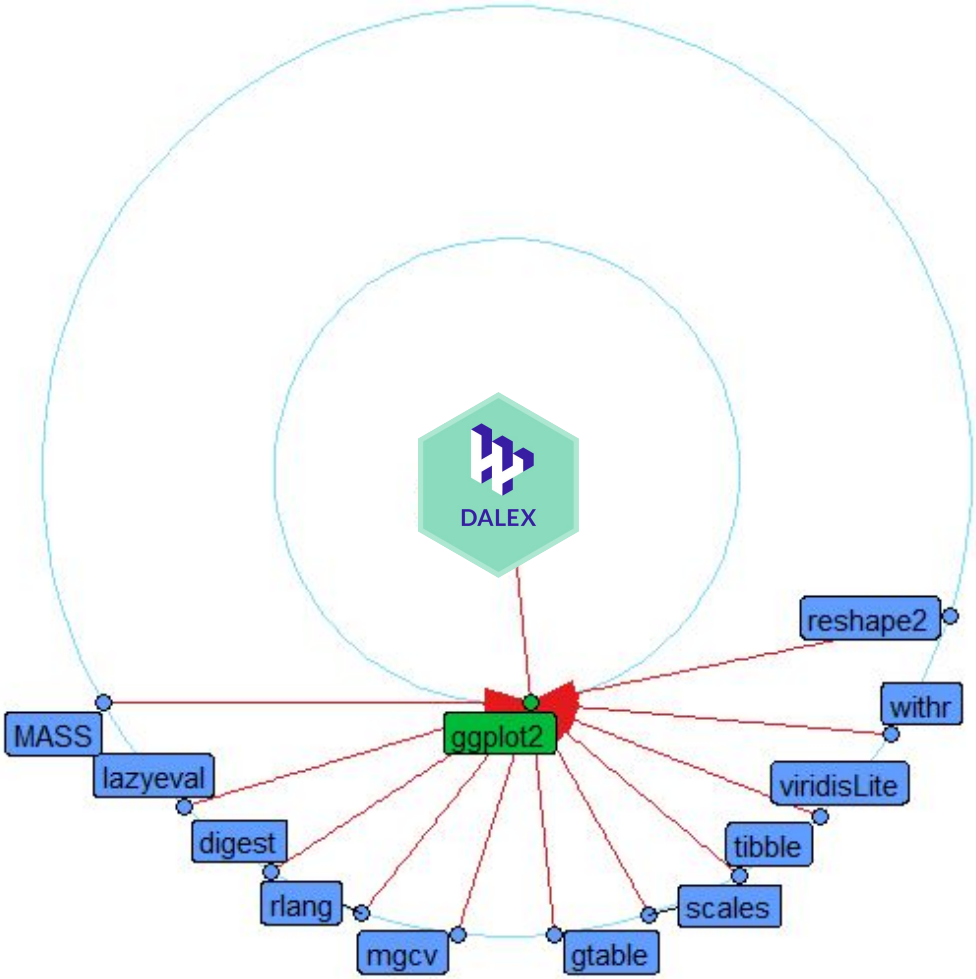
Descriptive mAchine Learning EXplanations

Tomasz Klonecki
Aleksandra Łuczak

Opis techniczny paczki

- **Autorzy:** MI2DataLab
- **Start:** 2018 rok
- **Motywacja:** Kontrola i wyjaśnienie modeli które mogą wpływać na nasze życie



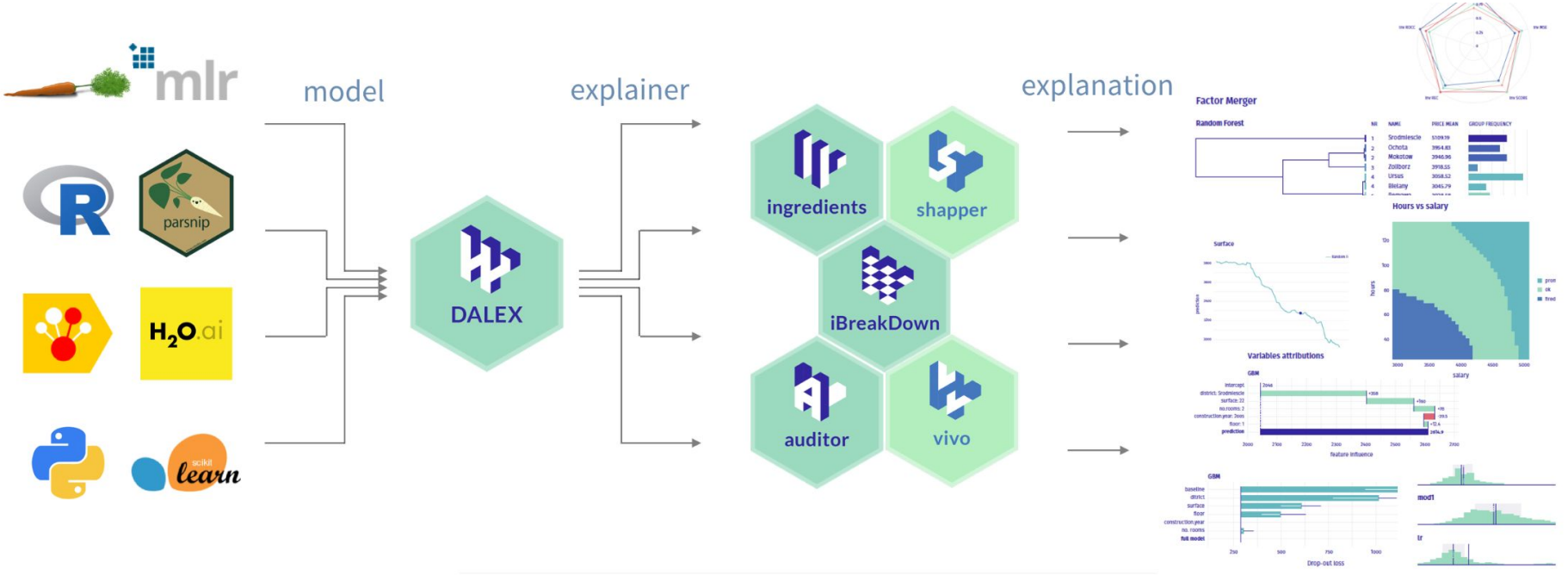
Możliwości

Ogólne:
- Wyjaśnianie modeli z różnych frameworków i JĘZYKÓW
- Wykorzystanie różnych bibliotek do wyjaśniania na wraperach DALEX

Model understanding:
- Residuals
- Feature importance
  - Model agnostic
  - Model specific
- Variable response
  - PD plots
  - ALE plots

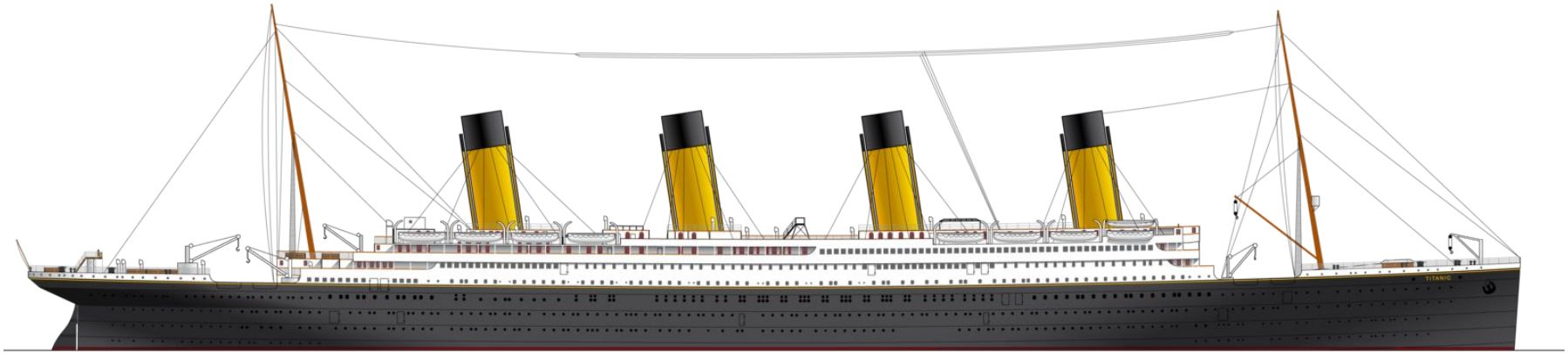
Prediction understanding:
- Outlier detection
- Single observation analysis:
  - Ceteris Paribus plot
  - BreakDown

Możliwości


mlr
model
explainer
explanation
parsnip
H2O.ai
scikit learn
DALEX
ingredients
shapper
iBreakDown
auditor
vivo
Factor Merger
Random Forest
Hours vs salary
Surface
Variables attributions
GBM
Drop-out loss

Use Case

Najlepiej zobaczyć jak paczka działa na konkretnym przykładzie. W prezentacji wykorzystamy popularny zbiór danych Titanic, gdzie zmienną celu jest przeżycie katastrofy.



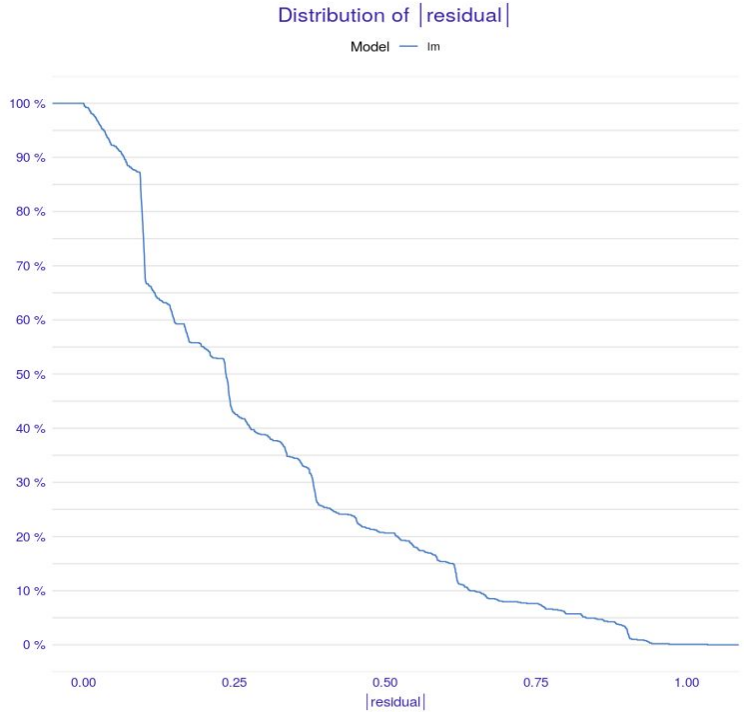
Stworzenie wrapera DALEX

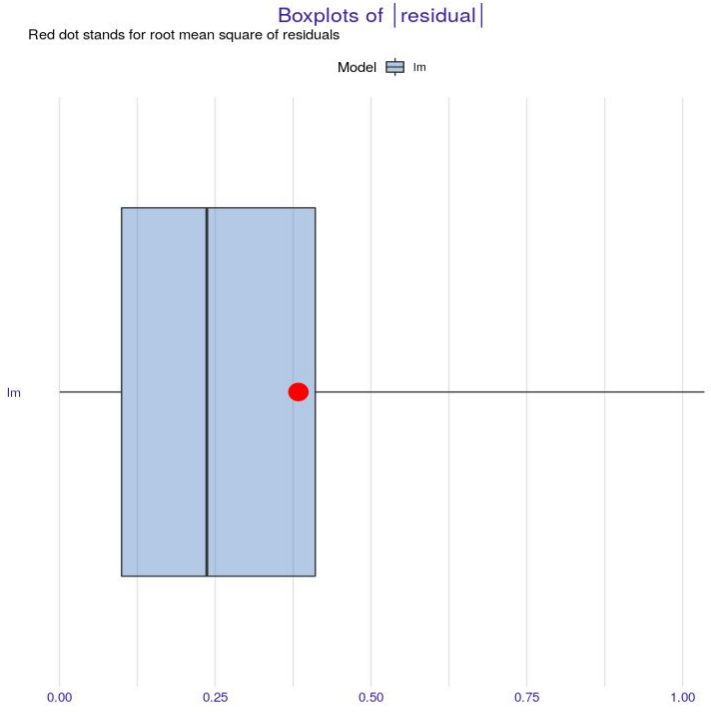
```
titanic_lm_model <-
        lm(Survived ~ . , data = train_data)

train_data$Survived<- NULL

explainer_lm <-
        explain(titanic_lm_model, data = train_data, y = y_train)
```

Model Understanding - Residuals

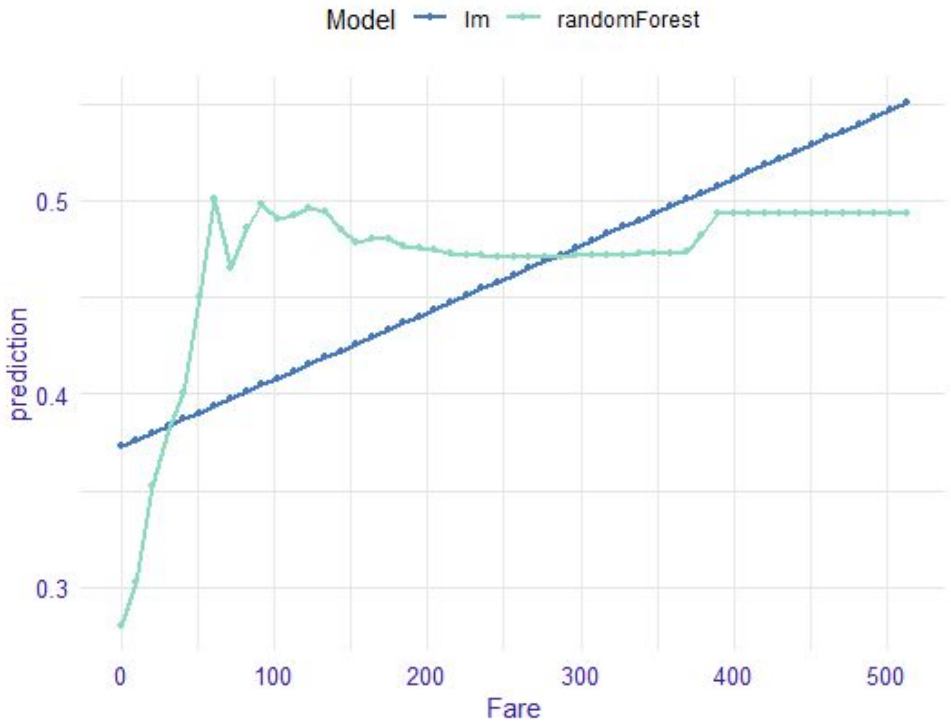


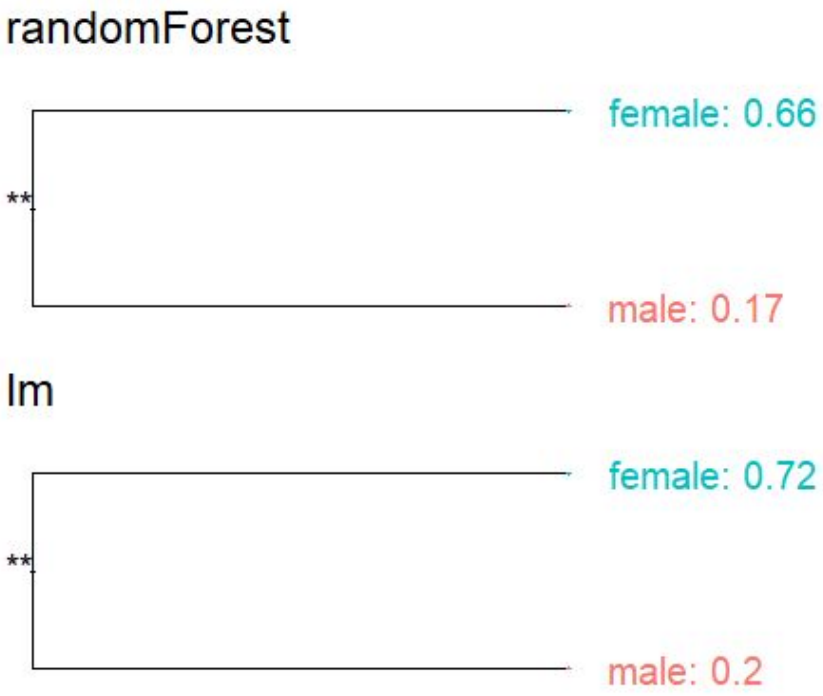



```
mp_lm <- model_performance(explainer_lm)
plot(mp_lm)
```

```
plot(mp_lm, geom = "boxplot")
```

Model Understanding - PD plots

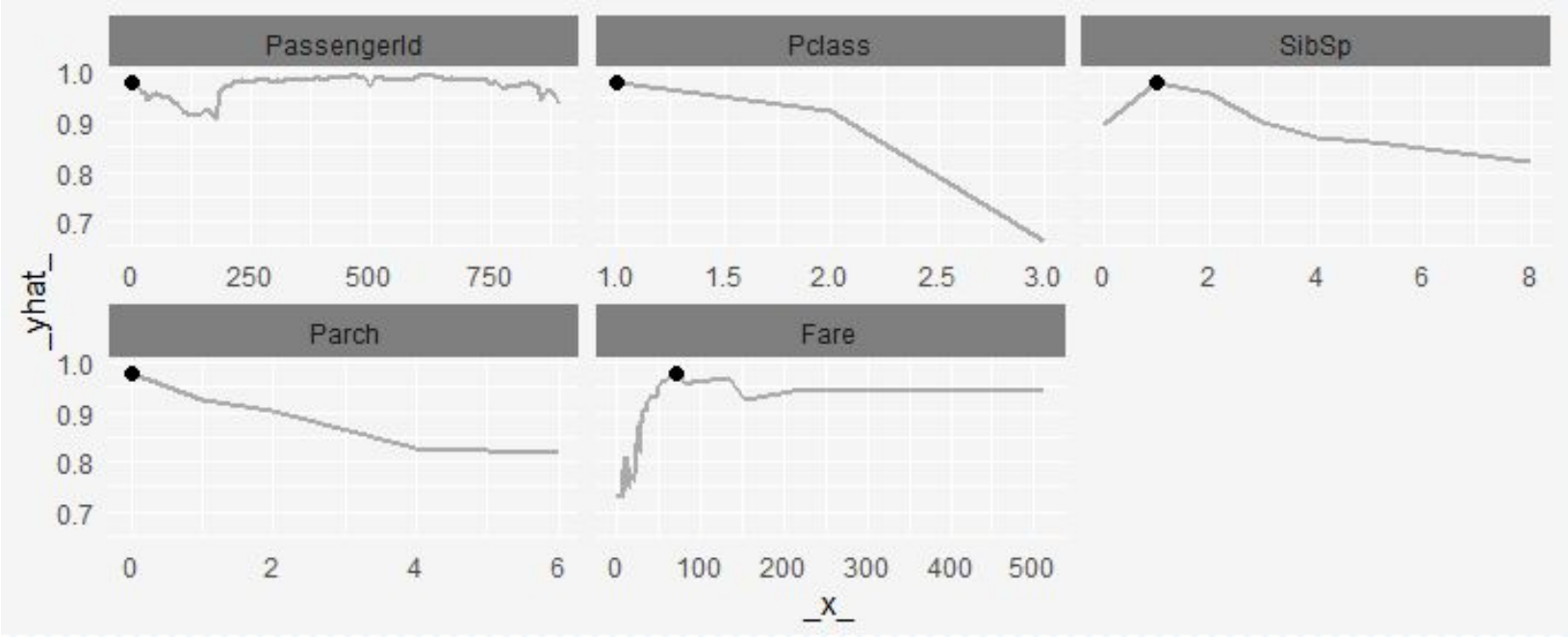




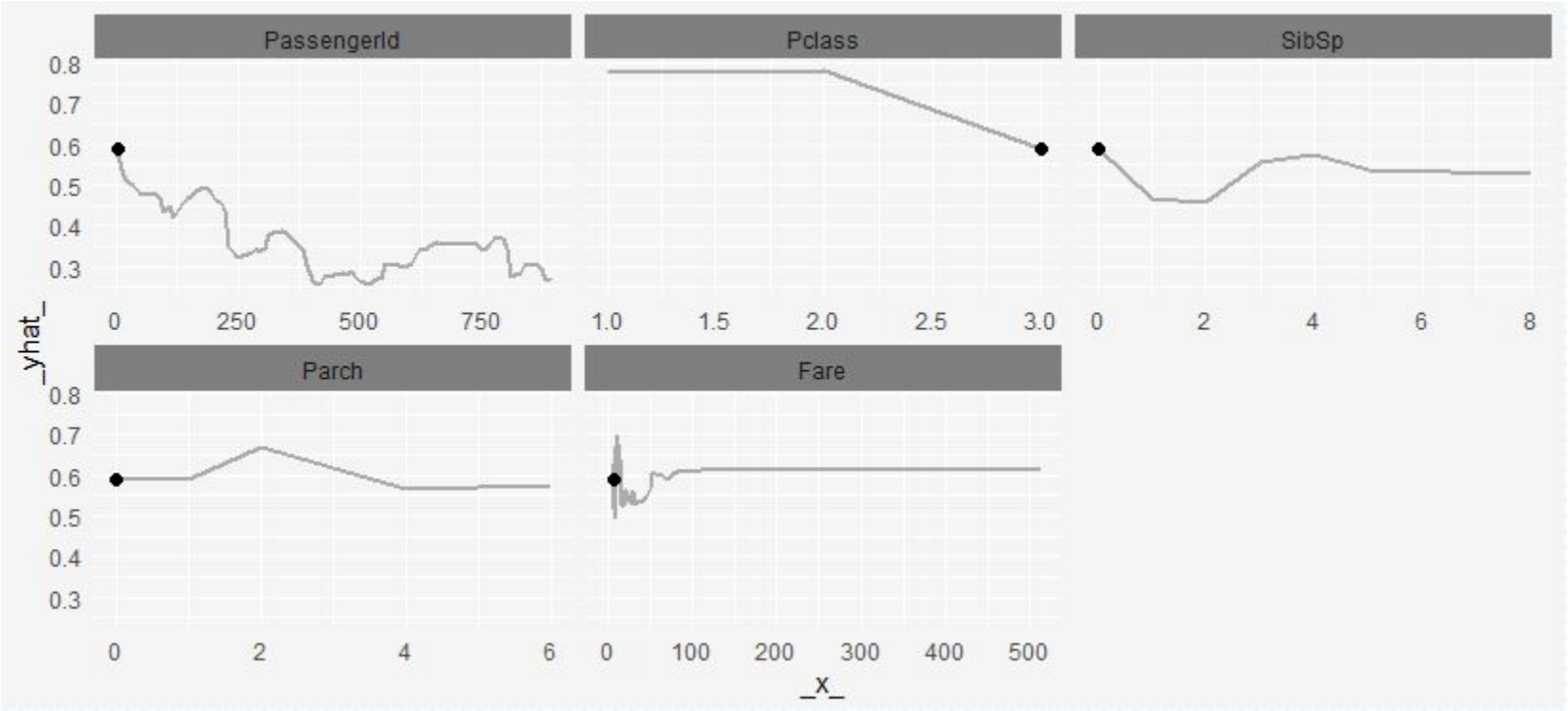
Prediction understanding

Pasażer	Klasa	Płeć	Ilość rodzeństwa	Ilość opiekunów	Cena za bilet
Biedny	3	kobieta	0	0	71.28
Bogaty	1	kobieta	1	0	7.92

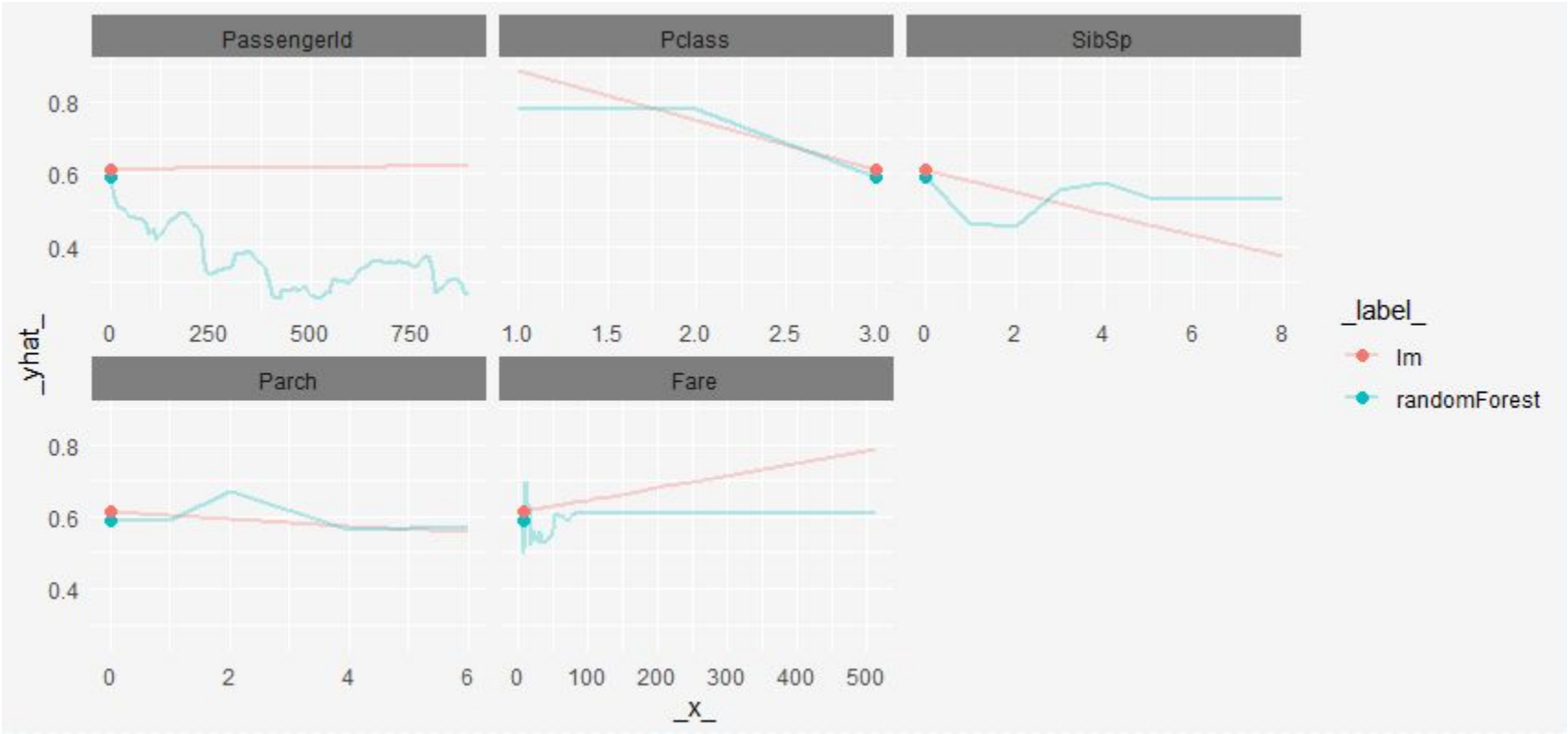
Prediction understanding - Ceteris Paribus - bogaty



Prediction understanding - Ceteris Paribus - biedny



Prediction understanding - bogaty - porónanie modeli



DALEX CHEAT SHEET

The DALEX package (Descriptive mAchine Learning EXplanations) helps to understand how complex models are working.

Main wrapper

explain(model, data, y, predict_function, residual_function)

Function turns models into explainers - wrappers with uniform structure. Then we can use various functions to turn explainers to explanations.

model
Object - a model to be explained

data
Data.frame or matrix - data that was used for fitting.

y
Numeric vector with outputs. If provided then it shall have the same size as data.

predict_function
Function that takes two arguments: model and new data and returns numeric vector with predictions.

residual_function
Function that takes three arguments: model, data and response vector y. It should return a numeric vector.

Model understanding

model_performance(explainer)

Prepare a data frame with model residuals.

explainer
Object - a model to be explained, preprocessed by the explain function.

variable_importance(explainer, loss_function)

Calculate model agnostic variable importance.

loss_function
Function that will be used to assess variable importance.

single_variable(explainer, variable, type)

Calculates the average model response as a function of a single selected variable.

variable
character - name of a single variable

type
'pdp' for Partial Dependency and 'ale' for Accumulated Local Effects

Prediction analysis

prediction_breakdown(explainer, observation)

Calculate Break Down Explanations.

observation
A a new observarvation for which predictions need to be explained

ceteris_paribus(explainer, observations)

This function calculate ceteris paribus profiles for selected data points.

observations
set of observarvation for which profiles are to be calculated

Install the package with install_packages('DALEX')

Dziękujemy za uwagę