

# Wyjaśnianie modeli ML

Mateusz Staniak  
STWUR 29.05.2018

# Rozkład jazdy

- 1) Zarządzanie modelami (archivist)**
- 2) Wyjaśnianie modeli:**
  - globalne wyjaśnienia (**DALEX**)
  - lokalne wyjaśnienia (**DALEX + live**)
- 3) Podsumowanie i różności**
- 4) Zadania**

# Archivist

DOI 10.5281/zenodo.47154 CRAN 2.3.1 downloads 1402/month downloads 52K  
repo status Active build passing pending pull-requests 0 open issues 1 coverage unknown

## A set of tools for datasets and plots archiving

Everything that exists in R is an object. `archivist` is an R package that stores copies of all objects along with their metadata. It helps to manage and recreate objects with final or partial results from data analysis.

Use the `archivist` to record every result, to share these results with future you or with others, to search through repository of objects created in the past but needed now.

### Installation

To get started, install the latest version of `archivist` from CRAN:

```
install.packages("archivist")
```

or from GitHub:

```
devtools::install_github("pbiecek/archivist")
```

### Cheatsheet

#### archivist : record, restore and govern your R objects

Everything that exists in R is an object. `archivist` is an R package that stores copies of all objects along with their metadata. It helps to manage and recreate objects with final or partial results from data analysis.

Use the `archivist` to record every result, to share these results with future you or with others, to search through repository of objects created in the past but needed now.

Key functionalities include:  
i) management of local and remote repositories which contain R objects and their meta-data (objects' properties and relations between them).

#### Record artifacts

R objects with partial or final results are called artifacts. They may be either tables, models, plots or any other structures. Artifacts are stored in repositories. One repository may be either local or remote.

- local repository is a folder with write/read access,

- remote repositories are usually available through http and have only read access.

Use the `createLocalRepo()` function to create an empty local repository.

```
library(archivist)
```

```
createLocalRepo("arepo")
```

Use the `saveToRepo()` function to store selected R objects in a local repository. Each object is stored

#### Share artifacts

Use `loadFromLocalRepo()` or `loadFromRemoteRepo()` or more compact `aread()` functions to retrieve artifacts from repositories. It's a good idea to attach hooks to key artifacts in reports or articles. The command below restores a ggplot2 object from GitHub pbiecek/Eseje.

```
aread("pbiecek/Eseje/arepo/65e430c41")
```



The code below downloads and calculates BIC scores for all artifacts with tag `class:lm` (linear models) from the remote GitHub repository `pbiecek/graphGallery`.

```
models <- asearch("pbiecek/graphGallery",
```

```
    patterns = "class:lm")
```

```
models$BIC <- sapply(models, BIC)
```

```
sort(models$BIC)
```



# Dlaczego?

- 1) Reprodukowalność**
- 2) Dzielenie się modelami**
- 3) Zarządzanie modelami**
- 4) Modele zmieniają się w czasie**

# **Archivist: prezentacja pakietu**

# Ćwiczenia

# **Wyjaśnianie modeli ML**

# DALEX

## DALEX

[CRAN 0.2.2](#) [downloads 2014](#) [build passing](#) [coverage 92%](#)

Descriptive mAchine Learning EXplanations

### DALEX Stories

- [A gentle Introduction to DALEX with examples](#)
- [How to use DALEX with caret](#)
- [How to use DALEX with mlr](#)
- [How to use DALEX with xgboost package](#)
- [Talk about DALEX at Complexity Institute / NTU February 2018](#)
- [Talk about DALEX at SER / WTU April 2018](#)
- [How to use DALEX for teaching. Part 1](#)

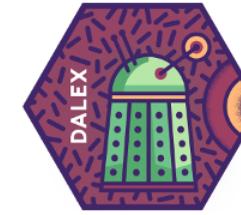
### Install

From GitHub

```
# dependencies  
devtools::install_github("MI2DataLab/factorMerger")  
devtools::install_github("pbiecek/breakDown")  
  
# DALEX package  
devtools::install_github("pbiecek/DALEX")
```

or from CRAN

```
install.packages("DALEX")
```



# Dlaczego?

IBM Watson Health   Life sciences   **Oncology**   Value-based care   Government   Imaging   Blog

Oncology and Genomics

## Bringing confident decision-making to oncology

Provide evidence-backed cancer care to each patient, by understanding millions of data points

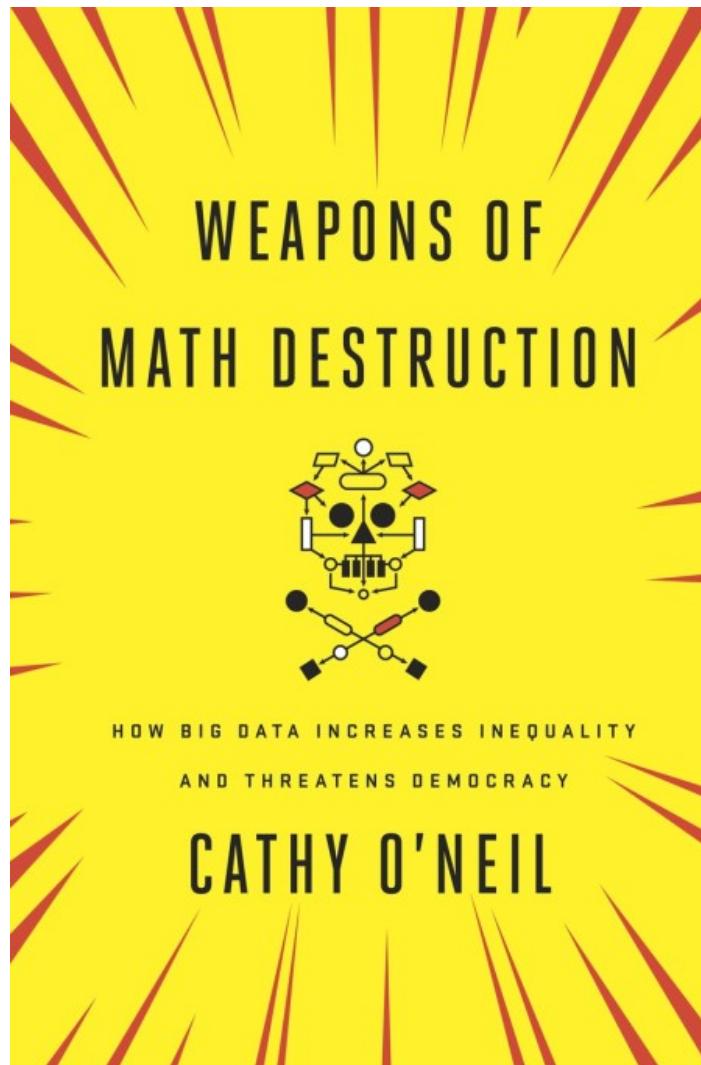
[Meet at an upcoming event](#)   [Understand the data](#)



Put Watson for Oncology to the test with your tumor board today

Watson for Oncology: 90% concordance with tumor board recommendation

# Dlaczego?



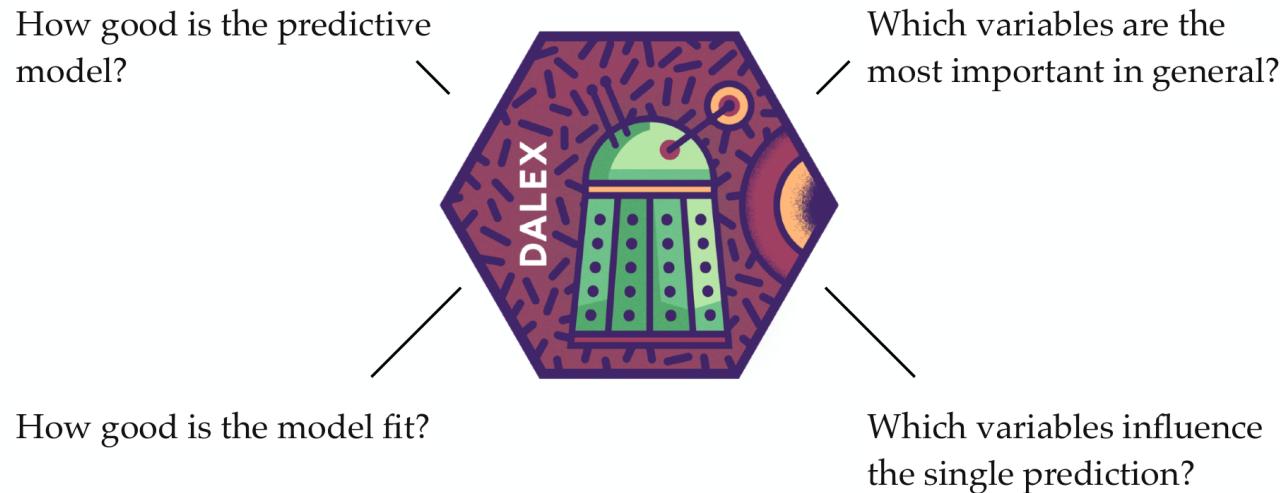
# Dlaczego?

The screenshot shows the homepage of the TrackML Particle Tracking Challenge. At the top, there's a banner for a 'Featured Prediction Competition' with a \$25,000 prize. Below the banner, the challenge is described as 'High Energy Physics particle tracking in CERN detectors'. It mentions 'CERN · 315 teams · 3 months to go (2 months to go until merger deadline)'. The main navigation bar includes links for Overview, Data, Kernels, Discussion, Leaderboard (which is underlined), and Rules. Below the navigation, there are two tabs: Public Leaderboard (selected) and Private Leaderboard. A note states that the leaderboard is based on 29% of test data and that final results will be based on 71%. There are download and refresh buttons. The legend indicates four categories: In the money (green), Gold (orange), Silver (grey), and Bronze (brown). The main content is a table showing the top 10 teams:

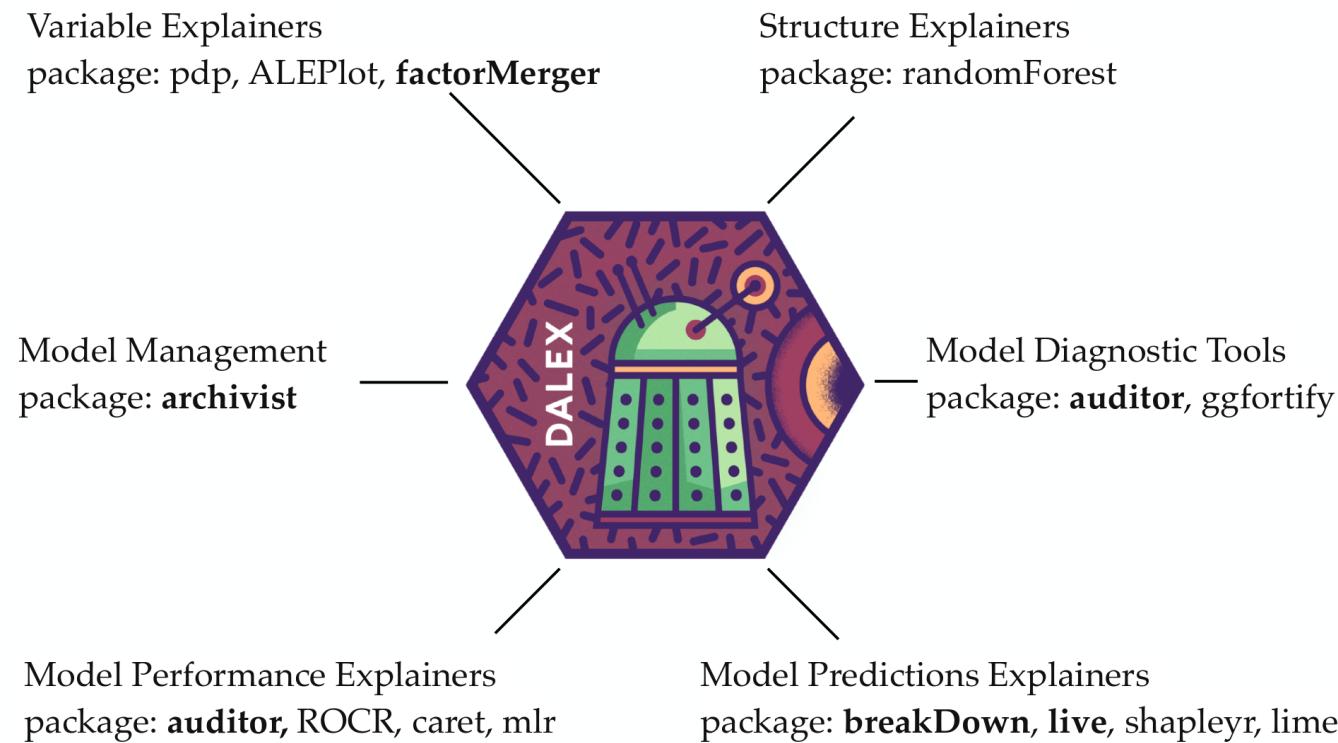
#	△1w	Team Name	Kernel	Team Members	Score	Entries	Last
1	—	Mickey			0.6405	7	9d
2	▲ 257	Zidmie			0.5829	5	3d
3	▼ 1	Vicens Gaitan			0.4969	2	6d
4	▲ 1	Lin12345			0.4775	16	20h
5	▲ 5	HomerJSimpson			0.4306	11	11h
6	▼ 3	Grzegorz Sionkowski			0.4225	12	1d
7	▲ 28	Andrea			0.4219	4	2d
8	new	Vivek			0.4207	4	1d
9	new	dohlee			0.4107	1	1d
10	▲ 129	all_random			0.4063	13	18h

# Czym jest DALEX?

DALEX is a set of tools that helps to understand the way complex predictive models work



# Czym jest DALEX?



# Wpływ pojedynczej zmiennej: idea

- 1) Zaniedbać wpływ części zmiennych**
- 2) Obliczyć predykcje dla różnych wartości ustalonej zmiennej / zmiennych**
- 3) Narysować zależność predykcji od zmiennej**

$$\phi_j(x) = \frac{1}{N} \sum_{k=1}^N F(x_{1,k}, \dots, x_{j-1,k}, x, x_{j+1,k}, \dots, x_{p,k}).$$

# Wpływ pojedynczej zmiennej

**1) PDP dla pojedynczej obserwacji:**

→ ICEbox (Individual Conditional Expectation)

→ pomaga w wykrywaniu interakcji

**2) Rozkład warunkowy zamiast  
brzegowego:**

→ ALEPlot (Accumulated Local Effect)

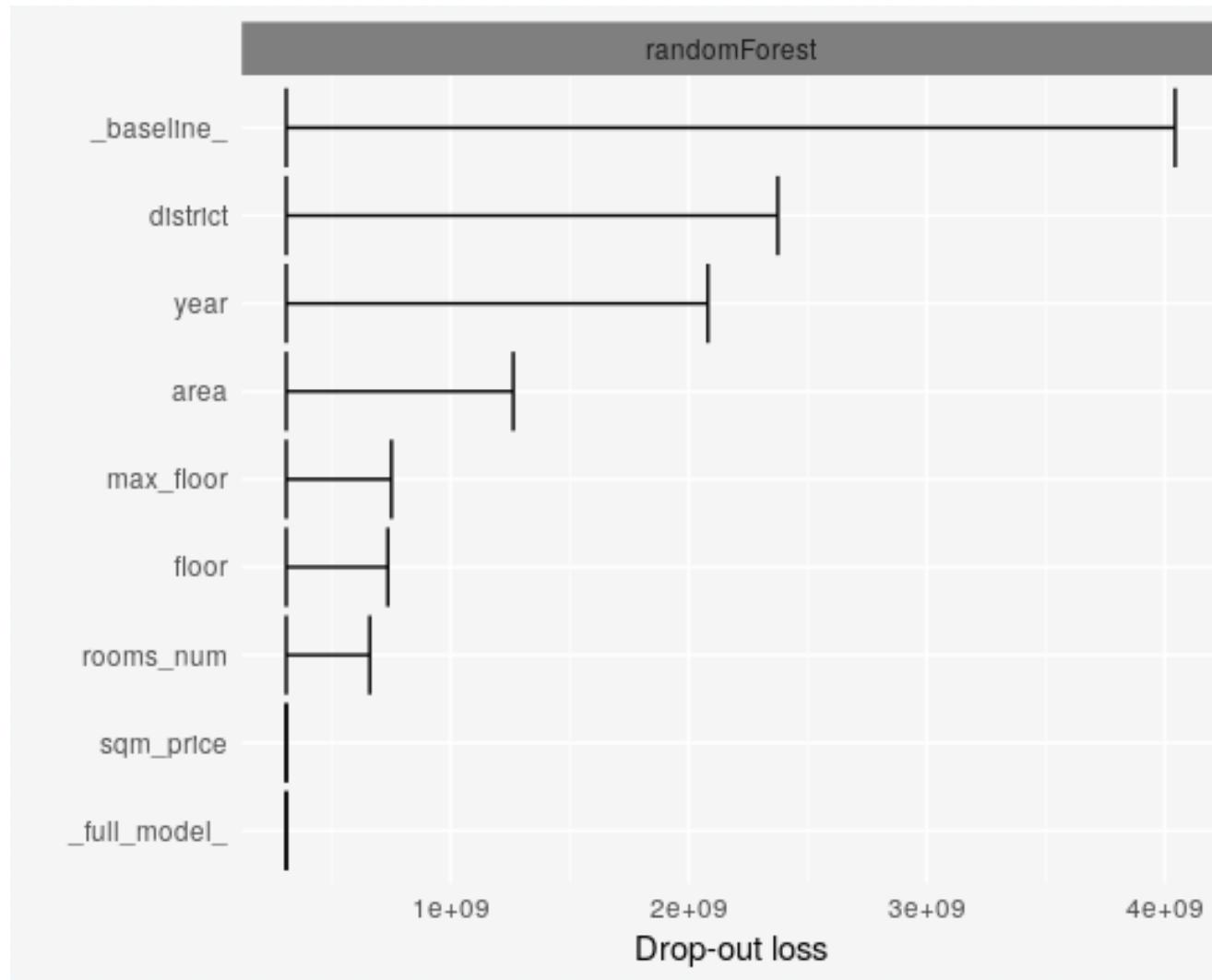
# **Funkcja single\_variable**

# Ćwiczenia

# Ważność zmiennych: idea

- 1) Model oceniamy na podstawie funkcji straty**
- 2) Zakładamy, że jeśli zmienna jest ważna, jest usunięcie z modelu spowoduje duży wzrost funkcji straty**
- 3) „Usunięcie” zmiennej jest realizowane poprzez jej losową permutację**

# Ważność zmiennych



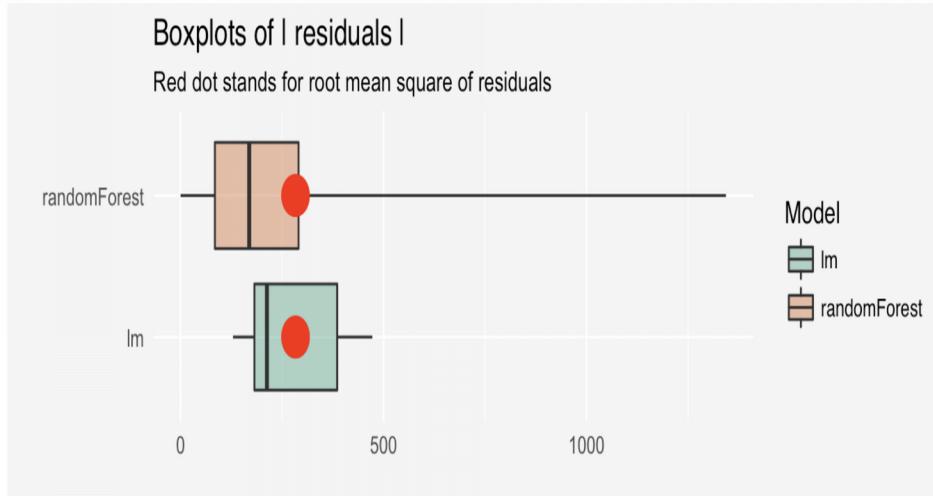
# Funkcja variable\_dropout

# Ćwiczenia

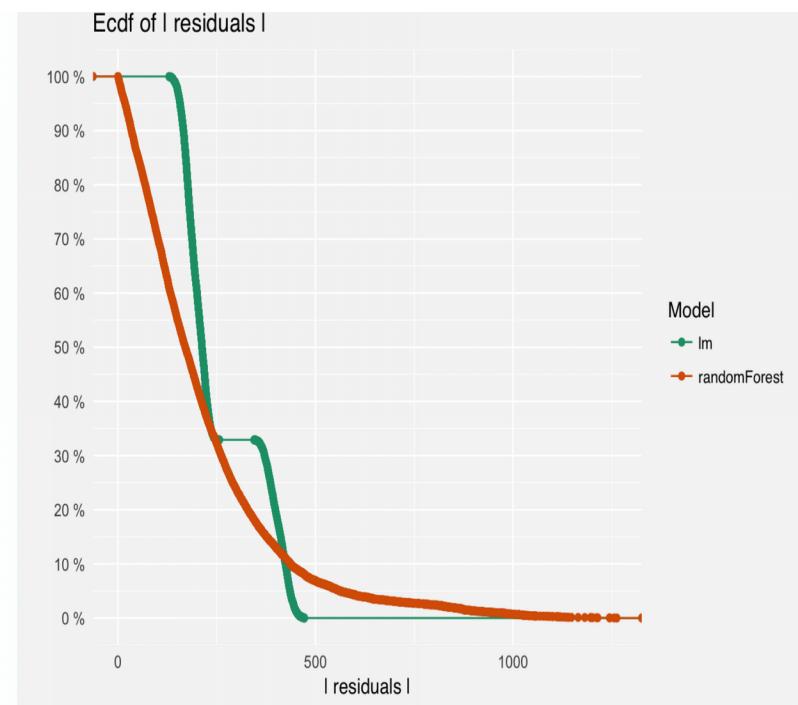
# Jakość modelu: idea

- 1) Miara jakości modelu: np. MSE, AUC**
- 2) Pojedyncza liczba to za mało!**
- 3) Więcej informacji uzyskamy, patrząc na rozkład reszt**

# Jakość modelu



46



# **Funkcja model\_performance**

# Ćwiczenia

# Diagnostyka modelu

## 1) Analiza reszt

→ wielkość

→ autokorelacja

→ rozkład

## 2) Analiza predykcji

→ zaobserwowane vs przewidziane

## 3) Więcej: obserwacje odstające, porównanie modeli ...

# Diagnostyka modelu

## ROC plots with auditor :: CHEAT SHEET



### Basics

Package **auditor** provides several methods for model verification and validation.

This includes both, graphical methods and scores.

In this cheatsheet, we present ROC curves and their extensions for regression problem.

ROC analysis is a very popular tool for the assessment of classifier performance. So far there have been several approaches to adapt ROC curves to regression.

In the auditor package, there is a possibility to use two approaches: Regression Receiver Operating Characteristic (RROC) and Regression Error Characteristic (REC).

### MODEL PREPARATION

We will show the use of a package for a logistic regression model.

The example uses the *Pima Indian Diabetes* dataset.

```
library(mlbench)
data("PimaIndiansDiabetes")
mod_glm <- glm(diabetes~.,
  family=binomial,
  data=PimaIndiansDiabetes)
```

In order to analyze the model performance, we need to convert the model into a uniform structure readable by the auditor package.

```
library(auditor)
audit_glm <- audit(mod_glm)
```

An object created with the `audit()` function can be used to draw different diagnostic plots. We will show some of them in this cheatsheet.

More detailed description and additional functionalities are presented in the package vignettes which can be found on the auditor website: <https://mi2-warsaw.github.io/auditor>

### REFERENCES

- Bi J., Bennett K.P. (2003). Regression error characteristic curves, in: Twentieth International Conference on Machine Learning (ICML-2003), Washington, DC.
- Hernández-Orallo, J. (2013). ROC curves for regression. Pattern Recognition, 46, 3395-3411.



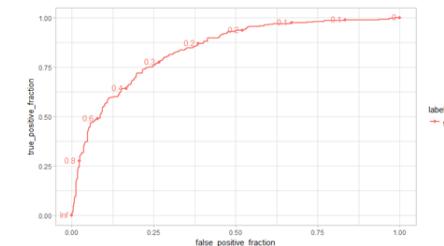
### ROC curves

The `plot()` function can be used to draw several different graphs. Use `type=plot name` to draw different types of diagnostic plots.

#### RECEIVER OPERATING CHARACTERISTICS (ROC)

Receiver Operating Characteristic Curve is a plot of the true positive rate (TPR) against the false positive rate (FPR) for the different thresholds. The area under the curve (AUC) is a measure of model accuracy.

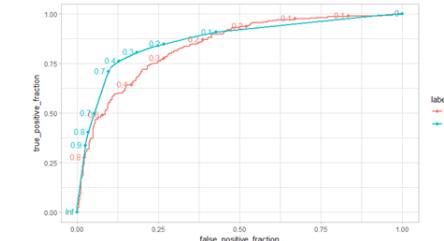
```
plot(audit_glm, type="ROC")
```



#### OVERLAYING RESPONSES FROM DIFFERENT MODELS

A very useful feature of the auditor is the possibility to overlay performance of different models on a single plot. Below we present results for logistic regression model and classification tree.

```
library(rpart)
mod_tree <- rpart(diabetes~.,
  data=PimaIndiansDiabetes)
p.fun <- function(model, data){predict(mod_tree)[,2]}
audit_tree <- audit(mod_tree, data=PimaIndiansDiabetes,
  y = PimaIndiansDiabetes$diabetes,
  predict.function = p.fun)
plotROC(audit_glm, audit_tree, type = "ROC")
```



CC BY Alicja Gosiewska • alicjagosiewska@gmail.com • <https://github.com/agosiewska> • Learn more at <https://github.com/mi2-warsaw/auditor> • package version 0.1.1 • Updated: 2018-03

### Regression ROC curves

auditor package provides functions which are adaptations of ROC curves for regression. Below we present them for a linear model.

As for ROC curves, for regression curves, there is a possibility to overlay performance of different models.

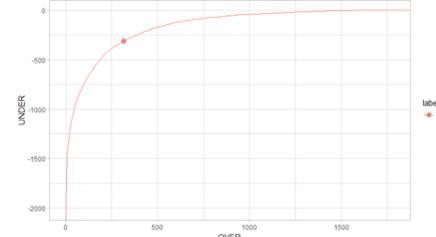
```
library(car)
model_lm <- lm(prestige~education + women + income,
  data = Prestige)
audit_lm <- audit(model_lm)
```

#### REGRESSION RECEIVER OPERATING CHARACTERISTIC (RROC)

The basic idea of the RROC is to show model asymmetry. The RROC is a plot where on the x-axis we depict total over-estimation and on the y-axis total under-estimation.

For RROC curves we use a shift, which is an equivalent to the threshold for ROC curves. For each observation we calculate new prediction:  $\hat{y}' = \hat{y} + s$  where  $s$  is the shift. The shift equals 0 is represented by a dot.

```
plot(audit_lm, type = "RROC")
```

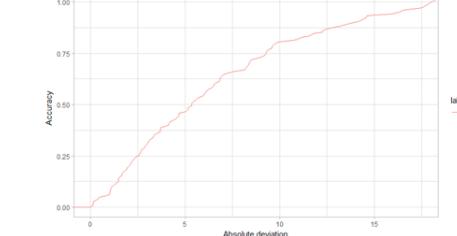


#### REGRESSION ERROR CHARACTERISTIC (REC)

REC curves illustrate the accuracy allowing for a certain level of error tolerance.

On the x-axis of the plot there is an error tolerance and on the y-axis, there is a percentage of observations predicted within the given tolerance. The REC curve estimates the cumulative distribution function (CDF) of the error.

```
plot(audit_lm, type = "REC")
```



**auditor:  
prezentacja  
pakietu**

# Ćwiczenia

# **Objaśnianie pojedynczej predykcji**

# Dlaczego?

**Gdy na bazie modelu ML podejmowane są ważne decyzje, model musi być godny zaufania**

- zrozumienie modelu budzi zaufania
- rośnie zapotrzebowanie na interpretowalne algorytmy  
**(przykłady: Weapons of math destruction, kontrowersje wokół algorytmów Facebooka itd.)**

# Dlaczego?

**Zapotrzebowanie na zrozumiałe modele powoduje powstawanie regulacji prawnych**

→ RODO!

# Dlaczego?

- 1) Zrozumienie predykcji pozwala wykryć problemy w modelu**
- 2) Wyjaśnienie predykcji pozwala porównać modele i wybrać lepszy**
- 3) Zrozumienie predykcji może pozwolić na poprawę modelu**

# BreakDown: idea

- Dekompozycja predykcji
- Założenie: predykcja jest sumą składowych
- Przypisujemy wagi do zmiennych
  - wizualizacja
  - warto rozkładać różnicę pomiędzy predykcją i średnią predykcji
  - podobny pomysł: Shapley values

# breakDown

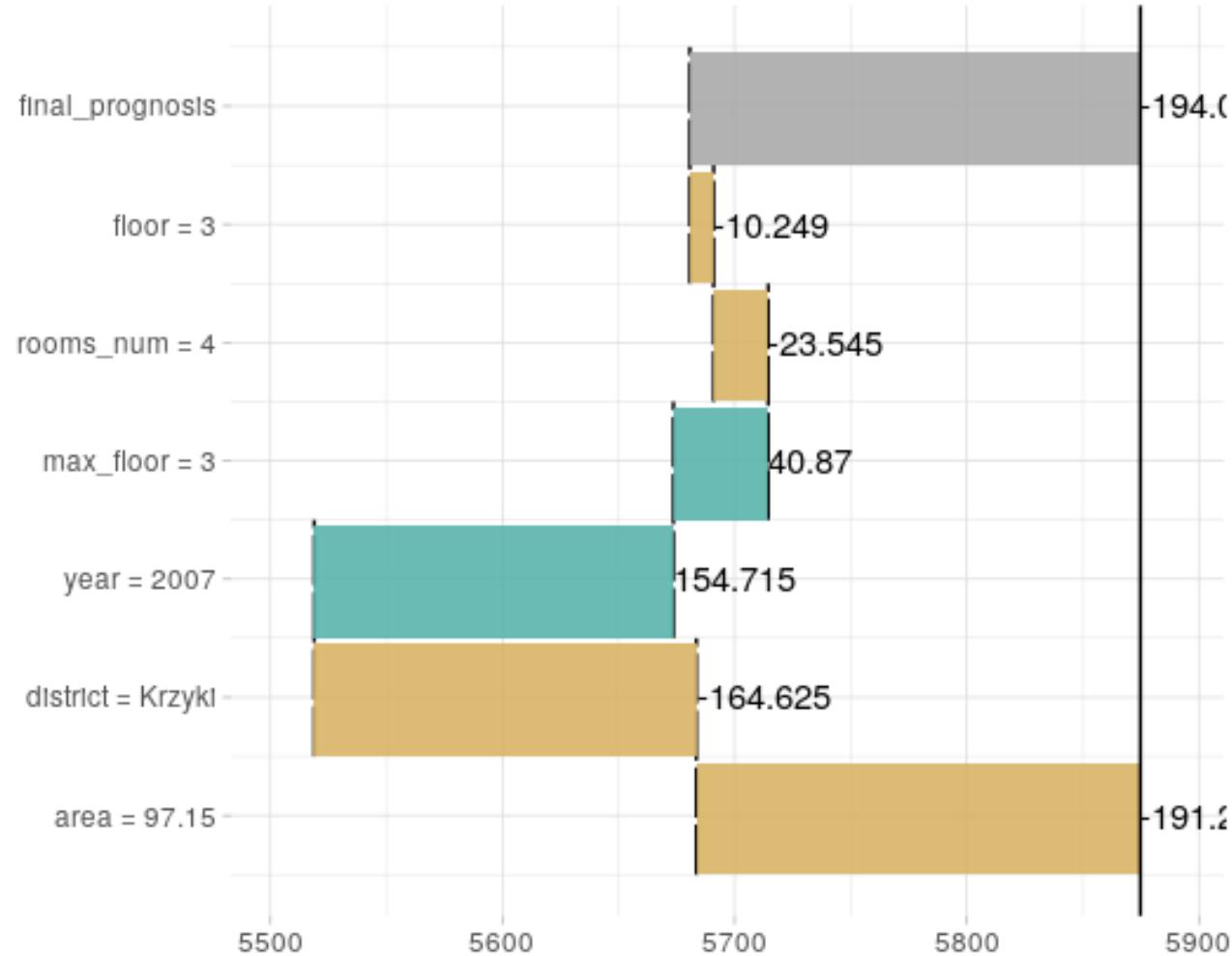
- **Model liniowe**

$$f(x^{new}) = (1, x^{new})(\mu, \beta)^T = baseline + (x_1^{new} - \bar{x}_1)\beta_1 + \dots + (x_p^{new} - \bar{x}_p)\beta_p$$

$$baseline = \mu + \bar{x}_1\beta_1 + \dots + \bar{x}_p\beta_p.$$

- **Dowolny model**  
→ **algorytm zachłanny**

# breakDown



# **breakDown: prezentacja pakietu**

# Ćwiczenia

# Live

## live: Local Interpretable (Model-agnostic) Visual Explanations

CRAN 1.5.4 downloads 646/month downloads 821 build passing coverage 41% [Tweet](#)

### Installation

Install stable CRAN version:

```
install.packages("live")
```

or the development version:

```
devtools::install_github("MI2DataLab/live")
```

[See the latest changes.](#)

Features coming up next:

- more methods of sampling,
- better support for comparing explanations for different models / different instances,
- Improved Shiny application (see `live_shiny` function in development version).

If you have any bug reports, feature requests or ideas to improve the methodology, feel free to leave an issue.

### Materials

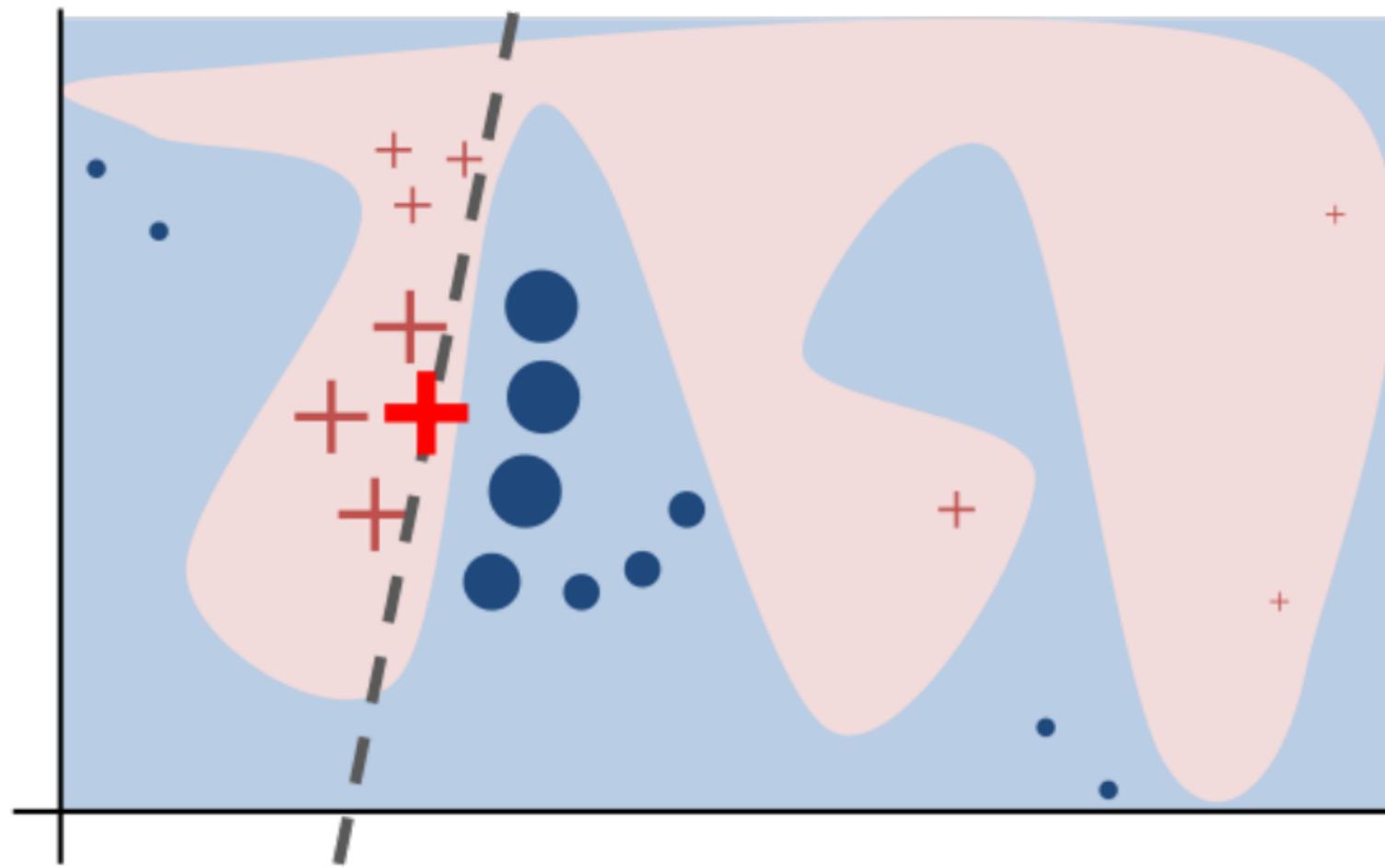
Find the paper about `live` and `breakDown` on arXiv.

Website: <https://mi2datalab.github.io/live/>

Conference talk on `live`: [https://github.com/mstanlak/Berlin\\_2017](https://github.com/mstanlak/Berlin_2017)

Cheatsheet

# Live: idea - metoda LIME



# Lokalne aproksymacje: jak to zrobić?

- 1) Symulacja nowego zbioru danych**
- 2) Dodanie predykcji objaśnianego modelu dla nowych obserwacji**
- 3) Dopasowanie prostego modelu do predykcji czarnej skrzynki**
- 4) Wizualizacja modelu i jego analiza**

**LIVE: prezentacja  
pakietu**

# Ćwiczenia

# Alternatywny soft

**1) Zbiór metod: pakiet `iml`**

**2) Poszczególne metody:**

- `pdp`
- `ICEbox`
- `ALEPlot`
- `lime`
- `shapleyr`
- `auditor`

# Model-specific

## 1) Pakiety

- **randomForestExplainer**
- **xgboostExplainer**
- **condvis**
- **forestmodel / forestplot / sjPlot**
- **randomForest i inne implementacje**

# Więcej

[https://pbiecek.github.io/DALEX\\_docs/](https://pbiecek.github.io/DALEX_docs/)

<https://christophm.github.io/interpretable-ml-book/>

# Zaproszenie: Why R?

1) Zapisy do 15.06!

The header features a wide-angle photograph of the Wroclaw Old Town skyline at dusk or night. The sky is a deep blue with scattered clouds. In the foreground, the ornate facades of several historic buildings are illuminated, with warm yellow lights coming from windows and street lamps. A prominent red brick tower with a clock is visible on the right side. On the left, there's a black circular logo for the conference, containing the letters 'WHY R?' and the word 'Conference' below it, with 'WROCŁAW 2018' written underneath.

ABOUT   SPEAKERS   PROGRAMME   PRE-MEETINGS   VENUE   HACKATON   SPONSORS   ORGANISING COMMITTEE

WHY R?  
Conference  
WROCŁAW 2018

Why R?

Wrocław, 2-5 July 2018

Submit your abstract   Register now

# Zaproszenie: DSS

dssconf.pl

The screenshot shows the homepage of the Data Science Summit (DSS) conference website. At the top, there are navigation icons for search, user profile, and login. Below the header, there are sections for diamond sponsors (Accenture), silver sponsors (Samsung, Lingaro, NetPhone, Vertica), and a logo featuring three interlocking hexagons. The main title "DATA SCIENCE SUMMIT" is displayed prominently. A call-to-action section encourages users to register for the second edition of the event in Warsaw on June 8th. A large graphic on the right side features a 3D perspective of overlapping hexagonal shapes in purple and cyan. At the bottom, there are two callout boxes: one for becoming a sponsor/partner and another for presenting a talk.

sponsor diamentowy

sponsorzy srebrni

accenture

SAMSUNG lingaro NetPhone VERTICA

DATA SCIENCE SUMMIT

Zarejestruj się na II edycję największego wydarzenia Data Science w Polsce

› adresowanego do profesjonalistów, jak i zainteresowanych tym obszarem  
› przygotowanego przez uznane społeczności eksperckie we współpracy z czołowymi ośrodkami akademickimi

piątek, 8 czerwca, Warszawa

ZAREJESTRUJ SIĘ →

Poleć znajomym lub przypomnij później

○ Chcesz zostać Sponsorem lub Partnerem?  
    > Skontaktuj się z nami.

□ Chcesz wygłosić prezentację na DSS?  
    > Skontaktuj się z nami.

# Podziękowanie

KLIENCI ZADŁUŻENI | PARTNERZY BIZNESOWI | RELACJE INWESTORSKIE | KARIERA | DLA PRASY | Logowanie e-KRUK

JAK MOŻEMY CI POMÓC? PORADNIK O NAS DO POBRANIA KONTAKT PL

Masz dług w KRUKe?  
Zarządzaj swoim  
długiem przez internet.

**Sprawdź nas**

PORADY [ZOBACZ WSZYSTKIE PORADY >](#) KONTAKT [KONTAKT >](#)

Kruk - Pomoc dla osób zadłużonych

71 8888 000

Pracujemy od poniedziałku do piątku.  
Biuro czynne w godz. od 8:00 do 16:00.  
Infolinia czynna w godz. od 8:00 do 21:00.  
Każdy rok połączony jest z nowym kontaktem.

49

**Dziękuję za uwagę!**

**Zapraszamy na piwo w  
Cybermachinie!**