Case Study on Business Analytics and Big Data: Developing a Rating Model for Companies using Real Scrapped Yahoo Data and Machine Learning in Python.

Date: 07/February/2018
Author: Prof. Manoel Gadi
Many thanks to my student Rita Alcalde (IE MBD 2017-2018) for helping with the translation into English.

Content

9.4.1 Company Rating Fundamentals.

The analysis of the activity and solvency of a company is one of the most classic applications of statistical and analytical methods. When deciding to give a loan, expanding or cutting financing, financial institutions no longer apply subjective criteria, instead they use scientific tools, one of them called Rating. The Rating can be understood as a set of algorithms applied at the time of the decision. To build such a rating, institutions look, among other indices, at quantitative aspects, which are mainly financial ratios combined in a certain proportion. This could be compared with a cooking recipe in which there are some ingredients are used in a certain proportion, ad others are not used at all. The challenge from Analytics is to find what ingredients to use, the proportions and how to combine them. By doing so, statistical methods and / or Machine Learning techniques are used.

9.4.1.1 What is a rating report and what is it used for?

> "With this tool, it is possible to actively manage your balance and achieve more ways of accessing financing sources and better conditions. "

In a very direct and probably very boring way, one could say that a Financial Rating Report is a document that contains an analysis and an opinion of the financial community (banks, investors or other agents) about the credit quality or insolvency risk of a company. The opinion is summarized in a credit risk score or rating following a standardized numeric or letter scale. The credit quality of credit risk is the result of a process of analysis of quantitative and qualitative factors that affect not only the company, but also its sector and its country.

However, the famous economist Michael Porter, author of the book "Competitive Advantage", considered the bible of business thinkers, would not have achieved the same impact, a real Eureka moment among business experts, if he had explained the Value Chain as a diagram that organizes the interconnected activities in the company, or maybe he would.

Michael Porter presents the idea of the Value Chain as the map of the gold mine to understand how a certain company puts all its machinery to transform money into more money. A true X-ray that reveals a diagram of interconnected activities of the bone structure of the company, where one sees how it transforms money into raw material, then into final product, going on sale and finally resulting in profit. The Value Chain is this X-Ray that allows us to understand the operational functioning of a company.

Following the medical metaphor, we can say that the Rating Report is the blood analysis of the company. The blood analyzer measures, for example, the volume of platelets, lymphocytes, monocytes and other names that sometimes seem to us palaver without much sense. However, we are able to understand that we are health if the measured value is within the lower and upper reference limits that appear in the report. These limits in general, are benchmarks of healthy people with similar characteristics to us, how they can be the same gender and similar age.

The Rating Reports have a very similar structure to the blood analysis. Here blood is the company's money and within the annual accounts it appears mainly in the shape of cash, debt and equity. Therefore, we measure the adequacy of ratios such as the leverage ratio, liquidity ratio, profitability ratio and the capital structure to reference values of similar companies, often from the same sector and size as the company being analyzed. We call this first part of the report Quantitative Rating. But laboratories result by themselves are not everything we need. It is always recommended to visit the doctor so that he interprets the analytical report and asks some questions to have a global evaluation of the person. In the Rating Report, this visit is part of an interview and the answer of a qualitative questionnaire. Although very structured and regulated, this interview tries to map issues as diverse as the evolution of demand and the market, understand the history of the partners and the management team, the strategy and its business plan as a whole. Today it is impossible to imagine a company that does not know its operative X-Ray through the Chain of Value of its business and its sector. In the same way, it should be unimaginable to conceive a company that does not know the "blood analysis" of its money. A company that does not have the correct tools to understand its insolvency risk and does not know how the bank perceives it. A company with these limitations consequently does not have the ability to compare their creditworthiness with that of their competition.

After all, the Rating Report is the basis for accessing funding sources and their terms and conditions. Once in possession of this tool, one can actively manage its balance to improve the credit rating and, therefore, achieve more access to sources of financing, better conditions and, finally, greater benefits.

9.4.1.2 Who are the main players of the Rating industry?

The main rating agencies in the world are:

- Moody's - https://www.moodys.com/

- Standard and poor's (S&P) - https://www.standardandpoors.com
- Fitch - https://www.fitchratings.com

In Spain:

- Axesor - https://www.axesor.es

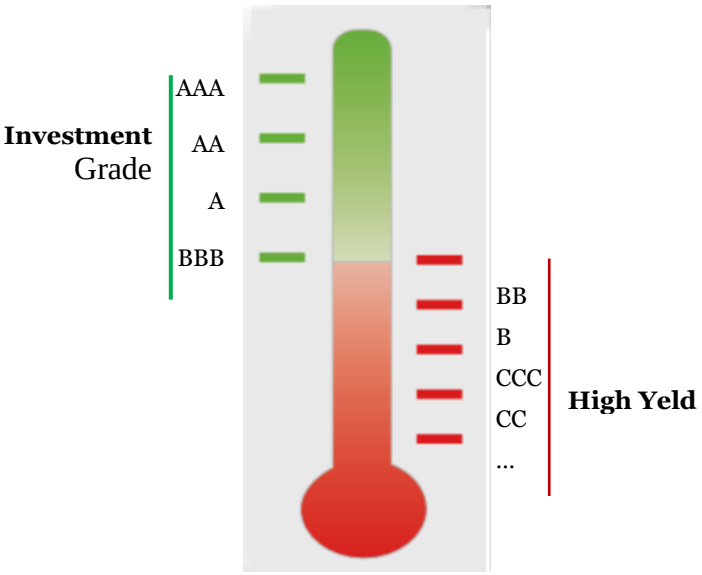
Moody's and S & P can be considered the most influential agencies because of their great coverage worldwide.

Rating agencies in general get their income in two ways:

1. The collection of fees from the evaluated companies / counterparts. These fees are usually charged at the time of evaluation and annual payments for the renewal of the note. The Rating´s prices depend very much on the size and market in which a company operates. As a reference, we can say that the prices are in a range that goes from 30 thousand to 200 thousand euros, approximately.
2. Through the sale of research projects, consulting services, software, or proprietary information.

9.4.1.3 How is the rating note calculated? Alphabetic scale and its numerical interpretation.

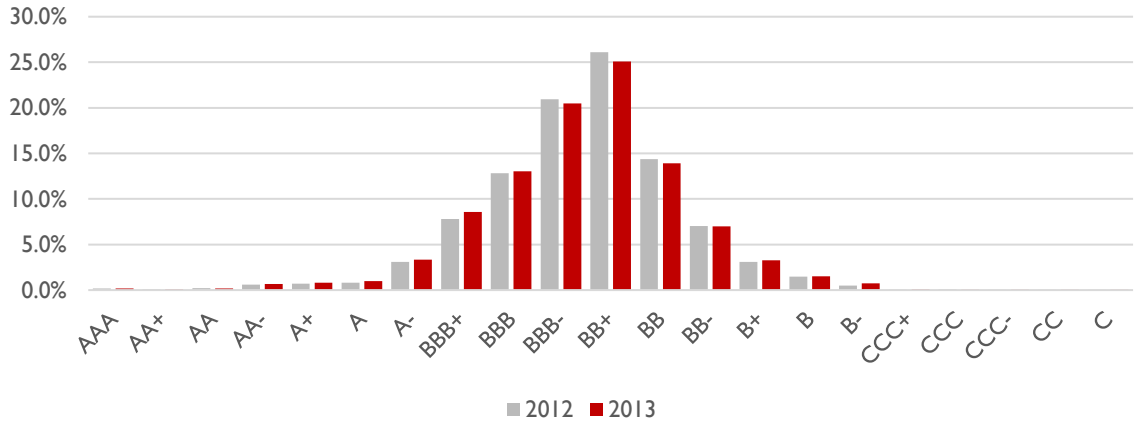
Numeric Interpretation	Moodys	S&P	Fitch
9.9 ≤ r < 10	Aaa	AAA	AAA
9.6 ≤ r < 9.9	Aa1	AA+	AA+
9.2 ≤ r < 9.6	Aa2	AA	AA
9.5 ≤ r < 9.2	Aa3	AA-	AA-
8 ≤ r < 8.5	A1	A+	A+
7.6 ≤ r < 8	A2	A	A
6.9 ≤ r < 7.6	A3	A-	A-
6.2 ≤ r < 6.9	Baa1	BBB+	BBB+
5.6 ≤ r < 6.2	Baa2	BBB	BBB
5.0 ≤ r < 5.6	Baa3	BBB-	BBB-
4.4 ≤ r < 5.0	Ba1	BB+	BB+
3.9 ≤ r < 4.4	Ba2	BB	BB
3.3 ≤ r < 3.9	Ba3	BB-	BB-
2.7 ≤ r < 3.3	B1	B+	B+
2.2 ≤ r < 2.7	B2	B	B
1.6 ≤ r < 2.2	B3	B-	B-
1.4 ≤ r < 1.6	Caa1	CCC+	CCC+
1.2 ≤ r < 1.4	Caa2	CCC	CCC
1.0 ≤ r < 1.2	Caa3	CCC-	CCC-
0.8 ≤ r < 1.0	Ca	CC	CC
0.6 ≤ r < 0.8	C	C	C



0.4 ≤ r < 0.6	C	DDD	DDD
0.2 ≤ r < 0.4	C	DD	DD
0 ≤ r < 0.2	C	D	D

9.4.2 Context

According to the rating report of Bravo Capital, a financing company, the note of the companies follows a Normal distribution:

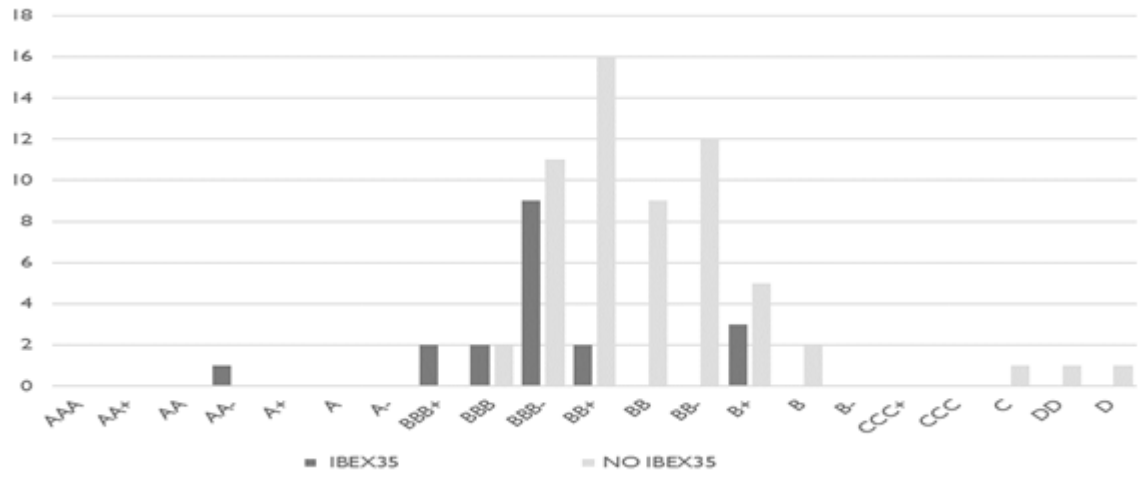


45% of Spanish companies are between BBB- (investment grade limit) and BB +.

The main detriment of the companies that obtain high notes resides in the low profitability for the shareholder when being in such note. However, an AAA company has the potential to acquire much more debt than it has, for example, to grow organically in new markets or through acquisitions. On the opposite side of the bell, the drawback is the lack of liquidity and solvency. This situation results in the bank demanding the repayment of debts, a phenomenon that leads many companies to tender or liquidation to repay the above-mentioned debts.

9.4.2.1 Listed company´s rating

Comparison between IBEX 35 companies and companies in the continuous market.



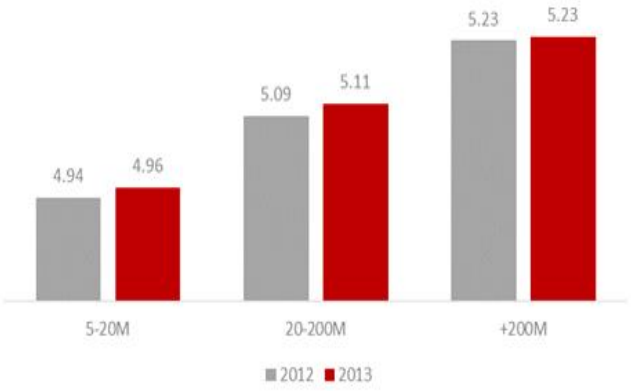
The IBEX 35 companies have a better distribution than the Spanish average, while the NO IBEX 35 worse than the Spanish average.

9.4.2.2 Is size an important variable to obtain a better rating?

The Spanish business environment is composed, mainly, by small SMEs (Small and Medium Enterprises). Taking into account that larger companies get a better Rating, Spanish companies are penalized by this factor.

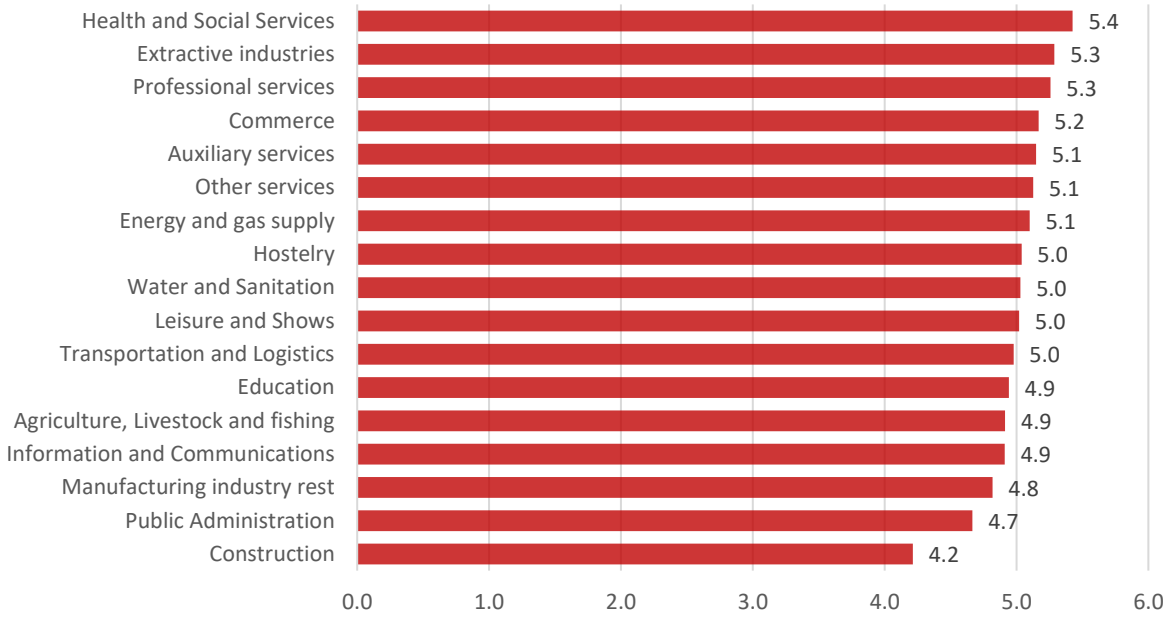
But does it make sense that a bigger company get a better Rating just for being bigger? It actually does, this is known as a portfolio effect, that is, it is less likely that a large company will lose 20% of its clients in a given period of time, than a small company find itself in this situation during the same period. Therefore, time, statistically speaking, happens to play a key role in the phenomenon of the loss of customers.

Rating average by range of sales (in millions of Euros €)



9.4.2.3 Is the risk different depending on the sector?

Rating by sectors of Spanish companies



Claudia Tarragona, CEO of the company InfoEmpresarial Spain, realized in 2014 that there were very few companies that offered Rating services in Spain. Given the banking crisis that was taking place in the country, Claudia observed that the Rating tool could be key to reduce the risk balance sheet of Spanish SMEs; and consequently, bring the whole Spanish market to a more comfortable position, in terms of systemic risks.

However, Claudia had a big problem. She was not an expert in Rating and believed that no one in her company was. One morning, Claudia invited José Campos, the company's financial director, to grab coffee and asked him about his understanding on the subject of Rating. His response was direct. Like Claudia, José was not knowledgeable about the topic. However, as a financial director, he perfectly knew everything regarding capital and debt models, such as the Capital Asset Pricing Model and many other valuation models. Additionally, he knew that the company's bank used internal models to evaluate them, but he did not know how to dig deeper into the matter either. Sometime into the conversation he had a great idea, perhaps, they could talk to Nakamura (employee at the Takashi partners fund) whom he already knew and who was, in addition, an expert investment analyst in Rating.

The next morning, Nakamura showed up at the InfoEmpresarial office, invited by his friend José Campos. Without exactly knowing what the reason for his visit was, Nakamura began to complain about the low effectiveness of the expansive economic policy. According to Nakamura, the investment banking market was in an extremely difficult position. The crisis in which the companies were immersed had worsened their rating scores, increased the risks for investments and, therefore, decreased margins. Also, as if this was not hard enough, interest rates at historic lows tightened, even more, the profit margin. But, for both the Japanese investment analyst and for the financial director, in the current scenario there were situations that were very bizarre. From his point of view, the flow of money from the central bank did not seem to be really reaching the companies. In short, the banking industry also had to fulfill its obligations, and these met the new capital needs imposed by Basel; according to which, lending money to companies in a bad financial situation was not an idea that they were passionate about.

Having reached this point, Claudia (CEO) commented that everything the Takashi fund analyst was telling them was very interesting. However, the reason for his call was related to a proposal. Claudia wanted to offer a job to him in order to build a rating system for InfoEmpresarial, with the main goal of helping Spanish companies to improve their credit ratings. After commenting on the proposal, Nakamura, in a very polite way, opted to decline her offer. Nakamura claimed that although he was a great user of the Rating tool, his knowledge limited to the use of Porter's 5 forces[1], some profit, activity ratios and debt ratios and, finally, the Rating grade to guide his investment decisions. In a simple way, his job was to calculate the valuation of companies. He does that by using EBITDA factors, adjusting this valuation for the predictions of growth or decrease, then adjust it again using the 5 forces of Porter, and finally adjusting it once again depending on how risky the business is, and the riskiness comes out of the Rating (above BBB- good, below BB+ risky). The investment analyst reiterated that he was not the right person for that job and that, unfortunately, he did not know anyone in the international agencies or in a bank that could help them.

A few moments later Nakamura retreated; he remembered that he knew a person who could help them. It was a headhunter named Fred Asterix who was dedicated to the search of managerial candidates in the Risk industry and who, according his understanding, had work for Rating agencies in some occasions. So, Nakamura, after recommending Fred, facilitated his contact information to Claudia and José, who greatly appreciated it since they would need to contract his services to be able to conduct an efficient search for a candidate. Finally, both the CEO and the CFO thanked Nakamura one more time as they were sure they had found the right contact that would help them hire this "extraterrestrial" professional. couple of weeks later, as a result of some emails and telephone conversations with the headhunter Fred Asterix to whom Claudia transmitted all her needs, Fred provided his first candidate.

The first candidate was Joaquin Quintana, a true star professional in the corporate banking sector, with more than 15 years' experience as an Enterprise Risk Analyst and who currently held the position of Enterprise Risk director at the VVBA bank.

Used to dealing with the enterprise world, Joaquin quickly began to explain to Claudia what methods he used when evaluating a company. Furthermore, since he was a very practical person and also a professor at a well-known business school, he quickly put aside the subjective field of conversation. Thus, he invited Claudia to do some real calculations with Ibex35 companies.

He kindly asked Claudia to enter the web https://finance.yahoo.com, look for the Telefonica company (the ticker is TEF.MC) and click on Financials. Once there, they saw Telefonica's profit and loss account. From the analysis of this Financial Statement they could extract that as of December/31/2016 the Net Profit Margin had been 4.47%. This result was reached by dividing Net Income / Total Revenue.

Net Profit Margin = Net Income / Total Revenue = 2369000 / 52903000 = 4,47%

Claudia found it very interesting. For her, as for any professional linked to the financial world, Net Profit Margin is a very important ratio, in order to know the performance of a company.

However, Joaquín stressed that this ratio only used variables from the income statement and that it therefore biased the analysis for a single period, in this case a single year. To avoid this bias, Joaquín suggested that it would also be convenient to look at the balance sheet ratios; which they did by clicking on the Balance Sheet tab. In this way, they calculated one of the possible debt ratios, Total Debt / Total Asset. For the numerator they needed to add the short debt and the long-term debt that

[1] Porter's Five Forces - A Practical Example: https://www.youtube.com/watch?v=OCnlArFuU-E

appeared in the balance sheet. Short-term debt appeared as [Short / Current Long Term Debt] and long-term debt [Long Term Debt].

$$\frac{\text{Total Debt}}{\text{Total Assets}} = \frac{[\text{Short Current Long Term Debt}] + [\text{Long Term Debt}]}{\textbf{Total Assets}}$$
$$= \frac{60{,}361{,}000 + 45{,}612{,}000]}{\mathbf{123{,}641{,}000}} = 85{,}71\%$$

After observing the result of the debt ratio, Claudia thought that Telefonica was quite indebted.

To which Joaquín replied that only with this information he was not able to tell much about their financial status. "Here is where my technical limitation appears", commented the risk analyst. To analyze this company and know if a Net Profit Margin of 4.47% is reasonable or if Total Debt / Total Assets = 85.71% is a highly indebted company, it would be required to take into account the whole industry. That is, compare Telefonica with other companies in the same economic sector. He also let Claudia know the enormous amount of time it would require to manually calculate those ratios for not more than 100 companies. At this point in the conversation, Joaquín honestly told Claudia that he was not the professional she was looking for. To carry out the sectoral analysis and, finally, to build an own Rating model, what she needed was a Model Analyst or, even better, a professional who was being born in that exact moment. Finally, Joaquín asked Claudia: "Have you ever heard about Data Scientists or Big Data professionals?"

Given the refusal of the CEO, Joaquin replied that he knew a Big Data course professor named Sasha Popov and that, probably, he was the person she was looking for. Joaquín went on to note that "our" Big Data professor had worked for many years in the areas of Models, Methodology and Analytics in the Banking sector. And explained that in order for her to understand how a data scientist could help her, she would have to think about what they have just done by taking the yahoo information manually to calculate the ratios. Finally, he stressed that Sasha could build a program called Web Scraper that would be able in a matter of minutes to collect all the data from the income statements and bance sheets of hundreds of companies. With this tool he could compare Telefonica with numerous companies in the sector. He also mentioned that Sasha would be able to build the model not only with two ratios but using hundreds of them.

On top of that, Sasha would use Machine Learning methods and Model Development techniques to discover which ratios are the most relevant. He would finally build a Rating model giving weight to each of these ratios. "Do not be scared if he starts calling the ratios: variables." Joaquin said.

"Well, I'll call Professor Sasha today," Claudia answered. To which she concluded: "It has never been so difficult to hire a profile. These people from Big Data need to sell themselves better, because I had heard about them, but I thought they were computer experts, or people that knew how to deal only with "big databases", like our IT team.

So that is what Claudia did, and after two interviews in which Claudia became very impressed with Sasha, and Sasha became very excited with the project, the agreement was imminent. Sasha's only requirement to sign the contract was that he wanted to also hire a student of his. Claudia had no problem accepting his request. Thus, on March 1, 2014, Sasha and Rita (student of Sasha) began their job challenge in the newly created Rating department.

When the two of them arrived on their first day of work, they already had a dual-core Windows laptop and 16 Gigabytes of RAM for each of them. It was the configuration that Sasha had requested before starting to work. He knew that he would not need more than 2 processors, but since he was going to

use a Python library (called pandas and that uses data in memory) his bottleneck would be memory. For this reason, he asked for the maximum that could be given, which was 16 gigs of RAM. One of the first tasks that Sasha entrusted to Rita, her student and now employee, was to ask the IT department to install the most advanced version of Anaconda Python, since the last time he had a look they had the 3.6 one. As it was possible that the IT department did not know where to download the program and its most up-to-date version, to avoid any confusion, she sent the following links:

- https://www.anaconda.com/download/
- https://docs.continuum.io/anaconda/install/windows

Moreover, to ensure that there were no incidents, in case they had questions, she also attached a video, which Sasha sent her once: https://www.youtube.com/watch?v=EbYGBANqDdY. However, as they were computer scientists they had no problem installing, quickly and easily, Anaconda Python.

While our Director of Rating was still arranging his table, Rita approached him and told him that he had been preparing a small list of companies in various sectors at home. This list was from companies in the United States and each of them had their ticker symbols from Yahoo Finance next to the name. Rita consulted her superior if he would like her to start writing a Python code to go and Scrape some information from Yahoo Finance.

Sasha said that he would like to see the list before, so Rita gave him the following link:

1. https://www.dropbox.com/s/2x1rmt2ma96j2my/yahoo_ticker_sample.xlsx?dl=1

As soon as the file was opened, Sasha was happy to see that Rita had prepared 5 sectors with a reasonable number of companies. Thus, they could be able to rank them according to several ratios that they would build. He directly asked Rita to work on the code. However, he asked her not to complicate her life at that stage downloading income statements and balance sheets, and simply focus her efforts on downloading the information from the Key Statistics (KS) of Yahoo; He gave her the example to download the Google KS, which was just a matter of using the 'GOOG' ticker he had in his file on the link below:

2. https://finance.yahoo.com/quote/GOOG/key-statistics?p=GOOG

Rita, who had not had any doubt, clicked on the link and using the right button of his mouse clicked again 'View source code of the page'. Although Rita was not an expert in HTML, she began to alternate between the original page and the source code. She started looking for the values that she saw on the original page in the source code, with the intention of understanding how Yahoo's HTML was built, it did not take long to identify that she simply had to look for the concept she wants with a ">" at the beginning and then the tags </ td> </ tr> at the end. For example: "> Market Cap" and "</ td> </ tr>".

In the afternoon of the same day, with Anaconda Python installed on her "supercomputer", Rita started working on her Python code. She had previously worked with a Python library called Scrapy. However, this time she preferred to write the code that would "peel the onion" of the Yahoo Finance Key Statistics HTML from scratch, because she realized that Yahoo's HTML had small variations in the tags that could make her life a hell if she used Scrapy.

After two intense days of work, Rita had already built the Python code to scrap Yahoo Finance Key Statistics data:

3. https://www.dropbox.com/s/0jdut0jl23tn05h/YahooFinanceKeyStatisticsScraper.py?dl=1

She also had the Excel file in her hands with the results:

4. https://www.dropbox.com/s/66vnpxeiesqcg78/yahoo_ticker_sample_scraped.xlsx?dl=1

Rita's first contact with modeling had been the Machine Learning course in Finance of Professor Sasha, who was now her boss. The professor told him in several sessions that the equation that was used for modeling was:

$$y = f(X)$$

Where:

- X, are the independent variables or input (the variables that Rita had just scraped).
- f, was the predictive method applied to adjust a model and later predict the "y" using X (Rita already knew the Logistic and Linear Regression, Decision Tree and other methods, but she was not sure which one to apply).
- y, was the objective variable, variable target or variable output, which is the variable that requires prediction.

The "y" was a mystery for Rita, since she had never worked in a model before. In addition, she did not know how to get it and how to define it, so she did not hesitate to ask her boss on how he should proceed.

Sasha told him that, in Rating, the variable to be predicted is the bankruptcy of a company in a certain future period of time. The bankruptcy in most of the cases entails the closure of the company with debt. However, that is not it, to that we must add other undesired situations, such as for example: a company can generate unpaid percentages, because many times the bank prefers to make a 30% reduction of the original debt (and recover 70%) rather than taking the company to court (and in many cases to the subsequent liquidation) and risk losing an even greater percentage. In what concerns the time window, Sasha commented that it would depend on the time of the credit product to which the model is to be applied to. That's why the Rating models are not one, but many with different time windows. The normal thing is to find at least two versions, the short-term version (of one year) and the long-term version (of 5 years). Once this point was made, the head of the Rating department went on to say that the first objective variable they could create was:

1 – (bad) companies with debt in the next 12 months.
0 – (good) anything else.

Rita interrupted: "In this case we would apply a classification method, since the output variable is binary, right?"

Sasha replied affirmatively. However, the problem that had at that time InfoEmpresarial Spain was that they did not have internal data of defaulted companies, so, to create a model close to reality, they had to use external data, probably from a Credit Bureau such as Experian or Equifax. These data would cost them money, but without it, they will not be able to build a model. Getting such data would take a long time, so Sasha asked Rita to start building a code for the model with an approximate objective variable, that is, with what we call a Proxy.

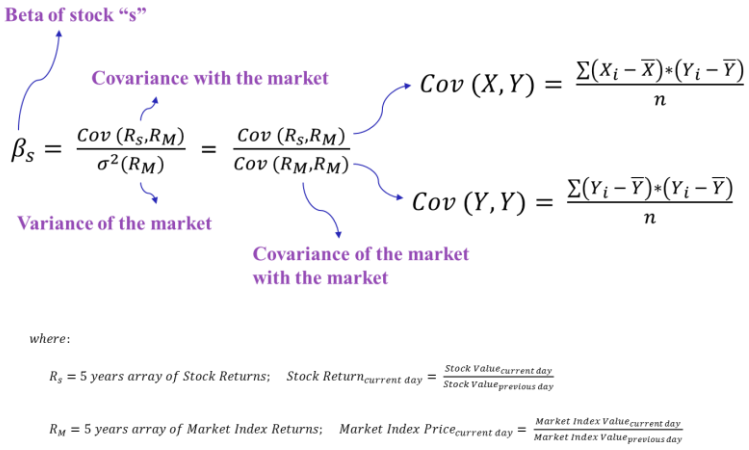
This Proxy could be the Rating score of other agencies or the probability of default given to them by the Bloomberang system, which due to the observations Sasha believed it was based on the behavior of the stock market.

One more time Rita interrupted to add: "But now we would be talking about a regression method, because the variable output has many possible values, right?

Sasha replied that there was more than one answer to her question, but that for the test they were going to do a Linear Regression would be enough.

So as not to complicate further their first development, they would use something they already had. It was the same principle that the Bloomberang company used, a measure of the company's volatility. This would be their "y". The idea was the confidence in the wisdom of the crowds ("The Wisdom of Crowds"), that is, in the coming days the market would become more volatile if there was more risk in a company or would become less volatile if there was less risk. "Have you heard about the Beta market?" Sacha asked, and he continued "If you have not seen it yet, you have it in your yahoo finance scraped information."

If Rita really wanted to understand what the Beta was, Sasha suggested to take a look at slide 21 of his time series course. After a few moments, Rita looked for it and found it:



Sasha continued: "Now, since I know you'll be interested to learn how to calculate it in Python, take a look at this blog, it can help you": http://gouthamanbalaraman.com/blog/calculating-stock-beta.html

When Rita saw that he had all the elements for this simulated development, she told her boss that she would "get down to work" with the code. Nevertheless, Sasha, who would be at her side in the development and warned her that the most important thing for a Data Scientist is not to capture the data and apply methods, but to understand the problem that is being modeled and how that translates into actions that should be taken on the data.

Rita did not understand, but she sat enthusiastically next to her boss to work on her first project outside the Master in Big Data.

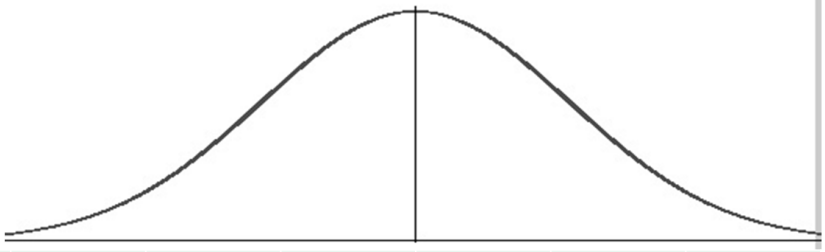
Sasha saw Rita's doubtful face, so he gave her the example of what would happen if Dr. Pepper, a direct competitor of Coca Cola and Pepsi, ceased to exist. Later, he added that one possibility was for people would stop drinking cola, but he was convinced that most Dr. Pepper consumers would migrate on to consuming cola from the competition, which are, Coca Cola and Pepsi. For our Big Data professor, the situation was comparable to two individuals who were escaping from a lion in the African savanna. In this situation the individuals did not need to run more than the lion, it was enough

to be faster than his opponent. Sasha continued: "Rita, in terms of data, these anecdotes are translated in that for all the variables that we have scraped, we need to create a new transformed variable using its order, its ranking".

Sasha added that he had seen cases in which transformations of the "percentile" type were applied (something easily achieved using the "percentile" function in Python) or Normalization (applying the formula (Variable - median) / standard deviation); but he saw theoretical and practical problems in these transformations and commented that his preferred transformation technique is Slot or Bucketing Normalization.

Sasha turned again to a slide of his course to explain to Rita how the transformation of the variable Return On Assets (ROA) would look like in 6 ranges.

ROA = Net Income / Total Assets



Percentage of companies	7.14%	14.29%	28.57%	28.57%	14.29%	7.14%
ROA to be in the group	ROA < -17%	-17% <= ROA < -2.3%	-2.3% <= ROA < 0.25%	0.25% <= ROA < 1.6%	1.6% <= ROA < 15%	ROA > 15%
Bucket Associated	5	4	3	2	1	0
Translation of Bucket	D	CC	B	BB	BBB	A

The new variable would receive the values of the "Bucket Associated" row, which are 0, 1, 2, 3, 4, 5 and 6 according to the value of the ROA of the company.

Sasha explained that if Telefonica's ROA was 2.93%, his bucket would be 1 (which would translate into a "BBB" if the Rating formula considered only one variable.) He added that the sizes of the ranges change to create a normal distribution of the new variable, where the first rank has 7.14% of the population and the second has 14.29%, then 28.57%, and finally again 28.57%, 14, 29% and 7.14%.

Sasha concluded by saying that the sample of companies scraped by Rita (yahoo_ticker_sample_scraped.xlsx) were composed by between 16 and 18 companies for each sector. However, some variables would not have data for all the companies. Thus, what should be done was to translate the original variables into a score between 0 and 5 attributing to the worst company a score of 5, the following two a score of 4, the next 4-6 companies a score of 3, the following 4-6 a 2, the next two a score of 1 and, finally, the best company would score a 0.

Grades	5	4	3	2	1	0
#companies	1	2	4-6	4-6	2	1

Rita, after having reviewed all the codes she worked on in the classes of his Big Data master, and consulting some forums, she found the answer of how to implement the transformation and develop a first simple model using Ordinary Least Square of Statsmodels in Python.

5. https://www.dropbox.com/s/40n8qo1glfulnim/YahooFinance_Grouping_ModelDevelopment.py?dl=1

OLS Regression Results

Dep. Variable:	Beta	R-squared:	0.932
Model:	OLS	Adj. R-squared:	0.873
Method:	Least Squares	F-statistic:	15.75
Date:	----------	Prob (F-statistic):	1.78e-16
Time:	15:01:15	Log-Likelihood:	-15.993
No. Observations:	86	AIC:	112.0
Df Residuals:	46	BIC:	210.2
Df Model:	40		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Market Cap_group	0.0255	0.110	0.232	0.818	-0.196	0.247
Enterprise Value_group	0.0075	0.092	0.082	0.935	-0.177	0.192
Trailing P/E_group	-0.1675	0.076	-2.208	0.032	-0.320	-0.015
Forward P/E_group	0.1638	0.076	2.153	0.037	0.011	0.317
PEG Ratio_group	-0.0261	0.048	-0.547	0.587	-0.122	0.070
Price/Sales_group	0.0474	0.147	0.323	0.748	-0.248	0.342
Price/Book_group	0.0791	0.058	1.353	0.183	-0.039	0.197
Enterprise Value/Revenue_group	0.0131	0.137	0.096	0.924	-0.262	0.288
Profit Margin_group	-0.0770	0.100	-0.772	0.444	-0.278	0.124
Operating Margin_group	0.1314	0.097	1.357	0.181	-0.064	0.326
Return on Assets_group	0.0462	0.085	0.545	0.588	-0.124	0.217
Return on Equity_group	-0.0442	0.072	-0.610	0.545	-0.190	0.102
Revenue_group	0.2286	0.114	2.006	0.051	-0.001	0.458
Revenue Per Share_group	0.0319	0.070	0.456	0.651	-0.109	0.173
Quarterly Revenue Growth_group	0.0524	0.050	1.054	0.297	-0.048	0.152
Net Income Avi to Common_group	0.0288	0.094	0.308	0.760	-0.160	0.218
Diluted EPS_group	-0.1729	0.097	-1.783	0.081	-0.368	0.022
Quarterly Earnings Growth_group	-0.0678	0.044	-1.527	0.133	-0.157	0.022
Total Cash_group	-0.0398	0.069	-0.576	0.567	-0.179	0.099
Total Cash Per Share_group	-0.0238	0.065	-0.368	0.715	-0.154	0.106
Total Debt_group	-0.0654	0.077	-0.855	0.397	-0.219	0.089
Book Value Per Share_group	-0.0359	0.059	-0.606	0.548	-0.155	0.083
Operating Cash Flow_group	0.0050	0.056	0.090	0.929	-0.108	0.118
52-Week Change_group	-0.1167	0.058	-2.017	0.050	-0.233	-0.000
52 Week High_group	0.0569	0.236	0.241	0.811	-0.418	0.532
52 Week Low_group	0.0920	0.127	0.724	0.473	-0.164	0.348
50-Day Moving Average_group	0.1212	0.224	0.541	0.591	-0.330	0.572
200-Day Moving Average_group	0.0125	0.127	0.098	0.922	-0.244	0.269

Avg Vol (3 month)_group	0.1619	0.169	0.960	0.342	-0.178	0.501
Avg Vol (10 day)_group	-0.1539	0.147	-1.044	0.302	-0.451	0.143
Shares Outstanding_group	0.0794	0.084	0.950	0.347	-0.089	0.248
Float_group	-0.0295	0.091	-0.323	0.748	-0.213	0.154
% Held by Institutions_group	0.0592	0.052	1.135	0.262	-0.046	0.164
Shares Short_group	0.0392	0.147	0.266	0.791	-0.257	0.336
Short Ratio_group	0.1243	0.054	2.308	0.026	0.016	0.233
Shares Short (prior month)_group	-0.2159	0.136	-1.585	0.120	-0.490	0.058
Forward Annual Dividend Rate_group	-0.1644	0.119	-1.387	0.172	-0.403	0.074
Forward Annual Dividend Yield_group	0.2229	0.088	2.538	0.015	0.046	0.400
Trailing Annual Dividend Rate_group	-0.0233	0.125	-0.186	0.853	-0.276	0.229
Trailing Annual Dividend Yield_group	0.0111	0.085	0.130	0.897	-0.161	0.183

Omnibus:	2.363	Durbin-Watson:	1.629
Prob(Omnibus):	0.307	Jarque-Bera (JB):	1.728
Skew:	-0.154	Prob(JB):	0.422
Kurtosis:	2.378	Cond. No.	130.

However, Rita faced multiple questions:

1. Was this model reasonable? How can I evaluate it?
2. Assuming that the model is improvable, how can I improve the model? What can I look at, adjust or change in my process to improve the model?
3. Assuming that my model is good:
    a) How can I transform my prediction into the letters AAA, BBB, BB, etc?
    b) Where can an SME apply this model once it has been built?
    c) If we decide to sell our Rating for thousands of Euros, who pays the Rating, the analyzed company or the company that wants to consult the Rating? Can there be any conflict of interest? Did we achieve the objective of giving access to Rating reports to SMEs and make them available to SMEs?
    d) If we charge our Rating report cheaply, how can we make the product profitable knowing that it may cost more to produce it than to sell it? Does this change the possible conflict of interest? Did we achieve the objective of giving access to Rating reports to SMEs and make them available to SMEs?

When Rita, very interested and intrigued, asked Sasha her questions, Sasha asked her to answer these questions herself and recommended her to go to his next class, where precisely this case would be discussed. This way they would be able to try to answer all these unknowns. In addition, he pointed out that since Rita had followed all the steps indicated by him, she would be very interested in the class.