**Data Description**

There are 25 variables:

* **ID**: ID of each client
* **LIMIT\_BAL**: Amount of given credit in New Taiwan 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 (-2=No consumption,-1=pay duly,0=The use of Revolving card, 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)

*Based on how it works in real-time*

* *-2 = Balance paid in full and no transactions this period (we may refer to this credit card account as having been 'inactive' this period)*
* *-1 = Balance paid in full, but account has a positive balance at end of period due to recent transactions for which payment has not yet come due*
* *0 = Customer paid the minimum due amount, but not the entire balance. I.e., the customer paid enough for their account to remain in good standing, but did revolve a balance*
* **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.month**: Default payment (1=yes, 0=no)

List of Categorical variables:

Sex

Education

Marriage

PAY\_0, PAY\_2, PAY\_3, PAY\_4, PAY\_5, PAY\_6

Default payment next month – target variable

* PAY\_0=0&&PAY\_2=3 means the bill was paid

Some of the issues to be dealt with:

* Where did the default payment next month column even come from?
* Duplicate rows
* Can I use one separate column to list the difference between all the payments and bill amounts and use it for analysis? – total amount including interest

Because only based on that value default is calculated in the dataset

* PAY\_0,1… have -2 and 0 as the values, what do they mean?
* ID 27 PAY\_0=1 and rest all are negative meaning payment delay for one month and rest all are negative, bill amount=-394 and payment amt = 2500, bill negative meaning? Why did he even pay and why did it default?
* ID 46, PAY\_0…6 all -2, bill\_amt=0, payment amt=0 yet defaulted
* ID 32, PAY\_0=2 and PAY\_1…6 are all 0’s which is illogical because when PAY\_0=2 mean 2 months delay in payment so logically PAY\_1 should be 1, similarly with ID 90

*Negative bill amount meaning*

*It is generally possible for a credit card customer to overpay their bill and temporarily carry a negative balance. E.g., say my bill this month is $100 but I pay $250. Assuming I have no other recent purchases, my balance will be -$150.*

*As for why a customer might do this, maybe their autopay is set up to pay the same amount every month regardless of balance, maybe they forgot if they paid their bill one month and accidentally paid twice, or maybe they just like to pay down some future expected purchases in advance. Payment behavior like this may not always appear rational, but it is not uncommon.*

Drop all the PAY columns because they do not make sense

Drop the rows which have all the bill amts 0 and all the payments 0’s

Drop the rows which have bill amts <-100

Avg of all the columns to see which factor is affecting the default payment method more

Suppose males having less education are defaulting

Payment has much more effect on the default rather than bill

Check to see which is affecting the default more

Whether the payments or bill or credit limit or (pay-bill) etc

Bill amt1-pay\_amt1 etc

Run a linear model/logistic model to find put which is affecting the prob of default