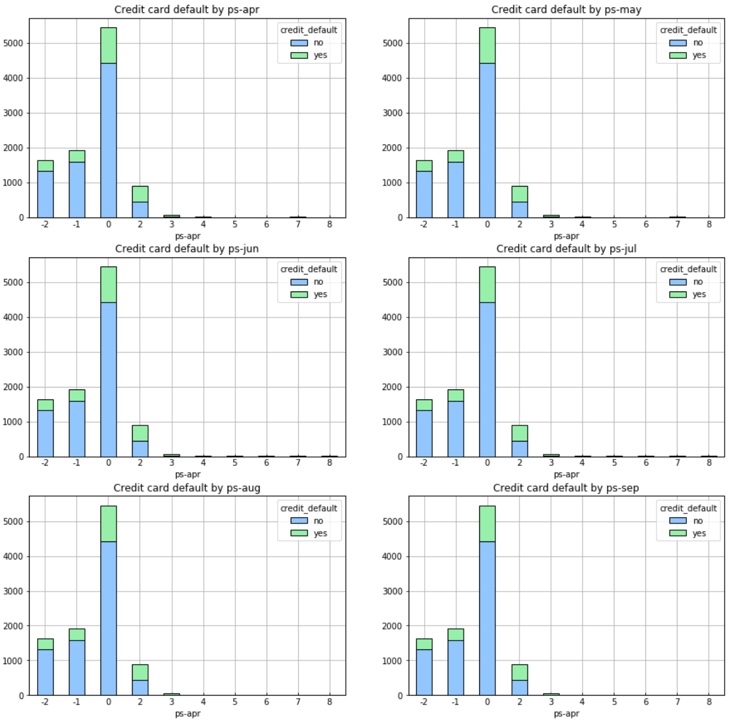
It is the maximum amount the credit card company will let borrow on the account, a continuous attribute that can assume values in $[10000, 780000]$ (all values are multiples of 10000).

The average is 167197 and the standard deviation is 128975, 50% of the ages lie in $[50000, 240000]$. The bin with most elements is 50000.

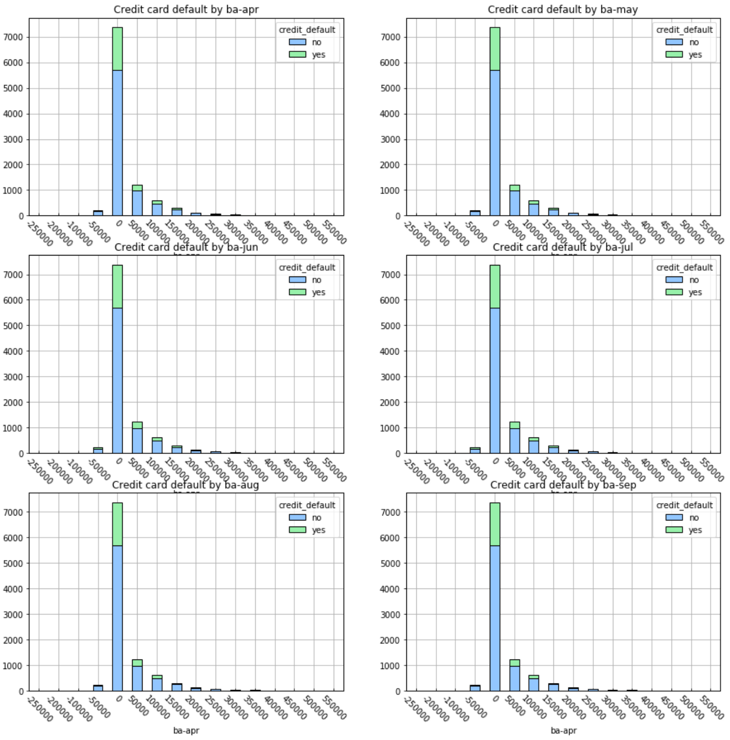
The default rate in this case is very different for each bin as it decrease with the limit: the first bin has a default rate of 38%, the second one of 26% and around 10% for the higher bins.



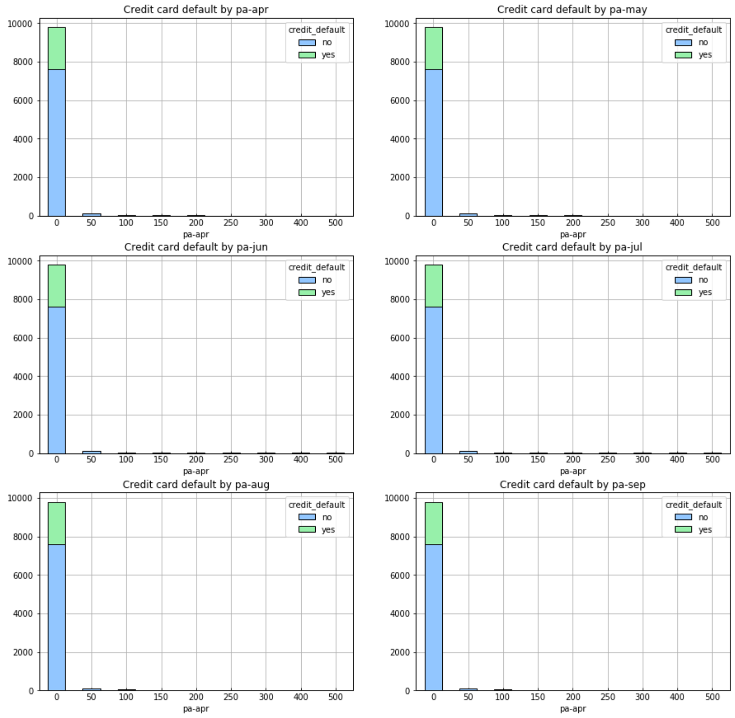
Payment status



Bill Amount



Payment amount



2.2 Assessing data quality

Assessing data quality (missing values, outliers)

2.3 Variables transformations

modifica variabili

2.4 Correlations and redundant variables

Cose Cose cose Cose cose Cose cose

Cose Cose cose Cose cose Cose cose

3. Clustering

3.1 K-means

3.1.1. Choice of attributes and distance function

3.1.2. Choise of the best value of k

3.1.3. Cluster analysis

3.2 DBSCAN

3.2.1. Choice of attributes and distance function

3.2.2. Study of the clustering parameters

3.2.3. Characterization and interpretation of the obtained clusters

3.3 Hierarchical clustering

3.3.1. Choice of attributes and distance function

3.3.2. Discussion of dendograms using different algorithms

3.4 Evaluation of clustering approaches and comparison of the clustering obtained

4. Association Rules Mining

4.1 Frequent patterns extraction with different parameters

4.2 Discussion of the most interesting frequent patterns

4.3 Association rules extraction with different values of confidence

4.4 Discussion of the most interesting rules

4.5 Use the most meaningful rules to replace missing values

4.6 Use the most meaningful rules to predict credit card defaults

5. Classification

5.1 Choice of attributes for the decision trees

5.2 Decision trees interpretation and validation with test and training set

5.3 Discussion of the best prediction model

6. Conclusion