Computing Abductive Explanations for Boosted Regression Trees

Gilles Audemard\textsuperscript{1}, \textbf{Steve Bellart}\textsuperscript{1}, Jean-Marie Lagniez\textsuperscript{1} \\
Pierre Marquis\textsuperscript{12}

\textsuperscript{1}Univ. Artois, CNRS, CRIL, Lens, France \\
\textsuperscript{2}Institut Universitaire de France

July 03, 2023
Explainable Artificial Intelligence (XAI) is a subfield of AI that aims to make the decisions made by AI systems transparent and understandable to human users.
Black Box Explainability Tools

- LIME (Local Interpretable Model-Agnostic Explanations)
- SHAP (SHapley Additive exPlanations)
- Partial Dependence Plots (PDP)
- Anchors

All these tools aim to increase the transparency and interpretability of black-box models, though they each have different strengths and limitations.
Main Limitations

- Focus on Classification
- Lack of Robustness
Our Approach

Motivation
Focus on Regression
Robust explanation

Our Approach
We decided to pursue a **model-specific** approach for explaining **boosted regression trees**.
We presented and evaluated two anytime algorithms $G$ and $E$ for generating and evaluating abductive explanations for boosted regression trees.

The datasets used for learning the boosted trees can be based on mixed type data, including categorical and numerical attributes.
Some preliminaries
Attributes

- Set of attributes $\mathcal{A} = \{A_1, \ldots, A_n\}$ with each attribute $A_i$ taking a value in domain $D_i$
- Types of attributes: Numerical, Categorical, Boolean
- Instance $x$ is a vector $(v_1, \ldots, v_n)$ where each $v_i$ is an element of $D_i$.
- Each pair $A_i = v_i$ is a characteristic of the instance $x$.

Example

Let us consider a loan application scenario with $\mathcal{A} = \{A_1, A_2, A_3\}$:

- $A_1$: **Numerical** - income per month
- $A_2$: **Categorical** - employment status: "employed", "unemployed" or "self-employed"
- $A_3$: **Boolean** - is married or not
A regression tree over \( \mathcal{A} \) is a binary tree \( T \), with internal nodes labeled with Boolean conditions on an attribute from \( \mathcal{A} \) and leaves labeled by real numbers.

The value \( T(x) \) of \( T \) for an instance \( x \) is given by the real number labeling the leaf reached from the root.
A boosted regression tree over $\mathcal{A}$ is an ensemble of trees $F = \{T_1, \ldots, T_m\}$, where each $T_i$ is a regression tree over $\mathcal{A}$.

The value $F(x)$ of $F$ for an instance $x$ is given by $F(x) = \sum_{i=1}^{m} T_i(x)$. 
Boolean Conditions and Constraints

- Let $\mathcal{B}$ denote the set of all Boolean conditions used in $F$.
- The Boolean conditions used in $F$ are not necessarily independent.
- Some constraints $\Sigma$ over $\mathcal{B}$ must be exploited to characterize the truth assignments over $\mathcal{B}$. 
Example of Boosted Regression Trees

Example

$F$ is built upon Boolean conditions: $B = \{B_1^1, B_1^2, B_3^1, B_1^2, B_2^2, B_3^2, B_3^3\}$:

- $B_1^1$, $B_2^1$ and $B_3^1$: are respectively $A_1 > 1000\$, $A_1 > 2000\$ and $A_1 > 3000\$.
- $B_1^2$, $B_2^2$ and $B_3^2$: are respectively $A_2 = "employed"$, $A_2 = "unemployed"$ and $A_2 = "self-employed"$.
- $B_3^3$: $A_3 = 1$ (is married).
Definition of Abductive Explanations for Boosted Regression Trees

Let $F$ be a boosted regression tree over $\mathcal{A}$, $x \in X$ an instance, and $I$ an interval over the reals such as $F(x) \in I$.

**Abductive Explanation**

A term $t$ over $\mathcal{B}$ is an *abductive explanation* for $x$ given $F$ and $I$ if and only if $t$ covers $x$ and for every instance $x' \in X$ that is covered by $t$, we have $F(x') \in I$.

**Subset-Minimal Abductive Explanation**

A term $t$ is a *subset-minimal abductive explanation* for $x$ given $F$ and $I$ if and only if $t$ is an abductive explanation for $x$ given $F$ and $I$ and no proper subset of $t$ is an abductive explanation for $x$ given $F$ and $I$.
Example of Subset-Minimal Abductive Explanation

Example of a prediction knowing an instance

Suppose that the applicant is described by $x_{ex} = (2200\$, "self-employed", 1)$. Then, $F(x_{ex}) = 1500 + 250 + 250 = 2000\$.

A simple explanation is then $\{B_1^1, B_2^1, \neg B_3^1, \neg B_1^2, B_3^2, B^3\}$ or in simpler terms $\{B_2^1, \neg B_3^1, B_3^2, B^3\}$.
Example of a subset-minimal abductive explanation knowing an instance ans an interval

On the same applicant, if we consider \( I = [2000, 2250] \) then \( \{B_1^2, B_3^2, B_3^3\} \)
Example of Subset-Minimal Abductive Explanation

\[ t_1 = \{ B_2^1, B_3^1, B_3^2, B_3^3 \} \quad I_{t_1} = [2000, 2000] \]
\[ t_2 = \{ B_3^1, B_3^2, B_3^3 \} \quad I_{t_2} = [500, 2000] \]
\[ t_3 = \{ B_2^1, B_3^2, B_3^3 \} \quad I_{t_3} = [2000, 2250] \]
\[ t_4 = \{ B_2^1, B_3^1, B_3^3 \} \quad I_{t_4} = [1500, 2000] \]
\[ t_5 = \{ B_2^1, B_3^1, B_3^2 \} \quad I_{t_5} = [1850, 2250] \]
\[ t_6 = \top \quad I_{t_6} = [−100, 2500] \]
From boosted trees to MILP
General constraints

Constraints over $B$ encoding the corresponding domain theory $\Sigma$:

$$\forall A_i \in A_N, \forall j \in [k_i - 1], B^i_j - B^i_{j+1} \geq 0$$
$$\forall A_i \in A_C, \forall B^i_j, B^i_k \in \tau(A_i), j \neq k, B^i_j + B^i_k \leq 1$$

(1)

$t$ is represented by:

$$\forall B^i_j \in t, B^i_j = 1$$
$$\forall B^i_j \in t, B^i_j = 0$$

(2)

For each leaf of each tree, we define $L^i_{t^i_j}$ to know if the leaf is active. By definition, only one must be set to true by tree:

$$\forall i \in [m], \sum_{t^i_j \in T_i} L^i_{t^i_j} = 1$$

(3)
General constraints

For all $i \in [m]$, the following set of constraints indicates how each $L_{t_j}^i$ is connected to the Boolean variables of $B$:

$$\forall t_j^i \in T_i, \sum_{B_j^i \in t_j^i} B_j^i + \sum_{\overline{B}_j^i \in t_j^i} (1 - B_j^i) - L_{t_j}^i \leq |t_j^i| - 1$$

(4)

We define each $W_i$ ($i \in [m]$) as:

$$\forall i \in [m], \sum_{j \in [p_i]} L_{t_j}^i \times w_j^i = W_i$$

(5)

Let $FW$ be a continuous variable that represents the value of the regression tree for any truth assignment over $B$:

$$\sum_{W_i \in \mathcal{W}} W_i = FW$$

(6)
Given a non-empty interval $I = (lb, ub)$, we add:

\[(IL = 1) \rightarrow (FW \leq lb)\] \hspace{1cm} (7)
\[(IU = 1) \rightarrow (FW \geq ub)\] \hspace{1cm} (8)

\[IL + IU = 1\] \hspace{1cm} (9)
**Generation: main algorithm**

**Input:** An instance $x$

**Output:** A subset-minimal abductive explanation $t$

$t_{tot} = \text{toBoolean}(x)$

$t = t_{tot}$

**for** cond **in** $t_{tot}$ **do**

- $\text{assignToMILP}(t \setminus \{\text{cond}\})$
- $solution = \text{solveMILP}()$
- $\text{unassignFromMILP}(t \setminus \{\text{cond}\})$
- **if** $solution$ **is** UNSAT **then**
  - $t = t \setminus \{\text{cond}\}$ **end**

**end**
Evaluation: example on a lower bound

- **Initial setup:**
  - $M_e$: a constraint-based model containing all $M_g$ constraints except Equation (9).
  - $lower$: initially set to $F(x_t)$, where $x_t$ satisfies $t \land \Sigma$.
  - $lower_b$: initially set to $m_F = \sum_{i=1}^{n} \min(T_i)$.

- **Binary search strategy to determine or estimate $m_t$:**
  - Compute $mid = \frac{lower + lower_b}{2}$.
  - If $M_e \land (FW \leq mid)$ is inconsistent, $lower_b$ is set to $mid$.
  - If $M_e \land (FW \leq mid)$ is consistent, $lower$ is set to $FW$.

- Repeat the binary search with updated bounds.

- This approach provides a boost to the binary search process.
Experiments
Experimental protocol

<table>
<thead>
<tr>
<th>Name</th>
<th>#A</th>
<th>#N</th>
<th>#C</th>
<th>#B</th>
<th>#I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winequality-red</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>1599</td>
</tr>
<tr>
<td>Winequality-white</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>4898</td>
</tr>
<tr>
<td>CreditcardFraudDet.</td>
<td>29</td>
<td>0</td>
<td>0</td>
<td>29</td>
<td>284807</td>
</tr>
<tr>
<td>l4d2-player-stats-final</td>
<td>112</td>
<td>111</td>
<td>1</td>
<td>0</td>
<td>20830</td>
</tr>
<tr>
<td>Houses-prices</td>
<td>46</td>
<td>26</td>
<td>20</td>
<td>0</td>
<td>2919</td>
</tr>
<tr>
<td>Steel ind. energy cons.</td>
<td>9</td>
<td>6</td>
<td>3</td>
<td>0</td>
<td>35040</td>
</tr>
<tr>
<td>Bike sharing: hour</td>
<td>15</td>
<td>13</td>
<td>0</td>
<td>2</td>
<td>17379</td>
</tr>
<tr>
<td>Bike sharing: daily</td>
<td>13</td>
<td>11</td>
<td>0</td>
<td>2</td>
<td>731</td>
</tr>
<tr>
<td>NASA airfoil self-noise</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1503</td>
</tr>
<tr>
<td>abalone</td>
<td>9</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>4177</td>
</tr>
</tbody>
</table>

\[ I_{F,x}^r = [F(x) - (\frac{r}{100} \cdot L_F), F(x) + (\frac{r}{100} \cdot L_F)] \]

with

\[ L_F = \sum_{i=1}^{n} \max(T_i) - \sum_{i=1}^{n} \min(T_i) \]
Experiments on Generation

Size explanation (nb bool)

Imprecision

0 10 20 30 40 50 60 70

Figure: Empirical results about algorithm G on the houses-prices dataset.
Experiments on Evaluation

Figure: Empirical results about algorithm E on the houses-prices dataset.
Conclusion
Most of the time, our algorithms can generate and evaluate abductive explanations within a few seconds.

The explanations generated using G are generally significantly smaller than the initial instance descriptions.

Notably, E’s reduction of the imprecision can be very significant.
These algorithms don’t require any specific assumption about the learning method of the input regression tree ensemble. Therefore, they are applicable to general machine learning decision tree ensemble models. However, the size of the explanations produced by $G$ can be quite large in certain cases, and even simplified explanations may not be intelligible enough for some users.
Future Directions

- This work sets the stage for focusing on applications where human expertise can be utilized to evaluate the quality of the generated explanations.
- The possibility of computing $l_t$ given $t$ and $F$ can be leveraged to design interaction protocols with an explainee, aiming to provide explanations with a good generality/precision trade-off.