Concept Tree: High-Level Representation of Variables for More Interpretable Surrogate Decision Trees

Xavier Renard[1] – Nicolas Woloszko[2] – Jonathan Aigrain[1] – Marcin Detyniecki[134]

[1] **AXA** - [2] **OECD** - [3] **Sorbonne University** - [4] **Polish Academy of Science**
Paris, France Warsaw, Poland

2019 Bank of England and King's College London joint conference on Modelling with Big Data and Machine Learning: Interpretability and Model Uncertainty, London, UK

Outline

1. Global, Post-Hoc & Model Agnostic Interpretability in Machine Learning
2. Concept Trees for more interpretable surrogate trees
3. Use-Case, Data & Experimentation
4. Conclusion

Home

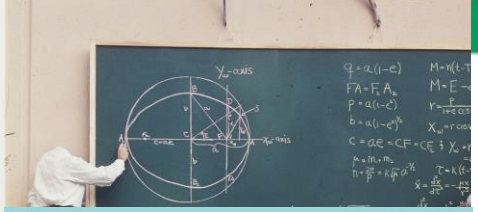
The New York Times Magazine

Share 197

Can A.I. Be Taught to Explain Itself?

As machine learning becomes more powerful, the field's researchers increasingly find themselves unable to account for what their algorithms know — or how they know it.

By CLIFF KUANG NOV



Why AI Is Still Waiting For Its Ethics Transplant

BUSINESS CULTURE GEAR IDEAS SCIENCE

SHARE

SHARE 894

TWEET

COMMENT

EMAIL

SCOTT ROSENBERG 11.01.17 08:45 AM

WHY AI IS STILL WAITING FOR ITS ETHICS TRANSPLANT

Kate Crawford, cofounder of AI N

Don't Make AI Artificially Stupid in the Name of Transparency

SHARE

DON'T MAKE AI ARTIFICIALLY STUPID IN THE NAME OF TRANSPARENCY

Le Monde.fr EDITION GLOBALE

INTERNATIONAL POLITIQUE SOCIÉTÉ ÉCO CULTURE IDÉES

M PIXELS CHRONIQUES DES (RÉ)VOLUTIONS NUMÉRIQUES

VIE EN LIGNE JEUX VIDEO BANC D'ESSAI

EDITION ABONNÉS

Ethique et intelligence artificielle : récit d'une prise de conscience

The New York Times

Opinion

OP-ED CONTRIBUTOR

Artificial Intelligence's 'Black Box' Is Nothing to Fear

By Vijay Pande

Jan. 25, 2018

Leer en español

Forbes Billionaires Innovation Leadership Money Consumer Industry

1,944 views | Jun 4, 2018, 05:14am

Is Explainability Enough? Why We Need Understandable AI

Rumman Chowdhury Contributor

co-authored with McCree Lake, Talent & Organization Strategy Senior Manager at Accenture

Artificial Intelligence is quickly becoming ubiquitous in personal and professional lives in ways we both observe and others we don't see as readily. Artificial Intelligence is used to influence life-changing decisions, such as whether or not you get hired to that dream job, who you will date, and whether or not you'll be approved for a loan for your first home. However, we have little insight into how critical decisions are made with AI. As a result, there is increasing demand (and legislation) to ensure the influence of these technologies is understood.

Does a Fair Algorithm Actually Look Like?

IDEAS SCIENCE

A FAIR ALGORITHM OOK LIKE?

Machine Learning Interpretability

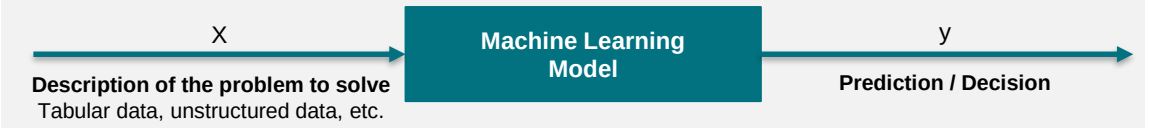
Evaluate machine learning models beyond accuracy scores

What has been learned by the model ?

Where is the model {correct ; wrong } ?

Why a particular prediction has been made ?

What can be done to change the prediction ?



Usually aggregated accuracy score (accuracy, AUC, f1-score)

Is the model robust?

How does the model behave in areas with few data ?

Is the model fair ?

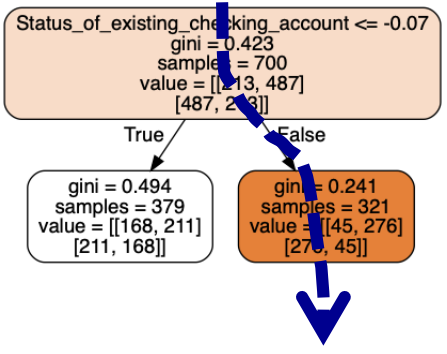
Is the model causal?

How the prediction is affected by small changes in input ?

Trade-off Interpretability-Accuracy

Accurate Machine Learning Models are **not Interpretable** (usually)

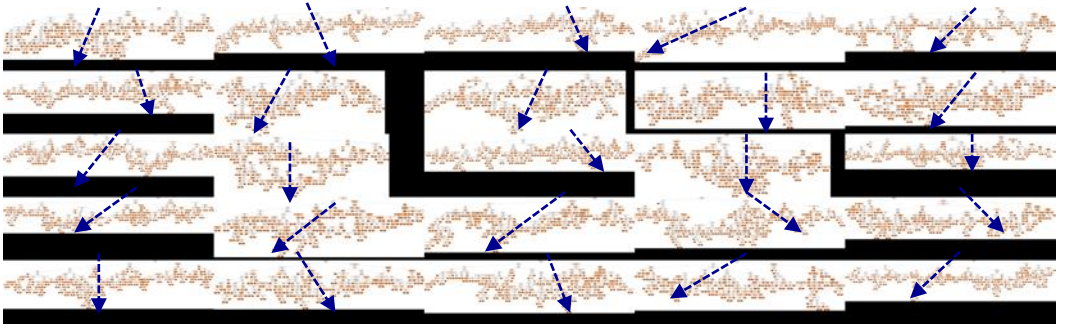
Simple machine learning model
e.g. Decision Tree
→ Interpretable
→ Less accurate



Decision: credit or not

One **path** → simple **explanation**

Blackbox machine learning model
e.g. Random Forest, CNN (Inception…)
→ Uninterpretable
→ More Accurate



One path → One decision **by** base model
Final decision: aggregation of each decision

Explanation: no consensus

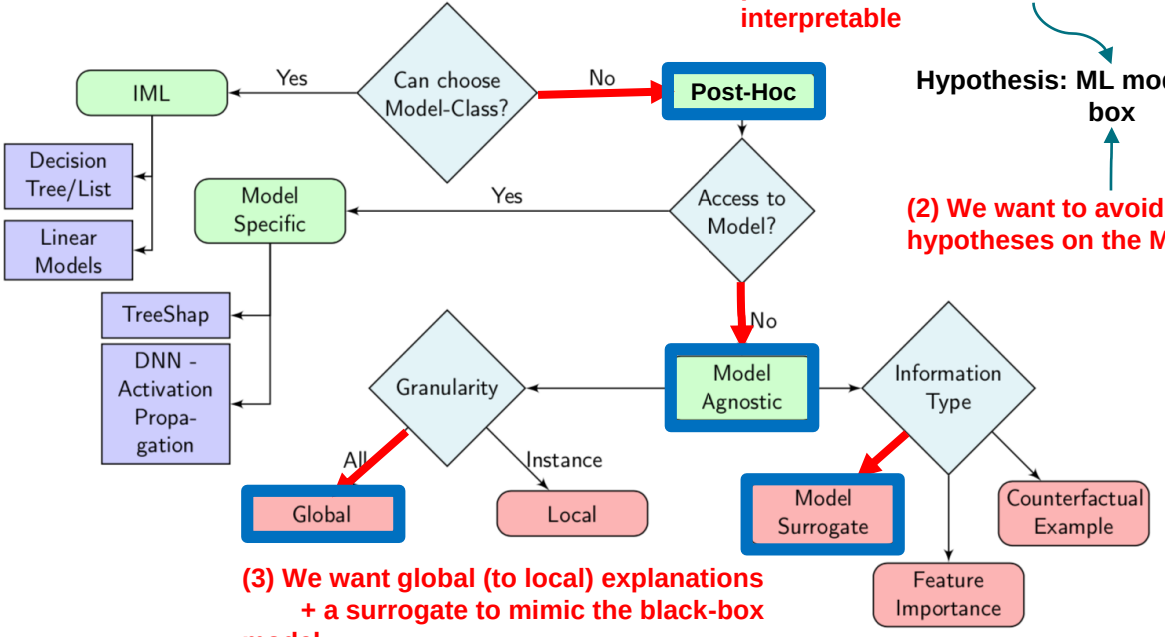
Global, Post-Hoc & Model Agnostic Interpretability in Machine Learning

Interpretability Approaches Considered in this Work

(1) We assume the best ML model in performance for a task is not necessarily interpretable

Hypothesis: ML model = black-box

(2) We want to avoid hypotheses on the ML model



(3) We want global (to local) explanations + a surrogate to mimic the black-box model

Principles of Post-Hoc, Model Agnostic Surrogates

Agnostic surrogate model pipeline

1) Train a “black-box””model $\hat{f}$ from the data (x, y)
2) Get predictions from the black-box model $\hat{y} = \hat{f}(x)$
3) Train an interpretable surrogate model S on $(x, \hat{y})$ to mimic the predictions of the black-box model
4) Gain insights from the trained surrogate model

Two performance metrics given a loss function L (say, MSE):

- **Fidelity** = $L(\hat{y}_S, \hat{y})$
- **Accuracy** = $L(\hat{y}_S, y)$

CONFIDENTIALITY LEVEL

Local v. Global explanation

Granularity of the explanation

- **Local interpretability**: explain a single prediction

➔ Ex: SHAP

➔ **(1) train the model, (2) predict instances, (3) interpret predictions**

➔ Drawback: possible lack of consistency between explanations

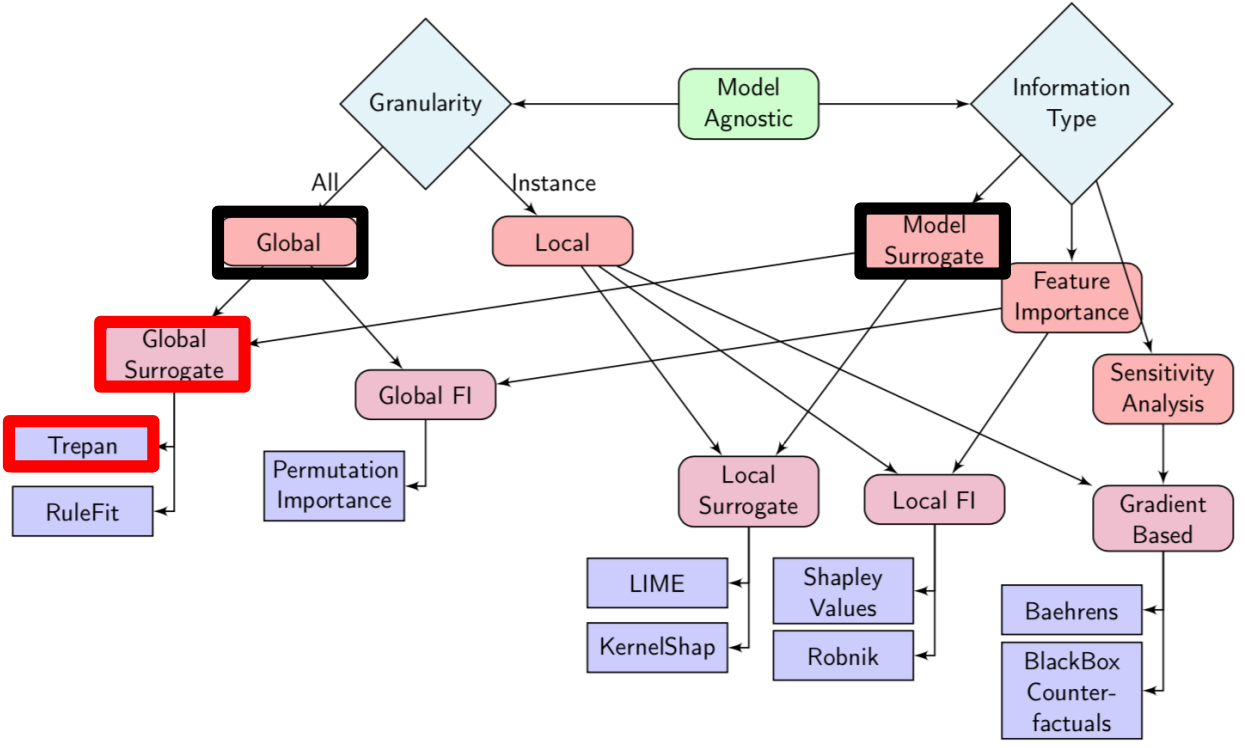
- **Global interpretability**: explain the whole ML model

➔ Ex: Feature importances

➔ **(1) train the model, (2) interpret the model**

➔ Advantage: gaining insights about what the model has learned during the training

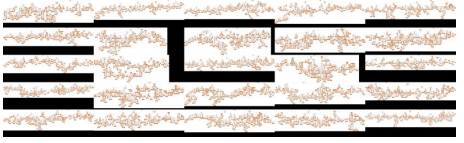
Overview of the ML Interpretability field



TREPAN: Model-Agnostic and Post-Hoc Surrogate

(Craven *et al.*, 1996)

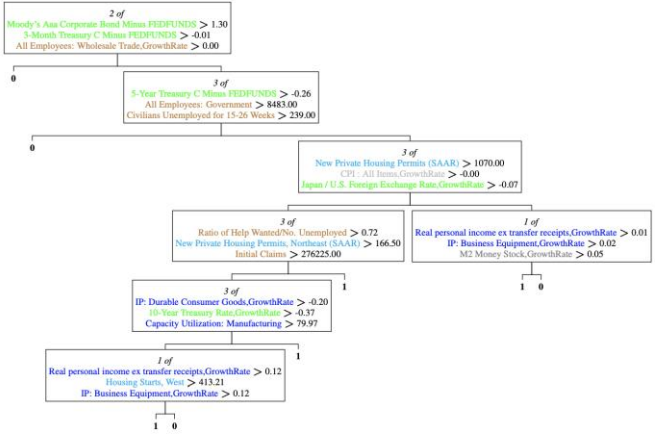
- Mimic the global predictive behaviour of any model
- Decision Tree trained to replicate predictions
  - High fidelity to the model



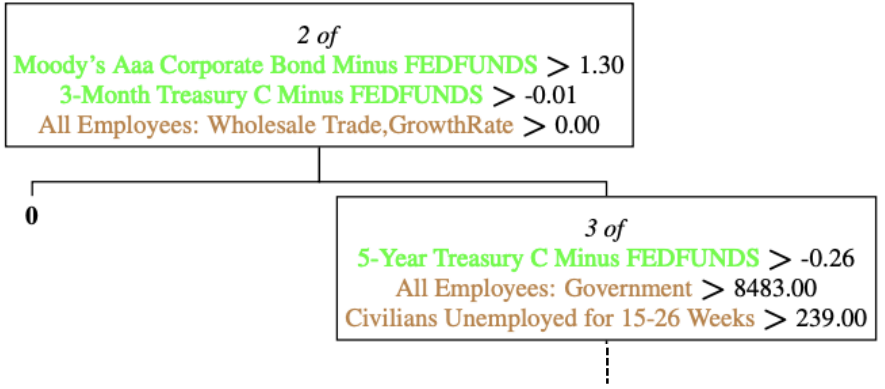
From black-boxes (Random Forests…) to a faithful Interpretable decision tree

- **Draw instances in low density subspaces**
  - Min number of instances for split selection
  - Important contribution for fidelity

- Expand most promising nodes to increase fidelity
  - $f(n) = reach(n) * \left(1 - fidelity(n)\right)$



A particularity: the ***m-of-n*** split rules



2-*of*-3 split rule: **pick 2** variables **among the 3** selected for the node

TREPAN uses ***m-of-n* split rules** at each node

- Inspired by *id2-of3* (Murphy & Pazzani, 1991)

$$2-of-\{x_1, \neg x_2, x_3\} \equiv (x_1 \vee \neg x_2) \wedge (x_1 \vee x_3) \wedge (\neg x_2 \vee x_3)$$

- **Improves fidelity**: learn more complex decision boundaries
- **Lower human-interpretability**: need to assess $\binom{n}{m}$ variable combinations

Concept Tree

Improve Interpretability with Groups of Dependent Variables

Impact of Dependent Variables on Interpretability

- Real world tabular datasets often involve **subgroups of dependent or related variables**
  - Similar meaning or similar origin (e.g. various measures of unemployment)
  - Result of multiple transformations to a variable (e.g. multiple lags of time series)

- **Known challenge** in ML and Interpretability

  Eg. robust interpretation of variable importance

  (Buhlmann et al., 2013; Gregorutti et al., 2017; Strobl et al., 2008; Tolosi & Lengauer, 2011)

- Subgroups of dependent variables **can impact interpretability**
  - *Issue*: **arbitrary selection** of one variable among a subgroup by surrogate models
  - *Opportunity*: **better, more comprehensive explanations**

Concepts: Grouping Dependent Variables into High-Level Representation

- **Grouping dependent variables** into high-level representation of variables

Big Picture		**Fine Grained Details**
High-Level Explanations	$\longrightarrow$	Low-Level Explanations
Subgroups of dependent variables		**Raw Variables**

- **Concepts** are **embedded into surrogates** by **constraining** their training

- Related works: group variables in ML
  - **Interpretability** (Kim et al., 2015 ; Kim et al., 2017; Ghorbani et al., 2019)
  - Some **ML sparse models**
    - Group-lasso (Yuan & Lin, 2006)
    - Sparse-group-lasso (Simon et al., 2013)
    - Subspace clustering (Vidal, 2011)

Concepts: Grouping Dependent Variables into High-Level Representation

- **Concept** definition: **subset of dependent features**
  - **Concepts** co-exist to form the set of **concepts**
  - Every feature belongs to one single **concept** only

- Build a **concept** with dependent features
  - **Expert-defined**: domain knowledge (e.g. documentation of FRED-MD data: 134 vars / 8 categories)
  - **Automatically-defined**: clustering algorithm based on Pearson correlation

$$\left|\rho(x_j, x_{j'})\right| \geq \epsilon \mid \forall j \in c_k \rightarrow c_k = c_k \cup j'$$

Any dependence measure

(every) feature in the concept considered

Feature candidate for inclusion

User-defined threshold

Concept considered for inclusion

Concept Tree: Embedding Concepts into Interpretable Surrogates

Model Architecture

- Objective: global explanations of complex ML models with interpretable surrogates
  - Decision Trees are interpretable: graphical structure, sparse subset of feature, hierarchy

- **Surrogate decision tree: TREPAN** (Craven & Shavlik, 1996)
  - Reassessment of its relevance up until recently (Sarkar et al., 2016)

- Our proposition: **Concept Tree**
  - Take advantage of TREPAN's efficient learning procedure to mimic a black-box classifier
  - In particular the additional instance-drawing procedure

Concept Tree: Embedding Concepts into Interpretable Surrogates

Node Learning to leverage **concepts**

- ***m-of-n* split rules** learning **constrained by concepts**
  - Whole feature space no longer considered
  - Rather subsets defined by **concepts**

- At each node, all **concepts** are assessed
  - The one with the **highest information gain** is selected

- Embed Concepts in a Surrogate Decision Tree is expected to **improve interpretability**
  - **Information is organized** at **concept** level: ease the interpretation of a node (*m-of-n* split rules)
  - Take into account **feature dependence** when producing explanations

Use Case, Data & Experimentation

Data

FRED-MD database

- Publicly released **macroeconomic** database
  - 134 monthly U.S. indicators
  - Our classification target: binary evolution of the *civilian unemployment rate*

- **Provide 8 subgroups** of dependent variables based on **domain knowledge**

Examples of subgroups in the database

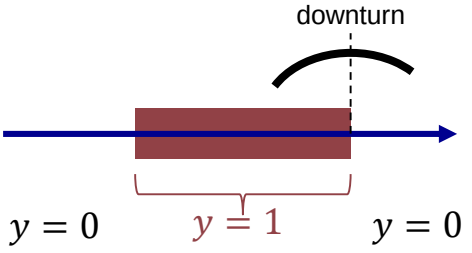
Concept 1: Output and Income	Concept 2: Labor Market	Concept 5: Money and Credit
Real Personal Income	Civilian Labor Force	Total Reserves of Depository Institutions
Real personal income ex transfer receipts	Civilian Employment	Commercial and Industrial Loans
IP: Consumer Goods	Civilian Unemployment Rate	Total Consumer Loans and Leases Outstanding
...	...	...

Raw variables belonging to the subgroups

Experimentation setting

Downturn prediction

- Downturn detection in US economy using variables from FRED-MD data (McCracken & Ng, 2016)
- Downturns defined from the NBER Business Cycle Chronology
- The target y_t is equal to 1 if a downturn occurs between t and $t + 12$ (months)



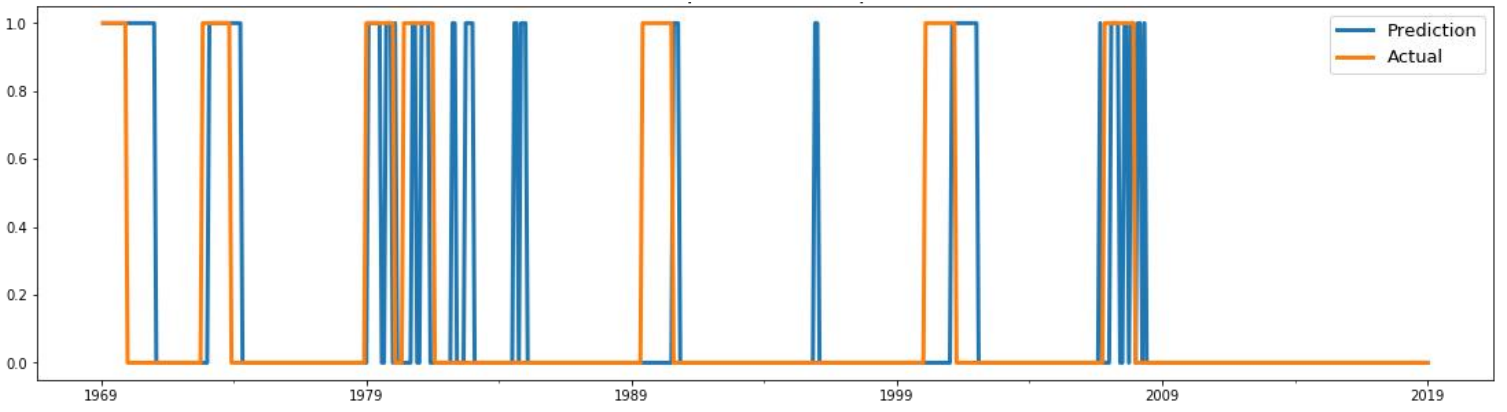
Black-box predictor

- Gradient Boosted Trees algorithm.
- Tree depth and learning rates are grid-searched (optimising the F2 score) using time series cross-validation.
- This is *not* a forecast simulation, but a retrospective exercise (we do not take release delays into account and use current series)

Experimentation setting

Performance in predicting downturns is reasonably good

Out-of-sample downturn prediction



- Recursive out-of-sample predictions
- Prediction performance is reasonably good with some false positives

Experimentation setting

Interpretation

- Objective: shedding light on patterns are captured by the downturn prediction black-box algorithm.

- **Concept Tree** vs. **TREPAN**
  - 2 flavors: Concept Tree-Expert and Concept Tree-Correlated (automatically defined concepts)

- **How to compare two surrogate models** ?
  - Out-of-sample **accuracy** = $accuracy(y, \hat{y}_{surrogate})$
  - Out-of-sample **fidelity** = $accuracy(\hat{y}_{black\ box}, \hat{y}_{surrogate})$
  - “Expert” assessment of interpretability

CONFIDENTIALITY LEVEL

Experimentation

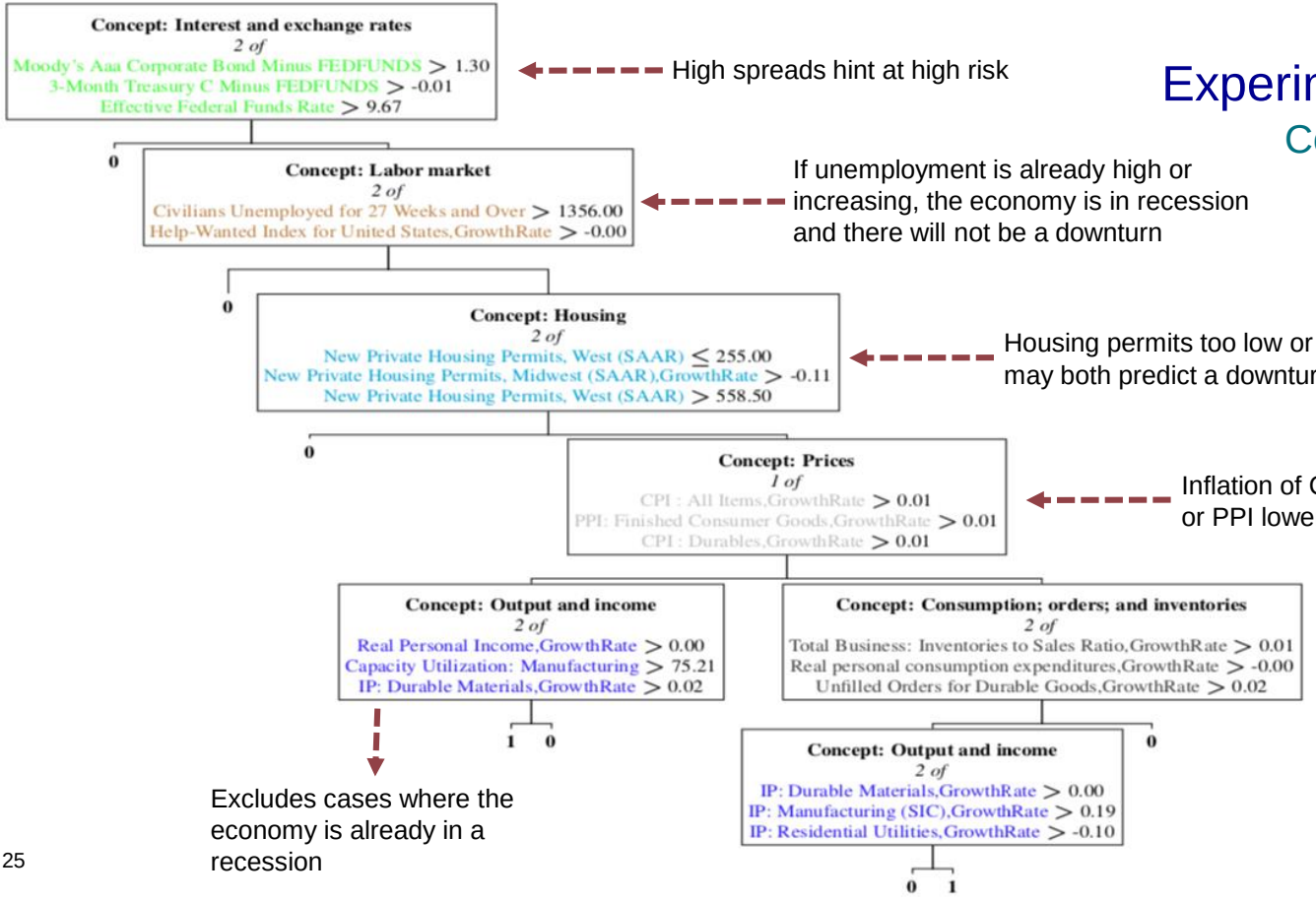
Quantitative Results

- Concept Tree **at least as accurate** as TREPAN

Algorithm	Concept Type	Split Rule	Surr. Accuracy	Surr. Fidelity
Concept Tree	Expert		$83\% \pm 29\%$	$85\% \pm 27\%$
Concept Tree	Correlation	$1-of-1$	$82\% \pm 30\%$	$85\% \pm 28\%$
TREPAN	/		$84\% \pm 29\%$	$86\% \pm 26\%$
Concept Tree	Expert		$78\% \pm 32\%$	$83\% \pm 28\%$
Concept Tree	Correlation	$3-of-3$	$83\% \pm 27\%$	$84\% \pm 25\%$
TREPAN	/		$84\% \pm 29\%$	$89\% \pm 22\%$
Concept Tree	Expert		$77\% \pm 31\%$	$82\% \pm 26\%$
Concept Tree	Correlation	$5-of-5$	$84\% \pm 25\%$	$82\% \pm 25\%$
TREPAN	/		$82\% \pm 30\%$	$86\% \pm 22\%$

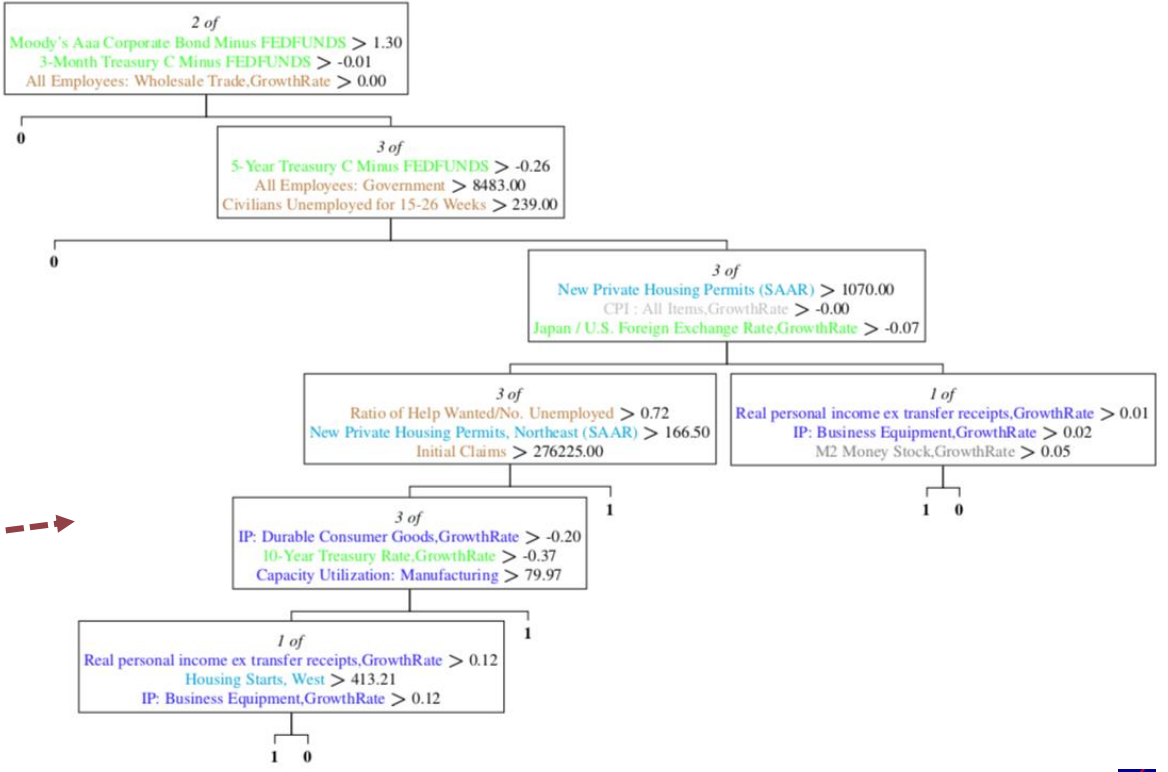
Experimentation

Concept Tree



Experimentation

TREPAN Tree



Decision rules from heterogeneous expert-defined concepts

Conclusion

Conclusion

- Discuss the important notion of **feature dependence in ML interpretability**
- Propose the idea of **concept** to tackle this issue
  - Expert-defined and automatic mode
- **Embed concepts in a TREPAN**, a notable global interpretable surrogate
  - Improved interpretability
  - No significant impact on surrogate fidelity
- Concept Trees **“better” structures information and enhances interpretability** at a constant complexity (quantity of information)
- Avenues for further research: implementing monotonicity, …

https://arxiv.org/abs/1906.01297

https://axa-rev-research.github.io/interpretability

xavier.renard@axa.com
nicolas.woloszko@oecd.org