Master 1 Ingénierie des risques

ISFA

Scoring et analyse financière

réalisé par
Pierre Lafon, Florian Robinet, Kacel Sofiane

Table des matières

1 Introduction

There are two types of investment : technical analysis of stock trends, and fundamental analysis.

These terms refer to two different stock-picking methodologies used for researching and forecasting the future growth trends of stocks. Like any investment strategy or philosophy, both have their advocates and adversaries. Here are the defining principles of each of these methods of stock analysis :

-Fundamental analysis is a method of evaluating securities by attempting to measure the intrinsic value of a stock. Fundamental analysts study everything from the overall economy and industry conditions to the financial condition and management of companies. We call value investors, investors that buy stocks that their price is under the intrinsic value of the company. One of the most famous and influential value investors are Warren Buffett, Benjamin Graham, or Walter Schloss. Warren Buffett is nowadays, one of the three wealthiest people on earth with a net worth between 60 billion dollars. He is considered the best investor in history. Benjamin Graham is considered the father of value investing. He wrote “Security Analysis” and “The intelligent investor” considered the bible of investment by a lot of value investors.

-Technical analysis is the evaluation of securities by means of studying statistics generated by market activity, such as past prices and volume. Technical analysts do not attempt to measure a security’s intrinsic value but instead use stock charts to identify patterns and trends that may suggest what a stock will do in the future. Charles Dow developed a series of principles for understanding and analyzing market behavior which later became known as Dow theory, the groundwork for technical analysis.

Fundamental analysis must be linked to a long term vision of investment, whereas technical analysis is rather used with very short period of time.

Now that we introduced the definition of both types of investment, we are going to tell our opinion. After the dot-come bubble in 2000 and the housing bubble in 2007, that led to dot-come collapse in 2001 and subprimes mortgage crisis in 2008, lot of stock exchange investors lost everything, because their reasoning were only based on graphics, without even knowing value of companies they were buying. As long as the market carried on going up, these investors were wining money, but when speculative bubbles collapsed, they all lost everything. Stock exchange is supposed to be a place to help companies to find financing, to be able to grow faster. But since 2000, and the arrival of Internet and non professional investors, Stock exchange has been the field of speculation and not investment. Traders only swear by technical analysis and we have seen the results. These reasons led us to choose fundamental analysis. By the way, we can see that the most successful investors use value investing, and made the biggest part of their wealth thanks to big crisis. The reason ? They buy when the market is at its lowest.

Before we get further, we need to explain the difference between “fundamental analysis” and “value investing”. Since we talk about investment strategies, we may tend to confuse the reader about these notions. Fundamental analysis is a method to try to measure the intrinsic value of a stock, whereas value investing is buying stocks under their intrinsic value. The difference is very important for the rest.

We have based the first step of our investment strategy on fundamental analysis, trying to choose companies that have good results, that we can say they are strong enough to keep their good results. But we didn't choose value investing. Why ? First, we need to wait that the price of a very good company is under its intrinsic value. And this happens rarely. Of course, very good companies are almost never rejected by the market. Secondly, if we want to find less known companies, but companies that have good results even though, we need to have at least balance sheets and income statement of the last 20 years, of the more companies possible. And this is the very complicated part. Gathering a huge amount of quality data is very long. We insist on the word ”quality”. A lot of websites collect some information about companies, but most of the time, there are some missing information, and you can hardly have more than years old information and the biggest problem is that if you compare websites, information are different.

The other problem is that we wanted to automate the research process. We don't have the time to analyse one by one every companies in the world to know which one we must buy. So, we decided to create an algorithm, that could rank companies. The aim was to go from a huge number of companies, good and bad ones, to a reduced number of companies, but only good ones. To rank them, we use a scoring algorithm. We don't care if the stocks are trading for more or less than their intrinsic value, we just rank companies by their financial results. The reason is that strong (financially speaking) companies will often remain strong and will often keep their good results. Like Warren Buffett, we are in fact trying to pick stocks that have a sustainable competitive advantage. People always carry on buying these stocks, and exchange rates continue to increase.

The second step of our investment strategy is based on technical analysis. We calculate the yield and the risk of the stocks we selected with fundamental analysis. Then, we choose what portfolio could be the best for our asset management.

2 Logistic model

In the context of population dynamics studies, Pierre François Verhulst, published in 1840 a new model which is a direct answer to Malthus who was considering the growth of the population as an exponential model. This model suppose that the population growth is only exponential for small groups. When the size of the group increase, some limiting factors appears like the quantity of food available. That assumption imply the existence of a maximal population size linked to its habitat. The resolution of the Verhulst model involved the creation of logistic functions in continuous time and logistic series in discrete time.

This discovery allowed to develop logistic law which is the source of logit model. Historically the logistic regression was the first method used to model a binary variable.

That logistic regression is really often used today despite the fact that other methods could lead more precise results. It is a particular case of the generalized linear model and it take back most of the usages of this family : estimation by maximum likelihood, statistic tests, etc...

2.1 Logistic model

2.1.1 Verhulst's model

The Malthusian model of exponential growth of the population is questioned in 1840 by François Verhulst who introduce the concept of limiting factor which imply the existence of a maximal size M for the population. The growth is no longer proportional to the population size x but to the quantity $x(M - x)$.

This model given him a prediction for the French population size in 1930 of 40 million and it is not so bad because in reality it was 41,5 million in 1931. The Verhulst model is still used today and is quite effective.

The model in continuous time leads us to a differential equation where all the solutions are known but in discrete time for particular parameters the series comportment could be complicated or chaotic.

2.1.2 Logistic function

In continuous time the functions are defined positive on $[0, +\infty[$ and verify the two following conditions :

$y(0) = y_0$

$y' = ry\left(1 - \frac{y}{K}\right)$ (1) avec $r > 0$ et $K > 0$

The substitution $z = \frac{1}{y}$ in (1), which work for $y > 0$, lead us to the differential equation $z' = -r\left(z - \frac{1}{K}\right)$

The solution of this equation is g and is define as $g(t) = \lambda e^{-rt} + \frac{1}{K}$

The function f must verify $f(t) = \frac{1}{g(t)} = K\frac{1}{1+\lambda K e^{-rt}}$

The initial condition $y(0) = y_0$ leads us to the unique solution $f(t) = K\frac{1}{1+\left(\frac{K}{y_0}-1\right)e^{-rt}}$

2.1.3 Logistic law

The law is defined by its cumulative distribution function $F(x) = \frac{1}{1+e^{-\frac{x-\mu}{s}}}$

The density function is $f(x) = \frac{e^{-\frac{x-\mu}{s}}}{s\left(1+e^{-\frac{x-\mu}{s}}\right)^2}$

The name of logistic law is due to the cumulative distribution function which belongs to the logistical family.

The mean and the variance are given by :

$\mathbb{E}(X) = m$

$\mathbb{V}(x) = \frac{s^2\pi^2}{3}$

The standard logistic law is the one with parameters 0 and 1. The repartition function is the sigmoid : $F(x) = \frac{1}{1+e^{-x}}$

We have $\mathbb{E}(X) = 0$ et $\mathbb{V}(x) = \frac{\pi^2}{3}$

2.2 Logistic regression

The classic generalized linear regression allow us to predict a continuous variable with other significant variables. The logistic regression allows us to predict a binary variable with other significant variables. The notion of odds is introduce.

2.2.1 Logistic regression

In order to be able to calculate the quantity $\mathbb{P}(Y = y_k/X)$, we must introduce some hypothesis under the distribution. We are using “semi-parametric” methods and we make hypothesis on the distribution rate.

The logistic regression is based on probabilistic considerations.

Bayes theorem : $\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$

That theorem allow us to estimate the conditional probability $\mathbb{P}(Y = y_k/X)$:

$\mathbb{P}(Y = y_k/X) = \frac{\mathbb{P}(Y=y_k)\mathbb{P}(X/Y=y_k)}{\mathbb{P}(X)} = \frac{\mathbb{P}(Y=y_k)\mathbb{P}(X/Y=y_k)}{\sum_k \mathbb{P}(Y=y_k)\mathbb{P}(X/Y=y_k)}$

When there are only two classes we must compare
$\mathbb{P}(Y = +/X)$ and $\mathbb{P}(Y = -/X)$

We built the fraction : $\frac{\mathbb{P}(Y=+/X)}{\mathbb{P}(Y=-/X)} = \frac{\mathbb{P}(Y=+)}{\mathbb{P}(Y=-)}\frac{\mathbb{P}(X/Y=+)}{\mathbb{P}(X/Y=-)}$

The decision rule becomes :

If $\frac{\mathbb{P}(Y=+/X)}{\mathbb{P}(Y=-/X)} > 1$

the probability to be positive is higher than to be negative so it's why $Y = +$

The logistic regression introduce the following fundamental hypothesis :

$ln\left[\frac{\mathbb{P}(X/Y=+)}{\mathbb{P}(X/Y=-)}\right] = b_0 + b_1X_1 + ... + b_JX_J$

This hypothesis covers a large range of data distribution laws : the normal distribution, discrete distributions, etc. But also a mix of binary and continuous explanatory variables, this property is very important because it often makes the logistical regression operational.

2.2.2 LOGIT law

Logistical regression can be described in a different way. Then, the general idea is not to predict Y anymore, but to give probabilities to $Y = -$ and $Y = +$ on condition of explanatory variables X = x :

$\pi(x) = \mathbb{P}(Y = +|X = x)$ and $1 - \pi(x) = \mathbb{P}(Y = -|X = x)$

For a ω individual, we call LOGIT transformation of $\pi(\omega)$ the expression :

$ln\left[\frac{\pi(\omega)}{1-\pi(\omega)}\right] = a_0 + a_1X_1 + ... + a_JX_J$

Odds is defined like that :

$odds(\omega) = \frac{\pi(\omega)}{1-\pi(\omega)} = \frac{P(Y=+/X)}{P(Y=-/X)}$ it's a chance relation.

For example, if an individual has a two odds, it means that he is two times more likely to be positive than to be negative.

Similarly, we define odds-ratio by :

$odds_{ratio}(\omega_i, \omega_j) = \frac{odds(\omega_i)}{odds(\omega_j)}$.

We set $C(X) = a_0 + a_1 X_1 + ... + a_J X_J$, and we can come back to π with logistic function :

$\pi = \frac{e^{C(X)}}{1+e^{C(X)}} = \frac{1}{1+e^{-C(X)}}$

It's the cumulative distribution function of the logistic distribution.

2.2.3 Property

According to the last π expression, we see it is possible to model π by $\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$ whose appearance correspond to representation of scatter plot (x_i, y_i) in binary observation case.

Therefore, the model take on its full meaning if LOGIT function allow to linearise observations.

Theorem : LOGIT function is a canonical link function for binomial variables.

Proof :

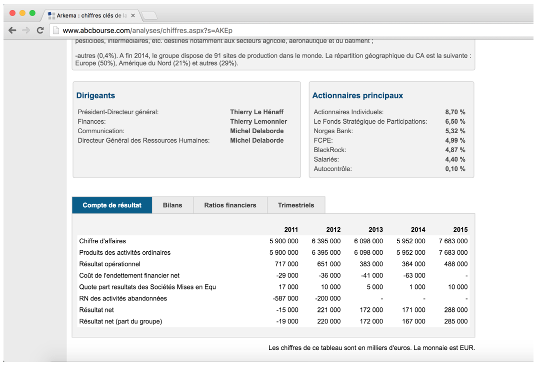
$$\begin{aligned} logit\left(\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}\right) &= ln\left(\frac{\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}}{1-\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}}\right) \\ &= ln\left(\frac{\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}}{\frac{1+e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}-\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}}\right) \\ &= ln\left(\frac{\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}}{\frac{1}{1+e^{\alpha+\beta x}}}\right) \\ &= ln\left(e^{\alpha+\beta x}\right) \\ &= \alpha+\beta x. \end{aligned}$$

3 Scoring

3.1 Database elaboration

One of our first challenge was to create a database with every variables that we needed. Indeed, we had to bring a lot of information together about companies, like their revenues, their earnings, and around ten financial data.

To elaborate the database, we used ABC Bourse. This website collect information of companies. Each company has its own page. We can find balance sheets, income statements, financial ratios and three-monthly information.



Arkema : chiffres clés de la ×
www.abcbourse.com/analyses/chiffres.aspx?s=AKEp

pesticides, intermédiaires, etc. destinés notamment aux secteurs agricole, aéronautique et du bâtiment ;

-autres (0,4%). A fin 2014, le groupe dispose de 91 sites de production dans le monde. La répartition géographique du CA est la suivante : Europe (50%), Amérique du Nord (21%) et autres (29%).

Dirigeants

Président-Directeur général:	**Thierry Le Hénaff**
Finances:	**Thierry Lemonnier**
Communication:	**Michel Delaborde**
Directeur Général des Ressources Humaines:	**Michel Delaborde**

Actionnaires principaux

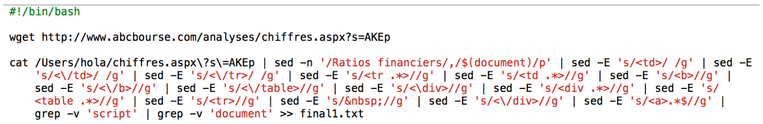
Actionnaires Individuels:	**8,70 %**
Le Fonds Stratégique de Participations:	**6,50 %**
Norges Bank:	**5,32 %**
FCPE:	**4,99 %**
BlackRock:	**4,87 %**
Salariés:	**4,40 %**
Autocontrôle:	**0,10 %**

	2011	2012	2013	2014	2015
Chiffre d'affaires	5 900 000	6 395 000	6 098 000	5 952 000	7 683 000
Produits des activités ordinaires	5 900 000	6 395 000	6 098 000	5 952 000	7 683 000
Résultat opérationnel	717 000	651 000	383 000	364 000	488 000
Coût de l'endettement financier net	-29 000	-36 000	-41 000	-63 000	-
Quote part resultats des Sociétés Mises en Equ	17 000	10 000	5 000	1 000	10 000
RN des activités abandonnées	-587 000	-200 000	-	-	-
Résultat net	-15 000	221 000	172 000	171 000	288 000
Résultat net (part du groupe)	-19 000	220 000	172 000	167 000	285 000

Les chiffres de ce tableau sont en milliers d'euros. La monnaie est EUR.

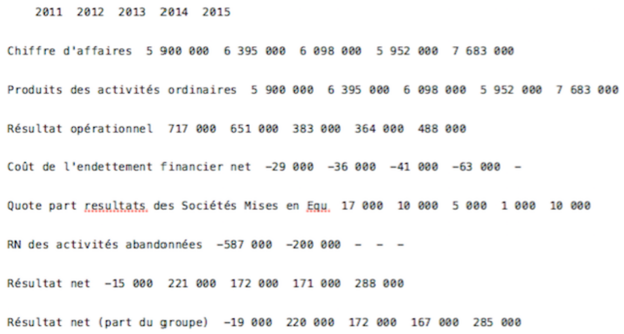
In order to collect data more easily, we wrote a bash script, to do it automatically.

The starting point is a sticker list of companies desired. It is easy to download it on the website. Each company page correspond to URL :
http ://www.abcbourse.com/analyses/chiffres.aspx ?s=**STICKER**p
Applying wget command for each url, we download the whole source code of wanted pages. Then, it is sufficient to clean each source code automatically with regular expressions and to send the result to a corresponding text file. The image below correspond to the script of a company.



```
#!/bin/bash

wget http://www.abcbourse.com/analyses/chiffres.aspx?s=AKEp

cat /Users/hola/chiffres.aspx\?s\=AKEp | sed -n '/Ratios financiers/,/$(document)/p' | sed -E 's/<td>/ /g' | sed -E
    's/<\/td>/ /g' | sed -E 's/<\/tr>/ /g' | sed -E 's/<tr .*>//g' | sed -E 's/<td .*>//g' | sed -E 's/<b>//g' |
    sed -E 's/<\/b>//g' | sed -E 's/<\/table>//g' | sed -E 's/<\div>//g' | sed -E 's/<div .*>//g' | sed -E 's/
    <table .*>//g' | sed -E 's/<tr>//g' | sed -E 's/&nbsp;//g' | sed -E 's/<\/div>//g' | sed -E 's/<a>.*$//g' |
    grep -v 'script' | grep -v 'document' >> final1.txt
```

The result of the corresponding text file is in this form. It contains the whole information of every selected companies, one after the other.



```
    2011  2012  2013  2014  2015

Chiffre d'affaires  5 900 000  6 395 000  6 098 000  5 952 000  7 683 000

Produits des activités ordinaires  5 900 000  6 395 000  6 098 000  5 952 000  7 683 000

Résultat opérationnel  717 000  651 000  383 000  364 000  488 000

Coût de l'endettement financier net  -29 000  -36 000  -41 000  -63 000  -

Quote part resultats des Sociétés Mises en Equ  17 000  10 000  5 000  1 000  10 000

RN des activités abandonnées  -587 000  -200 000  -  -  -

Résultat net  -15 000  221 000  172 000  171 000  288 000

Résultat net (part du groupe)  -19 000  220 000  172 000  167 000  285 000
```

3.2 Logistic regression analysis

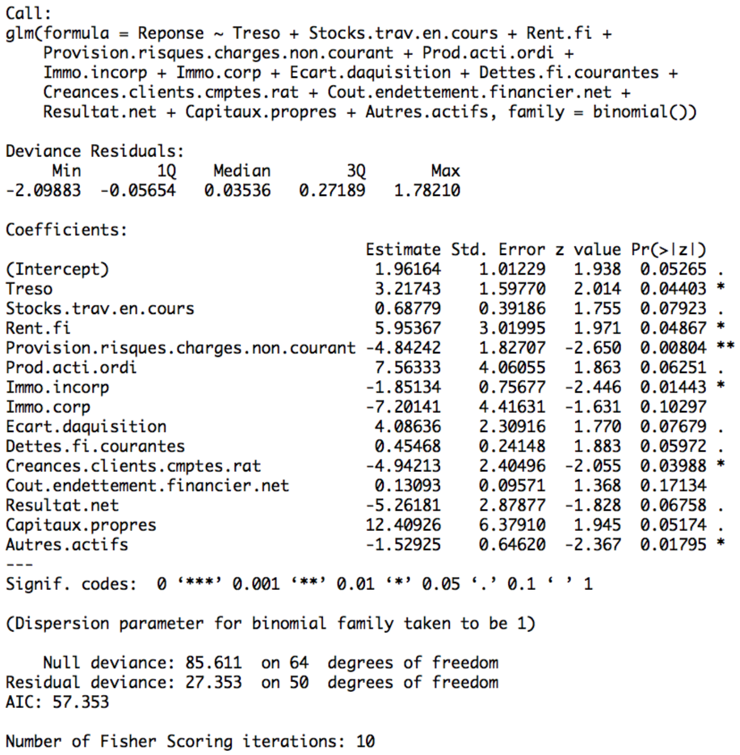
3.2.1 Regression

Scoring is a data ranking method[2] that enable to estimate with a mark or a score, the probability that an individual answer to a solicitation or belongs to a intended target.

In general, score is achieved from quantitative and qualitative data available on the individual in which a scoring model is applied.

Generally, the modeling method used is logistic regression. It is part of overseen learning techniques, in other words, we generally want to explain affiliation to a category from descriptors collected from a population sample in order to generalise learning.

For our study, we use a database that contains data of 65 companies that we want to foresee the stock growth. Each company is characterised by 23 variables. Each variable correspond to the five years change sum. Responding variable is a binary variable that correspond to a more than 10% increase.
Variable selection through significance threshold corresponding to p-value, bring us to build the following model.



```
Call:
glm(formula = Reponse ~ Treso + Stocks.trav.en.cours + Rent.fi +
    Provision.risques.charges.non.courant + Prod.acti.ordi +
    Immo.incorp + Immo.corp + Ecart.daquisition + Dettes.fi.courantes +
    Creances.clients.cmptes.rat + Cout.endettement.financier.net +
    Resultat.net + Capitaux.propres + Autres.actifs, family = binomial())

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.09883  -0.05654   0.03536   0.27189   1.78210

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                            1.96164    1.01229   1.938  0.05265 .
Treso                                  3.21743    1.59770   2.014  0.04403 *
Stocks.trav.en.cours                   0.68779    0.39186   1.755  0.07923 .
Rent.fi                                5.95367    3.01995   1.971  0.04867 *
Provision.risques.charges.non.courant -4.84242    1.82707  -2.650  0.00804 **
Prod.acti.ordi                         7.56333    4.06055   1.863  0.06251 .
Immo.incorp                           -1.85134    0.75677  -2.446  0.01443 *
Immo.corp                             -7.20141    4.41631  -1.631  0.10297
Ecart.daquisition                      4.08636    2.30916   1.770  0.07679 .
Dettes.fi.courantes                    0.45468    0.24148   1.883  0.05972 .
Creances.clients.cmptes.rat           -4.94213    2.40496  -2.055  0.03988 *
Cout.endettement.financier.net         0.13093    0.09571   1.368  0.17134
Resultat.net                          -5.26181    2.87877  -1.828  0.06758 .
Capitaux.propres                      12.40926    6.37910   1.945  0.05174 .
Autres.actifs                         -1.52925    0.64620  -2.367  0.01795 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 85.611  on 64  degrees of freedom
Residual deviance: 27.353  on 50  degrees of freedom
AIC: 57.353

Number of Fisher Scoring iterations: 10
```

We notice that :

-Provisions for risks and charges are very significant. They affect with a negative relation of 4,84.

-The most of other variables are significant at some lower threshold with coefficient that are in accordance with their nature. Only tangible assets and cost of net debt are not significant. However, we keep these variables because their significance threshold is not that low.

-The biggest coefficient is for the equity variable, with 12,4.

-The net result is correlated negatively with the model, appearing surprising and counter-intuitive.

3.2.2 Tests

How well our model fits depends on the difference between the model and the observed data. One approach for binary data is to implement a Hosmer-Lemeshow goodness of fit test.

The Hosmer–Lemeshow[3] test is a statistical test for goodness of fit for logistic regression models. It is used frequently in risk prediction models. The test assesses whether or not the observed event rates match expected event rates in subgroups of the model population. The Hosmer–Lemeshow test specifically identifies subgroups as the deciles of fitted risk values. Models for which expected and observed event rates in subgroups are similar are called well calibrated.

The Hosmer–Lemeshow test statistic is given by :

$$\sum_{g=1}^{G} \frac{(O_{1g}-E_{1g})^2}{E_{1g}} + \frac{(O_{0g}-E_{0g})^2}{E_{0g}} = \sum_{g=1}^{G} \frac{(O_{1g}-E_{1g})^2}{N_g \pi_g} + \frac{(N_g-O_{1g}-(N_g-E_{1g}))^2}{N_g(1-\pi_g)}$$

Which is equals to :

$$\sum_{g=1}^{G} \frac{(O_{1g}-E_{1g})^2}{N_g \pi_g (1-\pi_g)}.$$

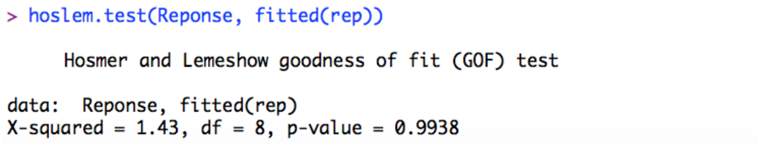
Here O_{1g}, E_{1g}, O_{0g}, E_{0g}, N_g, and π_g denote the observed $Y = 1$ events, expected $Y = 1$ events, observed $Y = 0$ events, expected $Y = 0$ events, total observations, predicted risk for the g_{th} risk decile group, and G is the number of groups. The test statistic asymptotically follows a χ^2 distribution with G 2 degrees of freedom. The number of risk groups may be adjusted depending on how many fitted risks are determined by the model. This helps to avoid singular decile groups.

So hypotheses will be :

H0 : Actual and predicted event rates are similar across 10 deciles
H1 : they are different

Hence if p-value is less than .05, they are not well distributed and you need to refine your model.

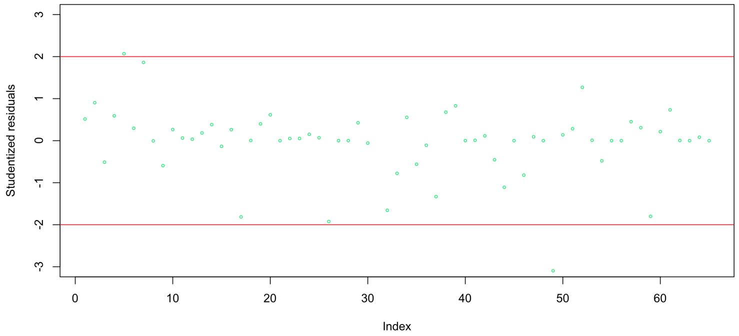
We make this test with hoslem.test command with R, what gives us :



```
> hoslem.test(Reponse, fitted(rep))

	Hosmer and Lemeshow goodness of fit (GOF) test

data:  Reponse, fitted(rep)
X-squared = 1.43, df = 8, p-value = 0.9938
```

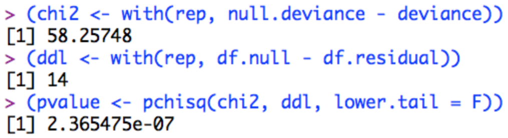
Our model appears to fit well because we have no significant difference between the model and the observed data (i.e. the p-value is above 0.05).

Then, we analyse studentized residuals. It is importe to notice that with a logistic regression, most of the time, we take an interest in deviance residuals. In general, their value fluctuates between -2 and 2.



We notice that the test is conclusive because only one residual value company is aberrant.

It is also possible to take an interest in the model deviance : likelihood relation tests and p-value calculation, including degrees of freedom between the model reduced to the constant and model retained, give global significance of the model.



```
> (chi2 <- with(rep, null.deviance - deviance))
[1] 58.25748
> (ddl <- with(rep, df.null - df.residual))
[1] 14
> (pvalue <- pchisq(chi2, ddl, lower.tail = F))
[1] 2.365475e-07
```

The result is very good because the p-value is very low.

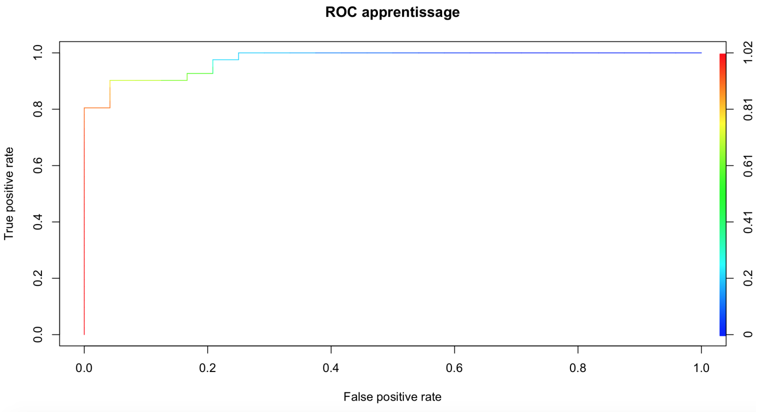
3.2.3 Obtained predictive probabilities

We calculate the ROC curve. This curve, or the area under the curve, represents the model sensitivity/specificity. A model is good if values has been foreseen. Generally, we look at the same time at the curve form and the area under it :

1 ideal model
0,5 random model

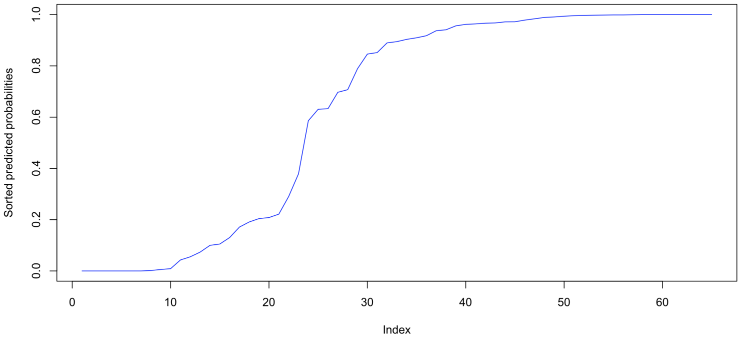
ROC curve principle : if the test gives a numeric result with a t threshold such as the prediction is positive if $x > t$ and negative if $x < t$, so as t increase, specificity increase but sensitivity decrease.

The ROC curve represents the evolution of sensibility (true positive rate) depending on 1 - specificity (false positive rate) when we change the t threshold.



We see quality results, the area under the curve is close to 1 with a value of 0,9756098.

It is also possible to represent graphically probabilities that an exchange rate increase to observe probabilities distribution.

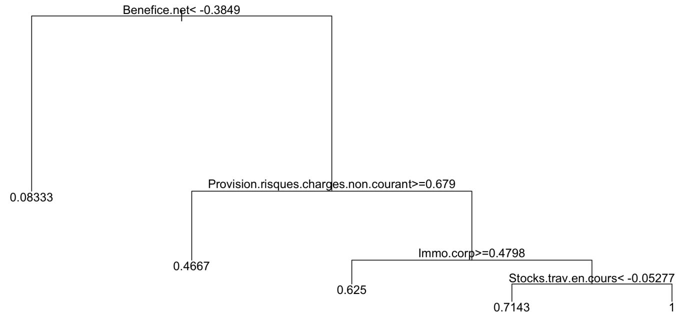


The probabilities correspond to ROC curve representation and support the view that the model is efficient.

3.3 Analysis by classification

Recursive Partitioning trees[1] offer various benefits to credit scoring : quick, simple logic which can be converted into rules and credit policies, non parametric and can deal with interactions. The down side of trees is that they unstable, small changes in data can lead to large deviations in models, and can overfit if not built using cross validation and pruning. R's Rpart is a performing and robust tree algorithms.

We use this function on our data set and we obtain the following tree :

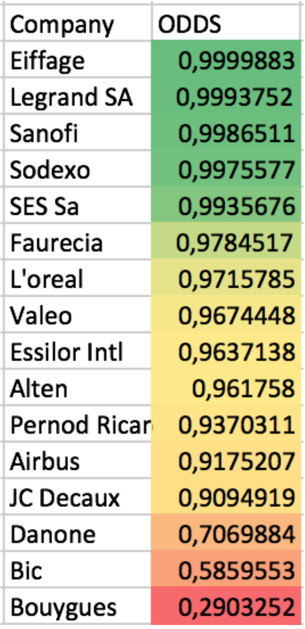

Benefice.net< -0.3849
0.08333
Provision.risques.charges.non.courant>=0.679
0.4667
Immo.corp>=0.4798
0.625
Stocks.trav.en.cours< -0.05277
0.7143
1

We see from this tree that according to this method, 4 variables are considered as significant. These variables don't necessarily correspond to the ones brought to light with logit regression, what enable us to bring more information to our model.

The values correspond to proportion of companies with an increasing exchange rate. It leads us to choose 16 favorable companies.

3.4 Results

Once we have selected the 16 companies, we look at their probability from logit regression to be increasing and we classify them according to this value. The results are :



Company	ODDS
Eiffage	0,9999883
Legrand SA	0,9993752
Sanofi	0,9986511
Sodexo	0,9975577
SES Sa	0,9935676
Faurecia	0,9784517
L'oreal	0,9715785
Valeo	0,9674448
Essilor Intl	0,9637138
Alten	0,961758
Pernod Ricar	0,9370311
Airbus	0,9175207
JC Decaux	0,9094919
Danone	0,7069884
Bic	0,5859553
Bouygues	0,2903252

Thanks to this rank, we select the five best companies and we start analysing their exchange rate history.

4 Investments

4.1 Portfolio theory

Modern portfolio theory (MPT) was introduced by Harry Markowitz in 1952 in his paper ≪ Portfolio Selection ≫ , for which he was awarded by Nobel Prize in economics . Its a statistical method for create a ≪ good ≫ portfolio of different assets only thanks to the return and the variance (risk). Today, it is one of the most important and influential economic theories.

The return at time t is : $Rt = \frac{(V(t)-V(t-1))}{/V(t-1)}$, with V(t) the value of the portfolio a time t.

We suppose that an investor is risk – averse, he wants a small variance (small risk) and a high expected return. If we have a portfolio with n different assets, were asset number i will give the return R_i. Let i be the mean of R_i. Suppose the the relative amount of the value of the portfolio invested in asset i is xi . If R is the return of the whole portfolio then :

-$\mu = \mathbb{E}[R] = \sum \mu_i * xi$, the expected return of the portfolio

-$\sigma^2 = \mathbb{V}[R] = \sum\sum xi * xj * \mathrm{Cov}(R_i, R_j)$, the variance of the portfolio

-$\sum xi = 1$

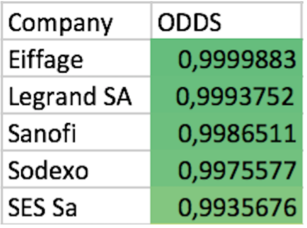
If we modify $(x_1, \ldots, x_n)$, we modify the couple (σ^2, μ). So if we want to minimize the risk and maximize the return expected you have to choose the right $(x_1, \ldots, x_n)$.

The diversification is the fact of build a portfolio with assets low/negatively correlated. Actually, the more the correlation between two assets is close to -1, the more you decrease the risk of the portfolio.

For our portfolio, we are going to take five companies obtained thanks to the scoring. The diversification can be effective if we take companies from different area. However, it is not necessary to take a lot of assets because a lot of asset does not mean a lower risk. Furthermore, the more assets we add, the more difficult will be the handling of the portfolio with the software.

4.2 Portfolio creation

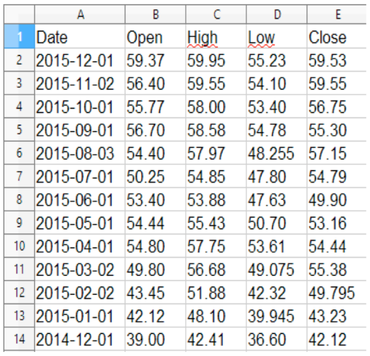
After the application of scoring on our data, we select only the following companies, the five with the highest score :



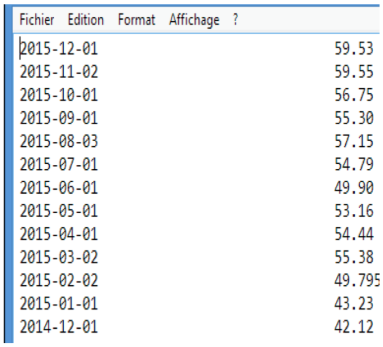
Company	ODDS
Eiffage	0,9999883
Legrand SA	0,9993752
Sanofi	0,9986511
Sodexo	0,9975577
SES Sa	0,9935676

Therefore, we apply the process of the Modern Portfolio Theory (MPT) to this five companies. First, we get the stock price of each company from 01/01/2011 to 31/12/2015. We can find them easily on the website "https ://fr.finance.yahoo.com" and download them as OpenOffice/Excel tables. We take the monthly values at the closing of the Paris's market . Then, we transfer the data in text files, which are easier to use with the R software.

Here an example with Eiffage's stock price :

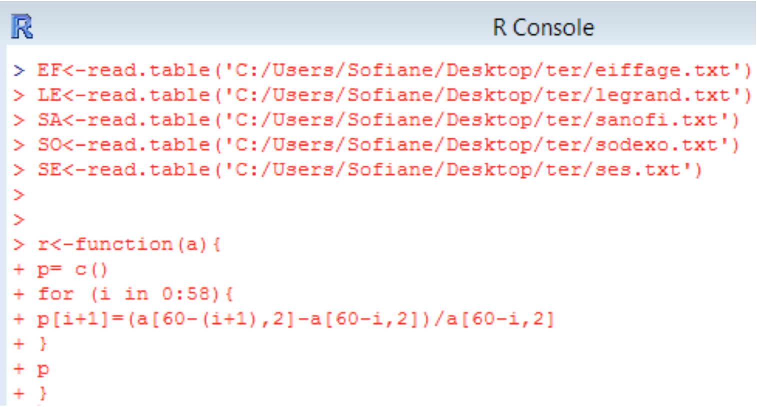


	A	B	C	D	E
1	Date	Open	High	Low	Close
2	2015-12-01	59.37	59.95	55.23	59.53
3	2015-11-02	56.40	59.55	54.10	59.55
4	2015-10-01	55.77	58.00	53.40	56.75
5	2015-09-01	56.70	58.58	54.78	55.30
6	2015-08-03	54.40	57.97	48.255	57.15
7	2015-07-01	50.25	54.85	47.80	54.79
8	2015-06-01	53.40	53.88	47.63	49.90
9	2015-05-01	54.44	55.43	50.70	53.16
10	2015-04-01	54.80	57.75	53.61	54.44
11	2015-03-02	49.80	56.68	49.075	55.38
12	2015-02-02	43.45	51.88	42.32	49.795
13	2015-01-01	42.12	48.10	39.945	43.23
14	2014-12-01	39.00	42.41	36.60	42.12



```
2015-12-01                            59.53
2015-11-02                            59.55
2015-10-01                            56.75
2015-09-01                            55.30
2015-08-03                            57.15
2015-07-01                            54.79
2015-06-01                            49.90
2015-05-01                            53.16
2015-04-01                            54.44
2015-03-02                            55.38
2015-02-02                            49.795
2015-01-01                            43.23
2014-12-01                            42.12
```

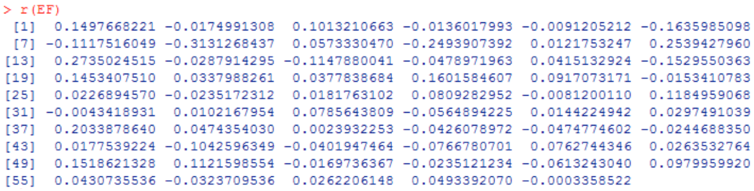
Now, we have the stock prices and we import them in our R file as vectors. We create five others vectors which will contain the historical returns of each company's assets. (We are talking about monthly return, of course). We are going to use these vectors during all the process.



```
> EF<-read.table('C:/Users/Sofiane/Desktop/ter/eiffage.txt')
> LE<-read.table('C:/Users/Sofiane/Desktop/ter/legrand.txt')
> SA<-read.table('C:/Users/Sofiane/Desktop/ter/sanofi.txt')
> SO<-read.table('C:/Users/Sofiane/Desktop/ter/sodexo.txt')
> SE<-read.table('C:/Users/Sofiane/Desktop/ter/ses.txt')
>
>
> r<-function(a){
+ p= c()
+ for (i in 0:58){
+ p[i+1]=(a[60-(i+1),2]-a[60-i,2])/a[60-i,2]
+ }
+ p
+ }
```

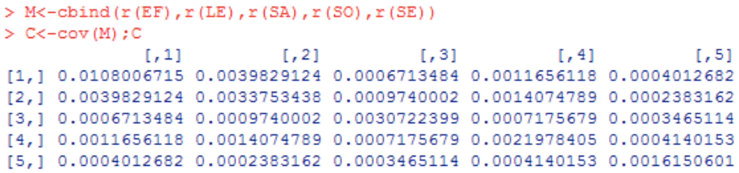
We have created a function called r to calculate the historical returns. This function is based on a loop where we calculate every yield. The stock prices are given in a decreasing order, so the last price in the stock price vector is the first price in time. We invert the order in the function r. Finally, we get a vector of historical returns with a normal order.

Here an Example with Eiffage :



```
> r(EF)
 [1]  0.1497668221 -0.0174991308  0.1013210663 -0.0136017993 -0.0091205212 -0.1635985098
 [7] -0.1117516049 -0.3131268437  0.0573330470 -0.2493907392  0.0121753247  0.2539427960
[13]  0.2735024515 -0.0287914295 -0.1147880041 -0.0478971963  0.0415132924 -0.1529550363
[19]  0.1453407510  0.0337988261  0.0377838684  0.1601584607  0.0917073171 -0.0153410783
[25]  0.0226894570 -0.0235172312  0.0181763102  0.0809282952 -0.0081200110  0.1184959068
[31] -0.0043418931  0.0102167954  0.0785643809 -0.0564894225  0.0144224942  0.0297491039
[37]  0.2033878640  0.0474354030  0.0023932253 -0.0426078972 -0.0474774602 -0.0244688350
[43]  0.0177539224 -0.1042596349 -0.0401947464 -0.0766780701  0.0762744346  0.0263532764
[49]  0.1518621328  0.1121598554 -0.0169736367 -0.0235121234 -0.0613243040  0.0979959920
[55]  0.0430735536 -0.0323709536  0.0262206148  0.0493392070 -0.0003358522
```

Then, we create an M matrix whose column vectors are the vectors of historical return of each asset. Thanks to this matrix, we can create the covariance matrix called C. We will use this symmetric matrix to calculate the portfolio's risk because the elements on the diagonal are the variances of the assets and the other elements are the covariances of each coupe of asset.



```
> M<-cbind(r(EF),r(LE),r(SA),r(SO),r(SE))
> C<-cov(M);C
             [,1]         [,2]         [,3]         [,4]         [,5]
[1,] 0.0108006715 0.0039829124 0.0006713484 0.0011656118 0.0004012682
[2,] 0.0039829124 0.0033753438 0.0009740002 0.0014074789 0.0002383162
[3,] 0.0006713484 0.0009740002 0.0030722399 0.0007175679 0.0003465114
[4,] 0.0011656118 0.0014074789 0.0007175679 0.0021978405 0.0004140153
[5,] 0.0004012682 0.0002383162 0.0003465114 0.0004140153 0.0016150601
```

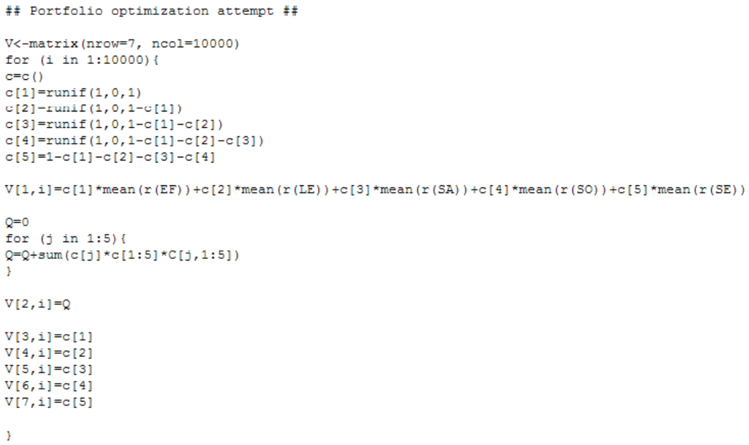
We can already see that the eiffage asset has the highest variance (and so the highest risk because $\sigma(EF) = (\mathbb{V}(EF))^{1/2}$). But, if we look at the expected return, we can also see that the EF asset has the highest :

$\mathbb{E}(R(EF)) = 1,33\% > \mathbb{E}(R(LE)) = 1,14\% > \mathbb{E}(R(SO)) = 1,10\% > \mathbb{E}(R(EF)) = 0,93\% > \mathbb{E}(R(SE)) = 0,71\%$

4.3 Portfolio optimization

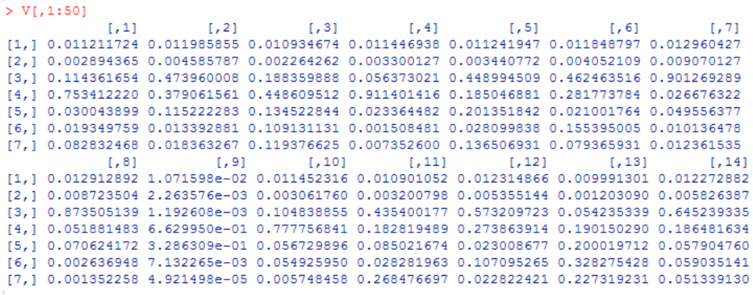
Now, we have to find the four coefficients α, β, γ and δ. Actually, we need five coefficients, but the last one, , can be found with only the four others. Indeed, we saw in the theory part that $\alpha + \beta + \gamma + \delta + \epsilon = 1$ so $\epsilon = 1 - (\alpha + \beta + \gamma + \delta)$. Theoretically, we should find coefficients which minimize the portfolio's risk and which maximize the return. But in our case, we deal with best rated companies from our scoring, so, we are ready to accept more risk. We know that Eiffage has the highest risk and the highest return, so, we are going to try to minimize this risk while keeping high return (close to the highest as possible) thanks to the diversification.

Our method is based on the fact of generate a lot of random vectors $(\alpha, \beta, \gamma, \delta, 1 - \alpha - \beta - \gamma - \delta)$. In each iteration, we evaluate the risk and the return of the portfolio and we put these values in a matrix called V . The columns of this matrix will contain the portfolio's return in first line, the risk in the second line, and in the five next lines, the coefficients. We do 10000 iterations, so, the matrix will have 7 lines and 10000 columns.



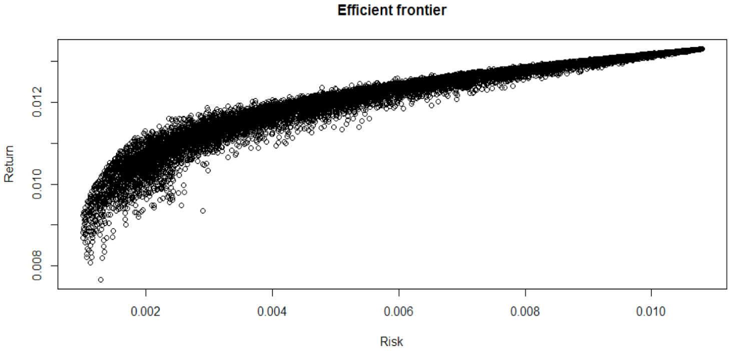
```
## Portfolio optimization attempt ##

V<-matrix(nrow=7, ncol=10000)
for (i in 1:10000){
c=c()
c[1]=runif(1,0,1)
c[2]=runif(1,0,1-c[1])
c[3]=runif(1,0,1-c[1]-c[2])
c[4]=runif(1,0,1-c[1]-c[2]-c[3])
c[5]=1-c[1]-c[2]-c[3]-c[4]

V[1,i]=c[1]*mean(r(EF))+c[2]*mean(r(LE))+c[3]*mean(r(SA))+c[4]*mean(r(SO))+c[5]*mean(r(SE))

Q=0
for (j in 1:5){
Q=Q+sum(c[j]*c[1:5]*C[j,1:5])
}

V[2,i]=Q

V[3,i]=c[1]
V[4,i]=c[2]
V[5,i]=c[3]
V[6,i]=c[4]
V[7,i]=c[5]

}
```

The head of the V matrix :

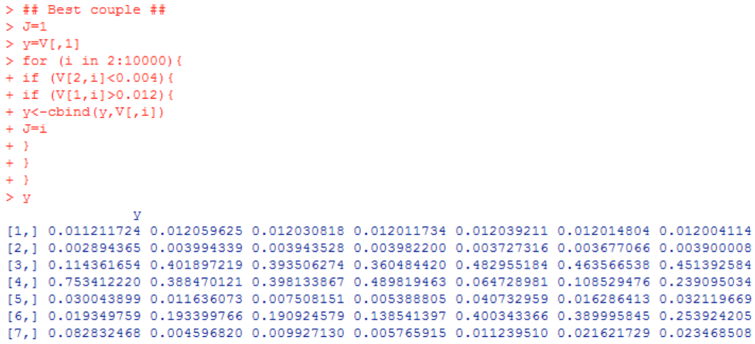


```
> V[,1:50]
            [,1]        [,2]        [,3]        [,4]        [,5]        [,6]        [,7]
[1,] 0.011211724 0.011985855 0.010934674 0.011446938 0.011241947 0.011848797 0.012960427
[2,] 0.002894365 0.004585787 0.002264262 0.003300127 0.003440772 0.004052109 0.009070127
[3,] 0.114361654 0.473960008 0.188359888 0.056373021 0.448994509 0.462463516 0.901269289
[4,] 0.753412220 0.379061561 0.448609512 0.911401416 0.185046881 0.281773784 0.026676322
[5,] 0.030043899 0.115222283 0.134522844 0.023364482 0.201351842 0.021001764 0.049556377
[6,] 0.019349759 0.013392881 0.109131131 0.001508481 0.028099838 0.155395005 0.010136478
[7,] 0.082832468 0.018363267 0.119376625 0.007352600 0.136506931 0.079365931 0.012361535
            [,8]         [,9]       [,10]       [,11]       [,12]       [,13]       [,14]
[1,] 0.012912892 1.071598e-02 0.011452316 0.010901052 0.012314866 0.009991301 0.012272882
[2,] 0.008723504 2.263576e-03 0.003061760 0.003200798 0.005355144 0.001203090 0.005826387
[3,] 0.873505139 1.192608e-03 0.104838855 0.435400177 0.573209723 0.054235339 0.645239335
[4,] 0.051881483 6.629950e-01 0.777756841 0.182819489 0.273863914 0.190150290 0.186481634
[5,] 0.070624172 3.286309e-01 0.056729896 0.085021674 0.023008677 0.200019712 0.057904760
[6,] 0.002636948 7.132265e-03 0.054925950 0.028281963 0.107095265 0.328275428 0.059035141
[7,] 0.001352258 4.921498e-05 0.005748458 0.268476697 0.022822421 0.227319231 0.051339130
```

We can use these values of returns and variances to build an efficient frontier :



We see on this graphic that there are some portfolios with a return close to 1,2% (close to 1,33%, the highest return possible) and a variance less than 0,004 ($\sigma < 6,32\%$). So we are going to find them with a little loop.



```
> ## Best couple ##
> J=1
> y=V[,1]
> for (i in 2:10000){
+ if (V[2,i]<0.004){
+ if (V[1,i]>0.012){
+ y<-cbind(y,V[,i])
+ J=i
+ }
+ }
+ }
> y
               y
[1,] 0.011211724 0.012059625 0.012030818 0.012011734 0.012039211 0.012014804 0.012004114
[2,] 0.002894365 0.003994339 0.003943528 0.003982200 0.003727316 0.003677066 0.003900008
[3,] 0.114361654 0.401897219 0.393506274 0.360484420 0.482955184 0.463566538 0.451392584
[4,] 0.753412220 0.388470121 0.398133867 0.489819463 0.064728981 0.108529476 0.239095034
[5,] 0.030043899 0.011636073 0.007508151 0.005388805 0.040732959 0.016286413 0.032119669
[6,] 0.019349759 0.193399766 0.190924579 0.138541397 0.400343366 0.389995845 0.253924205
[7,] 0.082832468 0.004596820 0.009927130 0.005765915 0.011239510 0.021621729 0.023468508
```

In this example, we chose the fifth column. But in fact, this is not this vector that we used. We used the same method, but we did not find exactly the same values. Every time we compile the program, the values are different, because we take random values. We used these following values for our investment :

$$(\alpha, \beta, \gamma, \delta, 1-\alpha-\beta-\gamma-\delta) = (0,507273; 0,005545; 0,010806; 0,452410; 0,025966)$$

$\mathbb{E}(Rt) = 1.23\%$, the portfolio's return

$\sigma(Rt) = 6.12\%$, the portfolio's risk

5 Results

At the beginning of the project, we supposed we had 1000000 euros to invest. With this amount, we did, the 1/01/2016 the following trade :

-We bougth 8497 stocks of Eiffage at 59.7 eur

-We bougth 105 stocks of Legrand at 52.62 eur

-We bougth 136 stocks of Sanofi at 79.71 eur

-We bougth 4947 stocks of Sodexo at 91.46 eur

-We bougth 1001 stocks of SES Sa at 25.93 eur

After three months, the 31/03/2016, at the opening, the stock prices are :

-Eiffage , 68 eur : +14%

-Legrand, 49 .69 eur : -5.6%

-Sanofi, 71 .23 eur : -10.6%

-Sodexo, 95.34 eur : +4.24%

-SES Sa, 25.76 eur : -0.66%

We sold all our stocks qt the opening of the market and we obtained 1088102 eur. The portfiolo's return is $R_{3t} = 8,8102\%$ for three months, so the monthly return is $Rt = 2,937\%$ (more than $2x\mathbb{E}(R_t)$). We can see that the Scoring was consistent, the best rated company was Eiffage and the stock won 14%. The fact that we invested a lot in that asset allowed us to compensate the other losses.

6 Conclusion

We obtain relatively good results with this method. It shows us that a company could be evaluate in short and long time by different methods with effective results. We could have work more on various ranking method.

Références

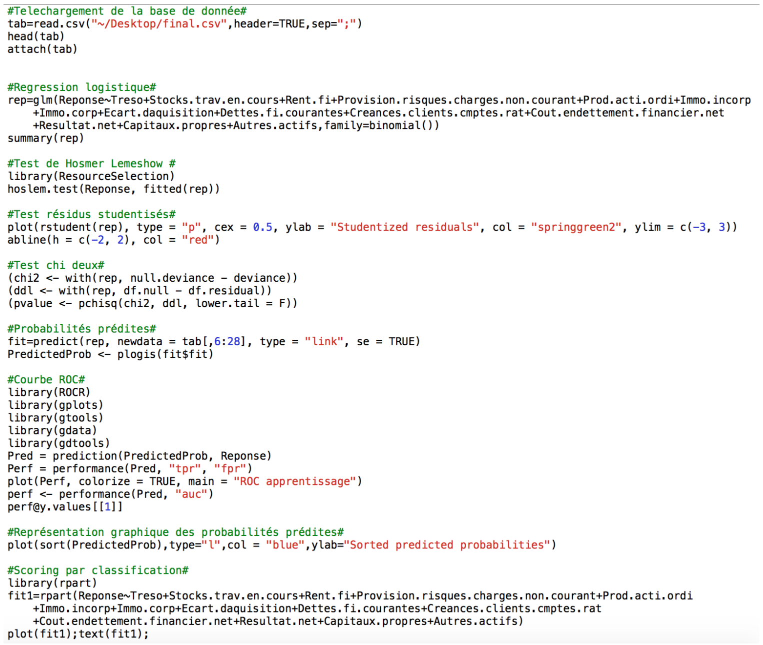
[1] Ross Gayler. *Guide to Credit Scoring in R.* url : `https://cran.r-project.org/doc/contrib/Sharma-CreditScoring.pdf`.

[2] *Scoring avec R.* url : `http://rstudio-pubs-static.s3.amazonaws.com/5267_0156db47a0604aa9818143e7d2db226e.html`.

[3] Hosmer David W. Lemeshow Stanley. *Applied Logistic Regression.* 2013. isbn : 978-0-470-58247-3.

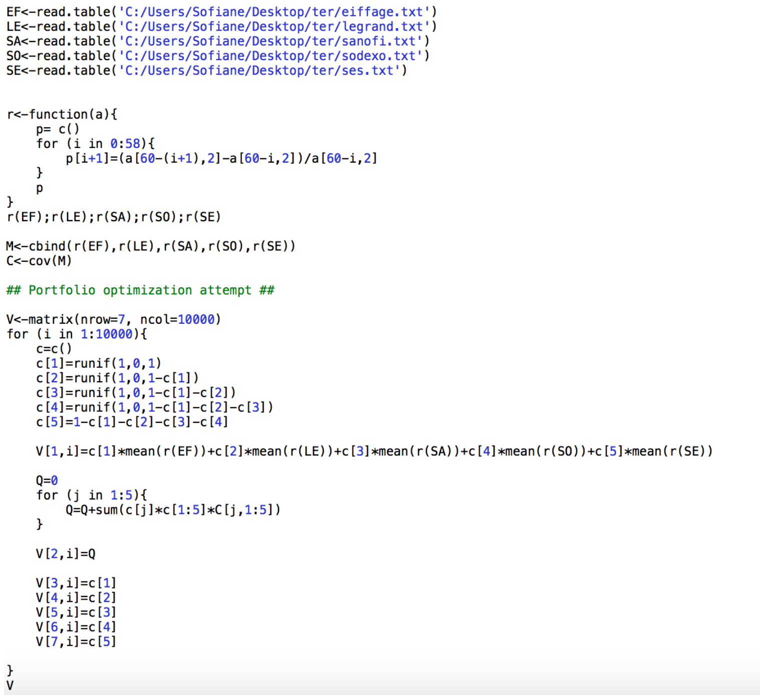
7 Annexes

Scoring code :

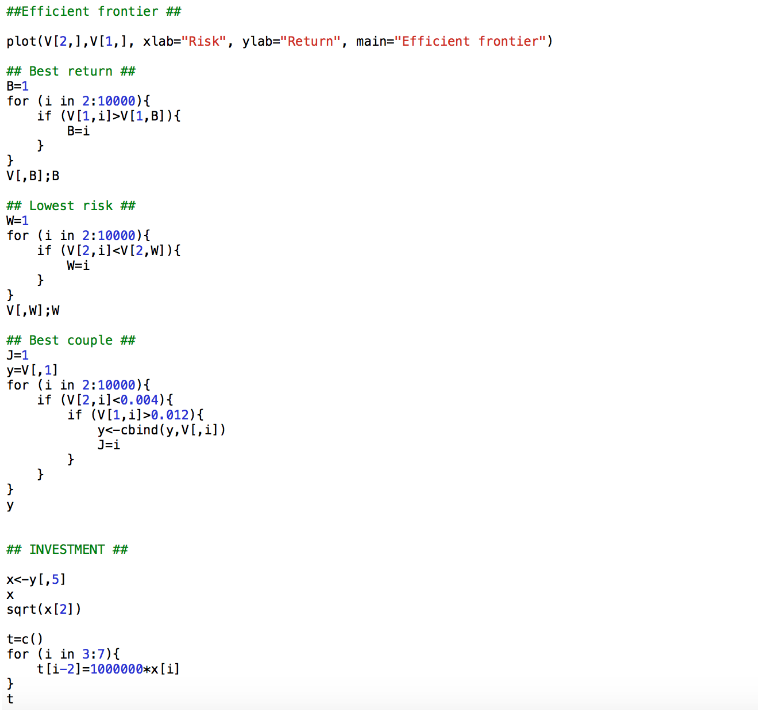


```r
#Telechargement de la base de donnée#
tab=read.csv("~/Desktop/final.csv",header=TRUE,sep=";")
head(tab)
attach(tab)


#Regression logistique#
rep=glm(Reponse~Treso+Stocks.trav.en.cours+Rent.fi+Provision.risques.charges.non.courant+Prod.acti.ordi+Immo.incorp
    +Immo.corp+Ecart.daquisition+Dettes.fi.courantes+Creances.clients.cmptes.rat+Cout.endettement.financier.net
    +Resultat.net+Capitaux.propres+Autres.actifs,family=binomial())
summary(rep)

#Test de Hosmer Lemeshow #
library(ResourceSelection)
hoslem.test(Reponse, fitted(rep))

#Test résidus studentisés#
plot(rstudent(rep), type = "p", cex = 0.5, ylab = "Studentized residuals", col = "springgreen2", ylim = c(-3, 3))
abline(h = c(-2, 2), col = "red")

#Test chi deux#
(chi2 <- with(rep, null.deviance - deviance))
(ddl <- with(rep, df.null - df.residual))
(pvalue <- pchisq(chi2, ddl, lower.tail = F))

#Probabilités prédites#
fit=predict(rep, newdata = tab[,6:28], type = "link", se = TRUE)
PredictedProb <- plogis(fit$fit)

#Courbe ROC#
library(ROCR)
library(gplots)
library(gtools)
library(gdata)
library(gdtools)
Pred = prediction(PredictedProb, Reponse)
Perf = performance(Pred, "tpr", "fpr")
plot(Perf, colorize = TRUE, main = "ROC apprentissage")
perf <- performance(Pred, "auc")
perf@y.values[[1]]

#Représentation graphique des probabilités prédites#
plot(sort(PredictedProb),type="l",col = "blue",ylab="Sorted predicted probabilities")

#Scoring par classification#
library(rpart)
fit1=rpart(Reponse~Treso+Stocks.trav.en.cours+Rent.fi+Provision.risques.charges.non.courant+Prod.acti.ordi
    +Immo.incorp+Immo.corp+Ecart.daquisition+Dettes.fi.courantes+Creances.clients.cmptes.rat
    +Cout.endettement.financier.net+Resultat.net+Capitaux.propres+Autres.actifs)
plot(fit1);text(fit1);
```

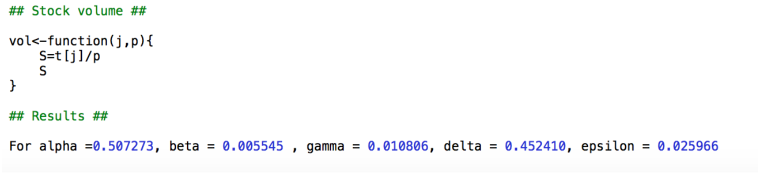
Portfolio code :



```
EF<-read.table('C:/Users/Sofiane/Desktop/ter/eiffage.txt')
LE<-read.table('C:/Users/Sofiane/Desktop/ter/legrand.txt')
SA<-read.table('C:/Users/Sofiane/Desktop/ter/sanofi.txt')
SO<-read.table('C:/Users/Sofiane/Desktop/ter/sodexo.txt')
SE<-read.table('C:/Users/Sofiane/Desktop/ter/ses.txt')


r<-function(a){
    p= c()
    for (i in 0:58){
        p[i+1]=(a[60-(i+1),2]-a[60-i,2])/a[60-i,2]
    }
    p
}
r(EF);r(LE);r(SA);r(SO);r(SE)

M<-cbind(r(EF),r(LE),r(SA),r(SO),r(SE))
C<-cov(M)

## Portfolio optimization attempt ##

V<-matrix(nrow=7, ncol=10000)
for (i in 1:10000){
    c=c()
    c[1]=runif(1,0,1)
    c[2]=runif(1,0,1-c[1])
    c[3]=runif(1,0,1-c[1]-c[2])
    c[4]=runif(1,0,1-c[1]-c[2]-c[3])
    c[5]=1-c[1]-c[2]-c[3]-c[4]

    V[1,i]=c[1]*mean(r(EF))+c[2]*mean(r(LE))+c[3]*mean(r(SA))+c[4]*mean(r(SO))+c[5]*mean(r(SE))

    Q=0
    for (j in 1:5){
        Q=Q+sum(c[j]*c[1:5]*C[j,1:5])
    }

    V[2,i]=Q

    V[3,i]=c[1]
    V[4,i]=c[2]
    V[5,i]=c[3]
    V[6,i]=c[4]
    V[7,i]=c[5]

}
V
```



```
##Efficient frontier ##

plot(V[2,],V[1,], xlab="Risk", ylab="Return", main="Efficient frontier")

## Best return ##
B=1
for (i in 2:10000){
    if (V[1,i]>V[1,B]){
        B=i
    }
}
V[,B];B

## Lowest risk ##
W=1
for (i in 2:10000){
    if (V[2,i]<V[2,W]){
        W=i
    }
}
V[,W];W

## Best couple ##
J=1
y=V[,1]
for (i in 2:10000){
    if (V[2,i]<0.004){
        if (V[1,i]>0.012){
            y<-cbind(y,V[,i])
            J=i
        }
    }
}
y


## INVESTMENT ##

x<-y[,5]
x
sqrt(x[2])

t=c()
for (i in 3:7){
    t[i-2]=1000000*x[i]
}
t
```



```
## Stock volume ##

vol<-function(j,p){
    S=t[j]/p
    S
}

## Results ##

For alpha =0.507273, beta = 0.005545 , gamma = 0.010806, delta = 0.452410, epsilon = 0.025966
```