**Data Science – Trading Houses case**

I choose to write this little note to explain concretely what is data science, as there is no closed definition, and to show some real cases application for Commodities Trading Houses.

This trendy field englobes lot of concepts, and data science is often viewed like long and costly projects for geeks, while it can have day to day implications in all parts of the business. It gives tool while the final objective is the optimization of the business: saving time and cost, bringing rigor in some aspects, and even turning some support functions into profit center.

***Definition – Data at the center***

(*Wikipedia*) Data science, is an interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured.

This introduction from the free encyclopedia gives a broad overview of what we call data science. This term became a buzzword after the Harvard Business Review called Data Scientist the “sexiest job of the 21st century”, but we see that there is no real academic definition of that. What we can take away from Wikipedia is that data are in the center of the stage, and the objective is to extract knowledge from that.

It is more the practical intersects of several fields like Computer science, Economics, Statistics, etc. that gives him its definition.

Often compared with statistics, which is its “parent” the main difference based on its current usage is:

* Statistics are more used to look at the past, for explanation
* Data science, is used for forecasting power

An article from *The Ecomomist* has recently put the light on data: they are now more valuable than what oil used to be. Companies know they are sitting on a gold mine of information, but as of today they are often not able to exploit it, due to processes, organization failure or lack of technical knowledge.

Within trading houses the information is often reported through phone calls or unstructured mails, so even if the information travels quickly through these ways, they are not the most rigorous to track them and structure them for later analysis.

This new field, even if trending among lot of industrial sectors, is under exploited among the commodity trading industry: these companies do not use the data they are sitting on. It is often considered to be in the scope of the market intelligence analysts, or economists, to deal with it but the potential usage of it is much wider.

Indeed, the intensive use of data can empower any part of the business, as illustrated below.

***Concrete use cases***

In this part, we will try to highlight what a more data oriented approach can add to various aspects of the business, some being close to classic analysis, some other closer to this modern approach of machine learning.

/ Processes improvement & cost reduction

A very significant part of internal processes in a trading house are made manually by people. This increases costs and errors. A part of the workforce is employed to do these repetitive tasks, while they can be reallocated in tasks with more value added for the company, leaving the simple ones to computers.

The automation doesn’t belong directly to this data science stuff, but is more a co-product.

*Example (to be developed):*

Position reporting: we do not know how all trading houses are monitoring their cash and derivatives positions, but generally it is quite archaic.

In the best case there is an extraction process from a global system, reporting positions in excel files and then an email is manually sent to the appropriate trader for him to check if what he has is matching with the system. Errors are common as human is implicated in lot of parts of this process.

There should be some routines doing it daily without any human intervention, and that requires some coding skills from the computer science area.

/ Risk management

Estimating and managing the risk is critical for trading houses. Apart from the processes, there is often a lack of information regarding physical prices. Even if ongoing prices can be checked through brokers to estimate a mark to market value, the issue is more with historical data, used to compute classic risk metrics.

*Example (to be developed):*

The classic example is when a company takes a position in a less liquid market: let’s say a barley in Ukraine. Apart from in house data, there is no easily available data to estimate the risk from the management standpoint. In that case lot of hypothesis need to be done to still be able to compute some metrics..

One solution to estimate it could be to collect some data related to Ukrainian Currency, Futures Markets, Cash prices available, and to perform a Principal Component Analysis (PCA) to estimate the risk factors. It is a statistical procedure that convert a set of observations into a set of linearly uncorrelated variables, which are the factor we want to highlight.

/ Market Intelligence

The market intelligence is the first example of use of Data Science that should come in mind. Developing econometric models to analyze markets. The idea is to be able to take decisions based on analysis, or more structured analysis at least.

*Example 1 (to be developed):*

The USDA proposed a yield model based on weather. The model description is available here:

<https://www.usda.gov/oce/forum/past_speeches/2013_Speeches/Westcott.pdf>

The model is a basic multilinear regression model, and has been developed by the USDA after some droughts in US to help the USDA analyzing the impact of climate on weather. What we highlight here is that by using some small improvements, based on a small set of data, we can improve the way of working.

A model, as the previous understanding, was for explanation. Machine Learning approach is based on the cross validation, i.e. how well the model

behaved in a (simulated) real situation

*Example 2 (to be developed):*

Use of new satellite image to develop monitoring of crops, to estimate production data like acreage or even yield, with such data like NDVI, or climate impacts like floods.

/ Investment

There are several ways of using data in a valuation process for an asset transaction. The classic methods for a valuation in the economic theory are:

* Comparables
  + Raw analysis, for example what can be found in some document produced by banks, are the valuation reported to a certain metric (like Valuation / Tones of mineral as Reserve, very common in mining)
  + In order to improve that, we can use Heuristic models, where all the metrics that were previously evoked, are considered at the same time
* Cash Flow
  + Accouting analysis, making some growing assumptions:
    - Previous experience in a company where some inflation hypothesis where applied to the cash flow, based also on some price assumptions.
  + More complex approach, where the different items of balance sheet can be modelized
    - Ex: the production future of a mine, based on some theoretical models
* Real Options

/ Database

As discussed in the introduction, one of the main problem for the long run in the trading houses, is the ability to keep and track the information.

The internal information travels through emails or phone calls, the external information is often through pdfs reports that are deleted once read.

This information at some point will probably be a huge asset for these companies, we can imagine they could even sell it. Before that they need to find a way to structure nd organize it with some rigors.

*Example (to be developed):*

We can imagine that some techniques derived from Natural Language Processing would allow to extract sentiments or data from the different reports.

The data can be stored in an internal system, for posterior analysis, and the reports sentiments for backtest: what was the take away of this report and what was the market reaction. Is this companies providing accurate forecasts ?

Cash prices contain all information and are proprietary of the company. These is very important to structure them and keep track of them

***References***

Wikipedia – Data Science

<https://en.wikipedia.org/wiki/Data_science>

Data is the new oil

<https://www.economist.com/news/leaders/21721656-data-economy-demands-new-approach-antitrust-rules-worlds-most-valuable-resource>