ECONOMIC MODELLING & MACHINE LEARNING

A PROOF OF CONCEPT

NICOLAS WOLOSZKO, OECD

NAEC – APRIL 16 2019

OECD
BETTER POLICIES FOR BETTER LIVES

Economic forecasting with Adaptive Trees

I Motivation

II Method

III Results

IV Perspectives

I. Motivation

Linear models are constrained where economic complexity is concerned

- Non-linearities
- Structural change

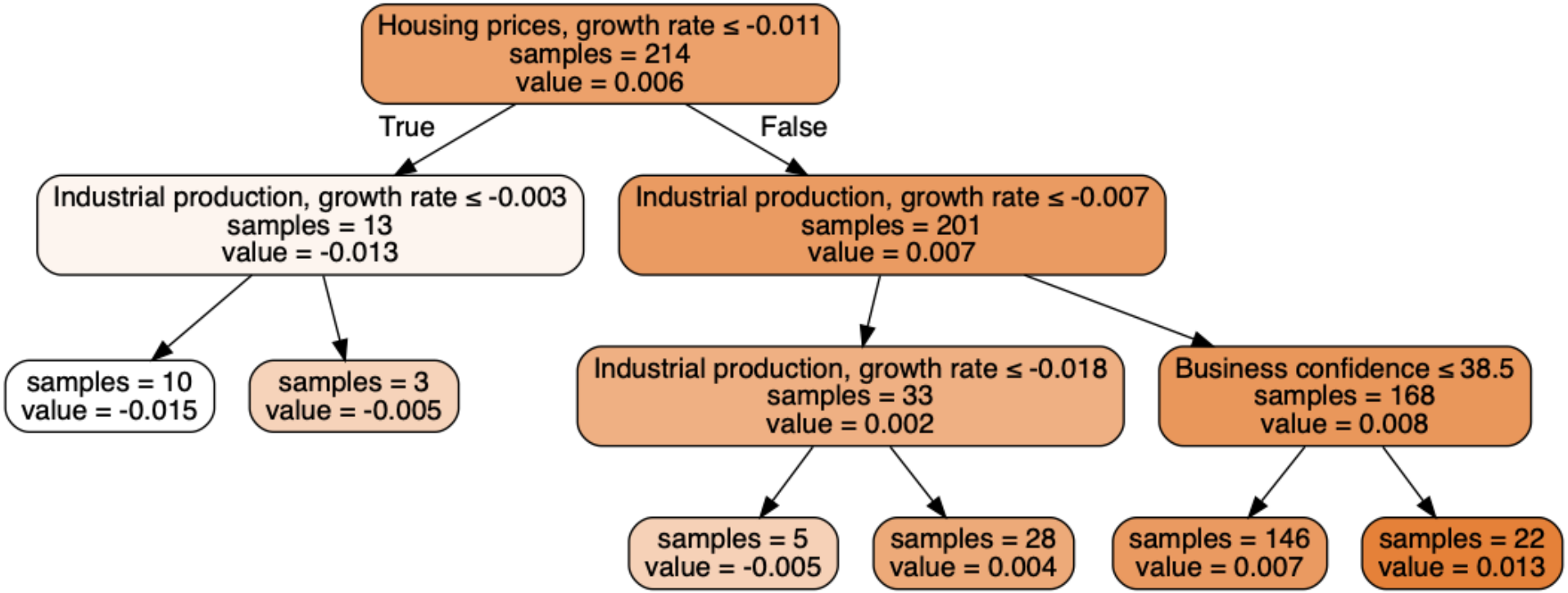
Machine learning can provide relevant tools to tackle these challenges

- Modelling without a model: no prior knowledge is required
- Algorithms designed to capture complex patterns in the data
- Use of cross-validation to prevent over-fitting

II. ADAPTIVE TREES: A METHOD FOR ECONOMIC FORECASTING

1. Regression trees
2. Gradient Boosted Trees
3. Adaptive Boosting

1. Regression trees



At each node, the algorithm selects the splitting variable + splitting point that minimises sub-group variance of GDP growth

2. Gradient Boosted Trees

- Simple regression trees lack robustness, hence the resort to ensemble methods.

- **Gradient Boosted Trees** (Freidman, 2002):

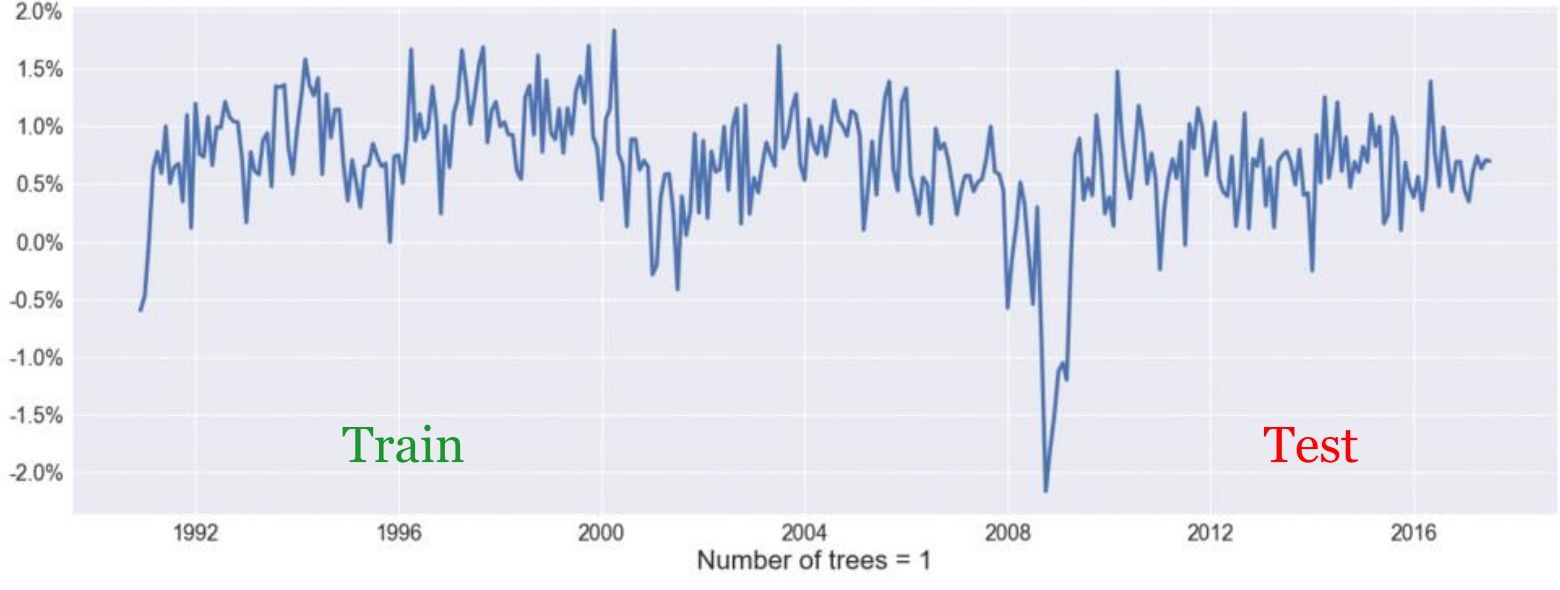
$$F_m(x) = F_{m-1}(x) + \nu h_m(x)$$

$h_m(x)$: regression tree trained on the residual from $F_{m-1}(x)$

- Gives more and more weight to observations harder to predict

2. Gradient Boosted Trees

XGBoost trained on US data, GDP growth shifted by 6 months



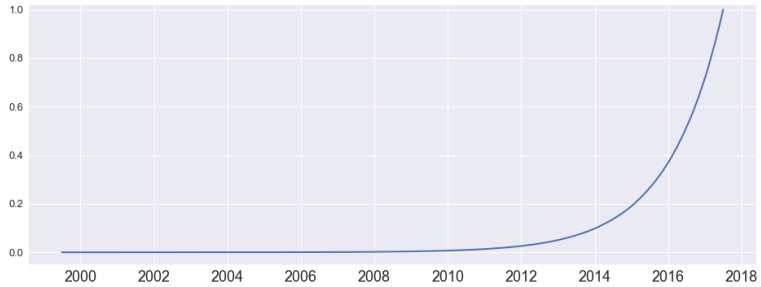
Gradient Boosted Trees end up giving more weight to observations harder to predict (larger residuals)

3. Adaptive Trees

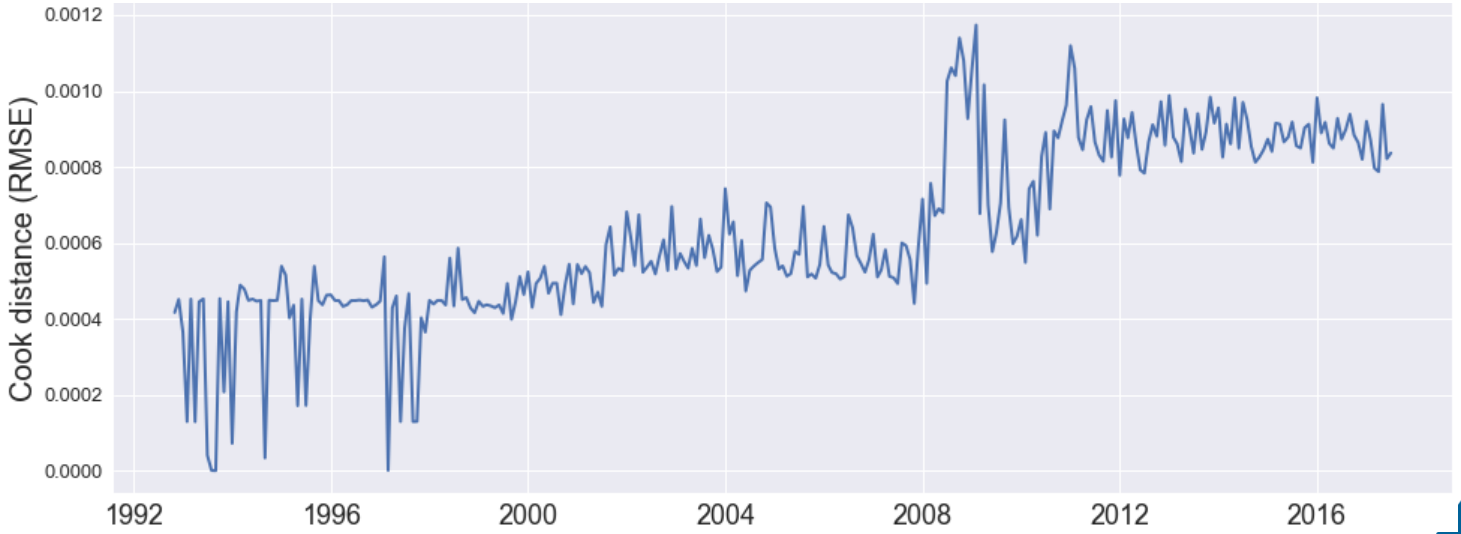
Adaptive Trees = Gradient Boosting + increasing *ex ante* observation weights

***Ex ante* observation weights:**

$$w(t) = e^{-\gamma(\frac{t}{N}-1)}$$



***Ex post* observation weights:**



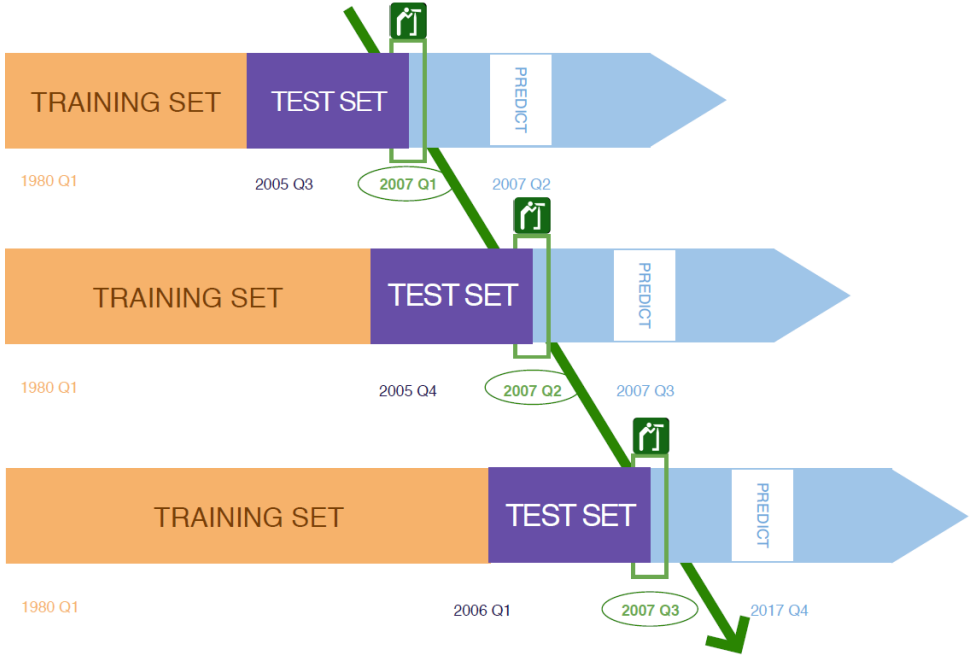
III. RESULTS

FORECAST OF GDP GROWTH

Setting of forecast simulations

- Simulations in ***pseudo***-*real time* of a forecast of GDP growth in G6 countries

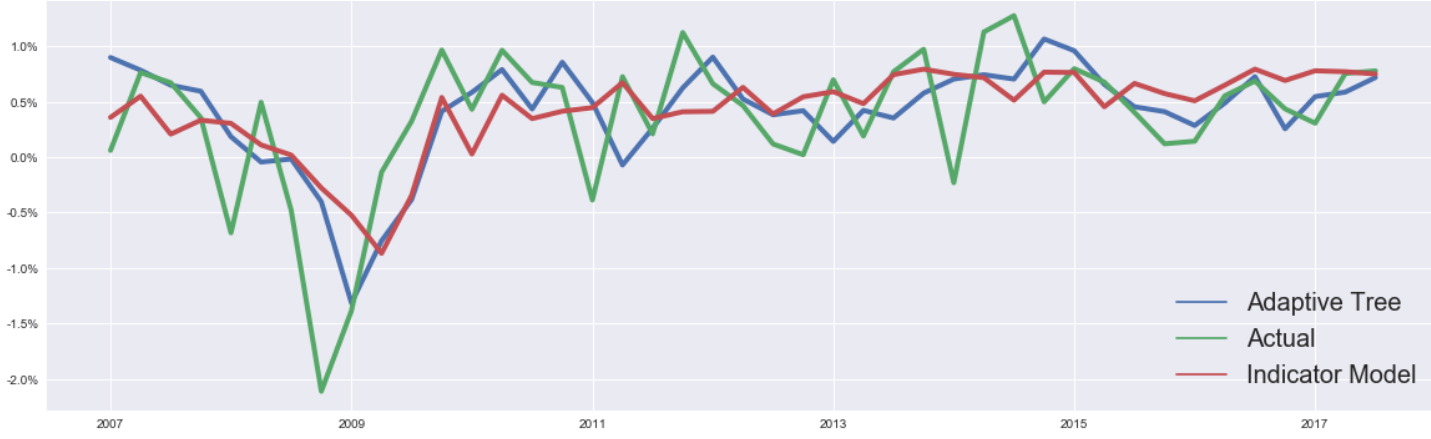
- Using the exact same data as benchmark OECD Indicator Model (housing prices, indutrial production, PMI…) so as to provide a *methodological benchmark*



Comparison with OECD Indicator Model

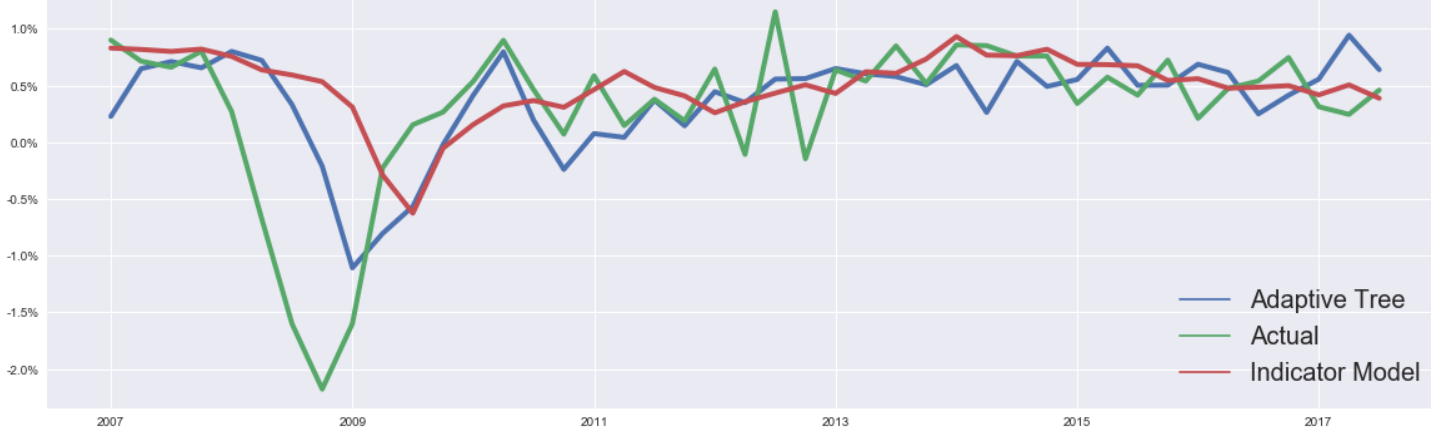
1. USA, M+3

Accuracy: + 5 %



2. UK, M+6

Accuracy: + 22 %



Perspectives

- The method could be extended to broader sets of variables, as it can be applied in high dimension
- That may include financial indicators or big data
- Machine learning also has promising applications in inference and causal analysis. Exisiting methods address non-linearities and causal heterogeneity

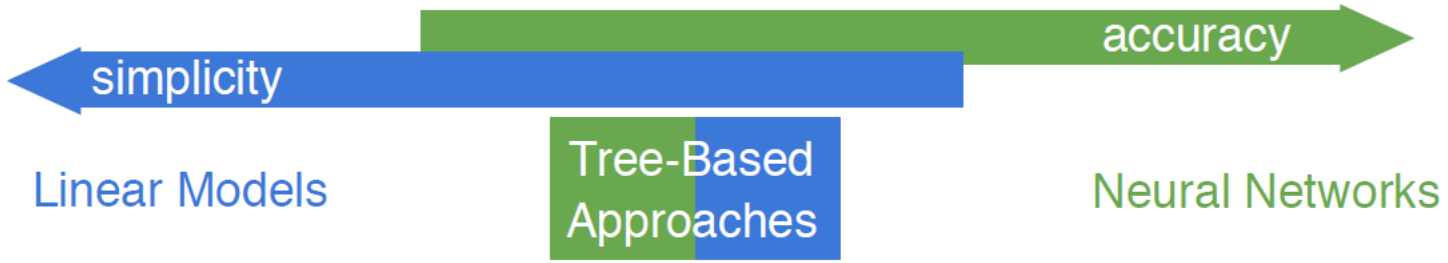
THANK YOU

OECD
BETTER POLICIES FOR BETTER LIVES

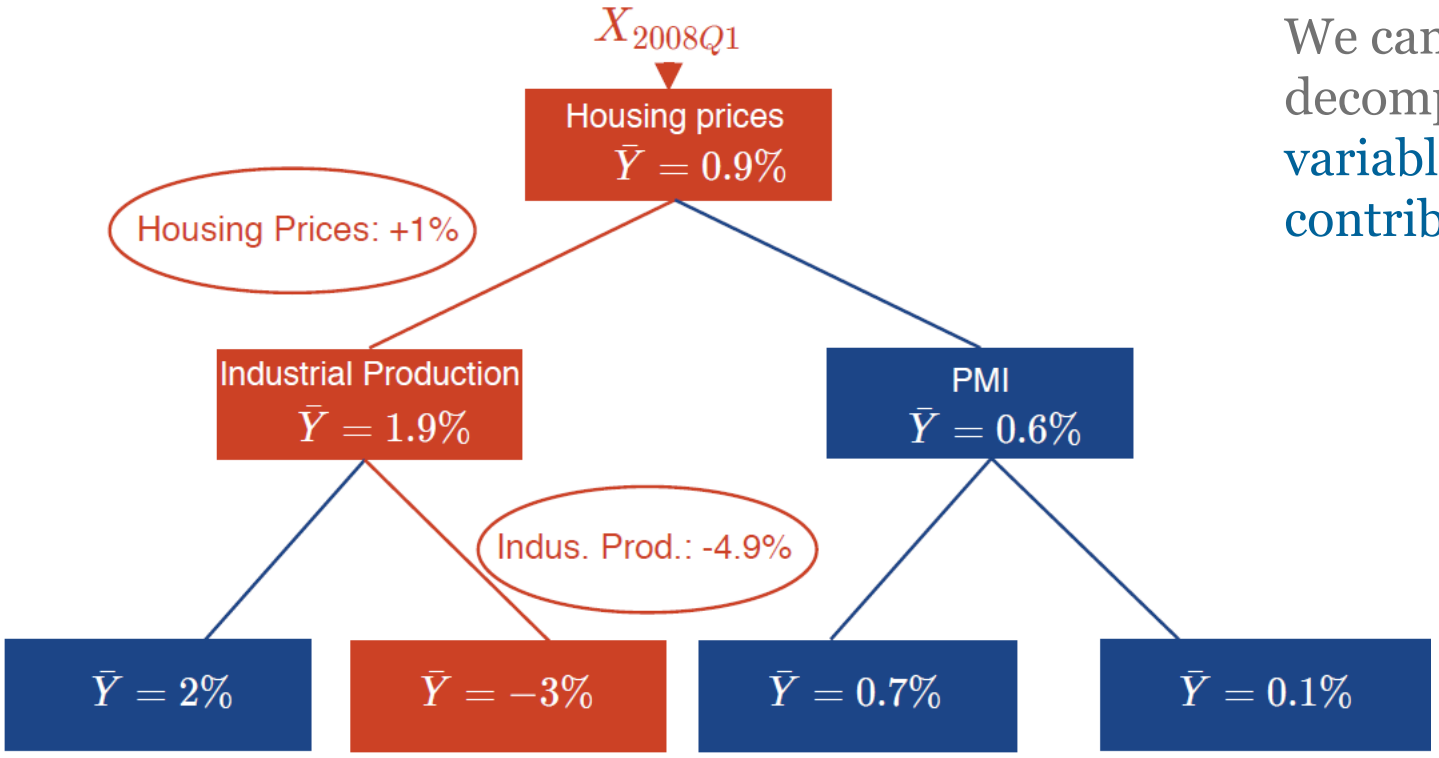
ADDITIONAL MATERIAL

Problem: interpretability

- Modelling complexity requires more complex models
- Trade off simplicity/accuracy:
  - Too much simplicity: fail to capture important variations
  - Too much complexity: fail to produce a sensible story



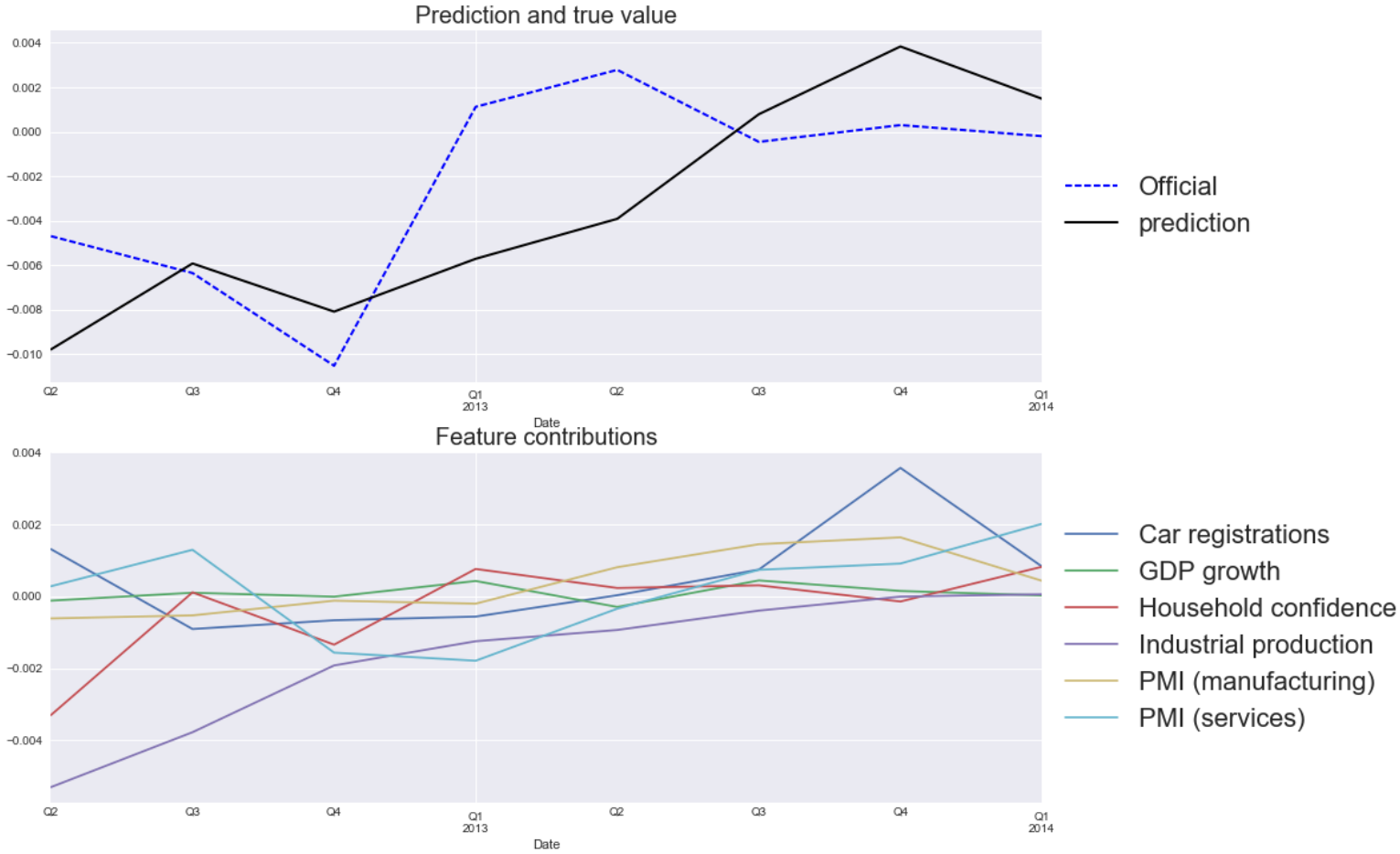
Interpretability



We can easily decompose in variable's contribution

$$\hat{Y} = 0.9\% + \sum Feature\ Contributions$$

Variable contributions, Italy M+3



Variable selection

- For each variable:
  - What relevant lag : M-1, M-2, M-12, M-24 ?
  - In level ? In growth rate ?

- Data-driven variable selection:
  - Based on variable importance
  - Variable importance: a variable is all the more important that it is **high in the tree**, close to the root
  - Accounts for multiple interactions (can keep a variable that is loosely correlated with the GDP but that provides relevant interactions. Ex: price of gold)

Complexity v. Bayesian econometrics

- In a regression with 10 variables, should we want to test all possible multiple interactions : 10^{10} possibilities

- With tree-based approaches, we explore all possible interactions with 120 variables

- Amount of prior knowledge:
  - Linear econometrics: we know the form of the relationship
  - Bayesian econometrics: we know the relationship can take any of the know forms
  - Machine learning: we know nothing