GBEx – towards Graph-Based Explanations

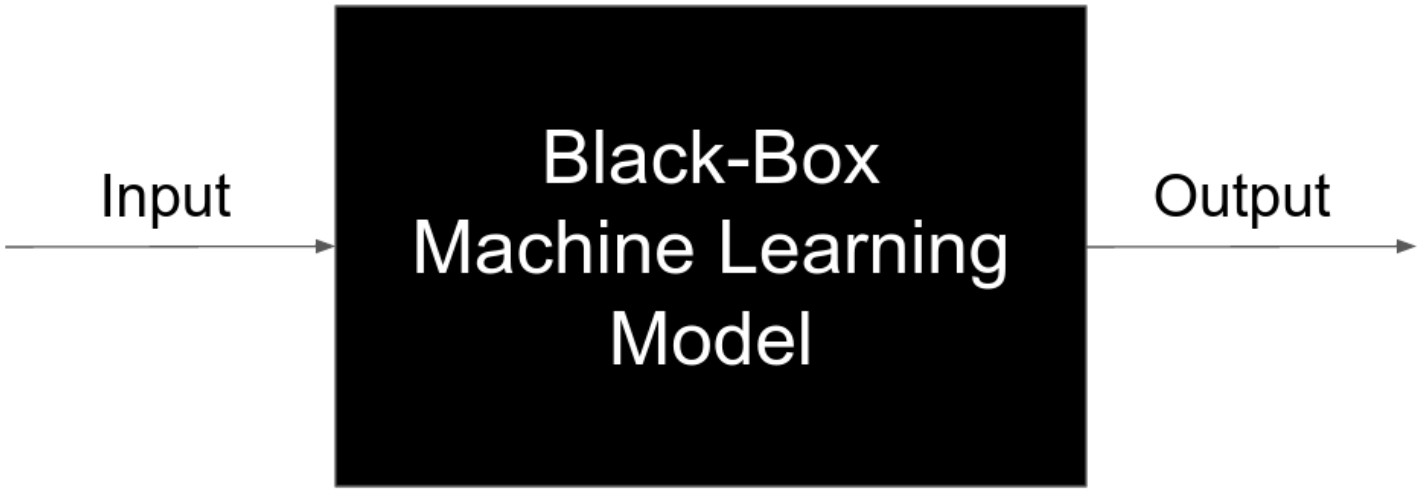
Paweł Mróz, Alexandre Quemy, Mateusz Ślażyński,
Krzysztof Kluza, Paweł Jemioło

IBM Krakow Software Lab, Cracow, Poland
Faculty of Computing, Poznań University of Technology, Poznań, Poland
AGH University of Science and Technology, Krakow, Poland

32th International Conference on Tools with Artificial Intelligence

Outline

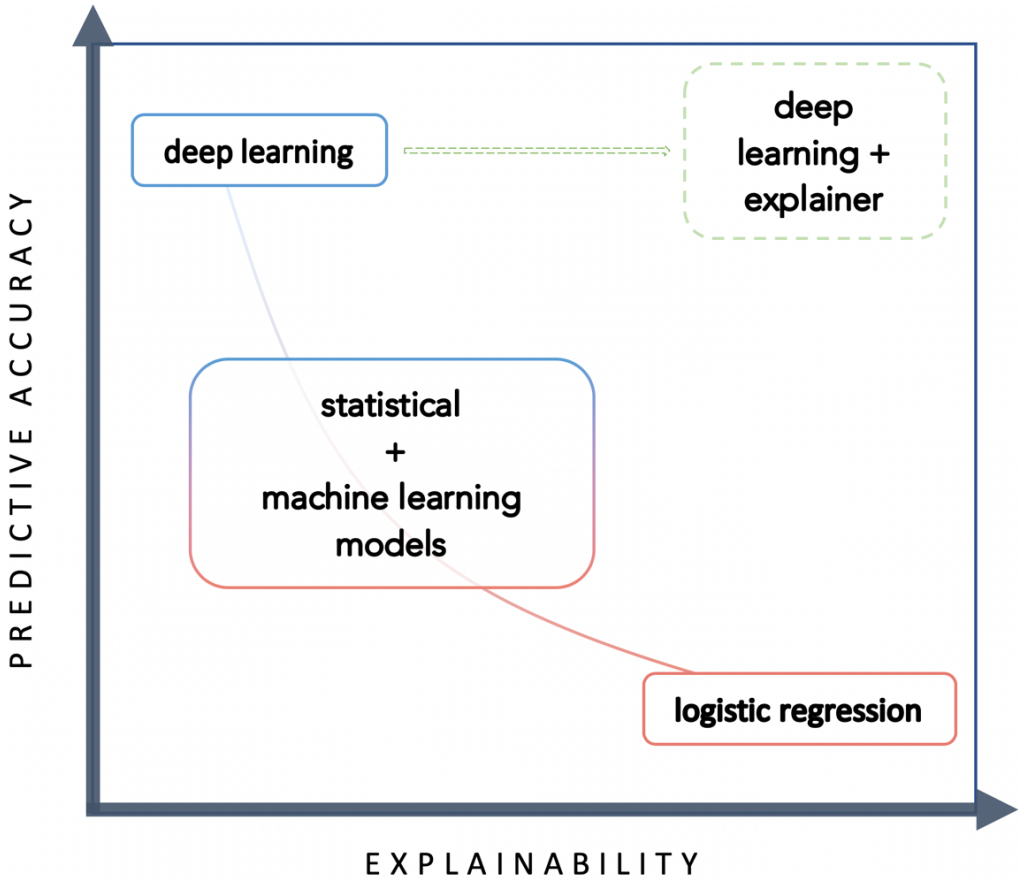
Black-box models



Why do we need explanations?

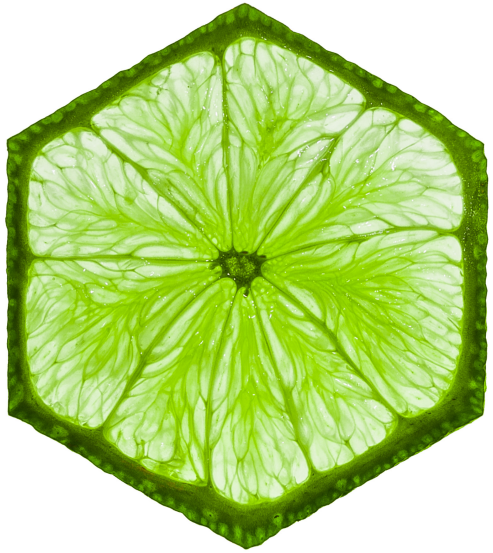
Source: *"Why Should I Trust You?": Explaining the Predictions of Any Classifier.*

Trade-off between explainability and accuracy



Source: *Characterising risk of in-hospital mortality following cardiac arrest using machine learning: A retrospective international registry study.*

State-of-the-art methods



Implementation and experiments

Main equation

$$\hat{y} = W^1\mu^1 + W^2\mu^2 + \beta \tag{1}$$

where:

- $\hat{y}$ – the vector to approximate or explain.
- W^1 – the input matrix of arguments.
- μ^1 – the vector of nodes importance.
- W^2 – the input matrix of connections.
- μ^2 – the vector of edge importance
- β – the base value.

Solving the equation

- Splitting equation in two parts
- Using heuristic methods
- Combine it into one big equation to solve

$$W^0 = [W^1 \;\; W^2] \tag{2}$$

$$\mu^0 = \begin{bmatrix} \mu^1 \\ \mu^2 \end{bmatrix} \tag{3}$$

The Equation 1 could be transformed to the following form:

$$\hat{y} = W^0 \mu^0 + \beta \tag{4}$$

Preparing data and presenting results

Preparing data

- GBEx can handle only categorized data
- To approach real numbers as input we performed clustering

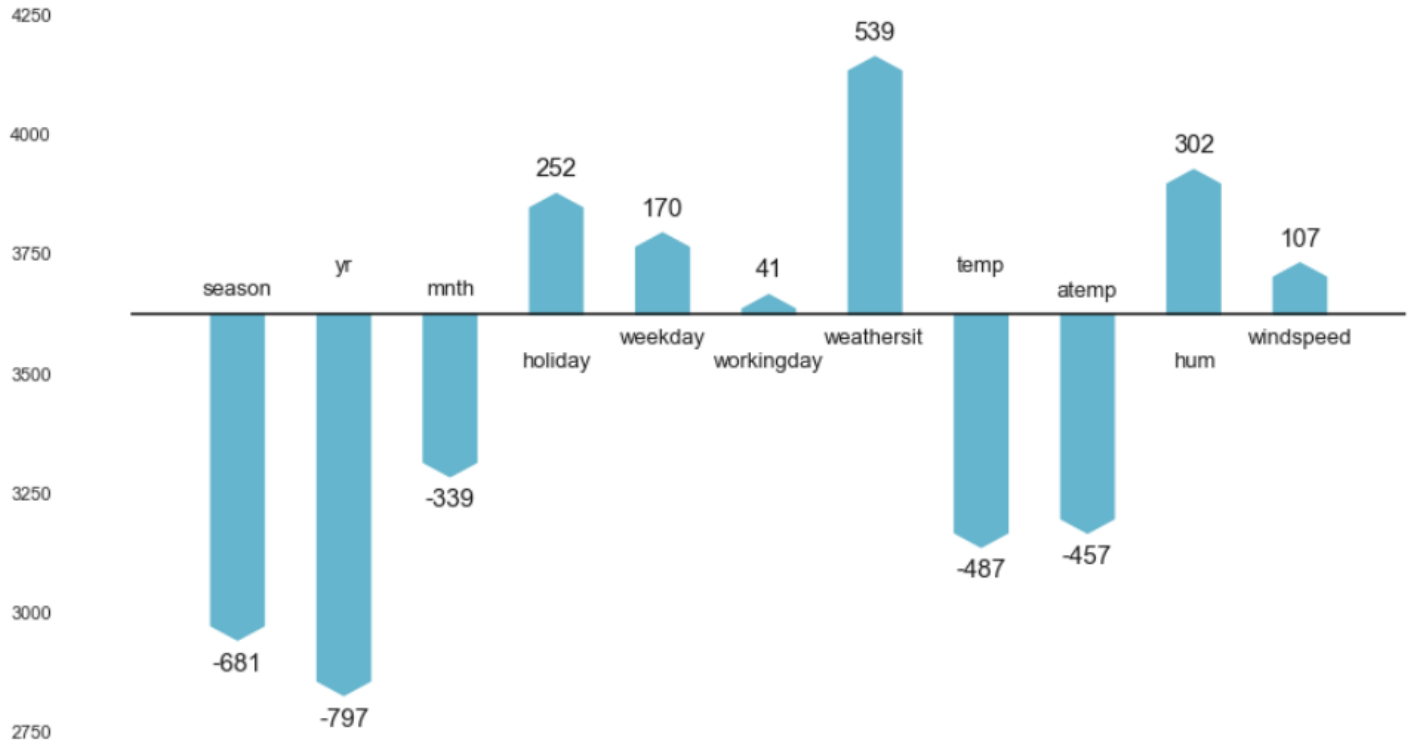
Presenting results

- Single case - bar-chart, heatmap
- Single case - graph
- General explanation - pie-chart, heatmap
- General explanation - graph

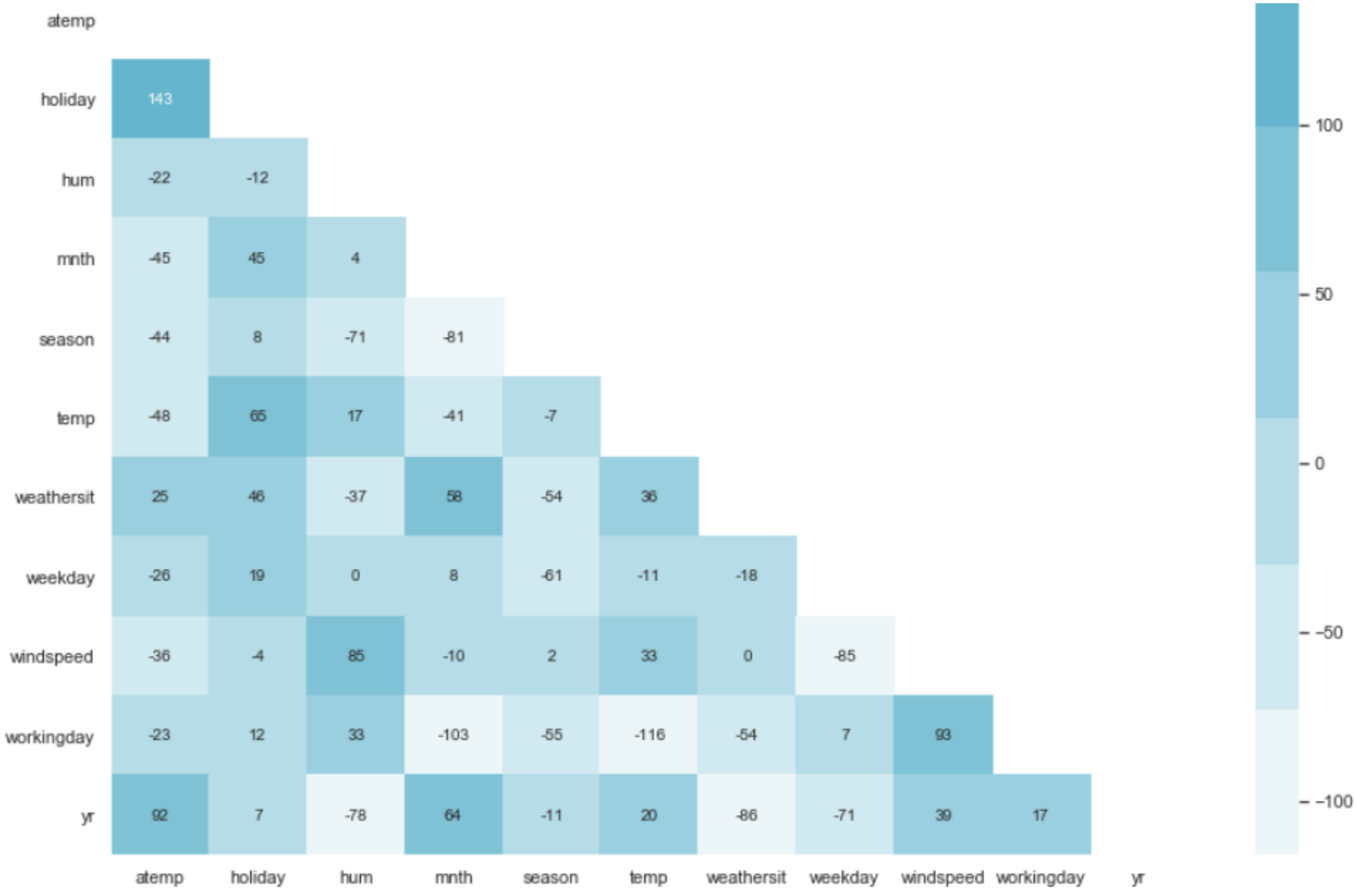
Values achieved in experiment

- Date = 24.12.2011
- Ground truth = 1011.
- MLP Regressor = 2026.
- Base value = 3624.
- GBEx = 1939.
- GBEx (without interactions) = 2274.

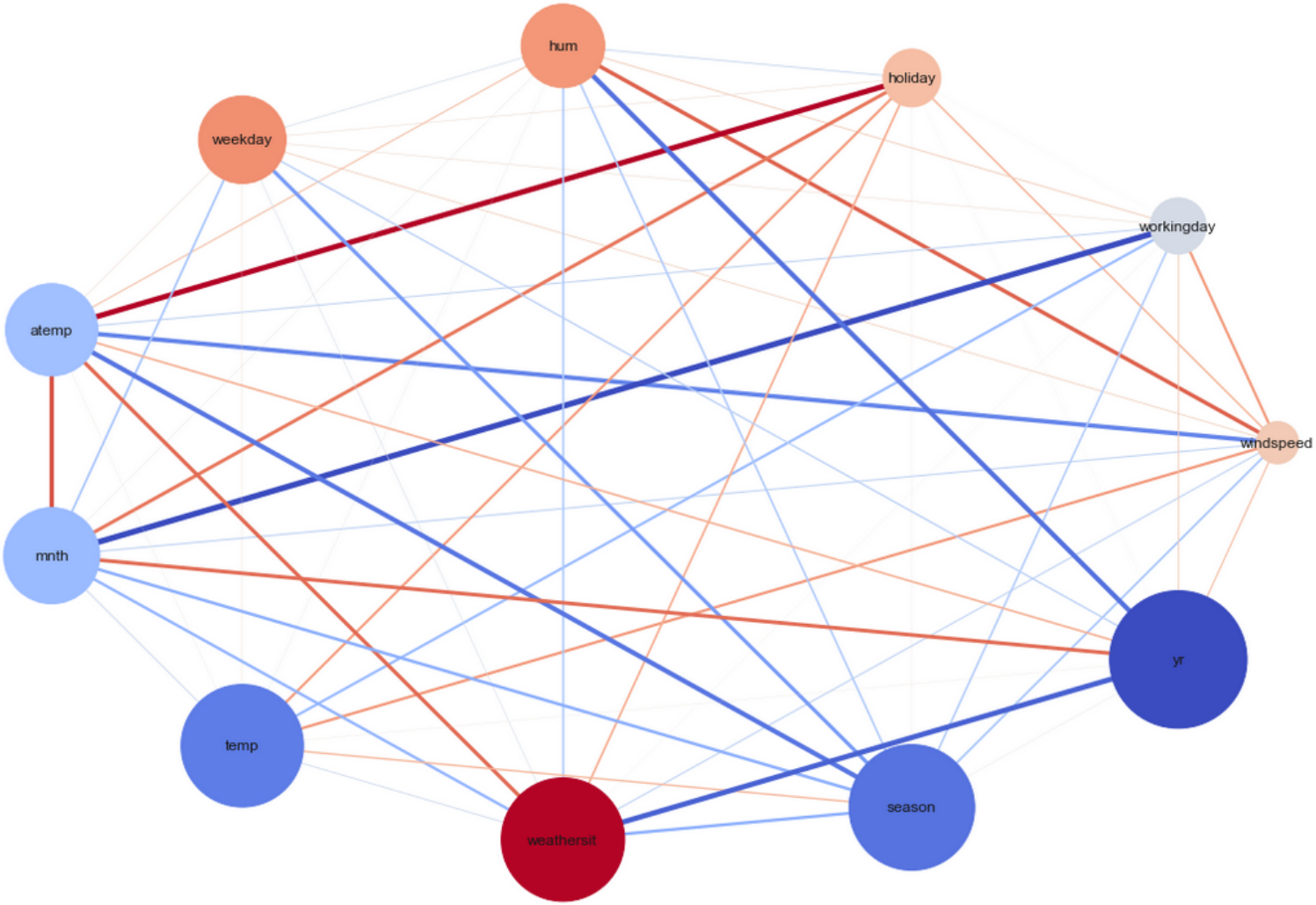
Single case explanation - bar chart



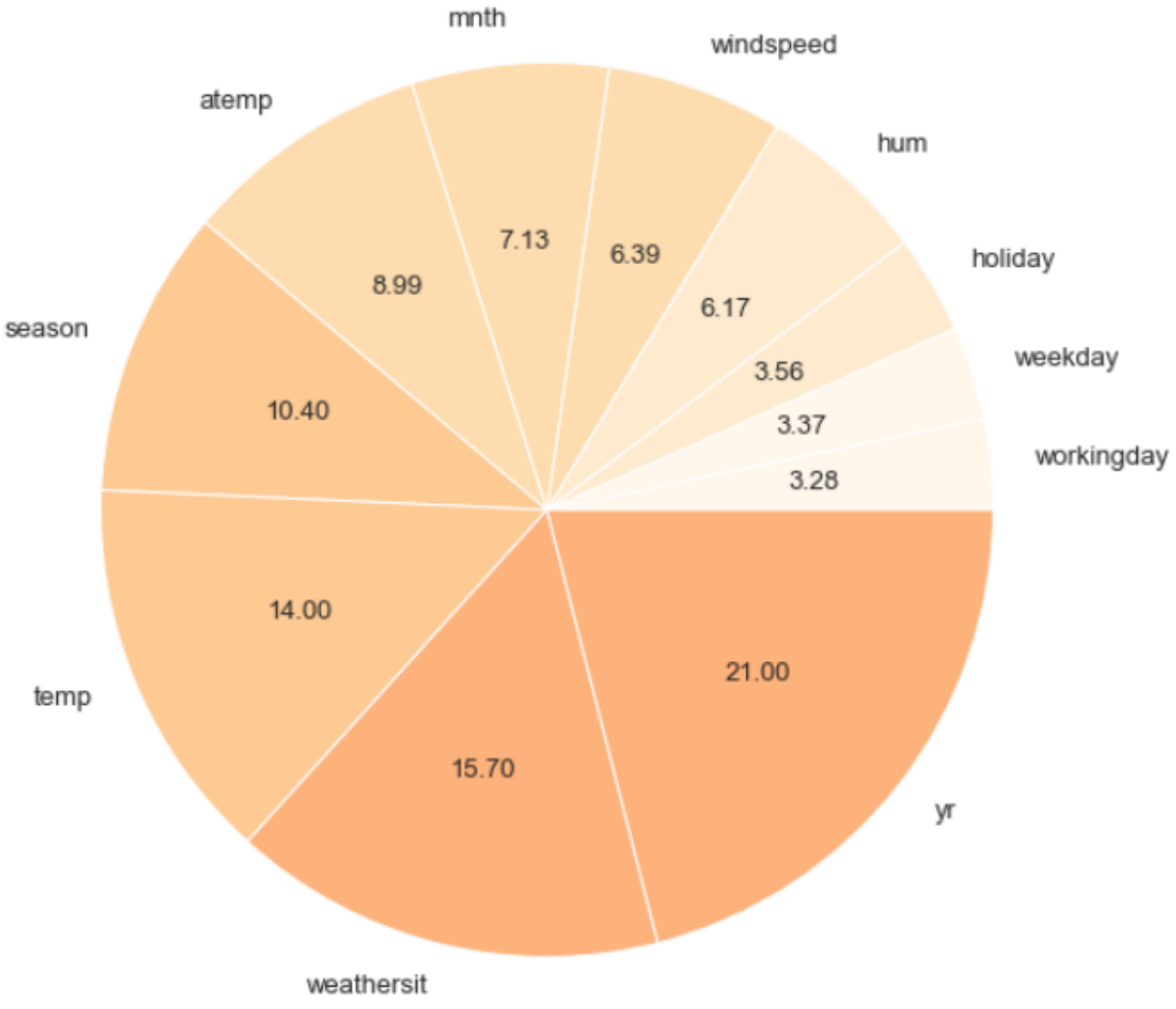
Single case explanation - heatmap



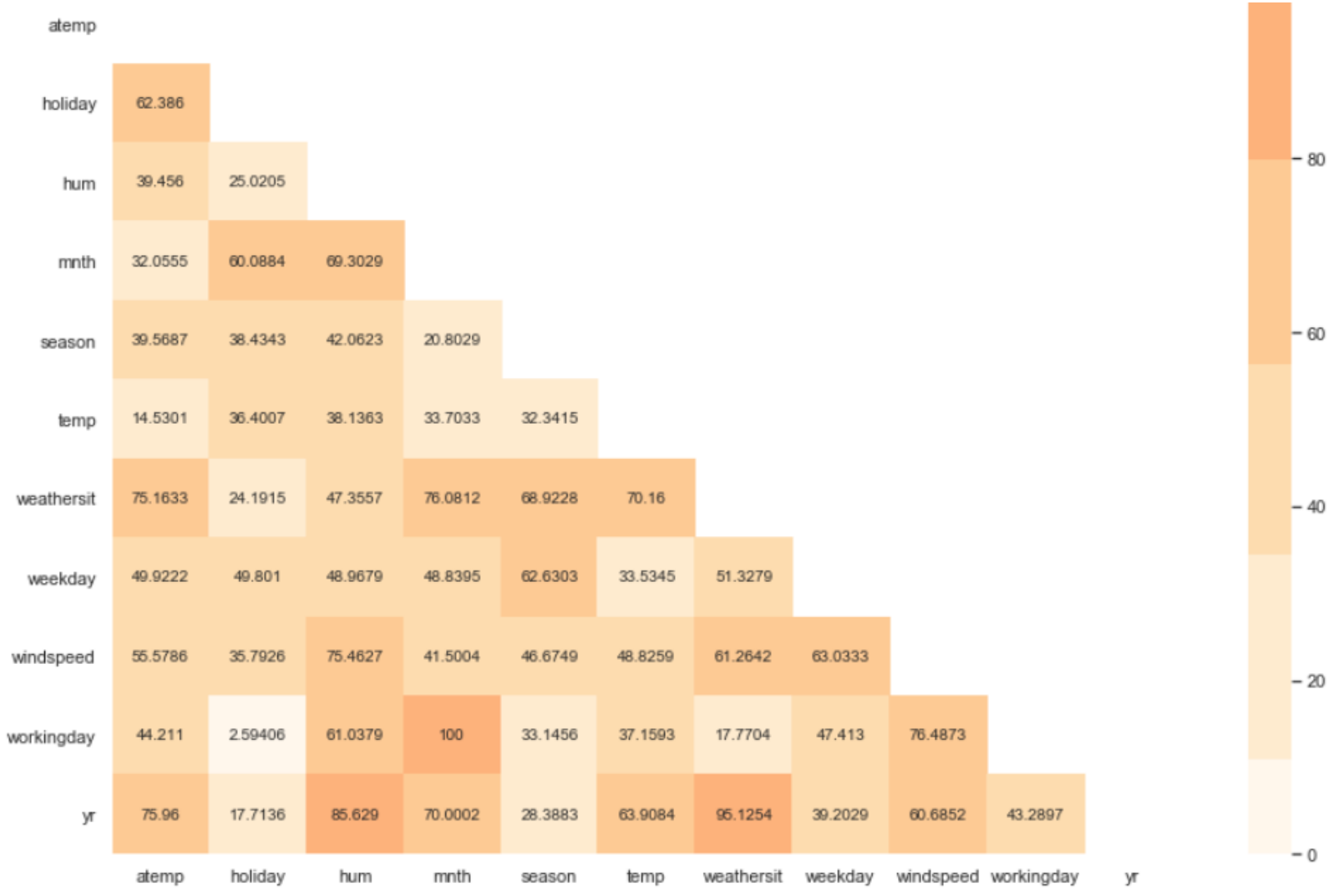
Single case explanation - graph



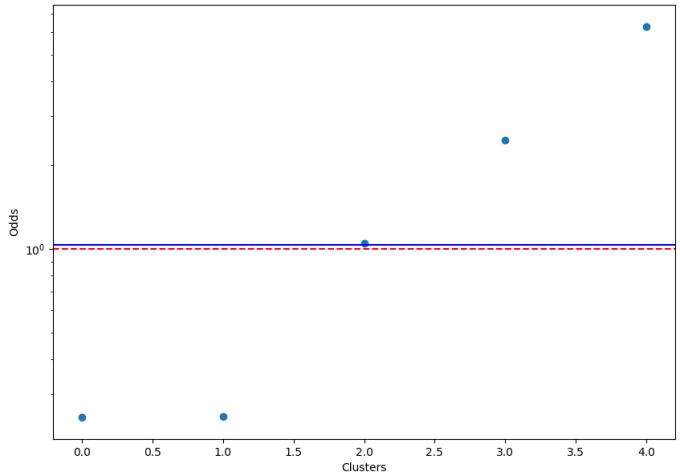
General explanation - pie chart



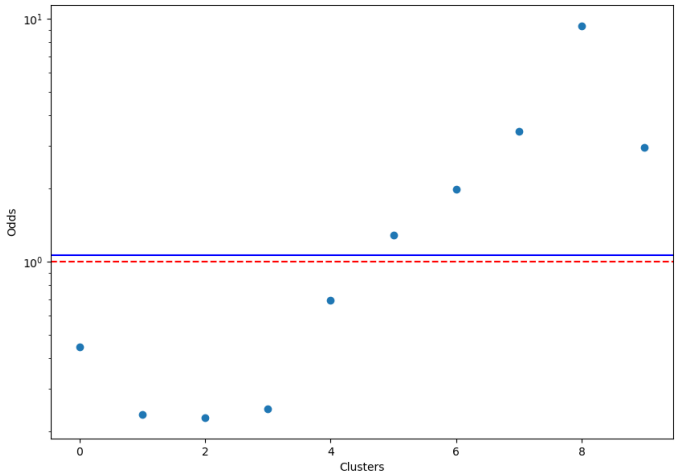
General explanation - heatmap



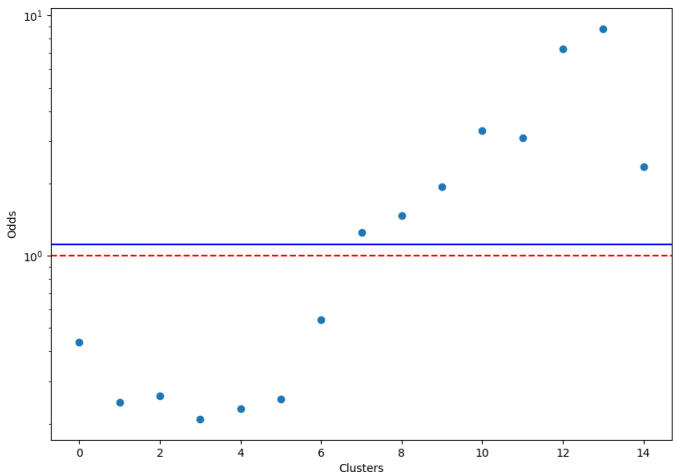
Feature analysis



(a) 5 clusters.



(b) 10 clusters.



(c) 15 clusters.

Conclusions and Future Work

Future Work

- Work on scalability and efficiency of the algorithm
- Verify explainability on some real life example with users
- Research some methods to obtain optimal number of clusters

The End

Thank you for your attention!
Any questions?

Powered by LaTeX