DSPT
DATA SCIENCE PORTUGAL

Avoiding the Machine Learning blackbox

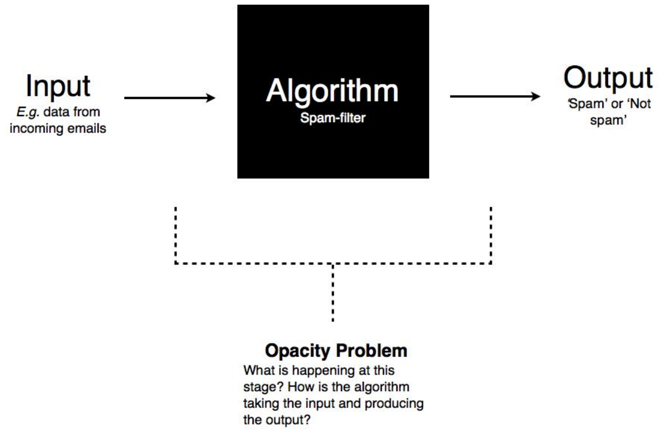
Miguel José Monteiro

Outline

- Why explain our models
- How we can do it
- Limitations of “feature importance”
- SHAP (with code)
- Plots and explanations
- Next steps

Blackbox algorithms

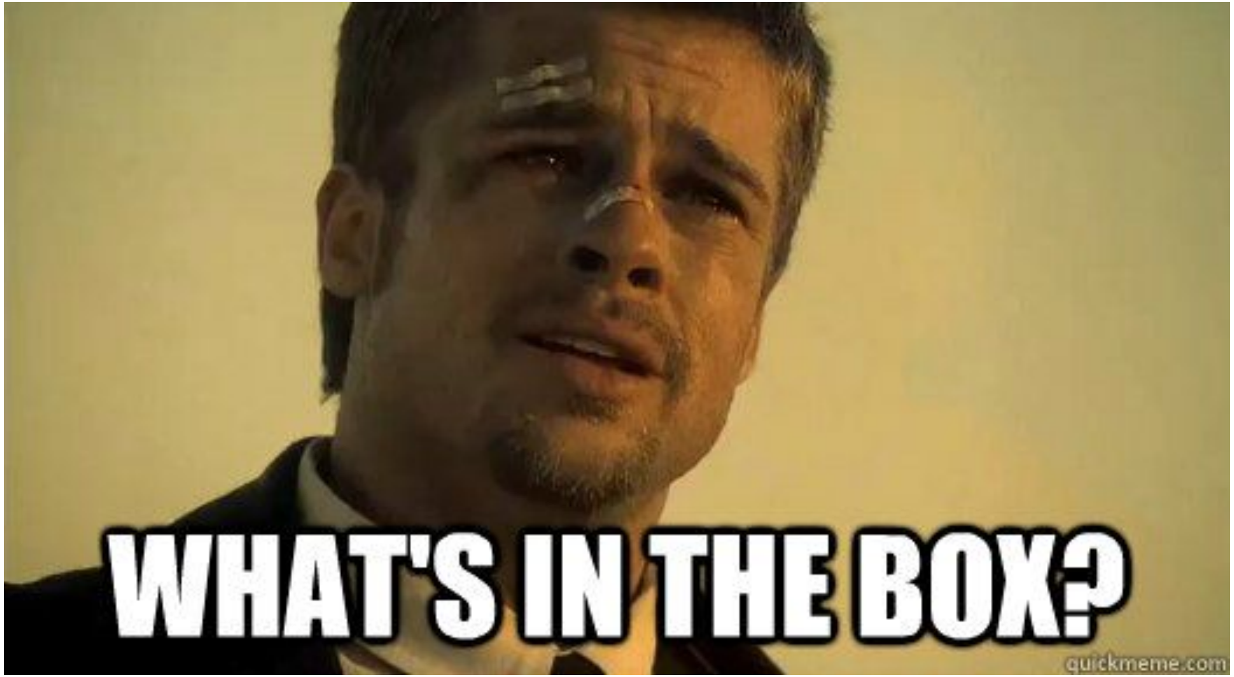
Device or system of which we only know its inputs and outputs but have no knowledge of its internal workings



Why explain our models

DSPT
DATA SCIENCE PORTUGAL

So the question is...



DSPT
DATA SCIENCE PORTUGAL

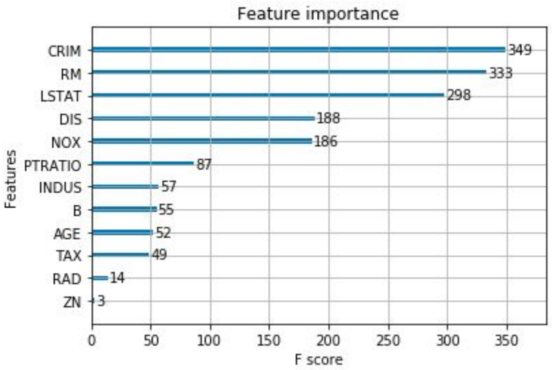
How we can do it

01	**LIME** *March 2016*	• https://github.com/marcotcr/lime
02	**SHAP** *November 2016*	• https://github.com/slundberg/shap
03	**eli5** *September 2016*	• https://github.com/TeamHG-Memex/eli5
04	**Interpret** *May 2019*	• https://github.com/Microsoft/interpret

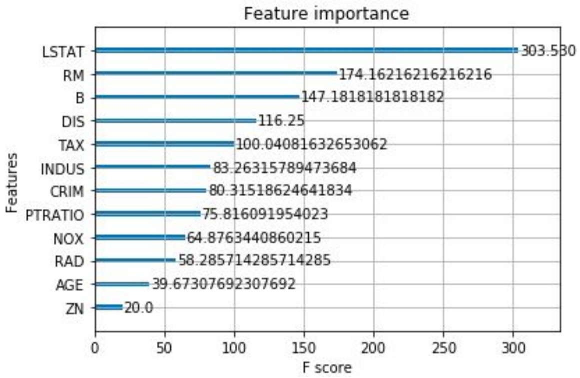
DSPT
DATA SCIENCE PORTUGAL

Wait...what about good old "feature importance?"

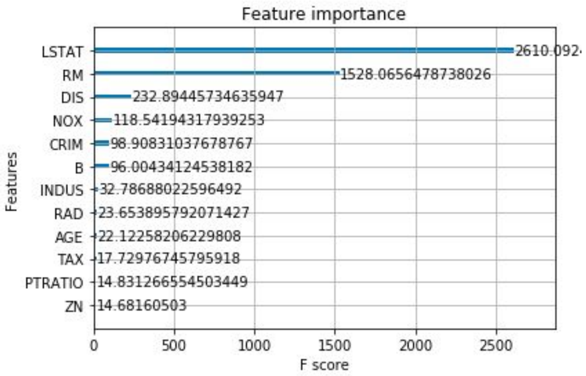
Limitations of “feature importance”



weight

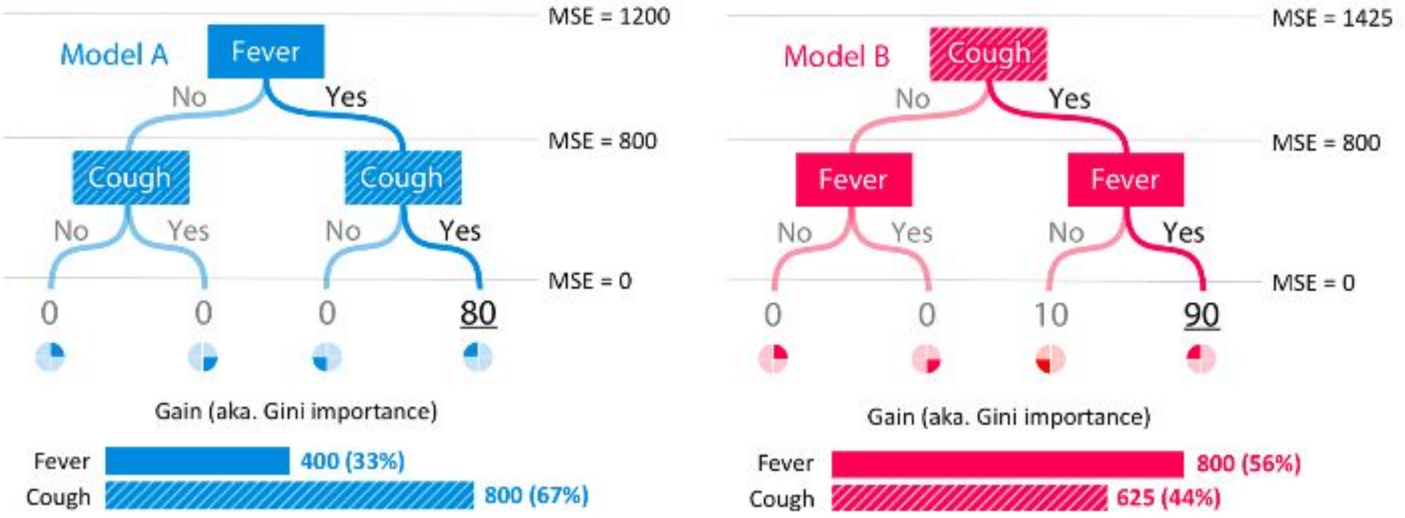


cover



gain

Limitations of “feature importance”



Computation of the gain (aka. Gini importance) scores for model A and model B.

What do we want in an explainer

- Interpretability
- Local fidelity
- Model-agnosticism
- Global perspective

DSPT
DATA SCIENCE PORTUGAL

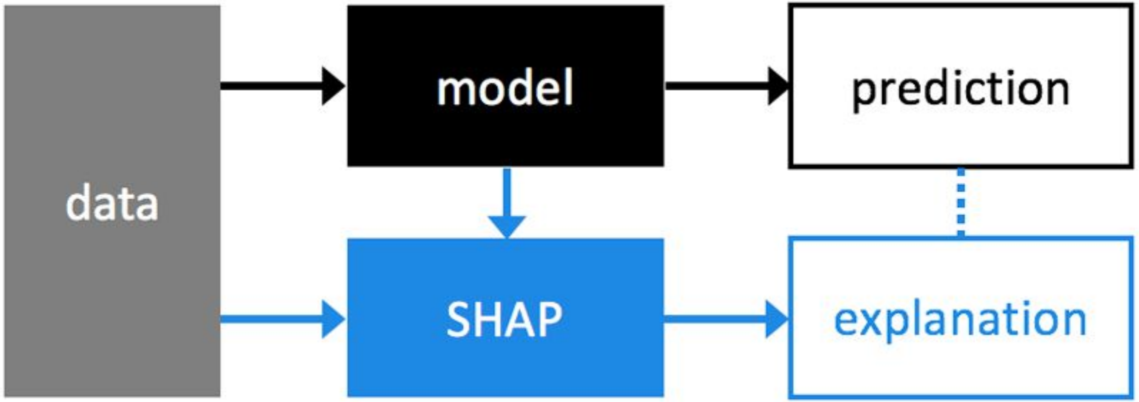
Let's talk about SHAP

About SHAP

- Full name is **SHapley Additive exPlanations**
- Created by Scott Lundeberg (PhD in Univ Washington)
- Paper published at NIPS 2017

How it works (high-level)

- **Method**: Using a less complex model as approximation



http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

DSPT
DATA SCIENCE PORTUGAL

Now let's play a bit

Boston house prices dataset

Information collected by the U.S Census Service concerning houses in the area of Boston

- 13 features
- 1 numeric target
    - Regression task
    - Predict house price

1. CRIM - per capita crime rate by town
2. ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS - proportion of non-retail business acres per town.
4. CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
5. NOX - nitric oxides concentration (parts per 10 million)
6. RM - average number of rooms per dwelling
7. AGE - proportion of owner-occupied units built prior to 1940
8. DIS - weighted distances to five Boston employment centres
9. RAD - index of accessibility to radial highways
10. TAX - full-value property-tax rate per $10,000
11. PTRATIO - pupil-teacher ratio by town
12. B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. LSTAT - % lower status of the population
14. MEDV - Median value of owner-occupied homes in $1000's

https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html

Training a model on Boston

Just your *usual suspects*:

- XGBoost
- GridSearch for hyperparameter optimization

Nothing special here!

Run SHAP on trained model

```
# explain the model's predictions using SHAP values
explainer = shap.TreeExplainer(model)
shap_values = explainer.shap_values(X)
```

class **`shap.TreeExplainer`**(*model, data=None, model_output='margin', feature_dependence='tree_path_dependent'*)

Uses Tree SHAP algorithms to explain the output of ensemble tree models.

`shap_values`(*X, y=None, tree_limit=None, approximate=False*)

Estimate the SHAP values for a set of samples.

DSPT
DATA SCIENCE PORTUGAL

SHAP's explanations

DSPT
DATA SCIENCE PORTUGAL

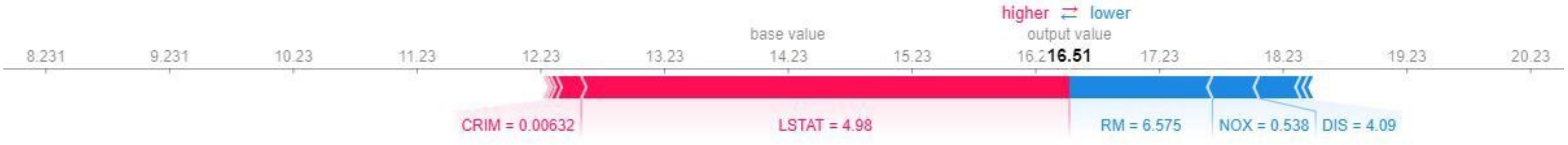
Prediction Explainer

```
# load JS visualization code to notebook
shap.initjs()

# visualize the first prediction's explanation
i = 0
shap.force_plot(explainer.expected_value, shap_values[i,:], X.iloc[i,:])
```

shap.force_plot(*base_value, shap_values, features=None, feature_names=None, out_names=None, link='identity', plot_cmap='RdBu', matplotlib=False, show=True, figsize=(20, 3), ordering_keys=None, ordering_keys_time_format=None, text_rotation=0*)

Visualize the given SHAP values with an additive force layout.

Prediction Explainer



DSPT
DATA SCIENCE PORTUGAL

Model Explainer

```
# load JS visualization code to notebook
shap.initjs()

# visualize the training set predictions
shap.force_plot(explainer.expected_value, shap_values, X)
```

shap.force_plot(*base_value, shap_values, features=None, feature_names=None, out_names=None, link='identity', plot_cmap='RdBu', matplotlib=False, show=True, figsize=(20, 3), ordering_keys=None, ordering_keys_time_format=None, text_rotation=0*)

Visualize the given SHAP values with an additive force layout.

Model Explainer

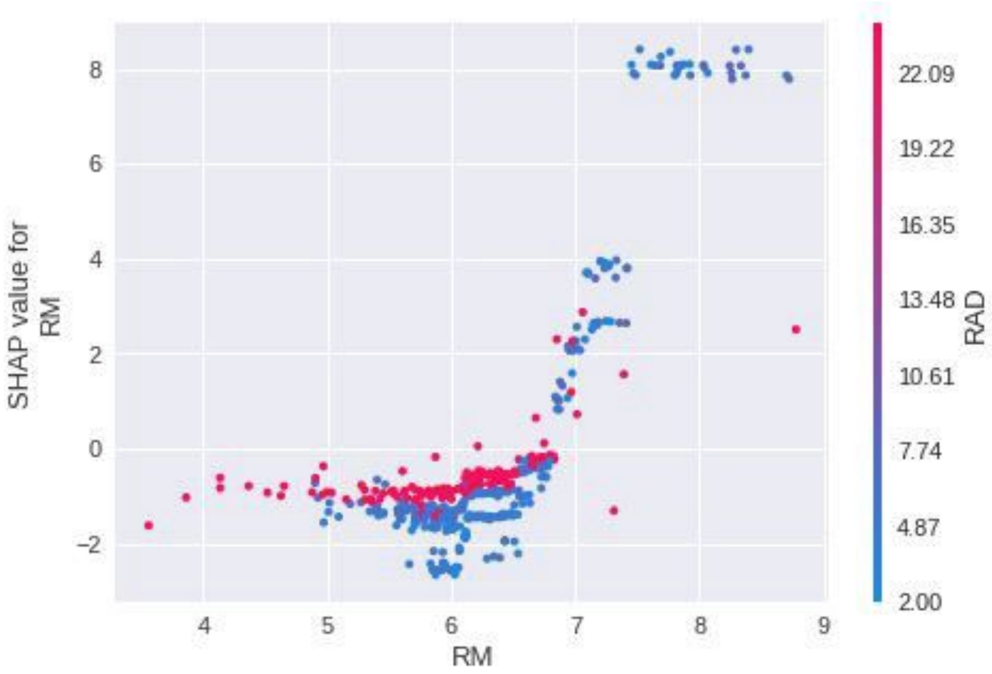
Dependence Plot

```
# load JS visualization code to notebook
shap.initjs()

# create a SHAP dependence plot to show the effect of a single feature across the whole dataset
shap.dependence_plot("RM", shap_values, X, interaction_index="LSTAT")
```

shap.dependence_plot(*ind, shap_values, features, feature_names=None, display_features=None, interaction_index='auto', color='#1E88E5', axis_color='#333333', cmap=<matplotlib.colors.LinearSegmentedColormap object>, dot_size=16, x_jitter=0, alpha=1, title=None, xmin=None, xmax=None, show=True*)

Create a SHAP dependence plot, colored by an interaction feature.

Dependence Plot



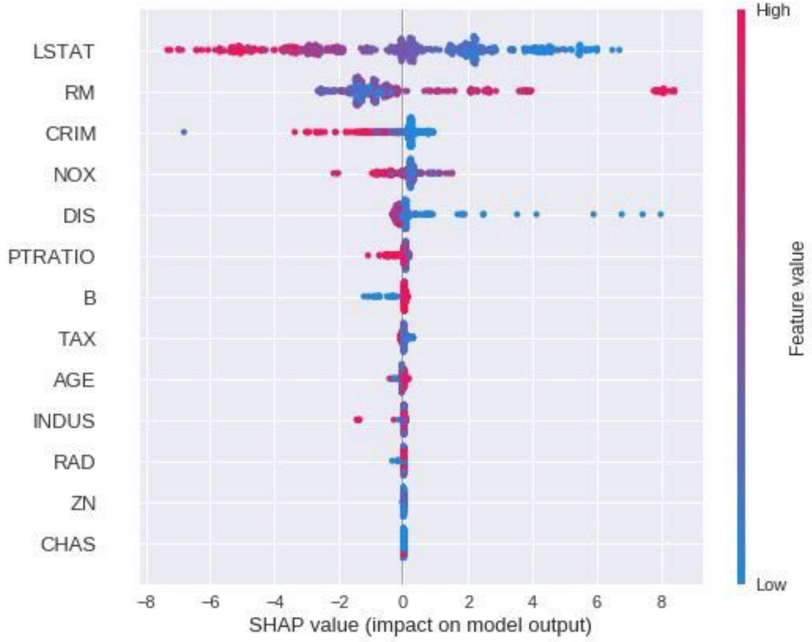
Summary Plot

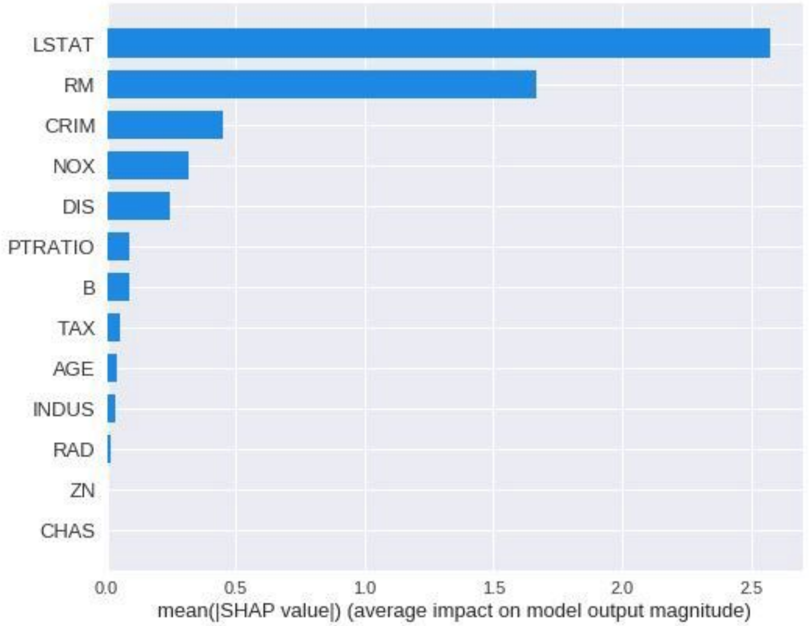
```
# load JS visualization code to notebook
shap.initjs()

# summarize the effects of all the features
shap.summary_plot(shap_values, X)
```

`shap.summary_plot`(*shap_values, features=None, feature_names=None, max_display=None, plot_type='dot', color=None, axis_color='#333333', title=None, alpha=1, show=True, sort=True, color_bar=True, auto_size_plot=True, layered_violin_max_num_bins=20, class_names=None*)

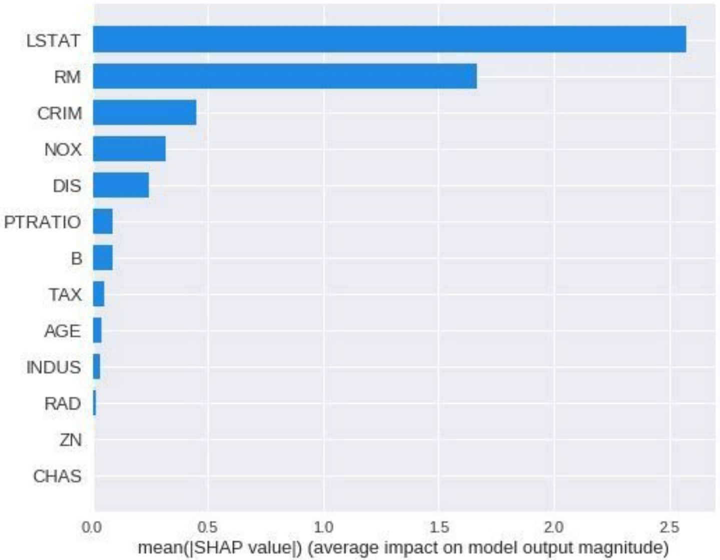
Create a SHAP summary plot, colored by feature values when they are provided.

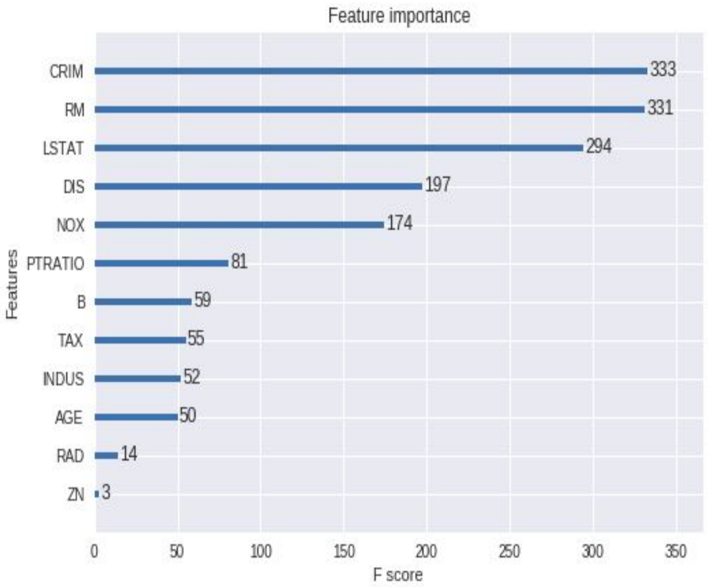
Summary Plot





Compare with "feature importance"





… also, check out "feature permutation"

DSPT
DATA SCIENCE PORTUGAL

Now for some sandbox

Next steps

- **SHAP for classification**: Iris dataset
  - https://archive.ics.uci.edu/ml/datasets/iris
- SHAP official documentation
  - https://shap.readthedocs.io/en/stable/#
- My article on TDS
  - http://bit.do/ml-blackbox-shap
- Scott Lundberg’s article on TDS
  - http://bit.do/lundberg-shap
- Feature importance article
  - https://explained.ai/rf-importance/

DSPT
DATA SCIENCE PORTUGAL

And that's that!

Miguel José Monteiro

miguelmarantes@gmail.com