Capstone Project:
Taiwan Customer Defaults

Project Notes 1

Amit KULKARNI

Table of Contents:

Project Approach.....3

Discovery.....3

1.1.1: Objective.....3

1.1.2: The hypothesis.....4

1.1.3: Potential Data Sources.....4

1.1.3: Applications list.....4

2: Data – Basic Exploratory Analysis.....4

2.1.1: Variable definition as provided.....5

2.1.2: Dimensions of the dataset provided.....5

2.1.3: Glimpse of first 6 records of the data set using the "Head" function.....5

2.1.4: Total number of defaulted customers 5

2.1.5: Structure of the data 6

2.1.6: Basic Visualization of the data 7

3: Univariate Analysis.....10

3.1.1: Regression Analysis on each variable individually is done with the dependent variable "default".....10

3.1.1: Regression Analysis on bivariate or multi-variate is done with the dependent variable "default" 10

1: Project Approach

A typical developmental Lifecycle can be adopted for this project which consists the following

1. Discovery

1.1.1: Objective

Banks extend credits and advances to customers to earn their core income of interest. Variety of parameters are studied before such an advance is done like the education, current working organization, historical payment trends etc. Additionally demographic parameters like marital status, gender, Age etc. are also considered while advancing a loan. However even after exhaustive study of the parameters which fairly would have indicated that the customer may not default, customers do default. Banks operate with this type of risk which is called as **credit risk** and would prefer that this risk be mitigated or kept as lower as possible.

In our case, data pertaining to 30,000 customers has been provided with ~24 variables like the repayment trend for past 6 months, bills for the last 6 months, average balance limit and payments made with their demographics like Sex, Marriage status and education. With these ~24 variables another variable pertaining to their default status has been provided where

1 = defaulted customer

0 = not a default customer

The overall objective of this project is to have the below developmental cycle

i. Study the variables provided
ii. Perform an in depth Exploratory Data Analysis
iii. Check for Multicollinearity amongst variables
iv. Perform PCA and FA if required and check if there is a possibility of dimension reduction
v. Perform the regression analysis on reduced dimensions
vi. Develop various applicability models which could be Supervised learning or unsupervised learning models
vii. Divide the data into Train and Test data and test each of the models with their accuracies, confusion matrices etc.
viii. Compare amongst models to identify the best model that can be used to predict the default for customers

This entire project is divided into 3 project notes that will cover the above mentioned 8 steps. **For the purpose of project notes 1, only point (i) will be applicable**

1.1.2: The hypothesis: To check if the default of customer is dependent on all the provided variables.

Subsequently a holistic view of macro environment variables can also be considered to see if that also could cause customers to default. Some macro environment variables would be the overall economy and its impact on the customers, government rulings etc.

1.1.3: Potential Data Sources: The data source referred in this project is a .csv file which has the requisite information for the modelling purpose.

Certain tweaking of the data without losing its value has been done so as to accommodate visualization in Tableau

1.1.4: Applications list:

i. R
ii. Tableau
iii. MS Office

2: Data – Basic Exploratory Analysis

2.1.1: Variable definition as provided

Column Name	Column Definition
X1	Amount of the given credit (NT dollar) it includes both the individual consumer credit and his/her family (supplementary) credit.
X2	Gender (1 = male; 2 = female).
X3	Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
X4	Marital status (1 = married; 2 = single; 3 = others).
X5	Age (year).
X6 - X11	History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is -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.
X12-X17	Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.
X18-X23	Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.
X18-X23:	Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

2.1.2: Dimensions of the dataset provided

No of Rows	No of Columns
30000	25

There are 30000 rows and 25 variables provided in the dataset

2.1.3: Glimpse of first 6 records of the data set using the "Head" function

LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	Rep0905	Rep0805	Rep0705	Rep0605	Rep0505	Rep0405	Bill0905
20000	2	2	1	24	2	2	-1	-1	-2	-2	3,913.00
120000	2	2	2	26	-1	2	0	0	0	2	2,682.00
90000	2	2	2	34	0	0	0	0	0	0	29,239.00
50000	2	2	1	37	0	0	0	0	0	0	46,990.00
50000	1	2	1	57	-1	0	-1	0	0	0	8,617.00
50000	1	1	2	37	0	0	0	0	0	0	64,400.00

Bill0805	Bill0705	Bill0605	Bill0505	Bill0405	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	default
3,102.00	689.00	-	-	-	-	689.00	-	-	-	-	1
1,725.00	2,682.00	3,272.00	3,455.00	3,261.00	-	1,000.00	1,000.00	1,000.00	-	2,000.00	1
14,027.00	13,559.00	14,331.00	14,948.00	15,549.00	1,518.00	1,500.00	1,000.00	1,000.00	1,000.00	5,000.00	0
48,233.00	49,291.00	28,314.00	28,959.00	29,547.00	2,000.00	2,019.00	1,200.00	1,100.00	1,069.00	1,000.00	0
5,670.00	35,835.00	20,940.00	19,146.00	19,131.00	2,000.00	36,681.00	10,000.00	9,000.00	689.00	679.00	0
57,069.00	57,608.00	19,394.00	19,619.00	20,024.00	2,500.00	1,815.00	657.00	1,000.00	1,000.00	800.00	0

2.1.4: Total number of defaulted customer are

```
> sum(data$default)
[1] 6636
> x= sum(data$default)
> x/30000
[1] 0.2212
> x/30000*100
[1] 22.12
```

~ 22% of the customers have defaulted in their payments

2.1.5: Structure of the data

Column Name	Column Structure
ID :	int 1 2 3 4 5 6 7 8 9 10 ...
LIMIT_BAL:	int 20000 120000 90000 50000 50000 50000 500000 100000 140000 20000 ...
SEX :	int 2 2 2 2 1 1 1 2 2 1 ...
EDUCATION:	int 2 2 2 2 2 1 1 2 3 3 ...
MARRIAGE :	int 1 2 2 1 1 2 2 2 1 2 ...
AGE :	int 24 26 34 37 57 37 29 23 28 35 ...
Rep0905 :	int 2 -1 0 0 -1 0 0 0 0 -2 ...
Rep0805 :	int 2 2 0 0 0 0 0 -1 0 -2 ...
Rep0705 :	int -1 0 0 0 -1 0 0 -1 2 -2 ...
Rep0605 :	int -1 0 0 0 0 0 0 0 0 -2 ...
Rep0505 :	int -2 0 0 0 0 0 0 0 0 -1 ...
Rep0405 :	int -2 2 0 0 0 0 0 -1 0 -1 ...
Bill0905 :	int 3913 2682 29239 46990 8617 64400 367965 11876 11285 0 ...
Bill0805 :	int 3102 1725 14027 48233 5670 57069 412023 380 14096 0 ...
Bill0705 :	int 689 2682 13559 49291 35835 57608 445007 601 12108 0 ...
Bill0605 :	int 0 3272 14331 28314 20940 19394 542653 221 12211 0 ...
Bill0505 :	int 0 3455 14948 28959 19146 19619 483003 -159 11793 13007 ...
Bill0405 :	int 0 3261 15549 29547 19131 20024 473944 567 3719 13912 ...
PAY_AMT1 :	int 0 0 1518 2000 2000 2500 55000 380 3329 0 ...
PAY_AMT2 :	int 689 1000 1500 2019 36681 1815 40000 601 0 0 ...
PAY_AMT3 :	int 0 1000 1000 1200 10000 657 38000 0 432 0 ...
PAY_AMT4 :	int 0 1000 1000 1100 9000 1000 20239 581 1000 13007 ...
PAY_AMT5 :	int 0 0 1000 1069 689 1000 13750 1687 1000 1122 ...
PAY_AMT6 :	int 0 2000 5000 1000 679 800 13770 1542 1000 0 ...
default :	int 1 1 0 0 0 0 0 0 0 0 ...

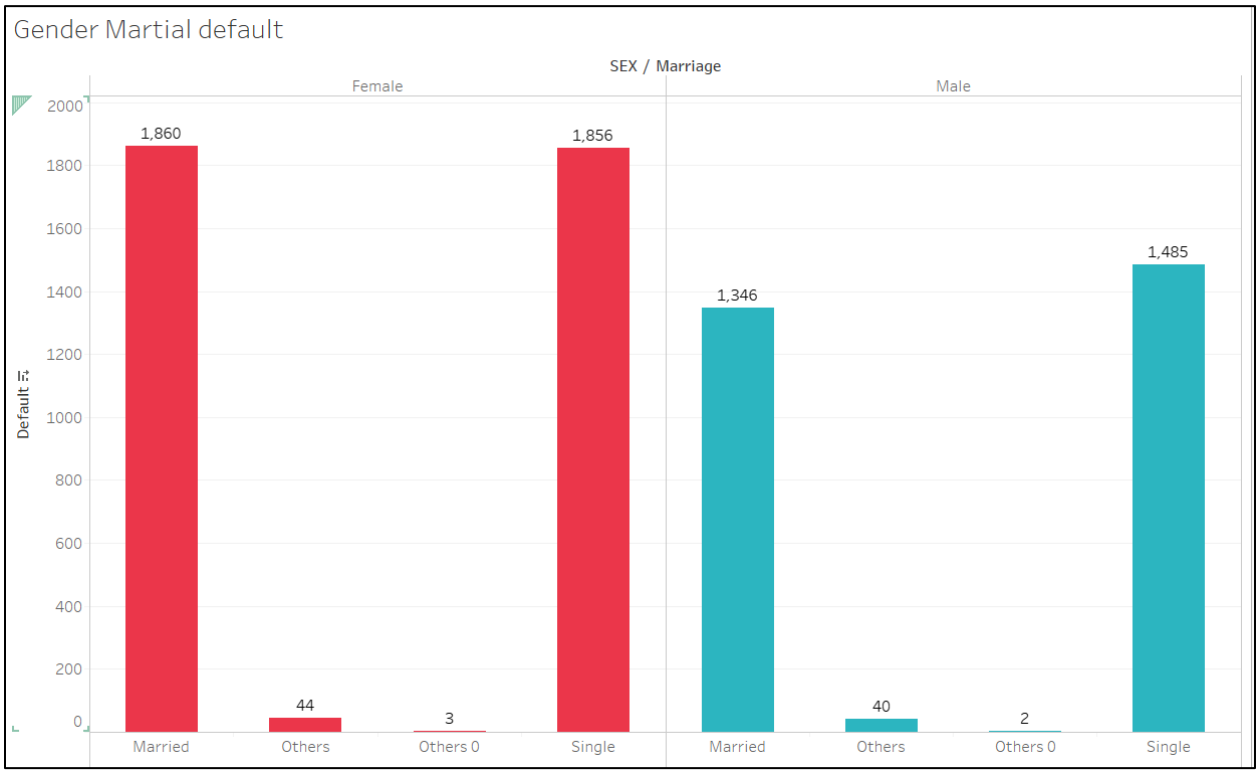
2.1.6: Basic Visualization of the data

Following variables have been converted to categorical variables so that they can be further used as dimensions to provide basic visualization:

i. Marriage
ii. Education
iii. Sex
iv. Marital Status

All the above variables were studied with the default measure to bring insights

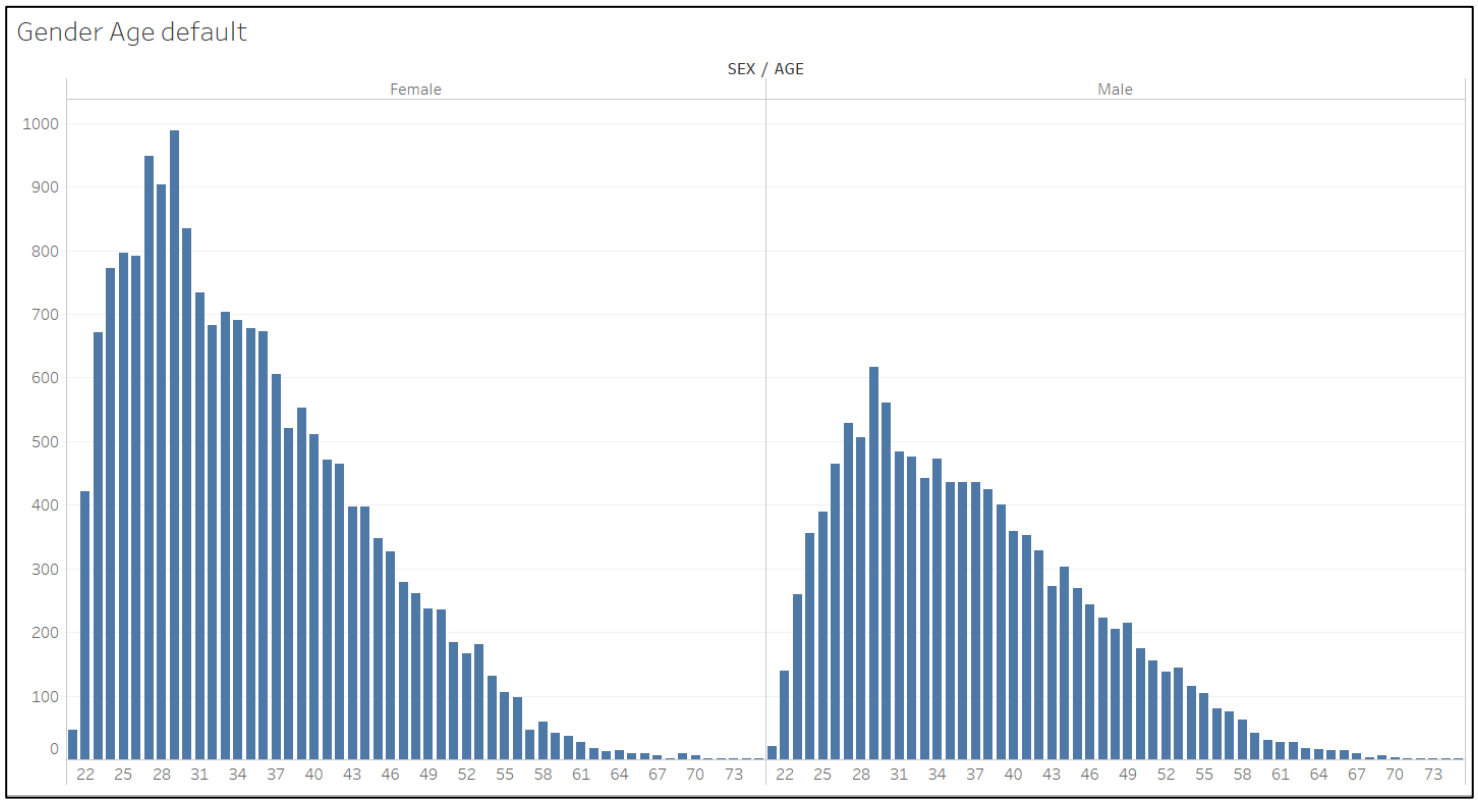
1. The below graphs shows the split at gender and their marital status who have defaulted



It is observed that Female irrespective of their marital status seem to default in their payments however there are slightly higher chance of Male being single and defaulting

2. Gender Education default ratio: It can be observed Females who have done their Universities default the maximum followed by Male who have done their University. It can be observed that irrespective of the Gender but the type of education is driving the default pattern for the customers

3. Gender to Age to Default ratio: It can be observed that women in the age bracket of 26-30 years in Females and 29-31 years in Males have the highest defaults. As the age progresses it can be observed that the defaults have a declining trend



3: Univariate Analysis

3.1.1: Regression Analysis on each variable individually is done with the dependent variable "default"

```
lm1 = lm(default ~ LIMIT_BAL, data = data)
summary(lm1)
lm2 = lm(default ~ SEX, data = data)
summary(lm2)
lm3 = lm(default ~ EDUCATION, data = data)
summary(lm3)
lm4 = lm(default ~ MARRIAGE, data = data)
summary(lm4)
lm5 = lm(default ~ AGE , data = data)
summary(lm5)
lm6 = lm(default ~ Rep0905, data = data)
summary(lm6)
lm7 = lm(default ~ Rep0805, data = data)
summary(lm7)
lm8 = lm(default ~ Rep0705, data = data)
summary(lm8)
lm9 = lm(default ~ Rep0605, data = data)
summary(lm9)
lm10 = lm(default ~ Rep0505, data = data)
summary(lm10)
lm11 = lm(default ~ Rep0405, data = data)
summary(lm11)
lm12 = lm(default ~ Bill0905, data = data)
summary(lm12)
```

```
lm13 = lm(default ~ Bill0805, data = data)
summary(lm13)
lm14 = lm(default ~ Bill0705, data = data)
summary(lm14)
lm15 = lm(default ~ Bill0605, data = data)
summary(lm15)
lm16 = lm(default ~ Bill0505, data = data)
summary(lm16)
lm17 = lm(default ~ Bill0405, data = data)
summary(lm17)
lm18 = lm(default ~ PAY_AMT1, data = data)
summary(lm18)
lm = lm(default ~ PAY_AMT1, data = data)
summary(lm)
lm19 = lm(default ~ PAY_AMT1, data = data)
summary(lm19)
lm20 = lm(default ~ PAY_AMT2, data = data)
summary(lm20)
lm21 = lm(default ~ PAY_AMT3, data = data)
summary(lm21)
lm22 = lm(default ~ PAY_AMT4, data = data)
summary(lm22)
lm23 = lm(default ~ PAY_AMT5, data = data)
summary(lm23)
lm24 = lm(default ~ PAY_AMT6, data = data)
summary(lm24)
```

The results of all the commands given above is given the table below

(Intercept)	Adjusted R2
LIMIT_BAL	0.02354
SEX	0.00156
EDUCATION	0.00075
MARRIAGE	0.00056
AGE	0.00016
Rep0905	0.10550
Rep0805	0.06943
Rep0705	0.05531
Rep0605	0.04689
Rep0505	0.04164
Rep0405	0.03489
Bill0905	0.00035
Bill0805	0.00017
Bill0705	0.00016
Bill0605	0.00007
Bill0505	0.00000
Bill0405	0.00529
PAY_AMT1	0.00529
PAY_AMT2	0.00340
PAY_AMT3	0.00313
PAY_AMT4	0.00320
PAY_AMT5	0.00301
PAY_AMT6	0.00280

Correlation between default and various other 24 variables seems to be very insignificant except Rep0905 which close to 10% which also is not significant enough.

3.1.1: Regression Analysis on bivariate or multi-variate is done with the dependent variable "default"

lm25 = lm (default ~ ., data = data)

summary (lm25)

(Intercept)	Estimate	Std. Error	t value	Pr(>|t|)
LIMIT_BAL	-9.05E-08	2.16E-08	-4.19E+00	2.80e-05 ***
SEX	-1.45E-02	4.64E-03	-3.13E+00	0.00177 **
EDUCATION	-1.51E-02	3.02E-03	-5.01E+00	5.56e-07 ***
MARRIAGE	-2.38E-02	4.77E-03	-5.00E+00	5.76e-07 ***
AGE	1.41E-03	2.75E-04	5.13E+00	2.95e-07 ***
Rep0905	9.57E-02	2.77E-03	3.46E+01	< 2e-16 ***
Rep0805	1.95E-02	3.34E-03	5.83E+00	5.56e-09 ***
Rep0705	1.17E-02	3.59E-03	3.25E+00	0.00115 **
Rep0605	3.40E-03	3.98E-03	8.55E-01	0.39272
Rep0505	5.67E-03	4.31E-03	1.32E+00	0.1881
Rep0405	7.90E-04	3.52E-03	2.24E-01	0.82237
Bill0905	-6.22E-07	1.14E-07	-5.45E+00	4.99e-08 ***
Bill0805	1.58E-07	1.60E-07	9.85E-01	0.32464
Bill0705	3.02E-08	1.51E-07	2.00E-01	0.84146
Bill0605	-6.50E-08	1.58E-07	-4.12E-01	0.68013
Bill0505	-2.29E-08	1.85E-07	-1.24E-01	0.90121
Bill0405	1.16E-07	1.46E-07	7.91E-01	0.42889
PAY_AMT1	-7.44E-07	1.77E-07	-4.20E+00	2.68e-05 ***
PAY_AMT2	-2.10E-07	1.46E-07	-1.44E+00	0.14975
PAY_AMT3	-2.83E-08	1.69E-07	-1.68E-01	0.86696
PAY_AMT4	-2.50E-07	1.84E-07	-1.36E+00	0.17361
PAY_AMT5	-3.42E-07	1.91E-07	-1.79E+00	0.07334 .
PAY_AMT6	-9.80E-08	1.37E-07	-7.18E-01	0.47279

The overall R-squared value is ~12.5% making it not a significant linear relationship amongst dependent and independent variables

End of Report