H_2O.ai

Low Frequency Categories

- Most real world datasets contain categorical data.

- Use knowledge about hierarchical data to collapse categories.
- Use Cross-Validated Mean Target Encoding.
- Use Cross-Validated Weight of Evidence Encoding when modeling binary outcome.
- Use H2O’s **`categorical_encoding`** and **`nbins_cat`** decision tree tuning arguments

Real Data

Solutions

Too Many Categories

- Problems can arise if you have **too many categories:**
  - Computational complexity during estimation
  - Infrequent categories can lead to overfitting

Target Mean Encoding

What?

Replace categorical variables with the mean of the response

Why?

Categorical variables increase the number of features (dummy encoding) and can cause us to overfit

Low Frequency Categories

Real Data

- Most real world datasets contain categorical data.

Too Many Categories

- Problems can arise if you have **too many categories**:
  - Computational complexity during estimation
  - Infrequent categories can lead to overfitting

Solutions

- Use knowledge about hierarchical data to collapse categories.
- Use Cross-Validated Mean Target Encoding.
- Use Cross-Validated Weight of Evidence Encoding when modeling binary outcome.
- Use H2O’s `categorical_encoding` and `nbins_cat` decision tree tuning arguments