**Information Value (IV) and Weight of Evidence (WOE)**

**Information value** is a very useful concept for variable selection during model building.

Chi Square value, an extensively used measure in statistics, is a good replacement for IV (information value).

However, IV is a popular and widely used measure in the industry. The reason for this is some very convenient rules of thumb for variables selection associated with IV – these are really handy as you will discover later in this article.

The formula for information value is shown below.

IV = \sum (Distribution Good_{i}-Distribution Bad_{i})\times ln(\frac{\normalsize Distribution Good_{i}}{\normalsize Distribution Bad_{i}})   

**Weight of Evidence (WOE),** is the log component in information value.

Weight of Evidence = ln(\frac{Distribution Good_{i}}{Distribution Bad_{i}})   

Hence, IV can further be written as the following.

IV = \sum (Distribution Good_{i}-Distribution Bad_{i})\times WOE_{i} 

If you examine both information value and weight of evidence carefully then you will notice that both these values will break down when either the distribution good or bad goes to zero. A mathematician will hate it.

The assumption, a fair one, is that this will never happen while a scorecard development because of the reasonable sample size.

A word of caution, if you are developing non-standardized scorecards with smaller sample size use IV carefully.

**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

**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.

**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**

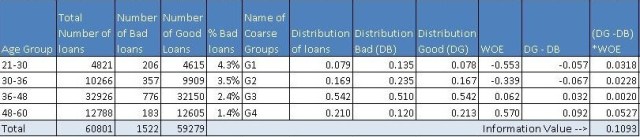
1. The WOE should be monotonic i.e. either growing or decreasing with the bins. You can plot WOE values and check linearity on the graph.

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.  
  
**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.

**Case Study**

Let us calculate both information value and weight of evidence for these coarse classes.

[](https://i0.wp.com/ucanalytics.com/blogs/wp-content/uploads/2013/10/IV-WOE.jpg)

Let us examine this table. Here, distribution of loans is the ratio of loans for a coarse class to total loans. For the group 21-30, this is 4821/60801 = 0.079.

Similarly, distribution bad (DB) = 206/1522 = .135 and distribution good = 4615/59279 (DG) = 0.078.

Additionally, DG-DB = 0.078 – 0.135 = – 0.057.

Further, WOE = ln(0.078/0.135) = -0.553.

Finally, component of IV for this group is (-0.057)\*(-0.553) = 0.0318.

Similarly, calculate the IV components for all the other coarse classes. Adding these components will produce the IV value of 0.1093 (last column of the table).

Now the question is how to interpret this value of IV?  The answer is the rule of thumb described below.

|  |  |
| --- | --- |
| **Information Value** | **Predictive Power** |
| < 0.02 | useless for prediction |
| 0.02 to 0.1 | Weak predictor |
| 0.1 to 0.3 | Medium predictor |
| 0.3 to 0.5 | Strong predictor |
| >0.5 | Suspicious or too good to be true |

Typically, variables with medium and strong predictive powers are selected for model development.

However, some school of thoughts would advocate just the variables with medium IVs for a broad-based model development.

Notice, the information value for age is 0.1093 hence it is barely falling in the medium predictors’ range.

**Logistic Regression with Weight of Evidence (WOE)**

Let us create a logistic regression model with weight of evidence of the coarse classes as the value for the independent variable age.

The following are the results generated through a statistical software.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Logistic Regression Results (Age Groups and Bad Rates)** | | | | | |
| **Predictor** | **Coefficient** | **Std. Er** | **Z** | **P** | **Odds Ratio** |
| **Constant** | -3.66223 | 0.0263162 | -139.16 | 0 |  |
| **WOEAge** | -1 | 0.0796900 | -12.55 | 0 | 0.37 |

If we estimate the value of bad rate for the age group 21-30 using the above information.

P(Bad Loan)=\frac{\large e^{(-1\times -0.553)-3.66223}}{\large 1+e^{(-1\times -0.553)-3.66223}}=4.3\%   