Credit Team Project Proposal
Give me some credit

Eden Belouadah, Chahir Bensmail, Adrien Pavao,
Ghiles Sidi Said, Taycir Yahmed

October 2017

1 Background

In an economic and financial world full of risks and challenges, the task of detecting potential threats has become an urgence, especially in the banking sector.

The goal of this project is to be able to decide if a loan should be granted or not by taking into account some informations about the borrower. The project we propose is a closed kaggle challenge named *Give me some credit (Sep 19, 2011)*. We have chosen this challenge because it is related to the finantial domain. Furthermore, it treats a serious real-world problem. And because we are interested in fraud and risk detection, we found that this topic is ideal to work on. Resolving such problems helps improve the security in the financial domain, and leads to more confidence between customers and companies.

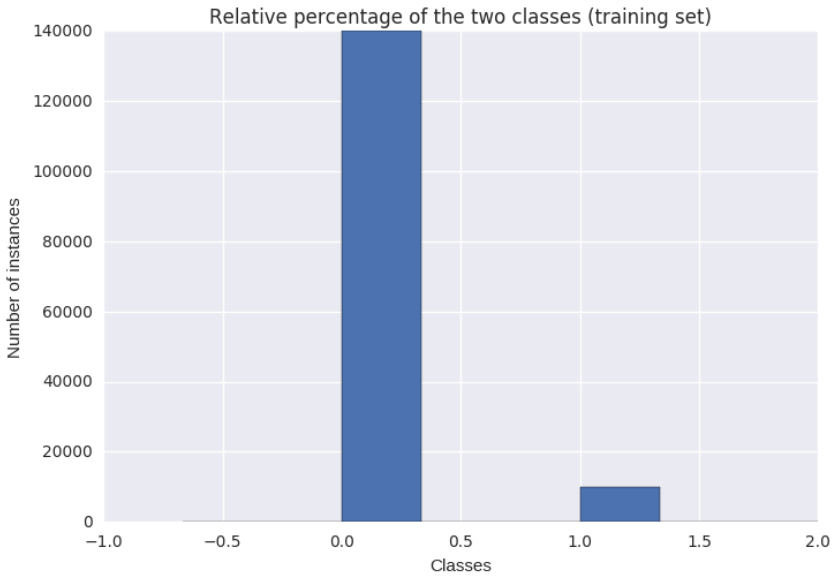
2 Material and method

The main task is to exploit informations about borrowers expressed as 10 features of different types (integer, pourcentage, real) to predict the class whose type is binary, it means that it can take the values *true* or *false*. The description of the different features is as follows:

- *RevolvingUtilizationOfUnsecuredLines* : Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits,
- *age* : Age of borrower in years,
- *NumberOfTime30-59DaysPastDueNotWorse* : Number of times borrower has been 30-59 days past due but no worse in the last 2 years,
- *DebtRatio* : Monthly debt payments, alimony,living costs divided by monthy gross income,
- *MonthlyIncome* : Monthly income

- *NumberOfOpenCreditLinesAndLoans* : Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards),
- *NumberOfTimes90DaysLate* : Number of times borrower has been 90 days or more past due,
- *NumberRealEstateLoansOrLines* : Number of mortgage and real estate loans including home equity lines of credit,
- *NumberOfTime60-89DaysPastDueNotWorse* : Number of times borrower has been 60-89 days past due but no worse in the last 2 years,
- *NumberOfDependents* : Number of dependents in family excluding themselves (spouse, children etc.),
- ***SeriousDlqin2yrs* : Person experienced 90 days past due delinquency or worse. It's the feature we want to predict.**

Figure 1: Classes distributions



We have 150000 examples in the training set and 101503 examples in the test set. Besides, we notice that the two variables *MonthlyIncome* and *NumberOfDependents* have a considerable percentage of missing values. To handle this issue, we can use various methods including statstical imputation and regression imputation.

Figure 2: Percentage of NaN by attribute

% of NaN	training set	test set
MonthlyIncome	19.82	19.80
NumberOfDependents	2.62	2.59

Moreover, we choose the *Area Under the ROC curve (AUC)* as an evaluation metric of the model, because the classes are unbalanced (less than 7% of the dependent variable values are equal to 1, more than 93% are equal to 0).

3 Preliminary results

Below are the results for different algorithms using a simple pipeline.

Figure 3: AUC scores by algorithm

% Algorithms	AUC score
LogisticRegression	0.50
RandomForestClassifier	0.57
GradientBoostingClassifier	**0.59**
ExtraTreesClassifier	0.56
MLPClassifier	0.50
AdaBoost	0.59
GaussianNB	0.50

From the table, we see that the algorithm that gives the best results is GradientBoosting-Classifier with an AUC score of 0.59. In the Kaggle challenge, the best AUC score obtained is 0.869558, so there is a large margin of progress.