Financial Machine Learning practitioners have been using the wrong candlesticks: here's why

In this article we will explore why traditional time-based candlesticks are an inefficient method to aggregate price data, specially under two situations: (a) highly volatile markets such as cryptocurrencies and (b) when using algorithmic or automatic trading. To prove this point, we will analyze the behavior of the Bitcoin-USD historical price, we will look at why markets do not follow sunlight cycles anymore and why the type of data we use can be an advantage respect competitors. Finally, we will briefly introduce alternative and state-of-the-art price aggregation methods, such as volume or tick imbalance bars, that aim to mitigate the shortcomings of traditional candlesticks.

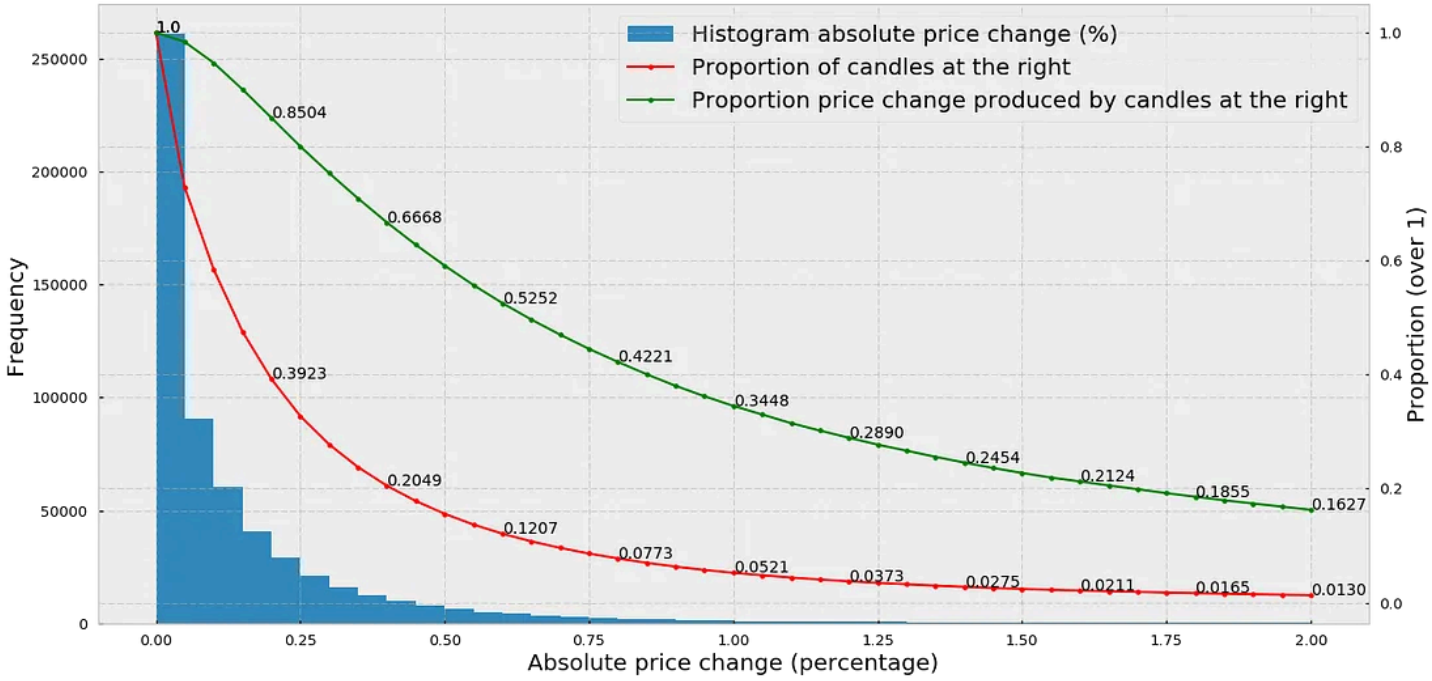
Gerard Martínez · Follow
Published in Towards Data Science
6 min read · Apr 21, 2019

Over-sampling and under-sampling in highly volatile markets

Cryptocurrency markets are extremely volatile. Prices change rapidly and is rather common to see the price moving side-ways for hours before being pumped or

dumped 5–20% in a matter of minutes. While long-term trading strategies may be still profitable while ignoring the intra-day volatility, any strategy in the middle or short term (not to mention high-frequency trading) will necessarily have to address the issue of volatility in one way or another.

In the following plot we analyze the volatility of the Bitcoin-US dollar pair since March 2013 until April 2019 for the Bitfinex exchange (data obtained from CryptoDatum.io) using 5-minute candlesticks. Specifically we show: (1) the histogram of absolute price changes (calculated as percentage of change of the close price respect the open price), (2) the proportion of candles at the right of the histogram from a specific point (red line) and (3) the proportion of total amount of price change produced by the candles at the right (green line).



Thus, we can observe that:

1. The majority of 5-minute candles (around 70%) experience price changes below 0.25%, with the largest part of them experiencing virtually no price change (first histogram peak at 0.00–0.05%).
2. 20% (0.2049) of the candlesticks explain almost 67% (0.6668) of the total amount of price change.
3. 2% of the candlesticks explain 21% of the total amount of the price change, indicating that in 2% of the BTC-USD candlesticks huge price variations occurred in the short time frame of 5 minutes (high volatility).

What we can conclude from point 1 is that *time-based candlesticks clearly over-sample periods of low activity* (activity understood as price change). In other words, in 70% of the candlesticks nothing is really going on so the question is: if we want to train a ML-based algorithm, do we need all these candles where no change is observed? Would finding a way to remove or discard most of the meaningless candles be useful to enrich our datasets?

Time-based candlesticks over-sample low activity periods and under-sample high activity periods

On the other hand, points 2 and 3 indicate that most of the price changes happen in a few percentage of candlesticks suggesting that *time-based candlesticks under-sample high activity periods*. What this means is that if price changes 10% in 5 minutes and we are using 5-minute candlesticks, our algorithm will not be able to see anything happening in between the opening and the closing of the time-based candlestick, potentially missing a good trading opportunity. Therefore, ideally we would like to find a way to sample more candles whenever market activity increases and sample less candles when the market activity decreases.

Markets probably do not follow human daylight cycles anymore

The main reason why we are using time-based candlesticks is because we humans live embedded in time and, therefore, time is something we find tremendously convenient to organize ourselves and to synchronize our biological rhythms with. Furthermore, the sunlight cycle is of utmost importance for humans because it determines the awake-asleep cycle, which is of biological relevance for our survival. As a consequence of the daylight cycle, traditional stock exchanges still open at 9:30 AM (so that actual humans can trade while they are awake) and close at 4 PM (so that traders can sleep in peace — but… do they?).

With the advent of technology, automatic trading bots have started to displace real human traders and, particularly in cryptocurrencies, markets no longer follow the daylight cycle since they remain open 24/7. In these circumstances, does it make sense to keep using time-based candlesticks, a mere standard, consequence of human convenience? López de Prado, summarizes it well in his book *Advances in Financial Machine Learning*:

> Although time bars are perhaps the most popular among practitioners and academics, they should be avoided. [...] Markets do not process information at a constant time interval. [...] As biological beings, it makes sense for humans to organize their day according to the sunlight cycle. But today's markets are operated by algorithms that trade with loose human supervision, for which CPU processing cycles are much more relevant than chronological intervals

Everyone's data is no one's advantage

The reason why most good trading algorithms are a well kept secret is because money is made when from an out-of-equilibrium situation we can anticipate going to another equilibrium. And, generally speaking, equilibrium means everyone is already aware of what's going on and there are enough positive and negative forces to keep the newly reached equilibrium in balance.

In order to anticipate a change in equilibrium we must be contrarian and correct at the same time: this is, we must know something that the rest doesn't know and be correct in our assertion. We must find out-of-equilibria that the majority of other traders are unaware of, otherwise we would be already in equilibrium again. It is often referred to as "zero-sum" game, although I don't particularly like this definition.

> Everyone's data is no one's advantage

In order to be contrarian, we must look at the data and analyze it in new creative ways that allow us to gain a certain advantage respect the others. Here's when small details such as the type of data we use to train our algorithms can make a big difference. In practice, this means that if everyone uses time-based candle-sticks, why would we use the same as everyone else? If a minoritarian — better — alternative may exist, why would we still use time-based candlesticks?

Alternative candlesticks

Few times I felt more enlightened than after reading the first chapters of de Prado's book *Advances in Financial Machine Learning*. In his book, this experienced fund manager reveals common practices and mathematical tools he has been using to manage multimillionaire funds for more than 20 years. Particularly, he knows from first hand that the behavior of markets has changed dramatically over the years and that competing with trading bots is rather the rule than the exception. In this context, de Prado describes a number of alternative types of candlesticks that aim to replace the traditional time-based candlesticks and that bring the necessary creativity and novelty to the financial space.

Here are some examples of alternative candlesticks or bars proposed by de Prado:

- **Tick bars**: we sample a bar every time a predefined number of transactions — i.e. trades — takes place. For instance, every time 200 trades take place in the exchange, we sample a OHLCV bar by calculating its Open-High-Low-Close-Volume values.
- **Volume bars**: we sample a bar every time a predefined volume is exchanged. For instance, we create a bar every time 10 Bitcoins are traded in an exchange.
- **Tick Imbalance bars**: We analyze how imbalanced is the sequence of trades and we sample a bar every time the imbalance exceeds our expectations.

In future posts I will analyze each of the alternative candlesticks described in the book. We will look at how to build them, we will compare their statistical properties, such as normality and serial correlation, with traditional candlesticks and we will analyze how exactly these bars overcome the shortcomings of traditional candlesticks.

Thanks for reading, more content is coming up so stay tuned!

Would you like to try the alternative bars we mentioned in the article by yourself? At CryptoDatum.io we strive to offer state-of-the-art cryptocurrency datasets to plug in your own ML-based trading algorithms. Check us out at https://cryptodatum.io .

The information provided is for educational purposes only. By no means it represents any financial advise and the information must be taken "as is" without guarantee of any kind.

CryptoDatum.io

Finance | Machine Learning | Bitcoin | Algorithmic Trading | Cryptocurrency

Follow

Written by Gerard Martínez

1.4K Followers · Writer for Towards Data Science

Trading strategy developer — Founder of CryptoDatum.io

More from Gerard Martínez and Towards Data Science

Gerard Martínez in Towards Data Science

Autoencoders for the compression of stock market data

A Pythonic exploration of diverse neural-network autoencoders to reduce the dimensionality of Bitcoin price time series

8 min read · Jan 18, 2019

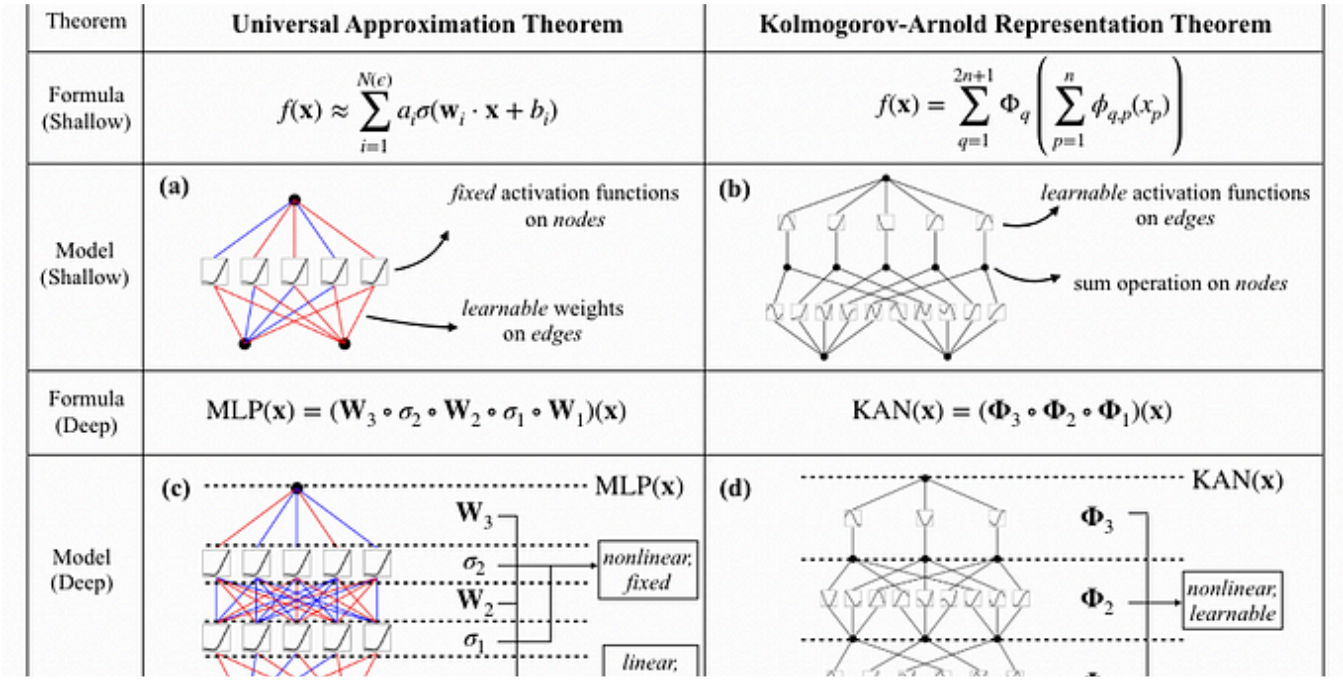
Torsten Walbaum in Towards Data Science

What 10 Years at Uber, Meta and Startups Taught Me About Data Analytics

Advice for Data Scientists and Managers

9 min read · May 30, 2024

5.3K 83



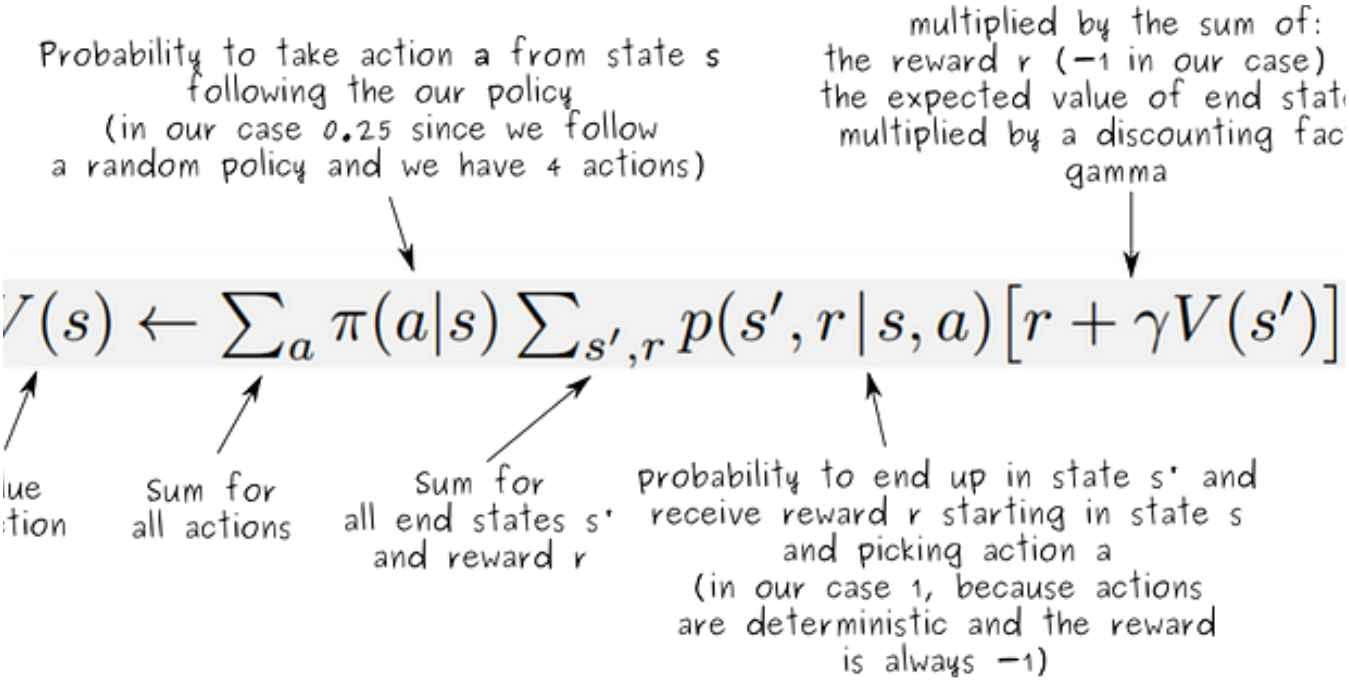
Theo Wolf in Towards Data Science

Kolmogorov-Arnold Networks: the latest advance in Neural Networks, simply explained

The new type of network that is making waves in the ML world.

· 9 min read · May 12, 2024

2.1K 18



Gerard Martínez in Towards Data Science

Reinforcement learning (RL) 101 with Python

Iterative policy evaluation and Monte Carlo simulations to solve the gridworld state-value function

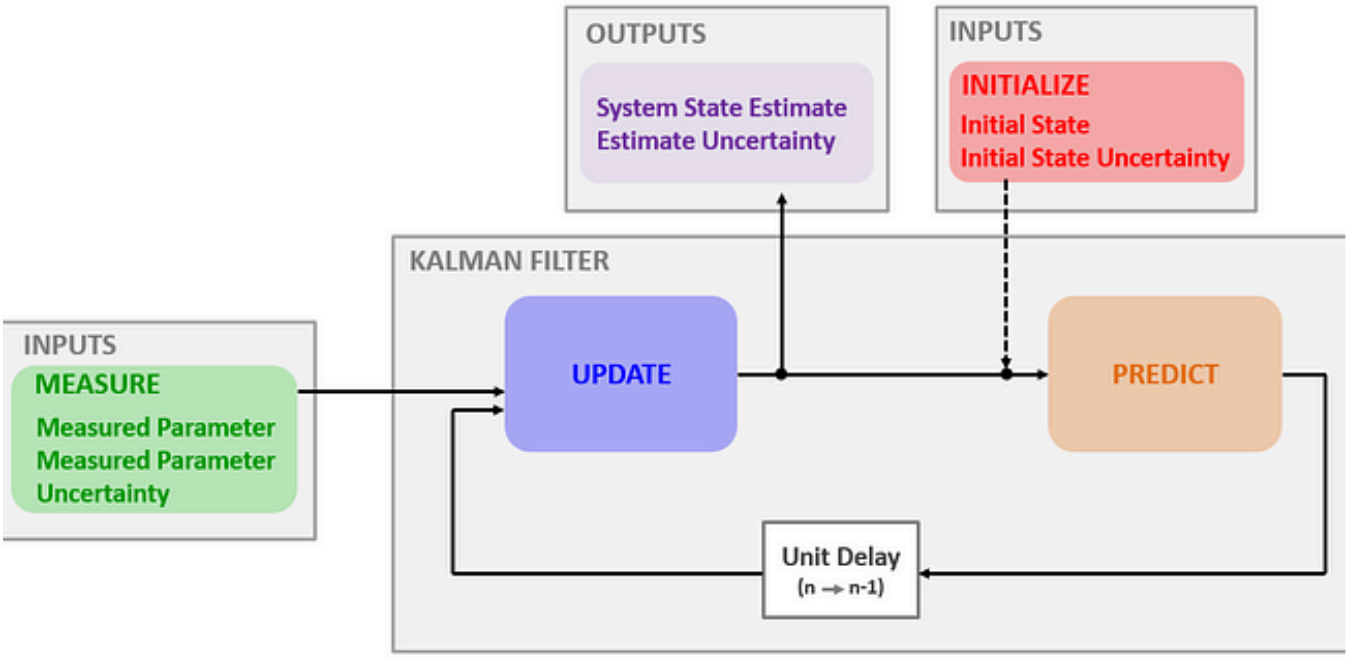
8 min read · Dec 20, 2018

716 7

See all from Gerard Martínez

See all from Towards Data Science

Recommended from Medium



Sausan

The Magic of Kalman Filters: How They Transform Data into Accurate Predictions

In the world of data analysis and signal processing, Kalman filters are nothing short of magical. These mathematical marvels can sift...

4 min read · Jun 15, 2024

Fisher Lok in InsiderFinance Wire

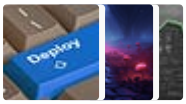
Trading from First Principle 3: Building a Machine Learning-Based Trading Strategy

This article proposed a standardized approach to produce a trading strategy based on machine learning models. A systematic approach is...

13 min read · Feb 13, 2024

177

Lists



Predictive Modeling w/ Python

20 stories · 1316 saves

Practical Guides to Machine Learning

10 stories · 1580 saves

Natural Language Processing

1537 stories · 1071 saves

The New Chatbots: ChatGPT, Bard, and Beyond

12 stories · 412 saves

Anastasija Draganic

Post-Event Analysis and ROI with AI in Event Management

Utilizing Advanced AI for In-Depth Insights and Accurate Financial Assessments

· 3 min read · Feb 15, 2024

Tanvi Lal

Tanvi's Take: Fintech in 2024

My 2 cents on what's coming up in fintech this year as well as the key areas I'll be digging into

7 min read · Jan 2, 2024

243

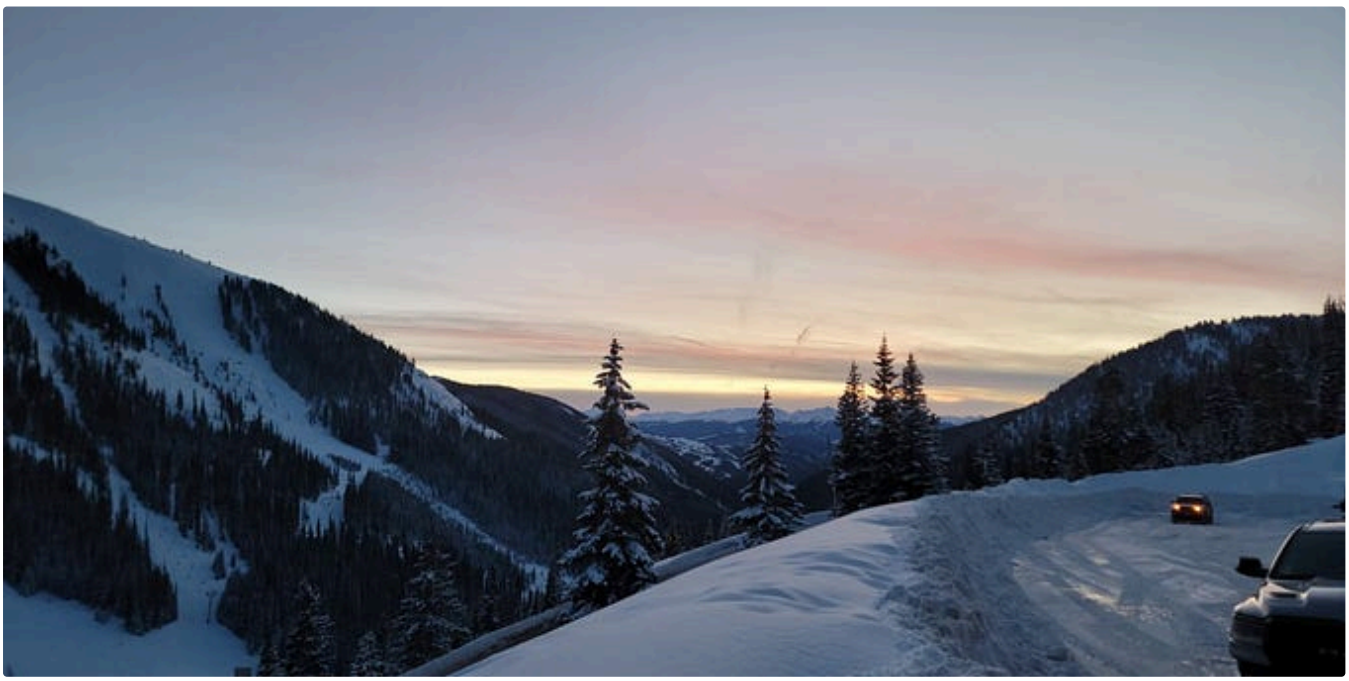
Adam in Call For Atlas

Unsupervised Learning as Signals for Pairs Trading and StatArb

In this article, we will leverage clustering algorithms to detect pairs in a universe of tradable securities. We will train three...

23 min read · Mar 6, 2024

202 2



Isaac Godfried in Deep Data Science

Advances in Deep Learning for Time Series Forecasting/Classification Winter 2024

Has time series forecasting had its GPT moment? Have transformers overcome their limitations on time series and do they out perform...

 · 18 min read · Jan 29, 2024

 522  4 

See more recommendations