Why Did the Model Decide X?

Interpretability of Controls & Machine Learning

Rain Vagel, Data Scientist, Wise

Who Am I?

Who Am I?

- **Over two years experience in mitigating financial crime**
  First in Fraud and then Anti-Money Laundering (AML)

Who Am I?

- **Over two years experience in mitigating financial crime**
  First in Fraud and then Anti-Money Laundering (AML)

- **Introducing machine learning as a control in a highly regulated environment**
  A wealth of information in managing external stakeholders

Who Am I?

- **Over two years experience in mitigating financial crime**
  First in Fraud and then Anti-Money Laundering (AML)

- **Introducing machine learning as a control in a highly regulated environment**
  A wealth of information in managing external stakeholders

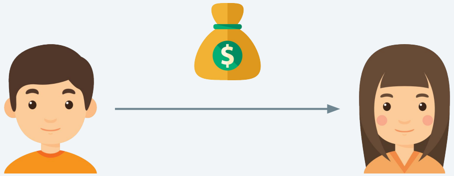
- **Managing a shift from something regulators are familiar with to something they are not**
  Using machine learning has some additional challenges in comparison to rules

What Will We Talk About?

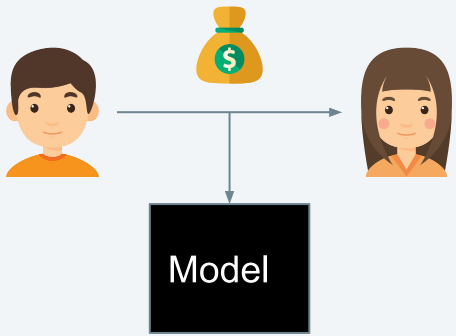
- Why do we need interpretability?
- What are glassbox and blackbox models?
- Interpretability of different controls and models.
- Interpreting controls to different end-consumers

Why Interpretability?

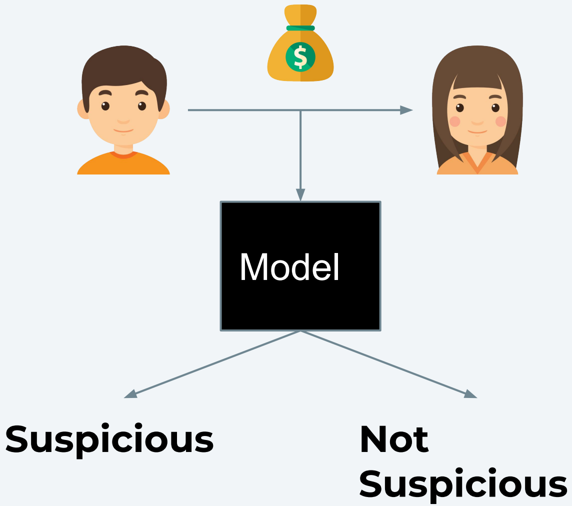
Why Interpretability?



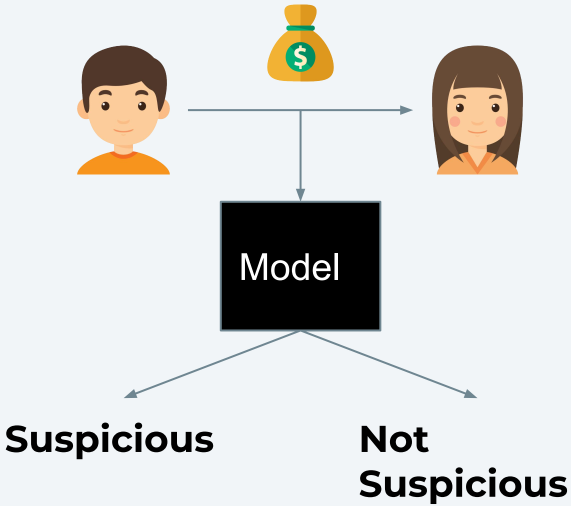
Why Interpretability?



Why Interpretability?



Why Interpretability?



- **Why did the model decide so?**
- **What would happen if you change something about the behaviour?**
- **Important in high-risk areas**
  Such as stopping terrorism financing

Duck Rule

If it looks like a duck,
swims like a duck,
and quacks like a duck,
then it probably is a duck

Duck Rule

If it looks like a duck,] - Condition
swims like a duck,] - Condition
and quacks like a duck,] - Condition
then it probably is a duck] - Result

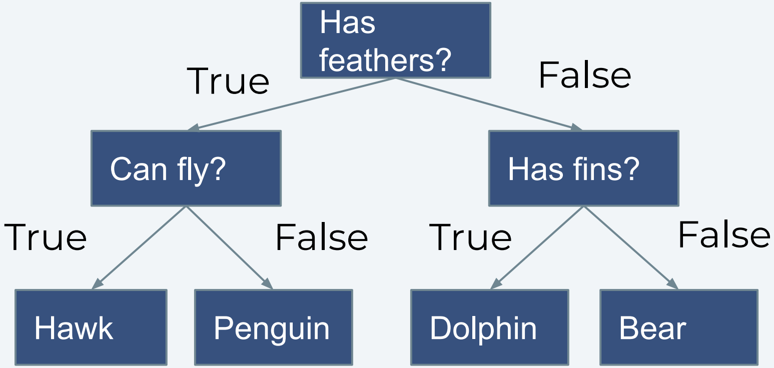
Glassbox vs. Blackbox.

What is a Glassbox vs. Blackbox Model?

Glassbox - Can follow the internal logic

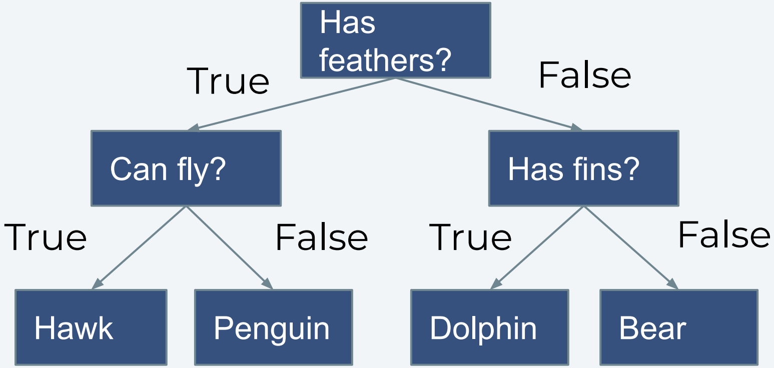
What is a Glassbox vs. Blackbox Model?

Glassbox - Can follow the internal logic



What is a Glassbox vs. Blackbox Model?

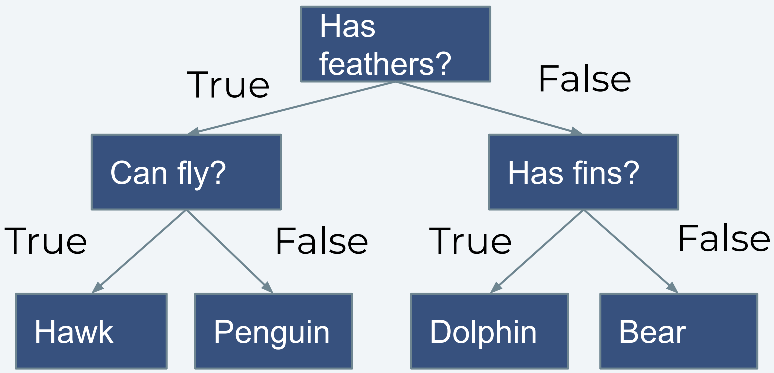
Glassbox - Can follow the internal logic



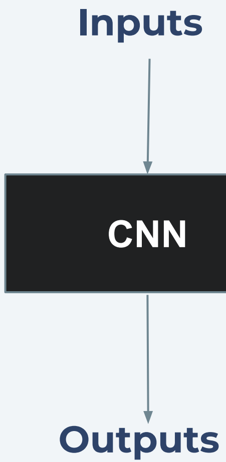
Blackbox - Can only observe the inputs and outputs

What is a Glassbox vs. Blackbox Model?

Glassbox - Can follow the internal logic

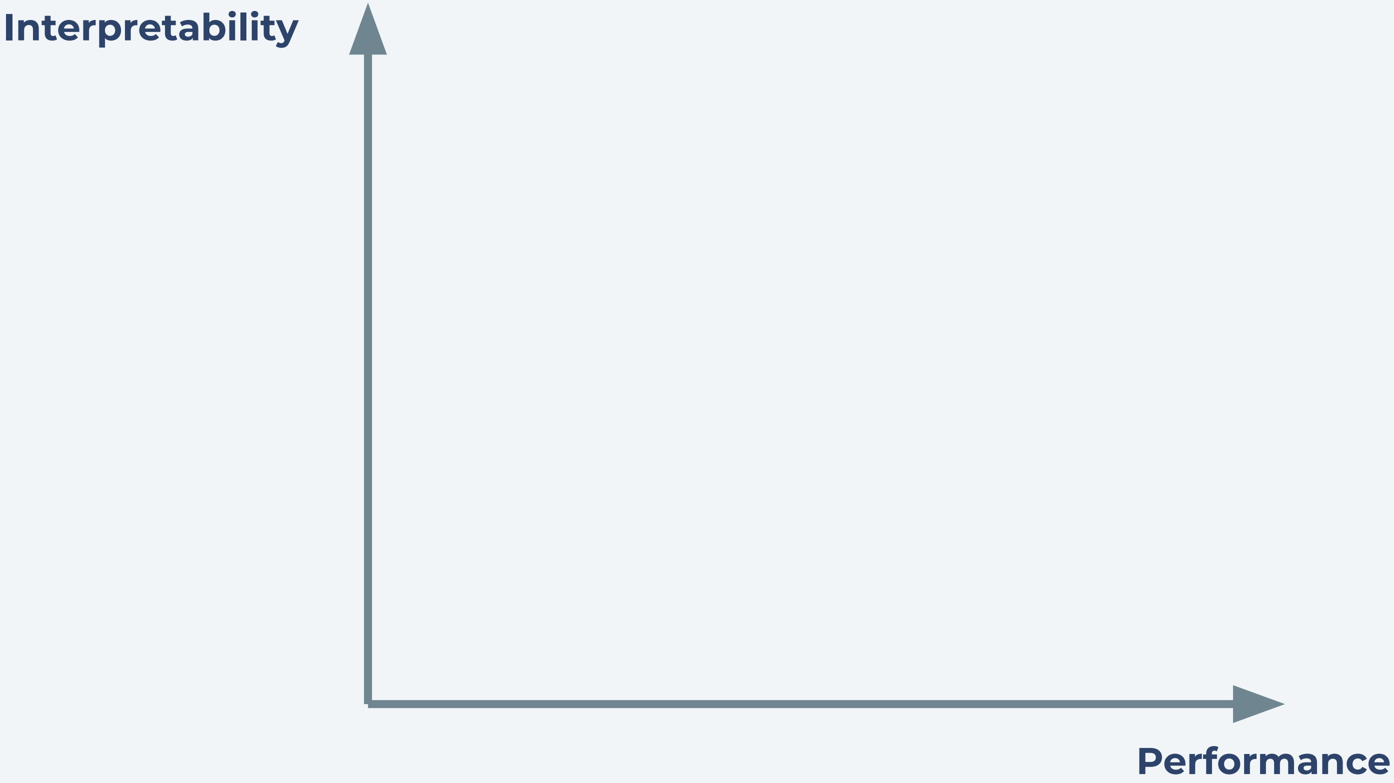


Blackbox - Can only observe the inputs and outputs



Interpretability of Different Controls.

Performance vs. Interpretability



Performance vs. Interpretability

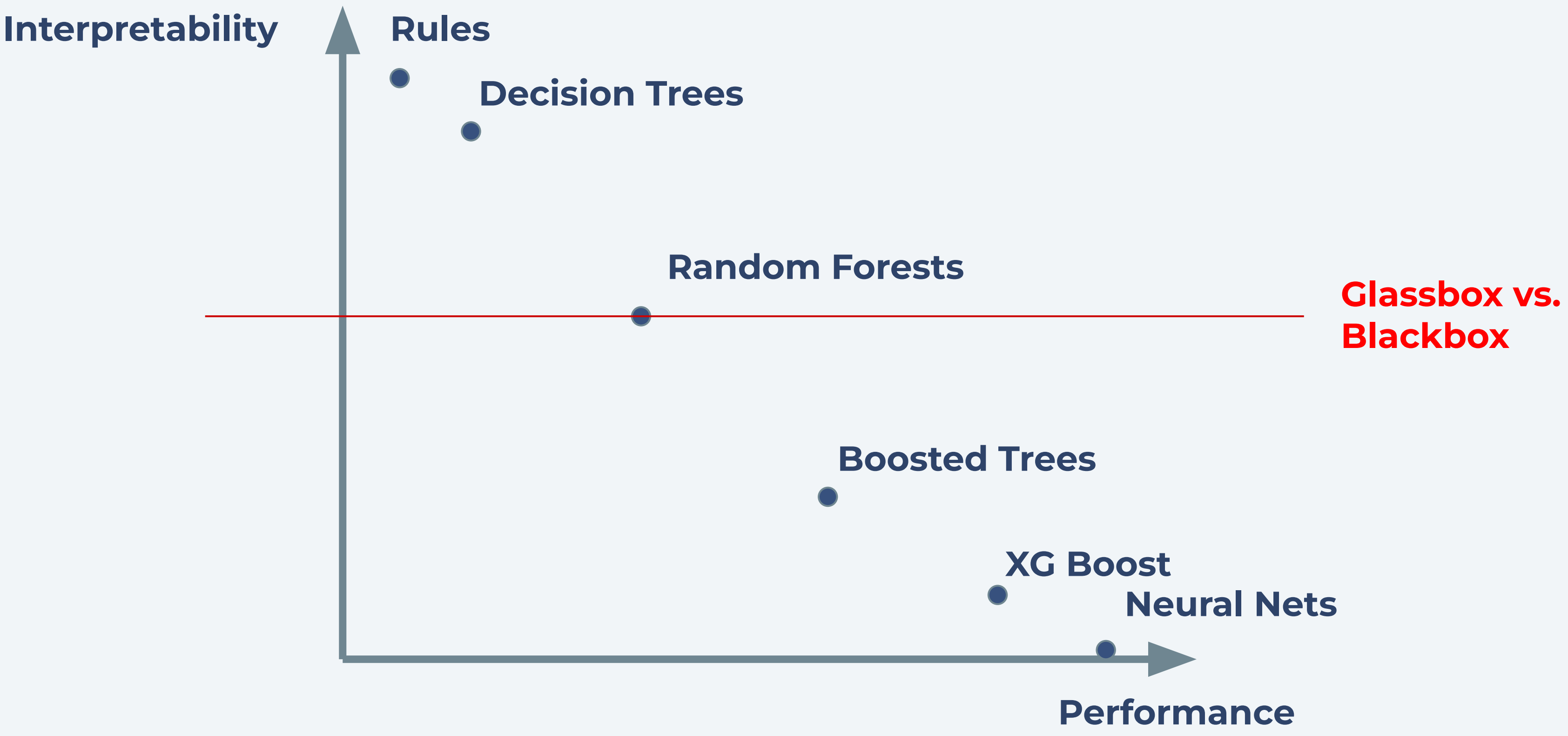
Performance vs. Interpretability

Performance vs. Interpretability

Performance vs. Interpretability

Performance vs. Interpretability

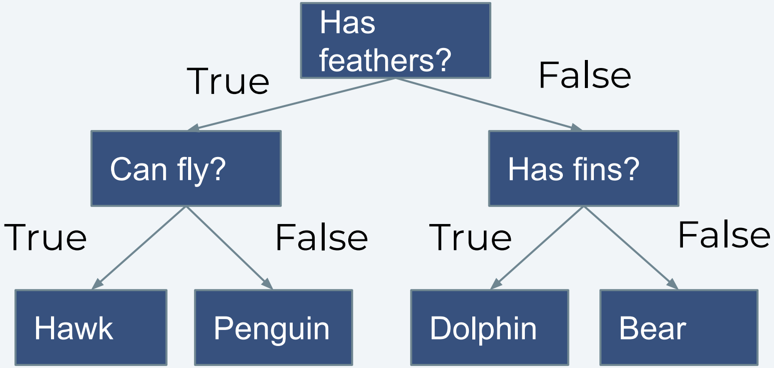
Performance vs. Interpretability



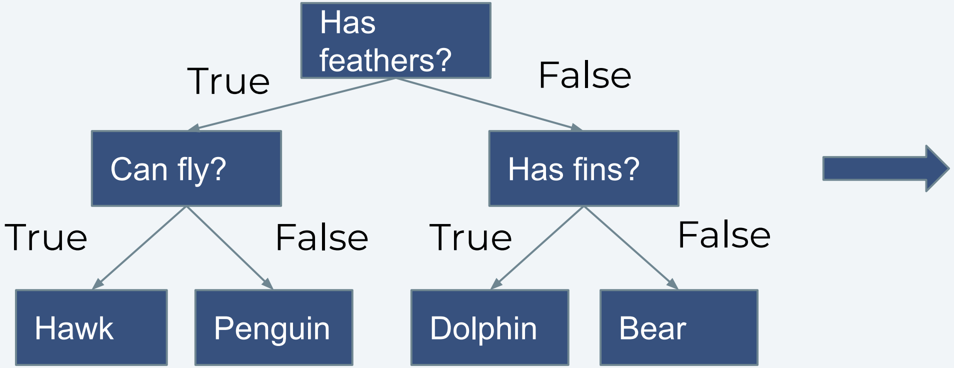
Duck Rule

If it looks like a duck,] - Condition
swims like a duck,] - Condition
and quacks like a duck,] - Condition
then it probably is a duck] - Result

Decision Tree Interpretability



Decision Tree Interpretability



It has feathers AND can not fly. It is a penguin.

Decision Tree Interpretability - Not Always Practical

Tree With 100 Nodes

Decision Tree Interpretability - Not Always Practical

Tree With 100 Nodes

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut leo justo, aliquam id viverra vel, euismod vel eros. In ultrices sapien lectus, ac convallis neque facilisis in. Mauris vitae urna turpis. Nullam lobortis pellentesque nulla non viverra. Praesent gravida tortor at lobortis sollicitudin. Aliquam nec posuere lorem, non tristique felis. Sed gravida viverra nulla, eget bibendum urna pulvinar sagittis. Vivamus hendrerit, lacus a tristique euismod, nisi augue facilisis mi, id elementum magna eros in quam. Cras imperdiet, sapien in condimentum faucibus, ipsum nisl congue tortor, euismod molestie lorem nulla et odio. Sed egestas condimentum augue. Suspendisse a ante luctus massa sagittis pulvinar. Sed rutrum quam ut metus egestas, quis facilisis leo elementum. Aliquam vitae ex eleifend, iaculis tellus vel, pellentesque ligula. Quisque laoreet ultricies est, non pretium arcu pellentesque et. Morbi id mi eu ipsum consequat eleifend in a mauris. Nulla vitae blandit odio. Sed tristique iaculis accumsan. Vivamus lobortis enim eu purus tincidunt, sit amet vehicula est accumsan. Aliquam gravida ullamcorper consequat. Praesent varius metus viverra commodo aliquam. Etiam semper dignissim turpis, ut volutpat sem molestie imperdiet. Integer non dapibus ante. Sed ante quam, commodo eget consectetur id, condimentum ut libero. Mauris id sollicitudin mi. Nulla sit amet nunc quis sapien varius venenatis. Mauris porta justo ante. Nullam sit amet urna ut nibh auctor hendrerit quis eget nisl. Sed elit sem, blandit ut metus in, sollicitudin porta leo. Nulla vehicula urna vitae vehicula faucibus. Nunc placerat venenatis dolor, a congue mi.

Interpreting Blackboxes

- **Shapley**
  Calculate how to fairly distribute outcome to features

Interpreting Blackboxes

- **Shapley**
  Calculate how to fairly distribute outcome to features

- **LIME - Local Interpretable Model-agnostic Explanations**
  Train local interpretable models to explain global mode

Interpreting Blackboxes - Shapley

Park Nearby	Floor Area	Cats Allowed	Prediction
Yes	50	Yes	310 000

Interpreting Blackboxes - Shapley

Park Nearby	Floor Area	Cats Allowed	Prediction
Yes	50	Yes	310 000

Park Nearby	Floor Area	Cats Allowed	Prediction
Yes	50	No	300 000

Interpreting Blackboxes - Shapley

Park Nearby	Floor Area	Cats Allowed	Prediction
Yes	50	Yes	310 000

Park Nearby	Floor Area	Cats Allowed	Prediction
Yes	50	No	300 000

310 000 - 300 000 = 10 000

Interpreting Blackboxes - Shapley

Park Nearby	Floor Area	Cats Allowed	Prediction
Yes	50	Yes	310 000

310 000 - 300 000 = 10 000

Park Nearby	Floor Area	Cats Allowed	Prediction
Yes	50	No	300 000

Park Nearby	Floor Area	Cats Allowed	Prediction
No	50	Yes	280 000

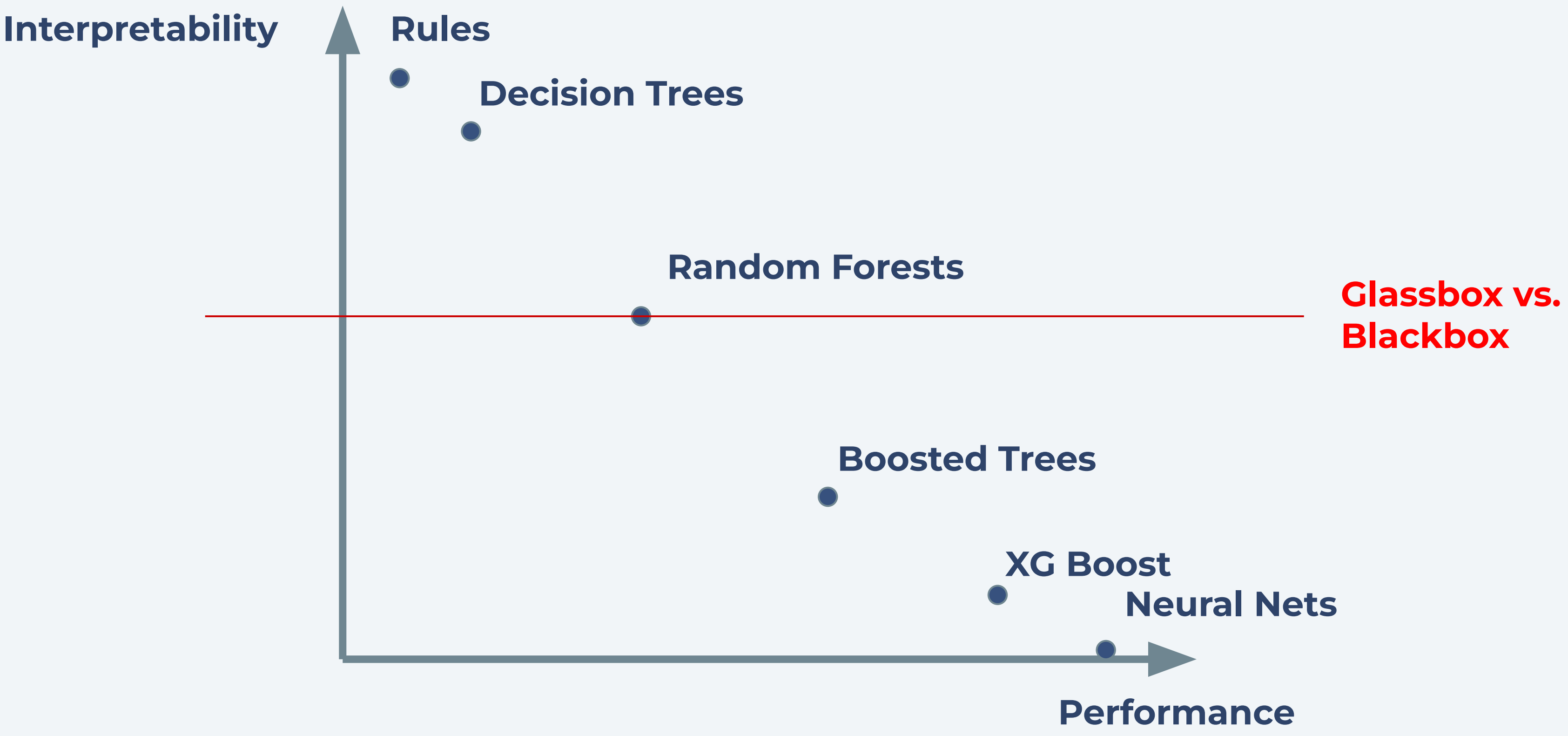
280 000 - 260 000 = 20 000

Park Nearby	Floor Area	Cats Allowed	Prediction
No	50	No	260 000

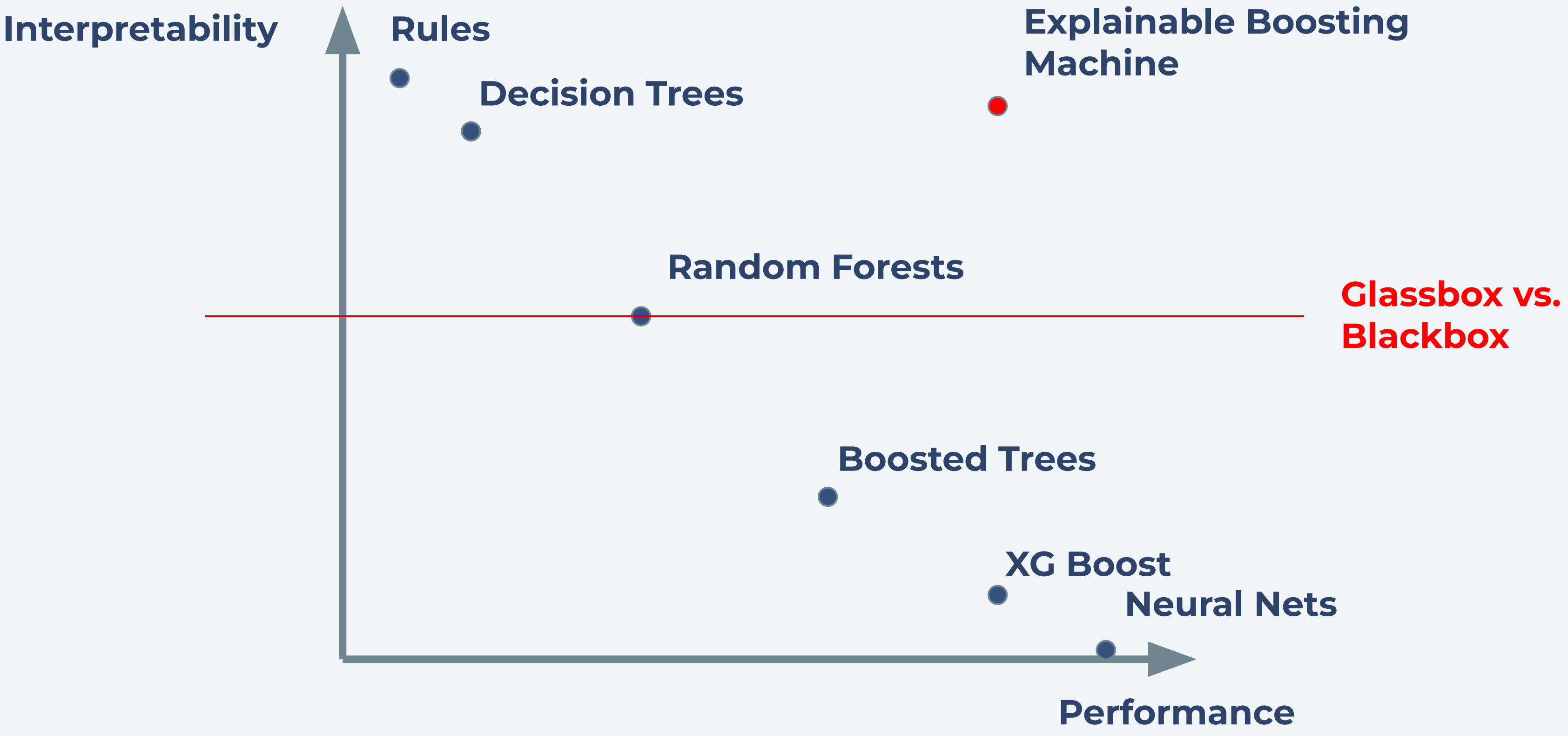
Interpreting Blackboxes - Shapley

- Marginal feature contribution
- Average contributions over all possible coalitions
- Monte-Carlo sampling to keep it efficient

Performance vs. Interpretability



Performance vs. Interpretability



Presenting to End-Consumers.

What Types of Consumers?

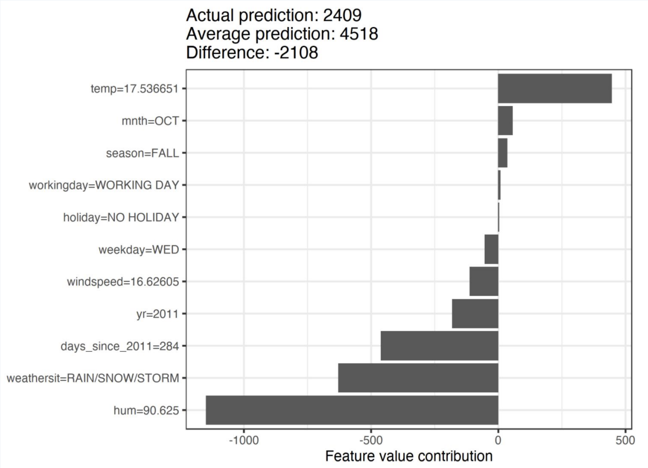
- **Data scientists**
  The most technical
- **Other functions working closely with the data scientists**
  For example, product managers, operational agents etc.
- **External consumers**
  Auditors, regulators etc.

Evolution of Interpreting Outcomes

Feature	Cont
Cat: Allowed	20 000
Cat: Banned	-15 000
Apt Size	200 000
Park: Near	40 000
Park: Far	- 25 000
Garage Size	100 000
Nbr of Rooms	45 000

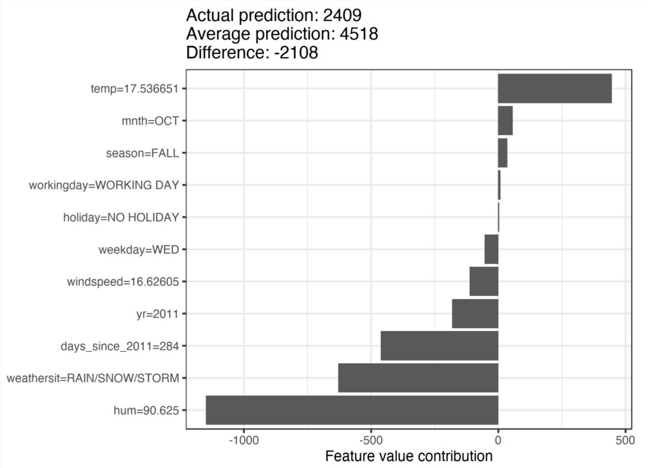
Evolution of Interpreting Outcomes

Feature	Cont
Cat: Allowed	20 000
Cat: Banned	-15 000
Apt Size	200 000
Park: Near	40 000
Park: Far	- 25 000
Garage Size	100 000
Nbr of Rooms	45 000



Evolution of Interpreting Outcomes

Feature	Cont
Cat: Allowed	20 000
Cat: Banned	-15 000
Apt Size	200 000
Park: Near	40 000
Park: Far	- 25 000
Garage Size	100 000
Nbr of Rooms	45 000



It has feathers AND can not fly. It is a penguin.

What makes a good slide deck?

- **Less is more**
  Keep slides as simple as possible

- **Space makes things look good**
  Avoid cramming slides with too much information

- **Tell a story**
  Break up lists and engage the viewer

And you will read this last

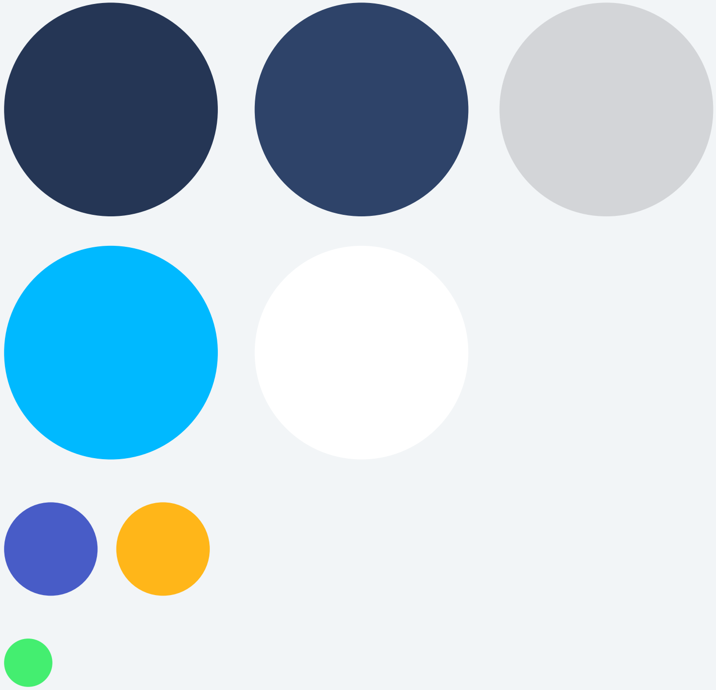
You will read this first.

And then you will read this

Then this one

The only colours you should be using in your deck. The purple, yellow and brand blue are highlights colours — use sparingly.

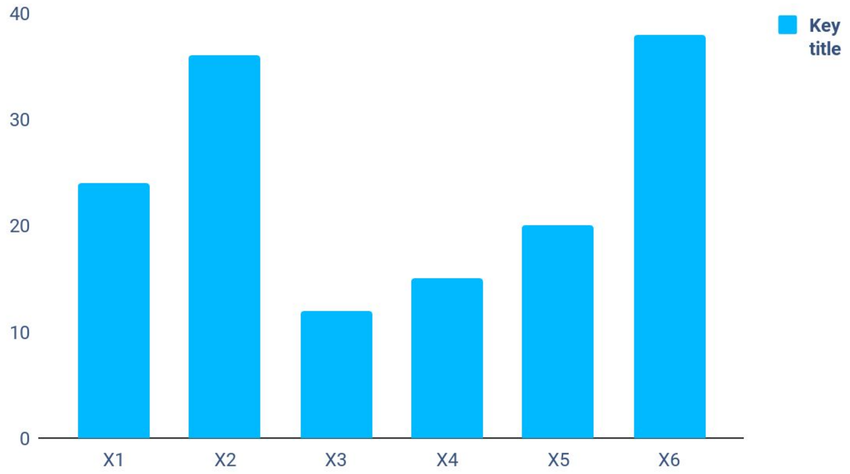
Green is only for borderless.



Section title here.

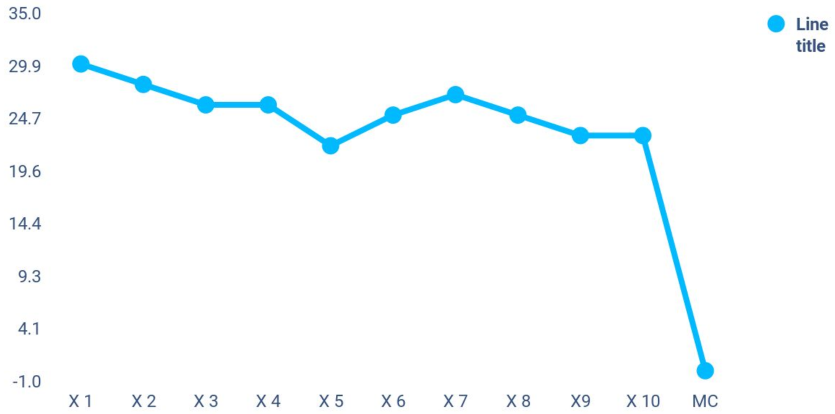
This bar chart is editable in Google sheets.

(Click on the chart, then the dropdown)

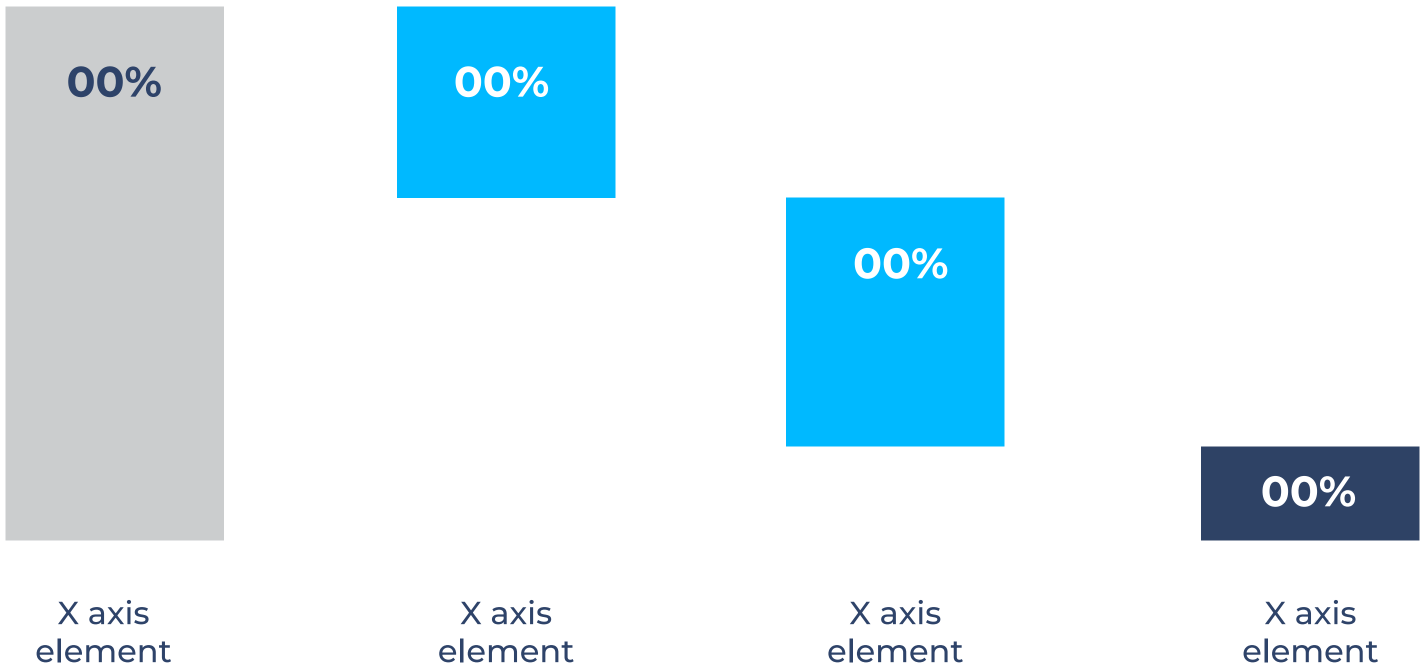


This line chart is editable in Google sheets.

(Click on the chart, then the dropdown)

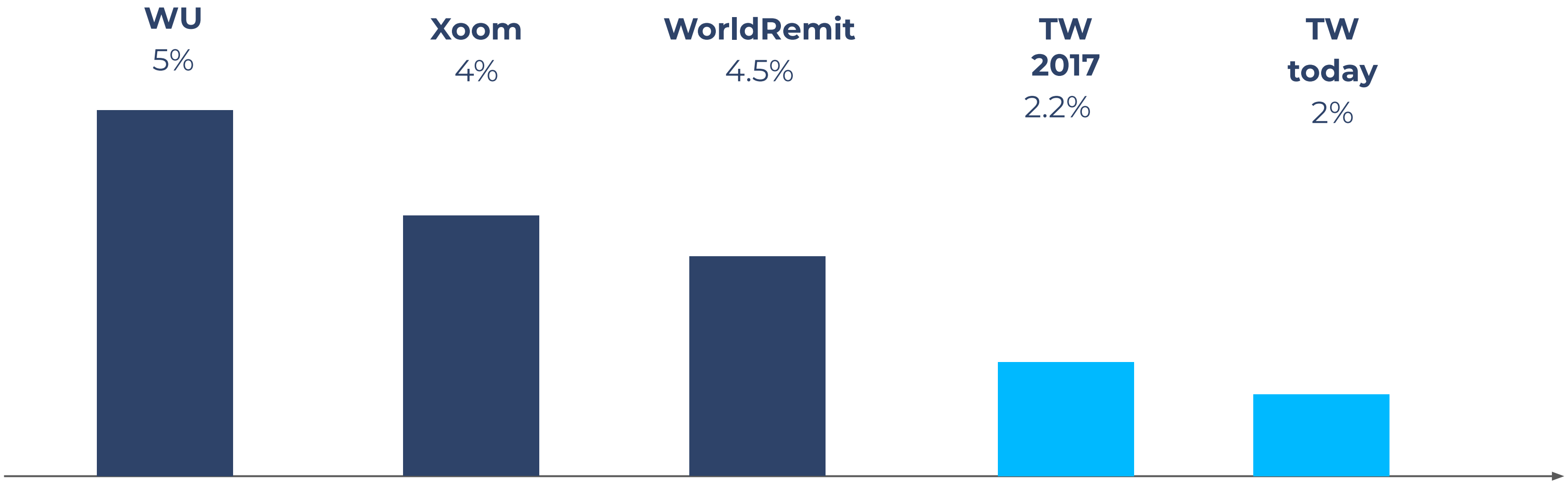


Title of the **page**

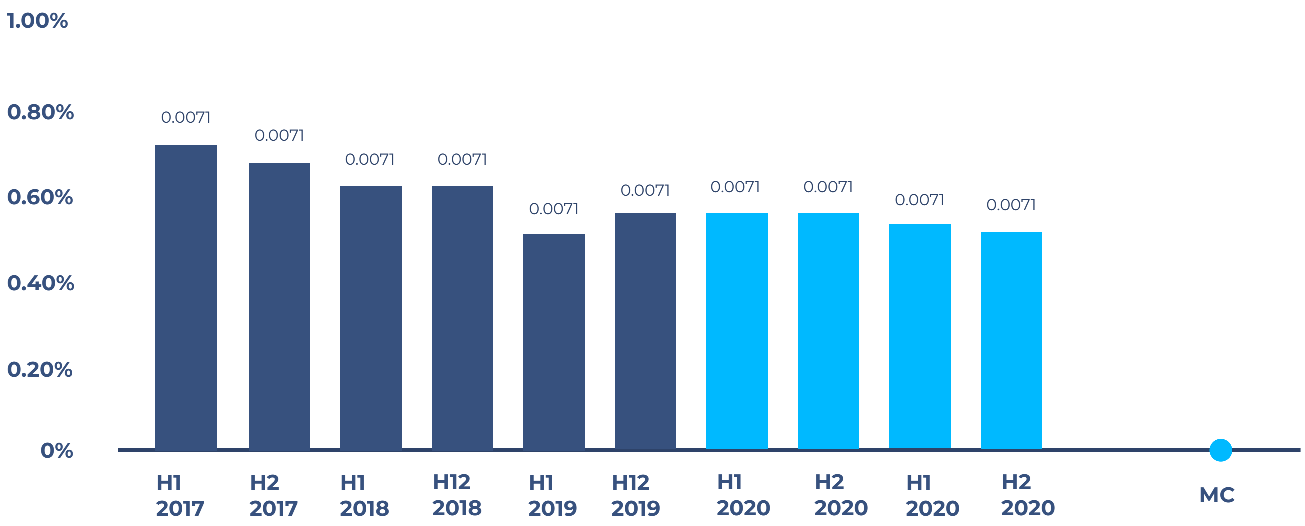


Title of the page

Graph title



Title of the page.



Main point goes here.

Section title here.

Header

Header

100%

Pillars.

Price

Subtitle

Convenience

Subtitle

Speed

Subtitle

Transparency

Subtitle

23.5%

```
if (this) {
    // then that
}
```