Achieving algorithmic transparency with Shapley Additive Explanations

QUANTUMBLACK
A MCKINSEY COMPANY

Contents:

- Trade-off between Explainability & Performance
- Shapley additive explanations
- SHAP: illustrative example
- XAI use case 1: clinical operations
- XAI use case 2: driver genome
- Future developments

Explainability vs Performance trade-off

Performance

- Provide highly accurate solutions for our clients

Explainability

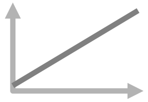
- Explain why our models perform as they do. How can we interpret the contribution of drivers in black-box models?
- How can we get drivers for an individual prediction?

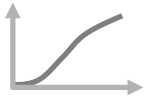
Natural trade-off between these two concepts

- Local interpretations of black-box models

Neural networks 

Random forests 

Linear regression 

Logistic regression 

Explainability ~~vs~~ Performance ~~trade-off~~ integration

Performance

- Provide highly accurate solutions for our clients

Explainability

- Explain why our models perform as they do. How can we interpret the contribution of drivers in black-box models?
- How can we get drivers for an individual prediction?

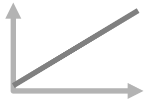
Natural trade-off between these two concepts

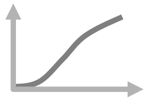
- Local interpretations of black-box models

Neural networks 

Random forests 

+

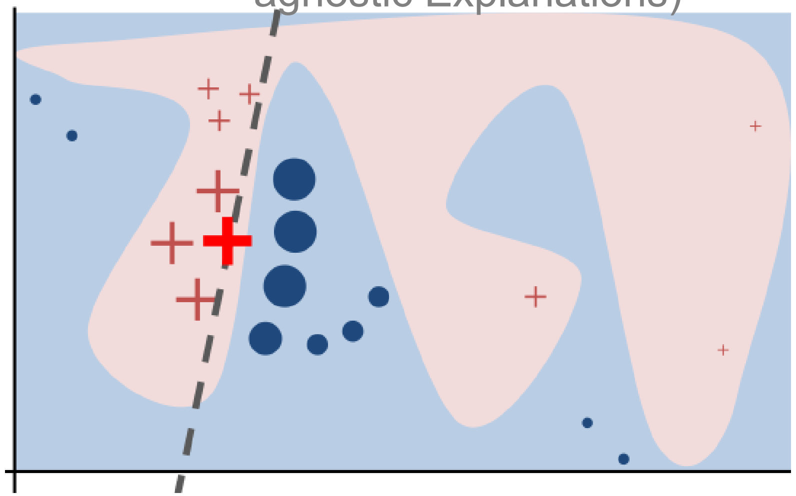
Linear regression 

Logistic regression 

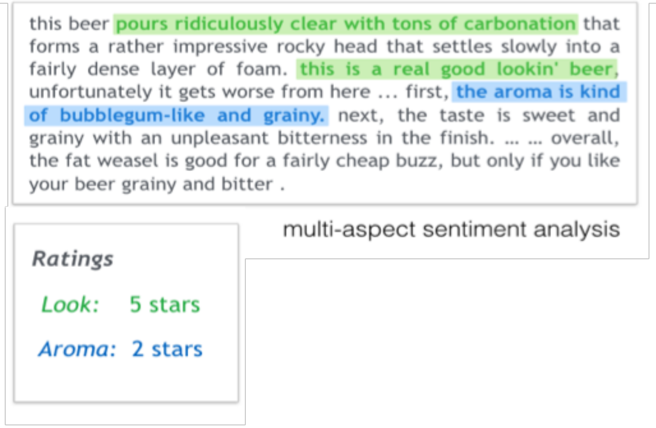
= ?

Methods for XAI

LIME (Locally Interpretable Model-agnostic Explanations)[1]



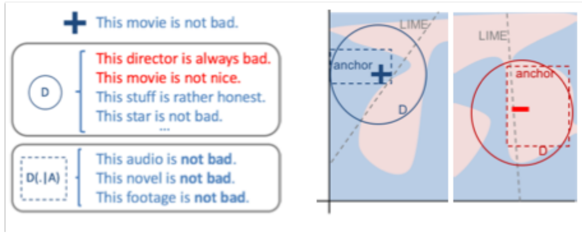
Rationalizing Neural Predictions[2]



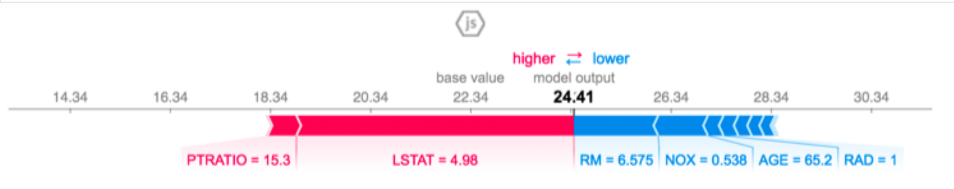
Bayesian Rule Lists[3]

If male and adult, then survival probability	21% (19%-23%)
else if 3rd class then survival probability	44% (38%-51%)
else if 1st class then survival probability	96% (92%-99%)
else survival probability	88% (82%-94%)

Anchors[4]



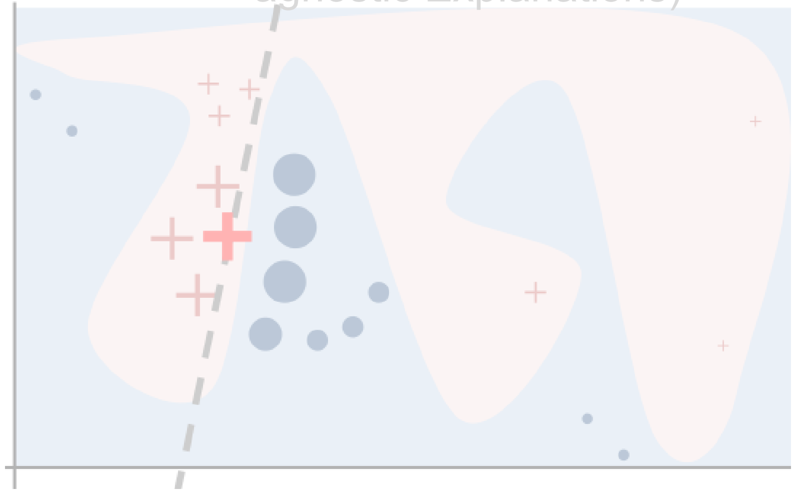
SHAP (Shapley Additive exPlanations)[5]



1. Ribeiro et al., "Why Should I Trust You?": Explaining the Predictions of Any Classifier, https://arxiv.org/abs/1602.04938
2. Lei et al., Rationalizing Neural Predictions, https://arxiv.org/abs/1602.04938
3. Letham et al., Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model, https://arxiv.org/abs/1511.01644
4. Lundberg and Lee, A unified approach to interpreting model predictions, http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions
5. Ribeiro et al., Anchors: High-Precision Model-Agnostic Explanations, https://homes.cs.washington.edu/~marcotcr/aaai18.pdf
6. All images taken from respective publications/github repos

Methods for XAI

LIME (Locally Interpretable Model-agnostic Explanations)[1]



Rationalizing Neural Predictions[2]

this beer pours ridiculously clear with tons of carbonation that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam. this is a real good lookin' beer, unfortunately it gets worse from here ... first, the aroma is kind of bubblegum-like and grainy. next, the taste is sweet and grainy with an unpleasant bitterness in the finish. ... ... overall, the fat weasel is good for a fairly cheap buzz, but only if you like your beer grainy and bitter .

multi-aspect sentiment analysis

Ratings

Look: 5 stars

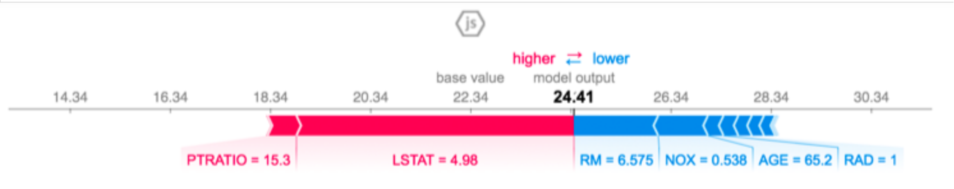
Aroma: 2 stars

Bayesian Rule Lists[3]

If male and adult, then survival probability	21% (19%-23%)
else if 3rd class then survival probability	44% (38%-51%)
else if 1st class then survival probability	96% (92%-99%)
else survival probability	88% (82%-94%)

Anchors[4]

SHAP (Shapley Additive exPlanations)[5]



- Trade-off between Explainability & Performance
- Shapley additive explanations
- SHAP: illustrative example
- XAI use case 1: clinical operations
- XAI use case 2: driver genome
- Future developments

SHAP (**Sh**apley **A**dditive Ex**p**lanations)

Explanation model[1]**:**

$$g(x') = \varphi_0 + \sum_{i=1}^{M} \varphi_i x'_i$$

(class of additive feature attribution models)

Shapley regression values[2], $\varphi_i \in \mathbb{R}$

- unified measure of additive feature attributions

$$\varphi_i = \sum_{S \in F \setminus \{i\}} \frac{|S|!\,(M - |S| - 1)!}{M!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_s) \right]$$

3 properties

Local accuracy: require the output of the local explanation model $g(x')$ to match the original model $f(x)$, where x' is the simplified input, i.e:

$$f(x) = g(x')$$

Missingness: require features missing in the original input to have no attributed impact, i.e.

$$x'_i = 0 \Longrightarrow \varphi_i = 0$$

Consistency: stipulates $\varphi_i(f', x) \geq \varphi_i(f, x)$ for any two models f and f' if the feature's contribution in $f' \geq$ the feature's contribution in f.

1. Lundberg and Lee (2017) „A unified approach to interpreting model predictions", https://arxiv.org/abs/1705.07874
2. Lloyd S Shapley (1953) "A value for n-person games", In: *Contributions to the Theory of Games,* 2:28, pp.307-317

Computing SHAP values

SHAP values - unified measure of additive feature attributions, $\varphi_i \in \mathbb{R}$:

$$\varphi_i = \boxed{\sum_{S \in F \setminus \{i\}} \frac{|S|!\,(M - |S| - 1)!}{M!}} \left[\underbrace{f_{S \cup \{i\}}(x_{S \cup \{i\}})}_{\text{output with } i^{\text{th}} \text{ feature}} - \underbrace{f_S(x_s)}_{\text{output without } i^{\text{th}} \text{ feature}}\right]$$

weighted average of all possible subsets of S in F

where

F = {all input features}

S = {subset of input features}

M = |F| = number of input features

Computing SHAP values:

- $f_{S \cup \{i\}}$ is trained with the i^{th} feature present
- f_S is trained without the i^{th} feature
- compute difference $f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_s)$ for the current input
- retrain the model on all feature subsets $S \in F \setminus \{i\}$
- take weighted average of all possible differences

SHAP in practice

Implementations:

Kernel SHAP[1] = LIME + Shapley Values

where loss function L, weighting kernel π, and regularisation term Ω are computed so that LIME meets Shapley properties:

Theorem 2 (Shapley kernel) *Under Definition 1, the specific forms of $\pi_{x'}$, L, and Ω that make solutions of Equation 2 consistent with Properties 1 through 3 are:*

$$\Omega(g) = 0,$$

$$\pi_{x'}(z') = \frac{(M-1)}{(M \text{ choose } |z'|)|z'|(M-|z'|)},$$

$$L(f, g, \pi_{x'}) = \sum_{z' \in Z} \left[f(h_x^{-1}(z')) - g(z')\right]^2 \pi_{x'}(z'),$$

where $|z'|$ is the number of non-zero elements in z'.

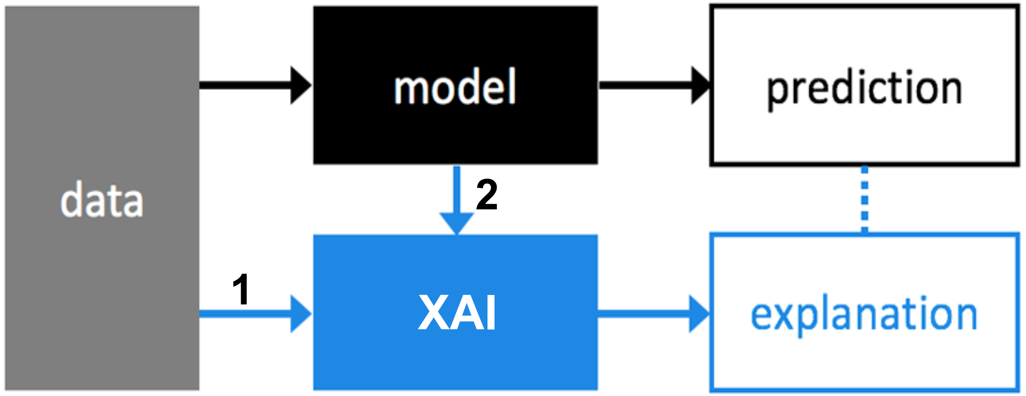
Deep SHAP [1] = DeepLIFT + Shapley Values

Linear SHAP[1], Low-Order SHAP[1], Max SHAP[1]

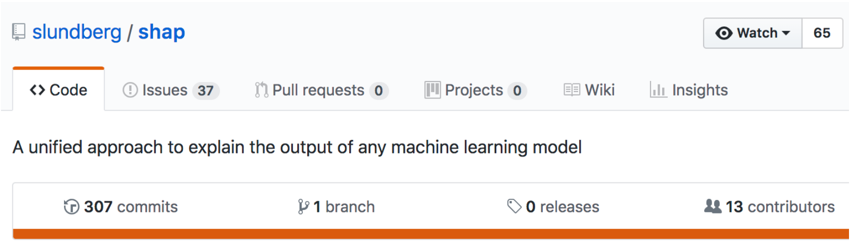
Tree SHAP[2]

- speed-up from $O(TL2^M)$ to $O(TLD^2)$

Integration




```
pip install shap
```



+ beautiful out-of-box JS visualisations

1. Lundberg and Lee (2017) „A unified approach to interpreting model predictions“, https://arxiv.org/abs/1705.07874
2. Lundberg, Erion, Lee (2018) “Consistent Individualized Feature Attribution for Tree Ensembles”, https://arxiv.org/abs/1802.03888

SHAP: illustrative example

Task: given demographic data, classify adult income groups

Dataset: US 1994 Census data[1]

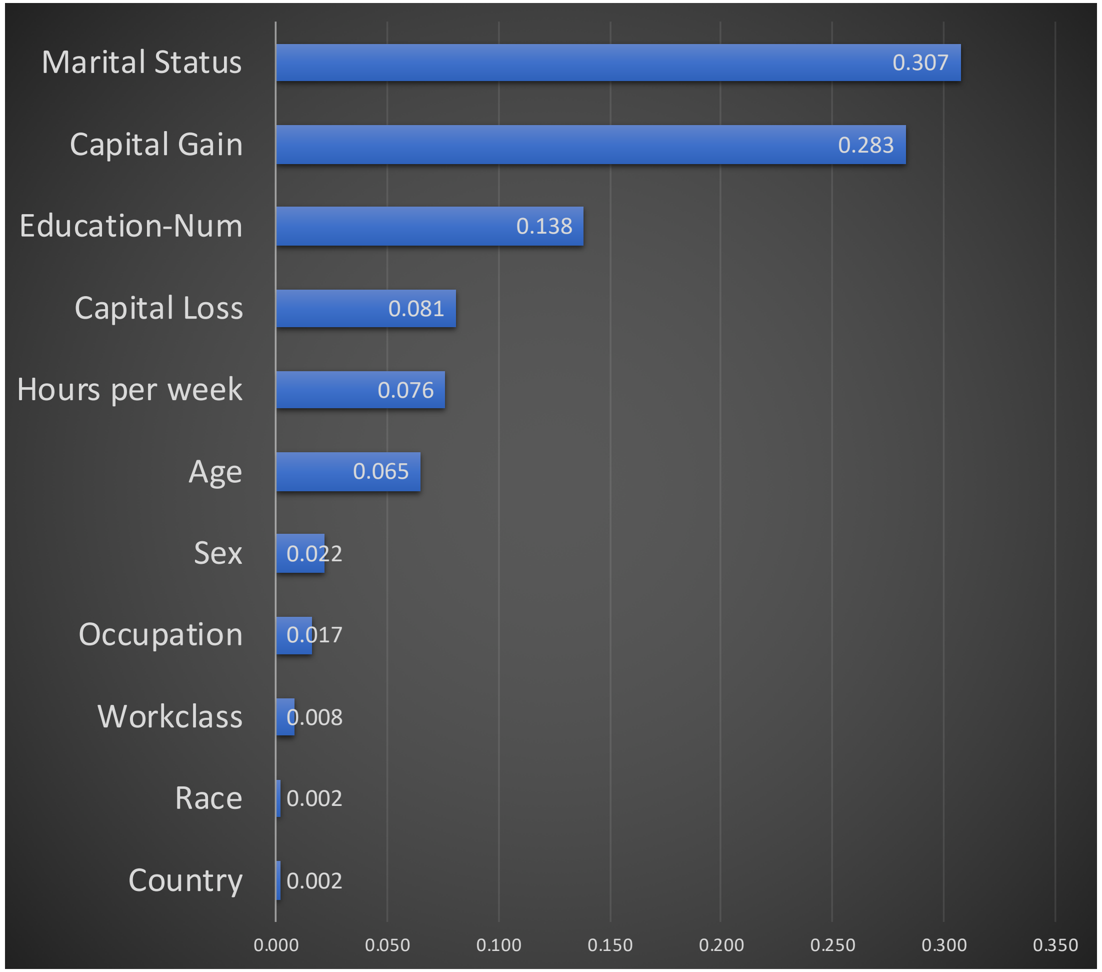
Base model: sklearn Random Forest classifier

Explanation model: TreeSHAP

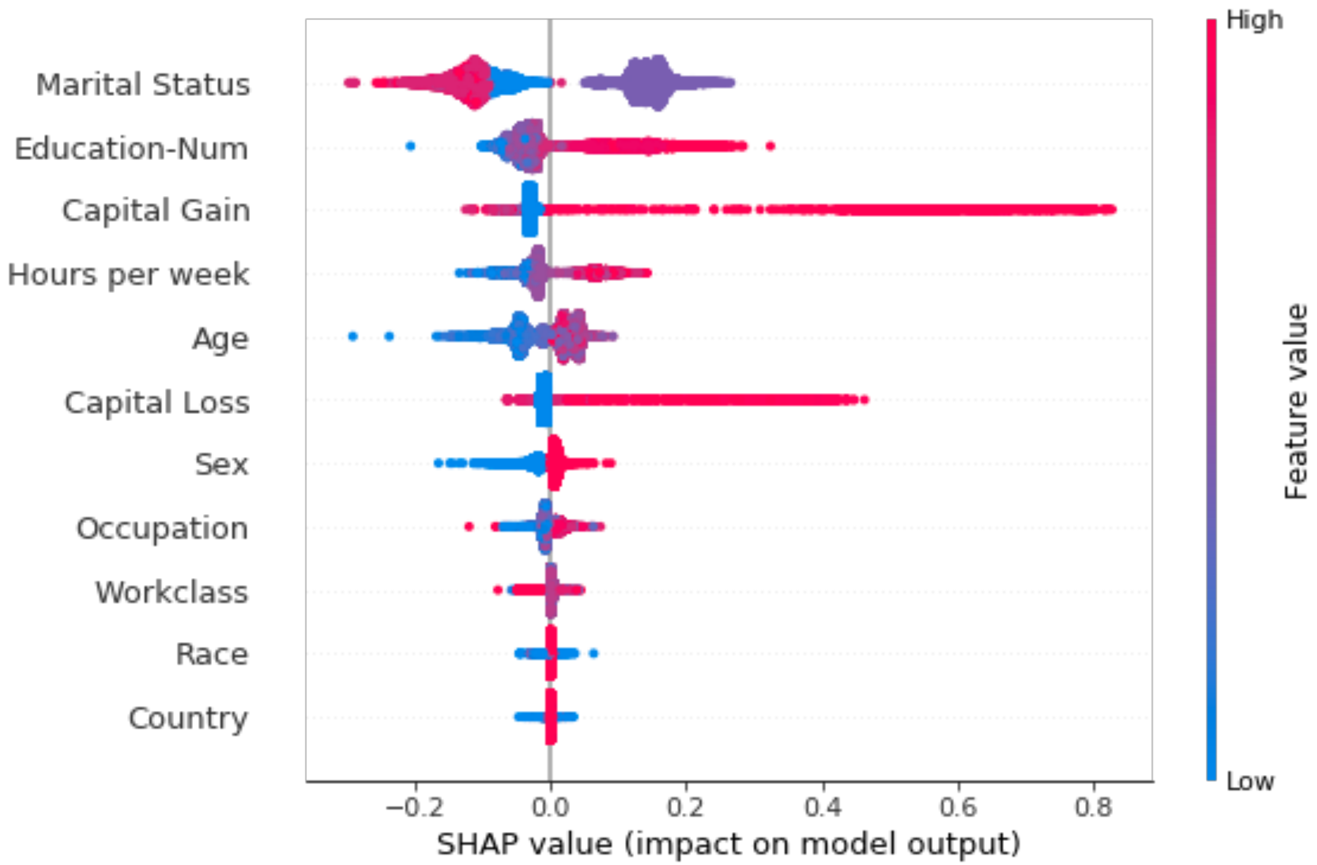
1. 1994 Census database, donated by B. Becker: https://archive.ics.uci.edu/ml/datasets/Adult

RF importance scores vs. SHAP's explanations

'Native' RF feature importance scores

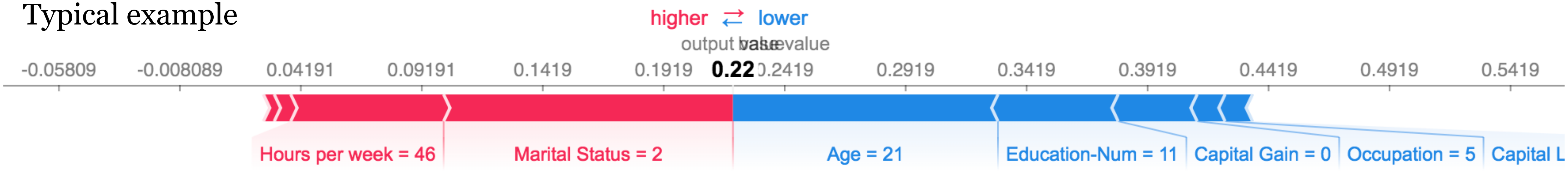


SHAP summary plot

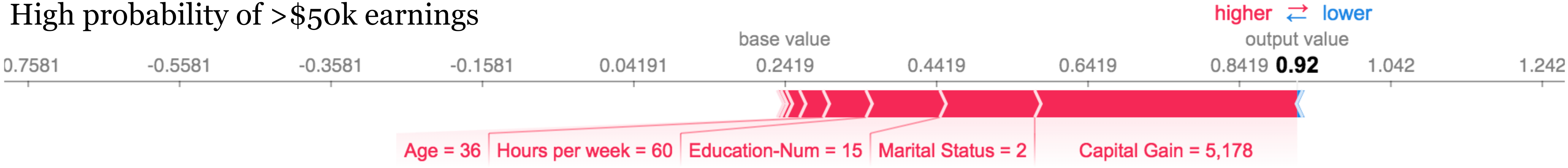


SHAP: individualised explanations

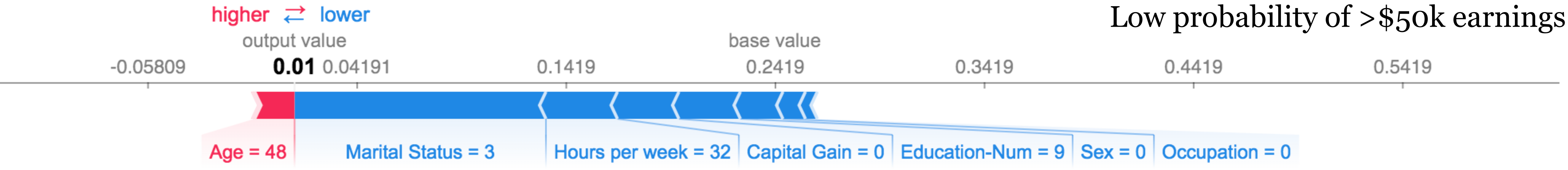
Typical example



High probability of >$50k earnings



Low probability of >$50k earnings



Legend to categorical value:
Sex: 0 = F, 1 = M | Marital status: 2 = never married, 3 = married, spouse absent
Occupation: 0 = Adm-clerical, 5 = Sales

SHAP: individualised explanations across the cohort

Contents:

Use case: Clinical Operations

Task: given a new drug trial, predict which hospitals will enroll more patients (high enrolment rate).

Two questions raised frequently by our client

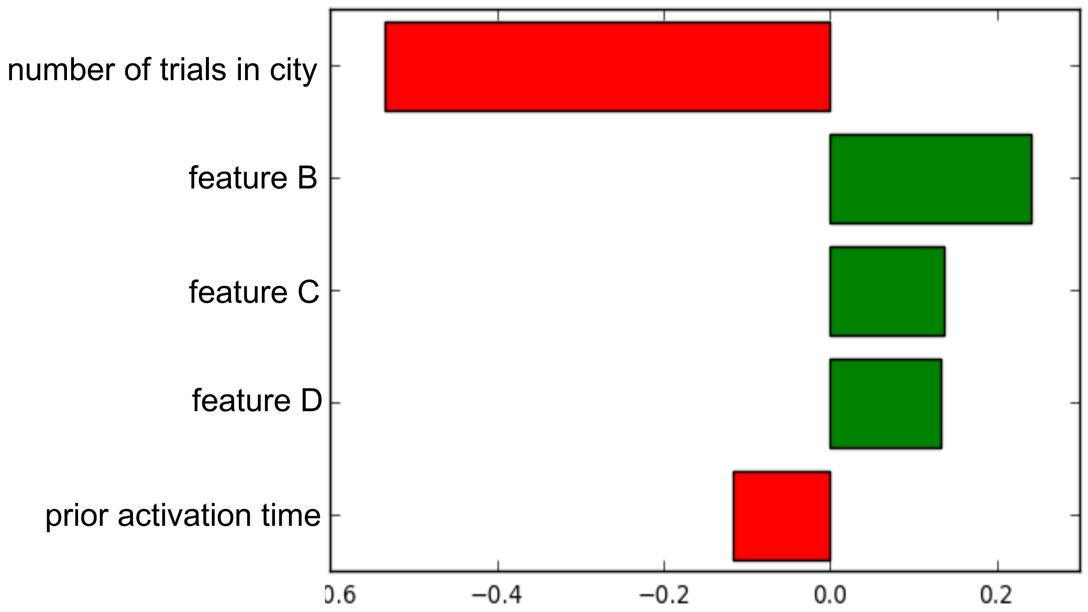
How can we interpret the predictions given by your black-box model? What drives the **direction** (high or low) of the predicted enrollment rate?

How can we get **personalised** explanations? Why hospital B enrolls more patients than hospital A?

Local explanations (Enrolment Rate)

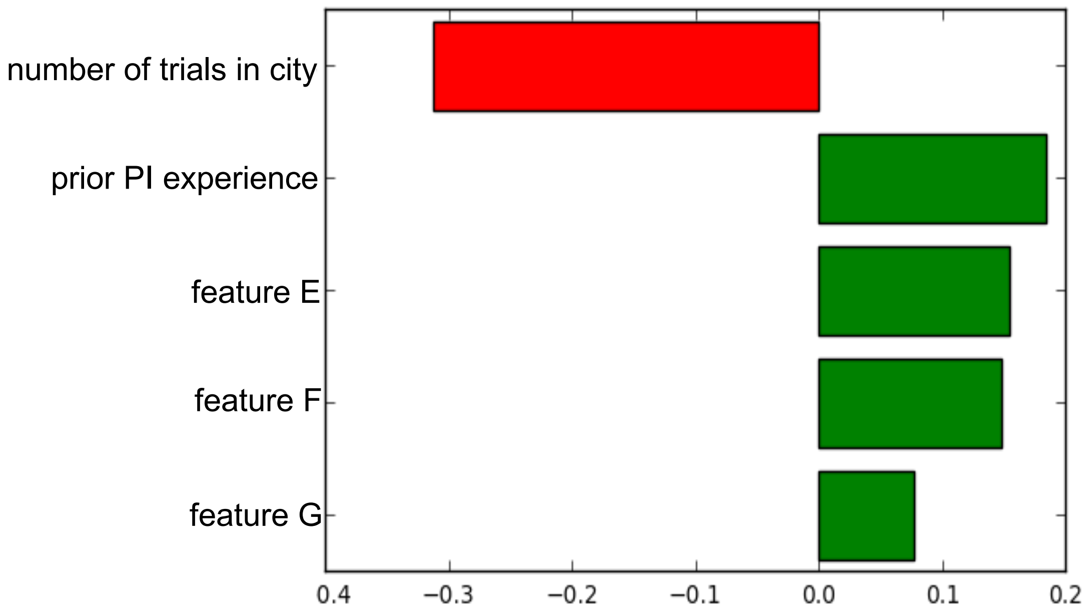
Hospital A

Local prediction = 2.07 pt/month
Black-box prediction = 2.19 pt/month



Hospital B

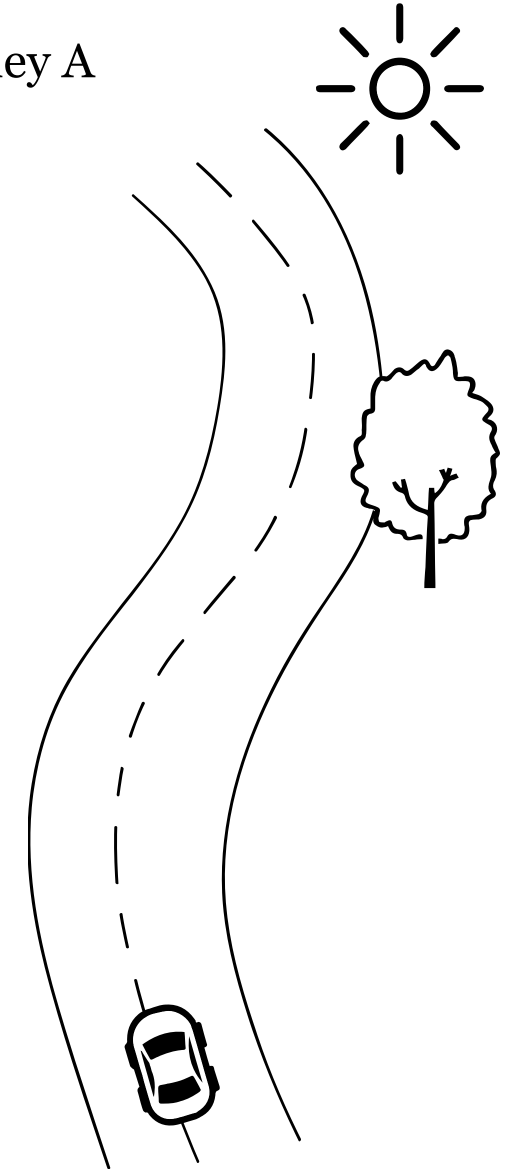
Local prediction = 4.96 pt/month
Black-box prediction = 4.99 pt/month



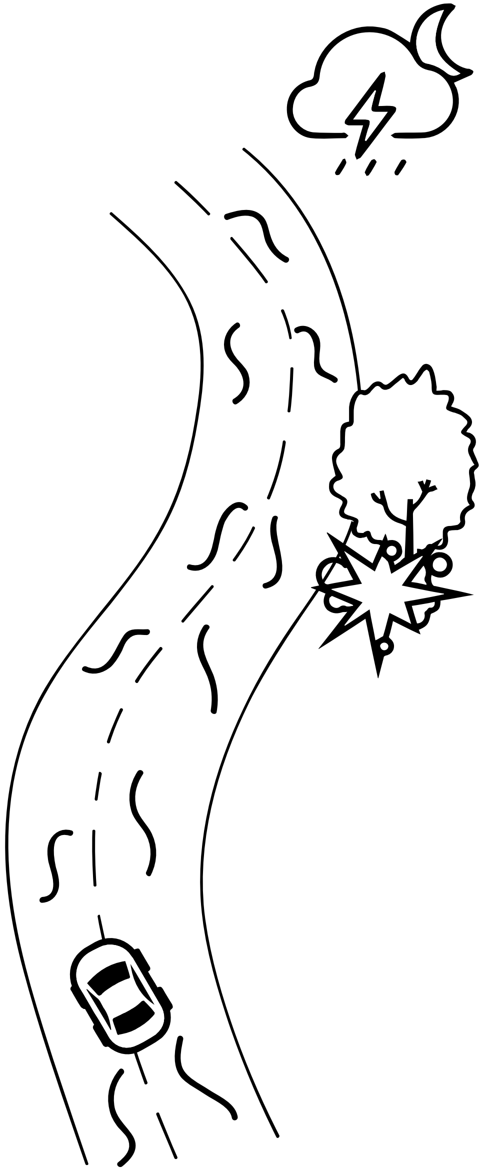
Contents:

Use case: Driving Safety

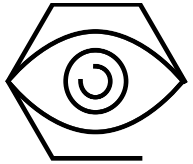
Journey A



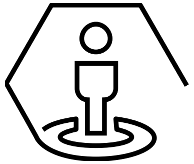
Journey B



Two questions raised frequently by our client



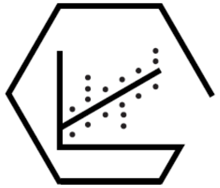
How can we interpret the predictions given by your black-box model? What drives the **direction** (high or low) of the predicted probability of an accident?



Drivers can use **personalised** explanations to improve their driving behaviour. Also, in case of a change in their insurance premium, they have the right to know what is the cause

The proposed solution uses Logistic Regression and Random Forest plus LIME for interpretability

~1500 features



~100 features

Feature engineering

Creation of single and multi-dimensional features based on hypotheses tree

Logistic regression

Use of elastic net regularization to prevent overfitting to the training set

Random Forest

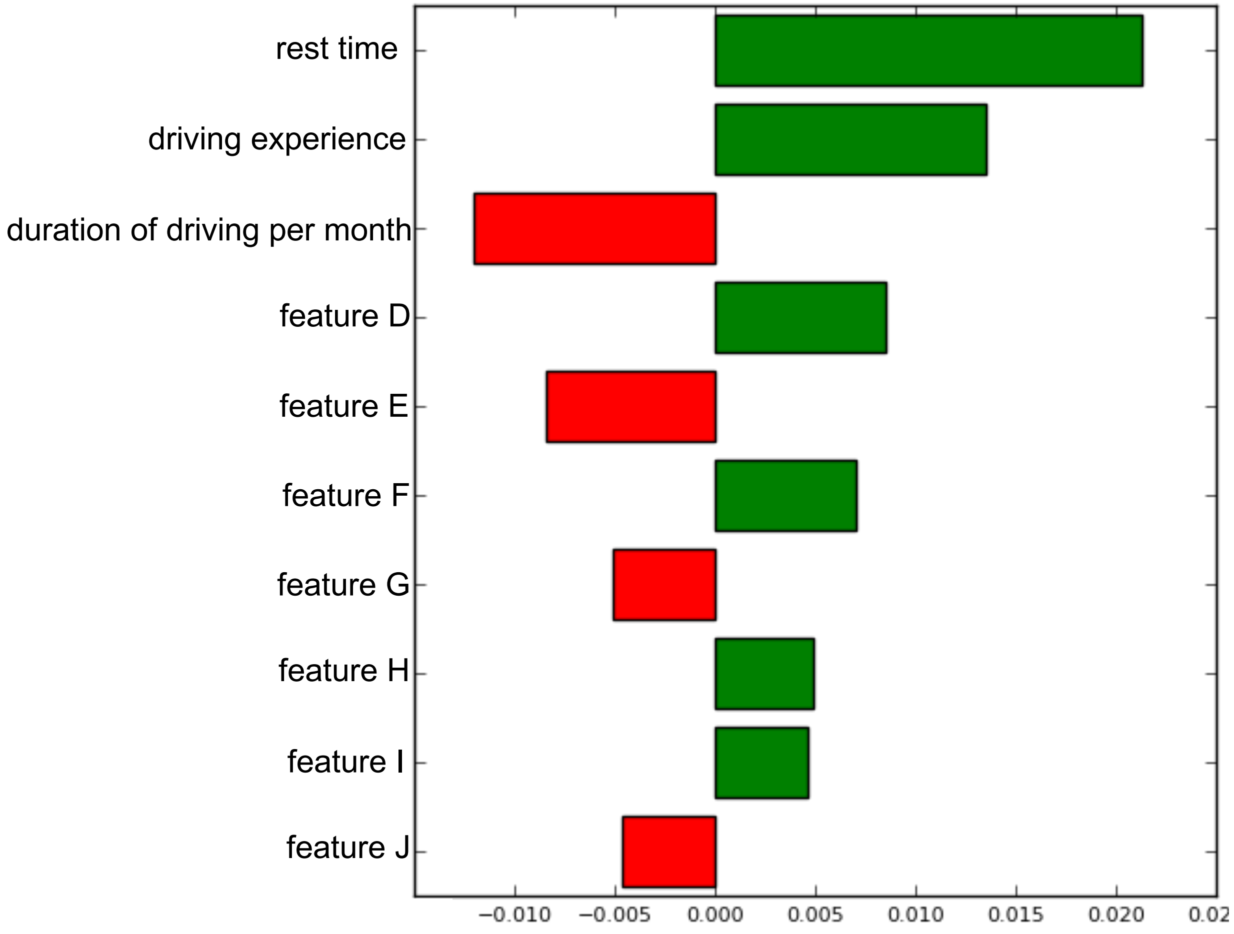
A random forest can discover complex non-linear relationships between features

LIME/SHAP

Personalized interpretations of the RF results

SOURCE: Team analysis

Example of a driver profile



Contents:

Future developments

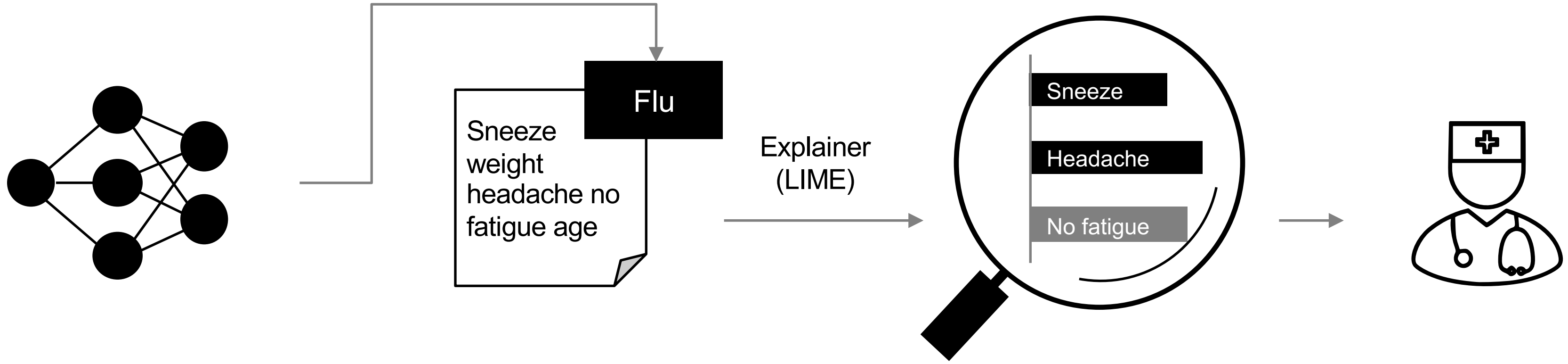
Model agnostic: **LIME, Kernel SHAP <- drive development & community interest**

Model-specific: **DeepLIFT (NNs), TreeSHAP (XGBoost, RFs) <- drive implementation**

SOURCE: Team analysis

Individualised explanations ≠ Transparency

Explain one instance using explanation model



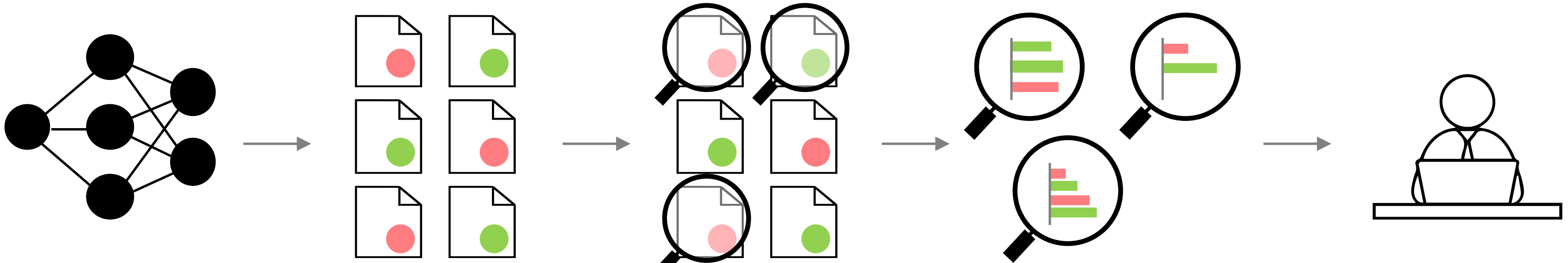
Model

Data and prediction

Explanation

Human makes decision

Pick a number of representable examples from a dataset



Model

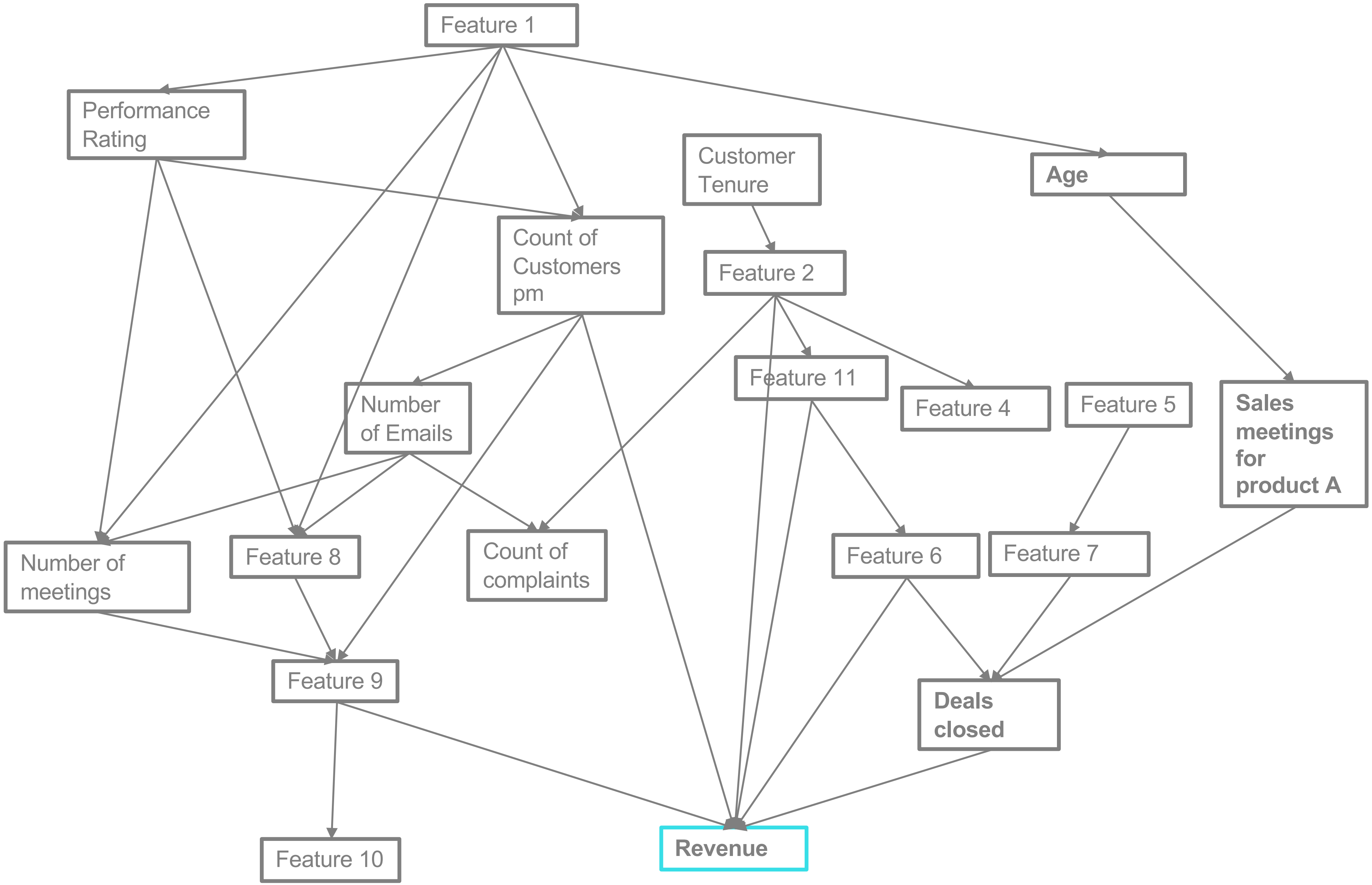
Dataset and prediction

Pick step

Explanations

Human makes decision

Explanation model ≠ causality



SOURCE: Team analysis

Questions?

