H_2O.ai

Binning in H2O Decision Trees

- Binning Numeric Features
  - Traditionally split points are chosen by sorting the each feature and inspecting an induced split.
  - For big data even when running parallel and distributed this can be computationally expensive so we approximate sorting with binning.
  - **More Bins, More Accurate** The number of bins can be specified by the user and it is the minimum number of bins required in a histogram built for each feature.
- Binning Categorical Features
  - **High Cardinality Features** slow model builds by inducing splits by each level.
  - Bin the levels in a categorical column according to “nbins_cat” parameter.
  - **More Bins, More Likely To Overfit** Increasing the number of bins can lead to splits on a single category, which can lead to overfitting.

Supervised Learning:

TREE ENSEMBLES

Binning in H2O Decision Trees

- Binning Numeric Features
  - Traditionally split points are chosen by sorting the each feature and inspecting an induced split.
  - For big data even when running parallel and distributed this can be computationally expensive so we approximate sorting with binning.
  - **More Bins, More Accurate** The number of bins can be specified by the user and it is the minimum number of bins required in a histogram built for each feature.
- Binning Categorical Features
  - **High Cardinality Features** slow model builds by inducing splits by each level.
  - Bin the levels in a categorical column according to “nbins_cat” parameter.
  - **More Bins, More Likely To Overfit** Increasing the number of bins can lead to splits on a single category, which can lead to overfitting.