An interpretable machine learning workflow with an application to economic forecasting

Andreas Joseph

Bank of England

Joint work with Marcus Buckmann (BoE) and Helena Robertson (FCA)

11th ECB Conference on Forecasting Techniques, virtual, 15. June 2021

Pros & Cons of machine learning (ML) relative to 'standard' econometric approach

Pros

- Often higher accuracy
- Lower risk of misspecification
- Return richer information set

Cons

- Higher model complexity ("black box critique")
- Less analytical guarantees, e.g. risk of overfitting
- Often larger data requirement

The ML workflow

1. Comparison of model predictions ("horse race" if accuracy is the goal)

   $\Rightarrow$ Is there gain in using ML, should I continue?

2. Model decomposition into Shapley values

   $\Rightarrow$ Identify important features & uncover learned functional forms

3. Statistical testing: "Shapley regression"

   $\Rightarrow$ Establish confidence & standard *communication*

Forecasting setup

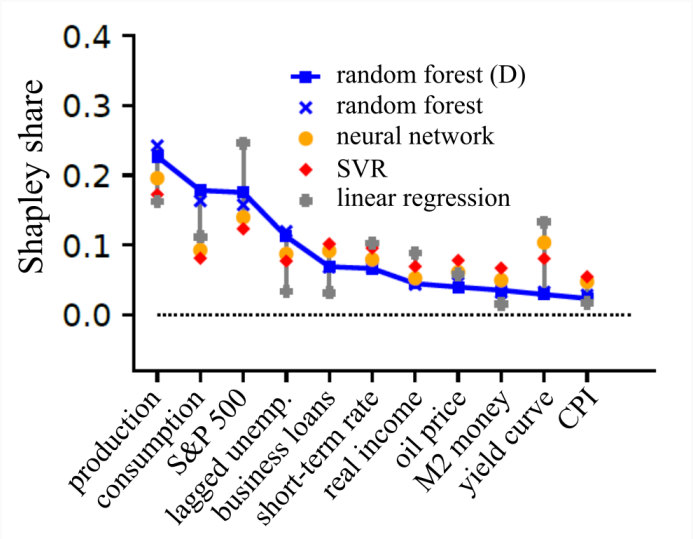
- Target: YoY change in US unemployment on a 1 year horizon
- Predictors: FRED-MD data base, McCracken and Ng (2016); 9 selected variables, lagged target
- Sample period: 1962:M2 - 2019:M11 (no Covid, no stress)
  - validation & training (yearly): Until 1989:M12
  - Testing: 1990:M1–2019:M11 (pseudo real-time), out-of-bag (full)
- Models:
  - *classical ML model:* Artificial neural networks (MLP), random forest, support vector regression (SVR)
  - *linear regressions:* OLS, Ridge, Lasso
  - *auto-regressions:* AR(1), AR(p) with $p \leq 12$ by AIC
- Hyper-parameters: (time series) 5-fold cross-validation, every 3 years
- Model-aggregation: Bootstrap aggregation ('bagging' over 100 draws)

Step 1: Horse race results

	Correlation	MAE	RMSE (normalised by first row)			
Time period	01/1990–11/2019		01/1990–11/2019	01/1990–12/1999	01/2000–08/2008	09/2008–11/2019
Random forest	0.609	1.000	1.000	1.000	1.000	1.000
Neural network	0.555	1.009	1.049	0.969	0.941	1.114**
Linear regression	0.521	1.094***	1.082**	1.011	0.959	1.149***
Lasso regression	0.519	1.094***	1.083***	1.007	0.949	1.156***
Ridge regression	0.514	1.099***	1.087***	1.019	0.952	1.157***
SVR	0.475	1.052	1.105**	1.000	1.033	1.169**
AR(p)	0.383	1.082(*)	1.160(***)	1.003	1.010	1.265(***)
AR(1)	0.242	1.163***	1.226***	1.027	1.057	1.352***

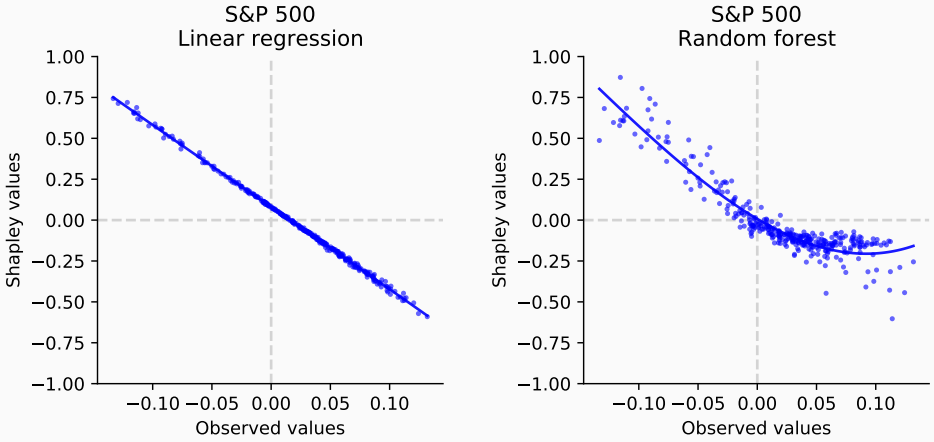
Models are ordered by decreasing RMSE on the whole sample with the errors of the random forest set to one. The forest's MAE and RMSE (full period) are 0.574 and 0.763, respectively. The asterisks indicate the statistical significance of the Diebold-Mariano test, comparing the performance of the random forest, with the other models,

Step 2-A: Shapley variable importance



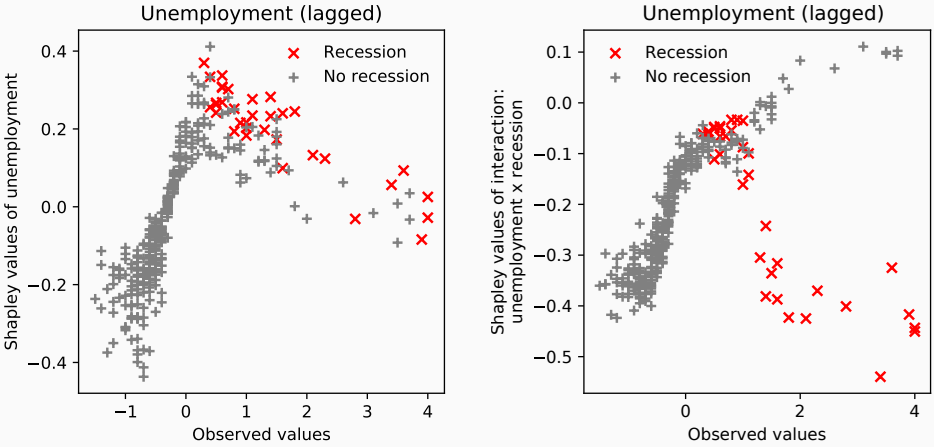
Fraction of absolute feature Shapley values within test period 1990–2019 for all full-information models.

Step 2-B: Learning non-linearities



Lines shows a polynomial fit of Shapley values (dots). Shapley values are computed on the out-of-bag predictions (look-ahead bias, but no model drift). Extreme values, below 2.5% and above 97.5% quantile, are excluded.

Step 2-C: Learning the economic cycle



Interaction between lagged unemployment and recessions (red) as learned by the random forest. LEFT: Baseline model. RIGHT: Unemployment-recession interaction with a recession dummy in the model.

Step 3: Statistical inference and standard communication

	Random forest			Linear regression		
	β^S	p-value	Γ^S	β^S	p-value	Γ^S
Industrial production	0.626	0.000	−0.228***	0.782	0.000	−0.163***
S&P 500	0.671	0.000	−0.177***	0.622	0.000	−0.251***
Consumption	1.314	0.000	−0.177***	2.004	0.000	−0.115***
Unemployment (lagged)	1.394	0.000	+0.112***	2.600	0.010	+0.033***
Business loans	2.195	0.000	−0.068***	2.371	0.024	−0.031**
3-month treasury bill	1.451	0.008	−0.066***	-1.579	1.000	−0.102
Personal income	-0.320	0.749	+0.044	-0.244	0.730	+0.089
Oil price	1.589	0.018	−0.040**	-0.246	0.624	−0.052
M2 Money	0.168	0.363	−0.034	-4.961	0.951	−0.011
Yield curve slope	1.952	0.055	+0.029*	0.255	0.171	+0.132
CPI	0.245	0.419	−0.024	-0.790	0.673	−0.022

Shapley regression of random forest (LEFT) and linear regression (RIGHT) for the forecasting predictions between 1990–2019. Significance levels: *p<0.1; **p<0.05; ***p<0.01.

Take-away messages

- We propose an interpretable ML workflow
  1. Model test evaluation (e.g. "horse race")
  2. Shapley decomposition of individual predictions
  3. Shapley regression for statistical inference
- Perform macro forecasting exercise of US unemployment
- ML models outperform conventional ones and endogenously learn
  - nuanced, meaningful and stable functional forms
  - to identify different points in the business cycle (recessions vs normal times)

In summary, this approach effectively addresses the black box critique and opens the door to many other applications.

Thanks for listening

contact: andreas.joseph@bankofengland.co.uk.

The machine learning (ML) setting

Everything here is about supervised learning, i.e. minimising an error

$$\min_{\theta} \ \mathbb{E}_{\Omega}\left[||y - \hat{f}_{\theta}||_{l}\right] .$$

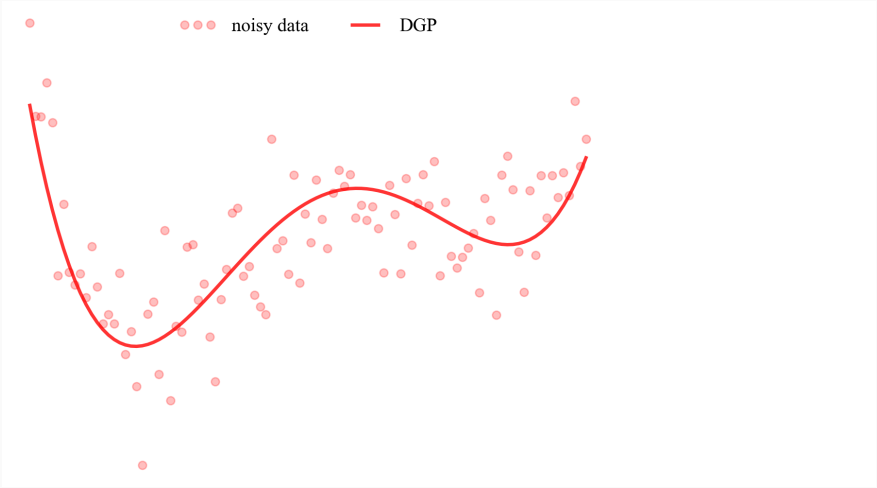
However, many aspects can be transferred to unsupervised or reinforcement learning.

Examples of supervised ML models

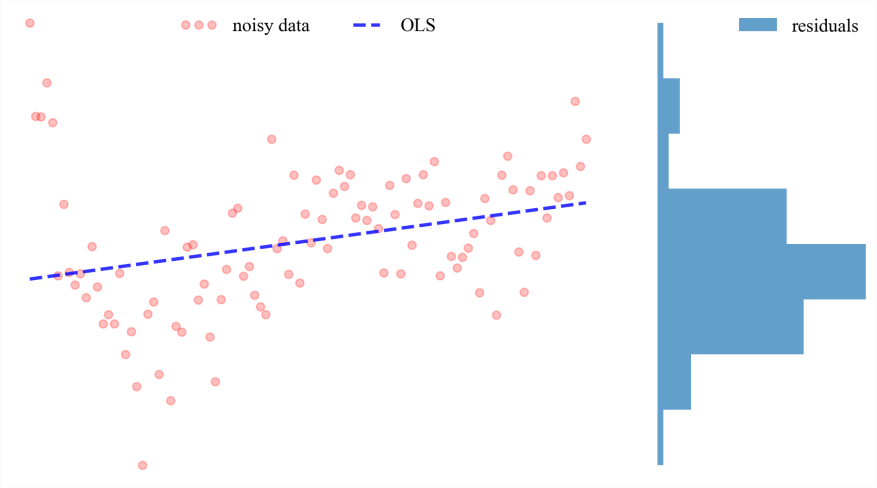
- Artificial neural networks: MLP, LSTM, CNN, etc.
- Support vector machines (SVM) with different kernels
- Tree-based models: decision tree, random forest, boosted trees, etc.

All of those are universal function approximators, meaning they can learn most relations of interest given enough data and appropriate levels of noise.

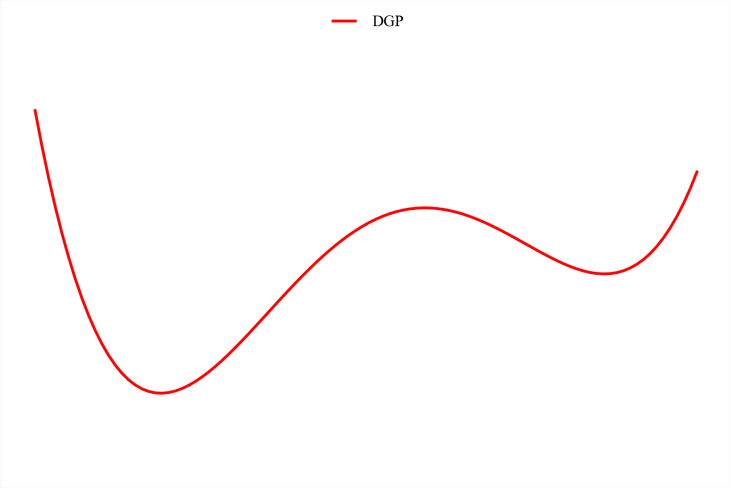
Toy example: unknown data data generating process (DGP)



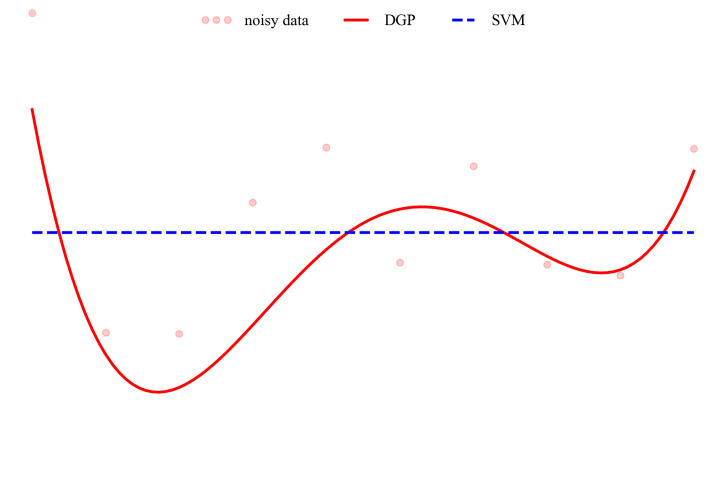
Toy example: naïve linear model



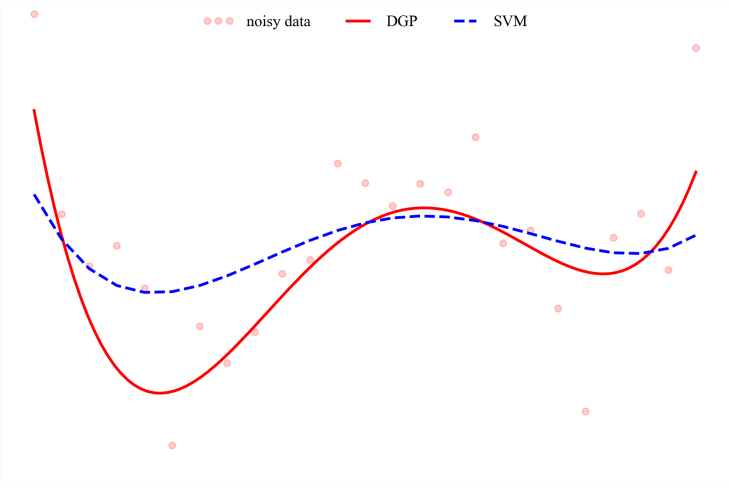
Toy example: ML error convergence



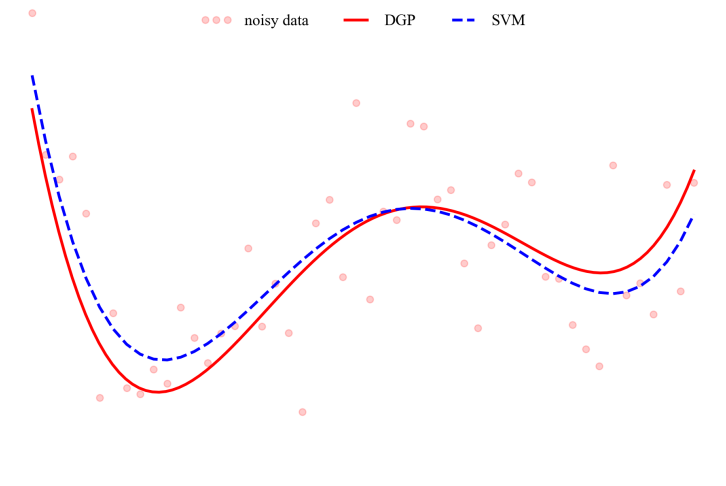
Toy example: 10 observations



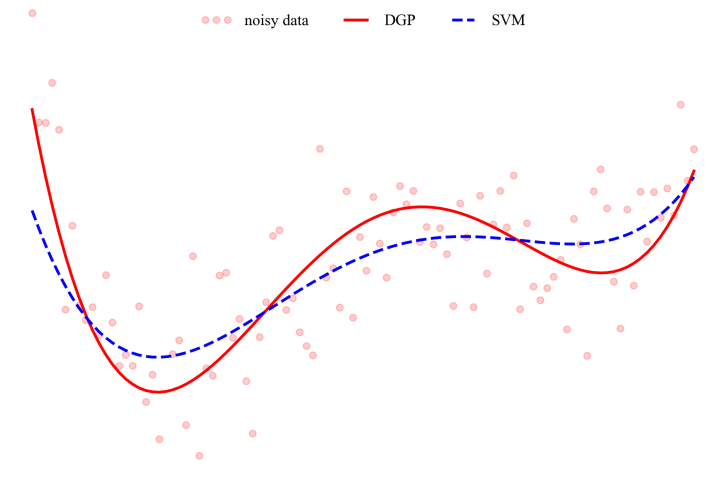
Toy example: 25 observations



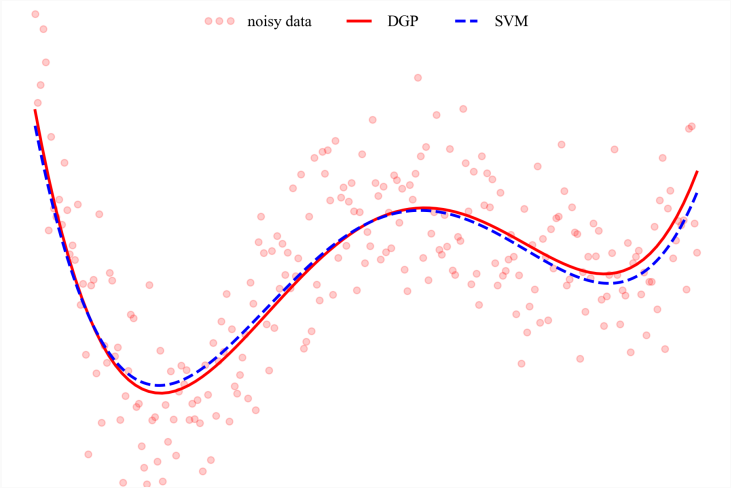
Toy example: 50 observations



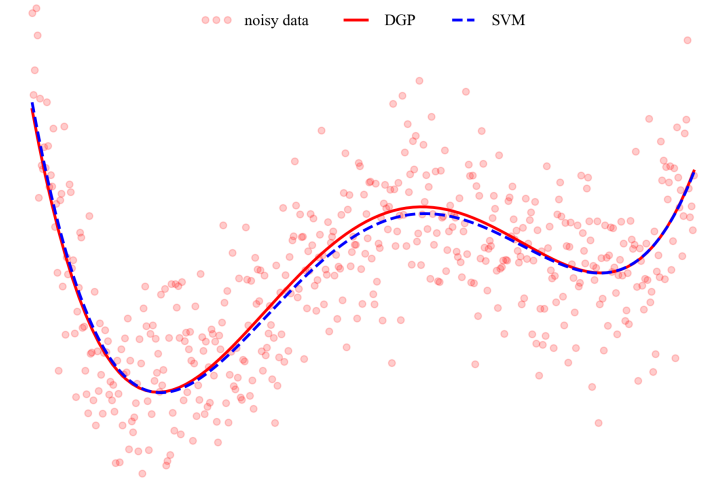
Toy example: 100 observations



Toy example: 250 observations



Toy example: 500 observations



The linear regression model (LR)

$$(I) : \hat{f}(x_i) = x_i\hat{\beta} = \sum_{k=0}^{n} x_{i,k}\hat{\beta}_k + \hat{\epsilon} \quad \text{with} \quad (II) : \mathcal{H}_0^k : \beta_k = 0 \tag{1}$$

- Workhorse of econometric analysis
- Special: *local* and *global* model inference ($\hat{\beta} = const.$)
- Widely accepted to be interpretable (if not too many regressors)
- Belongs to class of additive local variable attributions

$$\Phi(x_i) \equiv \phi_0 + \sum_{k=1}^{n} \phi_k(x_i) = \hat{f}(x_i) \tag{2}$$

Shapley values as analogy between game theory and (ML) models

	Cooperative game theory	Machine learning
n	Players	Predictors / variables
$\hat{f}/\hat{y}$	Collective payoff	Predicted value for one observation
S	Coalition of players	Group of predictors in model
Source	Shapley (1953)	Štrumbelj and Kononenko (2010)
		Lundberg and Lee (2017)

Model Shapley decomposition: $\hat{f}(x_i) = \phi_0 + \sum_{k=1}^{n} \phi_k^S(\hat{f}; x_i)$

Why Shapley values? Because they are the only attribution scheme which is *local, linear, exact, respects the null, is consistent, and allows for interactions* (Agarwal et al., 2019).

Shapley regression (SR) for statistical inference (Joseph, 2019)

Auxiliary inference analysis on $\hat{f}$ in the space of Shapley values:

$$y_i = \sum_{k=0}^{n} \phi_{ki}^S \hat{\beta}_k^S + \hat{\epsilon}_i \qquad \text{with} \qquad \mathcal{H}_0^k(\Omega) : \beta_k^S \leq 0 \tag{3}$$

Universality: $\hat{f}$ can be any model.

Interpretation: $\hat{\beta}^S$ measures the alignment of model components with the target.

Validity: Eq. 3 relates to generated regressors (Pagan (1984)) imposing minor conditions. Inference generally only valid on test set (standard in ML) and some consideration on convergence rates (cross-fitting helpful, Chernozhukov et al. (2018)).

Shapley share coefficients (SSC)

Normed summary statistic for the importance of x_k to the model $\hat{f}$ within a region Ω.

$$\Gamma_k^S(\hat{f},\Omega) \equiv \left[sign(\hat{\beta}_k) \left\langle \frac{|\phi_k^S(\hat{f})|}{\sum_{l=1}^m |\phi_l^S(\hat{f})|} \right\rangle_\Omega \right]^{(*)} \in [-1,1]$$

$$\overset{\hat{f}(x)=x\hat{\beta}}{=} \hat{\beta}_k^{(*)} \cdot \left\langle \frac{|(x_k - \langle x_k \rangle)|}{\sum_{l=1}^m |\hat{\beta}_k (x_l - \langle x_l \rangle)|} \right\rangle_\Omega \tag{4}$$

3 parts: **sign** (alignment of x_k and y), **size** (model fraction attributed to x_k) and **significance level** of $\hat{\beta}_k^S$ against $\mathcal{H}_0^k(\Omega)$.

$\Gamma_k^S(\hat{f},\Omega)$ is proportional to the coefficient of the linear model in the LR case.

Full recipe

Detour: Shapley values in cooperative game theory

- How much does player A contribute a collective payoff f obtained by a group of n? (Shapley, 1953).
- Observe payoff of the group with and without player A.
- Contribution depends on the other players in the game.
- All possible coalitions S need to be evaluated.

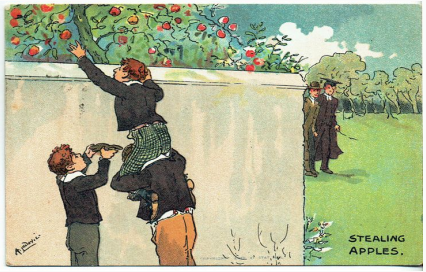
$$\phi_A = \sum_{S \subseteq n \setminus A} \frac{|S|!(|n| - |S| - 1)!}{|n|!} [f(S \cup \{A\} - f(S)] \tag{5}$$

$2^{|n|-1}$ coalitions are evaluated.

Computationally complex!

Intuitive Shapley value example: the Victorian bad boys

- Three siblings (strong [S], tall [T] & smart [M]) set off to nick some apples A (pay-off) from the neighbour's tree
- For each sibling, sum over marginal contribution to coalitions of one and two
- So, the Shapley value of the strong sibling [S] is then:



Source: 6oxgangsavenueedinburgh

$$\phi_S = \frac{1}{6}[A(S) - A(\emptyset)] + \frac{1}{6}[A(T,S) - A(T)] + \frac{1}{6}[A(M,S) - A(M)] + \frac{1}{3}[A(T,M,S) - A(T,M)] \tag{6}$$

Back

Numerical calculation of Shapley component for a math. model

The Shapley value of a feature is the weighted sum of marginal contributions to all possible coalitions of other features (players):

$$\phi_k^S(\hat{f}, x_i) = \sum_{S \subseteq \mathcal{C} \setminus \{k\}} \frac{|S|!(n - |S| - 1)!}{n!} \left(\hat{f}(x_i | S \cup \{k\}) - \hat{f}(x_i | S) \right) \tag{7}$$

$$= \sum_{S \subseteq \mathcal{C} \setminus \{k\}} \omega_S \left(\mathbb{E}_b[\hat{f}(x_i) | S \cup \{k\}] - \mathbb{E}_b[\hat{f}(x_i) | S] \right) \tag{8}$$

$$\text{with} \qquad \mathbb{E}_b[\hat{f}(x_i) | S] \equiv \int \hat{f}(x_i) \, \mathrm{d}b(\bar{S}) = \frac{1}{|b|} \sum_b \hat{f}(x_i | \bar{S}) \tag{9}$$

"Excluded" features are integrated out over background b, which is an informative dataset determining ϕ_0. E.g. training dataset or sample of untreated population.

There are some challenges (and solutions) to the calculation of (1)–(3).

Challenges in calculating model Shapley values

- Computational complexity: Generally intractable for large feature sets ($n!$ in 1)
  $\Rightarrow$ *Solutions:*
  - Coalition sampling
  - Feature grouping: important and 'others'
  - Model specific algorithms (e.g. Lundberg et al. (2018))

- Feature dependence: Equation 8 assumes independence
  $\Rightarrow$ *Solutions:*
  - Use exact method for trees and compare
  - Calculate higher-order terms of Shapley-Taylor index (Agarwal et al., 2019) and compare relative magnitudes
- Expectation consistency: Integration in (9) can break consistency
  $\Rightarrow$ *Solutions:* When comparing models, their background values ϕ_0 need to coincide (or close). Mostly the case in practical applications. See Joseph (2019).

Back

SR properties (proofs in Joseph (2019))

- SR identical to LR in case of LR (reassuringly the wheel was not reinvented)
- Inference only strictly valid locally within input region Ω (non-linearity of ML models)
- SR coefficients $\hat{\beta}^S$ gauge the learning process of $\hat{f}$:
  - $\mathcal{H}_0^k$ rejected: useful information contained x_k
  - And $\mathcal{H}_1^k$ *not* rejected: x_k robustly learned (perfect alignment, asymptotic limit)
  - Generally, $\hat{\beta}_k^S > / < 1$ measure under/over-reliance on x_k, respectively
- $\beta^S \in \{0, 1\}^m$ only possible true values, corresponding to the "no-signal" ($\mathcal{H}_0^k$) or "signal" ($\mathcal{H}_1^k$) cases, respectively
- SR allow to control for different error structures within ML models.
- SR coefficients $\hat{\beta}^S$ not really useful for communication (no scale information).

ML inference recipe (full details in Joseph (2019))

1. Cross-validation, training and testing of model $\hat{f}$
2. Model decomposition
   - 2.1 Shapley value decomposition $\Phi^S(\hat{f})$ [Eq. 7]
   - 2.2 (optional) Mapping of Φ^S to desired decomposition $\hat{\Psi}(\Phi^S(\hat{f}))$
3. Model inference
   - 3.1 Shapley regression [Eq. 3] with appropriate standard errors.
   - 3.2 Assessment of model bias and component robustness based on $\hat{\beta}^S$ over region Ω: [VEINs may be appropriate (Chernozhukov et al. (2018))]

     Robustness: $\mathcal{H}_0^c : \{\hat{\beta}_c^S = 0 | \Omega\}$ rejected and $\mathcal{H}_1^c : \{\hat{\beta}_c^S = 1 | \Omega\}$ not rejected for individual components

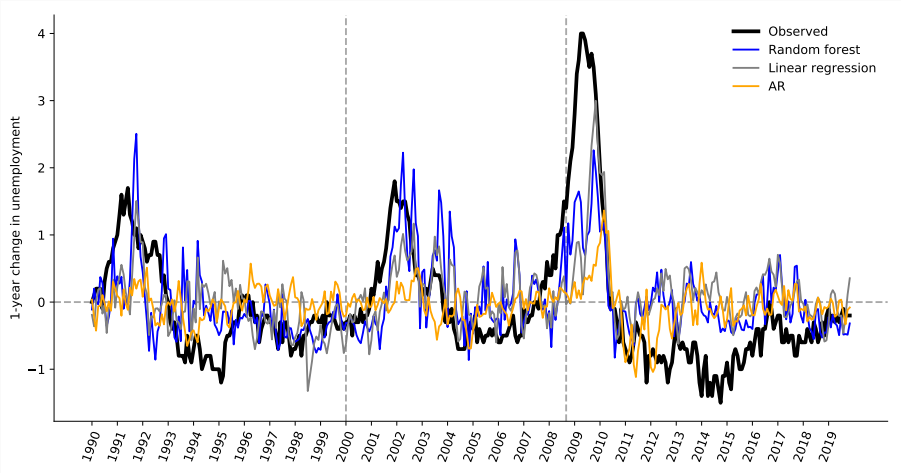
     Unbiasedness: $\mathcal{H}_1^c : \{\hat{\beta}_c^S = 1 | \Omega\}$ not rejected $\forall c \in \{1, \ldots, C\}$, or inclusion condition
   - 3.3 Calculate Shapley share coefficients (SSC) $\Gamma^S(\hat{f}, \Omega)$ [Eq. 4] and their standard errors

Variable selection: Capture different economic channels

Variable	Transformation	Name in Source
Unemployment (target+lag)	changes	UNRATE
3-month treasury bill	changes	TB3MS
Slope of the yield curve	changes	-
Real personal income	log changes	RPI
Consumption	log changes	DPCERA3M086SBEA
Industrial production	log changes	INDPRO
S&P 500	log changes	S&P 500
Business loans	second order log changes	BUSLOANS
CPI	second order log changes	CPIAUCSL
Oil price	second order log changes	OILPRICEx
M2 Money	second order log changes	M2SL

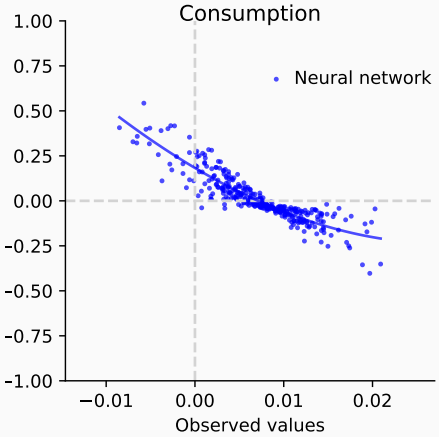
Transformations as suggested in McCracken and Ng (2016), using quarterly changes.

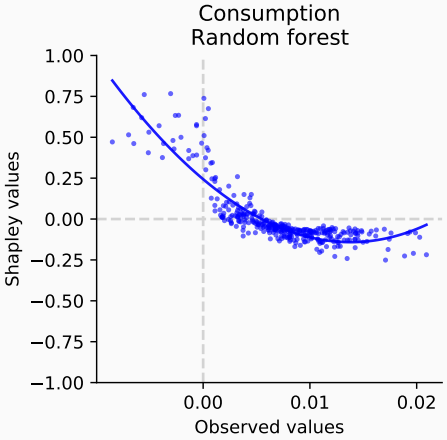
Eyeballing the horse race



Observed and predicted 1-year change in unemployment for the whole forecasting period. Source: McCracken and Ng (2016) and authors' calculation.

Differing functional forms for consumption





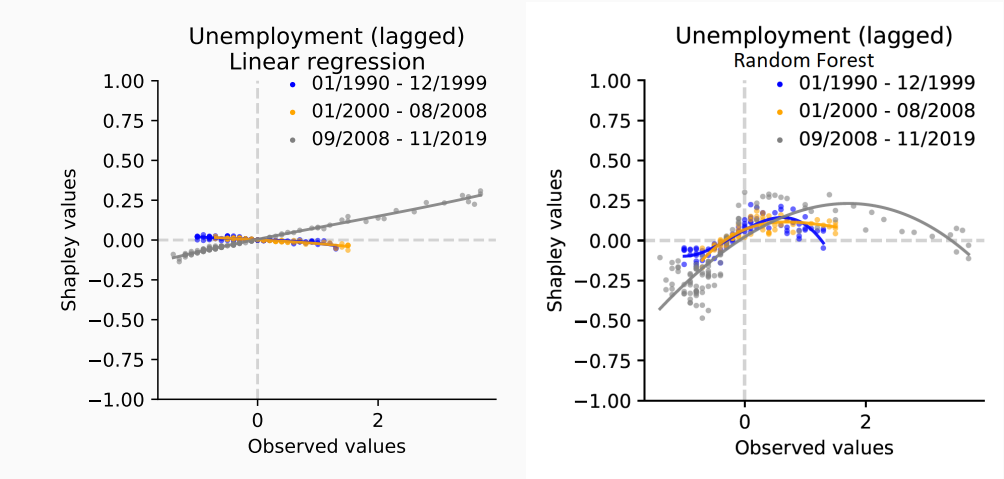
Lines shows a polynomial fit of Shapley values (dots). Shapley values are computed on the out-of-bag predictions (look-ahead bias, but no model drift). Extreme values, below 2.5% and above 97.5% quantile, are excluded.

Robustness analysis of horse race

	Random forest	Neural network	Linear regression	SVR	AR(p)	AR(1)
Training set size (in months)						
max (baseline)	1.000	1.049	1.082	1.105	1.160	1.226
60	1.487	1.497	1.708	1.589	2.935	1.751
120	1.183	1.163	1.184	1.248	1.568	1.257
240	1.070	1.051	1.087	1.106	1.304	1.198
Change horizon (in months)						
3 (baseline)	1.000	1.049	1.082	1.105	1.160	1.226
1	1.077	1.083	1.128	1.148	-	-
6	1.043	1.111	1.142	1.162	-	-
9	1.216	1.321	1.251	1.344	-	-
12	1.345	1.278	1.336	1.365	-	-
Bootstrapped models						
no	1.000	1.179	1.089	1.117	1.160	1.226
100 models	-	1.049	1.082	1.105	-	-

Performance for different parameter specifications using RMSE divided by the RMSE of the random forest in the baseline set-up. Source: Authors' calculation.

Step 2: Learned functional forms (II): Stability



Lines shows a polynomial fit of Shapley values (dots). Shapley values are computed on the out-of-bag predictions (look-ahead bias, but no model drift). Extreme values, below 2.5% and above 97.5% quantile, are excluded.

References i

Agarwal, A., Dhamdhere, K., and Sundararajan, M. (2019). A new interaction index inspired by the taylor series. *arXiv e-prints*, 1902.05622.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Demirer, M., Duflo, E., and Fernandez-Val, I. (2018). Generic machine learning inference on heterogenous treatment effects in randomized experiments. *NBER Working Paper Series*, (24678).

Joseph, A. (2019). Parametric inference with universal function approximators. *arXiv preprint arXiv:1903.04209*.

Lundberg, S., Erion, G., and Lee, S. (2018). Consistent individualized feature attribution for tree ensembles. *ArXiv e-prints*, 1802.03888.

References ii

Lundberg, S. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774.

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

Pagan, A. (1984). Econometric issues in the analysis of regressions with generated regressors. *International Economic Review*, 25(1):221–47.

Shapley, L. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2:307–317.

Štrumbelj, E. and Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11:1–18.