Machine learning in Finance

the “deep web” of data-driven business

About me

Practitioner

~7 years developing
data-driven products

Entrepreneur

building a distributed
consulting company

Educator

blogging, workshops,
teaching at universities

About you?

- …

what this workshop is not about

machine learning algorithms
deep stochastic mathematics
making bots for crypto trading
easy ways to hack the market

what this workshop is about

how to doubt everything in your data modeling activities
from now and forever

Plan

- What people in Finance do and what AI in Finance can do
- **Practice 1-2**. “Classical” ML in Finance. Fixing mistakes
- Why ML in Finance is different from “normal ML”
- **Practice 3**. Fixing mistakes again
- Reflexion and deep disappointment
- **Practice 4**. Fixing some more mistakes

People in Finance

	Front Office	Back Office
Buy Side	**Asset management** at a big bank. **Hedge fund** (strategies constrained to prospectus) **Prop trading** (fastest moving) *Matlab, Java, Functional Languages.* 70-100k+large bonus	**Data scraping and maintenance, Execution, Server administration** *Bash, SQL, SVN, Linux, C++.* 90-100k+small bonus
Sell Side	**Sales & Trading** at a big bank (taking & executing orders, creating derivatives by client reques, execution algos) *Excel.* 70-80k+medium bonus	**Technology**, **Operations**, or **Risk Management** at a big bank (hard to transition to front office) *C++, internal language, hacky code.* 90-100k+small bonus

http://isomorphisms.sdf.org/maxdama.pdf

Applied directions

Banking	Asset Management	Trading
Retail operations	Representation learning	Price forecasting
P2P operations	Portfolio optimization	Optimal execution
Lending, credit scoring	Risk management	Optimal strategy development

Trading roles

Analysts

Data curators

Feature analysts

Strategists

Backtesters

Deployment team

Portfolio managers

AI in Finance

Supervised learning: regression and classification

Unsupervised learning: clustering and representation learning

Reinforcement learning: optimal strategy, “learning” from the expert

Regression:
earnings prediction
credit loss forecasting
price forecasting

Classification:
rating prediction
default modeling
credit card fraud
anti-money laundering

Clustering:
customer segmentation
stock segmentation

Representation learning:
factor modeling
de-noising
regime change detection

Banking data: clients information, transactions…

Fundamental data: assets, liabilities, sales, earnings…

Market data: price, volume, dividends, open interest…

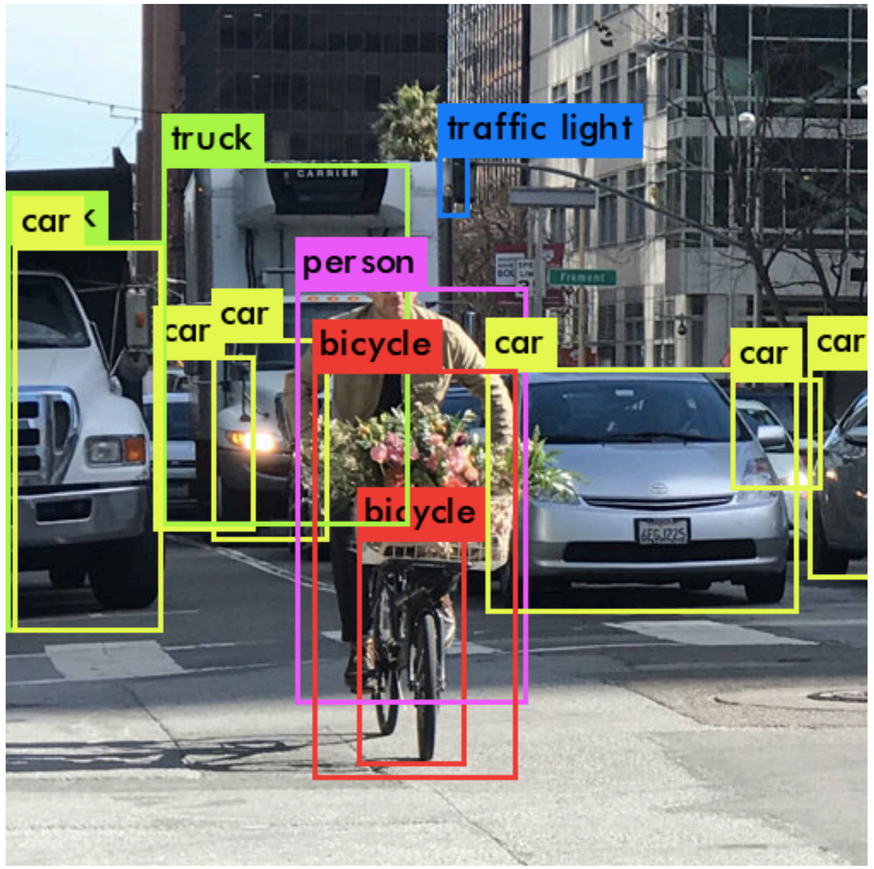
Analytics data: recommendations, credit rankings, new sentiment…

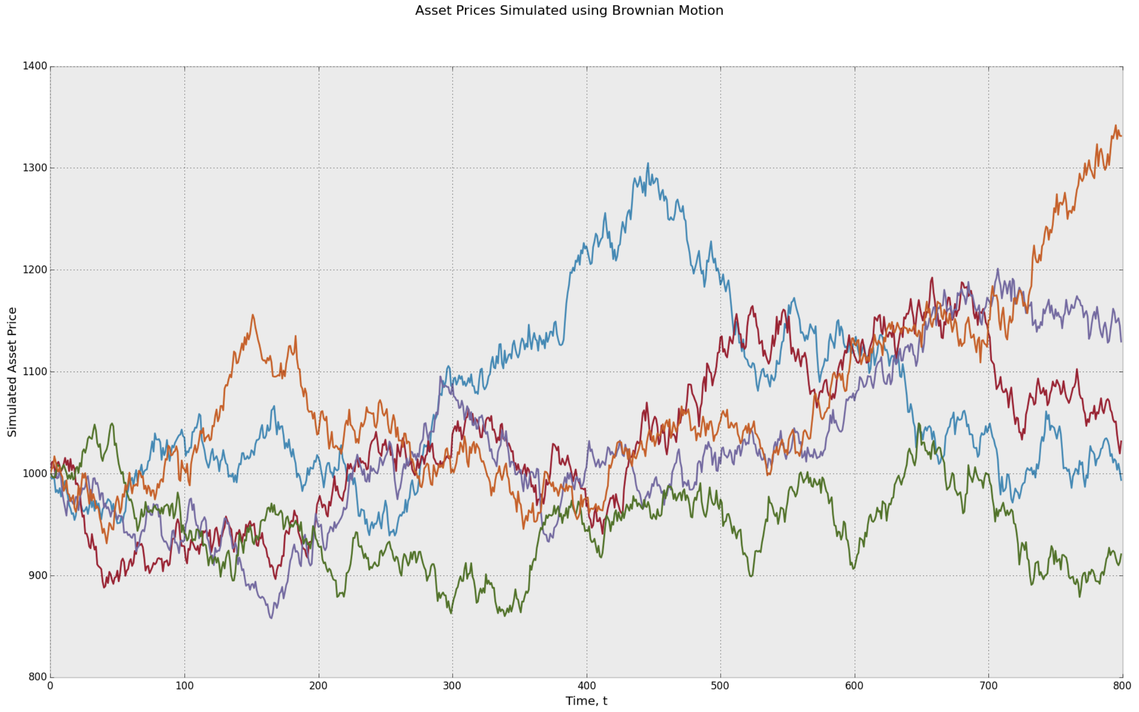
Alternative data: CCTV, satellite images, Twitter…

Practice time #1

Finance ML vs “Normal” ML

Treating it as a regular ML exercise



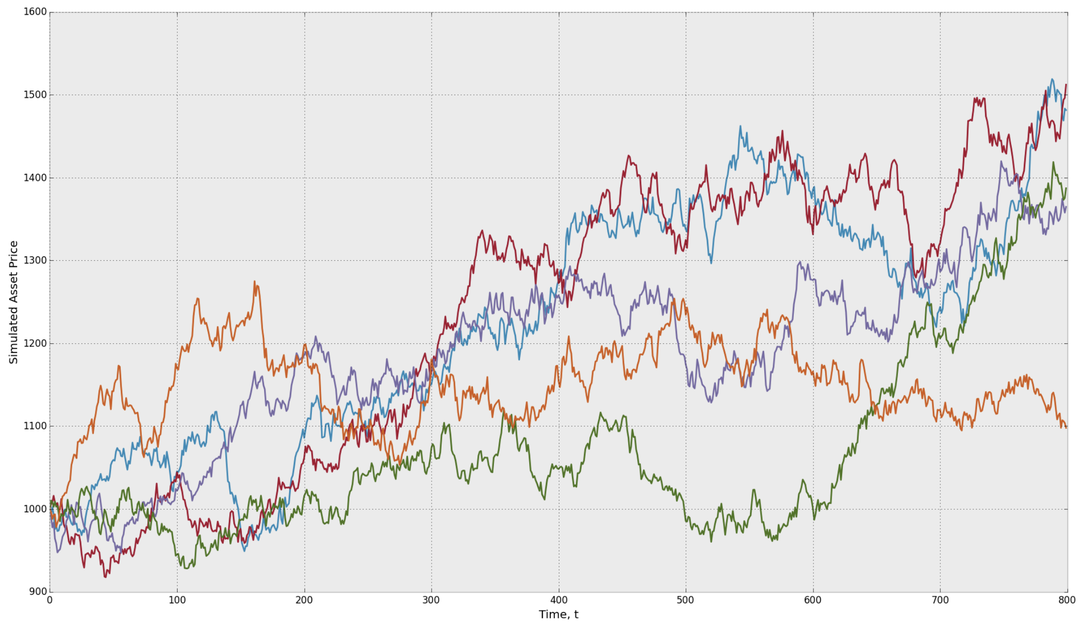


http://www.turingfinance.com/random-walks-down-wall-street-stochastic-processes-in-python/

Treating it as a regular ML exercise

$$dS_t = \mu S_t dt + \sigma dS_t W_t$$

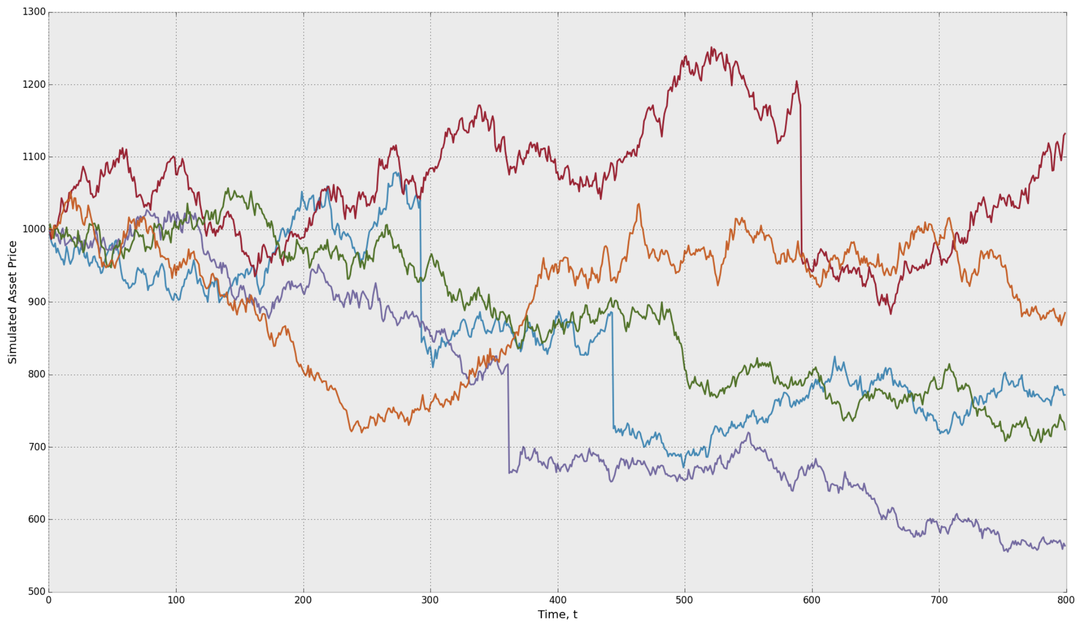
Asset Prices Simulated using Geometric Brownian Motion



$$dS_t = \mu S_t dt + \sigma S_t dW_t + dJ_t$$

$$dJ_t = S_t d\left(\sum_{i=0}^{N_t} (Y_i - 1) \right)$$

Asset Prices Simulated using the Merton Jump Diffusion Model



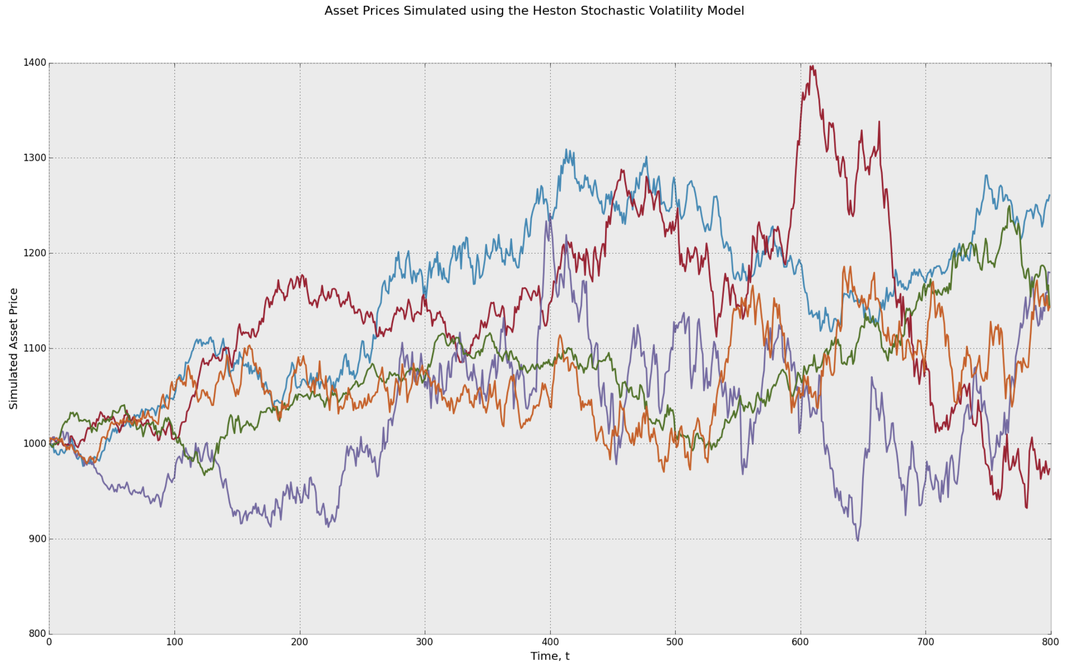
http://www.turingfinance.com/random-walks-down-wall-street-stochastic-processes-in-python/

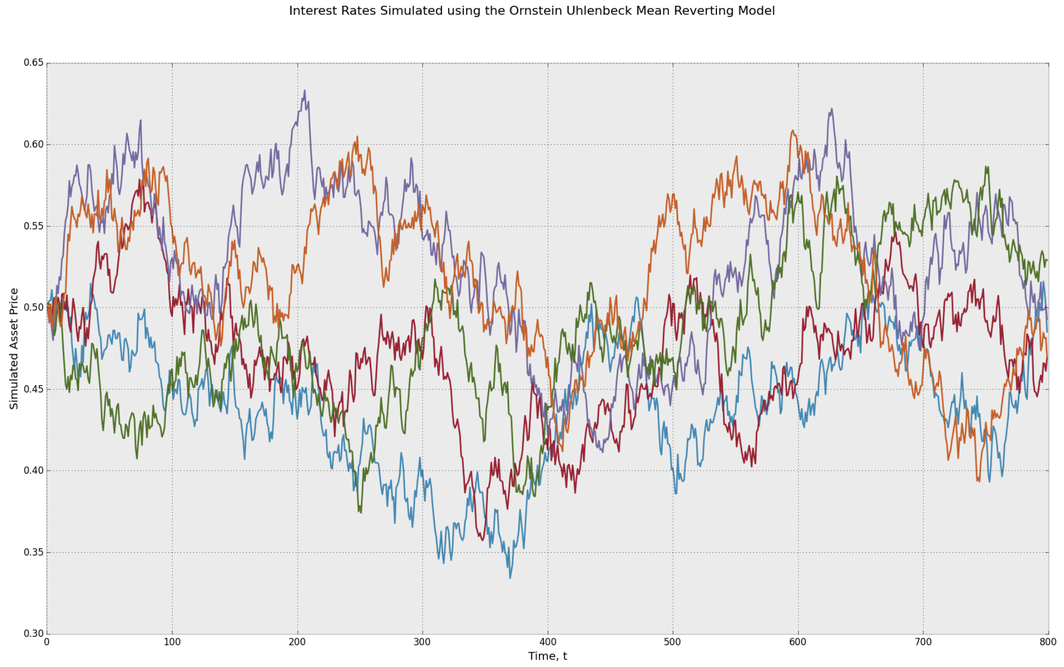
Treating it as a regular ML exercise

$$dS_t = \mu S_t dt + \sqrt{v_t} S_t dW_t^S \text{ where}$$

$$dv_t = a(b - v_t)dt + \sigma\sqrt{v_t} dW_t^v \text{ (Cox Ingersoll Ross),}$$

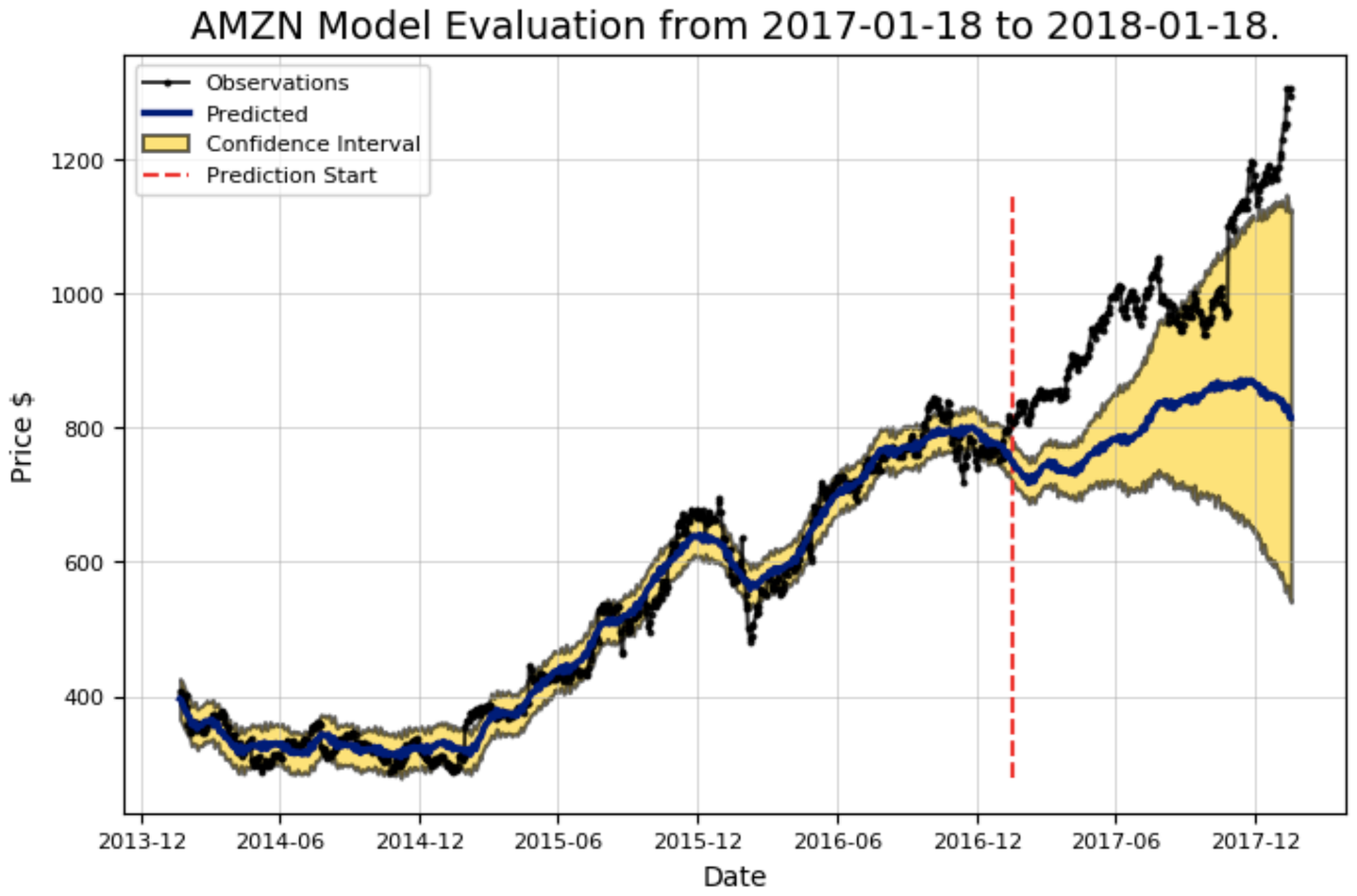
$$dr_t = a(b - r_t)dt + \sigma\sqrt{r_t} dW_t$$





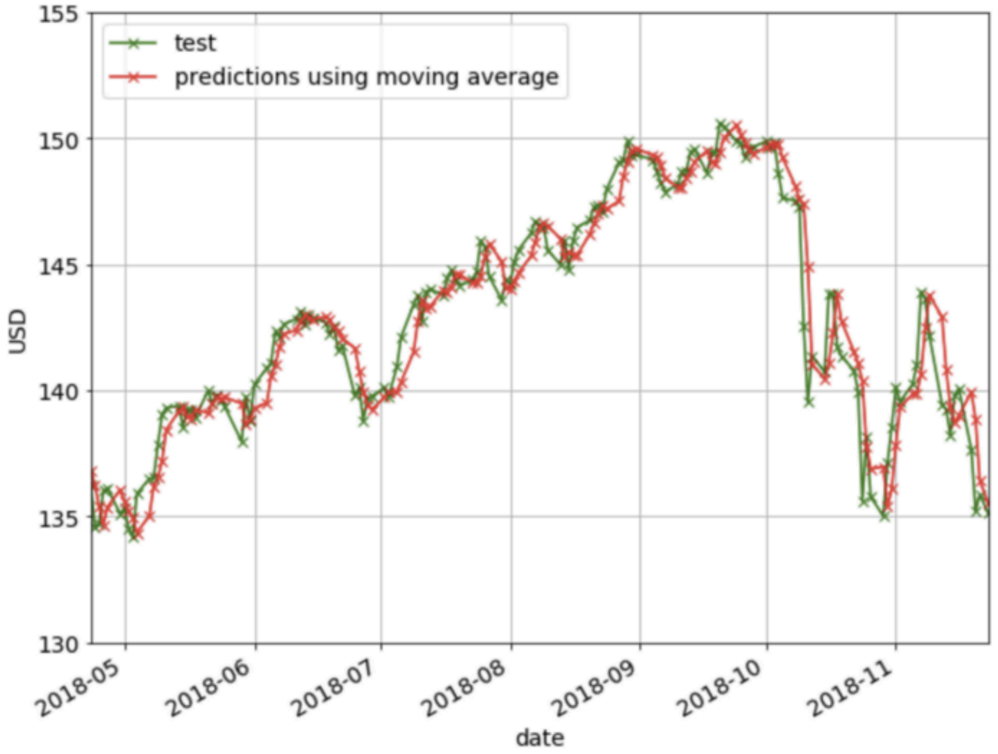
http://www.turingfinance.com/random-walks-down-wall-street-stochastic-processes-in-python/

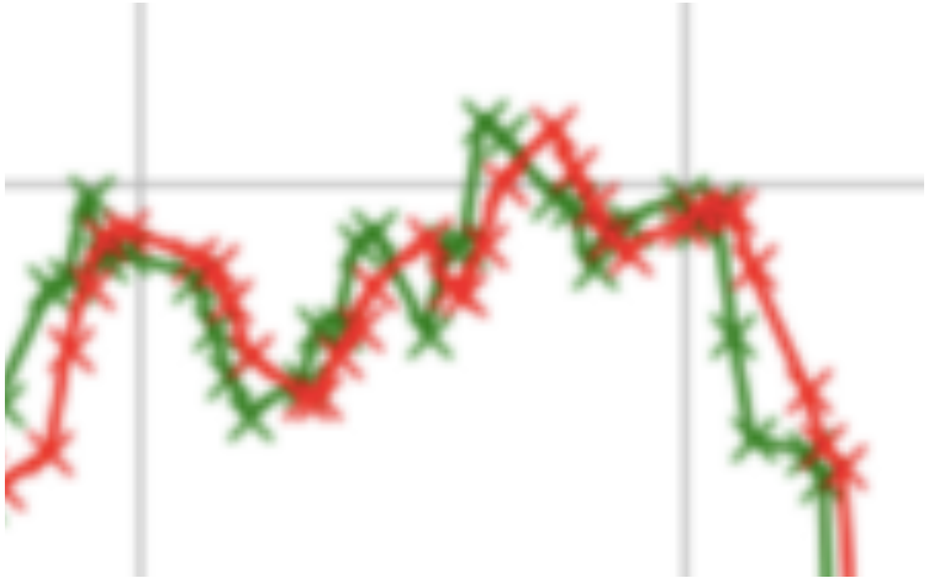
Forecasting the price in the future



https://towardsdatascience.com/stock-prediction-in-python-b66555171a2

Forecasting the price in the future

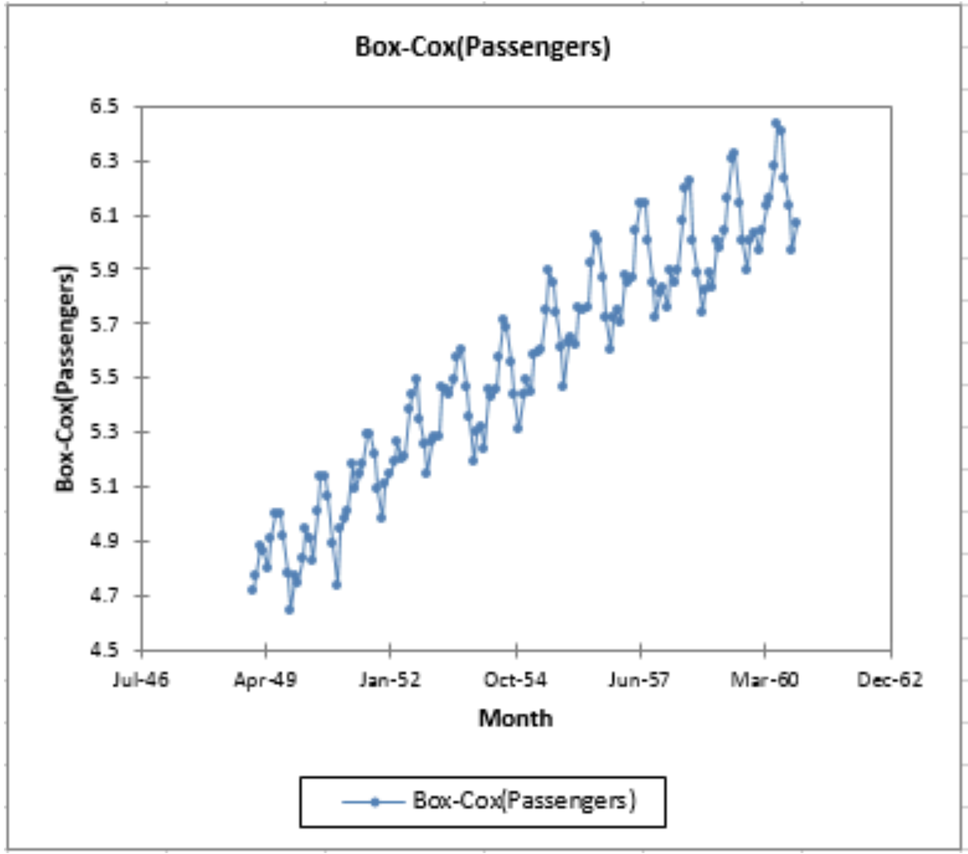


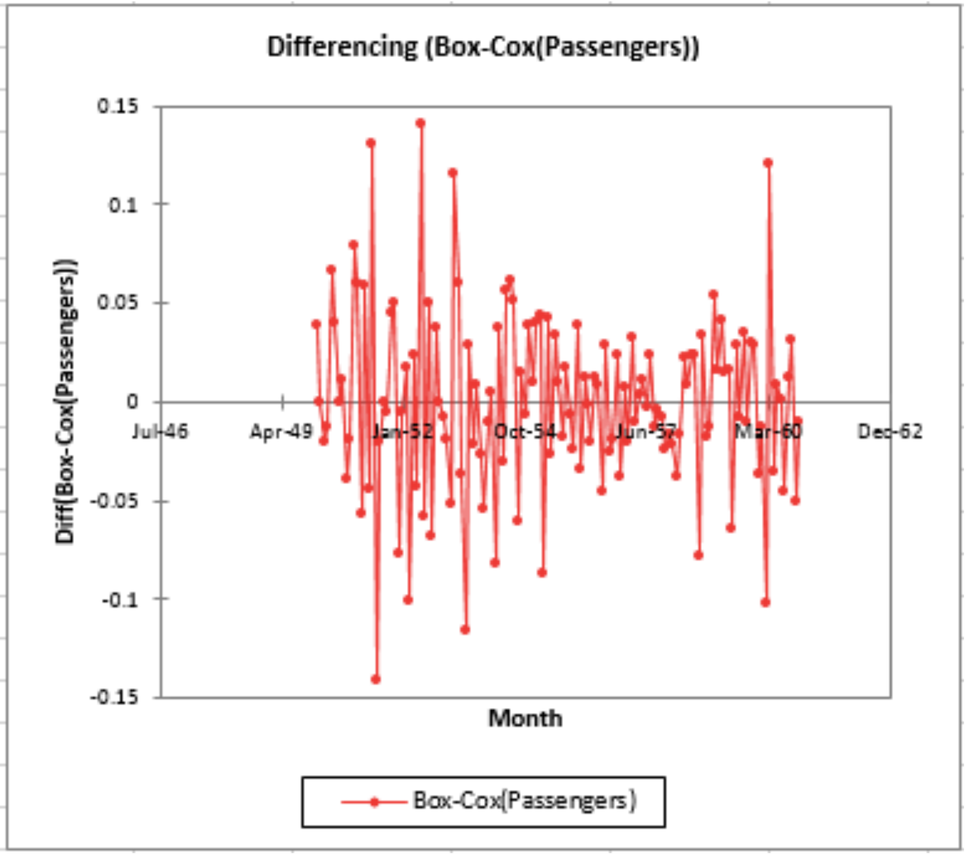


https://towardsdatascience.com/machine-learning-techniques-applied-to-stock-price-prediction-6c1994da8001

Practice time #2

Treating it as a regular ML exercise





Treating it as a regular ML exercise

Difference operator [edit]

Main article: Finite difference

In time series analysis, the first difference operator :∇

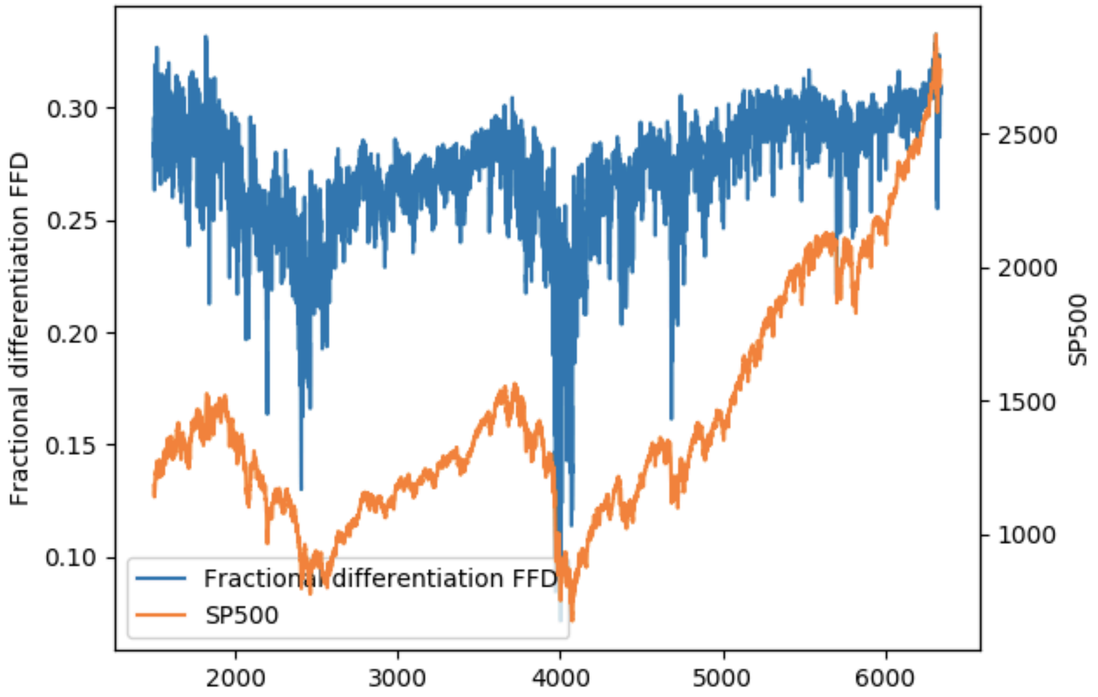
$$\begin{aligned}\nabla X_t &= X_t - X_{t-1} \\ \nabla X_t &= (1 - L)X_t \ .\end{aligned}$$

Similarly, the second difference operator works as follows:

$$\begin{aligned}\nabla(\nabla X_t) &= \nabla X_t - \nabla X_{t-1} \\ \nabla^2 X_t &= (1 - L)\nabla X_t \\ \nabla^2 X_t &= (1 - L)(1 - L)X_t \\ \nabla^2 X_t &= (1 - L)^2 X_t \ .\end{aligned}$$

The above approach generalises to the *i*-th difference operator $\nabla^i X_t = (1 - L)^i X_t$.

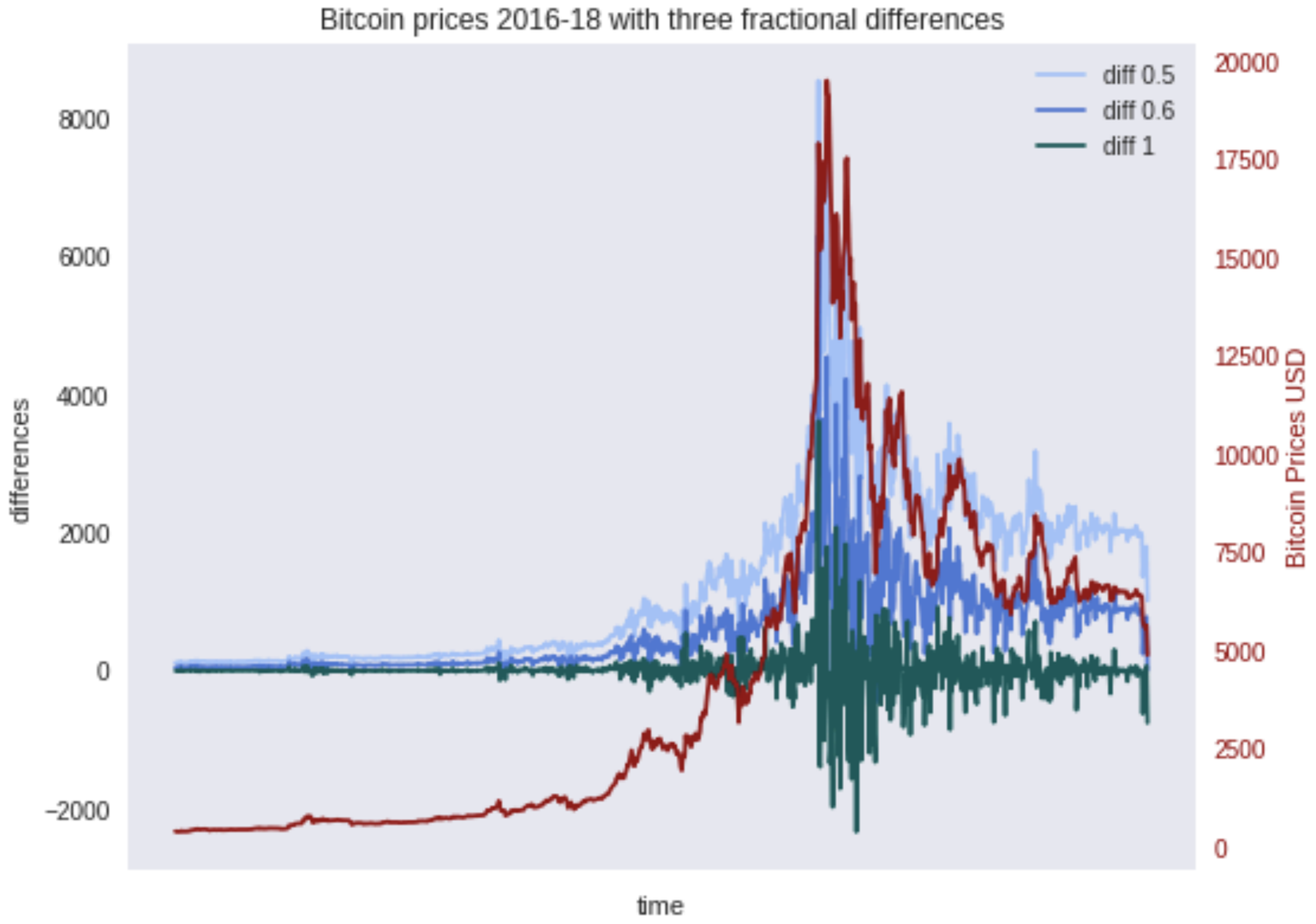
Treating it as a regular ML exercise



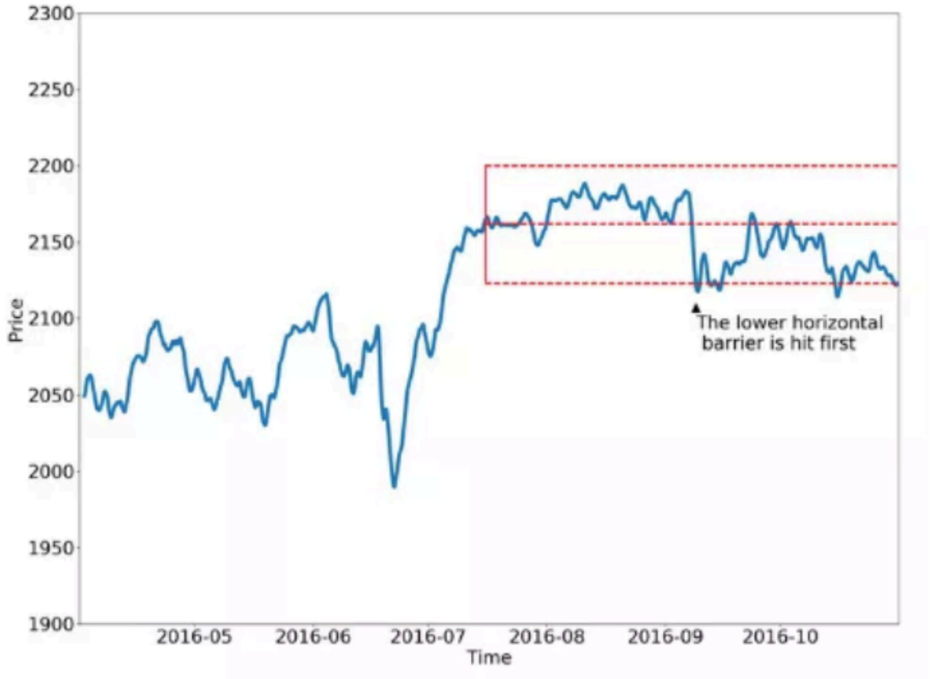
$$(\mathrm{I}-\mathrm{B})\boldsymbol{X}_t = \boldsymbol{X}_t - \boldsymbol{B}\boldsymbol{X}_t = \boldsymbol{X}_t - \boldsymbol{X}_{t-1}$$

$$\begin{aligned}(1-B)^d &= \sum_{k=0}^{\infty} \binom{d}{k} (-B)^k \\ &= \sum_{k=0}^{\infty} \frac{\prod_{a=0}^{k-1}(d-a)\,(-B)^k}{k!} \\ &= 1 - dB + \frac{d(d-1)}{2!}B^2 - \cdots .\end{aligned}$$

Treating it as a regular ML exercise



Forecasting the price in the future



Backtesting is not a research tool

SNIPPET 8.1 MARCOS' FIRST LAW OF BACKTESTING—IGNORE AT YOUR OWN PERIL

"Backtesting is not a research tool. Feature importance is."

—Marcos López de Prado

Advances in Financial Machine Learning (2018)

SNIPPET 11.1 MARCOS' SECOND LAW OF BACKTESTING

"Backtesting while researching is like drinking and driving.
Do not research under the influence of a backtest."

—Marcos López de Prado

Advances in Financial Machine Learning (2018)

Practice time #3

EDA: correlations won't work

Sampling: all this time we were violating the IID hypothesis

Validation: all this time we even didn't try cross-validation

Feature Importance: all this time there were 5 random features in the dataset

Backtesting: all this time we were testing on one random sample from big universe

Probabilistic interpretation: what to do with that Random Forest?

Hyper-parameters: … there are some specialties too :)

EDA: correlations won't work

Sampling: all this time we were violating the IID hypothesis

Validation: all this time we even didn't try cross-validation

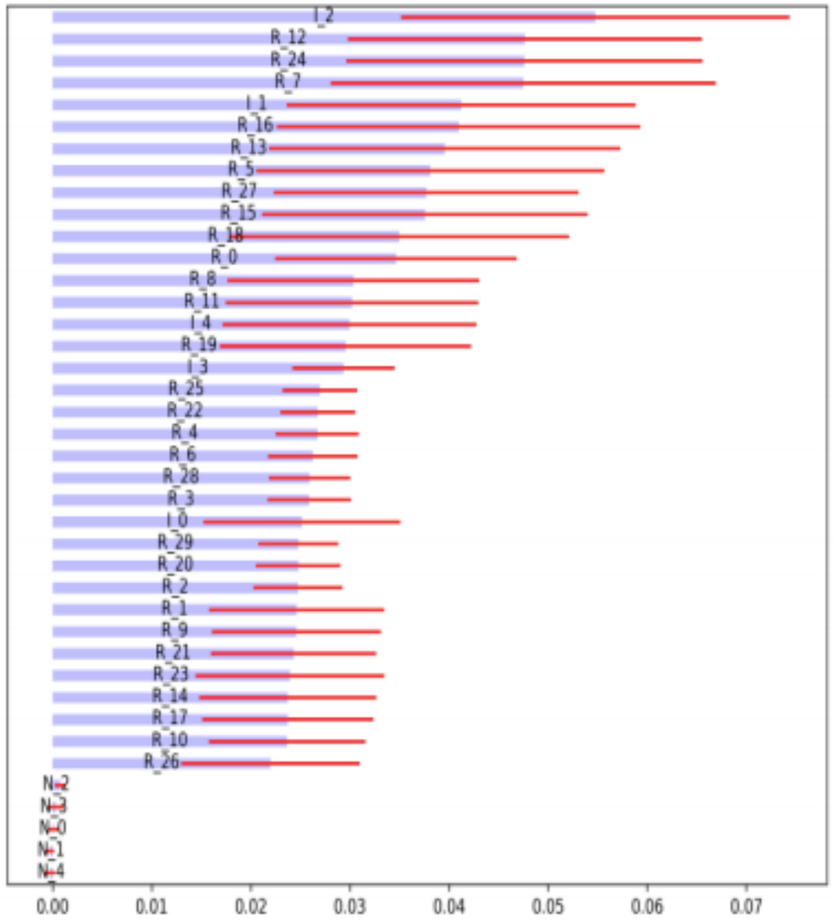
Feature Importance: all this time there were 5 random features in the dataset

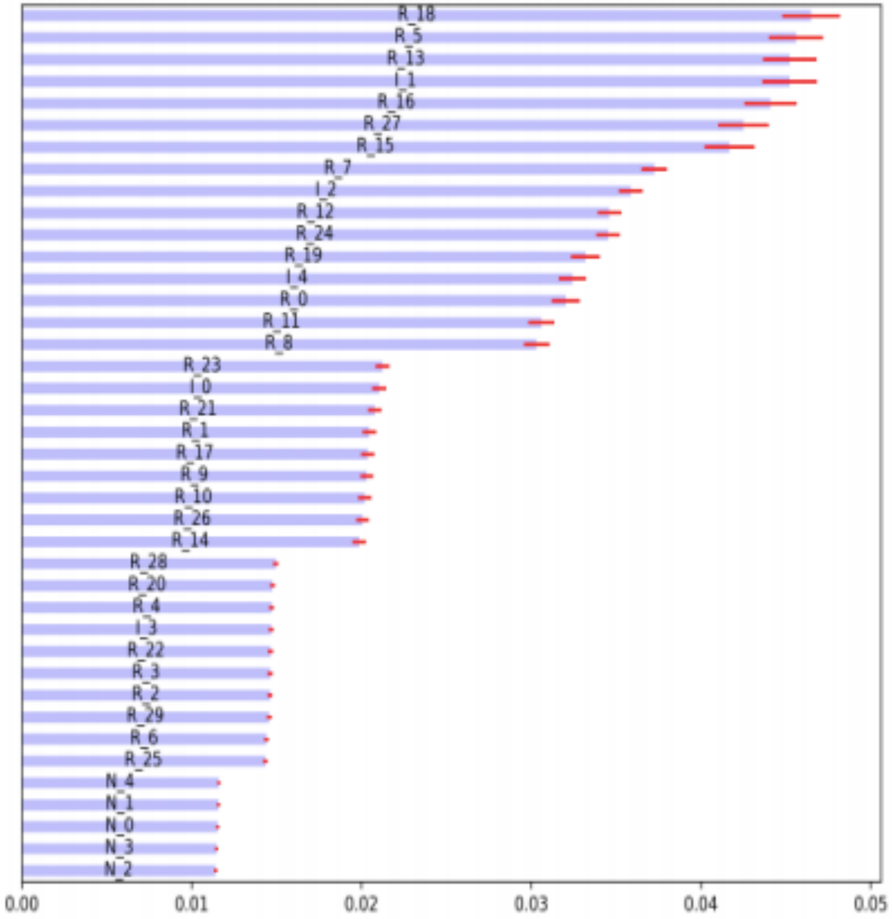
Backtesting: all this time we were testing on one random sample from big universe

Probabilistic interpretation: what to do with that Random Forest?

Hyper-parameters: … there are some specialties too :)

Feature Importance is a Research tool





Practice time #4

Homework

Inputs:

obviously, input data is still not perfect, fix the variables, maybe create new ones

Modeling:

why bagging of linear models works better than boosting of trees?

Feature Importance:

what are the redundant variables? how to detect them?

Literature analysis:

read 5+ papers on ML in stock price prediction, find all the mistakes ;)

GOAL:

to maximize the trust to the ML pipeline itself

HINT:

backtesting is not a research tool!

ML research directions

Simulations: scenario generations for the strategy tests

Investment management: beyond classical optimization, goal-based investments

Factor decomposition: reconstructing real market structure

Clustering: regime switching detection

Alternative data analysis: adding external variables to the game

Materials

Quantopian Lectures

Lectures

Lecture 1	Introduction to Research	A simple tut
Lecture 2	Introduction to Python	Some basic to
Lecture 3	Introduction to NumPy	How to use Nu
Lecture 4	Introduction to pandas	An introductio
Lecture 5	Plotting Data	A brief primer.
Lecture 6	Means	Measures of centrality.
Lecture 7	Variance	Measures of dispersion.

QuantStart

Successful Algorithmic Trading

A step-by-step guide to developing systematic trading strategies using the Python programming language

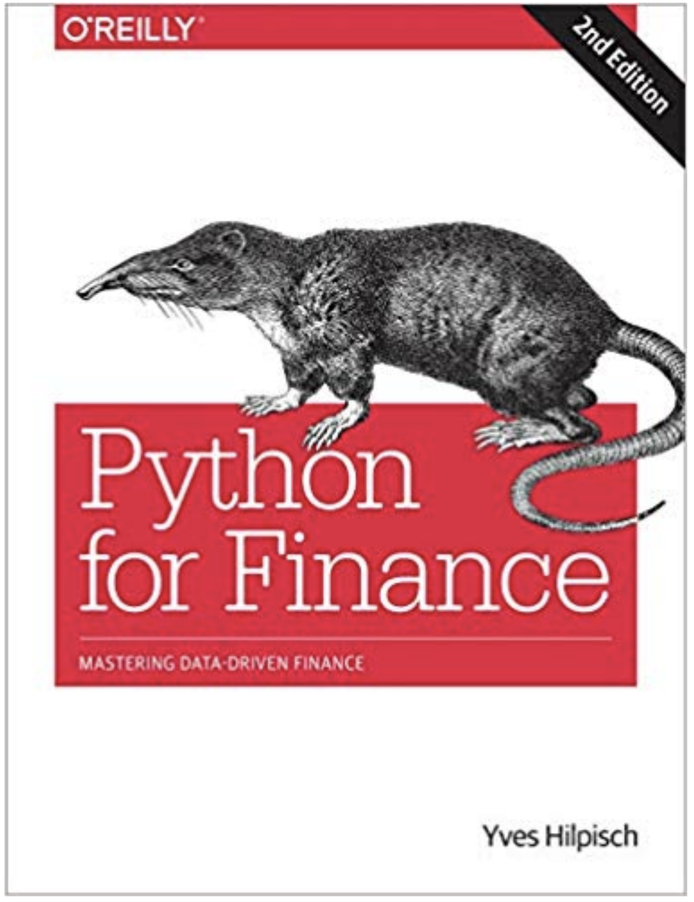
Michael L. Halls-Moore, PhD.

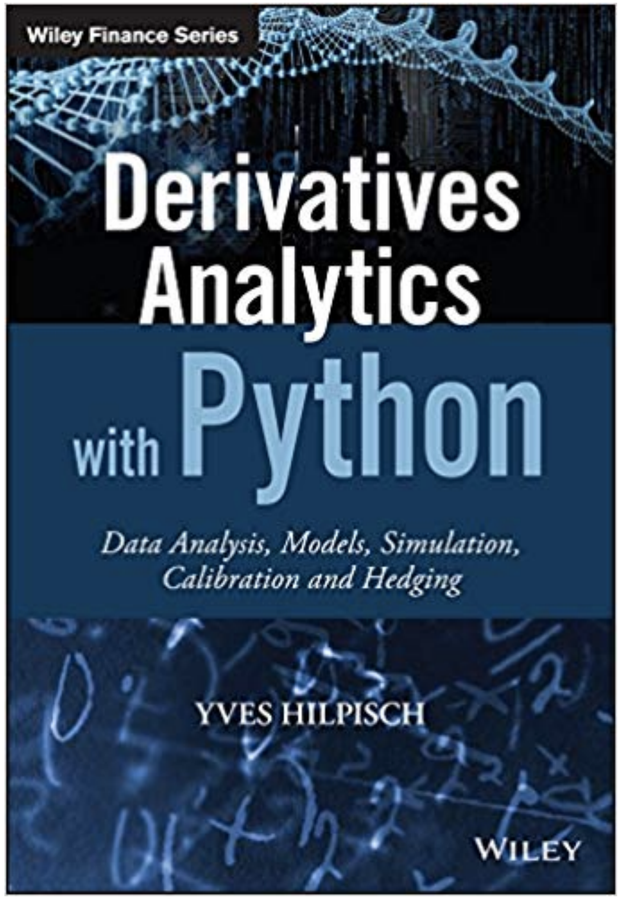
QuantStart

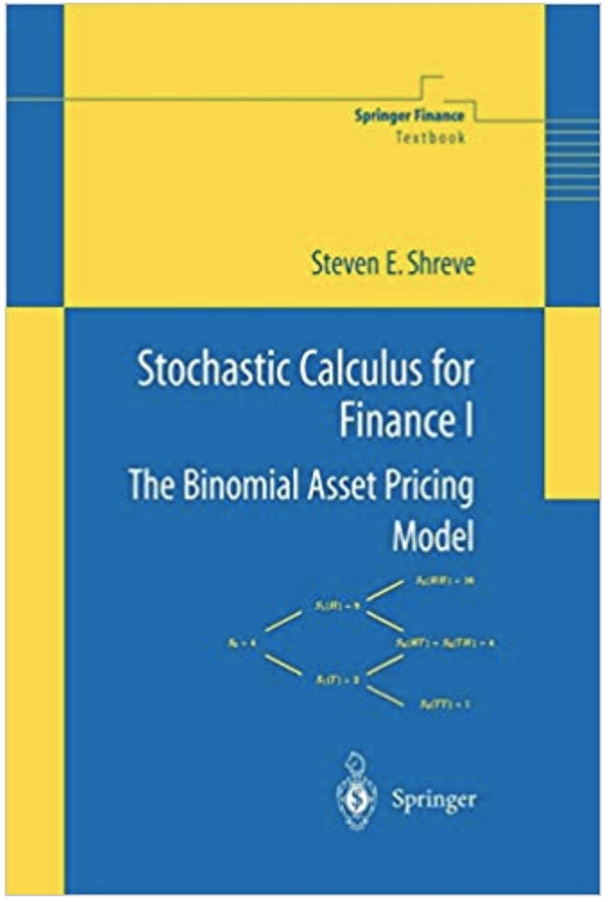
Advanced Algorithmic Trading

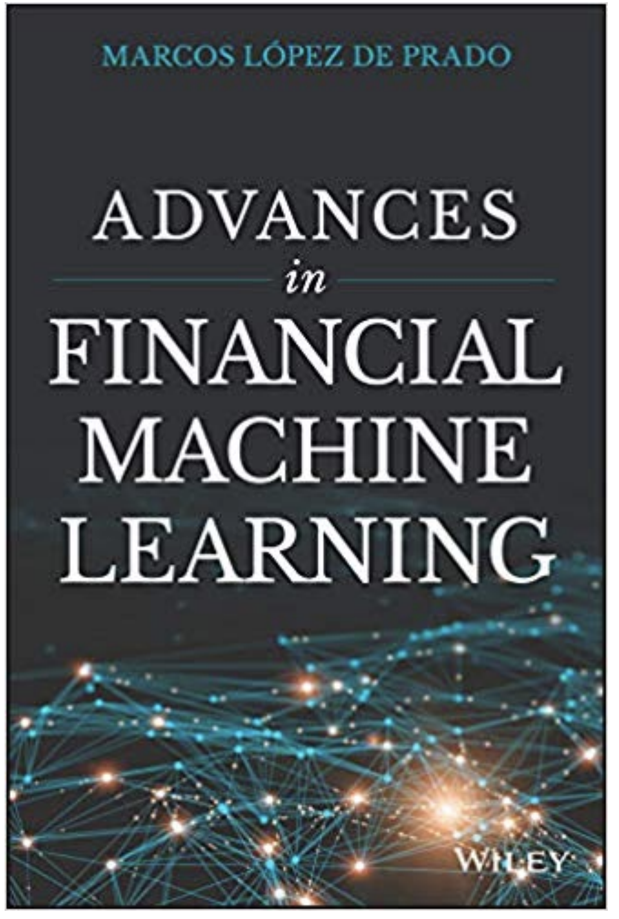
Bayesian statistics, time series analysis and machine learning for profitable systematic trading strategies

Michael L. Halls-Moore, PhD.


O'REILLY
2nd Edition
Python for Finance
MASTERING DATA-DRIVEN FINANCE
Yves Hilpisch


Wiley Finance Series
Derivatives Analytics with Python
Data Analysis, Models, Simulation, Calibration and Hedging
YVES HILPISCH
WILEY


Springer Finance
Textbook
Steven E. Shreve
Stochastic Calculus for Finance I
The Binomial Asset Pricing Model
Springer



Browse > Data Science > Machine Learning

Machine Learning and Reinforcement Learning in Finance Specialization

Reinforce Your Career: Machine Learning in Finance. Extend your expertise of algorithms and tools needed to predict financial markets.

3.8 (822 ratings)

Igor Halperin

Enroll
Starts Feb 29

Premium Access: €34/mo
Financial aid available

9,092 already enrolled

id	era	feature1	. . .	feature310	target
n2b2e3dd163cb422	era1	0.75	. . .	0.00	0.25
n177021a571c94c8	era1	1.00	. . .	0.25	0.75
n7830fa4c0cd8466	era1	0.25	. . .	1.00	0.00
nc584a184cee941b	era1	0.25	. . .	0.00	1.00
nc5ab86679019 46a	era1	0.75	. . .	0.25	0.25
n84e624e4714a7ca	era1	0.00	. . .	0.75	1.00

http://numer.ai/