# A Machine Learning Score for non listed Companies Based on Balance Sheets

***Juan Barragan***
*Haussman Consulting*

## Contents



## 1. Introduction

According to the Banks Federation (Fédération des Banques Françaises) The amount of loans given to Non Listed Small Caps over one year (from July 2015 to July 2016) in France stands at 385,2 billions Euros. And that liquidity helps to finance more than one million businesses. One of the challenges for giving credit to non listed companies is to assess their credit risk. On this paper we propose a *scoring model* for measuring the *Default Risk* using preceding balance sheets.

We are going to use big data techniques, namely *Machine Learning* for selecting key financial indicators allowing to predict default. We synthesize these indicators on a unique number, the *score*. This score is going to reflect, for a given enterprise, the default risk. In addition, the score combined with a zero coupon curve for a given country can be easily correlated to a rate of return for loans.

## 2. The data set

The French central bank have a data set of French enterprises, but this data base is not generally available [1]. In looking for data-sets, we have found that

[1] This has changed as of July 2017. French government has release through the INPI a data set of balance sheets. But

the Central Bank of Belgium through the Balance Sheet Center has a complete data-set with several decades of historical data.

### 2.1. Description of the Enterprises

Among all the balance sheets, we have selected the enterprises which are appropriate for a financial study. These are:

- European Cooperative Society
- Cooperative Society with Unlimited Liability
- Cooperative Society with Limited Liability
- Participating Cooperative Society with Limited Liability
- Private Limited Company
- Partnership Company
- Partnership Company by Actions

To these societies, we consider the status Either the society is in normal situation, either she is in default:

- Default (Declaration, opening)
- Default ended excusably.
- Default ended non excusably.

This will allow us to analyze the societies with healthy status or with real default. We avoid enterprises who disappear by fusion, or wanted dissolution, scission, etc. Still among that subset we want to concentrate our attention to small and medium enterprises, these are societies whose size:

- Have less than 250 employees,
- Either annual earning are less than 50 millions Euros, either their balance sheet is less than 40 millions euro's. And for avoiding very small enterprise, which are really fragile,
- We excluded enterprises whose number of employees is less than or equal 40.

this includes only one year of historical data, not enough for building a model

## 3. Tools

### 3.1. Software Tools

For realizing this note, we have used techniques from *Machine Learning*, through software tools as python and scikit. A parallel computing engine Spark. The date we have analyzed is nos structured neither fixed in size so we choose a NoSQL data base mongo.

Machine Learning is a branch with origins in Engineering and Computer Science which focuses in finding patters for classification using tools from Statics, Analysis and Geometry. Machine Learning will be very nicely adapted for us. We are going to explore several millions of accounting data looking for patterns allowing us to predict default given evidence in the balance sheet. This is made usually by financial analysts by hand with empirical criteria for determining critical thresholds. Our goal here is to formalize these methodologies by finding precise critical thresholds.

### 3.2. Principal Component Analysis

*Principal Component Analysis* is a Statistical methodology allowing us to transform a collection of vectors $X_1, X_2, \dots X_k$, $X_i \in \mathbb{R}^n$ on another collection $Y_1, Y_2, \dots Y_k$ $Y_i \in \mathbb{R}^m$, with $m \leq n$ and zero correlation among them. Vectors $Y_i$ are called *Principal Components*. The fact their dimension can be lowered keeping the variance is of great help whenever we look for visualizing a big number of observables.

Technically we can calculate these variables as follows, let $\mathbf{m}$ be the mean vector, $\mathbf{m} = 1/k \sum_{i=1}^{k} X_i$, let $\mathbf{S}$ and $\mathbf{S} = 1/k \sum_{i=1}^{k} (X_i - \mathbf{m})(X_i - \mathbf{m})^t$, the co-variance matrix, $X^t$ being the transpose. of $X$. For reducing an $m$ dimensional space is enough to calculate the $m$ eingenvectors of $\mathbf{S}$ $v_1, \dots v_m$, having the biggest eigenvalues. let $\mathbf{W} = (v_1, \dots v_m)$, be matrix of dimension $n \times m$. The space we are looking for $Y$, is given by $Y = \mathbf{W}^t X$.

### 3.3. K-Means Lloyd algorithm

$K-$Means is an algorithm for finding aggregates on unlabeled data-sets. More precisely, given $m$ values $X_1, X_2, \dots X_m$ and $k$ an integer, we look for a partition of the values into $k$ classes, $S_1, S_2, \dots, S_k$ and their centers $c_1, c_2, \dots c_k$ such that

$$\sum_{i=1} \sum_{X \in S_i} \|X - c_i\|$$

be a minimum. This can be done with the Lloyd algorithm:

**Figure 1:** *K-Means classes*

0 Pick arbitrary different centers $c_1^0, c_2^0, \dots c_k^0$.

1 Define

$$S_i^n = \{X : \|X - c_i^n\| \leq \|X - c_j^n\| \quad \forall i \neq j\}$$

2 Then the new centers are

$$c_i^{n+1} = \frac{1}{|S_i|} \sum_{X \in S_i} X$$

with $|S_i|$ the cardinality of $S_i$.

3 repeat steps 1 and 2 until a given tolerance threshold.

See 1 for a $K-$means example.

### 3.4. Logistic Regression

Let again $X_1, X_2, \dots X_k$, $X_i \in \mathbb{R}^n$ be a set of points and $\lambda : \mathbb{R}^n \to \{0, 1\}$ a *labeling system*. Let $Y_i = \lambda(X_i)$, we can consider $Y$ as a Bernoulli Random Variable. Given a value $X_0$, we write the probabilities:

$$\mathbb{P}(Y = 1|X = X_0) = p,$$

$$\mathbb{P}(Y = 0|X = X_0) = 1 - p$$

or more compactly:

$$\mathbb{P}(Y = Y_0|X = X_0) = p^{Y_0}(1 - p)^{1-Y_0}. \qquad (1)$$

We have here a classification problem, If the labeling is separable, there exist an hyper-plan $L : \mathbb{R}^n \to \mathbb{R}$, $L_{\theta,\theta_0}(X) = \theta X + \theta_0$ such that

$$L_{\theta,\theta_0}(X_i) > 0, \quad \text{if} \quad Y_i = 1, \quad \text{and}$$

$$L_{\theta,\theta_0}(X_i) < 0, \quad \text{if} \quad Y_i = 0.$$

Using the distance to the separating hyper-plan we can smooth the probability of the Bernoulli random variable using the *sigmoid* function, $s(x) = 1/(1 + e^{-x})$. For this is enough to write for each observation $X_i$,

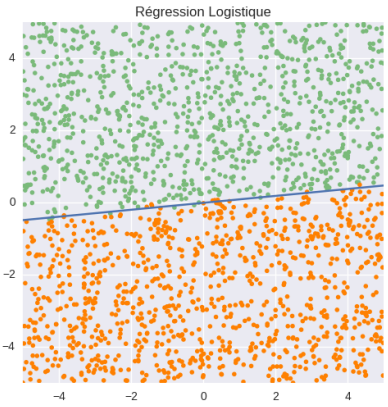


**Figure 2:** *Logistic Regression*

$$p = \mathbb{P}(Y = 1|X_i) = \frac{1}{1 + e^{-L_{\theta,\theta_0}(X_i)}}$$

So the Bernoulli random variable 1 can be written as

$$\left(\frac{1}{1 + e^{-L_{\theta,\theta_0}(X_i)}}\right)^{Y_i} \left(1 - \frac{1}{1 + e^{-L_{\theta,\theta_0}(X_i)}}\right)^{1-Y_i}$$

or

$$\frac{e^{-L_{\theta,\theta_0}(X_i)^{1-Y_i}}}{1 + e^{-L_{\theta,\theta_0}(X_i)}}$$

Now, for finding the optimal parameters, $\theta, \theta_0$, we maximize the *log-likelihood* over all the observations:

$$\mathbf{argmax}_{\theta,\theta_0} \log \prod_{i=1}^{k} \frac{e^{-L_{\theta,\theta_0}(X_i)^{1-Y_i}}}{1 + e^{-L_{\theta,\theta_0}(X_i)}}$$

Figure 2 shows an example of Logistic Regression and its separating plane.

### 3.5. Support Vector Machines (SVM)

A similar methodology is provided by the *Support Vector Machines*. Suppose we are exactly on the same situation than the preceding section, with a set of observations $X_1, X_2, \dots X_k$, $X_i \in \mathbb{R}^n$. But now we change slightly our labels $\lambda : \mathbb{R}^n \to \{1, -1\}$. If a separating plane does exists, $L : \mathbb{R}^n \to \mathbb{R}$, $L_{\theta,\theta_0}(X) = \theta X + \theta_0$ such that $\text{sign}(L_{\theta,\theta_0}(X_i)) = \lambda(X_i)$, then we have for all $i$

$$\frac{\lambda_i(X_i)L(X_i)}{\|\theta\|} > 0 \qquad (2)$$

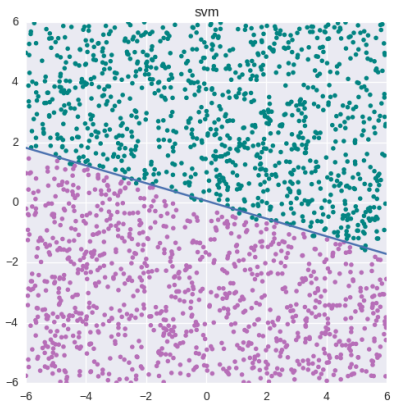


**Figure 3:** *Support Vector Machine*

We recognize on equation 2 the distance from point $X_i$ to the hyper-plan $L_{\theta,\theta_0}$. The support Vector Machine looks for maximizing:

$$\mathbf{argmax}_{\theta,\theta_0}\left\{\frac{1}{\|\theta\|}\mathbf{min}_i\lambda(X_i)L_{\theta,\theta_i}(X_i)\right\} \quad (3)$$

In order to simplify 3, we change the scale of the plan as $\theta \to \kappa\theta$ and $\theta_0 \to \kappa\theta_0$ so that if $X_p$ is the point realizing the minimum $\mathbf{min}_i\lambda(X_i)L_{\theta,\theta_i}(X_i)$, then for that point $X_p$ we have $\lambda(X_p)L_{\theta,\theta_0}(X_p) = 1$. Hence the optimization program 3 can be written:

$$\mathbf{argmax}_{\theta,\theta_0}\frac{1}{\|\theta\|}, \quad i = 1,\dots k$$

or equivalently:

$$\mathbf{argmin}_{\theta,\theta_0}\frac{1}{2}\|\theta\|^2, \quad i = 1,\dots k$$

with the constraints

$$\lambda(X_i)L_{\theta,\theta_0}(X_i) \geq 1 \quad i = 1,\dots k$$

Figure 3 shows an example of separating plane calculated with support vector machines.

# 4. Analysis of Financial Ratios

Creditworthiness analysis is realized by looking at financial ratios on the balance sheet, experts select the ones being relevant for financial health.

A pioneer in doing scoring is Altman see [1]. For detecting when an enterprise is near default, the author proposes a criteria using discriminant on financial ratios. Altman's discriminant is an hyper-plan $Z = V_1X_1 + V_2X_2 + \dots V_nX_n$, where $V_i$ are the discriminant coefficients and $X_i$ the independent variables coming from balance sheet financial ratios. Altman score is directly linked to the creditworthiness of the enterprise, the actual hyper-plan found by Altman is

$$\begin{aligned} Z = \; & 0.012X_1 + 0.014X_2 + \\ & 0.033X_3 + 0.006X_4 + 0.999X_5 \end{aligned}$$

with

$X_1$ Working capital over total assets. This measures liquid assets in relation to the size of the company.

$X_2$ Retained earnings over total assets. Which measures profitability.

$X_3$ EBDITA over total assets. EBDITA being earnings before interest and taxes, which measures operating efficiency apart from tax.

$X_4$ Market value of equity over book value of liabilities

$X_5$ Sales over total assets.

Altman found that the ratio profile for the bankrupt group fell at $-0.25$ in average, and for the non-bankrupt group at $+4.48$ in average. Our aim is to generalize this discriminant hyper-plane using Machine Learning, we are going notably to use Support Vector Machines (SVM) and Logistic regressions for estimating the coefficients.

## 4.1. Ratios of Belgium Central Bank

The number of ratios we are going to use are 21 and they are calculated by the Belgium Central Bank. A presentation of these ratios can be found on the site Centrale des Bilans

Altman's method cannot be used here as most Small Caps are not trade publicly. Our sample of Belgian Small Caps which have a balance sheet not exceeding 40 million Euros and with at least 40 employees is of size 4796. Among these 45 did default, 22 on year 2015, 22 on year 2014 an one in year 2013.

We remark an immediate difficulty here, defaults are rare, we have less than 1%. Another difficulty can be appreciated using Principal Component Analysis on the ratios of year 2013. See figure 4, defaulted enterprises on 2014 (in red) are really scattered. We can illustrate the difficulties of being rare and scattered using naively Machine Learning tools to that sample. So using the 2013 ratios and separating enterprises healthy on 2014 and in default on 2014, we

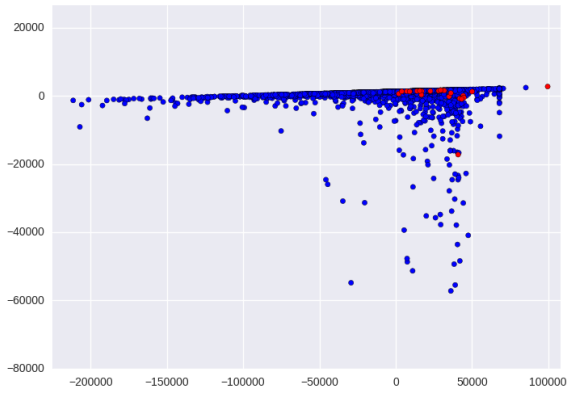


**Figure 4:** *PCA 2D, ratios 2014*

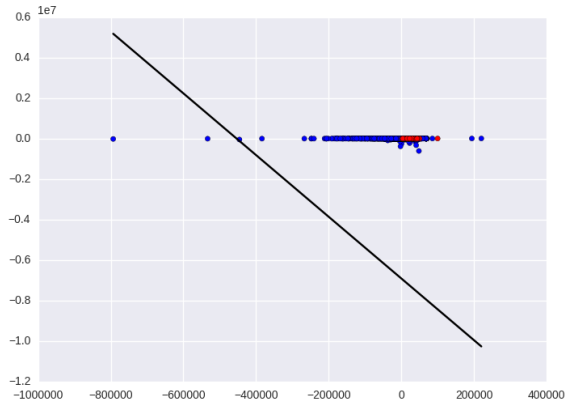


**Figure 5:** *Logistic Regression*

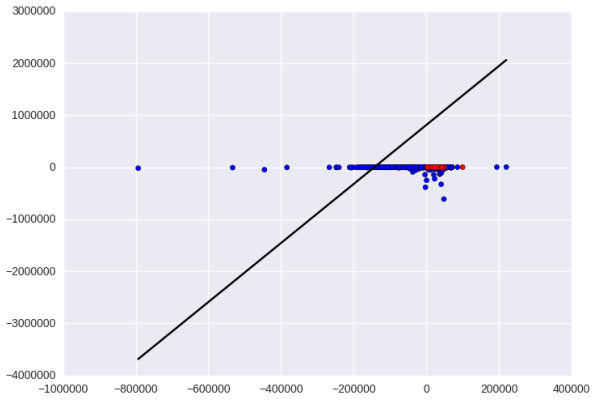


**Figure 6:** *Support Vector Machine*

use logistic regression 5 and support vector machines 6 projected on two dimensional plan.

As we can see results are mediocre, no method can separate the ratios.

### 4.2. Selecting a better sample

Inspired by the methodology of Altman we want to get ride of the seen difficulties by restricting our attention to a better sample. We take the whole defaulted enterprises on 2014 and working with 2013 ratios, we are going to better select the healthy ones. Let $\mathcal{E}$ be the set of all enterprises, $\mathcal{S}$ the healthy ones and $\mathcal{D}$ those being on default on 2014. Associated to each enterprise $e$ there exists a set of ratios $r_i(e)$. Let

$$R_i = \max\{r_i(e) | e \in \mathcal{D}\},$$
$$E_i = \{e \in \mathcal{S} : r_i(e) \geq R_i\}$$

We look for a subset of ratios, $r_i$ such that $\cap E_i \neq \varnothing$ and such that the cardinality be comparable to the enterprises on default $\# \cap E_i \sim \#\mathcal{D}$.

Using Spark we have calculated all the intersections. Additionally, for helping our analysis, we have done bidimentional scatterplots and histograms for selecting the best discriminants, (See appendix C) For a concrete example, we have selected defaulted enterprises on 2014 and we have selected healthy enterprises with the following best 2013 ratios:

- Net margin over sales
- Added value by employee,
- Liquidity (Treasury, accounts payable, available values, value of stocks)
- Capital over total asset.

We have found there are 29 healthy enterprises. Applying logistic regression and realizing a contour plot representing the probability of survival, all projected on two dimensions using principal components. See figure 7. This is very encouraging.

## 5. Scoring

Now, we would like to explore all smalls data set, and for doing so we we would like to extend the logistic regression to all enterprises. We want a *Score* $R = R(e)$, such that if $\mathbb{P}_t(e|r) = \mathbb{P}_t(e|r_i, i = 1, \dots n)$ is the survival probability on future year $t$ of the enterprise $e$ given its actual ratios $r = r_1, \dots r_n$, then $R$ should be that if

$$R(e_1) \leq R(e_2)$$

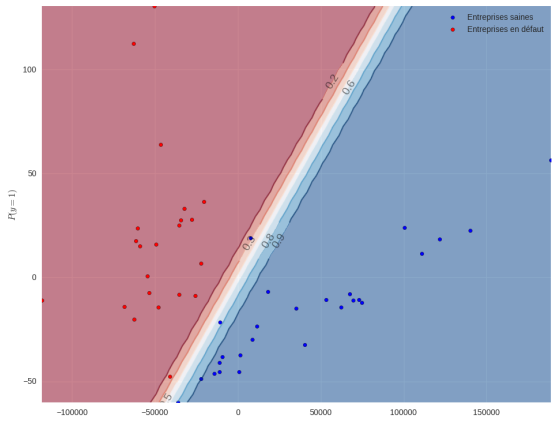


**Figure 7:** *Logistic Regression and Survival Probability*

then

$$\mathbb{P}_t(e_1|r) \leq \mathbb{P}_t(e_2|r)$$

In words: a better score means a better probability of survival.

Given a classificator in Machine Learning we traditionally measure its classification error. Usually, good classificators made less than 5% errors. But this measure is meaningless for us. Defaulted enterprises on a year are less than 0.5%. A dumb classification putting every enterprise as healthy will made only 0.5% errors.

So better than simple percentage of error, we are interested on what is called the *True Positives* and *False Positives* namely:

> True Positives (TP), which are healthy enterprises well classified,
>
> False Positives (FP), which are defaulted enterprises well classified.

A way of synthesizing this concept is a *Receiver Operating Characteristic*, ROC curve. Let us suppose we have a score $R = r(e)$ ans say that the *default threshold* is $R_0$. Meaning if $R(e) \leq R_0$ then enterprise $e$ will default next year. Then the ROC curve is the parametric curve

$$R_0 \to (FPR(R_0), TPR(R_0)$$

Where $TPR(R_0$ is the *True Positive Ratio* (True Positives over All Positives) and $FPR(R_0)$ is the *False Positive Ratio* (False Positives over all Negatives). A cautious investor is more concerned by the False Positive Rate as it measures how well bad enterprises are detected. Let $E$, be the set of all enterprises and $R = R(e)$, a score. Associate to this score the distribution function

$$F_R(x) = \frac{1}{|E|}|\{e \in E|R(e) \leq x\}|$$

Our metric $M$ for looking for a good score assigns to the set of all defaulted enterprises $D$ a low number

$$M = 1 - \frac{1}{|D|}\sum_{e \in D} F_R(e)$$

## 5.1. Natural Score

Our Score is inspired by Altman. Consider the plane obtained by logistic regression:

$$L_{\theta,\theta_0}(X) = \sum_{i=1} \theta_i X_i + \theta_0$$

This plan separates good enterprises form bad ones. The distance from a point to this plan is given by

$$d(X, L_{\theta,\theta_0}) = \frac{L_{\theta,\theta_0}(X)}{\|\theta\|}$$

This distance is hence a good indicator of the proximity of a given enterprise to the risk zone.

So if $r_1, \ldots r_n$ are the ratios of the enterprise, our score is defined by:

$$R(e) = \sum_{i=1}^{n} \theta_i r_i$$

with $\theta = (\theta_1, \ldots, \theta_n)$ the discriminant coefficients of the logistic regression. We have 21 discriminant (ratios) resulting in $2^{21} = 2097152$ possible combinations. Certainly there are correlations among the ratios. In order to reduce the number of ratios we are going to use our preceding metric $M$ on the selected intersections. Again this is a parallel calculus done using Spark. We have around 3400 non empty intersections.

## 5.2. Results

Our best metric comes out to be obtained using the following ratios for selecting healthy enterprises, we get 3:

- Added value by employee
- Debt cost under Added Value
- Rotation of Inventories and work in progress

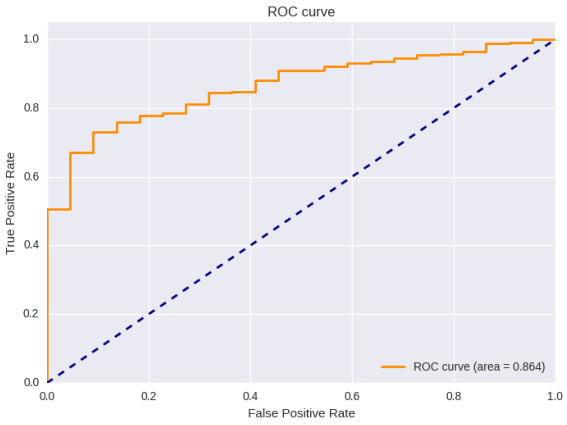


**Figure 8:** *Quality of classifications for 2014*

Using these ratios we select a hand-full of enterprises. Applying logistic regressions tho the whole combination of ratios, we discover the following best discriminants:

- Added value by employee
- net Profit of net equity after taxes.
- Gross profit of total assets before taxes and financial expenses.
- Net profit of total assets before taxes and financial expenses.
- Acquisition of tangible assets over Added Value.

For these discriminants, the coefficients of the logistic regression are:

$$HS(X) = 0.000026X_1 + 0.0035X_2 + 0.0507X_3 - 0.0248X_4 + 0.084X_5 \quad (4)$$

And we remark something interesting and surprising: the fourth ratio, Net profit of total assets before taxes and financial expenses, has a *negative, impact*. We resume the results with two graphs of ROC curves. The first one, see figure 8 shows the quality of the classifier on 2014, the learned data. And the second one, see 9 measures the quality of the classifier for predicting outcomes on 2015, our test data. With the *Area Under ROC Curve* at around 80% our classifier looks as a good one. As an illustration, lets say we place our threshold for default at the score 2.50. Then False Positive Rate is around 18% and True Positive Rate is at 58%, look at the point $(0.18, 0, 58)$ in figure 9. This means that for this threshold, we neglected $100\% - 58\% = 42\%$ of healthy enterprises. But on the other hand, we only have 18% chances of selecting enterprises who are going to default.

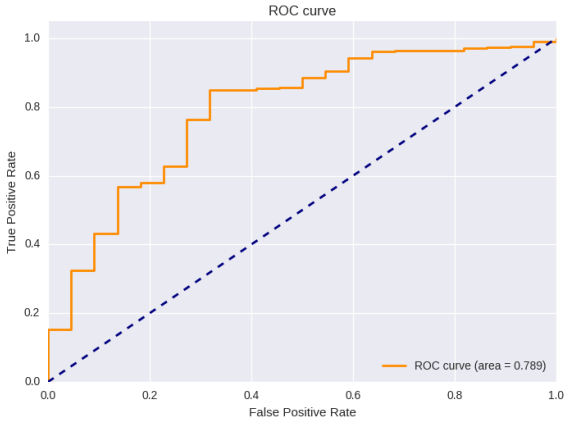


**Figure 9:** *Predictive quality of classificator for 2015*

# A. Histograms

We show aggregation of ratios for the year 2014. This is a first look into distributions of the ratios. On the height axis we plot density, and on the horizontal axis we have plot the ratio values of the defaulted enterprises. We ideally look for discriminants capturing on a given region all the defaulted enterprises. A good example can be see on fig , added value by employee. In this case, ahead of perhaps one outlier all red points are to the left of the histogram.

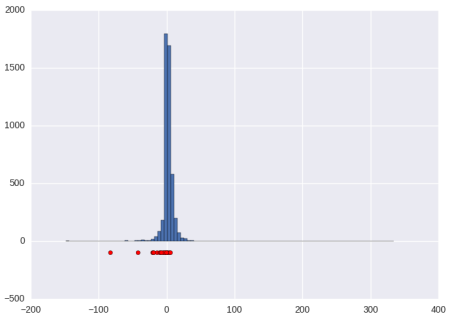


**Figure 10:** *Net margin over sales*

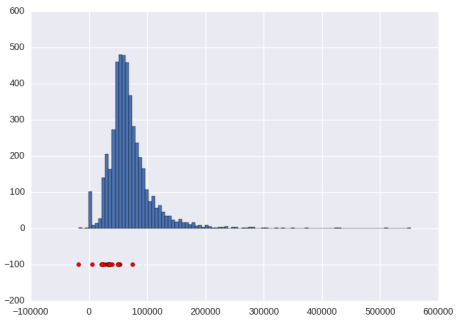


**Figure 11:** *Added value by employee*

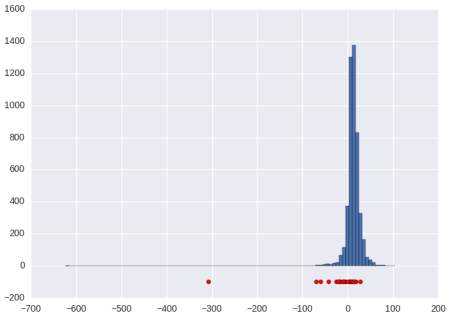


**Figure 12:** *Gross profit of net equity before taxes*

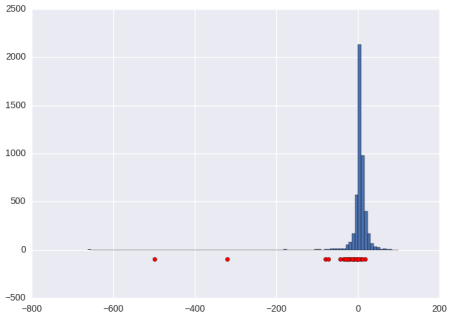


**Figure 13:** *Net Profitability of net equity before taxes*

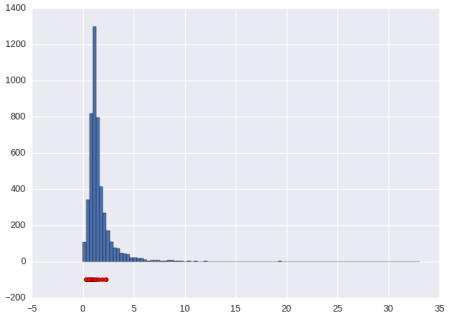


**Figure 14:** *Liquidity, wide*

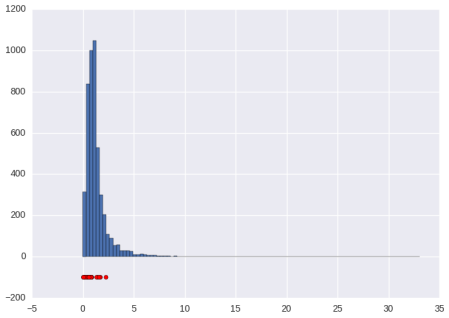


**Figure 15:** *strict Liquidity*

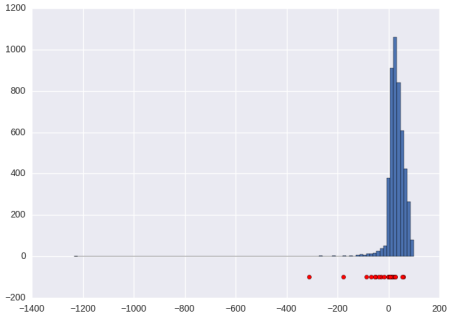


**Figure 16:** *Net Equity / Total Assets*

# B. Dispersion Graphics raster plots

These scatterplots give us a sense of how we can try to capture defaulted enterprises.

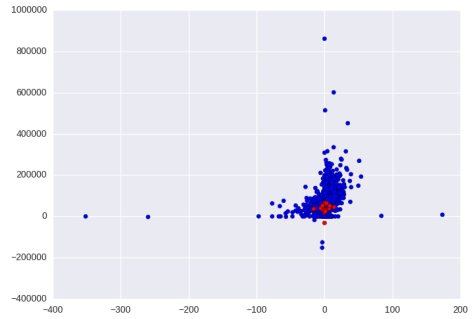


**Figure 17:** *Net margin over sales versus added value by employee*

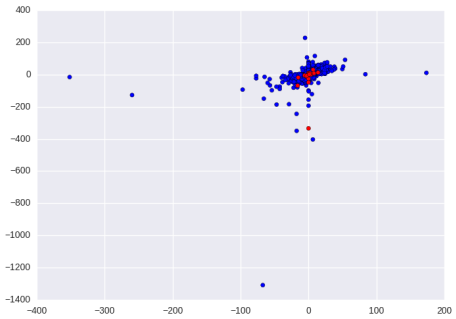


**Figure 18:** *Sales Margin versus Net Profitability of total assets before taxed and debt charges*

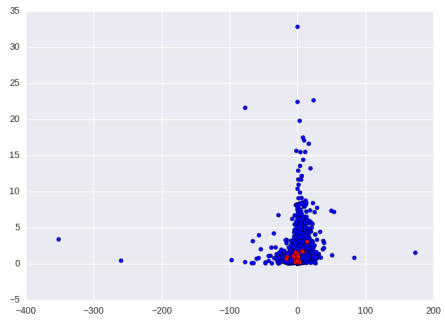


**Figure 19:** *Net Margin over sales versus Strict liquidity*

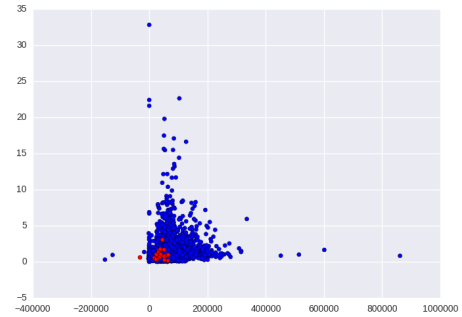


**Figure 20:** *Added value by employee versus Strict Liquidity*

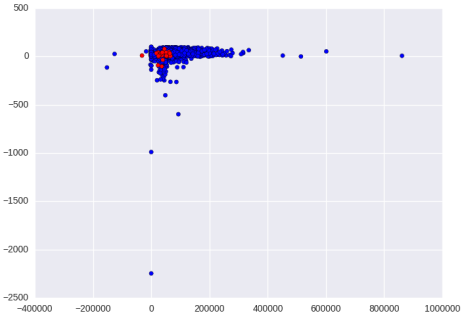


**Figure 21:** *Added value by employee versus Net Equity over Total Assets*

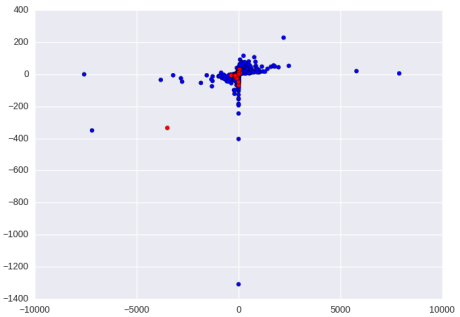


**Figure 22:** *Net Profitability Net Profitability of total assets before taxes over debt cost*

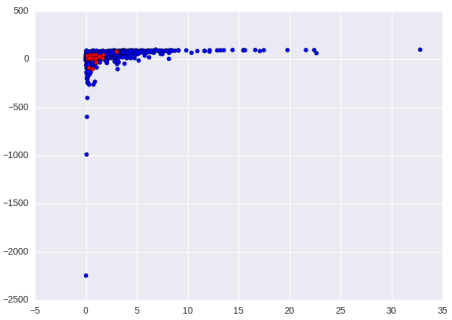


**Figure 23:** *Strict Liquidity versus Equity over Total Assets*

## References


[1] E. Altman Predicting Financial Distress of Companies: revisiting the Z-Score and Zeta Models. *New York University*

[2] C. M. Bishop. Pattern Recognition and Machine Learning *Springer*

[3] T. Hastie, R. Tibshirani, J. Friedman The Elements of Statistical Learning *Springer*

[4] S. M. Ross Introduction to Probability Models *Academc Press*

[5] C. Thibierge Analyse Financière *Vuibert*

[6] L. Wasserman All of Statistics *Springer*