SHAP: Interpreting ML Models with IML

March 7, 2020
SatRday
Who am I?

Drikus du Toit

Data Scientist
Decision Science
Capitec Bank
Resource

What is ML interpretability and why is it important?
The Client

Applies for loan, gets rejected.

Client Questions:
- Why did I not get the loan?
- What should I do to improve my credit score?

The Business

Build a credit default model. High accuracy usually goes with high complexity.

Business Questions:
- Is there bias in our model?
- Do we understand our underlying data?
- What will cause the model to not perform as expected?
- Are we within regulatory framework?
Context..
Context..
Context..

Complex Model

Simple Model

Interpretable

Accurate

Scott Lundberg H2O World
Context..

Complex Model

Simple Model

Interpretable

Accurate

Scott Lundberg H2O World
Context..

Complex Model: **Interpretable** (X)

Simple Model: **Interpretable** (✓)  **Accurate** (✓)

Scott Lundberg H2O World
Context..

<table>
<thead>
<tr>
<th>Complex Model</th>
<th>Simple Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Interpretable</td>
<td>Yes</td>
</tr>
<tr>
<td>Accurate</td>
<td>Not Accurate</td>
</tr>
</tbody>
</table>
Context..

Complex Model

Simple Model

Interpretable

Accurate
Context..

Complex Model

Simple Model

Interpretable

Accurate

Scott Lundberg H2O World
Machine Learning Model

Importance of Interpretability (what or why)

1. Human curiosity and learning
2. Goal of science
3. Safety measures
4. Detecting bias
5. Manage social interactions
6. Debugged and audited
Interpretability Techniques
- **Reveal its internal mechanisms**
- Fully understood by looking at their parameters
- Also called interpretable models

- **Does not reveal its internal mechanisms**
- Cannot be understood by looking at their parameters (e.g. a neural network)
<table>
<thead>
<tr>
<th>Model Agnostic Methods</th>
<th>Example Based Explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partial Dependence Plot (PDP)</td>
<td>Counterfactual Explanations</td>
</tr>
<tr>
<td>Individual Conditional Expectation (ICE)</td>
<td>Adversarial Examples</td>
</tr>
<tr>
<td>Accumulated Local Effects (ALE) Plot</td>
<td>Prototypes and Criticisms</td>
</tr>
<tr>
<td>Feature Interaction</td>
<td>Influential Instances</td>
</tr>
<tr>
<td>Permutation Feature Importance</td>
<td></td>
</tr>
<tr>
<td>Global Surrogate</td>
<td></td>
</tr>
<tr>
<td>Local Surrogate (LIME)</td>
<td></td>
</tr>
<tr>
<td>Scoped Rules (Anchors)</td>
<td></td>
</tr>
<tr>
<td>Shapley Values</td>
<td></td>
</tr>
<tr>
<td>SHAP (SHapley Additive exPlanations)</td>
<td></td>
</tr>
</tbody>
</table>
# Black Box Models (interpretability techniques)

<table>
<thead>
<tr>
<th>Model Agnostic Methods</th>
<th>Example Based Explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partial Dependence Plot (PDP)</td>
<td>Counterfactual Explanations</td>
</tr>
<tr>
<td>Individual Conditional Expectation (ICE)</td>
<td>Adversarial Examples</td>
</tr>
<tr>
<td>Accumulated Local Effects (ALE) Plot</td>
<td>Prototypes and Criticisms</td>
</tr>
<tr>
<td>Feature Interaction</td>
<td>Influential Instances</td>
</tr>
<tr>
<td>Permutation Feature Importance</td>
<td></td>
</tr>
<tr>
<td>Global Surrogate</td>
<td></td>
</tr>
<tr>
<td>Local Surrogate (LIME)</td>
<td></td>
</tr>
<tr>
<td>Scoped Rules (Anchors)</td>
<td></td>
</tr>
<tr>
<td><strong>Shapley Values</strong></td>
<td></td>
</tr>
<tr>
<td><strong>SHAP (SHapley Additive exPlanations)</strong></td>
<td></td>
</tr>
</tbody>
</table>
Shapley Values
Average Pred
15%
Average Pred 15% $E[f(x)]$

Joe 23% $f(x)$

Scott Lundberg H2O World
Average Pred
15%
$E[f(x)]$

Base
$\Phi(0)$
Average Pred 15%  
$E[f(x)]$

Base $\Phi(0)$  
Income not verified $\Phi(1)$

0  
19%

Scott Lundberg H2O World
Average Pred 15\% E[f(x)]

0

Base $\phi(0)$

Income not verified $\phi(1)$

DTI $\geq 30$ $\phi(2)$

19\%

21\%
Average Pred
15%
$E[f(x)]$

0

19%

19%

21%

23%

Base
$\Phi(0)$

Income not verified
$\Phi(1)$

DTI = 30
$\Phi(2)$

Delinquent 10 months ago
$\Phi(3)$

Scott Lundberg H2O World
Average Pred 15% $E[f(x)]$

- Base $\Phi(0)$
- Income not verified $\Phi(1)$
- DTI = 30 $\Phi(2)$
- Delinquent 10 months ago $\Phi(3)$
- No recent account openings $\Phi(4)$
Average Pred 15% $E[f(x)]$

- Base $\Phi(0)$
- Income not verified $\Phi(1)$
- DTI = 30 $\Phi(2)$
- Delinquent 10 months ago $\Phi(3)$
- No recent account openings $\Phi(4)$
- 40 years of credit history $\Phi(5)$
What about the order?

- Base: $\Phi(0)$
- Income not verified: $\Phi(1)$
- DTI = 30: $\Phi(2)$
- Delinquent 10 months ago: $\Phi(3)$
- No recent account openings: $\Phi(4)$
- 40 years of credit history: $\Phi(5)$

Average Pred 15% $E[f(x)]$
What about the order?

- Average Pred: 15% $E[f(x)]$
- Base: $\Phi(0)$
- Income not verified: $\Phi(1)$
- DTI = 30: $\Phi(2)$
- Delinquent 10 months ago: $\Phi(3)$
- 40 years of credit history: $\Phi(5)$
- No recent account openings: $\Phi(4)$
Shapley values results from averaging over all $N!$ possible orderings.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Shapley Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>$\phi(0)$</td>
</tr>
<tr>
<td>Income not verified</td>
<td>$\phi(1)$</td>
</tr>
<tr>
<td>DTI = 30</td>
<td>$\phi(2)$</td>
</tr>
<tr>
<td>Delinquent 10 months ago</td>
<td>$\phi(3)$</td>
</tr>
<tr>
<td>40 years of credit history</td>
<td>$\phi(5)$</td>
</tr>
<tr>
<td>No recent account openings</td>
<td>$\phi(4)$</td>
</tr>
</tbody>
</table>

Average Pred $E[f(x)]$ = 15%
Examples & Interpretation

- First we fit a machine learning model on the Boston housing data.

```r
set.seed(42)
library("randomForest")
data("Boston", package = "MASS")
rf = randomForest(medv ~ ., data = Boston, ntree = 50)
X = Boston[-which(names(Boston) == "medv")]
mod = Predictor$new(rf, data = X)
# Then we explain the first instance of the dataset with the Shapley method:
x.interest = X[1,]
shapley = Shapley$new(mod, x.interest = x.interest)
# plot
plot(shapley)
```

- Actual prediction: 25.75
  Average prediction: 22.56
SHAP (Shapley Additive Explanations)

KernelSHAP

An alternative, kernel-based estimation approach for Shapley values inspired by local surrogate models

TreeSHAP

An efficient estimation approach for tree-based models

SHAP comes with many global interpretation methods based on aggregations of Shapley values
The future of interpretability

The focus will be on model-agnostic interpretability tools

Robots and programs will explain themselves

Ethical Issues


**Package iml**

https://www.youtube.com/watch?v=ngOBhhINWb8


Questions?