Proposal

Lin Liuyu 1701213060

1. Project descrition

   Credit card default is commonly problem and how to it will influenced the banks' Rate of Bad Account, so it is important for the bank to make a good decision whether or not continue to provide a credit line to a person. This project is to analyze the dataset which contains some information of the historical behavior of people and try to get out a prediction model to figure out whether or not one person will default next month.

2. Project process

   a) Data analysis: Do some statistical analysis on the data as well as visualize the data to get the intuitively idea that which feature may be useful to clarify these two kinds of people. For example, we can see the distribution of different features in these two kinds of people.

   b) Data processing: Deal with the NA data. Consider about delete some data that is meaningless and only a small fraction of the whole data. For example, in feature 'education', the data of type 4, 5, 6 (which mean 'others' or 'unknown') is only 1.5% of the whole data, which do not means the higher level of education and will lead to bias

when predict by this feature and we should drop it out without too much information loss.

c) Feature construction: Try to build some new features by the raw data to describe the behavior of this person more specific.

d) Spilt the data into training data and test data randomly.

e) Use the in-sample data to train and validate the model, I may use perception, logistic regression, SVM, decision tree and combine different models for ensemble learning to improve the performance.

f) Try to use PCA or LDA and redo step e).

g) Do some visualization for the outcome.

h) Evaluate the clarification model.

3. Dataset description

Dataset source: https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset

Dataset Information: This dataset 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.

Data Variables:

User information

ID	ID of each client
LIMIT_BAL	Amount of given credit in NT dollars (includes individual and family/supplementary credit)
SEX	Gender (1=male, 2=female)
EDUCATION	(1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
MARRIAGE	Marital status (1=married, 2=single, 3=others)
AGE	Age in years

User beahavior

PAY_0	Repayment status in September, 2005 (-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 (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 (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 or not.next month

default.payment.next.month	Default payment (1=yes, 0=no)