

### 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, wherein one of the plurality of linear
  - 5 surrogate models corresponds to one of the plurality of clusters, wherein a linear surrogate model is configured to output a corresponding prediction based on input data associated with a corresponding cluster; and
  - outputting the prediction data associated with the machine learning model and prediction data associated with the plurality of linear surrogate models.
- 10 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
- 15 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
- 20 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.
- 25 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.
- 30 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.

15. 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.

16. The method of claim 15, 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. 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.

18. A system, comprising:  
a processor configured to:  
classify input data associated with a machine learning model into a plurality of clusters;  
generate a plurality of linear surrogate models, wherein one of the plurality of linear

surrogate models corresponds to one of the plurality of clusters, wherein a linear surrogate model is configured to output a corresponding prediction based on input data associated with a corresponding cluster; and

5           output the prediction data associated with the machine learning model and  
prediction data associated with the plurality of linear surrogate models; and  
a memory coupled to the processor and configured to provide the processor with  
instructions.

19.   The system of claim 18, wherein the processor is further configured to receive the input data associated with the machine learning model.

10   20.   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;  
          generating a plurality of linear surrogate models, wherein one of the plurality of linear surrogate models corresponds to one of the plurality of clusters, wherein a linear surrogate model is  
15   configured to output a corresponding prediction based on input data associated with a corresponding cluster; and  
          outputting the prediction data associated with the machine learning model and prediction data associated with the plurality of linear surrogate models.

21.   A method, comprising:  
20       receiving an indication of a selection of an entry associated with a machine learning model;  
and  
          dynamically updating one or more interpretation views associated with one or more machine learning models based on the selected entry.

22.   The method of claim 21, wherein the one or more machine learning models include one or  
25   more non-linear models.

23.   The method of claim 22, wherein one of the one or more non-linear surrogate models includes a feature importance model.

24.   The method of claim 23, wherein the feature importance model is configured to output one or more features, wherein the one or more features have a corresponding global feature importance  
30   value and a corresponding local feature importance value.

25.   The method of claim 24, wherein the corresponding global feature importance value associated with a feature is based at least in part on a number of times the feature is used in a

random forest model.

26. The method of claim 25, wherein the corresponding global feature importance value associated with the feature is based at least in part on a level of the random forest model that the feature was used to split the random forest model.

5 27. The method of claim 24, wherein the corresponding local feature importance value is computed using a leave-one-covariate out mechanism.

28. The method of claim 24, further comprising:  
comparing the corresponding global feature importance value associated with a feature with the corresponding local feature importance value associated with the feature; and

10 determining whether a difference between the corresponding global feature importance value associated with the feature and the corresponding local feature importance value is greater than or equal a threshold value.

29. The method of claim 28, in response to determining that the difference between the corresponding global feature importance value associated with the feature and the corresponding  
15 local feature importance value is greater than or equal to a threshold value, investigating the feature importance model.

30. The method of claim 28, in response to determining that the difference between the corresponding global feature importance value associated with the feature and the corresponding  
20 local feature importance value is less than a threshold value, forgoing an investigation of the feature importance model.

31. The method of claim 22, wherein one of the one or more non-linear surrogate models includes a decision tree surrogate model.

32. The method of claim 31, wherein a plurality of branches associated with the decision tree surrogate model are based on input data associated with the machine learning model, wherein the  
25 input data associated with the machine learning model includes a plurality of entries, wherein each entry has a one or more features and one or more corresponding feature values.

33. The method of claim 31, wherein dynamically updating the one or more interpretation views associated with the one or more machine learning models includes highlighting a path of the decision tree surrogate model, wherein the highlighted path is specific to the selected entry.

30 34. The method of claim 31, wherein a width of a path of the decision tree surrogate model indicates a frequency of which the path is used by the decision tree surrogate model.

35. The method of claim 32, wherein one of the one or more non-linear surrogate models includes a partial dependence plot.

36. The method of claim 35, wherein the partial dependence plot indicates a dependence of a prediction label of the partial dependence plot on a feature having a particular value.

5 37. The method of claim 35, wherein the partial dependence plot indicates an average prediction label based on all entries associated with the partial dependence plot having a corresponding feature with a same particular value.

38. The method of claim 31, wherein the one or more interpretation views associated with one or more machine learning models includes a view associated with a feature importance surrogate  
10 model, a view associated with a decision tree surrogate model, and a view associated with a partial dependence plot.

39. A system, comprising:  
a processor configured to:  
receive an indication of a selection of an entry associated with a machine learning  
15 model; and  
dynamically update one or more interpretation views associated with one or more machine learning models based on the selected entry; and  
a memory coupled to the processor and configured to provide the processor with instructions.

20 40. A computer program product, the computer program product being embodied in a non-transitory computer readable storage medium and comprising computer instructions for:  
receiving an indication of a selection of an entry associated with a machine learning model;  
and  
dynamically updating one or more interpretation views associated with one or more  
25 machine learning models based on the selected entry..