TDS Archive · Follow publication

Information-driven bars for financial machine learning: imbalance bars

8 min read · May 21, 2019

Gerard Martínez Follow

Listen Share More

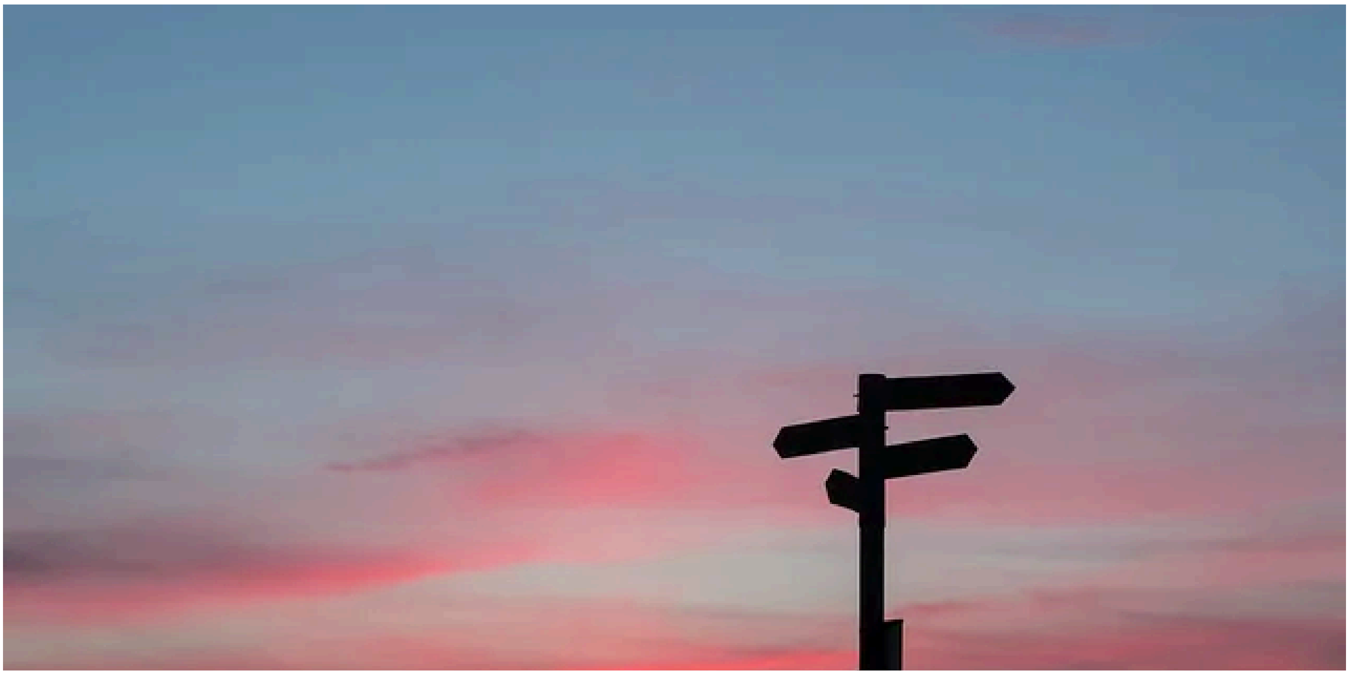


Photo by Javier Allegue, at Unsplash

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.

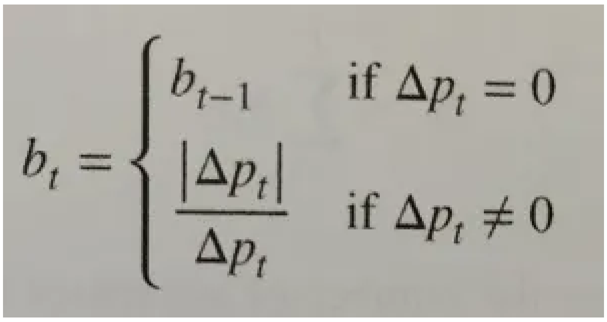
The main idea behind imbalance bars is that, based on the imbalance of the sequence of trades, we generate some expectation or threshold and we sample a bar every time the imbalance exceeds that threshold/expectation. But how do we calculate the imbalance? And how do we define the threshold? Let's try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

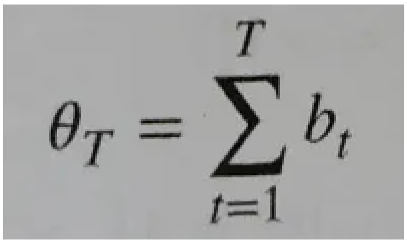
3. if the price is the same as in the previous trade, we set the signed tick equal to the previous signed tick.



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

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.

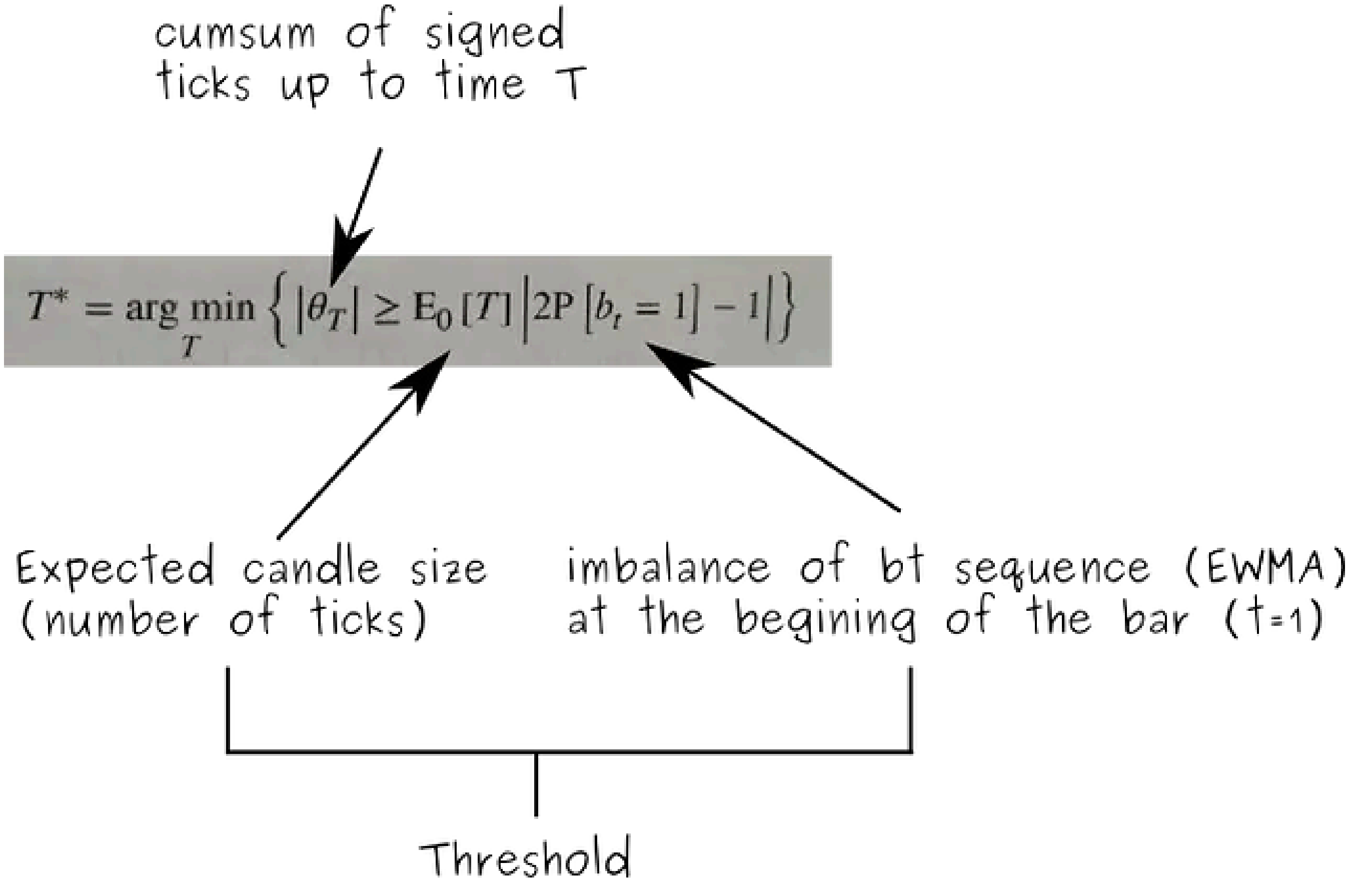
The intuition behind the signed tick imbalance is that we want to create a metric to see how many trades have been done towards a "higher price" direction (+1) or towards a "lower price" direction (-1). In the tick imbalance definition we assume that, in general, there will be more ticks towards a particular 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:



$$T^* = \arg\min_T \left\{ |\theta_T| \geq \mathrm{E}_0[T] \left| 2\mathrm{P}[b_t = 1] - 1 \right| \right\}$$

cumsum of signed ticks up to time T
Expected candle size (number of ticks)
imbalance of bt sequence (EWMA) at the begining of the bar (t=1)
Threshold

Formula 3. Tick imbalance bar definition.

A visual example

Let’s look at a visual example:

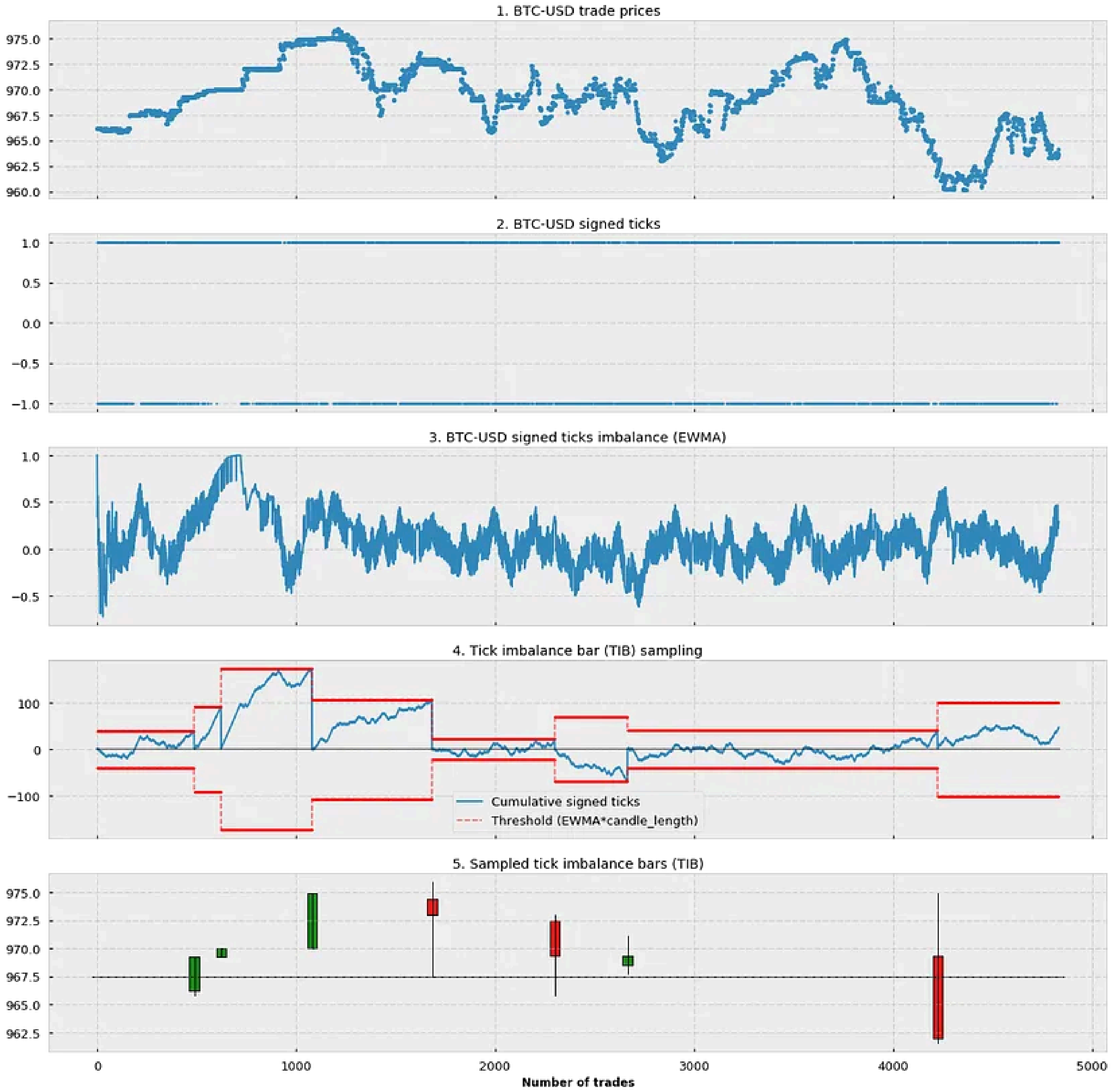


Figure 1. Example of tick imbalance bars for the BTC-USD pair.

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.

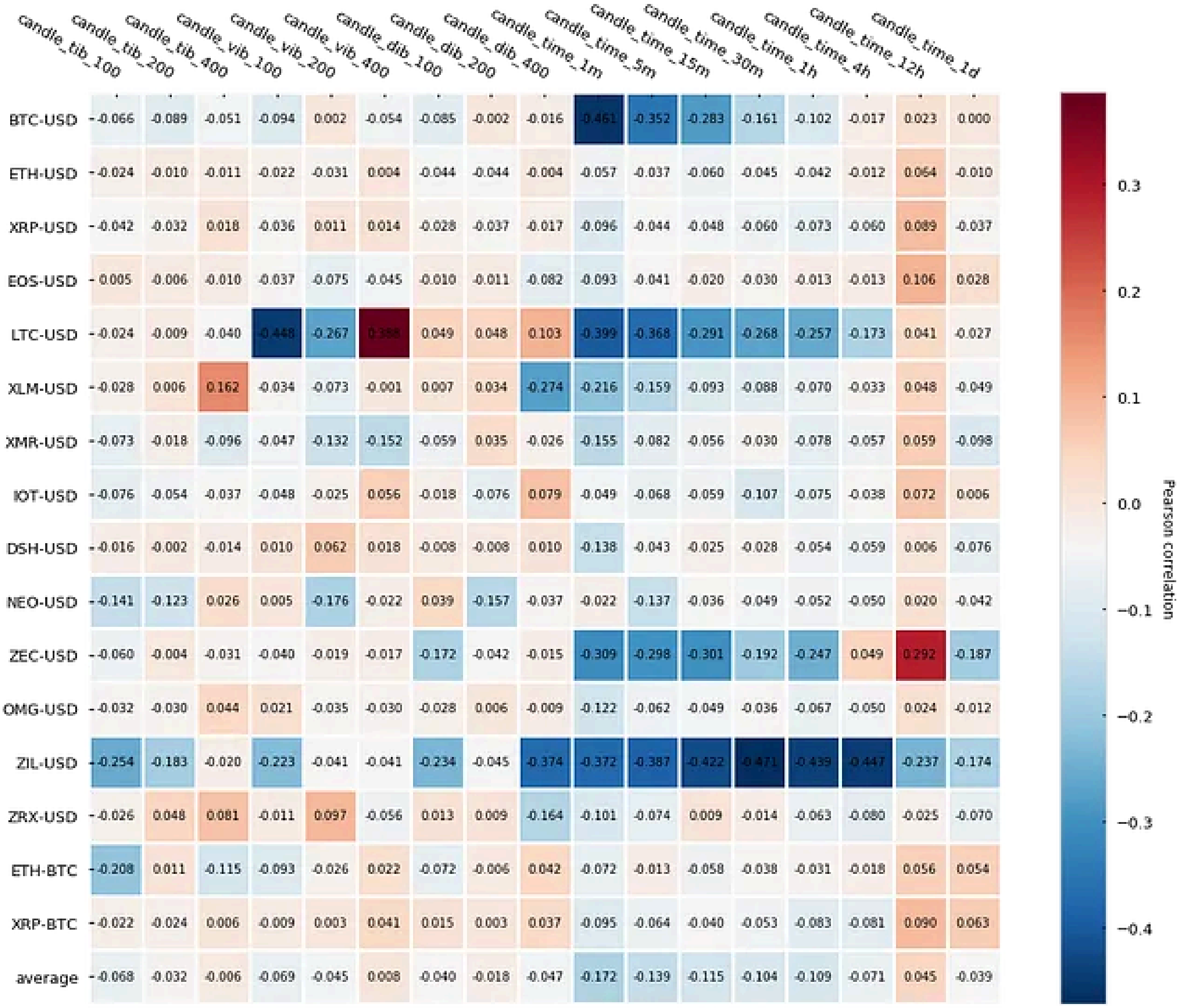
Volume and Dollar imbalance bars

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:



	candle_tib_100	candle_tib_200	candle_tib_400	candle_vib_100	candle_vib_200	candle_vib_400	candle_dib_100	candle_dib_200	candle_dib_400	candle_time_1m	candle_time_5m	candle_time_15m	candle_time_30m	candle_time_1h	candle_time_4h	candle_time_12h	candle_time_1d
BTC-USD	-0.066	-0.089	-0.051	-0.094	0.002	-0.054	-0.085	-0.002	-0.016	-0.461	-0.352	-0.283	-0.161	-0.102	-0.017	0.023	0.000
ETH-USD	-0.024	-0.010	-0.011	-0.022	-0.031	0.004	-0.044	-0.044	-0.004	-0.057	-0.037	-0.060	-0.045	-0.042	-0.012	0.064	-0.010
XRP-USD	-0.042	-0.032	0.018	-0.036	0.011	0.014	-0.028	-0.037	-0.017	-0.096	-0.044	-0.048	-0.060	-0.073	-0.060	0.089	-0.037
EOS-USD	0.005	-0.006	-0.010	-0.037	-0.075	-0.045	-0.010	-0.011	-0.082	-0.093	-0.041	-0.020	-0.030	-0.013	-0.013	0.106	0.028
LTC-USD	-0.024	-0.009	-0.040	-0.448	-0.267	0.398	0.049	0.048	0.103	-0.399	-0.368	-0.291	-0.268	-0.257	-0.173	0.041	-0.027
XLM-USD	-0.028	0.006	0.162	-0.034	-0.073	-0.001	0.007	0.034	-0.274	-0.216	-0.159	-0.093	-0.088	-0.070	-0.033	0.048	-0.049
XMR-USD	-0.073	-0.018	-0.096	-0.047	-0.132	-0.152	-0.059	0.035	-0.026	-0.155	-0.082	-0.056	-0.030	-0.078	-0.057	0.059	-0.098
IOT-USD	-0.076	-0.054	-0.037	-0.048	-0.025	0.056	-0.018	-0.076	0.079	-0.049	-0.068	-0.059	-0.107	-0.075	-0.038	0.072	0.006
DSH-USD	-0.016	-0.002	-0.014	0.010	0.062	0.018	-0.008	-0.008	0.010	-0.138	-0.043	-0.025	-0.028	-0.054	-0.059	0.006	-0.076
NEO-USD	-0.141	-0.123	0.026	0.005	-0.176	-0.022	0.039	-0.157	-0.037	-0.022	-0.137	-0.036	-0.049	-0.052	-0.050	0.020	-0.042
ZEC-USD	-0.060	-0.004	-0.031	-0.040	-0.019	-0.017	-0.172	-0.042	-0.015	-0.309	-0.298	-0.301	-0.192	-0.247	0.049	0.292	-0.187
OMG-USD	-0.032	-0.030	0.044	0.021	-0.035	-0.030	-0.028	0.006	-0.009	-0.122	-0.062	-0.049	-0.036	-0.067	-0.050	0.024	-0.012
ZIL-USD	-0.254	-0.183	-0.020	-0.223	-0.041	-0.041	-0.234	-0.045	-0.374	-0.372	-0.387	-0.422	-0.471	-0.439	-0.447	-0.237	-0.174
ZRX-USD	-0.026	0.048	0.081	-0.011	0.097	-0.056	0.013	0.009	-0.164	-0.101	-0.074	0.009	-0.014	-0.063	-0.080	-0.025	-0.070
ETH-BTC	-0.208	0.011	-0.115	-0.093	-0.026	0.022	-0.072	-0.006	0.042	-0.072	-0.013	-0.058	-0.038	-0.031	-0.018	0.056	0.054
XRP-BTC	-0.022	-0.024	0.006	-0.009	0.003	0.041	0.015	0.003	0.037	-0.095	-0.064	-0.040	-0.053	-0.083	-0.081	0.090	0.063
average	-0.068	-0.032	-0.006	-0.069	-0.045	0.008	-0.040	-0.018	-0.047	-0.172	-0.139	-0.115	-0.104	-0.109	-0.071	0.045	-0.039

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:

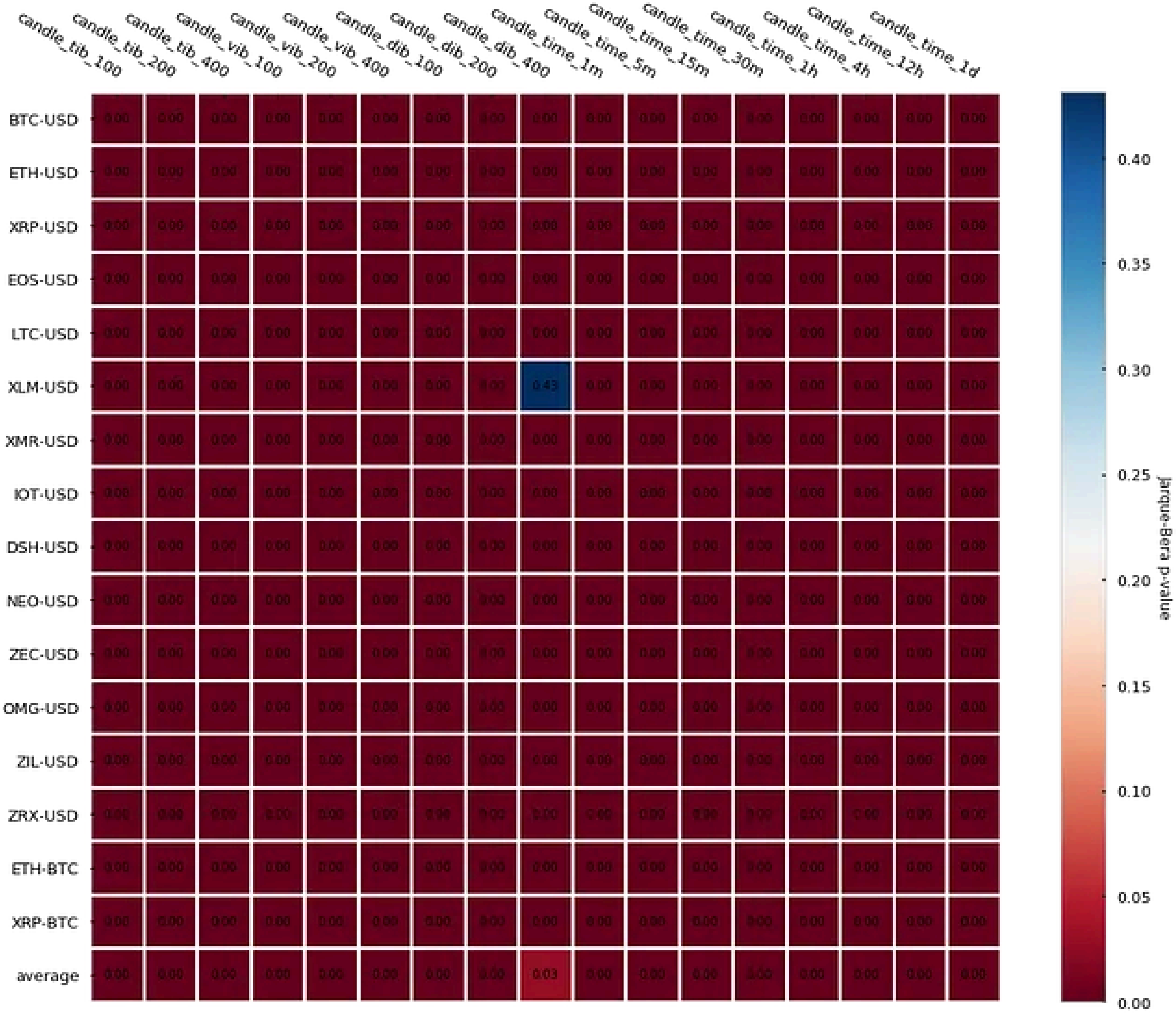


Figure 3. Jarque-Bera normality test

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

Data Science

Follow

Published in TDS Archive

820K Followers · Last published Feb 4, 2025

An archive of data science, data analytics, data engineering, machine learning, and artificial intelligence writing from the former Towards Data Science Medium publication.

Follow

Written by Gerard Martínez

1.5K Followers · 184 Following

Trading strategy developer — Founder of CryptoDatum.io

Responses (4)

Wangchuyin

What are your thoughts?

Brendan Blanchard
Jan 3, 2022

This is a shameful ripoff of copyrighted content in order to promote a product. Equations are literal scans from Lopez's book, with a vague attempt to add some commentary that doesn't improve upon what's in the book.

9 Reply

Frank Dai
Jun 5, 2020

Hi Gerard, how many bars(not ticks) do you use to calculate the EWMA of bt? since you set E[T] to fixed sizes of 100, 200, 400, you don't need to calculate EWMA of T anymore, but according to the AdvFinML book, the expectation of bt is the EWMA of bt from piror bars, your article didn't mention this windows size

3 Reply

Madhurgupta
Oct 15, 2022

I have a more basic question.

Why do we go for Information bars (or volume & dollar bars, for that matter) over time bars?

How are models made on Information Bars better than those made on time bars? Also, if we had the ability to capture trading data… more

Reply

See all responses

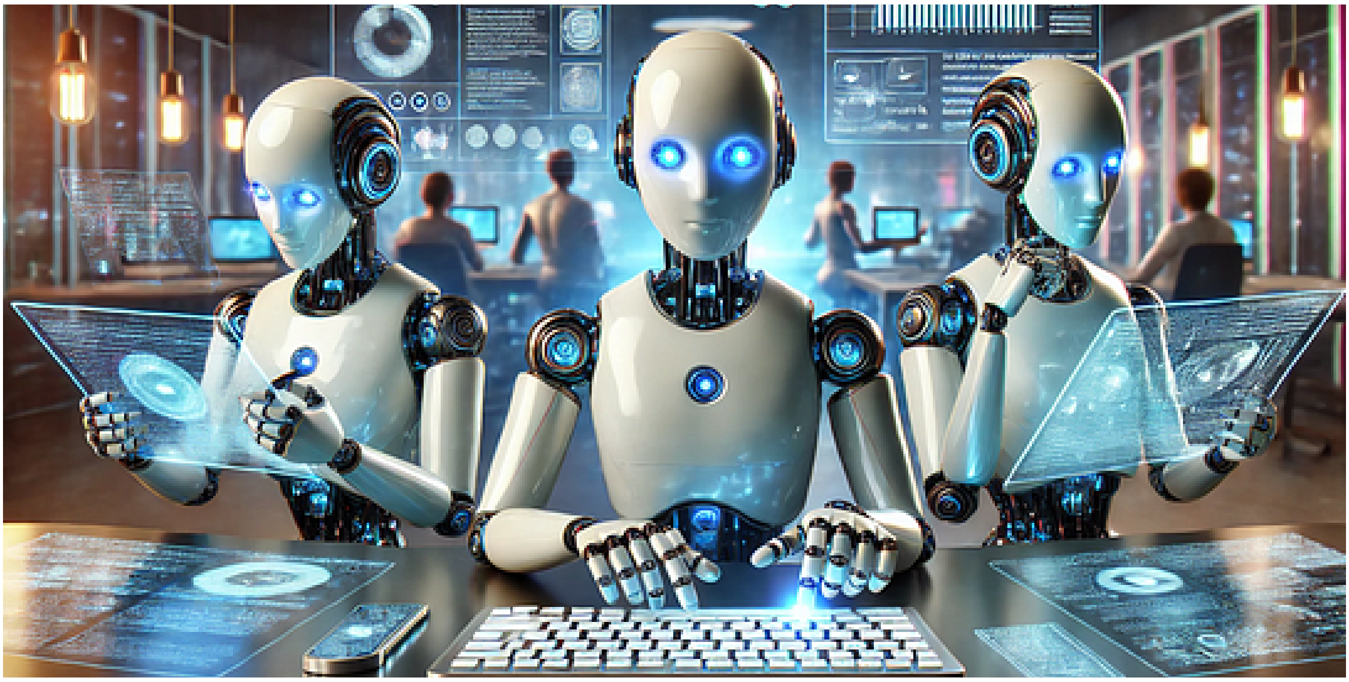
More from Gerard Martínez and TDS Archive

In TDS Archive by Gerard Martínez

Advanced candlesticks for machine learning (ii): volume and dollar bars

In this article we will learn how to build volume and dollar bars and we will explore what advantages they offer in respect to traditional…

May 2, 2019 384 5



In TDS Archive by Luís Roque

Agentic AI: Building Autonomous Systems from Scratch

A Step-by-Step Guide to Creating Multi-Agent Frameworks in the Age of Generative AI

Dec 13, 2024 1K 22

In TDS Archive by Chengzhi Zhao

How to Build an AI Agent for Data Analytics Without Writing SQL

Create a comprehensive AI agent from the ground up utilizing LangChain and DuckDB

Jan 8 478 8

In CryptoDigest by Gerard Martínez

RenkoTrading: an embarrassingly simple algorithm for cryptocurrency trading that will deprecate the...

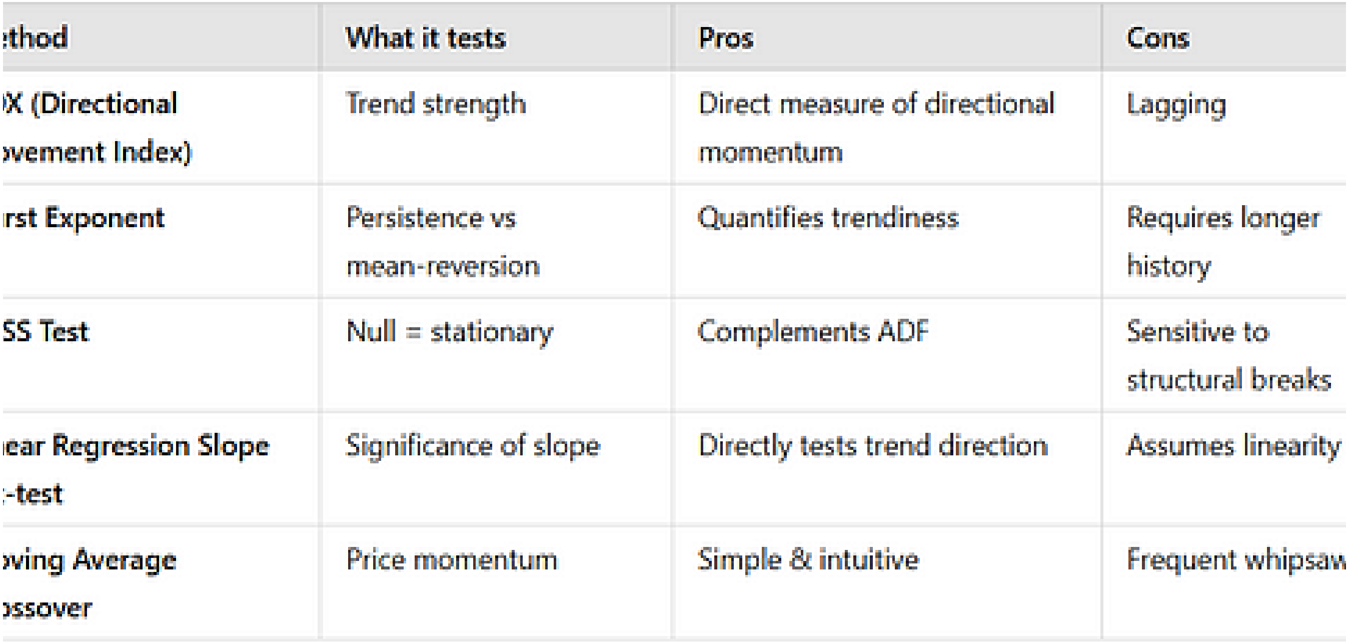
Introduction

Jun 2, 2018 397 1

See all from Gerard Martínez

See all from TDS Archive

Recommended from Medium



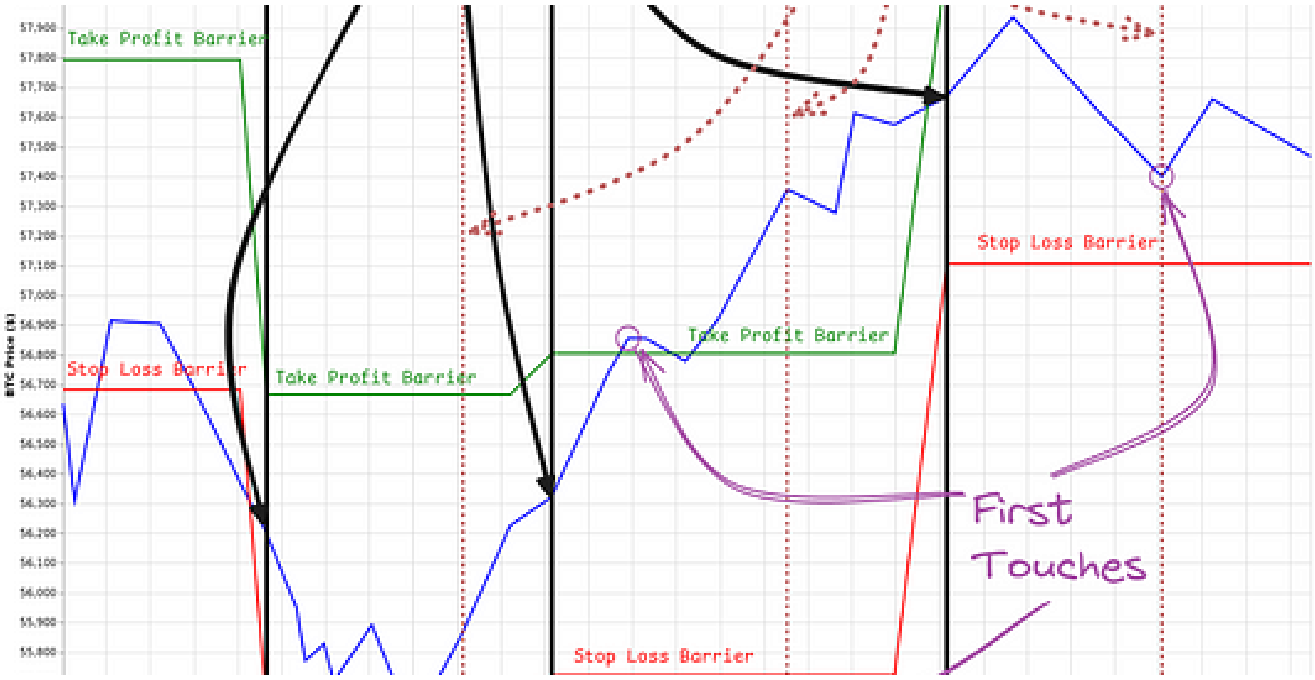
ethod	What it tests	Pros	Cons
X (Directional ovement Index)	Trend strength	Direct measure of directional momentum	Lagging
rst Exponent	Persistence vs mean-reversion	Quantifies trendiness	Requires longer history
SS Test	Null = stationary	Complements ADF	Sensitive to structural breaks
ear Regression Slope -test	Significance of slope	Directly tests trend direction	Assumes linearity
oving Average ossover	Price momentum	Simple & intuitive	Frequent whipsaw

PyQuantLab

ADF-Filtered Weekly Breakout Strategy

Momentum-driven breakouts are a cornerstone of systematic trading. I thought it might be a good idea to try a weekly breakout strategy that…

Mar 27 1



Yair Oz

The Triple Barrier Method: Labeling Financial Time Series for ML in Elixir

Assuming we have a trading strategy that generates entry points, how do we mark success or failure for Machine Learning Training? We can …

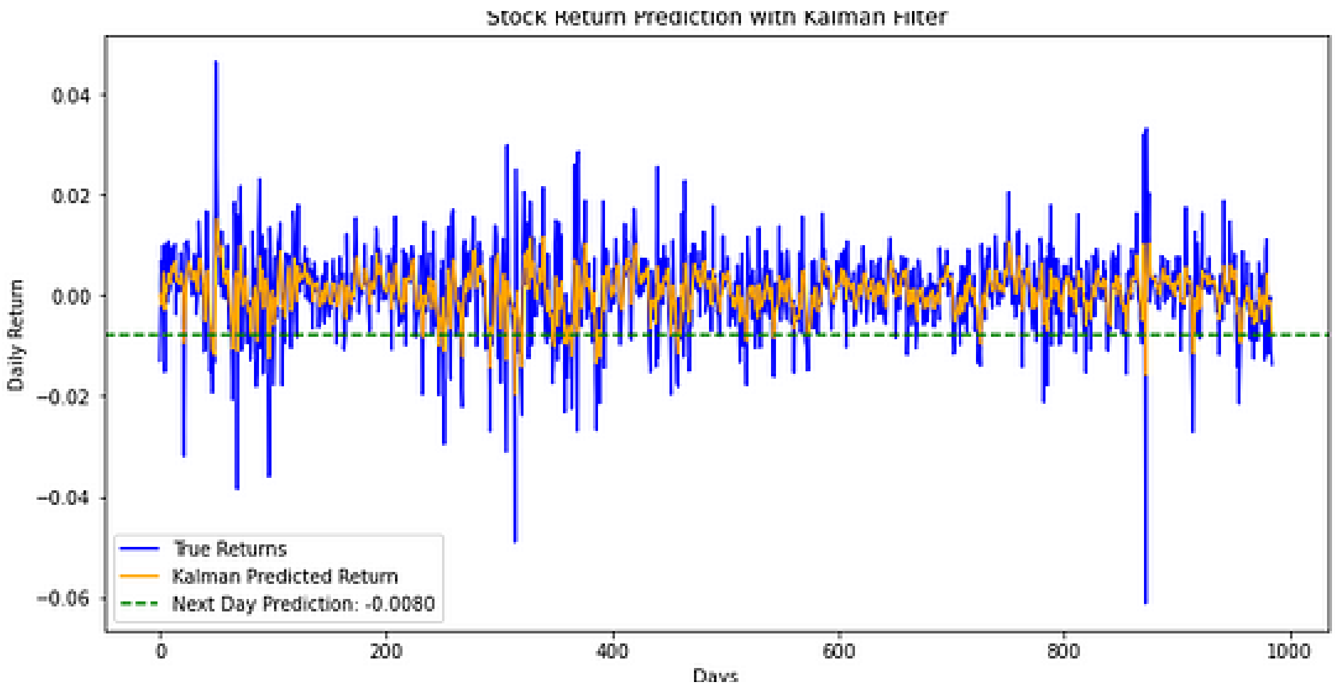
Nov 9, 2024 2

Alexzap

AAPL ML/AI Stock Price Prediction: Fusion of Tsfresh Feature Engineering & Voting Scenarios with...

Supervised ML Stock Price Predictions with Tsfresh Automated Feature Extraction, Weighted Voting Scenarios & 26 Accepted Regressors

Nov 4, 2024 161



In GoPenAI by Abhay Dodiya

Kalman Filter to Predict Next Day Return Nifty Example

A Kalman Filter is a technique from mathematical branch which helps in estimating the state of dynamic system from a noisy data. There are...

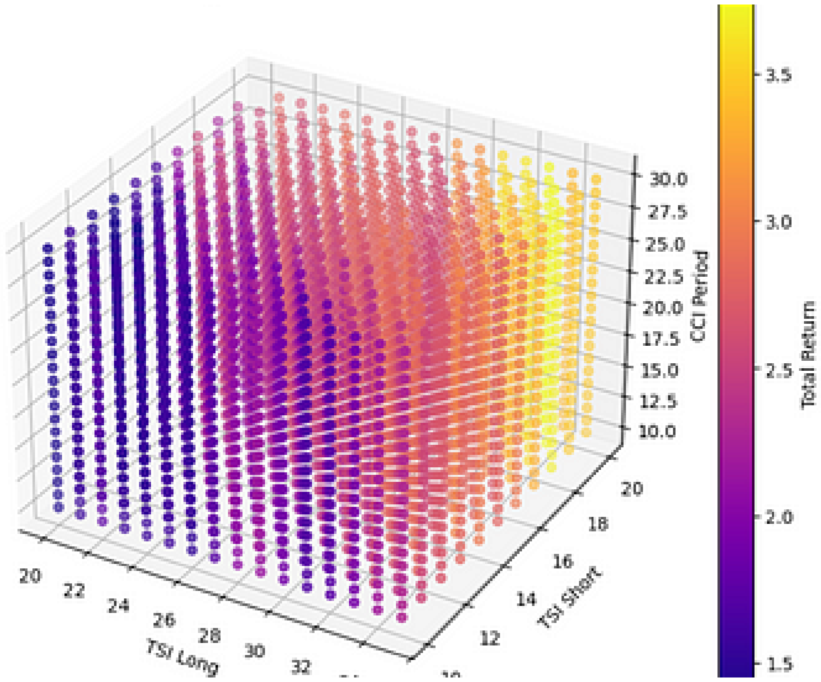
✦ Nov 16, 2024 67 2

Sayedali

The Best Buy/Sell Signals for Scalping with EMA and CPR: A Complete Guide

Learn how to combine EMA and CPR with Range Filter for more accurate buy and sell signals.

✦ Jan 15 12



Kridtapon P.

From Code to Cash: Building a Trading Strategy with TSI & CCI

Parameter tuning and visual analysis unlock a high-performing algorithmic system for trading META stock.

Apr 21 2

See more recommendations