

CLAIMS

1. A method, comprising:

classifying input data associated with a machine learning model into a plurality of clusters;
generating a plurality of linear surrogate models and a plurality of non-linear surrogate

models, wherein one of the plurality of linear surrogate models corresponds to one of the plurality
of clusters, wherein ~~a~~each linear surrogate model is configured to output a corresponding prediction
based on input data associated with a corresponding cluster, wherein each of the plurality of linear
surrogate models and each of the non-linear surrogate models approximate an output of the
machine learning model; ~~and~~

outputting ~~the~~ prediction data associated with the machine learning model, prediction data
associated with the plurality of linear surrogate models, and prediction data associated with the
plurality of non-linear surrogate models;

debugging each of the plurality of linear surrogate models until the prediction data
associated with each of the plurality of linear surrogate models correlates with a corresponding
non-linear surrogate model;

receiving production data, wherein the production data comprises one or more entries;

determining a corresponding cluster of the plurality of clusters for the one or more entries
based at least in part on a corresponding distance between a corresponding feature space location
for the one or more entries and a corresponding feature space location of a cluster centroid;

determining a corresponding linear surrogate model of the plurality of linear surrogate
models for each of the one or more entries based on the determined corresponding cluster; and

outputting, using the corresponding linear surrogate model, prediction data associated with
the one or more entries.

2. The method of claim 1, further comprising receiving the input data associated with the
machine learning model.

3. The method of claim 1, wherein the input data associated with the machine learning model
comprises one or more entries, wherein the one or more entries are sorted into training data and
validation data, wherein each entry of the one or more entries is associated with one or more
features having corresponding feature values, a corresponding prediction label, and a corresponding
actual outcome.

4. The method of claim 1, wherein the input data associated with the machine learning model
is classified into the plurality of clusters using a k-means clustering technique.

5. The method of claim 1, further comprising ranking the prediction data associated with the plurality of linear surrogate models.

6. The method of claim 5, wherein outputting the prediction data associated with the machine learning model and prediction data associated with the plurality of linear surrogate models includes plotting the ranked prediction data associated with the plurality of linear surrogate models with respect to a corresponding prediction label.

7. The method of claim 1, further comprising receiving a selection of a data point of the prediction data associated with the plurality of linear surrogate models.

8. The method of claim 7, in response to receiving the selection of the data point of the prediction data associated with the plurality of linear surrogate models, providing one or more reason codes associated with a prediction value associated with the data point.

9. The method of claim 8, wherein the one or more reason codes associated with the prediction value associated with the data point indicate a top threshold number of reasons the corresponding linear surrogate model made a prediction associated with the selected data point.

10. The method of claim 8, wherein the one or more reason codes have a corresponding contribution value.

11. The method of claim 10, wherein a sum of contribution values associated with the one or more reason codes is equal to a prediction value associated with the data point.

12. The method of claim 8, wherein the one or more reason codes correspond to one or more features associated with the input data.

13. The method of claim 1, further comprising generating a global surrogate model of the machine learning model based at least in part on the input data associated with the machine learning model.

~~14. The method of claim 1, further comprising:~~

~~receiving production data, wherein the production data comprises at least one entry;~~

~~determining a cluster of the plurality of clusters for the at least one entry based at least in part on a centroid associated with the cluster;~~

~~determine a linear surrogate model corresponding to the determined cluster; and~~

~~output, using the determined linear surrogate model, prediction data associated with the at least one entry.~~

~~14.~~ 14. The method of claim 1, wherein the input data associated with the machine learning model

comprises one or more entries, the method further comprising sorting the one or more entries of the input data into one or more groups, wherein a group corresponds to one of the plurality of clusters, wherein an entry is associated with a group based at least in part on a distance between feature values associated with the entry and a cluster centroid associated with the group.

5 ~~16,15.~~ The method of claim ~~15~~14, wherein a linear surrogate model of the plurality of linear surrogate models is trained using one or more entries associated with one of the one or more groups.

~~17,16.~~ The method of claim 1, wherein the plurality of linear surrogate models are trained to predict an actual value associated with the machine learning model.

10 ~~18,17.~~ A system, comprising:

a processor configured to:

classify input data associated with a machine learning model into a plurality of clusters;

15 generate a plurality of linear surrogate models and a plurality of non-linear surrogate models, wherein one of the plurality of linear surrogate models corresponds to one of the plurality of clusters, wherein ~~each~~ linear surrogate model is configured to output a corresponding prediction based on input data associated with a corresponding cluster, wherein each of the plurality of linear surrogate models and each of the non-linear surrogate models approximate an output of the machine learning model; and

20 output ~~the~~ prediction data associated with the machine learning model, prediction data associated with the plurality of linear surrogate models, and prediction data associated with the plurality of non-linear surrogate models;

25 debug each of the plurality of linear surrogate models until the prediction data associated with each of the plurality of linear surrogate models correlates with a corresponding non-linear surrogate model;

receive production data, wherein the production data comprises one or more entries;
determine a corresponding cluster of the plurality of clusters for the one or more entries based at least in part on a corresponding distance between a corresponding feature space location for the one or more entries and a corresponding feature space location of a
30 cluster centroid;

determine a corresponding linear surrogate model of the plurality of linear surrogate models for each of the one or more entries based on the determined corresponding cluster;
and

output, using the corresponding linear surrogate model, prediction data associated with the one or more entries; and

a memory coupled to the processor and configured to provide the processor with instructions.

5 ~~19.18.~~ The system of claim ~~17~~, wherein the processor is further configured to receive the input data associated with the machine learning model.

~~20.19.~~ A computer program product, the computer program product being embodied in a non-transitory computer readable storage medium and comprising computer instructions for:

classifying input data associated with a machine learning model into a plurality of clusters;

10 generating a plurality of linear surrogate models and a plurality of non-linear surrogate models, wherein one of the plurality of linear surrogate models corresponds to one of the plurality of clusters, wherein ~~a~~each linear surrogate model is configured to output a corresponding prediction based on input data associated with a corresponding cluster, wherein each of the plurality of linear surrogate models and each of the non-linear surrogate models approximate an output of the

15 machine learning model; and

outputting ~~the~~ prediction data associated with the machine learning model, prediction data associated with the plurality of linear surrogate models, and prediction data associated with the plurality of non-linear surrogate models;

debugging each of the plurality of linear surrogate models until the prediction data

20 associated with each of the plurality of linear surrogate models correlates with a corresponding non-linear surrogate model;

receiving production data, wherein the production data comprises one or more entries;

determining a corresponding cluster of the plurality of clusters for the one or more entries based at least in part on a corresponding distance between a corresponding feature space location

25 for the one or more entries and a corresponding feature space location of a cluster centroid;

determining a corresponding linear surrogate model of the plurality of linear surrogate models for each of the one or more entries based on the determined corresponding cluster; and

outputting, using the corresponding linear surrogate model, prediction data associated with the one or more entries.