The breakDown package explains predictions from black-box models, such as random forest, xgboost, svm or neural networks (it works for lm and glm as well). As a result you gets decomposition of model prediction that can be attributed to particular variables.

The version 0.3 has a new function break\_down. It identifies pairwise interactions of variables. So if the model is not additive, then instead of seeing effects of single variables you will see effects for interactions.  
It’s easy to use this function. See an example below.  
HR is an artificial dataset. The break\_down function correctly identifies interaction between gender and age. Find more examples in the documentation.

#

# Create a model for classification

library("DALEX")

library("randomForest")

library(“breakdown”)

model <- randomForest(status ~ . , data = HR)

#

# Create a DALEX explainer

explainer\_rf\_fired <- explain(model,

data = HR, y = HR$status == "fired",

predict\_function = function(m,x) predict(m,x, type = "prob")[,1])

#

# Calculate variable attributions

new\_observation <- HRTest[1,]

library("breakDown")

bd\_rf <- break\_down(explainer\_rf\_fired,

new\_observation,

keep\_distributions = TRUE)

bd\_rf

#> contribution

#> (Intercept) 0.386

#> \* hours = 42 0.231

#> \* salary = 2 -0.216

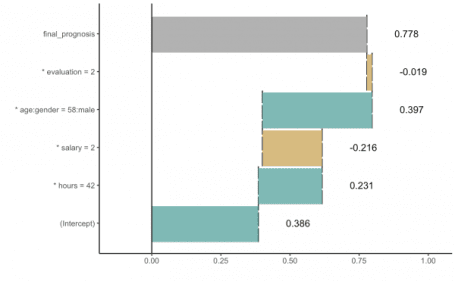
#> \* age:gender = 58:male 0.397

#> \* evaluation = 2 -0.019

#> final\_prognosis 0.778

#> baseline: 0

plot(bd\_rf)

Figure below shows that a single prediction was decomposed into 4 parts. One of them is related to the interaction between age and gender.  


BreakDown – collection of tools for visualisation, exploration and explanation of complex machine learning models.

Till the end of September I am visiting UC Davis and UC Berkeley. Happy to talk about DALEX explainers, XAI and related stuff.

