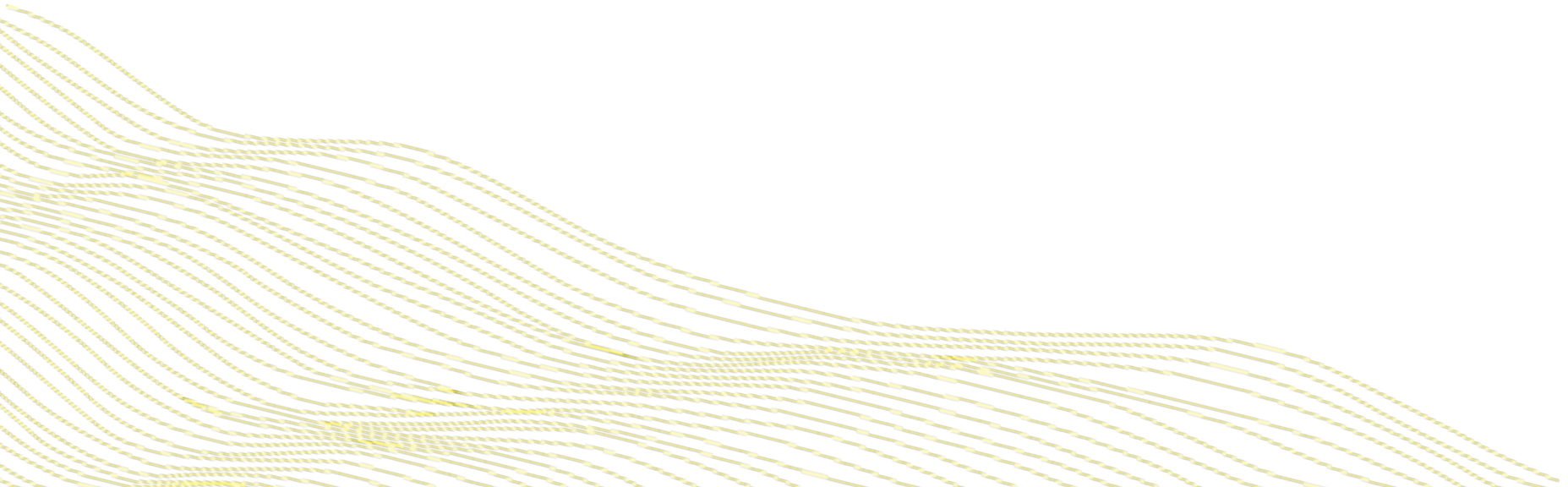
Automatic & Explainable Machine Learning with H2O

USF Seminar Series in Data Science
February 2021

Erin LeDell Ph.D.
@ledell

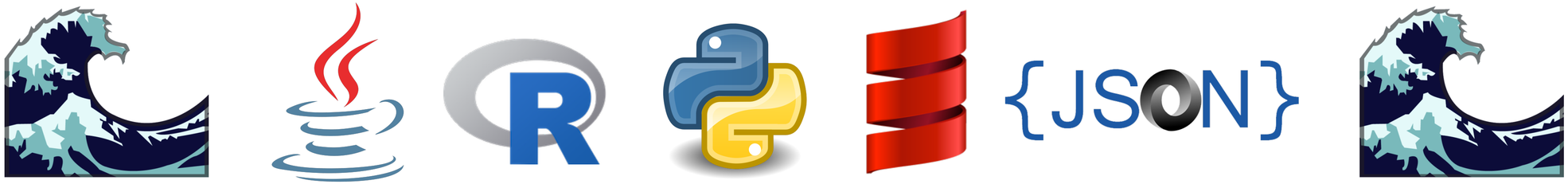
H2O.ai

H2O Platform



H2O Machine Learning Platform

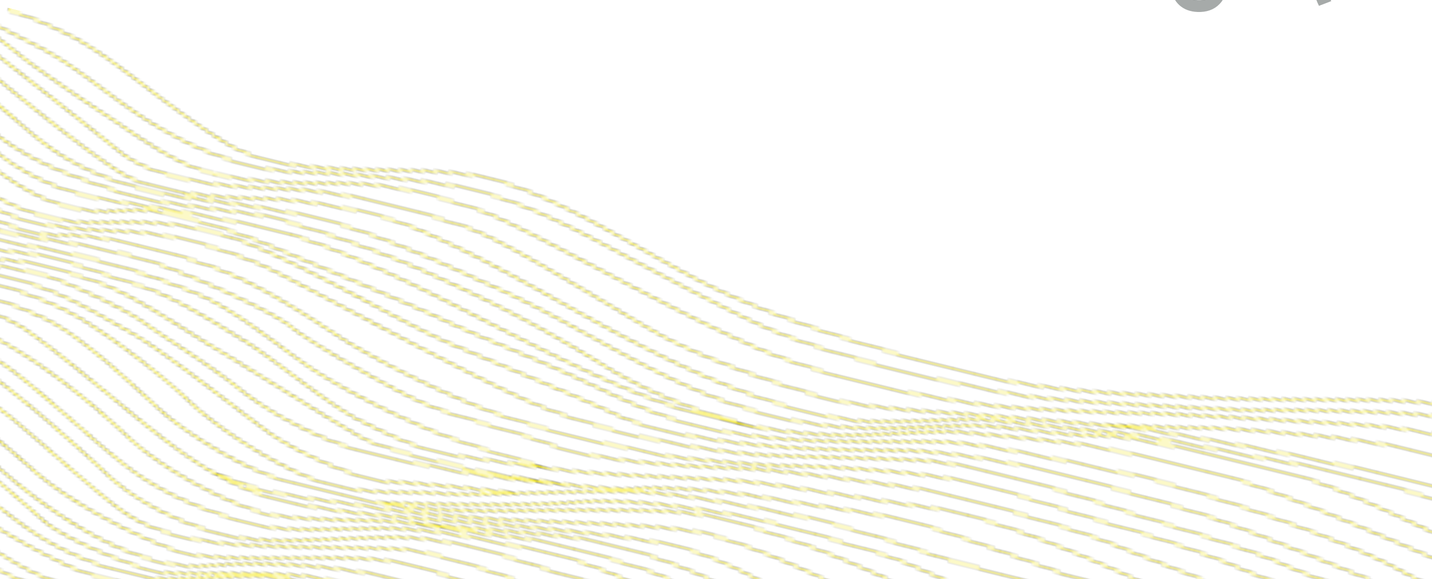
- Distributed (multi-core + multi-node) implementations of cutting edge ML algorithms.
- Core algorithms written in high performance Java.
- APIs available in R, Python, Scala; web GUI.
- Easily deploy models to production as pure Java code.
- Works on Hadoop, Spark, EC2, your laptop, etc.



H2O Machine Learning Features

- Supervised & unsupervised machine learning algos (GBM, RF, DNN, GLM, Stacked Ensembles, etc.)
- Automatic data pre-processing (imputation, encodings)
- Automatic early stopping (auto-tuning)
- Cross-validation, grid search & random search
- Variable importance, model performance metrics, plots
- Model explainability (PD plots, SHAP, residual analysis)

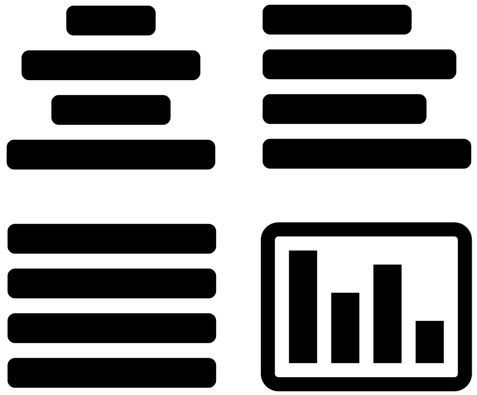
Automatic Machine Learning (AutoML)



Goals & Features of AutoML

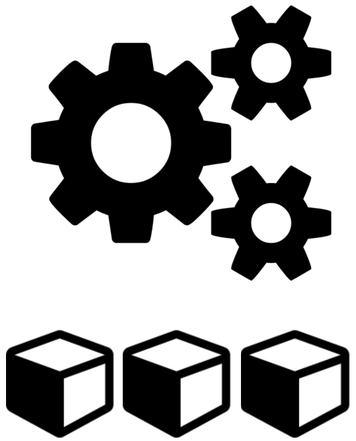
- 🏆 Train the best model in the least amount of time.
- 📉 Reduce the human effort & expertise required in machine learning.
- 📈 Improve the performance of machine learning models.
- 🔄 Increase reproducibility & establish a baseline for scientific research or applications.

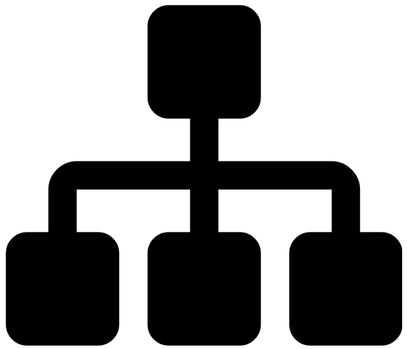
Aspects of Automatic Machine Learning



Data Prep

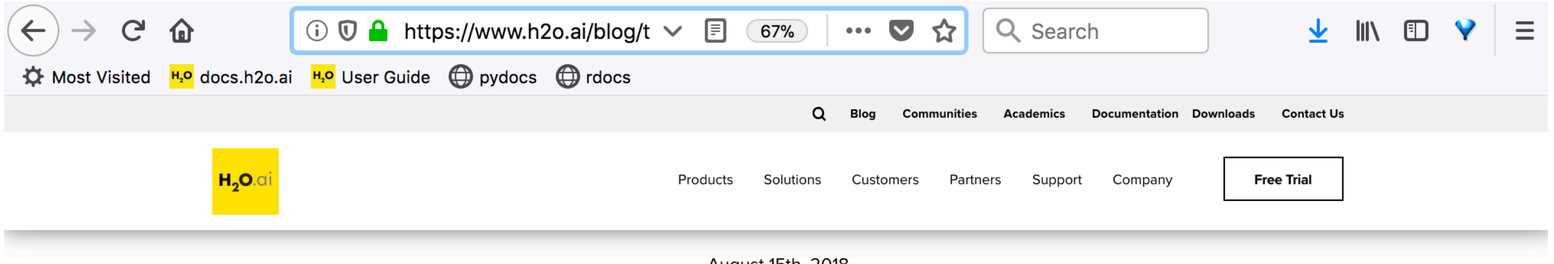
Model Generation





Ensembles

Different Flavors of AutoML



August 15th, 2018

The different flavors of AutoML

Share **Category:** AutoML, Data Science, Driverless AI, H2O

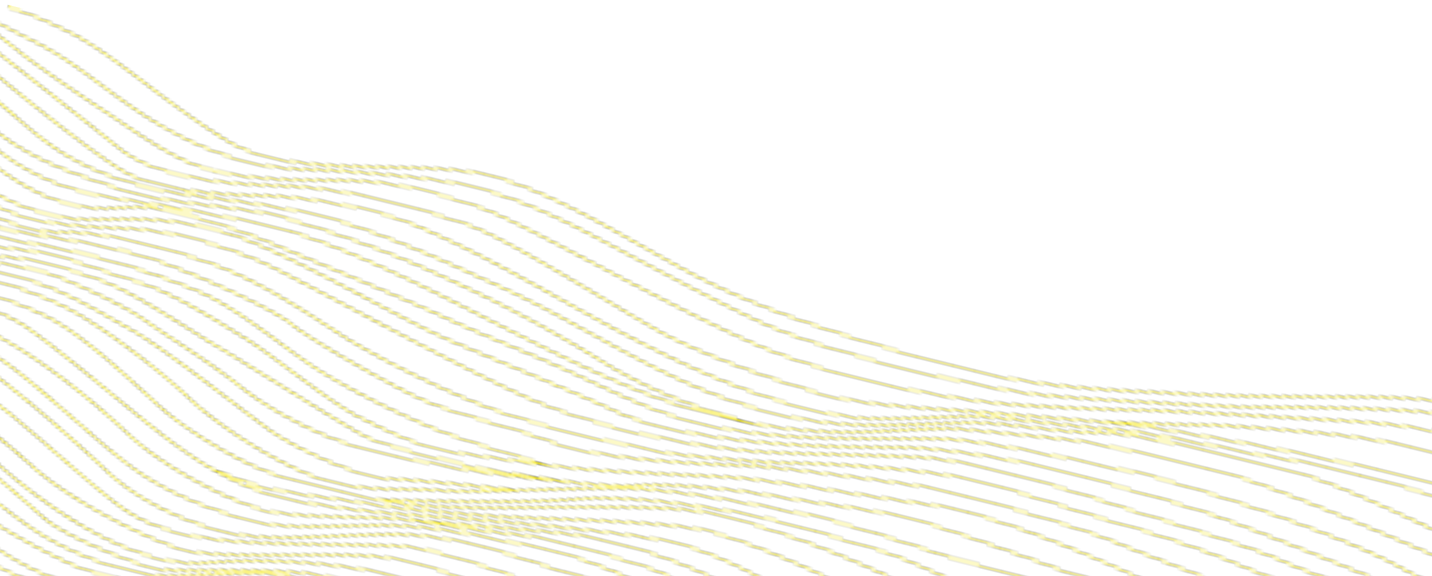
By: Erin LeDell

In recent years, the demand for machine learning experts has outpaced the supply, despite the surge of people entering the field. To address this gap, there have been big strides in the development of user-friendly machine learning software (e.g. H2O, scikit-learn, keras). Although these tools have made it easy to train and evaluate machine learning models, there is still a good amount of data science knowledge that's required in order to create the *highest-quality* model, given your dataset. Writing the code to perform a hyperparameter search over many different types of algorithms can also be time consuming and repetitive work.

What is AutoML?

https://tinyurl.com/flavors-of-automl

H2O AutoML



H2O AutoML (v3.32)

Data Preprocessing

- Imputation, one-hot encoding, standardization
- Feature selection and/or feature extraction (e.g. PCA)
- Count/Label/Target encoding of categorical features

Model Generation

- Cartesian grid search or random grid search
- Bayesian Hyperparameter Optimization
- Individual models can be tuned using a validation set

Ensembles

- Ensembles often out-perform individual models:
- Stacking / Super Learning (Wolpert, Breiman)
- Ensemble Selection (Caruana)

Random Grid Search & Stacking

- Random Grid Search combined with Stacked Ensembles is a powerful combination.
- Ensembles perform particularly well if the models they are based on (1) are individually strong, and (2) make uncorrelated errors.
- Stacking uses a second-level metalearning algorithm to find the optimal combination of base learners.

H2O AutoML

- Basic data pre-processing (as in all H2O algos).
- Trains a random grid of GBMs, DNNs, GLMs, etc. using a carefully chosen hyper-parameter space.
- Individual models are tuned using cross-validation.
- Two Stacked Ensembles are trained ("All Models" ensemble & a lightweight "Best of Family" ensemble).
- Returns a sorted "Leaderboard" of all models.
- All models can be easily exported to production.

H2O AutoML in Python

Example

```
import h2o

from h2o.automl import H2OAutoML

h2o.init()

train = h2o.import_file("train.csv")

aml = H2OAutoML(max_runtime_secs = 600)
aml.train(y = "response_colname",
          training_frame = train)

lb = aml.leaderboard
```

H2O AutoML in R

Example

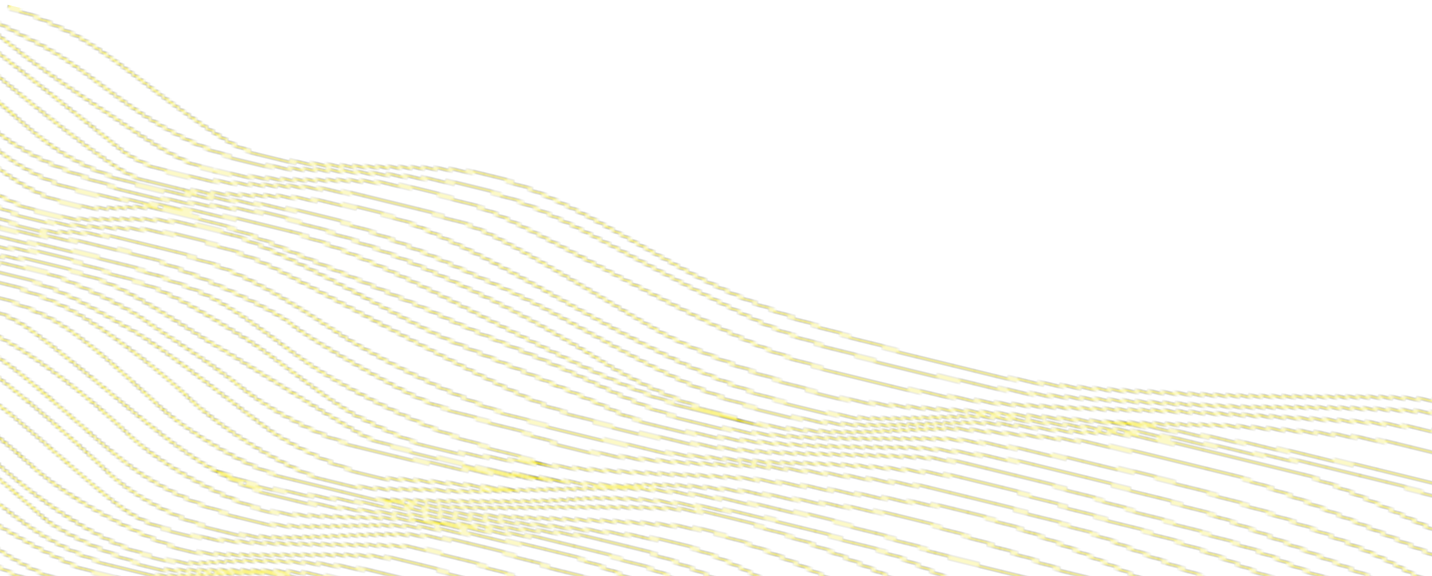
```
library(h2o)

h2o.init()

train <- h2o.importFile("train.csv")


aml <- h2o.automl(y = "response_colname",
                  training_frame = train,
                  max_runtime_secs = 600)

lb <- aml@leaderboard
```

H2O AutoML Leaderboard

	model_id	auc	logloss	aucpr	mean_per_class_error	rmse	mse
1	StackedEnsemble_AllModels_AutoML_20200709_004…	0.8378355	0.2866370	0.4481733	0.2498460	0.2913039	0.08485799
2	StackedEnsemble_BestOfFamily_AutoML_20200709_0…	0.8369381	0.2869531	0.4462222	0.2500683	0.2914670	0.08495302
3	XGBoost_3_AutoML_20200709_004415	0.8366588	0.2809896	0.4502926	0.2552901	0.2894478	0.08378002
4	GBM_4_AutoML_20200709_004415	0.8330289	0.2848382	0.4239271	0.2593298	0.2919957	0.08526147
5	GBM_3_AutoML_20200709_004415	0.8325824	0.2852444	0.4195761	0.2552272	0.2922670	0.08542002
6	GBM_2_AutoML_20200709_004415	0.8323248	0.2855498	0.4185351	0.2589230	0.2924915	0.08555129
7	GBM_1_AutoML_20200709_004415	0.8322315	0.2855884	0.4200573	0.2622791	0.2922375	0.08540278
8	XGBoost_1_AutoML_20200709_004415	0.8317490	0.2858897	0.4326282	0.2618297	0.2923182	0.08544993
9	GBM_5_AutoML_20200709_004415	0.8296069	0.2874258	0.4040567	0.2569593	0.2938664	0.08635746
10	XGBoost_2_AutoML_20200709_004415	0.8277037	0.2899311	0.4265391	0.2624847	0.2943874	0.08666391
11	DRF_1_AutoML_20200709_004415	0.8120043	0.3008964	0.3722857	0.2731671	0.2991530	0.08949252
12	GLM_1_AutoML_20200709_004415	0.6873574	0.3510707	0.2172795	0.3673990	0.3194751	0.10206432

Example Leaderboard for binary classification

AutoML Pro Tips!



AutoML Pro Tips: Customize

- Control time limit using `max_runtime_secs` or limit the number of models using `max_models`.
- You can turn off cross-validation for big datasets by setting `nfolds=0`. CV is required for Stacked Ensembles so that will be disabled.
- Turn on/off certain algorithms using `exclude_algos` or `include_algos`.

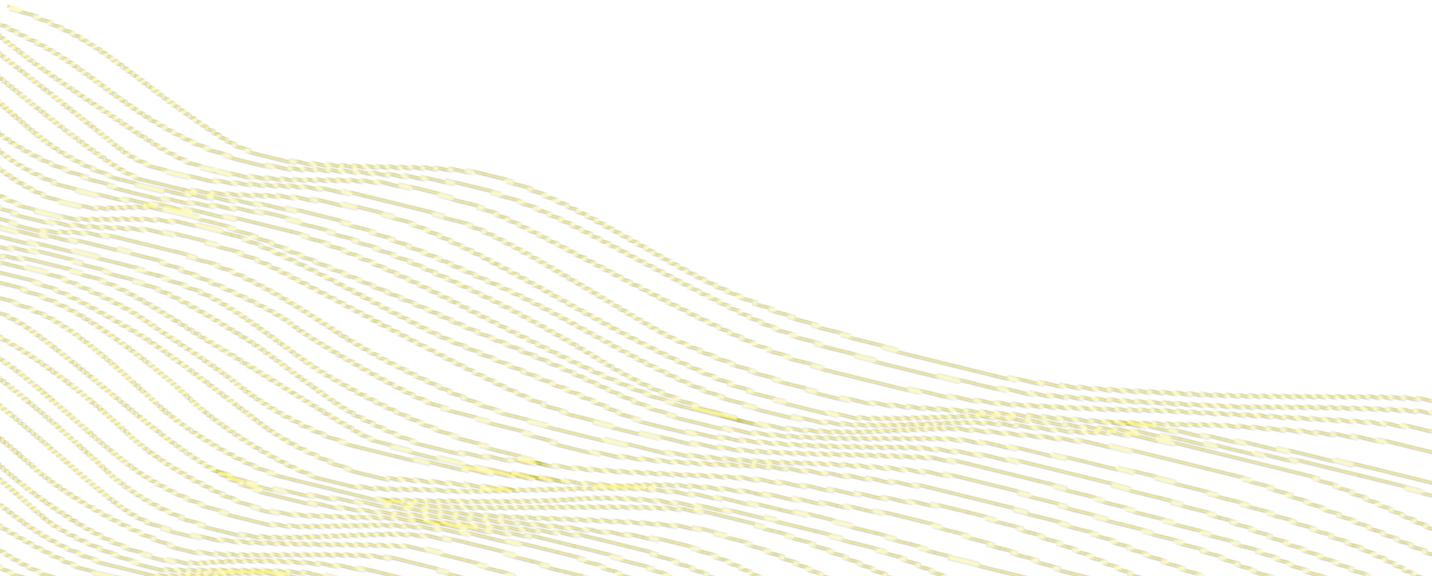
AutoML Pro Tips: Cluster memory

- Reminder: All H2O models are stored in H2O Cluster memory.
- Make sure to give the H2O Cluster a lot of memory if you're going to create hundreds or thousands of models.
- e.g. `h2o.init(max_mem_size = "80G")`

AutoML Pro Tips: Add More Models

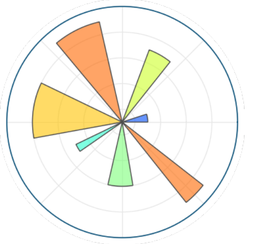
- If you want to add (train) more models to an existing AutoML project, just make sure to use the same training set and `project_name`.
- If you set the same seed twice it will give you identical models as the first run (not useful), so change the seed or leave it unset.

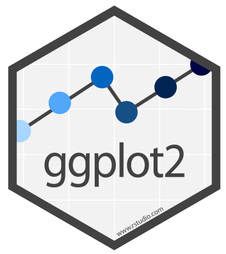
Explainable ML



H2O AutoML Explainability

- The new `h2o.explain()` interface automatically generates many explanations (annotated visualizations) for a single model or a group of models (e.g. AutoML leaderboard).
- Row-wise explanations are available via the `h2o.explain_row()` companion function.
- Visualizations are created with `ggplot2` in R and `matplotlib` in Python, and can be customized.





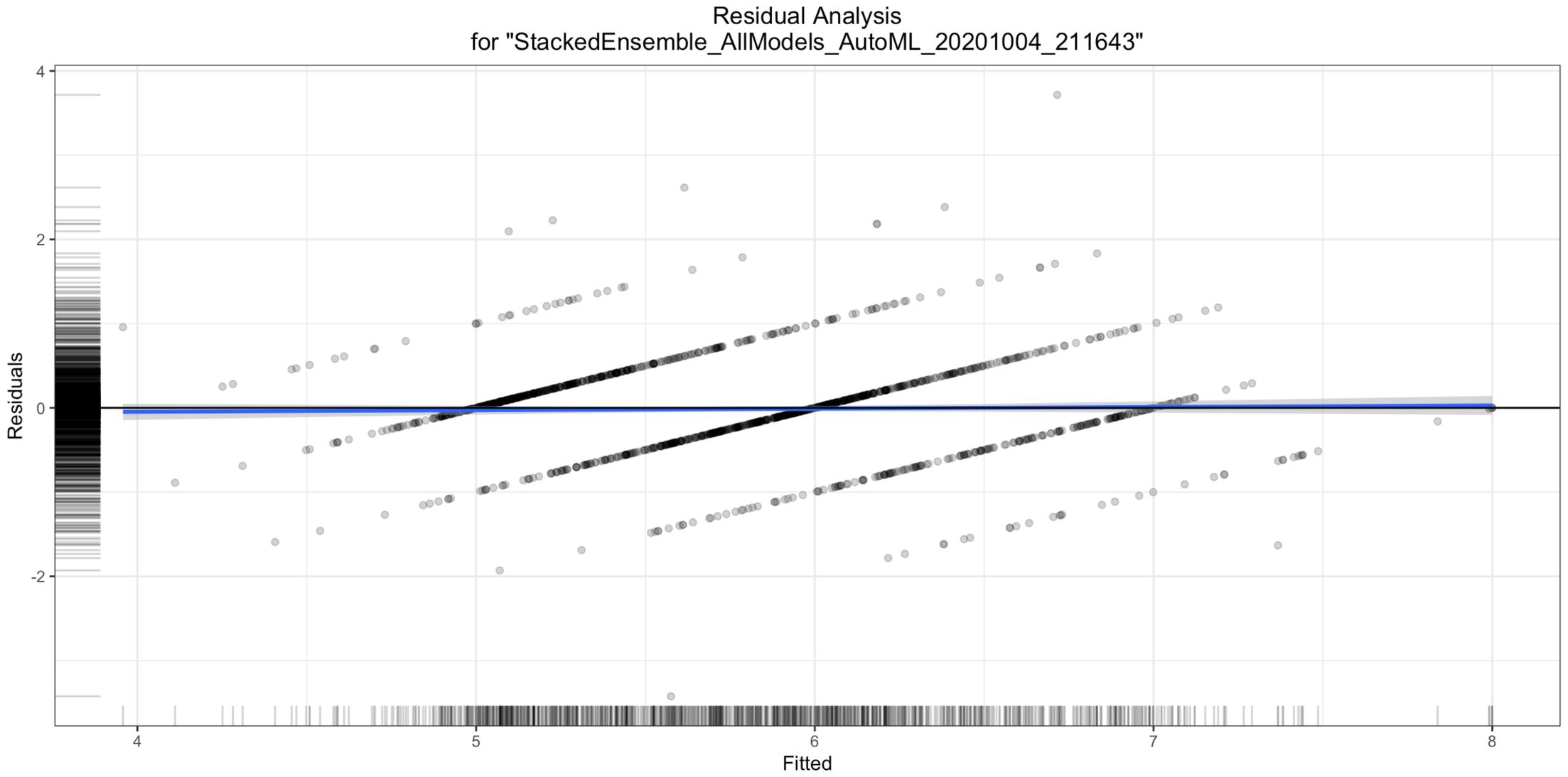
H2O AutoML Explainability

New in H2O v3.32: 🔬 🛠️

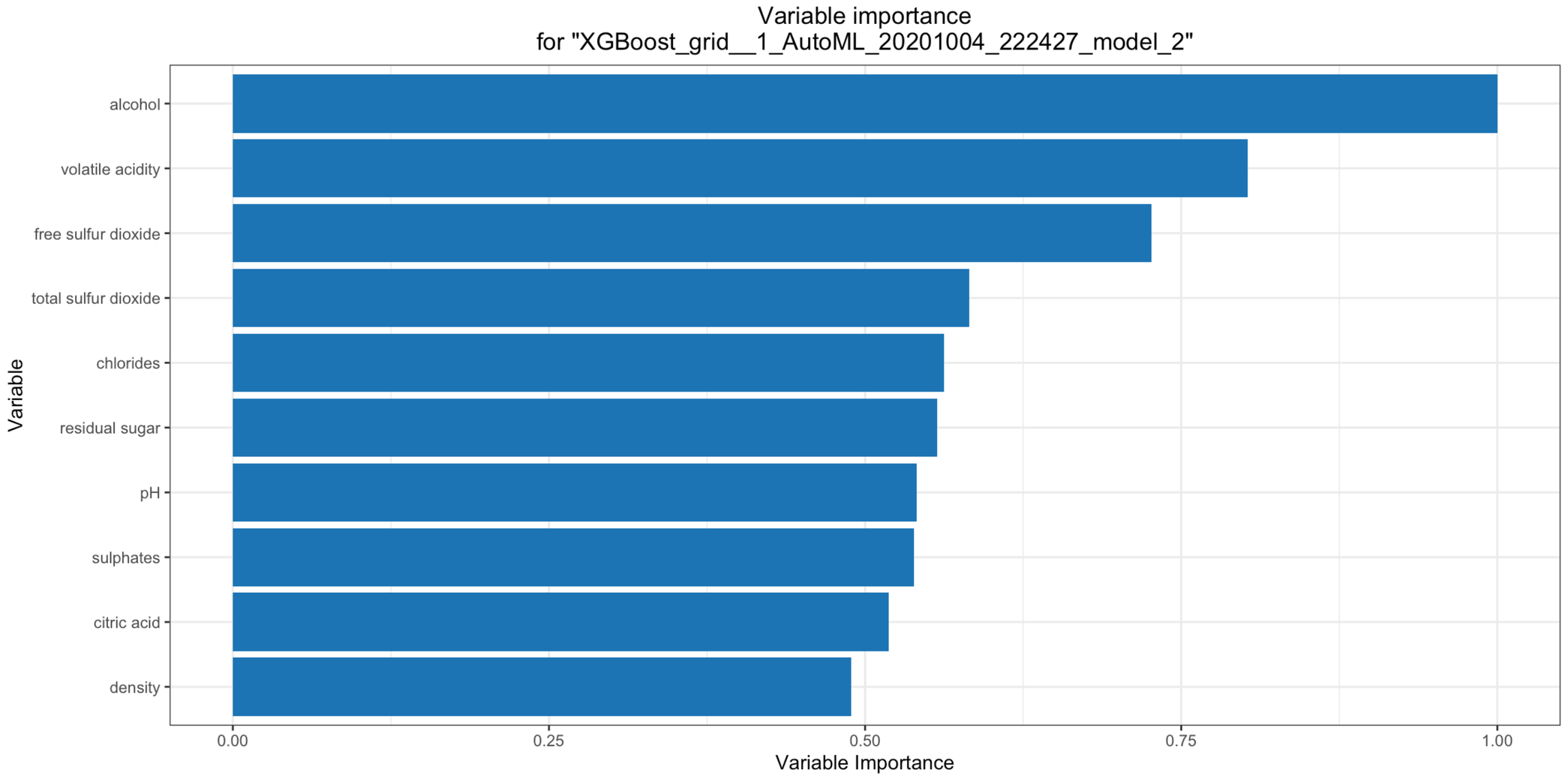
- Variable importance comparisons
- Model correlation heatmap
- SHAP contributions for tree-based models
- Partial dependence (PD) plots
- Individual Conditional Expectation (ICE) plots
- Residual Analysis

http://docs.h2o.ai/h2o/latest-stable/h2o-docs/explain.html

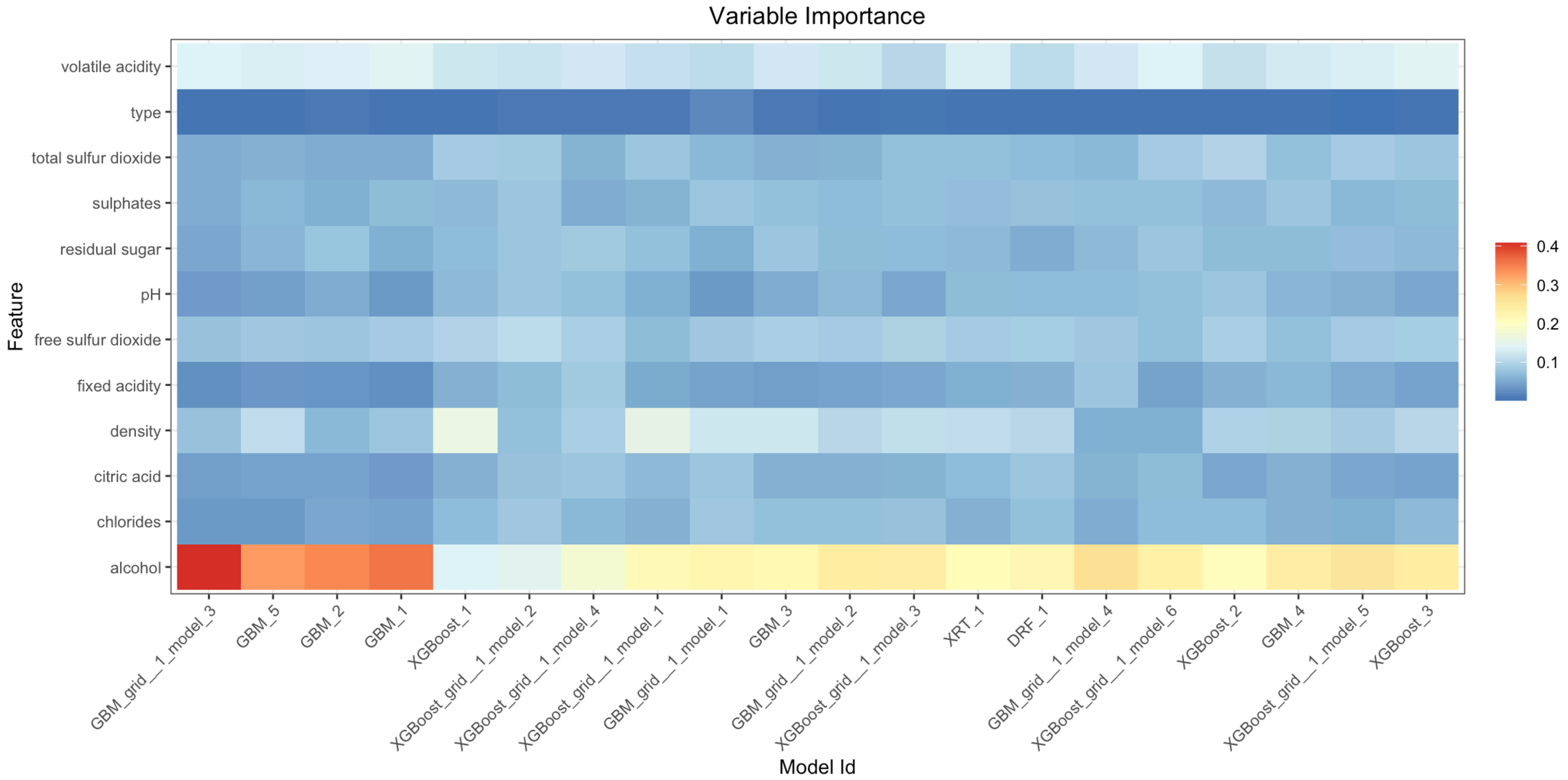
Residual Analysis



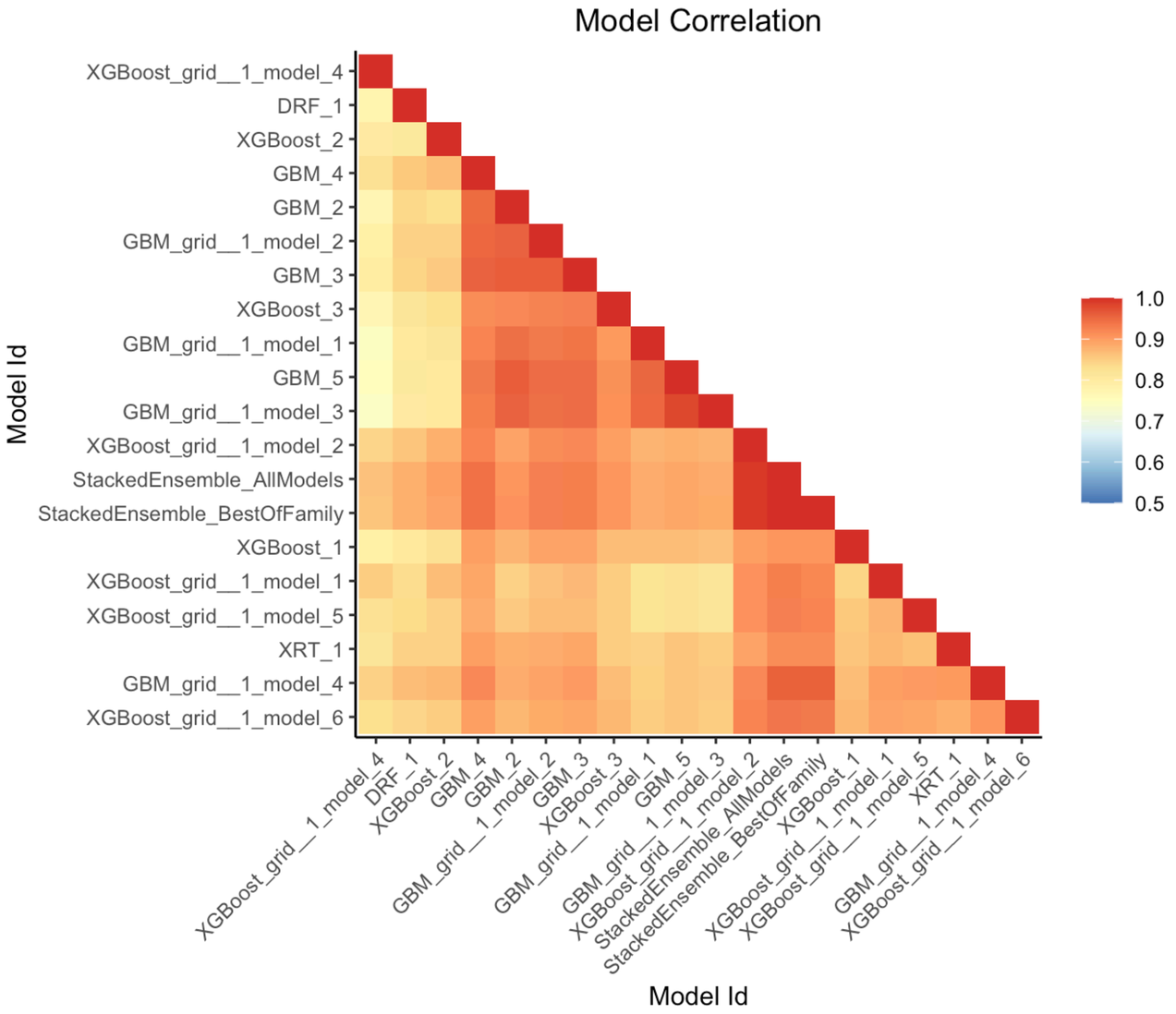
Variable Importance



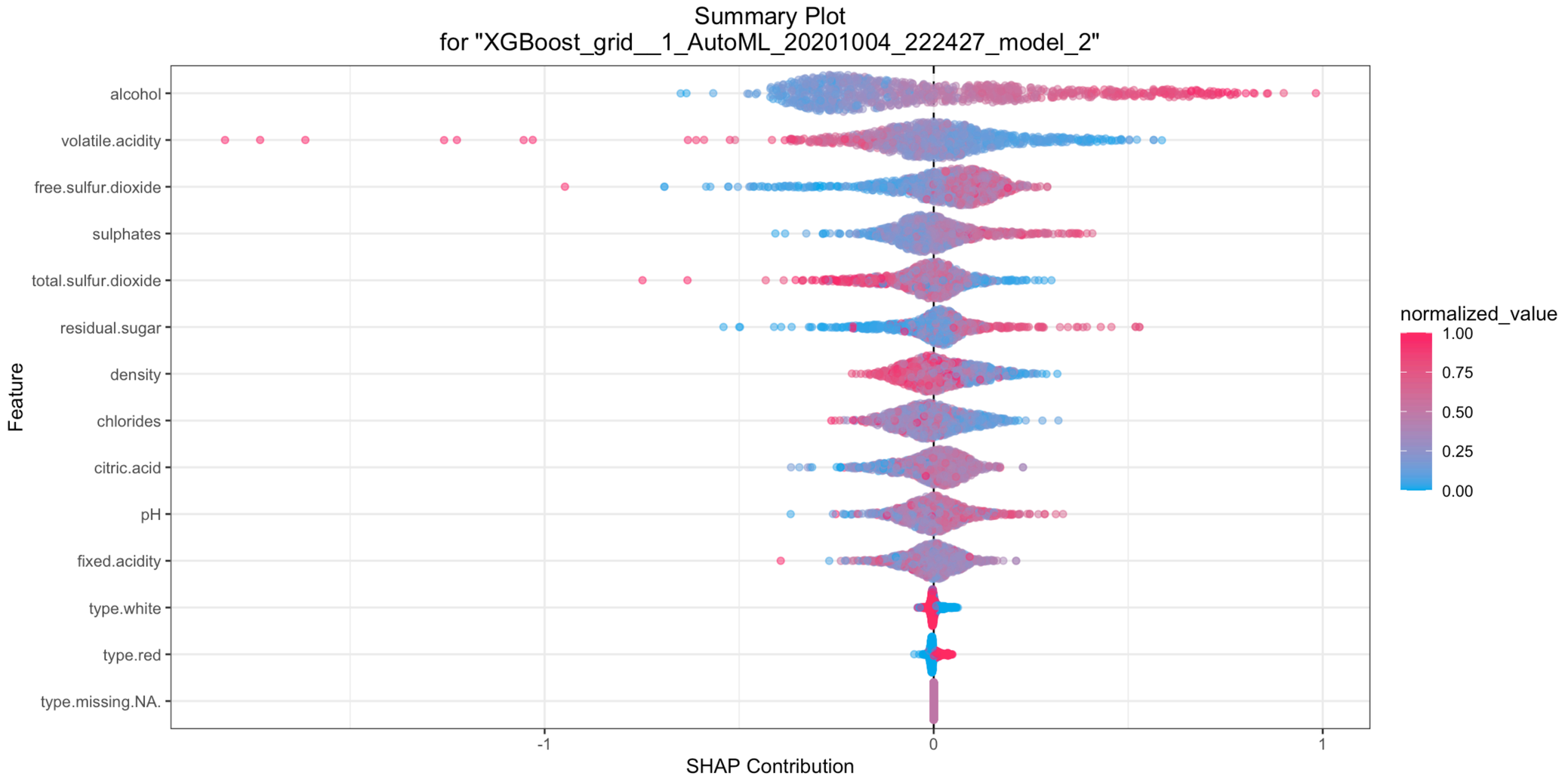
Variable Importance Heatmap



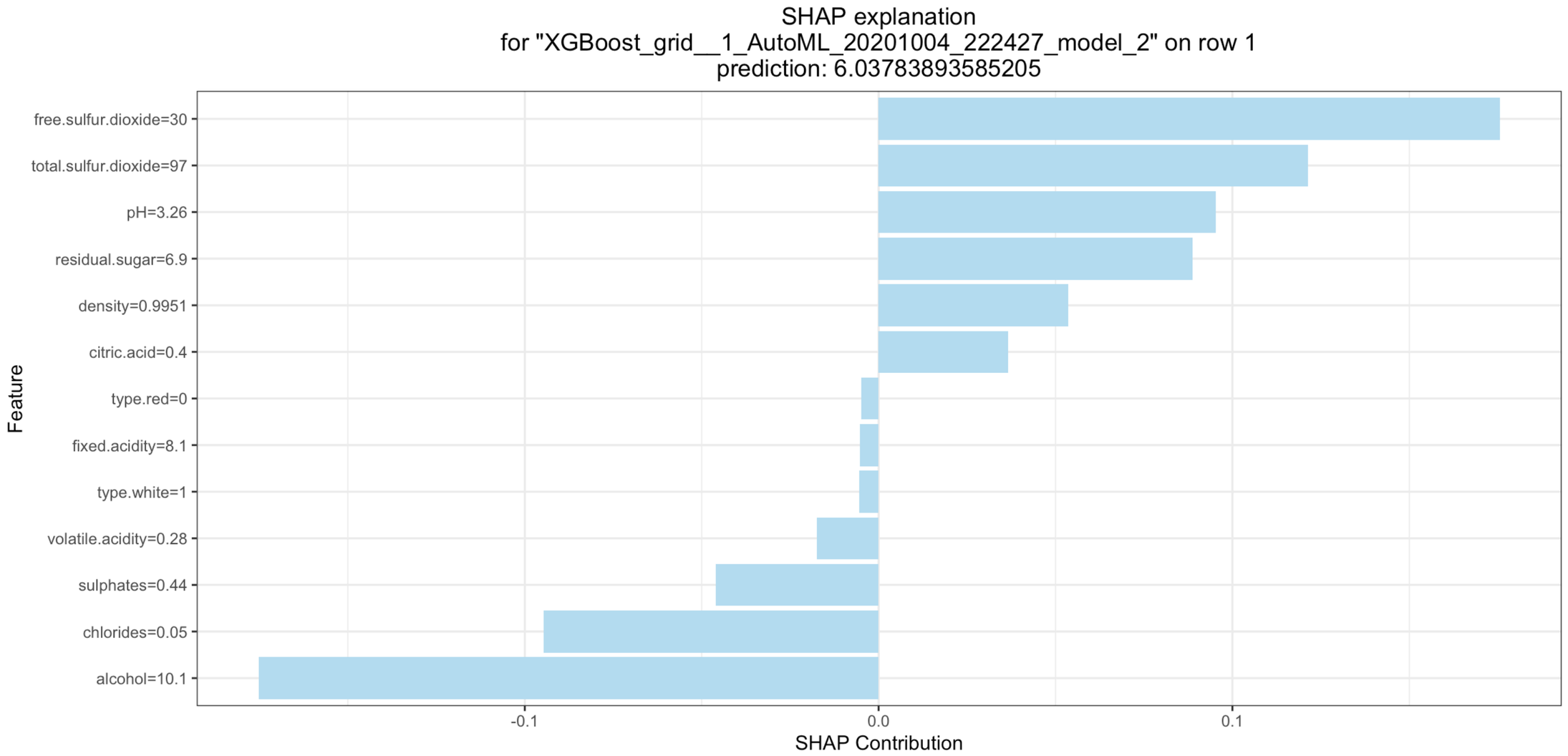
Model Correlation Heatmap



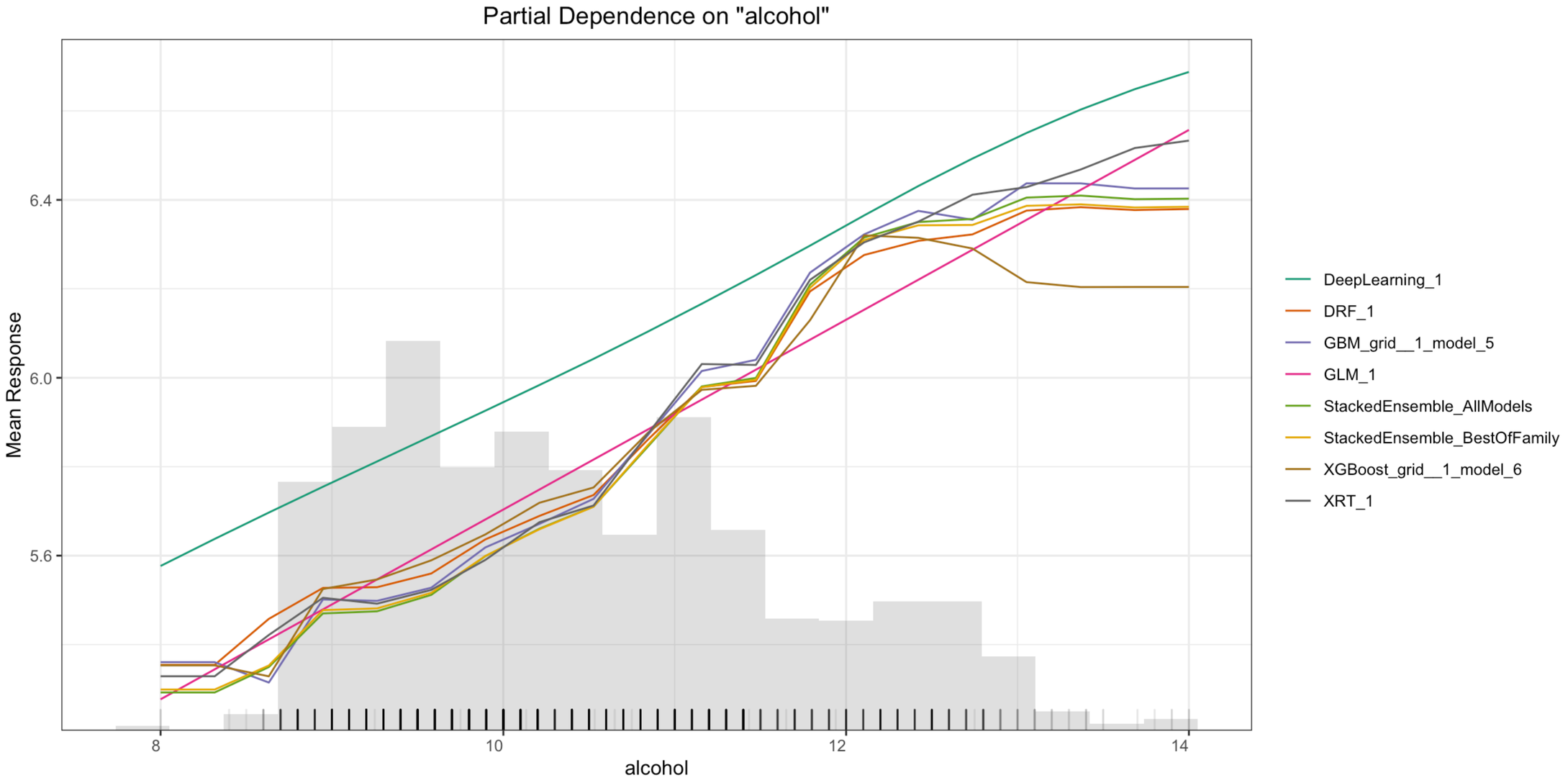
SHAP Summary



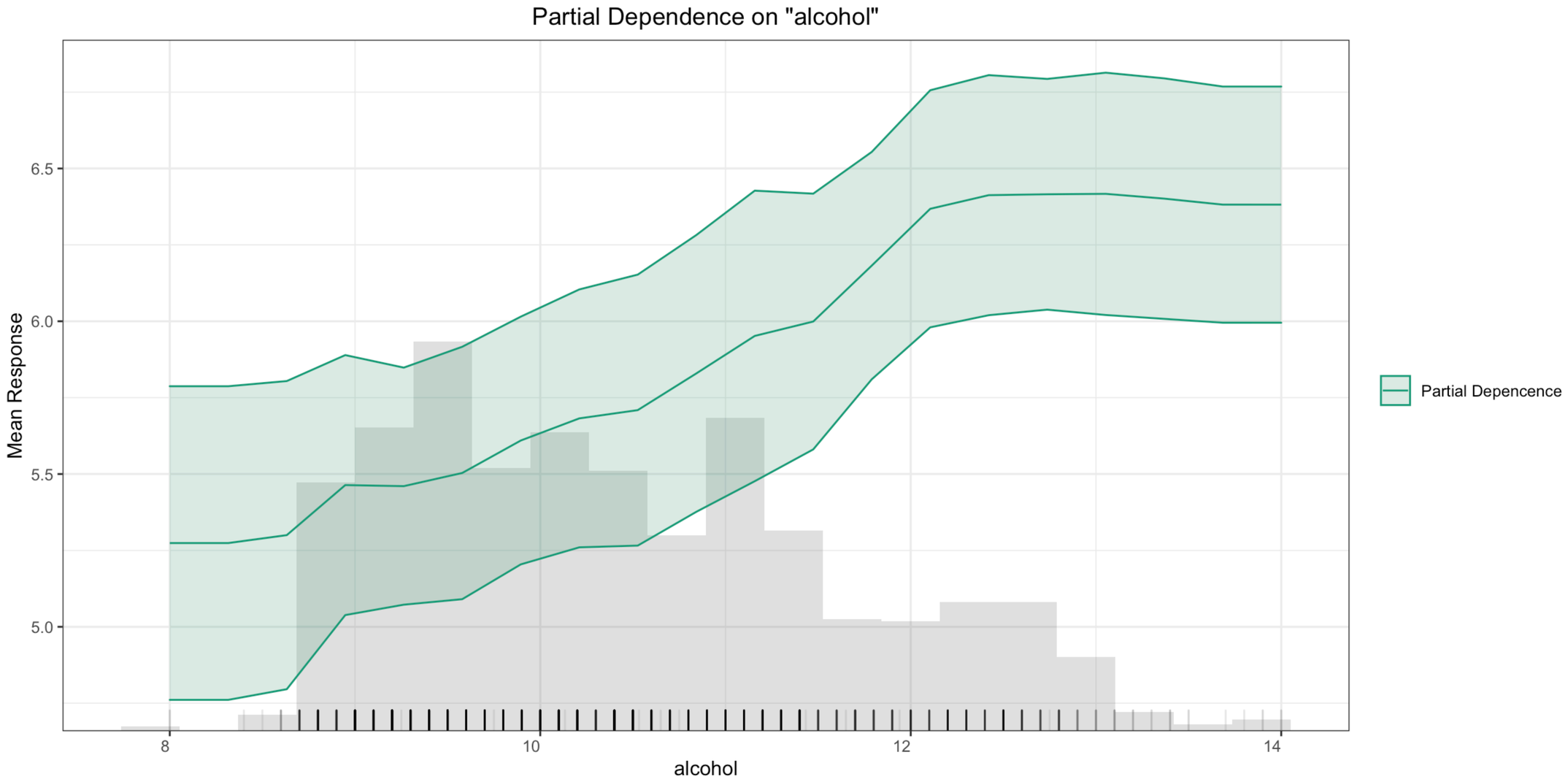
SHAP Local Explanation



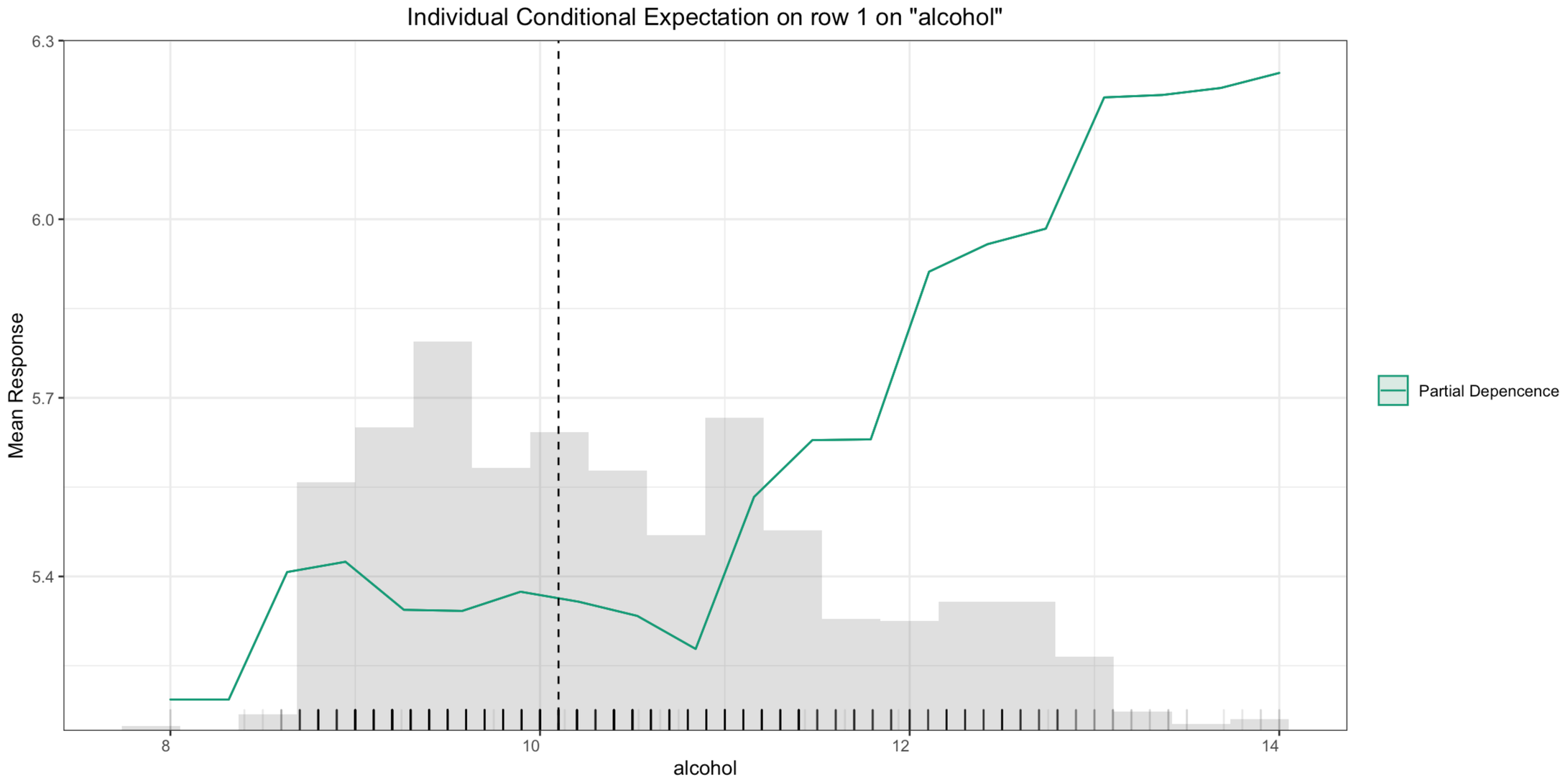
Partial Dependence (PD) Plots



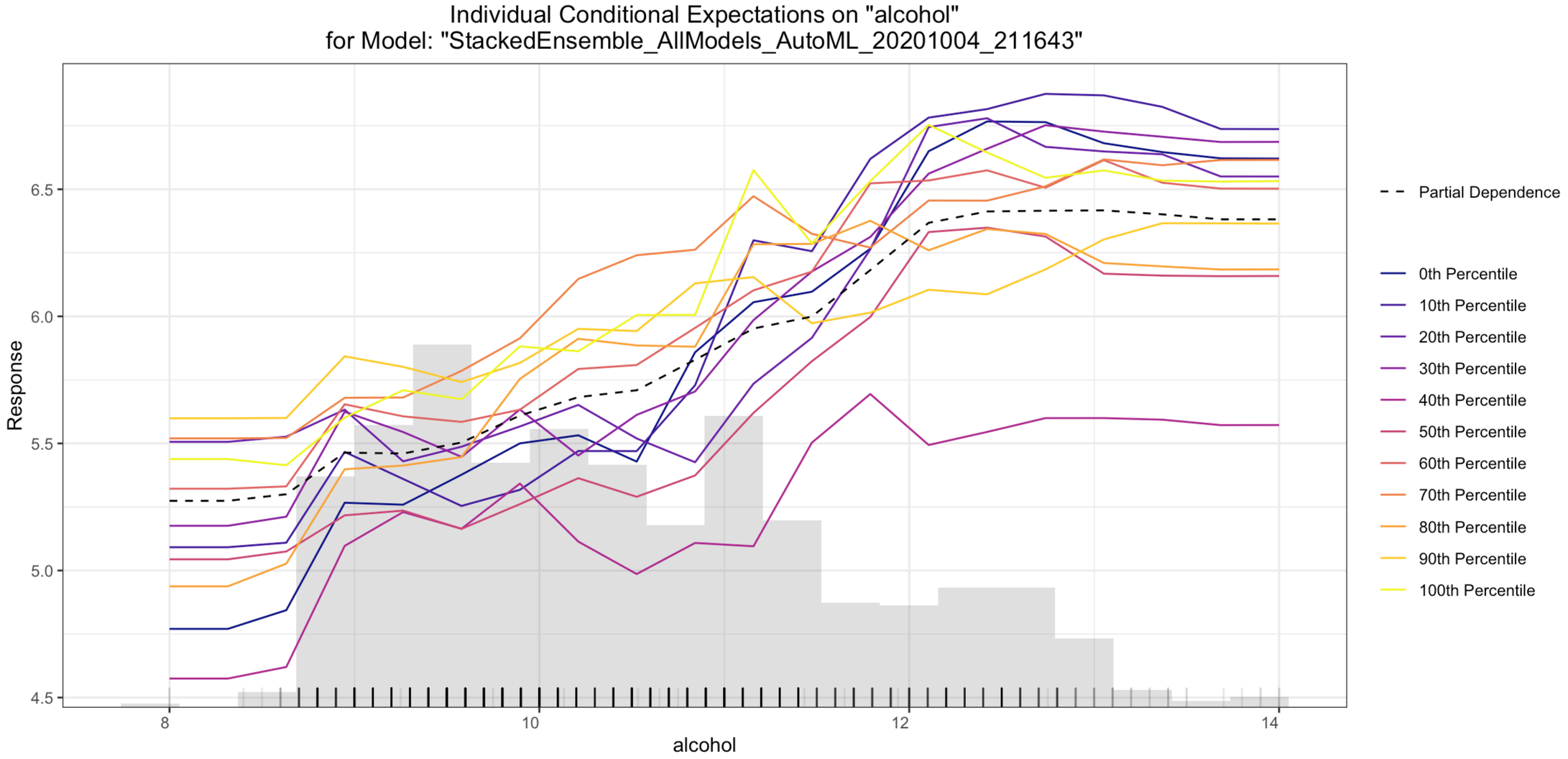
Partial Dependence (PD) Plots



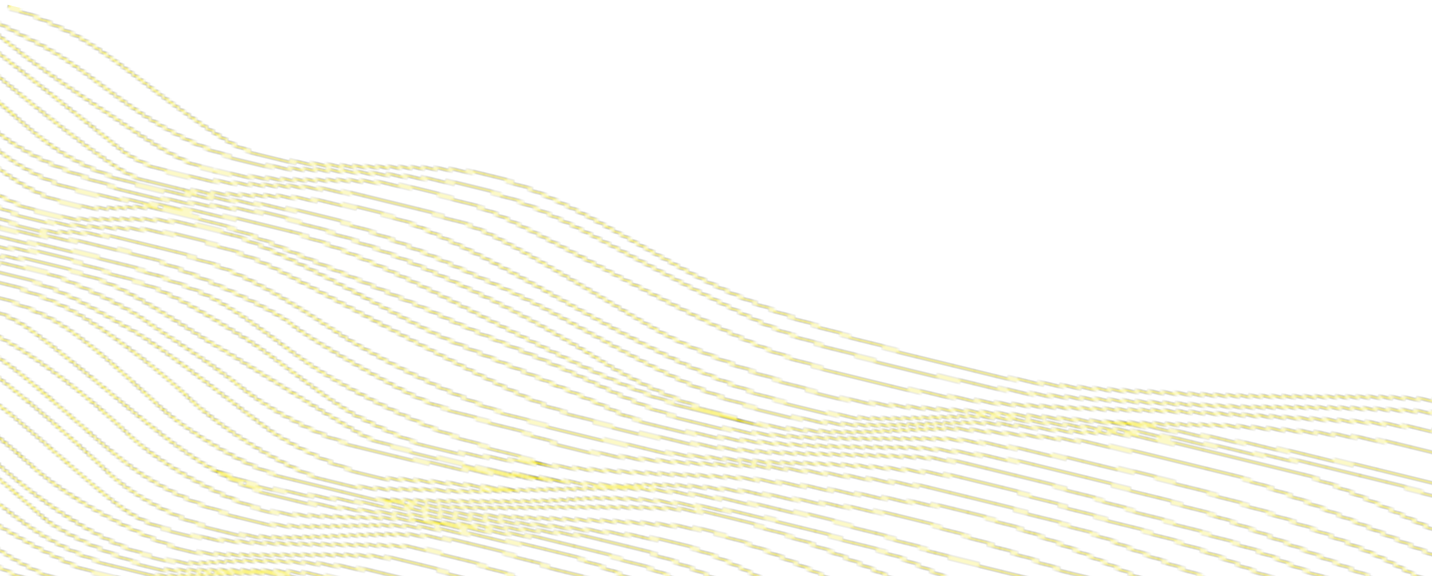
Individual Conditional Expectation (ICE)



Individual Conditional Expectation (ICE)

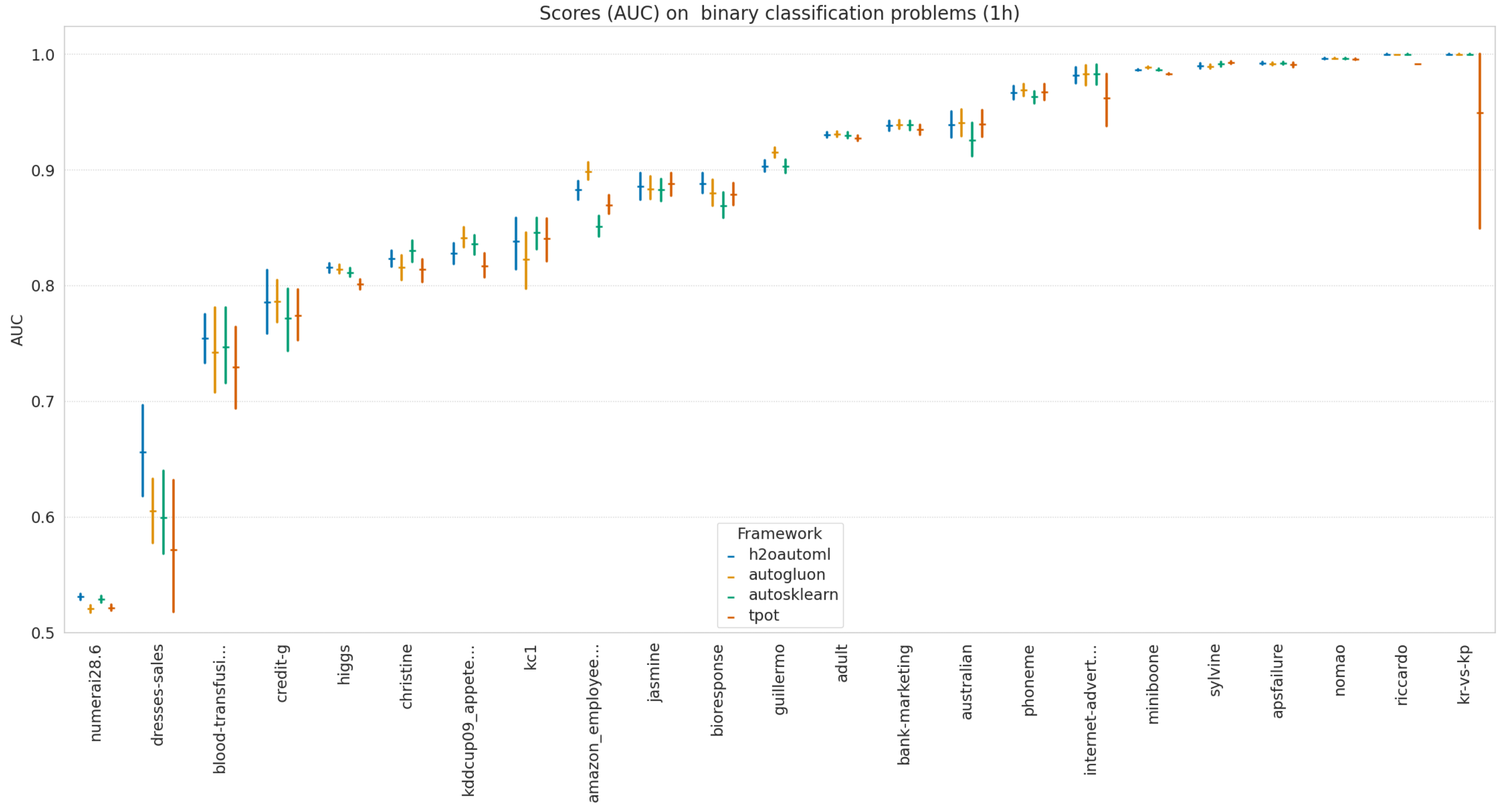


H2O Resources



H2O AutoML paper

The H2O AutoML paper was accepted at ICML 2020 AutoML Workshop



- *Official H2O AutoML paper*
- *Updated benchmarks*
- *Stacking study*
- *Scalability study (10K - 100M rows)*

ICML paper ⬇️ https://tinyurl.com/icml2020-h2o-paper

Learn H2O AutoML!

- Docs: https://tinyurl.com/h2o-automl-docs
- R & Py tutorials: https://tinyurl.com/h2o-automl-tutorials
- useR! 2020: https://github.com/ledell/useR2020-automl

H2O Resources

- Documentation: http://docs.h2o.ai
- Learning Center: https://training.h2o.ai
- Tutorials: https://github.com/h2oai/h2o-tutorials
- Slidedecks: https://github.com/h2oai/h2o-meetups
- Videos: https://www.youtube.com/user/0xdata
- Stack Overflow: https://stackoverflow.com/tags/h2o

Thank you!

@ledell on
Github, Twitter