.

Analysis of the dataset *Credit Card Default*

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

January 4, 2019

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

Contents

1. Introduction

This report is aimed to illustrate the phases and the results of the analysis that we have conduced regarding the customers default payments in Taiwan. In particular our target is to better understand under which conditions we should consider a client credibile or not.

Each customer is modelled by a record in the dataset, which is composed by 24 attributes that describes its personal information and its banking data (like the credit limit, payment amount and others).
The analysis is composed in 4 phases:

- Semantical analysis and data manipulations of each customer (data cleaning, variables transformation, redundant variables, ...)
- Use of 3 clustering algorithms (K-Means, DBSCAN, Hierarchical clustering) to group customers according to similarity properties in order to formulate hypotheses about customers credibility.
- Verification of the hypotheses given in the previous phase and determination of association rules, in order to find co-occurrences between attributes.
- Classification of the customers between who fail to make a payment by time and regular customers.

2. Data Understanding

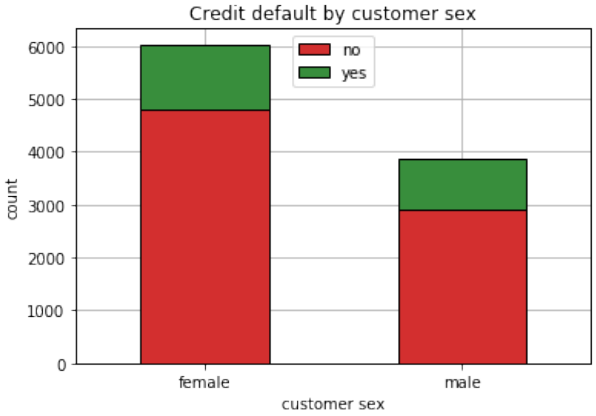
The dataset is composed by 10000 records. Each record represents a customer, described by 24 different attributes.

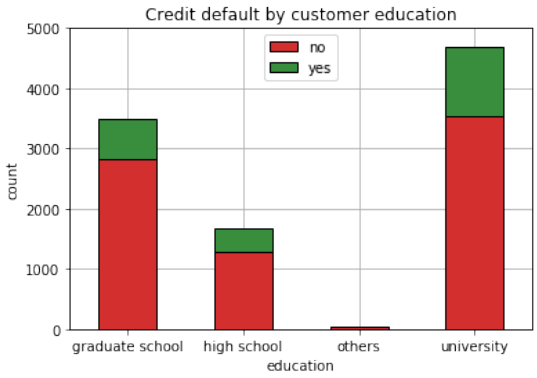
2.1 Data semantics, distribution and statistics

In this section we will analyze, for each attribute, its semantic and we will show interesting statistic and plot. We have used two different colors for who went in credit default (green) and not (red) in order to better visualize their distribution among the different attributes.
We have discretized the continuous attributes by using the natural binning method.
For these attributes the mode of the bin has also been reported as it is more representative.

Sex

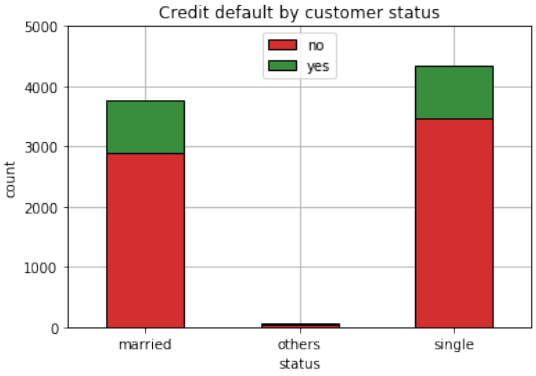
Gender of the customer.
A binary attribute that can assume the values of *male* (3868 of 10000) or *female* (6032 of 10000).
Both of the gender values have a similar credit default rate (25% for males and 20% for females).





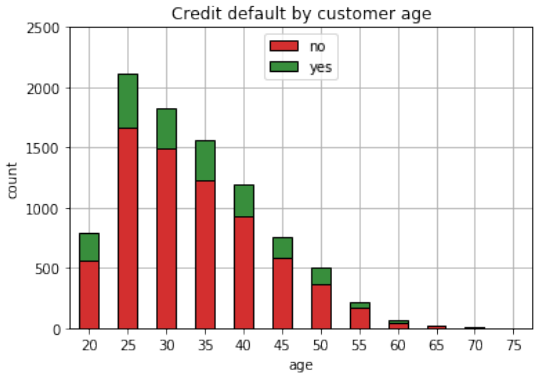
Education

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%).

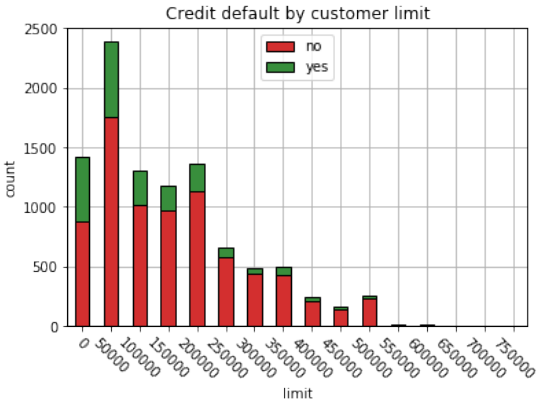


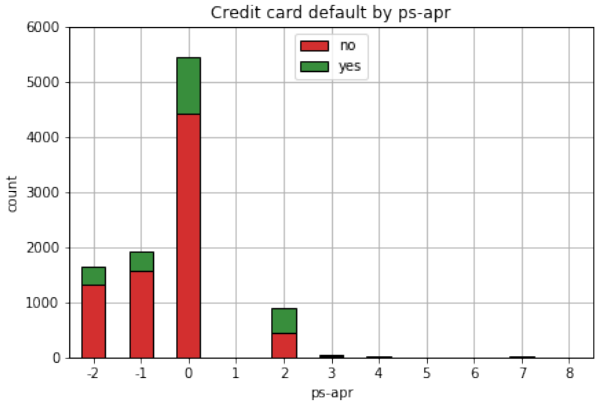
Age

Age of the customer.
An attribute that in the dataset assume integer values in $[21, 75]$, the lower limit to 21 is due that in Taiwan the age of majority is 20, on the other hand the upper limit can be as high as humanly possible. The average age is 35.49, the standard deviation is 9.22. The mode is 29 and the median is 34. The 50% of the ages lie in $[28, 41]$.
We decide to set a bin to 5 year as it represent a good trade-off between the size and number of bins. 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 last bin.

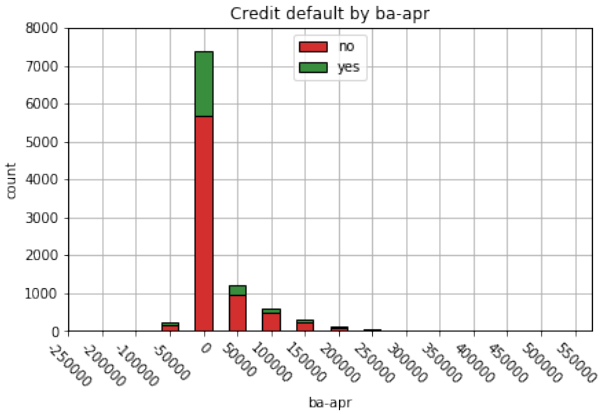


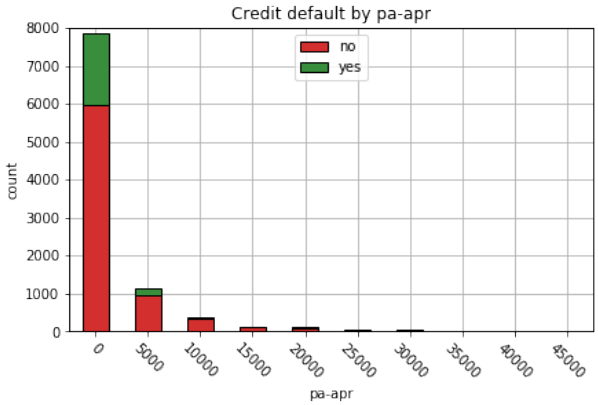


Payment status

History of past payments.
Six categorical attributes that represent the repayment status, one for each month between April and September. A payment status is an integer number in the range $[-2, 9]$ where:
-2 = no consumption
-1 = paid in full
0 = the use of revolving credit
1 = payment delay for one month
2 = payment delay for two months
...
The distribution of the payment status over the six months are all very similar to each other so we only illustate the first one for simplicity. The most frequent bin is always the 0 and around 90% of the customers always lies between -2 and 0.

Bill Amount

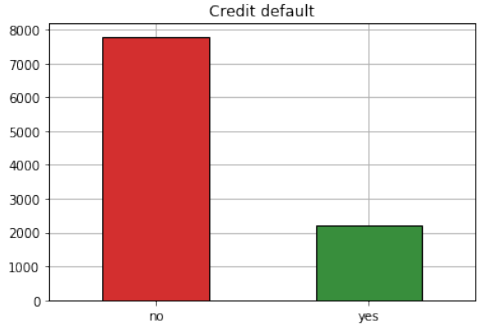




Payment amount

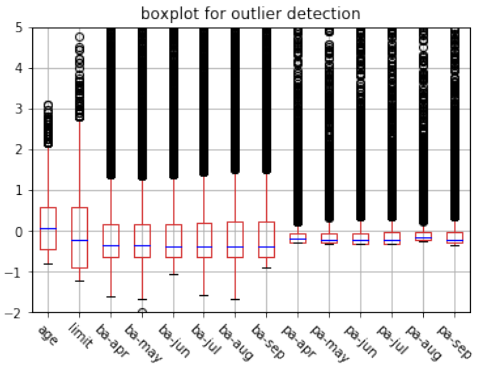
Credit default

prova



2.2 Assessing data quality

In this part we have checked for duplicates and outliers, in order to reduce the quantity of data and to avoid that their presence could negatively affect

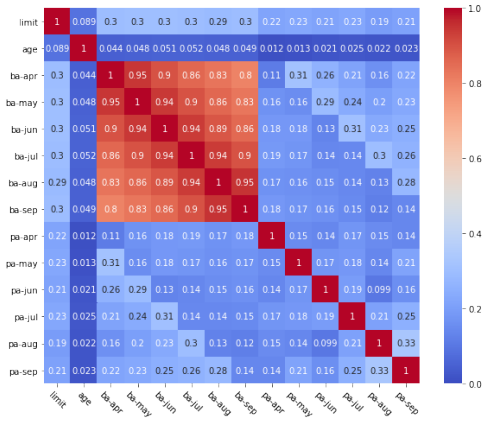


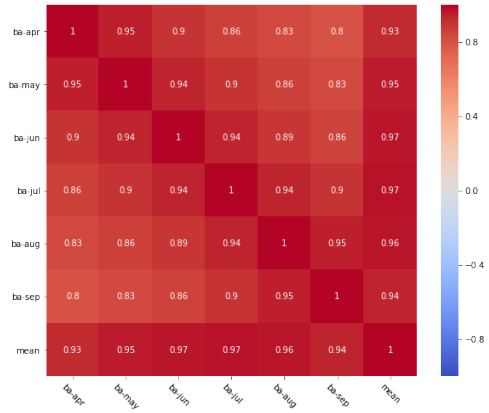
2.3 Variables transformations

modifica variabili
modifica variabili
modifica variabili

2.4 Correlations and redundant variables

Analyzing the correlation matrix of all the continuous attribute we can clearly see that all the attributes related to the bill amount are strongly correlated. All the





We add a new attribute to the dataset called *ba_mean* which is the mean of all the bill amount for each customer. We plot now a new correlation matrix rescricted only to the attributed of.

3. Clustering

3.1 K-means

prova

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

Cose

4.1 Frequent patterns extraction with different parameters

Cose

4.2 Discussion of the most interesting frequent patterns

Cose

4.3 Association rules extraction with different values of confidence

Cose

4.4 Discussion of the most interesting rules

Cose

4.5 Use the most meaningful rules to replace missing values

Cose

4.6 Use the most meaningful rules to predict credit card defaults

Cose

5. Classification

Classificazione

5.1 Choice of attributes for the decision trees

cose

5.2 Decision trees interpretation and validation with test and training set

cose

5.3 Discussion of the best prediction model

cose

6. Conclusion

Conclusione dove parlo di cose.