Spotfire Statistica®

Weight of Evidence Formula Guide

Version 14.2.0 | March 2024

Contents

Weight of Evidence Module

The purpose of the Weight of Evidence (WoE) module is to provide flexible tools to recode the values in continuous and categorical predictor variables into discrete categories automatically, and to assign to each category a unique WoE value. This recoding is conducted in a manner that produces the largest differences between the recoded groups with respect to the WoE values. In addition, other constraints are observed while the program determines solutions for the optimal binning of predictors.

Optimal Coding of Predictors

Specifically, the goal of the algorithms implemented in the automated WoE module is to identify the best groupings for predictor variables that results in the greatest differences in WoE between groups.

For continuous variables the automated WoE module identifies the best recoding to weight-of-evidence values. For categorical predictors or interactions between coded predictors, users can combine groups with similar observed WoE to create new coded predictors with continuous weight-of-evidence value.

Continuous Variables

For continuous predictors, first a default coding is derived using the Classification and Regression Trees (C&RT) algorithm. For default categories with fewer than 20 groups Statistica explicitly searches through all possible combinations of default groups to achieve the least numbers of groups with the greatest Information Value (IV). When the number of groups is greater than 20, Statistica uses the CHAID approach. The CHAID approach is a modification to the CHAID algorithm where instead of the customary X^2 criterion, the change in WoE is used as the criterion.

Three types of constrained WoE recoding solutions are provided subject to their existence:

- Monotone solutions, where the WoE values of all adjacent recoded groups (intervals) either increase (positive monotone relationship of predictor intervals to WoE), or the WoE values of all adjacent recoded groups always decrease (negative monotone relationship of predictor intervals to WoE).
- Quadratic solutions, where the relationship between the coded value ranges (intervals) to WoE can have a single reversal so that the resulting function is either U-shaped or inverse-U-shaped.
- Cubic solutions, where the relationship between the coded value ranges (intervals) to WoE values can have two reversals so that the resulting function is S-shaped.

Two types of unconstrained WoE recoding solutions are provided:

- Custom coding is based on the default binning scheme with either C&RT or 10 equal groups of approximately equal size.

- The no restrictions coding is based on the custom solution after running either the exhaustive search or the CHAID algorithm.

Note that the initial bins might be adjusted prior to the algorithm in order to make sure that each bin satisfies the minimum N and minimum Bad N user specified parameters.

Categorical Variables

For categorical (discrete) predictors, the default (original) grouping is further refined using the modified CHAID approach.

Two types of unconstrained WoE recoding solutions are provided:

- Custom coding is based on the default binning of the group.
- The no restrictions coding is based on the default categorization provided by the modified CHAID algorithm.

Note that the initial bins might be adjusted prior to the algorithm in order to make sure that each bin satisfies the minimum N and minimum Bad N user specified parameters.

Interactions

For pairs of coded predictors the modified CHAID approach is implemented using interaction coding of the two-way interaction table or user-defined coding.

Statistics

Chi-square

$$X^2 = \sum_{i=1}^{K} \frac{(Observed_i - Expected_i)^2}{Expected_i}$$

This statistic is distributed according to a chi-square distribution with degrees of freedom equal to the difference between the number of parameters under the alternative hypothesis and the number of parameters under the null hypothesis.

Cramer's V

$$V = \sqrt{\frac{X^2/N}{min_{(i-1)(j-1)}}}$$

Notation:

N = Total number of observations

$min_{(i-1)(j-1)}$ = Minimum of row dimension minus 1 and column dimension minus 1

F-test

$$F = \frac{\sum_i n_i \left(\bar{Y}_{i.} - \bar{Y}\right)^2 / K - 1}{\sum_{ij} \left(Y_{ij} - \bar{Y}_{i.}\right)^2 / N - K}$$

Notation:

$\bar{Y}_{i.}$ = sample mean of the i^{th} group

n_i = number of observations in the i^{th} group

$\bar{Y}$ = overall mean of the data

K = denotes the number of groups

Y_{ij} = j^{th} observation in the i^{th} out of K groups

N = overall sample size

Gini

$$g = 2\left(\frac{Number\ of\ Bads}{N}\right)\left(\frac{Number\ of\ Goods}{N}\right)$$

Notation:

N = Total number of observations

Information Value (IV)

$$IV = \sum_{i-1}^{K} \left[\left(Relative\ Frequency\ of\ Goods_i - Relative\ Frequency\ of\ Bads_i\right) * ln\left(\frac{Relative\ Frequency\ of\ Goods}{Relative\ Frequency\ of\ Bads}\right)\right]$$

The IV of a predictor is related to the sum of the (absolute) values for WoE over all groups. Thus, it expresses the amount of diagnostic information of a predictor variable for separating the Goods from the Bads.

Kolmogorov-Smirnov (KS) test

For all Good observations, predicted probability of Bad is computed, that is the relative frequency of bad cases in the bin a Good observation is placed. This process is repeated for all Bad observations. The KS test is then completed with the Good/Bad indicator as the group variable and the predicted probability of Bad as the response.

$$Z = max_j^{|D_j|} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} j$$

Significance level (p) approximation is based on the formula:

$$p = 2 \sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 \left(KS \sqrt{\frac{n_1}{n_1+n_2}+0.12+0.11} \Big/ \sqrt{\frac{n_1}{n_1+n_2}} \right)^2}$$

Logit Transformation (Logg Odds)

$$Logit = ln\left(\frac{\frac{Number\ of\ Goods}{N},}{\frac{Number\ of\ Bads}{N},}\right)$$

Mean

$$\bar{x} = \frac{\sum x}{n}$$

Somer's D

If ties are present:

$$d = \frac{(n_c - n_d)}{t}$$

If ties are not present:

$$d = 2\,c\, - 1 \text{ where } c = {(n_c + 0.5(t - n_c - n_d))}/{t}$$

Note: Sorting of cases for calculation of Somer's d is based on the relative frequency of bad, that is, estimated probably of bad.

Notation:

t = total number of pairs with different responses of good/bad

nc = number of pairs of cases where the case with the lower ordered response value has a lower predicted mean score than the case with the higher ordered response value.

nd = number of pairs of cases where the case with the lower ordered response value has a higher predicted mean score than the case with the higher ordered response value.

Weight of Evidence (WoE)

$$WoE = \left[ln\left(\frac{Relative\ Frequency\ of\ Goods}{Relative\ Frequency\ of\ Bads}\right)\right] * 100$$

The value of WoE is 0 if the odds of Relative Frequency of Goods / Relative Frequency Bads is equal to 1. If the Relative Frequency of Bads in a group is greater than the Relative Frequency of Goods, the odds ratio is less than 1 and the WoE is a negative number; if the Relative Frequency of Goods is greater than the Relative Frequency of Bads in a group, the WoE value is a positive number.

Notes

The WoE recoding of predictors is particularly well suited for subsequent modeling using Logistic Regression. Specifically, logistic regression fits a linear regression equation of predictors (or WoE- coded continuous predictors) to predict the logit-transformed binary Goods/Bads dependent or Y variable. Therefore, by using WoE-coded predictors in logistic regression, the predictors are all prepared and coded to the same WoE scale, and the parameters in the linear logistic regression equation can be directly compared, for example, when using the new modeling tools for Marginal Stepwise Logistic Regression.

Spotfire Documentation and Support Services

For information about this product, you can read the documentation, contact Spotfire Support, and join Spotfire Community.

How to Access Spotfire Documentation

Documentation for Spotfire products is available on the Product Documentation website, mainly in HTML and PDF formats.

The Product Documentation website is updated frequently and is more current than any other documentation included with the product.

Product-Specific Documentation

The documentation for this product is available on Spotfire Statistica® Product Documentation page and Spotfire Statistica®.

How to Contact Support for Spotfire Products

You can contact the Support team in the following ways:

- To access the Support Knowledge Base and getting personalized content about products you are interested in, visit our product Support website.
- To create a Support case, you must have a valid maintenance or support contract with a Cloud Software Group entity. You also need a username and password to log in to the product Support website. If you do not have a username, you can request one by clicking **Register** on the website.

How to Join Spotfire Community

Spotfire Community is the official channel for Spotfire customers, partners, and employee subject matter experts to share and access their collective experience. Spotfire Community offers access to Q&A forums, product wikis, and best practices. It also offers access to

extensions, adapters, solution accelerators, and tools that extend and enable customers to gain full value from Spotfire products. In addition, users can submit and vote on feature requests from within the Spotfire Ideas Portal. For a free registration, go to Spotfire Community.

Legal and Third-Party Notices

SOME CLOUD SOFTWARE GROUP, INC. (“CLOUD SG”) SOFTWARE AND CLOUD SERVICES EMBED, BUNDLE, OR OTHERWISE INCLUDE OTHER SOFTWARE, INCLUDING OTHER CLOUD SG SOFTWARE (COLLECTIVELY, “INCLUDED SOFTWARE”). USE OF INCLUDED SOFTWARE IS SOLELY TO ENABLE THE FUNCTIONALITY (OR PROVIDE LIMITED ADD-ON FUNCTIONALITY) OF THE LICENSED CLOUD SG SOFTWARE AND/OR CLOUD SERVICES. THE INCLUDED SOFTWARE IS NOT LICENSED TO BE USED OR ACCESSED BY ANY OTHER CLOUD SG SOFTWARE AND/OR CLOUD SERVICES OR FOR ANY OTHER PURPOSE.

USE OF CLOUD SG SOFTWARE AND CLOUD SERVICES IS SUBJECT TO THE TERMS AND CONDITIONS OF AN AGREEMENT FOUND IN EITHER A SEPARATELY EXECUTED AGREEMENT, OR, IF THERE IS NO SUCH SEPARATE AGREEMENT, THE CLICKWRAP END USER AGREEMENT WHICH IS DISPLAYED WHEN ACCESSING, DOWNLOADING, OR INSTALLING THE SOFTWARE OR CLOUD SERVICES (AND WHICH IS DUPLICATED IN THE LICENSE FILE) OR IF THERE IS NO SUCH LICENSE AGREEMENT OR CLICKWRAP END USER AGREEMENT, THE LICENSE(S) LOCATED IN THE “LICENSE” FILE(S) OF THE SOFTWARE. USE OF THIS DOCUMENT IS SUBJECT TO THOSE SAME TERMS AND CONDITIONS, AND YOUR USE HEREOF SHALL CONSTITUTE ACCEPTANCE OF AND AN AGREEMENT TO BE BOUND BY THE SAME.

Statistica, Spotfire, Process Tree Viewer, Process Data Explorer, Predictive Claims Flow, Live Score, Electronic Statistics Textbook, and Data Health Check, are either registered trademarks or trademarks of Cloud Software Group, Inc. in the United States and/or other countries.

Java and all Java based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

All other product and company names and marks mentioned in this document are the property of their respective owners and are mentioned for identification purposes only. You acknowledge that all rights to these third party marks are the exclusive property of their respective owners. Please refer to Cloud SG’s Third Party Trademark Notices (https://www.cloud.com/legal) for more information.

This document includes fonts that are licensed under the SIL Open Font License, Version 1.1, which is available at: https://scripts.sil.org/OFL

Copyright (c) Paul D. Hunt, with Reserved Font Name Source Sans Pro and Source Code Pro.

Cloud SG software may be available on multiple operating systems. However, not all operating system platforms for a specific software version are released at the same time. See the “readme” file

for the availability of a specific version of Cloud SG software on a specific operating system platform.

THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, OR NON-INFRINGEMENT.

THIS DOCUMENT COULD INCLUDE TECHNICAL INACCURACIES OR TYPOGRAPHICAL ERRORS. CHANGES ARE PERIODICALLY ADDED TO THE INFORMATION HEREIN; THESE CHANGES WILL BE INCORPORATED IN NEW EDITIONS OF THIS DOCUMENT. CLOUD SG MAY MAKE IMPROVEMENTS AND/OR CHANGES IN THE PRODUCT(S), THE PROGRAM(S), AND/OR THE SERVICES DESCRIBED IN THIS DOCUMENT AT ANY TIME WITHOUT NOTICE.

THE CONTENTS OF THIS DOCUMENT MAY BE MODIFIED AND/OR QUALIFIED, DIRECTLY OR INDIRECTLY, BY OTHER DOCUMENTATION WHICH ACCOMPANIES THIS SOFTWARE, INCLUDING BUT NOT LIMITED TO ANY RELEASE NOTES AND "README" FILES.

This and other products of Cloud SG may be covered by registered patents. For details, please refer to the Virtual Patent Marking document located at https://www.tibco.com/patents.