New data, new methods: highlights from the Inquire quant seminar – Part 1

A major quant seminar in the UK brought out the latest academic thinking on the advantages – and dangers – of using new data techniques like machine learning, AI and text analysis to improve investment outcomes

- Machine learning – useful for FX, commodities and finding patterns in less-invested asset classes
- Backtesting in the machine learning era needs to be honest, precise and scientific
- Finding value from financial text

The recent Inquire Europe and Inquire UK joint Spring 2019 seminar in Windsor, UK, included presentations and discussion of a broad range of topical investment-related issues.

Machine learning and AI

Sandy Rattray (https://uk.linkedin.com/in/sandy-rattray-7bb1628), CIO at Man Group, viewed machine learning and artificial intelligence as the biggest fad in 2018, but he believes machine learning techniques are not particularly useful for forecasting asset returns at horizons of six months and longer.

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Accept (#)  Read More (http://www.bnpparibas-ip.com/central/cookies-policy.page?l=eng)
Where Sandy sees useful applications is in trading, in particular FX, where Man Group collects most big data, and in commodity markets. He also sees it as useful for text mining and for discovering patterns in less exploited asset classes.

In terms of new trends in quantitative investing, Sandy highlighted credit multi-factor as something to watch. He also wants to increase interaction between quantitative and fundamentals fund managers but perceives the poor quantitative skills of some fundamental managers as a difficulty to overcome. The fact that managing concentrated portfolios is difficult for quantitative managers does not help either, in his view.

When it comes to new data, Sandy sees many people creating new datasets for fundamental managers, but he is unsure of their value, in particular taking into account their (often high) cost and the trend for lower management fees, even if he certainly does not think data budgets are likely to go down.

Finally, when it comes to long-term investing, he believes that there are not many factors adding value. He advises investors to stick with those they know.

A backtesting protocol in the era of machine learning


The first is to avoid HARKing (https://journals.sagepub.com/doi/abs/10.1207/s15327957pspr0203_4) (Hypothesizing After the Results are Known). Data-mined factors should be treated with much more scepticism than a factor from economic theory. The second is to be aware of the multiple testing problem (https://en.wikipedia.org/wiki/Multiple_comparisons_problem), i.e. that when we run a hypothesis test there is only a small chance of finding a bogus significant result but if we run thousands of tests, typical of machine learning, then the number of false alarms increases dramatically. Researchers should not stop researching as soon as they find a good model. All variables set out in the research agenda should be investigated.

The third is to be aware of data integrity, to what extent outlier exclusions and data transformations were set in advance and make sense, and to what extent results are resilient to minor changes in transformations.
The fourth is to be honest with cross-validation, avoiding later modifying an in-sample model to fit out-of-sample data, making sure that the out-of-sample analysis is representative of live trading, and that realistic transactions costs were accounted for.

The fifth is awareness of model dynamics – making sure the model is resilient to structural changes and that steps were taken to minimise overfitting of the model dynamics, the tweaking of the live model and the risk of overcrowding.

The sixth is to avoid complexity, i.e. that models avoid the curse of dimensionality, that the simplest practicable model specification is retained, and that results from machine learning can be interpreted rather than used as black box. Regularisation (https://en.wikipedia.org/wiki/Regularization_(mathematics)), i.e. introducing constraints to achieve model simplification that prevents overfitting is good practice.

Finally, the seventh is the need to have a scientific research culture, i.e. to reward quality rather than just the finding of a positive back-test of a strategy.

A good illustration of how violating these principles can lead dangerous results based on overfitting was that of a strategy that invests in an equally weighted portfolio of stocks with “S” as the third letter of their ticker symbol and shorts an equally weighted portfolio of stocks with “U” as the third letter of their ticker symbols. This strategy, cited in their paper (https://jfds.ijournals.com/content/1/1/64), was found after trying thousands of combinations. The strategy would have generated a significantly high risk-adjusted return when applied to US stocks between Jan-63 and Dec-15, even in the Global Financial Crisis of 2008, and with a low turnover of less than 10% per year.

Where’s the value in unstructured data?

Steven Young from Lancaster University (https://www.lancaster.ac.uk/lums/people/steven-young) gave a tutorial on the value of unstructured data for investing. The motivation for using and modelling with text is simple: many facts are not easily translated into summary numbers, and nuances are better expressed in text.

Steve’s objective was to propose a framework for using financial text and thinking about the sources of value associated with text. Clearly, even before starting, researchers should ask themselves about the comparative advantage of their research. Text mining packages such as those found in R, Python or SAS mean that the low hanging fruit has already been found, picked and eaten.

For example, when it comes to text analysis of 10-K and 10-Q reports for US firms, more than 60 papers have been produced. Once the added value of the project is clear, Steve proposes a four-step framework:
1. Corpus Creation, corresponding to the Definition of the Problem

2. Cleaning & Pre-Processing

3. Annotation

4. Processing, corresponding to the Search of Meaning.

For the Corpus Creation, the three broad genres of finance-related textual content are

1. forum, blogs and wikis

2. ii) news and research reports

3. iii) content generated by firms.

Analysing multiple genres to contrast content views and triangulate results is likely to add more value. It is also important to decide whether to use the entire text or just part of it. Sharper conclusions may be possible by focusing on particular sections. Then one must harvest the textual data, clean it and pre-process it removing unwanted contents and organising the relevant unstructured text into structured text or numerical data arranged in tables.


Manual tagging is more likely subjective but can play an important role for training in Big Text applications. In turn, automated tagging may use a number of available resources for Part-Of-Speech (POS) (https://nlp.stanford.edu/software/tagger.shtml), for morphology, grammar and syntax (http://nlp.stanford.edu:8080/parser/), for semantics (http://ucrel.lancs.ac.uk/usas/) and for pragmatic annotation (https://aclweb.org/anthology/papers/L/L14/L14-1287/).

Finally, processing may be based on simple word frequency counts using general dictionaries such as DICTIEN (https://www.dictionaries.com/dictionaries-overview) or based on extract from the Inquire (http://www.wh.harvard.edu/~inquirer/). LWC (http://lwc.wpeng.com) and others, or based
on more refined approaches based on domain-specific lexicons such as Netlingo (https://www.netlingo.com/dictionary/all.php) or Provalis (https://provalisresearch.com/products/content-analysis-software/wordstat-dictionary).


In Steven’s view, it is clear that automated text mining is a tool of increasing importance for quantitative investment.

For more articles by Raul Leote de Carvalho, click here > (https://investors-corner.bnpparibas-am.com/author/raul-leotedecarvalho/)
