Statistical & Machine Learning
Credit Card Default
Individual Assignment

IÉSEG School of Management – March 31st, 2022

Maria Arques

Table of Contents

1. INTRODUCTION

The project aims to predict which client will default for next month payment based on some factors. The first type of factors are the intrinsic ones, they belong to the person itself and do not show any kind of behavior, in this case we have the age, the marital status, gender and education. However, there are other factors that can be considered extrinsic and are related to the actions that the individual took in the past such as the status of his past payments or the amount previously paid. Both these factors might influence the default answer.

To predict whether the client will default for any given value of the other variables, some models will be created using machine learning algorithms. Later on, a benchmark will be set up to be able to decide in between the best model performance.

In our case, the response variable that we want to predict isn't quantitative but qualitative. Therefore, all the approaches taken for predicting this categorical variable will be used for classification. If our response variable would have been quantitative, regression models would have been applied.

We have a set of training observations that we use to build a classifier. The classifier needs to show a good performance not only on the training data but also on the testing.

2. DATA CLEANING

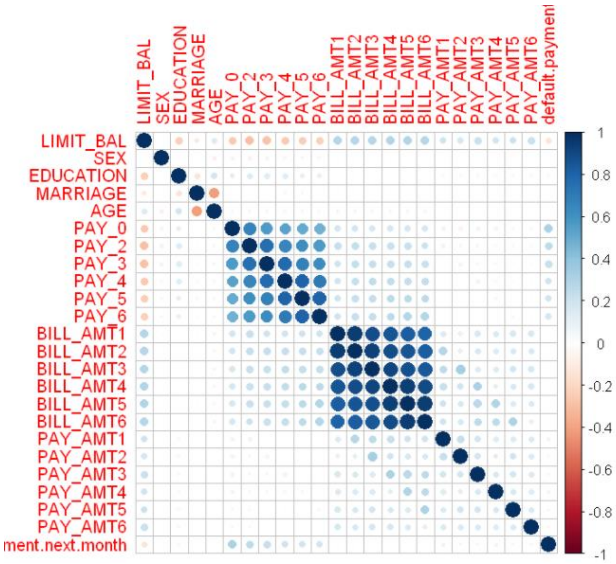
Before starting to build the models, data was prepared. The performance of the following steps was made to improve the accuracy of the models. The first step taken into consideration was to split the data into train and test.

Missing values: One important step in the data cleaning was handling the missing values. In this case, after checking the number of rows containing missing values, I realized that they couldn't be removed as they represented 80% of the information. The only column that didn't contain missing values was the response variable.

How were the missing values treated? The qualitative variables such as Sex, Education, Marriage and the Status Repayment that contained missing values were replaced by the mode. A big remark is that these variables were replaced both in the train and test parts by the mode from the train dataset. The unique values were checked for each of the qualitative variables to make sure that those correspond to the possible answers given in the excel description. For example, in the gender variable the options were 1 for male and 2 for female but when checking the unique values, it appeared that the variable contained some number 4. These "outliers" were replaced by the mode.

Similarly, quantitative variables having missing values were replaced by the mean of the train dataset.

Correlation matrix: another important consideration is the understanding of how variables are correlated. If there are highly correlated variables it can lead to an unstable model solution, and it would be better to remove them for overfitting problems.



We can observe that the only variables that are highly correlated are the amounts of bill statements. It is a good sign for our data to apply future models.

Variable transformation: one last preprocessing step was taken into consideration, categorical variables such as the gender, education, marriage and the target variable were converted from numerical to factors to avoid the hierarchy between different classes.

Scaling: was applied to all the numerical variables to make it easy for the model to understand the problem.

3. MODEL APPLICATION

3.1 Logistic regression

Model definition:

Logistic regression is a classification algorithm that calculates probabilities of being part of a class. Each observation is matched with the class where it has the highest probability score of belonginess.

Logistic regression is used for binary classification where our output predictions can only take 2 possible outputs: default or not default. Logistic regression doesn't fit a line to the data as the linear regression does, it fits an "S" shaped "logistic function" called Sigmoid curve. In this case, we can't use linear regression as we could predict values that are higher than 1 (default) or less than 0 (non-default) and it wouldn't be accurate (it doesn't fit all the data). Therefore, the logistic model

solves the problem by using the logistic function where the results range from 0 to 1. The formula that explains the model is the following:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.$$

or it can also be written as: p = exp(y) / [1 + exp(y)]

Where "y" is the linear regression function and "p" is the probability of an event to occur given x. In our case, as we have more than one predictor variables, so our "y" looks like this: b0 + b1*x1 + b2*x2 + ... + bn*xn (James et al., 2021).

Once we have defined our model, it is necessary to fit the S shaped line in the data, for doing so the maximum likelihood method is used as the statistical properties are higher than the least square approach. This method allows to find the coefficients beta 0, beta 1, beta 2… that when substituted into the logistic regression function, they result into a number close to 1 for people that defaulted and a number close to 0 for the ones that didn't. After the coefficients are estimated, the only thing missing is the computation of the probability of next payment for any given level of the other variables (James et al., 2021).

Result interpretation

When looking at the results obtained after fitting the model, there are some variables highly related to the outcome, those variables are the ones that have 3* in the p-value column as they are the ones with the smallest p-values. More specifically these variables are: being single (number 2 in the marital status column), the payment that the person did on time and the amount paid in September 2005. There are other variables less significant as having a university education that also played a role in the result.

The positive coefficient estimate means that an increase in the variable is associated with the increase in the probability of defaulting. Oppositely, when the coefficient is negative, an increase in the value of the variable leads to a reduction of the probability of defaulting. (*Logistic Regression Essentials in R*, 2018)

The variables that have larger p-values are not statistically significant. For this reason, stepwise regression has been applied to select the best features.

```
Call:  glm(formula = default ~ LIMIT_BAL + SEX + EDUCATION + MARRIAGE +
    PAY_0 + PAY_2 + PAY_3 + PAY_4 + BILL_AMT1 + BILL_AMT3 + BILL_AMT4 +
    PAY_AMT1 + PAY_AMT2 + PAY_AMT4 + PAY_AMT5, family = binomial,
    data = train)
```

Stepwise regression selected the variables in the picture as the best predictors and hasn't taken into consideration the other 12 variables. The main advantage is to reduce the complexity of the model and at the same time increase the accuracy (*Logistic Regression Essentials in R*, 2018). The model was then run with the new feature selection.

0.7985 (Accuracy after stepwise)

Finally, cross-validation was applied as a non-biased metric to obtain the AUC of the model and be able to compare it with other models to select the best one.

Advantages and disadvantages: (*Logistic Regression Essentials in R*, 2018)
+ Easy implementation, no high computation power.

+ It gives not only the accuracy of each predictor (it tells the coefficient size) but also the association relationship.

+ Gives high accuracy for simple data sets

+ "Logistic Regression **outputs well-calibrated probabilities** along with classification results. This is an advantage over models that only give the final classification as results"

- It constructs linear boundaries

- It assumes that a linear relationship exists between the dependent and independent variables. Nonlinear problems can't be solved.

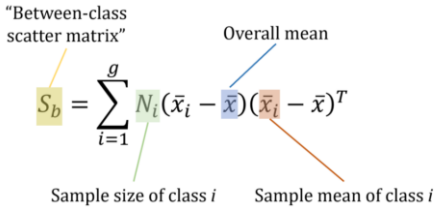
- Only used for discrete functions and sensitive to outliers

3.2 Linear Discriminant Analysis

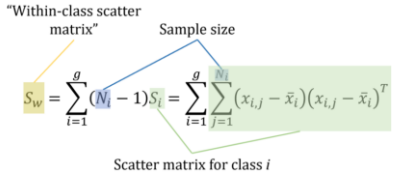
Model definition:
LDA is a classification algorithm that predicts the class of an observation by making use of linear combinations of the predictor variables. It is an alternative to the logistic regression method, and it estimates probabilities "by modeling predictor's distribution for each of the classes". Later, it uses Bayes' theorem to transform them into estimates. One big assumption that it does is based on the fact that the predictors are normally distributed, and the classes have identical covariances matrices (James et al., 2021).

The main idea of this model is to reduce dimensionality to increase model performance. Basically, what the algorithm does is to get rid of the redundant predictors and try to convert the dataset into a lower-dimensional space while keeping most of the data. To apply the linear discriminant analysis, 3 steps need to be taken into consideration:

1- Computation of the "separability" between classes. It is calculated by taking the mean of different classes to have an indication for the algorithm of how hard the problem is. In this case, if the means are closer, the problem gets harder (Ye, 2021).



2- Computation of the within-class variance: the higher the variance within classes, the more difficult it gets. (Ye, 2021)



3- Construction of a lower-dimensional space that maximizes the separability between classes and reduces the variance within classes. (Ye, 2021)

Result interpretation

```
Prior probabilities of groups:
         0         1
0.7800625 0.2199375
```

The results that we obtained express that there is 78% of the training observations that didn't default while the remaining 28% are the ones that defaulted.

```
Group means:
  LIMIT_BAL      SEX2 EDUCATION2 EDUCATION3   EDUCATION4 MARRIAGE2   MARRIAGE3
0  177093.5 0.6163769  0.4818524  0.1570387 0.0049675507 0.5462703 0.009614614
1  131089.0 0.5751634  0.5143507  0.1841432 0.0008525149 0.5140665 0.013071895
       AGE      PAY_0      PAY_2      PAY_3      PAY_4      PAY_5      PAY_6
0 35.43758 -0.2062335 -0.2898806 -0.3060652 -0.3491707 -0.3828219 -0.3994872
1 35.60023  0.6510372  0.4512646  0.3677181  0.2537653  0.1591361  0.1023018
    BILL_AMT1    BILL_AMT2   BILL_AMT3    BILL_AMT4    BILL_AMT5   BILL_AMT6
0  0.01034348  0.006740016  0.00621478  0.005535992  0.004954422  0.00410359
1 -0.03668570 -0.023905125 -0.02204225 -0.019634760 -0.017572079 -0.01455439
     PAY_AMT1    PAY_AMT2    PAY_AMT3    PAY_AMT4    PAY_AMT5    PAY_AMT6
0  0.03978868  0.03830653  0.03321629  0.03368027  0.03364782  0.02793977
1 -0.14112034 -0.13586355 -0.11780977 -0.11945538 -0.11934028 -0.09909527
   LIMIT_BALL
0  0.07876532
1 -0.27936062
```

When we fit the model, we obtain these results that represent the mean or separability between classes (step 1).

In the results we can see some of the probabilities and classification for the first six observations in the test set.

	0	1
10382	0.9153081	0.08469194
8806	0.4408607	0.55913930
4721	0.9331993	0.06680068
8533	0.7832661	0.21673393
11257	0.7430734	0.25692656
8992	0.7874096	0.21259035

Backward stepwise regression was applied then to the model for the automatic feature selection. Basically, what it does is to include or delete variable per variable from the model to improve its quality. I used backward elimination, so I started with all the features and only removed the ones that make the model worst. After applying the stepwise, the best features were obtained and plugged into the model. From there I got a model accuracy of 0.8015. Finally, cross validation was performed for a future benchmarking study.

Advantages and disadvantages:

Advantages:

- It is simple and fast, outperforms logistic regression methos
- It has a linear decision boundary

Disadvantages:

- The assumption of a normal distribution needs to fulfill
- It is only applicable for binary classification
- It is unstable with well-separated classes

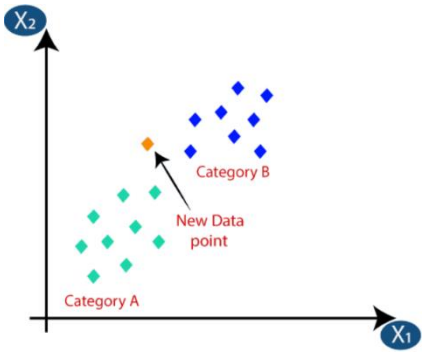
3.3 KNN

Model definition:

K-nearest neighbors is a supervised machine learning algorithm that is build on the assumption that proximity implies similarity. In other words, "similar things are close to each other". The goal of the algorithm is to predict which class does the data from the test dataset belong to by calculating the distance between the test data and the points from the train dataset (Christopher, 2021).

The purpose of our dataset will be to be able to predict or classify a client that will default next month payment. The knn algorithm will find similar features of both groups to be able to correctly classify it.

How does this method work?



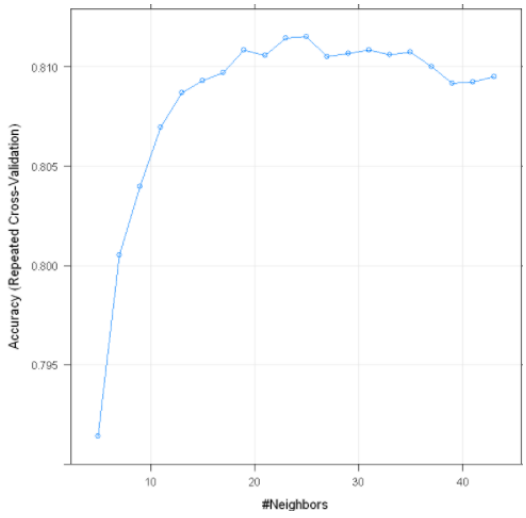
Let's imagine in this case that we need to assign the orange point to one of the two categories; in our dataset it can be extrapolated to assess if the "new person" will fall into the default category or into the non-default category. The algorithm will select at first the number of K neighbors that will be taken into consideration. In my jupyter notebook I gave it a random number at first but then I used a function to calculate the optimal number of neighbors. Once this step is done, Euclidean distance is calculated between points (Christopher, 2021).

$$\text{Euclidean Distance between } A_1 \text{ and } B_2 = \sqrt{(X_2-X_1)^2+(Y_2-Y_1)^2}$$

Thanks to the Euclidean distance we are able to find the closest neighbors (of the new data point) belonging to each of the class. Once the neighbors are identified, the count of the number of neighbors for each class is performed. Finally, we will assign this new data point to the category with the highest number of neighbors selected in the previous steps.

Result interpretation

After running the model with a random number of k and in order to choose the right value for the number of neighbors I decided to apply gridsearch that checked for 20 values of neighbors and gave me the optimal number where the accuracy was higher. This idea can be seen in this graph:



As we can see at the beginning, as we increase the value of K the predictor becomes more stable, and accuracy increases too. The optimal value corresponds to 25, after this value, if we keep increasing the number, the accuracy drops.

Finally, as for the other models, cross validation was applied for the benchmarking.

Advantages and disadvantages:

Advantages:

- It is simple and easy to implement, only 2 parameters to specify: the k value and distance function.
- No need to build a model, it is a lazy learner model. It learns from the train dataset only when making predictions.

Disadvantages

- It is not accurate for large datasets; it takes a lot of time to calculate the distance between the new point and existing one.
- It is not accurate with high dimensions
- Scaling of the data is needed

3.4 Decision Tree

Model definition:

Decision tree is an algorithm that corresponds to the family of the supervised machine learning algorithms. Basically, what the model does is split the data set into smaller subsets, the final result is a tree with decision nodes and leaf nodes.

It is a two-step process; the first step corresponds to the learning part and the second one is used for predictions. The goal is to predict a class by "learning simple decision rules inferred from prior data". Stratification and segmentation of the predictors are involved. When the prediction step occurs, the mode of the training observations is used. "Each observation belongs to the most commonly occurring class". (James et al., 2021)

In our dataset, the classification trees are the ones that I'll use as I need to predict a qualitative response. When building the classification tree, binary splitting is used. The splitting process is the process of dividing the data into smaller subsets on a particular variable. The second step corresponds to the pruning part which consists of reducing the tree size by removing some leaf nodes. The final step corresponds to selecting the smallest tree that fits the data. (James et al., 2021)

Based on what are the binary splits done? There are 3 possible criteria; classification error rate which corresponds to the % of training observations in the region that don't correspond to the "mode" of the class. The problem with this criterion is that it isn't enough sensitive. The 2 other alternatives are the Gini index and the Entropy index. (James et al., 2021)

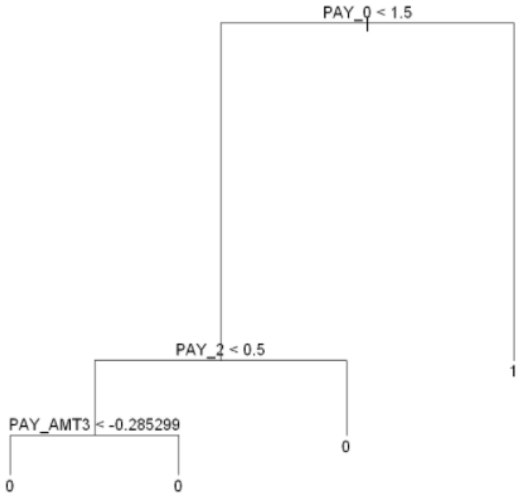
$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}),$$ Gini index

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}.$$ Entropy index

Gini index measures the variance between the classes, a small value is a good sign as it means that within a class most of the values are correctly allocated. Therefore, when we want to evaluate the quality of a split, these 2 criteria are the ones used.

Result interpretation

After fitting the model, the result that we get is the following:



Basically, what it represents is that when the April payment is delayed for more than one month, people will default. If the payment of April is not delayed, then the next node to look at is the payment of June. If the payment of June is not delayed, then the person won't default. Finally, the last variable taken into consideration was the amount paid in July.

Grid search was also applied to find the best feature for the model.

Advantages and disadvantages:

Advantages:

- It is simple and easy to explain and display graphically
- No need to create dummy variables

Disadvantages

- Not the best predictive accuracy.
- They are non-robust, small change in the data causes a large change in the final estimated tree.

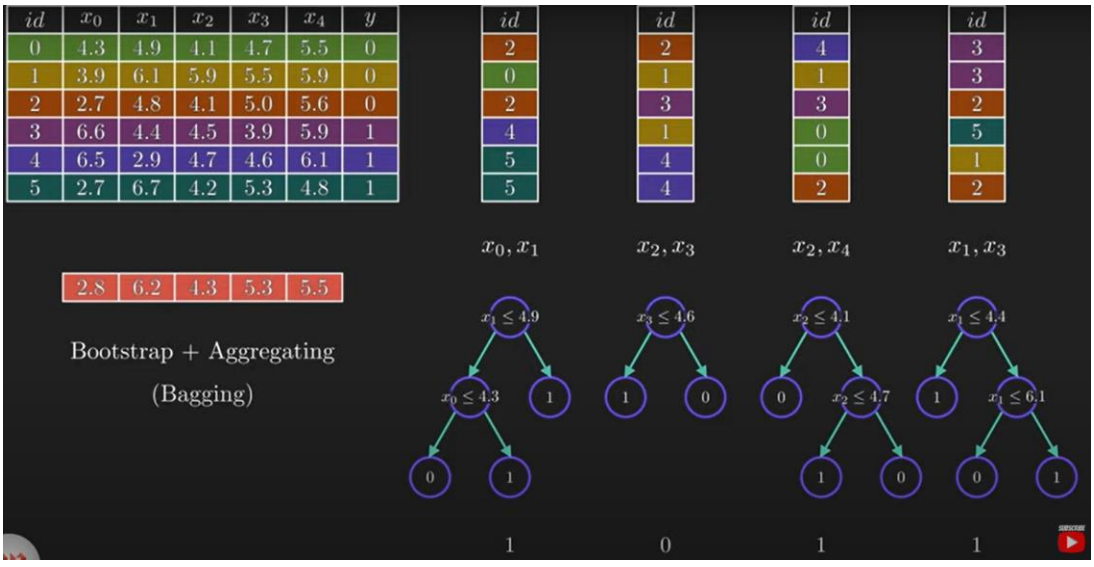
3.5 Random Forest

Model definition:

Random Forest is an algorithm that corresponds to the family of the supervised machine learning algorithms. The problem with decision trees is that they are sensitive to the training data and therefore it causes variance. For this reason, the random forest is useful: it builds a group of decision trees by training them with the "bagging" method. In other words, it sticks multiple decision trees to get a better accuracy and better prediction (Donges, 2021).

One of the good points about the random forest is that it adds the randomness factor; it looks for the best feature in between a subset of features contrary as looking for the best feature when splitting the node as decision tree does (Donges, 2021).

First step when building a random forest model is to create new datasets from the original data by taking some of the rows. We need to always keep in mind that we will take the same number of rows as the original dataset. The process of creating new datasets is called bootstrapping. The next step is to train a decision tree in each of the new datasets. The difference is that all the features won't be selected to be applied on each of the trees. A subset of features will randomly be selected to each tree. For example, if we build new datasets from the main one, feature 1,2 and free can be used in the first dataset, feature 5,8 and 9 can be used in the second one, and so on. Trees are built on each of the dataset applying the selected features and that's why each tree will look different than each other. Subsequently, a prediction needs to be made, therefore we will perform the prediction of the datapoint for each of the decision trees previously build. Finally, as it is a classification problem, the majority voting prediction will be taken. This last step is called aggregating. Bootstrapping makes sure that the same data is not used for every tree (less sensitive) and the random feature selection improves the correlation between trees (*Random Forest Algorithm Clearly Explained!*, 2021).



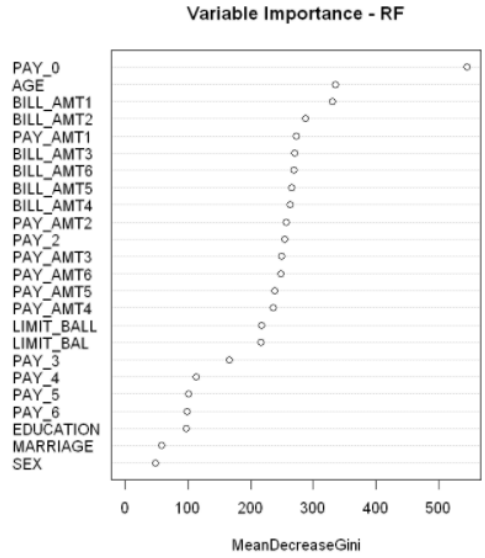
Result interpretation

At first, the model was fitted, and this were the results obtained:

```
Call:
 randomForest(formula = default ~ ., data = train, mtry = 6, importance = T)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 6

        OOB estimate of  error rate: 18.44%
Confusion matrix:
      0    1 class.error
0 11751  730   0.0584889
1  2220 1299   0.6308610
```

From there, best features were checked according to the ones with higher decrease Gini,



The model was then built with the best features and cross validation was applied.

Advantages and disadvantages:

Advantages:

- Stable model
- Robust to outliers

Disadvantages

- More complex model
- Long training Period

4. BENCHMARKING

	Logistic Regression	LDA	KNN	Decision Tree	Random Forest
AUC	0.723	0.7214284	0.7608	0.74911	0.726

After applying cross validation to understand which was the best model, I can conclude that the KNN algorithm is the one that better predicts if a client from the bank will default or not as the accuracy is the highest.

5. REFERENCES

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics)* (2nd ed. 2021 ed.). Springer.

Logistic Regression Essentials in R. (2018, March 11). Articles - STHDA. http://www.sthda.com/english/articles/36-classification-methods-essentials/151-logistic-regression-essentials-in-r/

Ye, A. (2021, December 15). *Linear Discriminant Analysis, Explained in Under 4 Minutes*. Medium. https://medium.com/analytics-vidhya/linear-discriminant-analysis-explained-in-under-4-minutes-e558e962c877

Shah, P. (2021). *What is Linear Discriminant Analysis(LDA)?* Knowledge Hut. https://www.knowledgehut.com/blog/data-science/linear-discriminant-analysis-for-machine-learning

A. (2019, November 15). *What are the Advantages and Disadvantages of KNN Classifier?* I2tutorials. https://www.i2tutorials.com/advantages-and-disadvantages-of-knn-classifier/

Christopher, A. (2021, December 29). *K-Nearest Neighbor - The Startup*. Medium. https://medium.com/swlh/k-nearest-neighbor-ca2593d7a3c4#:%7E:text=K%2Dnearest%20neighbors%20(KNN),closet%20to%20the%20test%20data.

Donges, N. (2021, September 17). *A Complete Guide to the Random Forest Algorithm*. Built In. https://builtin.com/data-science/random-forest-algorithm

Random Forest Algorithm Clearly Explained! (2021, April 21). YouTube. https://www.youtube.com/watch?v=v6VJ2RO66A