



Residual Gradient Booster

A gradient booster that starts from your model's own log probabilities and learns only what it missed.

Multiclass gradient boosting with the fitted model's log probabilities as the base margin, Newton leaf values, early stopping, and the option to pool leagues so a single division cannot over-fit itself.

SPORT

Football

FAMILY

The machine layer

LICENCE

£549

one-off, perpetual

RUNNING IN THE PLATFORM

<https://edgewortharena.com/arena/api/ml/league?comp=E0>

Inputs

NAME	TYPE	UNITS	WHERE IT COMES FROM
base_probabilities	dict per match	probability	any match model
features	array per match	mixed	rest, congestion, recent shots, form, the season so far
results	list	-	
depth, rounds, learning rate (optional)	ints and float	-	your choice; depth 2 is what survives out of sample

Outputs

NAME	TYPE	UNITS	NOTES
boosted probabilities	dict per match	probability	
importances	dict	share of split gain	
held_out_log_loss	float	nats	beside the base model's and the market's

Method

- 1 Set the base margin to log p from the fitted model, so round zero already prices the match.

THIS IS THE SAMPLE

The method runs to 4 steps; one is shown. The assumptions, the limits, the module layout and the reference implementation come with the licence. Everything on the page below this -- what the model was measured at -- is the same in both versions, because that is the part you should not have to take on trust.



What it was measured at

Trained on earlier seasons, early stopped on a validation slice, scored on the last complete season.

MEASURE	RESULT	READING
Leagues with a per-league booster	7	
Per-league gain over the base model	-2.02% to +0.25%	held-out season; as often negative as positive
What a suspicious number looks like	several percent	over a good base model that is a leak, not a model

Assumptions

What has to be true for the numbers to mean what they say.

- The base model is already close to right, so the trees are learning a small correction rather than the signal itself.

Limits

Where it is known to be wrong. Published because a model without stated limits is one nobody has pushed on.

- Per league on a few thousand matches the honest gain runs from about -4% to +1% of log loss; pooled across thirteen leagues it is a few tenths of a percent, positive.

LICENCE

Issued to Evaluation copy. One perpetual licence to use and modify this model inside your own organisation. Redistribution of the specification or the reference implementation is not included. No warranty is given: the validation above is the whole of the claim being made.