1. Financial Machine Learning as a Distinct Subject

2. Financial Data Structures

2.1. Motivation

2.2. Essential Types of Financial Data

1. **Fundamental Data:**
    - Stuff that gets published in Financial reports
    - need to be careful abou the information that the market knows at the time when training models
    - Revisions and delayed releases can change the data the market has, so it important to account when giving a model data.
    - the 2nd quarter results are published until a few months after the 2nd quarter ends.
2. **Market Data:**
    - Highly unstructured,
    - FIX Messages, BWIC (bids wanted in competition) responses.
    - able to examine how others in the market interact
3. **Analytics:**
    - ”derivated data”
    - data based on an orginal source, that is simplified
    - CONS: $$ and your not the only one with the insights that have been brought to the surface.
4. **Alternative Data:**
    - social media, news, web searches, satellites, geo-location, etc.
    - ’primary information’ - before the crash you can see the tanker get stuck in the ocean.
    - hopefully truly unique hard to process, makes you have novel information from the rest of the market

2.3. Bars

Bars are basically the rows of a table with you financial data. see types below.

2.3.1. Standard Bars

Time Bars

- Time Bars are obtained by sampling the market at tixed time intervals
- CONS: markets are not constant time (the trading has more and less dense points in the day)
- CONS: time series have nasty statistical properties, which require nasty tools to deal with (like GARCH), but some the other tools have better propertiesd

Tick Bars

- sample the market every fixed number of transactions (ie $1,000$ ticks)
- There are outlyers like auction in which 1 tick will have many trades, and that could mess stuff up

Volume Bars

- The problem with tick bars, is that 1 order for 10 stocks will have less of an impact then 10 orders for 1 stock and often times software will break down larger order for operational convience so it isn't a good measure of how fast information arrives.
- so we use the volume bar after x of number of stocks are traded.
- better statistical properties (close to IID Gaussian)

Dollar Bars

- the dollar bar is after a certain dollar amount of value gets traded.
- this is useful because in the previous strategy if the prices changes a lot it may be that $1,000$ or a stock can go from 0.25 stocks to 5 stocks depending on the price.

2.3.2. Information Driven Bars

These "Information Driven Bars" try to sample more frequently as more well informed

Tick Imbalance Bars

- We are trying to find tick bars, meaning how much time should each sample represent.
- We are looking at the time from $t = 0$ until $t = T^*$
- We can T^* is the first time that θ_t exceeds $E[\theta_t|F_0]$, where F_0 represents the information known at $t = 0$.

  So what is θ_t?

  $$b_i = \left\{ \begin{array}{cc} +1, & \text{if price increased} \\ -1, & \text{if price deceased} \\ b_{i-1} & \text{if price is the same} \end{array} \right\}$$

  (the b_i are sampled at regular time intervals) and θ_t is the sum of the b_is.
- In order to calculate $E[\theta_t|F_0]$: this turns into a level crossing problem.
- Insert **very** hand-wavy math: $E[\theta_t|F_0] \approx E[T]\dot{E}[b_i]$
  $E[b_i] = (+1 \cdot P[b = +1] - 1 \cdot P[b = -1]) = P[b = +1] - (1 - P[b = +1]) = 2P[b = +1] - 1$
- $E[T]$ is the expected value of the tick size (calculated via exponential weighted moving averages of the previous tick sizes)
- So the threshold is $|E[\theta_t]| = |E[T]||2P[b = 1] - 1|$

Volume/Dollar Imbalance Bars

- An equivalent idea to the Tick Imbalance Bars, but using either volume or dollars to measure the imbalance instead of transactions.
- this is done by having $\theta_t = \sum b_t v_t$, where v_t is either the volume or dollar amount traded.
- this propagates through the math, until get a new threshold $E[\theta_t|F_0] = E[T|F_0] \max\{P(b_t = 1), 1 - P(b_t = 1)\}$
- this solves the issue of corporate actions (like stock splits, etc) which can effect expected value of ticks, but not true dollar amounts.

Tick Runs Bars

- Used to monitor *sequence* of buys in the overall volume, and take samples when that sequence diverges from our expectation.
- To accomplish we define $theta_t$ as:
$$\theta_t = \max\{\sum_{t|b_t=1}^{T} b_t, -\sum_{t|b_t=-1}^{T} b_t\}$$
- the threshold is defined in the same way as $E[\theta_t|F_0]$ which turns into
$$E[\theta_t|F_0] = E[T|F_0] \cdot \max\{P[b_t = 1], 1 - P[b_t = -1]\}$$

Volume/Dollars Bars

- Extends the above to dollars/volume, we want to sample whenever the volumes or dollars traded by one side exceed our expectation for a bar.
- We extend the definition of θ_t
$$\theta_t = \max\{\sum_{t|b_t=1}^{T} b_t v_t, -\sum_{t|b_t=-1}^{T} b_t v_t\}$$
where v_t is either volume or dollars
- the cutoff $E[\theta_t|F_0]$ is:
$$E[\theta_t|F_0] = E_0[T]\max\{P[b_t = 1]E_0[v_t|b_t = 1], (1 - P[b_t = 1])E_0[v_t|b_t = -1]\}$$

2.4. Dealing With Multi-Product Series

Its a pain to deal with futures contract, and dividends and all the other stuff that isn't just straight forward changes in price. He introduced the "ETF Trick", which simplifies.

2.4.1. ETF Trick

Let's say we wanted to develop a strategies that trades a spread of futures. Our data (any of the bars which we described before), has the following properties:

1. $o_{i,t}$ Open Price of item $i_0, i_1 \ldots I$, and bar $t_0, t_1 \ldots, T$
2. $p_{i,t}$ Close Price of item $i_0, i_1 \ldots I$, and bar $t_0, t_1 \ldots, T$
3. $\phi_{i,t}$, the USD value of one point instrument of item $i_0, i_1 \ldots I$, and bar $t_0, t_1 \ldots, T$
4. $v_{i,t}$, the volume of the instrument $i_0, i_1 \ldots I$, and bar $t_0, t_1 \ldots, T$
5. $d_{i,t}$ is the carry, dividend, or coupon paid by the security. This could also be used to charge operating costs.

Even if everything wasn't tradable during the entire bar, it was at least tradable at the beginning and enad.