

Machine Learning - Project (Dare Nome piu decante)

Kasela Pranav¹, Pagani Miriam Beatrice², Saviano Marco³, Zaccaria Antonella⁴

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

¹Matricola: 846965, Department of Informatics, University of Bicocca

²Matricola: 794274, Department of Informatics, University of Bicocca

³Matricola: 793516, Department of Informatics, University of Bicocca

⁴Matricola: 848647, Department of Informatics, University of Bicocca

Contents

Introduction	1
1 Preprocessing	2
2 Classification	2
3 Feature Selection	2
References	3

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 0,5,6 are unknown and there are only 345 values, so we decide to proceed with the record removal. The *BILL_AMT1*, ..., *BILL_AMT6* are heavily correlated; the best option is to create a new feature which is the mean of these columns, 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 0=others. 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.

To start studying 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, 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.

Classifier	Recall	Precision	F_1 —measure	Accuracy	AUC
MLP	0.417	0.620	0.50	0.813	0.749
SPegasos	0	NaN	NaN	0.777	0.5
NaiveBayes	0.519	0.465	0.490	0.760	0.718
NBTree	0.460	0.562	0.505	0.800	0.744
BayesNet(TANB)	0.438	0.602	0.507	0.810	0.753
J48	0.331	0.615	0.431	0.805	0.657
RandomForest	0.343	0.614	0.441	0.806	0.724
DecisionTree	0.411	0.408	0.410	0.736	0.625
Logistic	0.172	0.649	0.272	0.795	0.737

The J48's method behaves like a ZeroRule, it classifies all clients as non defaulters. For the Recall Measure NaiveBayes is the optimal classifier, but it has a low score of 0.519, while using the F_1 —measure or AUC the best classifiers are MLP, NBTree and BayesNet.

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 with 10 folds using as evaluation measure AUC. The wrapper selects 12 attributes:

- LIMIT_BAL
- EDUCATION
- PAY_0
- PAY_2
- PAY_3
- PAY_4
- PAY_6
- PAY_AMT1
- PAY_AMT2
- PAY_AMT4
- AVG_BILL_AMT
- SEX_MARRIAGE

References

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