# Capstone Project

## Credit Card Default Prediction

TEAM MEMBER

[ NARAYAN SINGH PARMAR , SUMIT GAIKWAD, SAGAR JAIN ]

# Contents

- Introduction
- Exploratory Data Analysis (EDA)
- Data Visualization
- Conclusions

# Introduction

The aim of this study is to exploit some supervised machine learning algorithms to identify the key drivers that determine the likelihood of credit card default, underlining the mathematical aspects behind the methods used. Credit card default happens when you have become severely delinquent on your credit card payments. In order to increase market share, card-issuing banks in Taiwan over-issued cash and credit cards to unqualified applicants. At the same time, most cardholders, irrespective of their repayment ability, the overused credit card for consumption and accumulated heavy credit and debts

The goal is to build an automated model for both identifying the key factors, and predicting a credit card default based on the information about the client and historical transactions. The general concepts of the supervised machine learning paradigm are later reported, together with a detailed explanation of all techniques and algorithms used to build the models. In particular, Logistic Regression, Random Forest, Classifires , Conpusion metrix , Recall, KNN , ENSEMBLES Voting and Support Vector Machines algorithms have been applied.

# Exploratory Data Analysis (EDA)

The dataset used in this study is the Default of credit card clients from the UCI machine learning repository, available at the following link.
It consists of 30000 observations that represent distinct credit card clients. Each observation has 24 attributes that contain information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

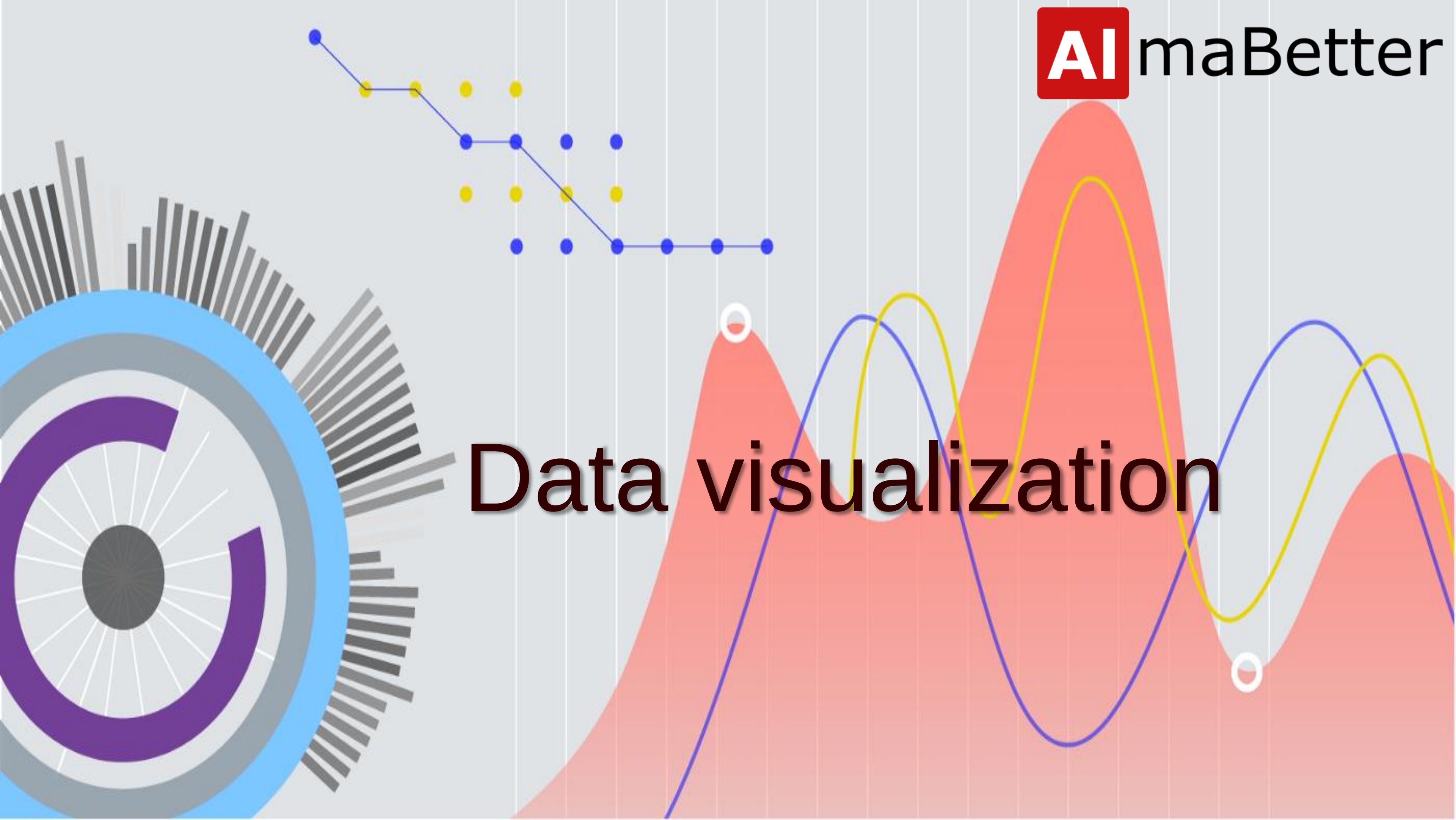
The first group of variables contains information about the client personal information:

ID: ID of each client, categorical variable
LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit)
SEX: Gender, categorical variable (1=male, 2=female)
EDUCATION: level of education, categorical variable (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
MARRIAGE: Marital status, categorical variable (1=married, 2=single, 3=others)
AGE: Age in years, numerical variable
The following attributes contains information about the delay of the past payment referred to a specific month:
PAY_0: Repayment status in September 2005 (-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)
PAY_2: Repayment status in August 2005 (same scale as before)

PAY_3: Repayment status in July 2005 (same scale as before)
PAY_4: Repayment status in June 2005 (same scale as before)
PAY_5: Repayment status in May 2005 (same scale as before)
PAY_6: Repayment status in April 2005 (same scale as before)
Other variables instead consider the information related to the amount of bill statement (i.e. a monthly report that credit card companies issue to credit card holders in a specific month):

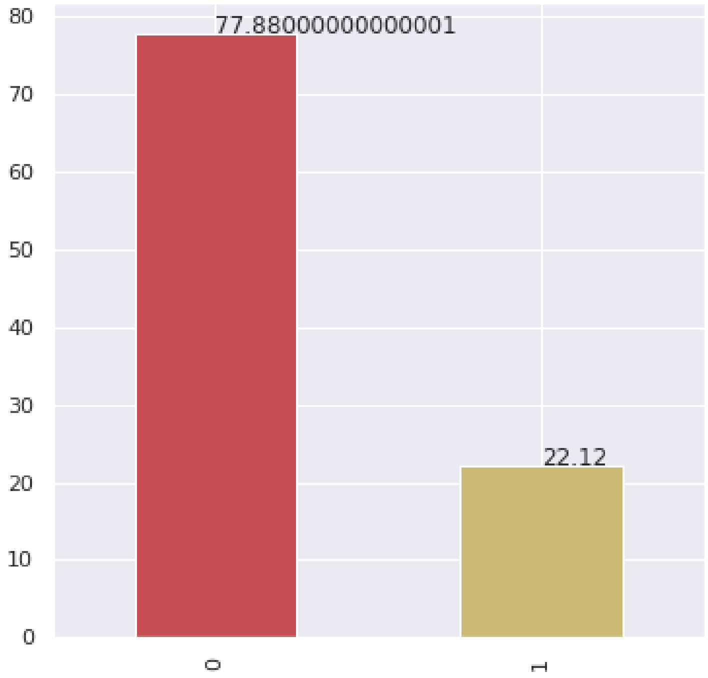
BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)
BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)
BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)
BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)
BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar)
BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)
The following variables instead consider the amount of previous payment in a specific month:

PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)
PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)
PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar)
PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar)
PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar)
PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)
The last variable is the one to be predicted:



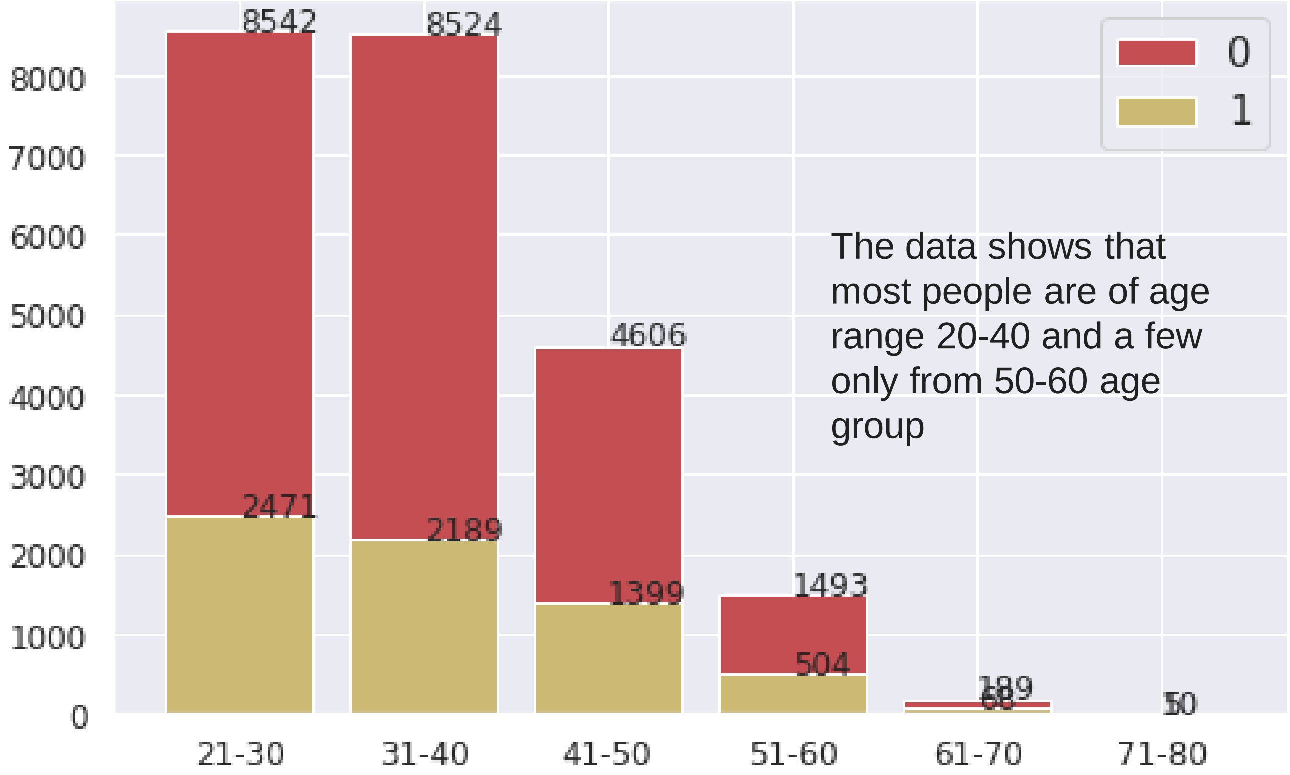
# Data visualization

# Defaulters Percentage



which is set to “0” for non-defaulters and “1” for defaulters so we have 22.12% defaulters in our dataset and 77.88% persons are non defaulters

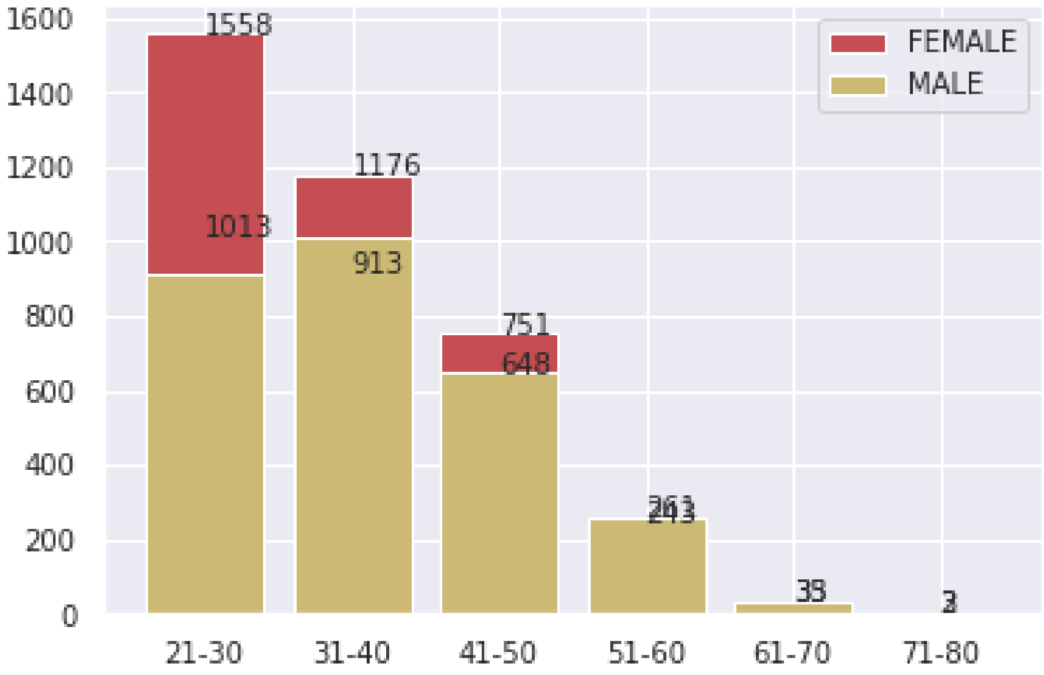
# Number of clients in each age group



The data shows that most people are of age range 20-40 and a few only from 50-60 age group

maximum clients from 21-30 age group followed by 31-40. Hence with increasing age group the number of clients that will default the payment next month is decreasing.

# Number of defaulters order by Sex

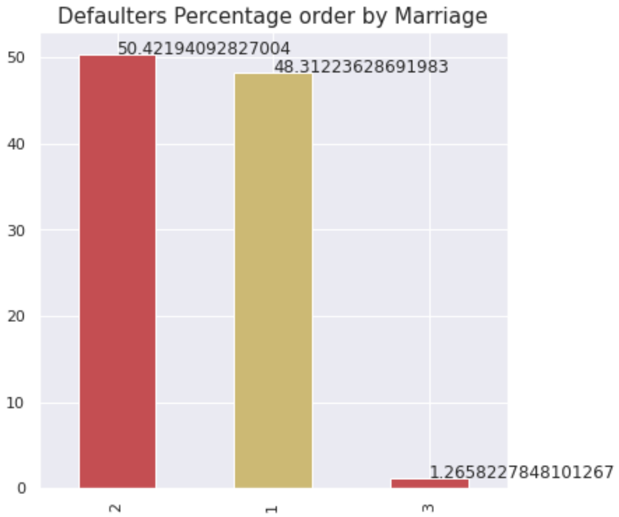


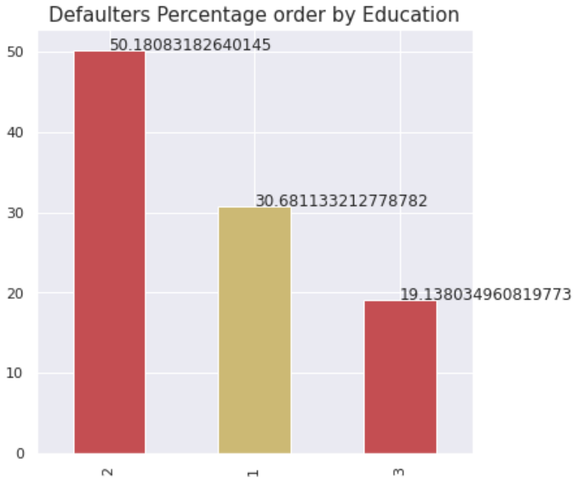
- Number of defaulters order by sex show 21-30 female is more the other group of age and other group of age are ratio is low

# Defaulters Percentage order by Marriage and Education

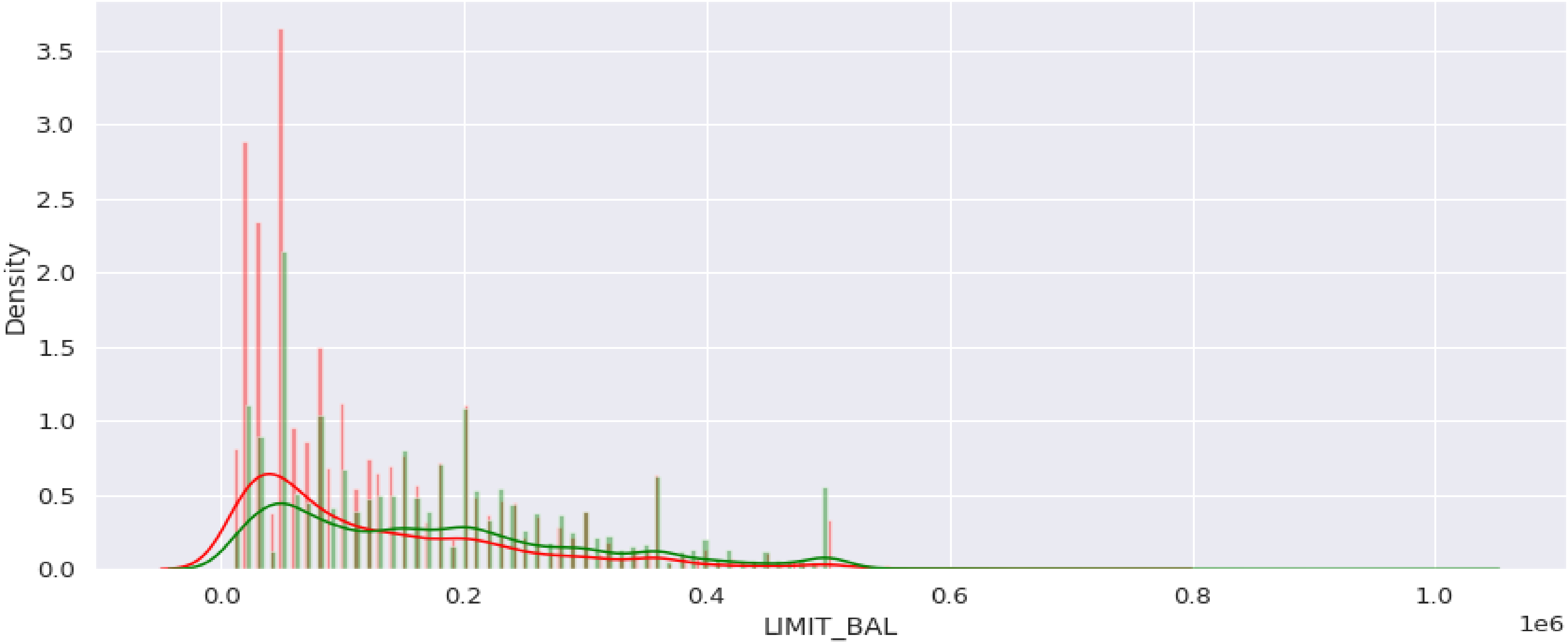
1 Regarding the attribute EDUCATION there are three categories not listed in the description of the dataset provided by the UCI website that corresponds to 0, 5, and 6.

2 While for MARRIAGE we can notice the presence of category 0 that does not correspond to any categories previously described


Defaulters Percentage order by Marriage

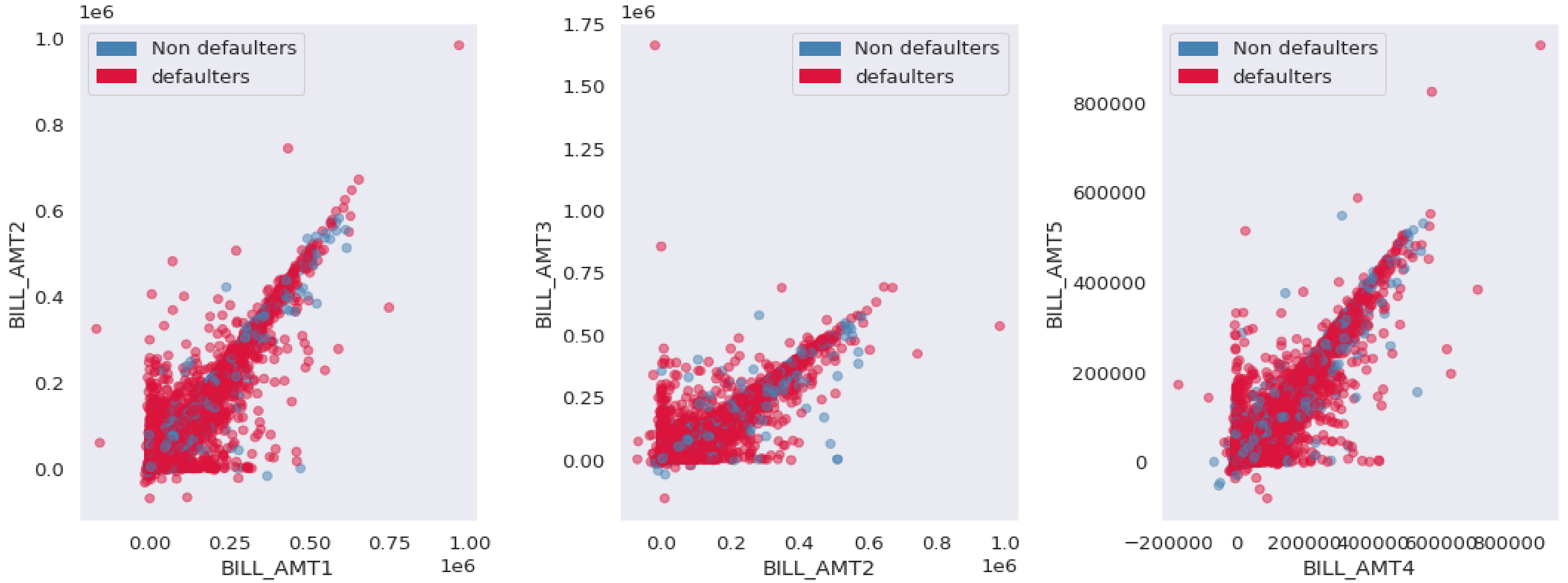

Defaulters Percentage order by Education

# Defaulters amount of credit limit - Grouped by Payment Next Month (Density Plot)



- The very high value of standard deviation has been further investigated. As can be seen, most of defaults are for credit limits 0-100,000 (and density for this interval is larger for defaults than for non-defaults)

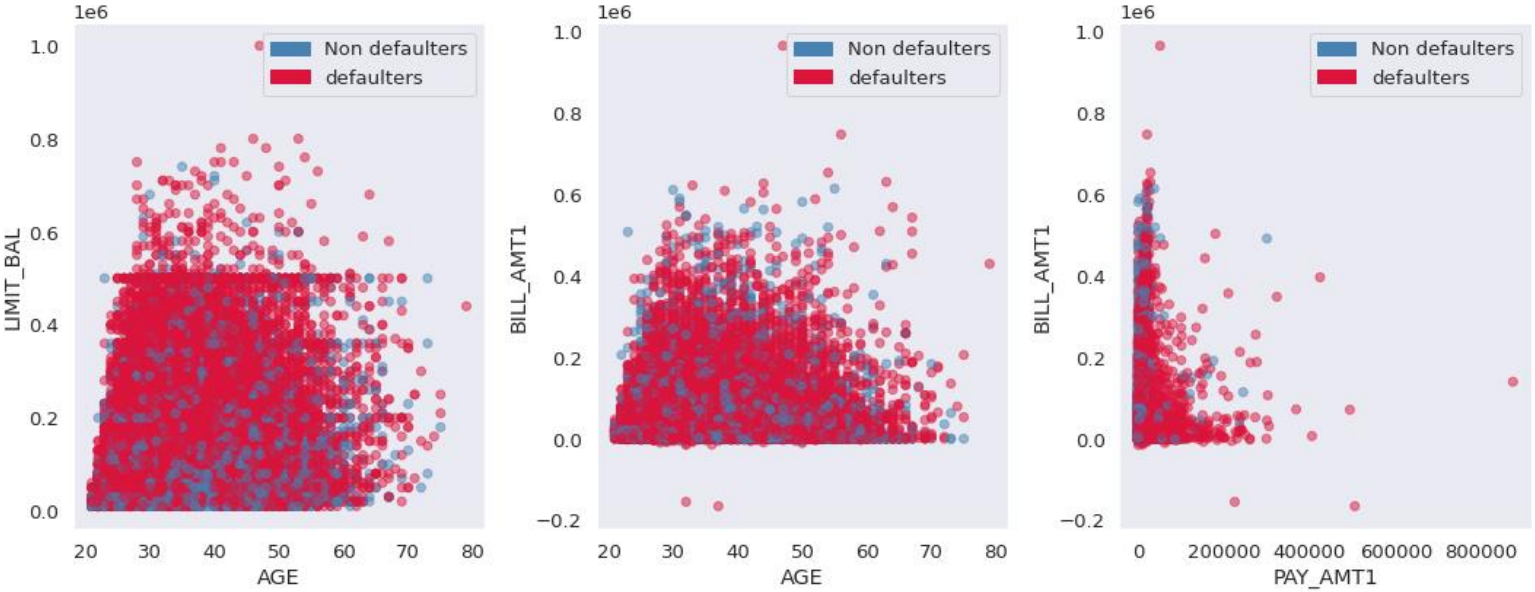
# Distribution of correlated features (scatter plot)



As shown in the correlation matrix above, some features show high correlations with each other. In particular, there exist an high positive correlation among the BILL_AMTn features, for example:

BILL_AMT1 and BILL_AMT2 have ρ = 0.95
BILL_AMT2 and BILL_AMT3 have ρ = 0.93
BILL_AMT4 and BILL_AMT5 have ρ = 0.94

# *Distribution of uncorrelated features*



The charts confirm what expected, the features in the same graph shows a linear trend as the Pearson coefficient suggested, indicating they encode pretty similar information.For completness, also charts of non-strongly correlated features are reported

AGE and LIMIT_BAL ρ=0.14

AGE and BILL_AMT1 ρ=0.055

PAY_AMT1 and BILL_AMT1 ρ= 0.099

# Conclusions

In this study different supervised learning algorithms have been inspected and presented with their mathematical details, and finally used on the UCI dataset to build a classification model that is able to predict if a credit card clients will default in the next month. Data preprocessing makes algorithms perform slightly better than when trained with original data: in particular, PCA results are approximately the same, but the computational cost has been lowered. Oversampling and undersampling techniques has been combined with PCA to assess the dataset imbalance problem. Oversampling as mentioned performed slightly better w.r.t. the undersampling, this is likely because the model is trained on a large amount of data. However, all the models implemented achieved comparable results in terms of accuracy.