Quantitative Research on Market Predictability:
Machine Learning and the Efficient Market Hypothesis

Aurouet Lucas

Master EKAP
IAE Nantes
Under the supervision of Prof. O.DARNE 01/03/2021

Contents

1 Introduction

Trading is futile. This is an oversimplification of the Efficient Market Hypothesis. According to Eugene Fama who formulated this hypothesis in 1970[1], prices at a certain date incorporate all the available and relevant information at that date. There are never overvalued nor undervalued assets, fundamental and/or technical analysis are therefore pointless as all the signals are already reflected by prices on the market. One cannot beat the market, is another way of expressing the EMH, which brings us to our initial statement; trading is futile. As trading basically consists in predicting an asset performance, under the EMH, trading is a wasted effort. It is quite hard to understand how, if the EMH holds, financial institutions have poured significant amounts of money into trying to outperform the market. Even less understandable is how some have succeeded. I think it is reasonable to define what 'outperforming the market' means. 'Outperforming the market' encompasses the achievement of greater returns than a portfolio containing every asset available on the market, with the same level of risk offered by such portfolio. Alternatively, it could mean achieving the same return, with less risk. Eugene Fama states that the only way to expect returns greater than the market is to take additional risks, or that you cannot achieve that same return on investment with a risk inferior to the market risk. We are now able to ask the following question: does the EMH holds with real life data ? Is trading futile ? Can you achieve greater returns than the market for a corresponding market risk ? Or is one constantly better off holding a large index like the S&P 500 for instance ? In my opinion, questioning the EMH is the most riveting topic in quantitative finance as it answers questions from a very theoretical point of view, typically what is observed in academics but also from a very practical point of view, the kind we observe on a trading floor. Econometrics is the science of combining these two aspects, using theoretical tools to shine lights on very down to earth problems. Therefore this work will try to reflect both dimensions. There are many ways to prove or disprove the EMH, but consistently beating the market is one of them (consistency is essential here). The objective here is to use data science to try and outperform the market. The results should definitely not be taken as an attempt to generate profits but rather as an exercise where the objective is to make use of quantitative techniques to elucidate a question. Machine Learning is a very powerful tool when a problem comes to identifying patterns and regularities in the data. As we will discuss later on, it is particularly adapted in our case. One fact often played down is that Machine Learning boils down to a list of algorithms applied in a particular order, and it is very easy to make mistakes when using automated techniques. We hope to have sufficient knowledge to question our own results, especially when working in quantitative finance. Although we are not, as mentioned above, attempting to concretely apply our findings, We think it is only natural to add some perspective when working with financial data. Finance is connected to the economy by many pipelines and misleading Artificial Intelligence can induce disastrous consequences. Artificial Intelligence is a very broad term and Machine Learning is one of its component, the later relies on the idea of being able to find patterns without explicitly defining these patterns, whereas Artificial Intelligence relies on the idea of writing a program capable of the same thought process as a human. They can sometimes be interchangeable and the frontier between the two is often blurry, here we are focusing on learning patterns from data, hence we will almost exclusively refer to Machine Learning as opposed to Artificial intelligencve. We will also refer to Statistical Learning, which could be seen as the gap between traditional statistical

[1]Efficient Capital Markets: A Review of Theory and Empirical Work - E.Fama

models and Machine Learning models. Statistical Learning solves data related problems by formulating them in a formal statistical problem which will then be solved using parts of Machine Learning algorithms. This gives us a large panel of tools to extract useful information from the data, coerce this information into a functional form that can then be exploited to infer the behavior of unseen data. We will first gaze over the Statistical and Machine Learning methods we will be using ranging from penalized regression to neural networks, we will then detail the financial side of the problem, going over technical analysis, fundamental analysis and the necessary financial mathematics. The second part is a step-by-step journal of the progress of the overall project and what modifications have been made to improve the data processing[2].

This project is coded in Python 3.8 and is associated with a Jupyter Notebook, available on GitHub. The integrity of the code is open source.

I often compare open source to science. To where science took this whole notion of developing ideas in the open and improving on other peoples' ideas and making it into what science is today and the incredible advances that we have had. - Linus Torvalds, developer of Linux and Git

2 Theory

2.1 Statistics & Machine Learning

2.2 Finance

3 Application

4 Conclusion

[2] Although data processing often refers to the pre-screening of the data before feeding it into a model, here we refer to the entire process from the data gathering to the diagnosis of the results