Data Science for Actuaries (ACT6100)

Arthur Charpentier

Supervisé # 4 (Interprétation)

automne 2020

https://github.com/freakonometrics/ACT6100/

Interpretation

Consider a regression model

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i$$

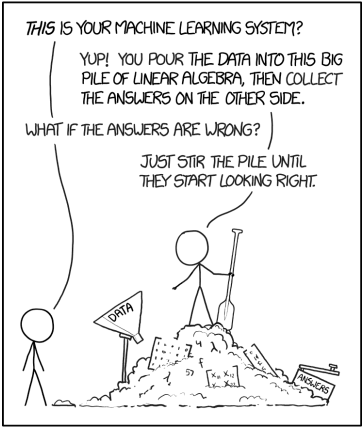
with standard econometric notations, holding x_2 fixed (*Ceteris Paribus* interpretation),

$$\beta_1 = \frac{\partial y}{\partial x_1}$$

assuming $\mathbb{E}[\varepsilon|x_1, x_2] = 0$, or with notions used so far

$$\beta_1 = \frac{\partial m(\boldsymbol{x})}{\partial x_1} \; (= \text{constant})$$

(source Randall Munroe (xkcd, 2016))



Interpretation

Ceteris paribus can be translated into “all other things being equal” or “holding other factors constant.”
In a linear regression, $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i$

$$\beta_1 = \frac{\partial y}{\partial x_1}$$

In a logistic linear regression, $\text{logit}(p_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i}$

$$\exp[\beta_1] = \frac{\partial y}{y \partial x_1}$$

Mutatis mutandis approximately translates as “allowing other things to change accordingly” or “the necessary changes having been made.”

Interpretation

What do we mean by interpreting a machine learning model, and why do we need it? Is it to trust the model? Or try to find causal relationships in the analyzed phenomenon? Or to visualize it?

- Lipton (2017, The Mythos of Model Interpretability),
- Lakkaraju *et al.* (2019, Faithful and Customizable Explanations of Black Box Models),
- Molnar (2019. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable),
- Guidotti *et al.* (2018, A Survey of Methods for Explaining Black Box Models)
- Gilpin *et al.* (2019, Explaining Explanations: An Overview of Interpretability of Machine Learning)
- Lundberg & Lee (2017, A Unified Approach to Interpreting Model Predictions)

Interpretation

For works that describe machine learning models as black boxes, transparency and interpretability are closely related, if not the same concept.

We can open the black box either

- by explaining the model,
- by explaining the outcome
- by inspecting the black box internally
- by providing a transparent solution.

"*Neural nets and random forests are considered as black boxes*", Ribeiro *et al.* (2016, "Why Should I Trust You?": Explaining the Predictions of Any Classifier).

Following Guidotti *et al.* (2018, A Survey of Methods for Explaining Black Box Models) In this general setting, a (black box) model is $m : \mathcal{X} \mapsto \mathcal{Y}$ (neural nets, SVM, etc), explanators ϵ will be described after (Features Importance, Sensitivity Analysis, Partial Dependence Plot, etc).

Interpretation

- Black-box model explanation

Given a black box predictor m and a dataset $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i)\} \in (\mathcal{X} \times \mathcal{Y})^n$, the black box model explanation problem consists in finding a function $f : (\mathcal{X} \mapsto \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y})^n \to (\mathcal{X} \mapsto \mathcal{Y})$ that returns a comprehensible global predictor c_g, i.e., $f(m, \mathcal{D}_n) = c_g$, such that c_g is able to mimic the behavior of m, and exists a global explanator function $\epsilon_g : (\mathcal{X} \mapsto \mathcal{Y}) \to \mathcal{E}$ that can derive from c_g a set of explanations $E \in \mathcal{E}$ modeling in a human understandable way the logic behind c_g, i.e., $\epsilon(c_g) = E$.

Interpretation

- Black-box outcome explanation

Given a black box predictor b and a dataset $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i)\}$, the black box outcome explanation problem consists in finding a function $f : (\mathcal{X} \mapsto \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y})^n \to (X \mapsto Y)$ that returns a comprehensible local predictor c_ℓ, i.e., $f(m, \mathcal{D}_n) = c_\ell$, such that c_ℓ is able to mimic the behavior of m, and exists a local explanator function $\epsilon_\ell : (\mathcal{X} \mapsto \mathcal{Y}) \times (\mathcal{X} \mapsto \mathcal{Y}) \times \mathcal{X} \to \mathcal{E}$ that can derive from the black box model m, the comprehensible local predictor c_ℓ, and a data record $\boldsymbol{x}$, a human understandable explanation $e \in E$ for the data record $\boldsymbol{x}$, i.e., $\epsilon_\ell(m, c_\ell, \boldsymbol{x}) = e$

Interpretation

- Black box inspection explanation

Given a black box predictor b and a dataset $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i)\}$, the black box inspection problem consists in finding a function $f : (\mathcal{X} \mapsto \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y})^n \to V$ that returns a visual representation of the behavior of the black box, $f(m, \mathcal{D}_n) = v$ with V being the set of all possible representations.
The transparent box design problem consists in providing a model which is locally or globally interpretable on its own,

Interpretation

- Transparent box design problem
Given a dataset $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i)\}$, the transparent box design problem consists in finding a learning function $L : (\mathcal{X} \times \mathcal{Y})^n \to (\mathcal{X} \mapsto \mathcal{Y})$ that returns a (locally or globally) comprehensible predictor c, i.e., $L(\mathcal{D}_n) = c$. This implies that there exists an explanator function, local ϵ_ℓ or global ϵ_g, that takes as input the comprehensible predictor c and returns a human understandable explanation $e \in E$, or a set of explanations E.

Interpretation

- Partial Dependence Plot

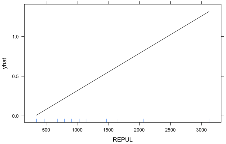
Introduced in Friedman (2001, Greedy function approximation: A gradient boosting machine). Let $\boldsymbol{x}$ be splited in two parts : $\boldsymbol{x}_s$ (variable(s) of interest) and $\boldsymbol{x}_c$ the complementary, $\boldsymbol{x} = (\boldsymbol{x}_s, \boldsymbol{x}_c)$. Partial dependence of $\boldsymbol{x}_s$ is

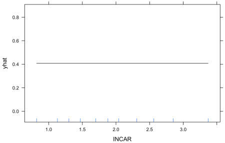
$$p(\boldsymbol{x}_s) = \mathbb{E}\big[m(\boldsymbol{x}_s, \boldsymbol{S}_c)\big] \text{ and } \widehat{p}(\boldsymbol{x}_s) = \frac{1}{n}\sum_{i=1}^{n}\widehat{m}(\boldsymbol{x}_s, \boldsymbol{x}_{i,c})$$

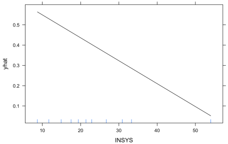
See pdp package

```
pdp::partial(model, pred.var = "REPUL", plot = TRUE)}
```

Consider a (standard) linear model on the myocarde dataset

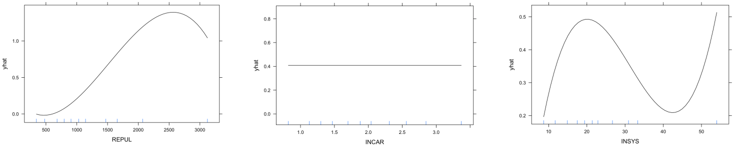




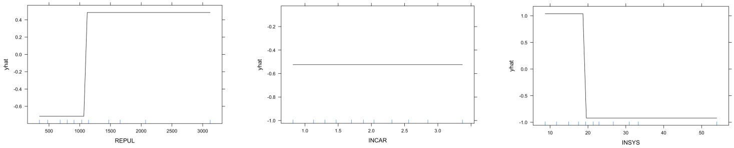


Interpretation

Consider an additive model (GAM) on the myocarde dataset



or a classification tree

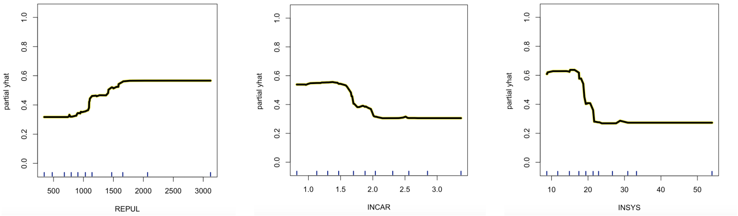


Interpretation

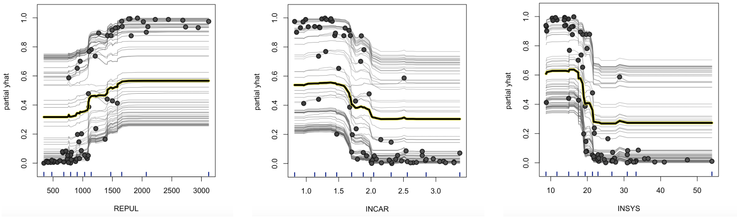
- Individual Conditional Expectation

Extention of Partial Dependence Plots, introduced in Goldstein *et al.* (2013, Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation), "*Visually, ICE plots disaggregate the output of classical PDPs. Rather than plot the target covariates'average partial effect on the predicted response, we instead plot the n estimated conditional expectation curves*"

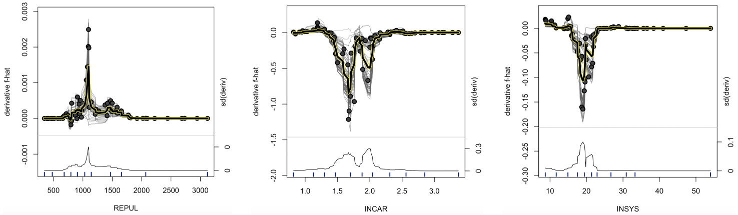
See ICEbox package. Here m is a random forest,



Interpretation



One can also plot the derivative of ICE functions



Interpretation

- Accumulated Local Effects

Introduced in Apley (2016, Visualizing the effects of predictor variables in black box supervised learning models). Partial dependence of $\boldsymbol{x}_s$ is

$$p(\boldsymbol{x}_s) = \mathbb{E}\big[m(\boldsymbol{x}_s, \boldsymbol{S}_c)\big] \text{ and } \widehat{p}(\boldsymbol{x}_s) = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}(\boldsymbol{x}_s, \boldsymbol{x}_{i,c})$$

Here, we focus on

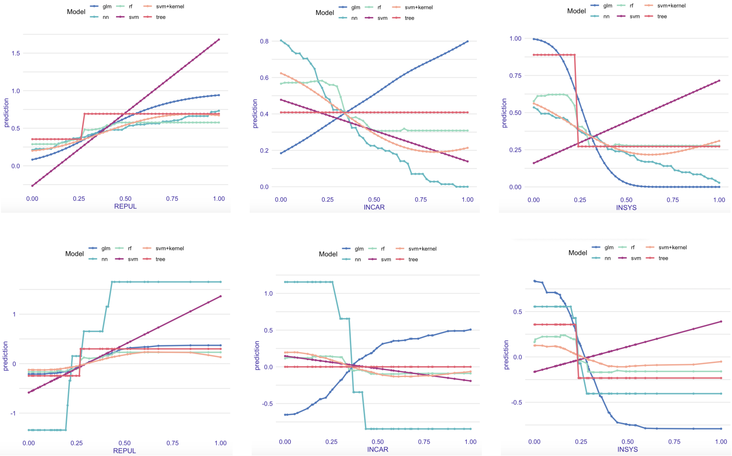
$$a(\boldsymbol{x}_s) = \int_{-\infty}^{\boldsymbol{x}_s} \mathbb{E}\left[\frac{\partial m(\boldsymbol{z}_s, \boldsymbol{S}_c)}{\partial \boldsymbol{x}_s}\right] d\boldsymbol{z}_s$$

See DALEX and

```
DALEX::single_variable(explain(m), variable =  x, type
    = "ale")
DALEX::single_variable(explain(m), variable =  x, type
    = "pdp")
```

Interpretation

Partial Dependence Plot (on top) and Accumulated Local Effects (below)



Interpretation

- Feature Interaction

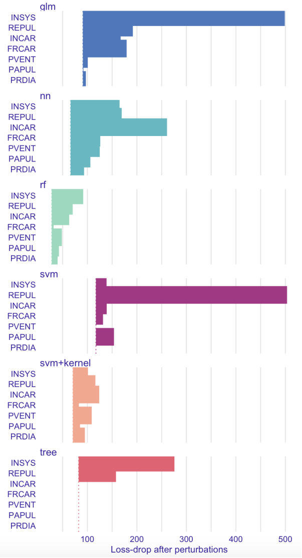
Friedman & Popescu (2008, (Predictive learning via rule ensembles), see Greenwell *et al.* (2018, A simple and effective model-based variable importance measure)

```
1 > iml::Interaction
```

- Feature (Variable) Importance

Breiman (2001, Random Forests)

```
1 > iml::FeatureImp
```



Interpretation

- Local Surrogate (LIME) - Local Interpretable Model-Agnostic Explanations
Alvarez-Melis & Jaakkola (2018, On the robustness of interpretability methods)
see the lime or ceterisParibus packages (and the *what-if* plot)