auditor + DALEX

a powerful duet for validation and explanation of machine learning models

Alicja Gosiewska
Mi2DataLab
PhD Student at Warsaw University of Technology

Tomasz Mikołajczyk
Mi2DataLab, Ministry of Health

Instructions

1. Find the repository:

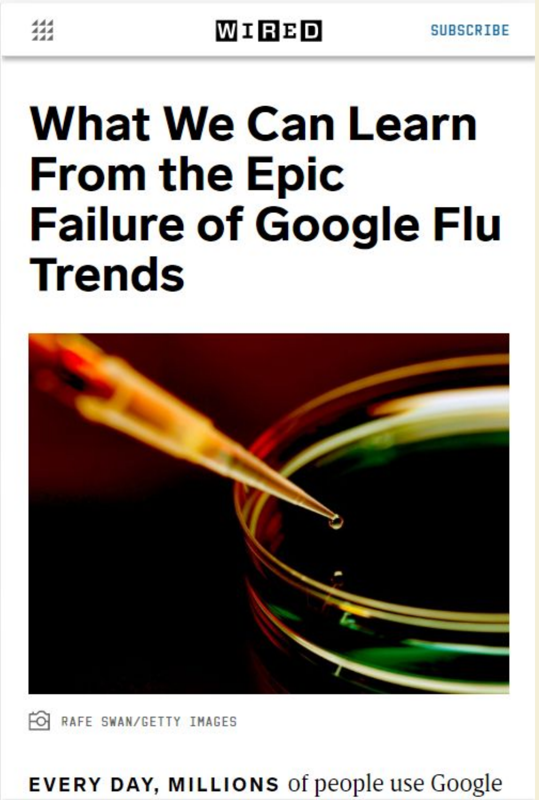
bit.ly/auditorWhyR

or

https://github.com/agosiewska/auditor-whyr2019

2. clone repository to the local folder or download zip file
3. open `part_1_explain.R` file

Do we need XAI? 1/3



What We Can Learn From the Epic Failure of Google Flu Trends

RAFE SWAN/GETTY IMAGES

EVERY DAY, MILLIONS of people use Google

Google Flu Trends

Researchers from Google claimed that they could predict outbreaks of flu based on people’s searches (paper stated that the Google Flu Trends obtains 97% accuracy).

BUT

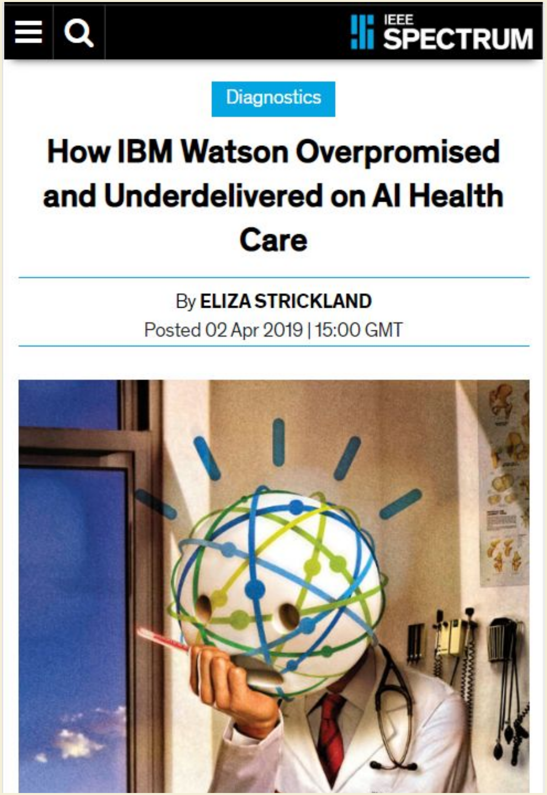
Subsequent reports asserted that Google Flu Trends' predictions have sometimes been very inaccurate

- in the 2012-2013 flu season predicted twice as many doctors' visits as it was recorded

By re-assessing the original model, it was uncovered that the model was aggregating queries about different health conditions.

https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

Do we need XAI? 2/3



IEEE SPECTRUM

Diagnostics

How IBM Watson Overpromised and Underdelivered on AI Health Care

By ELIZA STRICKLAND
Posted 02 Apr 2019 | 15:00 GMT

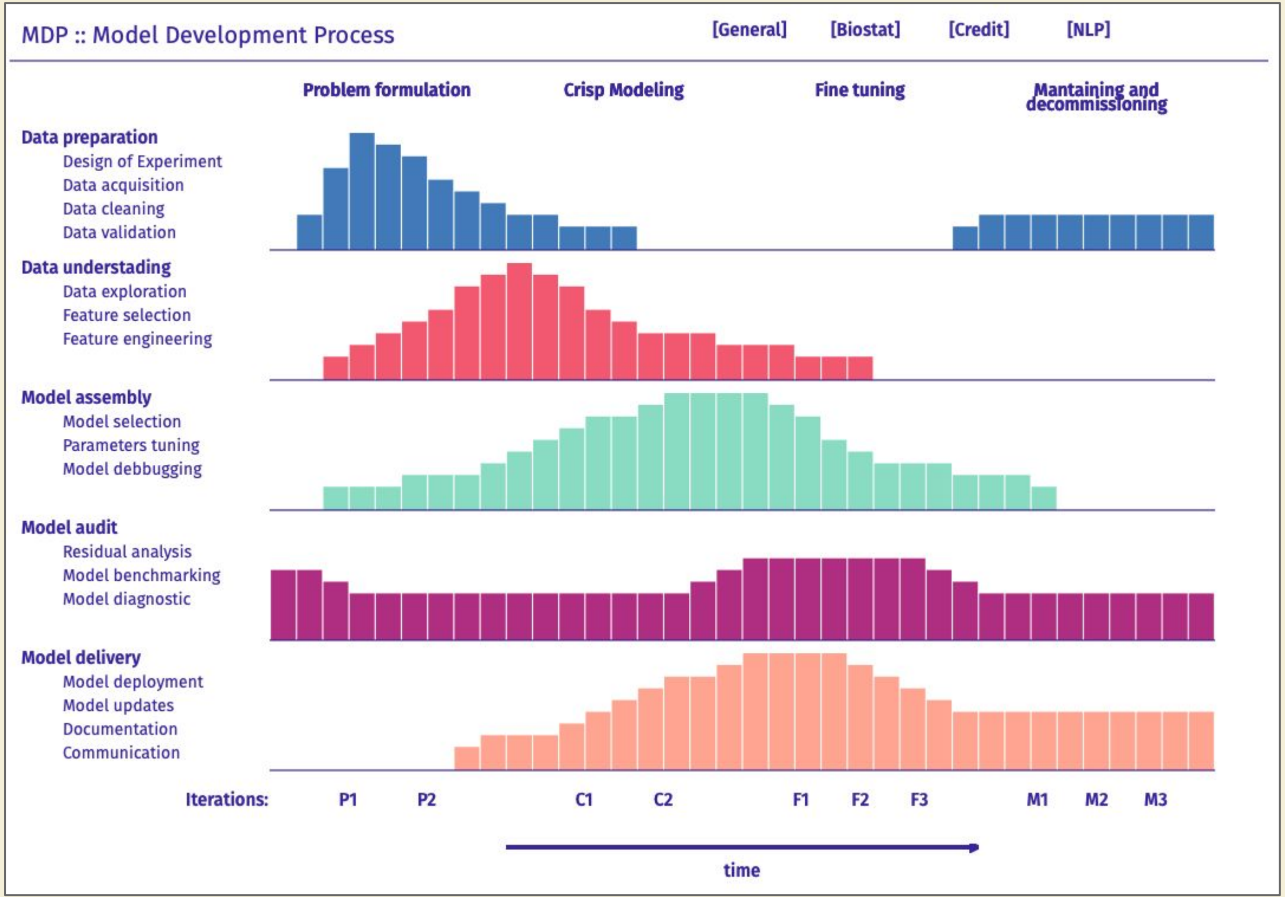
IBM Watson

Researchers from IBM took a collection of patient symptoms and came up with a list of possible diagnoses, each annotated with confidence level and links to supporting medical literature.

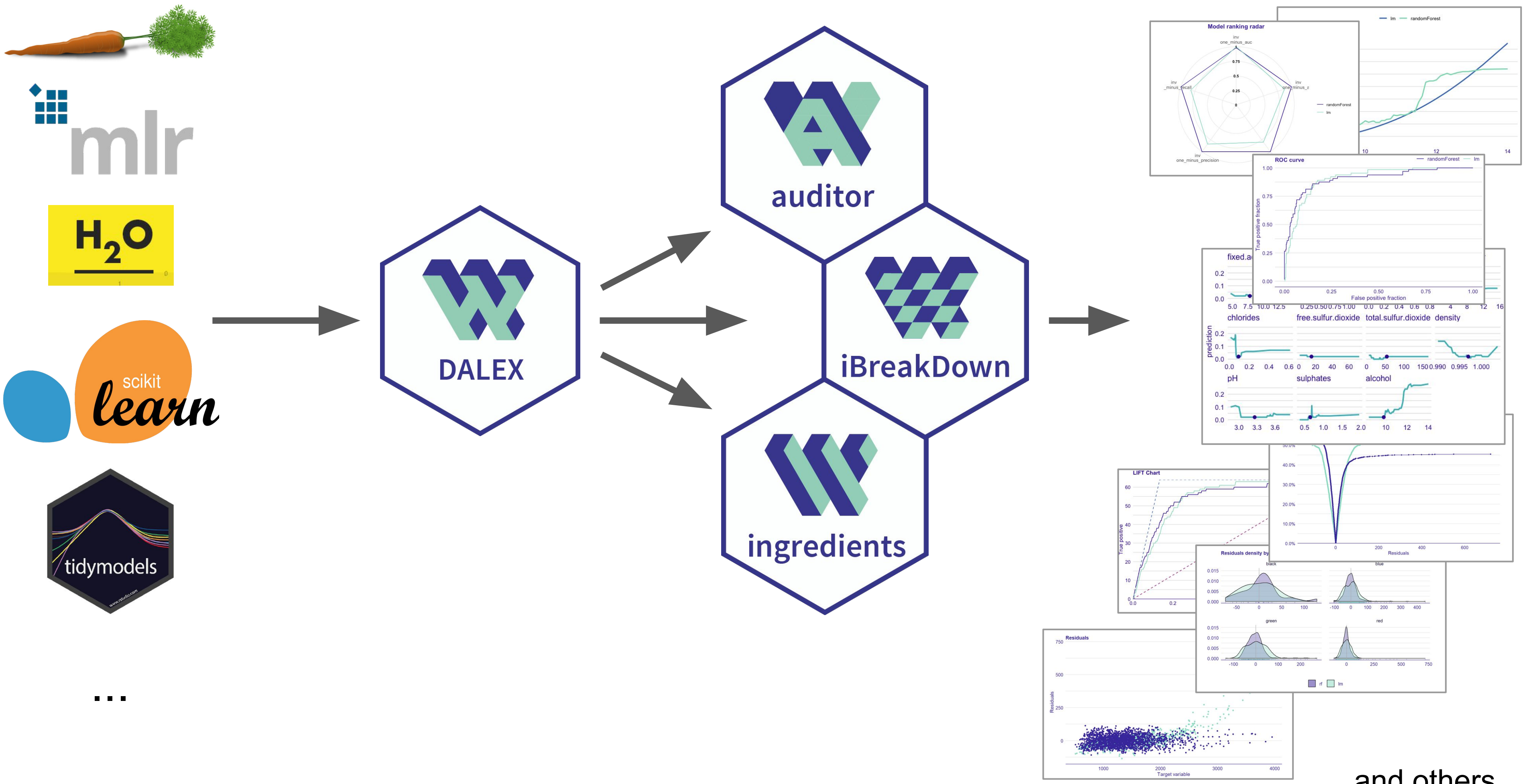
BUT
Watson's predictions did not earn the trust of physicians.
Watson did not achieve as good results as expected.

https://spectrum.ieee.org/biomedical/diagnostics/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care

Do we need XAI? 3/3



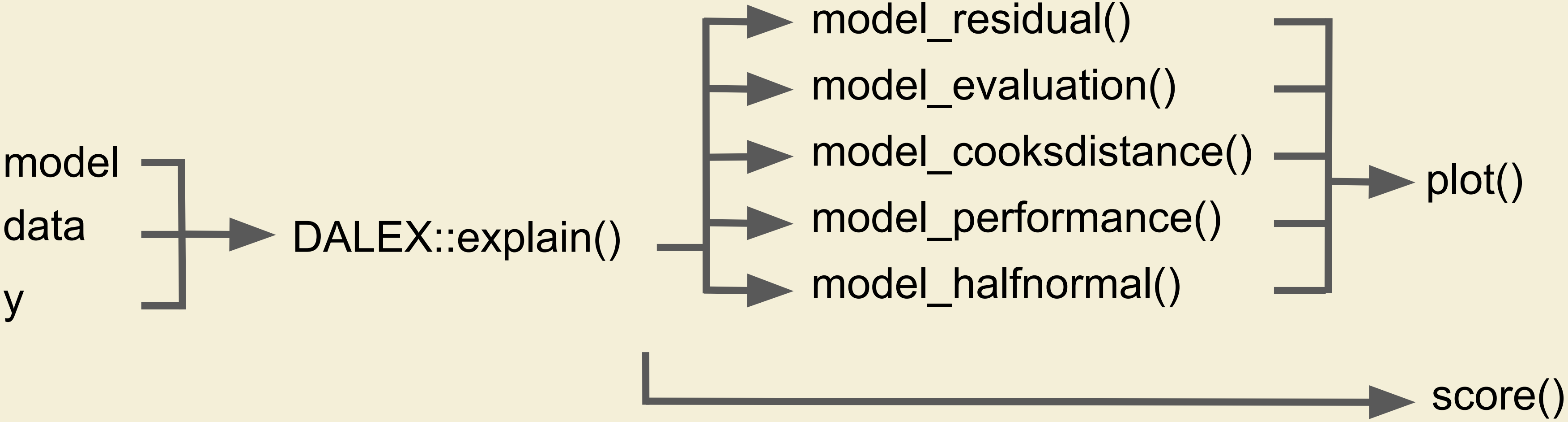
https://modeloriented.github.io/ModelDevelopmentProcess/


mlr
H2O
scikit learn
tidymodels
...
DALEX
auditor
iBreakDown
ingredients
Model ranking radar
lm
randomForest
ROC curve
True positive fraction
False positive fraction
fixed.acidity
chlorides
free.sulfur.dioxide
total.sulfur.dioxide
density
pH
sulphates
alcohol
prediction
LIFT Chart
True positive
Residuals density by
rf
lm
Residuals
Target variable

and others...

Workflow

auditor



model

data

y

DALEX::explain()

model_residual()

model_evaluation()

model_cooksdistance()

model_performance()

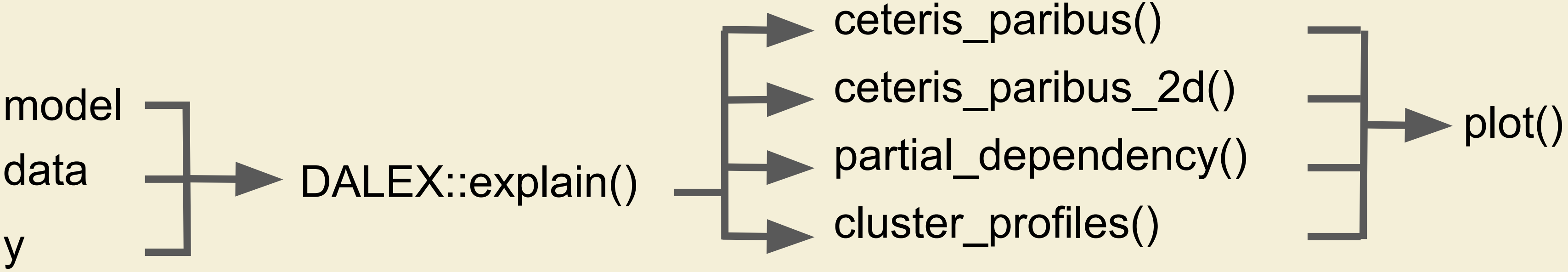
model_halfnormal()

plot()

score()

Workflow

ingredients



model

data

y

DALEX::explain()

ceteris_paribus()

ceteris_paribus_2d()

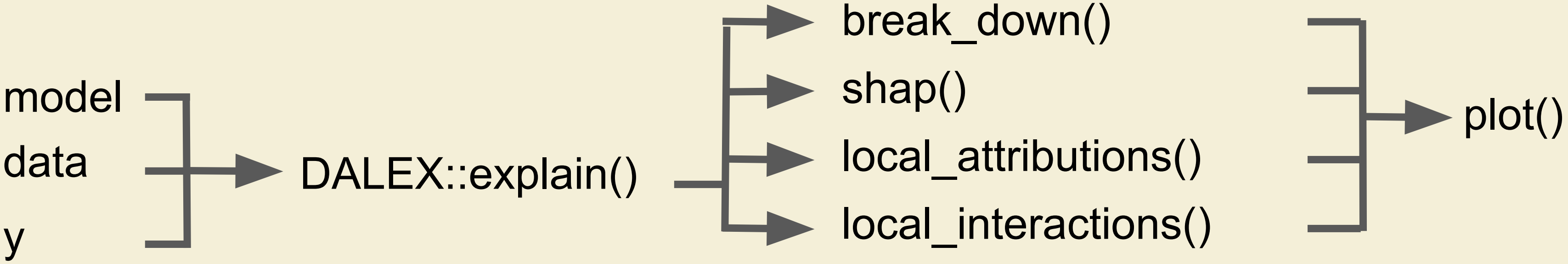
partial_dependency()

cluster_profiles()

plot()

Workflow

iBreakDown



model

data

y

DALEX::explain()

break_down()

shap()

local_attributions()

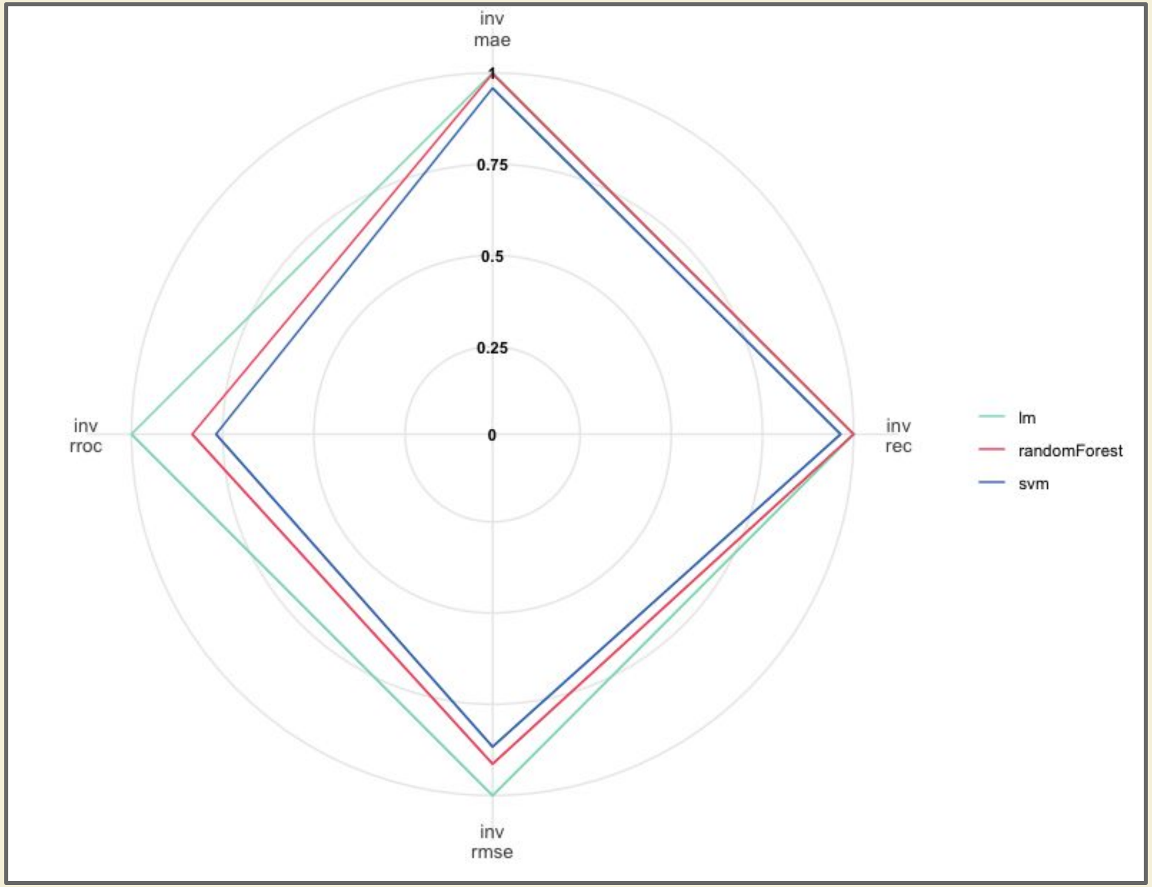
local_interactions()

plot()

Plots

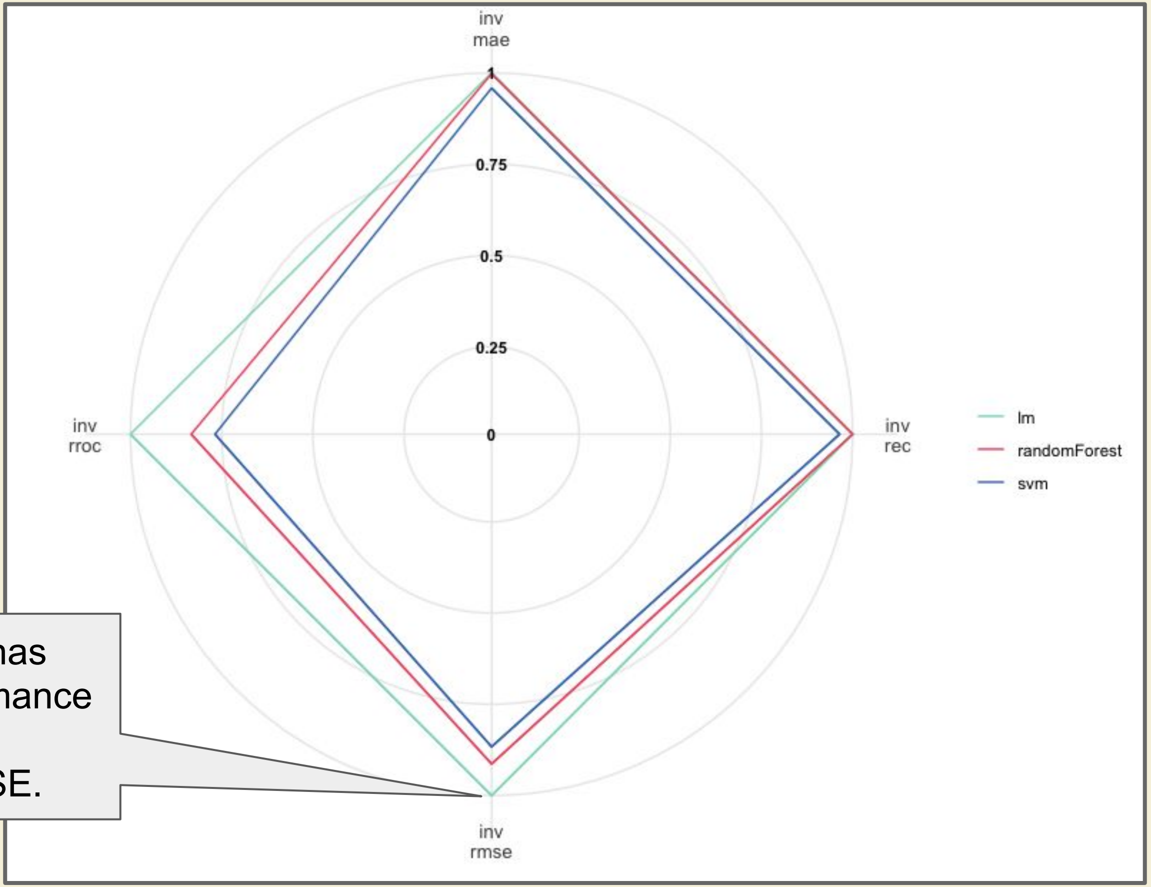
Radar plot

`plot_radar()`



Radar plot

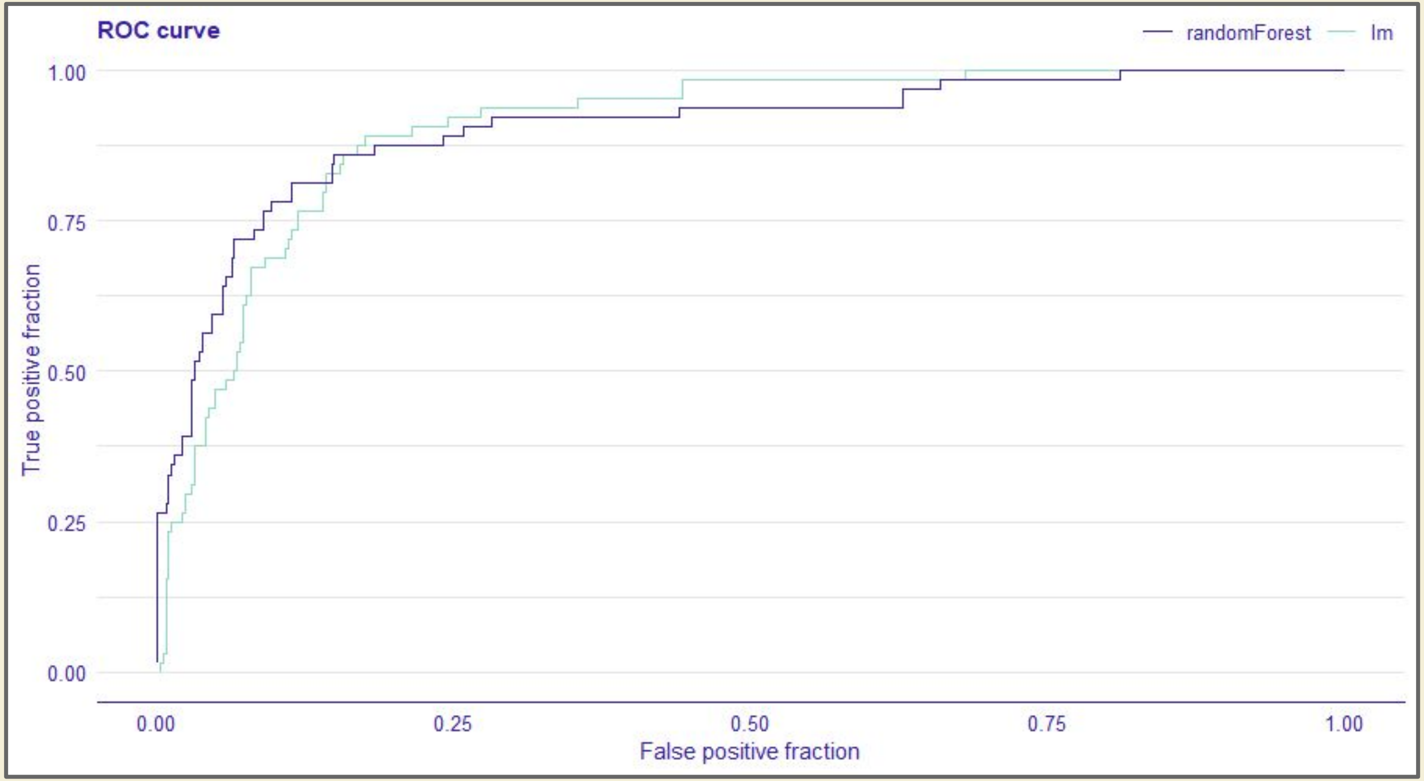
`plot_radar()`



The linear model has the largest performance when models are measured by RMSE.

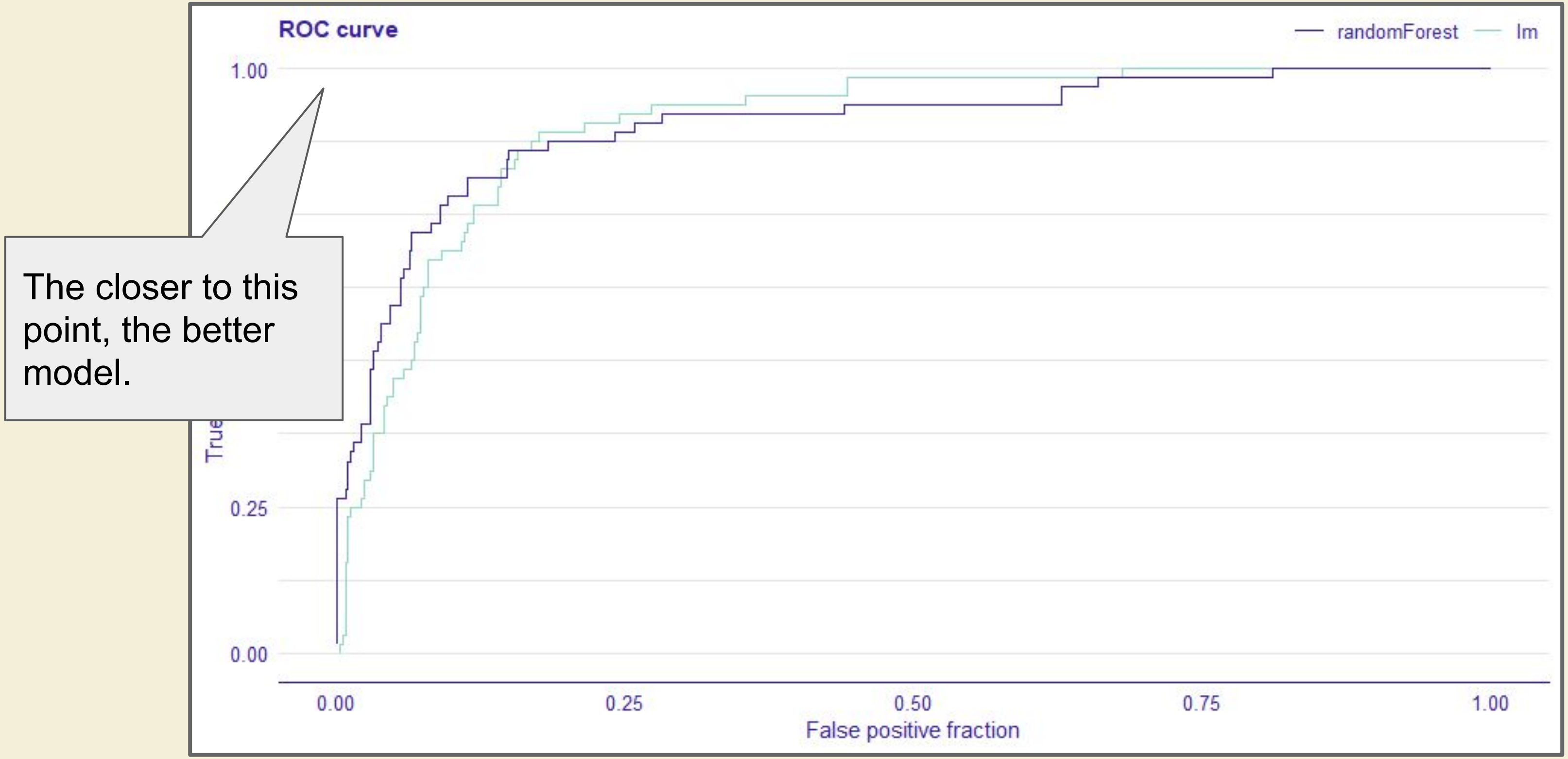
Receiver operating characteristic (ROC)

`plot_roc()`



Receiver operating characteristic (ROC)

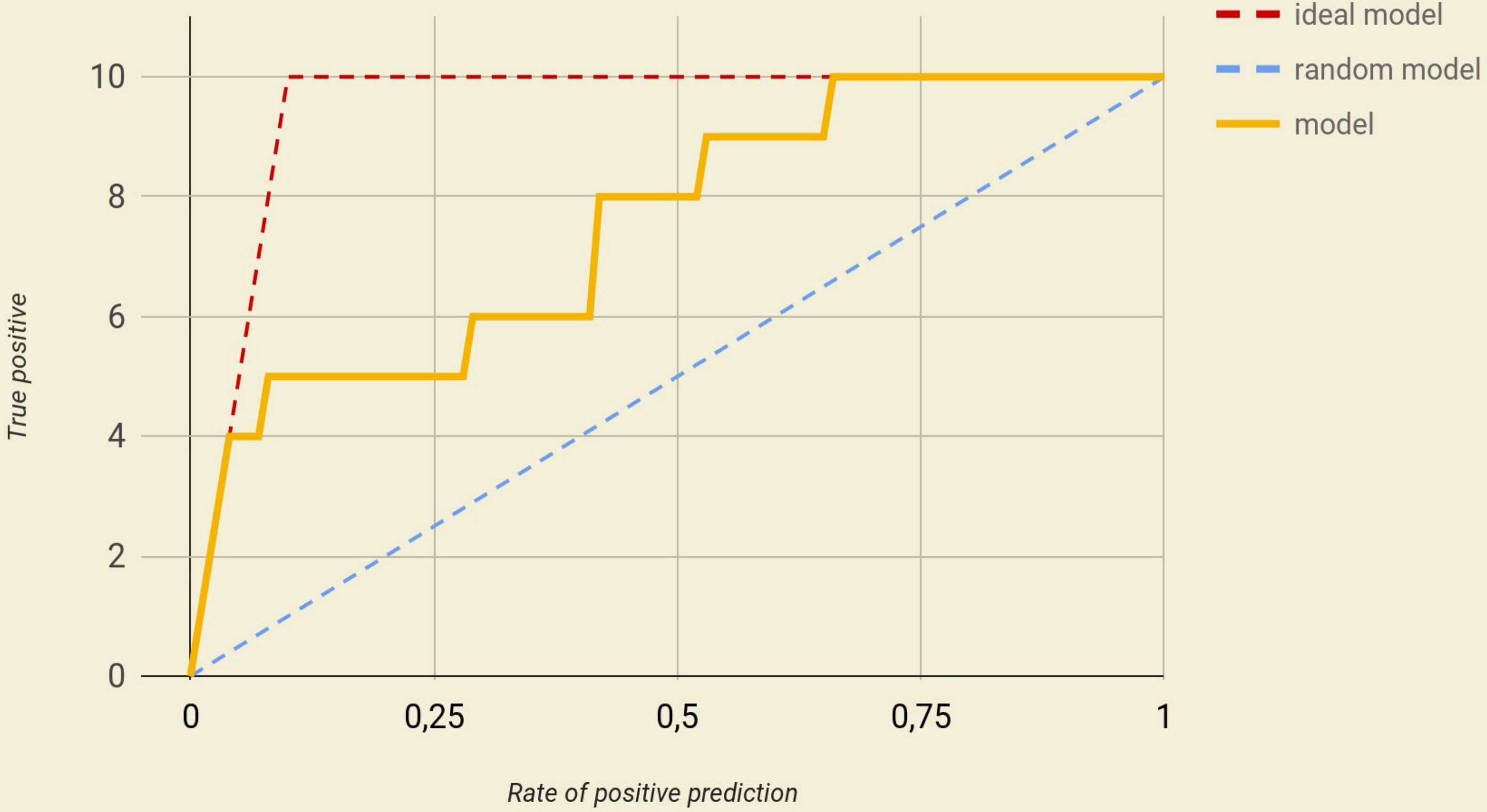
`plot_roc()`



The closer to this point, the better model.

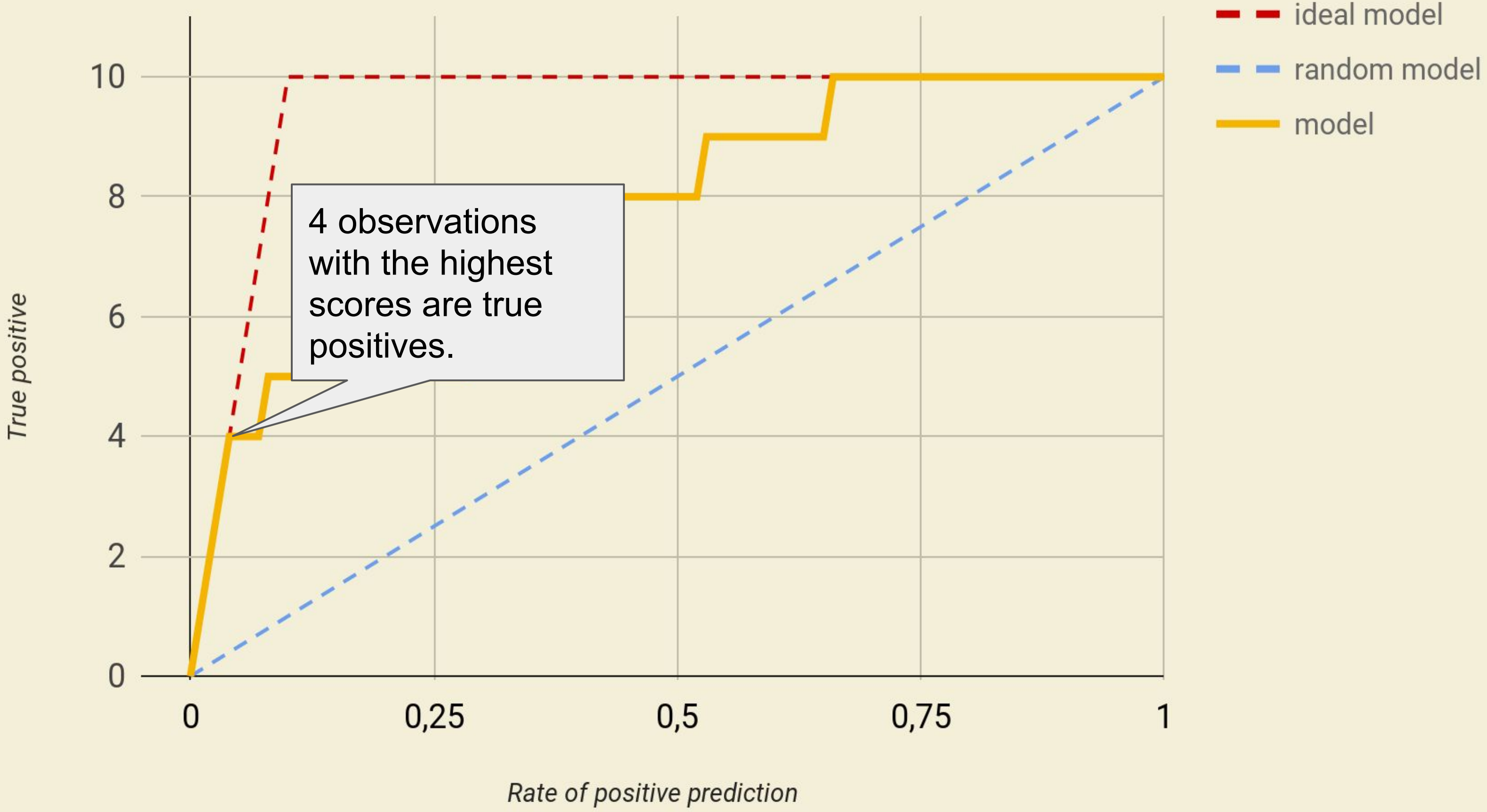
LIFT (toy example)

`plot_lift()`



LIFT (toy example)

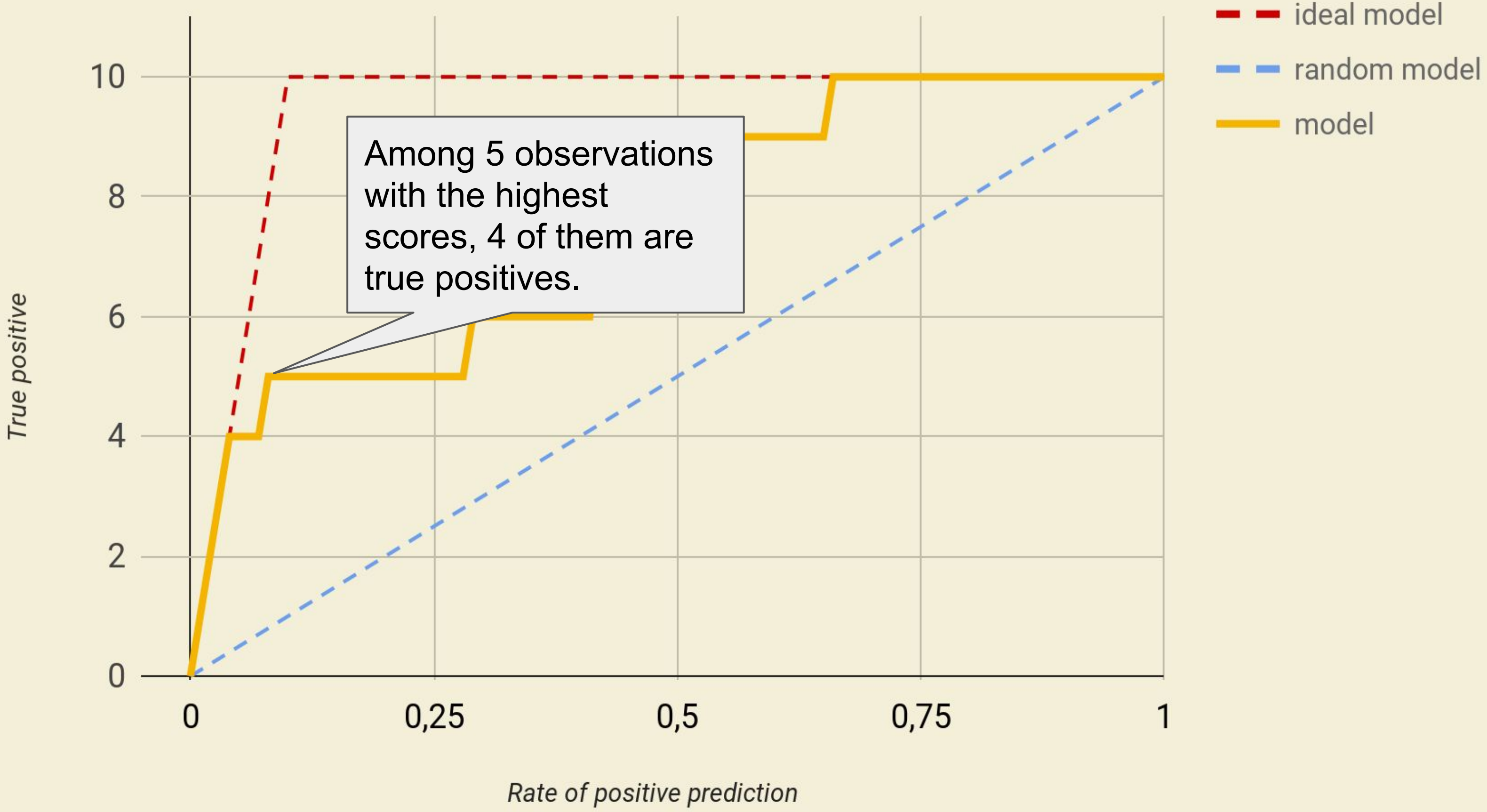
`plot_lift()`



4 observations with the highest scores are true positives.

LIFT (toy example)

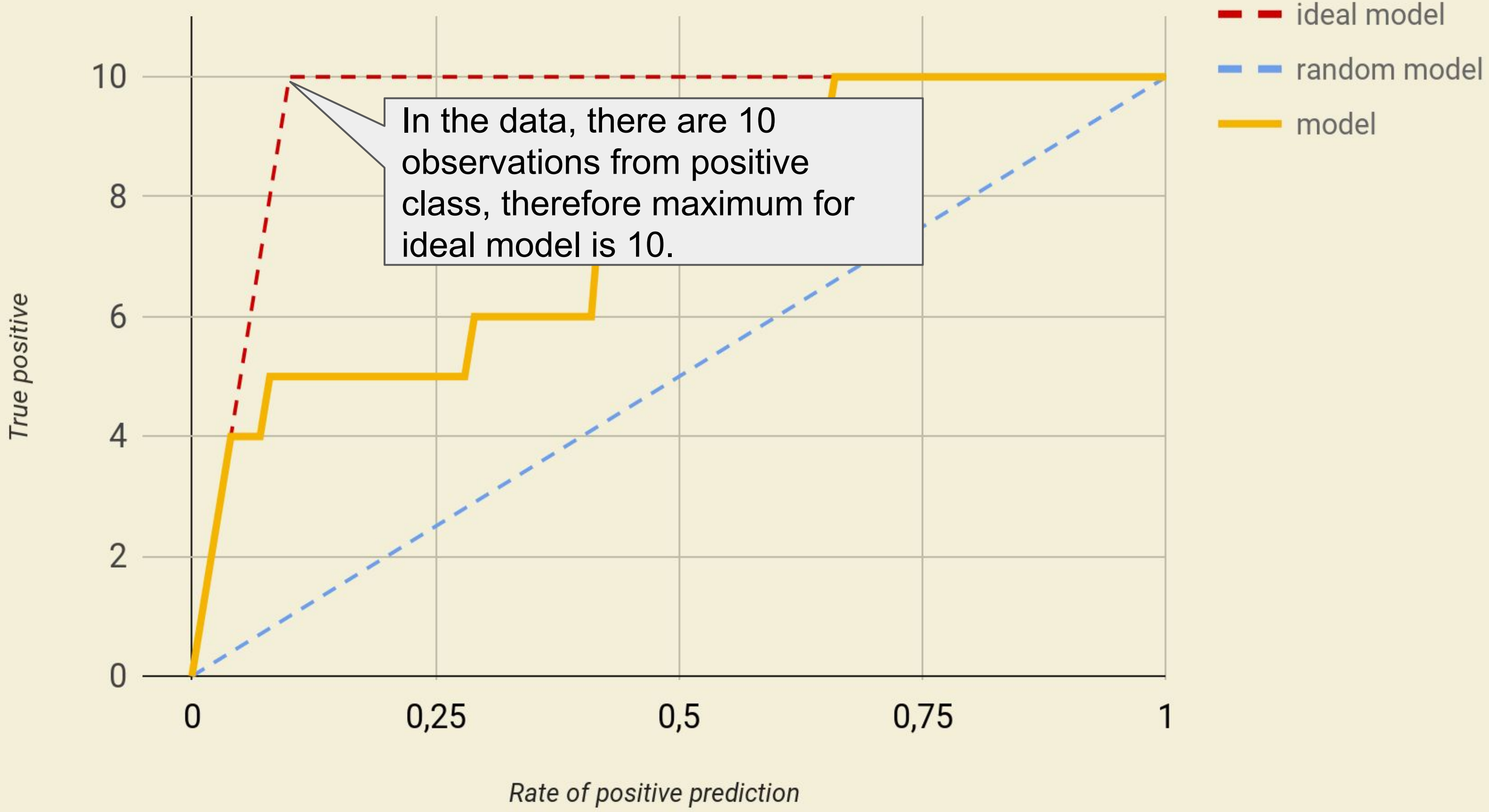
`plot_lift()`



Among 5 observations with the highest scores, 4 of them are true positives.

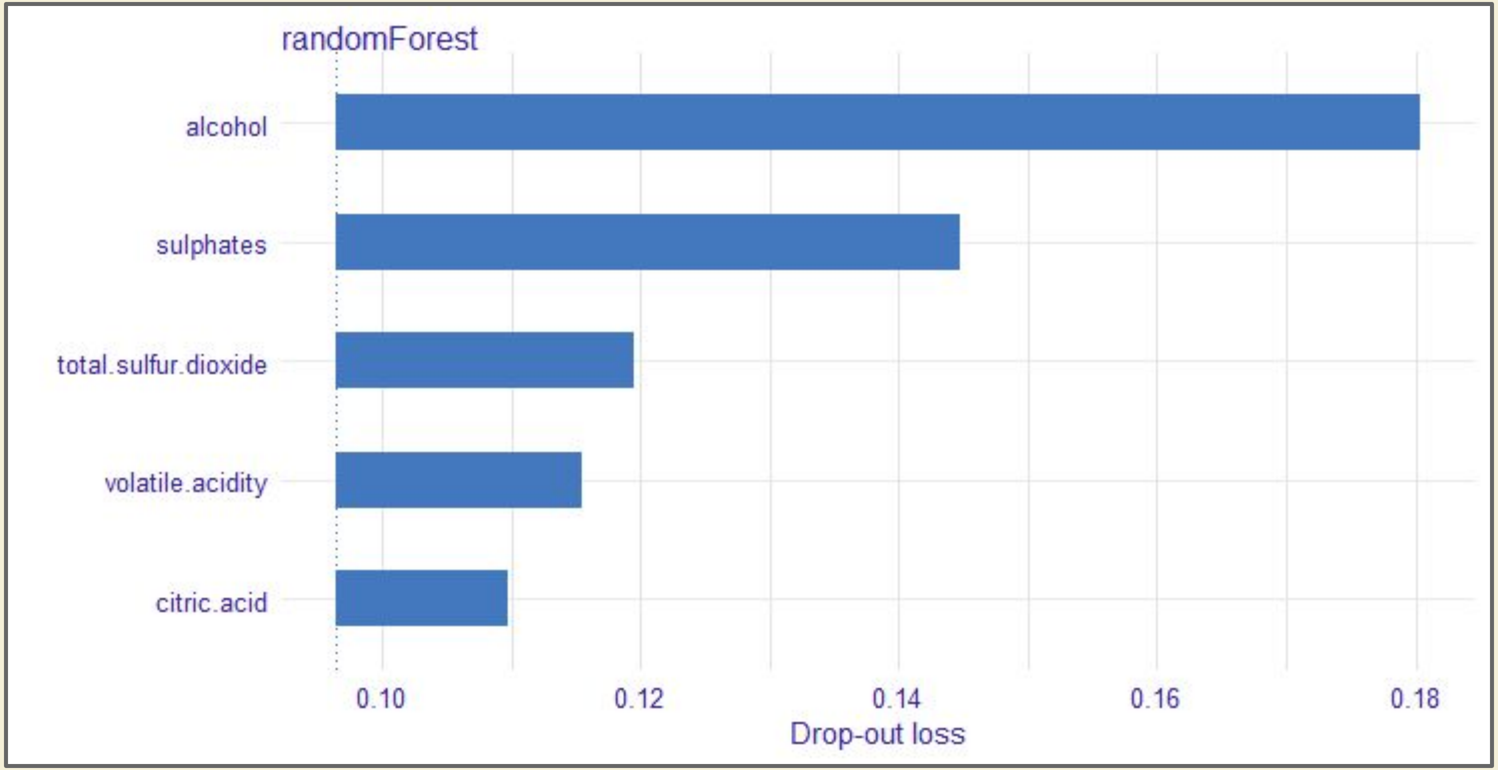
LIFT (toy example)

`plot_lift()`



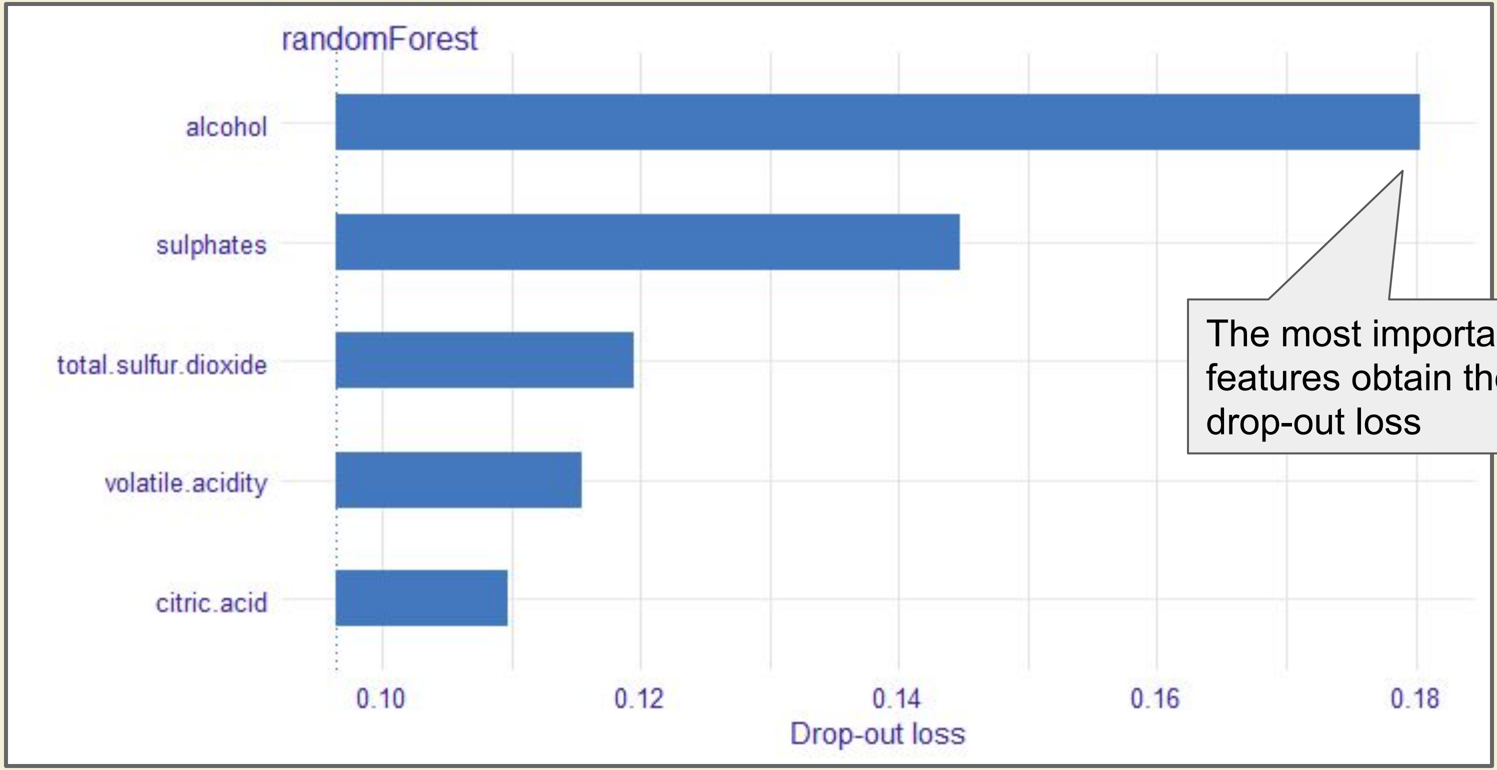
In the data, there are 10 observations from positive class, therefore maximum for ideal model is 10.

Feature importance



https://pbiecek.github.io/PM_VEE/featureImportance.html

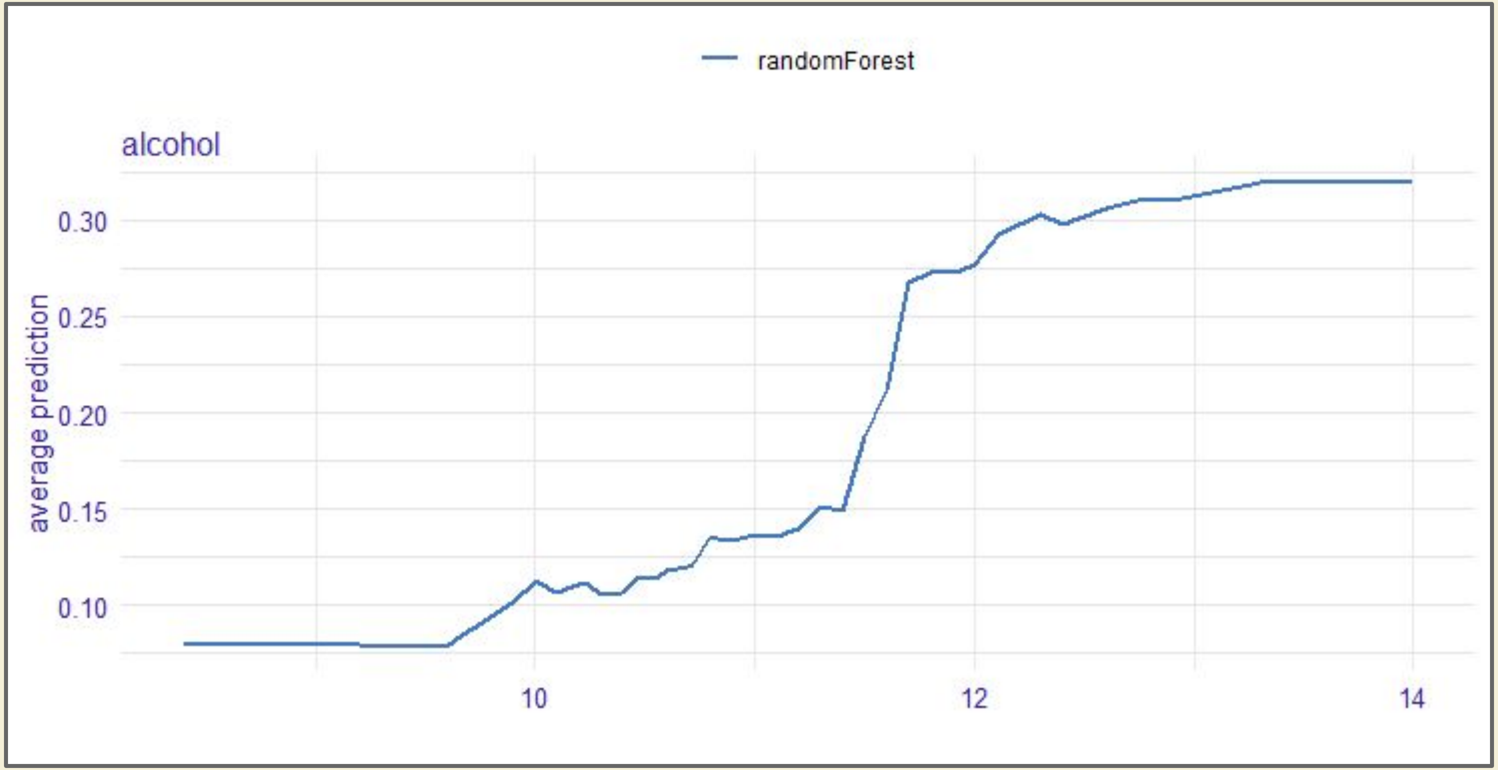
Feature importance



The most important features obtain the highest drop-out loss

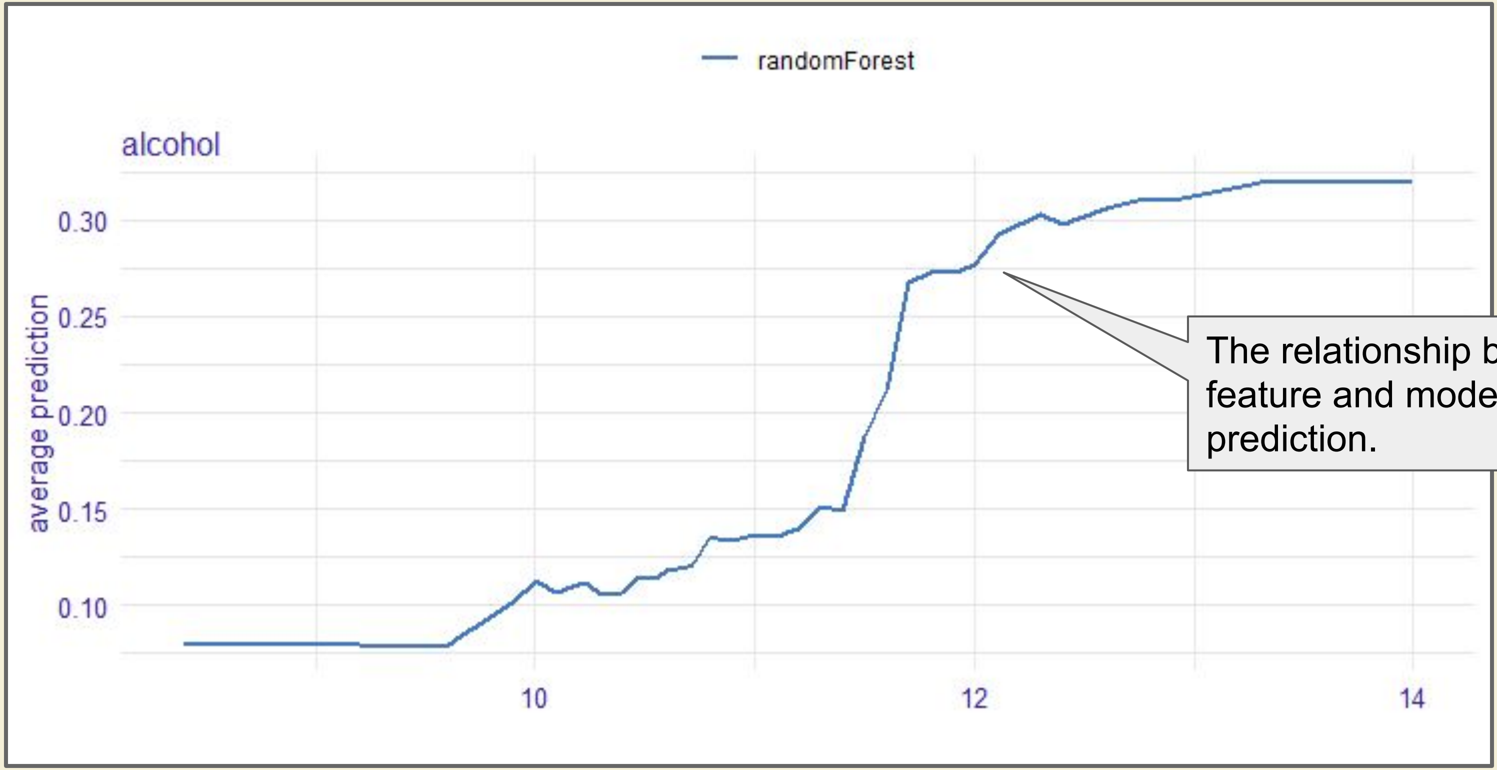
https://pbiecek.github.io/PM_VEE/featureImportance.html

Partial Dependence Plot



https://pbiecek.github.io/PM_VEE/partialDependenceProfiles.html

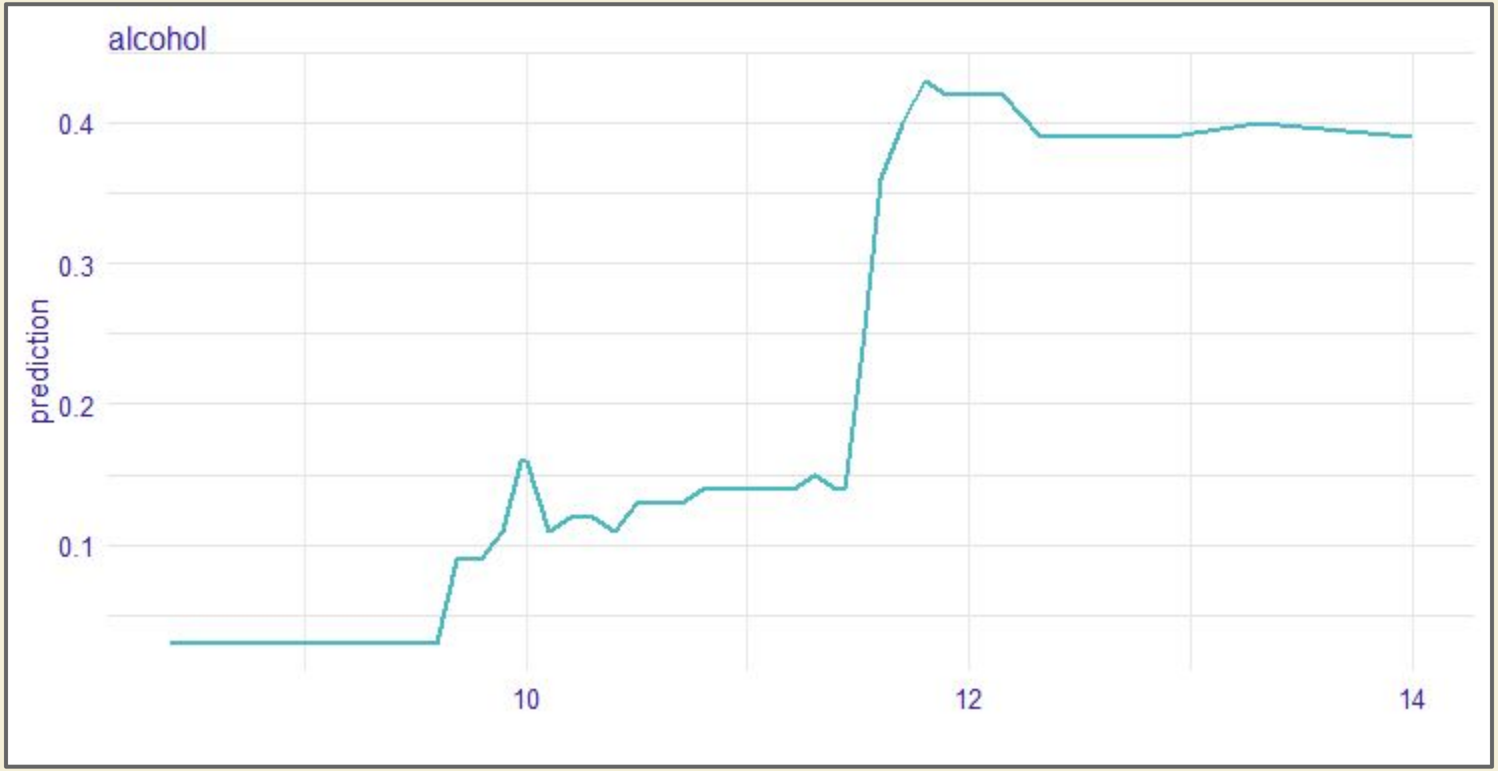
Partial Dependence Plot



The relationship between feature and model’s prediction.

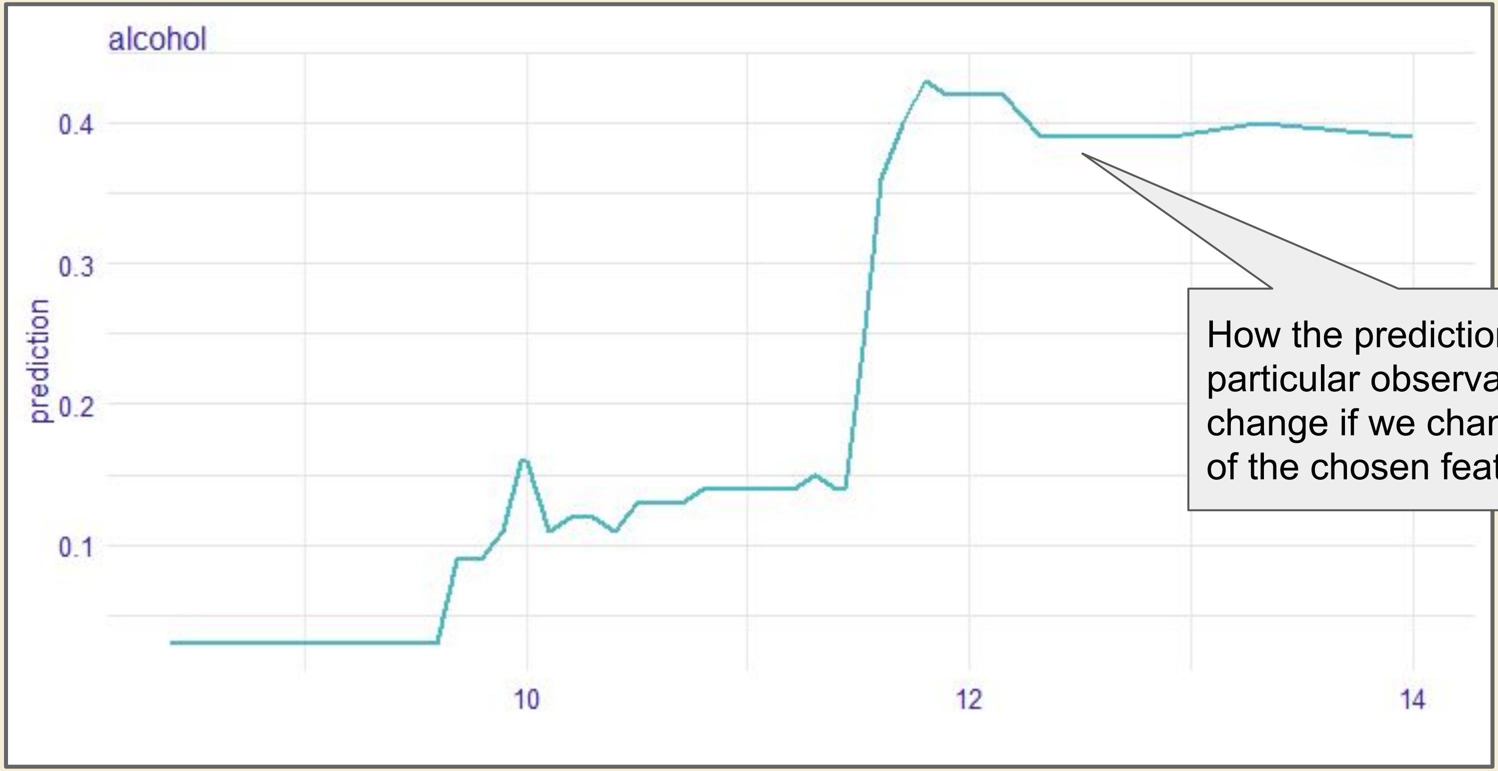
https://pbiecek.github.io/PM_VEE/partialDependenceProfiles.html

Ceteris Paribus Profile



https://pbiecek.github.io/PM_VEE/ceterisParibus.html

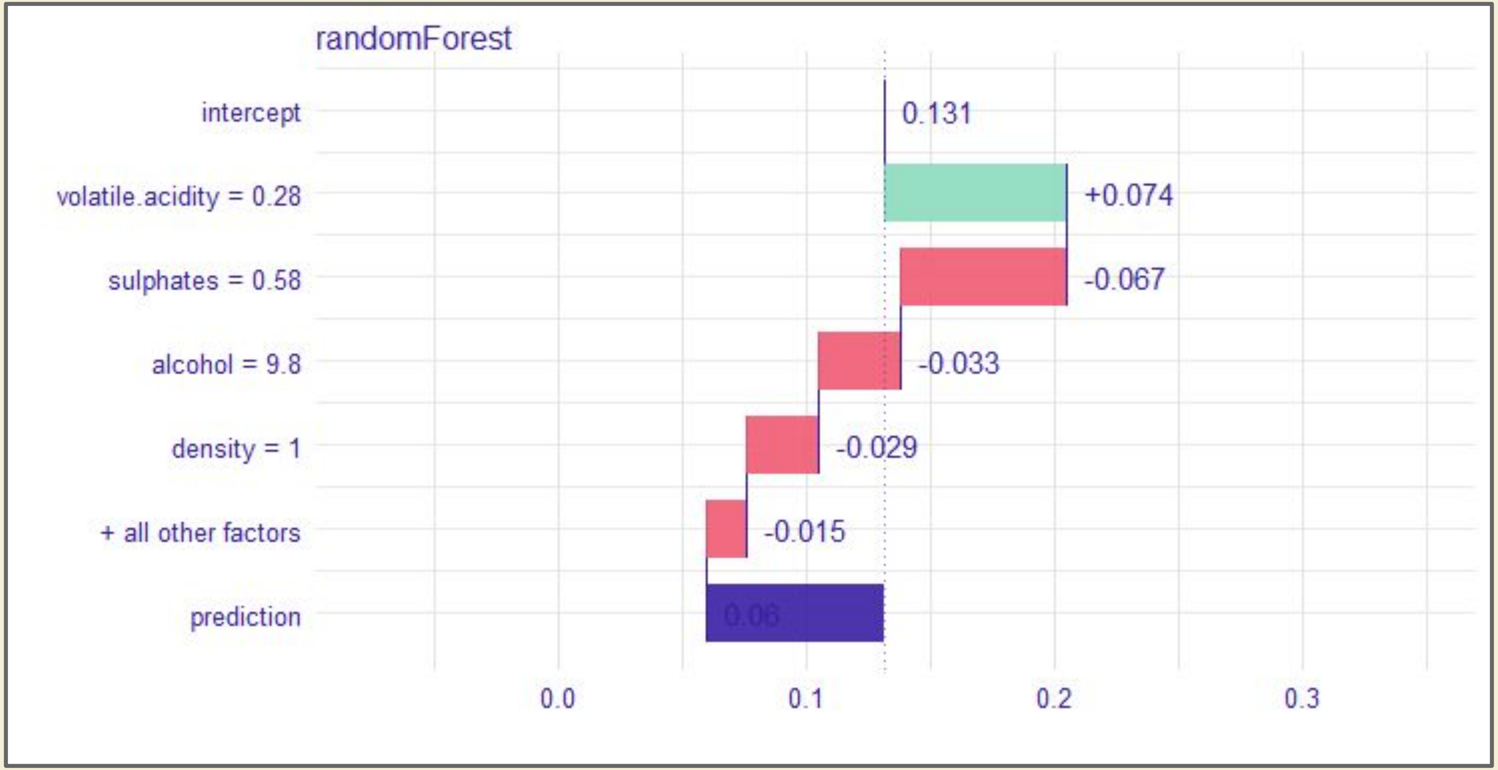
Ceteris Paribus Profile



How the prediction for this particular observation would change if we change value of the chosen feature.

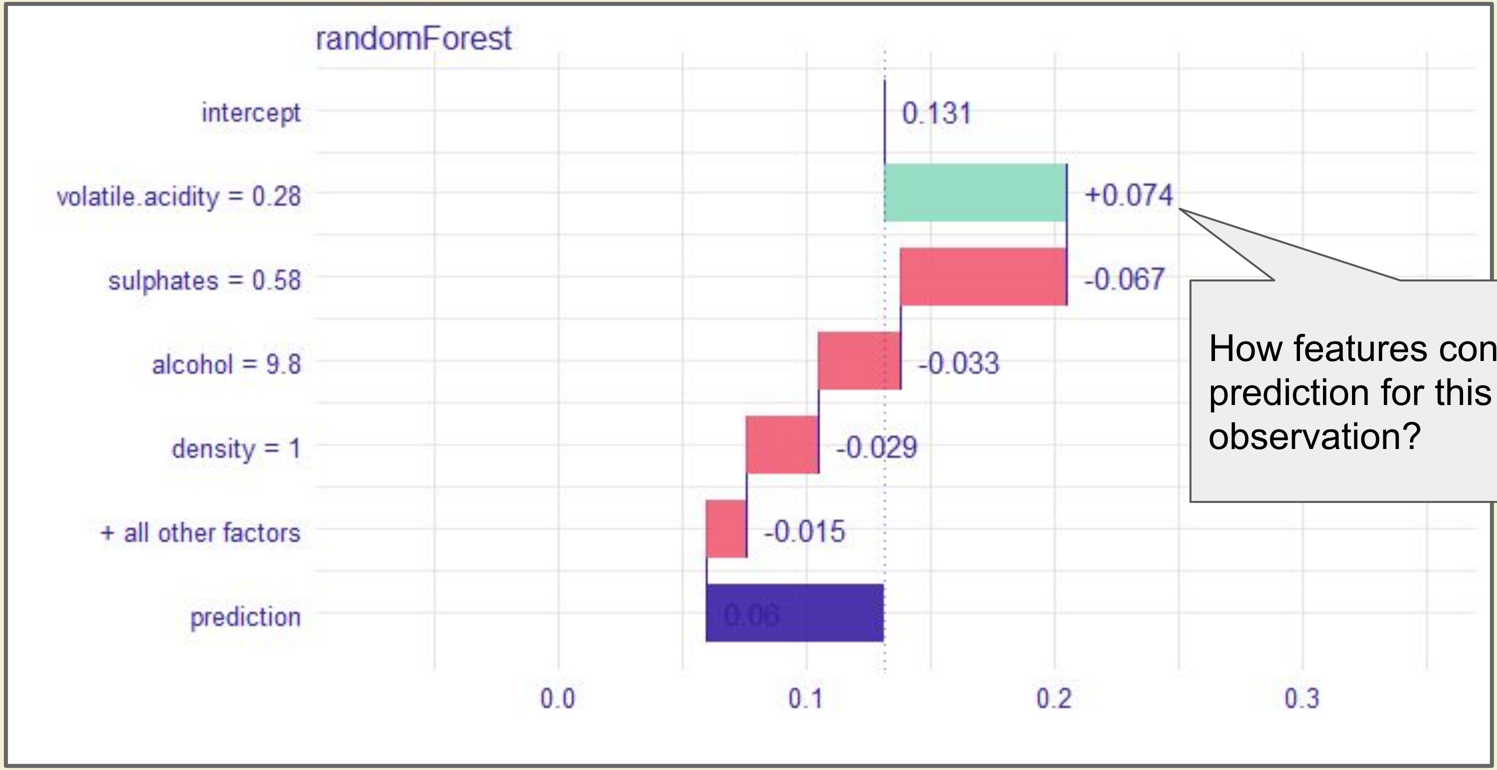
https://pbiecek.github.io/PM_VEE/ceterisParibus.html

Break Down



https://pbiecek.github.io/PM_VEE/breakDown.html

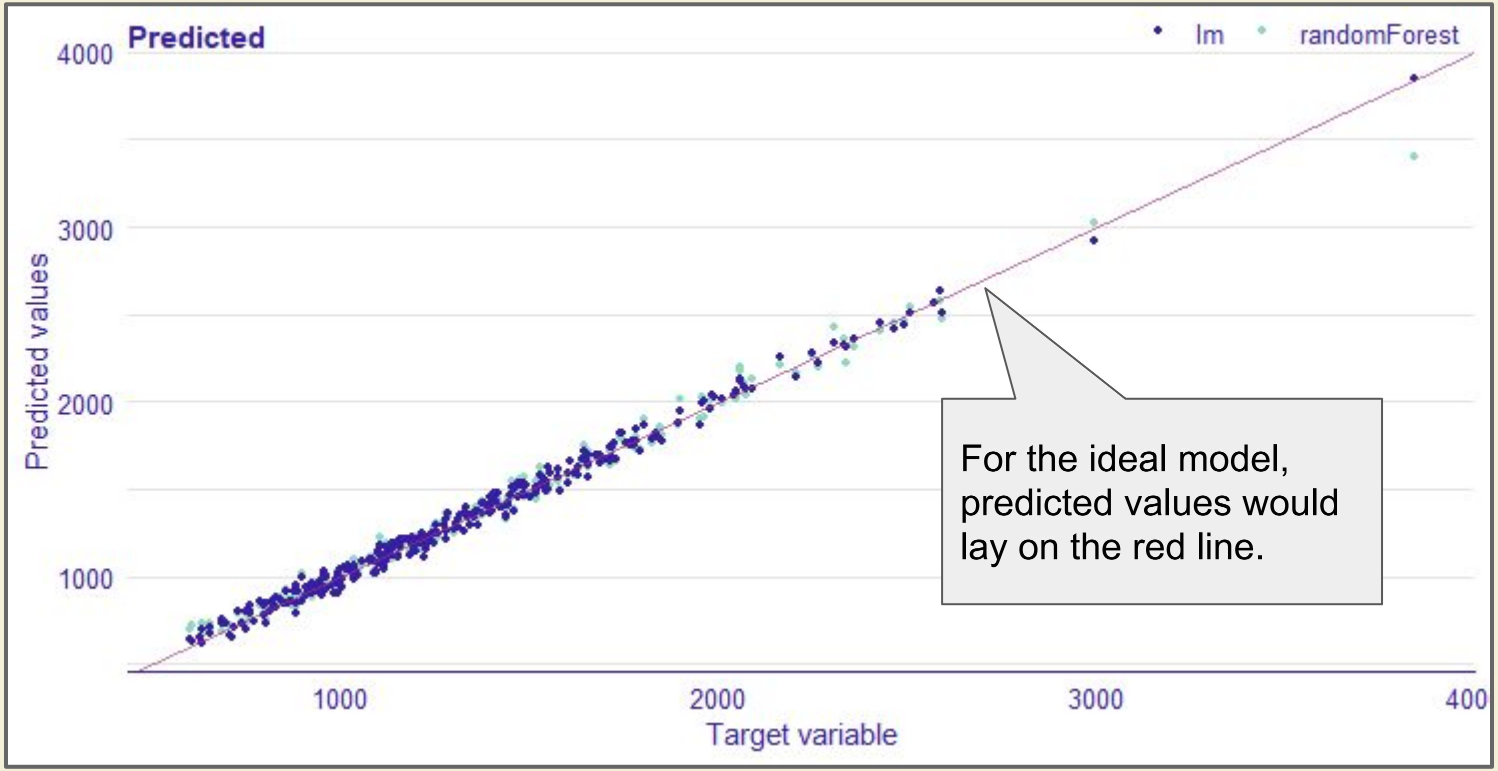
Break Down



How features contribute into prediction for this particular observation?

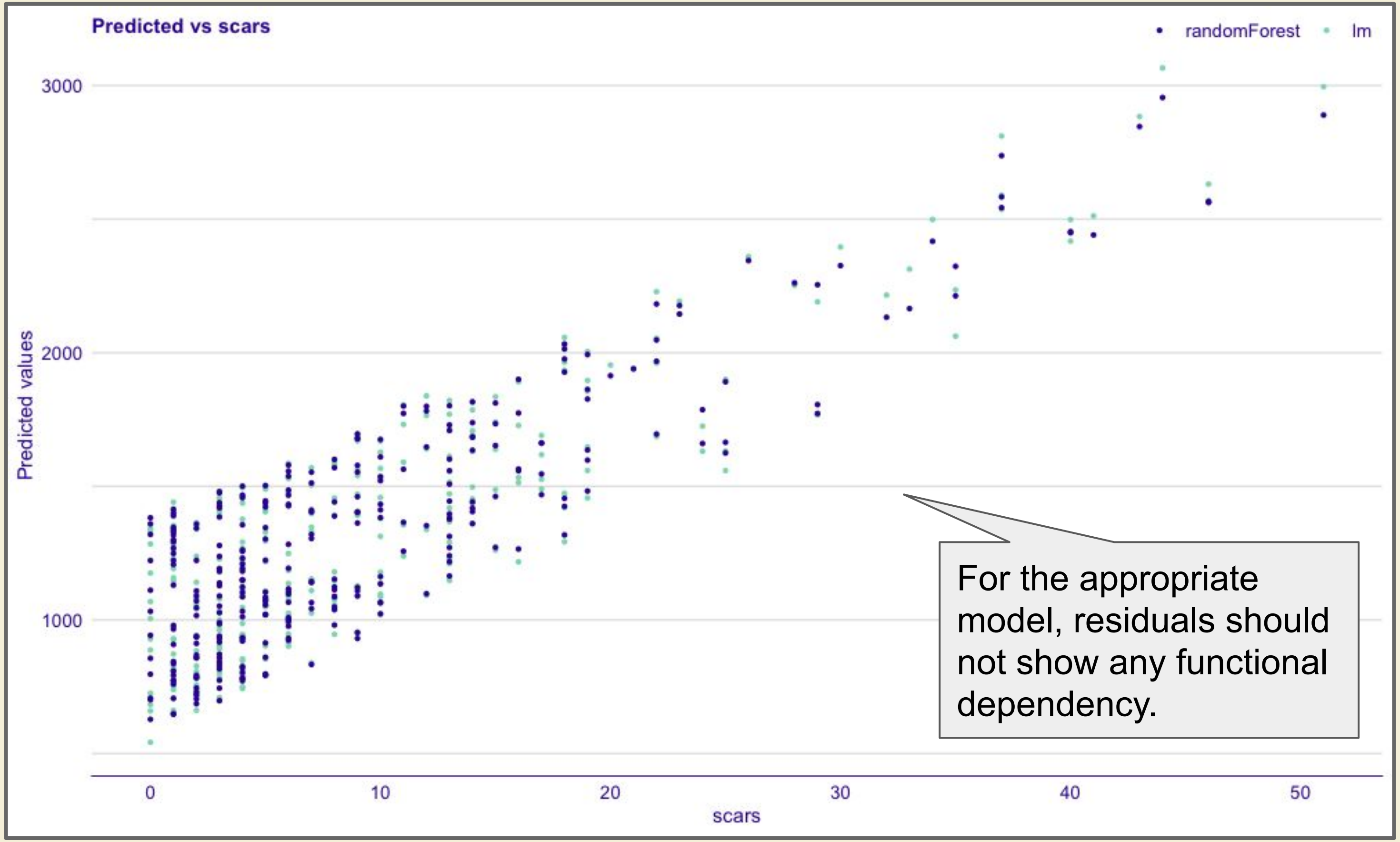
https://pbiecek.github.io/PM_VEE/breakDown.html

Plot predictions (against target variable)



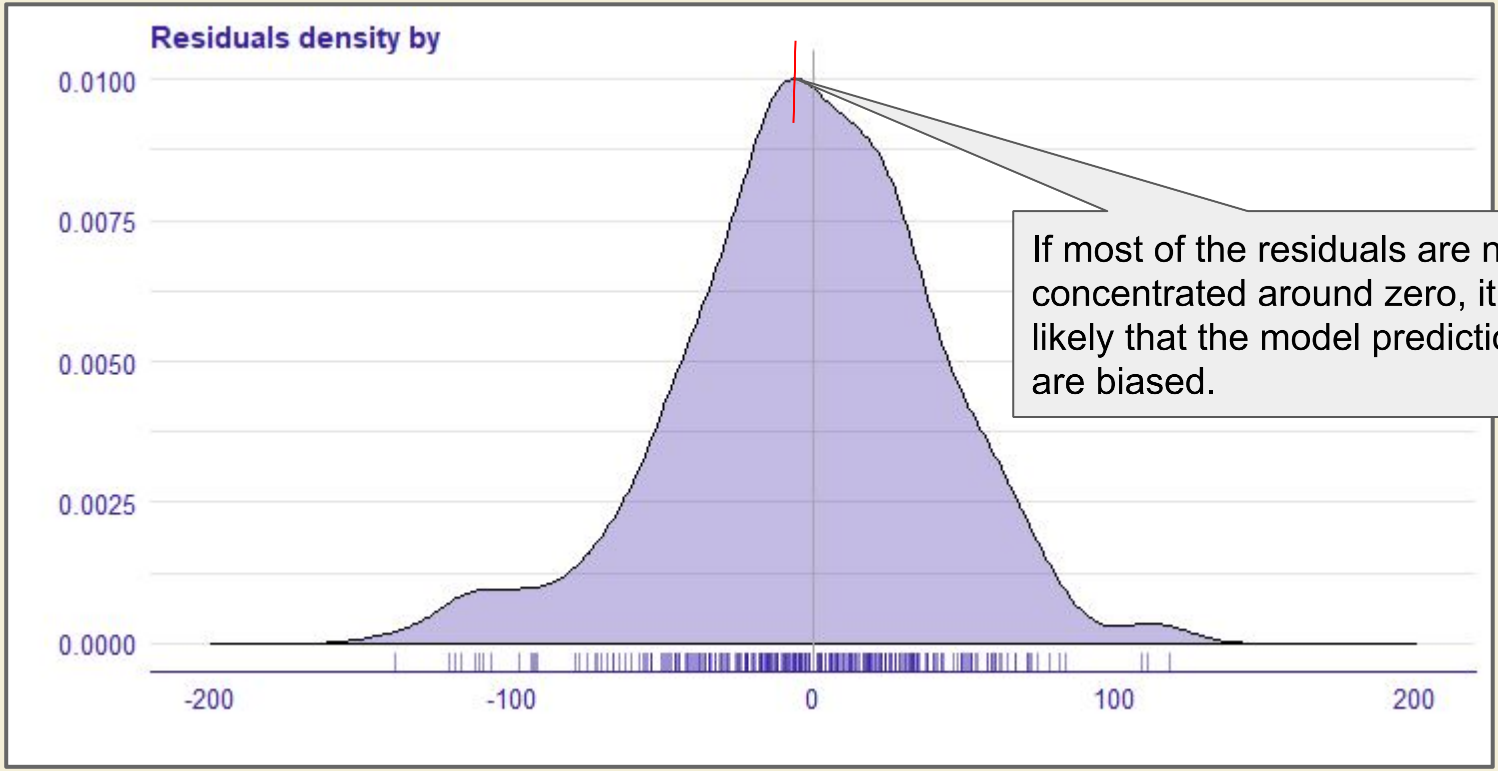
For the ideal model, predicted values would lay on the red line.

Plot predictions (against any variable)



For the appropriate model, residuals should not show any functional dependency.

Residual Density



If most of the residuals are not concentrated around zero, it is likely that the model predictions are biased.

Data sets

Wine quality (1)

Data set used to predict human wine taste preferences (part of dataset of red and white variants of the Portuguese "*Vinho Verde*" wine).

Contains physicochemical (inputs) and sensory (the output) variables (no data about grape types, wine brand, wine selling price, etc.).

The datasets can be viewed as classification or regression tasks.

Wine quality (2)

The classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines.

Interesting in testing feature selection methods, as it is not clear if all input variables are relevant.

The details are described in: P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. *Modeling wine preferences by data mining from physicochemical properties*. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.

Wine quality - variable description (1)

Fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily)

Volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste

Citric acid: found in small quantities, citric acid can add ‘freshness’ and flavor to wines

Residual sugar: the amount of sugar remaining after fermentation stops, it’s rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet

Chlorides: the amount of salt in the wine

Free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine

Wine quality - variable description (2)

Total sulfur dioxide: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine

Density: the density of water is close to that of water depending on the percent alcohol and sugar content

pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale

Sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant

Alcohol: the percent alcohol content of the wine

Quality: output variable (based on sensory data, score between 0 and 10)

Dragons data

Values are generated in a way to: have nonlinearity in **year_of_birth** and height and have **concept drift** in the test set

year_of_birth:	year in which the dragon was born. Negative year means year BC, eg: -1200 = 1201 BC
year_of_discovery:	year in which the dragon was found.
height:	height of the dragon in yards.
weight:	weight of the dragon in tons.
scars:	number of scars.
colour:	colour of the dragon.
number_of_lost_teeth:	number of teeth that the dragon lost.
life_length:	life length of the dragon.

MI2DataLab research lab is looking for you!

Are you intreseted in XAI, AutoML, AutoEDA other innovations in the next generation of ML?

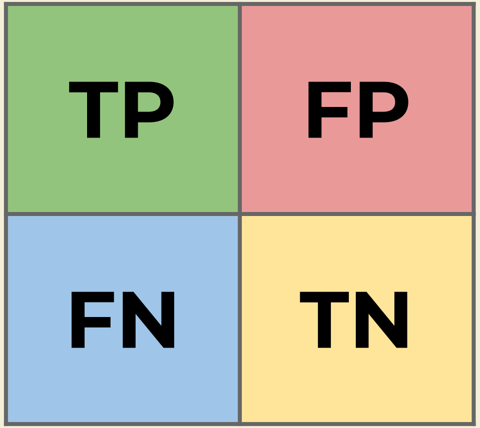
Come to us http://bit.do/MI2isHiring

Appendix

Confusion matrix

	Truth	
Prediction	**True-Positive**	**False-Positive**
	False-Negative	**True-Negative**

Confusion matrix - some scores



TP	FP
FN	TN

Accuracy: (TP + TN) / (TP + TN + FP + FN)

Recall: TP / (TP + FN)

Precision: TP / (TP + FP)

F-measure: (2 * Recall * Precision) / (Recall + Precision)