Introduction to Business Analytics

PAN Baoqian, Kris

ISOM 5610

Topic 5b: The Credit Card Default Case

Case study – Default of credit card clients

- Background: 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
- The loan officer can deny the loan request from potentially bad clients, and offer the loan to potentially good clients

* From https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset

- Goal: predict whether credit card clients will have default payment or not

- Method: build a logistic regression model

Description of Data File

- 30,000 observations
- 23 explanatory variables and 1 binary response (default payment, Yes=1, No=0)

Variable	Description
LIMIT_BA	Amount of the given credit (NT dollar)
SEX	1 = male, 2 = female
EDUCATION	Uni=University, GS=Graduate Study ,HS=High School ,others=other/unknown
MARRIAGE	1 = married; 2 = single; 3 = divorce; 0=others/unknown
AGE	Age in years

Variable	Description
PAY_1	Repayment status in September 2005
PAY_2	Repayment status in August 2005
PAY_3	Repayment status in July 2005
PAY_4	Repayment status in June 2005
PAY_5	Repayment status in May 2005
PAY_6	Repayment status in April 2005

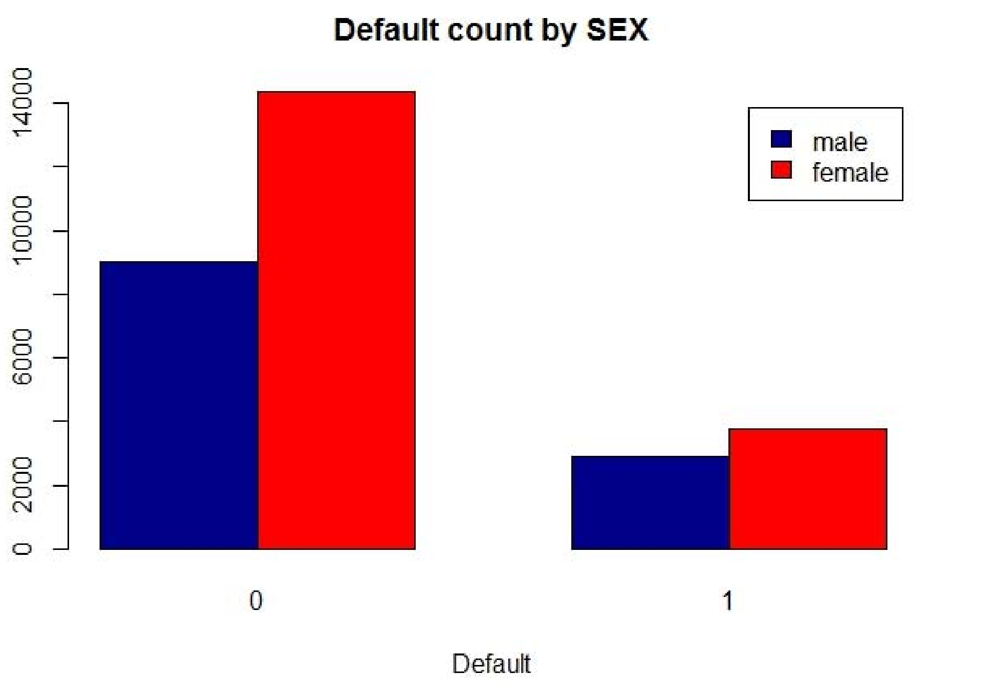
Category of above variables
No_consumption
Paid_in_full
Payment_delay
Revloving_credit

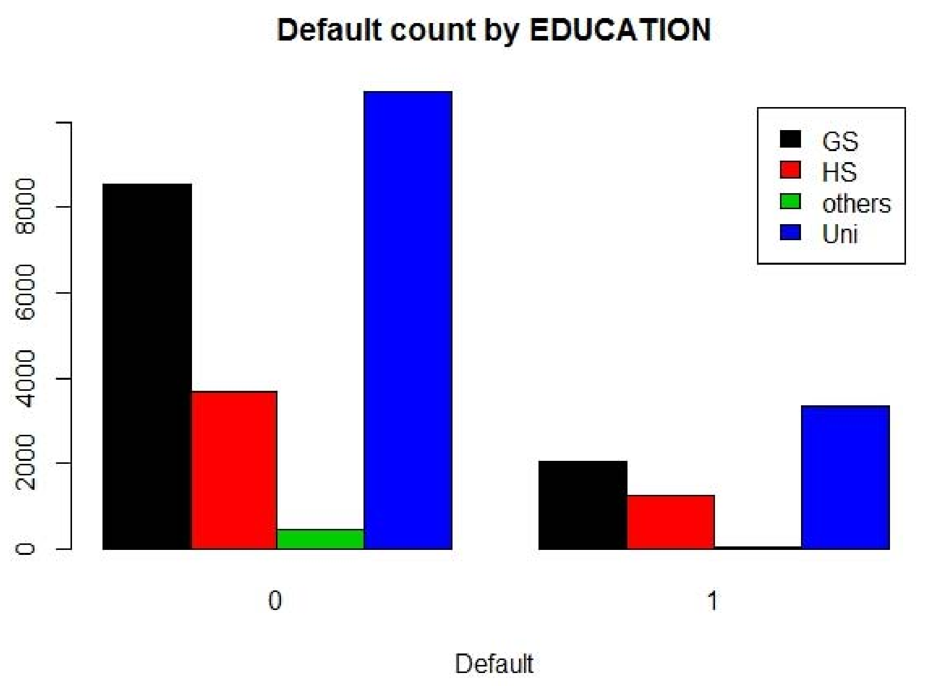
Variable	Description
BILL_AMT1	Amount of bill statement in September 2005
BILL_AMT2	Amount of bill statement in August 2005
BILL_AMT3	Amount of bill statement in July 2005
BILL_AMT4	Amount of bill statement in June 2005
BILL_AMT5	Amount of bill statement in May 2005
BILL_AMT6	Amount of bill statement in April 2005
PAY_AMT1	Amount of previous payment in September 2005
PAY_AMT2	Amount of previous payment in August 2005
PAY_AMT3	Amount of previous payment in July 2005
PAY_AMT4	Amount of previous payment in June 2005
PAY_AMT5	Amount of previous payment in May 2005
PAY_AMT6	Amount of previous payment in April 2005

Descriptive analysis

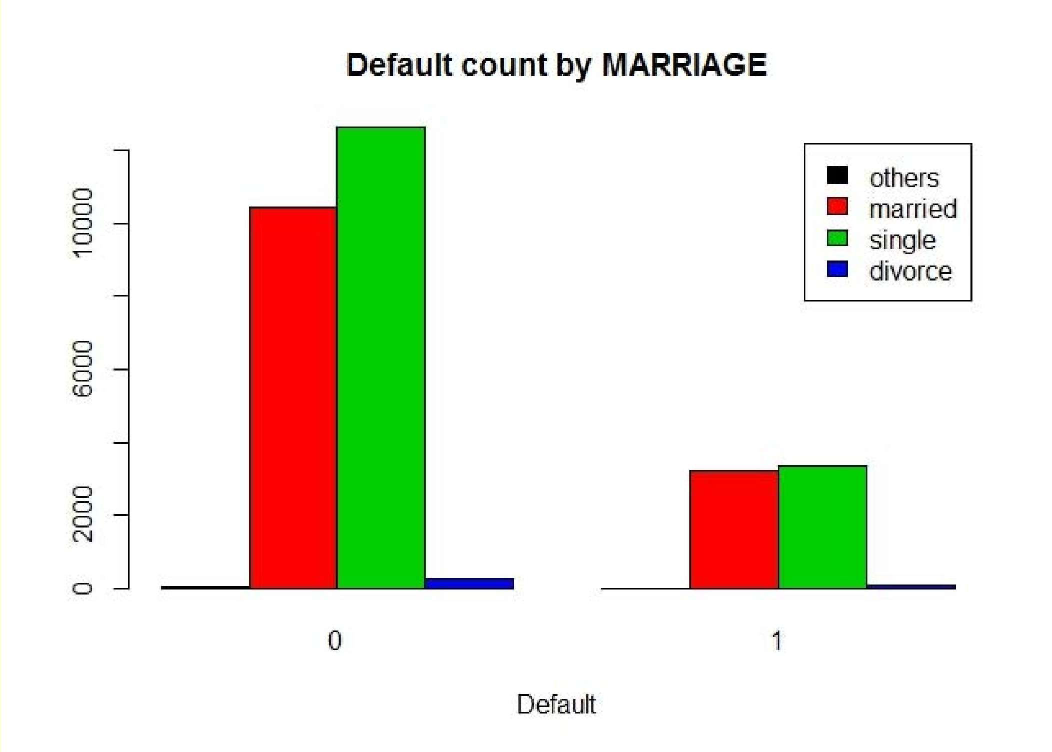
```
   LIMIT_BA          SEX        EDUCATION      MARRIAGE        AGE                        PAY_1                           PAY_2
 Min.    :  10000   1:11888   GS     :10585   0:    54   Min.   :21.00   no_consumption  : 2759   no_consumption  : 3782
 1st Qu.:  50000   2:18112   HS     : 4917   1:13659   1st Qu.:28.00   paid_in_full    : 5686   paid_in_full    : 6050
 Median : 140000             others:  468   2:15964   Median :34.00   payment_delay   : 6818   payment_delay   : 4438
 Mean   : 167484             Uni    :14030   3:   323   Mean   :35.49   revolving_credit:14737   revolving_credit:15730
 3rd Qu.: 240000                                         3rd Qu.:41.00
 Max.   :1000000                                         Max.   :79.00
             PAY_3                        PAY_4                        PAY_5                        PAY_6             BILL_AMT1          BILL_AMT2
 no_consumption  : 4085   no_consumption  : 4348   no_consumption  : 4546   no_consumption  : 4895   Min.   :-165580   Min.   :-69777
 paid_in_full    : 5938   paid_in_full    : 5687   paid_in_full    : 5539   paid_in_full    : 5740   1st Qu.:   3559   1st Qu.:  2985
 payment_delay   : 4213   payment_delay   : 3510   payment_delay   : 2968   payment_delay   : 3079   Median :  22382   Median : 21200
 revolving_credit:15764   revolving_credit:16455   revolving_credit:16947   revolving_credit:16286   Mean   :  51223   Mean   : 49179
                                                                                                     3rd Qu.:  67091   3rd Qu.: 64006
                                                                                                     Max.   : 964511   Max.   :983931
   BILL_AMT3          BILL_AMT4          BILL_AMT5         BILL_AMT6          PAY_AMT1          PAY_AMT2          PAY_AMT3
 Min.   :-157264   Min.   :-170000   Min.   :-81334   Min.   :-339603   Min.   :      0   Min.   :       0   Min.   :      0
 1st Qu.:   2666   1st Qu.:   2327   1st Qu.:  1763   1st Qu.:   1256   1st Qu.:   1000   1st Qu.:     833   1st Qu.:    390
 Median :  20089   Median :  19052   Median : 18105   Median :  17071   Median :   2100   Median :    2009   Median :   1800
 Mean   :  47013   Mean   :  43263   Mean   : 40311   Mean   :  38872   Mean   :   5664   Mean   :    5921   Mean   :   5226
 3rd Qu.:  60165   3rd Qu.:  54506   3rd Qu.: 50191   3rd Qu.:  49198   3rd Qu.:   5006   3rd Qu.:    5000   3rd Qu.:   4505
 Max.   :1664089   Max.   : 891586   Max.   :927171   Max.   : 961664   Max.   :873552   Max.   :1684259   Max.   :896040
    PAY_AMT4          PAY_AMT5             PAY_AMT6              default
 Min.   :     0   Min.   :     0.0   Min.   :     0.0   Min.   :0.0000
 1st Qu.:   296   1st Qu.:   252.5   1st Qu.:   117.8   1st Qu.:0.0000
 Median :  1500   Median :  1500.0   Median :  1500.0   Median :0.0000
 Mean   :  4826   Mean   :  4799.4   Mean   :  5215.5   Mean   :0.2212
 3rd Qu.:  4013   3rd Qu.:  4031.5   3rd Qu.:  4000.0   3rd Qu.:0.0000
 Max.   :621000   Max.   :426529.0   Max.   :528666.0   Max.   :1.0000
```

Descriptive analysis





Descriptive analysis



- Proportion of non-default and default case:

```
> prop.table(table(factor(credit$default)))

     0      1
0.7788 0.2212
```

  - Very unbalanced dataset

Logistic model

- $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 \ldots + \beta_k x_k$
- β_0: constant
- $\beta_1 \ldots \beta_k$: coefficients of predictors
- k: number of predictors
- p: probability of the event to happen i.e. P(Y=1)

Data Partition

- There are 30,000 observations
- Partition ~70% and ~30% of the data into training and testing set
- We simply put the first 21,000 (~70%) to be the training set
- And the reminding 9,000 (~30%) to be the testing set

The Full Model

- ***glm()*** is used to fit a logistics regression
  - By setting ***family=binomial***

```
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)           -4.185e+00  1.054e+00  -3.971 7.16e-05 ***
LIMIT_BA              -1.768e-06  2.051e-07  -8.619  < 2e-16 ***
SEX2                  -1.232e-01  3.762e-02  -3.274  0.00106 ** 
EDUCATIONHS           -4.296e-02  5.876e-02  -0.731  0.46472    
EDUCATIONothers       -1.359e+00  2.766e-01  -4.913 8.98e-07 ***
EDUCATIONUni          -1.781e-02  4.339e-02  -0.411  0.68141    
MARRIAGE1              2.989e+00  1.047e+00   2.855  0.00431 ** 
MARRIAGE2              2.808e+00  1.047e+00   2.681  0.00733 ** 
MARRIAGE3              2.947e+00  1.060e+00   2.781  0.00542 ** 
AGE                    2.024e-03  2.297e-03   0.881  0.37818    
PAY_1paid_in_full      1.314e-01  1.214e-01   1.082  0.27905    
PAY_1payment_delay     9.547e-01  1.007e-01   9.485  < 2e-16 ***
PAY_1revolving_credit -1.090e+00  1.242e-01  -8.775  < 2e-16 ***
PAY_2paid_in_full      2.415e-01  1.248e-01   1.935  0.05296 .  
PAY_2payment_delay     4.424e-01  1.304e-01   3.394  0.00069 ***
PAY_2revolving_credit  9.660e-01  1.458e-01   6.627 3.42e-11 ***
PAY_3paid_in_full     -1.162e-01  1.247e-01  -0.932  0.35132    
PAY_3payment_delay     3.221e-01  1.470e-01   2.192  0.02839 *  
PAY_3revolving_credit -1.095e-01  1.450e-01  -0.756  0.44979    
PAY_4paid_in_full      2.655e-03  1.275e-01   0.021  0.98339    
PAY_4payment_delay     2.435e-01  1.542e-01   1.579  0.11431    
PAY_4revolving_credit  3.851e-02  1.435e-01   0.268  0.78838    
PAY_5paid_in_full     -7.382e-02  1.241e-01  -0.595  0.55204    
PAY_5payment_delay     3.903e-01  1.536e-01   2.542  0.01104 *  
PAY_5revolving_credit  4.951e-02  1.379e-01   0.359  0.71969    
PAY_6paid_in_full     -1.193e-01  9.420e-02  -1.267  0.20532    
PAY_6payment_delay     5.187e-02  1.176e-01   0.441  0.65920    
PAY_6revolving_credit -2.979e-01  1.025e-01  -2.906  0.00366 ** 
BILL_AMT1             -2.059e-06  1.340e-06  -1.537  0.12433    
BILL_AMT2              2.604e-06  1.727e-06   1.508  0.13166    
BILL_AMT3              1.551e-06  1.564e-06   0.991  0.32150    
BILL_AMT4             -1.096e-08  1.702e-06  -0.006  0.99486    
BILL_AMT5              5.528e-07  1.868e-06   0.296  0.76729    
BILL_AMT6             -2.288e-07  1.412e-06  -0.162  0.87131    
PAY_AMT1              -1.579e-05  3.009e-06  -5.247 1.55e-07 ***
PAY_AMT2              -7.414e-06  2.315e-06  -3.202  0.00136 ** 
PAY_AMT3              -1.697e-06  2.142e-06  -0.792  0.42831    
PAY_AMT4              -3.032e-06  2.230e-06  -1.360  0.17389    
PAY_AMT5              -7.548e-07  1.956e-06  -0.386  0.69950    
PAY_AMT6              -2.870e-06  1.582e-06  -1.814  0.06971 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22567  on 20999  degrees of freedom
Residual deviance: 18898  on 20960  degrees of freedom
AIC: 18978
```

Stepwise selection

- ***step()*** is applicable on a glm object
- Starting from a full model, results in:

```
Coefficients:
                         Estimate Std. Error z value Pr(>|z|)    
(Intercept)            -4.120e+00  1.052e+00  -3.918 8.93e-05 ***
LIMIT_BA               -1.767e-06  2.020e-07  -8.750  < 2e-16 ***
SEX2                   -1.270e-01  3.731e-02  -3.404 0.000665 ***
EDUCATIONHS            -3.348e-02  5.766e-02  -0.581 0.561498    
EDUCATIONothers        -1.358e+00  2.763e-01  -4.914 8.92e-07 ***
EDUCATIONUni           -1.806e-02  4.337e-02  -0.416 0.677163    
MARRIAGE1               3.003e+00  1.047e+00   2.867 0.004141 ** 
MARRIAGE2               2.804e+00  1.048e+00   2.677 0.007425 ** 
MARRIAGE3               2.965e+00  1.060e+00   2.798 0.005141 ** 
PAY_1paid_in_full       1.354e-01  1.213e-01   1.116 0.264412    
PAY_1payment_delay      9.592e-01  1.005e-01   9.541  < 2e-16 ***
PAY_1revolving_credit -1.085e+00  1.241e-01  -8.739  < 2e-16 ***
PAY_2paid_in_full       2.421e-01  1.248e-01   1.939 0.052458 .  
PAY_2payment_delay      4.406e-01  1.304e-01   3.380 0.000725 ***
PAY_2revolving_credit   9.637e-01  1.457e-01   6.613 3.77e-11 ***
PAY_3paid_in_full      -1.128e-01  1.246e-01  -0.905 0.365228    
PAY_3payment_delay      3.203e-01  1.469e-01   2.179 0.029300 *  
PAY_3revolving_credit -1.114e-01  1.449e-01  -0.769 0.442092    
PAY_4paid_in_full      -8.103e-03  1.268e-01  -0.064 0.949037    
PAY_4payment_delay      2.504e-01  1.534e-01   1.632 0.102760    
PAY_4revolving_credit   4.138e-02  1.428e-01   0.290 0.771963    
PAY_5paid_in_full      -7.183e-02  1.235e-01  -0.581 0.560949    
PAY_5payment_delay      3.809e-01  1.522e-01   2.503 0.012308 *  
PAY_5revolving_credit   4.241e-02  1.366e-01   0.310 0.756303    
PAY_6paid_in_full      -1.243e-01  9.279e-02  -1.340 0.180391    
PAY_6payment_delay      5.598e-02  1.159e-01   0.483 0.629124    
PAY_6revolving_credit -2.968e-01  1.011e-01  -2.935 0.003332 ** 
```

```
BILL_AMT1              -2.185e-06  1.337e-06  -1.633 0.102395
BILL_AMT2               2.638e-06  1.726e-06   1.528 0.126422
BILL_AMT3               1.883e-06  1.215e-06   1.550 0.121123
PAY_AMT1               -1.621e-05  2.999e-06  -5.407 6.42e-08 ***
PAY_AMT2               -7.774e-06  2.277e-06  -3.414 0.000640 ***
PAY_AMT4               -2.962e-06  1.928e-06  -1.537 0.124359
PAY_AMT6               -2.981e-06  1.558e-06  -1.914 0.055615 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- So some of the payment related variables (e.g. PAY_AMT5) and AGE are excluded

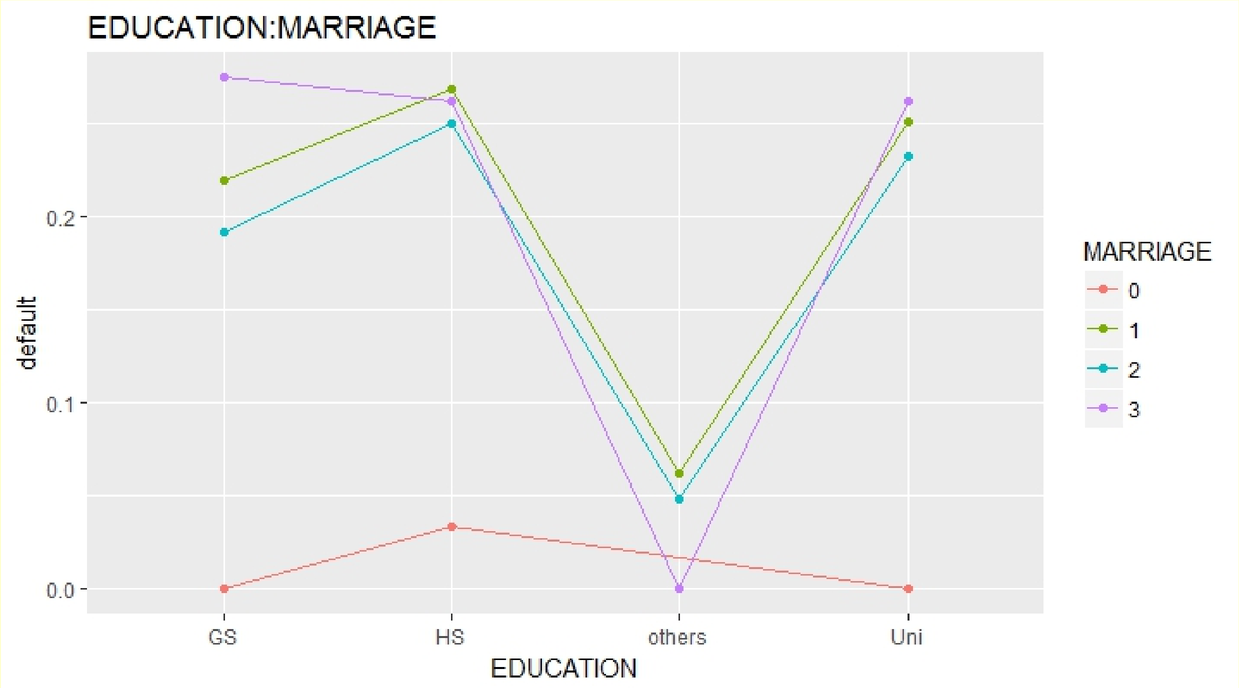
- Intuitively, those payment variables came in a sequential order of time, e.g. from April, May … and up to September
  - So, it dose not make sense to exclude some intermediate information
  - It is appropriate to keep them in sequential fashion
  - For instance, let's keep all the payment related variables (i.e. only exclude AGE)

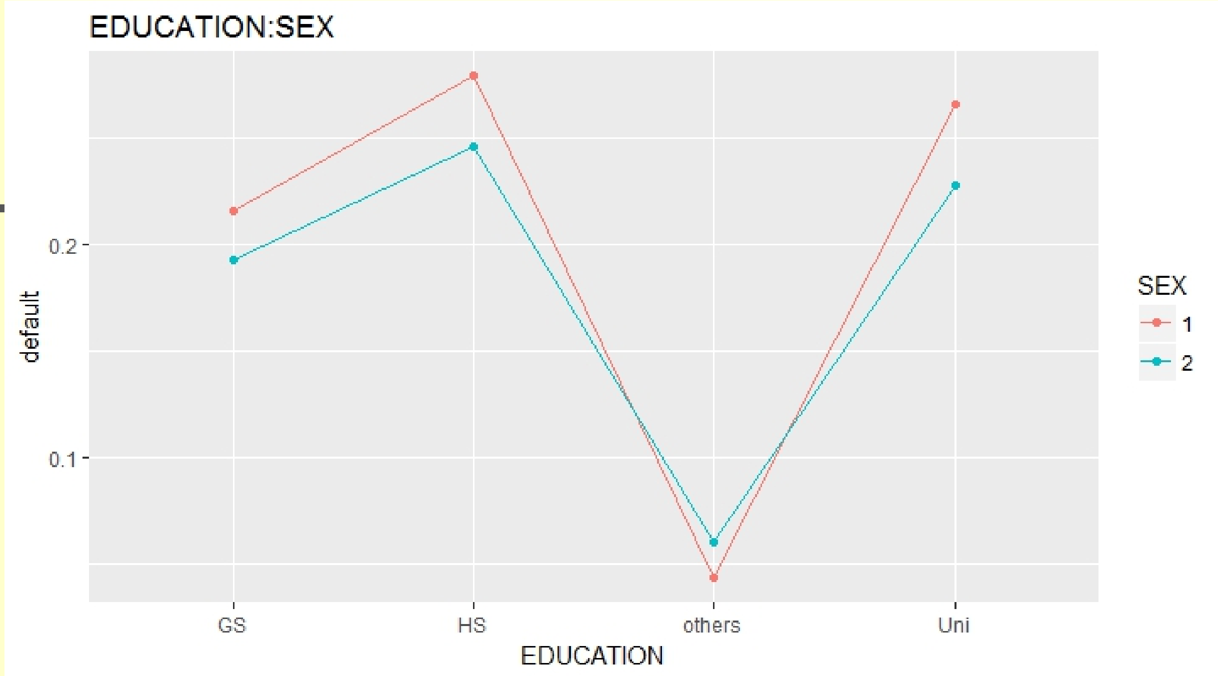
A model without AGE

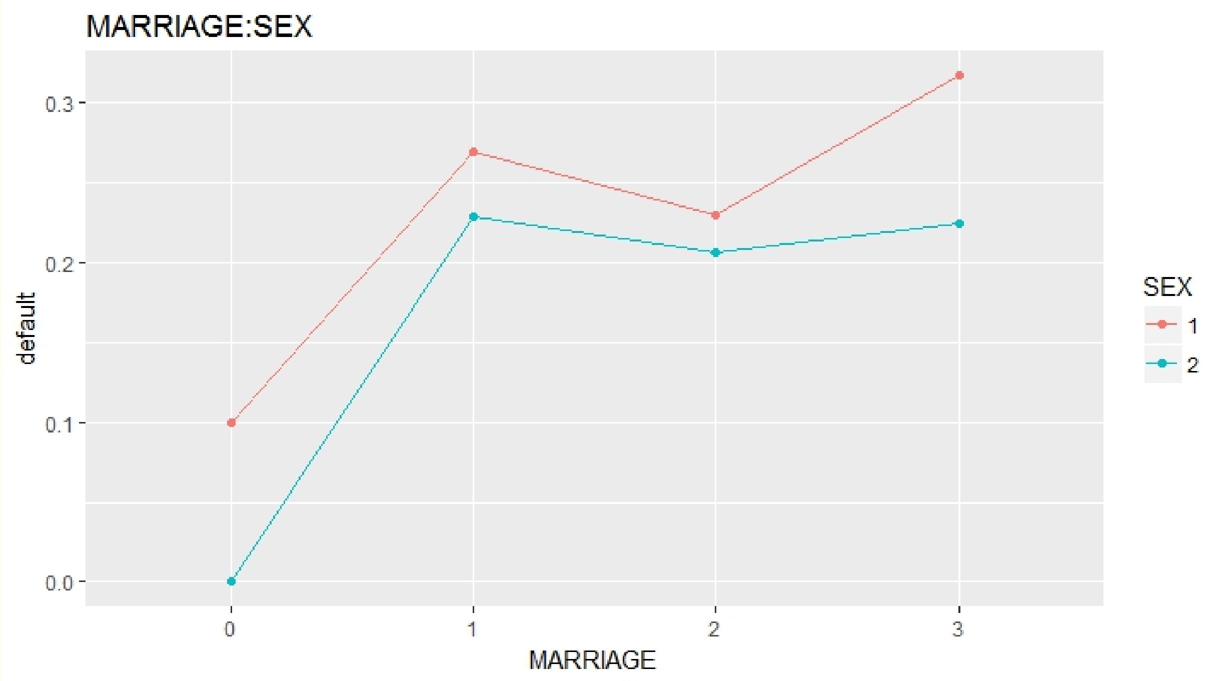
```
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)           -4.119e+00  1.051e+00  -3.918 8.94e-05 ***
LIMIT_BA              -1.749e-06  2.039e-07  -8.577  < 2e-16 ***
SEX2                  -1.273e-01  3.732e-02  -3.412 0.000646 ***
EDUCATIONHS           -3.307e-02  5.767e-02  -0.573 0.566354    
EDUCATIONothers       -1.355e+00  2.765e-01  -4.901 9.54e-07 ***
EDUCATIONUni          -1.769e-02  4.339e-02  -0.408 0.683413    
MARRIAGE1              3.004e+00  1.047e+00   2.868 0.004131 ** 
MARRIAGE2              2.806e+00  1.047e+00   2.679 0.007393 ** 
MARRIAGE3              2.968e+00  1.060e+00   2.801 0.005093 ** 
PAY_1paid_in_full      1.305e-01  1.214e-01   1.075 0.282171    
PAY_1payment_delay     9.539e-01  1.007e-01   9.478  < 2e-16 ***
PAY_1revolving_credit -1.090e+00  1.242e-01  -8.773  < 2e-16 ***
PAY_2paid_in_full      2.423e-01  1.248e-01   1.941 0.052200 .  
PAY_2payment_delay     4.425e-01  1.304e-01   3.394 0.000688 ***
PAY_2revolving_credit  9.654e-01  1.457e-01   6.624 3.50e-11 ***
PAY_3paid_in_full     -1.165e-01  1.247e-01  -0.934 0.350121    
PAY_3payment_delay     3.208e-01  1.469e-01   2.184 0.028997 *  
PAY_3revolving_credit -1.104e-01  1.449e-01  -0.762 0.446261    
PAY_4paid_in_full      2.605e-03  1.275e-01   0.020 0.983704    
PAY_4payment_delay     2.433e-01  1.542e-01   1.578 0.114627    
PAY_4revolving_credit  3.768e-02  1.435e-01   0.263 0.792786    
PAY_5paid_in_full     -7.515e-02  1.241e-01  -0.606 0.544828    
PAY_5payment_delay     3.880e-01  1.535e-01   2.527 0.011495 *  
PAY_5revolving_credit  4.768e-02  1.379e-01   0.346 0.729569    
PAY_6paid_in_full     -1.178e-01  9.418e-02  -1.251 0.211009    
PAY_6payment_delay     5.400e-02  1.176e-01   0.459 0.646025    
PAY_6revolving_credit -2.964e-01  1.025e-01  -2.892 0.003827 ** 
BILL_AMT1             -2.068e-06  1.341e-06  -1.543 0.122892    
BILL_AMT2              2.618e-06  1.728e-06   1.516 0.129613    
BILL_AMT3              1.559e-06  1.564e-06   0.997 0.318941    
BILL_AMT4             -1.629e-08  1.702e-06  -0.010 0.992362    
BILL_AMT5              5.544e-07  1.869e-06   0.297 0.766694    
BILL_AMT6             -2.293e-07  1.413e-06  -0.162 0.871107    
PAY_AMT1              -1.582e-05  3.011e-06  -5.255 1.48e-07 ***
PAY_AMT2              -7.420e-06  2.315e-06  -3.205 0.001352 ** 
PAY_AMT3              -1.687e-06  2.143e-06  -0.787 0.431137    
PAY_AMT4              -3.040e-06  2.229e-06  -1.364 0.172650    
PAY_AMT5              -7.562e-07  1.957e-06  -0.386 0.699207    
PAY_AMT6              -2.884e-06  1.583e-06  -1.823 0.068356 .  
---
```

Interaction terms

- We may further consider the interaction between the categorical variables
  - EDUCATION:MARRIAGE
  - EDUCATION:SEX
  - MARRIAGE:SEX
- Interaction plots:




EDUCATION:SEX
default
0.2
0.1
GS
HS
others
Uni
EDUCATION
SEX
1
2


MARRIAGE:SEX
default
0.3
0.2
0.1
0.0
0
1
2
3
MARRIAGE
SEX
1
2

- It seems that there are minor interaction effect for **EDUCATION:MARRIAGE** and **EDUCATION:SEX**
- We may further verify the significance of them by using LR test
  - Uses ***lrtest()*** from ***lmtest*** library

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
1	39	-9449.5			
2	50	-9446.0	11	7.114	0.7898

  - Interaction terms are not significant.
  - So, we keep using the model without AGE for classification

- the VIF of the majority of payment related variables are so large
- It is expected since they are payment history
  - E.g. repayment this month is mostly related to the bill amount previously

```
> vif(new_fit)
                GVIF
LIMIT_BA    1.689491
SEX         1.011586
EDUCATION   1.189251
MARRIAGE    1.082602
PAY_1      12.452604
PAY_2      54.259863
PAY_3      41.513605
PAY_4      44.240181
PAY_5      41.879204
PAY_6      15.051068
BILL_AMT1  24.080735
BILL_AMT2  38.242821
BILL_AMT3  28.416455
BILL_AMT4  29.211652
BILL_AMT5  33.089912
BILL_AMT6  18.121010
PAY_AMT1    1.482329
PAY_AMT2    1.536907
PAY_AMT3    1.493976
PAY_AMT4    1.553844
PAY_AMT5    1.571602
PAY_AMT6    1.124503
```

Prediction Example

- Once we have the model, we can predict the probability of a default case.
- Take the first client from the testing set as an example:

LIMIT_BA	SEX	EDUCATION	MARRIAGE	AGE
30000	1	Uni	1	36

PAY_1	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6
Payment_delay	Paid_in_full	Paid_in_full	Paid_in_full	Revolving_credit	Revolving_credit

BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6
0	780	0	1170	780	0

PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6
780	0	1170	0	0	0

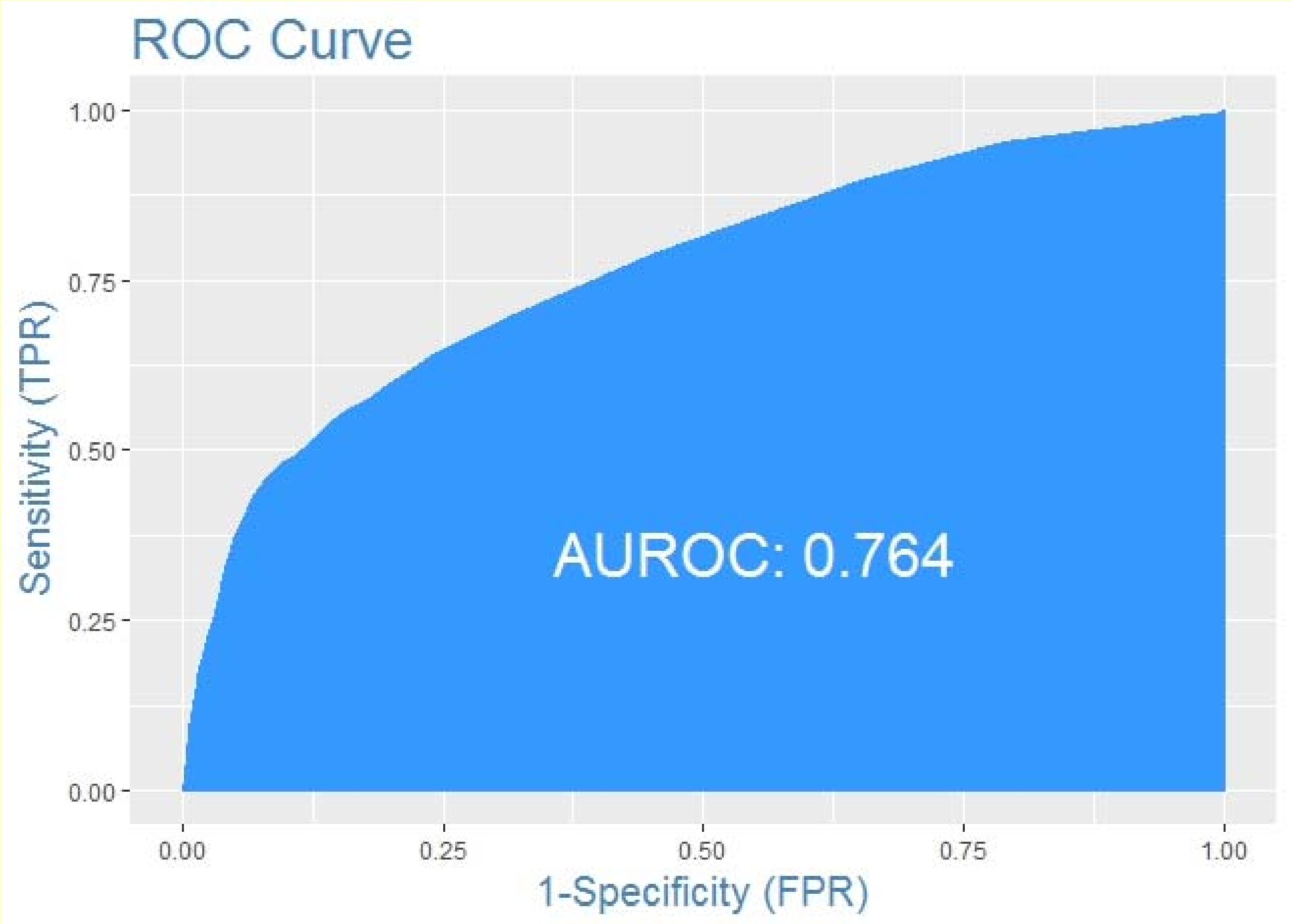
- ***predict()*** is used to predict the default probability
  - By setting type='response'
  - Results in 0.40996, which is the default probability
- Classification rule
  - If we set 'default probability>0.5' to be default and non-default otherwise
  - Then, this customer is predicted as non-default next month and we may accept his loan application

Odds Ratio

		2.5 %	97.5 %
(Intercept)	0.01625688	0.0008823536	0.08425196
LIMIT_BA	0.99999825	0.9999978491	0.99999865
SEX2	0.88043441	0.8183704229	0.94731435
EDUCATIONHS	0.96747186	0.8638441873	1.08297529
EDUCATIONothers	0.25798343	0.1439329300	0.42845994
EDUCATIONUni	0.98246235	0.9024035901	1.06970915
MARRIAGE1	20.15631074	3.9378543754	369.85434468
MARRIAGE2	16.53614338	3.2293549660	303.46325141
MARRIAGE3	19.45156851	3.6627945710	361.21138931
PAY_1paid_in_full	1.13945383	0.8991524111	1.44717112
PAY_1payment_delay	2.59592023	2.1331568760	3.16540273
PAY_1revolving_credit	0.33627180	0.2637691930	0.42927169
PAY_2paid_in_full	1.27415864	0.9976489423	1.62727600
PAY_2payment_delay	1.55656917	1.2054271817	2.00961846
PAY_2revolving_credit	2.62587807	1.9730653619	3.49377853
PAY_3paid_in_full	0.89004800	0.6977299110	1.13749732
PAY_3payment_delay	1.37829123	1.0340556012	1.83966240
PAY_3revolving_credit	0.89548852	0.6746500517	1.19081131
PAY_4paid_in_full	1.00260823	0.7813155060	1.28812695
PAY_4payment_delay	1.27542946	0.9430620405	1.72617929
PAY_4revolving_credit	1.03840285	0.7843314593	1.37640572
PAY_5paid_in_full	0.92760517	0.7277209876	1.18379054
PAY_5payment_delay	1.47406624	1.0914673699	1.99262025
PAY_5revolving_credit	1.04883573	0.8009164680	1.37540626
PAY_6paid_in_full	0.88887863	0.7392933628	1.06946950
PAY_6payment_delay	1.05548819	0.8384617102	1.32947548
PAY_6revolving_credit	0.74348027	0.6086763958	0.90969535
BILL_AMT1	0.99999793	0.9999952309	1.00000048
BILL_AMT2	1.00000262	0.9999992216	1.00000600
BILL_AMT3	1.00000156	0.9999984846	1.00000462
BILL_AMT4	0.99999998	0.9999965969	1.00000326
BILL_AMT5	1.00000055	0.9999969074	1.00000425
BILL_AMT6	0.99999977	0.9999970480	1.00000259
PAY_AMT1	0.99998418	0.9999780281	0.99998982
PAY_AMT2	0.99999258	0.9999878282	0.99999691
PAY_AMT3	0.99999831	0.9999938741	1.00000223
PAY_AMT4	0.99999696	0.9999923544	1.00000108
PAY_AMT5	0.99999924	0.9999952729	1.00000296
PAY_AMT6	0.99999712	0.9999938975	1.00000011

ROC Curve

- ROC curve can be constructed by ***plotROC()*** which is under ***InformationValue*** library
  - By using the testing set



Error Measure

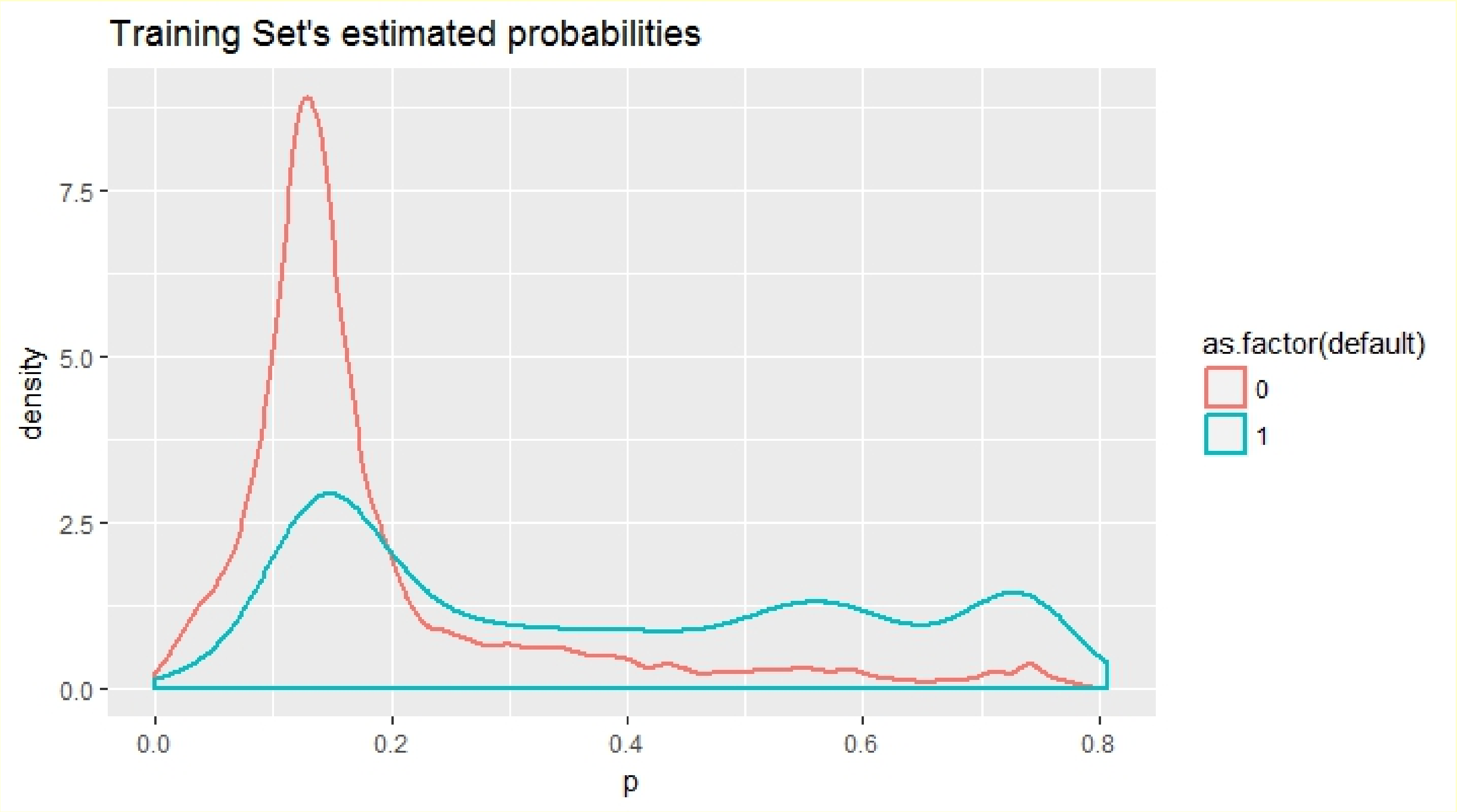
- Sensitivity, specificity and etc., can be computed under the ***InformationValue*** library too
  - Use 0.5 (default setting) as a cutoff

```
> sensitivity(test_set$default,prob)
[1] 0.3375
> specificity(test_set$default,prob)
[1] 0.9565642
> confusionMatrix(test_set$default,prob)
     0    1
0 6849 1219
1  311  621
> misClassError(test_set$default,prob)
[1] 0.17
```

- The reported confusion matrix is structured as follow: Actual is put vertically and predicted is put horizontally

Unbalance Response

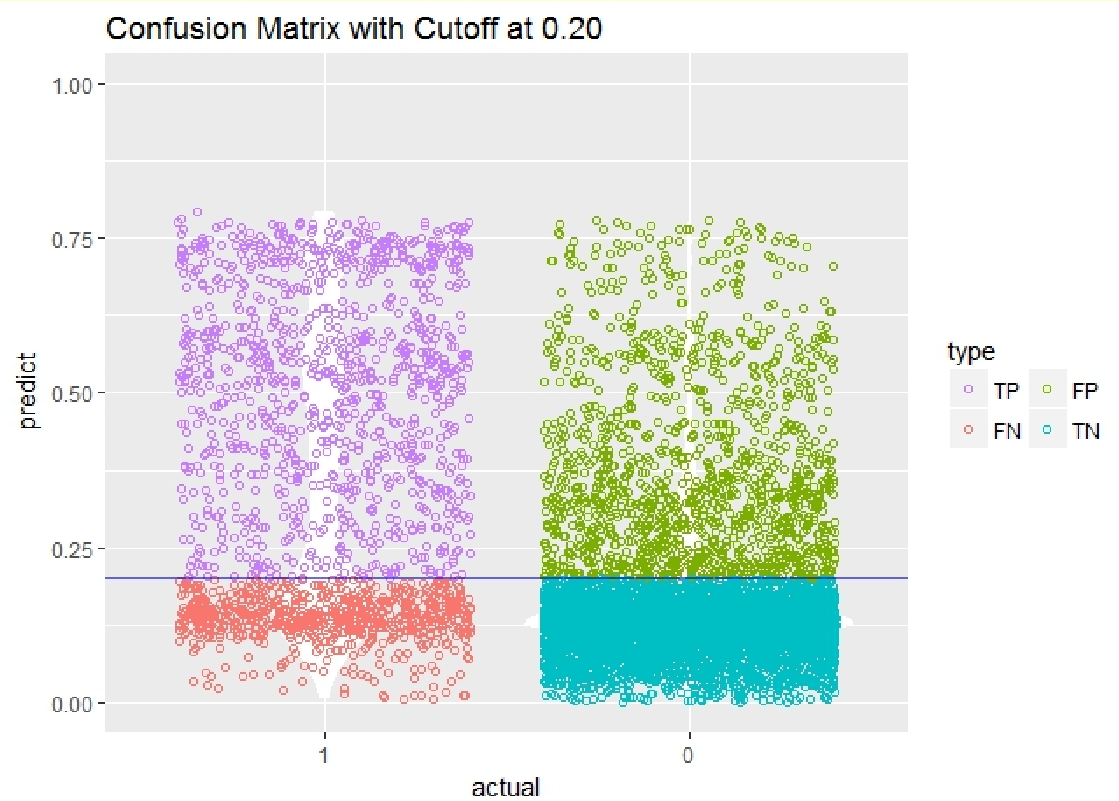
- This is a common feature of the credit data, such that there is a dominate group of response
  - There are ~80% response of non-default in our data
- It impacts on the classification cutoff
  - The double density plot shows the difficulty in determine the cutoff
  - An ideal double density plot should show two separated densities
    - Non-default on the left and default cases on the right
  - The worst case is they are close to each other

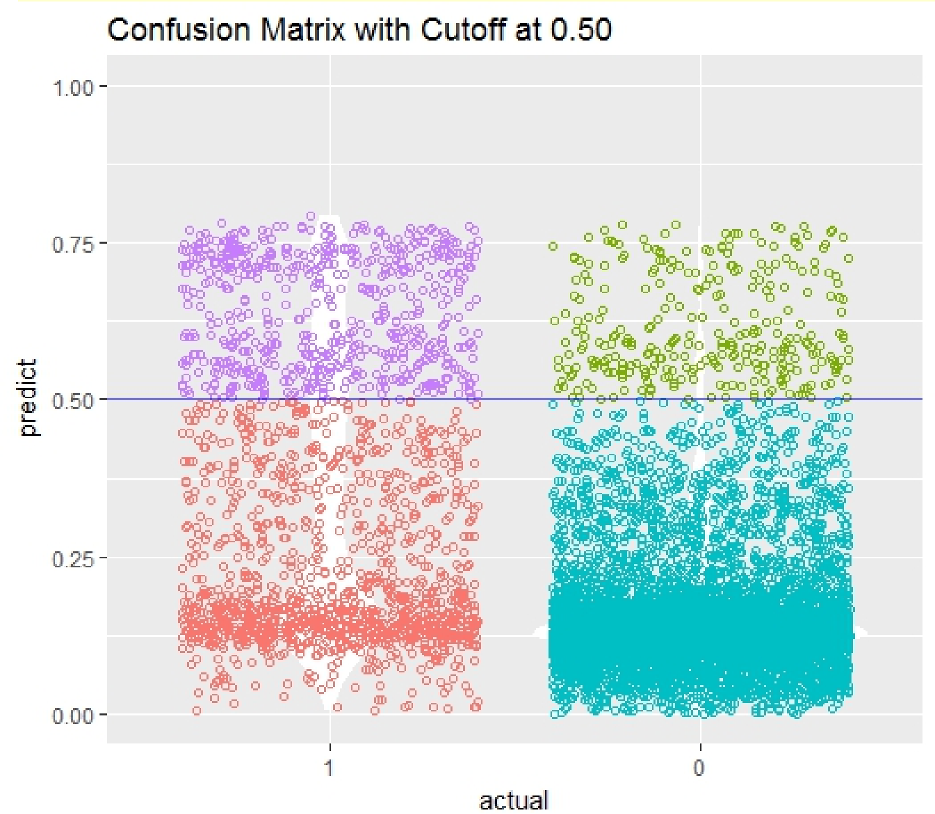

Training Set's estimated probabilities
7.5
5.0
2.5
0.0
density
0.0
0.2
0.4
0.6
0.8
p
as.factor(default)
0
1

- Obviously, the densities overlap
- The mode of default and non-default probabilities are around 0.15 and 0.12 respectively. They are close to each other
  - The reason for this is because our dataset only consists of ~20 percent of default cases
- And the variation for default probabilities is pretty large
- So it is hard to determine a cutoff

Optimal Cutoff

- The choice of optimal cutoff is sometimes subjective
  - Some may seek for an optimal cutoff by minimizing the misclassification rate
  - One may only maximize the True Positive Rate
  - Etc.,
- There is always a tradeoff between false negative and false positive
  - Take the testing set as an example


Confusion Matrix with Cutoff at 0.20
1.00
0.75
0.50
0.25
0.00
predict
1
0
actual
type
TP
FP
FN
TN


Confusion Matrix with Cutoff at 0.50
1.00
0.75
0.50
0.25
0.00
predict
1
0
actual

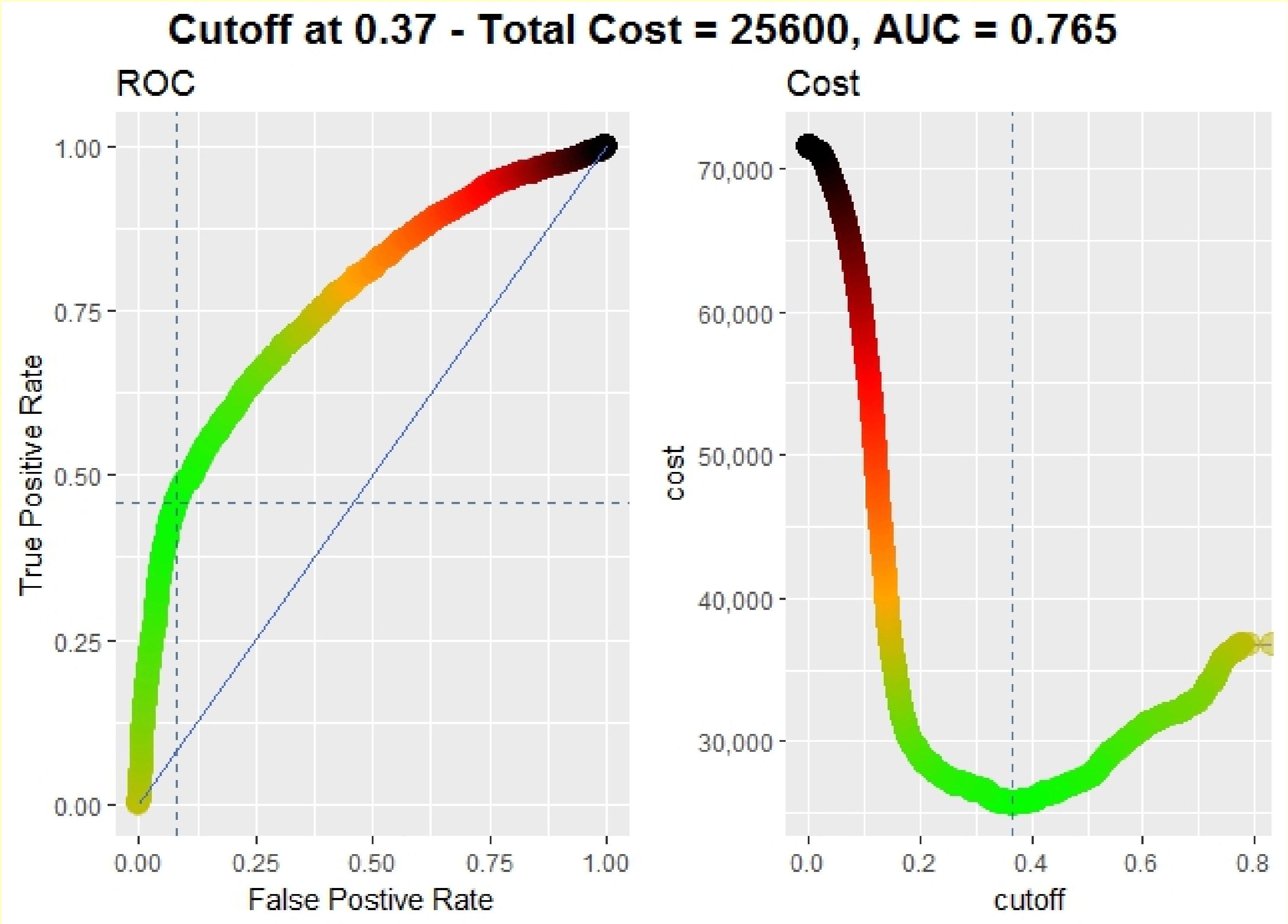
- As the cutoff line rises, the number of false positive reduces but at the same time the number of false negative increases
- In our case, the risk is lending loan to a client who will go default later
  - So we want to have a better control on the FALSE NEGATIVE case

- Assume that, on average lending into default (<u>false negative</u>) is two times as costly as not lending to a good debtor (<u>false positive</u>)
- Then we may set the cost for false negative and false positive and use this cost to find the optimal cutoff.

- Say, the cost of false positive = 10 and the cost of false negative = 20
  - i.e. the weighting of false negative is a double of false positive
- ***ROCInfo()*** from a third party function* compute the cutoff according to cost

* Load ‘unbalanced_function.R’ to invoke it

- Optimal cutoff =0.37



- The misclassification rate becomes a bit higher, but it results in a smaller false negative rate, which are two times more expensive than the false positive error.

```
> sensitivity(test_set$default,prob,threshold = roc_info$cutoff)
[1] 0.4586957
> specificity(test_set$default,prob,threshold = roc_info$cutoff)
[1] 0.9206704
> confusionMatrix(test_set$default,prob,threshold = roc_info$cutoff)
     0   1
0 6592 996
1  568 844
> misClassError(test_set$default,prob,threshold = roc_info$cutoff)
[1] 0.1738
```

Business Implications

- As a loan officer, to identify group of potential risk clients is far more important than just classification
- The odds ratio is an alert tool to risk clients
  - An odds ratio > 1 means the client has a relatively high default risk
  - Together with its CI, we may identify those risky clients in default

- Payment related variables with 95%CI above 1:

		2.5 %	97.5 %
PAY_1payment_delay	2.595920	2.133157	3.165403
PAY_2payment_delay	1.556569	1.205427	2.009618
PAY_2revolving_credit	2.625878	1.973065	3.493779
PAY_3payment_delay	1.378291	1.034056	1.839662
PAY_5payment_delay	1.474066	1.091467	1.992620

  - Obviously, those keep delaying payment are exposed to higher risk in default relative to those without consumption (the base group)
  - In particular, we have to stay alert to those revolving credit a month before.
  - So, the loan officer may pay attention to those risky clients. Probably to reject their loan application or charge them a higher management fee.