Loan default prediction using machine learning

Dipanjan Bandyopadhyay

Problem statement:

Loan defaults will cause huge losses for the banks, so they pay good amount of attention to this issue and apply various methods to detect and predict default behaviours of their customers. In this report, I am going to talk about the basic process of loan default prediction with machine learning algorithms.

Introduction:

Data mining includes the various techniques involved in exploring available data and recapitulating it into potentially useful information. Data Mining, in the field of computer science aims at detecting patterns and relationships among voluminous data in huge relational databases. These patterns then in turn act as input data for many machine-learning algorithms. Data mining, classifies, groups, separates the given data. This can be used for predictive analysis with a decision support system. There are numerous data mining algorithms available. These have been wisely used for feature selection, classification, rule framing and clustering.

Use of data mining in banking sector has been on constant rise. Machine learning involves the usage of many classification algorithms, which are used to separate the data into many suggested number of categories.

With constant need for computer usage in banking sector, the increasing number of bank accounts created and credit outflows, computer can play the role of assistive banker in ensuring default less credits and safe banking. Among this credit default has been a major challenge for bankers as it endangers their system and pose chances of losses that might be hard to revert. Credit card defaulters are on the rise too.

This project:

This report is commissioned to examine the payment behavior of credit card clients in Taiwan from April 2005 to September 2005 and make plausible prediction for client's future default status. The report applied logistic regression and random forest model, investigated critical factors and possible methods and determined the probability of certain client will be in default next month. The adapted variable that reflects the previous payment status seems to be the determinant factor.

. There are a number of factors affecting the payment behavior of credit card clients. The dataset we use contains information on demographic factors, credit data, history of payment behavior and bill statements of credit card clients. What we want to examine in this report is the factors that increase and decrease

the probability of paying the debts, thus looking into which group of clients would pay duly and the common features of these clients. The payment from credit cards are the main resource of revenue for the bank, and it is also important for banks to pick up potential qualified candidates to issue credit cards. Based on the dataset and algorithms, it is helpful for banks to obtain stable revenue from clients' payment and do better with interests. Additionally, we'll explore issues such as how the probability of default payment varies among different demographic variables and what are the strongest predictors of default payment.

.

Data

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. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = 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: Amount of bill statement (NT dollar). X12 = 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: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

.

Variables Definition

The data contains basic demographic variables like gender, education, marriage and age. Besides, 6 continuous months of credit card records are also included, details like repayment status, amount of bill statement and previous payment.

The variable *Default Prediction* is a variable from the original dataset; it represents the prediction for the payment status of a certain customer made by the credit card company. In the following analysis, we treat it as the real default status of the customer so that all results of our models are compared and checked with it.

In order to make the model more reasonable, we create some new direct variables based on the basic variables in the dataset. Namely, we take the average of variables including repayment status, bill amount and payment amount. Through these new variables, we can get a more clear and concise

picture of what the situation client is in for credit card usage.

During data cleaning, we find some values of pay status that couldn't be explained, such as 0, -2, and -3, which are not described in the data description. For those observations, we assume values no more than 0 imply either duly payment or payment in advance by several months. Therefore, the average of variables *Pay_1-6* describes the overall payment status of a customer during the last 6 months. For example, if *Ave_Pay_Stat* equals -2, it means that the customer made the payment 2 months in advance for his/her possible bills in future.

Descriptive Statistics

Summary Statistics

Default Prediction	Sex		Education					Marriage		
	1	2	1	2	3	4	5	1	2	3
	Male	Female	Graduate School	University	High School	Others	Unknown	Married	Single	Others
0	8648	13674	8100	10352	3513	109	248	10453	12623	239
1	2615	3446	1789	3116	1131	7	18	3206	3341	84

Figure 1

Figure 1 summarizes the scale of major nominal variables. According to the table, the core credit card users for the company are female since females reach a majority within either default payment group or non-default clients. From the proportion, we roughly know that people who get married are more likely to be in default of payment than singles. And among card holders, the more educated the client is (from high school to graduate school), the less possibility there is that he/she will default on payment.

Follow the default prediction; the population of repayment is about three to four times as many as that of default. In my opinion, the predicted default ratio is a little bit high for a profitable company and its prediction algorithm seems very conservative for the bank to avoid loss of potential default as much as possible.

Data Wrangling:

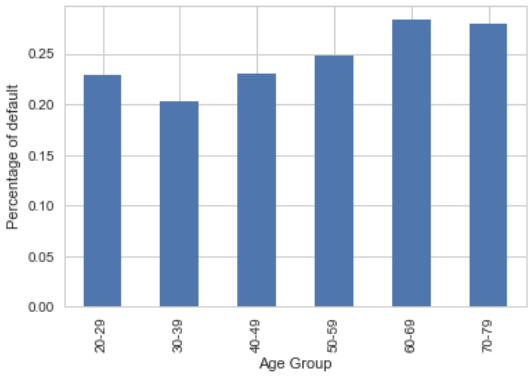
I converted the categorical variables to string. For example, Sex, Education and Marriage were changed. I also created a feature variable called 'balance to limit' which is the ratio of the average six monthly outstanding balance and Credit limit. I created a new field called 'Age Range' which put the continuous

variable age into six buckets or bins.

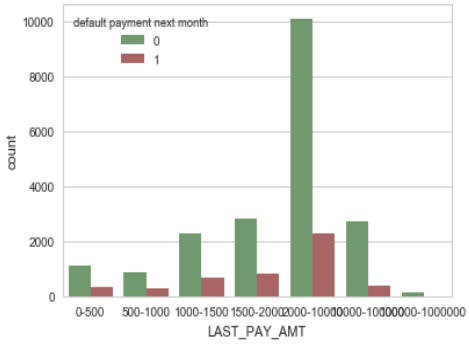
Data Visualization and data story:

The main goal of my investigation is to figure out, which variables among the 23 have higher impact in predicting probability of default next month. I took a series of steps and explored data to figure out the impacts:

1. Do age of the credit card holder have anything to do with default:
    a. The 60-69 age group was found to have higher percentage of defaulters that other age groups.
    b. The next two vulnerable groups are 70-79 and 50-59

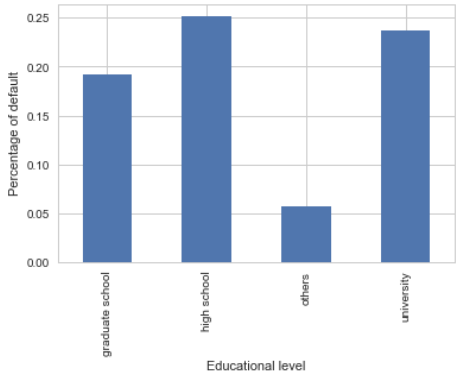


2. Next I went on to check the last bill payment of the credit card holder. Does this amount say anything?
    a. The proportion of default is interestingly higher when the last payment amount is low.
    b. The default percentage keeps on going down as the last payment amount increases.
    c. The default percentage is higher if the last payment amount is below 10,000

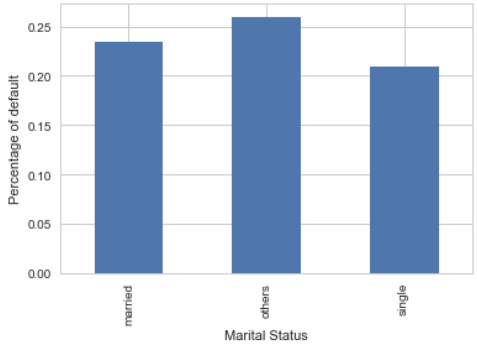


3. Education impacts default:
    a. Those with just high school education tends to default more.
    b. The next two groups are University and Graduate school.

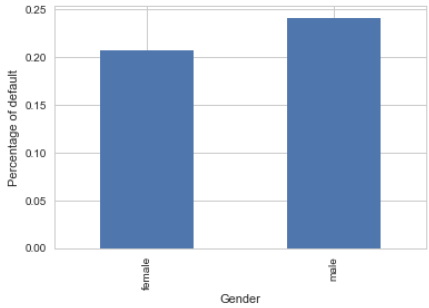
c. The less the educational level, more the default percentage.



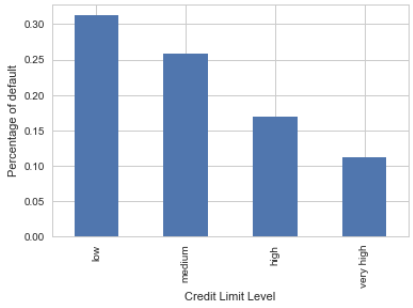
4. Does marital status have an impact on default?
    a. The marital status of 'others' have some interesting behavior of highest rate of default
    b. 'Married' people tend to default more than 'Single' folks.



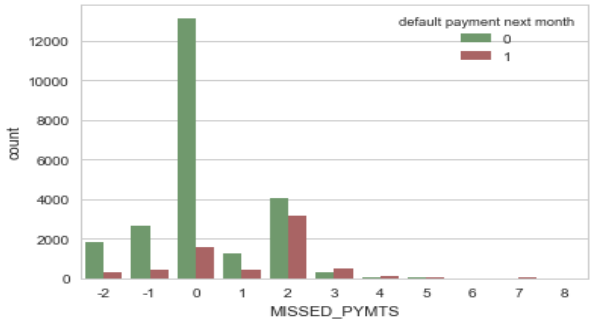
5. What about gender?
    a. Clearly, males tend to default more than females



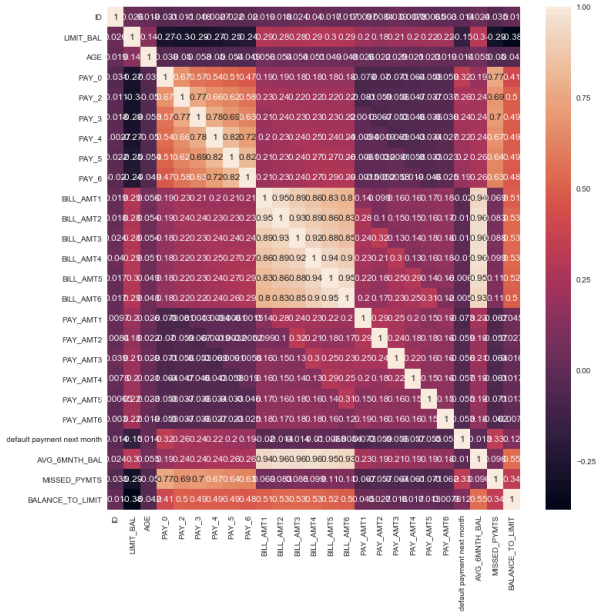
6. I was currious to know whether credit has any relationship with default
    a. The clear finding is: card holders with lower credit credit limits tend to default more



7. Is a large number of missed payments a precursor to default?
    a. It is evident that if there are two or more missed payments, the chance of default increases significantly
    b. But, there is a strange observation. I have observed that there are defaults when there has been zero missed payments.

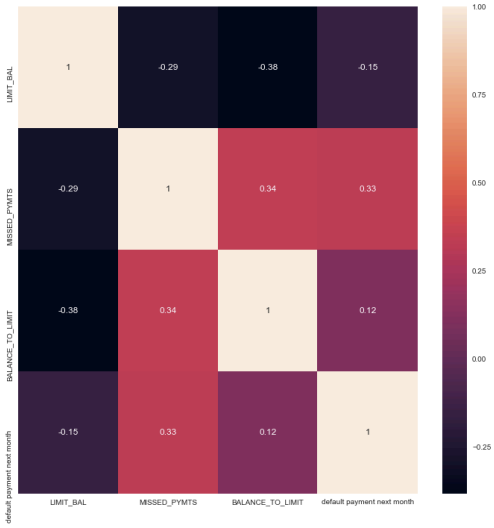


8. I also tried to find out correlations between all variables
    a. As expected, the missed payments and credit balance tend to indicate to some extent, that a default is coming.

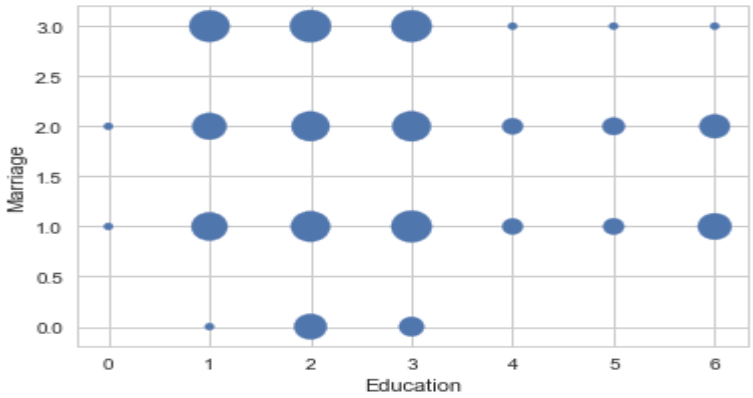


9. I constructed a correlation matrix for limited columns. Missed payments has a

positive correlation with default whereas Credit limit has a negative correlation with default



10. I was curious to see the composite effect of feature variables like Education and Marriage on default. In diagram below, the larger the circle , more is the probability of default.



Education:
1 – Graduate, 2- University, 3- High School, 4- Others , 5,6 – Unknown

Marriage:
1- Married, 2- Single, 3- Others, 0- Unknown

Inferential Statistics:

1. We checked whether there is significant difference between male and female default rates

Null Hypothesis: There is no significant difference between female and male default rate.

H0: μfemales−μmales=0 Significance Level: 95% Confidence

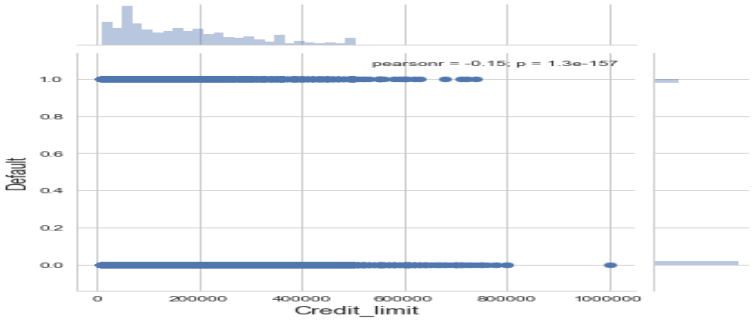
α=0.05

After conducting a two sample T test we concluded that there is significant difference between male and female default rates.

t-statistic: 6.92667700616 p-value: 4.39524880327e-12

P value is too low.

2. We created a joint plot of credit limit and default.



With p value being zero we can conclude that credit limit and default are independent of each other. But. there is little bit of negative correlation between the two. In other words, higher the credit limit, lower is the chance of default.

Use of replicates to determine the relationship between credit limit and default:

I used replicates created through random.choice of Numpy to determine the relationship between credit limit and default rate. The observation is interesting. It is evident that there is correlation between credit limit and default payment next month. If credit limit increases default reduces.

3. We analyzed the default rates at different educational levels.

Graduate sample size: 10585 Graduate sample mean: 0.192347661785 54558
University sample size: 14030 University sample mean: 0.2373485388 4533144

High School sample size: 4917 HighSchool sample mean: 0.25157616 43278422
Others sample size: 123 Others sample mean: 0.056910569105691054

Population size: 30000 Population mean: 0.2212

We tried to check whether there is significant difference between Gradu ate and University education level default rates.
We conducted a two sample t test.

Null Hypothesis: There is no significant difference between University and over dataset default rate.

t-statistic: -2.00728896952 p-value: 0.0447330796615

While we can reject the null hypothesis at 95% confidence level; we will have to accept the null hypothesis at 99% confidence level. So, our conclusion in this case is: the default rates of university educated and high school educated people are almost similar.

Machine learning Algorithms

One goal of this project is predictive analytics. We would like to predict which credit card holder is going to default. We used three different types of models to predict default.

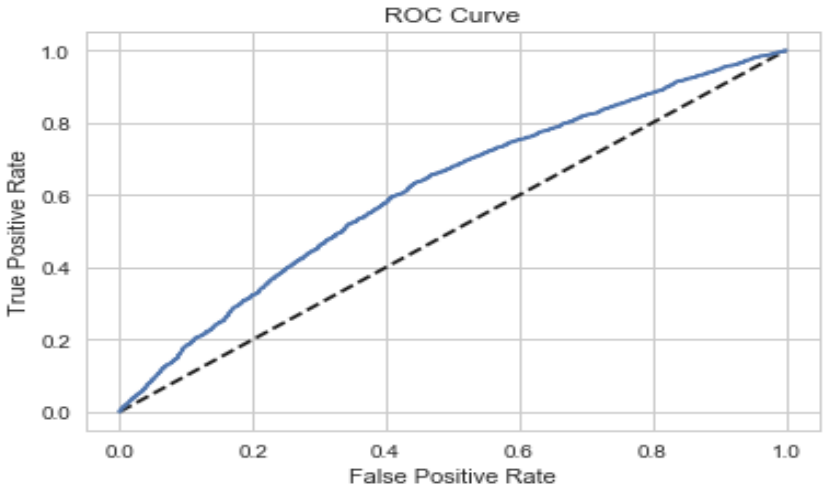
1. Logistic Regression using sklearn and train-test- split
    a. Feature set :
    LIMIT_BAL','SEX','EDUCATION','MARRIAGE','AGE','PAY_0','PAY_2','PAY_3','PAY_4','PAY_5','PAY_6','DUE_1'
    b. Training score: 0.777333333333 Testing score: 0.782222222222
    c. Coefficients [[-6.17570756e-06 -2.48807560e-04 -3.09325710e-04 -2.63706644e-04 -4.36734324e-03 3.28368937e-04 2.63650826e-04 2.32646107e-04 2.14305066e-04 2.02270811e-04 1.94074563e-04 3.56310401e-07]]
    Intercept [-0.00014417]
    d. Confusion matrix was created:

```
[[7040    0]
[1960    0]]
```

    e. Classification matrix was created:

	precision	recall	f1-score	support
0	0.78	1.00	0.88	7040
1	0.00	0.00	0.00	1960
avg / total	0.61	0.78	0.69	9000

f. ROC curve was constructed



2. Logistic Regression after taking action on skewed data of target variable:

   a. Feature set :
   LIMIT_BAL','SEX','EDUCATION','MARRIAGE','AGE','PAY_0','PAY_2','PAY_3','PAY_4','PAY_5','PAY_6','DUE_1'
   b. Training score: 0.692380952381 Testing score: 0.686333333333

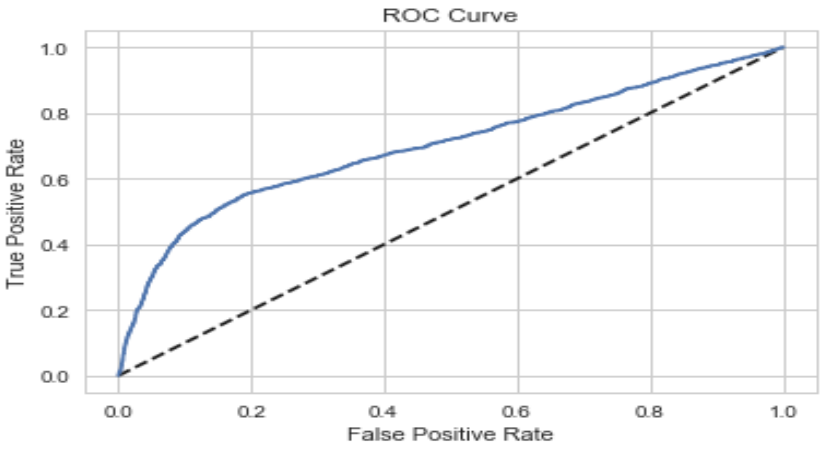
   c. Coefficients [[-1.01703445e-06 -2.17462378e-02 -2.82544400e-02 -2.48313532e-02 9.57661720e-03 3.04473624e-01 1.64981999e-01 1.09890795e-01 7.33049433e-02 5.19952471e-02 4.71070350e-02 -2.37348823e-06]]
   Intercept [-0.0015479]
   d. Confusion matrix was created:
   [[4991 2049]
   [774 1186]]

   e. Classification matrix was created:

	precision	recall	f1-score	support
0	0.87	0.71	0.78	7040
1	0.37	0.61	0.46	1960
avg / total	0.76	0.69	0.71	9000

   f. ROC curve was constructed:

The ROC curve actually improved , but accuracy score didn't improve.



g.

3. Apply KNN:
   K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNNhas been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique. Algorithm. We applied KNN with train test split. The following was the results:

   Training score: 0.805904761905 Testing score: 0.739888888889

4. Apply Random Forest:
   Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because it's simplicity and the fact that it can be used for both classification and regression tasks.

   Training score: 0.978619047619 Testing score: 0.799333333333

5. Apply Random Forest after taking action on skewed data for target variable
   Training score: `0.977857142857` `Testing` score: `0.802444444444`

Decision Trees tend to overfit the models and that is quite evident from the tarining score of both RF models of 0.98
However, after taking action on skewed data of target variable through the model, the test score improved.
The RF model which incorporates skewed target variable correction happens to be the best model for this problem.

Conclusion:

While there are several factors which contribute towards default of a credit card holder, with this report we are able to prove that we can predict the default with reasonable accuracy.
Our statistical analysis reveals that with a focussed approach on choosing the right kind of customer demographics, and credit limits a bank or financial institution can reduce the default rate.