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

# **1. Introduction**

Logistic regression is one of the most commonly used statistical techniques for solving binary classification problems and is widely accepted across various domains.  
The concepts of Weight of Evidence (WOE) and Information Value (IV) evolved from logistic regression and have been integral to credit scoring for over four to five decades.  
They are commonly used to explore data and screen variables, especially in credit risk modeling projects such as Probability of Default (PD).

# **2. What is Weight of Evidence (WOE)?**

Weight of Evidence (WOE) measures the predictive power of an independent variable relative to the dependent variable.  
Originating in credit scoring, WOE is typically described as a measure of how well a variable separates good customers (those who repay loans) from bad customers (those who default).

# **Formula:**

WOE=ln(Distribution of Bads/Distribution of Goods​)

Where:

* Distribution of Goods = % of good customers in a group
* Distribution of Bads = % of bad customers in a group
* ln = Natural logarithm

# **Key Inference:**

* Positive WOE: Distribution of Goods > Distribution of Bads
* Negative WOE: Distribution of Goods < Distribution of Bads

**fyi:**

* Natural log (ln) of a number greater than 1 is positive.
* Natural log (ln) of a number less than 1 is negative.

**Alternate definition:**WOE can also be expressed in terms of events and non-events:

WOE=ln(% of events% of non-events​)

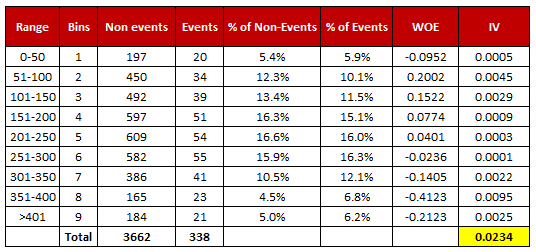
**How to Calculate Weight of Evidence (WOE)**

Follow these steps:

1. For continuous variables:
   * Split the data into 10 bins (or fewer depending on the distribution).
2. For each bin (or for each category in a categorical variable):

* Calculate the number of events and non-events.
* Calculate the % of events and % of non-events.
* Compute the WOE using the formula above.

Note:  
For categorical variables, no binning is required. Directly calculate WOE for each category.



**Terminologies Related to WOE**

**1. Fine Classing**

Fine classing involves creating 10–20 bins or groups for a continuous independent variable, followed by calculating the WOE and IV for each group.

**2. Coarse Classing**

Coarse classing combines adjacent bins or categories that have similar WOE values to simplify the model and reduce overfitting.

**Usage of WOE**

Weight of Evidence (WOE) transforms a continuous or categorical independent variable into a set of groups based on the distribution of the dependent variable (i.e., the number of events and non-events).

**For continuous variables:**

* First, create bins (groups/categories).
* Then, combine bins with similar WOE values.
* Replace the raw input values with their corresponding WOE values for modeling.

**For categorical variables:**

* Merge categories with similar WOE values.
* Replace raw categories with WOE values, transforming the variable into a continuous form suitable for modeling.

**Rules Related to WOE**

* Each bin should contain at least 5% of the total observations.
* Each bin must have a non-zero number of both events and non-events.
* WOE values should be distinct across bins; similar bins should be merged.
* WOE values should be monotonic (consistently increasing or decreasing) across bins.
* Missing values should be handled by placing them into a separate bin.

**Number of Bins (Groups)**

* Typically, 10 to 20 bins are used.
* Each bin should ideally contain at least 5% of the data.
* Fewer bins help capture significant patterns while smoothing out noise.
* Very small bins (e.g., bins with <5% of cases) can lead to unstable models.

**Why not use 1000 bins?**  
Using too many bins can capture random noise rather than real patterns, leading to overfitting and model instability.

**Handling Zero Events or Non-Events**

If a bin contains no events or no non-events, adjust the WOE calculation as follows:

**Adjusted WOE**=ln((Events in group+0.5)/Total events(Non-events in group+0.5)/Total non-events​)

Adding 0.5 ensures no division by zero.

**How to Validate Correct Binning with WOE**

1. **Monotonicity Check:**  
   Plot WOE values across bins. A good binning strategy shows WOE values that either consistently increase or decrease.
2. **Regression Check:**  
   After WOE transformation, run a logistic regression with the WOE-transformed variable as the only predictor.
   * The slope should be close to 1.
   * The intercept should be close to **ln(%non-events/%events)**.

If these conditions are not met, the binning may need revision.

**Benefits of Using Weight of Evidence**

1. **Outlier Handling:**  
   Extreme values (e.g., salaries > $500 million) are grouped into bins, preventing them from distorting the model.
2. **Missing Value Handling:**  
   Missing values can be assigned to a dedicated bin.
3. **Simplified Modeling:**WOE transformation removes the need for creating dummy variables for categorical inputs.
4. **Linear Relationship with Log Odds:**WOE transformation creates a strong linear relationship between the independent variable and the log odds of the event, which is critical for logistic regression.  
   Without WOE, achieving this linearity often requires multiple trial-and-error transformations (e.g., log, square root).

**What is Information Value (IV)?**

Information Value (IV) is a key metric for variable selection in predictive modeling. It ranks variables based on their predictive strength.

The formula for IV is:

**IV = ∑ (% of non-events - % of events) \* WOE**

|  |  |
| --- | --- |
| **Information Value** | **Variable Predictiveness** |
| Less than 0.02 | Not useful for prediction |
| 0.02 to 0.1 | Weak predictive Power |
| 0.1 to 0.3 | Medium predictive Power |
| 0.3 to 0.5 | Strong predictive Power |
| >0.5 | Suspicious Predictive Power |

**Important Points About Information Value**

1. **Impact of Binning on Information Value:**Information Value (IV) generally increases as the number of bins or groups for an independent variable increases. However, caution is needed when creating more than 20 bins, as some bins may end up with very few events or non-events, potentially leading to unreliable or unstable results.
2. **Suitability for Model Types:**Information Value is primarily designed for variable selection in binary logistic regression models, where the relationship between independent variables and the log odds of the dependent variable is critical.
   * It is not an optimal feature selection method for models like Random Forest, Support Vector Machines (SVM), or other non-linear classifiers.
   * Models like Random Forest naturally capture complex non-linear relationships, and relying on IV-based variable selection might limit their predictive performance.

In short, IV is ideal when the model assumes a linear relationship with log odds (like logistic regression), but not when non-linearity and interactions dominate the modeling approach.

**Code Files:**

1. 2. 

For more info go through:

1. Youtube – <https://www.youtube.com/watch?v=XVjq45YSjsY>
2. <https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html>
3. <https://www.analyticsvidhya.com/blog/2021/06/understand-weight-of-evidence-and-information-value/>