Machine Learning: Applications and Challenges in Finance and Economics

Georgios Sermpinis

Real World Applications

- Portfolio Management and Robo-Advisors
- Algorithmic Trading
- Fraud Detection
- Loan/ Insurance Underwriting
- Risk Management
- Chatbots
- Trade Settlements
- Sentiment Analysis (recently)

Real World Issues

- - The Eurekahedge AI Hedge Fund Index provides a broad measure of the performance of underlying hedge fund managers who utilize artificial intelligence and machine learning theory in their trading processes.

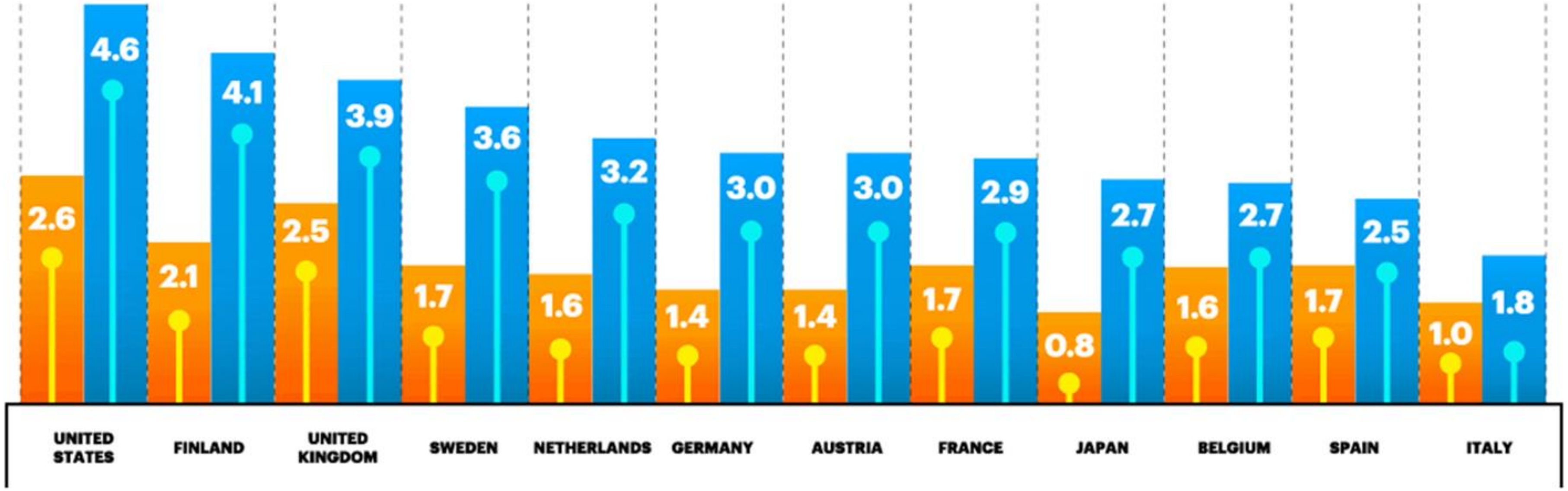
https://www.eurekahedge.com/Indices/IndexView/Eurekahedge/683/Eurekahedge_AI_Hedge_fund_Index#

The AI index lost 7.3% at February,2019

- In 2015 Javelin Strategy and Research reported that at least 15% of all cardholders had at least one transaction incorrectly declined in the previous year, which represented an annual revenue loss totalling nearly $118 billion.

- Increased Unemployment: https://m.economictimes.com/jobs/jobs-changing-with-artificial-intelligence-but-no-mass-unemployment-expected-un-labour-experts/articleshow/65682185.cms

The Effect of Machine Learning and AI



Annual growth rates in 2035 of gross value added (a close approximation of GDP), comparing baseline growth in 2035 to an artificial intelligence scenario where AI has been absorbed into the economy
Source: Accenture and Frontier Economics

Applications in Research

- - Trading
- - Forecasting (mainly economics)
- - Bankruptcy prediction
- - Risk Management
- - Volatility modelling
- - Assets (mainly derivatives) pricing

As machine learning models can perform regression and classification can be applied at any finance and economics issue. The results are promising compared to traditional methods. But:

- There is scepticism from many academics in the field

Research Issues

- Why machine learning does not dominate the research field in Finance and Economics as it does in the real world? Why researchers still rely in simple statistics?

1. Lack of connection with Finance and Economics theory:

Paul Samuelson (nobel prize in Economics):

"Economists stick to their theoretical models the majority of which are of little to no value in the real world because they just can't take account of the impact unpredictable and inherently irrational human behaviour on the state of affairs."

- The assumptions does not hold
- Research becomes irrational, markets have a time-varying behaviour
- The "zoo factor": Out of the 400 factors in asset pricing only 1-5 holds

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2528780

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3341728

Research Issues

2. Complexity

Similarly with traditional Finance and Economics, some of the research in machine learning becomes complex, irrational and lacks novelty

3. Data-snooping or p-hacking, no generalization and/or replication

No formal rules in most machine learning models

Parametrization issues, small samples

Some models can not inherently be exactly always replicated

It is possible to achieve good results with any complex model

Possible Remedies

- Increase the t-ratios threshold
- Robustness tests, multiple samples, sensitivity analysis
- Multiple Hypothesis Testing - multiple benchmarks
- Modern evaluation of the results (traditional metrics can be misleading – refer to statistical tests)
- Apply committees, no single models (for example for NNs)

Towards machine learning including full distribution of risk

Jens Perch Nielsen

Joint work with Ioannis Kyriakou, Parastoo Mousavi
and Michael Scholz

Cass Business School, City, University of London
&
Department of Economics, University of Graz

Machine learning workshop
Faculty of Actuarial Science and Insurance, Cass Business school
17 Jan 2020

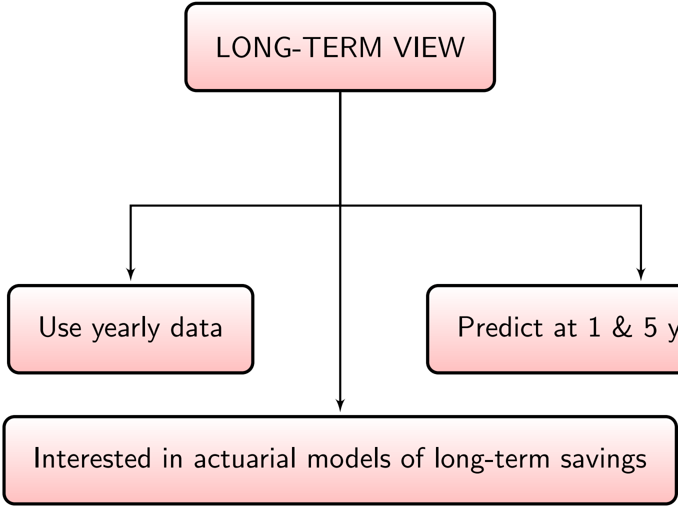
Machine learning I

Method of working involving certain key processes:

- **First**: Problem, audience and potential client must be articulated
- **Second**: Domain knowledge
- **Third**: New techniques must be qualified against earlier ones via validation
- **Fourth**: Channel prior knowledge to statistical model used for validation

CASS BUSINESS SCHOOL CITY UNIVERSITY OF LONDON EST 1894

Example: stock returns



Predictions based on annual US data by Robert Shiller (corecipient 2013 Nobel Memorial Prize Economic Sciences for empirical analysis of asset prices).
Use US market data consisting of annual series (period 1872–2015) of:

- S&P Composite Stock Price Index (P)
- Dividends Accruing to Index (D)
- Earnings Accruing to Index (E)
- One-year interest rate (R)
- Ten-year long government bond yield (L)
- Consumer Price Index (CPI).

Dataset:

- Updated and revised version of Shiller (1989, Chapter 26)
- Available from `http://www.econ.yale.edu/~shiller/data.htm`
- Further research $\Rightarrow$ 1872-2019 with some modifications

Financial model I

Stock returns:

$$S_t = \frac{P_t + D_t}{P_{t-1}}$$

- P_t: (nominal) stock price at the end of year t
- D_t: (nominal) dividends paid during year t

Financial model II

Stock returns in excess of some benchmark $B^{(A)}$:

$$Y_t^{(A)} = \ln \frac{S_t}{B_{t-1}^{(A)}} = \ln S_t - \ln B_{t-1}^{(A)}, \quad A \in \{R, L, E, C\}.$$

Benchmarks under study:

- Short-term rate:
$$B_t^{(R)} = 1 + \frac{R_t}{100}$$
- Long-term rate:
$$B_t^{(L)} = 1 + \frac{L_t}{100}$$
- Earnings-by-price:
$$B_t^{(E)} = 1 + \frac{E_t}{P_t}$$
- Inflation:
$$B_t^{(C)} = \frac{CPI_t}{CPI_{t-1}}.$$

Financial model III

A predictive nonparametric regression model:

$$Y_t^{(A)} = m(X_{t-1}) + \xi_t$$

- ξ_t: error terms
- $m(x) = \mathbb{E}(Y^{(A)}|X = x),\ x \in \mathbb{R}^q$: unknown smooth function
- X: predictive (lagged) variables

Financial model IV

Choice of predictive variables:

- Dividend-by-price ratio:

$$d_{t-1} = \frac{D_{t-1}}{P_{t-1}}$$

- Earnings-by-price ratio:

$$e_{t-1} = \frac{E_{t-1}}{P_{t-1}}$$

- Short-term interest rate:

$$r_{t-1} = \frac{R_{t-1}}{100}$$

- Long-term interest rate:

$$l_{t-1} = \frac{L_{t-1}}{100}$$

- Inflation:

$$\pi_{t-1} = \frac{CPI_{t-1} - CPI_{t-2}}{CPI_{t-2}}$$

- Term spread:

$$s_{t-1} = l_{t-1} - r_{t-1}$$

- Excess stock return:

$$Y_{t-1}^{(A)} = \ln \frac{S_{t-1}}{B_{t-2}^{(A)}}$$

CASS BUSINESS SCHOOL
CITY UNIVERSITY OF LONDON
EST 1894

Validation

- Nonparametric technique $\Rightarrow$ need for an adequate measure of predictive power.
- For model and optimal bandwidth selection $\Rightarrow$ leave-k-out cross-validation method (Nielsen and Sperlich, 2003):

$$R_V^2 = 1 - \frac{\sum_t (Y_t^{(A)} - \hat{m}_{-t})^2}{\sum_t (Y_t^{(A)} - \bar{Y}_{-t}^{(A)})^2}$$

where leave-k-out estimators are used:

  - $\hat{m}_{-t}$ for the nonparametric function m
  - $\bar{Y}_{-t}^{(A)}$ for the unconditional mean of $Y_t^{(A)}$
    Both are computed by removing k observations around the tth time point

Single benchmarking approach I

Recall original predictive nonparametric regression model:

$$Y_t^{(A)} = m(X_{t-1}) + \xi_t.$$

- Only dependent variable $Y^{(A)}$ adjusted according to one of the four benchmarks $B^{(A)}$, $A \in \{R, L, E, C\}$.
- Independent variable(s) is (are) measured on the original nominal scale.

Double (full) benchmarking approach I

Prediction problem reformulated as:

$$Y_t^{(A)} = m(X_{t-1}^{(A)}) + \xi_t.$$

- Independent and dependent variables adjusted according to same benchmark.
- Use transformed predictive variables

$$X_{t-1}^{(A)} = \begin{cases} \frac{1+X_{t-1}}{B_{t-1}^{(A)}}, & X \in \{d, e, r, l, \pi, s\} \\ Y_{t-1}^{(A)} & \end{cases}, \quad A \in \{R, L, E, C\}.$$

- Balanced dimension: additional economic structure in prediction without extra cost in the form of increasing problem dimensionality.

Single benchmarking approach - one year view

Table: Red : beat the historical mean. Green: the best prediction

Benchmark $B^{(A)}$	Explanatory variable(s) X_{t-1}						
	$Y^{(A)}$	d	e	r	l	π	s
Short-term rate	-1.5	-1.0	-0.3	4.0	-0.1	-1.4	13.2
Long-term rate	-1.8	-0.7	0.0	2.1	-0.1	-1.4	8.8
Earnings-by-price	-1.7	-1.4	-1.5	-0.1	-0.9	-1.2	8.7
Inflation	-1.3	-0.2	-1.5	1.1	-0.8	10.5	9.9
	$(Y^{(A)}, d)$	$(Y^{(A)}, e)$	$(Y^{(A)}, r)$	$(Y^{(A)}, l)$	$(Y^{(A)}, \pi)$	$(Y^{(A)}, s)$	
Short-term rate	-2.3	-2.0	1.9	-2.3	-2.8	9.0	
Long-term rate	-2.2	-1.8	0.0	-2.5	-3.0	4.4	
Earnings-by-price	-3.5	-3.7	-2.0	-2.8	-2.8	5.1	
Inflation	-1.2	-3.2	-0.6	-2.5	10.2	6.9	
	(d, e)	(d, r)	(d, l)	(d, π)	(d, s)		
Short-term rate	-2.7	3.0	-1.7	-2.4	12.0		
Long-term rate	-2.3	1.4	-1.4	-2.2	7.6		
Earnings-by-price	-3.8	-1.6	-2.3	-2.6	6.6		
Inflation	-2.0	0.9	-1.2	9.7	9.7		
	(e, r)	(e, l)	(e, π)	(e, s)			
Short-term rate	5.4	-0.8	-1.1	13.0			
Long-term rate	3.7	-0.3	-0.7	8.6			
Earnings-by-price	-1.5	-2.4	-2.7	6.4			
Inflation	0.0	-2.5	11.5	8.1			
	(r, l)	(r, π)	(r, s)				
Short-term rate	10.4	2.5	12.2				
Long-term rate	5.9	0.5	7.8				
Earnings-by-price	6.6	-1.6	8.2				
Inflation	6.6	9.7	8.8				
	(l, π)	(l, s)					
Short-term rate	-2.0	12.5					
Long-term rate	-2.1	8.1					
Earnings-by-price	-1.9	8.1					
Inflation	10.1	9.3					
	(π, s)						
Short-term rate	10.3						
Long-term rate	5.8						
Earnings-by-price	5.7						
Inflation	16.1						

Double (full) benchmarking approach - one year view

Table: Red : beat the historical mean. Green: the best prediction

Benchmark $B^{(A)}$	Explanatory variable(s) $X_{t-1}^{(A)}$						
	$Y^{(A)}$	$d^{(A)}$	$e^{(A)}$	$r^{(A)}$	$l^{(A)}$	$\pi^{(A)}$	$s^{(A)}$
Short-term rate	-1.5	4.2	7.0	–	13.2	-1.2	13.2
Long-term rate	-1.8	-0.2	1.0	8.9	–	-1.5	8.9
Earnings-by-price	-1.7	-2.3	–	0.3	-1.2	0.0	8.7
Inflation	-1.3	10.7	13.3	6.6	10.8	–	8.9
	$(Y^{(A)}, d^{(A)})$	$(Y^{(A)}, e^{(A)})$	$(Y^{(A)}, r^{(A)})$	$(Y^{(A)}, l^{(A)})$	$(Y^{(A)}, \pi^{(A)})$	$(Y^{(A)}, s^{(A)})$	
Short-term rate	2.7	5.0	–	8.7	-2.7	8.7	
Long-term rate	-1.9	-0.8	4.3	–	-3.1	4.3	
Earnings-by-price	-4.1	–	-1.7	-3.2	-2.5	5.0	
Inflation	11.2	12.3	5.8	10.1	–	6.0	
	$(d^{(A)}, e^{(A)})$	$(d^{(A)}, r^{(A)})$	$(d^{(A)}, l^{(A)})$	$(d^{(A)}, \pi^{(A)})$	$(d^{(A)}, s^{(A)})$		
Short-term rate	4.1	–	11.5	2.6	11.5		
Long-term rate	-1.2	7.0	–	-1.8	7.0		
Earnings-by-price	–	-2.8	-3.7	-3.2	4.9		
Inflation	11.2	9.8	10.3	–	16.2		
	$(e^{(A)}, r^{(A)})$	$(e^{(A)}, l^{(A)})$	$(e^{(A)}, \pi^{(A)})$	$(e^{(A)}, s^{(A)})$			
Short-term rate	–	12.9	5.2	12.9			
Long-term rate	8.4	–	-0.1	8.4			
Earnings-by-price	–	–	–	–			
Inflation	12.4	12.0	–	18.7			
	$(r^{(A)}, l^{(A)})$	$(r^{(A)}, \pi^{(A)})$	$(r^{(A)}, s^{(A)})$				
Short-term rate	–	–	–				
Long-term rate	–	5.7	–				
Earnings-by-price	4.9	-1.6	6.5				
Inflation	15.1	–	16.9				
	$(l^{(A)}, \pi^{(A)})$	$(l^{(A)}, s^{(A)})$					
Short-term rate	10.2	–					
Long-term rate	–	–					
Earnings-by-price	-2.4	6.3					
Inflation	–	16.7					
	$(\pi^{(A)}, s^{(A)})$						
Short-term rate	10.2						
Long-term rate	5.7						
Earnings-by-price	5.6						
Inflation	–						

Open questions:

1. How do we get towards "big-data" combining all the many results on previous slide (maybe adding one thousand times as many)?

2. Regression only works on the mean, what about full risk?

Current Cass business school work stream (with other universities as well):

1. Provide alternative to big-data algorithms such as regression-trees and Lasso with a more classical-statistical approach to modelling and validation

2. Consider full-risk models (regression approach only considers the mean)

Thank You For Your Attention

References

Nielsen, J. P., S. Sperlich. 2003. Prediction of stock returns: A new way to look at it. *ASTIN Bulletin* **33**(2) 399–417.

Shiller, R. J. 1989. *Market Volatility*. MIT Press, Cambridge.