AutoML Web Service Input and Ouput

Sam Johnson and Josh Elms

January 2023

1 Required Input

1.1 Data File

Currently, our program requires data in CSV format (with column names and prediction label in far right column), but we could build out our program to accommodate alternate file types.

1.2 Regressors

The user should be able to specify which of the available 43 regressors they would like to have included in their run. We should include some specific warnings about the time requirements of some of the regressors, and we should include an "all regressors" and "all suggested regressors" option.

The available regressors are as follows:
ARDRegression, AdaBoostRegressor, BaggingRegressor, BayesianRidge, CCA, DecisionTreeRegressor, DummyRegressor, ElasticNet, ExtraTreeRegressor, ExtraTreesRegressor, GammaRegressor, GaussianProcessRegressor, GradientBoostingRegressor, HistGradientBoostingRegressor, HuberRegressor, IsotonicRegression, KNeighborsRegressor, KernelRidge, Lars, Lasso, LassoLars, LassoLarsIC, LinearRegression, LinearSVR, MLPRegressor, MultiTaskElasticNet, MultiTaskLasso, NuSVR, OrthogonalMatchingPursuit, PLSCanonical, PLSRegression, PassiveAggressiveRegressor, PoissonRegressor, QuantileRegressor, RANSACRegressor, RadiusNeighborsRegressor, RandomForestRegressor, Ridge, SGDRegressor, SVR, TheilSenRegressor, TransformedTargetRegressor, and TweedieRegressor

2 Optional Input

2.1 Metrics to be Visualized

The user should have the freedom to pick metrics on which their models are scored and later visualized *Default: RMSE, MSE, MAE,* R^2

2.2 Metric Score Method

This metric will be the score on which the best model from each algorithm's CV run will be selected. *Default: RMSE*

2.3 Test Set Size

The proportion of the data reserved for testing the models. There should be text that recommends a range of 0.1 to 0.4. *Default: 0.2*

2.4 Number of CV folds

There should be text that recommends a range of 4 to 15 and warns of increased runtimes with more CV folds *Default: 10*

3 Graphing Input

These parameters will be optional input if we decide to handle visualization on the backend of the service. If visualization will be handled by the frontend of the website, we will remove the visualization module of our AutoML program

3.1 Top 'n' regressors visualized

The user should be able to determine how many of the regressors they want to include in their visualizations. *Default: 20*

3.2 Top 'n' models reported in table

The user should be able to determine how many of the models they want to include in the table that displays the test scores of the best-performing models. *Default: 10*

3.3 Style Dictionary

This container would store the style preferences of the user for their boxplot visualizations. The dictionary would require the following inputs:

- Box style - line style, line width, color
- Flier style - marker, marker face color, marker size, line style
- Median line style - line style, line width, color
- Whisker style - line style, line width, color
- Cap style - line width, color
- Box fill color

- Show grid T/F
- Dots per Inch

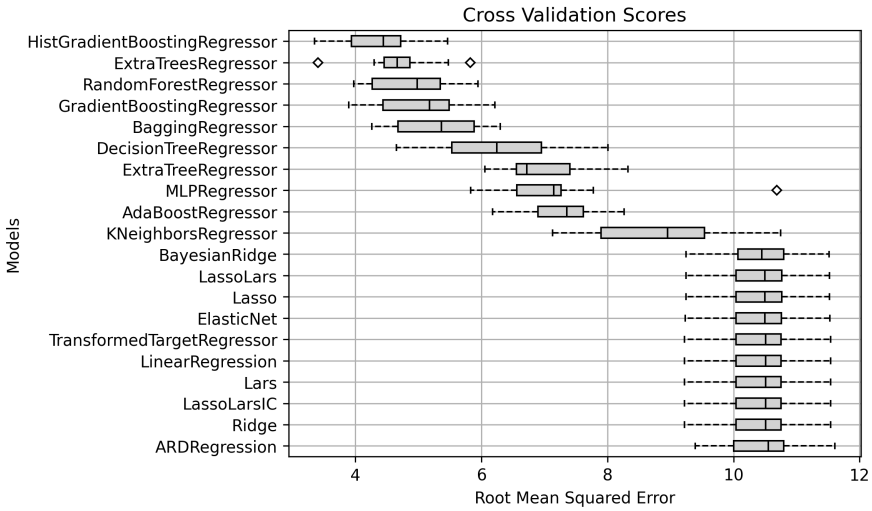
4 Output

4.1 Raw Data

We can deliver the data from our cross-validation runs and test predictions back to the user in either CSV or JSON format.

4.2 Boxplots

Using matplotlib, we currently construct boxplots for the performance of each regressor over the CV runs. We return a boxplot for every metric the user chooses. Example below.



4.3 Model Test Performance Table

After each algorithm has been trained using CV, we select the best-performing model of each algorithm to predict the instances of the test set. The performance on this prediction task is cataloged in a table and return. Example below.

	Root Mean Squared Error	Mean Squared Error	Mean Absolute Error	R-Squared
HistGradientBoostingRegressor	5.2231	27.2803	3.1768	0.9017
ExtraTreesRegressor	5.678	32.2399	3.272	0.8839
GradientBoostingRegressor	5.8767	34.5351	4.0895	0.8756
RandomForestRegressor	5.989	35.8682	3.8862	0.8708
BaggingRegressor	6.2613	39.2043	4.1124	0.8588
DecisionTreeRegressor	7.1254	50.7707	4.4298	0.8171
ExtraTreeRegressor	7.179	51.5385	4.7049	0.8144
MLPRegressor	7.8147	61.069	5.6595	0.78
AdaBoostRegressor	8.208	67.3716	6.4065	0.7573
KNeighborsRegressor	10.3257	106.6207	7.665	0.6159