Machine Learning Applications for Finance

ML in Quantitative Finance demystified, Part II

Industry Lecturer

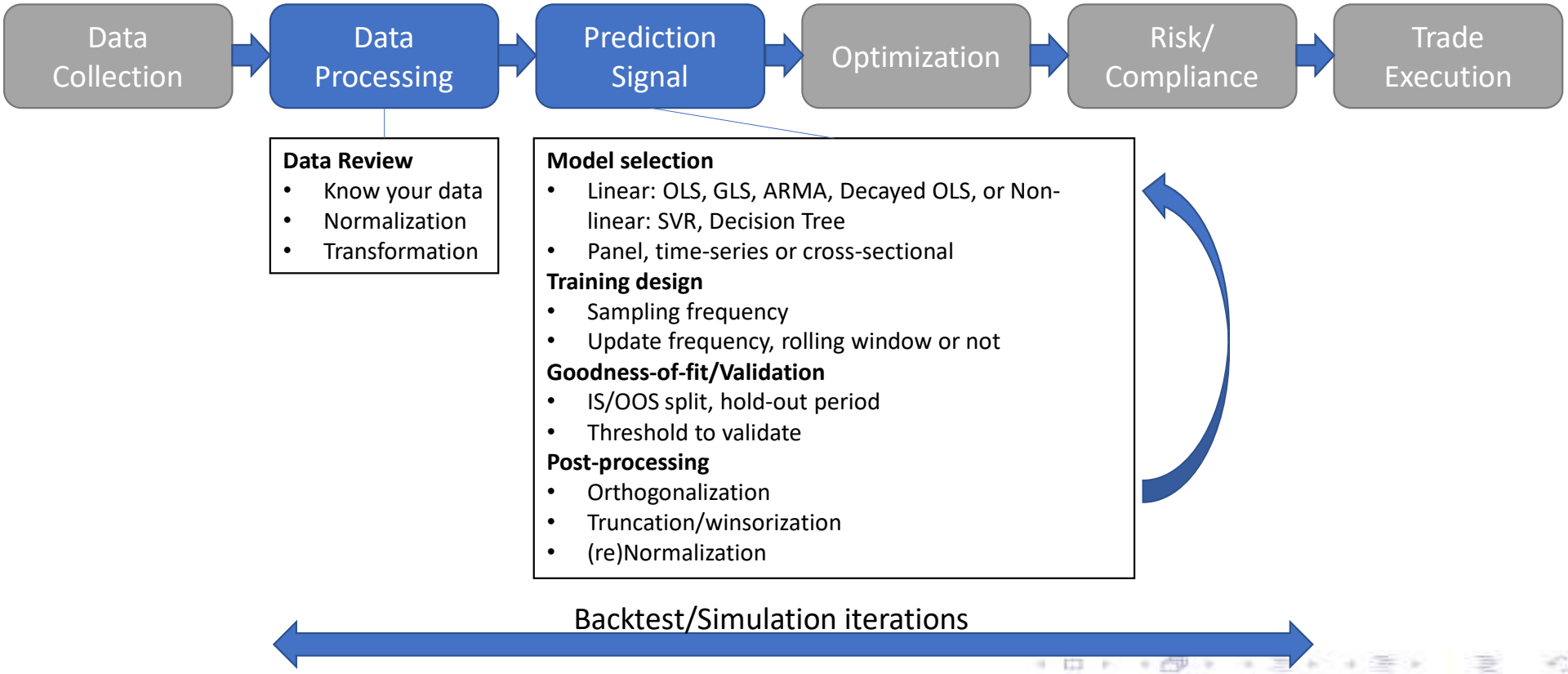
Jun Fan

junfan@bu.edu

Agenda – Part II

- Review Quantitative Investment Process
  - Assembly line process
  - Momentum
  - Training Design
- Modeling
  - OLS model
  - Support Vector Regression model
  - Decision Tree/Random Forest Model
- Landscape - Machine Learning applications in Finance industry
- Discussion and tips

Review - typical alpha research practices



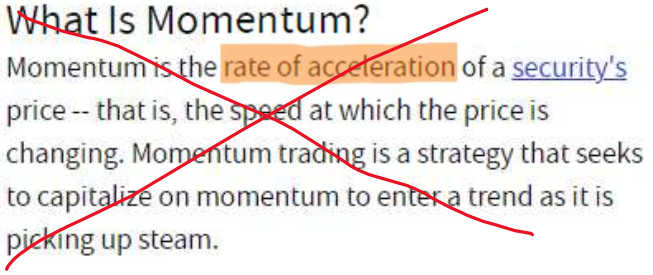
Review - Momentum Signal

Momentum is the **velocity** of price change of a security. In Quant Finance, it's usually constructed by summing up the past returns over a pre-defined period *n* months minus the recent *m* months. For example, Mike Carhart's 12-1 Momentum is constructed with *n=12* and *m=1*.

$$Mom_t = \sum_{i=t-n}^{t-m} Ret_i$$

In industry, various versions of Momentum are being explored and integrated into trading strategies. For hedge funds, the sampling frequencies are usually in days instead of months. e.g. *n=10 days*, *m=1 day*.

In our notebook, we choose to use 6-0 Momentum to demonstrate the process.



https://www.investopedia.com/terms/m/momentum.asp

Review - Training design

- Time-series (TS): Look back window length
  - Entire sample
  - Expanding window
  - Rolling window
- Cross-section (XS): Instruments or universe
  - Time-series (ignore XS, i.e. AR)
  - Cross-sectional (ignore TS, i.e. FMB 2nd regression)
  - pooled training (combine XS and TS into a panel data, a.k.a. flatten data)

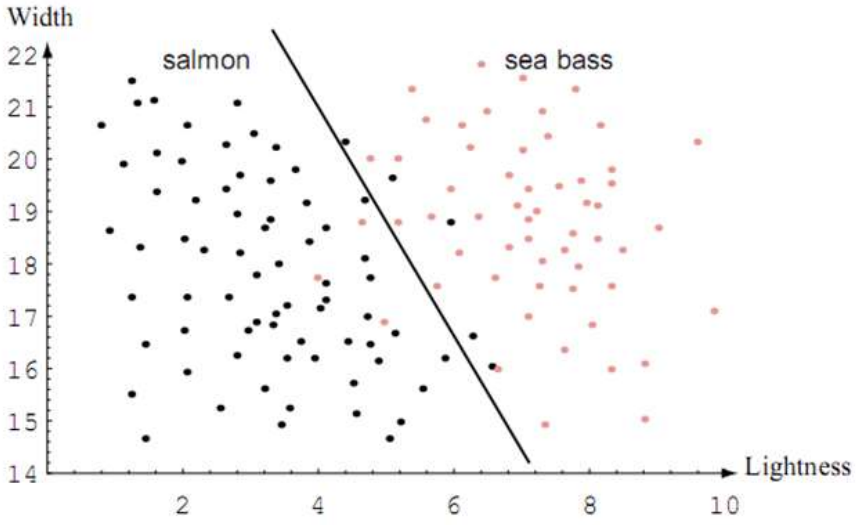
Cross-section

Time

DATE	USDJPY	GBPUSD	USDCAD	AUDUSD	NZDUSD	USDCHF	USDNOK	USDSEK	USDDKK	EURUSD
1/1/1971	358.0200	2.4058	1.0118	1.1181	1.1194	4.3053	7.1411	5.1639	7.4846	NaN
2/1/1971	357.5450	2.4178	1.0075	1.1238	1.1250	4.2981	7.1425	5.1726	7.4854	NaN
3/1/1971	357.5187	2.4187	1.0064	1.1243	1.1254	4.3003	7.1377	5.1628	7.4808	NaN
4/1/1971	357.5032	2.4179	1.0077	1.1238	1.1250	4.2987	7.1287	5.1630	7.4887	NaN
5/1/1971	357.4130	2.4187	1.0087	1.1243	1.1254	4.1242	7.1145	5.1660	7.4998	NaN
...	...	...	...	...	...	...	...	...	...	...
10/1/2020	105.2095	1.2980	1.3218	0.7121	0.6638	0.9124	9.2956	8.8346	6.3243	1.1768
11/1/2020	104.4061	1.3198	1.3073	0.7270	0.6856	0.9110	9.0999	8.6569	6.2968	1.1826
12/1/2020	103.7952	1.3434	1.2809	0.7532	0.7093	0.8884	8.7071	8.3631	6.1154	1.2168
1/1/2021	103.7883	1.3641	1.2725	0.7726	0.7200	0.8865	8.5096	8.2867	6.1082	1.2178
2/1/2021	105.3774	1.3867	1.2696	0.7753	0.7245	0.8979	8.5083	8.3437	6.1489	1.2094

Review - Scatter Plot

Classification (between two Xs)
a.k.a. Hyperplane



Y = 0.5*X1-0.1*X2

If Y<5, it's salmon, otherwise, it's sea bass

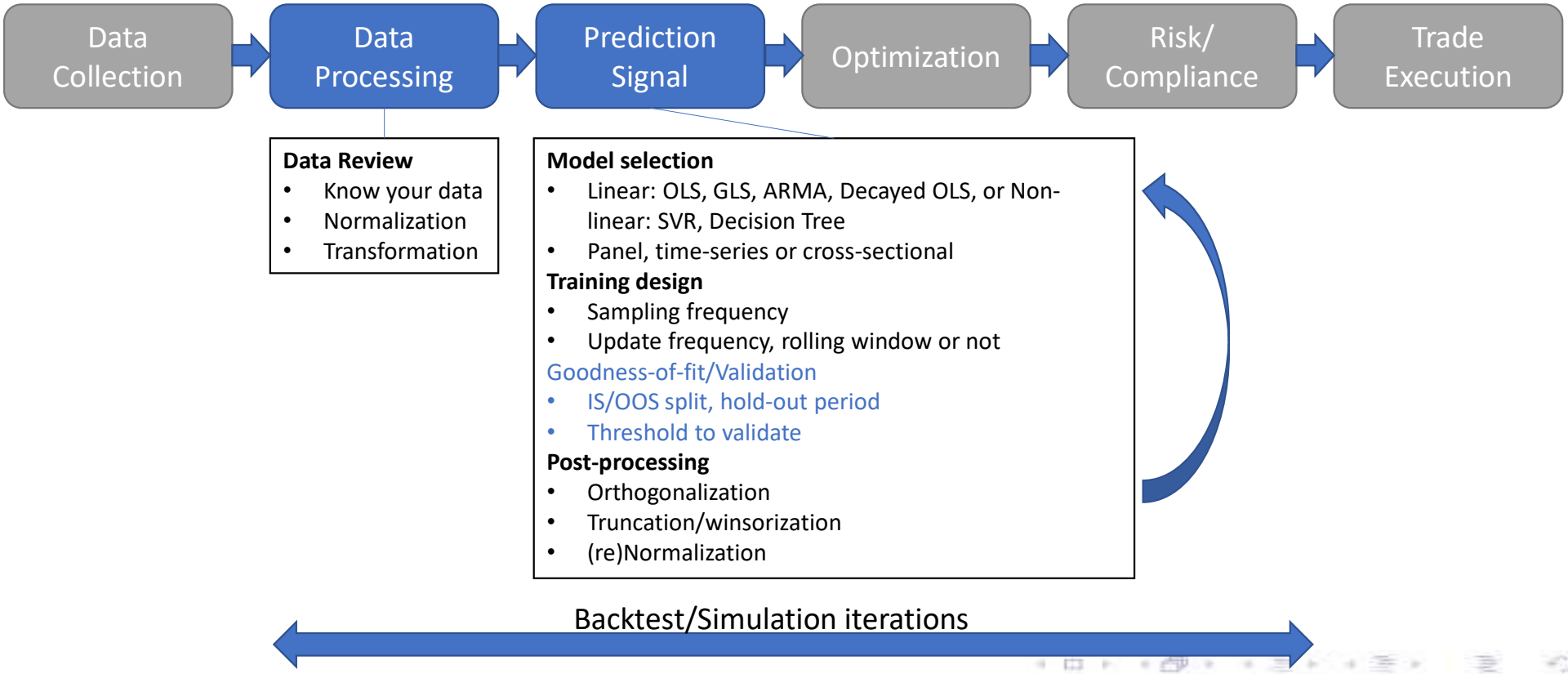
Line Fitting (between X and Y)

Y = 0.1*X1

Model Selection – iterative process

- Iteration 1, no regression or assume beta of 1
- Iteration 2, regression
  - Iteration 2.1, uni-variate regression
  - Iteration 2.2, Multi-variate regression
  - Iteration 2.3, GLS/WLS regression
- Iteration 3, SVR regression
- Iteration 4, Decision tree or Random Forest Regression
- Iteration 5, learn patterns from SVR/Decision Trees and consider picking or transforming the independent variables to use in linear models

Typical alpha research practices



Model review

- Goodness-of-fit
  - R2, t-stat, F-stat, volatility of prediction
    - Panel, Monthly/Yearly, Cross-sectional (bucketing)
    - in-sample and rolling out-of-sample
  - Autocorrelation
  - Dispersion
- Backtest/simulation performance
  - ex-post/ex-ante exposure to common factors
  - Time-series performance (subperiods)
  - Cross-sectional performance (smallcap v.s. largecap bucketing)
  - Forecast horizon/turnover

Why not skip GOF and go straight to simulation?

Goodness-of-fit

- Loosely track live performance
- No direct relationship to constraints
- Not path dependent
- Represents the overall model efficacy for **all stocks** regardless of their alpha
- Fast to run

Simulation

- Closest to live performance
- Can be heavily impacted by constraints
- Path dependent (turnover, DV)
- Represents the **small set of stocks** (usually of high alphas) and their performance
- Slow to run

SVM v.s. SVR

Support Vector Machine

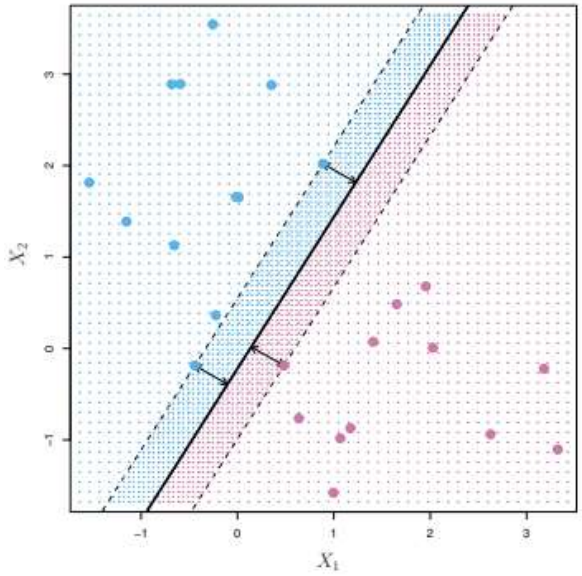
- Kernel
  - non-linear transformation to make linear **classification** feasible at higher dimension
- Hyper Plane
  - separation line between classes
- Boundary line
  - creates support vectors on the hyper plan that maximize the margin
- Support Vectors
  - data points within the boundary lines

Support Vector Regression

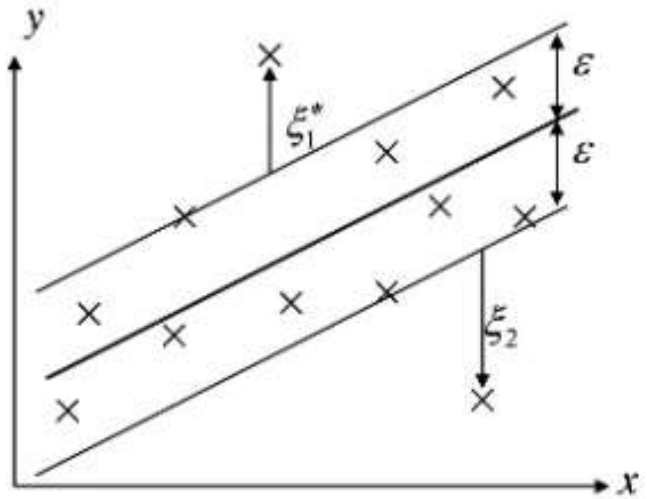
- Kernel
  - non-linear transformation to make linear **fitting** feasible at higher dimension
- Fitted Line
  - predict the continuous value
- Boundary line
  - creates support vectors around the prediction line that allows margin error
- Support Vectors
  - data points outside the boundary lines

SVM v.s. SVR

Support Vector Machine



Support Vector Regression

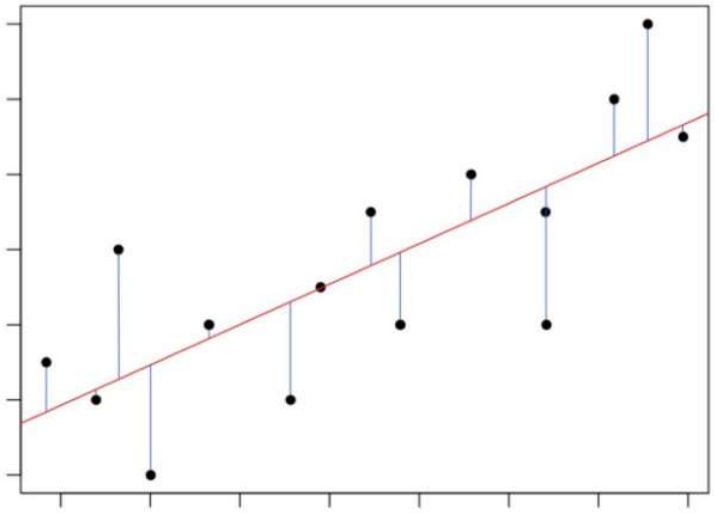


SVR v.s. OLS

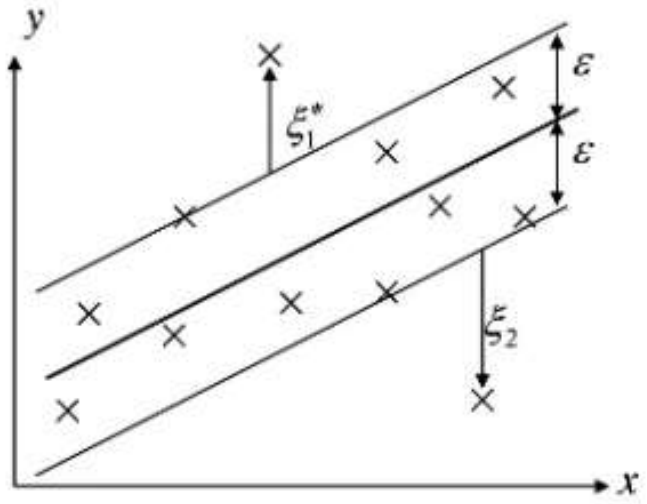
- Difference to Least Square methods
  - Decision boundary (epsilon)
    - Only pick the datapoints outside the margin
      - Balance between reduced #obs and datapoints outside the margin
  - Kernel Function (gamma)
    - Convert pattern from non-linear to linear (or non-linear fitting)
      - Why not truncation/winsorization instead?
      - What about polynomial regression?
- In what situation we should consider SVR?

OLS v.s. SVR

Ordinary Least Squares



Support Vector Regression



Decision Tree

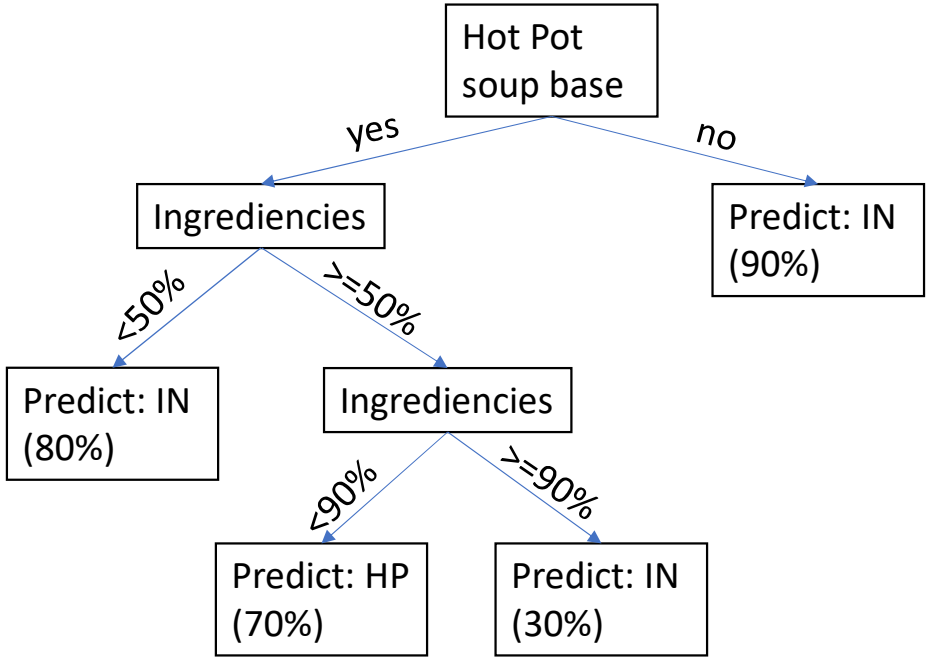
- Models target value using a set of IF-THEN rules
- At each node, choose the best split among all possible splits so that the child nodes are the best fit in training sample
- Recursive partitioning algorithm

Decision Tree Example

- Predict Jun's family dinner choice
  - Hot Pot (HP) or Instant Noodle (IN)
- Decisive variables
  - Hot Pot soup base
  - Ingrediencies
    - Meat, Vegetables, Egg, Tofu, Seafood, Mushroom

Decision Tree Example

- Predict Jun's family dinner choice
  - Hot Pot (HP) or Instant Noodle (IN)
- Decisive variables
  - Hot Pot soup base
  - Ingrediencies
    - Meat, Vegetables, Egg, Tofu, Seafood, Mushroom



Decision Tree

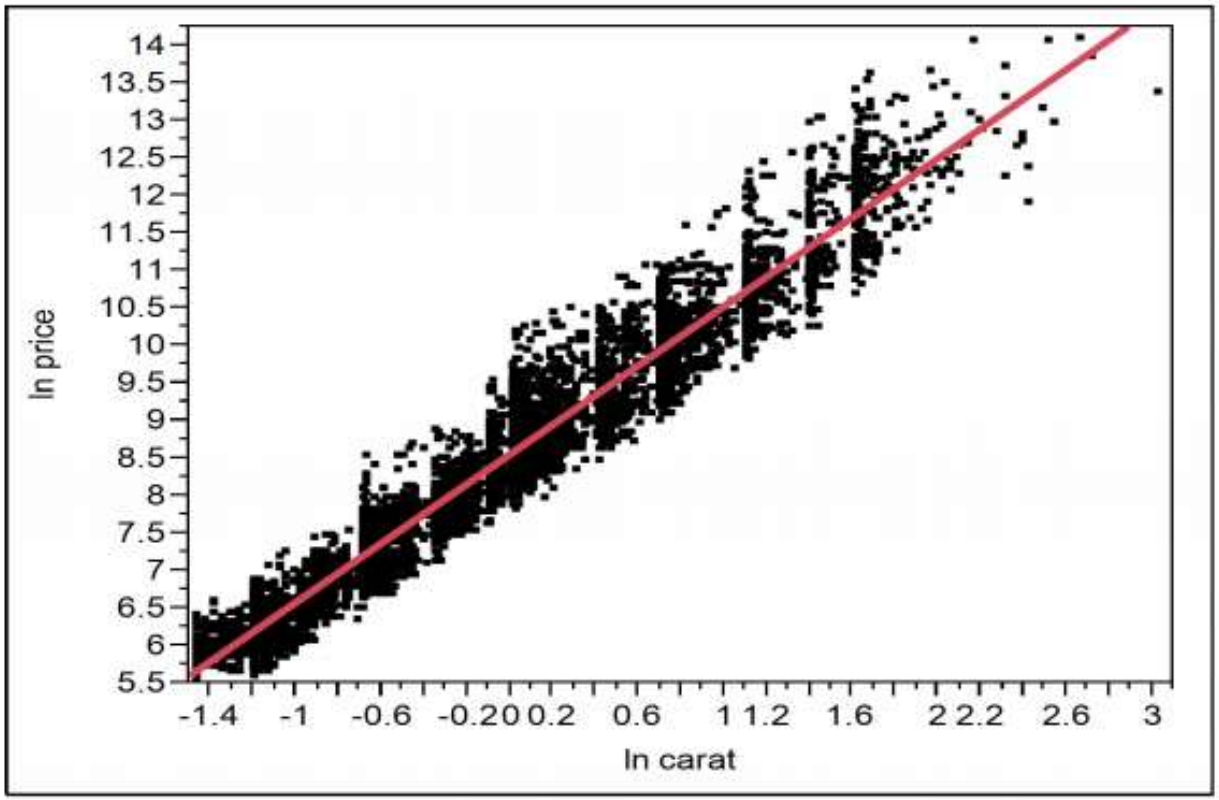
Pros

- Easy to interpret
- Correspond to decision making process
- No prior structure/knowledge needed
- Not sensitive to scale and outliers
- Can incorporate complex interactions

Cons

- They are not efficient if the relationship is in fact smooth
- No neat equation as OLS
- Can be quite unstable. Small change in data can lead to big change in tree output
- Sometimes too good to be true. Be careful not to overfit

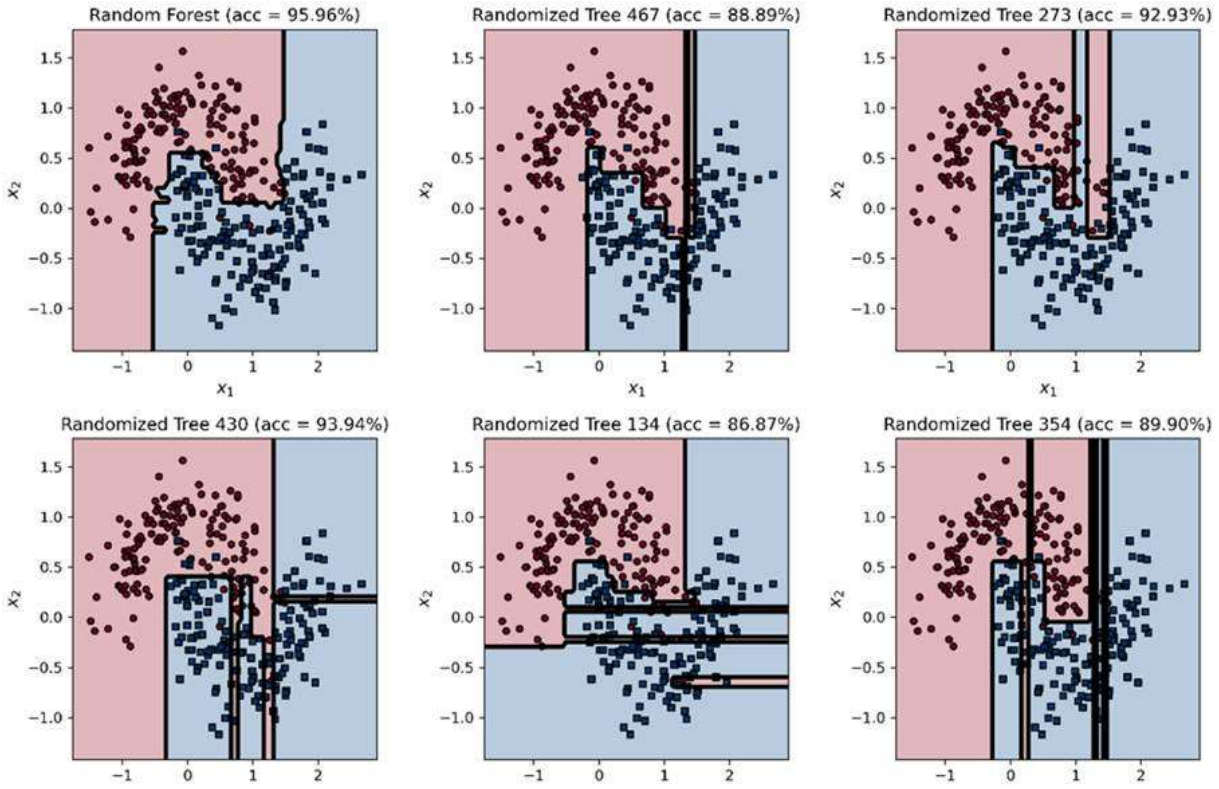
Continuous and smooth world view



Random Forest

- **Bootstrapped** decision trees, combines many weak trees into a strong tree
- Works very well as a goto prediction methodology in data science
- Tuning hyperparameters by cross-validation
- Calibration maybe needed
- Algorithm
  - Create bootstrap samples
  - Fit a small tree to each bootstrap sample
  - Final output is calculated by averaging over the predictions from each tree

Random Forest v.s. Decision Tree



https://livebook.manning.com/concept/machine-learning/random-forest

Takeaway

- Predictive models are centered around time-series modeling
- Industrial quantitative investment process is an assembly line
- Narrow AI models are widely applicable in each components of quantitative investment process
- Intuition and understanding of model (blackbox model is least preferred) is important in finance due to
  - the limited relevant data
  - dynamic nature of finance market

NLP applications

Measurement

- NLP based
  - Sentiment
  - Intensity
  - Agreement
  - Push notice
- Neural Nets based
  - Vector of related words

Methods

- Dictionary Based
  - Basic
  - Sentence embedded
- Neural Nets
  - Word2Vec
  - DOC2VEC

Applications

- Financial Statements
- Earnings call transcripts
- News articles
- Social Media

Foot Traffic tracking

Measurement

- # of workers in manufactures
  - Regular hours
  - Over time hours
- # of visitors to retail stores
  - YoY change
  - Relative to its peer stocks
- # of transportation activities
  - Between warehouses
  - Between warehouses and retail stores

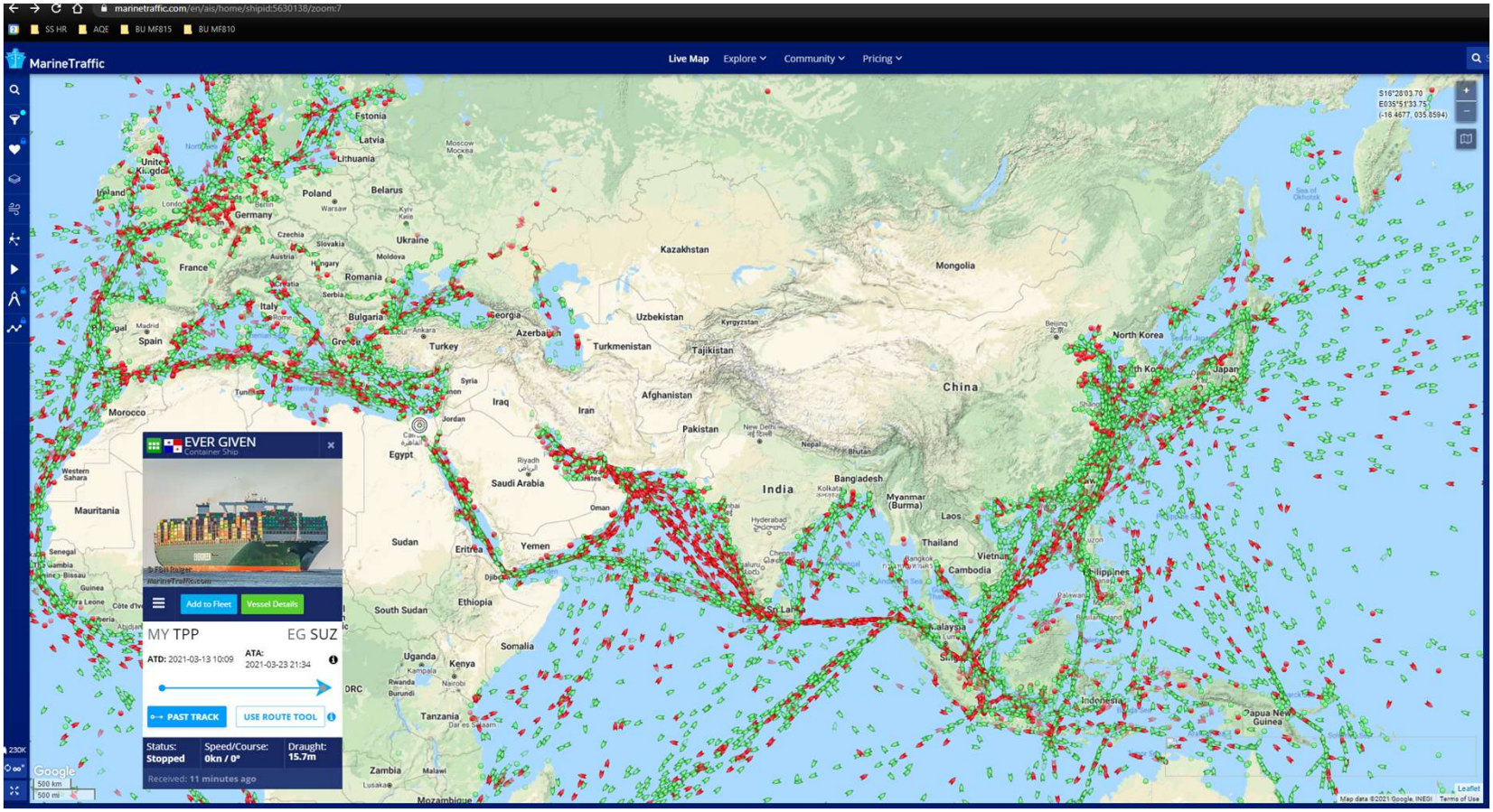
Methods

- Satellite Imaging
  - Parking lot
  - Transportation
- Cellphone Geolocation
  - By tower
  - By GPS
  - Via apps

Applications

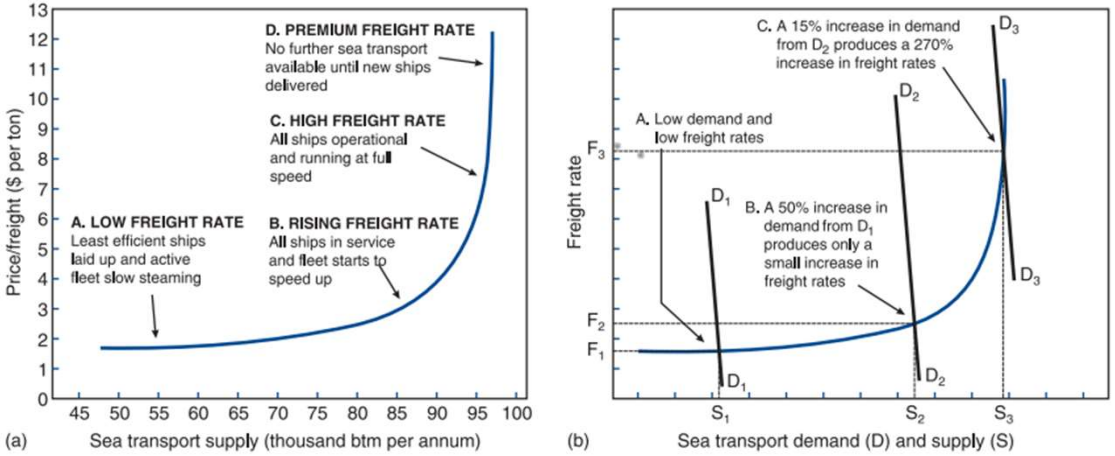
- Retail store visitors
- Supply Chain
- Manufactures workers

Deep-sea shipping intro



Maritime Economics

- Relative lower entry barrier – many family based ship owners
- Fragmentated market (top tanker ship companies own <10% of market share)
- Price is predominately driven by supply and demand



Note: The supply function shows the amount of sea transport offered at each freight rate

Figure 4.14
Short-run equilibrium: (a) short-run supply function; (b) short-run adjustment
Source: Martin Stopford 2007

Maritime Economics

- Use vessel geolocation to predict freight rate
  - Require heavy data lifting to get structured data
  - Most of the transformations are linear or sigmoid/logit
- Ship owner seeks to minimize the distance travelled, similar to Uber's routing
  - More of an optimization problem than a predictive problem
- Use maritime data to estimate the ex-ante beta between FX and commodities
  - Factor model and Bayesian prior
- Predict shocks/volatilities more than the price level
- Can be used to do companies earnings prediction/surprise

How to prepare for quant interviews

- Alpha research
  - Discuss and clarify the underlying driver of prediction
  - Rigorous hypothesis testing to reduce risk of data mining
  - Compare and contrast numerical models and describe their best use cases
- Risk
  - Understand the driver of risk factors
  - Know their influence on various models
  - Refresh your knowledge on cross-section factor regression, such as Fama-MacBeth Regression step 2
- Prepare for coding test
  - Refresh data structure and algorithm – use one of the coding prep website such as leetcode.com
  - Start with pseudo code and ask for syntax help if being asked to write in a specific language

- Ask for clarification questions if you don't fully understand the question (e.g. what you mean by model efficacy)
- Most of the answers aren't binary, so start your answer with "assume the situation A, we would do balabala"
- But Mathematics questions are most likely to be binary, such as $\beta = \rho \frac{\sigma_y}{\sigma_X}$

Intuition

- Be sensitive to numbers
  - But don’t be blindsided by number
  - back-of-the-envelope calculation
- Be mindful of mining risk
  - Data mining
  - Model mining
    - Understand why we choose to use this model
    - Parsimony (if simple model works, don’t choose more complicated models)
- Anticipate what you expect to see before running any model
- Know model characteristics
  - Exposure
  - Performance

Jargon used in quant finance industry

- Autocorrelation
- Orthogonal
- Root-cap weighting
- IC (correlation?)
- Vol/volatility (stddev or portfolio risk?)
- Data/model mining, not always a negative word
- Spread, is it
  - Bid/ask spread or
  - Bucket spread
  - Calendar spread in futures market
- Regions, such as DM, EM, DMxUS, DMxUSJP, EMxChinaA
- Risk Premium
- DV/ADV

ML in Quantitative Finance today

Mostly upstream application

- unstructured data to structured data conversion
  - NLP, Word2VEC, DOC2VEC
  - News articles, social media, Maritime data (BOL, AIS), ESG
- Push-notification (Dataminr, Social Media Analysis)
- Foot-traffic, supply chain flow tracking

A few downstream process such as

- Predictive alpha factor construction
  - with caveats discussed in our FX prediction model
- Non-linearly factor combination
- Non-linearly Optimization
  - Your next 2 classes
- Trade Decisions
  - What's the best execution algo for my order based on stock attribute?

FinTech

- Robo Advisor (not really a ML nowadays)
- Digital Marketing
- Blockchain

ML in Finance tomorrow

Moving downstream processes

- More alpha factor constructions that captures higher dimension patterns
- More Factor combination and portfolio optimization
- Widely applicable in entire quantitative process
- More automated trading recommendations or decisions directly from ML models

How about us human?

- Like the autopilot feature in airplanes, we design and build the auto piloting functions in airplanes. Most of the time we can let autopilot to fly the airplane, but we know the limitations of autopilots and know when to intervene the model when things are outside the normal conditions in order to take back the control and land the airplane safely to the ground.
  - Quant Researcher/Developer: Design and build the airplane and autopilot function
  - Quant Portfolio Manager: Fly the airplane using autopilot function, but be prepared to intervene when things go wrong

ML in Finance tomorrow

Cockpit of a Boeing 707

Cockpit of a Boeing 787

ML in Finance tomorrow