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

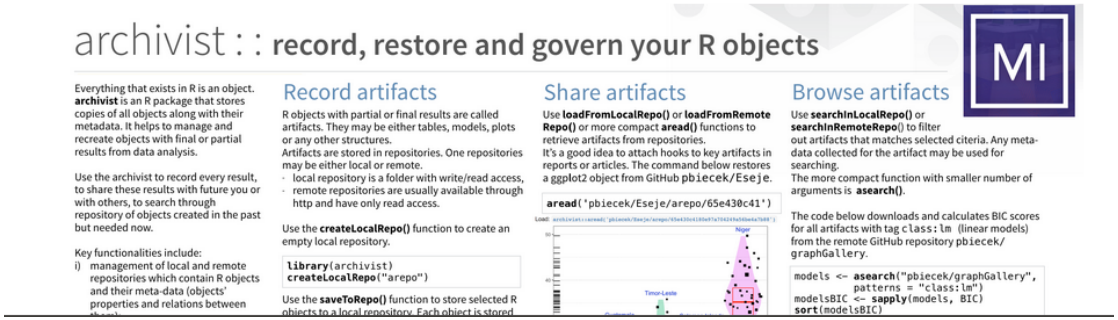
To get started, install the latest version of **archivist** from CRAN:

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

or from GitHub:

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

Cheatsheet



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

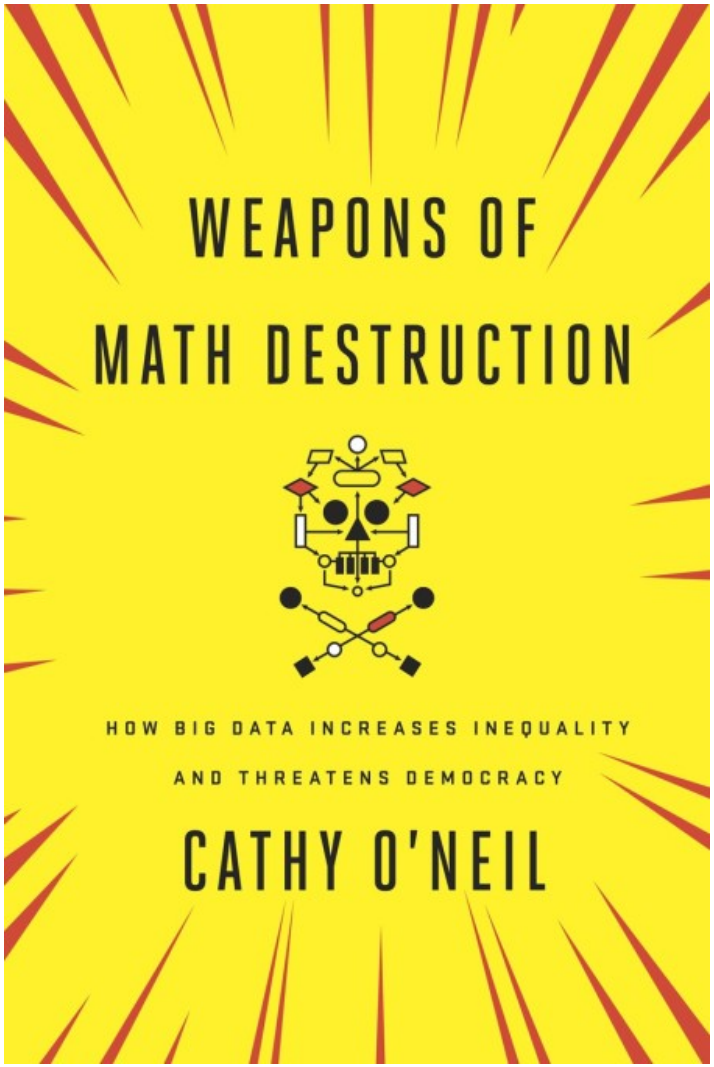
Provide evidence-backed cancer care to each patient, by understanding millions of data points

Meet at an upcoming event | Understand the data

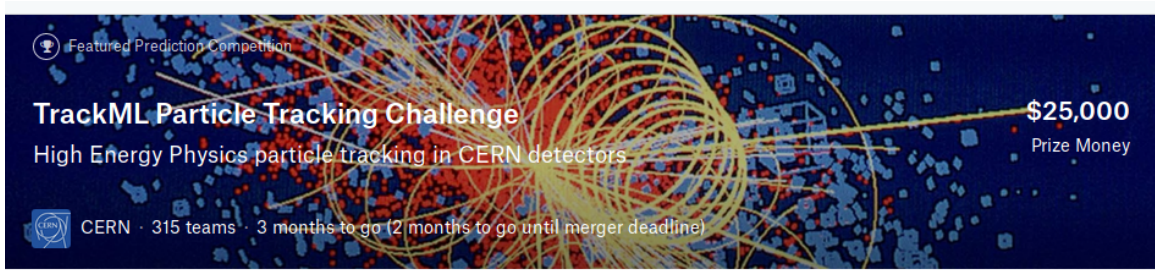
Put Watson for Oncology to the test with your tumor board today

Watson for Oncology: 90% concordance with tumor board recommendation

Dlaczego?


WEAPONS OF
MATH DESTRUCTION
HOW BIG DATA INCREASES INEQUALITY
AND THREATENS DEMOCRACY
CATHY O'NEIL

Dlaczego?



Overview Data Kernels Discussion Leaderboard Rules

Public Leaderboard Private Leaderboard

This leaderboard is calculated with approximately 29% of the test data.
The final results will be based on the other 71%, so the final standings may be different.

Raw Data Refresh

In the money Gold Silver Bronze

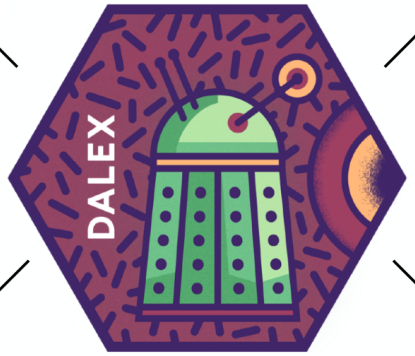
#	△1w	Team Name	Kernel	Team Members	Score	Entries	Last
1	–	Mickey			0.6405	7	9d
2	▲257	Zidmie			0.5829	5	3d
3	▼1	Vicens Gaitan			0.4969	2	6d
4	▲1	Lin12345			0.4775	16	20h
5	▲5	HomerJSimpson			0.4306	11	11h
6	▼3	Grzegorz Sionkowski			0.4225	12	1d
7	▲28	Andrea			0.4219	4	2d
8	new	Vivek			0.4207	4	1d
9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

How good is the predictive model?

Which variables are the most important in general?



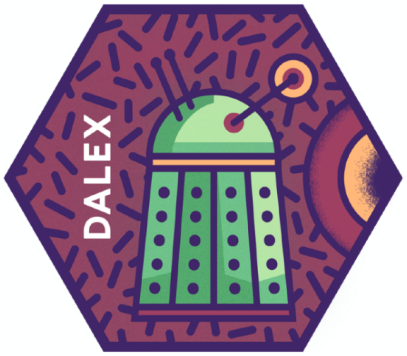
How good is the model fit?

Which variables influence the single prediction?

Czym jest DALEX?

Variable Explainers
package: pdp, ALEPlot, **factorMerger**

Structure Explainers
package: randomForest



Model Management
package: **archivist**

Model Diagnostic Tools
package: **auditor**, ggfortify

Model Performance Explainers
package: **auditor,** ROCR, caret, mlr

Model Predictions Explainers
package: **breakDown**, **live**, shapleyr, lime

Wpływ pojedynczej zmiennej: idea

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

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

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)

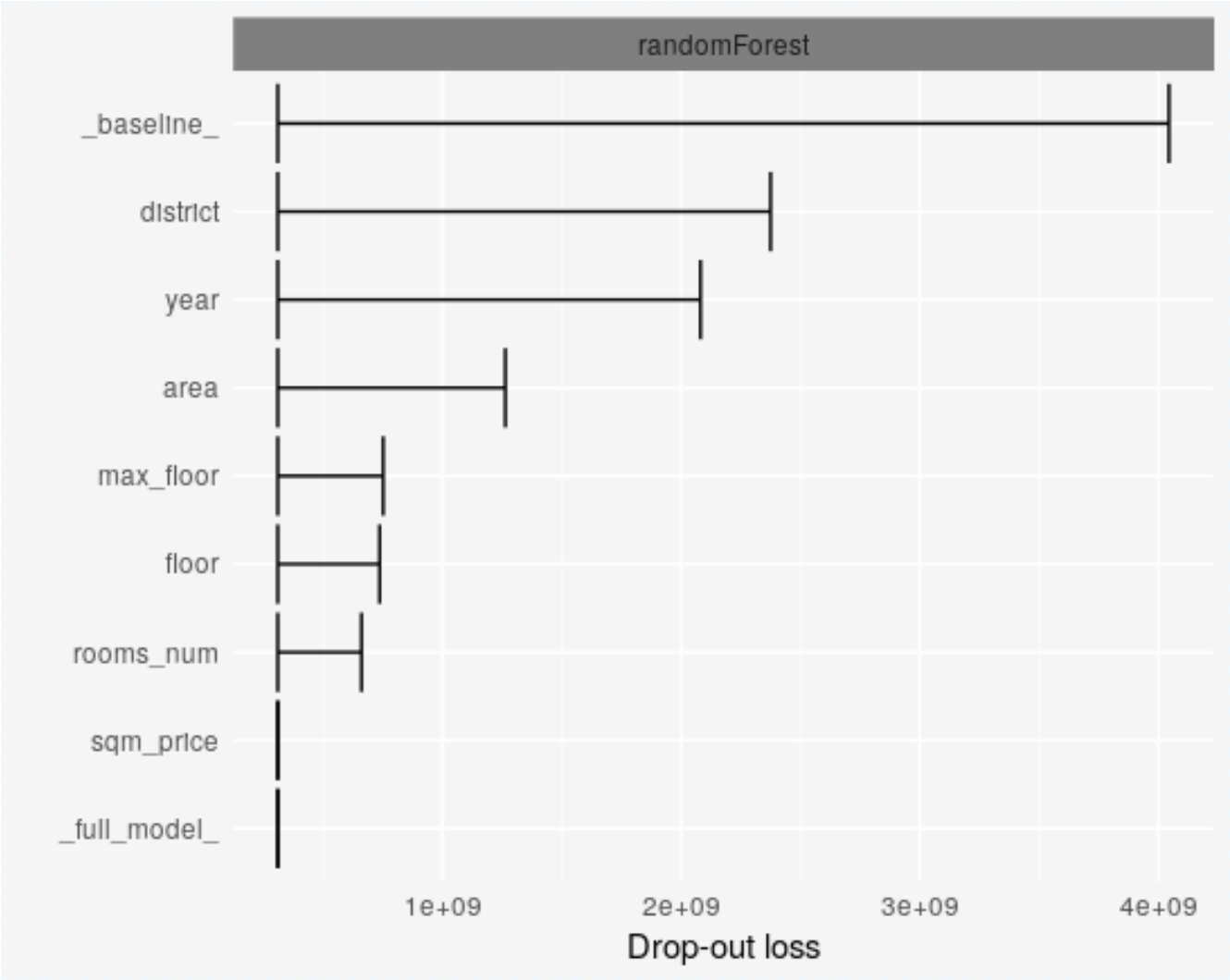
Funkcja single_variable

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



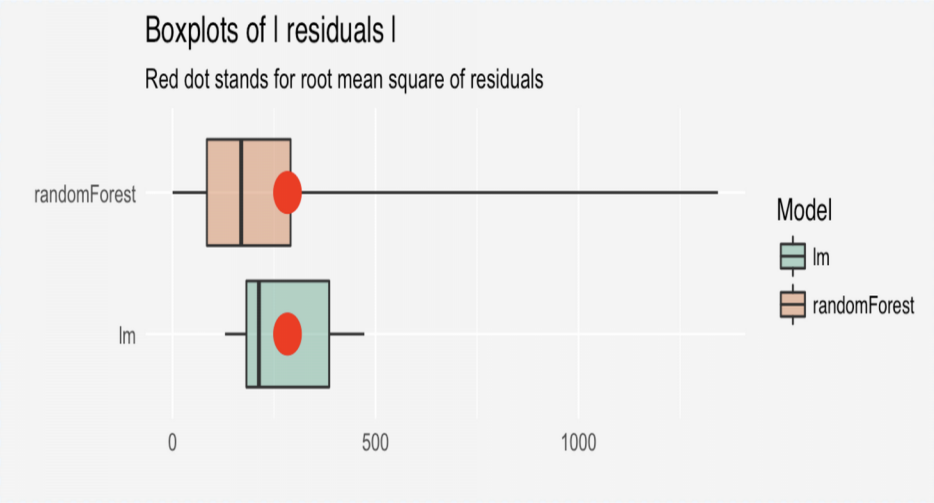
Funkcja variable_dropout

Ćwiczenia

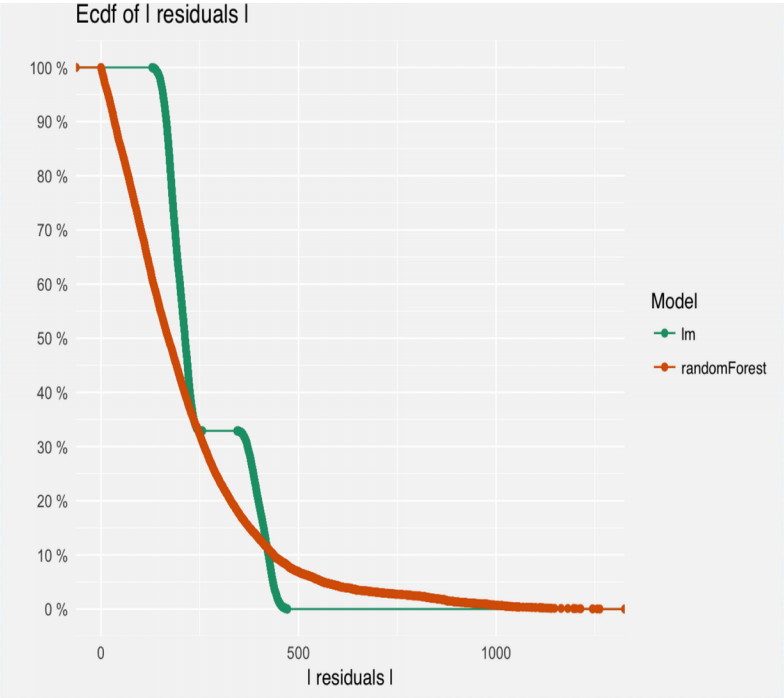
Jakość modelu: idea

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.*

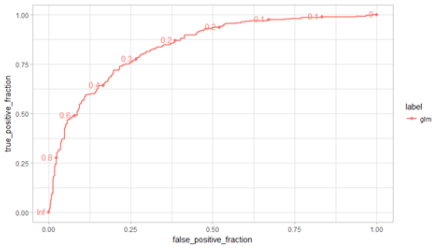
ROC curves

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.

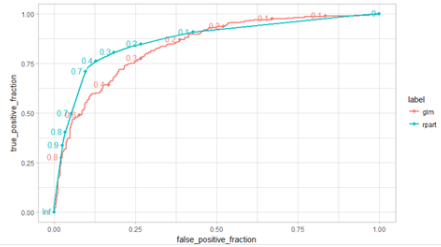
```
plot(audit_glm, type="ROC")
```



OVERLYING 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")
```



Regression ROC curves

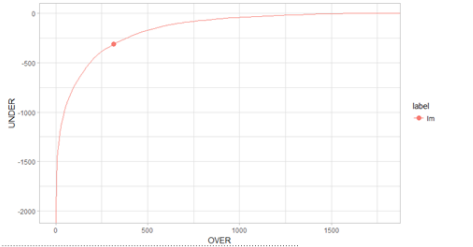
auditor package provides functions which are adaptations of ROC curves for regression. Below we present them for a linear model.
As for ROC curves, for regression curves, there is a possibility to overlay performance of different models.

```
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.

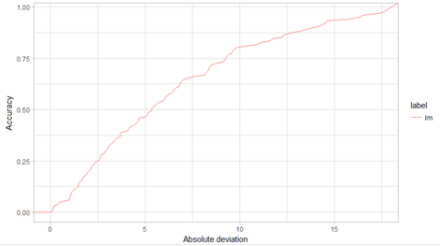
```
plot(audit_lm, type = "REC")
```



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 = "RROC")
```



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**

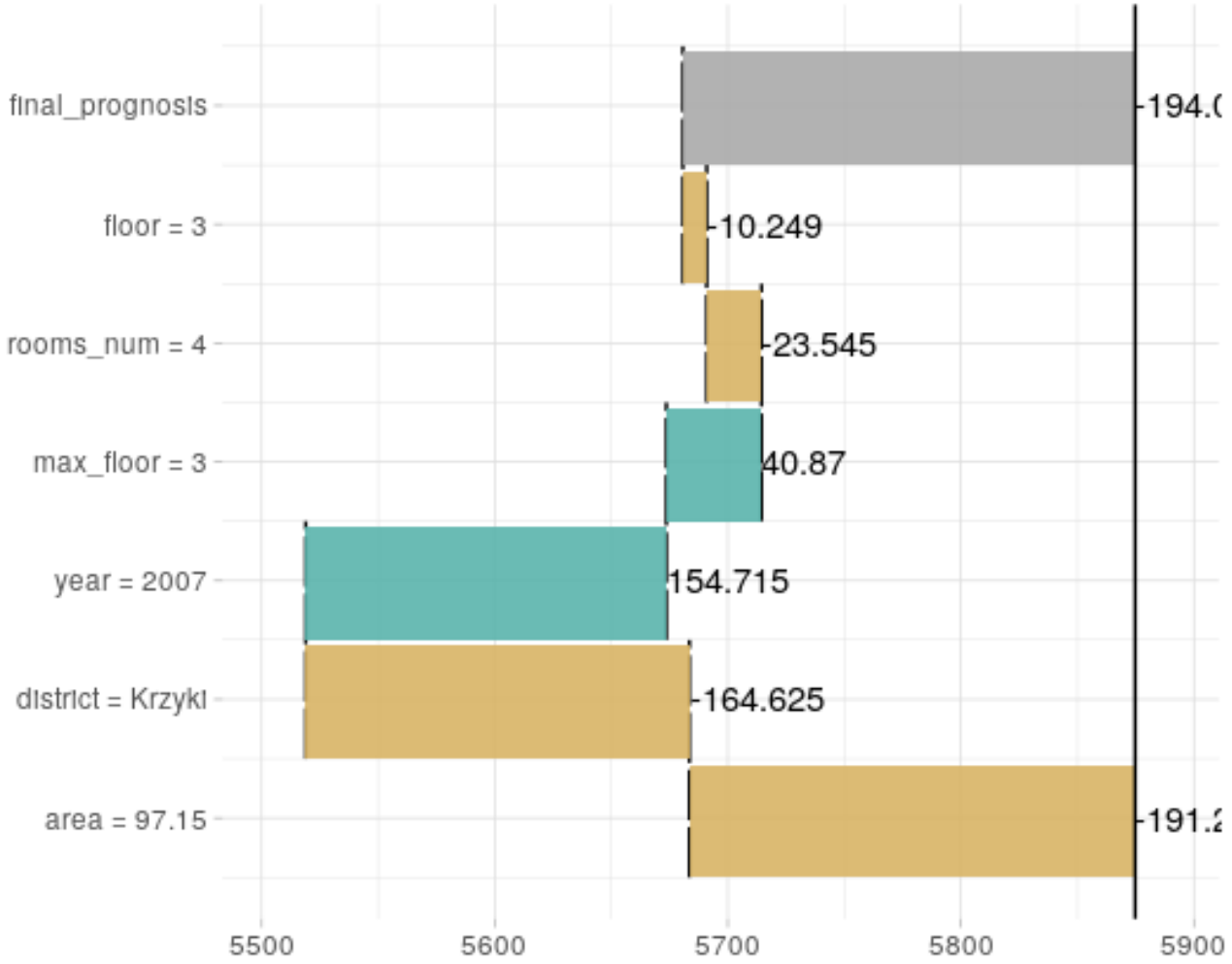
$$f(x^{new}) = (1, x^{new})(\mu, \beta)^T = baseline + (x_1^{new} - \bar{x}_1)\beta_1 + ... + (x_p^{new} - \bar{x}_p)\beta_p$$

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

- **Dowolny model**

  → algorytm zachłanny

breakDown


final_prognosis
floor = 3
rooms_num = 4
max_floor = 3
year = 2007
district = Krzyki
area = 97.15
194.(
10.249
23.545
40.87
154.715
164.625
191.2
5500
5600
5700
5800
5900

breakDown: prezentacja pakietu

Ćwiczenia

Live

live: Local Interpretable (Model-agnostic) Visual Explanations

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

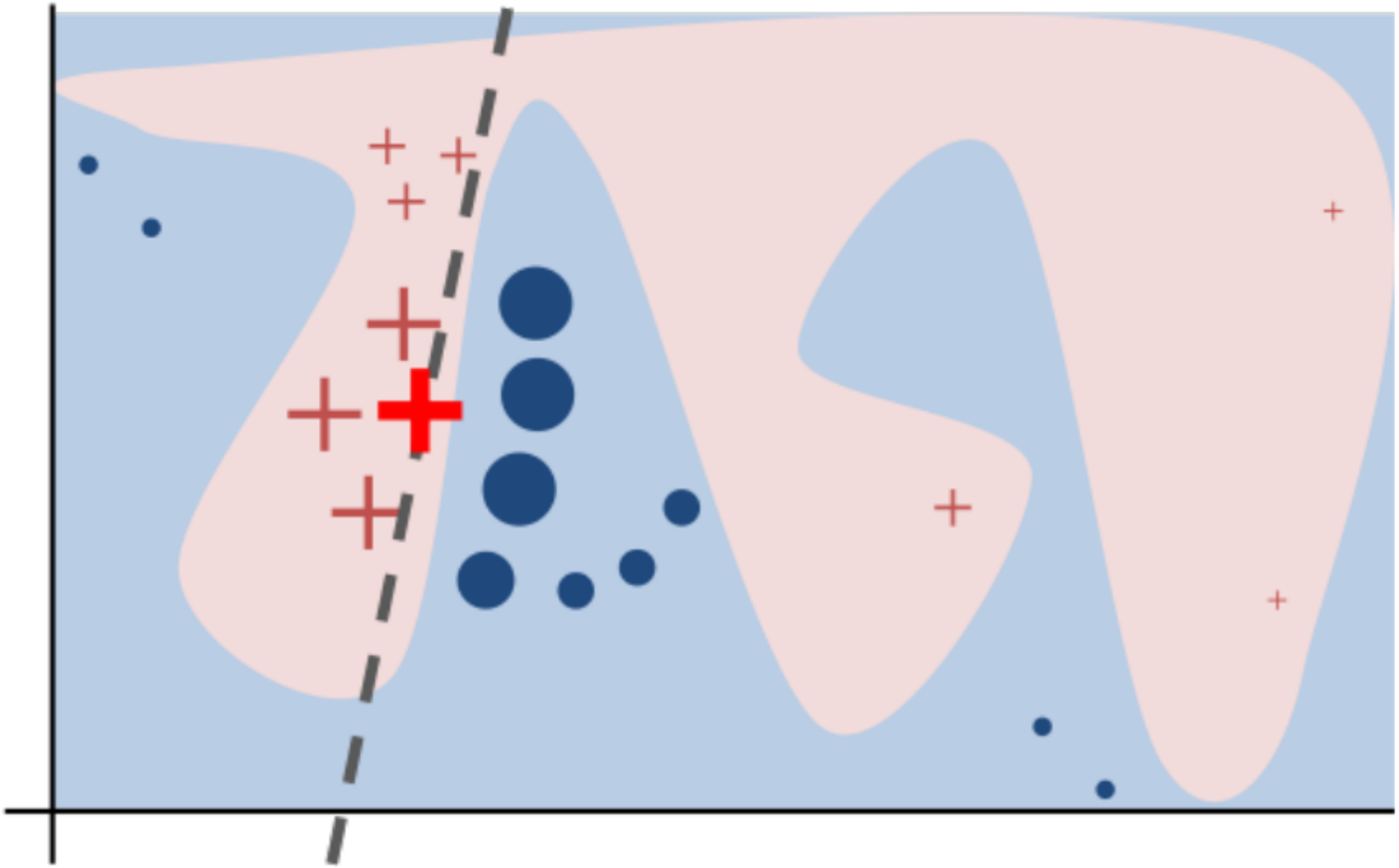
Find the paper about `live` and breakDown on arXiv.

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

Conference talk on `live` : https://github.com/mstaniak/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

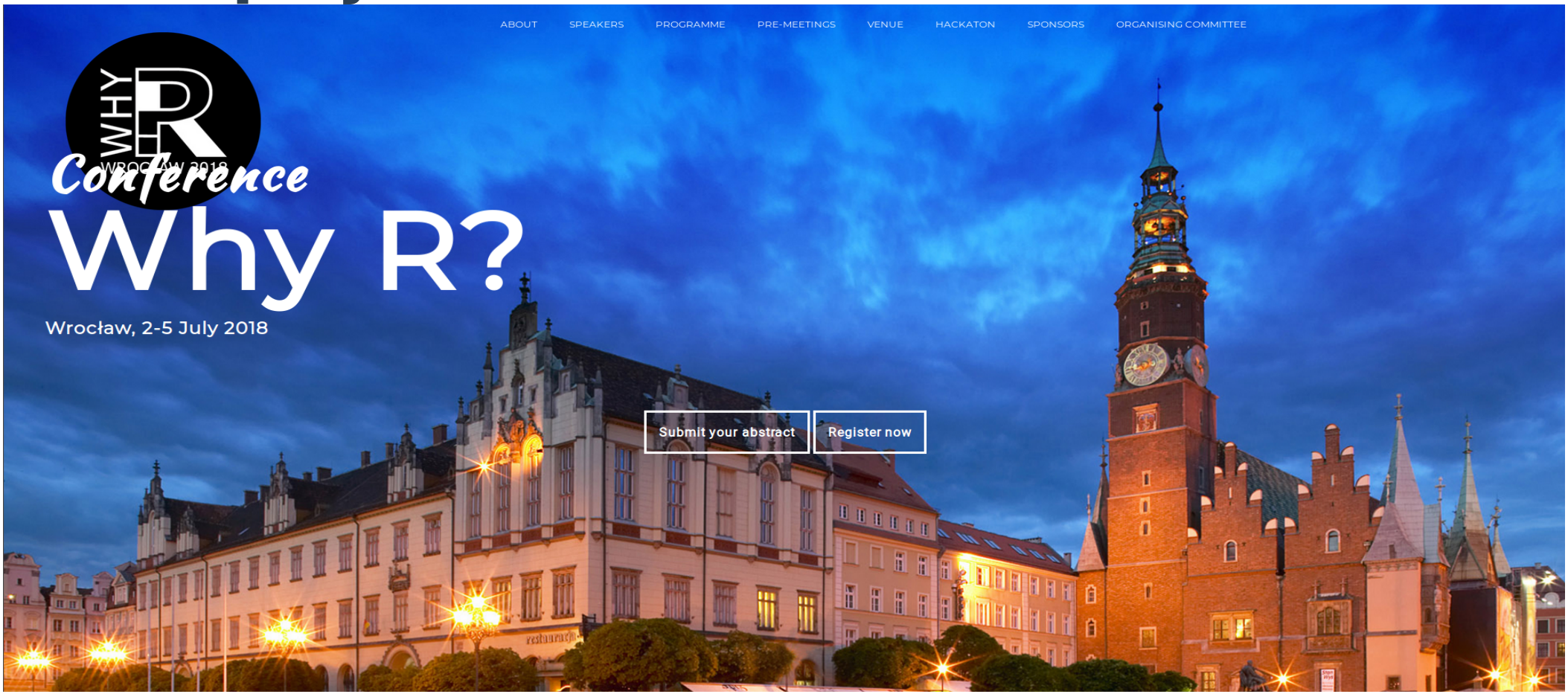
Więcej

https://pbiecek.github.io/DALEX_docs/

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

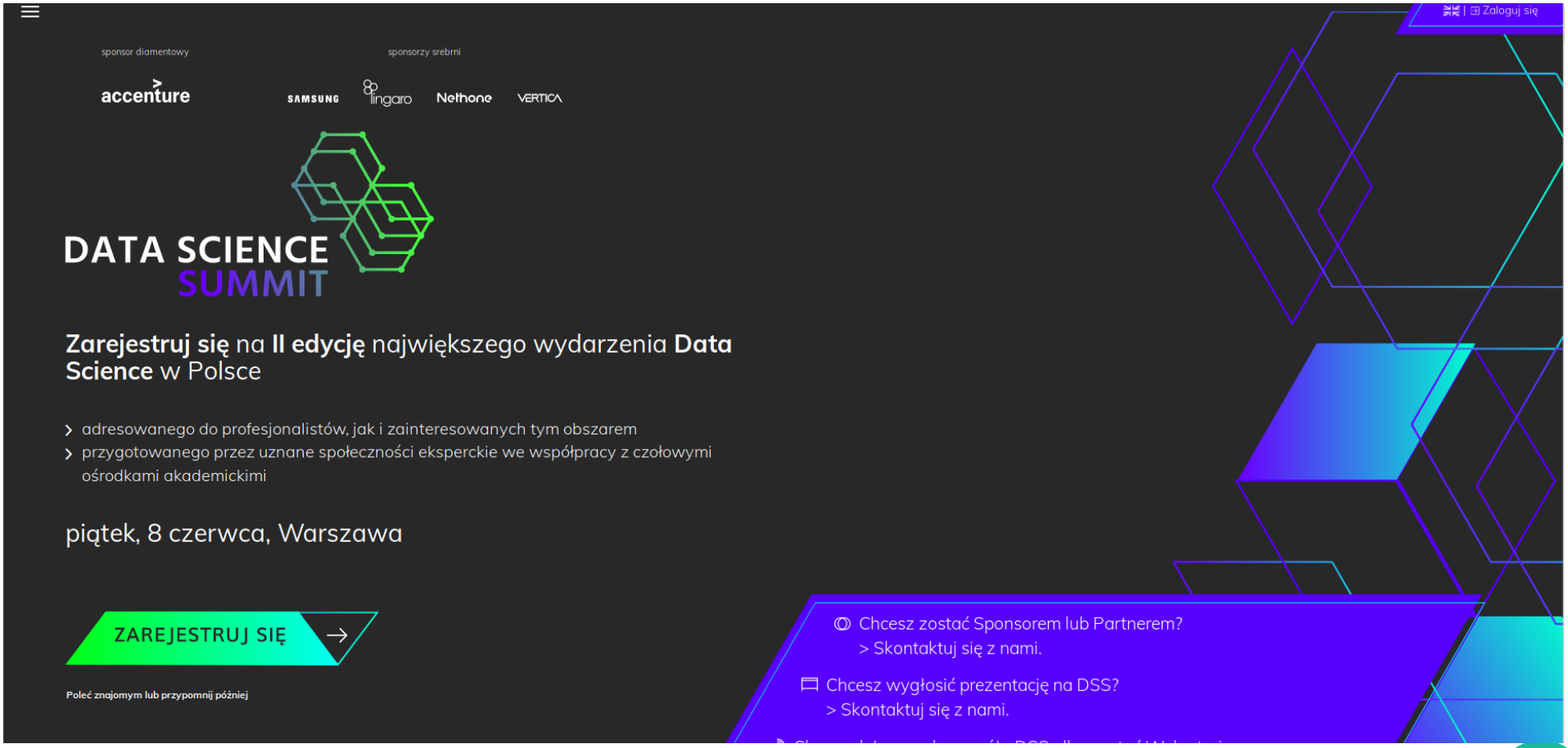
Zaproszenie: Why R?

1)Zapisy do 15.06!

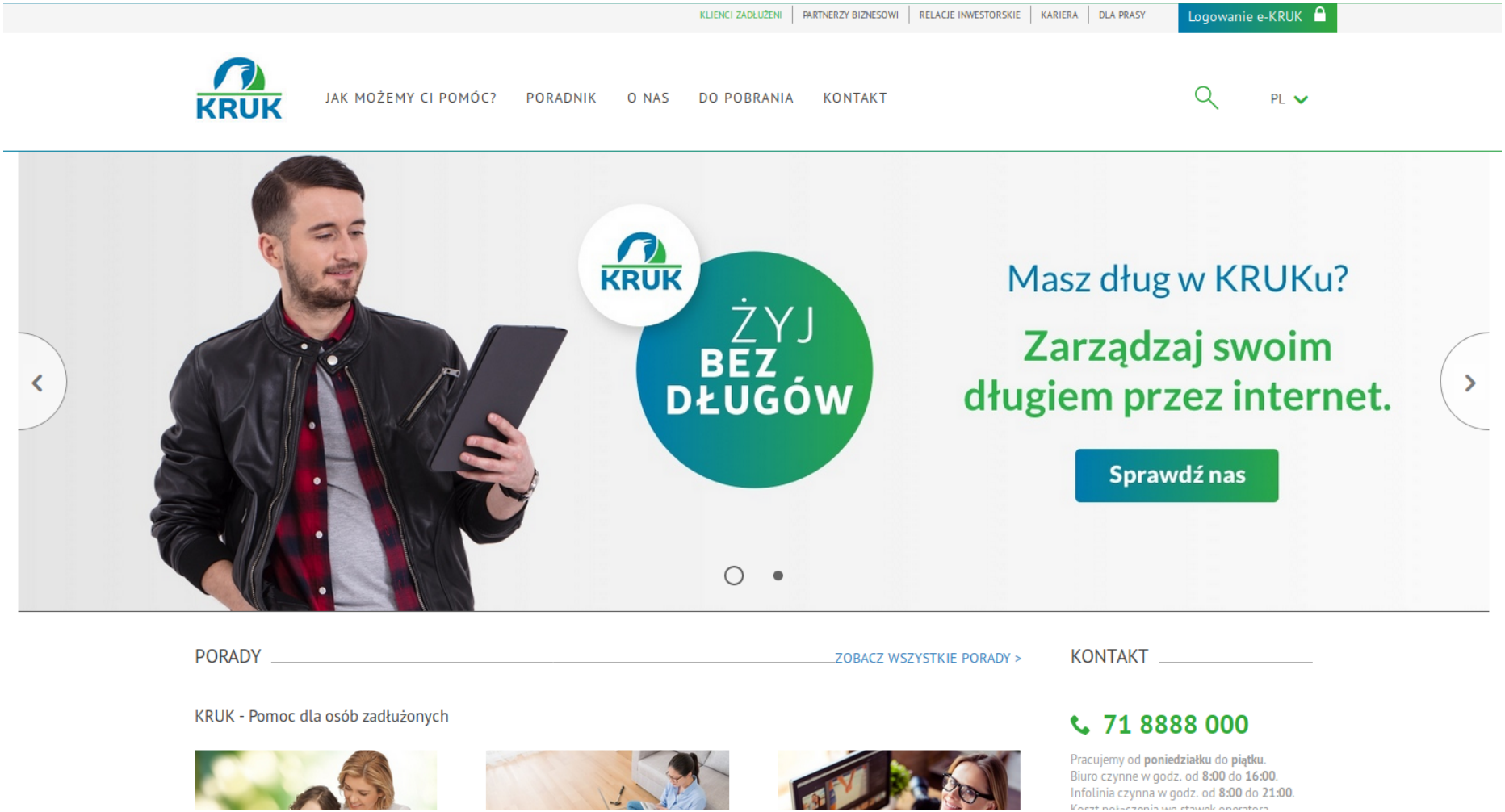


Zaproszenie: DSS

dssconf.pl



Podziękowanie



Dziękuję za uwagę!

Zapraszamy na piwo w Cybermachinie!