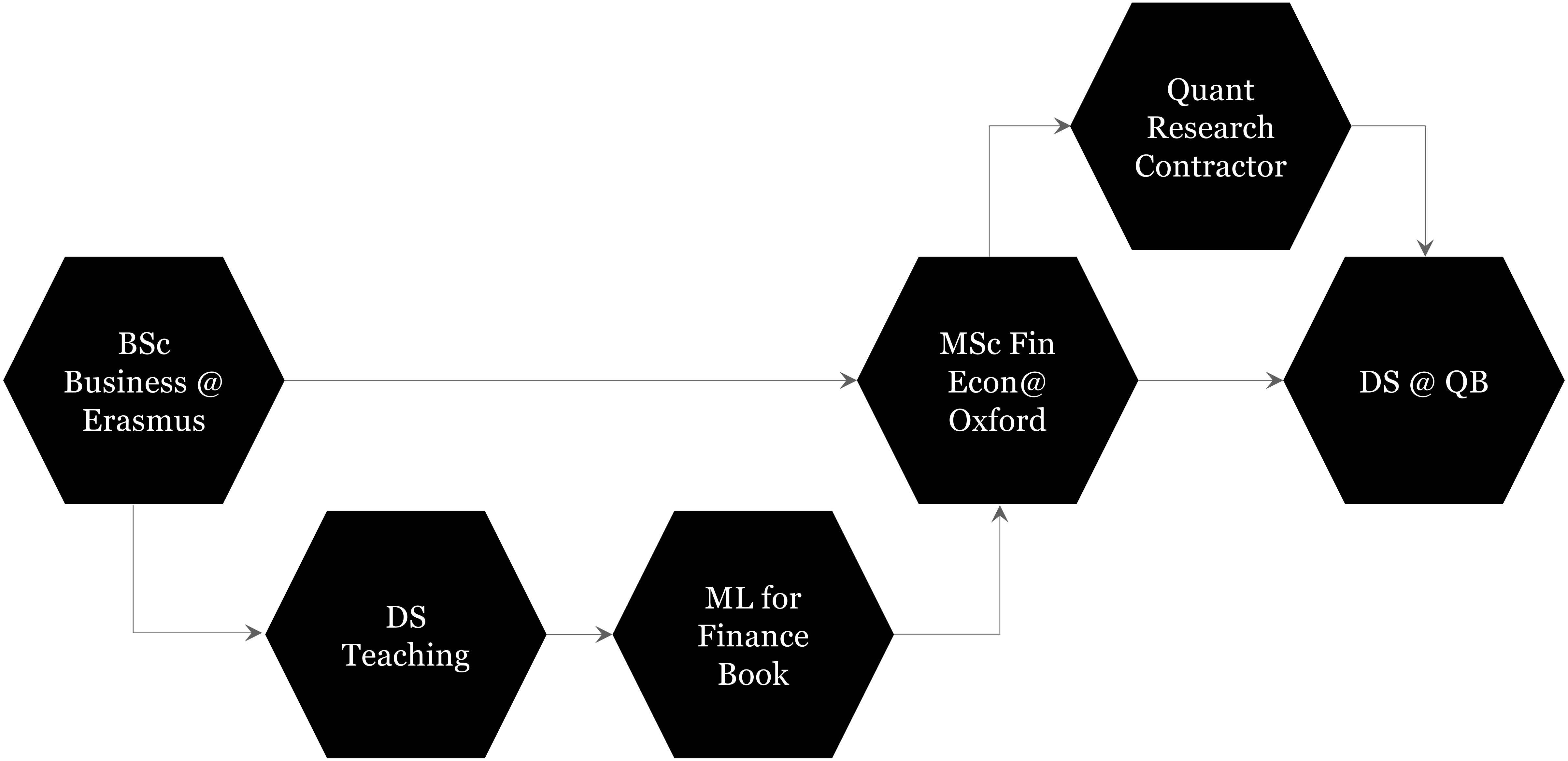
XAI in the Wild

Jannes Klaas – Imperial College 2020

QUANTUMBLACK
A MCKINSEY COMPANY

About Me



BSc Business @ Erasmus

DS Teaching

ML for Finance Book

MSc Fin Econ@ Oxford

Quant Research Contractor

DS @ QB

We exploit data, analytics and design to help our clients be the best they can be

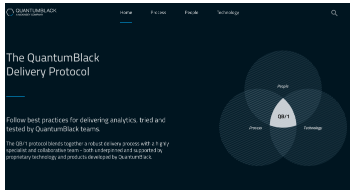
We were born and proven in Formula One, where the smallest margins are the difference between winning and losing and data has emerged as a fundamental element of competitive advantage

Tech: We build horizontal assets to improve end-to-end model lifecycle development and management

QB/Protocols

Protocols for delivering Analytics transformations

A codification of QB's best practices for delivering AI @ scale provided by technical practitioners that continues to evolve based on new learnings and experiences.



The QuantumBlack Delivery Protocol

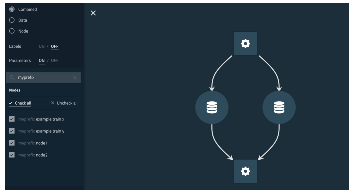
Follow best practices for delivering analytics, tried and tested by QuantumBlack teams.

The QB/1 protocol blends together a robust delivery process with a highly specialist and collaborative team - both underpinned and supported by proprietary technology and products developed by QuantumBlack.

Kedro

Data and analytics pipeline development framework

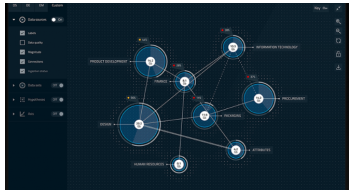
Kedro is a workflow development tool that helps you build data pipelines that are robust, scalable, deployable, reproducible and versioned.



Studio

Data management application

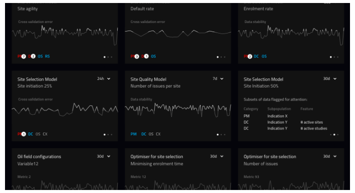
An application for capturing project knowledge, hypotheses, data source metadata, data landscapes, data ingestion status, and data quality metrics.



PerformanceAI

Tech tooling and playbooks for operational AI

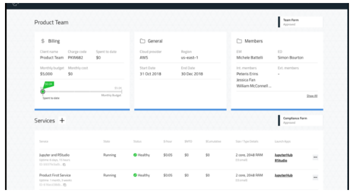
A product for sustaining model performance in a live environment and frameworks to accurately scope and consistently deliver models into production.



Platform McKinsey

Self-service cloud infra for analytics engagements

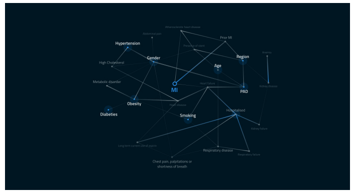
Automated provisioning of templated analytics environments with a single management layer for all security policies and monitoring, logging, audit.



CausalNex

Python library to help establish causality

A Python library that leverages Bayesian Networks to help data scientists to infer causation rather than observing correlation.



Note: We recently open sourced Kedro & CausalNex – video available here: https://youtu.be/KEdmJ2ADy_M

Contents:

- XAI use case 1: clinical operations
- XAI use case 2: driver safety
- Issues with Post-Hoc Explainers
- GA2M
- Causal Inference with CausalNex

Explainability vs Performance trade-off

Performance

- Provide highly accurate solutions for our clients

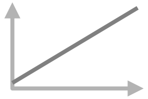
Neural networks

Random forests

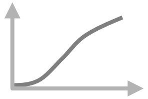
Explainability

- 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?

Linear regression



Logistic regression



Natural trade-off between these two concepts

- Local interpretations of black-box models

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

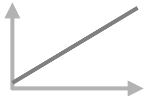
- Local interpretations of black-box models

Neural networks

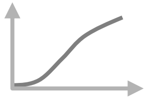
Random forests

+

Linear regression



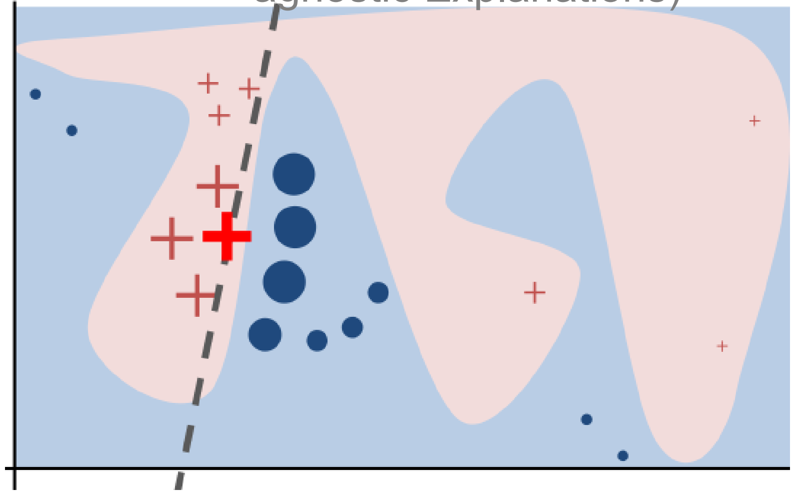
Logistic regression



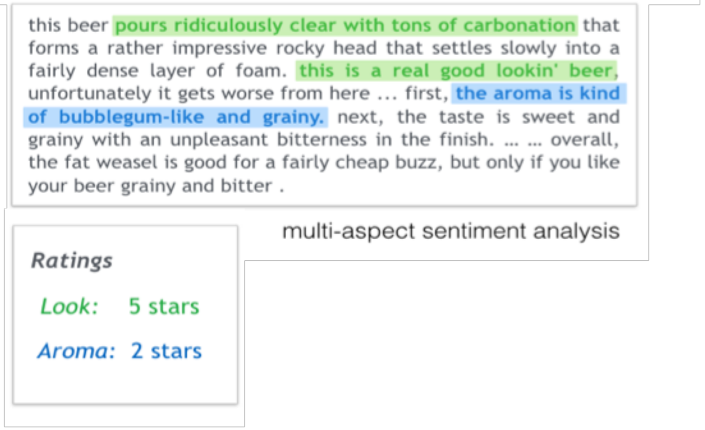
= ?

Methods for XAI

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



Rationalizing Neural Predictions[2]



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 .

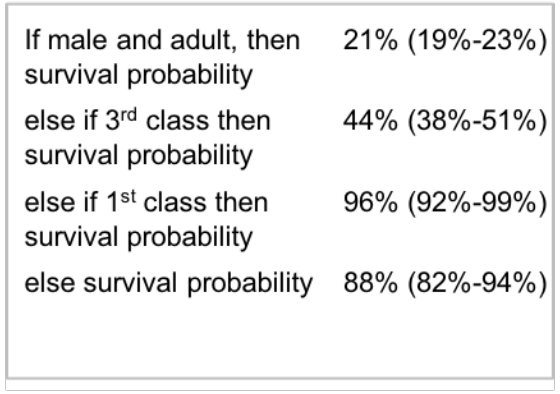
Ratings

Look: 5 stars

Aroma: 2 stars

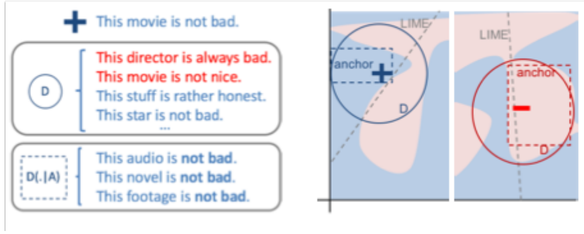
multi-aspect sentiment analysis

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]

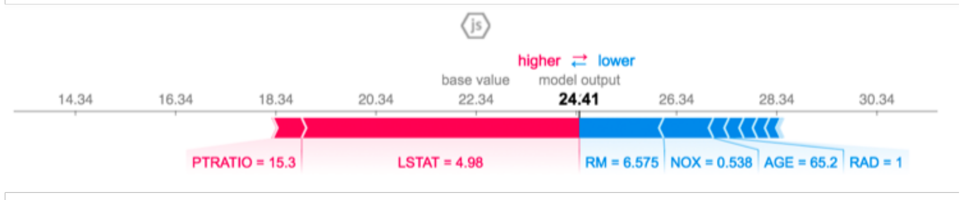


This movie is not bad.

This director is always bad.
This movie is not nice.
This stuff is rather honest.
This star is not bad.
...

This audio is not bad.
This novel is not bad.
This footage is not bad.

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

Use Case 1: Clinical Safety

01

Use case: Clinical Operations

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

Hospital A
Hospital B

Two questions raised frequently by our client

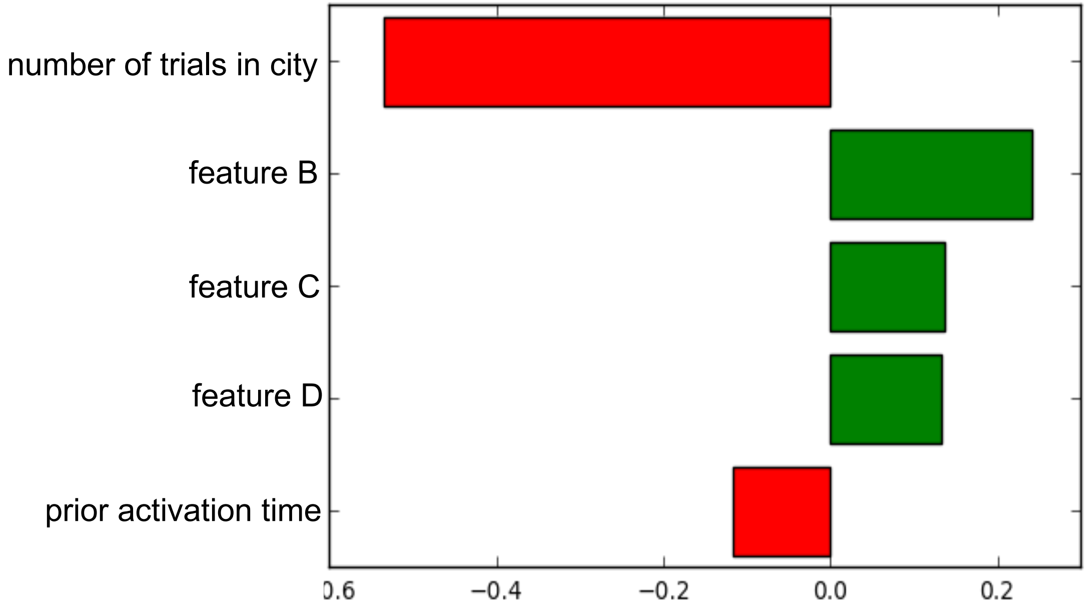
How can we interpret the predictions given by your black-box model? What drives the **direction** (high or low) of the predicted enrollment rate?

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

Local explanations (Enrolment Rate)

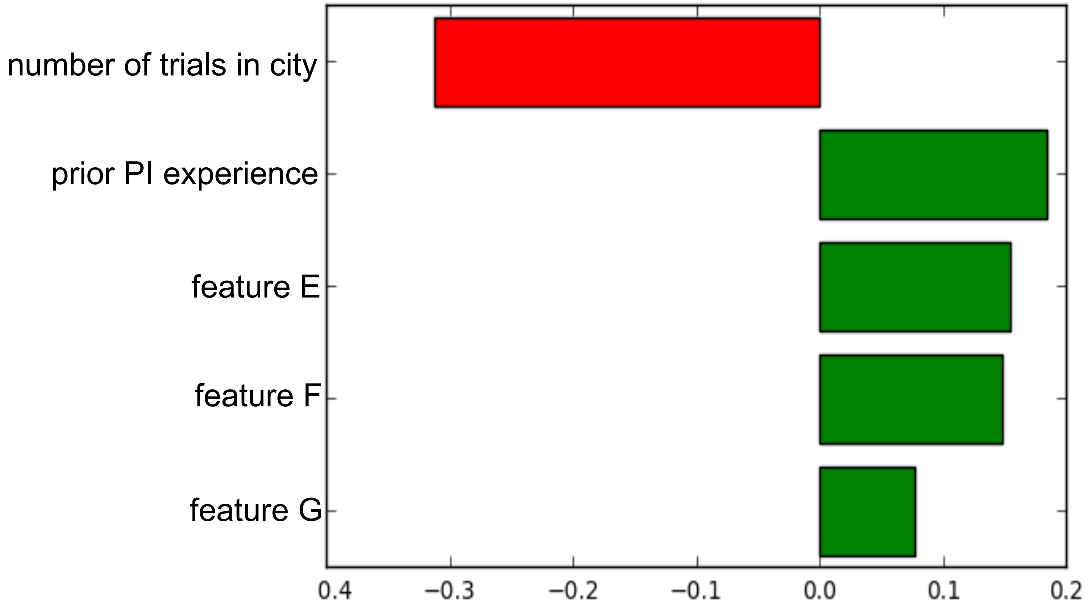
Hospital A

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

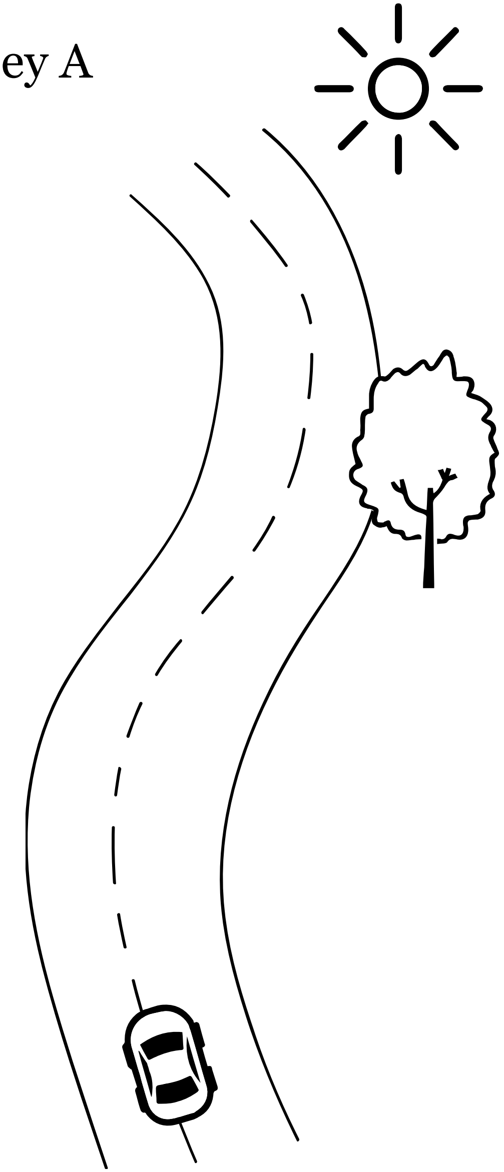


Use Case 2: Driver Safety

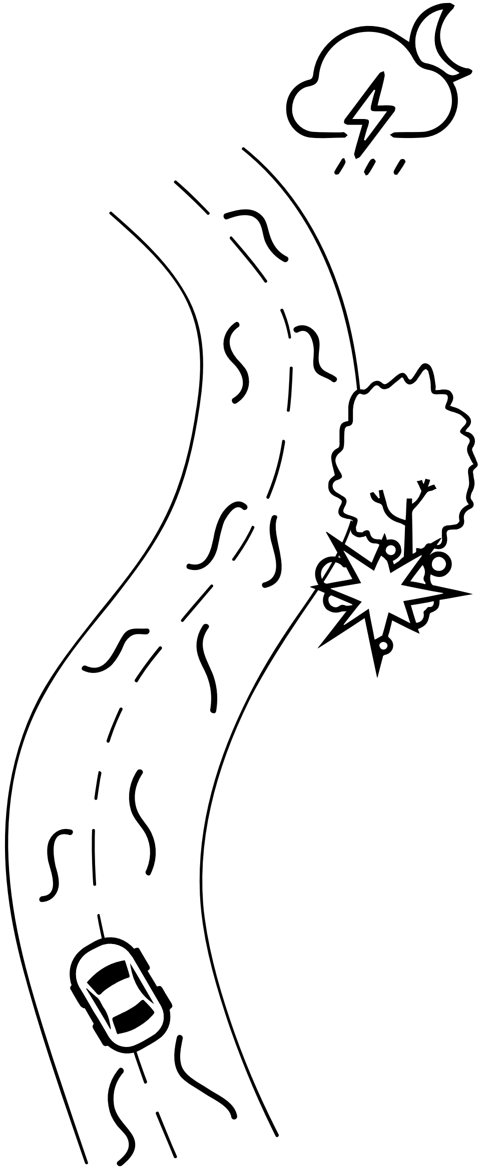
02

Use case: Driving Safety

Journey A



Journey B

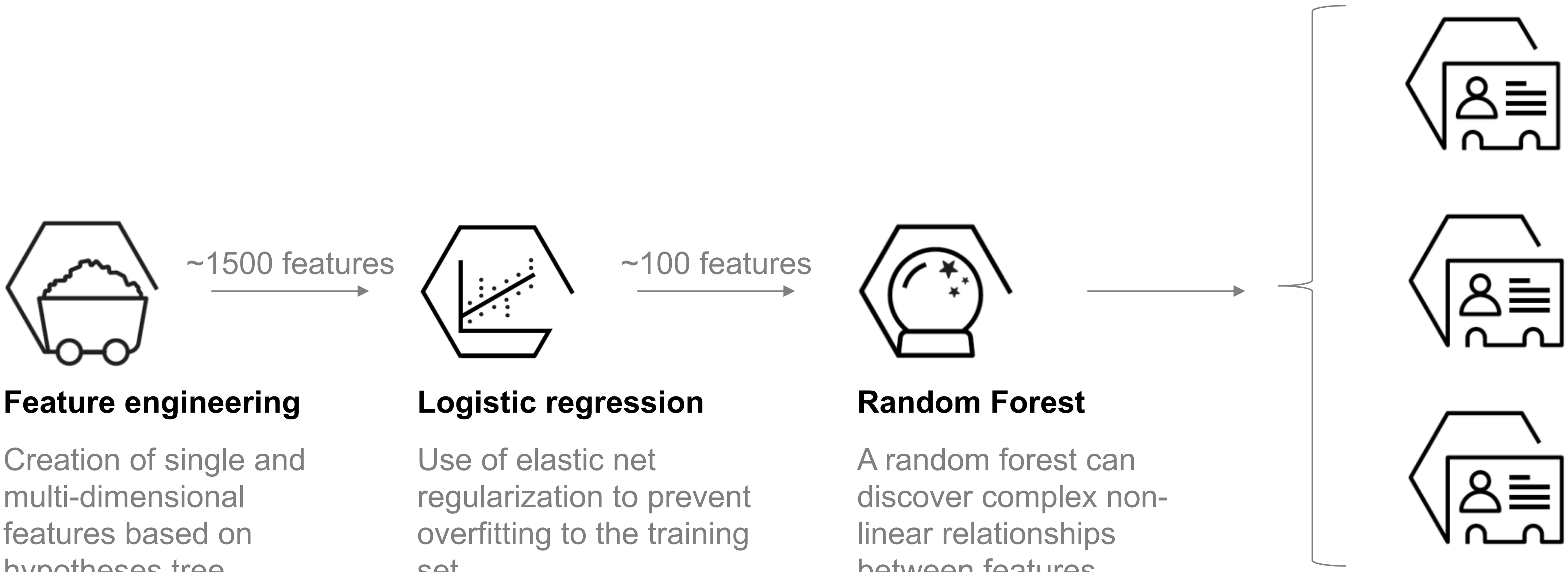


Two questions raised frequently by our client

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

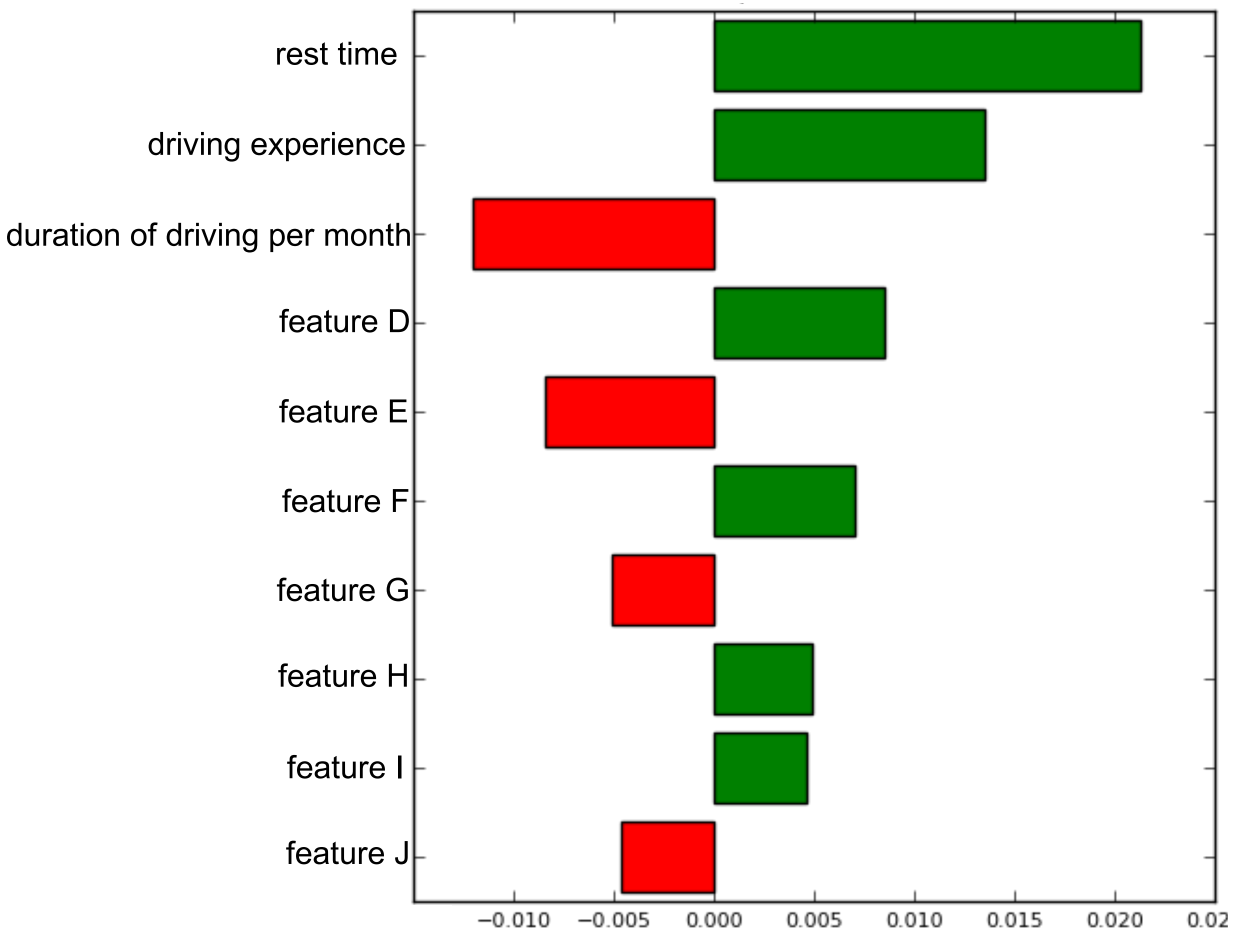
A random forest can discover complex non-linear relationships between features

LIME/SHAP

Personalized interpretations of the RF results

SOURCE: Team analysis

Example of a driver profile

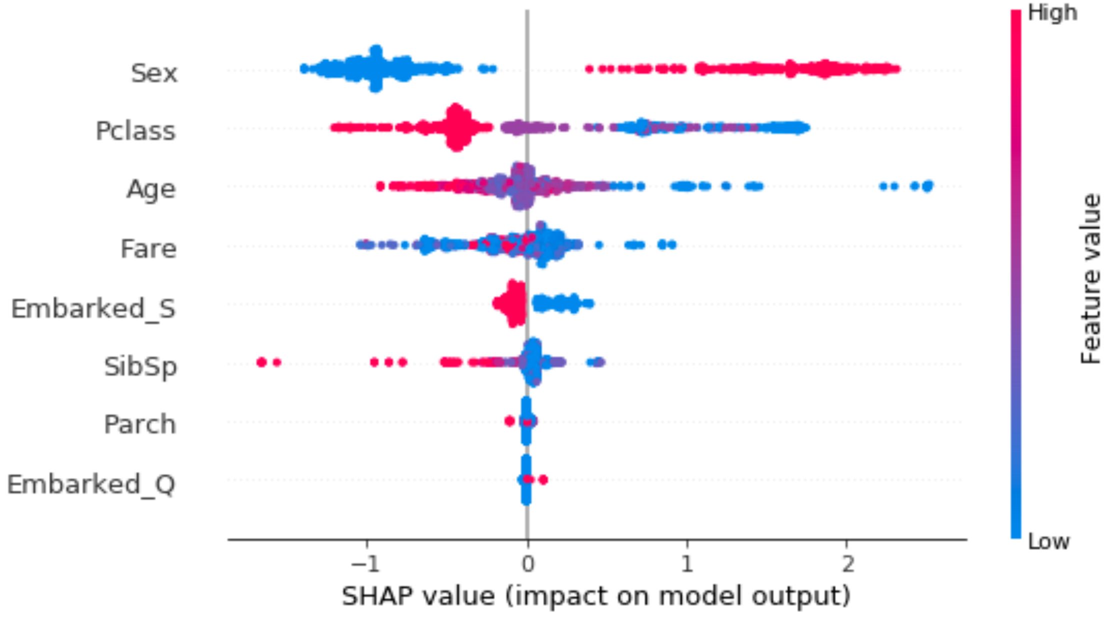


Explainers are useful but can lead to dangerous misinterpretations

03

LIME and SHAP provide ways to explain black-box models…

LIME are SHAP are two popular *model agnostic*, *local explanation* approaches to explain any black box classifier.

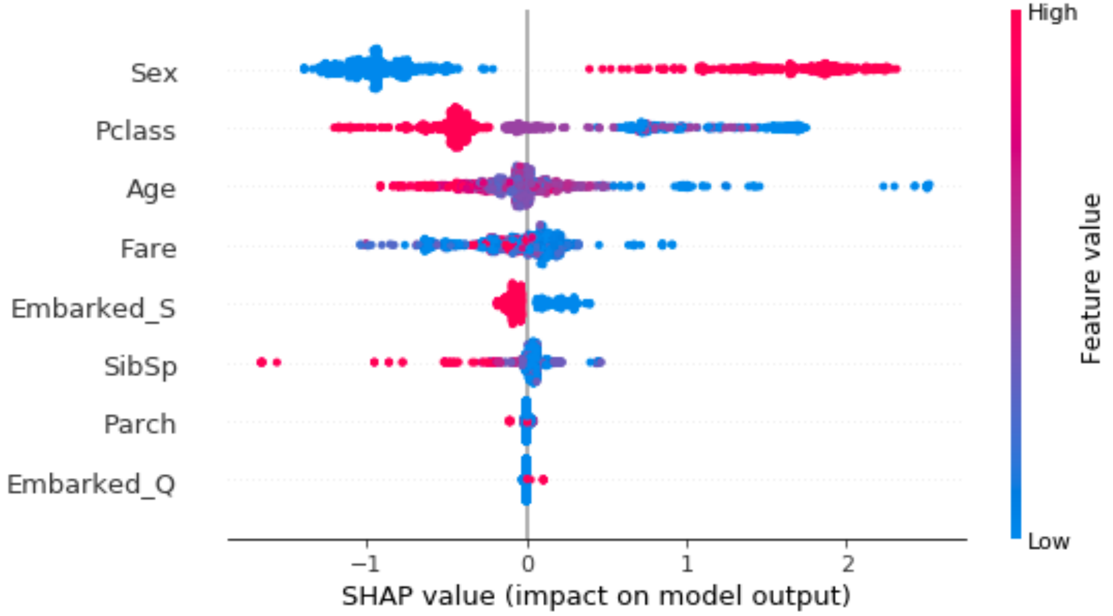


Sources:

- ["Why Should I Trust You?": Explaining the Predictions of Any Classifier, Ribeiro]
- [A Unified Approach to Interpreting Model Predictions, Lundberg]

LIME and SHAP provide ways to explain black-box models…

LIME are SHAP are two popular *model agnostic*, *local explanation* approaches to explain any black box classifier.



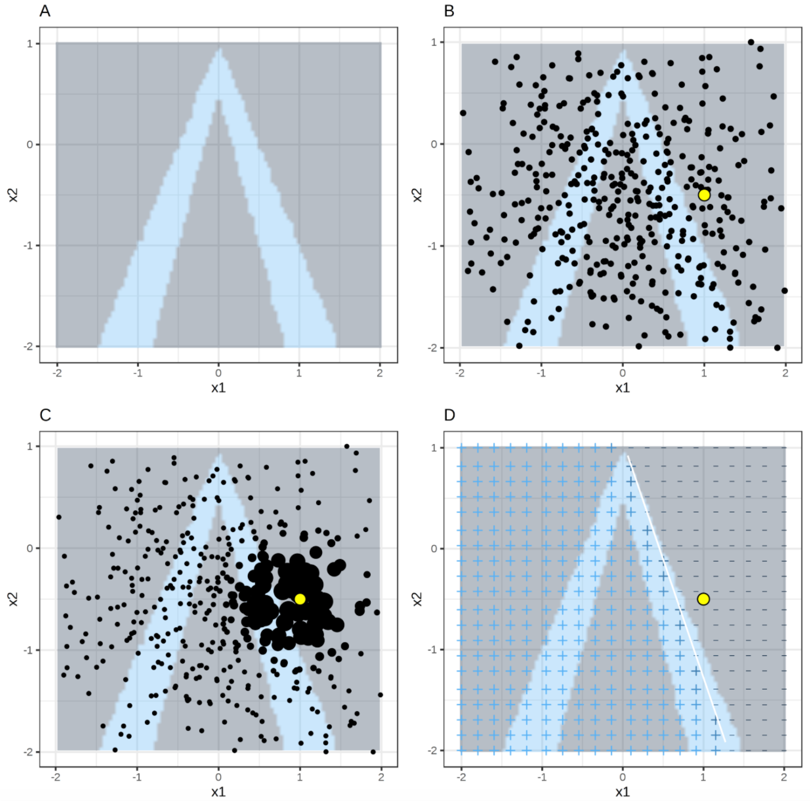
How to control that the explanations are trustworthy?

Sources:

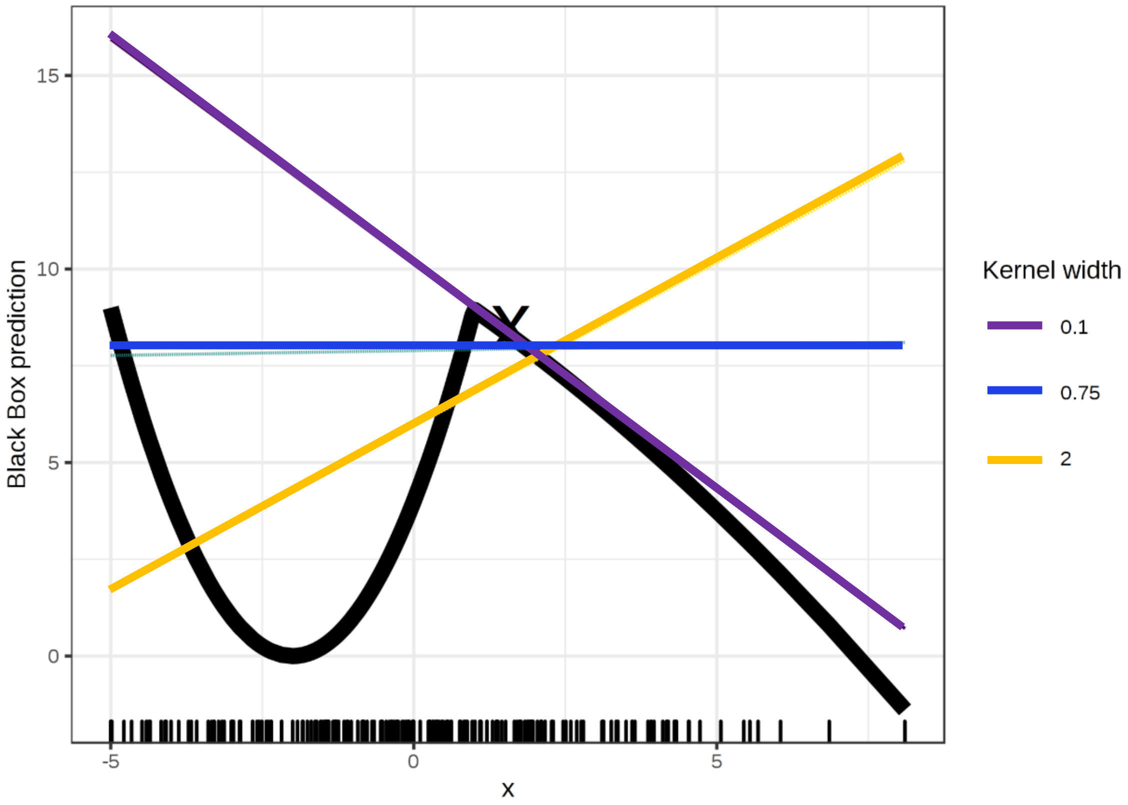
- "Why Should I Trust You?": Explaining the Predictions of Any Classifier, Ribeiro
- A Unified Approach to Interpreting Model Predictions, Lundberg

… but LIME and SHAP can be seen as black-boxes themselves (1/2)

They work by perturbing a local point and building an interpretable model and the sampled data.



The definition of "neighborhood" of a point can have a huge impact on the explanation



Source: [Interpretable Machine Learning, C. Molnar]()

… but LIME and SHAP can be seen as black-boxes themselves (2/2)

Adversarial attack on LIME and SHAP

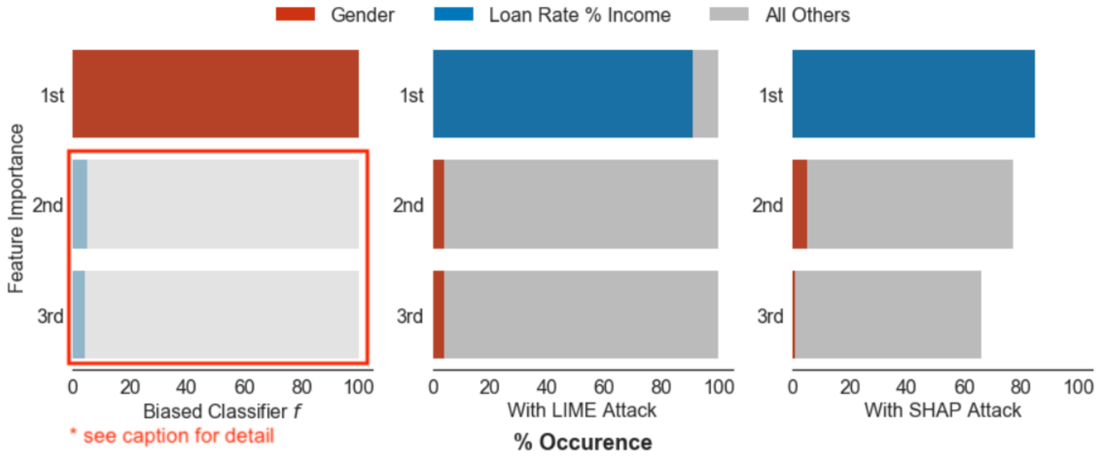
- For a loan-application, let's consider a classifier that is:
  - Fully biased on observed feature distribution
  - Unbiased elsewhere (i.e. on points that are unlikely to be real)
- Biased classifier:

$$f(x) = \begin{cases} 1\ if\ x\ is\ male \\ 0\ if\ x\ is\ female \end{cases}$$

- Unbiased classifier called $g(x)$
- Adversarial classifier:

$$e(x) = \begin{cases} f(x)\ if\ x\ close\ to\ training\ distribution \\ g(x) else \end{cases}$$

LIME and SHAP can be fooled to display a fully biased classifier as unbiased



% of data points for which each feature (color coded) shows up in top 3 (according to LIME and SHAP's ranking of feature attributions) for the biased classifier f (left), and adversarial classifier.

Source: How can we fool LIME and SHAP? Adversarial Attacks on Post hoc Explanation Methods, D. Slack

GA2M algorithm

04

GA2M idea

Prediction is written as a sum of interpretable functions

$$\hat{y} = \alpha + \underbrace{\sum_{i \in F} f_i(x_i)}_{\text{Univariate}} + \underbrace{\sum_{(i,j) \in S(F^2)} f_{i,j}(x_i, x_j)}_{\text{Bivariate}}$$

What’s different from classical GAM ?

1. Fit of univariate variables uses boosted trees (instead of splines)
2. Some bivariate features are fitted to capture interaction terms
3. Most relevant bivariate features are selected through a new method called FAST

Sources:

- Accurate Intelligible Models with Pairwise Interactions, Y. Lou, R. Caruana et al.
- Intelligible Models for Classification and Regression, Y. Lou, R. Caruana et al.

GA2M fitting steps

$$\hat{y} = \alpha + \underbrace{\sum_{i \in F} f_i(x_i)}_{\text{Univariate}} + \underbrace{\sum_{(i,j) \in S(F^2)} f_{i,j}(x_i, x_j)}_{\text{Bivariate}}$$

Main steps

1. Fit univariate functions with boosted trees

2. Use FAST to select most relevant pairs on the predict residuals

3. Fit pairwise functions on the residuals

Sources:

- Accurate Intelligible Models with Pairwise Interactions, Y. Lou, R. Caruana et al.
- Intelligible Models for Classification and Regression, Y. Lou, R. Caruana et al.

GA2M fitting is close to Gradient Boosting algorithm

Gradient boosting

```
$\alpha \leftarrow mean(y)$
$f \leftarrow \alpha$
$residual \leftarrow y - f(X)$

$for\ N\ iterations$:
    $f^{new} \leftarrow fit(X, residual)$
    $f \leftarrow f + \lambda f^{new}$
   $residual \leftarrow y - f(X)$



$return\ f$
```

Sources:

- Accurate Intelligible Models with Pairwise Interactions, Y. Lou, R. Caruana et al.
- Intelligible Models for Classification and Regression, Y. Lou, R. Caruana et al.

GA2M fitting is close to Gradient Boosting algorithm

Gradient boosting

```
$\alpha \leftarrow mean(y)$
$f \leftarrow \alpha$
$residual \leftarrow y - f(X)$

$for\ N\ iterations:$
    $f^{new} \leftarrow fit(X, residual)$
    $f \leftarrow f + \lambda\, f^{new}$
    $residual \leftarrow y - f(X)$

$return\ f$
```

GA2M boosting

```
$\alpha \leftarrow mean(y)$
$f \leftarrow \alpha$
$residual \leftarrow y - f(X)$

$for\ N\ iterations:$
    $for\ i\ in\ \{features\}:$
        $f_i^{new} \leftarrow fit(X_i, residual)$
        $f_i \leftarrow f_i + \lambda\, f_i^{new}$
        $f \leftarrow \alpha + \sum f_i$
        $residual \leftarrow y - f(X)$

$return\ f$
```

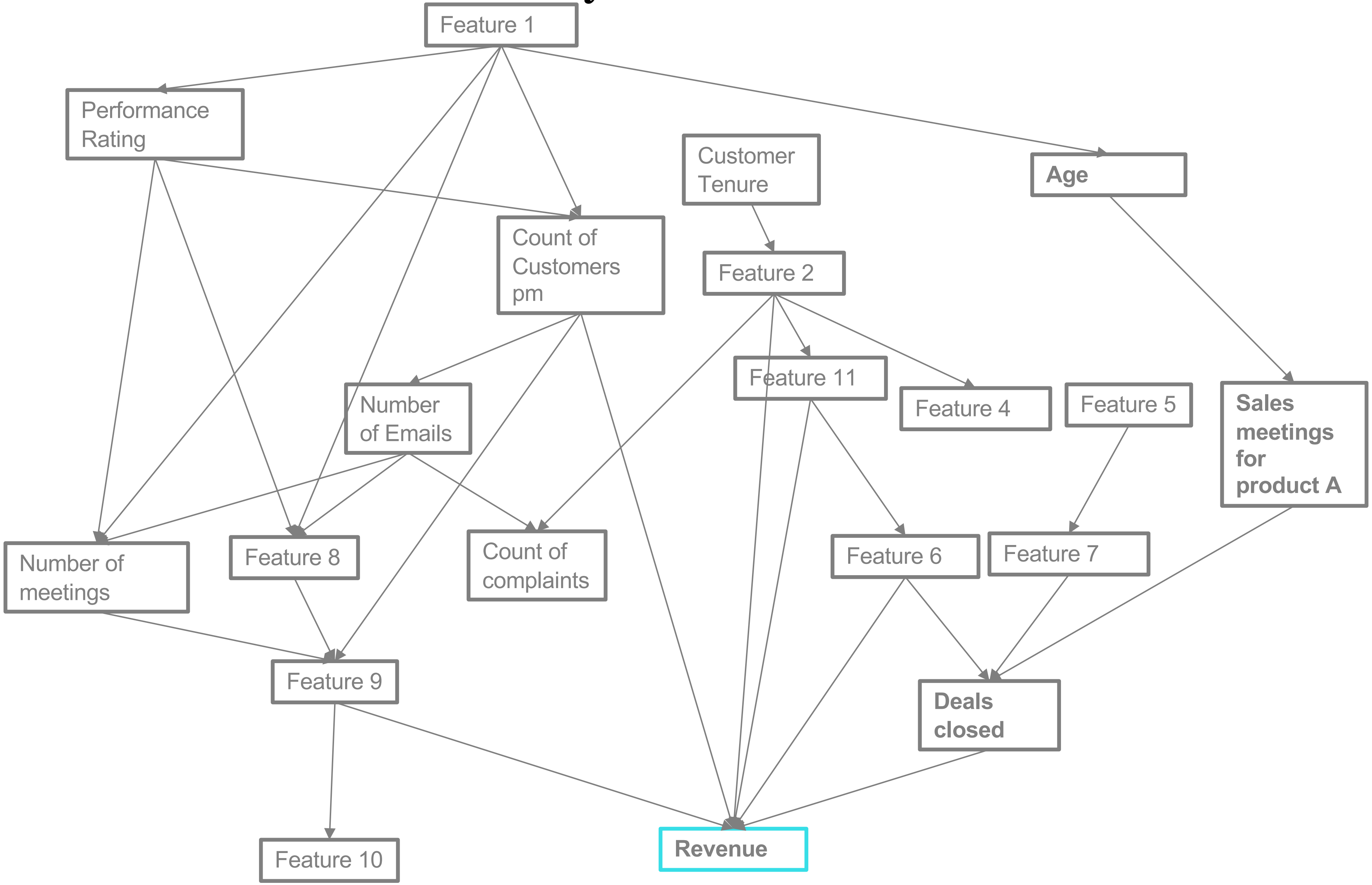
Sources:

- Accurate Intelligible Models with Pairwise Interactions, Y. Lou, R. Caruana et al.
- Intelligible Models for Classification and Regression, Y. Lou, R. Caruana et al.

Causal Inference with CausalNex

05

Explanation model ≠ causality

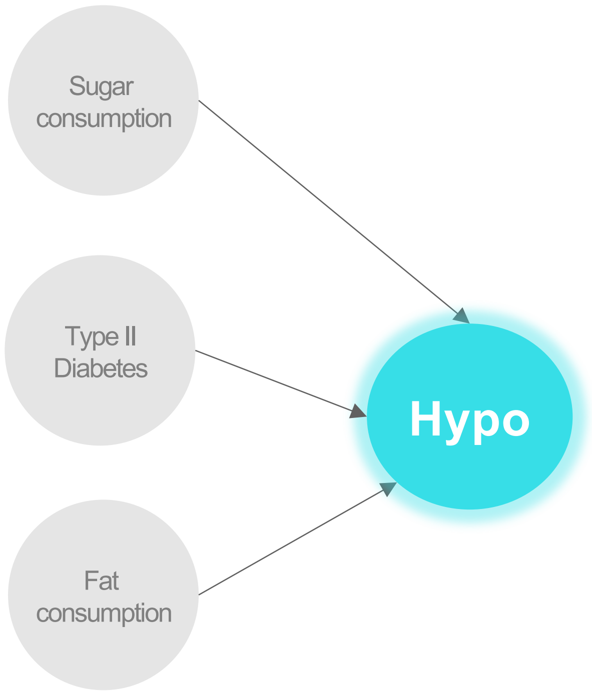


SOURCE: Team analysis

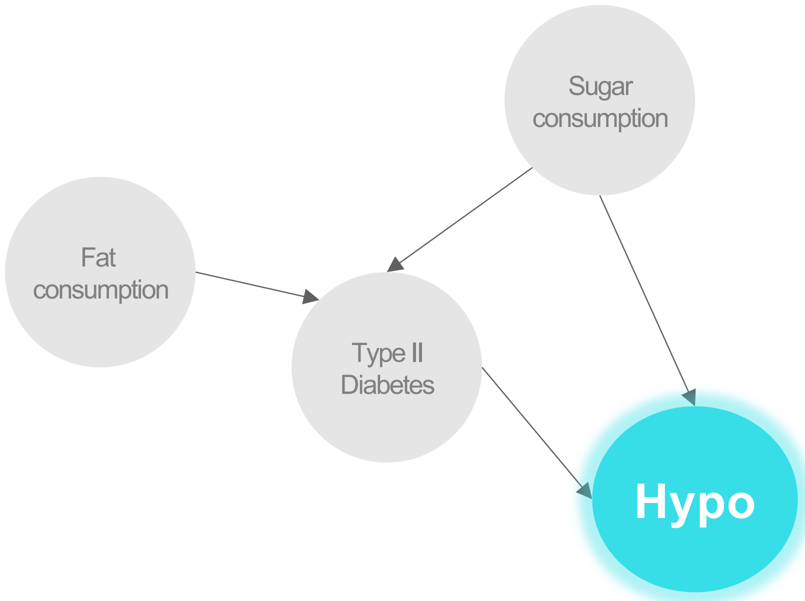
We can better identify the right intervention with CausalNex

- Establishing causal relationship is critical for us to recommend the right interventions
- Traditional models, like linear regressions, have simplistic assumptions, which may lead to the wrong recommendation.
- CausalNex uses graph models such as Bayesian Networks which can be more intuitive and allows domain expertise encoded with data to form a better understanding of relationships

Linear regression



Bayesian network



CausalNex is an **open-source** Python library that leverages **Bayesian Networks** to help data scientists to infer causation rather than observing correlation.

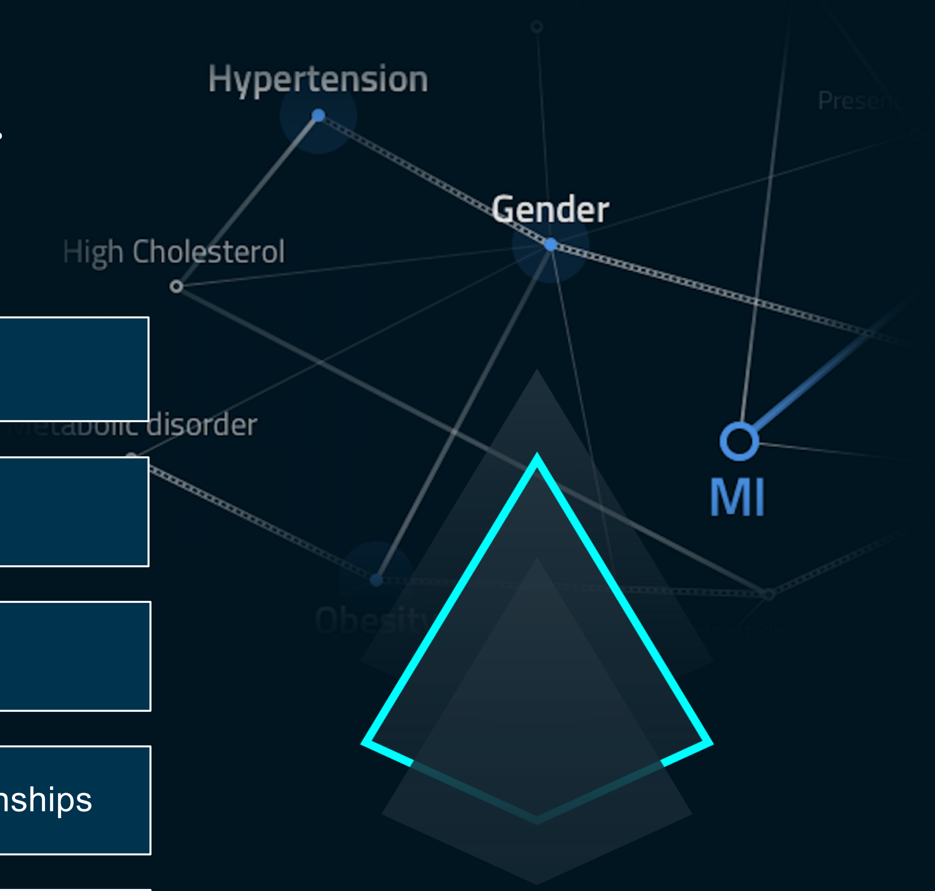
Structure learning – Learn relationships in data with NOTEARS, a state-of-the-art algorithm

Embed domain expertise – Enable experts to add and remove inaccurate correlation

Graph visualisation – Use extensions of NetworkX to understand causal relationships

Likelihood estimation – Estimate the probability distribution of variables based on their relationships

Counterfactual analysis – Infer what happens to *target Y* when *feature X* is changed



CausalNex

A data science research

Powered by QuantumBlack

Status:

7+ projects

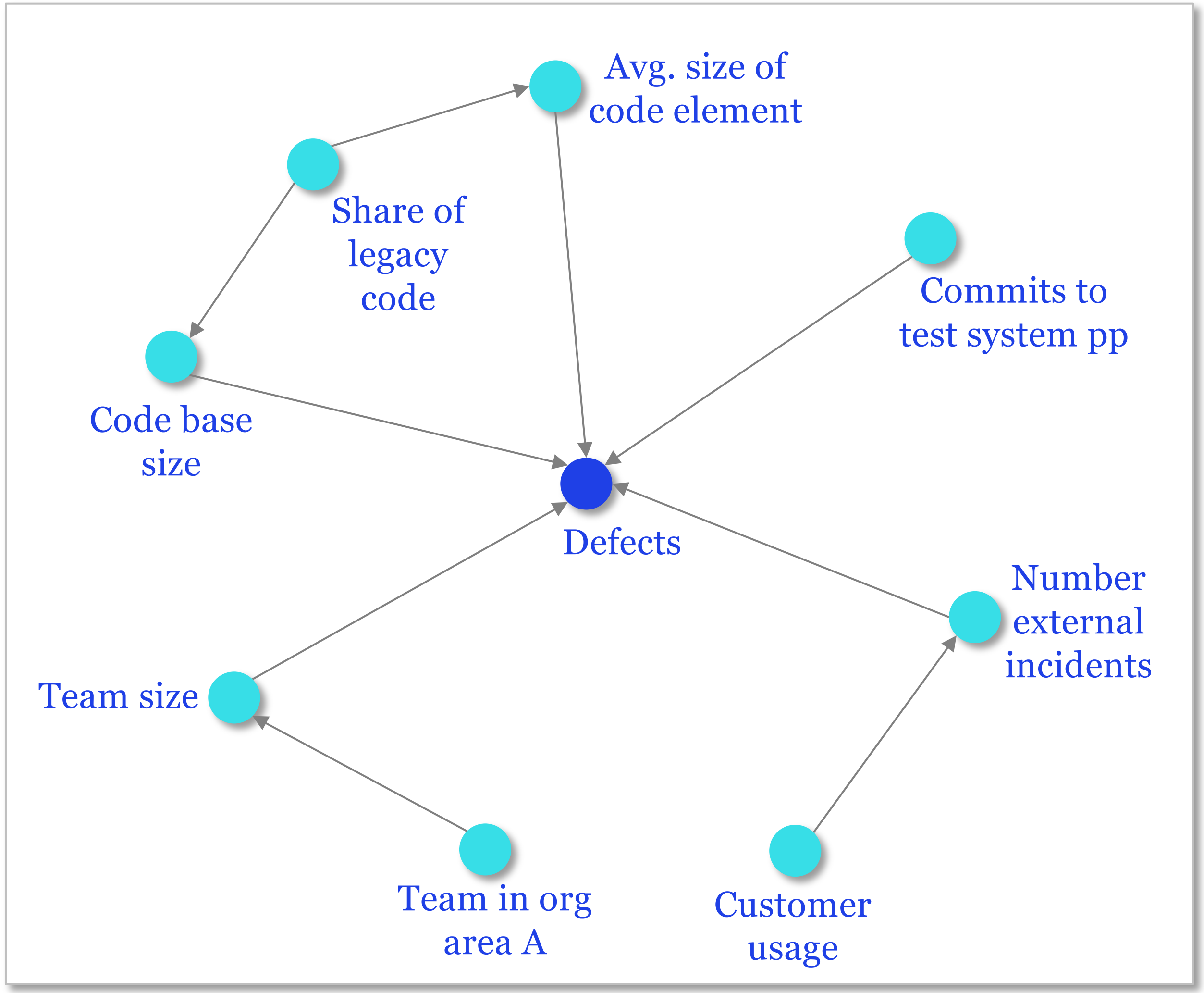
QuatumBlack Labs

PyPI

Read The Docs

ILLUSTRATIVE

Case Study 1. - Bayesian Networks can be used to identify drivers of quality and productivity in product development for a software company



- CausalNex/Bayesian Networks illustrate how different **drivers affect each other** as well as the **target variable** (conditional probabilities)
- The network structure enables **better estimation** of the **impact** of interventions based on specific drivers as the effect cascades through the network
- Specific connections can be set manually to better capture **business knowledge** of underlying processes

Questions?

QUANTUMBLACK

A MCKINSEY COMPANY