# Machine Learning Interpretability

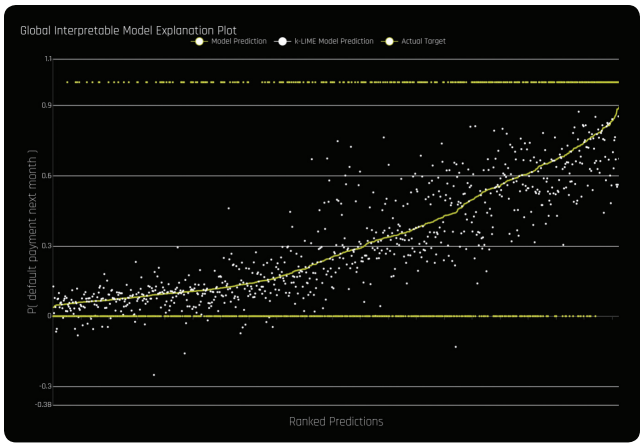
Machine learning interpretability by Driverless AI offers simple data visualization techniques for representing high-degree feature interactions and nonlinear model behavior. It also uses a contemporary linear model variant to generate reason codes, in comprehensible English, that may be appropriate for use in regulated industry.

## Why Does It Matter?

Driverless AI, by H2O.ai, provides a market leading Machine Learning Interpretability (MLI) component to address questions associated with machine learning fairness, accountability, and transparency through visualizations and measurements that clarify modeling results and the effect of features in a model.

The machine learning interpretability component of Driverless AI enables the data practitioner to get clear and concise explanations of the model results by generating four dynamic charts in a dashboard; these interactive charts can be used to visualize and debug a model by comparing the displayed global and local model decision-process, important variables, and important interactions to known standards, domain knowledge, and reasonable expectations.

### k-LIME:




k-LIME automatically builds linear model approximations to regions of complex Driverless AI models' learned response functions. These penalized GLM surrogates are trained to model the predictions of the Driverless AI model, and can be used to generate reason codes and English language explanations of complex Driverless AI models.

### BRINGING AI TO ENTERPRISE

**BUSINESS NEEDS:**

Interpretability aids business adoption of machine learning by:

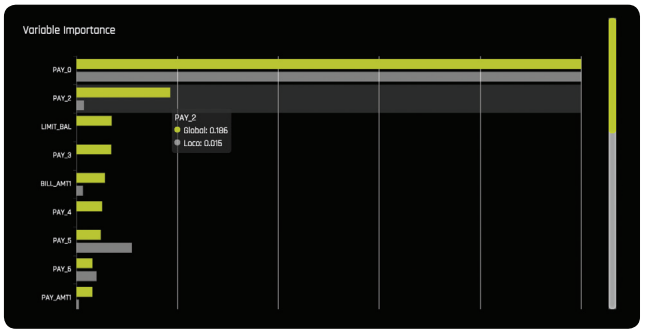
- Enhancing understanding of complex model mechanisms
- Increasing trust in model predictions and decisions

**BENEFITS:**

- Provides numerous charts and measurements for model debugging
- Near complete transparency & accountability for model mechanisms, predictions, and decisions
- Smart data visualization techniques portray high dimensional feature interactions in just two dimensions
- Reason codes in plain English, for easy understanding and regulatory compliance

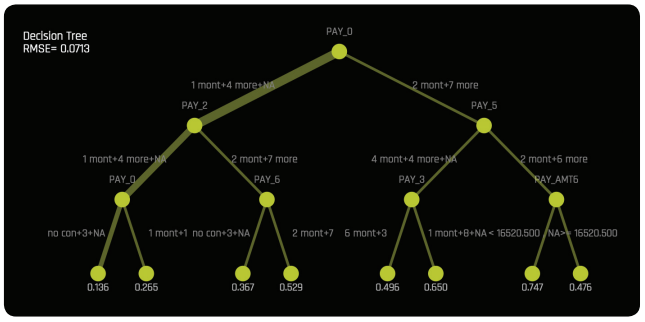
### Variable Importance:

Variable importance quantitatively measures the effect that an input variable has on the predictions of a model. Variable importance is most useful for machine learning models where traditional measures fall short of describing the relationship between an input variable and the target variable. Driverless AI presents both global and local, row-level variable importance values to tell users the most influential variables, and the variables' relative rank, in a model and in each model decision. Techniques such as TreeInterpreter and Shapley explanations are used to calculate local variable importance values.




Global (yellow) and local, i.e. row-wise (grey), variable importance for a Driverless AI model.

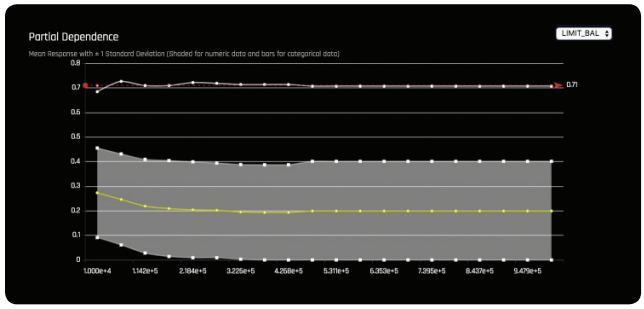
Decision Tree Surrogate:




A global decision tree surrogate model that summarizes the decision-process of a complex Driverless AI model.

The decision tree surrogate model provides insight into the Driverless AI model by displaying an approximate, overall flow-chart of the model's decision-making process. It also displays the most important variables and interactions in the Driverless AI model. Combined together with the partial dependence and ICE plots, these visualizations enable the understanding of complex, high-degree variable interactions with just two-dimensional plots.

Partial Dependence and ICE Plots:




A partial dependence plot that summarizes the decision-process of a complex Driverless AI model.

The partial dependence plot shows the average effect of changing one variable on the model prediction. Partial dependence plots are global in terms of the rows of a data set, but local in terms of the input variables. ICE plots can provide an even more local view of the model's decision-making process - an input variable in a single row is selected, and that input variable in the selected row can be toggled through the set of values for the input variable in the training data set, and then run through the model again for each value to display a range of possible predictions for rows similar to the selected row. Combined together with the decision tree surrogate plot, these visualizations enable the understanding of complex, high-degree variable interactions with just two-dimensional plots.

Interpretation of External Models:
The Driverless AI Interpretation model allows for the uploading of external model (Python, R, SAS, etc.) inputs and predictions to be interpreted using all described plots and explanation techniques.

H2O.ai is focused on bringing AI to businesses through software. Its flagship product is H2O, the leading open source platform that makes it easy for financial services, insurance and healthcare companies to deploy machine learning and predictive analytics to solve complex problems. More than 13,000 organizations and 130,000+ data scientists depend on H2O for critical applications like predictive maintenance and operational intelligence. The company accelerates business transformation for 222 Fortune 500 enterprises, 8 of the world's 12 largest banks, 7 of the 10 largest auto insurance companies and all 5 major telecommunications providers.

Follow us on Twitter @h2oai. To learn more about H2O customer use cases, please visit http://www.h2o.ai/customers/. Join the Movement.

Tel: +1.650.227.4572 sales@h2o.ai http://www.github.com/h2oai @h2oai