**WEIGHT OF EVIDENCE (WOE)**

The weight of evidence tells the predictive power of an independent variable in relation to the dependent variable.  
  
Since it evolved from credit scoring world, it is generally described as a measure of the separation of good and bad customers. **"Bad Customers"** refers to the customers who defaulted on a loan. and **"Good Customers"**refers to the customers who paid back loan.

|  |
| --- |
| [https://4.bp.blogspot.com/-X1m0w40w0xg/V9V_7LS1AQI/AAAAAAAAFWc/f4bgPvE1In8Q13kGGBghp98MeWma8KgqACLcB/s320/woe.png](https://4.bp.blogspot.com/-X1m0w40w0xg/V9V_7LS1AQI/AAAAAAAAFWc/f4bgPvE1In8Q13kGGBghp98MeWma8KgqACLcB/s1600/woe.png) |
| WOE Calculation |

***Distribution of Goods****- % of Good Customers in a particular group****Distribution of Bads -****% of Bad Customers in a particular group*

Many people do not understand the terms goods/bads as they are from different background than the credit risk. It's good to understand the concept of WOE in terms of **events and non-events**. It is calculated by taking the natural logarithm (log to base e) of division of % of non-events and % of events.

*WOE = In(% of non-events ➗ % of events)*

|  |
| --- |
| <https://3.bp.blogspot.com/-eqZJpJZ4Kig/VPnZUBaP7II/AAAAAAAADkc/yeW8XVL35dA/s1600/weight%2Bof%2Bevidence.png> |
| Weight of Evidence Formula |

**Steps of Calculating WOE**

1. For a continuous variable, split data into 10 parts (or lesser depending on the distribution).
2. Calculate the number of events and non-events in each group (bin)
3. Calculate the % of events and % of non-events in each group.
4. Calculate WOE by taking natural log of division of % of non-events and % of events

**Note :** For a categorical variable, you do not need to split the data (Ignore Step 1 and follow the remaining steps)

|  |
| --- |
| [Weight of Evidence and Information Value](https://1.bp.blogspot.com/-eNJ4G_DqhUs/XNRigoIXh2I/AAAAAAAAHiU/8Bt059tLpDoc6DKBUPCOf3ffOW2eOO2DQCLcBGAs/s1600/IV_WOE.png) |
| Weight of Evidence and Information Value Calculation |

**Download :**[**Excel Template for WOE and IV**](https://sites.google.com/site/pocketecoworld/WOE%20and%20IV.xlsx)  
  
  
**Terminologies related to WOE**  
 **1. Fine Classing**

*Create 10/20 bins/groups for a continuous independent variable and then calculates WOE and IV of the variable*

**2. Coarse Classing**

*Combine adjacent categories with similar WOE scores*

**Usage of WOE**

Weight of Evidence (WOE) helps to transform a continuous independent variable into a set of groups or bins based on similarity of dependent variable distribution i.e. number of events and non-events.

**For continuous independent variables :**First, create bins (categories / groups) for a continuous independent variable and then combine categories with similar WOE values and replace categories with WOE values. Use WOE values rather than input values in your model.

*data age1;  
set age;  
if age = . then WOE\_age = 0.34615;  
if age >= 10 then WOE\_age = -0.03012;  
if age >= 20 then WOE\_age = 0.07689;  
run;*

*proc logistic data=age1 descending;  
model y = WOE\_age;  
run;*

**For categorical independent variables :**Combine categories with similar WOE and then create new categories of an independent variable with continuous WOE values. In other words, use WOE values rather than raw categories in your model. The transformed variable will be a continuous variable with WOE values. It is same as any continuous variable.

**Why combine categories with similar WOE?**

It is because the categories with similar WOE have almost same proportion of events and non-events. In other words, the behavior of both the categories is same.

**Rules related to WOE**

1. Each category (bin) should have at least 5% of the observations.
2. Each category (bin) should be non-zero for both non-events and events.
3. The WOE should be distinct for each category. Similar groups should be aggregated.
4. The WOE should be monotonic, i.e. either growing or decreasing with the groupings.
5. Missing values are binned separately.

**Number of Bins (Groups)**  
  
In general, 10 or 20 bins are taken. Ideally, each bin should contain at least 5% cases. The number of bins determines the amount of smoothing - the fewer bins, the more smoothing. If someone asks you ' "why not to form 1000 bins?" The answer is the fewer bins capture important patterns in the data, while leaving out noise. Bins with less than 5% cases might not be a true picture of the data distribution and might lead to model instability.  
 **Handle Zero Event/ Non-Event**  
  
If a particular bin contains no event or non-event, you can use the formula below to ignore missing WOE. We are adding 0.5 to the number of events and non-events in a group.  
  
**AdjustedWOE**= ln (((Number of non-events in a group + 0.5) / Number of non-events)) / ((Number of events in a group + 0.5) / Number of events))

**How to check correct binning with WOE**

1. The WOE should be monotonic i.e. either growing or decreasing with the bins. You can plot WOE values and check linearity on the graph.

2. Perform the WOE transformation after binning. Next, we run logistic regression with 1 independent variable having WOE values. If the slope is not 1 or the intercept is not **ln(% of non-events / % of events)** then the binning algorithm is not good. **[Source :** [Article](http://www.m-hikari.com/ams/ams-2014/ams-65-68-2014/zengAMS65-68-2014.pdf)**]**  
  
**Benefits of WOE**

1. It can treat outliers. Suppose you have a continuous variable such as annual salary and extreme values are more than 500 million dollars. These values would be grouped to a class of (let's say 250-500 million dollars). Later, instead of using the raw values, we would be using WOE scores of each classes.
2. It can handle missing values as missing values can be binned separately.
3. Since WOE Transformation handles categorical variable so there is no need for dummy variables.