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?

Model
Why Interpretability?

Model

Suspicious
Not Suspicious
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, it swims like a duck, and it quacks like a duck, then it probably is a duck.
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

- Has feathers?
  - True: Can fly?
    - True: Hawk
    - False: Penguin
  - False: Has fins?
    - True: Dolphin
    - False: Bear
What is a Glassbox vs. Blackbox Model?

Glassbox - Can follow the internal logic

Blackbox - Can only observe the inputs and outputs

Has feathers?
- True
  - Can fly?
    - True: Hawk
    - False: Penguin
  - False: Dolphins

Has fins?
- True
  - Can fly?
    - True: Hawk
    - False: Penguin
  - False: Dolphins
- False
  - Can fly?
    - True: Hawk
    - False: Penguin

Can fly?
- True
  - Has feathers?
    - True: Hawk
    - False: Penguin
  - False: Dolphins
- False
  - Has feathers?
    - True: Hawk
    - False: Penguin
  - False: Dolphins
What is a Glassbox vs. Blackbox Model?

Glassbox - Can follow the internal logic

- Has feathers?
  - True
    - Can fly?
      - True (Hawk)
      - False (Penguin)
  - False
    - Has fins?
      - True (Dolphin)
      - False (Bear)

Blackbox - Can only observe the inputs and outputs

Inputs

CNN

Outputs
Interpretability of Different Controls.
Performance vs. Interpretability
Performance vs. Interpretability

Interpretability

Rules

Performance
Performance vs. Interpretability

Interpretability

Rules

Decision Trees

Performance
Performance vs. Interpretability

- **Interpretability**
  - Rules
- **Performance**
  - Decision Trees
  - Random Forests
  - Boosted Trees
Performance vs. Interpretability

Interpretability

Rules

- Decision Trees
- Random Forests
- Boosted Trees
- XG Boost
- Neural Nets

Performance
Duck Rule

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

- Has feathers?
  - True
    - Can fly?
      - True: Hawk
      - False: Penguin
  - False
    - Has fins?
      - True: Dolphin
      - False: Bear
Decision Tree Interpretability

- Has feathers?
  - True: Can fly? (Hawk)
  - False: Has fins?
    - True: It has feathers AND can not fly. It is a penguin.
    - False: Bear

- Penguin
- Dolphin
Decision Tree Interpretability - Not Always Practical

Tree With 100 Nodes
Decision Tree Interpretability - Not Always Practical

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

<table>
<thead>
<tr>
<th>Park Nearby</th>
<th>Floor Area</th>
<th>Cats Allowed</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>50</td>
<td>Yes</td>
<td>310 000</td>
</tr>
</tbody>
</table>
### Interpreting Blackboxes - Shapley

<table>
<thead>
<tr>
<th>Park Nearby</th>
<th>Floor Area</th>
<th>Cats Allowed</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>50</td>
<td>Yes</td>
<td>310 000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Park Nearby</th>
<th>Floor Area</th>
<th>Cats Allowed</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>50</td>
<td>No</td>
<td>300 000</td>
</tr>
</tbody>
</table>
## Interpreting Blackboxes - Shapley

<table>
<thead>
<tr>
<th>Park Nearby</th>
<th>Floor Area</th>
<th>Cats Allowed</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>50</td>
<td>Yes</td>
<td>310 000</td>
</tr>
</tbody>
</table>

310 000 - 300 000 = 10 000

<table>
<thead>
<tr>
<th>Park Nearby</th>
<th>Floor Area</th>
<th>Cats Allowed</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>50</td>
<td>No</td>
<td>300 000</td>
</tr>
</tbody>
</table>
## Interpreting Blackboxes - Shapley

<table>
<thead>
<tr>
<th>Park Nearby</th>
<th>Floor Area</th>
<th>Cats Allowed</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>50</td>
<td>Yes</td>
<td>310 000</td>
</tr>
</tbody>
</table>

\[
310 000 - 300 000 = 10 000
\]

<table>
<thead>
<tr>
<th>Park Nearby</th>
<th>Floor Area</th>
<th>Cats Allowed</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>50</td>
<td>No</td>
<td>300 000</td>
</tr>
</tbody>
</table>

\[
310 000 - 300 000 = 10 000
\]

<table>
<thead>
<tr>
<th>Park Nearby</th>
<th>Floor Area</th>
<th>Cats Allowed</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>50</td>
<td>Yes</td>
<td>280 000</td>
</tr>
</tbody>
</table>

\[
280 000 - 260 000 = 20 000
\]

<table>
<thead>
<tr>
<th>Park Nearby</th>
<th>Floor Area</th>
<th>Cats Allowed</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>50</td>
<td>No</td>
<td>260 000</td>
</tr>
</tbody>
</table>
Interpreting Blackboxes - Shapley

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

- Rules
- Decision Trees
- Random Forests
- Boosted Trees
- XG Boost
- Neural Nets

Glassbox vs. Blackbox
Performance vs. Interpretability

Interpretability

Rules

Decision Trees

Random Forests

Boosted Trees

XG Boost

Neural Nets

Explainable Boosting Machine

Glassbox vs. Blackbox

Performance
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

<table>
<thead>
<tr>
<th>Feature</th>
<th>Cont</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cat: Allowed</td>
<td>20 000</td>
</tr>
<tr>
<td>Cat: Banned</td>
<td>-15 000</td>
</tr>
<tr>
<td>Apt Size</td>
<td>200 000</td>
</tr>
<tr>
<td>Park: Near</td>
<td>40 000</td>
</tr>
<tr>
<td>Park: Far</td>
<td>-25 000</td>
</tr>
<tr>
<td>Garage Size</td>
<td>100 000</td>
</tr>
<tr>
<td>Nbr of Rooms</td>
<td>45 000</td>
</tr>
</tbody>
</table>
## Evolution of Interpreting Outcomes

<table>
<thead>
<tr>
<th>Feature</th>
<th>Cont</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cat: Allowed</td>
<td>20,000</td>
</tr>
<tr>
<td>Cat: Banned</td>
<td>-15,000</td>
</tr>
<tr>
<td>Apt Size</td>
<td>200,000</td>
</tr>
<tr>
<td>Park: Near</td>
<td>40,000</td>
</tr>
<tr>
<td>Park: Far</td>
<td>-25,000</td>
</tr>
<tr>
<td>Garage Size</td>
<td>100,000</td>
</tr>
<tr>
<td>Nbr of Rooms</td>
<td>45,000</td>
</tr>
</tbody>
</table>

![Graph showing feature value contribution](image-url)
### Evolution of Interpreting Outcomes

<table>
<thead>
<tr>
<th>Feature</th>
<th>Cont</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cat: Allowed</td>
<td>20 000</td>
</tr>
<tr>
<td>Cat: Banned</td>
<td>-15 000</td>
</tr>
<tr>
<td>Apt Size</td>
<td>200 000</td>
</tr>
<tr>
<td>Park: Near</td>
<td>40 000</td>
</tr>
<tr>
<td>Park: Far</td>
<td>-25 000</td>
</tr>
<tr>
<td>Garage Size</td>
<td>100 000</td>
</tr>
<tr>
<td>Nbr of Rooms</td>
<td>45 000</td>
</tr>
</tbody>
</table>

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
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

Graph title

- **WU**: 5%
- **Xoom**: 4%
- **WorldRemit**: 4.5%
- **TW 2017**: 2.2%
- **TW today**: 2%
Title of the page.
Main point goes here.
Pillars.

Price
Subtitle

Convenience
Subtitle

Speed
Subtitle

Transparency
Subtitle
23.5%
if (this) {
    // then that
}