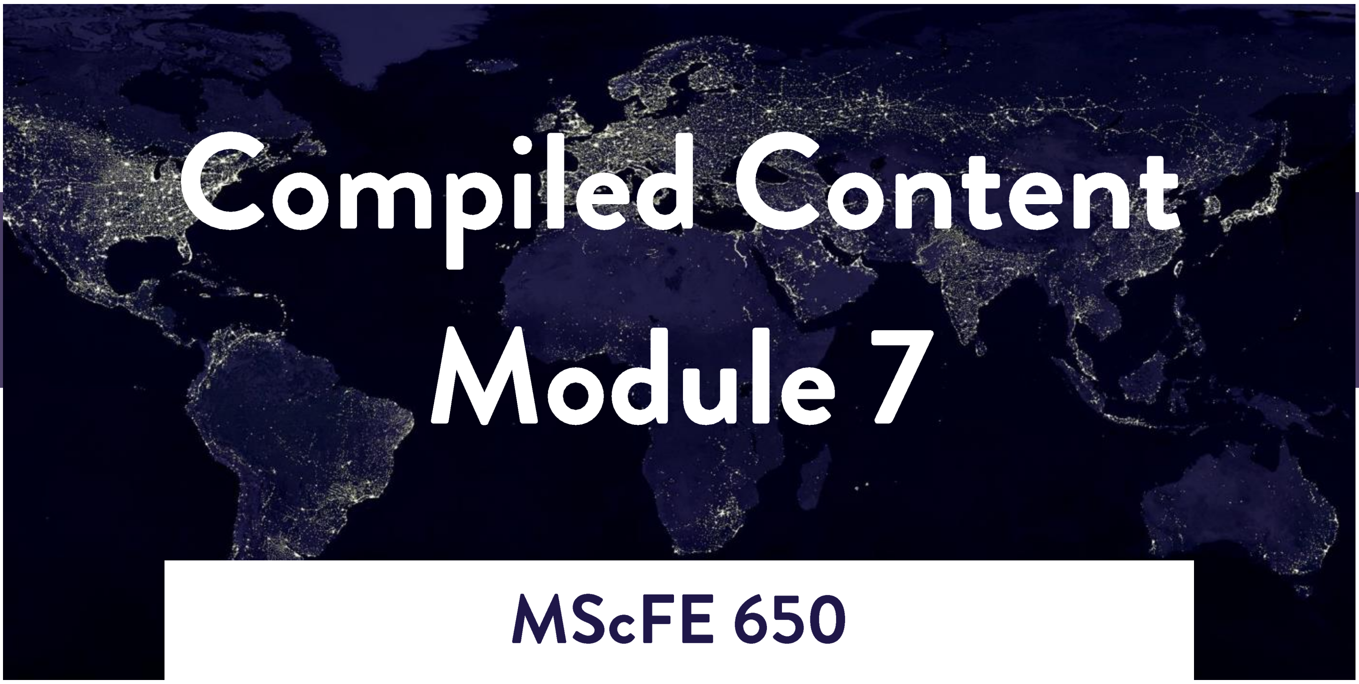
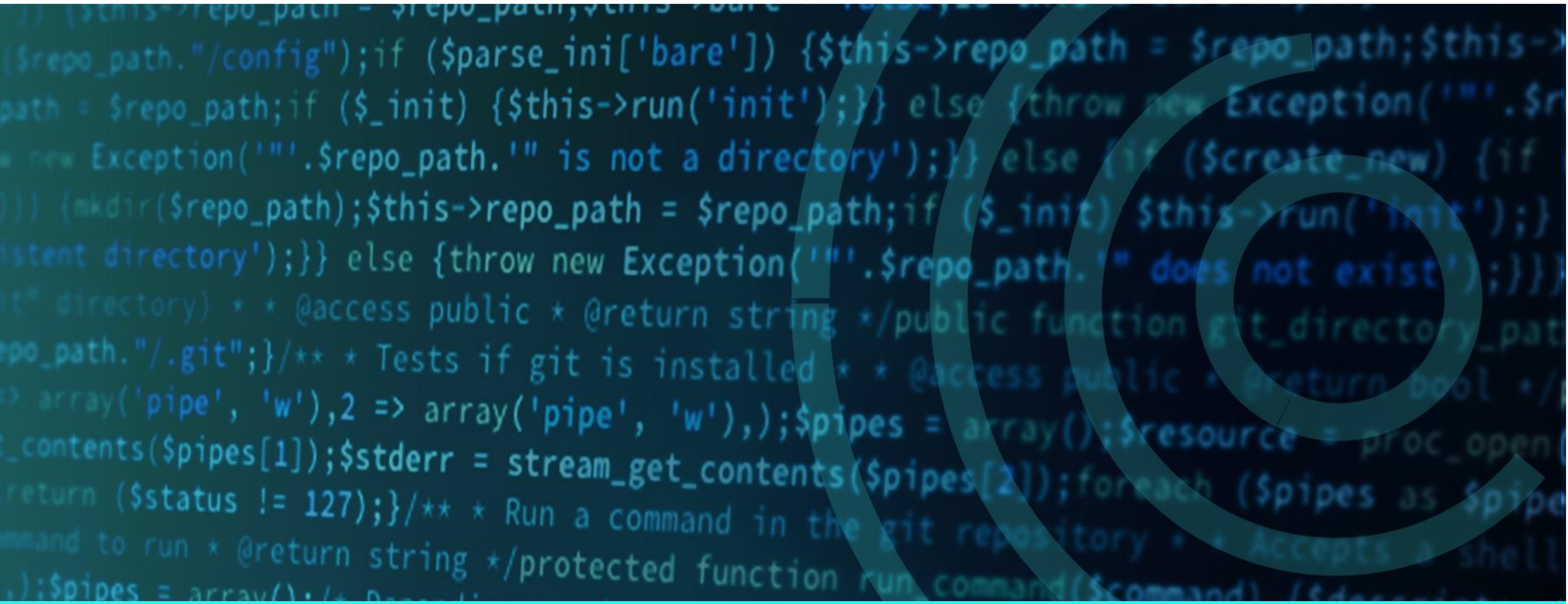


Compiled Content Module 7

MScFE 650

Machine Learning in Finance



Revised: 09/07/2019

Table of Contents

Module 7: Machine Learning for Finance

In this module students are exposed to the top reasons most machine learning funds fail. We introduce new financial data structures that have their roots in high-frequency trading, and we end the module by investigating various labeling techniques − including meta-labeling, which is used to boost the performance metrics of a primary model.

Unit 1: Introduction

Welcome to the final module of this course, where we will be discussing some of the latest advances in financial machine learning.

In this module, we will lightly cover some of the material in Dr Marcos Lopéz de Prado's book, Advances in Financial Machine Learning. The video lectures are made of Dr Lopéz de Prado's keynote lecture, from the 2018 Bloomberg Quant seminar series, on why most machine learning funds fail.

To complement the guest lectures, the three sets of notes in this module also cover, on a highly practical level, some of the latest advances in financial machine learning. We begin by introducing a data science workflow and solutions document; then we move on to cover new sampling techniques for creating financial data structures that have their roots in high-frequency trading (HFT). Finally, we address labeling techniques, as well as how to boost performance metrics using meta-labeling.

Unit 2: Workflows and Solution Documents

The first set of notes consist of two outlines, both inspired by interviews from top Kagglers: the first is a description of a workflow for data science projects; the second, of another very valuable tool – namely, a solutions document.

What a typical workflow looks like

Following much of the "agile" philosophy of project management, we recommend the following approach: fail fast, iterate, and pivot. This allows you to adapt your techniques to a given situation in real time. Click here for a lecture by Atlassian which introduces agile and some of its tools.

Now, outside of the typical structure of agile project management – which includes notions such as sprint planning, daily stand ups, and sprint reviews – the following is the typical structure we recommend you follow when tackling a new project:

1. Set up and update the solutions document.
2. Understand the problem you are trying to solve in more detail.
3. Do a quick data exploration to get a better understanding of the features.
4. Read academic papers and solutions on similar problems. (Try to look for some of the latest research to make sure you're not missing out on any new developments.)
5. Analyze the data and split it into cross validation sets.
6. Data pre-processing, feature engineering, and model training. (Don't spend too much time worrying about feature engineering. Train a model, check its performance and then iterate – you can work on feature engineering along the way.)
7. Performance analysis:
   a. Distributions,
   b. Performance metrics,
   c. Analysis on examples where your model failed.
8. Improve the model or create a new model.
9. Consider ensembles.
10. Repeat.

A solution document

When starting out on a new project, the first thing one should do is set up a solution document which we follow and update as time goes on. This concept has come up in many interviews with top Kagglers, and in particular, Shubin Dai:

What is your first plan of action when working on a new competition?

“Within the first week of a competition launch, I create a solution document which I follow and update as the competition continues on. To do so, I must first try to get an understanding of the data and the challenge at hand, then research similar Kaggle competitions and all related papers.”

What do you consider your most creative trick/find/approach?

“I think it is to prepare the solution document in the very beginning. I force myself to make a list that includes the challenges we faced, the solutions and papers I should read, possible risks, possible CV strategies, possible data augmentations, and the way to add model diversities. And, I keep updating the document. Fortunately, most of these documents turned out to be winning solutions I provided to the competition hosts.”

An example of a solutions document has been provided to help you with your group work project. It is important to remember that is just an example structure, and we encourage students to experiment and find a format that best suits their needs. We used a Word document, but others may find that a Jupyter Notebook, or a ReadMe.md file works better for these purposes.

One of the positive aspects of creating a solutions document is that it forms part of your solutions library, a source that can provide lasting inspiration and solutions to complex problems that you have already solved.

One such solutions library was created by a Kaggler called SRK, and can be found here. Note that this resource covers everything from financial machine learning to high-energy physics particle-tracking in CERN detectors.

Unit 3: Financial Data Structures

In this section, we will explore better ways of preparing your data for financial modeling. As you should know by now, there are different forms of useful financial data. The two most commonly used are:

1 Price data

This data is sampled using a fixed time interval, such as hourly, daily, weekly. It usually comes in the format of open, high, low, close, and sometimes volume. Below is an example.

Date	**Open**	**High**	**Low**	**Close**	**Volume**
1950-01-03	16.66	16.66	16.66	16.66	1260000.0
1950-01-04	16.85	16.85	16.85	16.85	1890000.0
1950-01-05	16.93	16.93	16.93	16.93	2550000.0
1950-01-06	16.98	16.98	16.98	16.98	2010000.0

Table 1: Standard price data

2 Fundamental data

This data primarily consists of financial accounting data. Typical examples are: book value, PE ratios, and market capitalization. We see a lot of this data used in traditional financial modeling – for example, the Fama French 3 Factor Model.

All fund managers hire data vendors like Bloomberg or Thomson Reuters to provide these sources of data; but, by making use of the same data as everyone else, you would only find what most other funds have already found. Now, it sounds very obvious, but often the question is then asked: what data should we use?

Interestingly, we can exploit and find new anomalies when we use the same data, albeit structured in a way that lends itself better to machine learning. It is for this purpose that students will learn to construct new structured datasets from raw unstructured tick data. An example of a raw set can be found below.

Symbol	**Date**	**Time**	**Price**	**Volume**	**Market Flag**	**Sales Condition**	**Exclude Record Flag**	**Unfiltered Price**
ESU13	09/01/2013	17:00:00.083	1640.25	8	E	0	NaN	1640.25
ESU13	09/01/2013	17:00:00.083	1640.25	1	E	0	NaN	1640.25
ESU13	09/01/2013	17:00:00.083	1640.25	2	E	0	NaN	1640.25
ESU13	09/01/2013	17:00:00.083	1640.25	1	E	0	NaN	1640.25
ESU13	09/01/2013	17:00:00.083	1640.25	1	E	0	NaN	1640.25

Table 2: Raw unstructured tick data

Standard bars

As mentioned, the standard way to sample price data is to take a sample using a fixed time interval − for example, once every hour, or at the end of each day, and then grouped and aggregated to determine the open, high, low, and closing prices.

Fixed time interval sampling should be avoided for 2 reasons:

1 Markets don't process information at fixed time intervals. Time bars oversample during quiet periods such as noon, and undersample during busy periods like the open and close. You will find that many large asset management companies try to make use of the open and closing auctions to execute large orders in an attempt to reduce market impact. Those interested can read up on algorithms for volume-weighted average price (VWAP) and time-weighted average price (TWAP).

2 Time-sampled bars often exhibit poor statistical properties:

   a. Serial Correlation
   b. Heteroscedasticity
   c. Non-normality of Returns

To get started, we will need raw tick data from which we can create three new types of bars, which have much better statistical properties.

- Tick bars
- Volume bars
- Dollar bars

Tick bars

An alternative is to sample once a pre-defined number of transactions have taken place – for example, every 10,000 ticks. This technique acts as a proxy for when information arrives to market. The idea is that the smart money places orders when new information becomes available. It is the first method under discussion that resolves the problem of inefficient sampling.

The following quote highlights the idea of sampling using a proxy for market information, although this particular quote is in reference to volume bars:

> "The idea that the time between trades was correlated with the existence of new information, providing our basis for looking at trade time instead of clock time. It seems reasonable that the more relevant a piece of news is, the more volume it attracts. By drawing a sample every occasion the market exchanges a constant amount of volume, we attempt to mimic the arrival to the market of news of comparable relevance. If a particular piece of news generates twice as much volume as another piece of news, we will draw twice as many observations, thus doubling its weight in the sample." (Easley, Lopéz de Prado & O'Hara, 2012a)

Mandlebrot & Taylor (1967) pointed out that sampling as a function of trading activity, i.e. the number of transactions, resulted in more desirable statistical properties. Multiple studies have confirmed this, and Ane & Geman (2000) showed that it results in returns closer to IID Normal. This is important as many statistical models rely on the assumption that observations are drawn from an IID Gaussian process.

Volume bars

The next option is to sample once a pre-defined volume (units of shares) have been executed – for example, every 50,000 shares traded. This technique offers even better statistical properties than sampling using ticks, confirmed by Clark (1973). Another reason to favor this technique over tick bars is that there are several market microstructure theories that study the relationship between price and volume.

Easley et al. (2012a) show that by using volume bars, the data has less heteroscedasticity:

> "We show how volume bucketing (time intervals selected so that each has an equal volume of trade) reduces the impact of volatility clustering in the sample. Because large price moves are associated with large volumes, sampling by volume is a proxy for sampling by volatility. The resulting time series of observations follows a distribution that is closer to normal and is less heteroscedastic than it would be if it were sampled uniformly in clock time."

In the paper titled "The Volume Clock" (Easley et al., 2012b), the authors display standardized distributions of fixed time intervals and volume bars. Figure 1 (below) is an example.

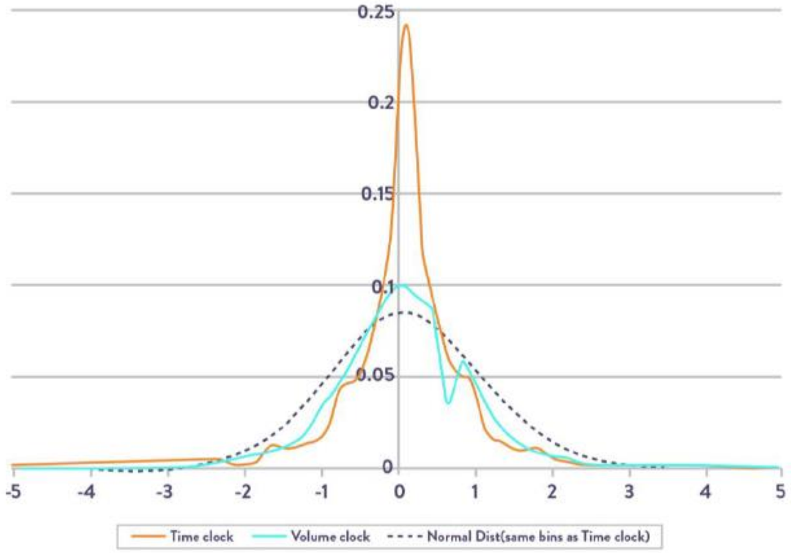


Figure 1: Partial recovery of normality through a price sampling process subordinated to a volume clock

(Adapted from Easley et al., 2012b)

Notice how the volume bars have a lower kurtosis. In the group project, students will be required to reconstruct this plot, and check the test statistics from the Jarque-Bera normality tests. Groups should find that the returns generated by sampling using volume bars are much closer to the normal distribution than fixed time interval bars.

Dollar bars

With dollar bars, one samples once a predetermined market value has taken place. It doesn't have to be dollars – it should be the local currency in which the instrument is priced (for example every 100,000 dollars or euros).

This technique has the advantage of not being susceptible to corporate actions such as stock splits or consolidations. In the case of a 2:1 stock split, the volume bars will suffer where the dollar bars will be unaffected.

Dollar bars are the most robust and stable of all the bars. They capture the information of the volume bars and line up with market microstructure theories as they capture the interaction between price and volume.

Further reading

For the remainder of this section, students are required to read the following three papers, ideally in the order in which they are listed. Please note that, during the final week of the course, they will form the basis of the multiple-choice quiz assessments.

1. Easley, D., Lopez de Prado, M. and O'Hara, M. (2012b). 'The Volume Clock: Insights into the High Frequency Paradigm. The Journal of Portfolio Management, 9.
2. Mandelbrot, B. and Taylor, H.M. (1967). 'On the Distribution of Stock Price Differences'. Operations Research, 15(6), pp.1057-1062.
3. Easley, D., López de Prado, M.M. and O'Hara, M. (2012). 'Flow Toxicity and Liquidity in a High-frequency World'. The Review of Financial Studies, 25(5), pp.1457-1493.

For those who wish to read further, here are another three recommendations:

1. Ané, T. and Geman, H. (2000). 'Order Flow, Transaction Clock, and Normality of Asset Returns'. The Journal of Finance, 55(5), pp.2259-2284.
2. Clark, P.K. (1973). 'A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices'. Econometrica: Journal of the Econometric Society, pp.135-155.

3. Easley, D. and O'Hara, M. (1992). 'Time and the Process of Security Price Adjustment'. The Journal of Finance, 47(2), pp.577-605.

An implementation of the various techniques can be found here:

1. Financial Data Structures
2. Advances in Financial Machine Learning

Unit 4: Labeling Techniques

Traditional labeling

The majority of the papers regarding machine learning in alpha design will focus on using classification rather than regression techniques. You will find that the typical paper will attempt to forecast the next period's direction move as either up or down.

For example, given the time series of Apple stock, one might attempt to predict if the next day's move (returns) will be positive or negative. In this setting, the target labels are {0, 1}, a binary classifier. However, you may find it more useful to add some kind of threshold level – for example, the 90-day rolling standard deviation. In this setting, you will label: a move above the threshold as a 1; a move below the negative of the threshold a -1; and if neither, then 0. This results in a {-1, 0, 1} set.

Remember that because you are using categorical variables for your target variable, you need to one-hot encode them and make use of a cost function such as categorical cross entropy, rather than mean-squared error.

Triple-barrier method

You are encouraged to read up on the triple-barrier method and, if possible, apply it in the group project. It is also briefly touched on in this module's lectures.

For an implementation of the triple-barrier method, students can find code on this Github repo, and a further description on this blog post by Quantopian.

Meta-labeling

Meta-labeling is a technique used to boost the performance metrics of binary classifiers. In a financial setting, it allows us to determine position sizing of a trade – for example, a primary model, technical trading strategy, or analyst rating, provides the position of the trade (i.e. long or short). Then, the secondary model uses meta-labeling to determine if the first model is correct or not.

This allows the primary model to set the side of the trade, and the secondary model outputs a value between 0 and 1 which acts as a confidence weighting on the first model. We can then map the output from the second model to different position sizes, and this allows us to weight more heavily the position in which we are confident – leading to greater profits.

A thorough explanation of meta-labeling can be found in the accompanying Jupyter Notebook.

Bibliography

References

Ane, T. and Geman, H. (2000). 'Order Flow, Transaction Clock, and Normality of Asset Returns'. *The Journal of Finance*, 55(5), pp. 2259-2284.

Easley, D., López de Prado, M. and O'Hara, M. (2012a). 'Flow Toxicity and Liquidity in a High-frequency World'. *The Review of Financial Studie*s, 25(5), pp.1457-1493.

Easley, D., Lopez de Prado, M. and O'Hara, M. (2012b). 'The Volume Clock: Insights into the High Frequency Paradigm'. *The Journal of Portfolio Management*, 9.

Lopéz de Prado, M. (2018). *Advances in Financial Machine Learning.* Wiley Publishers.

Mandelbrot, B. and Taylor, H.M. (1967). 'On the Distribution of Stock Price Differences'. *Operations Research*, 15(6), pp.1057-1062.