EB5103 Advanced Analytics

PCA / Cluster Analysis of Credit Card Clients of a Taiwanese Bank

Lecturer: Dr. Rita Chakravarti

Student ID	Name	E-mail
A0178551X	Choo Ming Hui Raymond	e0267862@u.nus.edu
A0056418A	Tan Zhi Wen David	e0267355@u.nus.edu
A0178329R	Wong Yeng Fai Edric	e0267640@u.nus.edu

Submission Date: 3rd March 2018

Table of Contents

1. Introduction

The credit department of a Taiwanese bank is interested in predicting loan default based on known data like demographic factors, credit data, history of payment and bill statements of its customers. The bank would also like to identify different segments of its clientele based on data collected from April to September 2005. This report describes the use of Principal Component Analysis (PCA), K-means clustering and Logistic Regression to achieve these objectives.

2. Data Preparation and Exploration

2.1 Dataset information

The dataset used for this analysis contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients from a Taiwanese Bank from April to September 2005. In total, there are 30,000 observations with 25 variables. The data dictionary is included in **Appendix A**.

2.2 Data Preparation

While no missing values in the dataset was detected, discrepancies were found. As they were assessed to negatively influence further analysis, actions were taken to resolve them and is summarised in Table 1 below.

Table 1: List of observations found during inspection

Variable Name	Possible values	Observation	Actions Taken
default.payment.next. month	1=yes, 0=no	209 instances where Pay_0 = 1, Pay_2, 3, 4, 5, 6 = -2, BILL_AMT1,2,3,4,5,6 = 0, but default = 1 108 instances where Pay_0, 2, 3, 4, 5, 6 = -2, BILL_AMT1,2,3,4,5,6 = 0, but default = 1	These instances describe clients with no debit in the preceding 6 months but defaults on the 7th month. This is illogical and are removed from the dataset after being assessed as data errors.
Education	1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown	15 instances of ‘0’, 280 instances of ‘5’ and 51 instances of ‘6’	There is no appreciable difference between ‘0’, ‘5’ and ‘6’. Hence, these entries are subsumed under “4 = others”
Marriage	1=married, 2=single, 3=others	There are 54 instances of ‘0’	There is no appreciable difference between ‘0’ and ‘3’. Hence, these entries are subsumed under “3=others”

2.3 Multi-collinearity

Multi-collinearity of the variables was tested first as it adversely impacts logistic regression. This was done via visual inspection of the scatterplots for all continuous variables, modelled pairwise in JMP. There were several groups of variables which demonstrates multi-collinearity, one of which is shown in Figure 1 below, involving the amount billed to the client over various months. As such, PCA was assessed to be a suitable dimension reduction technique as it also removes multi-collinearity, which facilitates subsequent logistic regression modelling.

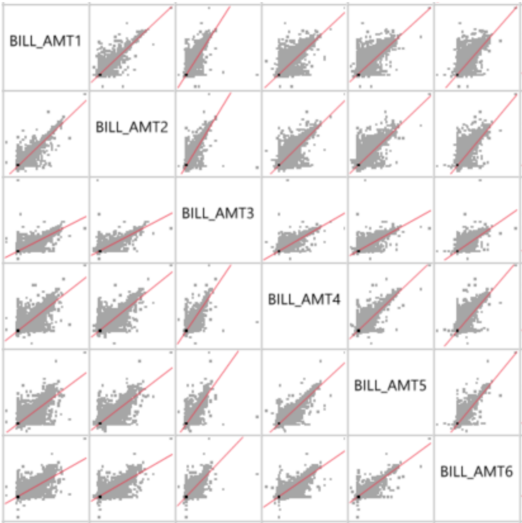

Figure 1: Selected Evidence of Multi-collinearity in the Dataset

3. Principal Component Analysis

3.1 Selection of Variables for Reduction

PCA is typically performed on continuous variables although ordinal variables with sufficient distinction (e.g. Likert Scale with sufficient intervals) may be considered. As such, the following variables were chosen:

Table 2: List of Variables Chosen for PCA

LIMIT_BAL	PAY_3	BILL_AMT1	BILL_AMT5	PAY_AMT3
AGE	PAY_4	BILL_AMT2	BILL_AMT6	PAY_AMT4
PAY_0	PAY_5	BILL_AMT3	PAY_AMT1	PAY_AMT5
PAY_2	PAY_6	BILL_AMT4	PAY_AMT2	PAY_AMT6

Due to insufficient resolution for the ordinal variable "EDUCATION", it was not considered in this PCA.

3.2 Selection of Principal Components

There were a few considerations when deciding the number of Principal Components to extract. The Eigenvalue Criterion, which suggests considering only components with values greater than 1, indicates that four components should be extracted. This is also supported by the Scree Test Criterion shown below, where the elbow is at the 4th component. Furthermore, the selected variables are able to explain about 65% of the variability which is consistent with heuristics on variance criterion where explained variability of 60% to 65% is considered acceptable. As such, the first four principal components were selected.

Eigenvalues

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	6.4979	32.490		32.490
2	4.0912	20.456		52.946
3	1.5045	7.522		60.468
4	1.0031	5.015		65.483
5	0.9244	4.622		70.106
6	0.8894	4.447		74.553
7	0.8747	4.373		78.926
8	0.7847	3.924		82.850
9	0.7337	3.668		86.518
10	0.7034	3.517		90.035
11	0.5907	2.953		92.989
12	0.3994	1.997		94.986
13	0.2632	1.316		96.302
14	0.2517	1.258		97.560
15	0.1927	0.963		98.524
16	0.1351	0.675		99.199
17	0.0705	0.352		99.551
18	0.0410	0.205		99.756
19	0.0254	0.127		99.883
20	0.0233	0.117		100.000

Figure 2: Eigenvalues and Cumulative Percentages

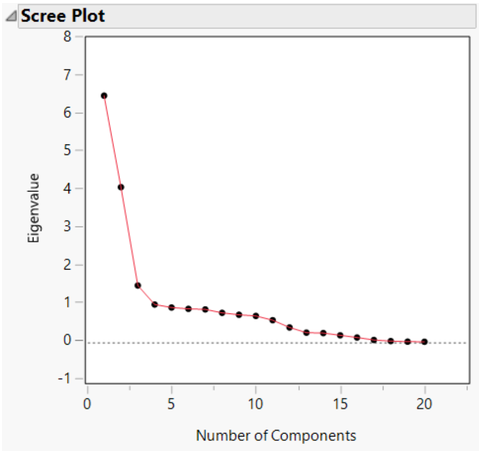

Figure 3: Scree Plot

3.3 Naming of Principal Components

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10
LIMIT_BAL	0.07520	0.31010	0.05618	0.19348	0.13537	0.01051	0.06369	-0.06486	-0.02726	-0.34485
AGE	0.01571	0.06513	-0.02836	0.96549	-0.13565	-0.01974	-0.03798	0.04089	0.00304	0.08793
PAY_0	0.16212	-0.30625	0.00561	0.03657	0.09405	-0.00986	0.01191	0.01465	-0.01597	0.53603
PAY_2	0.18858	-0.33910	0.04328	0.04344	0.11299	-0.00642	0.01822	0.00622	-0.01231	0.36630
PAY_3	0.19477	-0.34871	0.10003	0.04625	0.05332	0.02470	0.01720	0.02654	0.11183	0.07915
PAY_4	0.20180	-0.34983	0.13843	0.05227	0.00392	0.04581	-0.01927	0.04397	-0.00545	-0.23728
PAY_5	0.20592	-0.33686	0.15770	0.04149	-0.04719	0.01268	-0.01680	-0.08945	-0.02970	-0.38159
PAY_6	0.20155	-0.31055	0.15534	0.03232	-0.06421	-0.07744	0.03586	-0.03584	-0.04516	-0.41759
BILL_AMT1	0.33651	0.13674	-0.22459	-0.01912	0.05106	0.02971	0.05073	-0.00450	-0.00670	0.07019
BILL_AMT2	0.34809	0.13669	-0.19365	-0.03160	-0.01928	0.07902	0.04180	0.04294	0.13578	0.05360
BILL_AMT3	0.35218	0.14234	-0.13039	-0.04196	-0.06394	0.13726	-0.07763	0.12307	-0.09957	0.02820
BILL_AMT4	0.35641	0.14104	-0.11786	-0.04011	-0.09293	0.03272	-0.01678	-0.12487	-0.03271	0.00228
BILL_AMT5	0.35428	0.13906	-0.09944	-0.04062	-0.06353	-0.12586	0.06777	0.00523	-0.04795	-0.02994
BILL_AMT6	0.34695	0.13729	-0.08984	-0.03235	0.05922	-0.14979	-0.10146	-0.01387	0.00457	-0.04546
PAY_AMT1	0.09426	0.14964	0.40083	-0.05271	-0.33190	0.25501	-0.00393	0.24172	0.73083	0.05302
PAY_AMT2	0.07887	0.14031	0.43777	-0.05966	-0.29128	0.24610	-0.37550	0.27272	-0.59375	0.11545
PAY_AMT3	0.08658	0.15203	0.39763	-0.02490	-0.23752	-0.25525	0.04031	-0.77514	-0.00888	0.19294
PAY_AMT4	0.07779	0.14192	0.33666	-0.01953	0.10048	-0.67027	0.37388	0.45369	-0.07278	0.05617
PAY_AMT5	0.07533	0.13781	0.26208	0.04343	0.68801	-0.07731	-0.58002	-0.04503	0.18586	-0.00572
PAY_AMT6	0.06991	0.12991	0.29451	0.05359	0.41754	0.52838	0.58943	-0.08728	-0.15821	0.03424

	Prin11	Prin12	Prin13	Prin14	Prin15	Prin16	Prin17	Prin18	Prin19	Prin20
LIMIT_BAL	0.84226	0.01038	0.03651	-0.00354	0.00546	0.00624	-0.00903	-0.01464	-0.00111	0.00250
AGE	-0.18111	0.00200	0.00011	-0.00732	-0.00068	-0.00493	0.00020	0.00121	0.00007	0.00111
PAY_0	0.29723	0.62445	-0.27914	-0.01599	0.15428	-0.03708	-0.00578	-0.00041	-0.00284	0.00050
PAY_2	0.20630	-0.26707	0.56102	-0.01310	-0.49800	0.14258	-0.02998	-0.01189	0.00322	-0.00022
PAY_3	0.12911	-0.55346	-0.05182	-0.16199	0.59865	-0.30550	0.02463	0.02767	-0.00831	-0.00480
PAY_4	0.02120	-0.20729	-0.58300	0.02392	-0.21910	0.56727	0.01919	-0.00167	0.01386	-0.00146
PAY_5	-0.05955	0.15747	-0.09781	0.13959	-0.38477	-0.67353	0.02635	-0.02292	-0.00808	-0.00061
PAY_6	-0.11853	0.35912	0.49138	0.09552	0.38847	0.31965	-0.04232	0.01260	0.00006	0.00446
BILL_AMT1	-0.02982	-0.06666	-0.00140	0.55186	0.05266	0.01311	0.41450	0.43256	-0.18402	-0.31715
BILL_AMT2	-0.03758	-0.04722	-0.00641	0.37121	0.05221	-0.00047	0.03953	-0.34385	0.33019	0.64584
BILL_AMT3	-0.04410	-0.03431	-0.01840	0.10889	0.03928	-0.01736	-0.48355	-0.49690	-0.08665	-0.52605
BILL_AMT4	-0.05837	0.00561	-0.03377	-0.20667	-0.04099	-0.00838	-0.52138	0.48855	-0.36261	0.34624
BILL_AMT5	-0.06877	0.04995	-0.00006	-0.41284	-0.05846	-0.01687	0.06845	0.24972	0.71813	-0.22758
BILL_AMT6	-0.09252	0.06867	0.02413	-0.48002	-0.04913	0.02332	0.51419	-0.33820	-0.42654	0.07213
PAY_AMT1	0.03488	0.09310	0.05124	-0.04208	-0.06648	0.01889	0.04371	0.06756	-0.04481	-0.08473
PAY_AMT2	0.04853	-0.03668	0.03287	0.03469	0.04630	-0.02241	0.14692	0.07088	0.03837	0.12464
PAY_AMT3	-0.02574	-0.08023	-0.05574	0.12484	0.02915	0.05168	0.00112	-0.12373	0.02636	-0.06331
PAY_AMT4	-0.05503	-0.04166	-0.05895	0.12042	-0.02346	-0.04875	-0.11217	-0.00247	-0.08130	0.04224
PAY_AMT5	-0.15057	0.03061	0.00385	0.06632	0.01677	0.00061	-0.10066	0.06968	0.09510	-0.00844
PAY_AMT6	-0.20323	0.03674	0.00176	-0.09405	0.00488	0.00034	0.03481	-0.02735	-0.01727	0.00832

Figure 4: Eigenvector of each Principal Component

3.3.1 Prin1

Based on the Eigenvectors of the Principal Components shown in Figure 4, the linear combination of each principal component was derived. The first principal component was named as "Cumulative 6 months amount billed" because BILL_AMT1 to BILL_AMT6, which correspond to the billed amount for each of the last 6 months, has a higher loading as compared to other variables. The formula of Prin1 is given by:

$$
\begin{aligned}
Prin1 = {} & 0.07520LIMIT_{BAL} + 0.01571AGE + 0.16212PAY_0 + 0.18858PAY_2 + 0.19477PAY_3 \\
& + 0.2018PAY_4 + 0.20592PAY_5 + 0.20155PAY_6 + 0.33651BILL_{AMT1} \\
& + 0.34809BILL_{AMT2} + 0.35218BILL_{AMT3} + 0.35641BILL_{AMT4} \\
& + 0.35428BILL_{AMT5} + 0.34695BILL_{AMT6} + 0.09426PAY_{AMT1} + 0.07887PAY_{AMT2} \\
& + 0.08658PAY_{AMT3} + 0.07779PAY_{AMT4} + 0.07533PAY_{AMT5} + 0.06991PAY_{AMT6}
\end{aligned}
$$

3.3.2 Prin2

For the second principal component, LIMIT_BAL (credit limit), PAY_0 and PAY_2 to PAY_6 (payment status) has a higher loading as compared to the other variables. Since these are factors which are taken into consideration during credit limit reviews, this principal component was named as "Credit worthiness score".

It should be noted that since PAY_0 and PAY_2 to PAY_6 are ordinal in nature, and a positive value represents poor payment history, this means that negative values in these variables implies a better credit status of the customer.

The formula of Prin2 is given by:

$$\begin{aligned} Prin2 = {} & 0.31010LIMIT_{BAL} + 0.06513AGE - 0.30625PAY_0 - 0.3391PAY_2 - 0.34871PAY_3 \\ & - 0.34983PAY_4 - 0.33686PAY_5 - 0.31055PAY_6 + 0.13674BILL_{AMT1} \\ & + 0.13669BILL_{AMT2} + 0.14234BILL_{AMT3} + 0.14104BILL_{AMT4} \\ & + 0.13906BILL_{AMT5} + 0.13729BILL_{AMT6} + 0.14964PAY_{AMT1} + 0.14031PAY_{AMT2} \\ & + 0.15203PAY_{AMT3} + 0.14102PAY_{AMT4} + 0.13781PAY_{AMT5} + 0.12991PAY_{AMT6} \end{aligned}$$

3.3.3 Prin3

The third principal component was named as "Cumulative 6 months amount paid" because PAY_AMT1 to PAY_AMT6, which correspond to the payment amount over each of the last 6 months, has a higher loading as compared to other variables.

The formula of Prin3 is given by:

$$\begin{aligned} Prin3 = {} & 0.05618LIMIT_{BAL} - 0.02836AGE + 0.00561PAY_0 + 0.04328PAY_2 + 0.1003PAY_3 \\ & + 0.13843PAY_4 + 0.15770PAY_5 + 0.15534PAY_6 - 0.22459BILL_{AMT1} \\ & - 0.19365BILL_{AMT2} - 0.13039BILL_{AMT3} - 0.11786BILL_{AMT4} \\ & - 0.09944BILL_{AMT5} - 0.08984BILL_{AMT6} + 0.40083PAY_{AMT1} + 0.43777PAY_{AMT2} \\ & + 0.39763PAY_{AMT3} + 0.33666PAY_{AMT4} + 0.26208PAY_{AMT5} + 0.29451PAY_{AMT6} \end{aligned}$$

3.3.4 Prin4

The fourth principal component was named as "Age" because it has a higher loading as compared to other variables.

The formula of Prin4 is given by:

$$\begin{aligned} Prin4 = {} & 0.19348LIMIT_{BAL} + 0.96549AGE + 0.03657PAY_0 + 0.04344PAY_2 + 0.04625PAY_3 \\ & + 0.05227PAY_4 + 0.04149PAY_5 + 0.03232PAY_6 - 0.01912BILL_{AMT1} \\ & - 0.03160BILL_{AMT2} - 0.04196BILL_{AMT3} - 0.04011BILL_{AMT4} \\ & - 0.04062BILL_{AMT5} - 0.03235BILL_{AMT6} + 0.05271PAY_{AMT1} + 0.05966PAY_{AMT2} \\ & + 0.02490PAY_{AMT3} + 0.01953PAY_{AMT4} + 0.04343PAY_{AMT5} + 0.05359PAY_{AMT6} \end{aligned}$$

4. Cluster Analysis

4.1 Clustering Method

Cluster analysis was performed on the 4 principal components from PCA to profile customers into different segments based on distinct attributes. Hierarchical clustering was ruled out as the algorithm was not optimised to handle large dataset with more than thousands of records. K-Means was chosen instead due to its efficiency to handle bigger datasets, which allows quick comparison and optimisation on the number of clusters.

4.2 Number of Clusters

Model Summary

Algorithm	K-Means
Inputs	4
Clusters	4

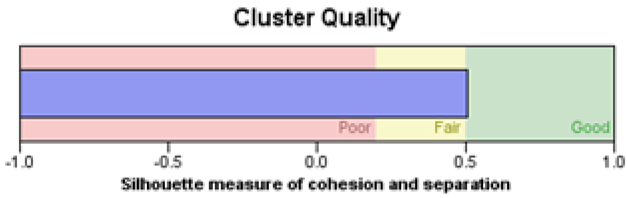

Figure 5: Cluster Quality for k=4 clusters

Upon running the stream in SPSS, it was found that 4 was the ideal number of clusters as it had the best silhouette measure of 0.5. A high silhouette measure tells us that the clusters are well spaced apart, with the elements within each cluster relatively close to each other with respect to the elements from other clusters. This can be summed up in the equation $(\boldsymbol{D}_{ave} - \boldsymbol{d}_{ave}) / \mathrm{Max}(\boldsymbol{D}_{ave}, \boldsymbol{d}_{ave})$.

4.3 Characteristics of Clusters

From Figure 6, Cluster 1 (53.6% of the sample) has a lower than average amount billed in the last 6 months (Prin1), lower than average credit worthiness score (Prin2), average amount paid in the last 6 months (Prin3) and a younger than average age group (Prin4). These characteristics matches typical young working adults who just entered the workforce. The lower credit scores are likely due to a lower disposable income and little personal assets. In addition, it is likely that they spend less on big ticket items with their credit cards at this stage of their lives, which translates to lower amounts billed and paid to the bank. This group of customers should be slowly grown as their net value will likely increase with time.

Clusters

Input (Predictor) Importance

1.0 0.8 0.6 0.4 0.2 0.0

Cluster	cluster-1	cluster-4	cluster-3	cluster-2
Label				
Description				
Size	53.6% (15901)	34.0% (10106)	12.4% (3671)	0.0% (5)
Inputs	Prin1 -0.80	Prin1 -0.53	Prin1 4.92	Prin1 20.01
	Prin2 -0.26	Prin2 -0.11	Prin2 1.40	Prin2 22.29
	Prin3 0.03	Prin3 0.09	Prin3 -0.42	Prin3 38.37
	Prin4 -0.65	Prin4 1.11	Prin4 -0.22	Prin4 -4.00

Figure 6: Centroid of Clusters

Cluster 2 (with only 5 elements) appears to be an extreme outlier group consisting of ultra-rich youngsters with much higher than average credit worthiness scores and lower average age. They are also very financial active with high bill amount and payment amount and should be maintained to ensure they continue their business with the bank.

Cluster 3 (12.4% of the sample) are high earners and spenders who are relatively young, have above average credit worthiness scores with high bill amounts, and lowest repayment to the bank. This group of customers appears to have the highest risk of defaulting as the incurred expense and repayment does not commensurate each other. Care should be placed on this group of customers to ensure their net debit does not spiral out of control.

Cluster 4 (34% of the sample) can be generalised as older working adults when comparing their credit worthiness scores, age, average bill and paid amounts against Cluster 1. This is a group of stable clients where their incurred expenses and payment exceeds those of Cluster 1 but remains relatively consistent. This group of customers should be grown as they may have other assets which can be capitalised on.

5. Logistic Regression

5.1 Model Building

The categorical variables not used in PCA such as EDUCATION, MARIAGE and SEX, along with the 4 principal components were used to develop the logistic regression model. The P Values (Figure 7) of the model suggest that all the variables used are significant (i.e. P<0.05). The Chi Square results (Figure 8) from the parameter estimates further provides resolution on the significance level for each variable.

Effect Summary

Source	LogWorth	PValue
Prin2	636.500	0.00000
Prin1	39.693	0.00000
Prin3	12.595	0.00000
EDUCATION	8.491	0.00000
MARRIAGE	7.635	0.00000
Prin4	3.399	0.00040
SEX	3.213	0.00061

Remove Add Edit FDR

Figure 7 : Effect Summary

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	1.45610173	0.0516108	795.98	<.0001*
Prin1	-0.0961593	0.0070291	187.15	<.0001*
Prin2	0.44594522	0.0100053	1986.6	<.0001*
Prin3	0.16685973	0.0242289	47.43	<.0001*
Prin4	-0.061333	0.0172881	12.59	0.0004*
SEX[1]	-0.0533027	0.0155347	11.77	0.0006*
EDUCATION[2-1]	0.0703309	0.035618	3.90	0.0483*
EDUCATION[3-2]	0.04031018	0.0428494	0.88	0.3468
EDUCATION[4-3]	0.91534329	0.1874529	23.84	<.0001*
MARRIAGE[1]	-0.1321767	0.0468178	7.97	0.0048*
MARRIAGE[2]	0.07339257	0.0481782	2.32	0.1277

Figure 8 : Parameter Estimates

5.2 Goodness of Fit

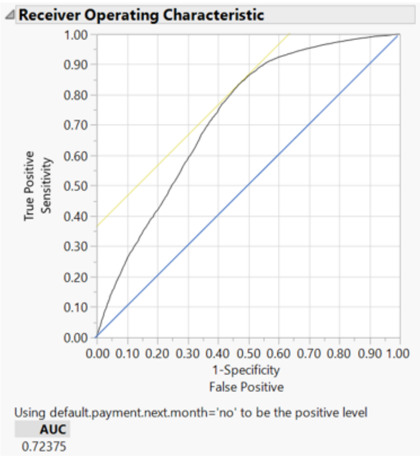

Figure 9: ROC Curve

From Figure 9, it was observed that the Receiver Operating Characteristic (ROC) curve has a positive predicted response as it is above the 45^{o} diagonal line. The Area Under the Curve (AUC) is ~0.72 which suggest the model is a relatively good fit for the data set.

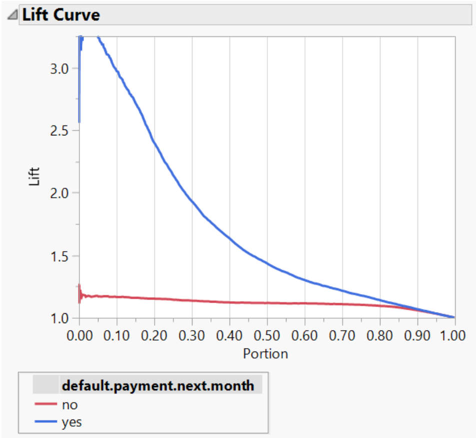

Figure 10 : Lift Curve

In addition, the lift curve gives a lift of 3.0 for the first 10% of the data set. This means the model demonstrate 3 times more responses than if the data were selected at random. This will be useful to compare model accuracy modeled from the same dataset in future studies.

5.3 Model Accuracy

Confusion Matrix

Training

Actual default.payment.next.month	Predicted Count no	yes
no	22785	579
yes	5097	1222

Figure 11 : Confusion Matrix

Accuracy	80.88%
True Positive Rate/Sensitivity or Recall	19.34%
False Positive Rate/ Type I error	2.48%
True Negative rate / Specificity	97.52%
False Negative rate/ Type II error	80.66%

Figure 12 : Model attributes

From Figure 11 and 12, it is noted that while model accuracy and specificity is relatively high at 80.88% and 97.52% respectively, sensitivity is very low at 19.34%. This means that while the model is capable of predicting non-defaulter, it is unable to do the same for defaulters. The low sensitivity also corresponds to a high type II error, which means that the model is unable to differentiate defaulters from non-defaulters. Hence, more data (with more variables) would likely be needed to improve model sensitivity.

5.4 Model Equation

From the model building process, the equation of the logistic regression model developed is shown in Figure 13.

$$
\begin{aligned}
&1.4561017338 \\
&+ -0.096159286 \cdot Prin1 \\
&+ 0.4459452241 \cdot Prin2 \\
&+ 0.1668597275 \cdot Prin3 \\
&+ -0.061333046 \cdot Prin4 \\
&+ \mathrm{Match}\left(SEX \right) \begin{pmatrix} 1 & \Rightarrow -0.053302716 \\ 2 & \Rightarrow 0.0533027159 \\ \text{else} & \Rightarrow . \end{pmatrix} \\
&+ \mathrm{Match}\left(EDUCATION \right) \begin{pmatrix} 1 & \Rightarrow 0 \\ 2 & \Rightarrow 0.0703308976 \\ 3 & \Rightarrow 0.1106410734 \\ 4 & \Rightarrow 1.0259843678 \\ \text{else} & \Rightarrow . \end{pmatrix} \\
&+ \mathrm{Match}\left(MARRIAGE \right) \begin{pmatrix} 1 & \Rightarrow -0.132176682 \\ 2 & \Rightarrow 0.0733925746 \\ 3 & \Rightarrow 0.058784107 \\ \text{else} & \Rightarrow . \end{pmatrix}
\end{aligned}
$$

Figure 13: Equation of Logistic Regression Model

6. Conclusion

In meeting the Bank's requirement to accurately predict the default rate of existing clients, a logistic regression model was proposed. PCA was first performed to simplify modelling and remove multi-collinearity in the dataset. Thereafter, K-means clustering was used, and 4 customer segments were identified for targeted marketing activities for the bank. With the logistic regression model, it was found that while it is capable of predicting non-defaulters, it is poor at detecting actual defaulters (or differentiate them from non-defaulters).

To improve the regression model, it is recommended that data with variables beyond this study be considered and incorporated into the modelling of an updated regression model. Furthermore, once a model with acceptable sensitivity is obtained, regression models can also be built for each customer segment identified from clusters analysis earlier for even better prediction capability.

Data Dictionary

Variable Name	Explanation
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=divorced, 0=others)
AGE	Age in years
PAY_0	Repayment status as on September 2005 (-2= No Consumption, -1=pay duly, 0=uses revolving credit 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)
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)