Machine Learning - Project (Dare Nome piu decente)

Kasela Pranav[1], Pagani Miriam Beatrice[2], Saviano Marco[3], Zaccaria Antonella[4]

Abstract
Credit card default happens when you've become severely delinquent on your credit card payment.
It's a serious credit card status that not only affects your standing with that credit card issuer, but also your credit standing in general and your ability to get approved for credit cards, loans, and other credit-based services.
When you accept a credit card, you agree to certain terms, e.g. you agree to make your minimum payment by the due date listed on your credit card statement.
If you miss the minimum credit card payment six months in a row, your credit card will be in default; your credit card issuer will likely close your account and report the default to the credit bureaus.
By the time your credit card defaults, you've likely accumulated hundreds of dollars in fees and interest charges.
Unfortunately, your options for clearing up the credit card default may be limited because of the number of payments you've missed on your account.
For this reason, assuming truthful the given data, we show the procedure used to create prevision algorithm aiming to foresee the default payment's client.

Keywords
Keyword1 — Keyword2 — Keyword3

[1] *Matricola: 846965, Department of Informatics, University of Bicocca*
[2] *Matricola: 794274, Department of Informatics, University of Bicocca*
[3] *Matricola: 793516, Department of Informatics, University of Bicocca*
[4] *Matricola: 848647, Department of Informatics, University of Bicocca*

Contents

Introduction[1]

This dataset is available on Kaggle under the name *Default of Credit Card Clients Dataset* and it contains information on default payments, demographic factors, credit data, history of payment and bill statements of 30,000 credit card clients in Taiwan from April 2005 to September 2005.

The 25 attributes and their characteristics are:

- ID: ID of each client
- LIMIT_BAL: Amount of given credit in NT dollars
- SEX: Gender (1=male, 2=female)
- EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 0,5,6=unknown)
- MARRIAGE: Marital status (1=married, 2=single, 3=others, 0=others)
- AGE: Age in years
- PAY_0: Repayment status in September 2005 (-2=no consumption, 0=use of revolving credit, -1=pay duly, 1=payment delay for one month, 2=payment delay for two months, ... 8=payment delay for eight months, 9=payment delay for nine months and above)
- PAY_2: Repayment status in August, 2005 (scale same as above)
- PAY_3: Repayment status in July, 2005 (scale same as above)
- PAY_4: Repayment status in June, 2005 (scale same as above)
- PAY_5: Repayment status in May, 2005 (scale same as above)
- PAY_6: Repayment status in April, 2005 (scale same as above)
- BILL_AMT1: Amount of bill statement in September 2005 respectively (NT dollar)
- BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)
- BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)
- BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)
- BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar)
- BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)
- PAY_AMT1: Amount of previous payment in September 2005 respectively (NT dollar)

- PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)
- PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar)
- PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar)
- PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar)
- PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)
- default.payment.next.month: Default payment (1=yes, 0=no)

Goal:
Looking at the problem we see a potential use in predicting month by month, the default of the clients.

1. Preprocessing

Before proceeding with Feature Engineering and the implementation of Machine Learning, we analyze the dataset with Descriptive Statistics in order to understand more the composition and tendencies of our data.

We notice that in the attribute *EDUCATION* the values 0,5 and 6 are unknown while the value 4 is other, so we decide to categorize 0,5 and 6 as other too.

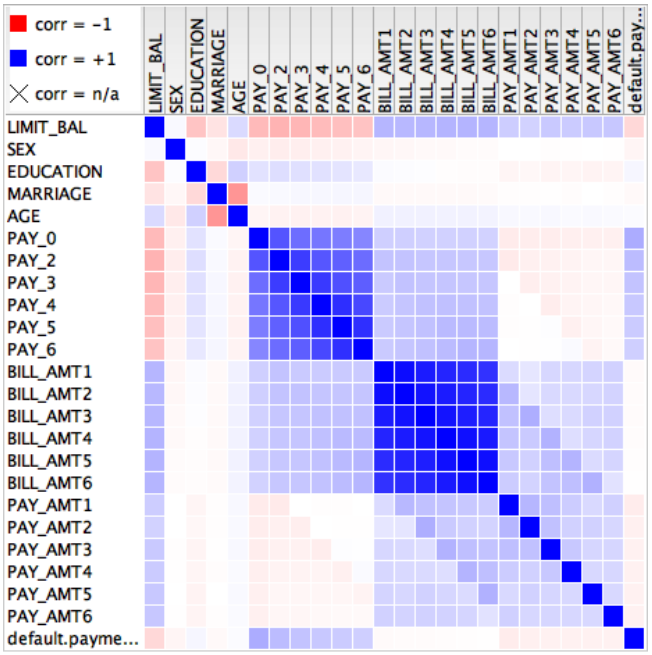

Figure 1. Correlation Plot of all the attributes

The *BILL_AMT1, . . .,BILL_AMT6* are heavely correlated (Figure 1); the best option is to create a new feature which is the mean of these colums, called *AVG_BILL*. We decide to compute the average because the bill indicates how much a person spends and it usually remains constant during the months. We remove the *BILL_AMT1, . . .,BILL_AMT6* attributes.

For the *AGE* attribute we decide to discretize, creating 4 groups: $\{[20,30],(30,40],(40,50],(50,+\infty)\}$. The choice to discretize is made to avoid the possible overfitting phenomena since the *AGE* attribute is imbalanced.

In the *EDUCATION* attribute we combine *0=others* with *3=others* (it is considered to be divorced or separated). We combine *SEX* and *MARRIAGE* to reduce the dimension using the following python code:

```
#df is our dataframe given in the input node
df.loc[((df.SEX == 1) & (df.MARRIAGE == 1))
    , 'SEX_MARRIAGE'] = 1 #married man
df.loc[((df.SEX == 1) & (df.MARRIAGE == 2))
    , 'SEX_MARRIAGE'] = 2 #single man
df.loc[((df.SEX == 1) & (df.MARRIAGE == 3))
    , 'SEX_MARRIAGE'] = 3 #divorced man
df.loc[((df.SEX == 2) & (df.MARRIAGE == 1))
    , 'SEX_MARRIAGE'] = 4 #married woman
df.loc[((df.SEX == 2) & (df.MARRIAGE == 2))
    , 'SEX_MARRIAGE'] = 5 #single woman
df.loc[((df.SEX == 2) & (df.MARRIAGE == 3))
    , 'SEX_MARRIAGE'] = 6 #divorced woman
```

Due to the distribution of *LIMIT_BAL, PAY_AMT1,. . ., PAY_AMT6, AVG_BILL_AMT* we use the logarithmic transformation. Since there are negative values, we translate the minimum to 1 and apply the logarithm to the later. We notice that the distribution of the *AVG_BILL_AMT* has one strong outlier so we remove it to avoid overfitting and standardize the latter for it to be more sparse.

To start studing the classifiers we split our data in training set(67%) and test set(33%), using a stratified sampling on *default.payment.next.month*.

2. Classification

The models selected for the classifications are:

- MLP
- Support Vector Machine: SPegasos
- NaiveBayes
- Bayesian Network:
  - NBTree
  - BayesNet
- Heuristic:
  - J48
  - RandomForest
  - DecisionTree
- Logistic

The target class has 77.7% of no defaulters, 22.3% of defaulters (Figure 2), so we have a class imbalance problem and we cannot use accuracy as an evaluation measure for the classifiers. The optimal measure could be either Recall or Precision, depending on bank's preference. We decide to use two measures: harmonic mean between Recall and Precision: F_1-measure and AUC of the ROC Curve.

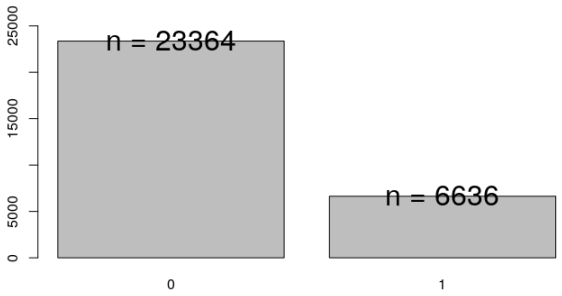

Figure 2. Target Class Distribution

Classifier	Recall	Precision	F_1-measure	Accuracy	AUC
MLP	0.474	0.565	0.515	0.803	0.722
SPegasos	0.202	0.721	0.316	0.806	0.59
NaiveBayes	0.486	0.488	0.487	0.773	0.742
NBTree	0.481	0.53	0.504	0.791	0.754
BayesNet(TANB)	0.421	0.586	0.49	0.806	0.754
J48	0.325	0.631	0.429	0.809	0.629
RandomForest	0.325	0.612	0.425	0.805	0.723
DecisionTree	0.401	0.402	0.401	0.736	0.635
Logistic	0.221	0.685	0.335	0.805	0.745

Table 1. Evaluation measures using Holdout

In the Table 1 we see that using holdout, for the Recall Measure NaiveBayes and NBTree are the optimal classifiers, but NaiveBayes has a low precision of 0.488. Using the F_1-measure or AUC the best classifiers are MLP, NBTree, BayesNet and NaiveBayes.
Since the Holdout method depends on the particular test set, we use the 3-Folds Cross Validation obtaining:

Classifier	Recall	Precision	F_1-measure	Accuracy	AUC
MLP	0.372	0.662	0.476	0.819	0.773
SPegasos	0.22	0.703	0.334	0.807	0.597
NaiveBayes	0.503	0.485	0.494	0.772	0.749
NBTree	0.456	0.561	0.501	0.8	0.748
BayesNet(TANB)	0.427	0.59	0.495	0.808	0.759
J48	0.351	0.63	0.451	0.811	0.651
RandomForest	0.326	0.608	0.424	0.804	0.729
DecisionTree	0.408	0.406	0.407	0.737	0.634
Logistic	0.241	0.673	0.355	0.806	0.748

Table 2. Evaluation measures using 3-Folds Cross Validation

In the Table 2 we see that using 3 Folds Cross Validation, for the Recall Measure NaiveBayes is the optimal classifiers. Using the F_1-measure the best classifiers are NBTree, BayesNet and NaiveBayes. While for the AUC the best one is the MLP.

3. Feature Selection

For the feature selection we use 5 filters:

- Gain Ratio feature evaluation
- Information Gain Ranking Filter
- Symmetrical Uncertainly Ranking Filter
- Relief Ranking Filter(k=10)
- Correlation Ranking Filter

But all these filters select all attributes. By using these filters we gain no additional information. The next approach is to use a Wrapper, we decide to use as the classifier BayesNet and NBTree with 10 folds using as evaluation measure AUC. The wrapper with BayesNet selects 10 attributes:

- LIMIT_BAL
- EDUCATION
- PAY_0
- PAY_2
- PAY_3
- PAY_4
- PAY_6
- PAY_AMT1
- PAY_AMT4
- SEX_MARRIAGE

While the wrapper with NBTree select the following 10:

- LIMIT_BAL
- EDUCATION
- PAY_0
- PAY_2
- PAY_3
- PAY_6
- PAY_AMT3
- PAY_AMT4
- AgeBin
- SEX_MARRIAGE

References

[1] UCI I-Cheng Yeh. https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset.

[2] Luca Basanisi. https://www.kaggle.com/lucabasa/credit-card-default-a-very-pedagogical-notebook.

[3] Latoya Irby. https://www.thebalance.com/what-is-credit-card-default-960209.

[4] ItalPress. https://www.italpress.com/lifestyle/mancato-pagamento-carta-credito/.