Tesina

Angelica Urbanelli S271114

Data Spaces, a.y. 2020/2021

Contents

1 Introduction

-spiegare il problema importanza di distinguere buoni e cattivi pagatori statistiche?
-interpretabilità? -falsi buoni peggio di falsi cattivi

2 Dataset description

2.1 History

The dataset used comes from the UCI Machine Learning Repository [2], under the name "South German Credit (UPDATE) Data Set" [4].
Ulrike Grömping, professor at the Beuth University in Berlin, in her paper [3] provides the history of this dataset, her considerations about the data and corrections on the code table.
Briefly, the data come from a large regional bank in the southern Germany that have been collected from 1973 to 1975, and have been originally provided to UCI in 1994 by Professor Dr. Hans Hofmann from Hamburg University [1] as part of a group of datasets in the context of the EU Statelog Project.
Because of many inconsistencies, found while trying to interpret the final results of her experiments, Grömping decided to research the story of this data, that she found in the German literature together with the same dataset with some differences. These informations helped her to fix the code table (a file that explains the encoding of categorical variables) of this dataset by providing a new one (now attached in the .zip downloadable from UCI).
Grömping also explained that it was worth it because, although the dataset contains very old data, it is widely used in many researches in the domain of interpretable machine learning, indeed there are various R packages that include this data, and it *is one of the few data sets on credit scoring that has a meaning attached to variables and their levels*, which is a very important feature when using this kind of data to do experiments whose interpretability is a key point of research.

2.2 General structure

The dataset contains 1000 samples, each one characterized by 20 features and classified as `good` or `bad credit risk`, in particular there are 700 good ones and 300 bad ones [Figure 1]. Customers with “good” credits perfectly complied with the conditions of the contract, while customers with “bad” credits did not comply with the contract as required.
As reported in the aforementioned paper [3], the actual percentage of bad credits was around 5%, and examples of bad credit risk have been heavily oversampled.

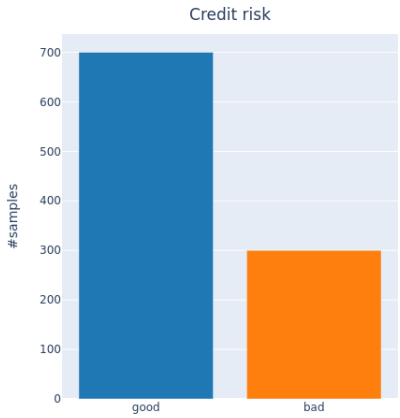

Figure 1: caption

2.3 Features description

Among the 20 features, there are 3 numerical discrete variables:

- **duration**: credit duration in months
- **amount**: credit amount in DM[1]; it is the result of an unknown monotonic transformation, thus original values are not available
- **age**: age of the debtor in years

7 additional numerical variables that have been aggregated into a fixed number of intervals, so that they can be treated as ordinal ones:

- **employment_duration**: duration of debtor's employment with current employer (unemployed; $<$ 1 year ; $\geq$ 1 and $<$ 4 years; $\geq$ 4 and $<$ 7 years; $\geq$ 7 years)
- **installment_rate**: : credit installments expressed as a percentage of debtor's disposable income ($\geq$ 35; $\geq$ 25 and $<$ 35; $\geq$ 20 and $<$ 25; $<$ 20)
- **present_residence**: from how many years the debtor lives in the present residence ($<$ 1 year ; $\geq$ 1 and $<$ 4 years; $\geq$ 4 and $<$ 7 years; $\geq$ 7 years)
- **number_credits**: number of credits including the current one the debtor has (or had) at this bank (1; 2 or 3; 4 or 5; $\geq$ 6)
- **people_liable**: number of people who financially depend on the debtor (i.e., are entitled to maintenance) (from 0 to 2; 3 or more)
- **status**: status of the debtor's checking account with the bank in DM (no checking account; $<$ 0; 0 $\leq$ ... $<$ 200; $\geq$ 200 / salary for at least 1 year)
- **savings**: debtor's savings in DM (unknown/no savings account; $<$ 100; 100 $\leq$ ... $<$ 500; 500 $\leq$ ... $<$ 1000; $\leq$ 1000)

10 categorical variables:

- **credit_history**: history of compliance with previous or concurrent credit contracts (delay in paying off in the past; critical account/other credits elsewhere; no credits taken/all credits paid back duly; existing credits paid back duly till now; all credits at this bank paid back duly)
- **purpose**: purpose for which the credit is needed (others; car (new); car (used); furniture/equipment; radio/television; domestic appliances; repairs; education; vacation; retraining; business)
- **personal_status_sex**: combined information on sex and marital status; sex cannot be recovered from the variable because male singles and female non-singles are coded with the same code; in addition, female widows are not listed in any of the categories (male divorced/separated; female non-single or male single; male married/widowed; female single)
- **other_debtors**: whether there is another debtor or a guarantor for the credit (none; co-applicant; guarantor)
- **property**: the debtor's most valuable property, i.e. the highest possible code is used (unknown / no property; car or other [savings don't fall into this category]; building society savings agreement (mortgage)/life insurance; real estate)

[1] stands for Deutsche Mark, was the official currency of West Germany from 1948 until 1990 and later the unified Germany from 1990 until 2002 [5]

- **other_installment_plans**: installment plans from providers other than the credit-giving bank (bank; stores; none)
- **housing**: type of housing the debtor lives in (for free; rent; own)
- **job**: quality of debtor's job (unemployed/unskilled - non-resident; unskilled - resident; skilled employee/official; manager/self-employed/highly qualified employee)
- **telephone**: whether there is a telephone landline registered on the debtor's name; of course this variable would have no meaning nowadays, but this data come from 1970s (yes; no)
- **foreign_worker**: whether the debtor is a foreign worker (yes; no)

Among these categorical features, a further distinction can be done, by dividing ordinal and nominal ones. In the first group I considered the ones whose values can be ranked, that are **credit_history**, **property** and **job**.

3 Data preprocessing

3.1 Labels encoding

All the features' labels are expressed as integers. Indeed, both nominal and ordinal values have been previously mapped to integers using label encoding from 0 to N-1 (only for **purpouse** and **credit_history**) or from 1 to N (all the other variables) where N is the number of labels for a certain feature. Apparently there is no particular reason for which the donor of the data did this distinction. For what concerns class labels (`credit risk`), they have been mapped to 0 for `bad` and 1 for `good`.
In order to make variables more uniform, two small changes have been done:

- the variable **installment_rate** is the only one among the ordinal variables having a decreasing order, so its mapping has been inverted;
- **credit_history**'s mapping has been changed from $0 \rightarrow 4$ to $1 \rightarrow 5$. Also **purpouse** has the same problem but it will be treated later.

In addition, all nominal variables have to be encoded in a different way with respect to the actual one ($1 \rightarrow N$). Indeed this encoding gives an arbitrary ranking to features that don't have one, and this is a problem when applying distance-based classification algorithms because those could potentially exploit this fictional structure created by the mapping itself.

3.2 One hot encoding vs binary encoding

In order to perform this mapping I took into account two

4 Missing values and outliers

5 Dataset normalization

6 Dataset Analysis

7 preprocessing

-trovare outliers e decidere che farne
-normalizzazione

8 Analysis

-distribuzione variabili -distribuzione variabili rispetto alle classi (scatter plot, violin plot) -correlazione variabili -t-SNE -clustering per vedere se mappa le label

9 dimensionality reduction

-eliminare/aggregare variabili correlate -pca per capire quali sono le feature più importanti -lineare -kpca -eventualmente di nuovo correlazione tra variabili

10 prepare data for training

-oversampling -undersampling -smote -divisione training-test

11 classificazione

-logistic regression (also + regularization ridge or lasso) -tree -random forest (less simple to explain results) -svm (linear, rbf, polinomial, sigmoid) -knn -fda -qda? -lda? bayes? -eventually simple mlp

12 results

-per ogni combinazione (dim-reduction - algoritmo) fare ROC, accuracy, confusion matrix, recall, precision, F1 (armonic mean between precision and recall) -istogrammi ?

References

[1] H. Hofmann. *Statlog (German Credit Data) Data Set*. 1994. URL: https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29.

[2] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.

[3] U. Grömping. *South German Credit Data: Correcting a Widely Used Data Set*. Tech. rep. 4/2019, Reports in Mathematics, Physics and Chemistry, Berlin: Department II, Beuth University of Applied Sciences, Apr. 2019. URL: http://www1.beuth-hochschule.de/FB_II/reports/Report-2019-004.pdf.

[4] U. Grömping. *South German Credit (UPDATE) Data Set*. 2020. URL: http://archive.ics.uci.edu/ml/datasets/South+German+Credit+%28UPDATE%29.

[5] Wikipedia, the free encyclopedia. *Deutsche Mark*. Last edit: 2021-01-3 Accessed: 2021-01-10. URL: https://en.wikipedia.org/wiki/Deutsche_Mark.