

## Instance-level explanations for fraud detection (poster)

**Citation for published version (APA):**

Collaris, D., van Wijk, J. J., & Vink, L. M. (2019). *Instance-level explanations for fraud detection (poster)*. Poster session presented at ICT Open 2019, Hilversum, Netherlands.

**Document status and date:**

Published: 19/03/2019

**Document Version:**

Accepted manuscript including changes made at the peer-review stage

**Please check the document version of this publication:**

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

**General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

[www.tue.nl/taverne](http://www.tue.nl/taverne)

**Take down policy**

If you believe that this document breaches copyright please contact us at:

[openaccess@tue.nl](mailto:openaccess@tue.nl)

providing details and we will investigate your claim.

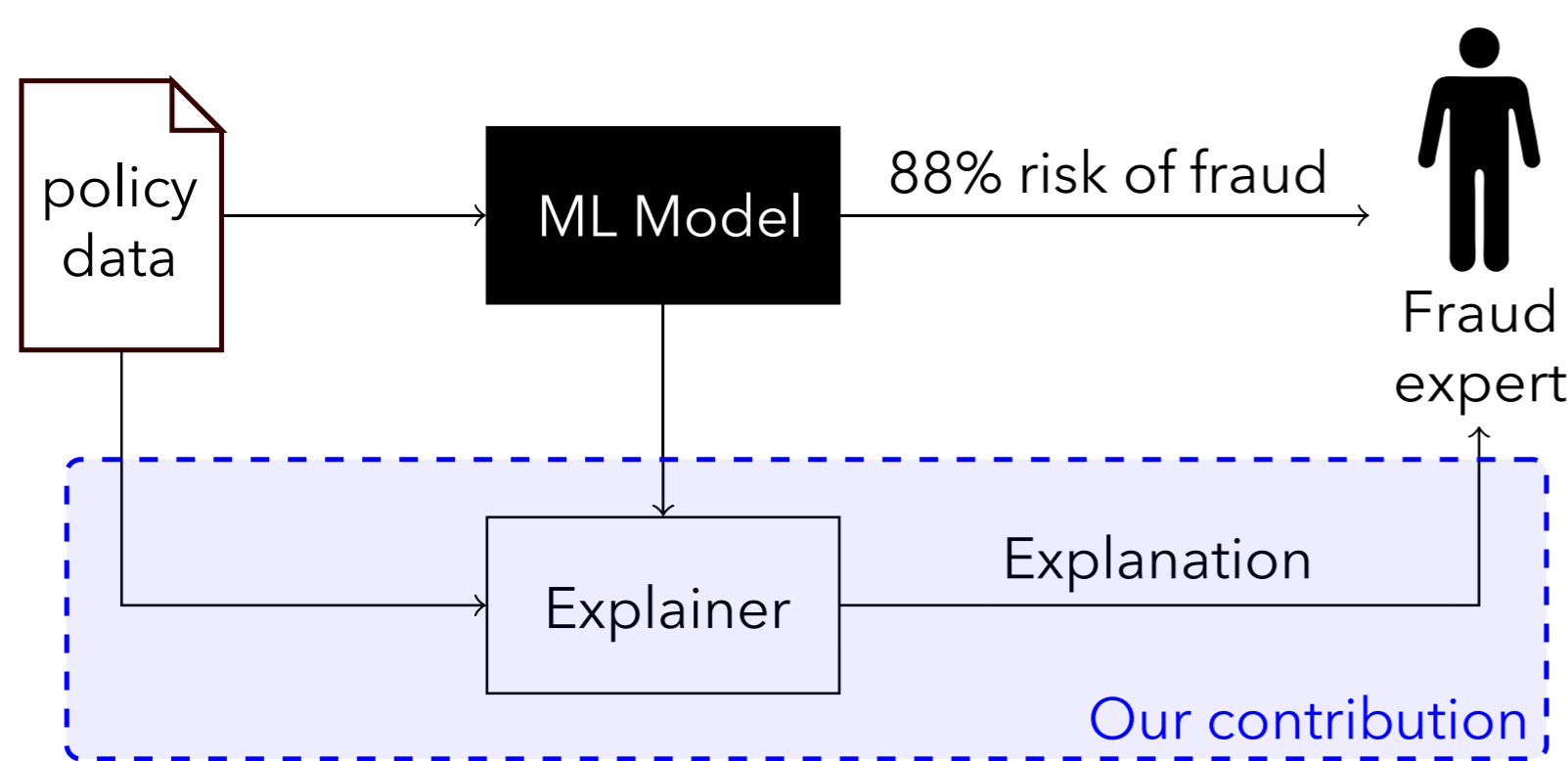
# INSTANCE-LEVEL EXPLANATIONS FOR FRAUD DETECTION

Dennis Collaris, Leo M. Vink, Jarke J. van Wijk



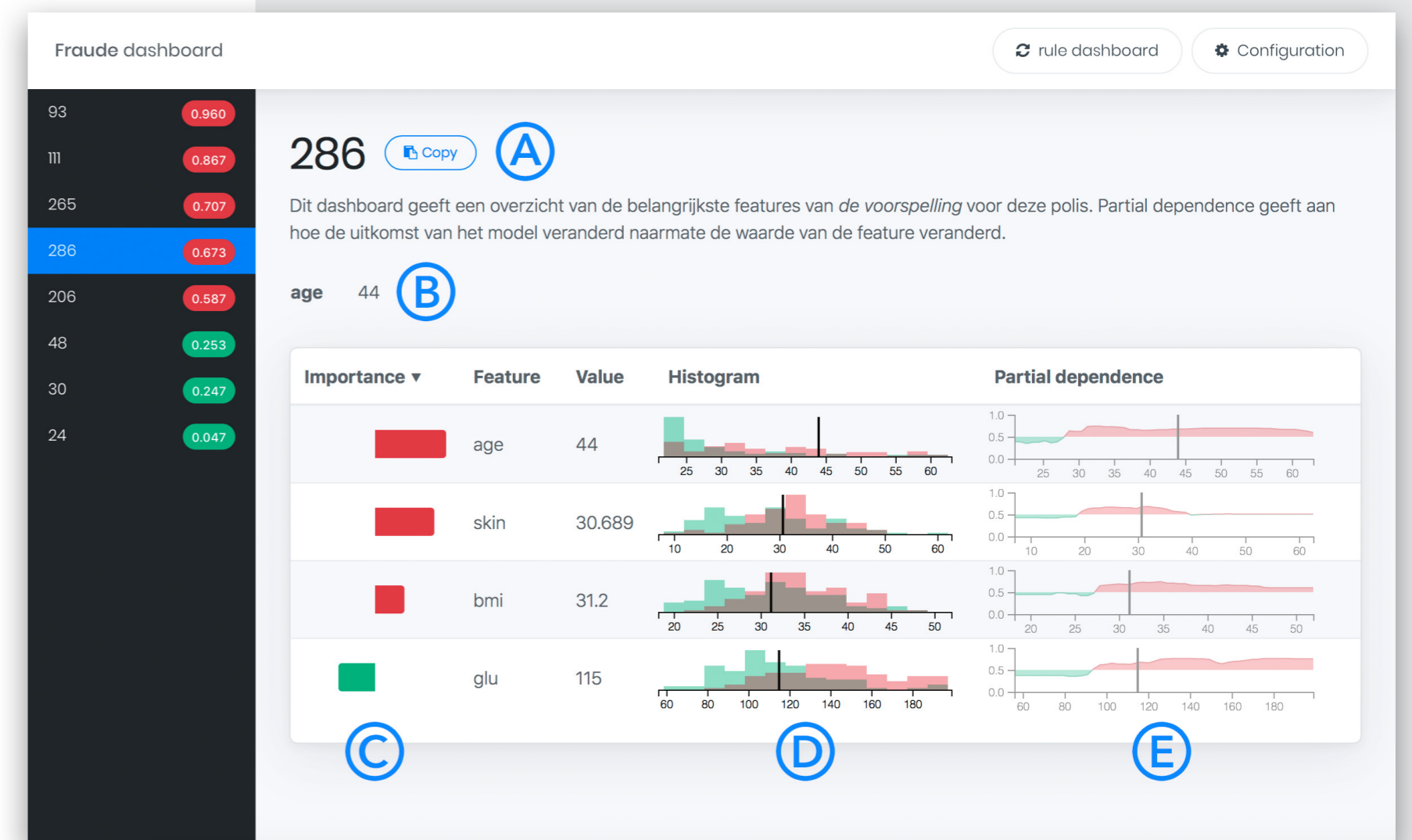
## THE PROBLEM

Fraud detection is a difficult problem that can benefit from predictive modeling. However, the verification of a prediction is challenging; for a single insurance policy, the model only provides a prediction score.



## THE SOLUTION

We designed two novel dashboards combining various state-of-the-art explanation techniques.



## FEATURE DASHBOARD

This dashboard shows bar charts (A) expressing the contribution of a feature to the prediction. Additionally, partial dependence plots (B) show the impact of changing the feature value indicated with a vertical line on the prediction.

### Feature contribution

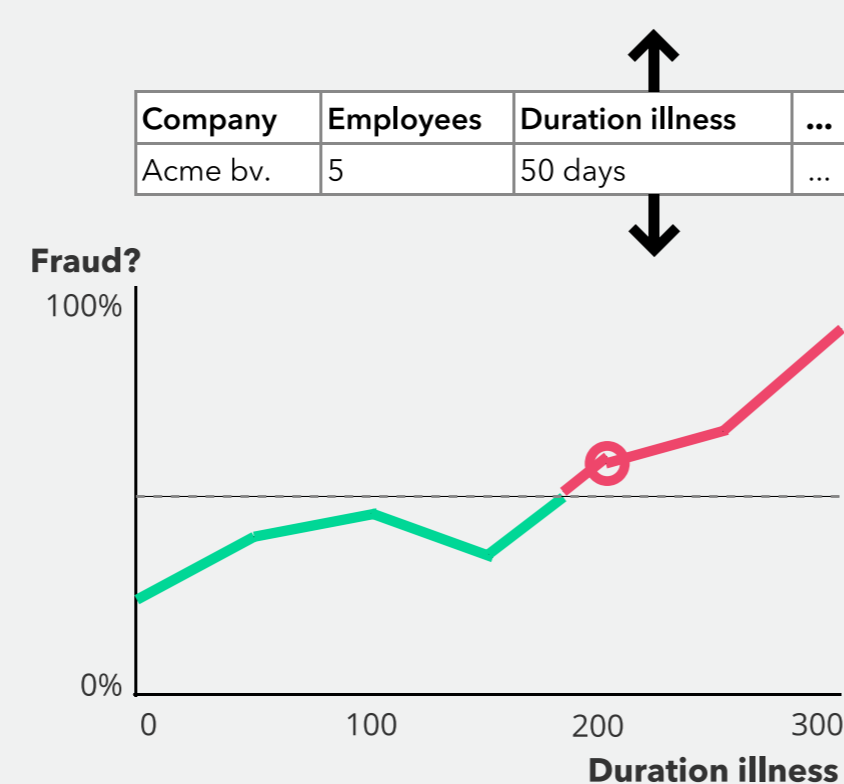
Local increment of feature  $f$ :

$$LI_f^c = \begin{cases} Y_{mean}^c - Y_{mean}^p, & \text{Parent splits on feature } f. \\ 0, & \text{Otherwise.} \end{cases}$$

Contribution of feature  $f$  in decision rule  $R$ :

$$FC_{i,t}^f = \sum_{N \in R_{i,t}} LI_f^N$$

### Partial dependence



### Local rule extraction

Synthetic pruning data set, uniform samples from an  $n$ -ball:

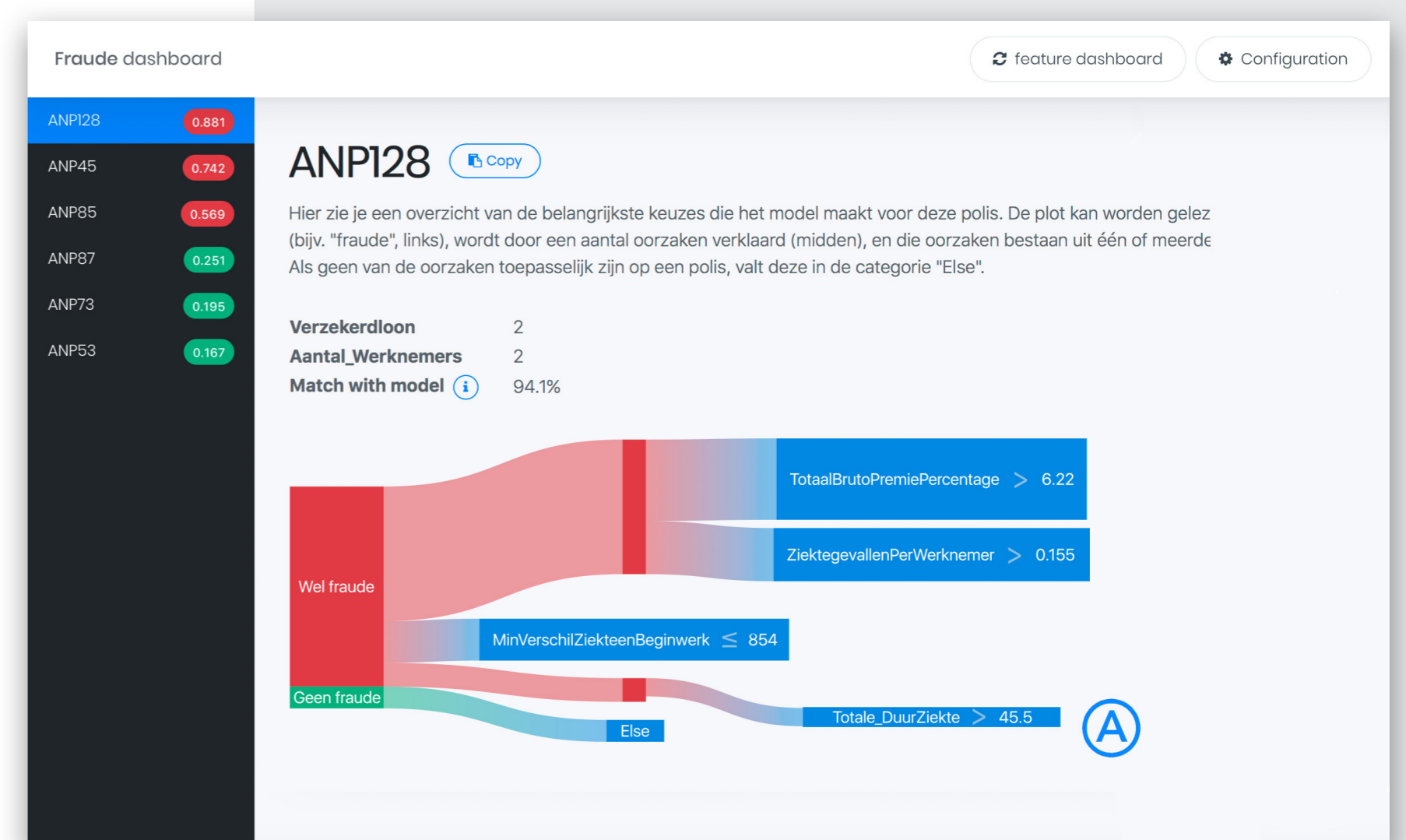
$$n\text{-ball uniform distribution} = \frac{Y * U^{\frac{1}{n}}}{\|Y\|}$$

with  $Y \sim N(0, 1)$  and  $U \sim U(0, 1)$

All decision rules applicable to instance  $i$  are extracted and pruned.

A Regularized Random Forest is trained on binary matrix of applicability of rules on the pruning dataset. Feature importance of that forest constitutes a metric of importance of individual decision rules.

For the example on the right, 1.300.000 rules are reduced to only 4, while still retaining 94.1% of the local fidelity of the reference model.



## RULE DASHBOARD

This dashboard shows a flow diagram (A) representation of locally extracted rules. Four rules are shown, and rule and feature importance is encoded by the width of the edges.

Check out the paper:  
arxiv.org/abs/1806.07129



Dennis Collaris  
d.a.c.collaris@tue.nl

Commit2Data