Information-driven bars for financial machine learning: imbalance bars

Gerard Martínez · Follow
Published in Towards Data Science
8 min read · May 20, 2019

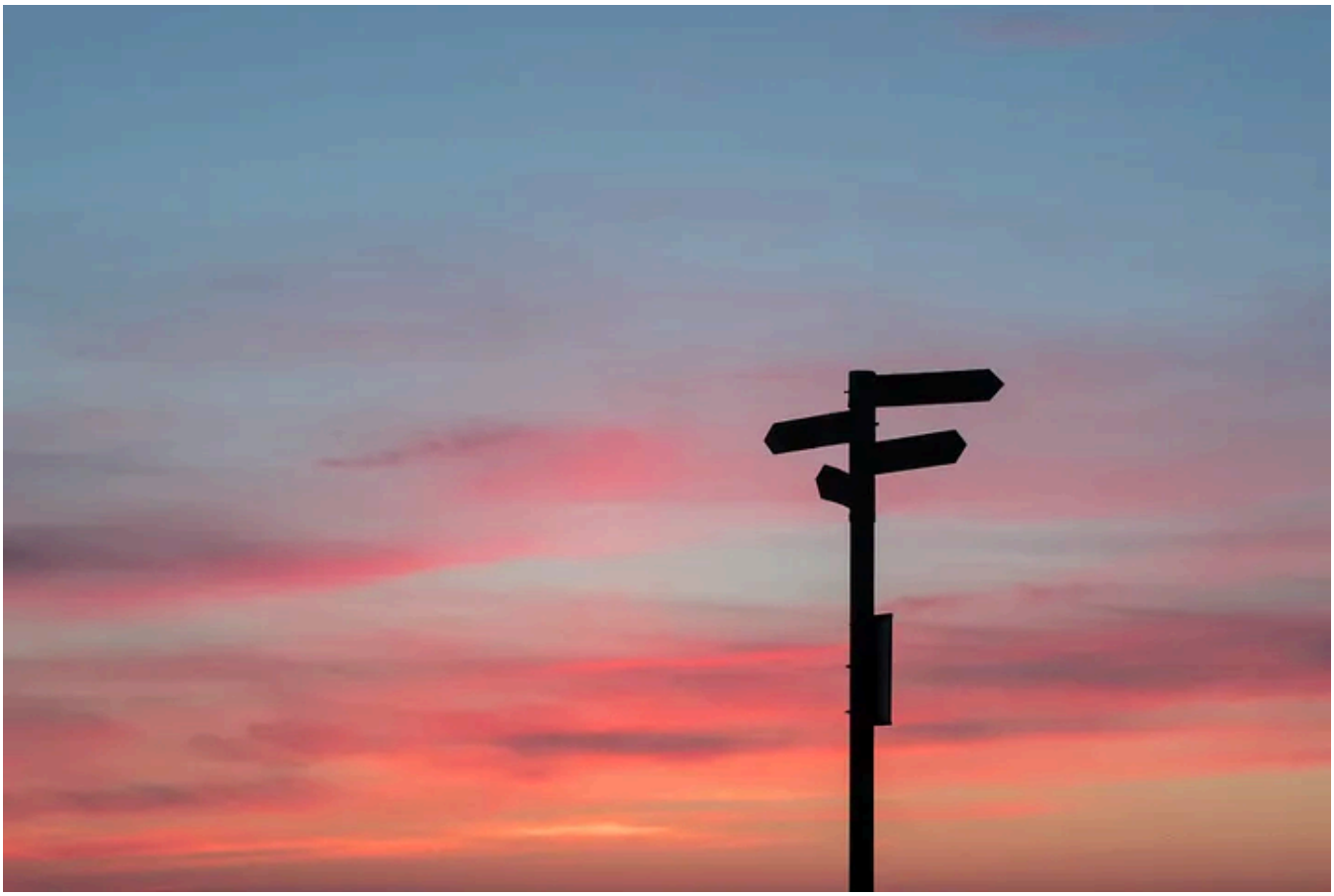


Photo by Javier Allegue, at Unsplash

In previous articles we talked about tick bars, volume bars and dollar bars, alternative types of bars which allow market activity-dependent sampling based on the number of ticks, volume or dollar value exchanged. Additionally, we saw how these bars display better statistical properties such as lower serial correlation when compared to traditional time-based bars. In this article we will talk

about information-driven bars and specifically about imbalance bars. These bars aim to extract information encoded in the observed sequence of trades and notify us of a change in the imbalance of trades. The early detection of an imbalance change will allow us to anticipate a potential change of trend before reaching a new equilibrium.

The concept behind imbalance bars

Imbalance bars were firstly described in the literature by Lopez de Prado in his book *Advances in Financial Machine Learning* (2018). In his own words:

> *The purpose of information-driven bars is to sample more frequently when new information arrives to the market. […] By synchronizing sampling with the arrival of informed traders, we may be able to make decisions before prices reach a new equilibrium level.*

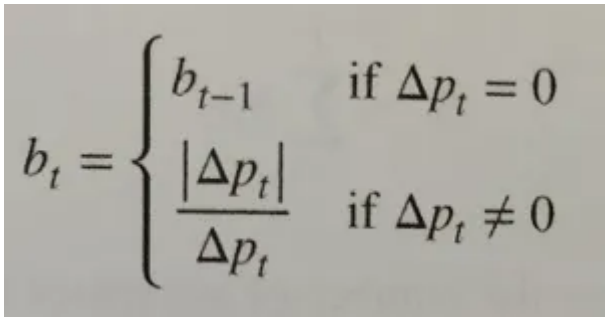
Imbalance bars can be applied to tick, volume or dollar data to produce tick (TIB), volume (VIB) and dollar (DIB) imbalance bars, respectively. Volume and dollar bars are just an extension of tick bars so in this article we will focus mainly on tick imbalance bars and then we will briefly discuss how to extend them to handle volume or dollar information.

The main idea behind imbalance bars is that, based on the imbalance of the sequence of trades, we generate some expectation or threshold and we sample a bar every time the imbalance exceeds that threshold/expectation. But how do we calculate the imbalance? And how do we define the threshold? Let's try to answer these questions.

What is tick imbalance?

Given a sequence of trades, we apply the so-called *tick rule* to generate a list of signed ticks (bt). You can see the tick rule in Formula 1. Essentially, for each trade:

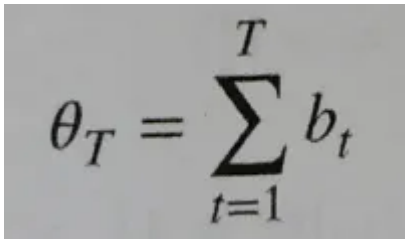
1. if the price is higher than in the previous trade, we set the signed tick as 1;
2. if the price is lower than in the previous trade, we set the signed tick as -1;
3. if the price is the same as in the previous trade, we set the signed tick equal to the previous signed tick.



$$b_t = \begin{cases} b_{t-1} & \text{if } \Delta p_t = 0 \\ \dfrac{|\Delta p_t|}{\Delta p_t} & \text{if } \Delta p_t \neq 0 \end{cases}$$

Formula 1. Tick rule to define signed ticks [1,-1]. pt is the price of trade t and delta-pt is the increment in price respect p(t-1). b(t-1) is the signed tick at t-1.

By applying the *tick rule* we transform all trades to signed ticks (either 1 or -1). This sequence of 1s and -1s can be summed up (cumulative sum) to calculate how imbalanced is the market (Formula 2) at any time T.



$$\theta_T = \sum_{t=1}^{T} b_t$$

Formula 2. Cumulative sum of signed ticks up to time T.

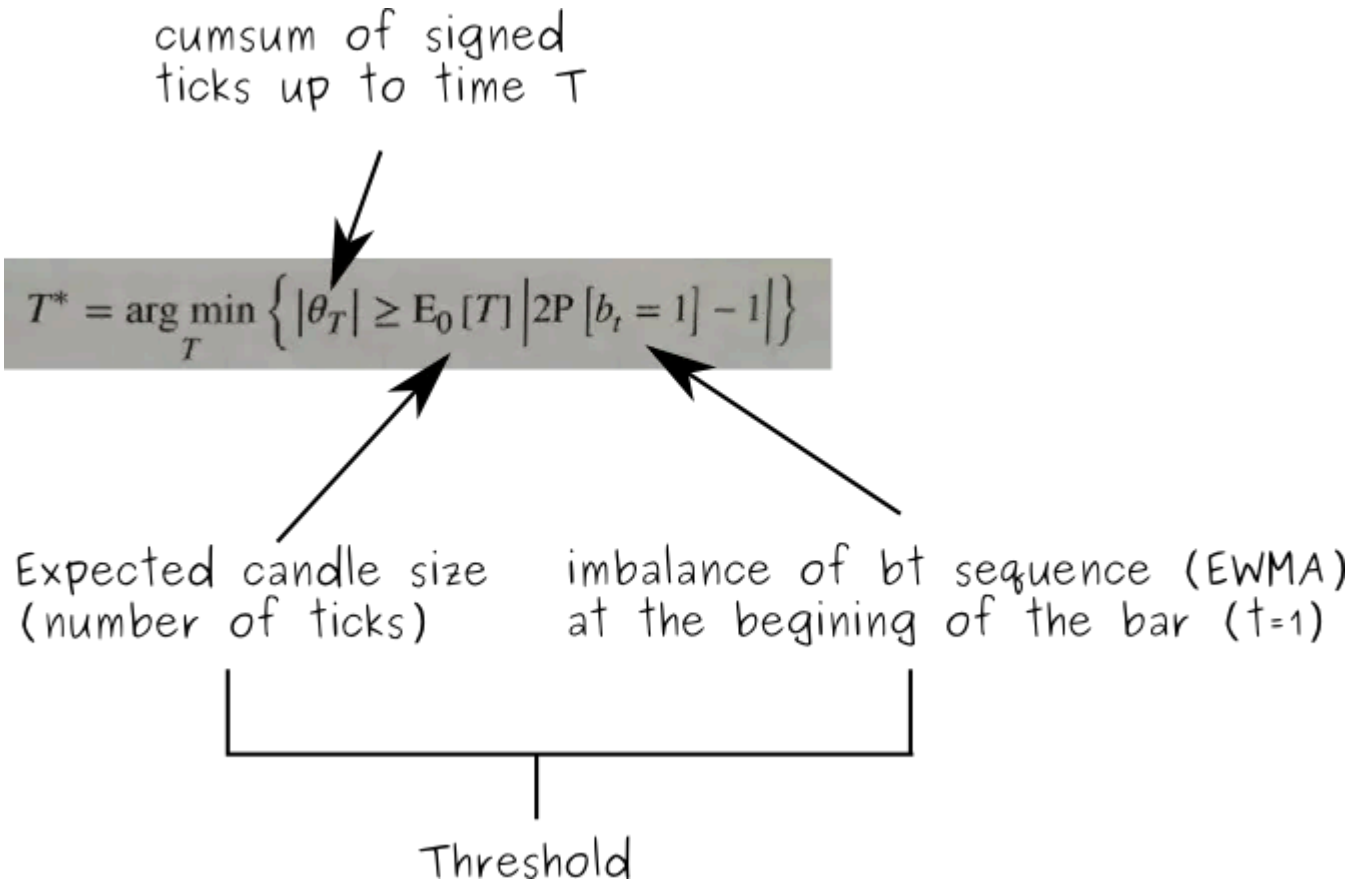
The intuition behind the signed tick imbalance is that we want to create a metric to see how many trades have been done towards a “higher price” direction (+1) or towards a “lower price” direction (-1). In the tick imbalance definition we assume that, in general, there will be more ticks towards a particular up/down direction if there are more informed traders that believe on a particular direction. Finally, we assume that the presence of a higher amount of informed traders towards a particular direction is correlated with information arrival (e.g. favorable technical indicators or news releases) that could lead the market to a new equilibrium. The goal of imbalance bars is to detect these inflows of information as early as possible so we can be notified on time of a potential trading opportunity.

How do we set the threshold?

At the beginning of each imbalance bar, we look at the sequence of old signed ticks and we calculate how much the signed tick sequence is imbalanced towards 1 or -1 by calculating an exponentially weighted moving average (EWMA). Finally, we multiply the EWMA value (the expected imbalance) by the expected bar length (number of ticks) and the result is the threshold or expectation that our cumulative sum of signed ticks must surpass (in absolute value) to trigger the sampling of a new candle.

How do we define a tick imbalance bar?

In mathematical terms we define a tick imbalance bar (TIB) as a contiguous subset of ticks that satisfy the following condition:



Formula 3. Tick imbalance bar definition.

A visual example

Let's look at a visual example:

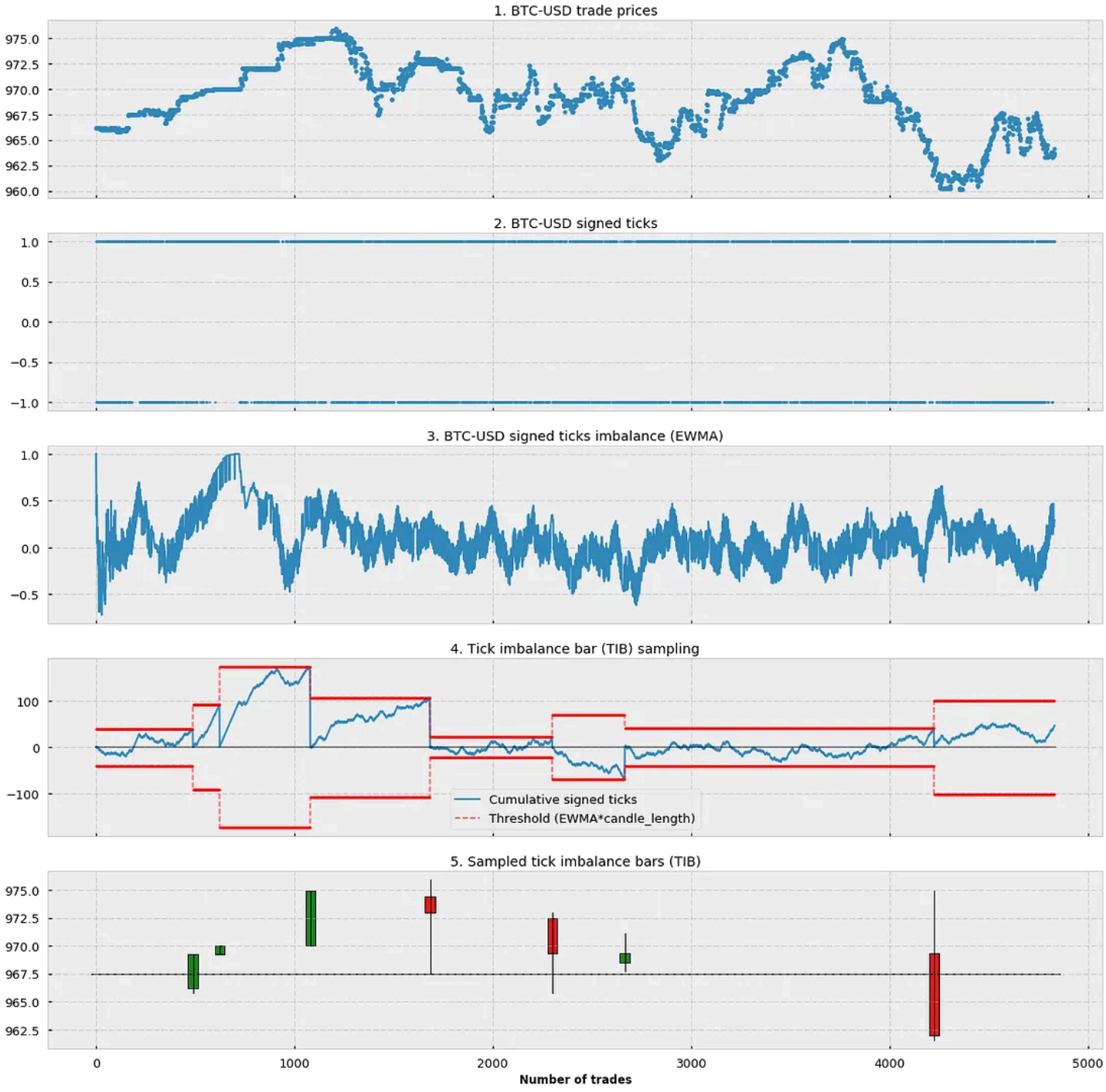


Figure 1. Example of tick imbalance bars for the BTC-USD pair.

In Figure 1.1 you can see the price of approx. 5000 trades starting from 31–01–2017 for the BTC-USD pair in the Bitfinex exchange (source: CryptoDatum.io).

In Figure 1.2 you can see how we applied the tick rule and transformed all the trades from 1.1. into signed ticks (1 or -1). Notice that there are more than 5000 signed ticks and most of the time they overlap with each other.

In Figure 1.3 we applied an exponential weighted moving average (EWMA) to the whole sequence of signed ticks. We can observe how the resulting EWMA is a stochastic oscillating wave between -1 and 1 that indicates the general trend/frequency of positive and negative signed ticks.

In Figure 1.4. we show, in red, the threshold or expectation as calculated in the last term of Formula 3. This threshold is calculated at the beginning of each bar. Notice that in the figure we show both the positive and negative threshold but, in practice, since we use the absolute value (Formula 3), we only care about the positive one. In blue, we show the cumulative sum of signed ticks at each particular point in time. Notice that the cumulative sum oscillates until reaching the lower or upper threshold, point in which a new candle is sampled, the cumulative sum is reset to 0 and a new threshold (expectation) is calculated based on the EWMA imbalance at that particular point.

Finally, in Figure 1.5 we represent the generated tick imbalance bars.

Implementation and observations

If you followed the explanation above, you may be wondering about:

1. Concrete implementations of the TIB.
2. How to calculate the “expected candle size”.

To answer the question 1, please refer to this Github issue, as well as the parent repository. They offer good starting content to understand and implement tick imbalance bars in Python but beware of errors and different interpretations of TIBs.

In the same Github issue, the question 2 is thoroughly discussed. The official definition by Lopez de Prado states that the expected candle size, much like the “expected imbalance” at time t=1, should be calculated as an EWMA of T values of previous bars. However, in my experience and like other people in the thread, the sizes of the bars end up exploding (very big sizes of thousands of ticks) after few iterations. The reason is simple: as a threshold grows, it takes more and more signed ticks to reach the threshold which, in turn, makes the “expected candle size” grow in a positive-feedback loop that keeps increasing the candle size until infinity. I have tried different solutions to fix this issue: (1) limiting the max. candle size and (2) fixing the candle size. It turns out the limiting the max. candle size to, for instance, 200 makes all expected candle sizes to become 200 after few iterations. Therefore,

both solutions work indistinctly and following the Occam's razor principle I went for the simplest one (solution 2). Since now the candle size becomes a variable to take into account, in CryptoDatum.io we decided to offer tick imbalance bars for three different candle sizes: 100, 200 and 400.

The way I interpret thresholds and these candle sizes is in terms of a "challenge". Every time you set a new expectation/threshold at the beginning of a bar, we are challenging the time series to exceed our expectations. In these "challenges", the candle size becomes one more parameter that allows us to specify "how big" we want this challenge to be. If we pick a larger candle size, we are essentially increasing the "challenge" difficulty and, as a result, we will end up with a lower amount of bar sampling although, in principle, with higher meaningfulness.

Volume and Dollar imbalance bars

Up to now we only talked about tick imbalance bars. It turns out that generating volume and dollar bars is trivial and it just involves adding a final multiplication term in Formula 2: either the volume (in case of volume imbalance bars — VIB) or the dollar/fiat value (in case of dollar imbalance bars — DIB).

Statistical properties

As we did with tick, volume and dollar bars, we will look at two statistical properties: (1) serial correlation and (2) normality of returns. We will analyze the first one by running the Pearson correlation test of the shifted series (shift=1) and we will analyze the latter by running the Jarque-bera test of normality.

Let's look at the Pearson correlation test:

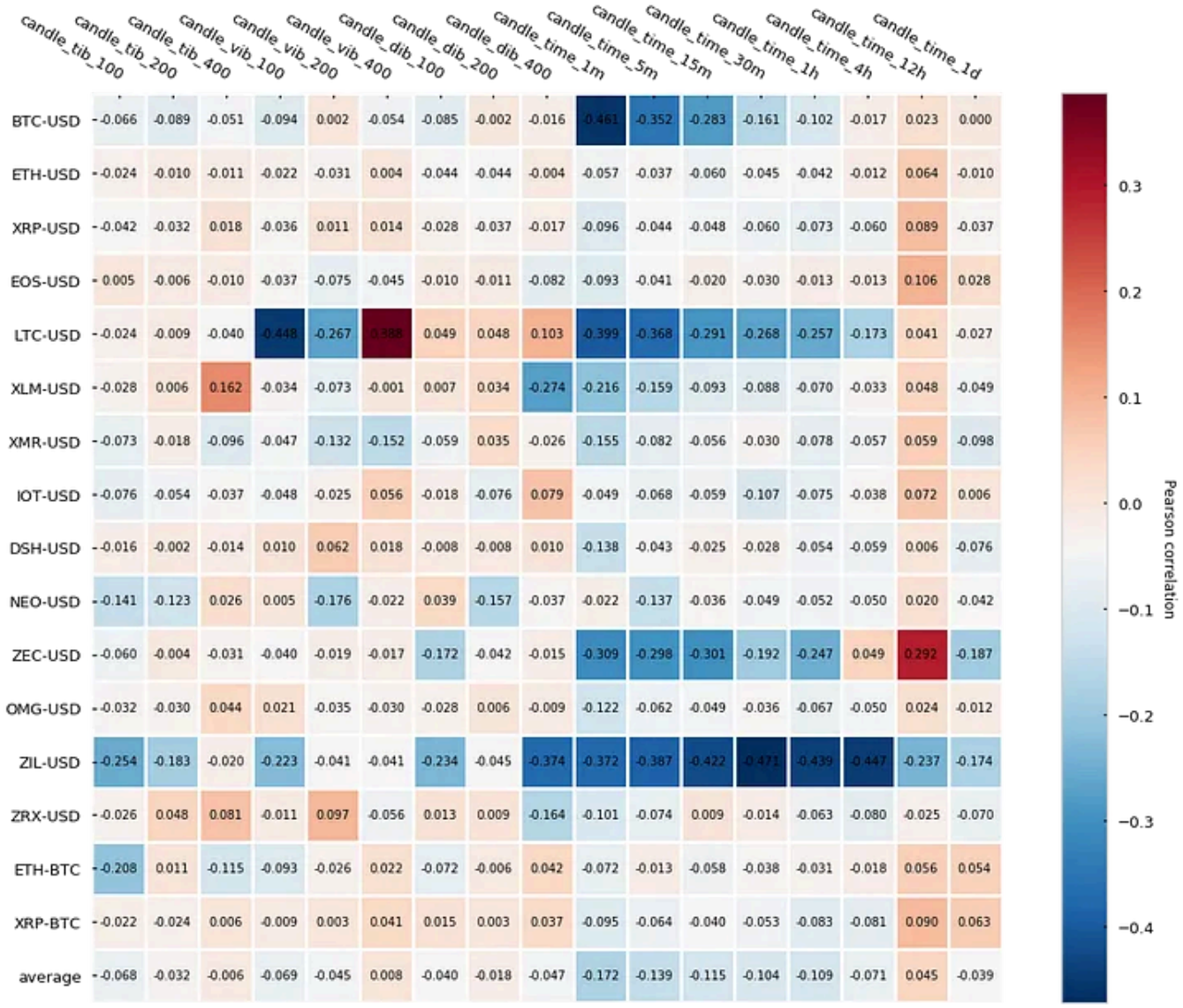


Figure 2. Pearson correlation of the shifted series of returns (shift=1)

Similar to other alternative bars (tick and volume bars) the overall auto-correlation is lower in imbalance bars than in traditional time-based candlesticks. As we explained in the original article, this is a good feature because it means data points are more independent of each other.

Now let's look at the Jarque-Bera normality test:

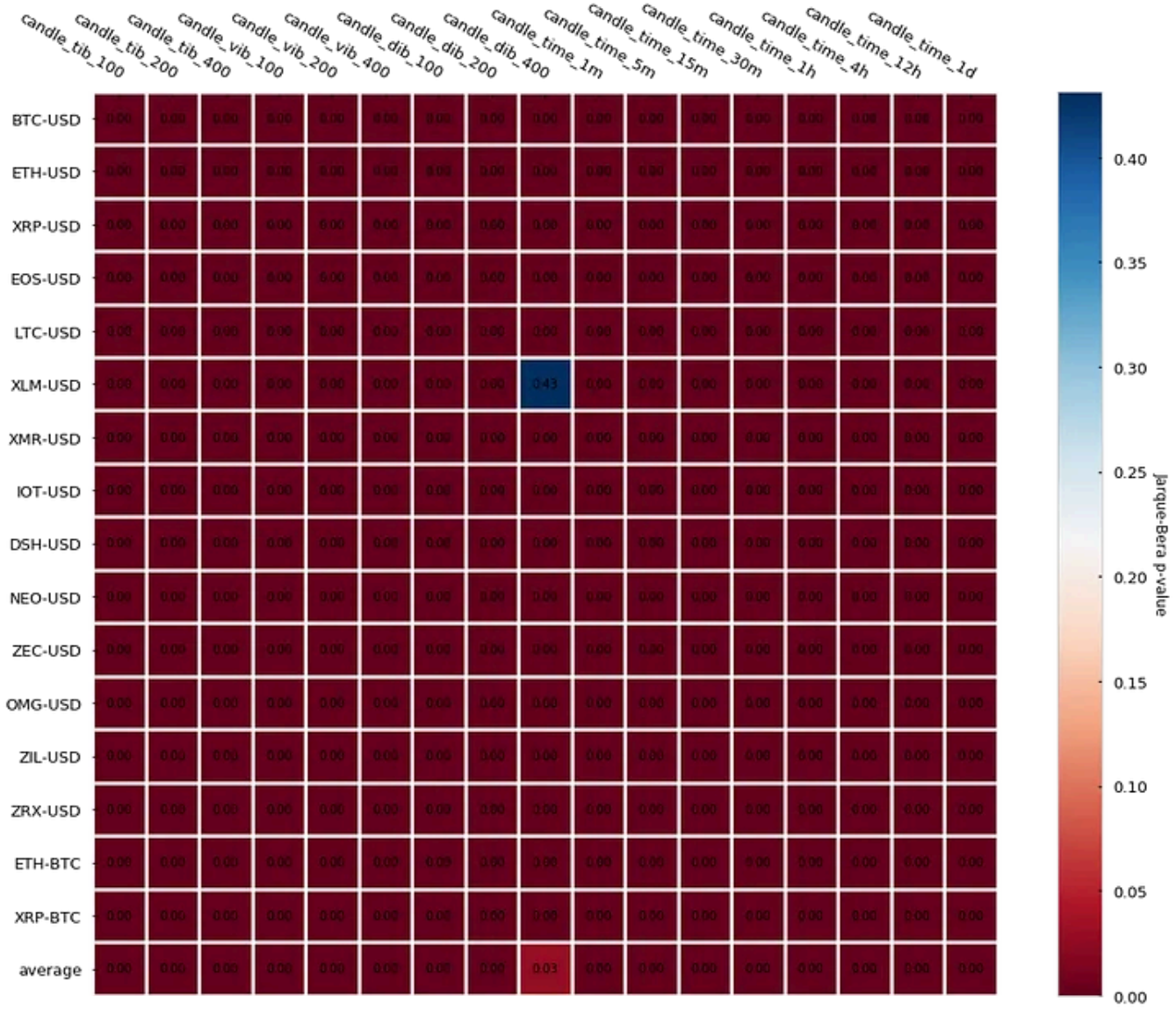


Figure 3. Jarque-Bera normality test

We reject the null hypothesis of normality in both imbalance bars and time-based bars. For good or for bad this does not come as a surprise as results are in line with what we saw in previous articles.

What have we learned?

- Imbalance bars are generated by observing the imbalance of the asset price.
- The imbalance is measured by the magnitude of the cumulative sum of signed ticks.
- Signed ticks are computed by applying the *tick rule*.
- Bars are sampled every time the imbalance exceeds our expectations (calculated at the beginning of each bar).

- The objective of imbalance bars is to early detect a shift in the directionality of the market before a new equilibrium in reached.
- Imbalance bars display lower autocorrelation compared to traditional time-based candlesticks and non-normality of results.

This project is part of our research at CryptoDatum.io, a cryptocurrency data API that aims to provide plug-and-play datasets to train machine learning algorithms. If you liked the data we showed in this article, get your free API key and play with it yourself at https://cryptodatum.io

Trading | Cryptocurrency | Towards Data Science | Algorithmic Trading | API

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

473 6

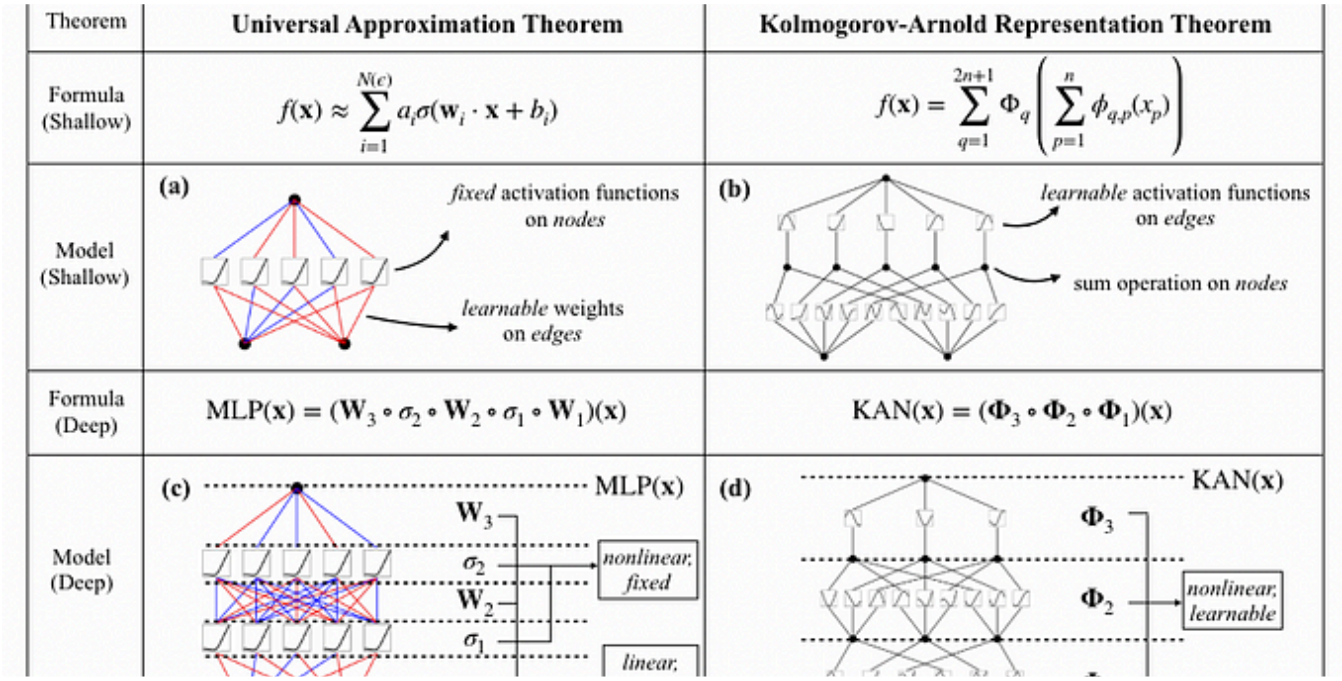
Torsten Walbaum in Towards Data Science

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



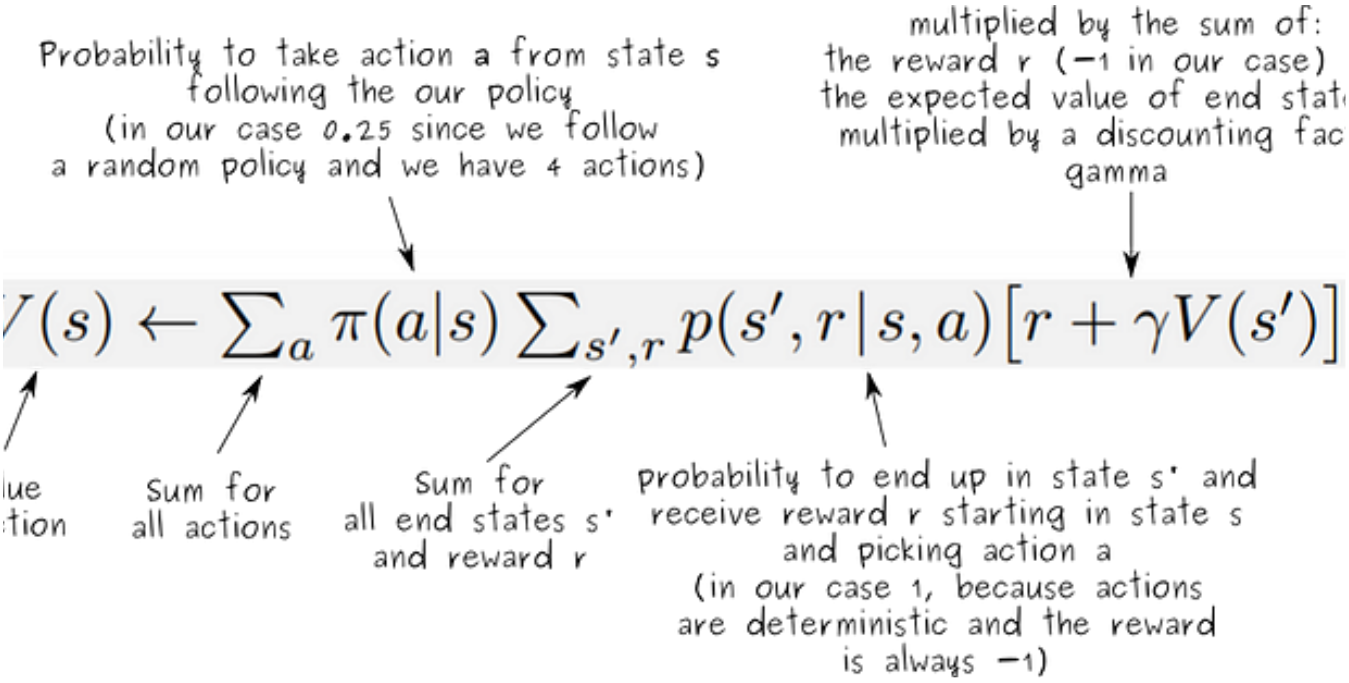
Theo Wolf in Towards Data Science

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

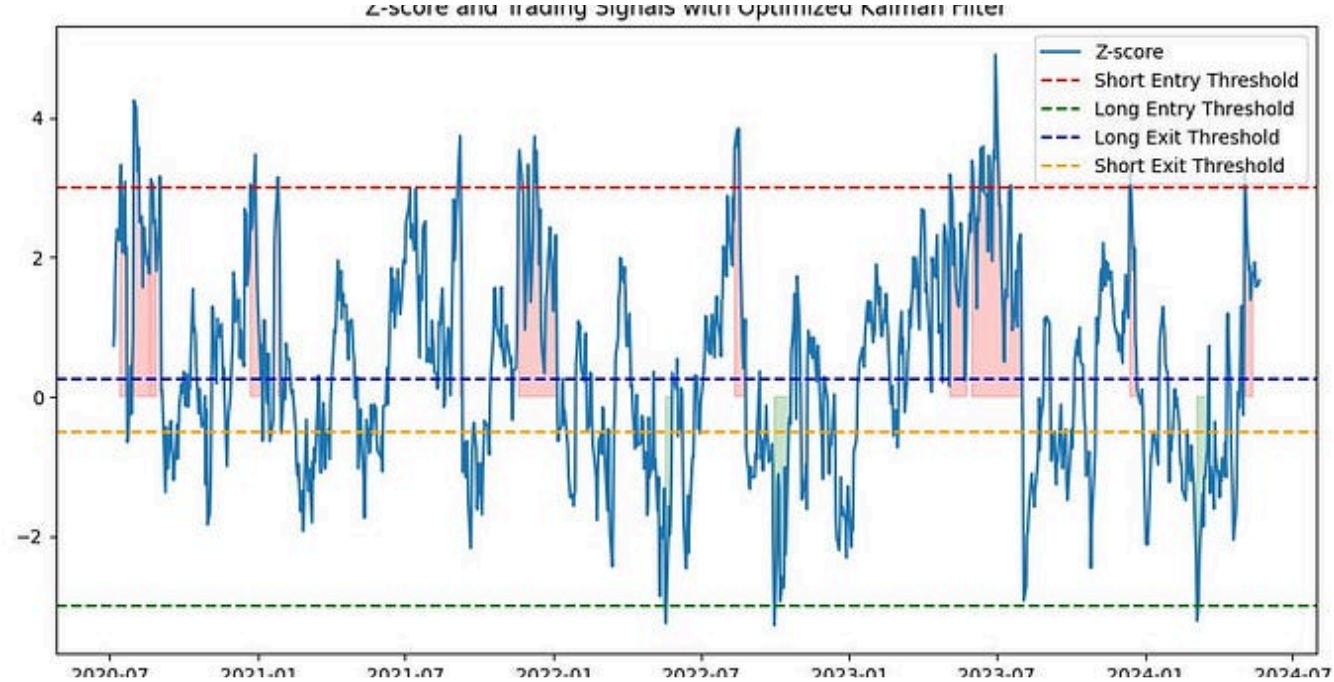
8 min read · Dec 20, 2018

716 7

See all from Gerard Martínez

See all from Towards Data Science

Recommended from Medium



Serdar İlarslan

Implementing a Kalman Filter-Based Trading Strategy

Financial markets are inherently noisy and unpredictable, making it challenging for traders and investors to identify and capitalize on...

7 min read · May 25, 2024

163 2

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...

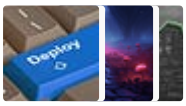
23 min read · Mar 6, 2024

202 2

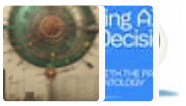
Lists

Coding & Development
11 stories · 664 saves

Company Offsite Reading List
8 stories · 130 saves



Predictive Modeling w/ Python
20 stories · 1316 saves



data science and AI
40 stories · 192 saves

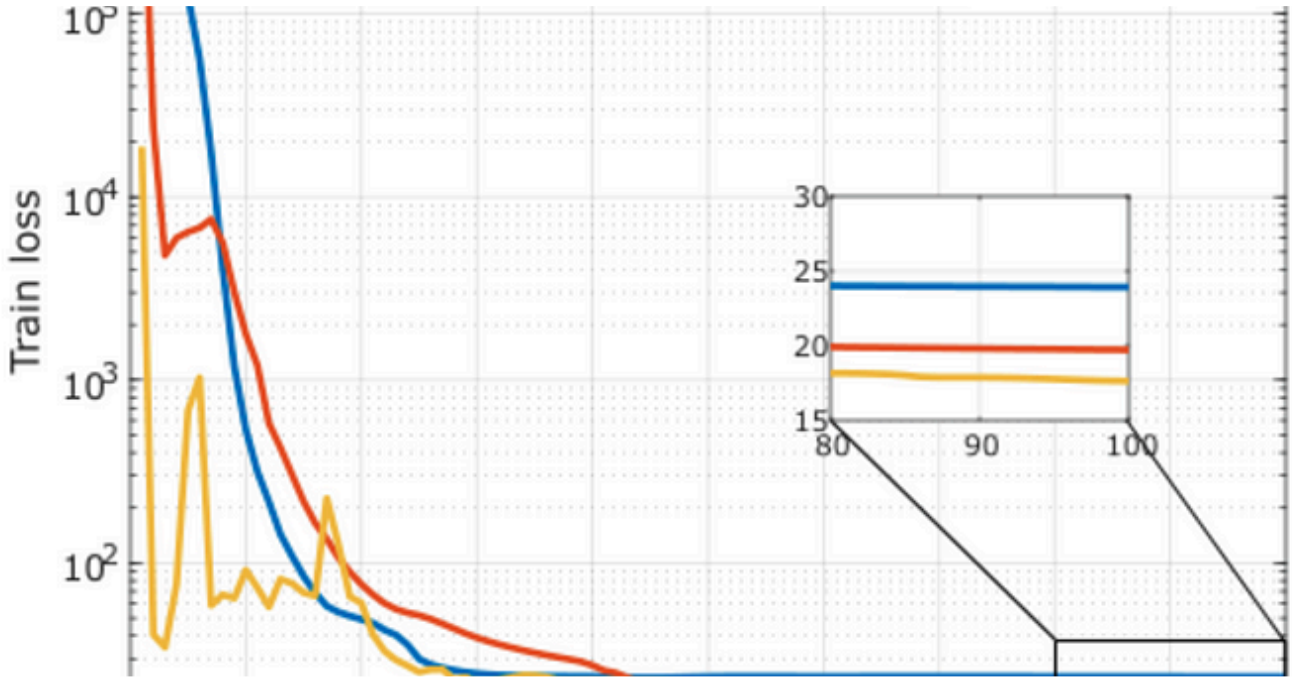
Scott Stockdale in The Startup

How Pieter Levels Makes (At Least) $210K a Month From His Laptop—With Zero Employees

I went through 126.8K tweets & found 7 patterns to Pieter's success

· 7 min read · May 17, 2024

4.4K 44



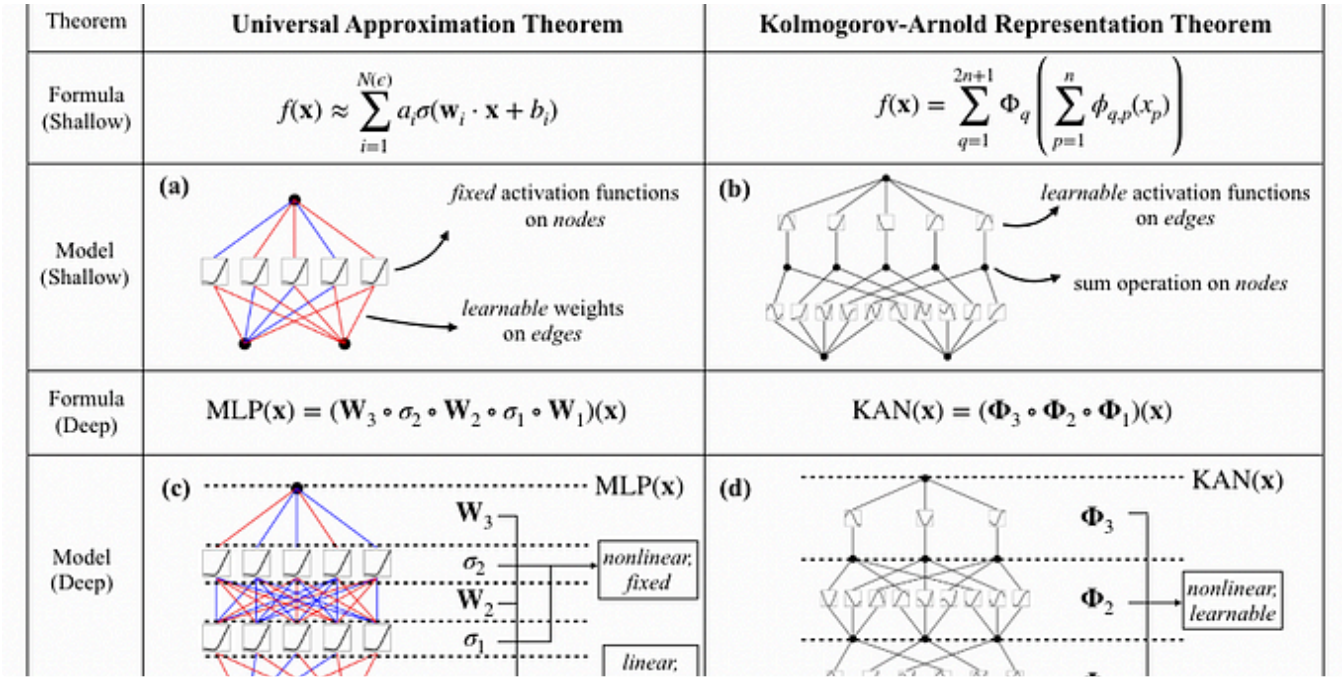
Sercan Bugra Gultekin

(2) Stock Price Prediction with ML in Python: GRU (Gated Recurrent Unit) Model

In our first series, I made a forecast with the LSTM model, now we will change our model and try the same experiment with GRU (Gated...

· 3 min read · May 31, 2024

1



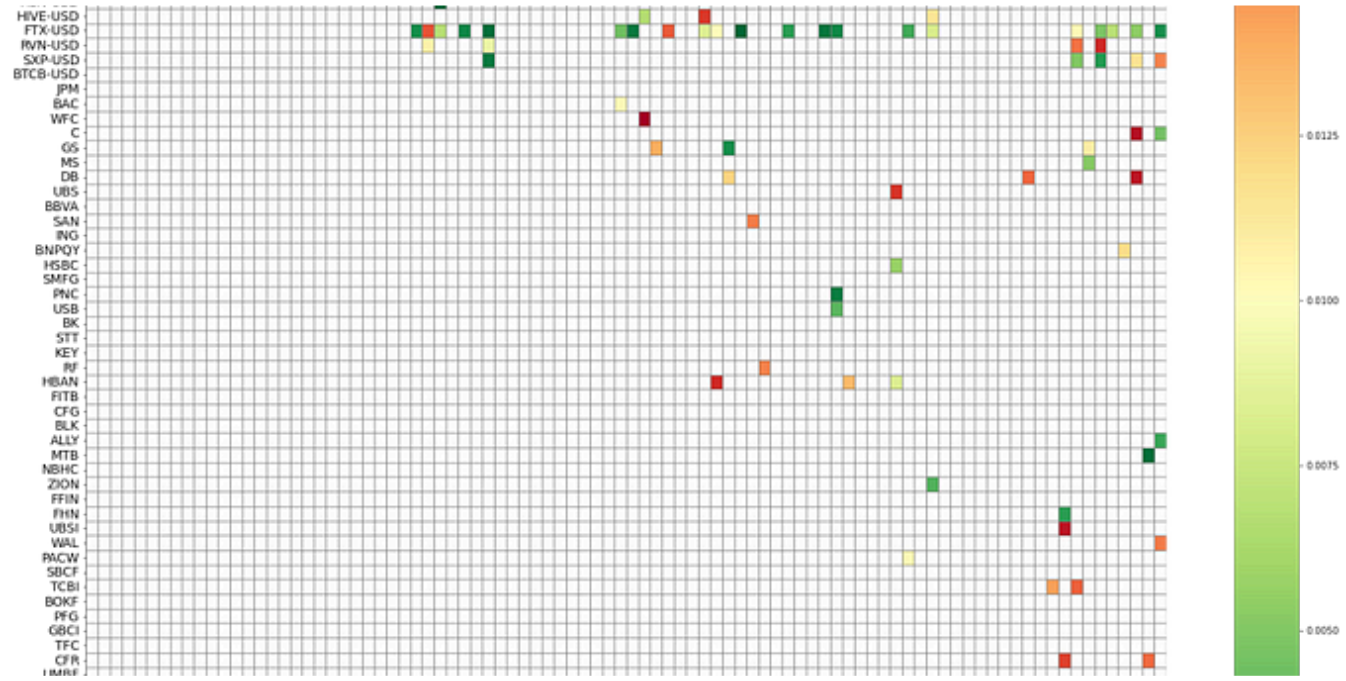
Theo Wolf in Towards Data Science

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



Ayrat Murtazin in DataDrivenInvestor

Citadel's Strategy Anyone Can Use—Pairs-Trading

Pairs trading is a sophisticated strategy often employed by quantitative traders to trade a portfolio, rather than focusing solely on...

 · 12 min read · Mar 6, 2024

 402  10 

See more recommendations