RuleFit and Explainable Boosting Machine

What is RuleFit?

RuleFit is a machine learning method that can generate clear and concise rules from complex data. The general implementation process involves extracting rules (if-then conditionals) learned from several trees and weighting them using a sparse linear model (with L1 regularization). The resulting model is a linear combination of rules, which has good nonlinear expressiveness and can explain the contribution of each rule individually.

How to achieve RuleFit?

1. Rule Generation

Gradient Boosting or Random Forest was recommended to use, which generate a variety of candidate if-then rules. A tree depth of 3-4 is recommended, as shorter rules have better generalization capabilities. After training, the RuleFit package can generate rules.

2. Rule pre-filtering

Before training the rules with the L1 model, you can do some cleaning to control the scale: remove duplicate, always true, or always false rules; set rule support to [0.5%, 99.5%]; limit rule length ($\leqslant$ 4 conditions); and limit the total number of candidates (max_rules = 2000-5000).

3. Training RuleFit (L1-Logistic)

The regular 0/1 features and the original features are linearly fed into the L1 linear model, which automatically performs sparse selection and weighting.

4. Get interpretability rules

get_rules() retrieves a list of rules with non-zero coefficients and reasonable support. It can select the top N high-value rules and directly trigger warnings, such as dispatching personnel to inspect the track if certain rules are hit.

Advantages and limitations

Advantages

Outputs if-then rules and coefficients, offering strong interpretability. For a single sample, each matching rule and linear term can be listed separately, indicating their contribution to the final score, as RuleFit's scoring is additive. Rules are stronger than pure linear models and more stable than deep models. L1 weights will reduce the weights of most rules to 0, resulting in only a small number of useful rules.

Limitations

Prediction accuracy is generally lower than deep models. It relies on rule quality, as candidate rules come from a set of trees. Too few trees or an inappropriate depth setting can lead to poor performance. The rule set size can easily explode.

What is Explainable Boosting Machine?

EBM is an interpretable additive model that represents the prediction score as the sum of the individual contribution of each feature and the interaction contributions of a small number of feature pairs.

How to achieve EBMS?

1. Feature binning and function initialization

Continuous features are binned using histograms, categorical features are binned using categories, and missing values are binned separately. Initialize each shape function.

2. Circulation Improvement

A cyclic alternation strategy is used, where only one feature is selected per round. A shallow tree (max_leaves = 2-3) is trained using the bins of that feature to fit the current residual. The learned increment is added to the shape function, and the next feature is selected, repeating the process for multiple rounds.

3. **Learning a small number of interacting feature functions**

A small number of important feature pairs are picked according to the decreasing loss, and the same shallow tree is trained on the two-dimensional bin grid to obtain the interactive feature function.

4. **Get interpretable output**

Globally: explain_global() returns a curve for each feature function and a two-dimensional heatmap for each interaction feature.
Single sample: explain_local() gives the corresponding values of each feature function and interaction feature function, sums them up, and then calculates the probability value through sigmoid.

Advantages and limitations

Advantages

EBMs are highly interpretable, with a single curve for each feature and a heatmap for small interactions. EBMs achieve high prediction accuracy, often comparable to complex models such as random forests and deep learning networks. EBMs are generally smoother and more resistant to noise. EBMs also facilitate what-if analysis; modifying a single feature value reveals changes in probability on the curve.

Limitations

EBMs primarily capture one-dimensional effects and a small number of binary interactions, and are not robust to high-order or non-axis-aligned complex boundaries. Binning selection can easily affect results: too small bins can lead to underfitting, while too large bins can cause jitter, requiring validation sets and early stopping. Larger features require increased memory usage and training time, increasing the burden of visualization and interpretation. The number of interactions should be controlled; excessive interactions impair interpretability and can lead to overfitting.

Reference

Yalcin, 2021, *Interpretable Machine Learning in 10 Minutes with RuleFit and Scikit Learn*, Medium, viewed 9 October 2025, <https://medium.com/data-science/interpretable-machine-learning-in-10-minutes-with-rulefit-and-scikit-learn-da9ebb925795>.

RuleFit, h2o.ai, viewed 9 October 2025, <https://h2o.ai/wiki/rulefit/>.

Indraneel Dutta Baruah, 2023, *How Do Inherently Interpretable AI Models Work? Explainable Boosting Machine*, Medium, viewed 10 October 2025, <https://medium.com/nerd-for-tech/how-do-inherently-interpretable-ai-models-work-explainable-boosting-machine-f9e3718b1a4>.

2024, *Explainable Boosting Machines (EBMs)*, GeeksforGeeks, viewed 10 October 2025, <https://www.geeksforgeeks.org/machine-learning/explainable-boosting-machines-ebms/>.

Microsoft Developer, YouTube, *The Science Behind InterpretML: Explainable Boosting Machine* ,2020, viewed 10 October 2025, <https://www.youtube.com/watch?v=MREiHgHgl0k>.