Applied Data Science in Economics

What can Data Science contribute to Economics?

Julian Oliver Doerr

November 4, 2019

Research project

Machine learning

Macroeconomic forecasting

Call for new approaches in economics

Blanchard (2014): 'The **techniques** we use [...] were **best suited** to a worldview in which **economic fluctuations** occurred but were **regular**, and essentially self correcting.'

Romer (2016): 'Post-real macro models [and their] predictions were wildly incorrect, and [...] the **doctrine** on which they were based is **fundamentally flawed**.'

Haldane (2016): 'Few forecasters foresaw even a slight downturn in GDP in 2008 and none foresaw a recession. [...] At root, these were failures of models, methodologies and mono-cultures. [...] The **methodological mono-culture** produced, unsurprisingly, the same crop.'

(Standard) econometrician forecasting toolbox

Forecasting by:

- history of target variable (e.g. ARIMA)
- other variables (e.g. VAR)
- factors (e.g. FAVAR)
- ...

Extending the toolbox

▷ calibration results

This project aims at forecasting real GDP growth by means of:

- Random Forest (RF)
- Gradient Boosting (GB)
- Support Vector Regression (SVR)

Data

FRED-QD: U.S. quarterly macroeconomic variables

Balanced panel with quarterly data of **202 features** from 1959Q3:2019Q2

Features comprise time series related to:

- labor market
- housing market
- stock market
- price indices
- interest rates
- sentiment surveys
- ...

Results

▷ benchmark ARIMA

Table: Forecasting performance

Model	1-quarter-ahead RMSE	1-quarter-ahead Accuracy increase	1-year-ahead RMSE	1-year-ahead Accuracy increase
RW	0.657		0.876	
ARIMA	0.597	9.1%	0.698	20.3%
RF	0.537	18.2%	0.628	28.4%
GB	0.504	23.3%	0.588	32.9%
SVR	0.478	27.2%	0.700	20.1%

Note: Performance based on forecasting errors in test set (2007Q2:2019Q2). Accuracy increase relative to RMSE of RW model.

OECD (2019) leading indicators

▷ more information

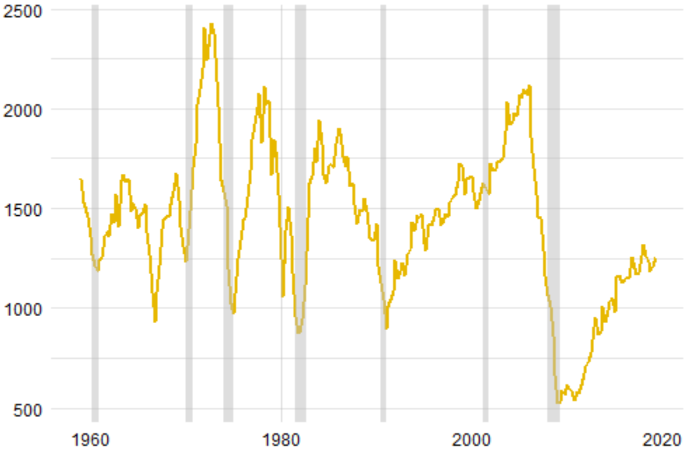
202 features

6 leading indicators

1 Housing starts
2 Manufacturer's new orders of durable good
3 S&P 500 stock price index
4 Consumer sentiment index
5 Weekly hours worked in manufacturing
6 Yield curve

OECD (2019) leading indicators

Figure: New privately owned housing units started (in thous of units)



Source: Time series from McCracken and Ng (2016), recession dates from NBER

Results ...

... based on leading indicators only

Table: Forecasting performance: Leading indicators

Model	1-quarter-ahead		1-year-ahead	
	RMSE	Accuracy increase	RMSE	Accuracy increase
VAR	0.643		0.680	
RF	0.608	5.6%	0.655	3.7%
GB	0.553	14.1%	0.628	7.6%
SVR	0.625	2.9%	0.667	2.0%

Note: Performance based on forecasting errors in test set (2007Q2:2019Q2). Accuracy increase relative to RMSE of VAR model.

Results

Would have machine learning models **predicted** the **last financial crisis?**

We do not know without real-time data.

Would have machine learning models yielded **more accurate forecasts?**

Results support it.

Reproducibility

machineLearning-economicForecasting

References

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, *31*(2), 3–32.

Blanchard, O. (2014). Where danger lurks. *Finance & Development*, *51*(3), 28–31. Retrieved from https://www.imf.org/external/pubs/ft/fandd/2014/09/pdf/blanchard.pdf

Haldane, A. (2016). The dappled world. *Shackle Biennial Memorial Lecture*. Retrieved from https://www.bankofengland.co.uk/-/media/boe/files/speech/2016/the-dappled-world.pdf?la=en&hash=5775E592A5F79791B5F883604FF9E817C5185631

McCracken, M. W., & Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, *34*(4), 574–589. doi:10.1080/07350015.2015.1086655

OECD. (2012). OECD system of composite leading indicators. Retrieved from http://www.oecd.org/sdd/leading-indicators/41629509.pdf

OECD. (2019). OECD composite leading indicators: Turning points of reference series and component series. Retrieved from http://www.oecd.org/sdd/leading-indicators/CLI-components-and-turning-points.pdf

Romer, P. (2016). The trouble with macroeconomics. *The American Economist*, *20*, 1–20. Retrieved from https://ccl.yale.edu/sites/default/files/files/The%20Trouble%20with%20Macroeconomics.pdf

Backup

◁ presentation

Training data	1959Q3:2007Q1
Test data	2007Q2:2019Q2
Model selection	BIC
Search method	Grid search
Final parameters:	
p	2
q	0

Models
VAR

Training data	1959Q3:2007Q1
Test data	2007Q2:2019Q2
Model selection	BIC
Search method	Grid search
Final parameters:	
p	2

Models

RF

Training data	1959Q3:2007Q1
Test data	2007Q2:2019Q2
Model selection	Blocked cross validation based on rolling-origin re-calibration for hyperparameter tuning
Search method	Random Search
Final parameters:	
M	16
d_{try}	54
$node_{min}$	36

Note: Parameter results based on full feature space

Training data	1959Q3:2007Q1
Test data	2007Q2:2019Q2
Model selection	Blocked cross validation based on rolling-origin re-calibration for hyperparameter tuning
Search method	Random Search
Final parameters:	
M	411
ν	0.074
$depth_{max}$	8

Note: Parameter results based on full feature space

Models

SVR

Training data	1959Q3:2007Q1
Test data	2007Q2:2019Q2
Model selection	Blocked cross validation based on rolling-origin re-calibration for hyperparameter tuning
Search method	Random Search

Final parameters:	
C	0.019
ϵ	0.393
kernel	sigmoid

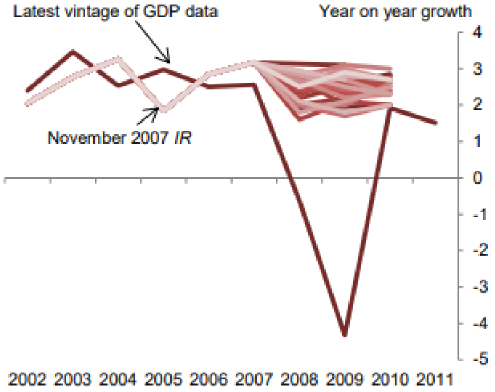
Final kernel parameters:	
γ	0.005
c	0

Note: Parameter results based on full feature space

Sigmoid kernel: $K(x_i, x) = \tanh\left(\gamma\langle x_i', x\rangle + c\right)$

Failure of macroeconomic forecasts

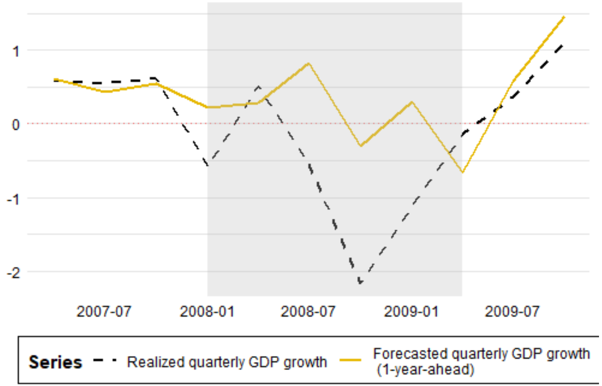
Figure: Range of forecasts for UK GDP growth from 2008 onwards produced by 27 economic forecasters in 2007



Source: Haldane (2016)

Best machine learning forecast

Figure: Gradient Boosting on-year-ahead forecast during global financial crisis



Results

◁ presentation

Table: Forecasting performance

Model	1-quarter-ahead		1-year-ahead	
	RMSE	Accuracy increase	RMSE	Accuracy increase
ARIMA	0.597		0.698	
RF	0.537	10.1%	0.628	10.0%
GB	0.504	15.6%	0.588	15.8%
SVR	0.478	19.9%	0.700	−0.3%

Note: Performance based on forecasting errors in test set (2007Q2:2019Q2). Accuracy increase relative to RMSE of ARIMA model.

Leading indicators

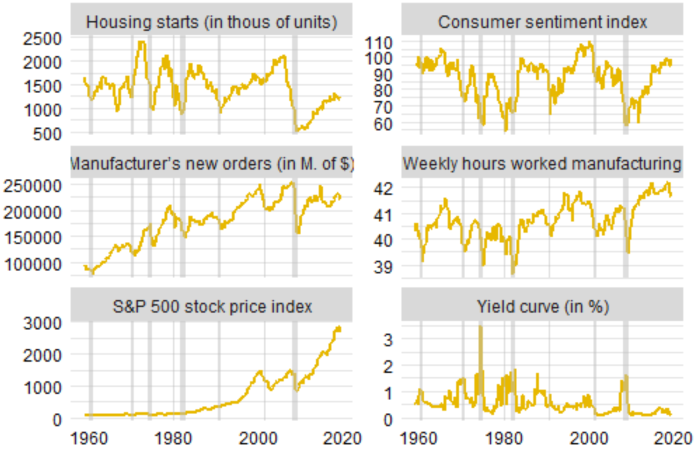
Table: U.S. leading indicators

Leading indicator	Explanation
Housing starts	Households spend substantial fractions of their income not only on their homes, but also what goes in them. This affects employment in the construction sector and money demand. Moreover, housing market contribues a substantial fraction to overall GDP.
Consumer sentiment	Reflects how well-off consumers expect to be in the future relative to today affecting future spending.
S&P 500 stock prices	Present value of expected future returns (incorporates expectations!)
New orders manufacturing	Increases in new orders for consumer goods and materials usually mean positive changes in actual production. The new orders decrease inventory and contribute to unfilled orders, a precursor to future revenue.
Hours worked manufacturing	Adjustments to the working hours of existing employees are usually made in advance of new hires or layoffs.
Interest rate spread	Also referred as yield curve which entails expected direction of short-, medium- and long-term interest rates. This is particularly true when the curve becomes inverted, that is, when the longer-term returns are expected to be less than the short rates.

Note: Leading indicators determined by OECD (2019). OECD (2012) defines leading indicators as time series which exhibit leading relationship with the reference series (GDP) at turning points.

Leading indicators

◁ presentation



Source: Time series from McCracken and Ng (2016), recession dates from NBER

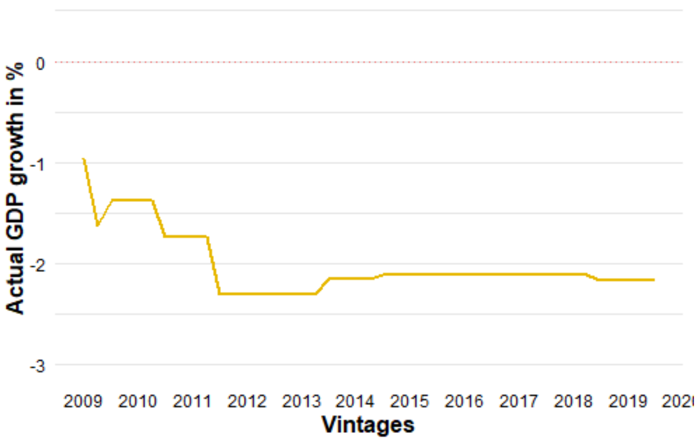
Research project

Open topics

- structural VAR models
- econometric factor models (dimension reduction)
- neural networks
- turning point analysis
- interval forecast (uncertanity)
- Diebold-Mariano Test
- writing

Data revisions

Figure: Revision of 2008Q4 U.S. GDP growth at different vintage dates



Source: The Real-Time Data Research Center of the Federal Reserve Bank of Philadelphia

Limitations of machine learning

Econometric models

Curse of dimensionality

Machine learning methods

Curse of interpretability

Causality as one of the most important research fields in social sciences.

Usually no experimental/laboratory setup in economics. Without **randomized experiments** challenge to find causation.

Econometric solutions:

- Regression discontinuity design (RDD)
- Difference in difference (DD)
- ...

But economic research is developing fast.

Athey and Imbens (2017): ‘Machine learning methods provide important new tools to improve estimation of causal effects in high-dimensional settings, because in many cases it is important to flexibly control for a large number of covariates as part of an estimation strategy for drawing causal inferences from observational data.’

Back to slide 1

What can Data Science contribute to Economics?

Machine learning

- forecasting
- covariate selection in traditional econometric models
- ...

Natural Language Processing

- effect of press releases/social media contents on economic time series
- sentiment in newspaper articles and effect on investor behavior/bankruptcy (self-fulfilling prophecies)
- ...

Network analysis

- effect of social networks on employment and inequality
- networks of surviving companies
- ...