Universitat Politècnica de Catalunya

Machine Learning

2^{nd} Semester

Prediction of Credit Card Default

Authors:
Ricard Gardella Garcia

Albert Garriga Porqueras

David Sardà Campomanes

October 9, 2018

1 Introduction

In this project we will use different machine learning models on a dataset to predict the default of credit cards. The dataset is a publicly open dataset from the UCI Machine Learning repository.

First, in section 2 we explain some previous work that was done with the same dataset. In the following section, number 3 we explain the main characteristics and the variables of the dataset used. In section 4 the preprocessing done in the data is explained and later, in section 5 some first analysis to understand the dataset is done. Then, section 6 we present the different models that we use and the results obtained with each one. With the results obtained in section 7 we explain the model chosen and we estimate its generalization error using cross-validation. Finally, the conclusions are presented in section 8.

2 Previous work

There is a previous paper [YL09] that uses this dataset to predict the default of clients. In the paper six models are tested and their accuracies are computed. 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 evalutae the algorithms, the error rate (related with accuracy) and the area ration of the lift curve because othe dataset is unbalanced.

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

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

3 Data Used

The data used for this project is a data 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 dataset contains 30000 rows and 24 columns. It contains categorical and numerical variables. It does not contain missing values but it contains errors that will be corrected during the preprocessing.

3.1 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 unecessary as it do 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. NT$ is the currency of Taiwan, 1 NT$ is 0.03319$. The percentage of default in our data set is 22.12%.

So, the dataset 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. So, we will use more measures, the precision and the recall to evaluate the classifiers.

4 Pre-Processing

First of all, we checked if there are some missing values and we have seen that there are no missing values, but, we have detected wrong values in the variables MARRIAGE and EDUCATION. These wrong variables were imputated using the *mice* library.

In addition to this pre-processing, we create 6 new variables that tell if an individual has a good account status, paid duly, paid the minimum or has some delay in the payments. This values will be one per month, so, an individual will have 6 states.

After this pre processing, we computed the PCA to see how the data behaves and how the variables are correlated one to each other. In that PCA the variables PAY_X are so correlated that can be transformed into one variable. We can see also how the variables BILL_AMTXand PAY_AMTX are very correlated too, we have decided to unify those variables too. We also see how the variable LIMIT_BAL is slightly correlated with the payments represented by the variables PAY_AMTX which make sense.

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 STATUS that will show the most common status status of this individual.

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 the section of PCA. After the execution of this function, we transform the variable 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, because all the six variables for each one are very correlated. In addition, we have created a new variable, only for plotting purposes that tells to which decade

belongs each individual.

5 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 1 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 seem to have less default, which make sense due the higher economic status.

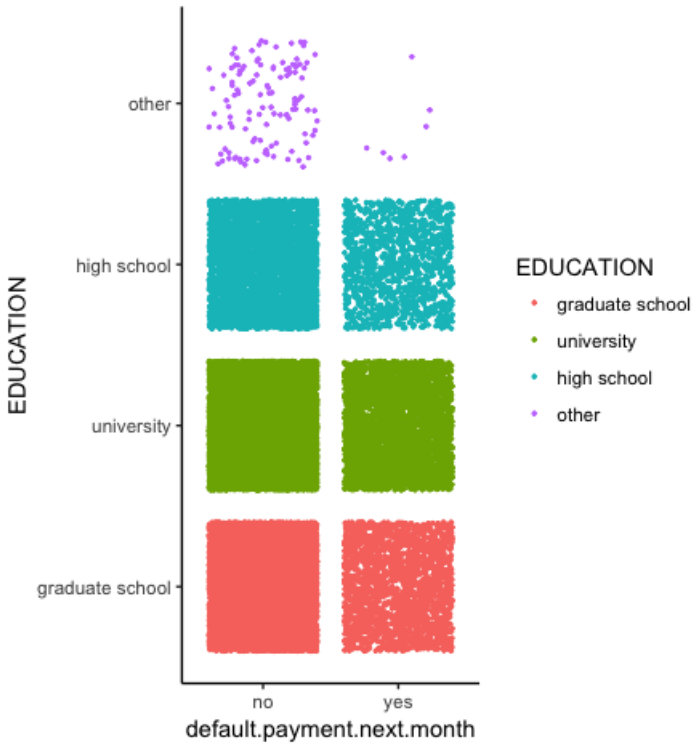

Figure 1: Relation between education and default

In the plot in figure 2 we can see the individuals ploted by his age grouped by decade and the default per age. We are also showing the individuals by sex, so we can see what are the sex more influent in the areas. The plot in figure 2 we

can also see all the individuals ploted by the account status. We also ploted 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.

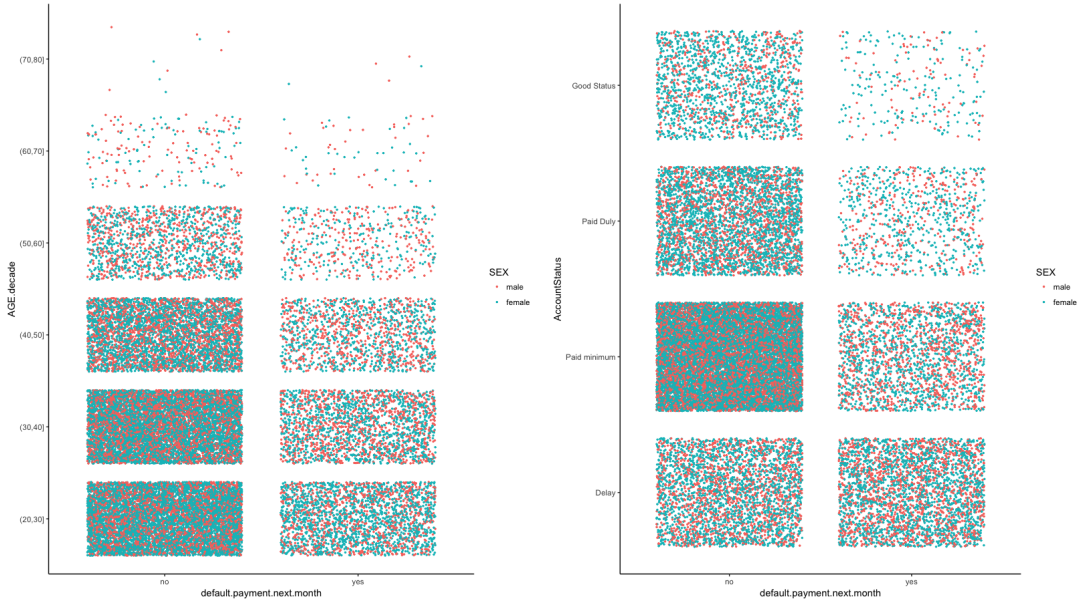

Figure 2: Plots showing the age decade and account status grouped by default.

In the plot of the figure 3 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

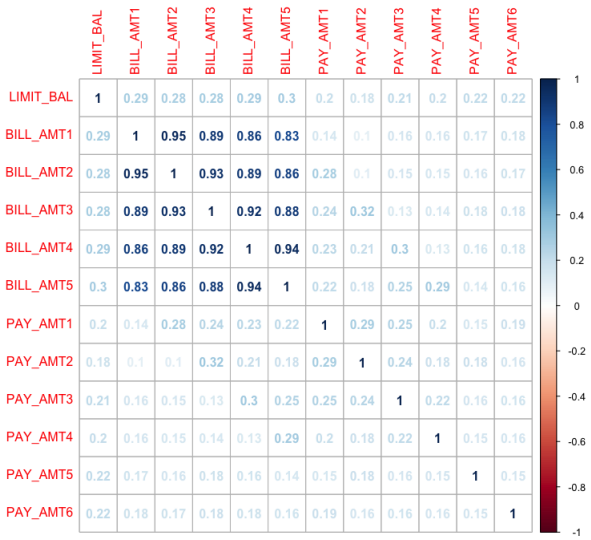

Figure 3: Correlations Between Limit Balance, Bill Amounts & Payments

In the plot of the figure 4 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 200k$ 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 200k$ 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.

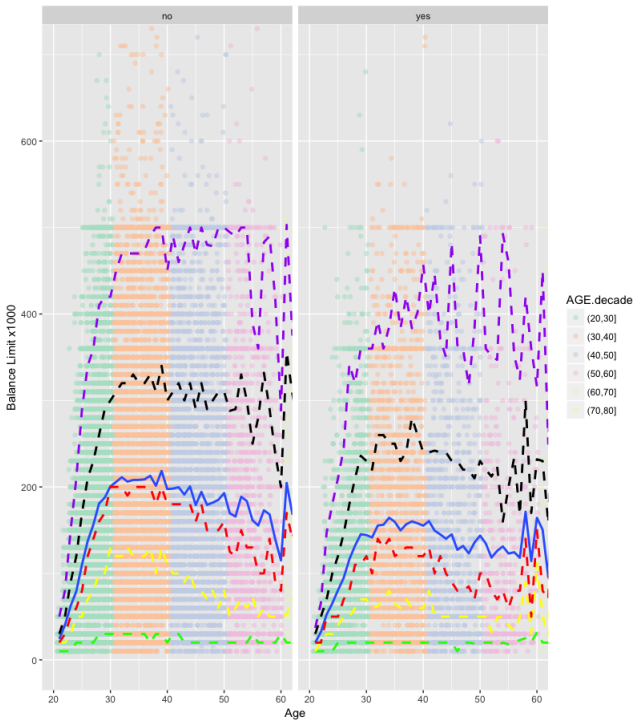

Figure 4: Personal Balance Limits Probabilities & Given Limits By Age

5.1 PCA

In this section we will deal with all 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 5

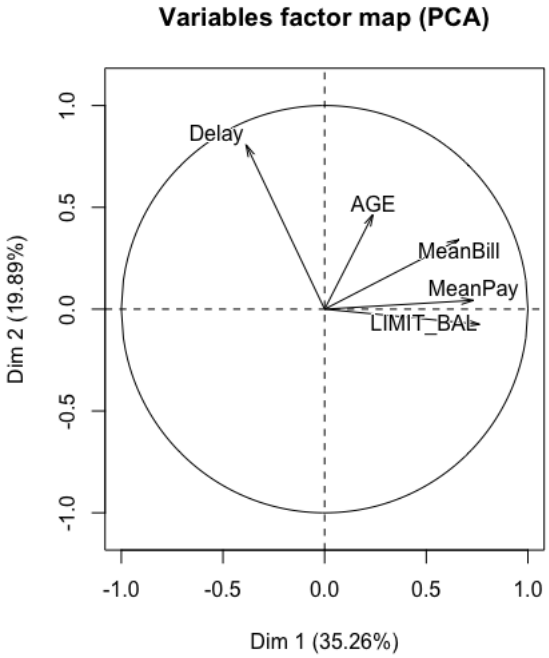

Figure 5: PCA

After executing the 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 5.1. For example, in the table 5.1 we can see that the most influential variable sin the first principal components are the limit of credit, the mean of payments and the mean of bills. That makes absolute sense in a data set of a bank. The most influential individuals, for example, we can see that the individual 28717 is an outlier in terms of payments and bills, having a meanBill of 182092NT$ and a meanPay of 627344.3NT$. Absolutely insane. It is important to remind that all this data is

real, so, this outlier is not an error, is a real individual that is a client of one of the banks that we are analyzing.

Operation	**Results**
Best represented individual	7581
Worst represented individual	25784
Most influential individuals in the first PC	28717, 5297, 2198
Most influential individuals in the second PC	25870, 4337, 1993
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	Delay, AGE , MeanBill

Table 1: Table PCA

5.2 Clustering

In this section we will deal how we have done the clustering in this data set. In this 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 due the number of individuals that we have. 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 centroid of every cell. Computing the matrix distances lets us to perform hierarchical clustering.

Now, we can see, in the figure 6 that the number of clusters that we should have is 3 or 7, we have decided to do it with 7 clusters, that plot is a plot of the last 40 aggregations.

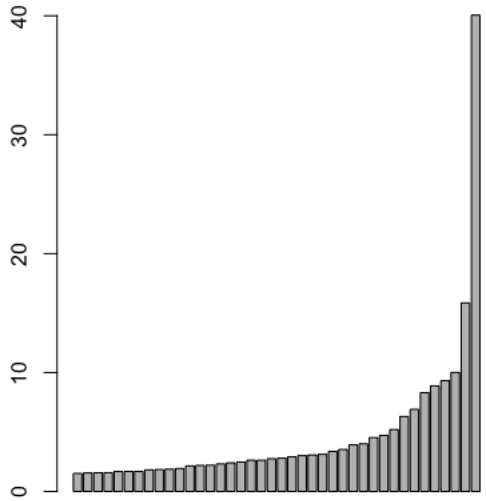

Figure 6: BarPlot.

After this, we cut the dendogram using the number of clusters obtained before,then we use again the function aggregate, compute the centroids. Finally, we compute the k-means. The k-means can be seen on the figure. In order to see the plot in a more clear way, we have done the same plot zooming the zone with more clusters.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. There are clear outliers, which are people, that spends and pays an amazing amount of money.

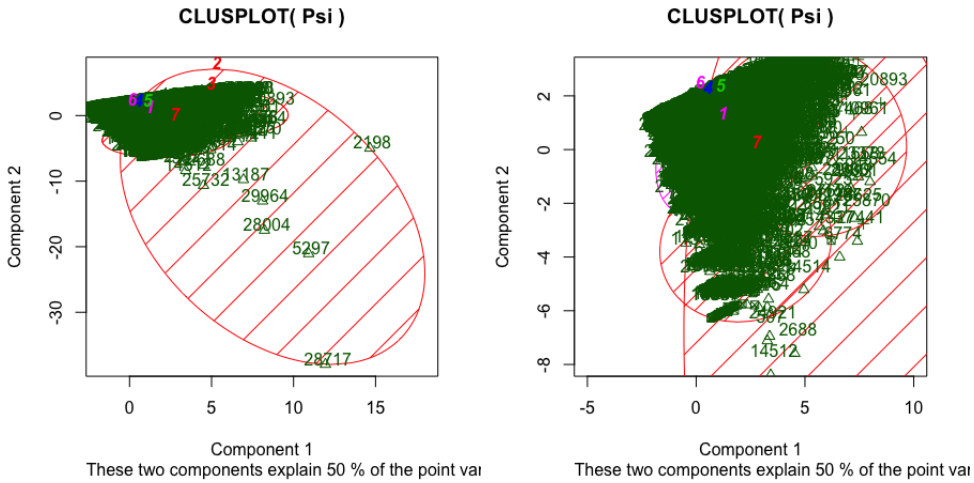

Figure 7: K-means clustering.

In the table 2 we can see a summary about the different clusters that we have obtained. We can see in a clear way, for example, how cluster 7 represents the people with more debt and default, 1637 of 2573 people will default next month, that's a lot considering that the default is a 22.15% of the total. Cluster 2 will represent the wealthiest people in the bank, having payments the mean of the payments more than 50000NT$ and bills of more than 120000NT$. The other clusters represent medium class individuals but divided by how good their status is and how much they spend and pay. We can see for example how cluster 3 has a mean of bills of more than 200000NT$, most of the individuals have a paid minimum status, which is logic, this will belong to middle class individuals with high debt, but we can also observe how cluster 5 have a mean of the bills of 27000NT$ and a mean of the pays of 2800NT$ and most of them have good status, this people should be middle class but with a higher economic stability. In the code attached, with the execution of the code, more information can be extracted.

Cluster	Size	Mean of MeanPay	Mean of MeanBill	Default	Account Status
1	4655	2380.5	25419	-no: 3006 -yes: 1649	-Delay: 2662 -Paid Minimum: 1323 -Paid Duly: 670 -Good Status: 0
2	477	58216	128429	-no: 432 -yes: 45	-Delay: 13 -Paid Minimum: 340 -Paid Duly: 83 -Good Status: 41
3	2457	9400.4	203782	-no: 2006 -yes: 451	-Delay: 542 -Paid Minimum: 1908 -Paid Duly: 2 -Good Status: 5
4	3756	12557	35510	-no: 3473 -yes: 283	-Delay: 183 -Paid Minimum: 737 -Paid Duly: 1561 -Good Status: 1275
5	9364	2865	27646	-no: 8170 -yes: 1194	-Delay: 1 -Paid Minimum: 6716 -Paid Duly: 1764 -Good Status: 883
6	6718	2528	24787	-no: 5341 -yes: 1377	-Delay: 842 -Paid Minimum: 3677 -Paid Duly: 1545 -Good Status: 654
7	2573	2072.2	42856.3	-no: 936 -yes: 1637	-Delay: 2542 -Paid Minimum: 31 -Paid Duly: 670 -Good Status: 0

Table 2: Summary clusters

6 Model selection

In this section we try the following models: naive-Bayes, nearest neighbours, decision trees, random forest and support vector machines. For each model we present the results obtained by the best parameters that we have found from the ones tested. As a measure to choose the best model we use the overall accuracy. However, we also report the precision and accuracy of each model because these measures are very relevant in a binary classification problem, specially if it is unbalanced like this case.

6.1 Naive Bayes

Naives Bayes is one of the best, easier and most famous methods to apply when facing a classification problem.

Before applying this method, we will shuffle our data and create two folds, one for training (75%) and another one for test (25%). We used the library *mlr* to compute the naive bayes model. The first model computed with bayes have and error of 22.13%. Considering that the percentage of default in our data set is 22.12% this result is not good at all.

In order to have better results, we will use cross validation on our data. Cross validation consists on dividing our data in a lot of folds, and creating different models, trained with different data in order to have the best model possible. We will do cross validation with 10 folds.

The results obtained with cross validation are very similar that the ones that we obtained without it. The best of the 10 models have an error of 22%, which is very bad.

After that result we tried to take out some variables from the data to see if an over-fitting was taken place, and we saw that if we keep only the general variables that we created (MeanBill,MeanPay,AccountStatus) and we take out the more concrete variables we get sligher better results. The best model in this case was of 20,2% which is slightly better than the previous ones.

6.2 Nearest neighbours

Nearest neighbors is also a predictor that comes easily in mind, so we wanted to try to see how accurate it was in our case. Using the same two folds as in naive bayes we applied it to our data. We used the *knn* function from the package *class*, and tested it for different number of neighbours. The best results we obtained were with a high number on neighbours, around 17-21 was where we obtained the best results, and the error obtained was of around 19%. We can see the accuracy of every number of neighbors in the next figure:

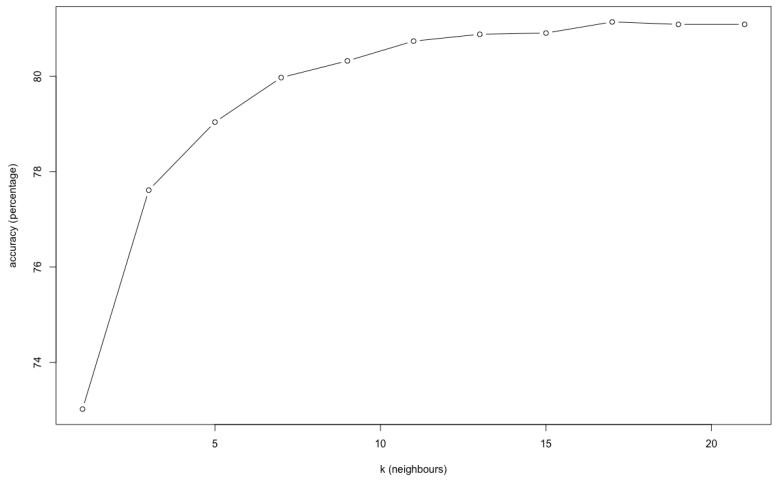

Figure 8: Accuracy for different number of nearest neighbors

We can see that taking into account the percentage of default, like we did in nearest neighbours, it's certainly an improvement over Naive Bayes, so nearest neighbours looked promising as a predictor for our case.

6.3 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 him 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 9. This model have an error of 17.6% which is not bad at all compared to the results that other researchers got in this same problem.

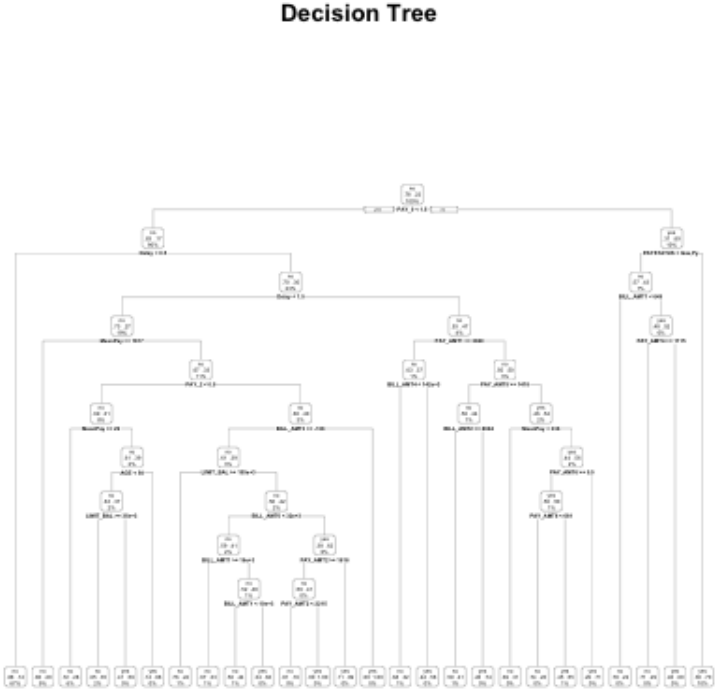

Figure 9: Decision tree.

6.4 Random Forests

Random Forests seemed like a logical step forward from Decision trees, so we gave it a chance. First we began by tuning its parameters, specifically the mtry parameter (the number of variables randomly sampled at each step as candidates), and we found out that actually 2 was the best value in our case.

Later, we executed our predictor at first using 500 trees, using the $randomForest$ function from the package of the same name, and the results were as follows:

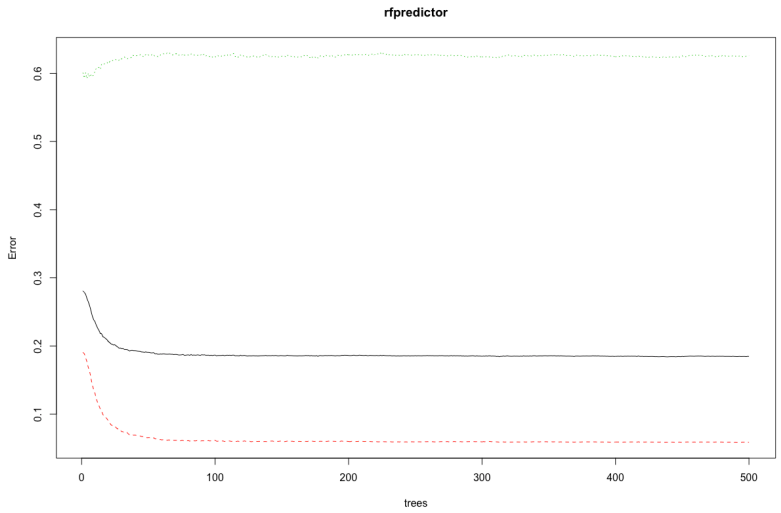

Figure 10: Random forest accuracy for different number of trees

As we can see, it seems that at 100 trees our predictor is as accurate as it can be. Still, the error obtained was of approximately 19%. Is interesting to see that a well-tuned decision tree gives a very similar accuracy, even a bit better than this one.

6.5 Support Vector Machines

Since this is a binary classification problem, we thought that Support Vector Machines (SVM) could be a good method. First, we have used a linear kernel and then a radial kernel. For the radial kernel several values of γ are tested: 0.0001, 0.001, 0.01, 0.1 and 1. From all the tests, the best accuracy is obtained by 82.27% for the linear kernel. The best radial kernel also showed a good accuracy of 82.02%, which was achieved with $\gamma = 0.01$. The code used for this model can be found in *svm.R*.

7 Model validation

The summary of the performance of all methods is:

	Accuracy	Precision	Recall
Bayes	75,98	55,91	46,91
Random Forest	81,03	65,32	36,56
KNN	81,24	64,25	36,44
Decision Tree	81,83	69,33	33,66
SVM lineal	82,27	37,49	69,12
SVM radial	82,02	35,86	68,85

Table 3: Performance of every model tested.

From all the models we see that the best one obtained is the linear SVM using as a criteria having the highest accuracy. In order to estimate its generalization error the test accuracy obtained is not useful because we have used this error to tune parameters of the model (choose the kernel function used). So, probably this value underestimates the true error. In order to estimate this error we are going to use 10-fold cross-validation with this model. The cross-validation metrics obtained are shown in the table 4.

Accuracy	Precision	Recall
82,12	36,38	67,94

Table 4: SVM Cross-validation

We can see that the values obtained are very similar to the ones obtained before. They are slightly worse, as we could expect because here we are not tunning the model. We can see that the precision of this model is very low, but the accuracy is quite high. This means that the model classifies an individual to default 'easily' than the others. This means that more individuals are classified, so the recall is higher, but at the cost of having a lowest precision. There is a trade-off between precision and recall.

8 Conclusions

Although we have 30000 rows and 24 columns, after the analysis of the data we have seen that a lot of columns were too correlated to give additional information. We have computed the machine learning models that we considered more adequate

to the problem that we were facing. We also have seen that all the models have a similar accuracy. However, they have a different behaviour. Some of them have a high precision (compared with the other methods) and a low recall while some other methods have the opposite behaviour. This is an example of a trade-off. A high recall means that more samples are labelled as positive (default), so the precision will be also be lower because we are less selective. On the contrary, if the recall is low, we are more selective on the samples so the precision is also higher.

For this particular dataset we observe that the results are not very good. It is difficult to classify the individuals correctly. The final model chosen, despite being the one with the highest accuracy, has a low precision and and the recall is not very high. For this particular problem, probably having a higher recall at the cost of a lower precision makes sense, because giving a credit to someone that is going to default have a higher cost for the company that not giving a credit to someone that could have paid (it just loses one client).

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.

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.