REFERENCES

1. Phik Correlation

The Phik matrix is a correlation matrix that measures the association between categorical variables. It is an extension of the phi coefficient, which is commonly used to measure the association between two binary variables. The Phik matrix takes into account the information contained in variables with more than two categories.

PhiK matrix has advantages like consistent usage for mixed variable types, capturing non-linear dependencies, and reverting to Pearson correlation for bi-variate normal input distribution.

To use PhiK correlation analyzer library, we can use this command: *pip install phik*

2. Weight of Evidence (WOE)

WOE (weight of evidence) is a widely used feature engineering and feature selection technique in scorecard modeling. This method ranks variables into strong, medium, weak, useless, etc. based on their ability to predict bad debt. The ranking standard is the information value index IV (information value), calculated from the WOE method.

The model also creates feature values for each variable, which measure the difference in distribution between good and bad. The WOE method processes continuous and categorical variables differently.

For continuous variables, WOE labels each observation based on the value label of the bin it belongs to. The bins are consecutive intervals determined from a continuous variable such that each bin has an equal number of observations. To determine the bins, we need to decide on the number of bins. We can imagine that the ends of bins are quantiles.

For categorical variables, WOE can consider each class as a bin or group several groups with few observations into one bin. Additionally, the degree of difference between the good/bad distribution measured through the WOE index can also be used to identify groups with the same categorical properties. If their WOE values are closer to each other, they will likely be grouped into one group. Null cases can also be considered a separate group if their number is significant or grouped into other groups if it is a minority.

Formula:

$$WOE = \ln\left(\frac{\%Good}{\%Bad}\right)$$

Where:

No observation: Number of observations in each bin. It will usually be divided equally between bins.

No Good: Number of bad debt records in each bin (label 1)

No Bad: Number of non-bad debt records in each bin (label 0)

Good/Bad: Ratio of Good/Bad records in each bin.

%Good: Distribution of good records across all bins

%Bad: Distribution of bad records across all bins

WOE has several advantages.. Firstly, it helps transform continuous independent variables into variables with a linear relationship. Secondly, outliers can be removed. Thirdly, the WOE value reflects the impact of each category on the dependent variable

However, WOE analysis can be challenging when determining the appropriate number of bins for a continuous variable and when to group or separate groups. WOE variables are always monotonic with the dependent variable, resulting in a correlation between the independent variables and a high risk of multicollinearity. Additionally, overfitting can occur when adjusting variables by grouping categories, leading to inaccurate results.

3. Information Value (IV)

Information value evaluates whether a variable has power in classifying bad debt or not

Formula:

$$IV = \sum_{i=1}^{n} (\%Good_i - \%Bad_i).WOE_i$$

Some documents provide standards for classifying the power of variables according to the IV value as below:

≤ 0.02: useless (feature has no effect in classifying good/bad records)

0.02 – 0.1: weak

0.1 – 0.3: medium

0.3 – 0.5: strong

≥ 0.5: suspicious