An Overview on Explainable AI using R

Gero Szepannek
Statistics, Business Mathematics & ML
Stralsund University of Applied Sciences

ML vs. Traditional Statistics…

1. In ML the type of the relationship between dependent and independent variables is typically not pre-specified in advanced but left to the algorithm.

2. …but the *search space* of potential functions is controlled by the algorithm’s *hyperparameters* (*regularization*).

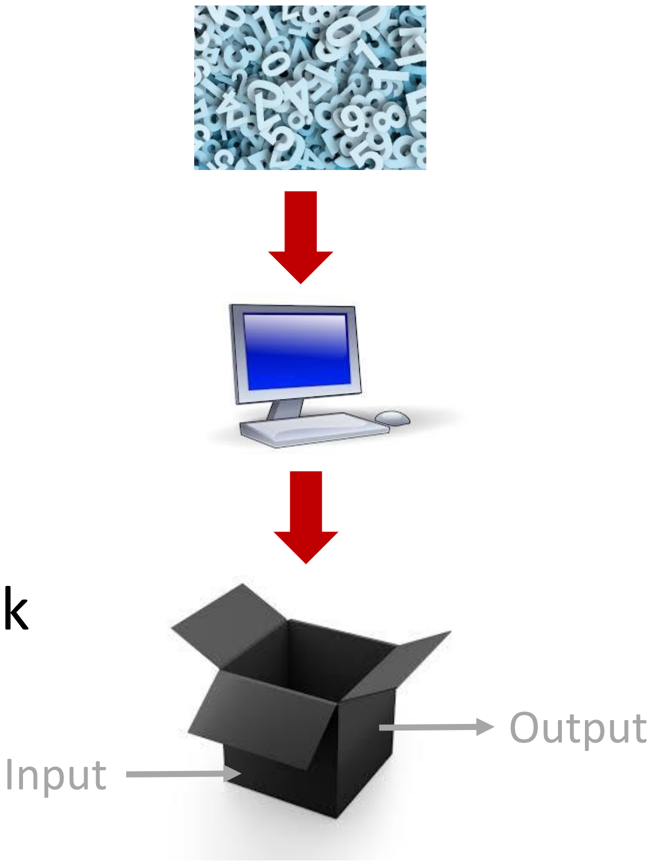
3. …As a consequence the resulting models typically are of black box type.

ML vs. Traditional Statistics…

1. In ML the type of the relationship between dependent and independent variables is typically not pre-specified in advanced but left to the algorithm.

2. …but the *search space* of potential functions is controlled by the algorithm’s *hyperparameters* (*regularization*).

3. …As a consequence the resulting models typically are of black box type.



Need For explainable AI

Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process

GDPR

"[...] the controller shall provide [...] the following information: the existence of automated decision-making and [...] meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject."

Art. 13-15 & 22 Regulation (EU) 2016/679

https://dsgvo-gesetz.de/

https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine

Need For explainable AI

Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process

https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine

Note: Different requirements on 'explainability'

GDPR

"[...] the controller shall provide [...] the following information: the existence of automated decision-making and [...] meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject."

Art. 13-15 & 22 Regulation (EU) 2016/679
https://dsgvo-gesetz.de/

Different Requirements on Explainability

Question:

Explain…

a) entire model or

b) single predictions?

Understand…

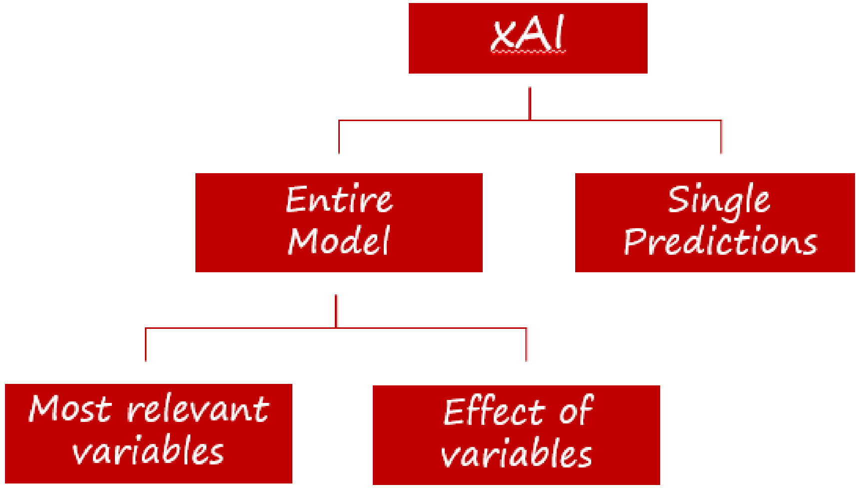
a) which are the most relevant variables or

b) effect of single variables on the output?

- xAI
  - Entire Model
    - Most relevant variables
      - Variable Importance
    - Effect of variables
      - Partial Dependency
      - Accumulated Local Effects (ALE)
  - Single Predictions
    - LIME / Local regression
    - Shapley Values / Breakdown Plots

Overview on Packages and Functionalities

Package	Variable Importance	PDP	ALE	ICE	2D	surrogate trees	merging path	LIME	local regression	Shapley Values	breakdown	live
randomForest	x											
gbm	x											
xgboost	x											
mlr	x	x			x							
caret	x											
DALEX	x	x	x				x				x	x
iml	x	x	x	x	x	x			x	x		
pdp		x		x	x							
ALEPlot		x	x		x							
ICEbox		x		x								
lime								x				
shapleyR										x		
	↑	↑	↑		↑			↑		↑		



Two General Frameworks

{ DALEX } (P. Biecek, 2019)

- Unique interface to already existing packages (dependencies), e.g.
  - Variable importance
  - PDP & ALE (1D), merging path plots
  - breakdown, live
- …supports models from: mlr, caret, parsnip, h2o, keras
- Based on `explainer` objects

…since version 0.4 additionally suggests:

- `ingredients`
- `iBreakDown`

{ iml } (C. Molnar, 2019)

- Reimplementation of many algorithms, e.g.
  - Variable importance
  - PDP & ALE , ICE
  - Shapley values, local regression
- …supports models from: mlr & caret
- …based on `Predictor` Objects

Good book to read!

1st Step: Explainers (...example using mlr) (Bischl et al., 2016)

```r
# load example data
library(DALEX)
data("titanic")
str(titanic)
titanic <- titanic[complete.cases(titanic),]


# train classifiers using mlr
library(mlr)
classif_task <- makeClassifTask(data = titanic, target = "survived", positive = "yes")

lrn_rf  <- makeLearner("classif.randomForest", predict.type = "prob")
lrn_glm <- makeLearner("classif.binomial", predict.type = "prob")
lrn_svm <- makeLearner("classif.ksvm", predict.type = "prob")

classif_rf  <- train(lrn_rf, classif_task)
classif_glm <- train(lrn_glm, classif_task)
classif_svm <- train(lrn_svm, classif_task)


# step 1: build explainers for subsequent analysis of the models
library(DALEX)

# define cutomized predcit function for mlr learners
custom_predict_classif <- function(object, newdata)
  return(predict(object, newdata = newdata)$data[,3])
 # function that takes two arguments: model and new data
 # ...and returns numeric vector with predictions

explainer_rf  <- explain(classif_rf, data = titanic, y = titanic$survived == "yes",
                         label = "rf", predict_function = custom_predict_classif)
explainer_glm <- explain(classif_glm, data = titanic, y = titanic$survived == "yes",
                         label="glm", predict_function = custom_predict_classif)
explainer_svm <- explain(classif_svm, data = titanic, y = titanic$survived == "yes",
                         label ="svm", predict_function = custom_predict_classif)
```

https://www.geo.de/geolino/mensch/10493-rtkl-geschichte-die-letzte-nacht-auf-der-titanic

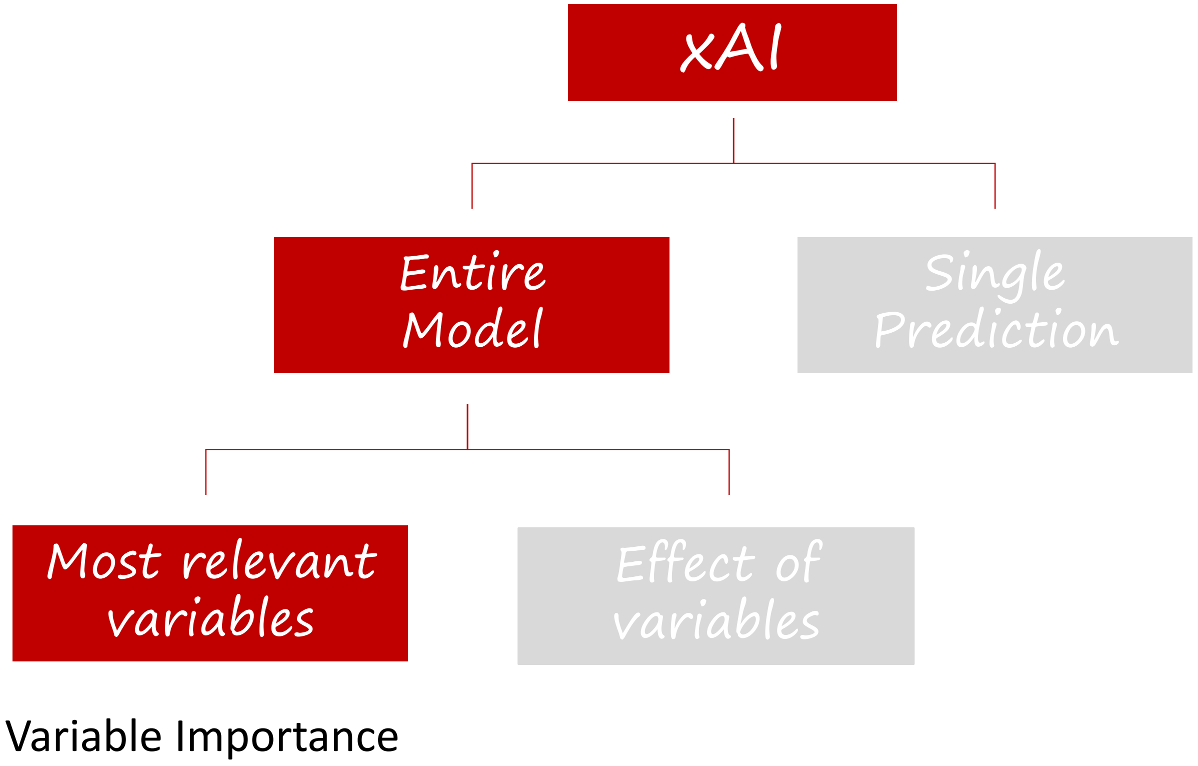
- Explainers provide unique access to different learners for further analysis, required:
  - Model
  - Data
  - Vector of true target values
- ...here: comparison of RF, SVM and logistic regression

Different Requirements on Explainability

Question:

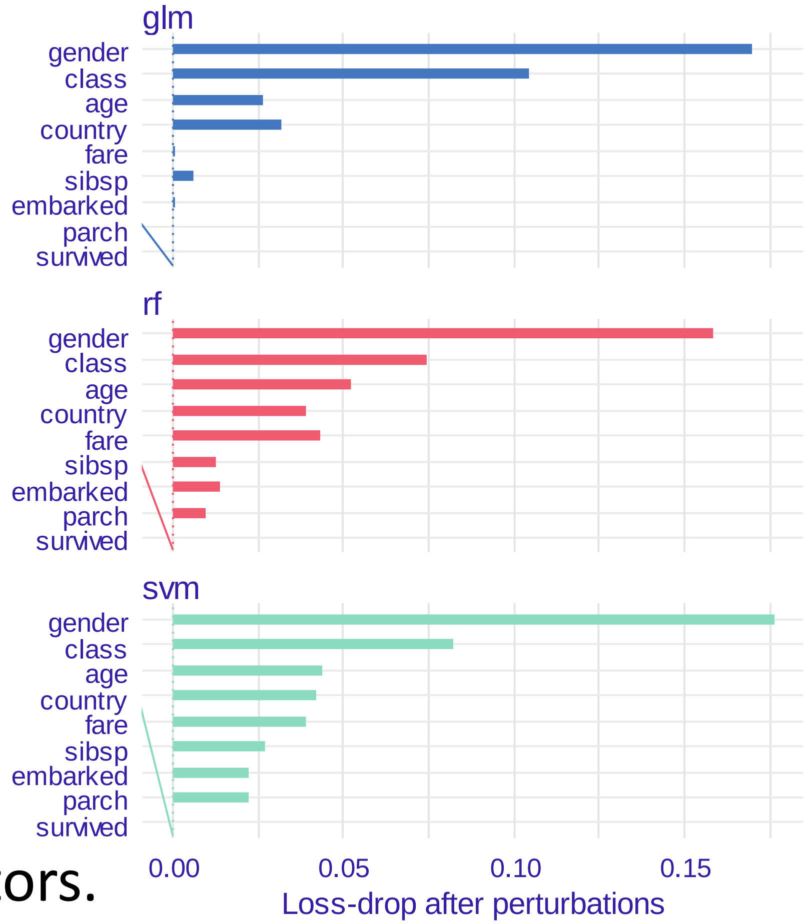
Explain…
a) entire model or
b) single predictions?

Understand…
a) which are the most relevant variables or
b) effect of single variables on the output?



Variable (Permutation) Importance
(Breiman, 2001)

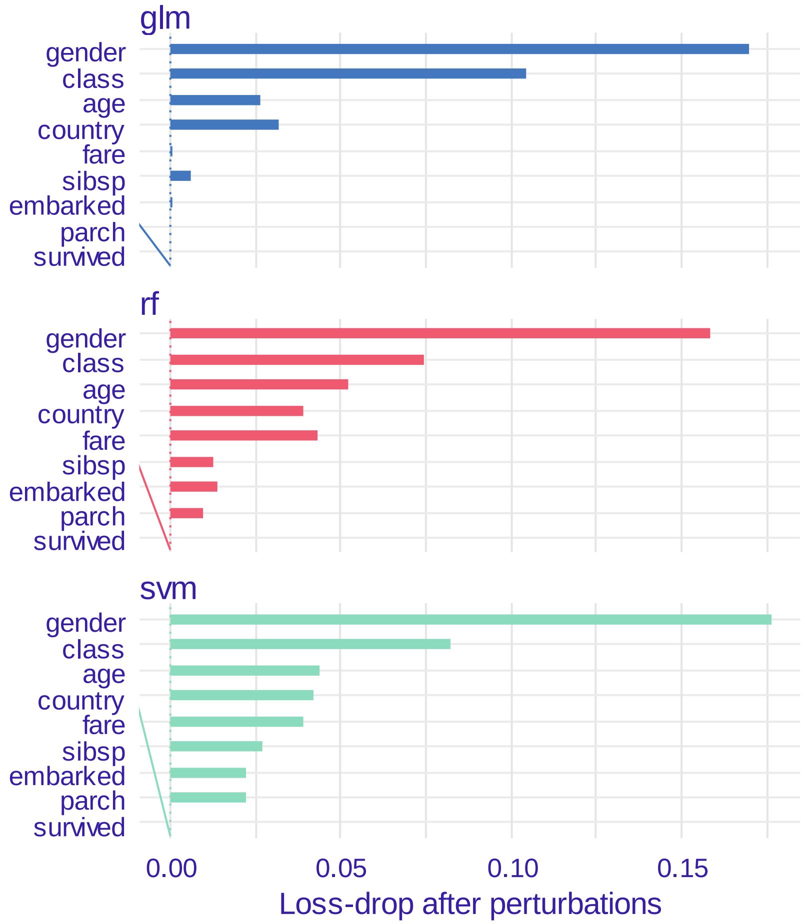
- **Idea: Compute loss in performance if a variable is randomly perturbed.**



- …biased towards # levels.
- …reflects only importance w.r.t. saturated levels (→ backward elimination).
- …be aware of the effect of potential correlated predictors.

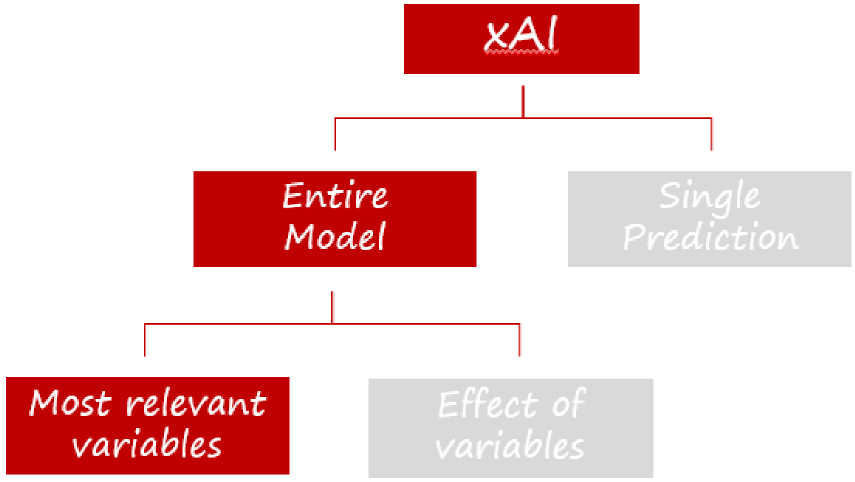
Variable (Permutation) Importance using DALEX

```
# Variable Importance
library(ingredients)

# define customized loss function for variable importance
# ...pre-implemented: loss_root_mean_square(), loss_sum_of_squares()
# ...example: loss_func = function(observed, predicted) sum((observed - logit(predicted))^2))

auc_loss <- function(observed, predicted){
  require(pROC)
  curve <- roc(observed, predicted, levels = c("FALSE","TRUE"), direction = "<")
  # REM: according to help: levels = c("controls", "cases"), direction = controls < cases
  1-auc(curve)
}

# calculate variable importance
vimp_rf  <- feature_importance(explainer_rf, loss_function = auc_loss, type = "difference")
vimp_glm <- feature_importance(explainer_glm, loss_function = auc_loss, type = "difference")
vimp_svm <- feature_importance(explainer_svm, loss_function = auc_loss, type = "difference")

vimp_rf
# ...note AUC _full_model_ = _baseline_ + 0.5
plot(vimp_rf, vimp_glm, vimp_svm, bar_width = 2) #, max_vars = 20)
```



Other Packages…

Package	Variable Importance	PDP	ALE	ICE	2D	surrogate trees	merging path	LIME	local regression	Shapley Values	breakdown	live
randomForest	x											
gbm	x											
xgboost	x											
mlr	x	x			x							
caret	x											
DALEX	x	x	x				x				x	x
iml	x	x	x	x	x	x			x	x		
pdp		x		x	x							
ALEPlot		x	x		x							
ICEbox		x		x								
lime								x				
shapleyR										x		



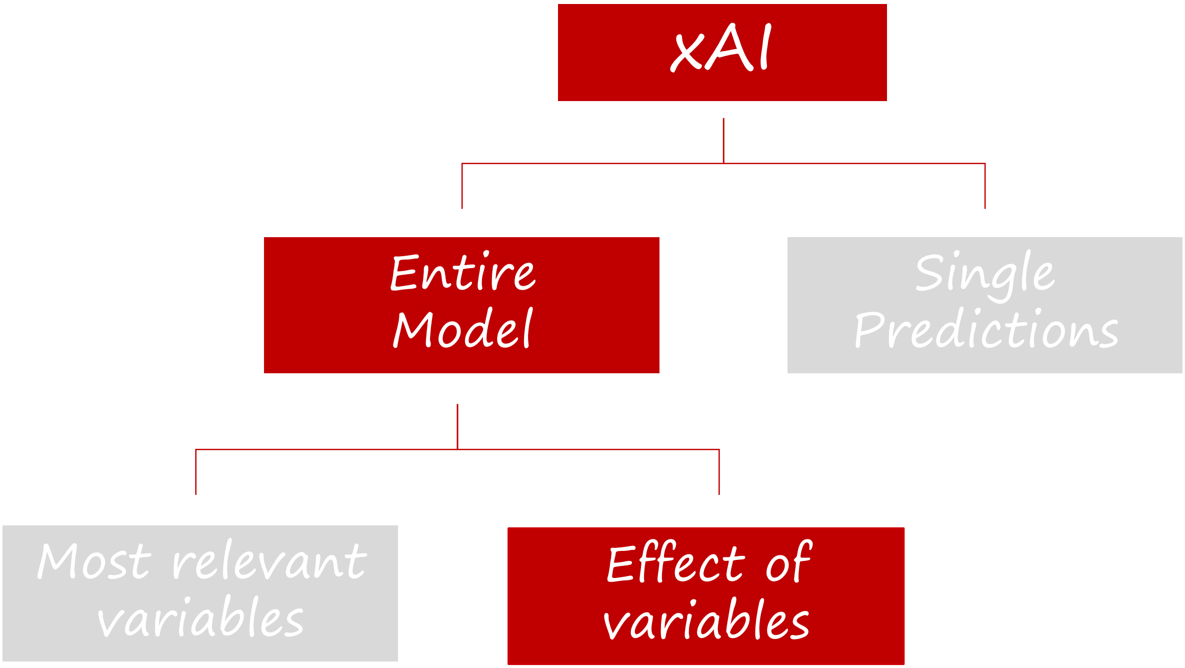
Variable Importance

Different Requirements on Explainability

Question:

Explain…
a) entire model or
b) single predictions?

Understand…
a) which are the most relevant variables or
b) effect of single variables on the output?



Partial Dependency
ALE Plots

Partial Dependency Plots
(Friedman, 2001)

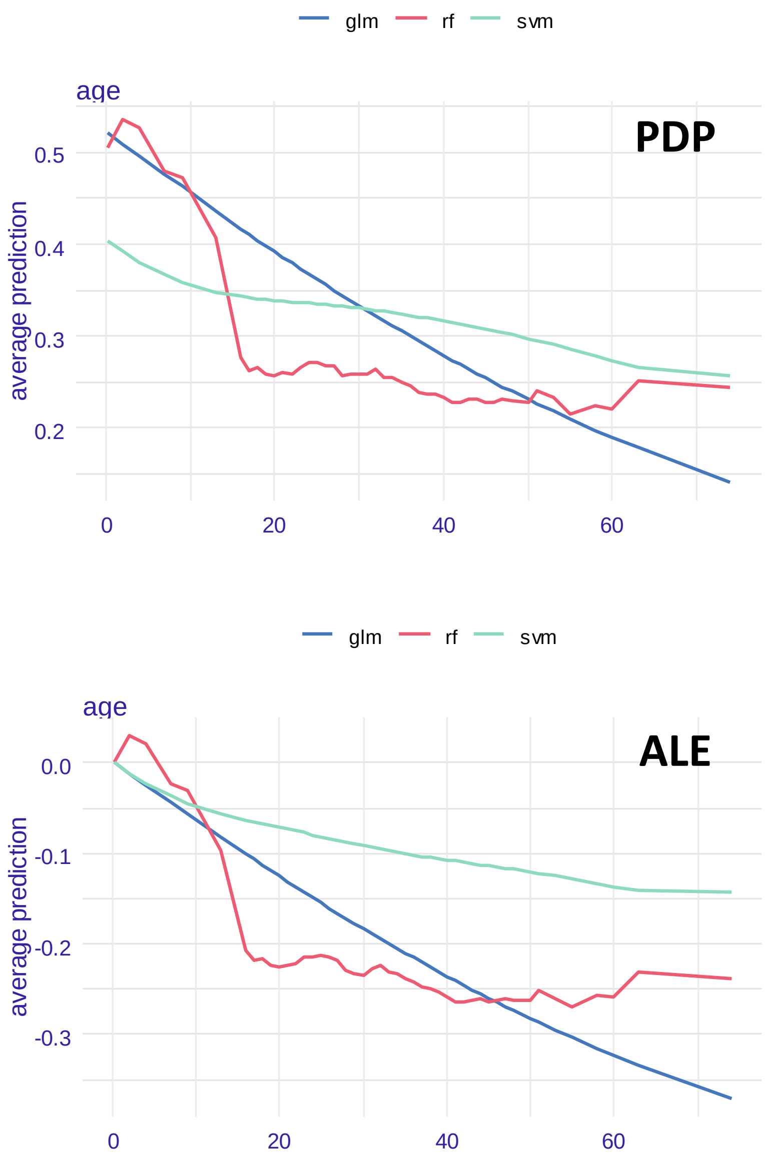
- **Idea of PDP: Compute average prediction depending on one (or several) variable(s) X_j:**

$$E_{x \setminus x_j}\left(\hat{f}\,(X_j = x)\right)$$

…in practice this is done via:

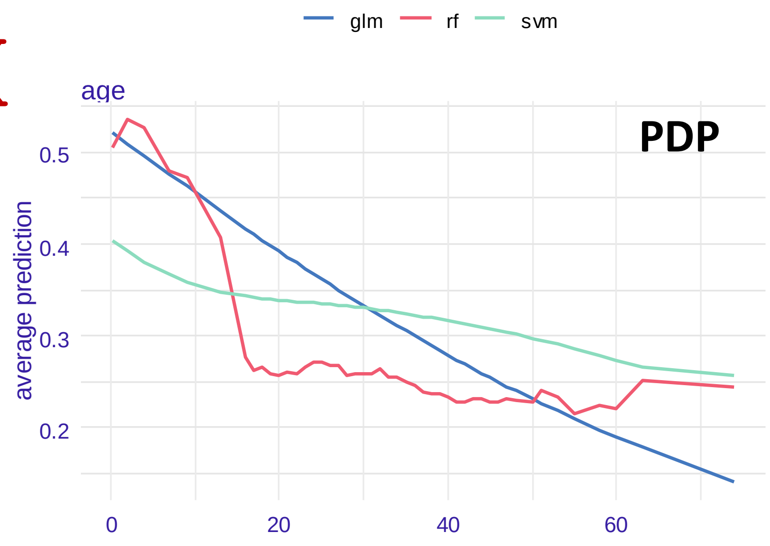
$$PD(X_j = x) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}\,(x_{i1}, \dots, x_{i(j-1)}, x, x_{i(j+1)}, \dots, x_{ip})$$

- …only partly explains the model.
- …averaging does not take into account for correlations between predictors. …ALE Plots are more appropriate (Apley, 2016).



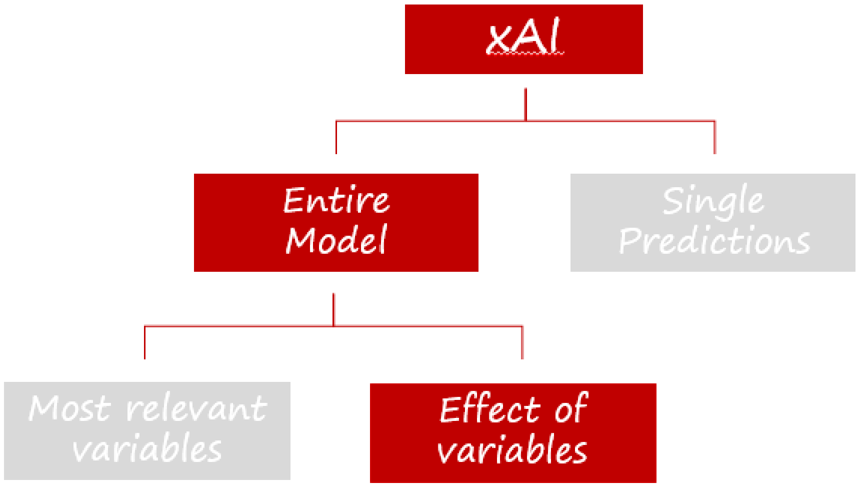
PDP and ALE Plots using DALEX

```r
# partial dependency for single variables (here: variable "age")
pdp_rf   <- partial_dependency(explainer_rf,  variables = "age", grid_points = 101)
pdp_glm  <- partial_dependency(explainer_glm, variables = "age")
pdp_svm  <- partial_dependency(explainer_svm, variables = "age")

ale_rf   <- accumulated_dependency(explainer_rf,  variables = "age")
ale_glm  <- accumulated_dependency(explainer_glm, variables = "age")
ale_svm  <- accumulated_dependency(explainer_svm, variables = "age")

plot(pdp_rf, pdp_glm, pdp_svm)
plot(ale_rf, ale_glm, ale_svm)
```



Other Packages…

Package	Variable Importance	PDP	ALE	ICE	2D	surrogate trees	merging path	LIME	local regression	Shapley Values	breakdown	live
randomForest	x											
gbm	x											
xgboost	x											
mlr	x	x			x							
caret	x											
DALEX	x	x	x				x				x	x
iml	x	x	x	x	x	x			x	x		
pdp		x		x	x							
ALEPlot		x	x		x							
ICEbox		x		x								
lime								x				
shapleyR										x		



Partial Dependency
ALE Plots

Different Requirements to Explainability

Question:

Explain…
- **a) entire model or**
- **b) single predictions?**

Understand…
- **a) which are the most relevant variables or**
- **b) effect of single variables on the output?**

xAI

- Entire Model
  - Most relevant variables
  - Effect of variables
- Single Predictions
  - Breakdown Plots
  - Shapley Values
  - LIME

Breaking Down Predictions

(Staniak and Biecek, 2018)

$x = (x_s, x_c)$ s, c: subsets of variables.

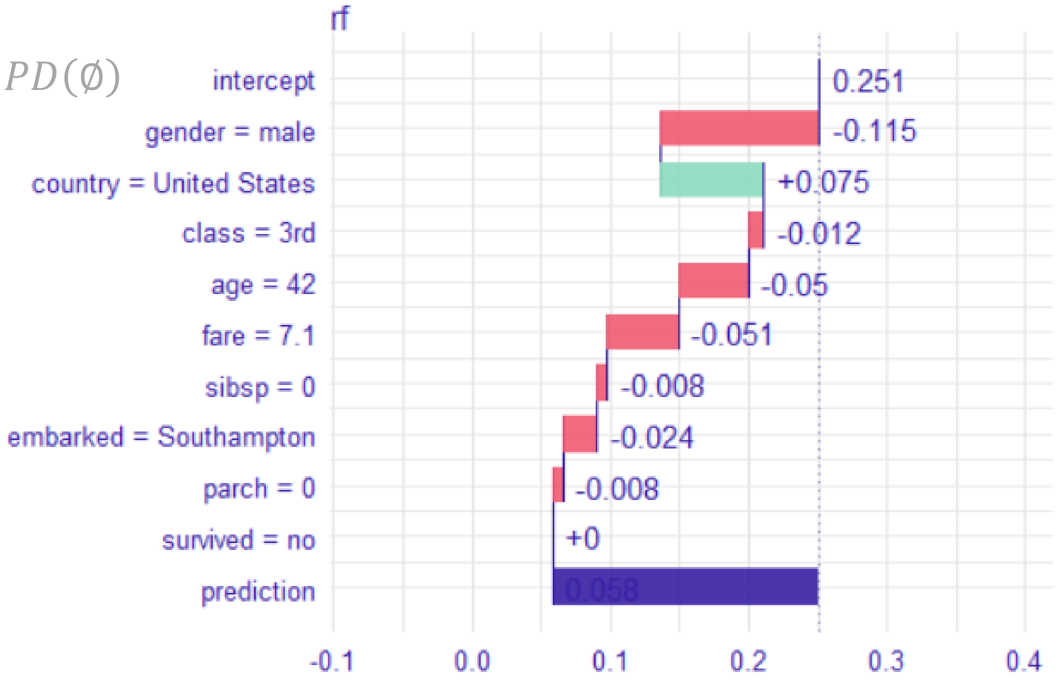
PDP for on a set of variables:

$$PD(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}\left(x_S, x_C^{(i)}\right)$$

$$\hat{f}(x) = PD(x) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}\left(x, x_{\emptyset}^{(i)}\right)$$

$E_X[\hat{f}(x)] = \beta_0 = PD(\emptyset)$

$\hat{f}(x) = PD(x)$



...Iteratively remove variable X_j such that the difference $\left|PD(x_S) - PD\left(x_{S \setminus X_j}\right)\right| \rightarrow min$

Breaking Down Predictions using DALEX

$x = (x_s, x_c)$ s, c: subsets of variables.

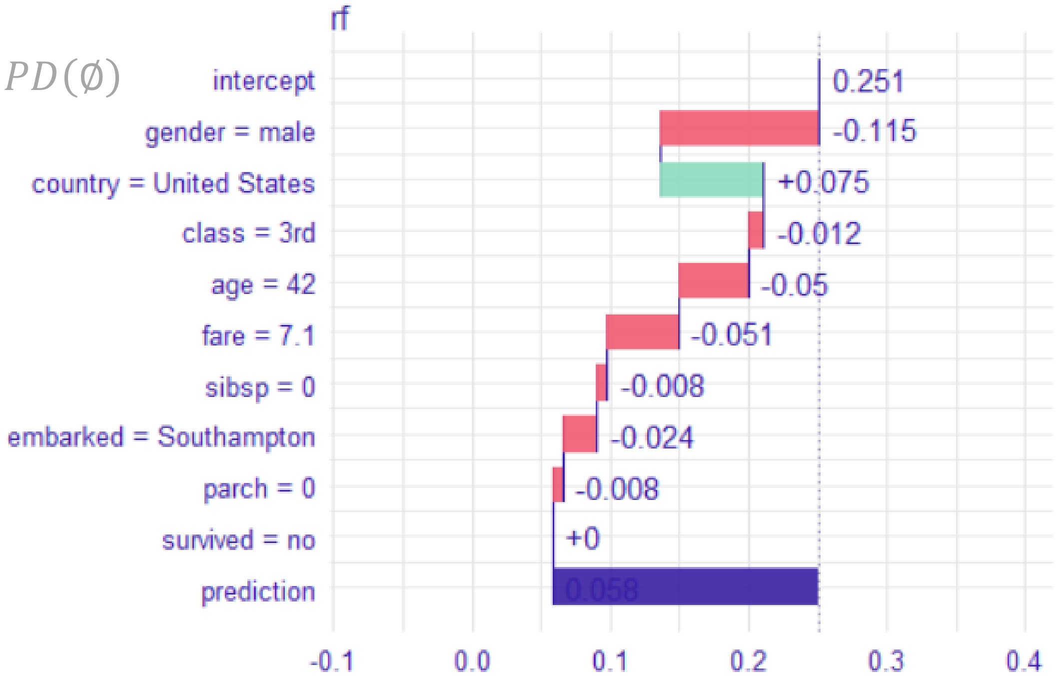
PDP for on a set of variables:

$$PD(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}\left(x_S, x_C^{(i)}\right)$$

$$\hat{f}(x) = PD(x) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}\left(x, x_{\emptyset}^{(i)}\right)$$

$E_X[\hat{f}(x)] = \beta_0 = PD(\emptyset)$

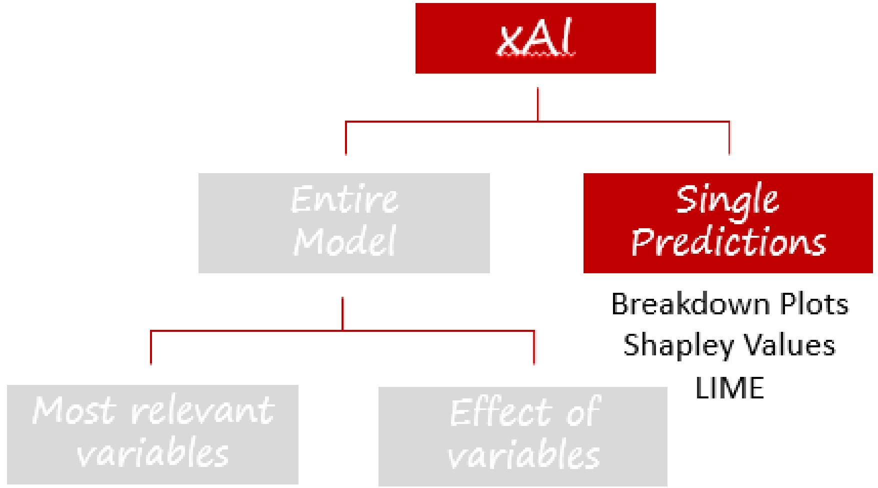
$\hat{f}(x) = PD(x)$




```r
# explaining a single prediction
new.obs <- titanic[1,]
predict(classif_rf, newdata = new.obs)

library(iBreakDown)
rf_brkdn <-  break_down(explainer_rf, new_observation = new.obs)
plot(rf_brkdn)
```

Other Packages…

Package	Variable Importance	PDP	ALE	ICE	2D	surrogate trees	merging path	LIME	local regression	Shapley Values	breakdown	live
randomForest	x											
gbm	x											
xgboost	x											
mlr	x	x			x							
caret	x											
DALEX	x	x	x				x				x	x
iml	x	x	x	x	x	x			x	x		
pdp		x		x	x							
ALEPlot		x	x		x							
ICEbox		x		x								
lime								x				
shapleyR										x		



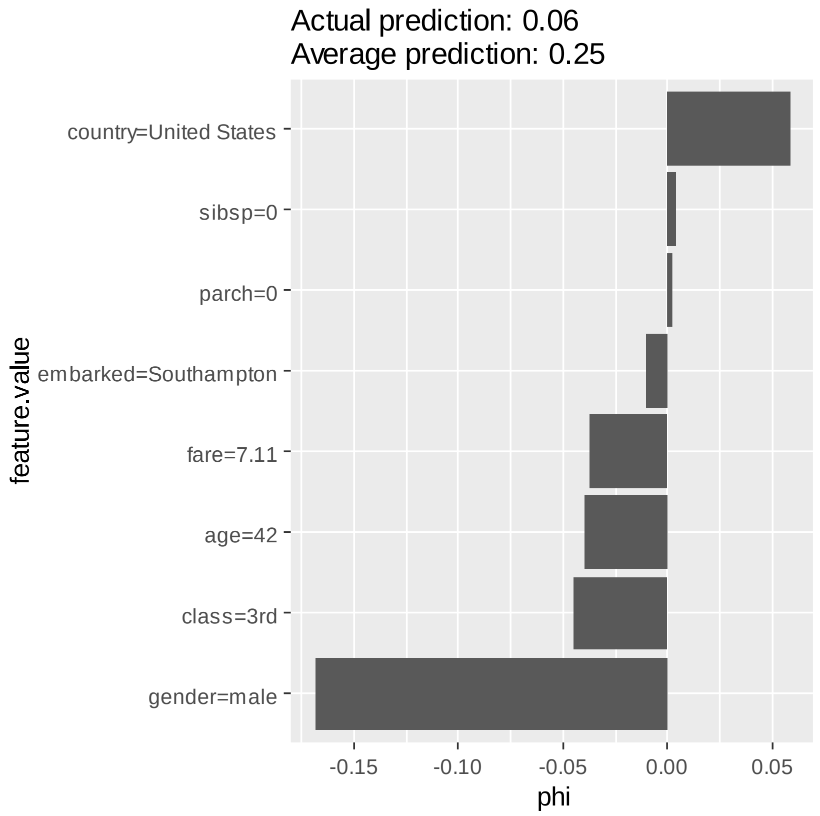
Shapley Values
(Strubelj and Kononenko, 2010)

- Conceptually similar to breakdown plots.
- Originates from game theory.

$$PD(x_S) = \sum_{i=1}^{n} \hat{f}\left(x_S, x_C^{(i)}\right)$$

$$\Delta_j(x_S) = PD\left(x_{S \cup X_j}\right) - PD(x_S)$$

...increase in explanation by variable X_j for a realization x_S of a variable set $S \ni j$.



$$SV_j(x) = \sum_{S \subseteq X \setminus X_j} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \Delta_j(x_S)$$

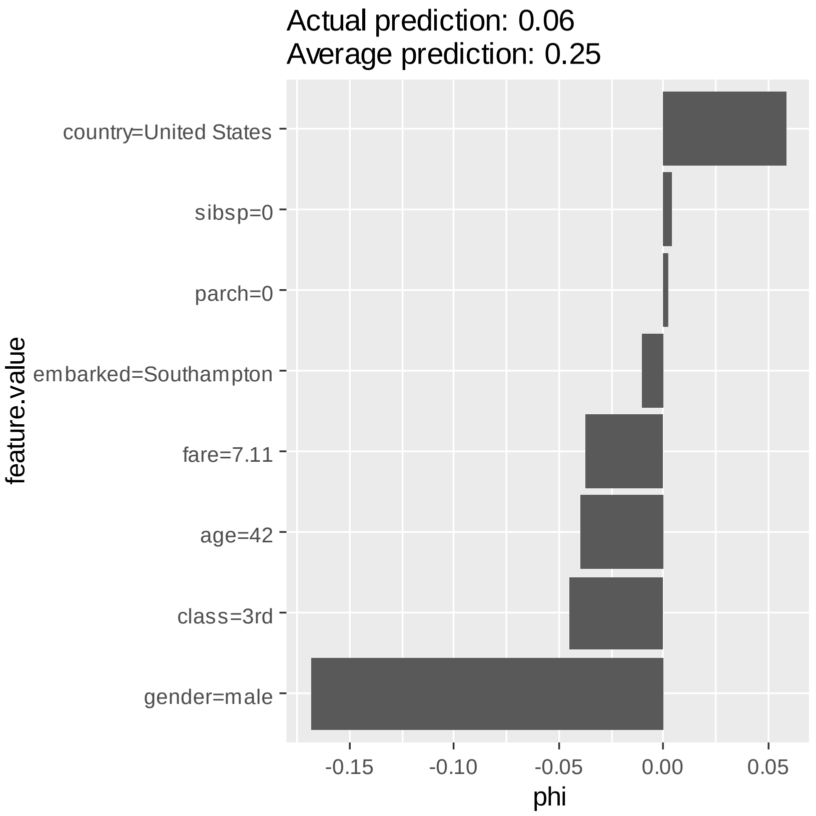
...average increase $\Delta_j(x_S)$ over all possible variable sets where variable j is added.

...with F being the set of all variables.

Shapley Values using iml

iml uses Predictor objects

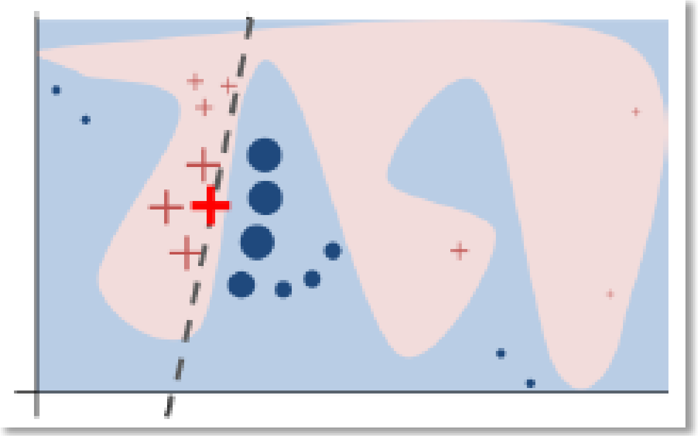
```
# shapley values
library(iml)
# create Predictor object
mod      <- Predictor$new(classif_rf, data = titanic, class = "yes")

# compute shapley values for obs. 1 based on 100 random feature subsets
shapley <- Shapley$new(mod, x.interest = titanic[1,], sample.size = 100)
shapley
plot(shapley)
```



Local Explanation (LIME)

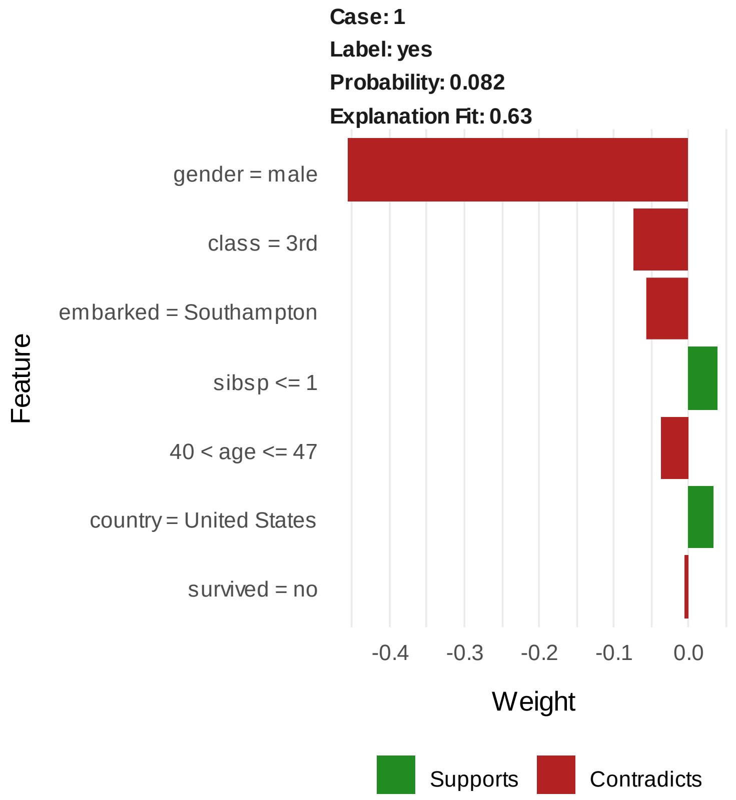
(Ribeiro, Singh and Guestrin, 2016)

1. Permute observation x n times.
2. Predict outcome for all permutations.
3. Calculate similarity (distance) of permutations to x.
4. (Discretize Variables.)
5. Select m best features (e.g. using LASSO).
6. **Fit a simple local (additive) model to the permuted data weighted by similarity.**



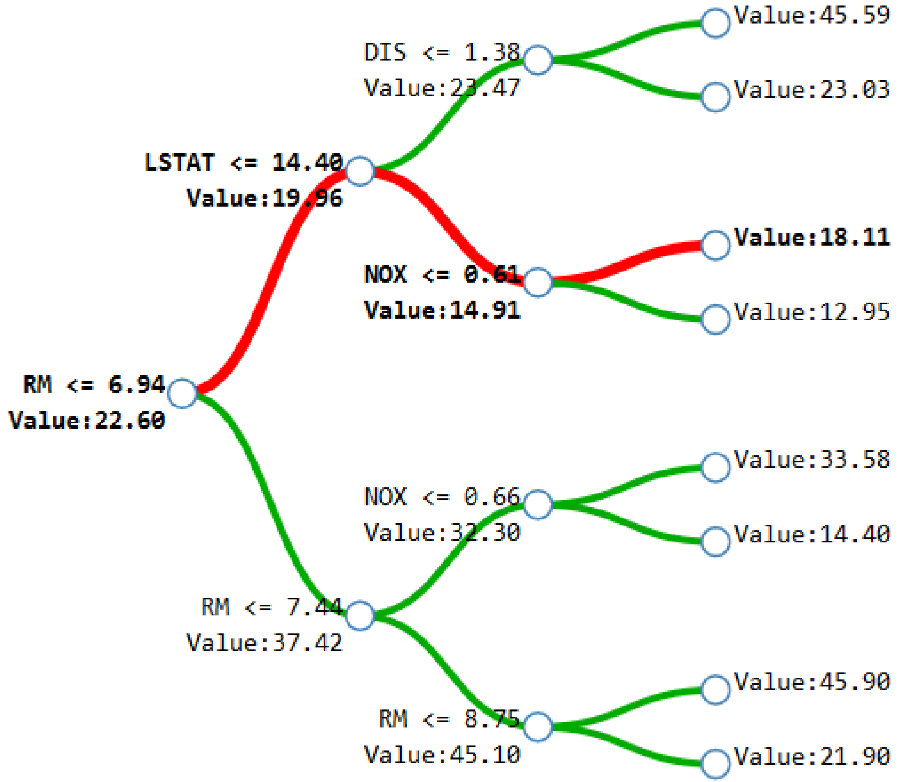
Local Explanation using lime

```
# LIME
library(mlr)
task    <- makeClassifTask(data = titanic, target = "survived", pos = "yes")
learner <- makeLearner("classif.randomForest", predict.type = "prob")
mod     <- train(learner, task)

library(lime)
explanation  <- lime(titanic, mod, n_bins = 10, quantile_bins = TRUE)
#explanation2 <- lime(titanic, mod, bin_continuous = FALSE)

explain(titanic[1,], explanation, n_features = 7, labels = "yes")
plot_features(explain(titanic[1,], explanation, n_features = 7, labels = "yes"))
```



Note on Explaining Tree Ensembles

(Saabas, 2014)



Decomposition of a single tree...

$$\hat{f}(x) = 18.11 = 22.6\ (\bar{y}) - 2.64(\text{RM}) - 5.04\ (\text{LSTAT}) + 3.2\ (\text{NOX})$$

$$= y_{root} + \sum_j contrib(x, j)$$

...of a random forest:

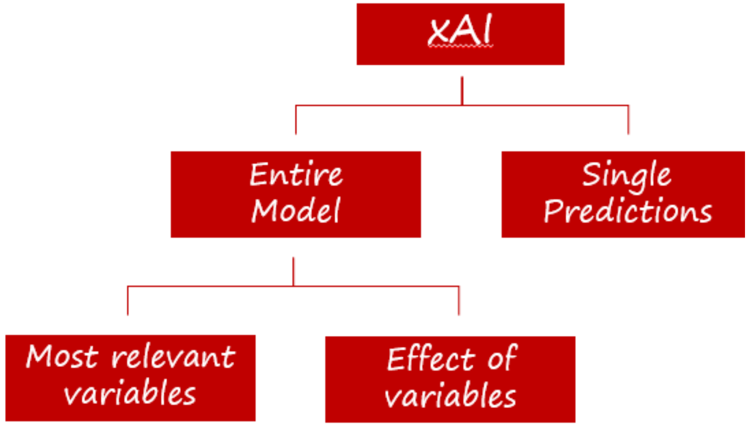
$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} y_{root,b} + \sum_j \frac{1}{B}\sum_{b=1}^{B} contrib_b(x, j)$$

...for xgboost this is offered by argument `predcontrib = TRUE` **of the** `predict.xgb.Booster` **or the package** `xgboostExplainer` **(Foster, 2018).**

Takeaways

- Different Requirements on ‘explainability’.
- A lot of packages providing different functionalities.
- Most popular methods
  - Variable importance
  - PDP and ALE Plots
  - LIME and Shapley values
- Two general frameworks: `DALEX` and `iml`.

- Open issues: Which method to chose? Are available approaches sufficient? There is an obvious need for explanation. *...there is still work to be done !*



Interpretable Machine Learning

References

- Apley, D. (2016): Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models, arXiv:1612.08468. (**ALE plots**)
- Biecek, P. (2018): DALEX: Explainers for Complex Predictive Models in R, JMLR 19, 1-5. (**DALEX**)
- Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G. and Jones Z. (2016): mlr – Machine Learning in R, JMLR 17, 1−5. (**mlr**)
- Breiman, L. (2001): Random Forests, Machine Learning 45(1), 5 – 32. (**Variable Importance**)
- Foster, D. (2018): xgboostExplainer. https://github.com/AppliedDataSciencePartners/xgboostExplainer.
- Friedman, J. (2001): Greedy function approximation: A gradient boosting machine.The Annals of Statistics,29:1189–1232. (**PDP**)
- Molnar, C., Casalicchio, G. and Bischl, B. (2018): iml: An R package for Interpretable Machine Learning. J. Open Source Software 3 (26): 786. (**iml**)
- Ribeiro, M. , Singh, S. and Guestrin, C. (2016): Why Should I Trust You?. ACMPress, 2016. doi: 10.1145/2939672.2939778. URL: https://arxiv.org/abs/1602.04938. (**LIME**)
- Saabas, A. (2014): Interpreting Random Forests. https://blog.datadive.net/interpreting-random-forests/,
- Staniak, M. and Biecek, P. (2018): Explanations of Model Predictions with live and breakDown Packages, R Journal 10/2, 395-409.
- Strubelj, E. and Kononenko, I. (2010): An Efficient Explanation of Individual Classifications using Game Theory, JMLR 11, 1-8. (**Shapley Values**)

Thank You