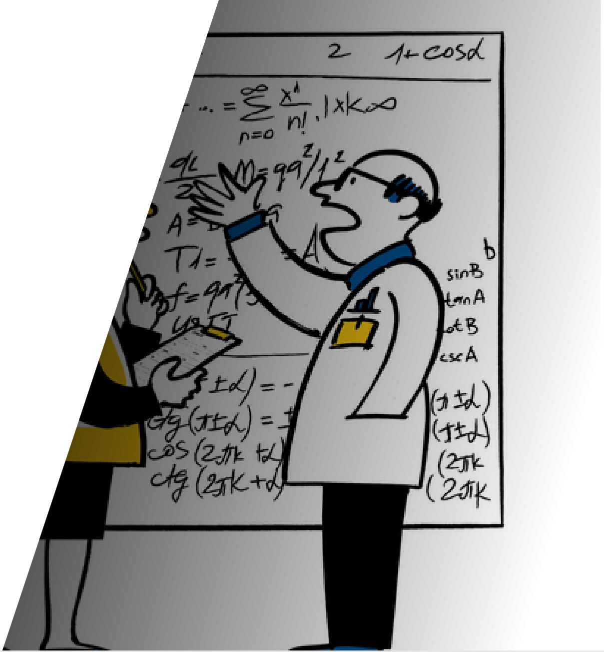
Global explanation of machine learning with sensitivity analysis

Francois Bachoc
Fabrice Gamboa
Max Halford
Jean-Michel Loubes
Laurent Risser

GdR MASCOT-NUM
Tuesday 10th March, 2020



Motivation

- Decisions are taken based on machine learning algorithms (most often black boxes).
- Used for recommendation systems, insurances, loans, human resources, education... but also other areas such as judicial systems, clinical diagnosis, security, political decisions...
- Understanding why decisions are taken is of paramount importance

Machine learning framework

- **Learning set**: $(X_1, Y_1), \dots, (X_n, Y_n)$ with distribution $\mathbb{P}$ learnt using empirical version $\mathbb{P}_n$
- **Parameter of interest**: $f^\star \in \arg\min \mathbb{E}_{\mathbb{P}}\{\ell(Y, f(X)) + \text{penalty}(f)\}$
- **Decision rule**: $\widehat{f}_n = \arg\min \frac{1}{n}\sum_{i=1}^n \{\ell(Y_i, f(X_i)) + \text{penalty}(f)\}$
- **Optimised** from a mathematical point of view and **generalised** for all new observations: $\widehat{Y} = \widehat{f}_n(X)$

Acceptability of AI

- Main assumption: PAC learning. The distribution of test samples is the same as the distribution used to learn the model.
- AI **generalises** the situation encountered in the learning sample to the whole population. It shapes the reality according to the learnt rule without questioning nor evolution.

Acceptability of AI requires that the algorithm should be **explainable** and **understandable**

Explainability

- A huge literature with exponential growth rate
- Several points of views:
  - **Local** explanation: fit locally a small regression model to understand local behaviours
  - **Global** explanation: rank the variables using importance scores (can be variable importances or Shapley values)
- Several scopes:
  - Explain individual predictions
  - Explain model behaviour on average

Our approach

- We propose to **stress the model** and study its response to controlled deformations of the input variables: **extension of sensitivity analysis**.
- Testing and stressing the algorithm (within the boundaries of its normal behaviour and without violation of the assumptions) to obtain understandability and robustness.
- Intuition: zoom into parts of the test set support and what says

Machine learning under stress

$f : \mathbb{R}^p \to \mathbb{R}$ on the input observation $X_i = (X_i^1, \ldots, X_i^p)$, and Y_i is the true output.
f is learnt using an independent test distribution

- Construct a **new sample** $\tilde{D}$ with approximately the same distribution $\mathbb{P}$ but with well chosen deformations
- Study the impact on the distribution of the algorithm

$$\mathcal{L}(f(\tilde{D})).$$

Why?

1. Quantify the specific influence of each one of the $p \geq 1$ variables.
2. Determine the global effect of each variable in the learning rule and how a particular variation of said variable affects the accuracy.

Twofold purpose: **Understand** how the predictions evolve when a characteristic of the observations is modified with the same distribution and **quantify the robustness** of the learnt algorithm.

Benefit: we don't need access to the model, only the predictions! Ideal for auditing.

Entropy projection: a theorem (1)

- **Kullback-Leibler information**. Let $(E, \mathcal{B}(E))$ be a measurable space and Q a probability measure on E. Kullback-Leibler information $KL(P, Q)$ is defined as equal to $\int_E \log \frac{\mathrm{d}P}{\mathrm{d}Q} \,\mathrm{d}P$, if $P \ll Q$ and $\log \frac{\mathrm{d}P}{\mathrm{d}Q} \in L^1(P)$, and equal to $+\infty$ otherwise.
- Let Φ be a measurable function representing the shape of the stress deformation.
- Let $\mathbb{P}_{\Phi,t}$ be the set of all probability measures P on $(E, \mathcal{B}(E))$ such that

$$\int_E \Phi(x) \,\mathrm{d}P(x) = t.$$

- $t_0 = \int_E \Phi(x) \,\mathrm{d}Q(x)$ (whenever it exists) be the parameter that represents no deformation.

Entropy projection: a theorem (2)

Theorem

The distribution under the stress modeled by Φ and quantified by t is the solution of

$$Q_t := \arg \inf_{P \in \mathbb{P}_{\Phi,t}} KL(P, Q)$$

$$Q_t = \frac{1}{n} \sum_{i=1}^{n} \lambda_i^{(t)} \delta_{X_i, \hat{Y}_i, Y_i}.$$

Explainability and stress model

1. First method to generate in a **fast and scalable way** counterfactual distributions without violation of the PAC assumption
2. Stress models to study the response of the algorithms to particular or rare events: **robustness with respect to stress conditions**
3. Explainability of the black-box models: from the reactions to certain type of stress, understanding the propagation of uncertainty in the algorithm: **Extension of Sensitivity Analysis to AI based models**

Outcome:

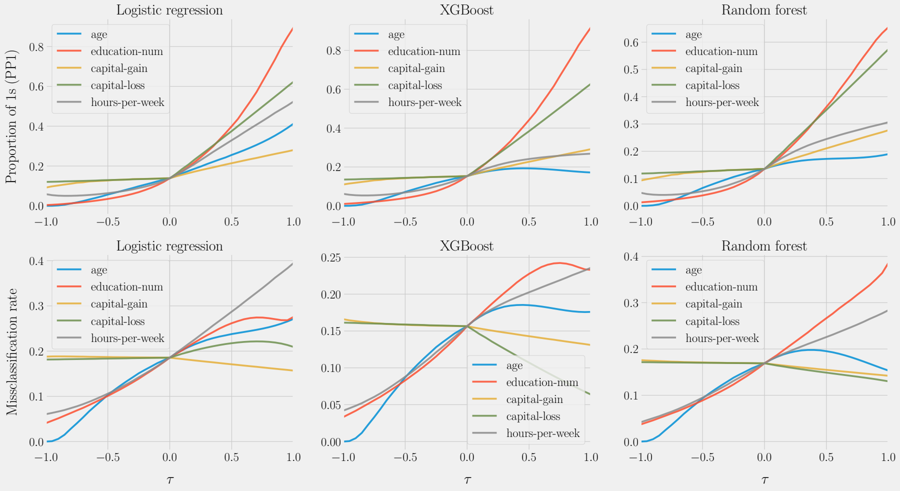
- Publications and a package actually tested by research engineers from DEEL
- Winner of CNRS innovation prize (2018)

Example: adult income (1)

Forecast a bank loan using $p = 14$ variables and $n = 32000$ observations.

- $Y = \begin{cases} 1 & \text{income exceeds \$ } 50.000/\textit{year} \\ 0 & \text{otherwise} \end{cases}$
- Features:
  - Age
  - Amount of education
  - Capital gain
  - Capital loss
  - Hours worked per week
  - etc.

Example: adult income (2)



Indicators for understanding a regression model

- Prediction score

$$\mathrm{M}_{p,\tau} = \frac{1}{N}\sum_{i=1}^{N}\lambda_i^{(p,\tau)} f_n(X_i),$$

- The variance criterion

$$\mathrm{V}_{p,\tau} = \frac{1}{N}\sum_{i=1}^{N}\lambda_i^{(p,\tau)} \left(f_n(X_i) - \mathrm{M}_{p,\tau}\right)^2$$

- The root mean square error (RMSE) criterion

$$\mathrm{RMSE}_{p,\tau} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\lambda_i^{(p,\tau)} \left(f_n(X_i) - Y_i\right)^2}$$

Example: Boston house prices

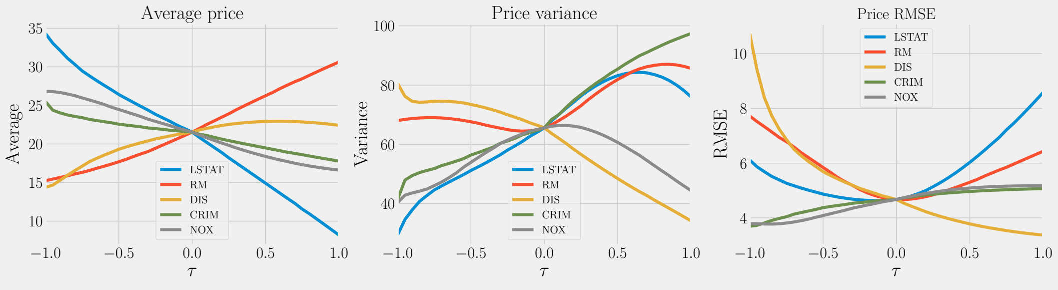
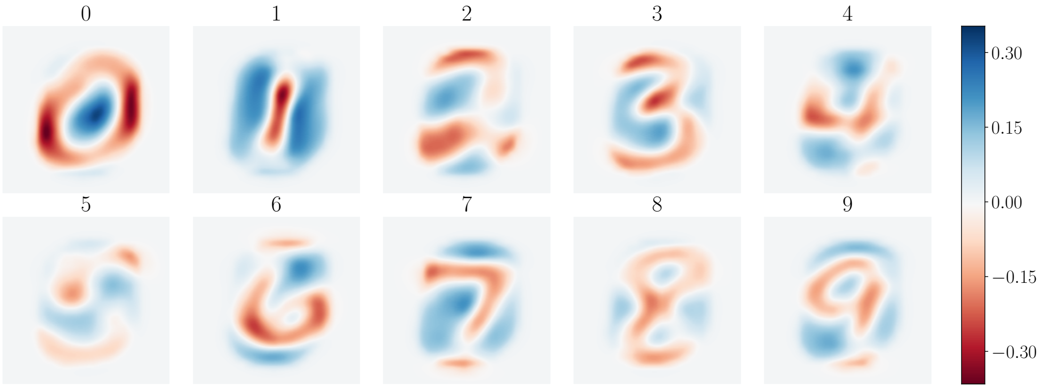


Figure: Results obtained on the *Boston Housing* dataset with Random Forests. The explanatory variable perturbation τ has the same signification as in Fig. 1.

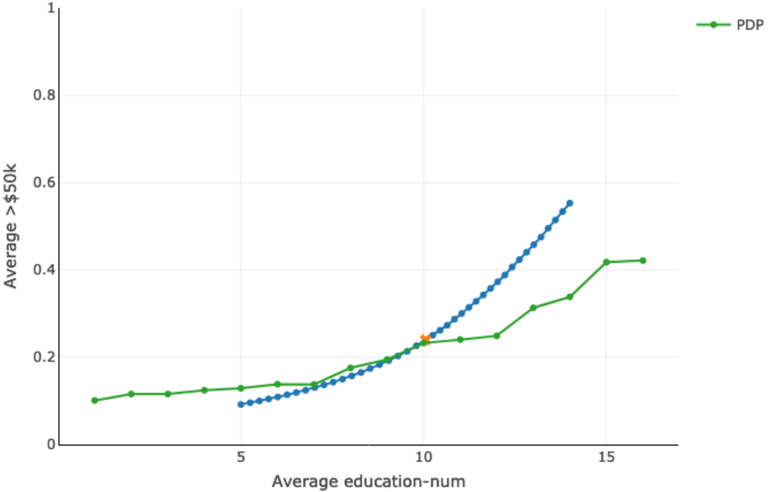
The usual suspects: MNIST digits

CNN on the MNIST dataset using a deep learning architecture (Keras). Training set 60,000 images whilst the predictions were made on another 10,000 images.

28×28 variables: pixels of image



Comparison with partial dependence plots



Computational burden

p	n	time (sec)
10	10000	0.46
100	10000	4.28
1000	10000	38.5
10	100000	1.93
10	1000000	12.3

Table: Computational times required on synthetic datasets, where 21 levels of stress (τ) were computed on each of the p variables.

Package usage

https://github.com/XAI-ANITI/ethik

```
import ethik
import lightgbm as lgb
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y)

model = lgb.LGBMClassifier().fit(X_train, y_train)
y_pred = model.predict(X_test)

explainer = ethik.ClassificationExplainer()

explainer.explain_influence(X_test['age'], y_pred)
explainer.plot_influence(X_test['age'], y_pred)
```

Future work

- Variational method to understand the influence of a variable (or a stress on variables) in the outcome of a machine learning algorithm
- Quick and easy but powerful tool: package submitted in (Bachoc, Gamboa, Halford, Loubes, Risser 2020) with CNRS research grant (winner of CNRS applied innovation grant 2018-2019)
- Only studied for testing sample $\mapsto$ Extension to understand the effect in the training sample (joint work with F. Bachoc, E. Pauwels, P. Zamolodtchikov)