University of California Santa Barbara

Course : DS 100

Building a Classifier for Credit Default Risk

June 13, 2020

Author	*Student ID*
Thiha Aung	9680729
Julia Chen	9529850

Contents

1 Abstract

In order to predict the clients who default and do not default next month, we used Principal Component Analysis to explore the relationship between Bill Amount and Payment Amount, indeed we used logistic regression and confusion matrix for building the classifier. Based on the Principal Component Analysis we reduced out 12 variables into 7 PCs and gives us 78% of the accuracy; for the logistic regression, we try to fit the model using simple logistic model, over-sampling, under-sampling and feature selection. All in all our best model is under-sampled model with feature selection.

2 Introduction

2.1 Goals & Questions

The data contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. What we interested in:

- To explore the relationship between bill amount and payment amount.
- To identify key features that could best predict credit card default.
- To build an efficient logistic classifier

2.2 Literature Review

It is a very common thing to over issue cash and credit cards in order to stimulate economy and increase market share. I-Cheng Yeh and Che-hui Lien [1], used six data mining techniques, such as k-nearest neighbor, logistic regression, classification trees and so on to predict the probability of default of credit card clients and concluded that the artificial neural network is the only one that can accurately estimate the data which gives 87.88% classification accuracy. Even though we did not learn artificial neural network, but we used other techniques we learned in the DS100 to analyze this data and build an efficient classifier.

3 Dataset

3.1 Description

The dataset consists of 30000 distinct credit card clients and 24 variables, and the descriptions of each variable are shown below:

- X1**(numeric)**: Amount of the given credit (NT dollar). (1 NT \$ $\approx$ 0.034 USD)
- X2**(categorical)**: Gender (1 = male; 2 = female).
- X3**(categorical)**: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
- X4**(categorical)**: Marital status (1 = married; 2 = single; 3 = others).
- X5**(numeric)**: Age (year).
- X6−X11**(categorical)**: History of past payment. X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005;...; 11 = the repayment status in April, 2005. The measurement scale for the repayment status is: 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.
- X12−X17**(numeric)**= amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005;...;X17 = amount of bill statement in April, 2005.

- X18−X23**(numeric)**: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005;...;X23 = amount paid in April, 2005.
- X24**(predict variable)**: Default payment next month: (1=yes, 0=no)

3.2 Ethical Considerations

We got this data set from the UC Irvine Machine Learning Repository, while the original was collected by Chung Hua University and Tamkang University, Taiwan. Besides, the usage or the analysis of this data should not cause any harm to those represented in the data since it is completely anonymous.However, the dataset contains sex and age as predictive features. While building the model, we will explore if the model is baised against any sex or age. It is very possible that human bias can be carried along to the regression model.

4 Methods & Exploratory Analysis

4.1 Data Cleaning

The data has no missing variables and missing data, so we will not handle the missing data. However, in order to a better understanding of the categorical variables, we checked the feature for "SEX", "EDUCATION", "MARRIAGE" and "PAY_ 0", "PAY_ 2" to "PAY_6".

- "SEX": Each variable is documented such that male = 1, female = 0. No action is needed further.
- "EDUCATION": Value 0 is not documented, value 4 is others 5,6 unknown. Hence, it is reasonable to assume that values 0, 4, 5, 6 are same and set all of them to 4.
- "MARRIAGE" : 0 is undocumented and 3 is others. For interpretability, we set both values to 3 to form a group of people who are either divorced or doesn't want to answer as 3.
- "PAY_ 0", "PAY_ 2" to "PAY_6": Note that we have data point -2 which has no description. Since -1 is described as pay duly, we assume -2 is also pay duly or pay one month ahead and set both to 1.
- Since "EDUCATION" feature is ordinal categorical data, we will do one-hot encoding.

4.2 Exploratory Data Analysis

4.2.1 Exploration of Response Variable

After counting the default payment next month based on the its value, we get that 23364 0's (No default) and 6636 1's (Default).

Our response variable is imbalanced, and the ratio of no-default payment and default payment instances is 78:22.

default payment next month	LIMIT_BAL	SEX	MARRIAGE	BILL_AMT1	BILL_AMT2	PAY_AMT1	PAY_AMT2
0	178099.726074	1.614150	1.564929	51994.227273	49717.435670	6307.337357	6640.465074
1	130109.656420	1.567058	1.530289	48509.162297	47283.617842	3397.044153	3388.649638

Figure 2: Group by Default Payment

- Overall, Sex and Education and Marriage status are relatively same whatever if there is a default payment next month or not.

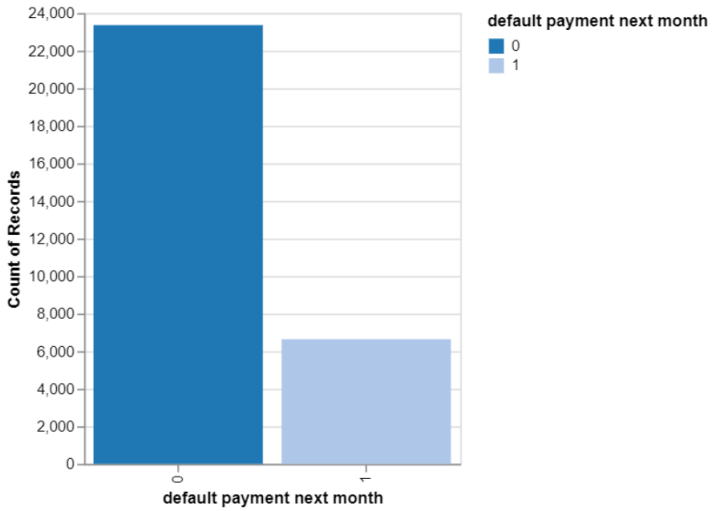

Figure 1: Count of Default Payment and No Default Payment

- Repayment status from April to September is highly correlated with default payment, if there is no default payment, the repayment status will be negative which means it is pay duly.
- Amount of bill statement of no default payment next month is slightly higher than default payment next month.
- Amount of previous payment of no default payment next month is significantly higher than default payment next month.

Indeed, we could groupby other categorical variable and visualize such as Sex, Education, Marital status to get a more detailed sense of our data.

4.2.2 Visualizations of Categorical Variables

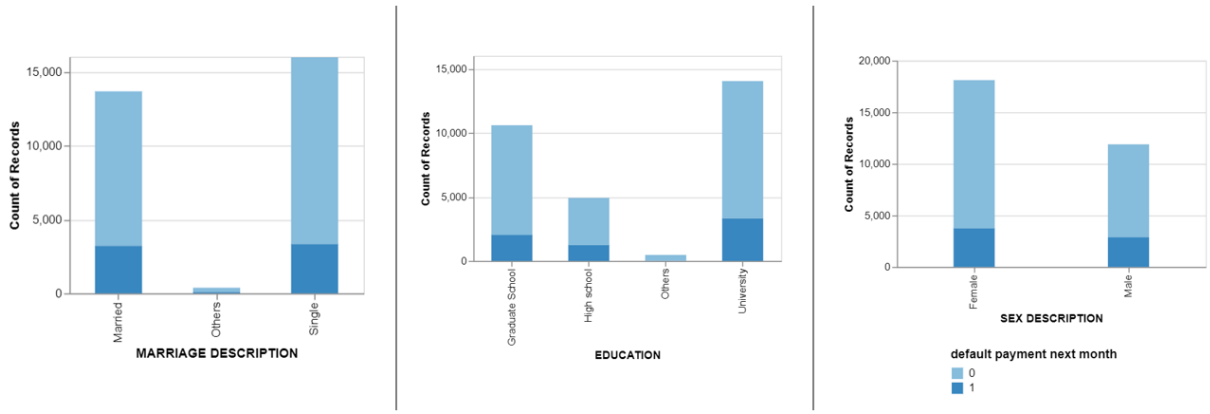

Figure 3: Visualizations of Marriage, Education and Sex

Since our data is imbalanced, so for all plot above, there are more default payment than no default payment. Some interesting thing we could find in those plots:

- There are more single clients compared to married clients. But only 20.9 percent of single clients default whereas 23.4 percent of clients in marriage default.

- Education level somehow affects the chances of defaulting credit. 25.1 percent of people with high school education default and 23.7 percent of people with University education default. However, it is very interesting to see only 19.23 percent of people with graduate school education default. Hence, the higher the education, we can see the less people default. This is a strong indicator that education is a good predictive feature.
- There are more females than males with credit cards in the data set. However, only 20.8 percent of female population default where as 24.1 percent of male population default. This suggests male clients have a higher likelihood of defaulting.

4.2.3 Visualizations of Numerical Variables

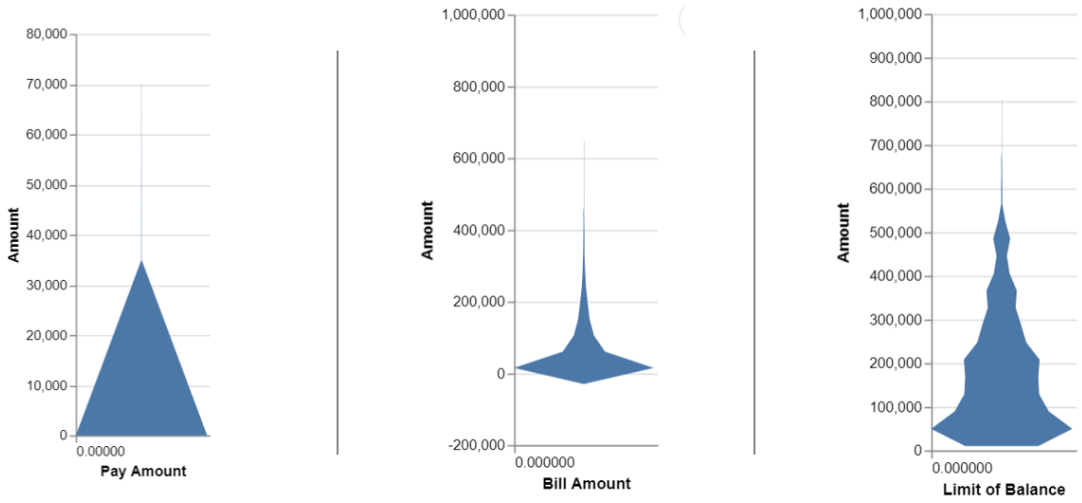

Figure 4: Visualizations of Marriage, Education and Sex

Instead of using Box plot, we use violin plot which combines the box plot with a kernel density plot. What we can see from those plot is payment amount, bill amount and limit of balance are all right skewed.

4.3 Principal Component Analysis

Our goal is to build a classifier for Credit Default, at the time our data contains 6 columns of Payment Amount (backward from September to April) and 6 columns of Bill Amount (backward from September to April). So there might be some redundancy. Thus, we are interested in exploring the relationship between the Payment Amount and Bill Amount and would like to apply Principal Component Analysis to reduce dimension. That is to say, we would like to keep the most relevant variables then build a classifier.

According to the literature review paper by Jolliffe1 and Cadima (2005) [2], PCA can work optimally under the situation that correlations are linear, Figure 3 below shows the correlations between each Payment Amount and Bill Amount. And we could easily found that some Bill Amount are highly correlated, eg. BILL_AMT1 and BILL_AMT2.

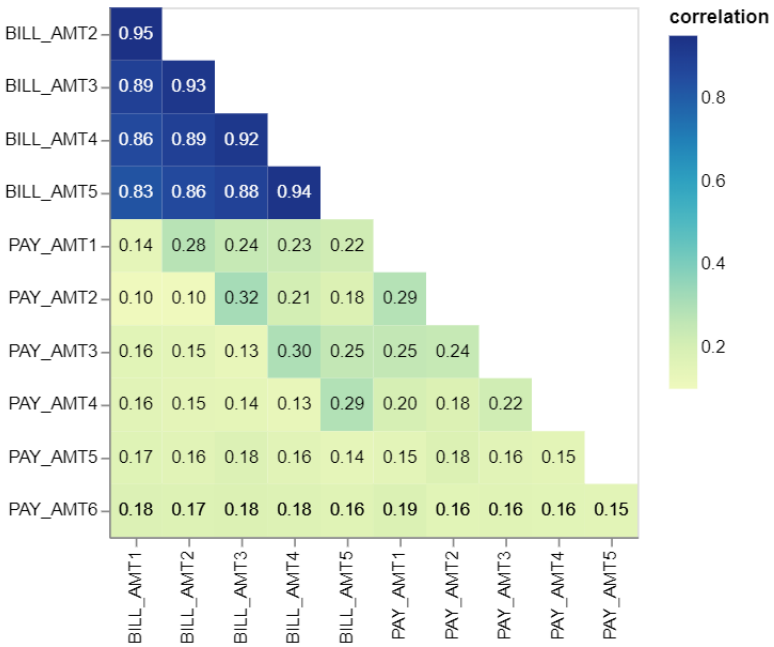

Figure 5: Correlation Matrix

Before doing PCA, we winsorized the outliers to prevent the extreme case. Then the data was split into 80% training data and 20% test data, and those only contain Bill Amount and Payment Amount, a total of 12 variables. Since we already know that the data is imbalanced, so separate minority and majority classes then match numbers in the majority class. So the training data set to be 18661 0's and 18661 1's.

In order to make all the observations on the same scale, we normalized the features. Below is the scree plot:

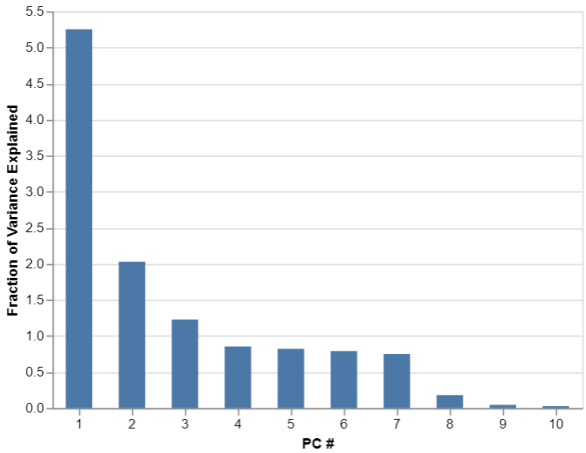

Figure 6: Scree Plot

Since principal component matrix P is simply the original data rotated in space so that it appears axis-aligned. We could just use P=XV, from the plot, there is a cutoff at PC7, so we will use the first 7 PCs to fit the logistic regression and see accuracy. See Figure 5 for visualization of PC1 and PC2. We observed there is a triangular shape.

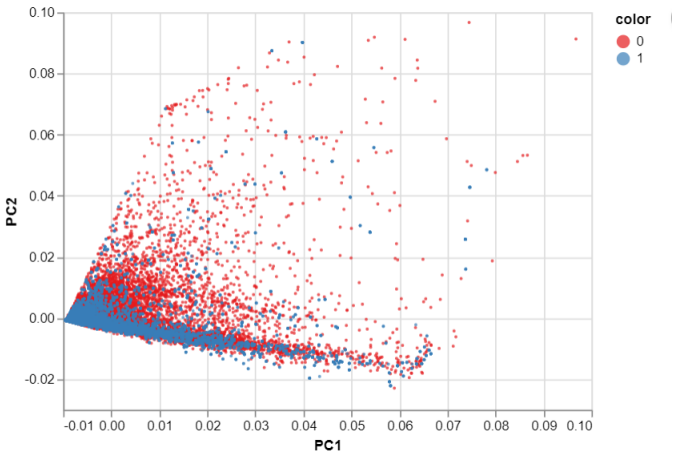

Figure 7: PC1 vs. PC2

The result is optimistic when we fit the model, which gives us about 78% of the accuracy. Hence we conclude that PCA does perform well in this case.

4.4 Feature Engineering

Further data visualizations suggest that there are effects of age and sex towards the limit balance of credit cards. Hence, we will create two new features before we start building our final model: one for effect of age on limit balance and one for effect of sex on limit balance. Details of the analysis can be seen in the Jupyter notebook.

5 Results

5.1 Attempt 1: Simple Logistic Model

After performing EDA, data cleaning and feature engineering, the data was split into 80 % training data and 20 % test data.Since the output is binary, we will use Multiclass Logistic Regression as our training model.

	Number of Clients
Not default	23364
Default	6636

Table 1: Number of clients who default and do not default

Recall the data is imbalance with 23364 clients not defaulting and 6636 clients defaulting. Hence, we will train three different models: Oversampled model, Undersampled model and simple Logistic model. We will use all the original features and newly engineered features while training our models. Our first attempt will be simple logistic model.

	Training	Testing
Accuracy	81.3 %	81.2 %

Table 2: Training and Testing Accuracy for Simple Logistic Regression

We used both test data and train data to perform 5 fold cross validation. By doing so, we obtained our training and testing accuracy. Since our testing accuracy was slightly smaller than training accuracy, we can

conclude the training and test sets were randomly selected. However, accuracy is not all what we should focus on. From the tesing data set, how much false positives (Type I error) and how much false negatives (Type II) error did the model give? So, let's take a look at the confusion matrix below.

	Predict Not Default	Predict Default
Actual Not Default	4579	154
Actual Default	908	359

Table 3: Confusion Matrix of Simple Logistic Model

From Table 3, the model predicted that 908 clients will not default given the clients defaulted and 154 clients will default given they did not default. Hence, we have much larger type II error (False Negatives) than type I error (False Positives). In the next sections, we will discover how we can improve these errors by oversampling and undersampling. We will also discuss about the best models tailored towards particular cooperation needs at the end.

5.2 Attempt 2: Logistic Model with oversampling

Our minority class is the number of clients who don't default. Hence, we will re-sample the minority class with replacement to obtain the same size as our majority class (number of clients who default). Using the re-sampled training data, we build the model again.

	Training	Testing
Accuracy	81.3 %	81.2 %

Table 4: Training and Testing Accuracy for Logistic Regression with over-sampled training data

From our logistic model with over-sampled training data, we obtained the same testing and training accuracy as Simple Logistic Model. Let's look at our confusion matrix.

	Predict Not Default	Predict Default
Actual Not Default	3585	1148
Actual Default	471	796

Table 5: Confusion Matrix of Logistic Model with over-sampled data

Now,the model predicted that 471 clients will not default given the clients defaulted and 1148 clients will default given they did not default. We have much less type II error than type I error in contrast to Simple Logistic Model above.

5.3 Attempt 3: Logistic Model with under-sampling

Since our majority class is the number of clients who default, we will re-sample it without replacement to obtain the same size as the minority class(the number of clients who don't default). We will use this under-sampled data to train our model. Now, the training and testing accuracy are the same as the previous two models. But the confusion matrix is slightly different this time.

	Predict Not Default	Predict Default
Actual Not Default	3463	1270
Actual Default	448	819

Table 6: Confusion Matrix of Logistic Model with under-sampled data

From Table 6,the model predicted that 448 clients will not default given the clients defaulted and 1270 clients will default given they did not default. Hence, we have slightly lower type II error and slightly higher type II error in comparison to the model with over-sampled training data. However, the model has much less type II errors compared to type I errors in contrast to the simple Logistic Model.

5.4 Attempt 4: Logistic Model with under-sampling and feature selection

Minimizing risk is one of the key factors in financial decisions. The logistic model with under-sampled data minimizes type II error(credit default risk) and maximizes type I error. Hence, this is one of the more important models we will need. Using this model, we will perform χ^2 test to eliminate unnecessary features. The χ^2 test from ANOVA table tells us that age, sex and amount of bill from two months or more until default are not predictive features. Hence, we will remove these features and train the model again with under-sampled training data. Our accuracy remained the same but the confusion matrix gives us an interesting result.

	Predict Not Default	Predict Default
Actual Not Default	3911	822
Actual Default	543	724

Table 7: Confusion Matrix of Under-sampled logistic model with feature selection

Unlike the previous three models, the number of type II errors is higher than over-sampled and under-sampled models but still lower than simple Logistic Model and vice versa for type I errors. To have a better understanding of these errors, we will take a look at the precision and recall table in next section.

6 Discussion

6.1 Summary

Now, we will take a look at precision and recall for all models and decide the best model for each different strategy.

TP = True Positive, FP = False Positive(Type I Error) , FN = False Negative(Type II Error)

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

Note that the less type I error(FP) , the better the precision . Similarly, the less type II error(FN) or credit default risk risk, the better the recall. From table 8,under-sampled Logistic Model has highest recall rate and lowest precision rate. Hence, for a financial institution planning to minimize credit default risk, our under-sampled Logistic Model is best fit to their interest.

	Precision	Recall
Oversample	0.3920	0.6211
Undersample	0.3849	0.6249
Simple	0.7244	0.2574
Undersample with feature selection	0.4423	0.567198

Table 8: Precision and Recall of four Logistic models

If an institution is robust to credit default risk and planning to maximize profits from interests, it must maximize the number of credit approvals. In other words, we need a model with highest precision or lowest recall. Hence, simple Logistic Model will be tailored towards its interest. However, notice that recall rate for this model is very low, hence there is now a high risk of default.

Lastly, if a financial institution wants to moderately minimize risk but still want to gain profits from interests by approving higher number of credit applications, we need a model with precision close to recall. In this case, under-sampled model with feature selection will be the best fit.

6.2 Principles of Measurement

Overall, this data is relevant because our goal is to build classifier for default payment, all the variables involves in the data is related with default payment. Since the data was collected in 2005, there's a high chance it may not precisely reflect current credit default behaviour. Moreover, this data set was collected before 2008, sub-prime mortgage financial crisis. In order to obtain the most precise credit default behaviour, we may need to obtain data collected in 2010 or later. The things may distort my data is we do not know how the data collected, we are afford of voluntary response sample, because it gives us bias. The cost of obtaining and creating this data is time, since the payment amount and bill amount are recorded for 6 months.

7 Conclusion

Overall, we were able to extract useful insights by visualizing relationships between predictive features. Also, PCA gave us interesting information regarding strong and weak correlations between bill amounts and payment amounts. Our most efficient model was under sampled model with feature selection. In this model, we were able to reduce from 23 predictors to 14 predictors . Moreover, using only 14 features, we were able to obtain same training and testing accuracy all other models. Hence, we can conclude we succeeded building a fairly efficient classifier. Since χ^2 test of Anova table suggests that age and sex are not useful predictors, there was no bias against male or female or any age range. Moreover, the final model doesn't assume any innate human characteristics as predictive features, so it is fair to assume the model doesn't add any bias. One way to maximize the accuracy is by training deep neural networks. However, this is beyond the scope of this class but we are excited to explore this in further study.

8 Acknowledgements

We would like to give our special thanks Professor Yekaterina Kharitonova and Alex Franks for their tremendous help in this class and to teaching assistants for their support.

References

[1] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[2] I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009.