Credit Card Default Research

by Sumaira Afzal, Viraja Ketkar, Murlidhar Loka, Vadim Spirkov

Abstract Credit card default might very well be a life altering event. It happens when a client have become severely delinquent on his/her credit card payment. It's a serious credit card status that not only affects person's standing with that credit card issuer, but also individual's credit standing in general and his/her ability to get approved for credit cards, loans, and other credit-based services. This research will make yet another attempt to predict if a client goint to default on the next payment. Employing verious machine learning technique we also will make an attemt to estime the amount a client would be able to pay when the bill comes. The authors of this study will try to discover who is more likely to default on the payment.

Background

Overdepandance on credit card debt has been an ongoing theme in many countries around the word. For example US consumers started 2018 owing more than $1 trillion in credit card debt (Ref: Comoreanu). It is projected that by the end of 2019 US consumers will increase their collective debt by another 60 billion dollars. Unfortunately many consumers overestimate their ability to pay the debt on time, or the unforeseen circumstances and luck of savings make people default on their payments. This is the least desirable outcome for all parties. Unpaid debt leads, in most cases, to default on the whole outstanding balance causing financial loss for the credit institutions. Majority of the clients go through tremendous emotional and financial stress, risking their credibility. The financial institution make significant efforts to evaluate the prospective client ability to sustain the debt and pay in time to avoid the credit default.

Objective

This study pursues two goals. The main objective is to predict as accurate as possible if the client makes the next month payment or defaults. We will employ CRISP-DM methodology (Ref: Jiawei Han (2012)) and supervised learning approach, to achieve this goal.

We are also motivated to profile, if possible, the customer base. The questions we will try to address are:

- If there are specific groups of cardholders that share similar features or behavioral patterns
- How theses groups, if exist, pay their debt

To achieve the second goal we will use unsupervised learning approach.

Data Analysis

This research employs the data set sourced from UCI Machine Learning Repository. This real-life data comprises 30000 observations of the credit card payment history of Taiwanese consumers.

Data Dictionary

Column Name	Column Description
ID	Customer ID
LIMIT_BAL	Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit
SEX	Gender (1 = male; 2 = female).
EDUCATION	Education (1 = graduate school; 2 = university; 3 = high school; 4 = others)
MARRIAGE	Marital status (1 = married; 2 = single; 3 = divorced; 0 - other)
AGE	Age (year)

Column Name	Column Description
PAY_1	PAY_1 - PAY_6 are payment statuses over a course of the last six months, where -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. Positive numbers denote payment delay in months. For example 1 = payment delay for one month; 2 = payment delay for two months; . . .; 9 = payment delay for nine months and above. PAY_1 - Payment status in September
PAY_2	Payment status in August
PAY_3	Payment status in July
PAY_4	Payment status in June
PAY_5	Payment status in May
PAY_6	Payment status in April
BILL_AMT1	BILL_AMT1 - BILL_AMT6 are bill amounts (NT dollar) from April till September. BILL_AMT1: September bill
BILL_AMT2	August bill
BILL_AMT3	July bill
BILL_AMT4	June bill
BILL_AMT5	May bill
BILL_AMT6	April bill
PAY_AMT1	Amount of previous payment (NT dollar). PAY_AMT1: paid in September (August bill)
PAY_AMT2	Amount paid in August (July bill)
PAY_AMT3	Amount paid in July (June bill)
PAY_AMT4	Amount paid in June (May bill)
PAY_AMT5	Amount paid in May (April bill)
PAY_AMT6	Amount paid in April (March bill)
DEFAULT	**Target label that denotes whether the client paid the next month bill (0) or did not (1)**

Statistics

We start our research with the feature exploration and understanding.

Table 2: Credit Card Payment Data Summary

No	Variable	Stats / Values	Freqs (% of Valid)	Missing
1	ID [integer]	Mean (sd) : 15000.5 (8660.4) min < med < max: 1 < 15000.5 < 30000 IQR (CV) : 14999.5 (0.6)	30000 distinct values (Integer sequence)	0 (0%)
2	LIMIT_BAL [integer]	Mean (sd) : 167484.3 (129747.7) min < med < max: 10000 < 140000 < 1e+06 IQR (CV) : 190000 (0.8)	81 distinct values	0 (0%)
3	SEX [integer]	Min : 1 Mean : 1.6 Max : 2	1 : 11888 (39.6%) 2 : 18112 (60.4%)	0 (0%)
4	EDUCATION [integer]	Mean (sd) : 1.9 (0.8) min < med < max: 0 < 2 < 6 IQR (CV) : 1 (0.4)	7 distinct values	0 (0%)
5	MARRIAGE [integer]	Mean (sd) : 1.6 (0.5) min < med < max: 0 < 2 < 3 IQR (CV) : 1 (0.3)	4 distinct values	0 (0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Missing
6	AGE [integer]	Mean (sd) : 35.5 (9.2) min < med < max: 21 < 34 < 79 IQR (CV) : 13 (0.3)	56 distinct values	0 (0%)
7	PAY_1 [integer]	Mean (sd) : 0 (1.1) min < med < max: -2 < 0 < 8 IQR (CV) : 1 (-67.3)	11 distinct values	0 (0%)
8	PAY_2 [integer]	Mean (sd) : -0.1 (1.2) min < med < max: -2 < 0 < 8 IQR (CV) : 1 (-8.9)	11 distinct values	0 (0%)
9	PAY_3 [integer]	Mean (sd) : -0.2 (1.2) min < med < max: -2 < 0 < 8 IQR (CV) : 1 (-7.2)	11 distinct values	0 (0%)
10	PAY_4 [integer]	Mean (sd) : -0.2 (1.2) min < med < max: -2 < 0 < 8 IQR (CV) : 1 (-5.3)	11 distinct values	0 (0%)
11	PAY_5 [integer]	Mean (sd) : -0.3 (1.1) min < med < max: -2 < 0 < 8 IQR (CV) : 1 (-4.3)	10 distinct values	0 (0%)
12	PAY_6 [integer]	Mean (sd) : -0.3 (1.1) min < med < max: -2 < 0 < 8 IQR (CV) : 1 (-4)	10 distinct values	0 (0%)
13	BILL_AMT1 [integer]	Mean (sd) : 51223.3 (73635.9) min < med < max: -165580 < 22381.5 < 964511 IQR (CV) : 63532.2 (1.4)	22723 distinct values	0 (0%)
14	BILL_AMT2 [integer]	Mean (sd) : 49179.1 (71173.8) min < med < max: -69777 < 21200 < 983931 IQR (CV) : 61021.5 (1.4)	22346 distinct values	0 (0%)
15	BILL_AMT3 [integer]	Mean (sd) : 47013.2 (69349.4) min < med < max: -157264 < 20088.5 < 1664089 IQR (CV) : 57498.5 (1.5)	22026 distinct values	0 (0%)
16	BILL_AMT4 [integer]	Mean (sd) : 43262.9 (64332.9) min < med < max: -170000 < 19052 < 891586 IQR (CV) : 52179.2 (1.5)	21548 distinct values	0 (0%)
17	BILL_AMT5 [integer]	Mean (sd) : 40311.4 (60797.2) min < med < max: -81334 < 18104.5 < 927171 IQR (CV) : 48427.5 (1.5)	21010 distinct values	0 (0%)
18	BILL_AMT6 [integer]	Mean (sd) : 38871.8 (59554.1) min < med < max: -339603 < 17071 < 961664 IQR (CV) : 47942.2 (1.5)	20604 distinct values	0 (0%)
19	PAY_AMT1 [integer]	Mean (sd) : 5663.6 (16563.3) min < med < max: 0 < 2100 < 873552 IQR (CV) : 4006 (2.9)	7943 distinct values	0 (0%)
20	PAY_AMT2 [integer]	Mean (sd) : 5921.2 (23040.9) min < med < max: 0 < 2009 < 1684259 IQR (CV) : 4167 (3.9)	7899 distinct values	0 (0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Missing
21	PAY_AMT3 [integer]	Mean (sd) : 5225.7 (17607) min < med < max: 0 < 1800 < 896040 IQR (CV) : 4115 (3.4)	7518 distinct values	0 (0%)
22	PAY_AMT4 [integer]	Mean (sd) : 4826.1 (15666.2) min < med < max: 0 < 1500 < 621000 IQR (CV) : 3717.2 (3.2)	6937 distinct values	0 (0%)
23	PAY_AMT5 [integer]	Mean (sd) : 4799.4 (15278.3) min < med < max: 0 < 1500 < 426529 IQR (CV) : 3779 (3.2)	6897 distinct values	0 (0%)
24	PAY_AMT6 [integer]	Mean (sd) : 5215.5 (17777.5) min < med < max: 0 < 1500 < 528666 IQR (CV) : 3882.2 (3.4)	6939 distinct values	0 (0%)
25	DEFAULT [integer]	Min : 0 Mean : 0.2 Max : 1	0 : 23364 (77.9%) 1 : 6636 (22.1%)	0 (0%)

Table 2 describes main statistical parameters of each column. It also outputs the values of the binary features. The first thing that jumps at us is that the data set has no missing data! We shall note that our target feature is not balanced. Almost **80%** of the clients do pay on time. Secondly female customers make **60%** of the data set. **Customer ID** column, as usual, will be dropped since it presents no analytical value. Here is a look at the data sample.

ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_1	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3
1	20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689
2	120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682
3	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559
4	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291
5	50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835
6	50000	1	1	2	37	0	0	0	0	0	0	64400	57069	57608
7	500000	1	1	2	29	0	0	0	0	0	0	367965	412023	445007
8	100000	2	2	2	23	0	-1	-1	0	0	-1	11876	380	601
9	140000	2	3	1	28	0	0	2	0	0	0	11285	14096	12108
10	20000	1	3	2	35	-2	-2	-2	-2	-1	-1	0	0	0
11	200000	2	3	2	34	0	0	2	0	0	-1	11073	9787	5535
12	260000	2	1	2	51	-1	-1	-1	-1	-1	2	12261	21670	9966

BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5
3913	3102	689	0	0	0	0	689	0	0	0
2682	1725	2682	3272	3455	3261	0	1000	1000	1000	0
29239	14027	13559	14331	14948	15549	1518	1500	1000	1000	1000
46990	48233	49291	28314	28959	29547	2000	2019	1200	1100	1069
8617	5670	35835	20940	19146	19131	2000	36681	10000	9000	689
64400	57069	57608	19394	19619	20024	2500	1815	657	1000	1000
367965	412023	445007	542653	483003	473944	55000	40000	38000	20239	13750
11876	380	601	221	-159	567	380	601	0	581	1687
11285	14096	12108	12211	11793	3719	3329	0	432	1000	1000
0	0	0	0	13007	13912	0	0	0	13007	1122
11073	9787	5535	2513	1828	3731	2306	12	50	300	3738
12261	21670	9966	8517	22287	13668	21818	9966	8583	22301	0

Table 3: Credit Card Payment Data Sample

Let's review demographic characteristics of the customer base, namely: *EDUCATION*, *MARITAL STATUS* and *AGE*. We immediately can observe some deficiencies in the data quality (Figure: 1). As we see the majority of the credit card holders have a university degree. There are three groups which are not supposed to be in the data set: **Unknown** - code **0**, **Unknown5** - code **5** and **Unkown6** - code **6**. We will assign these customers to the **Other** group, since the description for the aforementioned codes is not provided.

Number of single people is slightly higher than the number of the married ones.

Majority of the credit card holders are people between age of 25 and 50, which does not come as a surprise (Figure: 1)... Let's see if the *AGE* feature has outliers.

```
print(original %>% filter(AGE < 18 || AGE > 100) %>% summarise(COUNT = n()))
```

```
  COUNT
1     0
```

The *AGE* feature maintains perfect data.

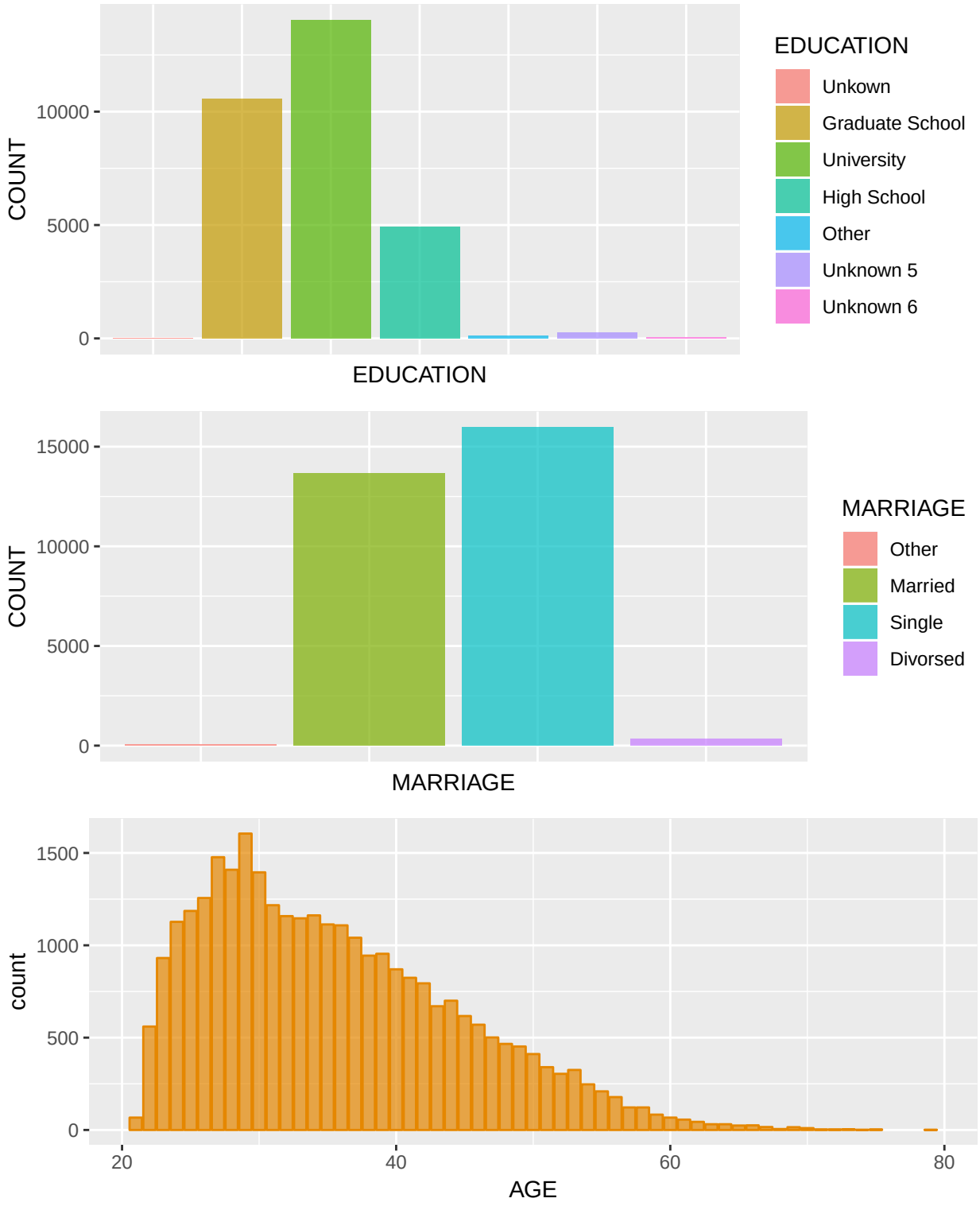

Figure 1: Customer Demographics

The next group of features we are going to explore is payment statuses. As per the data dictionary the payment statuses are supposed to have the status codes in the following range: **-2 : 9**. Let's verify the data integrity of the features.

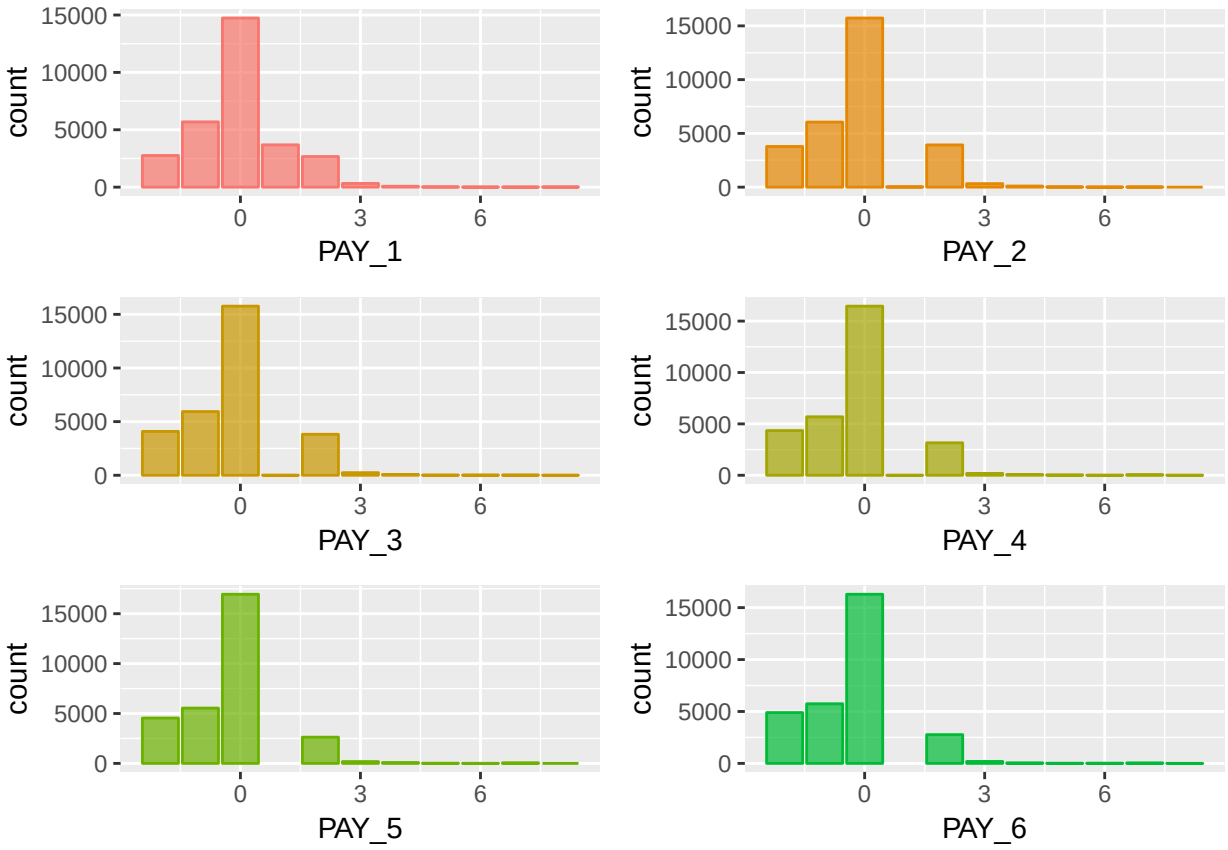

Figure 2: Customer Payment Statuses for the Last Six Months

As we have already noticed many of the credit card holders pay duly, codes: *-2 and -1* (see Figure: 2). Majority of the customers do maintain good standing. Noticeably though they **paid the required minimum or more but not the full balance** (code *0*). There is a rather significant group that falls behind with the payment by one or two months. The next group of features are the bill amounts for the last six months.

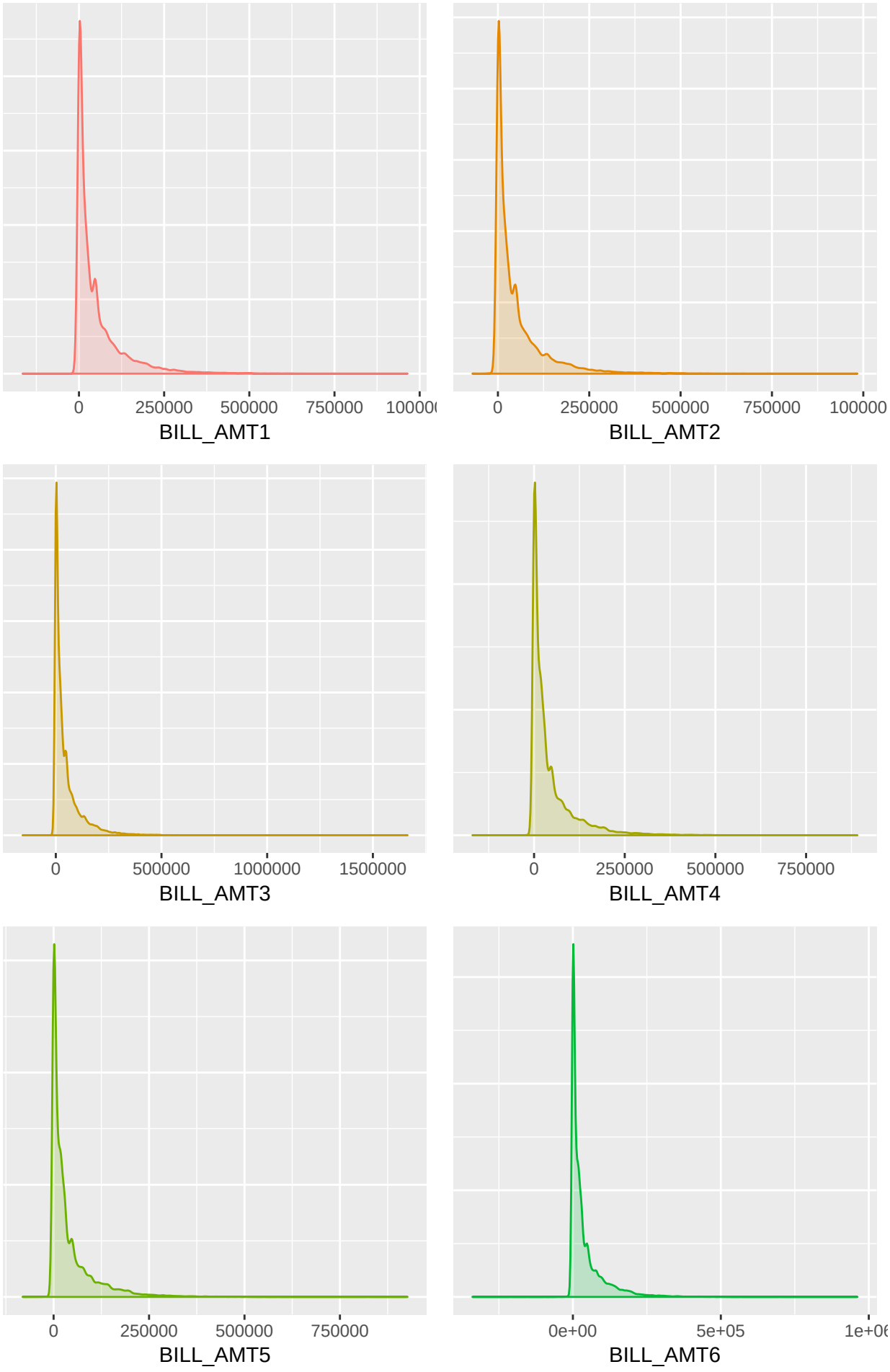

Figure 3: Customer Bill Amount Distribution for the Last Six Months

The bill payment amounts have negative values, significant amounts that reach at times **thens of thousands** of NT dollars! The negative amount on the credit card bill statements happen when a card holder overpaid his/her bill or were issued a credit after he/she already paid the bill.

Noticeably the bill amounts have very long tails. They average in tens of thousands of TN dollars, hovering around 50,000 dollars or so on average (see Table: 2). Thus it makes the negative bill amounts we previously observed more plausible.

The last group of the features is the client monthly payments. The data maintained in those columns appears to be integral (see Table: 2). Let's see how the customer payments are distributed. We employ normal and log-scaled visualization for better presentation.

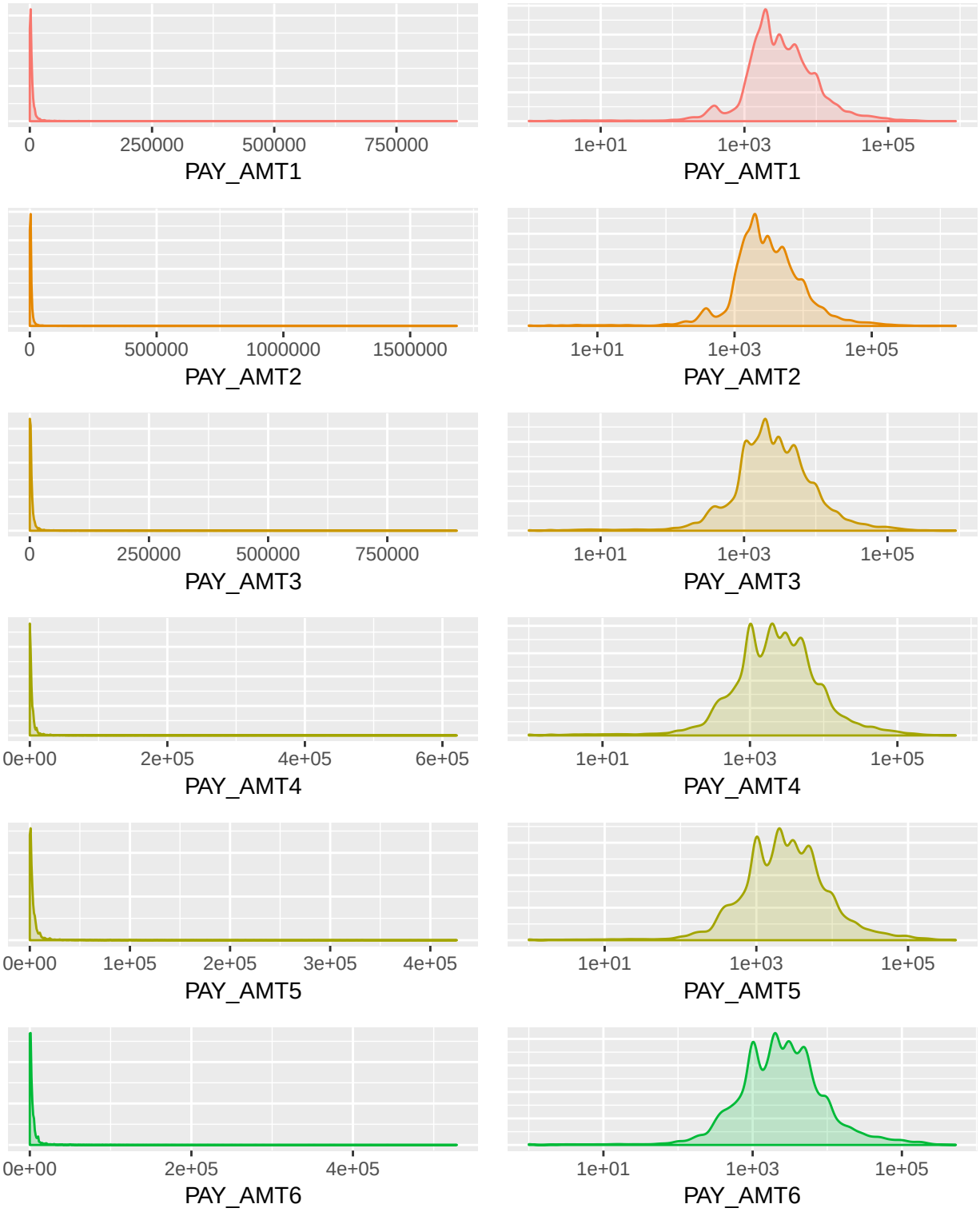

Figure 4: Customer Payment Amount Distribution for the Last Six Months

The pay amounts mirror in the distribution the bill amounts, which is expected. The charts have very long tails which imply that the amounts the card holders pay, very greatly. Most likely the payments that are way outside of the normal distribution curve are lump sum payments. More often than not the clients pay between 1000 and 10000 dollars monthly, which is still way below the average bill amount. This finding and the payment status statistics (see Figure: 2) make us believe that the majority of the credit card holders do have quite significant debt, despite the good standing. This hypotheses also explains the distribution of the payments. To keep the debt growth in check the customers pay lump sums whenever they accumulate some saving. Let's plot the delta between the

bill amounts and the payment amounts to support our theory. Figure 5

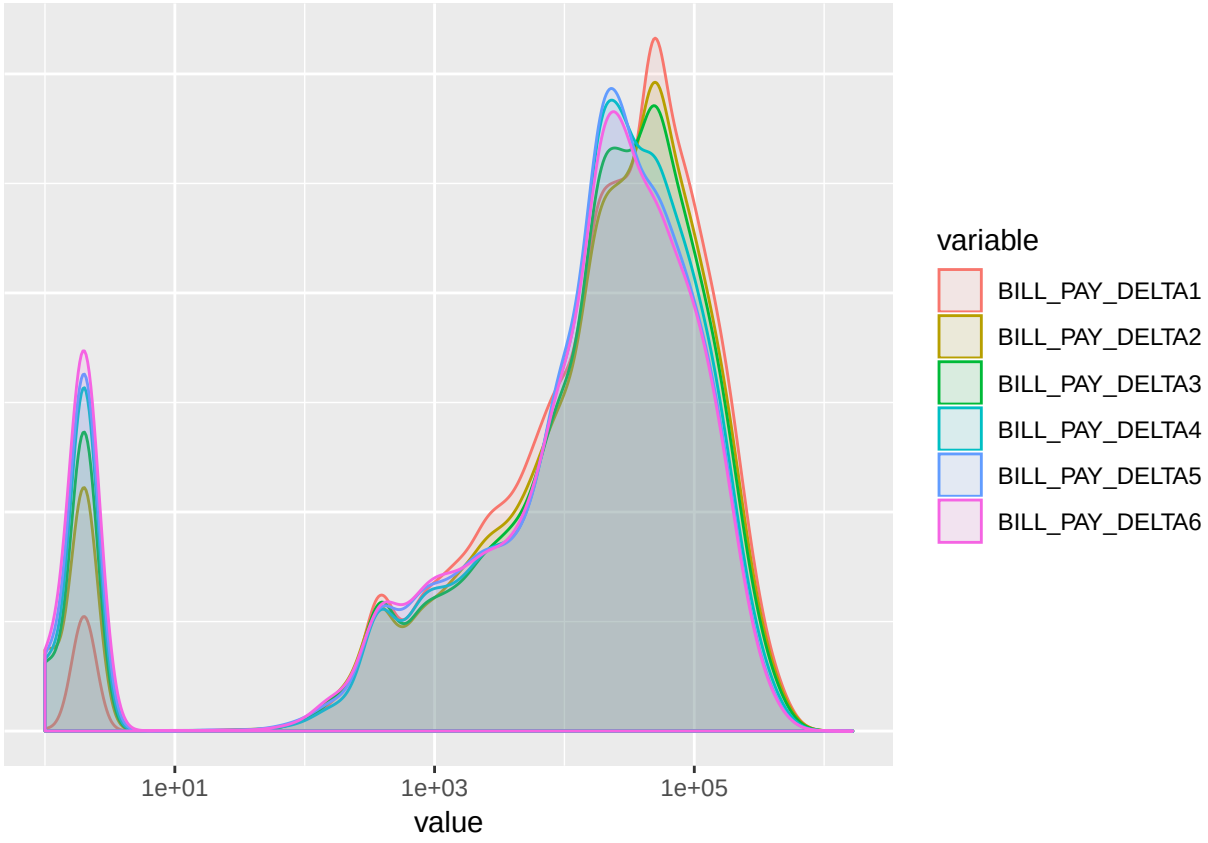

Figure 5: Customer Bill/Payment Amount Delta for Six Months

Data Transformation

Before we proceed further we are going to clean the data set as described in the previous paragraph, namely:

- We will remove *Customer ID* column
- We assign code **4 - Other** to the *EDUCATION* column values that fall out of the declared code range (1:4)

```
original$EDUCATION = with(original, ifelse(EDUCATION == 0 | EDUCATION == 5 | EDUCATION == 6 , 4, EDUCATION))
data = dplyr::select(original, -ID)
data %>% filter(EDUCATION == 0 | EDUCATION >4) %>% summarise(COUNT=n())
```

```
  COUNT
1     0
```

Data Correlation and Principal Component Analysis

In this section of our study we continue exploring the relations between various features of the data set. We put stress on finding the correlated features, the correlation between the features and the target label. We also are going to apply principal component analysis (PCA) to understand which attributes of the data set explain most variance of the original data (Ref: Kassambara (2017)). If our findings are fruitful we may design a model that requires smaller number of the input parameters without sacrificing the predictive power of the model.

Let's plot the correlation matrix first.

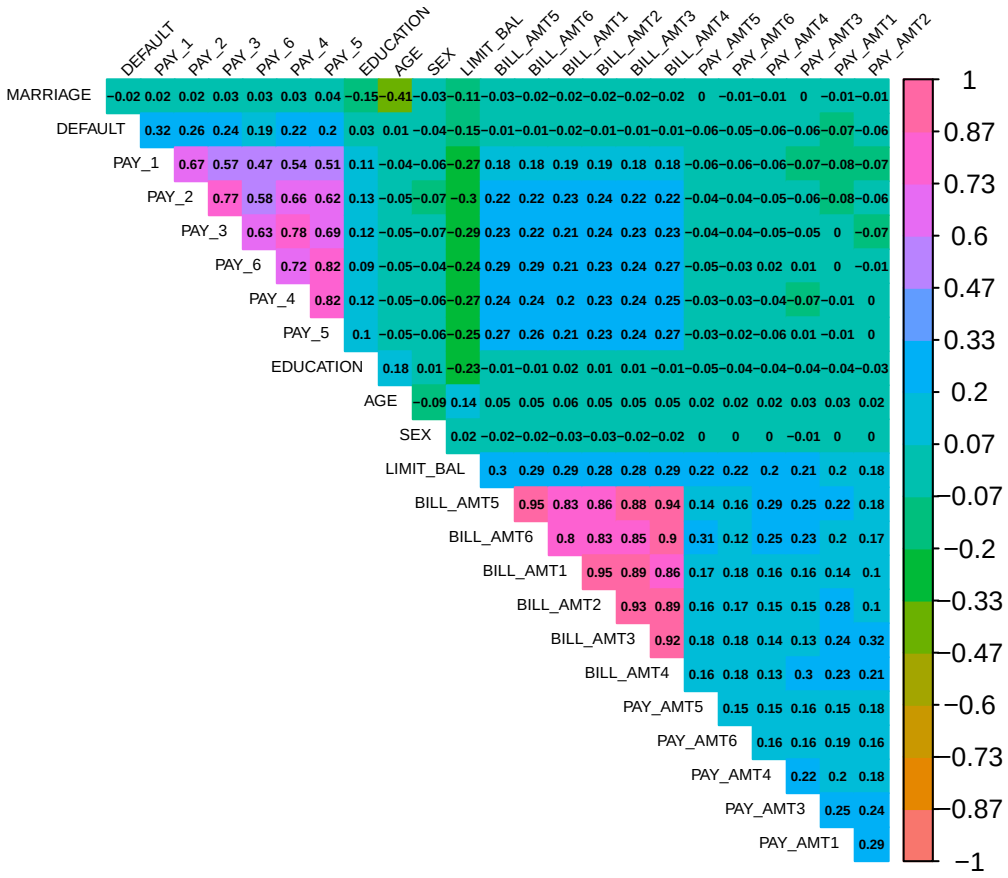

Figure 6: Data Correlation

The correlation matrix does not yield any surprises (see Figure: 6)). Bill payment amounts exhibit higher correlation as well as the payment status group. This does not give us much. The target label has no correlation with any other feature. We proceed with the PCA analysis now. We scale and center the data to achieve meaningful result. We also remove the target feature from the PCA computation. We are going to retain the 15 top components.

```
       eigenvalue variance.percent cumulative.variance.percent
Dim.1  6.54287678       28.4472904                    28.44729
Dim.2  4.10332133       17.8405275                    46.28782
Dim.3  1.55513542        6.7614583                    53.04928
Dim.4  1.47549730        6.4152057                    59.46448
Dim.5  1.02455764        4.4545984                    63.91908
Dim.6  0.95545872        4.1541683                    68.07325
Dim.7  0.90407409        3.9307569                    72.00401
Dim.8  0.88793791        3.8605996                    75.86461
Dim.9  0.87030881        3.7839514                    79.64856
Dim.10 0.78268998        3.4029999                    83.05156
Dim.11 0.73278811        3.1860353                    86.23759
Dim.12 0.68195482        2.9650210                    89.20261
Dim.13 0.57091319        2.4822313                    91.68484
Dim.14 0.51977719        2.2599008                    93.94474
Dim.15 0.40359547        1.7547629                    95.69951
Dim.16 0.25988392        1.1299301                    96.82944
Dim.17 0.24930179        1.0839208                    97.91336
Dim.18 0.18869146        0.8203976                    98.73376
Dim.19 0.13178740        0.5729887                    99.30674
Dim.20 0.07015088        0.3050038                    99.61175
Dim.21 0.04078476        0.1773250                    99.78907
Dim.22 0.02529347        0.1099716                    99.89905
Dim.23 0.02321955        0.1009546                   100.00000
```

Good news! The **top 10 components explain 83%** of the data variance. Let's review what the top

four components are made of.

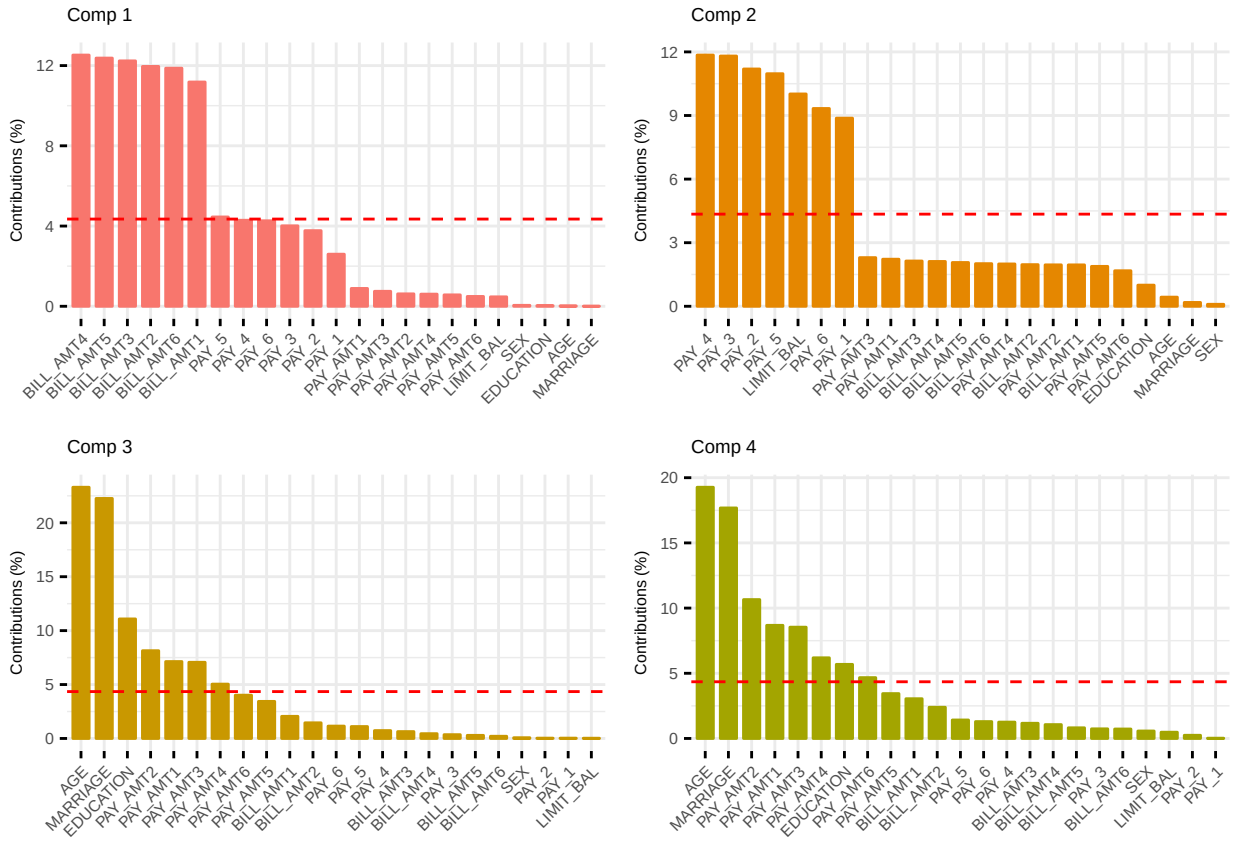

Figure 7: Feature Contribution to the Top Four Components

The red dashed line on the graph (see Figure: 7) indicates the expected average contribution. If the contribution of the variables were uniform, the expected value would about 4.3%. For a given component, a variable with a contribution larger than this cutoff could be considered as important in contributing to the component. The PCA analysis supports the correlation matrix (Figure: 6) in a sense that the bill amounts and the pay statuses are highly correlated and are subject to the dimentionality reduction due to the redundancy. So as we can see the first component is largely being dominated by the bill amounts and pay statuses. The second one mainly includes the payment status features and the credit card limit. The demographic features and the amount paid dominate the third and fourth components.

It would be very interesting to see how the top two components look like in the context of the target label. Very well, Figure: 9 shows that there is no clear separation between component one and two. As the previous chart highlighted (see Figure: 7) this shall be expected since the top two components largely comprise bill payments and pay statuses. We also observe two linear formations: one is located in the first quadrant; green (no default). The second formation is located in the fourth quadrant; it peppered with the red color that denotes the default green. to understand better what those formations are we will plot the feature vectors in the context of the component one and two.

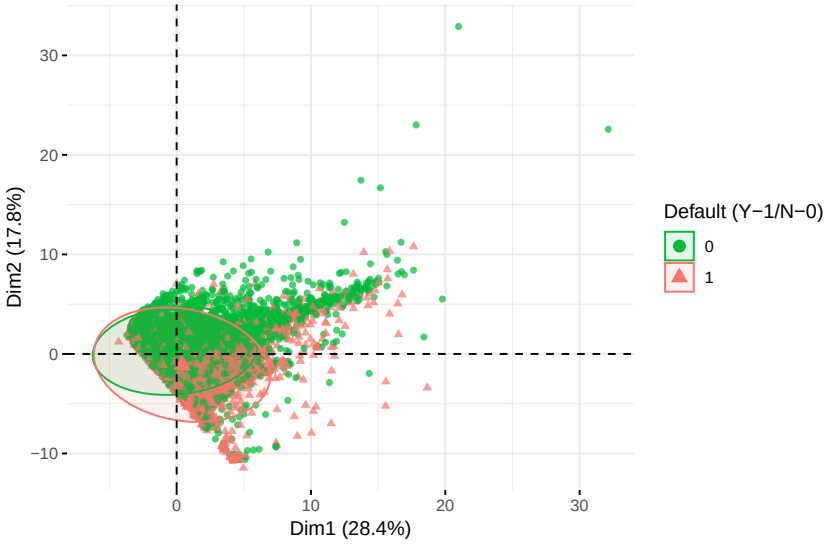

Figure 8: Feature contribution to the Top Two Componenets

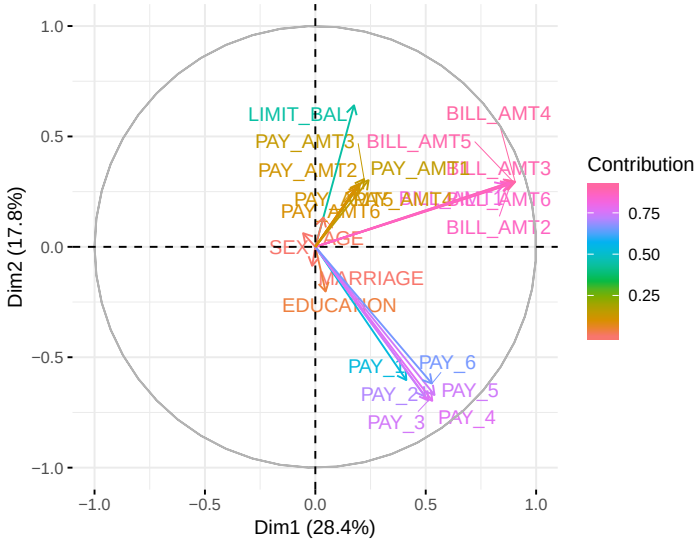

Figure 9: Components and Feature Contribution in the Context of Target Label

The plot submitted above clearly shows that the biggest contributor to the first quadrant is a bill amount group. And the contributors to the fourth quadrant are payments statuses.

Takeaways from Data Exploration Excersize

We have concluded the data exploration study. There are a few major points we would like to highlight. They are:

- The majority of the credit card holders maintain good standing (see Figure: 2).
- The clients do maintain debt. On regular basis they pay the amount that meets the minimal required payment but less then the bill amount (see Figure 5).
- The clients tend to reduce the amount of debt paying the outstanding balance in lump sums (Figure: 4).
- Demographically married and single people represented almost equally, women are represented better than men. Majority of the credit card holders have a university degree. The vast majority of the clients are between 25 and 50 year old.

- The data set is not balanced; the number of the customers in a good standing is much higher than the number of people who defaulted on the next payment.
- Principal component analysis showed that we can reduce the number of the features to 10 - 12 sustaining the data variance coverage at about **85%**. The top principal components comprise mainly bill amounts and pay statuses (Figure: 7). The components do not have clear separation and affect equally the target label (Figure: 8).

Classification Models and Model Performance Evalutation

In this section we will try reach our main goal - predict if the cardholder is going to default on the next payment or not.We will evaluate three supervised learning approaches: Naive Bayes, Random Forest and Logistic Regression (Ref: Max Kuhn). Each model will be cross-validated employing k-fold technique. We will analyse the ROC curve and confusion matrix of each model.

In addition to the said above we compare performance of the logistic regression algorithm on the whole data set and the data set that comprises 10-top components identified during PCA analysis.

In the end of this section we plot all tree model AUC stats.

Data Preparation

Prior to fitting the models with the data we would have to upsample the training set, because our data set is not balanced; the number of customers who defaults is much smaller than the number of the paying customers.

All models we are going to fit will benefit from the data scaling and centering, thus we are going to apply these transformations. Below numbers are the count of the deaults on the balanced data set.

```
    0     1
16340 16340
```

Feature Selection

We will be using all features of the data set. In the case of the logistic regression we will examine the algorithm performance on the PCA-reduced data set.

Naive Bayes Model

Naïve Bayes classification is a kind of simple probabilistic classification methods based on Bayes' theorem with the assumption of independence between features.

It is simple (both intuitively and computationally), fast, performs well with small amounts of training data, and scales well to large data sets. The greatest weakness of the Naïve Bayes classifier is that it relies on an often-faulty assumption of equally important and independent features which results in biased posterior probabilities. Although this assumption is rarely met, in practice, this algorithm works surprisingly well and accurate; however, on average it rarely can compete with the accuracy of advanced tree-based methods (random forests & gradient boosting machines) but is definitely worth having in our toolkit.

```
Naive Bayes

32680 samples
   23 predictor
    2 classes: 'no', 'yes'

Pre-processing: Box-Cox transformation (3), centered (23), scaled (23)
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 26144, 26144, 26144, 26144, 26144
Resampling results across tuning parameters:

  usekernel  ROC        Sens       Spec
  FALSE      0.7388411  0.3465116  0.8517748
   TRUE      0.7578969  0.9400245  0.3251530
```

```
Tuning parameter 'fL' was held constant at a value of 0
Tuning
 parameter 'adjust' was held constant at a value of 1
ROC was used to select the optimal model using the largest value.
The final values used for the model were fL = 0, usekernel = TRUE
 and adjust = 1.

Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  6664 1419
       yes  360  557

               Accuracy : 0.8023
                 95% CI : (0.794, 0.8105)
    No Information Rate : 0.7804
    P-Value [Acc > NIR] : 2.052e-07

                  Kappa : 0.2856
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9487
            Specificity : 0.2819
         Pos Pred Value : 0.8244
         Neg Pred Value : 0.6074
             Prevalence : 0.7804
         Detection Rate : 0.7404
   Detection Prevalence : 0.8981
      Balanced Accuracy : 0.6153

       'Positive' Class : no
```

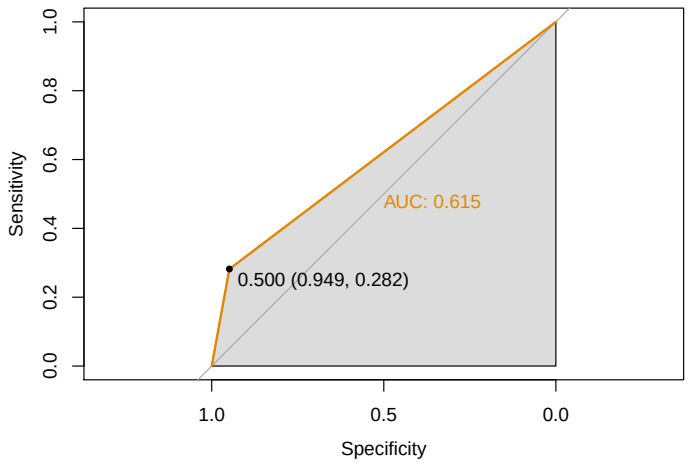

Figure 10: Naive Bayes Model AUC and ROC Curve

Random Forest Model

Random Forest is also considered as a very handy and easy to use algorithm, because it's default hyper parameters often produce a good prediction result. Random Forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model. The main limitation of Random Forest is that a large number of trees can make the algorithm to slow and ineffective for real-time predictions. In general, these algorithms are fast to train, but quite slow to create predictions once they are trained.

```
Random Forest

32680 samples
   23 predictor
    2 classes: 'no', 'yes'

Pre-processing: Box-Cox transformation (3), centered (23), scaled (23)
Resampling: Cross-Validated (3 fold)
Summary of sample sizes: 21787, 21786, 21787
Resampling results across tuning parameters:

  mtry  ROC        Sens       Spec
   2    0.9661183  0.8700126  0.9525092
  12    0.9699316  0.8876380  0.9504895
  23    0.9692160  0.8826198  0.9504283

ROC was used to select the optimal model using the largest value.
The final value used for the model was mtry = 12.

Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  6428 1132
       yes  596  844

               Accuracy : 0.808
                 95% CI : (0.7997, 0.8161)
    No Information Rate : 0.7804
    P-Value [Acc > NIR] : 7.726e-11

                  Kappa : 0.3792
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9151
            Specificity : 0.4271
         Pos Pred Value : 0.8503
         Neg Pred Value : 0.5861
             Prevalence : 0.7804
         Detection Rate : 0.7142
   Detection Prevalence : 0.8400
      Balanced Accuracy : 0.6711

       'Positive' Class : no
```

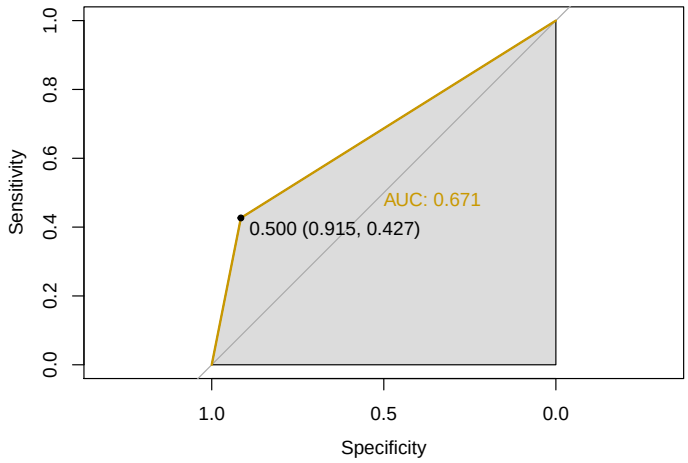

Figure 11: Random Forest Model AUC and ROC Curve

Logistic Regression Model

Logistic regression is an efficient, interpretable and accurate method, which fits quickly with minimal tuning. Logistic regression prediction accuracy will benefit if the data is close to Gaussian distribution. Thus we apply addition transformation to the training data set. We will also be employing 5-fold cross-validation re-sampling procedure to improve the model.

```
Generalized Linear Model

32680 samples
   23 predictor
    2 classes: 'no', 'yes'

Pre-processing: Box-Cox transformation (3), centered (23), scaled (23)
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 26144, 26144, 26144, 26144, 26144
Resampling results:

  Accuracy   Kappa
  0.6767748  0.3535496

Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  4935  711
       yes 2089 1265

               Accuracy : 0.6889
                 95% CI : (0.6792, 0.6984)
    No Information Rate : 0.7804
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2741
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7026
            Specificity : 0.6402
         Pos Pred Value : 0.8741
         Neg Pred Value : 0.3772
             Prevalence : 0.7804
         Detection Rate : 0.5483
   Detection Prevalence : 0.6273
      Balanced Accuracy : 0.6714
```

```
'Positive' Class : no
```

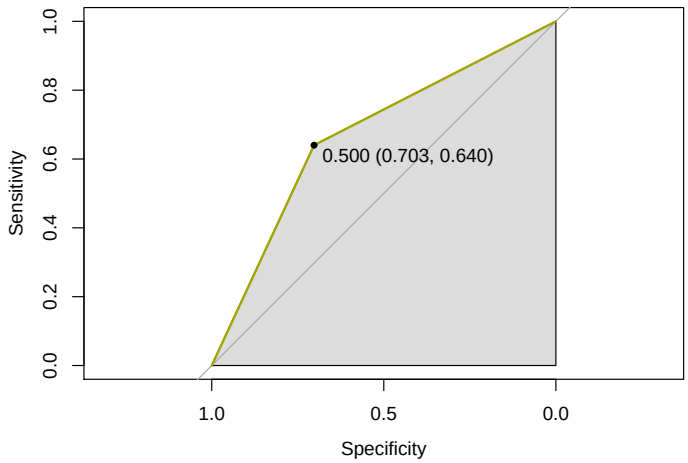

Figure 12: Logistic Regression Model AUC and ROC Curve

Confusion matrix and Figure 12 demonstrate the logistic model performance on the balanced data set. Using the proportion of positive data points that are correctly considered as positive (true positives) and the proportion of negative data points that are mistakenly considered as positive (false negative), we generated a graphic that shows the trade off between the rate at which the model correctly predicts the rain tomorrow with the rate of incorrectly predicting the rain. The value around 0.67 indicates that the model does a good job in discriminating between the two categories.

Let's now compare the performance of the PCA components to the performance of the full data set. Just to remind we selected have identified 10 top components that explain *83%* of data variance. We are going to reduce the dimentionality of the training and test data set multiplying them on the principal component weight matrix.

```
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  3718  619
       yes 3306 1357

               Accuracy : 0.5639
                 95% CI : (0.5536, 0.5742)
    No Information Rate : 0.7804
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1451
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.5293
            Specificity : 0.6867
         Pos Pred Value : 0.8573
         Neg Pred Value : 0.2910
             Prevalence : 0.7804
         Detection Rate : 0.4131
   Detection Prevalence : 0.4819
      Balanced Accuracy : 0.6080

       'Positive' Class : no
```

Well the result of the model fitted with the PCA components is inferior to the same very model. The balanced accuracy of the data set with lower dimentionality dropped by about **7%**. But we shall

not forget that the PCA components data set had less than a half of the features of the original data set. Tolerance of the reduction of the predictive power of the model depends on the business requirements. In some cases it might be Okay in others not. In our particular case we consider such precision drop is critical, thus we stick to the fully featured data set.

Model Comparison

Now it is time to compare the models side by side and pick a winner.

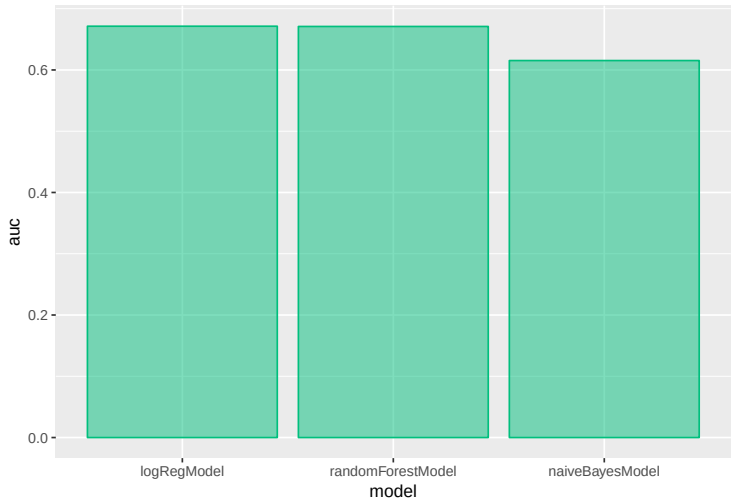

Figure 13: Model AUC Comparison

```
              model       auc
1       logRegModel 0.6713867
3 randomForestModel 0.6709549
2   naiveBayesModel 0.6153149
```

AUC - ROC perfomance

AUC stands for Area under the ROC Curve and ROC for Receiver operating characteristic curve. This is one of the most important KPIs of the classification algorithms. These two metrics measure how well the models distinguishing between the classes. The higher AUC the better model predicts positive and negative outcome.

Figures 10, 11, 12 and accompanying data show that on the test data set all the models demonstrated very close results. Random Forest has the highest overall accuracy (~80%) but performs poorly on the negative outcome - *paymant default* (43%), thus the balanced accuracy is lower (about 67%).

Naive Bayes model almost mirrors the random forest in the overall accuracy.But its specificity is even lower, 28%. The balanced accuracy of the model is about 62%.

Logistic regression model is the most blanaced one; it separates positive and negative outcomes equally well. It's balanced accuracy is **67%**.

Model Interpretibility

Logistic Regression and Naive Bayes are all highly interpretable models. It is easy to explain to the business what impact each input parameter has. The decision tree could be visualized (provided if it is not too large).

Random Forest on the other hand is a black-box model, complex algorithm which is difficult to explain in simple terms.

Data Pre-Processing

Random Forest and Naive Bayes can deal with missing data, outliers, numeric and alphanumeric values. Simply speaking they are not very demanding for data quality. It would be interesting to

see how they perform on the original data set without data cleaning. But this is subject of another research...

Logistic regression does require conversion of alphanumeric values to numeric, struggles dealing with the outliers and performs best when fitted with the data that have normal distribution.

Verdict

Despite sensitivity to data quality Logistic Regression outperforms Naive Bayes and Random Forest models in the ability to separate positive and negative classes. It is fast, scales well and highly interpretable. Thus this is our winner.

Undertanding the Client Base Employing Unsupervised Learning

Lastly we believe it would be beneficial to profile the client base. This would add additional insights into understanding of the credit card holders demographics, spending and borrowing habits. We will be employing *Clastering Large Applications (CLARA)* approach to attack this challenge. This study does not make its goal to compare various unsupervised model techniques. *CLARA* has been chosen because it is robust, relatively easy to understand, scales well, handles categorical and continuous features and fast.

It is based on The **Partitioning Around Medoids** (*PAM*) algorithm, which is a popular realization of k-medoids clustering method. We employ the silhouette coefficient measure to gauge how well the clusters are separated. The silhouette analysis also provides insights into the cluster density.

After many trials and errors we have selected a few features that describe the client financial and demographic profile quite well, namely:

- Bill amount history
- Credit standing history
- Cardholder education

Figure 14 shows that the clusters are fairly well defined. There is a minor overlap between clusters 2, 3 and 4.

```
  cluster size ave.sil.width
1       1  205          0.30
2       2   70          0.21
3       3  178          0.30
4       4   47          0.19
```

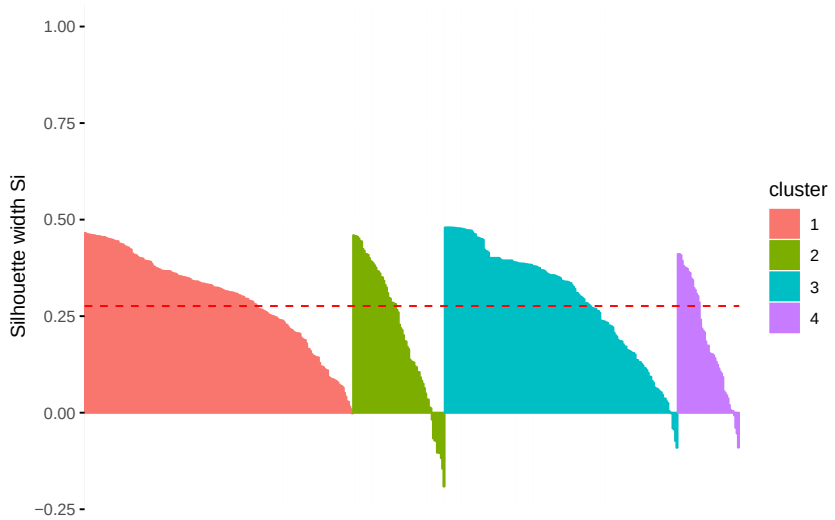

Figure 14: Cluster Silhouette

The plot 15 renders the clusters shape against the first two dimensions.

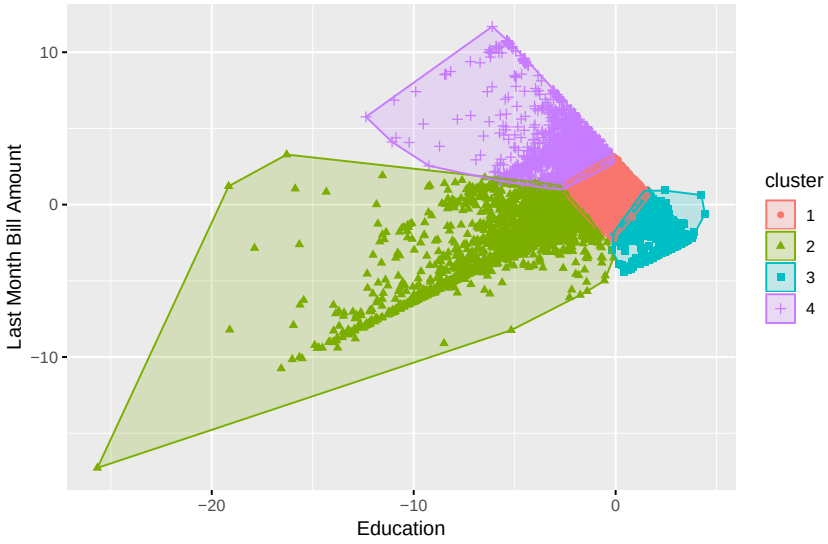

Figure 15: Clusters

So what are those clusters are. How to interpret them? The next few paragraphs visualize and describe the cluster contents.

Cluster #1 - The Regular Folks

Nothing particular stands out about this group. Majority of the clients in this group pay required minimum. small percentage of the group delay the payments but no more than 2 months. Their monthly bills rarely go beyond 75,000 Taiwanese dollars. Some members of this cluster have high credit limit, but it does not look like they take advantage of it. They are rather well educated, range between 20 and 40 years of age. This is the largest category that describes pretty accurately the main mass of the cardholder - **43.7%**.

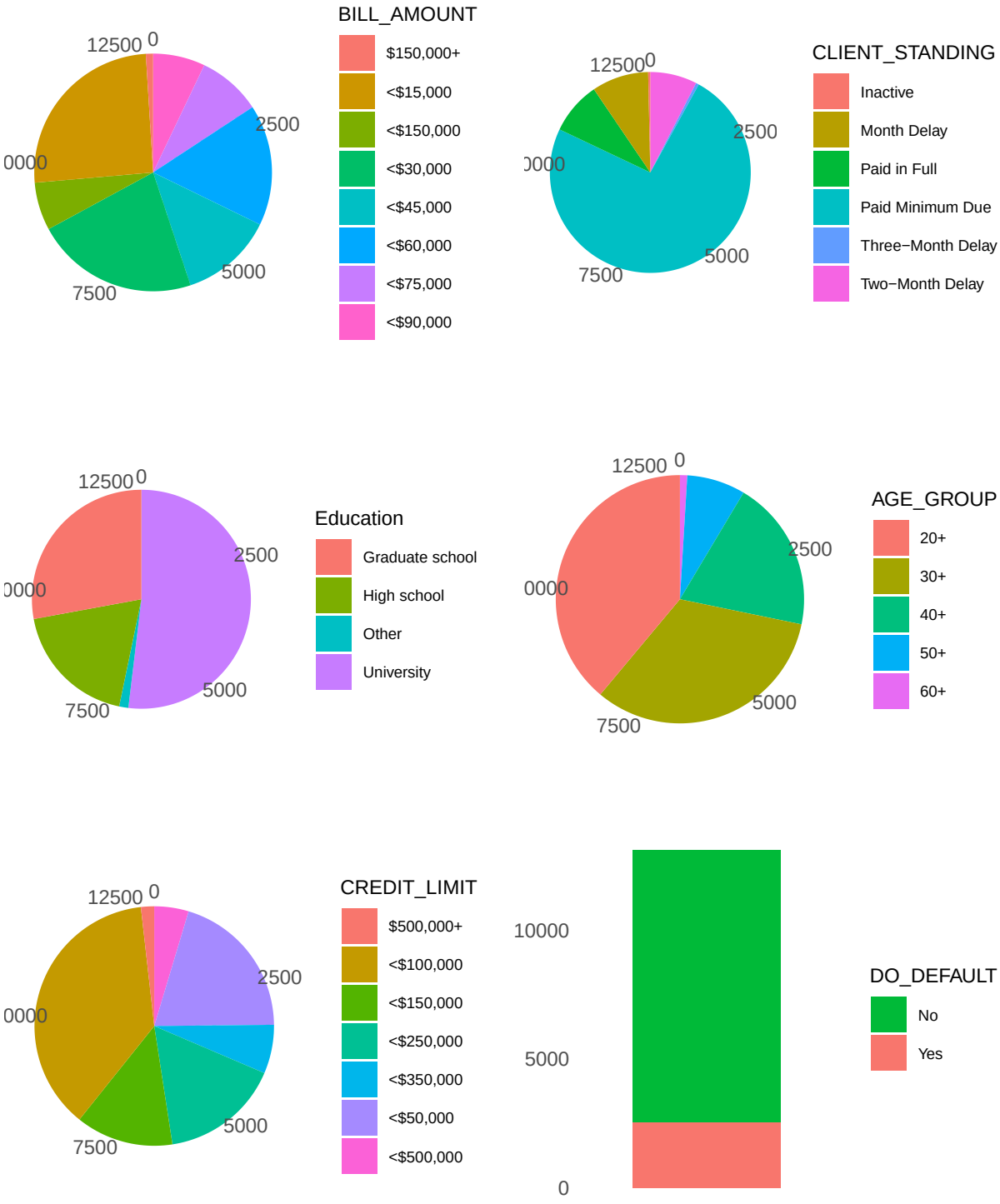

Figure 16: Cluster One - The Regular Folks

Cluster #2 - The Exuberant Spenders

People in this cluster live the life! Their monthly bills often go beyond $150,000 mark. Many group members have a credit limit in the range 250,000 or more. As the *Regular Folks* the cardholder of this class do not pay the full bill amount. There is a good sector of the clients that are 3 month behind with the payment. Education-wise the group almost equally split between the university and graduate school graduates. Just like in the case of the first cluster majority of the people in this group are between 20 and 40 years old, where **30+** age group dominates. This is the third largest group of the

credit card holder population - **14.8%**.

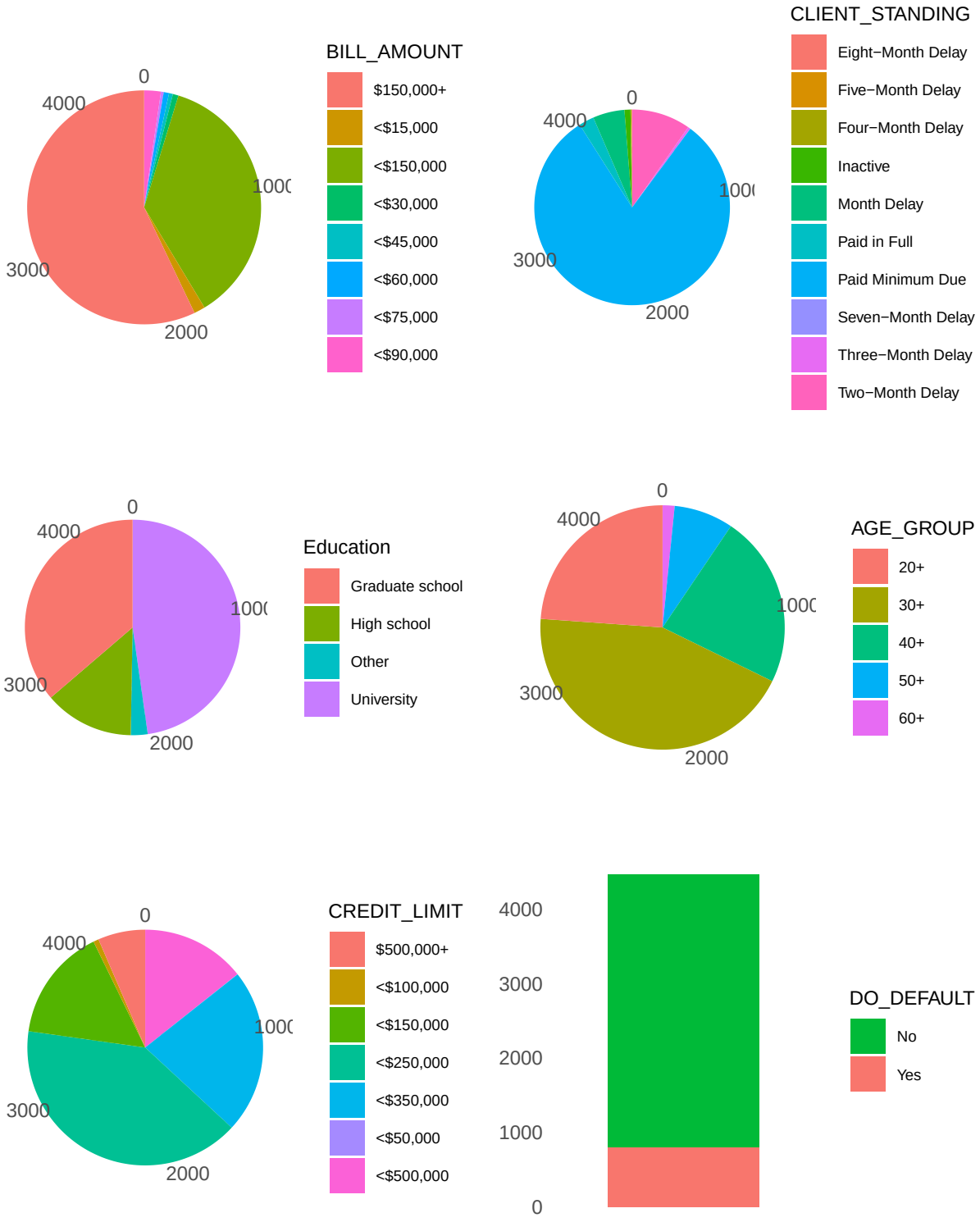

Figure 17: Cluster Two - The Exuberant Spenders

Cluster #3 - The Realists

The second largest group of the clients (**32.7%**) are people who really keep their spending in check. The vast majority of the people in this cluster have the bill amount less than $15,000 NT (about 650 CAD). Unlike the previous two groups almost half of the realists pay their bills in full. Only small percentage have delayed the payment by a month. The realists are generally older. Half of the group

finished the graduate school, the second largest education group are university graduates. The credit limit of the group varies and could be very high but they do not fall victims to seduction!

Figure 18: Cluster Three - The Realists

Cluster #4 - The Grinders

The striking difference between this cluster of people and the others is a default rate; it is about **65%**! The majority of the grinders have relatively low bills: $45,000 or less. Smaller credit limits, - less than 100,000. Yet, predominantly they are two or more month late with their payments. This is a

highly educated group, majority of which are university graduates. Wait a minute… Age-wise this is the youngest group of all, where over 40% a people in their twenties. Maybe they are still students? Thankfully this is the smallest group that make **8.6** of the total population of the cardholders.

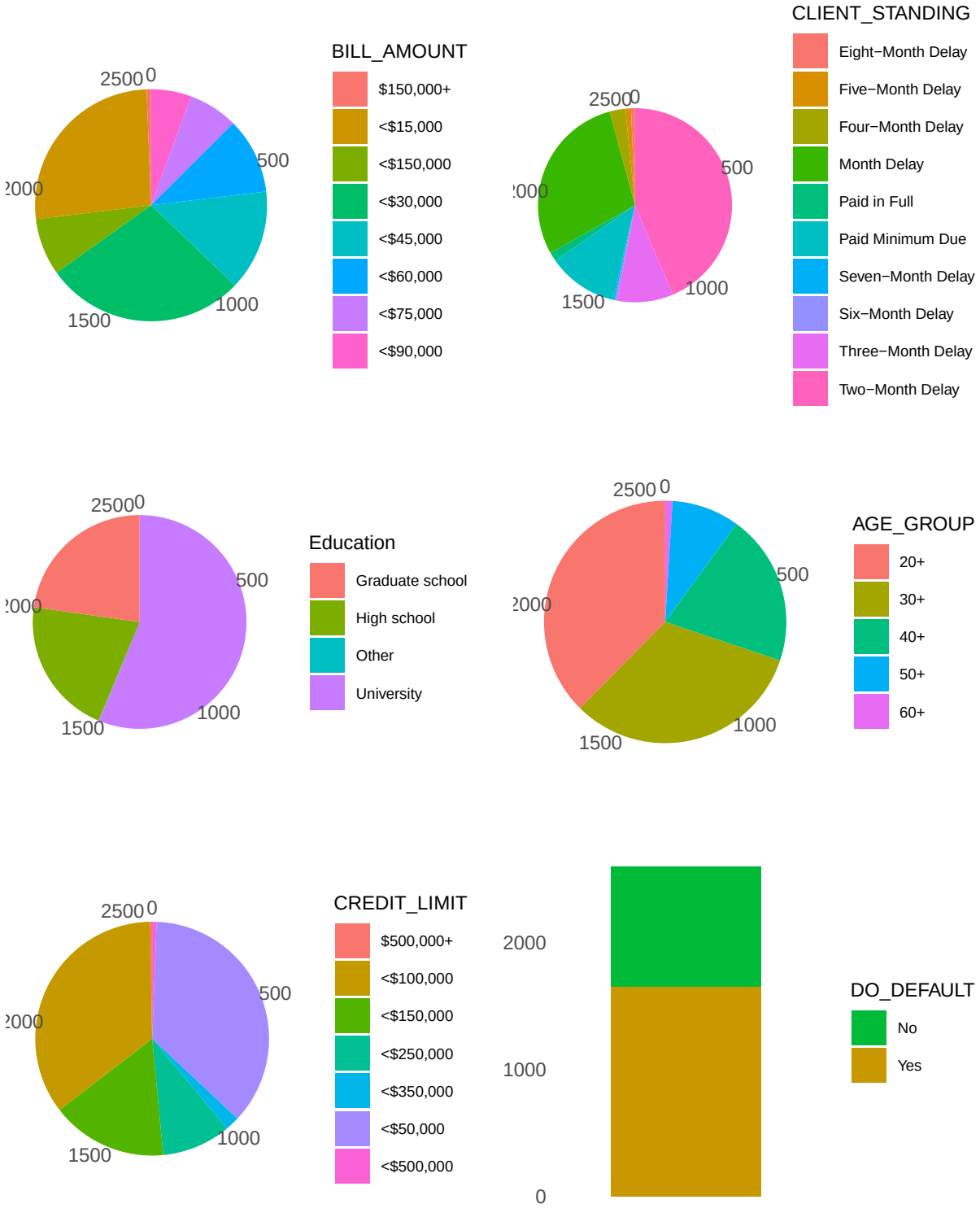

Figure 19: Cluster Four- The Grinders

Model Deployment

Our best classification model has predictive power of 68%. Is it good enough? It is hard to tell. We feel that the model meets our first objective, which is prediction of the default on the next month payment. Logistic regression model we picked is fast, equally sensitive to the positive and negative outcomes. It produces reasonably good result even on the lower dimension data sets. The model is fast and easy to deploy. Due to the nature of the business the model does require frequent data updates and re-training.

We are also very satisfied with accomplishment of our second goal, which is understanding of the customer base. We believe that we pretty accurately captured the main characteristics and behavioral patterns of the credit card holders. **CLARA** approach we employed provided very plausible result. This model is fast and easy to deploy. It does not require significant computing power to produce good result.

Conclusion

Through exploring credit card holders borrowing and spending patterns collected in 2006 in Taiwan we were able to come up with two models. One is a binary classifier, which predicts a default on the upcoming credit card bill. The second model provides in-depth view of the client base.

The study started with thorough analysis of the data set. At this phase we were able to identified many interesting patterns that insured the success of the whole project. We commenced our research providing descriptive stats on all available features of the data set. We also applied Principal Component Analysis approach to deduce which features carried the most information.

Then we applied and evaluated three supervised learning algorithms: Logistic Regression, Naive Bayes and Random Forest. The logistic regression model we tested with the whole feature set and with the component-based set, which we picked during the PCA phase. All three models were k-fold cross-validated and thoroughly evaluate employing ROC curve and confusion matrix approaches.

Lastly we applied unsupervised learning to understand who the credit card holders are, how they borrow and spend money, whether they are the clients in good standing or prone to default.

As a result of this study we fully understood the data we dealt with. We designed reasonably accurate credit card default prediction model. We managed to group the clients into meaningful, highly interpretable clusters that explain the customer behavioral patterns well.

Overall we believe we have achieved all our goals.

Bibliography

A. Comoreanu. Credit card debt study. trends and insights. URL https://wallethub.com/edu/credit-card-debt-study/24400/. [p1]

J. P. Jiawei Han, Micheline Kamber. *Data Mining. Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems, 225 Wyman Street, Waltham, MA 02451, USA, 2012. ISBN 978-0-12-381479-1. [p1]

A. Kassambara. *Practical Guide to Cluster Analysis in R*. CreateSpace Independent Publishing Platform, 2017. ISBN 91542462703. [p10]

S. W. Max Kuhn, Jed Wing. Caret (classification and regression training) project. URL https://cran.r-project.org/web/packages/caret/index.html. [p14]

Note from the Authors

This file was generated using *The R Journal* style article template, additional information on how to prepare articles for submission is here - Instructions for Authors. The article itself is an executable R Markdown file that could be downloaded from Github with all the necessary artifacts.

Sumaira Afzal
York University School of Continuing Studies

Viraja Ketkar
York University School of Continuing Studies

Murlidhar Loka
York University School of Continuing Studies

Vadim Spirkov
York University School of Continuing Studies