IRIS - Glossary

Overview

This web application will provide a visual representation for the .iris file output from Compellon's IRIS Informational Assessment Engine.
The .iris file is not uploaded to the web, it is processed using only the local systems browser resources in order to generate the visual display.

Pre-Analysis Cleaning

Non-Informative Columns

These columns have little or no relevant information for any analysis. Examples are columns that have overly unique values (such as id columns), too many missing values or columns that contain too many low frequency values.

Duplicate Columns

These columns are identified as having redundant information to each other and therefore only one of the columns are not represented in the Relationship Graph. It is suggested that all but one of the duplicate columns be excluded from analytical tasks such as selection of drivers, construction of models, etc.

Un-targeted Glossary

Un-targeted Analysis

The analysis is performed without the user specifying any target columns. The outputs are an information graph with a central target and alternate targets, as well as a list of Non-Informative columns. For each target identified, the application reports results such as information quality score, consistency score, proxies etc.

Information Graph

A connected directed acyclic graph representing hierarchies of relations of columns in the dataset.

Proxies to the Target

For each target column, these columns are identified as having redundant information relevant to the target and are not represented in the Relationship Graph. It is suggested that these columns be excluded from analytical tasks such as selection of drivers, construction of models, etc.

Central Target

The center of the information graph. The one column in the graph which only has incoming edges (no outgoing edges).

Alternate Targets

These columns, along with the central target are identified as good candidates to be targeted for analytical purposes (model construction, discovery of insights, etc).

Information Quality Score

For each target column, Information Quality indicates the amount of meaningful information (duplicate information removed) that exists in the dataset to construct a stable predictor of the future for the target column.

Information Consistency Scores

For each target column, this score indicates the consistency of information across different (random) sections in the dataset. A low score indicates varying patterns of relationships and suggests that attempting to use a single model for the whole dataset would be suboptimal, potentially resulting in poor accuracy and stability. Further analysis such as discovery of predictive drivers and construction of models should be done after identifying segments of homogenous behaviors.

Targeted Glossary

Targeted Analysis

The analysis is performed in the same way as an Un-targeted Analysis but allows the user to specify a list of additional target columns for which an information quality score, consistency score and proxies for each target are calculated. The Untargeted Analysis section will be the same as if no targets were specified.

Information Quality Score

For each user specified target column, Information Quality indicates the amount of meaningful information (duplicate information removed) that exists in the dataset to construct a stable predictor of the future for the target column.

Information Consistency Score

For each user specified target column, this score indicates the consistency of information across different (random) sections in the dataset. A low score indicates varying patterns of relationships and suggests that attempting to use a single model for the whole dataset would be suboptimal, potentially resulting in poor accuracy and stability. Further analysis such as discovery of predictive drivers and construction of models should be done after identifying segments of homogenous behaviors.

Proxies to the Target

For each user specified target column, these columns are identified as having redundant information relevant to the target. It is suggested that these columns are excluded from analytical tasks such as selection of drivers, construction of models, etc.