Data Manipulation

This section provides examples of common tasks performed when preparing data for machine learning. These examples are run on a local cluster.

Note: The examples in this section include datasets that are pulled from GitHub and S3.

- Uploading a File
- Importing a File
- Importing Multiple Files
- Downloading data
- Changing the Column Type
- Combining Columns from Two Datasets
- Combining Rows from Two Datasets
- Fill NAs
- Group By
- Imputing Data
- Merging Two Datasets
- Pivoting Tables
- Replacing Values in a Frame
- Slicing Columns
- Slicing Rows
- Sorting Columns
- Splitting Datasets into Training/Testing/Validating
- Tokenize Strings

Feature Engineering

H2O also has methods for feature engineering. Target Encoding is a categorical encoding technique which replaces a categorical value with the mean of the target variable (especially useful for high-cardinality features). Word2vec is a text processing method which converts a corpus of text into an output of word vectors.

- Target Encoding
- Word2vec