Capstone Proposal
Advances in Financial Machine Learning

Ashutosh. Singh, Jacques. Joubert

Abstract

The goal of our Capstone projects is to create a foundation of code on which various strategies can be built and tested. We derive our inspiration from the book Advances in Financial Machine Learning by Dr Marcos Lopez de Prado. The proposed project will deliver a open source package and implementation of the strategies from which the research community can build upon.

Keywords: Machine Learning, Finance, Alpha Design, Data Structures, Labeling, Meta labels

---

1. Introduction

Inspired by 2019 Quant of the Year Dr Marcos Lopez de Prado we propose an implementation and further research into the novel ideas published in his book Advances in Financial Machine Learning. Our project is split over two capstone sessions, the first 6 weeks sets the foundation by creating a published and open source python package which will enable further research into the field and an implementation of a trading strategy which illustrates the benefits of the techniques used. The second 16 weeks will focus on further implementation of de Prado's work as well as seeking a novel contribution to the literature.

The key contributions of part one will be the following:

1. An open source python package
2. Transformed data sets to promote further research
3. Prove by way of example that meta labeling works in improving the performance metrics of a primary model

This proposal only discusses the first 6 weeks and the rest is split up as follows: section 2 data and sources, section 3 methodology, section 4 deliverables, section 5 research beyond capstone, and section 6 additional benefits to WorldQuant University.

2. Data

We will source high quality tick data from Tick Data LLC at the cost of approximately 1,000 USD. The focus will be on SP 500 E mini futures, for the period 2008 - 2011. This set will encompass several market regimes including the global financial crises as well as the 2010 flash crash. The SP 500 E Mini futures data is the set which de Prado regularly references in his work and by using the same set we create a natural way to benchmark our implementations.

3. Methodology

3.1. Financial Data Structures

We will work with raw tick data and transform the unstructured data and form a structured data set amenable to machine learning algorithms. It has been shown that sampling data by means of fixed time intervals, e.g. end of day or hourly, exhibit poor statistical properties..

We will code up an implementation of the standard bars:

1. Time Bars
2. Tick Bars
3. Volume Bars
4. Dollar Bars

We will then write a small paper on the different bars and the various statistical properties they exhibit, as well as why these financial data structures are more amenable to machine learning algorithms.

At the time of writing, to the best of our knowledge there isnt a python package the implements these various financial data structures and also there are no sources that share data of this kind. Our two key contributions will be a python package the implements these various financial data structures along with a test data set.

3.2. Labeling Techniques

In this section we will discuss some of the ways to label financial data structures.

We introduce the typical fixed-time horizon method and explain why classification dominates the financial literature rather than regression.

Next we introduce the following with implementations:

3.2.1. The Triple Barrier Labeling method

This technique incorporates the idea that stock prices follow a Geometric Brownian Motion, random walk, described by a mean and variance. It very naturally fits into the framework of derivatives pricing and allows us to model a stock as path dependant time series.

3.2.2. Meta Labeling

Meta labeling can be used to improve the performance metrics of a binary classification model. The main idea is that a secondary model is used to determine if a primary model is correct or not. Meta labeling is a new technique that leaves lots of room for us to make a novel contribution to the literature.

3.2.3. Learning Direction and Size

Meta labeling enables us to not only learn the next direction but also the size of the bet. In the first implementation we will only use it to determine if we should trade or not, based on a given direction.

3.3. Quantitative Strategy

After implementing the principles discussed in this paper, we develop a trading strategy that would use the data and techniques discussed in section 3.1 and 3.2. The main idea is that we want to train a model that uses machine learning to determine not only the direction of a next move but also whether to trade or not (this can later be extend to optimal bet sizes). Meta labeling enables this but also allows us to strap on machine learning to other strategies such as discretionary fund managers and technical trading strategies.

The model we propose combines components of trend following as well as mean reversion. The features that describe these regimes as well as additional features are passed through to the model.

One of the key contributions of this project will be to show that meta labeling works in improving the performance metrics of a primary model.

A draft outline inspired by chapter 3 of the book follows:

1. Create structured data set from the raw tick data
   (a) Drop unnecessary labels
   (b) Create labels using the Triple Barrier method
2. Develop a trend following strategy based on a popular technical analysis rule (such as moving average crossover)
   (a) Derive Meta Labels
   (b) Train a SVM, Random Forest, and Neural Network to determine to trade or not

3. Develop a mean-reverting strategy based on bollinger bands to suggest a direction of the next move.
   (a) Derive Meta Labels
   (b) Add additional engineered features
   (c) Train a SVM, Random Forest, and Neural Network to determine to trade or not.
   (d) Discuss the performance metrics of the primary model if the secondary model doesnt filter the bets. Show standard classification metrics.
   (e) Now show the performance of the metrics with the additional layer of a secondary model.

The second part of the Capstone project (Jacques) would extend this architecture to include bet sizing, sample weights, and additional model inputs such as structural breaks, market microstructure, and entropy features. Our design would enable us to test new ideas and strategies post Capstone projects.

4. Deliverables

The following are a list of deliverables for the project:

1. Academic style paper written in Latex, detailing our techniques and results
2. Publish our Python package with the Python Package Index (PyPi)
   (a) This will enable future research within the field, as a package doesnt currently exist.
   (b) We will also open source the project and create a community via Github repo
3. All code hosted on Github to promote further research

5. Research beyond Capstone 1 (Ashutosh) and Capstone 2 (Jacques)

We think the next steps with our work could entail exploring newer data sampling methods such as information-driven bars (below) and other machine learning strategies involving regime change detection

Information-driven bars:

1. Tick Imbalance Bars
2. Volume / Dollar Imbalance Bars
3. Tick Runs Bars
4. Volume / Dollar Runs Bars

6. Additional Benefits for WQU

Jacques is the course designer and lecturer for the machine learning in finance module taught at WQU. Any additional packages, data transformations, and example projects can be added to that module. In particular, the transformed data that can be shared with other students to encourage similar research.