**Information Value (IV)**

Information value is one of the most useful technique to select important variables in a predictive model. It helps to rank variables on the basis of their importance. The IV is calculated using the following formula :

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

|  |
| --- |
| <https://2.bp.blogspot.com/-hkTX-LJoANY/VPnv5Wd3UoI/AAAAAAAADk4/SZFPuuecbkg/s1600/IV.png> |
| Information Value Formula |

|  |  |
| --- | --- |
| **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 |

**According to Siddiqi (2006), by convention the values of the IV statistic in credit scoring can be interpreted as follows.**

If the IV statistic is:

1. Less than 0.02, then the predictor is not useful for modeling (separating the Goods from the Bads)
2. 0.02 to 0.1, then the predictor has only a weak relationship to the Goods/Bads odds ratio
3. 0.1 to 0.3, then the predictor has a medium strength relationship to the Goods/Bads odds ratio
4. 0.3 to 0.5, then the predictor has a strong relationship to the Goods/Bads odds ratio.
5. > 0.5, suspicious relationship (Check once)

**Important Points**

1. Information value increases as bins / groups increases for an independent variable. Be careful when there are more than 20 bins as some bins may have a very few number of events and non-events.
2. Information value is not an optimal feature (variable) selection method when you are building a classification model other than binary logistic regression (for eg. random forest or SVM) as conditional log odds (which we predict in a logistic regression model) is highly related to the calculation of weight of evidence. In other words, it's designed mainly for binary logistic regression model. Also think this way - Random forest can detect non-linear relationship very well so selecting variables via Information Value and using them in random forest model might not produce the most accurate and robust predictive model.