DOMINO

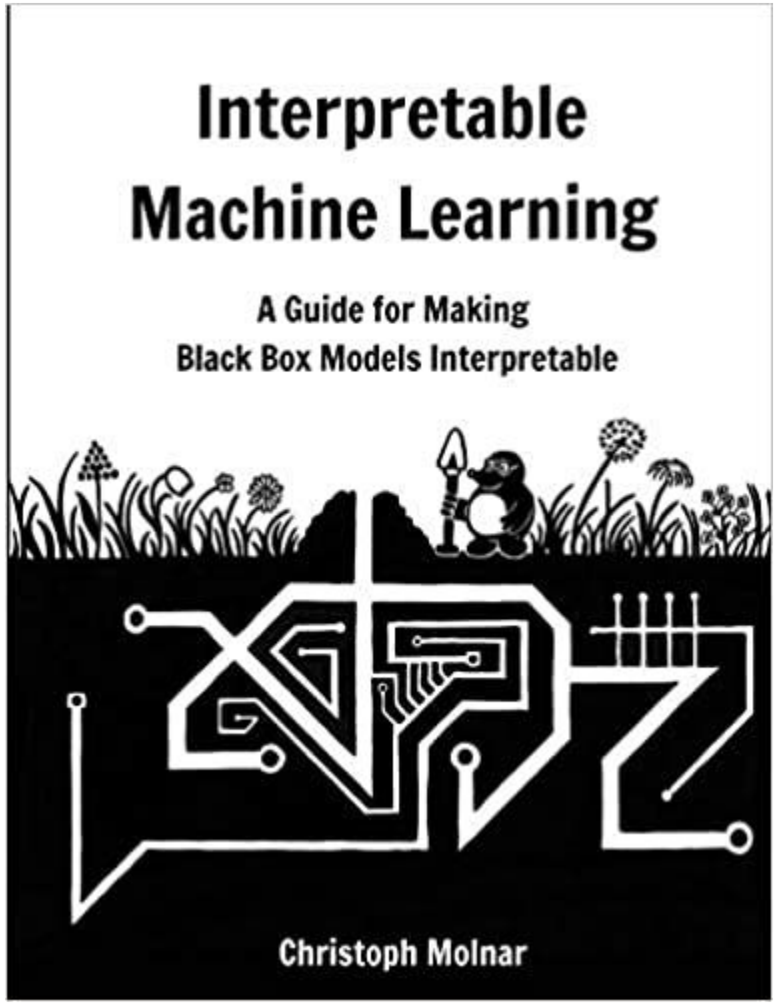
Nikolay Manchev
Principal Data Scientist for EMEA
Domino Data Lab
@nikolaymanchev

Interpreting Machine Learning Models

16 JULY 2020

Housekeeping

- Moving to virtual events until the Covid-19 situation is resolved
- Next event - 27th August
- I need your help
  - Speakers
  - Topics
- The slide deck will be available on github https://github.com/nmanchev/LondonDSML
- To get in touch with me use **@nikolaymanchev**
- Prize



DOMINO

Domino is “The Center of the ML Ecosystem”

The **gateway** to data science infrastructure and the **system of record** for work

Your favorite tools and IDEs

Use data from anywhere

Integrated with your workflow

DOMINO

- Infrastructure Orchestration
- Reproducibility & collaboration
- Model Ops (host, publish, monitor)
- Governance & management

Use any programming language

Any distributed compute framework

With any algorithm package

Running on any infrastructure

DOMINO

DOMINO

Nikolay Manchev
Principal Data Scientist for EMEA
Domino Data Lab
@nikolaymanchev

Interpreting Machine Learning Models

16 JULY 2020

Why Interpretable Models

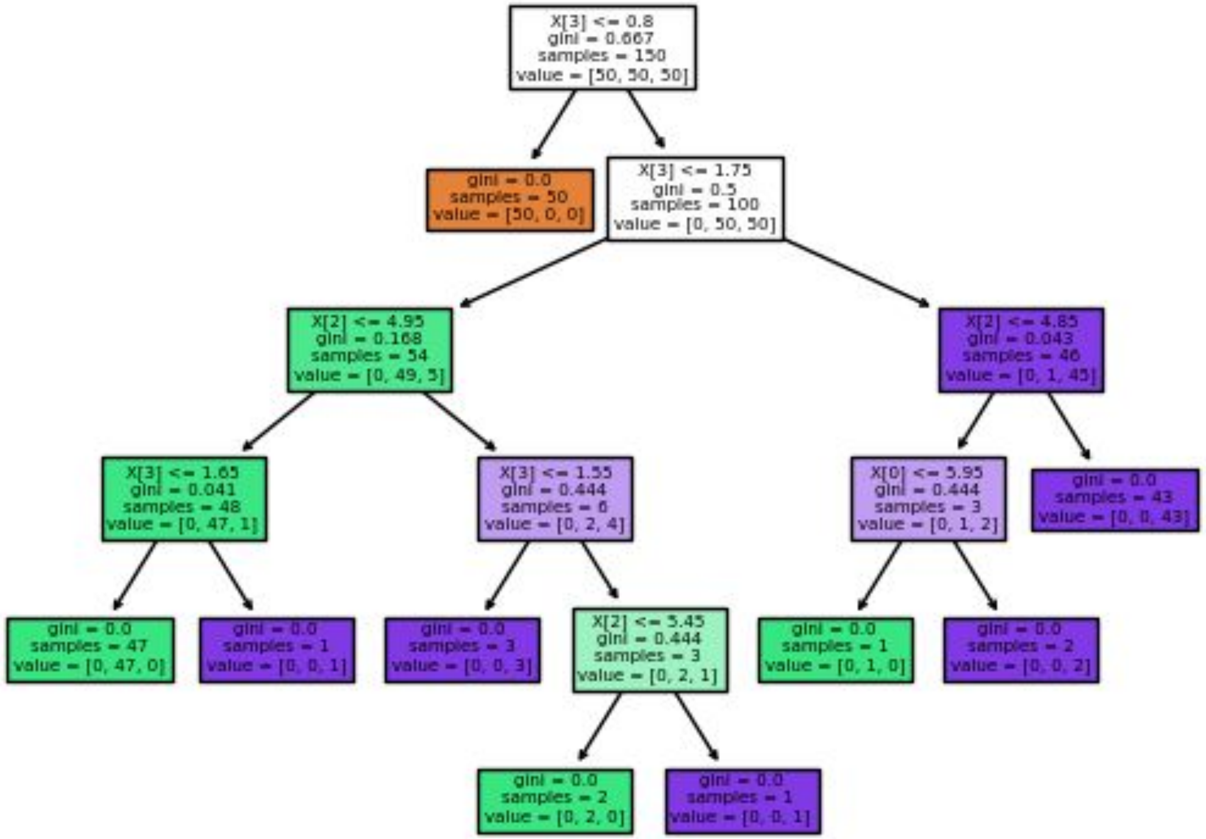
- No hard-coded rules
  - Difficult to explain to the business
  - Justifying decisions
- Biased models can have devastating effects
  - Predicting potential criminals

https://weaponsofmathdestructionbook.com/

https://xkcd.com/1838

Interpretable Models

- Linear Regression
- Logistic Regression
- GLM
- Decision Trees
- Naive Bayes



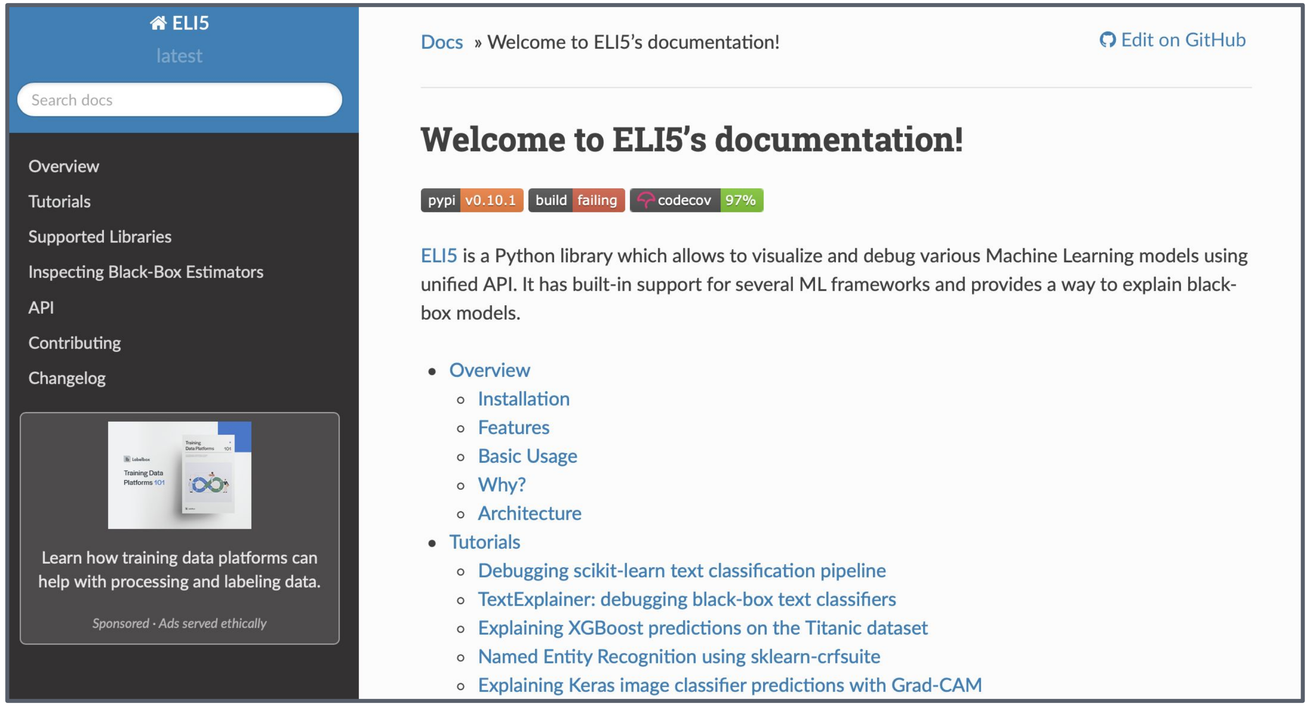
[Scikit-learn official documentation](), Decision Trees, Section 1.10

Ensemble Models (e.g. XGBoost)

- The **what vs. why** tradeoff:
  - The sole purpose of using an ensemble is to increase predictive performance
  - Will interpreting an average of models going to answer anything?
    - n_estimators=100 by default in XGBoost
- There are still things we can do
  - How important are the individual features?
    - Gain - average gain across all splits the feature is used in.
    - Cover - average coverage across all splits the feature is used in
    - Weight - number of times a feature is used to split the data across all trees
  - Examining individual predictions

ELI5

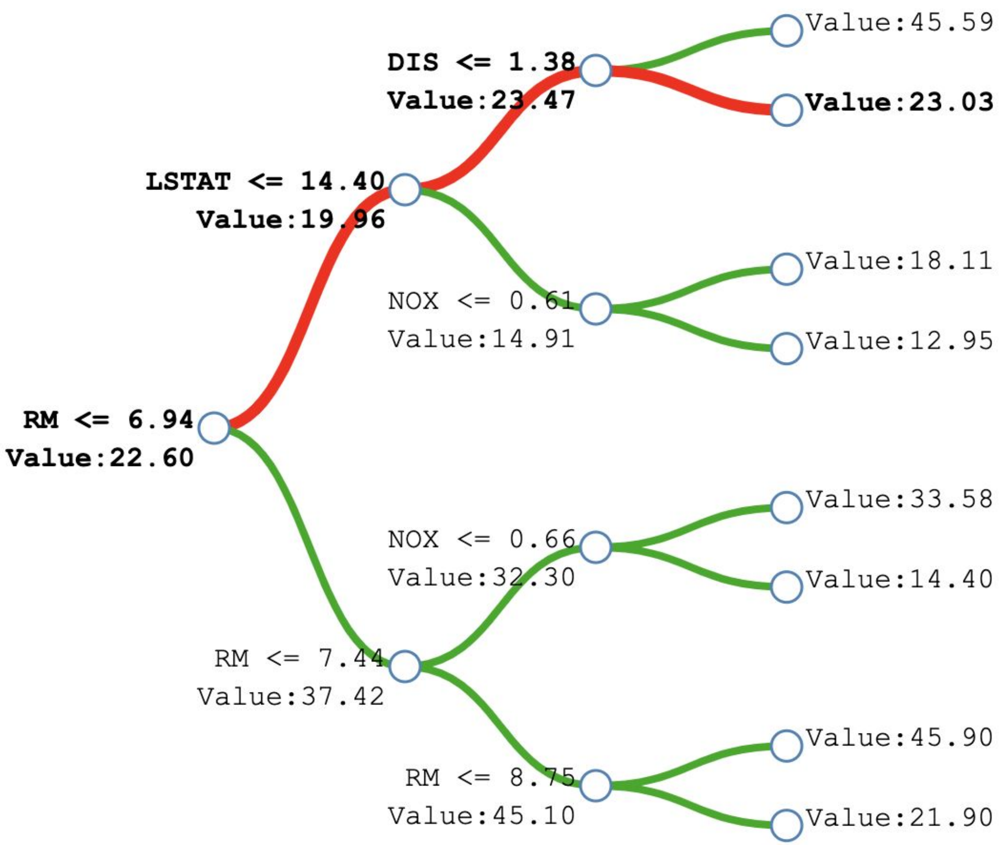
- ELI5 is a Python library which allows to visualize and debug various Machine Learning models
- Built-in support for several ML frameworks and provides a way to explain black-box models.
  - Scikit-learn
  - XGBoost
  - lightning
  - Keras
  - ...



DOMINO

[Interpreting random forests](), Diving into data blog, Oct 2014, visited 05/2020

ELI5's interpretation algorithm



Prediction: 23.03 ≈ 22.60 (trainset mean) - 2.64(loss from RM) + 3.52(gain from LSTAT) - 0.44(loss from DIS)

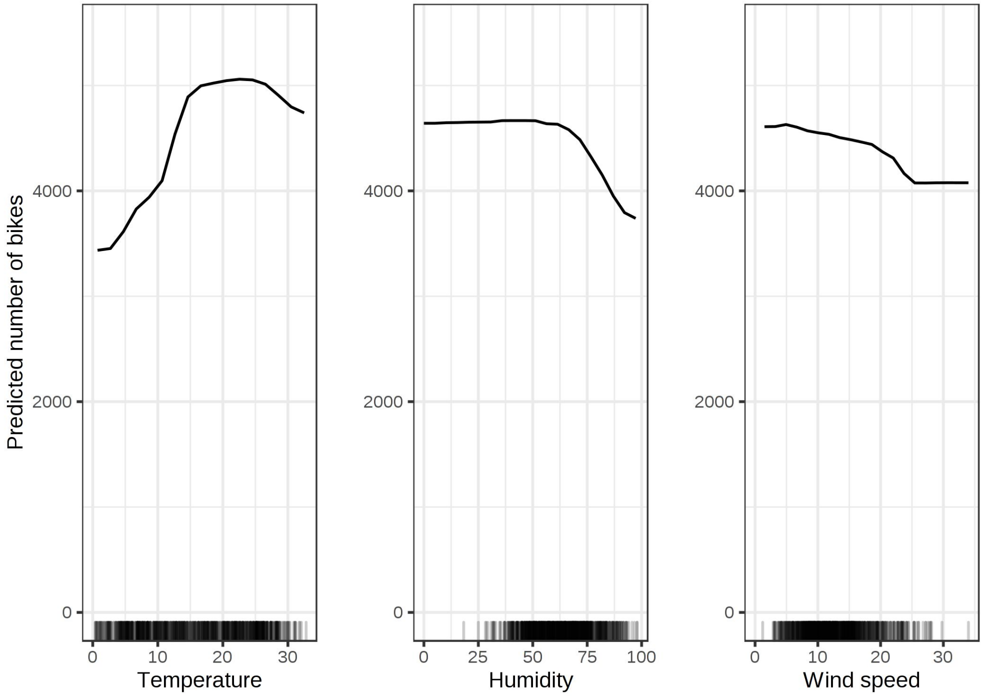
Partial Dependence Plot

- Model agnostic method
- Shows the marginal effect one or two features have on the predicted outcome
- Shows the dependence between the target and a set of "target" features, marginalizing over the values of all other features
- The target features set (S) is usually limited to 2

$$X = [x_s, x_c] \in \mathbb{R}^{n \times p}, y \in \mathbb{R}$$

$$\hat{y} = f(x) + \epsilon$$

$$\hat{f}_s(x) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}(x, x_c), x \in x_s$$



PDPs for the bicycle count prediction model, Interpretable Machine Learning, Christoph Molnar, Section 5.1

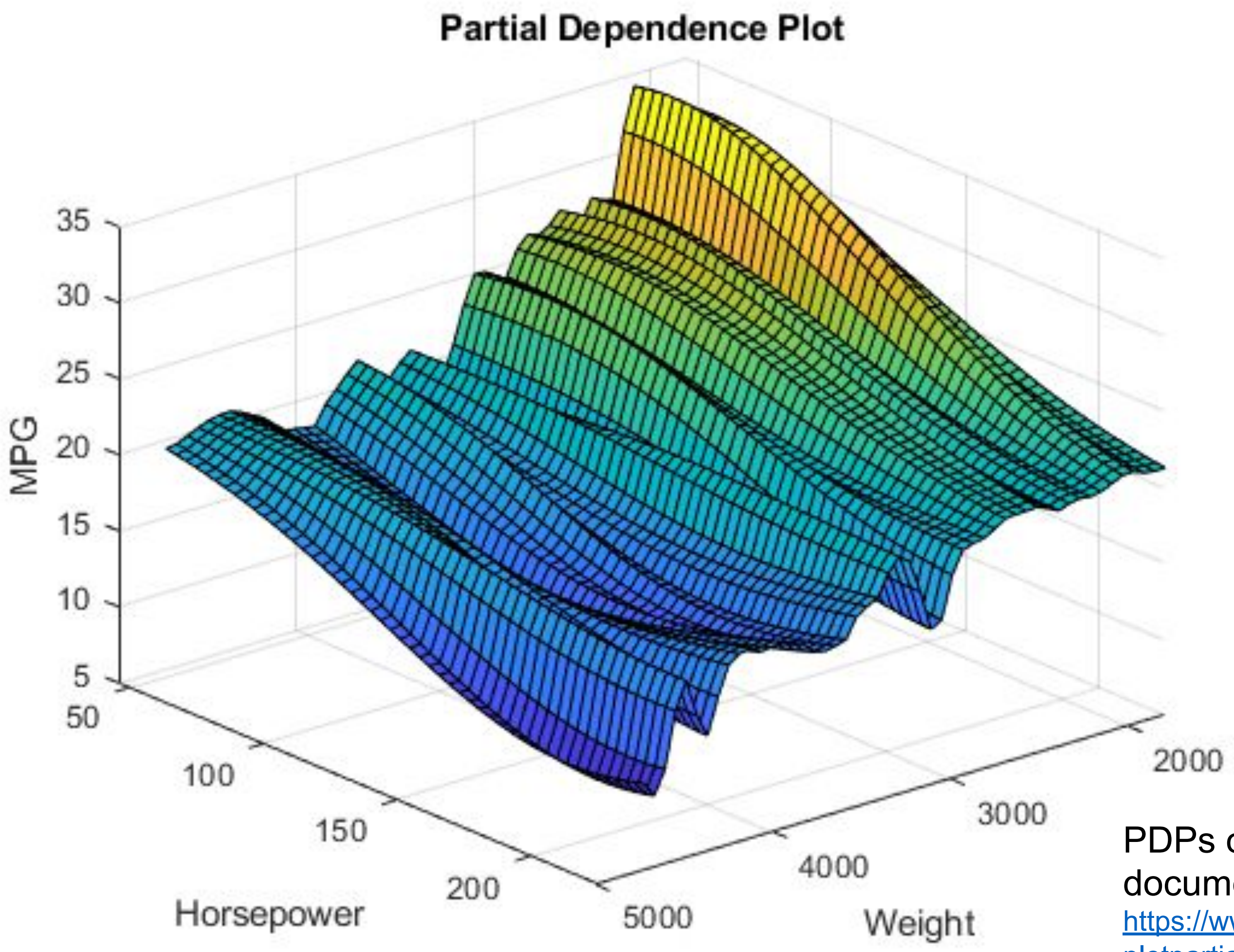
Skater

- Unified framework to enable Model Interpretation for all forms of model
- Supports local and operationalised models

<table>
<tr><th>Scope of Interpretation</th><th colspan="2">Algorithms</th></tr>
<tr><td>Global Interpretation</td><td colspan="2">Model agnostic Feature Importance</td></tr>
<tr><td>Global Interpretation</td><td colspan="2">Model agnostic Partial Dependence Plots</td></tr>
<tr><td>Local Interpretation</td><td colspan="2">Local Interpretable Model Explanation(LIME)</td></tr>
<tr><td>Local Interpretation</td><td>DNNs</td><td>• Layer-wise Relevance Propagation (e-LRP): image
• Integrated Gradient: image and text</td></tr>
<tr><td>Global and Local Interpretation</td><td colspan="2">• Scalable Bayesian Rule Lists
• Tree Surrogates</td></tr>
</table>

Skater documentation, Overview, Sep 2018, visited 5/2020

PDP on AutoMPG

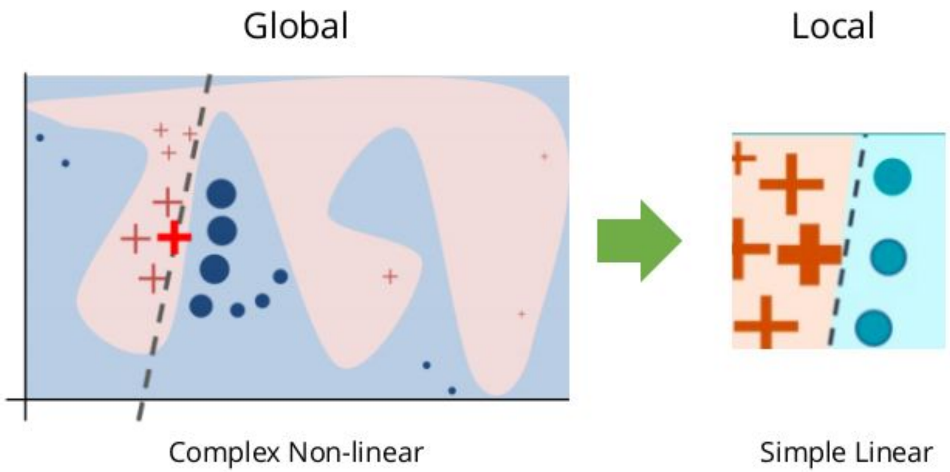


PDPs on the carsmall data set, MATLAB documentation, https://www.mathworks.com/help/stats/regressiontree.plotpartialdependence.html

DOMINO

Local Surrogate (LIME)

- Used to explain individual predictions of black box models*
- Highlights
  - Generate a new dataset consisting of permuted samples and their predictions
  - Train an interpretable model, weighted by the proximity of the sampled instances to the instance of interest



$$\text{explanation}(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

[Interpretability part 3: opening the black box with LIME and SHAP](), Manu Joseph, KDnuggets

* Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. “Why should I trust you?: Explaining the predictions of any classifier.” Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM (2016).

https://www.meetup.com/London-Data-Science-and-Machine-Learning

London Data Science
& Machine Learning