

CLAIMS

1. A computer-implemented method for generating performance predictions of a machine learning (ML) model, comprising:
 - obtaining a first model configured to perform a task and a production dataset comprising unlabeled data;
 - generating, using a second model, at least one metric predicting a performance of the first model at performing the task on the production dataset, wherein the second model is a meta-model associated with the first model;
 - generating, using one or more third models, at least one value predicting an uncertainty of the at least one metric predicting the performance of the first model at performing the task on the production dataset, wherein each of the one or more third models is a meta-meta-model associated with the second model; and
 - providing an indication of the at least one metric predicting the performance of the first model and the at least one value predicting the uncertainty of the at least one metric.
2. The computer-implemented method of claim 1, wherein the first model is a black-box model.
3. The computer-implemented method of claim 1, wherein the first model is a white-box model.
4. The computer-implemented method of claim 1, wherein at least one of the one or more third models is trained on (i) a training dataset for the first model and (ii) the second model.
5. The computer-implemented method of claim 1, wherein at least one of the one or more third models is trained on (i) a training dataset for the first model, (ii) one or more additional datasets, and (iii) the second model.
6. The computer-implemented method of claim 1, wherein at least one of the one or more third models is trained on (i) one or more additional datasets and (ii) the second model.
7. The computer-implemented method of claim 6, wherein the at least one of the one or more third models is trained offline on the one or more additional datasets and the second model.
8. The computer-implemented method of claim 6, wherein:
 - the one or more additional datasets comprises at least one dataset that includes a first set of labels for classifying data in the at least one dataset; and
 - the first set of labels are different from a second set of labels for classifying data in a dataset used to train the second model.

9. The computer-implemented method of claim 6, wherein the one or more additional datasets comprises at least one labeled dataset, the computer-implemented method further comprising:
generating a plurality of testing datasets and a plurality of production datasets, based on the labeled dataset; and
including the plurality of testing datasets and the plurality of production datasets in the one or more additional datasets.
10. The computer-implemented method of claim 9, wherein generating the plurality of testing datasets and the plurality of production datasets comprises resampling one or more features of the labeled dataset, such that each testing dataset and corresponding production dataset includes a different proportion of the one or more features relative to the labeled dataset.
11. The computer-implemented method of claim 6, further comprising generating at least one feature value for training the one or more third models, based at least in part on a first statistical distribution of a first dataset in the one or more additional datasets and a second statistical distribution of a second dataset in the one or more additional datasets, wherein the at least one of the one or more third models is further trained with the one or more feature values.
12. The computer-implemented method of claim 11, wherein the at least one feature value is further generated based on a divergence function between the first statistical distribution and the second statistical distribution.
13. The computer-implemented method of claim 11, wherein the at least one feature value is further generated based on a statistical hypothesis test between the first statistical distribution and the second statistical distribution.
14. The computer-implemented method of claim 1, wherein the one or more third models are a subset of a plurality of third models, the computer-implemented method further comprising selecting the one or more third models from the plurality of third models, based at least in part on one or more features of the one or more additional datasets.
15. The computer-implemented method of claim 1, wherein the at least one value predicting the uncertainty of the at least one metric comprises an interval range.
16. The computer-implemented method of claim 15, wherein the interval range is an assymetric interval range comprising a first upper range and a second lower range.
17. The computer-implemented method of claim 16, wherein the first upper range and the second lower range are generated via one of the one or more third models.

18. The computer-implemented method of claim 16, wherein:
the one or more third models comprise a plurality of third models;
the first upper range is generated via a first of the plurality of third models; and
the second lower range is generated via a second of the plurality of third models.
19. The computer-implemented method of claim 16, wherein:
the first upper range is generated via a first component of one of the one or more third models; and
the second lower range is generated via a second component of the one of the one or more third models.
20. A system comprising:
one or more computer processors; and
a memory containing a program, which when executed by the one or more computer processors performs an operation for generating performance predictions of a machine learning (ML) model, the operation comprising:
obtaining a first model configured to perform a task and a production dataset comprising unlabeled data;
generating, using a second model, at least one metric predicting a performance of the first model at performing the task on the production dataset, wherein the second model is a meta-model associated with the first model;
generating, using a third model, at least one value predicting an uncertainty of the at least one metric predicting the performance of the first model at performing the task on the production dataset, wherein the third model is a meta-meta-model associated with the second model; and
providing an indication of the at least one metric predicting the performance of the first model and the at least one value predicting the uncertainty of the at least one metric.
21. A computer-readable storage medium having computer-readable program code embodied therewith, the computer-readable program code executable by one or more computer processors to perform an operation for generating performance predictions of a machine learning (ML) model comprising:
obtaining a first model configured to perform a task and a production dataset comprising unlabeled data;
generating, using a second model, at least one metric predicting a performance of the first model at performing the task on the production dataset, wherein the second model is a meta-model associated with the first model;
generating, using a third model, at least one value predicting an uncertainty of the at least one metric predicting the performance of the first model at performing the task on the production dataset, wherein the third model is a meta-meta-model associated with the second model; and
providing an indication of the at least one metric predicting the performance of the first model and the at least one value predicting the uncertainty of the at least one metric.