Springboard Data Science Capstone Project   
Predicting the probability of default of credit card client

Machine Learning Practice

Jing Xu

Contents

1 Introduction

2 Data Acquisition and Cleaning

3 Data Exploration

4 Inferential Statistics

5 Modeling

6 Assumptions and Limitations

7 Conclusions

1. Introduction

This project is based on the paper “The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients” (Yeh, I. C., & Lien, C. H. 2009). The dataset was donated at 2016-01-26 and is available on the UC Irvine Machine Learning Repository website. The dataset includes historical payment data from a bank in Taiwan in October 2005. The purpose of the research paper is to predict the credit card default rate and compare the predictive accuracy by using different mining techniques. The purpose of my project is to use the open data and do a practice project about machine learning.

1. Data Description

The dataset contains 30,000 observations and each observation represents an individual credit card holder in a Taiwan bank in October 2005.

The response variable employed a binary value to represent default payment, in which y=1 represents default and y=0 represents non-default.

The explanatory variables include 23 factors, which are:

* X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
* X2: Gender (1 = male; 2 = female).
* X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
* X4: Marital status (1 = married; 2 = single; 3 = others).
* X5: Age (year).
* X6 - X11: History of past payment.

Data tracked the past monthly payment records from April to September 2005 and the corresponding values are the repayment statue:

* + - * + -1 = pay duly;
        + 1 = payment delay for one month;
        + 2 = payment delay for two months;
        + . . .;
        + 9 = payment delay for nine months and above.
* X12-X17: Amount of bill statement (NT dollar) from April to September 2005.
* X18-X23: Amount of previous payment (NT dollar) from April to September 2005.

1. Data wrangling
   1. Drop null values and duplicate records

For the accuracy of the data, I checked null values and duplicates in the dataset. There is no null value but have 35 duplicates in the dataset. Even it is a small percentage compared to 30,000 observations, I deleted these duplicates for consistency.

* 1. Data cleaning

In the dataset, there are 23 explanatory variables. Some are discrete values and others are dummy variables. To insure the quality of analysis, I need to check outliers before doing further statistics tests. It is hard for me to find outliers from past payments and billing statements, as financial situations vary largely from individual to individual. However, outliers in dummy variables and variables such as gender, education etc. are easy to detect. Therefore, my focus in this project is on variables from X2 to X11.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **sex** | **education** | **marriage** | **age** |
| **count** | 29964.000000 | 29964.000000 | 29964.000000 | 29964.000000 |
| **mean** | 1.603724 | 1.853624 | 1.551896 | 35.488353 |
| **std** | 0.489131 | 0.790424 | 0.521996 | 9.219374 |
| **min** | 1.000000 | 0.000000 | 0.000000 | 21.000000 |
| **25%** | 1.000000 | 1.000000 | 1.000000 | 28.000000 |
| **50%** | 2.000000 | 2.000000 | 2.000000 | 34.000000 |
| **75%** | 2.000000 | 2.000000 | 2.000000 | 41.000000 |
| **max** | 2.000000 | 6.000000 | 3.000000 | 79.000000 |

Table 1. Statistical Description

By checking minimum and maximum values of sex, education and marriage in Figure 1, I found that some values are not explained in the raw data.  The gender column looks good, but education has values equal to 0,5 and 6, and marriage has values equal to 0, which are not listed in the data description. In order know more about the characters of these outliers, I calculated each label's default rate.

|  |  |  |  |
| --- | --- | --- | --- |
| **default payment next month** | **0** | **1** | **default\_percentage** |
| **marriage** |  |  |  |
| **0** | 49 | 5 | 0.092593 |
| **1** | 10442 | 3200 | 0.234570 |
| **2** | 12605 | 3340 | 0.209470 |
| **3** | 239 | 84 | 0.260062 |

Table 2 The relationship between marriage and default rate

The default rate in the column 'marriage=0' is 0.09.I was thinking to classify them into the 'other' label in 'marriage', but the default rate in 'other' label is 0.26, which may violate the final outcome if I do so.

|  |  |  |  |
| --- | --- | --- | --- |
| *default payment next month* | *0* | *1* | *default\_percentage* |
| *education* |  |  |  |
| *0* | *14.0* | *0.0* | *0.000000* |
| *1* | *8531.0* | *2032.0* | *0.192370* |
| *2* | *10691.0* | *3327.0* | *0.237338* |
| *3* | *3678.0* | *1237.0* | *0.251679* |
| *4* | *116.0* | *7.0* | *0.056911* |
| *5* | *262.0* | *18.0* | *0.064286* |
| *6* | *43.0* | *8.0* | *0.156863* |

Table 3 The relationship between education and default rate

Similar situation happened in the column ‘education’. I checked the reference paper but there is no further expansion on data values which are no specified in the data description. Since there are 29964 data in the database, and the outliers are not that many, I decided to drop these unspecified data.

When checking outliers of historical payment, value ‘-2’ is ,apprantly, an outlier in past payments as it is not stated in data description.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | pay\_1 | pay\_2 | pay\_3 | pay\_4 | pay\_5 | pay\_6 |
| count | 29964.00000 | 29964.000000 | 29964.000000 | 29964.000000 | 29964.000000 | 29964.000000 |
| mean | -0.01682 | -0.131925 | -0.164364 | -0.218896 | -0.264451 | -0.289381 |
| std | 1.12345 | 1.196278 | 1.195888 | 1.168186 | 1.132194 | 1.149067 |
| min | -2.00000 | -2.000000 | -2.000000 | -2.000000 | -2.000000 | -2.000000 |
| 25% | -1.00000 | -1.000000 | -1.000000 | -1.000000 | -1.000000 | -1.000000 |
| 50% | 0.00000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 75% | 0.00000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| max | 8.00000 | 8.000000 | 8.000000 | 8.000000 | 8.000000 | 8.000000 |

Table 4. Statistical Description of past payment

I was trying to delete data in which past payment value equals to -2. However, I found that I would loss 20% data and it will be a big loss if I do that. Therefore, I tried to clarify ‘-2’ to another group. As the default rate of value=-2 is similar to value=0, I classified it to value 0. This method is less rigorous as I didn’t compare data similarity with other explanatory variables.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| default payment next month | Pay\_1 | PAY\_2 | pay\_3 | pay\_4 | pay\_5 | pay\_6 |
| -2 | 0.132364 | 0.182836 | 0.185450 | 0.192682 | 0.196899 | 0.200452 |
| -1 | 0.167899 | 0.159775 | 0.155908 | 0.158923 | 0.162060 | 0.169979 |
| 0 | 0.128113 | 0.159123 | 0.174512 | 0.183288 | 0.188529 | 0.188444 |
| 1 | 0.340333 | 0.178571 | 0.250000 | 0.500000 | 0.000000 | 0.000000 |
| 2 | 0.691182 | 0.555924 | 0.515580 | 0.533267 | 0.541889 | 0.506508 |
| 3 | 0.757764 | 0.616564 | 0.575000 | 0.611111 | 0.634831 | 0.641304 |
| 4 | 0.684211 | 0.505051 | 0.573333 | 0.661765 | 0.602410 | 0.625000 |
| 5 | 0.500000 | 0.600000 | 0.571429 | 0.514286 | 0. 588235 | 0.538462 |
| 6 | 0.545455 | 0.750000 | 0.608696 | 0.400000 | 0.750000 | 0.736842 |
| 7 | 0.777778 | 0.600000 | 0.814815 | 0.827586 | 0. 827586 | 0.826087 |
| 8 | 0.578947 | 0.000000 | 0.666667 | 0.500000 | 1.000000 | 1.000000 |

Table 5. The relationship between past payment and default rate

1. Story telling

In the last chapter, I explained outliers I found and implemented corresponding solutions. Data visualization can give us a vivid and further understanding to the dataset. In order to understand the relationship between education and default rate, I made a new table without outliers and plotted a barplot.

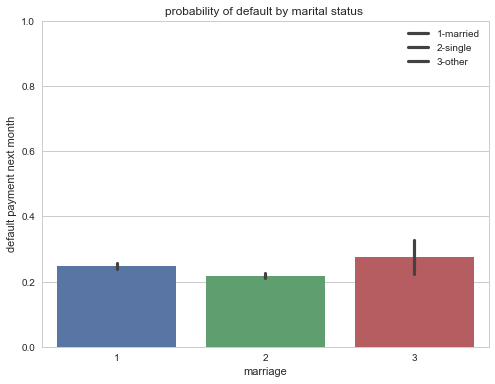
non-default default default\_percentage

married 7787 2564 0.247706

single 9802 2723 0.217405

other 195 74 0.275093

The adjusted graph is:



As we can see from the graph, there is no big difference between marital status and default rate, while ‘other’ column has higher default rate. The value=3 represents “other” in the dataset, which may represent divorce, but I couldn’t find reference.

Similarly, I got the adjusted table and graph which explained the relationship between education and default rate.

non-default default default\_percentage

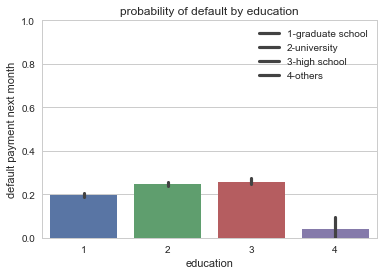
graduate 6084 1479 0.195557

graduate 8670 2846 0.247134

high school 2961 1033 0.258638

others 69 3 0.041667

The graph is :



1. Machine learning

The main purpose of this project is to use UCI Machine Learning dataset do practice of data mining techniques. In the practice, I will use K-NN and Logistic Regression methods and compare accuracy of different predicting methods.

* 1. K-NN

Step 1: Standardization

It is not difficult to notice that values of different explanatory variables vary largely. To avoid this problem, I standardized values in each column.

Step 2: Determine K value

In K-NN, k value represents the complexity of the model, because the larger the k value, the smoother the boundary is. On the other hand, smaller k can lead to overfitting. In order to avoid this problem, I visualized accuracy scores with different k values.

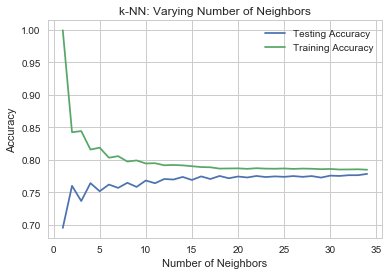


Figure 1. Accuracy scores in different k values

As we can see in the Figure 1, the training accuracy score gets very high at 18. Therefore, it is reasonable to choose k=18 in this project.

Step 3: Classification Report

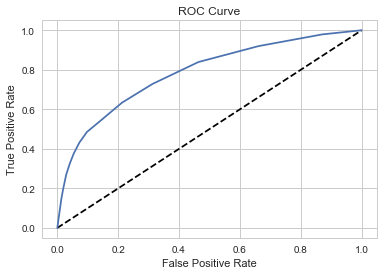
We can evaluate the performance of the k-NN classifier based on its accuracy. However, accuracy is not always an informative metric. In this project, I evaluated the performance of binary classifiers by generating a classification report.

precision recall f1-score support

0 0.83 0.95 0.89 4661

1 0.66 0.34 0.45 1332

avg / total 0.80 0.81 0.79 5993



AUC scores computed using 5-fold cross-validation: [ 0.72311916 0.72904602 0.75550898 0.76038856 0.75471059]

According to the result, ROC AUC of the KNN model is quite high at 0.74. We will do more investigation and comparison with other supervised machine learning models.

* 1. Logistic

In this project, the response variable is a dummy variable, in which 1 is default. The major advantage of this approach is that it can produce a simple probabilistic formula of classification. The weaknesses are that LR cannot properly deal with the problems of non-linear and interactive effects of explanatory variables.

Step 1 Standardization

Same step.

Step 2 Hyperparameter Tuning With GridSearch CV

Logistic regression also has regularization parameters. In this case, I chose C and penalty. C controls the *inverse* of the regularization strength. A large CC can lead to an *overfit* model, while a small CC can lead to an *underfit* model. P is used to specify the norm used in the penalization. The output of these two values are 3.73 and L2 respectively.

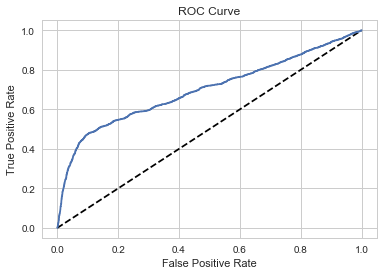
Step 3 Classification Report

precision recall f1-score support

0 0.86 0.92 0.89 9367

1 0.61 0.44 0.51 2619

avg / total 0.80 0.82 0.80 11986



AUC scores computed using 5-fold cross-validation: [ 0.69848033 0.70974544 0.729946 0.73764843 0.73257638]

According to the result, ROC AUC of the logistic model is at 0.72, which is slightly lower than KNN method.We will do more investigation and comparison with other supervised machine learning models.

1. Conclusions

In this project, we first explored the raw data and found outliers. Then, we applied some cleaning techniques to clean outliers. After that, we did data exploratory analysis and applied supervised machine learning model to fit the data. This is what we found from the research.

Data Exploration Conclusions:

1.

Modeling Conclusions:

In the machine learning part, I used two different supervised classification algorithms (Logistic Regression, KNeighbors) to train the predictive model by using 70% of the data. The remaining 30% was tested to evaluate the model.

1. Using performance evaluation metrics, we found that the KNN gives the best model performance in term of predictability.

2.The ROC AUC is 0.74. In terms of features, we found that we do not need do any feature modification since it didn’t improve any predictive score.