

## **TRAINING TREE-BASED MACHINE-LEARNING MODELING ALGORITHMS FOR PREDICTING OUTPUTS AND GENERATING EXPLANATORY DATA**

### Technical Field

[0001] The present disclosure relates generally to machine learning. More specifically, but not by way of limitation, this disclosure relates to machine learning using tree-based algorithms for emulating intelligence, where the tree-based algorithms are trained for computing predicted outputs (e.g., a risk indicator or other predicted value of a response variable of interest) and generating explanatory data regarding the impact of corresponding independent variables used in the tree-based algorithms.

### Background

[0002] Automated modeling systems can implement tree-based machine-learning modeling algorithms that are fit using a set of training data. This training data, which can be generated by or otherwise indicate certain electronic transactions or circumstances, is analyzed by one or more computing devices of an automated modeling system. The training data includes data samples having values of a certain output, which corresponds to a response variable of interest in the model developed by the automated modeling system, and data samples having values of various predictors, which correspond to independent variables in the model developed by the automated modeling system. The automated modeling system can be used to analyze and learn certain features or patterns from the training data and make predictions from “new” data describing circumstances similar to the training data. For example, the automated modeling system uses, as training data, sample data that contains at least one output and relevant predictors. The automated modeling system uses this training data to learn the process that resulted in the generation of response variables (i.e., the output or other response variable) involving transactions or other circumstances (i.e., the predictors or other independent variables). The learned process can be applied to other data samples similar to the training data, thereby to predicting the response variable in the presence of predictors or independent variables.

### Summary

[0002a] According to a first aspect of the present disclosure, there is provided a system including:

- a processing device; and

- one or more memory devices storing:

- instructions executable by the processing device, and

- a tree-based machine-learning model that is a memory structure comprising interconnected parent nodes and terminal nodes, wherein each parent node includes a respective splitting variable that causes the parent node to be connected via links to a respective pair of child nodes, wherein the terminal nodes includes respective representative response values based on values of the splitting variables;

- wherein the processing device is configured to access the one or more memory

devices and thereby execute the instructions to:

identify independent variables having relationships with a response variable, wherein (i) each independent variable corresponds to an action performed by an entity or an observation of the entity and (ii) the response variable has a set of outcome values associated with the entity,

generate, from one or more splitting rules, the tree-based machine-learning model that includes decision trees for determining a relationship between each independent variable and a predicted response associated with the response variable, wherein the predicted response indicates a predicted behavior associated with the entity,

iteratively adjust the tree-based machine-learning model to enforce monotonicity with respect to the representative response values of the terminal nodes by adjusting one or more decision trees such that one or more of the representative response values are modified and a monotonic relationship exists between each independent variable and the response variable, and output, based on the adjusted tree-based machine-learning model having the monotonic relationship between each independent variable and the response variable, explanatory data indicating relationships between (i) changes in the response variable and (ii) changes in one or more of the independent variables,

wherein the processing device is further configured to:

in a first iteration of adjusting the tree-based machine-learning model:

generate a decision tree by grouping data samples into terminal node regions in accordance with splitting rules,

select a terminal node region of the decision tree,

determine that a monotonicity constraint is violated for a set of representative response values that includes (i) a representative response value for the terminal node region, (ii) representative response values for upper neighboring regions of the terminal node region, and (iii) representative response values for lower neighboring regions of the terminal node region, and

modify, based on the monotonicity constraint being violated, one or more particular representative response values to enforce monotonicity by smoothing among the set of representative response values;

in a second iteration of adjusting the tree-based machine-learning model:

determine that the monotonicity constraint is satisfied for a set of modified representative response values that is generated by modifying the one or more particular representative response values, and

output, based on the monotonicity constraint being satisfied, the decision tree with the modified representative response values for inclusion in the tree-based machine-learning model.

[0002b] According to a second aspect of the present disclosure, there is provided a system including:

a processing device; and

one or more memory devices storing:

instructions executable by the processing device, and

a tree-based machine-learning model that is a memory structure comprising interconnected parent nodes and terminal nodes, wherein each parent node includes a respective splitting variable that causes the parent node to be connected via links to a respective pair of child nodes, wherein the terminal nodes includes respective representative response values based on values of the splitting variables;

wherein the processing device is configured to access the one or more memory devices and thereby execute the instructions to:

identify independent variables having relationships with a response variable, wherein (i) each independent variable corresponds to an action performed by an entity or an observation of the entity and (ii) the response variable has a set of outcome values associated with the entity,

generate, from one or more splitting rules, the tree-based machine-learning model that includes decision trees for determining a relationship between each independent variable and a predicted response associated with the response variable, wherein the predicted response indicates a predicted behavior associated with the entity,

iteratively adjust the tree-based machine-learning model to enforce monotonicity with respect to the representative response values of the terminal nodes by adjusting one or more decision trees such that one or more of the representative response values are modified and a monotonic relationship exists between each independent variable and the response variable, and output, based on the adjusted tree-based machine-learning model having the monotonic relationship between each independent variable and the response variable, explanatory data indicating relationships between (i) changes in the response variable and (ii) changes in one or more of the independent variables,

wherein the processing device is further configured to:

in a first iteration of adjusting the tree-based machine-learning model:

perform a first split in a decision tree by partitioning data samples into a first tree region and a second tree region in accordance with the one or more splitting rules,

compute (i) a first representative response value from first grouped data samples in the first tree region and (ii) a second representative response value from second grouped data samples in the second tree region,

identify a set of representative response values comprising the first representative response value, the second representative response value, representative response values for upper neighboring regions of the first region, representative response values for lower neighboring regions of the first region, representative response values for upper neighboring regions of the second region, and representative response values for lower neighboring

regions of the second region,

determine that a monotonicity constraint is violated for the set of representative response values, and

modify the one or more splitting rules based on the monotonicity constraint being violated;

in a second iteration of adjusting the tree-based machine-learning model:

modify the first split in the decision tree by regrouping the data samples into a first modified tree region and a second modified tree region in accordance with the modified one or more splitting rules;

compute (i) a first modified representative response value from first regrouped data samples in the first modified tree region and (ii) a second modified representative response value from second regrouped data samples in the second modified tree region, wherein the first regrouped data samples differ from the first grouped data samples and the second regrouped data samples differ from the second grouped data samples,

identify a modified set of representative response values comprising the first modified representative response value, the second modified representative response value, representative response values for upper neighboring regions of the first modified region, representative response values for lower neighboring regions of the first modified region, representative response values for upper neighboring regions of the second modified region, and representative response values for lower neighboring regions of the second modified region,

determine that the monotonicity constraint is satisfied for the modified set of representative response values, and

perform a second split in the decision tree based on the monotonicity constraint being satisfied.

[0002c] According to a third aspect of the present disclosure, there is provided a method performed by one or more processing devices, the method including:

accessing a tree-based machine-learning model that is a memory structure comprising interconnected parent nodes and terminal nodes, wherein each parent node includes a respective splitting variable that causes the parent node to be connected via links to a respective pair of child nodes, wherein the terminal nodes includes respective representative response values based on values of the splitting variables;

identifying independent variables having relationships with a response variable, wherein (i) each independent variable corresponds to an action performed by an entity or an observation of the entity and (ii) the response variable has a set of outcome values associated with the entity;

generating, from one or more splitting rules, the tree-based machine-learning model that includes decision trees for determining a relationship between each independent variable and a predicted response associated with the response variable, wherein the predicted

response indicates a predicted behavior associated with the entity;

iteratively adjusting the tree-based machine-learning model to enforce monotonicity with respect to the representative response values of the terminal nodes by adjusting one or more decision trees such that one or more of the representative response values are modified and a monotonic relationship exists between each independent variable and the response variable;

outputting, based on the adjusted tree-based machine-learning model having the monotonic relationship between each independent variable and the response variable, explanatory data indicating relationships between (i) changes in the response variable and (ii) changes in one or more of the independent variables, and

performing, by the one or more processing devices, operations comprising:

in a first iteration of adjusting the tree-based machine-learning model:

generating a decision tree by grouping data samples into terminal node regions in accordance with splitting rules,

computing a set of representative response values for the terminal node regions,

determining that a monotonicity constraint is violated for the set of representative response values, and

modifying, based on the monotonicity constraint being violated, one or more particular representative response values to enforce monotonicity by smoothing among the set of representative response values;

in a second iteration of adjusting the tree-based machine-learning model:

determining that the monotonicity constraint is satisfied for a set of modified representative response values that is generated by modifying the one or more particular representative response values, and

outputting, based on the monotonicity constraint being satisfied, the decision tree with the modified representative response values for inclusion in the tree-based machine-learning model.

[0002d] According to a fourth aspect of the present disclosure, there is provided a method performed by one or more processing devices, the method including:

accessing a tree-based machine-learning model that is a memory structure comprising interconnected parent nodes and terminal nodes, wherein each parent node includes a respective splitting variable that causes the parent node to be connected via links to a respective pair of child nodes, wherein the terminal nodes includes respective representative response values based on values of the splitting variables;

identifying independent variables having relationships with a response variable, wherein (i) each independent variable corresponds to an action performed by an entity or an observation of the entity and (ii) the response variable has a set of outcome values associated with the entity;

generating, from one or more splitting rules, the tree-based machine-learning model

that includes decision trees for determining a relationship between each independent variable and a predicted response associated with the response variable, wherein the predicted response indicates a predicted behavior associated with the entity;

iteratively adjusting the tree-based machine-learning model to enforce monotonicity with respect to the representative response values of the terminal nodes by adjusting one or more decision trees such that one or more of the representative response values are modified and a monotonic relationship exists between each independent variable and the response variable; and

outputting, based on the adjusted tree-based machine-learning model having the monotonic relationship between each independent variable and the response variable, explanatory data indicating relationships between (i) changes in the response variable and (ii) changes in one or more of the independent variables; and

performing, by the one or more processing devices, operations comprising:

in a first iteration of adjusting the tree-based machine-learning model:

performing a first split in a decision tree by partitioning data samples into a first tree region and a second tree region in accordance with the one or more splitting rules,

computing (i) a first representative response value from first grouped data samples in the first tree region and (ii) a second representative response value from second grouped data samples in the second tree region,

identifying a set of representative response values comprising the first representative response value, the second representative response value, a representative response values for a closest lower neighboring region of the first region from a set of lower neighboring regions of the first region, and representative response value for a closest upper neighboring region of the second region from a set of upper neighboring regions of the second region,

determining that a monotonicity constraint is violated for the set of representative response values, and

modifying the one or more splitting rules based on the monotonicity constraint being violated;

in a second iteration of adjusting the tree-based machine-learning model:

modifying the first split in the decision tree by regrouping the data samples into a first modified tree region and a second modified tree region in accordance with the modified one or more splitting rules;

computing (i) a first modified representative response value from first regrouped data samples in the first modified tree region and (ii) a second modified representative response value from second regrouped data samples in the second modified tree region, wherein the first regrouped data samples differ from the first grouped data samples and the second regrouped data samples differ from the second grouped data samples,

identifying a modified set of representative response values comprising the first modified representative response value, the second modified representative

response value, a representative response values for a closest lower neighboring region of the first modified region from a set of lower neighboring regions of the first modified region, and representative response value for a closest upper neighboring region of the second modified region from a set of upper neighboring regions of the second modified region,

determining that the monotonicity constraint is satisfied for the modified set of representative response values, and

performing a second split in the decision tree based on the monotonicity constraint being satisfied.

[0003] Various aspects of the present disclosure involve training tree-based machine-learning models used in automated modeling algorithms. The tree-based machine-learning models can compute a predicted response, e.g. probability of an event or expectation of a response, and generate explanatory data regarding how the independent variables used in the model affect the predicted response. For example, independent variables having relationships

## Claims

1. A system comprising:
  - a processing device; and
  - one or more memory devices storing:
    - instructions executable by the processing device, and
    - a tree-based machine-learning model that is a memory structure comprising interconnected parent nodes and terminal nodes, wherein each parent node includes a respective splitting variable that causes the parent node to be connected via links to a respective pair of child nodes, wherein the terminal nodes includes respective representative response values based on values of the splitting variables;
  - wherein the processing device is configured to access the one or more memory devices and thereby execute the instructions to:
    - identify independent variables having relationships with a response variable, wherein (i) each independent variable corresponds to an action performed by an entity or an observation of the entity and (ii) the response variable has a set of outcome values associated with the entity,
    - generate, from one or more splitting rules, the tree-based machine-learning model that includes decision trees for determining a relationship between each independent variable and a predicted response associated with the response variable, wherein the predicted response indicates a predicted behavior associated with the entity,
    - iteratively adjust the tree-based machine-learning model to enforce monotonicity with respect to the representative response values of the terminal nodes by adjusting one or more decision trees such that one or more of the representative response values are modified and a monotonic relationship exists between each independent variable and the response variable, and output, based on the adjusted tree-based machine-learning model having the monotonic relationship between each independent variable and the response variable, explanatory data indicating relationships between (i) changes in the response variable and (ii) changes in one or more of the independent variables,
  - wherein the processing device is further configured to:
    - in a first iteration of adjusting the tree-based machine-learning model:
      - generate a decision tree by grouping data samples into terminal node regions in accordance with splitting rules,
      - select a terminal node region of the decision tree,
      - determine that a monotonicity constraint is violated for a set of representative response values that includes (i) a representative response value for the terminal node region, (ii) representative response values for upper neighboring regions of the terminal node region, and (iii) representative response values for lower neighboring regions of the terminal node region, and
      - modify, based on the monotonicity constraint being violated, one or more particular representative response values to enforce monotonicity by

smoothing among the set of representative response values;  
 in a second iteration of adjusting the tree-based machine-learning model:

determine that the monotonicity constraint is satisfied for a set of modified representative response values that is generated by modifying the one or more particular representative response values, and

output, based on the monotonicity constraint being satisfied, the decision tree with the modified representative response values for inclusion in the tree-based machine-learning model.

2. A system comprising:

a processing device; and

one or more memory devices storing:

instructions executable by the processing device, and

a tree-based machine-learning model that is a memory structure comprising interconnected parent nodes and terminal nodes, wherein each parent node includes a respective splitting variable that causes the parent node to be connected via links to a respective pair of child nodes, wherein the terminal nodes includes respective representative response values based on values of the splitting variables;

wherein the processing device is configured to access the one or more memory devices and thereby execute the instructions to:

identify independent variables having relationships with a response variable, wherein (i) each independent variable corresponds to an action performed by an entity or an observation of the entity and (ii) the response variable has a set of outcome values associated with the entity,

generate, from one or more splitting rules, the tree-based machine-learning model that includes decision trees for determining a relationship between each independent variable and a predicted response associated with the response variable, wherein the predicted response indicates a predicted behavior associated with the entity,

iteratively adjust the tree-based machine-learning model to enforce monotonicity with respect to the representative response values of the terminal nodes by adjusting one or more decision trees such that one or more of the representative response values are modified and a monotonic relationship exists between each independent variable and the response variable, and output, based on the adjusted tree-based machine-learning model having the monotonic relationship between each independent variable and the response variable, explanatory data indicating relationships between (i) changes in the response variable and (ii) changes in one or more of the independent variables,

wherein the processing device is further configured to:

in a first iteration of adjusting the tree-based machine-learning model:

perform a first split in a decision tree by partitioning data samples into a first tree region and a second tree region in accordance with the one or more splitting rules,

compute (i) a first representative response value from first grouped data samples in the first tree region and (ii) a second representative response value from second grouped data samples in the second tree region,

identify a set of representative response values comprising the first representative response value, the second representative response value, representative response values for upper neighboring regions of the first region, representative response values for lower neighboring regions of the first region, representative response values for upper neighboring regions of the second region, and representative response values for lower neighboring regions of the second region,

determine that a monotonicity constraint is violated for the set of representative response values, and

modify the one or more splitting rules based on the monotonicity constraint being violated;

in a second iteration of adjusting the tree-based machine-learning model:

modify the first split in the decision tree by regrouping the data samples into a first modified tree region and a second modified tree region in accordance with the modified one or more splitting rules;

compute (i) a first modified representative response value from first regrouped data samples in the first modified tree region and (ii) a second modified representative response value from second regrouped data samples in the second modified tree region, wherein the first regrouped data samples differ from the first grouped data samples and the second regrouped data samples differ from the second grouped data samples,

identify a modified set of representative response values comprising the first modified representative response value, the second modified representative response value, representative response values for upper neighboring regions of the first modified region, representative response values for lower neighboring regions of the first modified region, representative response values for upper neighboring regions of the second modified region, and representative response values for lower neighboring regions of the second modified region,

determine that the monotonicity constraint is satisfied for the modified set of representative response values, and

perform a second split in the decision tree based on the monotonicity constraint being satisfied.

3. The system of claim 1, wherein modifying the one or more particular representative response values comprises:

computing an average of a first particular representative response value for a first terminal node region and a second particular representative response value for a second terminal node region; and

modifying the decision tree such that the computed average is a first modified

representative response value for the first terminal node region and the computed average is a second modified representative response value for the second terminal node region.

4. The system of claim 1, wherein the monotonicity constraint comprises an increase in independent variable values corresponding to an increase in response variable values or a decrease in independent variable values corresponding to a decrease in response variable values.

5. The system of claim 1, wherein the processing device is further configured to generate the explanatory data by determining, using the adjusted tree-based machine-learning model, a rank of each independent variable indicating an impact of the independent variable on the predicted response.

6. A method performed by one or more processing devices, the method comprising:  
accessing a tree-based machine-learning model that is a memory structure comprising interconnected parent nodes and terminal nodes, wherein each parent node includes a respective splitting variable that causes the parent node to be connected via links to a respective pair of child nodes, wherein the terminal nodes includes respective representative response values based on values of the splitting variables;

identifying independent variables having relationships with a response variable, wherein (i) each independent variable corresponds to an action performed by an entity or an observation of the entity and (ii) the response variable has a set of outcome values associated with the entity;

generating, from one or more splitting rules, the tree-based machine-learning model that includes decision trees for determining a relationship between each independent variable and a predicted response associated with the response variable, wherein the predicted response indicates a predicted behavior associated with the entity;

iteratively adjusting the tree-based machine-learning model to enforce monotonicity with respect to the representative response values of the terminal nodes by adjusting one or more decision trees such that one or more of the representative response values are modified and a monotonic relationship exists between each independent variable and the response variable;

outputting, based on the adjusted tree-based machine-learning model having the monotonic relationship between each independent variable and the response variable, explanatory data indicating relationships between (i) changes in the response variable and (ii) changes in one or more of the independent variables, and

performing, by the one or more processing devices, operations comprising:

in a first iteration of adjusting the tree-based machine-learning model:

generating a decision tree by grouping data samples into terminal node regions in accordance with splitting rules,

computing a set of representative response values for the terminal node regions,

determining that a monotonicity constraint is violated for the set of

representative response values, and

modifying, based on the monotonicity constraint being violated, one or more particular representative response values to enforce monotonicity by smoothing among the set of representative response values;

in a second iteration of adjusting the tree-based machine-learning model:

determining that the monotonicity constraint is satisfied for a set of modified representative response values that is generated by modifying the one or more particular representative response values, and

outputting, based on the monotonicity constraint being satisfied, the decision tree with the modified representative response values for inclusion in the tree-based machine-learning model.

7. A method performed by one or more processing devices, the method comprising:

accessing a tree-based machine-learning model that is a memory structure comprising interconnected parent nodes and terminal nodes, wherein each parent node includes a respective splitting variable that causes the parent node to be connected via links to a respective pair of child nodes, wherein the terminal nodes includes respective representative response values based on values of the splitting variables;

identifying independent variables having relationships with a response variable, wherein (i) each independent variable corresponds to an action performed by an entity or an observation of the entity and (ii) the response variable has a set of outcome values associated with the entity;

generating, from one or more splitting rules, the tree-based machine-learning model that includes decision trees for determining a relationship between each independent variable and a predicted response associated with the response variable, wherein the predicted response indicates a predicted behavior associated with the entity;

iteratively adjusting the tree-based machine-learning model to enforce monotonicity with respect to the representative response values of the terminal nodes by adjusting one or more decision trees such that one or more of the representative response values are modified and a monotonic relationship exists between each independent variable and the response variable; and

outputting, based on the adjusted tree-based machine-learning model having the monotonic relationship between each independent variable and the response variable, explanatory data indicating relationships between (i) changes in the response variable and (ii) changes in one or more of the independent variables; and

performing, by the one or more processing devices, operations comprising:

in a first iteration of adjusting the tree-based machine-learning model:

performing a first split in a decision tree by partitioning data samples into a first tree region and a second tree region in accordance with the one or more splitting rules,

computing (i) a first representative response value from first grouped data samples in the first tree region and (ii) a second representative response value from second grouped data samples in the second tree region,

identifying a set of representative response values comprising the first representative response value, the second representative response value, a representative response values for a closest lower neighboring region of the first region from a set of lower neighboring regions of the first region, and representative response value for a closest upper neighboring region of the second region from a set of upper neighboring regions of the second region,

determining that a monotonicity constraint is violated for the set of representative response values, and

modifying the one or more splitting rules based on the monotonicity constraint being violated;

in a second iteration of adjusting the tree-based machine-learning model:

modifying the first split in the decision tree by regrouping the data samples into a first modified tree region and a second modified tree region in accordance with the modified one or more splitting rules;

computing (i) a first modified representative response value from first regrouped data samples in the first modified tree region and (ii) a second modified representative response value from second regrouped data samples in the second modified tree region, wherein the first regrouped data samples differ from the first grouped data samples and the second regrouped data samples differ from the second grouped data samples,

identifying a modified set of representative response values comprising the first modified representative response value, the second modified representative response value, a representative response values for a closest lower neighboring region of the first modified region from a set of lower neighboring regions of the first modified region, and representative response value for a closest upper neighboring region of the second modified region from a set of upper neighboring regions of the second modified region,

determining that the monotonicity constraint is satisfied for the modified set of representative response values, and

performing a second split in the decision tree based on the monotonicity constraint being satisfied.

8. The method of claim 6, wherein modifying the one or more particular representative response values comprises:

computing an average of a first particular representative response value for a first terminal node region and a second particular representative response value for a second terminal node region; and

modifying the decision tree such that the computed average is a first modified representative response value for the first terminal node region and the computed average is a second modified representative response value for the second terminal node region.

9. The method of claim 6, wherein the monotonicity constraint comprises an increase in independent variable values corresponding to an increase in response variable values or a

decrease in independent variable values corresponding to a decrease in response variable values.

10. The method of claim 6, further comprising generating, by the one or more processing devices, the explanatory data by determining, using the adjusted tree-based machine-learning model, a rank of each independent variable indicating an impact of the independent variable on the predicted response.

11. The method of claim 2, wherein the monotonicity constraint comprises an increase in independent variable values corresponding to an increase in response variable values or a decrease in independent variable values corresponding to a decrease in response variable values.

12. The method of claim 2, wherein the processing device is further configured to generate the explanatory data by determining, using the adjusted tree-based machine-learning model, a rank of each independent variable indicating an impact of the independent variable on the predicted response.

13. The method of claim 7, wherein the monotonicity constraint comprises an increase in independent variable values corresponding to an increase in response variable values or a decrease in independent variable values corresponding to a decrease in response variable values.

14. The method of claim 7, further comprising, generating by one or more processing devices, the explanatory data by determining, using the adjusted tree-based machine-learning model, a rank of each independent variable indicating an impact of the independent variable on the predicted response.