major reasons why machine learning fails in stock prediction: part -01

Ved Prakash · Follow

6 min read · Jan 10, 2024

932 18

in this series of blogs i am going to discuss about the reasons why machine learning fails in predicting the stock prices or in general why machine learning based investment funds fail?. The content of this blog is taken from the book "**advances in financial machine learning by Marcos Lopez de Prado**". this is a must read book for everyone interested in finance. This book mention all the mistake done by practitioner of machine learning while dealing with finance data. through this blogs i am trying to summarize my learning from this book.

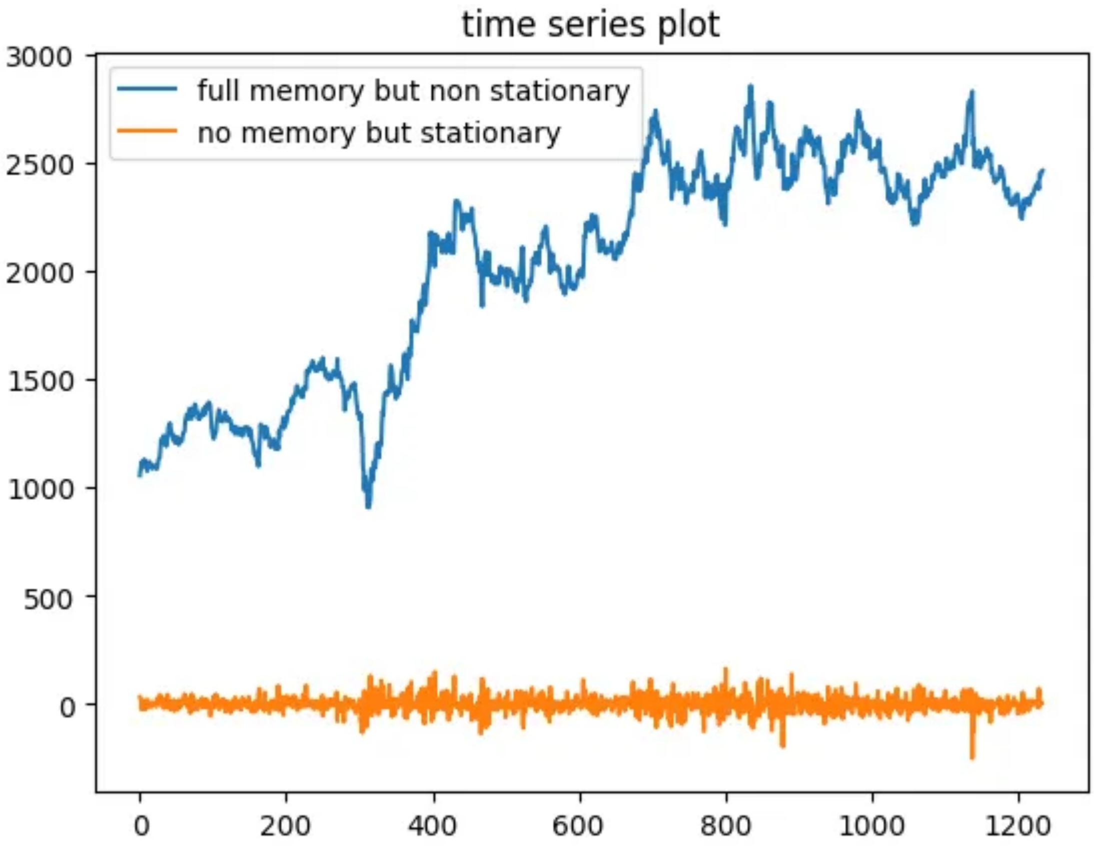
1. **Reason 1 : Memory vs stationarity trade off :**

let's say we want to predict the price of any stock. as we know all the traditional method like ARMA works on stationary data. loosely speaking stationary data mean constant mean and constant variance throughout the series. in order to make the data stationary we perform differencing in time series data. differencing of order 1 means we subtract the current value of stock from previous value.

$$X_{t+1} - X_t$$

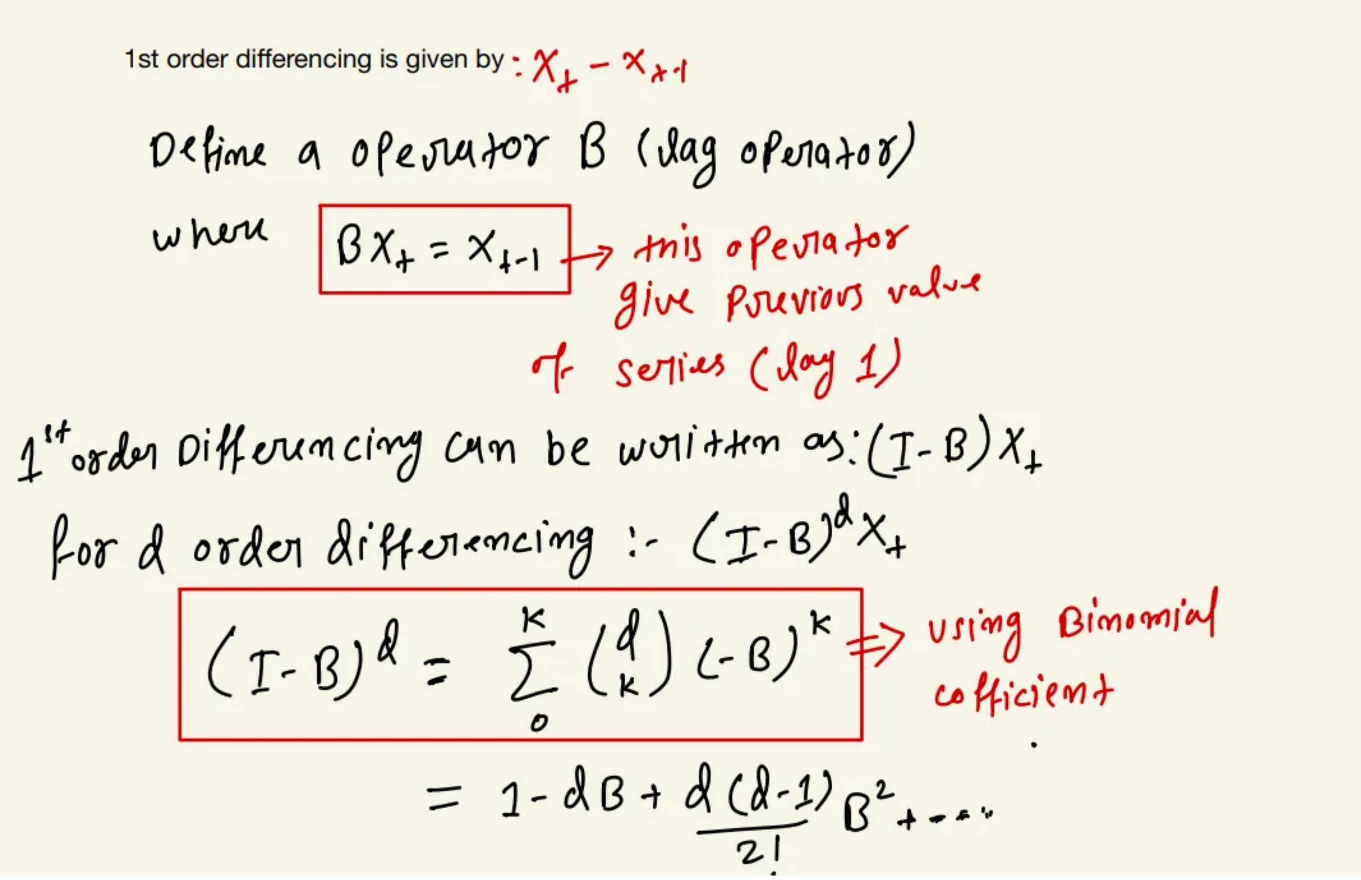
1st order differencing

differencing will make data stationary but when we are doing 1st order differencing then we loss all the historical pattern of data (see fig. 01) and hence loose it's memory. and memory is the predictive power of any model so while making data stationary we end up losing it's memory. this is very common mistake done across all the academic paper and by practitioner. so is there any way to make data stationarity while losings no information? answer is no. so here we have a trade off more stationary we want to make we have to lose more memory. let's understand this case with help of historical price of reliance.



in this plot we can see when we make data completely stationary we loose all the information about how the price moved in time?

as you can see from above plot when we make the data stationary we loose all the information about how price moved in time and hence we loose it's memory. and then we loose the predictive power of any model made with such data. so what we need to do? we need to find a way to achieve partial stationarity while loosing partial memory. then the concept of fractional differencing comes into picture. in fractional differencing instead of doing differencing with order 1 we do the differencing with some fraction value which will not loose all the information. so how to perform fractional differencing? let’s look into this:



1st order differencing is given by : $X_t - X_{t-1}$

Define a operator B (lag operator)

where $\boxed{BX_t = X_{t-1}}$ → this operator give previous value of series (lag 1)

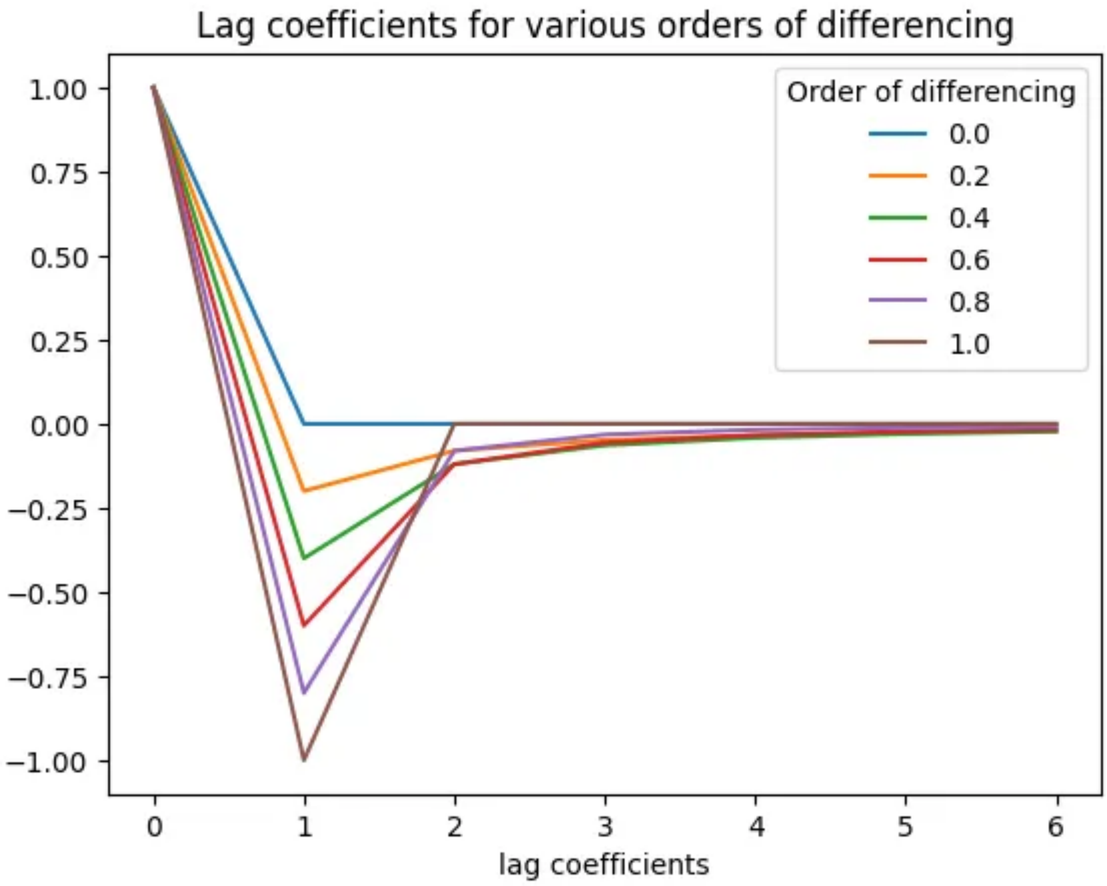
1st order Differencing can be written as: $(I-B)X_t$

for d order differencing :- $(I-B)^d X_t$

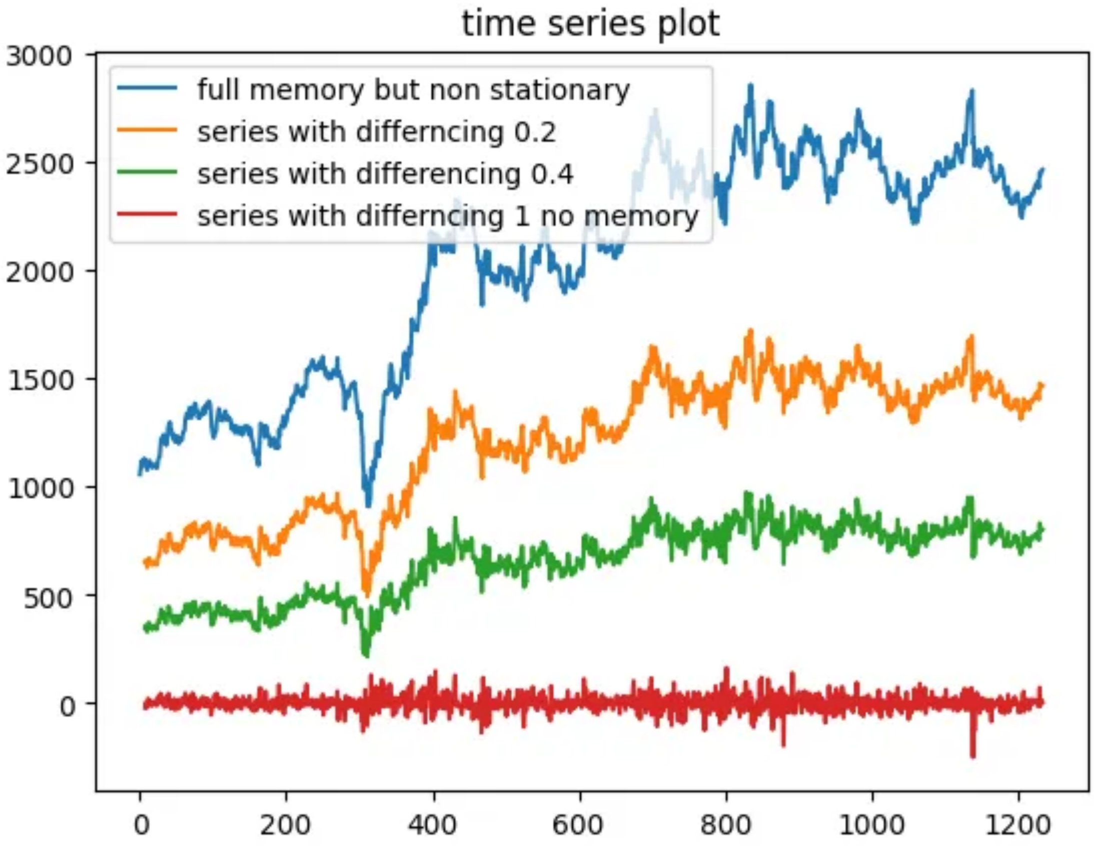
$$\boxed{(I-B)^d = \sum_{0}^{k} \binom{d}{k} (-B)^k} \Rightarrow$$ using Binomial cofficient

$$= 1 - dB + \frac{d(d-1)}{2!} B^2 + \dots$$

as you can see we expand d order differencing. now if we put the value of d=0.3 then it will give the value of fractional differencing of 0.2. now this is an infinite series so we can we take the value till some point and truncate the series after that because higher **B^n** coefficient value will be close to zero. below graph show that coefficient of **B^n** for different value of d after a point will become close to zero.

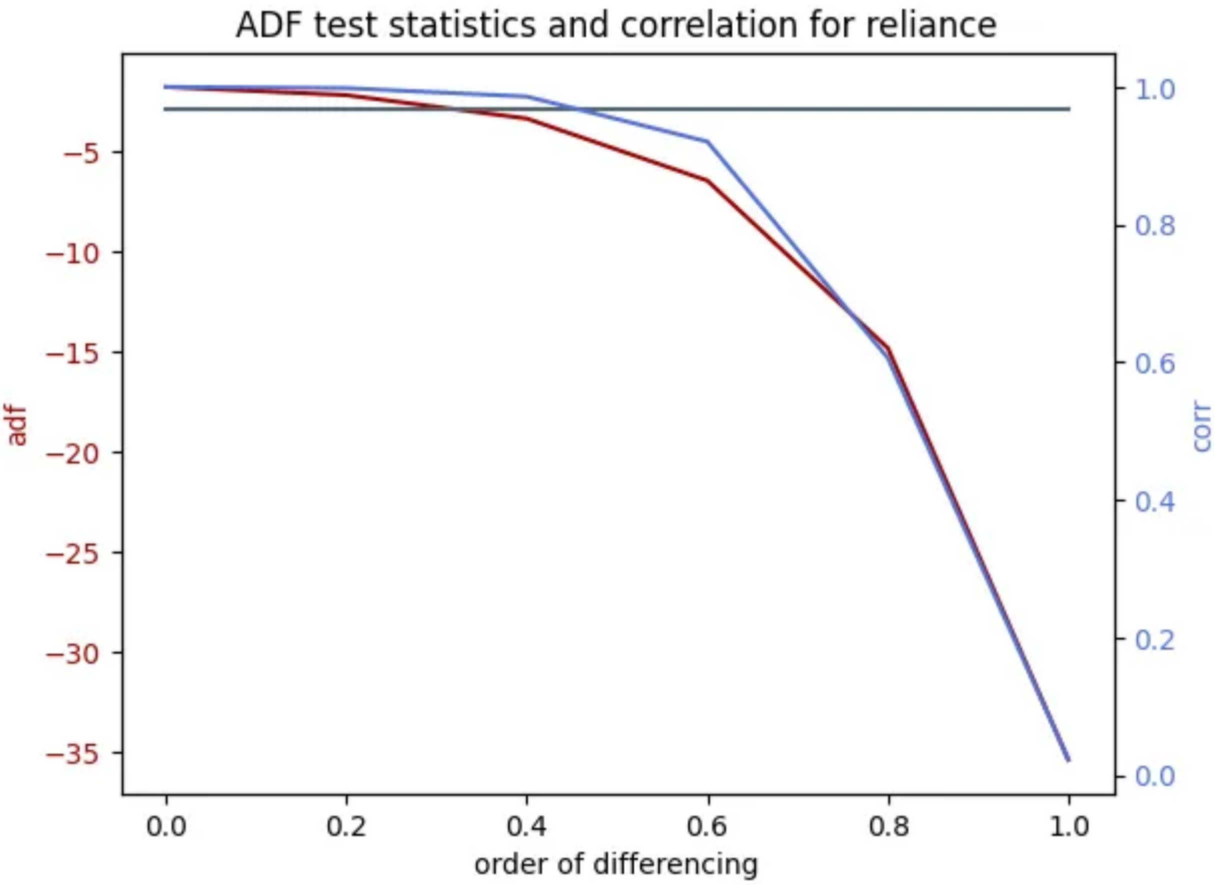


now let's look at the same plot with different fractional order of differencing.



above chart show that as we are increasing the order of differencing the we are loosing more and more memory and becoming more stationary. with differencing of 0 we have all the memory but series is not stationary and with differencing of 1 series has no memory but series is completely stationary. so we need to do a trade off b/w loosing some memory and while keeping the data stationary with significant confidence value. so now how to find that value of d?

ADF test for stationary come into picture. ADF test is used to test if the series is stationary or not. After performing the ADF test for different value of d, below red line plot is test statistics value with different value of d. if value of test statistics are less than the horizontal line then we can say that series is stationary. so from below chart we can say that d=0.4 is sufficient to make series stationary. so we can take d=0.4 to make series stationary while loosing minimum memory from the series.



2. Reason 2 : inefficient sampling

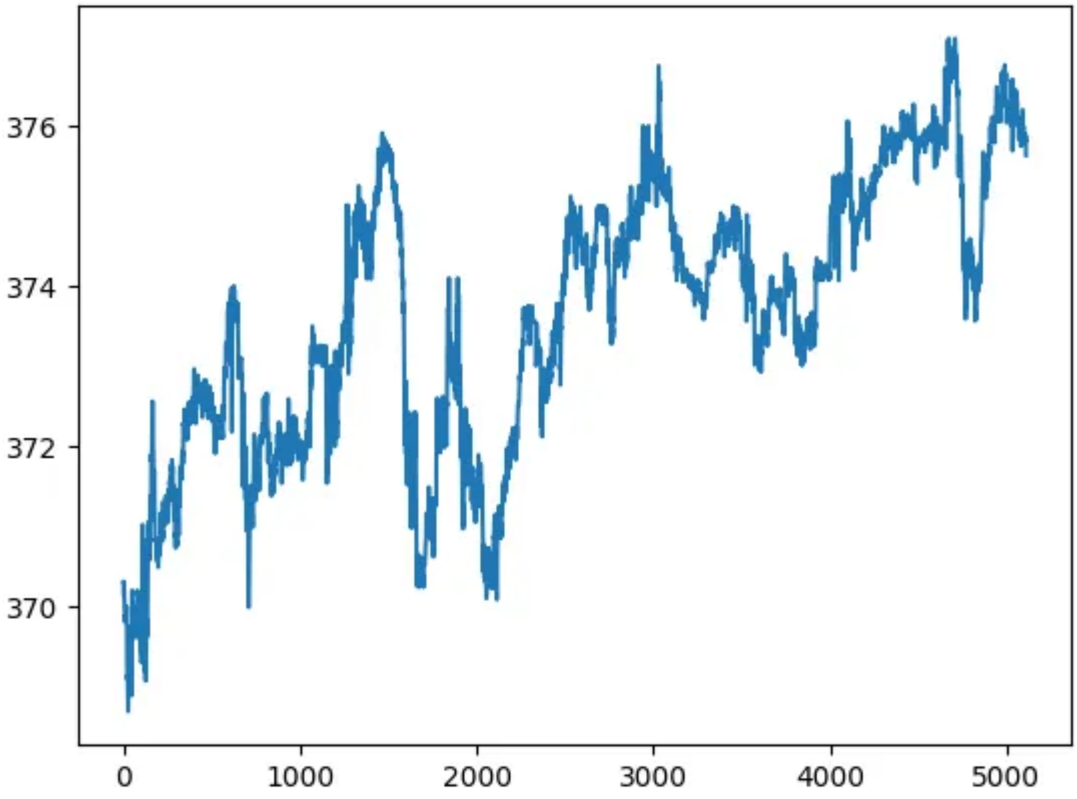
Another common mistake done by many practitioner and academic paper is of inefficient sampling of data. mostly they sample the data on every time interval. for example they sample the data each 5 min or 10 min. these are the major problem when sampling the data based on time frame.

- since market doesn't process information on regular time interval like the market is more active when it open compare to noon so it undersample the information during high activity time and oversample the information during low activity time.
- time sampled series exhibits poor statistical properties like serial correlation, heteroscedasticity, and non normality of returns.

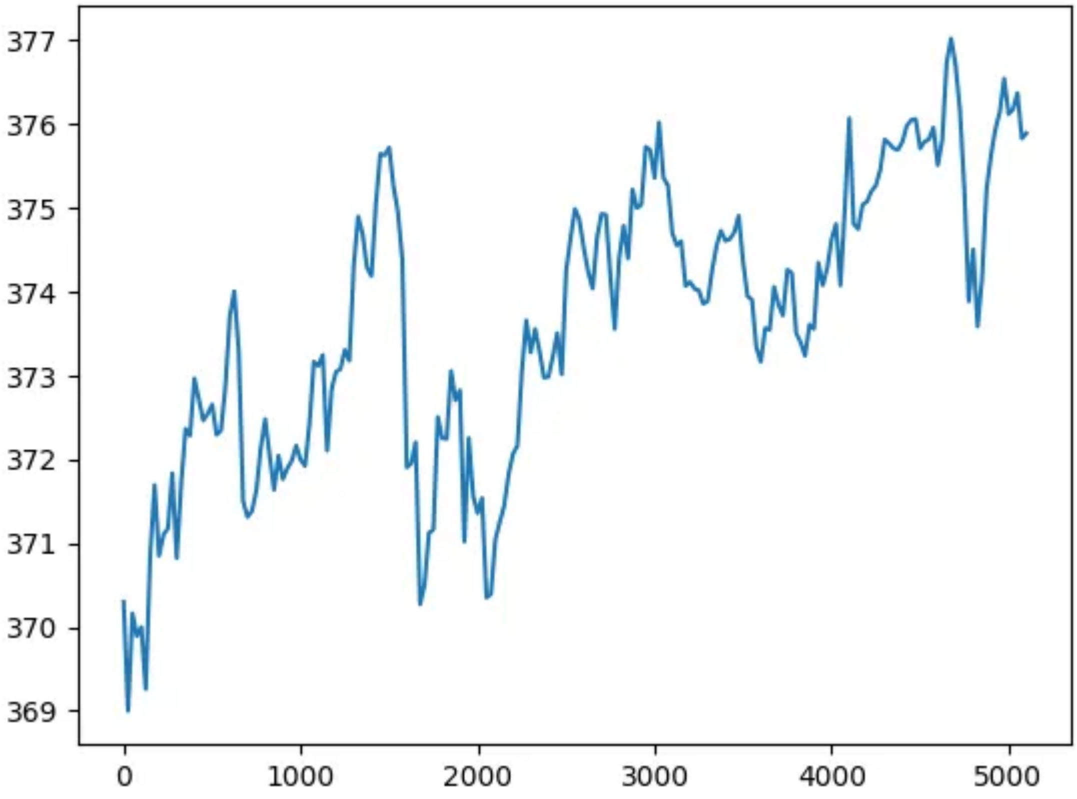
we will see next to overcome this issue different bars are being defined

- **Tick bar:** all the variable like timestamp, volume, open price, close e.g is extracted after a finite no of transactions take place. for example we sample all variable after 1000 transactions. Mandelbrot and Taylor [1967] were among the first to realize that sampling as a function of the number of transactions exhibited desirable statistical properties: "Price changes over a fixed number of transactions may have a Gaussian distribution.

> *" one important thing to take care of while constructing any bar is of outliers because many exchange for some period of time while opening of market and closing of market accumulates bid and ask offer in orderbook without matching them. so you can see a outlier behavior for some time while opening and closing of market"*



Time bar plot of a stock



Tick bar of same stock (we can see the plot is more smooth compare to time bar)

- **Volume bar :** tick bar has a issue of order fragmentation. which means let's say there we have 10 no of share for ask at some price and some one buy 10 share. then it is recorded as 1 tick but let's say if you buy 1 share 10 times then it will counted as 10 transactions. for solving this issue we sample the information after a particular volume of transaction happens. this is known as volume bar.

- **Dollar bar** : dollar bar are formed by sampling the information after a particular value of transaction happens. for example we sample the information like timestamp, open price, close price after 5000$ worth transaction happened. value doesn't necessarily mean to be in $ it can be in Rs, euro etc. why we need dollar bar? suppose that we wish to analyze a stock that has exhibited an appreciation of 100% over a certain period of time. Selling $1,000 worth of that stock at the end of the period requires trading half the number of shares it took to buy $1,000 worth of that stock at the beginning. In other words, the number of shares traded is a function of the actual value exchanged. Therefore, it makes sense sampling bars in terms of dollar value exchanged, rather than ticks or volume, particularly when the analysis involves significant price fluctuations. another argument is the no of share often change with bonus share, share split so it make more sense to sample based on price than on volume.

in next part we will be going to talk about the some advance information driven bar. it means we sample the information if some new information come into market. we will look into some of the information driven bar in next part of this blog series.

don't forget to share your thought or give a thumbs up if you like this blog.

Reference:

1) Advance in finance machine learning by Marcos López de Prado

Machine Learning Finance AI Stock Market Economics

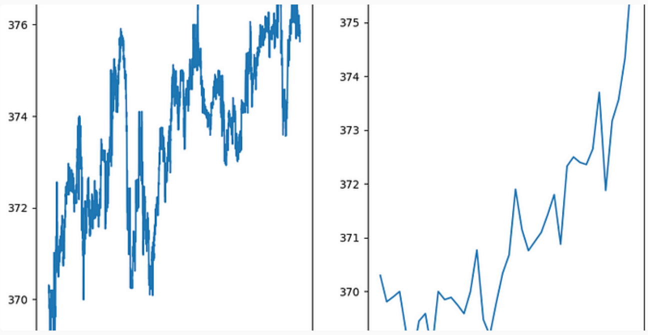
Written by Ved Prakash

Follow

382 Followers

Data Scientist at Michelin

More from Ved Prakash



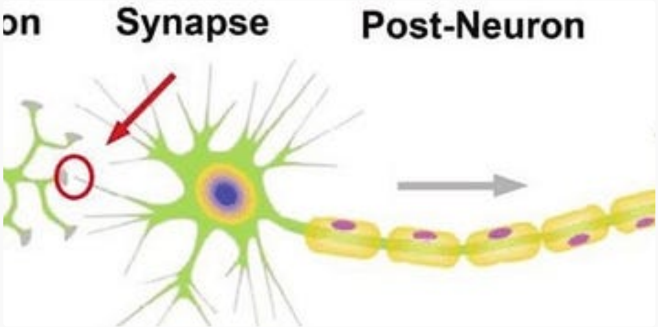
Ved Prakash

Major reasons why machine learning fails in stock prediction :...

This is continuation of previous article where I have discussed some of the reason why...

6 min read · Feb 17, 2024

147 5



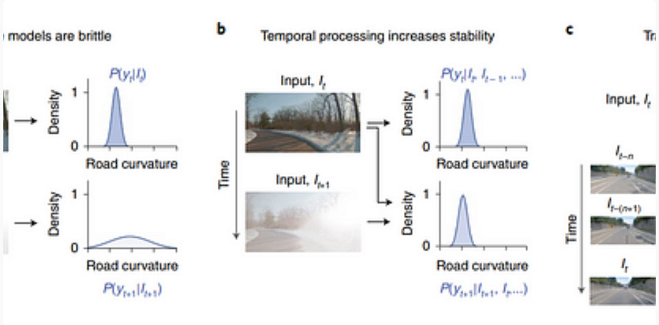
Ved Prakash

Liquid Neural Network : A adaptive way to train ML model

This is my 1st article in series of articles where I will review different research papers from t...

8 min read · Aug 7, 2023

180 2



Ved Prakash

Neural Circuit Policy : training a autonomous vehicles using mode…

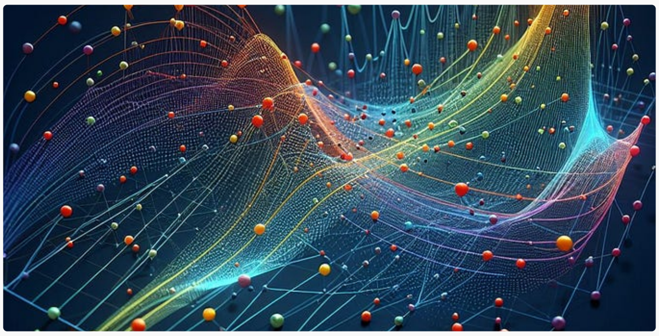
This is my 2nd article in series of articles where I will review the different research…

9 min read · Sep 16, 2023

56 1

See all from Ved Prakash

Recommended from Medium



Tim Sumner in Towards Data Science

A New Coefficient of Correlation

What if you were told there exists a new way to measure the relationship between two…

10 min read · Mar 31, 2024

1.3K 26

Somnath Singh in Level Up Coding

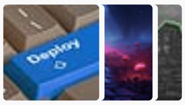
The Era of High-Paying Tech Jobs is Over

The Death of Tech Jobs.

✦ · 14 min read · Apr 1, 2024

4.3K 129

Lists

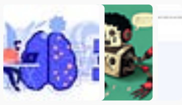


Predictive Modeling w/ Python

20 stories · 1079 saves

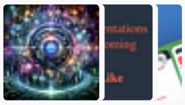
Practical Guides to Machine Learning

10 stories · 1302 saves



The New Chatbots: ChatGPT, Bard, and Beyond

12 stories · 353 saves



Natural Language Processing

1361 stories · 849 saves

Nikhil Adithyan in DataDrivenInvestor

Creating a Scalping Strategy in Python with a 74% Win Rate

Comprehensive backtesting of an unconventional trading strategy

12 min read · Apr 2, 2024

304 7

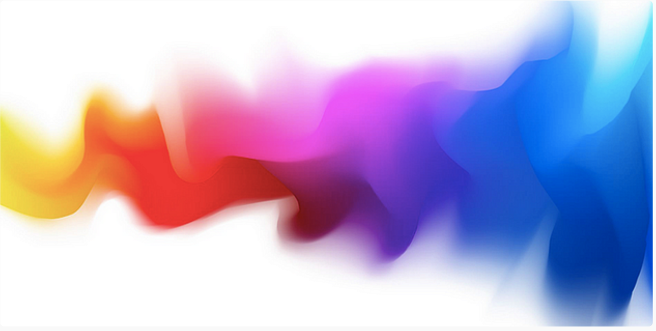
Boris Meinardus in Towards AI

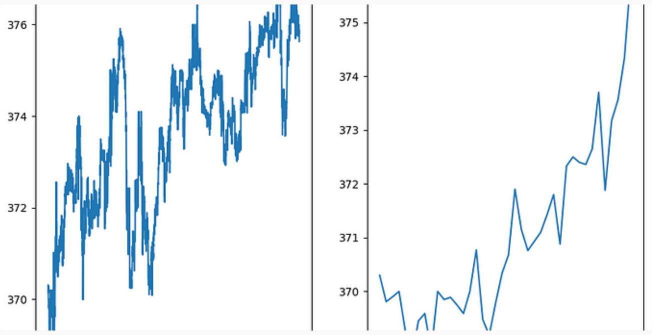
Machine Learning Was Hard Until I Learned These 5 Secrets!

The secrets no one tells you but make learning ML a lot easier and enjoyable.

✦ · 10 min read · Mar 29, 2024

1.5K 16





Chris Kuo/Dr. Dataman

Quantile Regression for Time Series Probabilistic Forecasting

We often hear “the only thing that is certain is that nothing is certain.” We do not like…

✦ · 6 min read · 5 days ago

260 2

Ved Prakash

Major reasons why machine learning fails in stock prediction :…

This is continuation of previous article where I have discussed some of the reason why…

6 min read · Feb 17, 2024

147 5

See more recommendations