Capstone Project

Bayesian Statistics: Techniques and models, Coursera

Rasmus Nyberg

August, 2020

Introduction

The goal of this project is to analyze credit risk and correlation between credit default and customer specfic information such as age and marital status. A bayesian logistic regression model will be estimated and be used to answer two questions.

1) What performace can we get in terms of ROC AUC
2) Does a women have a lower probability of default compared to a man given some input

Data

The 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 was downloaded from Kaggle [1]. This project is supposed to take about 10 hours, therefore some modifications and simplifications was done to the original data before any analysis was done. These actions might not be appropriate if the goal is to estimate the best predictive model.

1) Payment history variables replaced with max MAX_PAY_HIST (max of PAY_0 - PAY_6 > 0)
2) Variable EDUCATION replace by dummy D_EDU (1 if Graduate school or university, 0 otherwise)
3) Variable MARRIAGE replace by dummy D_MAR (1 if merried, 0 otherwise)
4) Bill statement variables (PAY_AMT1 - PAY_AMT6) were dropped
5) Variable default.payment.next.month renamed to DEFAULT

Table 1: Modified dataset

ID	LIMIT	AGE	SEX	D_MAR	D_EDU	MAX_PAY_HIST	DEFAULT
1	20000	24	2	1	1	2	1
2	120000	26	2	0	1	2	1
3	90000	34	2	0	1	0	0
4	50000	37	2	1	1	0	0
5	50000	57	1	1	1	0	0
6	50000	37	1	0	1	0	0

Summary statistics are shown below. We can see that there are 30k observations of which 22% have defaulted during the next month. The default rate, meaning that more than 1/5 will default during the next month, is surprisingly high.

Table 2: Summary statistics

ID	DEFAULT
Min. : 1	Min. :0.0000
1st Qu.: 7501	1st Qu.:0.0000

[1] https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset

ID	DEFAULT
Median :15000	Median :0.0000
Mean :15000	Mean :0.2212
3rd Qu.:22500	3rd Qu.:0.0000
Max. :30000	Max. :1.0000

Correlation matrix is shown below. We can see that the strongest correlation with DEFAULT is MAX_PAY_HIST which is expected because defaulted customers tend to have a history of late payments before the default event occurs. LIMIT have a negative correlation meaning that the probability of default decreases with higher limits. This could also make sense beacuse higher limits generally require more in terms of the customers payment history, salary etc. SEX also have a negative correlation meaning that the probabilty of DEFAULT decreases if the customer is a women. AGE and D_MAR have a small positive correlation indicating that an old or merried customer is associated with higher probability of default which is surprising. D_EDU have a negative correlation indicating that customers with a good education have a lower probabilty of default.

Table 3: Spearman correlation matrix

	LIMIT	AGE	SEX	D_MAR	D_EDU	MAX_PAY_HIST	DEFAULT
LIMIT	1.000	0.186	0.057	0.109	0.142	-0.319	-0.170
AGE	0.186	1.000	-0.092	0.480	-0.205	-0.074	0.005
SEX	0.057	-0.092	1.000	0.031	0.004	-0.051	-0.040
D_MAR	0.109	0.480	0.031	1.000	-0.119	-0.030	0.029
D_EDU	0.142	-0.205	0.004	-0.119	1.000	-0.043	-0.017
MAX_PAY_HIST	-0.319	-0.074	-0.051	-0.030	-0.043	1.000	0.321
DEFAULT	-0.170	0.005	-0.040	0.029	-0.017	0.321	1.000

To speed up the modelling process the data was aggregated by all independent variables. Also variable LIMIT was dropped and AGE was rounded to nearest tenth. These actions might be inappropriate if the goal is to estimate the best predictive model. Other considerations might be to try some other transformations of the variables, for example grouping of the variables MAX_PAY_HIST and AGE.

Model

The dependent variable is binary and therefore we will fit a logistic regression model using log likelihood. The configuration setup includes normal priors on the coefficients (mean 0 and variance 100), 10000 burn-in iterations and 100000 total iterations for the 3 chains.

From the modeling diagnostics (see Appendix A-C) we can observe autocorrelation and that all parameters have not converged. This indicates that we might have to run the model for more iterations and, eventually, change the model specification.

The coefficients from the estimation is shown below. We can see that two variables (AGE and D_EDU) are not statistically significant (Mean +/- 2*SD surrounds 0) and should be removed from the model.

Table 4: Coefficients (means), bayesian model

	Mean	SD	Naive SE	Time-series SE
B[1]	0.0026654	0.0016806	0.0000031	0.0000223
B[2]	-0.1189807	0.0305121	0.0000557	0.0003185
B[3]	0.1710364	0.0331975	0.0000606	0.0001833

	Mean	SD	Naive SE	Time-series SE
B[4]	0.0253153	0.0392443	0.0000716	0.0003327
B[5]	0.6226718	0.0118738	0.0000217	0.0000439
B0	-1.7013066	0.0911506	0.0001664	0.0015647

A standard logistic model is estimated as a baseline for comparison and quality check. The coefficients are very close to what we have observed from the bayesian model.

Table 5: Coefficients baseline model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.7038754	0.0902243	-18.8848847	0.0000000
AGE	0.0027217	0.0016735	1.6263839	0.1038680
SEX	-0.1183641	0.0302968	-3.9068258	0.0000935
D_MAR	0.1705662	0.0330512	5.1606640	0.0000002
D_EDU	0.0255290	0.0390676	0.6534574	0.5134614
MAX_PAY_HIST	0.6225534	0.0118773	52.4153746	0.0000000

Results

The performance of the model in terms of ROC AUC is acceptable. It should be possible to improve the performance by doing o more serious data analysis, better variable transformations and removing inappropriate variables.

Table 6: ROC AUC

x
0.7156763

Given the posterior predictive distribution we can estimate the probability that a woman compared to a man have a lower probability of default given some input. Given 40 years old, married, good education and good history of payments - the probability a woman compared to a man have a lower probability of default is almost 100%.

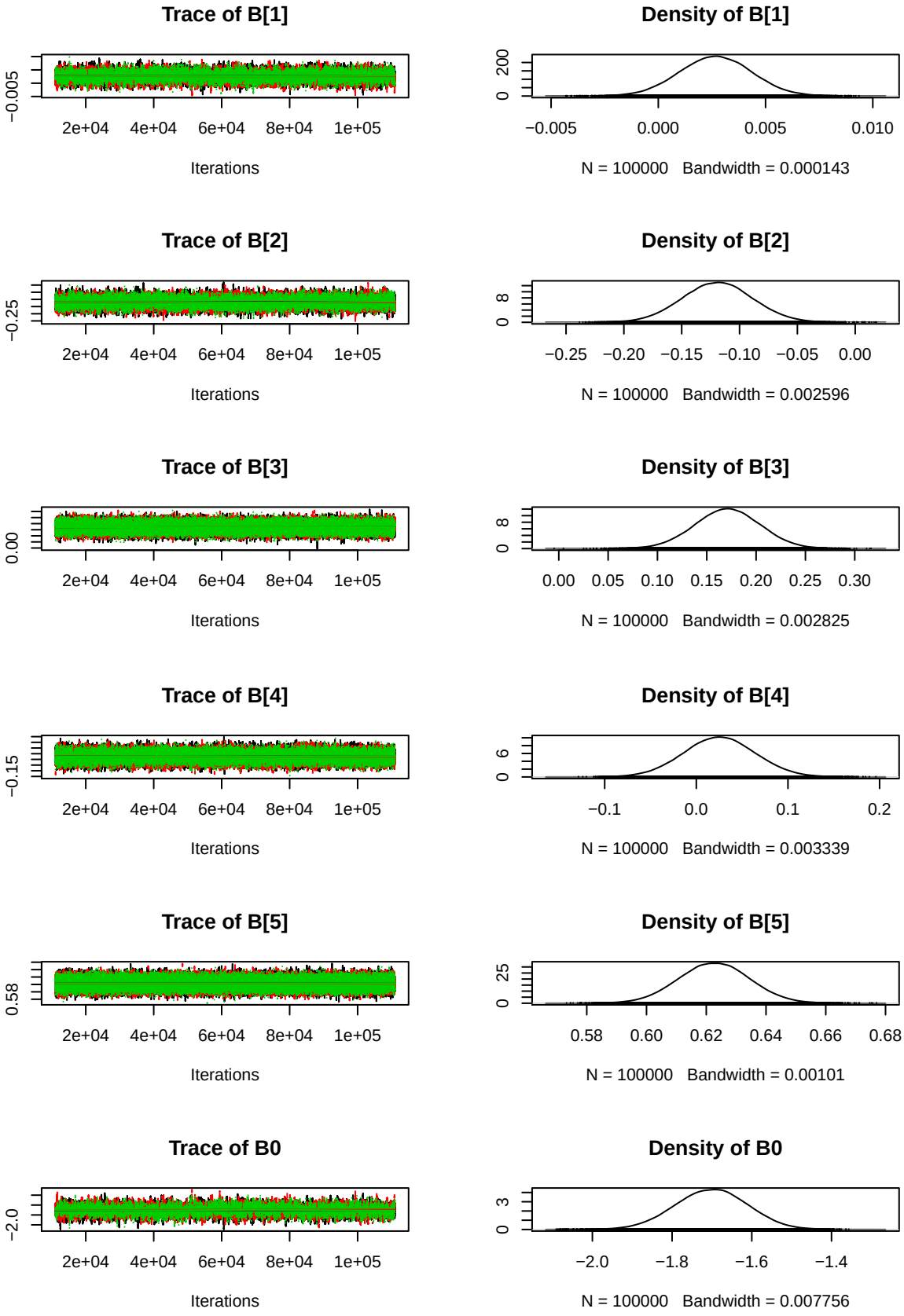
Table 7: Probability

x
0.9999133

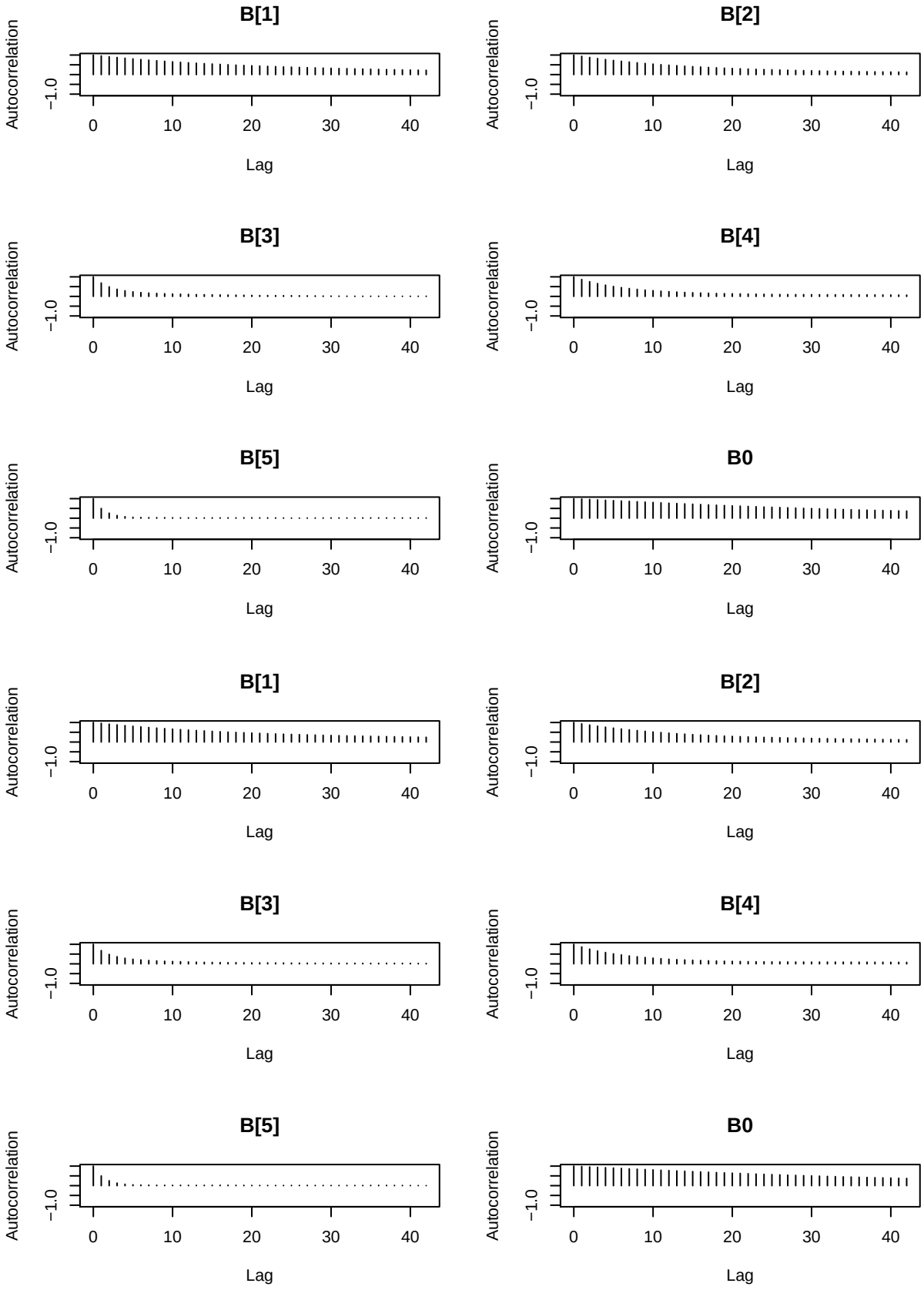
Conclusions

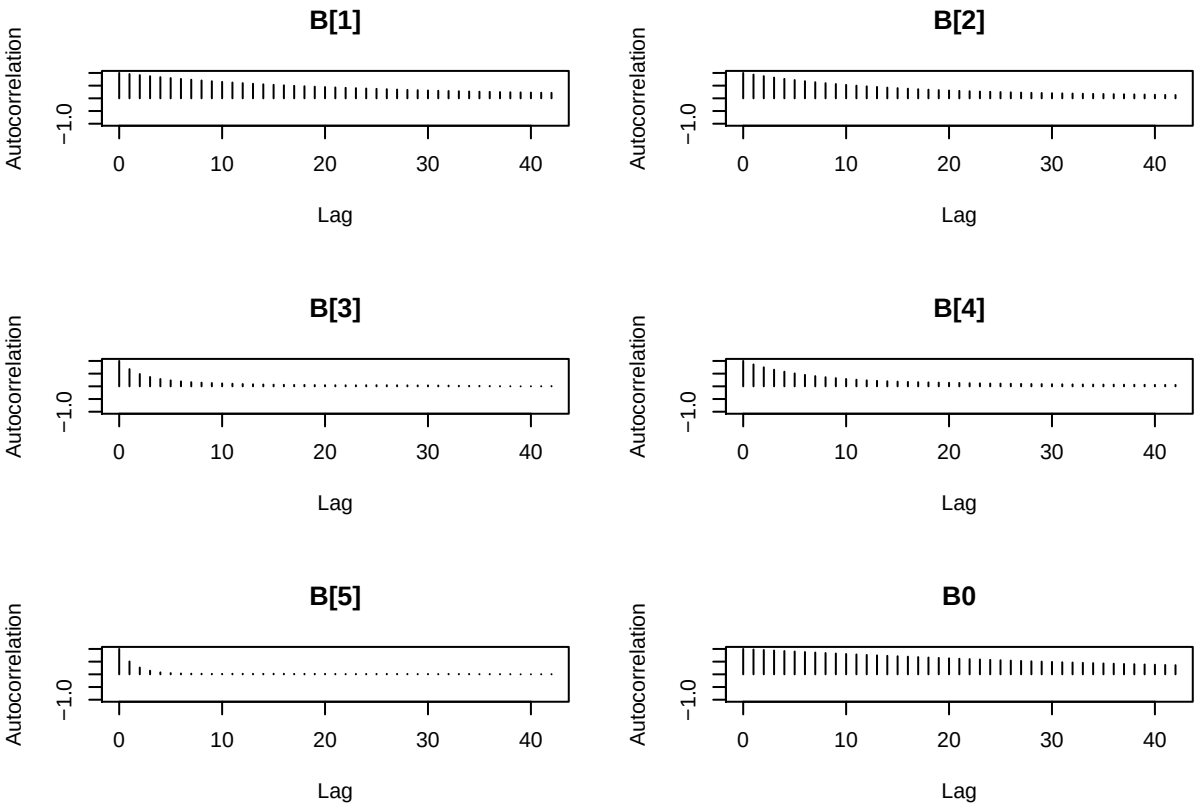
It is possible to estimate a relatively good predictive model that can be used for inference. At the same time it's important to understand that the model in this analysis need to be improved based on a more serious data analysis and that the dataset might not represent a random credit customer. The analysis has been done with a limited time frame and conclusions should be drawn with a lot of caution.

Appendix A: Trace plots



Appendix B: Autocorrelation plots




B[1]
B[2]
B[3]
B[4]
B[5]
B0
Autocorrelation
−1.0
0
10
20
30
40
Lag

Appendix C: Tables

Table 8: Raftery And Lewis's Diagnostic, MC 1

	M	N	Nmin	I
B[1]	54	64557	3746	17.20
B[2]	48	56752	3746	15.20
B[3]	12	17952	3746	4.79
B[4]	24	33064	3746	8.83
B[5]	8	9482	3746	2.53
B0	104	114075	3746	30.50

Table 9: Raftery And Lewis's Diagnostic, MC 2

	M	N	Nmin	I
B[1]	60	69520	3746	18.60
B[2]	45	55800	3746	14.90
B[3]	12	17872	3746	4.77
B[4]	24	25052	3746	6.69
B[5]	6	6755	3746	1.80
B0	96	99696	3746	26.60

Table 10: Raftery And Lewis's Diagnostic, MC 3

	M	N	Nmin	I
B[1]	60	65350	3746	17.40
B[2]	35	43610	3746	11.60
B[3]	12	17732	3746	4.73
B[4]	20	26240	3746	7.00
B[5]	8	9604	3746	2.56
B0	96	113776	3746	30.40

Table 11: Markov Chain Autocorrelation

	B[1]	B[2]	B[3]	B[4]	B[5]	B0
Lag 0	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000
Lag 1	0.9563440	0.9336445	0.6794220	0.8645659	0.5001274	0.9795401
Lag 5	0.8034313	0.7191199	0.2356132	0.5097727	0.0460450	0.8992930
Lag 10	0.6547082	0.5280597	0.1190076	0.2919425	0.0195175	0.8046680
Lag 50	0.1874898	0.0939201	0.0058904	0.0579661	0.0089490	0.3009236

Table 12: Gelman And Rubin's Convergence Diagnostic

	Point est.	Upper C.I.
B[1]	1.001758	1.006250
B[2]	1.001772	1.006438

	Point est.	Upper C.I.
B[3]	1.000066	1.000210
B[4]	1.001465	1.005378
B[5]	1.000207	1.000779
B0	1.003885	1.014120

Table 13: Effective Sample Size

	x
B[1]	5704.430
B[2]	9207.999
B[3]	32805.843
B[4]	14012.557
B[5]	73563.340
B0	3393.215

Appendix D: Original dataset

Table 14: Original dataset columns

Column	Description
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
PAY_0	Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, ..., 8=eight months, 9=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.payment.next	Default payment next month (1=yes, 0=no)