Model selection and assessment

Definition

Model selection - estimating the performance of different models in order to choose the best one.

Model assessment - estimating the generalization error of a given model.

Performance metrics

Summarizes the performance of a model in a real valued score (higher is better) or error (lower is better).

Most algorithms try to minimize a loss function; loss and error don't have to be same.

How to choose performance metric?

Business goal

The performance metric should be a good proxy of the business goal. Do you need to make decisions? Rank items? Provide accurate probabilities?

E.g. Ad placement and well calibrated click probabilities -> log-loss.

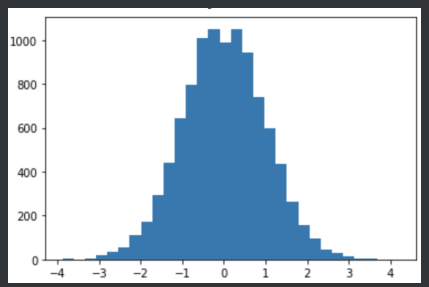
Response distribution

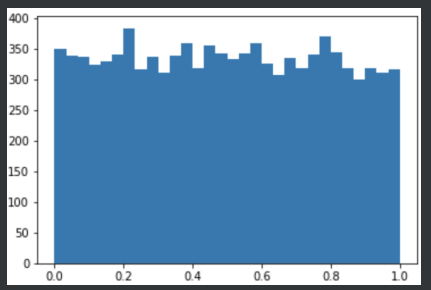
Is it positive count data?

Do you expect outliers in the response? How do you want them to be handled?

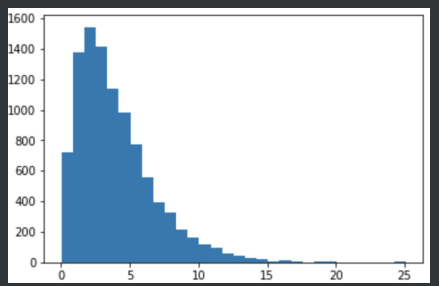
Based on distribution

Squared / absolute error?

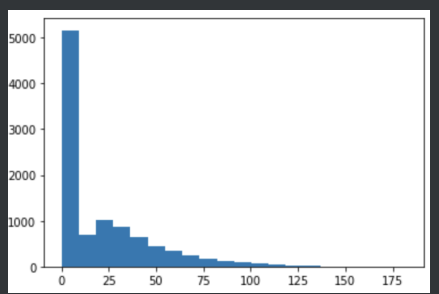




Poisson / gamma?



Tweedie?



Out-of-sample error

Out-of-sample data - data that wasn't used to fit model parameters.

Train-validation-holdout

It's common practice to partition data into there disjoin sets:

- Train: used to fit model parameters
- Validation: Used for model selection
- Holdout: Used to estimate generalization error

Partitioning

Random (stratified) partitioning

Split sets randomly. Pitfall: make sure you shuffle!

Grouped partitioning

Split sets by mutually exclusive groups (e.g. user IDs for recommendations)

Out-of-time partitioning

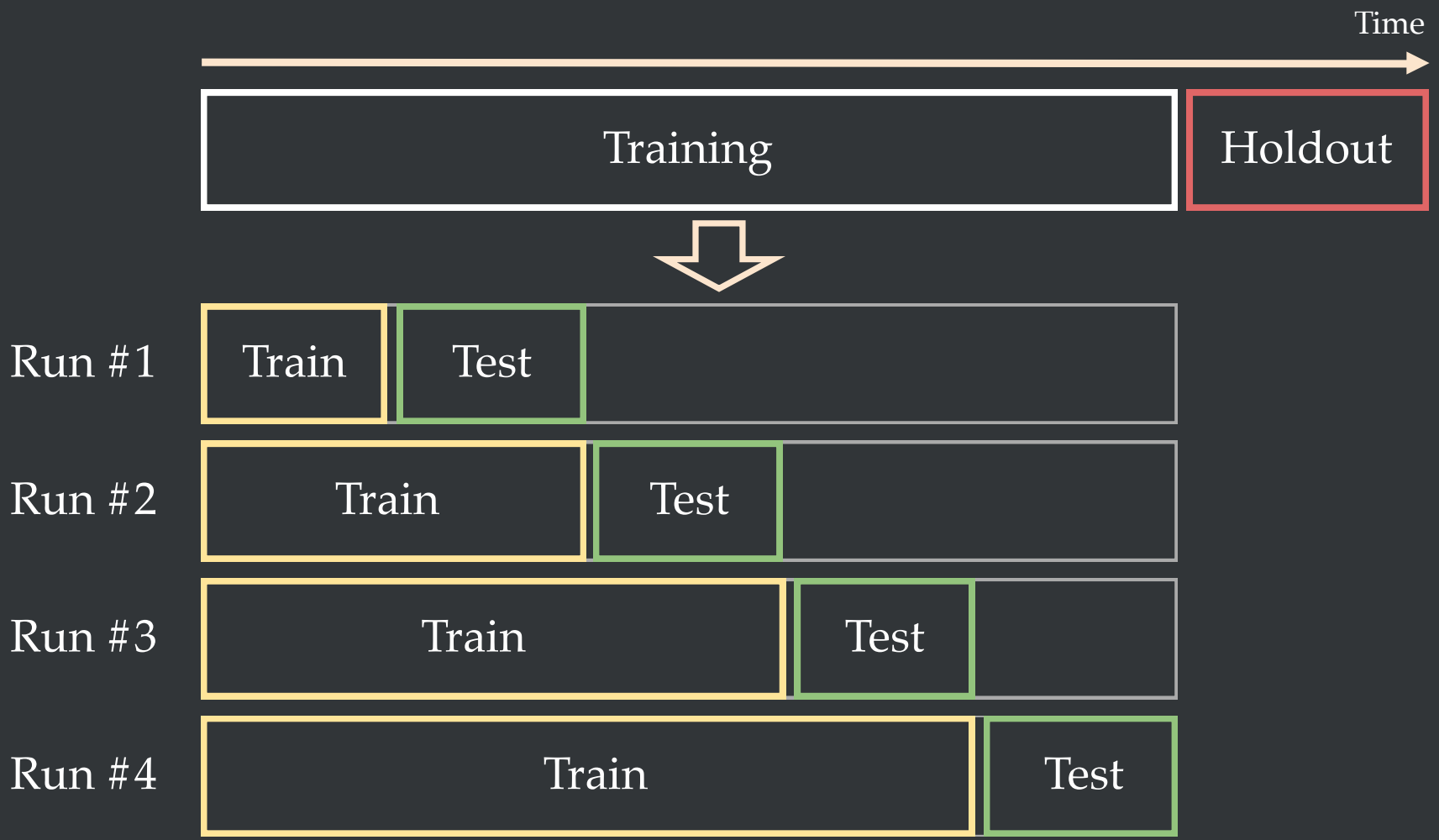
Split sets by chronological order. Note: extrapolation vs interpolation

Mind the gap: use gaps to forecast multiple steps ahead

When estimating generalization error: fit model on train+validation

Cross-validation

Out-of-time validation



Nested cross validation

Multiple layers of cross-validation for model tuning, selection and assessment.

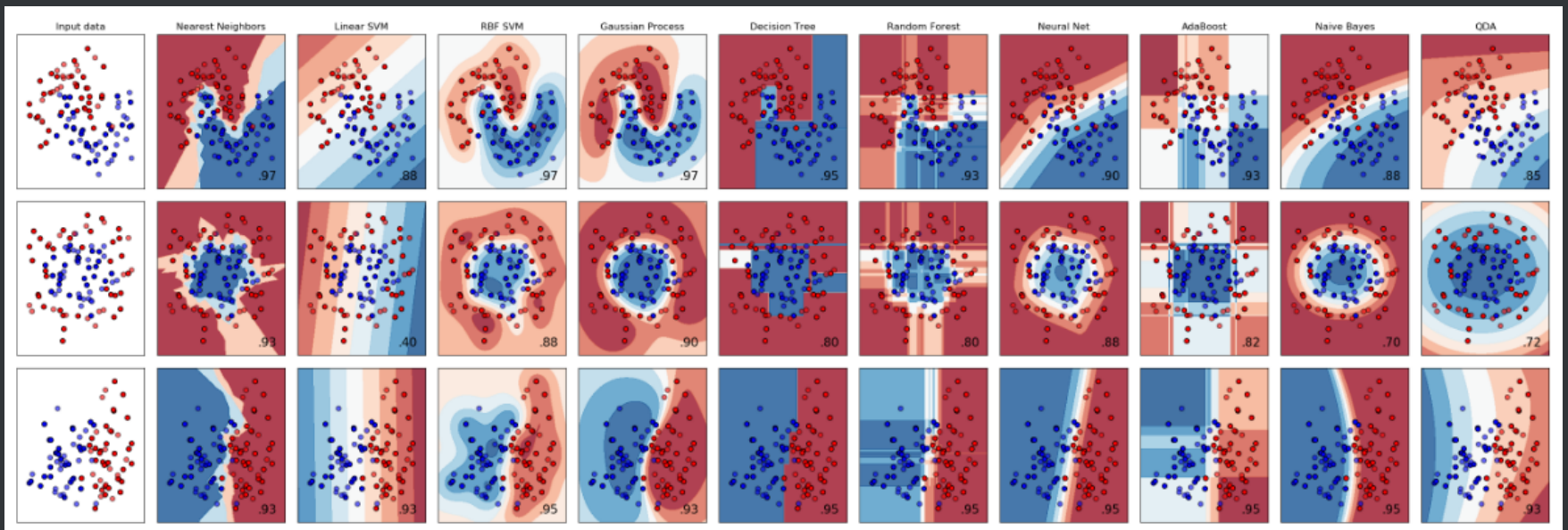
Links:

- scikit-learn user guide
- R. Hyndman TSCV
- Towards Data Science Article

No-free-lunch Theorem

Why do we need to do model selection? Aren't Deep Neural Networks the only algorithm we need?

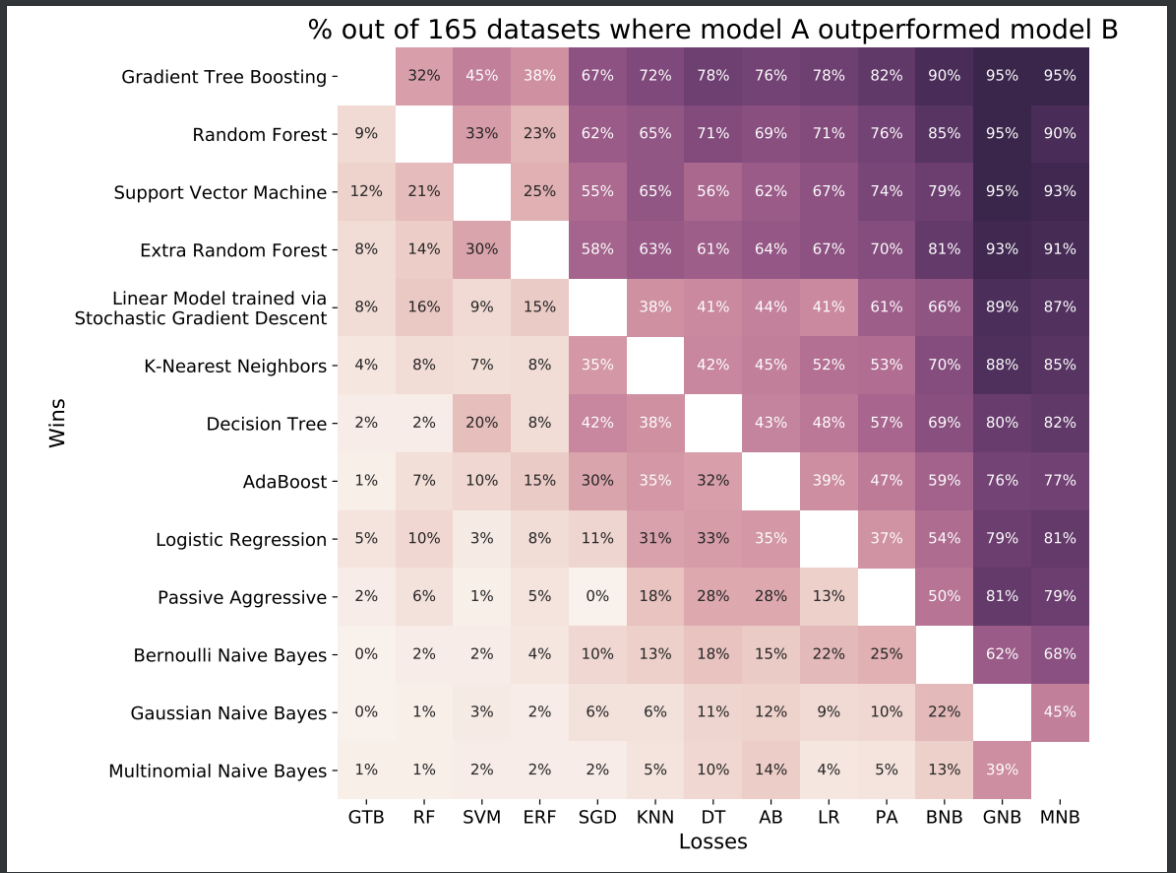
NFL: whenever a learning algorithm performs well on some function, as measured by out-of-training set generalization, it must perform poorly on some other(s).



Source: Scikit-learn user guide

NFL revised

There is no silver bullet...



Source: R. Olson et. al. (2017) "Data-drive advice for applying Machine learning to Bioinformatics Problems."