Course Capstone Project

Bayesian Statistics: Techniques and models, Coursera

Rasmus Nyberg

August, 2020

Abstract

This is the abstract.

Introduction

This is the introduction.

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 was done to the original data before any analysis was done. This might be a bad choice if one were to build 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 otherwisae)
3) Variable MARRIAGE replace by dummy D_MAR (1 if merried, 0 otherwisae)
4) Bill statement variables (PAY_AMT1 - PAY_AMT6) were dropped
5) Variable*default.payment.next.month renamed to DEFAULT
6) Variable LIMIT_BAL 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 is surprisingly high.

Table 2: Summary statistics

ID	DEFAULT
Min. : 1	Min. :0.0000
1st Qu.: 7501	1st Qu.:0.0000
Median :15000	Median :0.0000

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

ID	DEFAULT
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 often have a history of late payments. LIMIT_BAL have a negative correlation meaning that the probability of DEFAULT decreases with higher limit balance. This could also make sense beacuse higher limits generally would 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. The same is true for MARRIAGE meaning that

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

Model

This is the model.

Results

This is the results.

Conclusions

This is the conslusions.

Original dataset

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)