UNIVERSIDAD CENFOTEC

Data Analytics / Big Data 2018c-3

Program 5 Task 3

Data Science with Python

Build and Evaluate Models

PRESENTA:

Alberto Madrigal Jiménez

ASESOR:

Gustavo Rojas

San José, CRC. 2020

Credit One Report

Main Problem:

An increase in customer default rates is bad for Credit One since its business is approving customers for loans in the first place. This is likely to result in the loss of Credit One's business customers.

Proposed Solution:

Make predictions with the Default Payment Field, in order to understand the behavior of the customers answering the following questions.

1) How do you ensure that customers can/will pay their loans?

   With the information provided in the dataset as: sex, education, marriage, age, the last 6 month pays, and the last 6 billing amounts, we can predict the customer behavior, if they will default the loan or not.

2) Can we approve customers with high certainty?

   Based on the story provided by the data, we can understand the behavior of the current customer with at least six month with historical information. We can predict and say with 82% of accuracy the answer about the default loan.

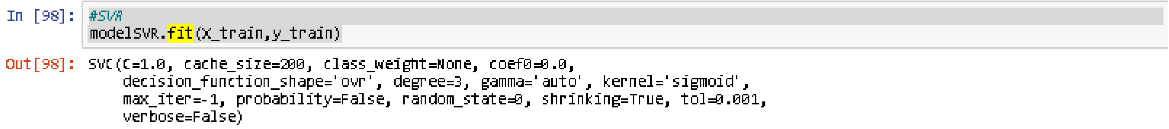
Cleaning and Pre-processing the Dataset

The following processes were applying over the data to clean and show the final dataset, we did not add a lot of technical information in this section, because, it information can be checked in the python doc.

1) Fit

In the following image we can see how using fit function, we can fit an estimator to be able to predict.

Example:



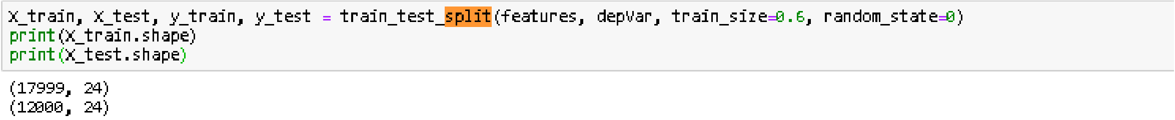
```
In [98]: #SVR
         modelSVR.fit(X_train,y_train)

Out[98]: SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
             decision_function_shape='ovr', degree=3, gamma='auto', kernel='sigmoid',
             max_iter=-1, probability=False, random_state=0, shrinking=True, tol=0.001,
             verbose=False)
```

2) Mapping to a Uniform distribution

Function: train_test_split is for splitting a single dataset for two different purposes: training and testing. The testing subset is for building your model. The testing subset is for using the model on unknown data to evaluate the performance of the model.

Example:

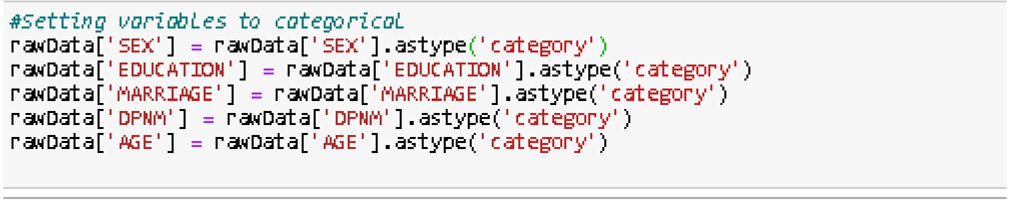


```
X_train, X_test, y_train, y_test = train_test_split(features, depVar, train_size=0.6, random_state=0)
print(X_train.shape)
print(X_test.shape)

(17999, 24)
(12000, 24)
```

3) Encoding categorical features

Categorical Data: Often in real-time, data includes the text columns, which are repetitive. Features like gender, country, and codes are always repetitive. These are the examples for categorical data.

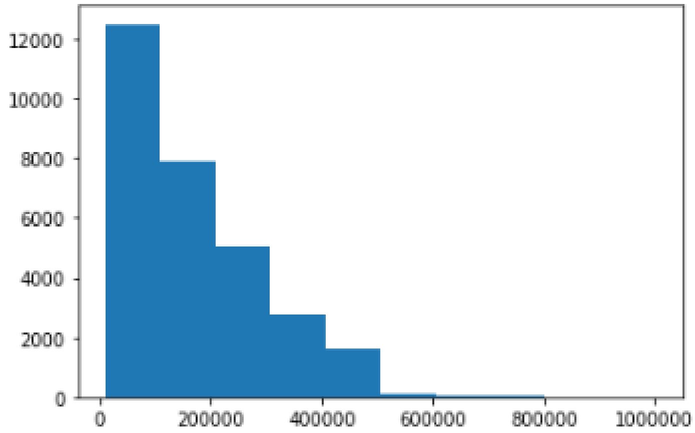


```
#Setting variables to categorical
rawData['SEX'] = rawData['SEX'].astype('category')
rawData['EDUCATION'] = rawData['EDUCATION'].astype('category')
rawData['MARRIAGE'] = rawData['MARRIAGE'].astype('category')
rawData['DPNM'] = rawData['DPNM'].astype('category')
rawData['AGE'] = rawData['AGE'].astype('category')
```

EDA

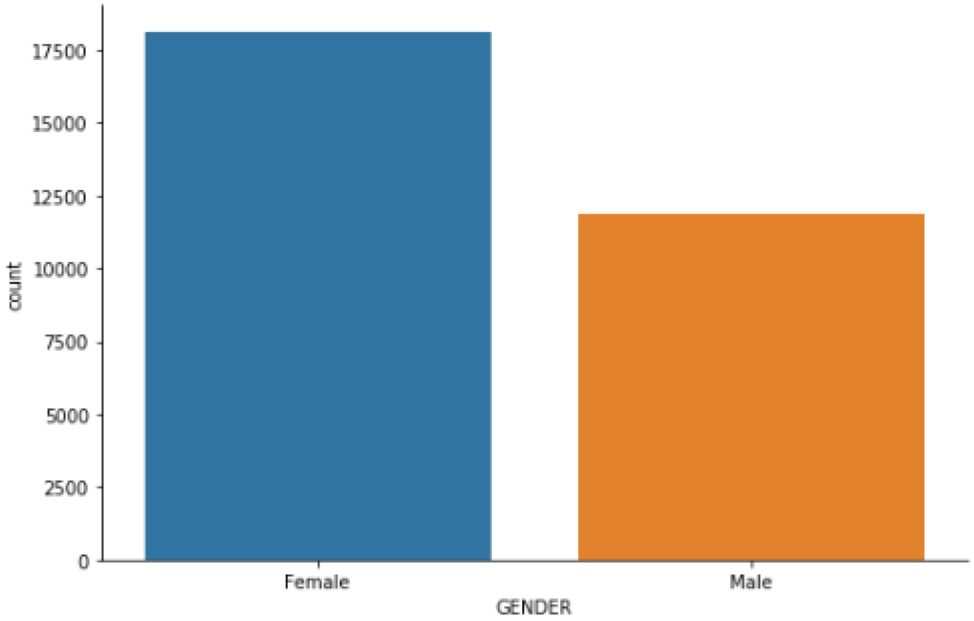
Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main features, often with visual methods. A statistical model can be used or not, but primarily EDA is for identifying what the data can tell us beyond the formal modeling or hypothesis testing task.

Balance of the limits:



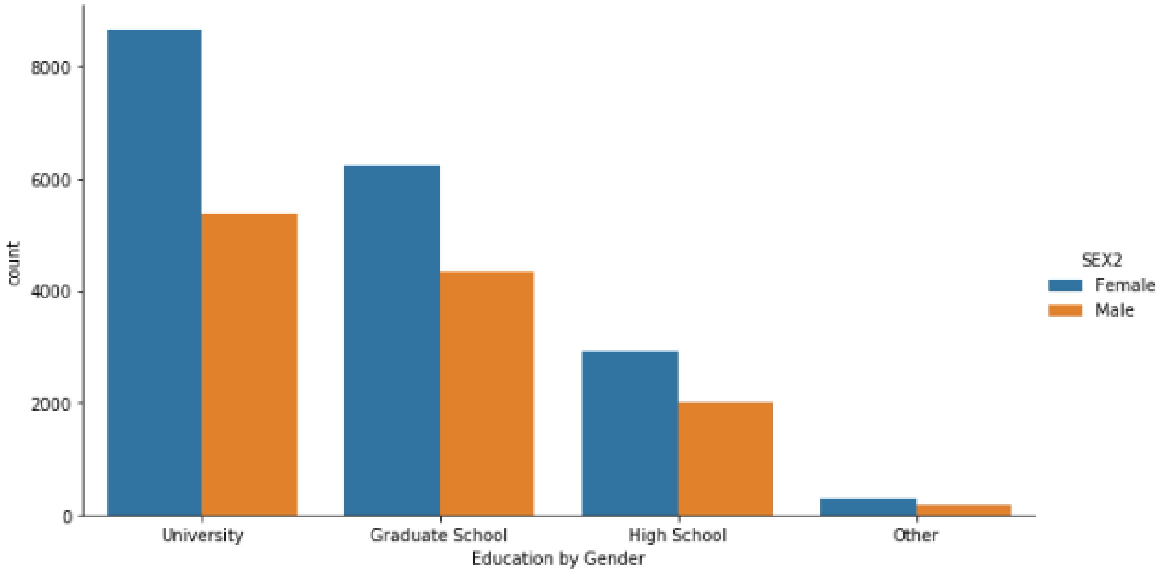
- ○ As we can see in the previous chart, the highest count of amount are between zero and 200000.

Count of rows by Gender:



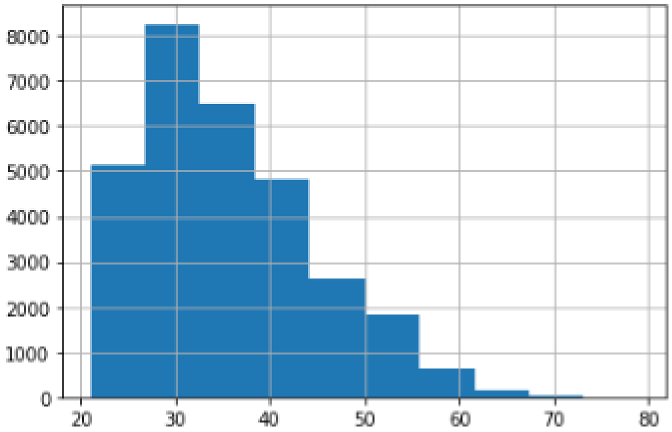
- ○ The Female count is higher than Male count

Analysis of Count by Education



- The previous image shows how the university grade school has the highest count of people, also, the female domains the volume.

Distribution Clients by Ages

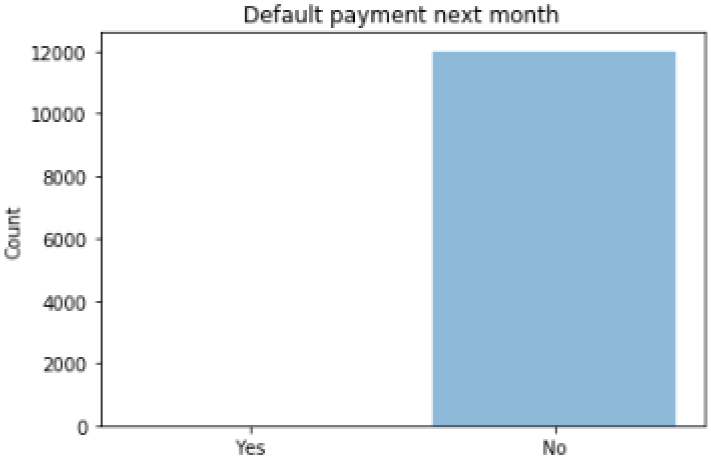


- The highest numbers of people are around of 30 years old.

Classification Methods Applied

SVM (SVC):

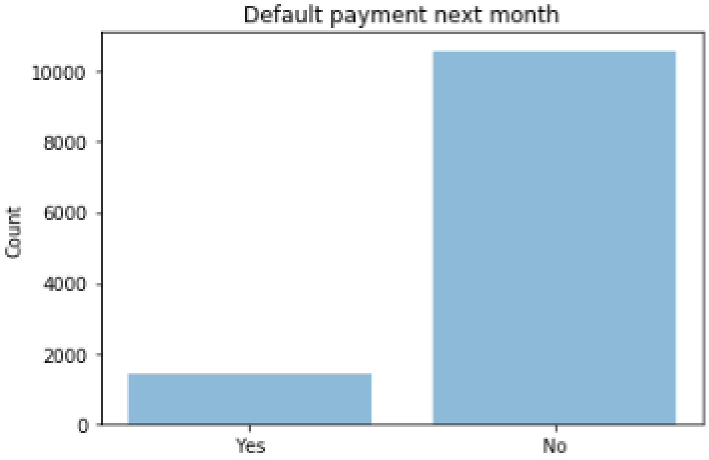
Vector support machines, support vector machines, or support vector machines are a set of supervised learning algorithms. These methods are properly related to classification problems.



Outcome: But the results obtained are not the best, because this method is not properly due to, all the values obtained are zero.

Random Forest Classifier (Classification Tree):

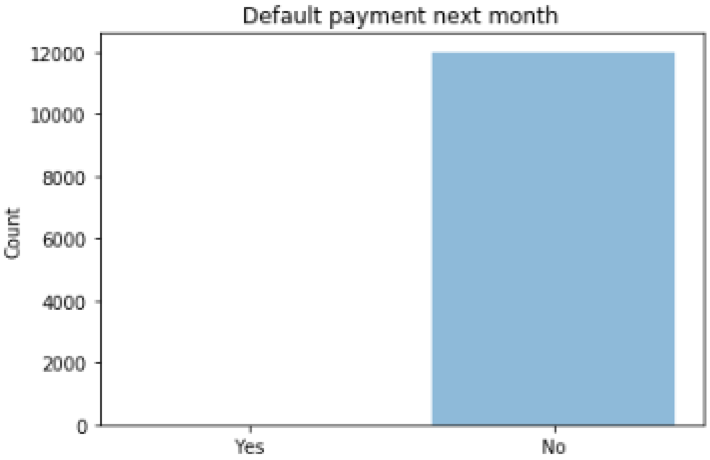
Random forest also known in Spanish as '"Random Forest"' is a combination of predictor trees such that each tree depends on the values of a random vector tested independently and with the same distribution for each of these.



Outcome: the results obtained look pretty nice, with high percent of accuracy on the results.

Logistic Regression:

Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.).

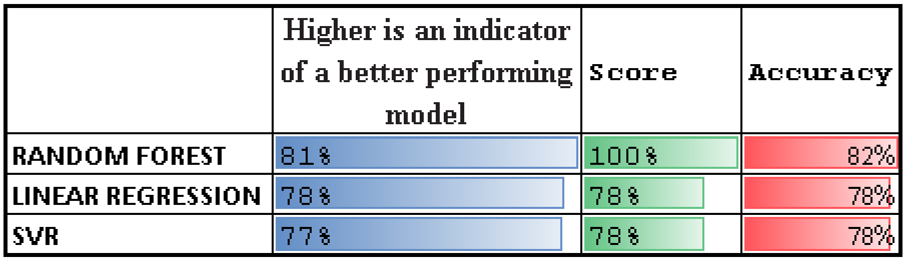


Outcome: But the results obtained are not the best, because this method is not properly due to, all the values obtained are zero.

Method Selected:

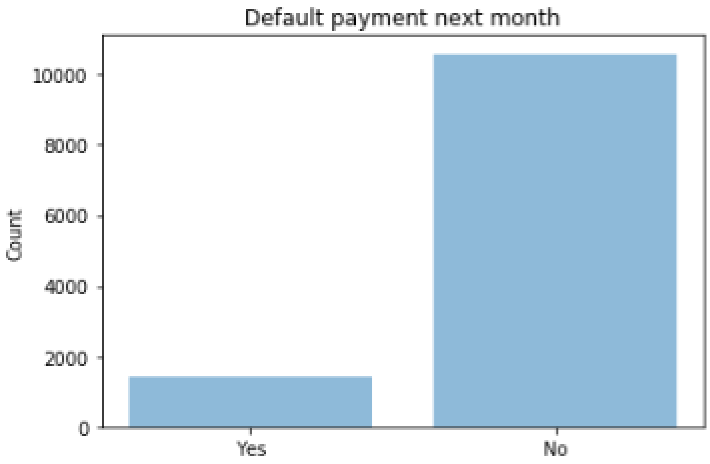
After applied all the methods, the best method selected was the Random Forest, we achieve with **82% of accuracy** the results, and also related with the higher indicator we got **81% of the better performance model**.

Model used through thru the process:



	Higher is an indicator of a better performing model	Score	Accuracy
RANDOM FOREST	81%	100%	82%
LINEAR REGRESSION	78%	78%	78%
SVR	77%	78%	78%

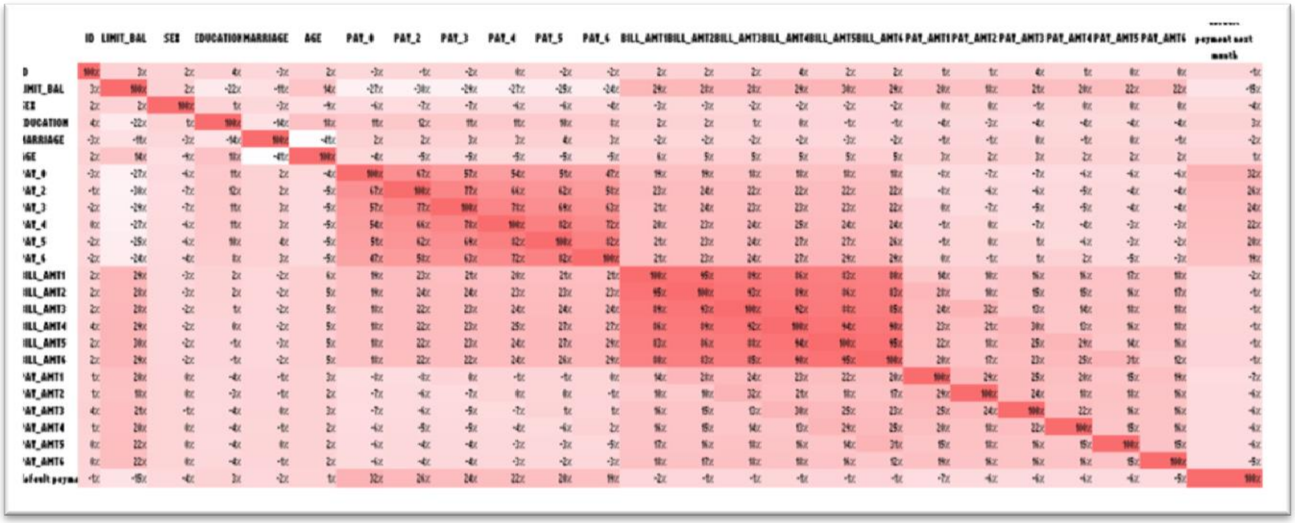
The results obtained with the test data shows the following results.



Observations:

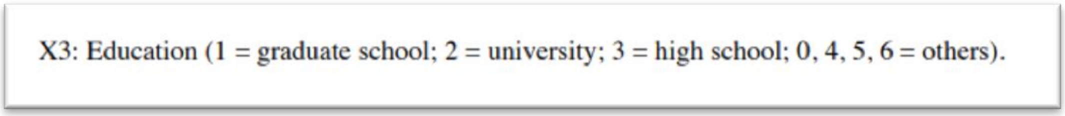
1) On this project we did not delete any column, based on the Correlation Matrix, we saw some weird behaviors, with high level of correlations between the independent variables, but, we did not consider drop them off due to the dataset lost the sense, basically, we need to understand the customer and the payment, when we deleted some columns, we lost the north of the investigation and we obtained abnormal results.

Correlation Matrix Image:



	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	payment next month
D	100%	3%	2%	4%	-3%	2%	-3%	-1%	-2%	0%	-2%	-2%	2%	2%	2%	4%	2%	2%	1%	1%	4%	1%	0%	0%	-1%
IMIT_BAL	3%	100%	2%	-22%	-11%	14%	-27%	-30%	-29%	-27%	-25%	-24%	29%	28%	28%	29%	30%	29%	20%	18%	21%	20%	22%	22%	-15%
EX	2%	2%	100%	1%	-3%	-9%	-6%	-7%	-7%	-6%	-6%	-4%	-3%	-3%	-2%	-2%	-2%	-2%	0%	0%	-1%	0%	0%	0%	-4%
DUCATION	4%	-22%	1%	100%	-14%	18%	11%	12%	11%	11%	10%	8%	2%	2%	1%	0%	-1%	-1%	-4%	-3%	-4%	-4%	-4%	-4%	3%
ARRIAGE	-3%	-11%	-3%	-14%	100%	-41%	2%	2%	3%	3%	4%	3%	-2%	-2%	-2%	-2%	-3%	-2%	-1%	-1%	0%	-1%	0%	-1%	-2%
GE	2%	14%	-9%	18%	-41%	100%	-4%	-5%	-5%	-5%	-5%	-5%	6%	5%	5%	5%	5%	5%	3%	2%	3%	2%	2%	2%	1%
AY_0	-3%	-27%	-6%	11%	2%	-4%	100%	67%	57%	54%	51%	47%	19%	19%	18%	18%	18%	18%	-8%	-7%	-7%	-6%	-6%	-6%	32%
AY_2	-1%	-30%	-7%	12%	2%	-5%	67%	100%	77%	66%	62%	58%	23%	24%	22%	22%	22%	22%	-8%	-6%	-6%	-5%	-4%	-4%	26%
AY_3	-2%	-29%	-7%	11%	3%	-5%	57%	77%	100%	78%	69%	63%	21%	24%	23%	23%	23%	22%	0%	-7%	-5%	-5%	-4%	-4%	24%
AY_4	0%	-27%	-6%	11%	3%	-5%	54%	66%	78%	100%	82%	72%	20%	23%	24%	25%	24%	24%	-1%	0%	-7%	-4%	-3%	-3%	22%
AY_5	-2%	-25%	-6%	10%	4%	-5%	51%	62%	69%	82%	100%	82%	21%	23%	24%	27%	27%	26%	-1%	0%	1%	-4%	-3%	-2%	20%
AY_6	-2%	-24%	-4%	8%	3%	-5%	47%	58%	63%	72%	82%	100%	21%	22%	24%	27%	29%	29%	0%	-1%	1%	2%	-5%	-3%	19%
ILL_AMT1	2%	29%	-3%	2%	-2%	6%	19%	23%	21%	20%	21%	21%	100%	95%	89%	86%	83%	80%	14%	10%	16%	16%	17%	18%	-2%
ILL_AMT2	2%	28%	-3%	2%	-2%	5%	19%	24%	24%	23%	23%	22%	95%	100%	93%	89%	86%	83%	28%	10%	15%	15%	16%	17%	-1%
ILL_AMT3	2%	28%	-2%	1%	-2%	5%	18%	22%	23%	24%	24%	24%	89%	93%	100%	92%	88%	85%	24%	32%	13%	14%	18%	18%	-1%
ILL_AMT4	4%	29%	-2%	0%	-2%	5%	18%	22%	23%	25%	27%	27%	86%	89%	92%	100%	94%	90%	23%	21%	30%	13%	16%	18%	-1%
ILL_AMT5	2%	30%	-2%	-1%	-3%	5%	18%	22%	23%	24%	27%	29%	83%	86%	88%	94%	100%	95%	22%	18%	25%	29%	14%	16%	-1%
ILL_AMT6	2%	29%	-2%	-1%	-2%	5%	18%	22%	22%	24%	26%	29%	80%	83%	85%	90%	95%	100%	20%	17%	23%	25%	31%	12%	-1%
AY_AMT1	1%	20%	0%	-4%	-1%	3%	-8%	-8%	0%	-1%	-1%	0%	14%	28%	24%	23%	22%	20%	100%	29%	25%	20%	15%	19%	-7%
AY_AMT2	1%	18%	0%	-3%	-1%	2%	-7%	-6%	-7%	0%	0%	-1%	10%	10%	32%	21%	18%	17%	29%	100%	24%	18%	18%	16%	-6%
AY_AMT3	4%	21%	-1%	-4%	0%	3%	-7%	-6%	-5%	-7%	1%	1%	16%	15%	13%	30%	25%	23%	25%	24%	100%	22%	16%	16%	-6%
AY_AMT4	1%	20%	0%	-4%	-1%	2%	-6%	-5%	-5%	-4%	-6%	2%	16%	15%	14%	13%	29%	25%	20%	18%	22%	100%	15%	16%	-6%
AY_AMT5	0%	22%	0%	-4%	0%	2%	-6%	-4%	-4%	-3%	-3%	-5%	17%	16%	18%	16%	14%	31%	15%	18%	16%	15%	100%	15%	-6%
AY_AMT6	0%	22%	0%	-4%	-1%	2%	-6%	-4%	-4%	-3%	-2%	-3%	18%	17%	18%	18%	16%	12%	19%	16%	16%	16%	15%	100%	-5%
default payme	-1%	-15%	-4%	3%	-2%	1%	32%	26%	24%	22%	20%	19%	-2%	-1%	-1%	-1%	-1%	-1%	-7%	-6%	-6%	-6%	-6%	-5%	100%

2) In the data source document where this explains every single field, we found some duplicated information in the Education field. To solve this situation, and improve our analysis, we replace all the values in 4, 5 and 6 by 0. With this change we could get better results and minimize the errors in the results.



X3: Education (1 = graduate school; 2 = university; 3 = high school; 0, 4, 5, 6 = others).

Recommendations:

- If the company make the decision and put in action the new model, we can improve the current process to define the default and no default customers.
- The machine learning is a continuous improvement process where it very important understands every single day the new features and improves our models.
- The method selected it has 82% of accuracy, and the final results are strong and consistent, consider that a new tool in the way to answer the current and the future questions, but remember to combine this results with the feedback of the Subject Matter Experts.

End.