Universitat Politècnica de Catalunya

Multivariate Analysis

2^{nd} Semester

Credit Card Default Analysis

Author:
Ricard Gardella

Author:
Alexander Parunov

July 10, 2018

Contents

1 Introduction to the problem

Using data about the default of credit cards of real clients, our goal is to inspect, analyze and perform techniques learned in the Multivariate Analysis class in order to get hidden information and transform this information into knowledge that could be used in a real project. In addition to that at the end of a project we should be able to predict payment capability of a client, whether he/she can pay next month or not.

2 Previous work

There is a previous paper [YL09] that uses this data set to predict the default of clients. Authors tested six models and computed their accuracy. The paper evaluates the performance of the different methods using a train and test sets instead of cross-validation or any other technique. It uses two methods to evaluate algorithms, the error rate (related with accuracy) and the area ration of the lift curve because the data set is unbalanced.

The paper aims not only to classify the card issuers, but also to estimate the probability of default. To do that, authors try to obtain a true probability of default using the Sorting Smoothing Method (SSM). We are not going to attempt this second task in our project.

With respect to the results obtained in the paper, the best error rate that was achieved with k-nearest neighbours algorithm is 17%.

3 Data Used

The data used for this project is a about the default of credit cards in real clients. The data is provided by the ICS and the origin of this data is banks located in Taiwan. The data set can be found here:

https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

The data set contains 30000 rows and 24 columns. Those columns are mixture of categorical and numerical variables. It does not contain missing values but it contains errors that will be corrected during the pre-processing.

4 Structure of the data

The variables of the dataset are the following ones:

- ID: ID of each client
- LIMIT_BAL: Amount of given credit in NT dollars
- SEX
- EDUCATION:
- MARRIAGE: Marital status
- AGE
- PAY_X: Repayment status (-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). There are six variables of this kind (X = 1,2...,6), one for each month from April to September 2005.
- BILL_AMTX: Amount of bill statement in the X month (the same ones than before).
- PAY_AMTX: Amount of previous payment in month X.
- default.payment.next.month: this is the target variable. It indicates if the client decleared default or not.

The variable ID is not necessary as it does not give any relevant information. The variables SEX, EDUCATION, MARRIAGE, AGE and DEFAULT.PAYMENT. NEXT.MONTH are categorical variables.
The variables PAY_X, BILL_ATMX and pay_ATMX, where X goes from 1 to 6, represent the different months. The percentage of default in our data set is 22.12%. So, the data set that we have is unbalanced towards not being in default. This is problematic to estimate the accuracy of a classifier, because any classifier, even a random one, will have 78% of accuracy.
The currency in this dataset is NT$ which is the currency of taiwan.
1NT$ = 0.032893$.

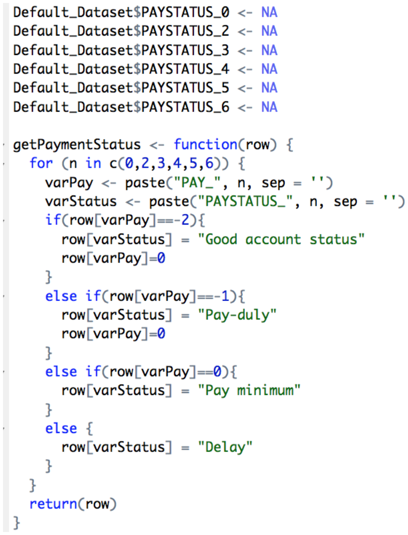
4.1 Pre-Processing

First of all, we checked if there are some missing values and we found out that there are none. However, we have detected wrong values in the variables MARRIAGE and EDUCATION. The wrong values were very few, so, we deleted the wrong values and set them NA in order to do an imputation using the *mice* library.

After that, we transformed this set of variables into factors and changed the factors for other ones that are better in order to analyze and comprehend the data.

- sex: male, female.
- marriage: married, single, other.
- education: Graduate school, university, high school, other.
- default.payment.next.month: yes,no.

In addition to this pre-processing, we created 6 new variables that tell if an individual have good account status, paid duly, paid the minimum or have some delay in the payments. This values will be one per month. So, an individual will have 6 states. We ended up modifying this function for another more efficient after computing a PCA. This function is the one that we show in figure 1



```r
Default_Dataset$PAYSTATUS_0 <- NA
Default_Dataset$PAYSTATUS_2 <- NA
Default_Dataset$PAYSTATUS_3 <- NA
Default_Dataset$PAYSTATUS_4 <- NA
Default_Dataset$PAYSTATUS_5 <- NA
Default_Dataset$PAYSTATUS_6 <- NA

getPaymentStatus <- function(row) {
  for (n in c(0,2,3,4,5,6)) {
    varPay <- paste("PAY_", n, sep = '')
    varStatus <- paste("PAYSTATUS_", n, sep = '')
    if(row[varPay]==-2){
      row[varStatus] = "Good account status"
      row[varPay]=0
    }
    else if(row[varPay]==-1){
      row[varStatus] = "Pay-duly"
      row[varPay]=0
    }
    else if(row[varPay]==0){
      row[varStatus] = "Pay minimum"
    }
    else {
      row[varStatus] = "Delay"
    }
  }
  return(row)
}
```

Figure 1: Pre processing old funtion

After this pre-processing, we computed simple PCA to see how the data behaves and how the variables are correlated one to each other. What we can see in this PCA is that the variables pay_x are so correlated that can be transformed into one variable. The PCA variable plot is showed in the figure 2. We can see also how the variables bill_amtxand pay_amtx are correlated too. We have decided to unify those variables too. We can see also how the variable limit_bal is slightly correlated with the payments represented by the variables pay_amtx which make sense.

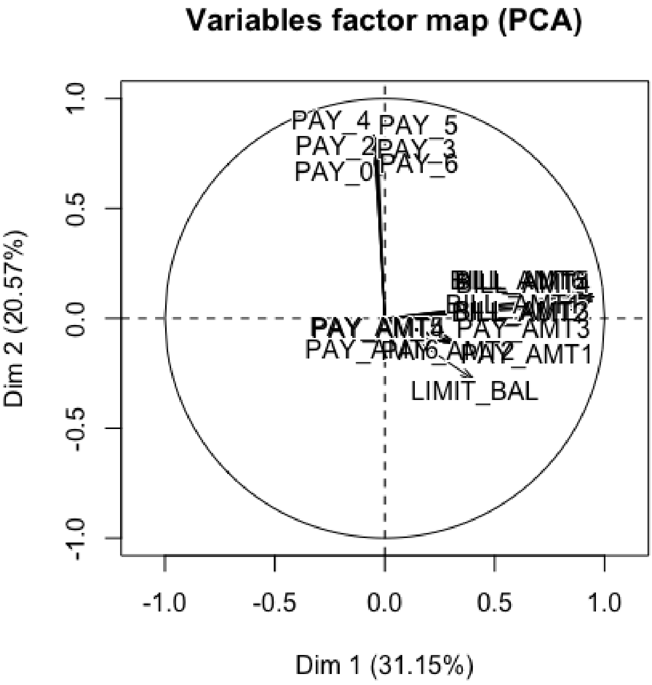
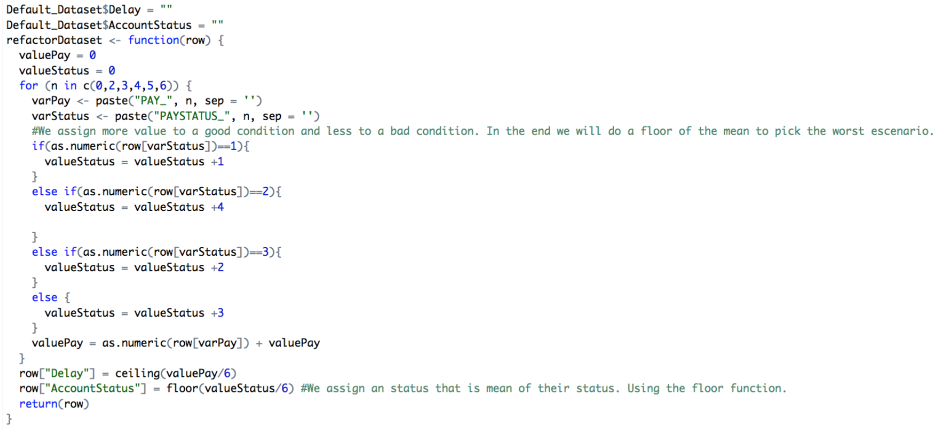


Figure 2: Old PCA

What we have decided is to transform the variables pay_x into one variables named delay that will show the mean delay of the individual. The variables paystatus_x will be transformed into a variable named Account Status that will show the most common status status of this individual.

The function used to perform this re-factor is showed in the figure 3. With this re-factor the PCA is exactly the same but the variables have been unified into one. More exploration on the PCA will be done further in this document in section 5.

After the execution of this function, we transform the variable ACCOUNT STATUS into a categorical variable. The old variables will not be deleted as will be useful for the prediction model. We also created two variables MEANBILL and MEANPAY that show the mean bill and the mean pay for each individual. We have done that because as we can see on the figure 2 the bills and the payments are immensely correlated. In addition, we have created a new variable, only for plotting purposes that tells to which decade belongs that individual.



```
Default_Dataset$Delay = ""
Default_Dataset$AccountStatus = ""
refactorDataset <- function(row) {
  valuePay = 0
  valueStatus = 0
  for (n in c(0,2,3,4,5,6)) {
    varPay <- paste("PAY_", n, sep = '')
    varStatus <- paste("PAYSTATUS_", n, sep = '')
    #We assign more value to a good condition and less to a bad condition. In the end we will do a floor of the mean to pick the worst escenario.
    if(as.numeric(row[varStatus])==1){
      valueStatus = valueStatus +1
    }
    else if(as.numeric(row[varStatus])==2){
      valueStatus = valueStatus +4

    }
    else if(as.numeric(row[varStatus])==3){
      valueStatus = valueStatus +2
    }
    else {
      valueStatus = valueStatus +3
    }
    valuePay = as.numeric(row[varPay]) + valuePay
  }
  row["Delay"] = ceiling(valuePay/6)
  row["AccountStatus"] = floor(valueStatus/6) #We assign an status that is mean of their status. Using the floor function.
  return(row)
}
```

Figure 3: New preprocessing function

4.2 Outlier detection and elimination

After the creation of the new variables, we will use robust mahalanobis distance to detect and eliminate outliers. As the variables are too correlated between them, we had to omit a lot variables for the mahalanobis distance method of the *chemometrics* library. The figure 4 shows the plots of both normal mahalanobis distance and robust mahalanobis distance. As we can see in the plot, there are clear outliers that should be eliminated. After looking and the distances we have decided to delete a 0.001% of the individuals, so, 30 individuals. Of course, we will delete the individuals that have greatest value in the robust mahalanobis distance. We only want to delete biggest outliers and lose the minimum amount of information in the process.

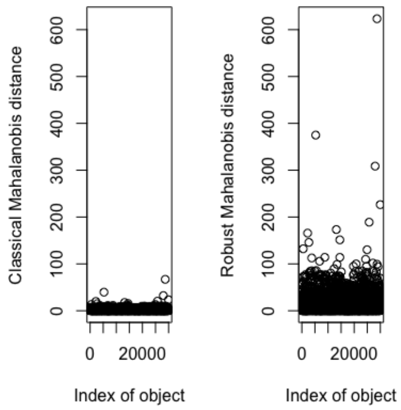


Figure 4: Mathalanobis plot

4.3 Descriptive statistics and analysis of the data.

In this section we will show different plots and analysis that we have done on the data after the pre-processing showed in the last section in order to get more understanding and knowledge of the data.

In figure 5 we can see the relation between education and default. We can see that there is no clear relation, individuals of all types of education can default.

The graduated individuals, however, seem to have less default, which make sense due the higher economic status.

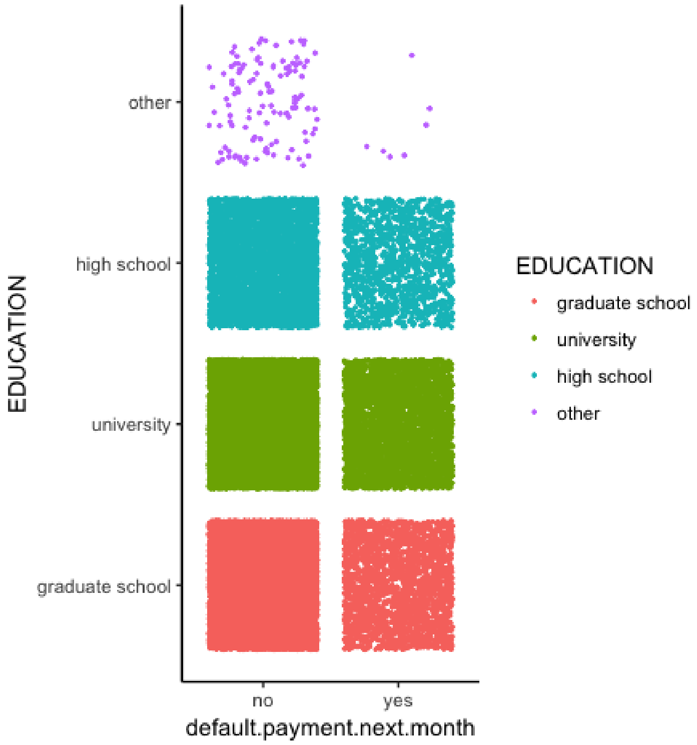


Figure 5: Relation between education and default

In the figure 6 we can see the individuals plotted by age grouped by decade and the default per age. We are also showing the individuals by sex, so we can see what sex in the areas. From the plot in figure 6 we can also see that all the individuals plotted by the account status. We also plotted them with the legend of sex to see if there is some difference. We can see that there is no significant difference by sex but we can see that the most of the individuals are located in the paid minimum status and in the no default. We can also see how the individuals that default, most of them have a delay status and, most of the individuals with a status of delay, default next month.

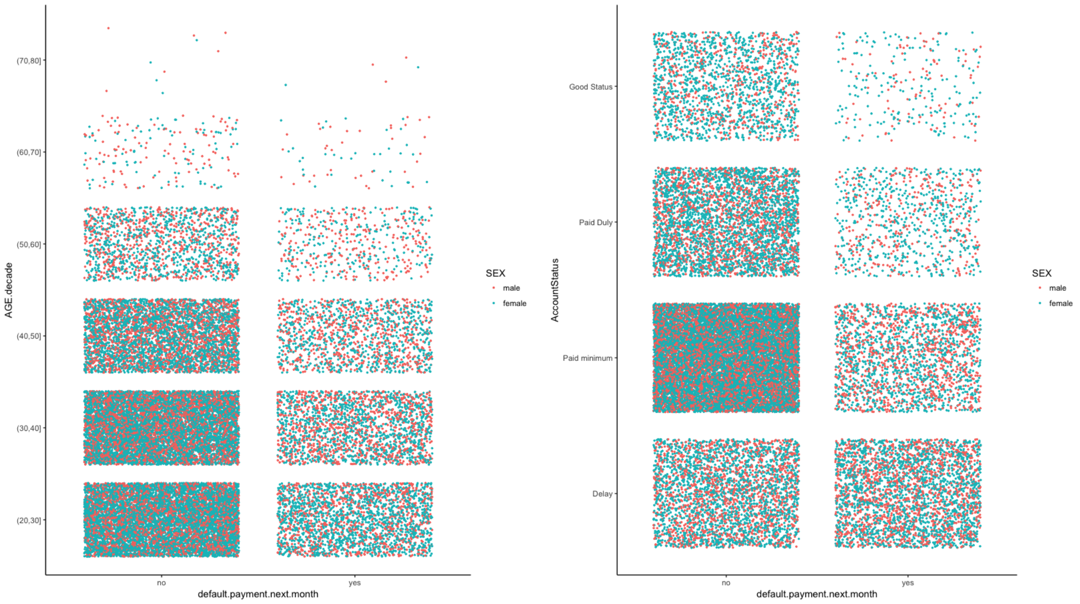
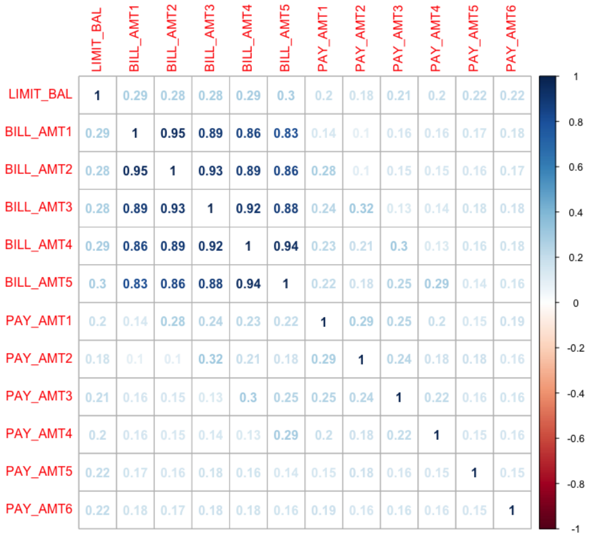


Figure 6: Plots showing the age decade and account status grouped by default.

In the plot of the figure 7 reflect the correlations between limit balances, bill amounts and payments amounts; it presents us that theres a low correlation between the limit balances and payments and bill amounts. However it can be seen that bill amounts has high correlation



	LIMIT_BAL	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6
LIMIT_BAL	1	0.29	0.28	0.28	0.29	0.3	0.2	0.18	0.21	0.2	0.22	0.22
BILL_AMT1	0.29	1	0.95	0.89	0.86	0.83	0.14	0.1	0.16	0.16	0.17	0.18
BILL_AMT2	0.28	0.95	1	0.93	0.89	0.86	0.28	0.1	0.15	0.15	0.16	0.17
BILL_AMT3	0.28	0.89	0.93	1	0.92	0.88	0.24	0.32	0.13	0.14	0.18	0.18
BILL_AMT4	0.29	0.86	0.89	0.92	1	0.94	0.23	0.21	0.3	0.13	0.16	0.18
BILL_AMT5	0.3	0.83	0.86	0.88	0.94	1	0.22	0.18	0.25	0.29	0.14	0.16
PAY_AMT1	0.2	0.14	0.28	0.24	0.23	0.22	1	0.29	0.25	0.2	0.15	0.19
PAY_AMT2	0.18	0.1	0.1	0.32	0.21	0.18	0.29	1	0.24	0.18	0.18	0.16
PAY_AMT3	0.21	0.16	0.15	0.13	0.3	0.25	0.25	0.24	1	0.22	0.16	0.16
PAY_AMT4	0.2	0.16	0.15	0.14	0.13	0.29	0.2	0.18	0.22	1	0.15	0.16
PAY_AMT5	0.22	0.17	0.16	0.18	0.16	0.14	0.15	0.18	0.16	0.15	1	0.15
PAY_AMT6	0.22	0.18	0.17	0.18	0.18	0.16	0.19	0.16	0.16	0.16	0.15	1

Figure 7: Correlations Between Limit Balance, Bill Amounts & Payments

In the plot of the figure 8 we explore the probability of having a higher balance limit by age and comparing the results of the limit of credit by default or not default. We can see that the no default individuals have more credit limit, but not that more, we can see also, that the blue line, which is the mean, it is 200000NT$ of limit of credit for all individuals, but with the age advance, the limit is lower. From this plot we can interpret that the bank prefers to give more credit and risk more, because we can see how people with more than 200000NT$ of credit limit defaults. It will be a decision of the bank to change the politics of credit limit, we an suggest to decrease the credit limit until the 40's.

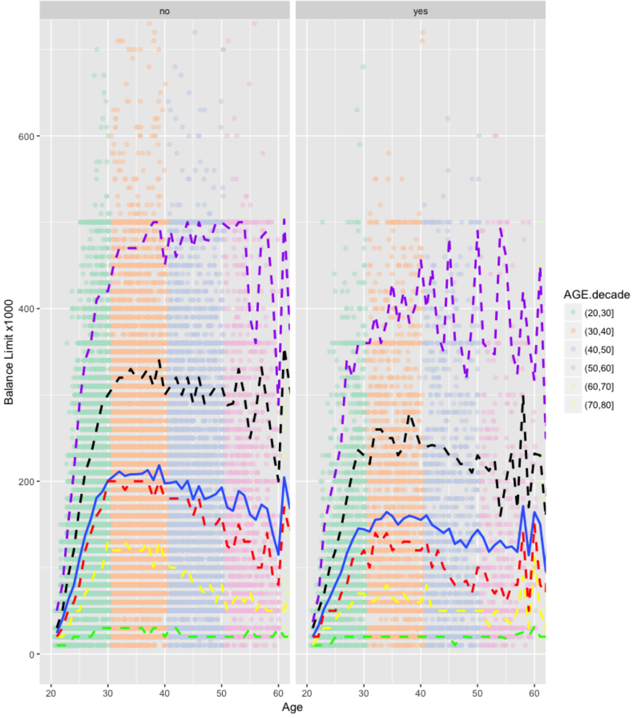


Figure 8: Personal Balance Limits Probabilities & Given Limits By Age

5 PCA

In this section we will deal with everything regarding the PCA in the default data set that we proposed for this project. We have executed the PCA with the variables that we have generated in the pre-processing section. After executing the PCA, we have the figure 9. We can see how the MeanBill and MeanPay are clearly correlated which make sense, if an individuals paid a lot with the credit card, for sure the individuals must have a big bill, so, the varialbes are correlated. Also the pay is correlated with the limit of credit (LIMIT_BAL), also make sense, if an individual pays more with the credit card, for sure, will have higher limit of credit. The delay is not that correlated with the MeanBill and MeanPay, it seems that an individuals can have delay regarding those variables, but seem quite correlated with the age.

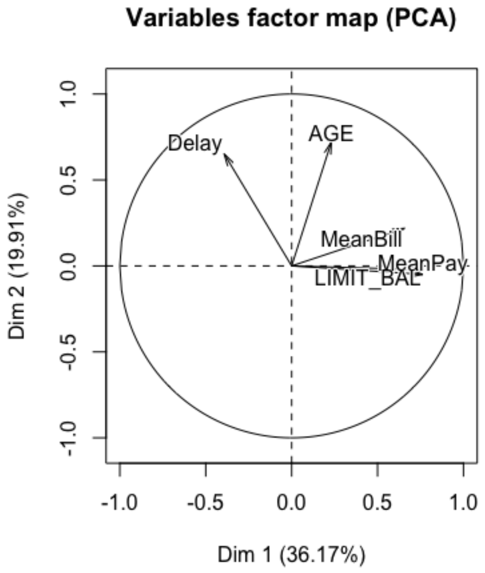


Figure 9: PCA

We can access some use full information about the individuals and the variables present in this PCA. For example, we can see what is the best and the worst represented individual, which are the most influential individuals and variables. We can see that information in the table 5.

Operation	**Results**
Best represented individual	7581
Worst represented individual	25790
Most influential individuals in the first PC	6913, 22851, 14554
Most influential individuals in the second PC	1993, 13262, 25870
Best represented variable	LIMIT_BAL
Worst represented variable	AGE
Most influential variables in the first PC	LIMIT_BAL, MeanPay, MeanBill
Most influential variables in the second PC	AGE, Delay , MeanBill

Table 1: Table PCA

6 Clustering

In this section we will describe how we have done the clustering in this data set. In our case we have performed a k-means clustering. During the first tests using a k-means we have realized that we need to reduce the number of individuals because it takes too much time to train. We have performed the k-means, using the matrix of individuals obtained from the PCA that we performed before, using that matrix we have performed two k-means, then, we identify to which cell every individual belongs and we compute the centroids of every cell. Computing the matrix distances lets us to perform hierarchical clustering. In the figure 10 we can see the dendrogram of the clustering performed before the cutting of the tree.

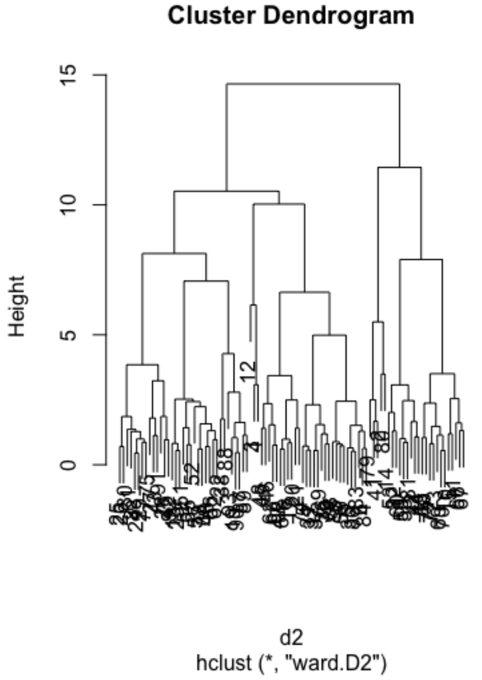


Figure 10: Cluster Dendrogram.

Now, we can see, in the figure 11 that the number of clusters that we should have is **3**, as is where the bigger "jump" is produced.

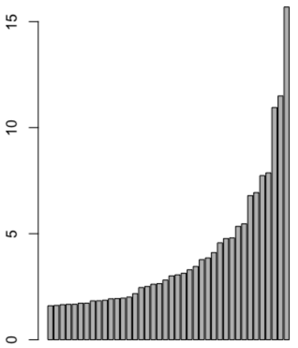


Figure 11: BarPlot.

After this, we cut the dendrogram using the number of clusters obtained before,then we use again the function aggregate, compute the centroids. Finally, we compute the k-means using those centroids obtain from hierarchical tree. The k-means can be seen on the figure. We can observe how in this clustering, most individuals are grouped in a single zone of the graphic. That's because most of the people that have a credit card (it is not given to everyone) have a more or less the same profile and they are not completely different. It is clear that the structure of this clustering is hierarchical structure.

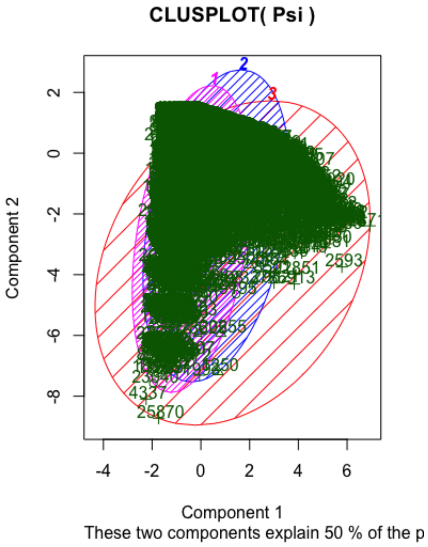


Figure 12: K-means clustering.

In the table 2 we can see a summary about the different clusters that we have obtained. It is clear how the data has been divided in 3 clear groups. We have a majority group of people which is the people that have more financial problems, as the percentage of default in the first cluster is 26.7%. We can also see how the Mean of the MeanBill is higher than cluster 2 despite of wasting less. That means that the individuals of this cluster tend to delay the payments, which can also be seen in the Mean of Delay, which is 0.64 months. The % of default is really high in this cluster for the previous reasons, the mean of default in the data set is 22.12% and this cluster have a default % of 26.7.

Cluster 2 is the cluster that will have the other majority of people, more middle-

upper class profile, that have better account status, more financial solvency. We can see how the delays of the cluster 2 are 1456, in cluster 1 are 4656, thats a huge difference. The mean of Delay has also decreased a lot, more than a 50% as now is of 0.29 months. Also the % of default has decreased a lot, which now is of a 19,2%.
Cluster 3 will include the individuals with the best economic status in the bank. We can see how the mean of the Mean Bill and the mean of the MeanPay have increased a lot. We can also see how the mean of Delay has decreased too. That means that this individuals are wasting and paying more to the bank, delay less and have less % of default at the end of the month, that clearly means that this individuals are the ones with the best economic status.

Cluster	Size	Mean of MeanPay	Mean of MeanBill	Mean Delay	Default	Account Status
1	14490	2813.8 NT$	30301NT$	0.64	Yes: 3870 No: 10620 %: 26,7	Delay: 4656 Paid Minimum: 7274 Paid Duly: 1949 Good Status: 611
2	11481	3500.4 NT$	25977NT$	0.29	Yes: 2215 No: 9266 %: 19,2	Delay: 1456 Paid Minimum: 5971 Paid Duly: 2965 Good Status: 1359
3	3999	18163NT$	151403NT$	0.21	Yes: 549 No: 3450 %: 13,7	Delay: 558 Paid Minimum: 3035 Paid Duly: 268 Good Status: 138

Table 2: Summary Clustering

7 Classification

7.1 Decision Tree

One of the predictors that we want to use in this case is the decision tree. Decision tree has been probed to one of the best predictors and we wanted to give it a try.

Before doing anything we shuffle all the data and we split the data in two parts, train(75%) and test(25%). We have tuned the tree in order to force it consider more leafs than the default ones, because with no tuning, the tree only considered one variable of decision. After the tuning, the tree obtained is the one showed at figure 13. This model have an error of 18.6% which is not bad at all compared to the results that other researchers got in this same problem.

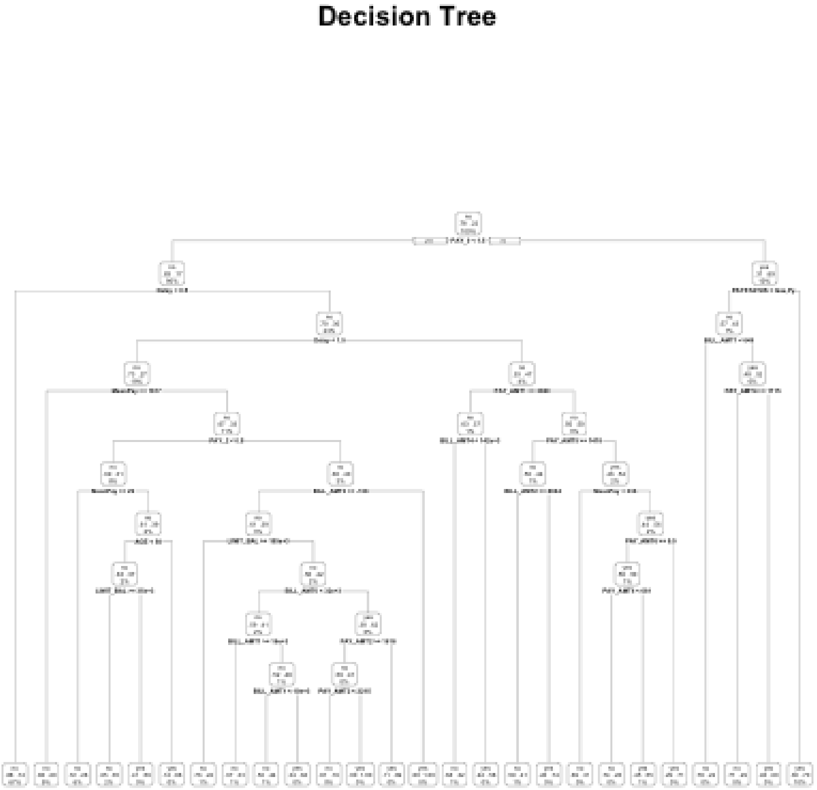


Figure 13: Decision tree.

7.2 Random Forest

Another classifier we wanted to give a try is Random Forest, since it has proven to be a good solution for classification problems. However, before applying the algorithm, it's important to find out optimal number of trees based on classification error. We ran random forest for several instances of trees and results are depicted in Table 3

ntrees	OOB
10	0.2308
16	0.2118
25	0.2020
40	0.1937
63	0.1890
100	0.1855
158	0.1855
251	0.1860
398	0.1850
631	0.1847
1000	0.1842
1585	0.1840

Table 3: Random Forrest error rate for number of trees

As we can see from Table 3, the optimal number of trees is **100**, since it has the lowest error and any number above that doesn't decrease error much. While it will take more time to train model with more number of trees.
As we can see, error rate of Decision Tree and Random Forest is almost same. Moreover, we need to perform cross validation for Random Forest to validate our final model

7.3 Validation Protocol

Besides splitting data into train and test with 75/25 % proportion in order to validate the model and estimate generalization error for test data, we decided to perform cross validation and be more confident in final model selection. The table

4 shows the result of **10-fold cross-validation**:

Fold	Training Error	Validation Error
1	0.1879506	0.1915556
2	0.1859259	0.1888889
3	0.1861728	0.1915556
4	0.1871605	0.1804444
5	0.1880000	0.1768889
6	0.1875556	0.1822222
7	0.1868642	0.1862222
8	0.1859259	0.1951111
9	0.1879012	0.1826667
10	0.1857778	0.1840000
Average	0.1869235	0.1859556

Table 4: 10-fold cross-validation for Random Forest

From table 4 we can see that the average *Validation Error* is **0.186** with **(0.180,0.191)** *95%* confidence interval. Which means our final model is performing expectedly good and we can use it as our final model.

8 Conclusions

In conclusion we can say that the data set was quite challenging and interesting at the same time. We had to clean data, perform feature selection, detect outliers and wrong data, set them as missing values and perform imputation. In addition to that we heavily performed feature extraction, which helped to perform clustering with very good and easily understandable results. And the accuracy was drastically increased from default one.
The Principal Component Analysis and Clustering helped us analyze latent concepts and understand the data better, trough interpretation of clusters. Although we have 30000 rows and 24 columns in the original data set, after the analysis of the data we have seen that a lot of columns were too correlated to give additional information and we created new ones.
After exploring the data we shuffled all the data and then used 75% train and

25% test validation protocol, as well as 10-fold cross validation to verify our final classification model. It is important to remark that even in the pre-processing we unified those variables, in the classification models we had to use all of them, if not we got very bad results. That's because those variables, in terms of understanding the correlations may be unnecessary, but they still give relevant information and deleting them lead to lose of information, that is crucial when building the machine learning model.

Overall, the dataset used was difficult to predict. One possible reason was that the data available was not complete. For example, we did not know the total credit that each individual owns, we only have the difference between what they pay and the bill and the total credit. We neither had information about the income of individuals, that would be a very helpful variable. Most likely this data would lead to a better results.

References

[YL09] I. Yeh and C. Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Experti Systems with Applications*, 36:2473–2480, 2009.