Statistical Binning of Categorical Risk Factors

Probability of Default Modeling

Andrija Djurovic

www.linkedin.com/in/andrija-djurovic

Statistical Binning in Credit Risk

- Statistical binning is a step typically applied in credit risk modeling.
- In simple terms, binning refers to the process of discretizing or grouping values into a certain number of bins.
- When performing binning, practitioners often follow several principles, the most common of which are:
    - Each bin should contain at least 5% of the observations.
    - Each bin should include at least one bad case.
    - Adjacent bins should represent different risk levels.
    - The risk levels across bins should follow either a monotonic or U-shaped trend.
    - The number of bins should not exceed ten.

Statistical Binning in Credit Risk cont.

- In credit risk modeling, statistical binning is almost exclusively associated with numeric risk factors and model outputs.
- Since the development of credit risk models typically involves both numeric and categorical risk factors, practitioners often face challenges in processing and engineering the categorical ones.
- In practice, categorical risk factors are usually subject to manual engineering. However, is this always the optimal way to prepare categorical variables for model development?
- Although less common in practice, statistical binning of categorical risk factors offers significant benefits. When performing binning on categorical variables, practitioners can, if needed, follow the same principles recognized as good practice for binning numeric risk factors.
- This presentation illustrates one possible approach to statistical binning of categorical risk factors for Probability of Default modeling. Practitioners can find further details and examples in Anderson (2007), The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation.

Categorical Risk Factors

- The development of credit risk models typically involves handling both numerical and categorical risk factors.
- Categorical risk factors may arise directly from data collection or be derived through the discretization or grouping of numerical variables.
- A key characteristic of categorical risk factors is that they can contain too many categories or unique values to be directly usable in model development.
- Another important aspect is whether the categories can be ordered according to their relationship with the target variable. For some variables, such an order is expected, while for others, it is not. A simple example is the days past due (DPD) buckets used as a risk factor, with the one-year default rate as the target variable. In this case, we expect categories with lower average DPD values to correspond to lower average one-year default rates, allowing us to impose an order on the categories based on their relationship with the target variable. Other categorical risk factors, however, may not exhibit such a natural ordering.
- Regardless of these characteristics, practitioners often face challenges when preparing categorical risk factors for further use in model development.
- One of the most common challenges is grouping the categories in a way that ensures meaningful and stable modeling results. This process is often performed manually and relies heavily on expert judgment.
- The following slides illustrate one approach for reducing the number of categories of a categorical risk factor while minimizing information loss in Probability of Default (PD) modeling.

Simulation Study

The following steps demonstrate a simplified process for reducing the number of bins of a categorical risk factor while ensuring minimal information loss:

1. Select the target variable and the categorical risk factor.
2. Define the ordering of the categories.
3. Summarize the data by category and calculate the Weight of Evidence (WoE) and Information Value (IV) for each bin, as well as the overall IV for the risk factor.
4. For each adjacent pair of bins, compute the sum of their individual IVs.
5. For each adjacent pair of bins, calculate the IV assuming the two bins were merged.
6. Identify the pair of bins with the smallest difference between the IV values obtained in steps 4 and 5, and merge those bins.
7. Recalculate the summary statistics for the risk factor, reflecting the newly merged bins.

Notes:

- Step 2: The categories can be ordered either alphabetically (based on their labels) or according to the observed default rate per category. In practice, alphabetical ordering is used when the categorical factor results from the binning of a numerical risk factor or when the categories have been defined based on expert judgment. In such cases, the category labels are typically designed to reflect a meaningful order. When ordering by the observed default rate, practitioners usually group categories that exhibit similar default rates.
- Step 3: The WoE and the overall IV are supportive metrics and are not directly required for the subsequent steps.
- The presented steps outline a general framework that can be further refined by incorporating additional constraints, such as a minimum number of observations per bin, a minimum number of defaults, or specific treatment of missing or special-case categories.

Simulation Results

The simulation results presented in this and the following slides are based on the data available here.

The objective of the simulation is to reduce the number of bins for the analyzed risk factor from five to four, while ensuring minimal information loss.

1. The target variable is contained in the column y, and the risk factor in the column x.
2. Since the risk factor results from the binning of a numerical variable, the order of its bins is determined by the labels assigned to the bins.
3. The following table provides a summary where:
    - bin denotes the risk factor category;
    - no represents the number of observations;
    - ng indicates the number of "good" cases (binary indicator equal to 0);
    - nb represents the number of "bad" cases (binary indicator equal to 1);
    - woe is the Weight of Evidence;
    - iv.b is the Information Value (IV) of each bin;
    - iv.s is the overall IV of the risk factor.

```
##          bin  no  ng  nb     woe   iv.b   iv.s
## 01 (-Inf,8)  87  78   9  1.3122 0.1068 0.2826
##   02 [8,16) 344 264  80  0.3466 0.0383 0.2826
##  03 [16,36) 399 270 129 -0.1087 0.0048 0.2826
##  04 [36,45) 100  58  42 -0.5245 0.0300 0.2826
## 05 [45,Inf)  70  30  40 -1.1350 0.1027 0.2826
```

Simulation Results cont.

4. For each adjacent pair of bins, compute the sum of their individual IVs (`iv.a`).

```
##          bin  no  ng  nb     woe   iv.b   iv.s   iv.a
##  01 (-Inf,8)  87  78   9  1.3122 0.1068 0.2826     NA
##    02 [8,16) 344 264  80  0.3466 0.0383 0.2826 0.1451
##   03 [16,36) 399 270 129 -0.1087 0.0048 0.2826 0.0431
##   04 [36,45) 100  58  42 -0.5245 0.0300 0.2826 0.0348
##  05 [45,Inf)  70  30  40 -1.1350 0.1027 0.2826 0.1327
```

5. For each adjacent pair of bins, compute the IV assuming the two bins were merged (`iv.m`).

```
##          bin  no  ng  nb     woe   iv.b   iv.s   iv.a   iv.m
##  01 (-Inf,8)  87  78   9  1.3122 0.1068 0.2826     NA     NA
##    02 [8,16) 344 264  80  0.3466 0.0383 0.2826 0.1451 0.0957
##   03 [16,36) 399 270 129 -0.1087 0.0048 0.2826 0.0431 0.0060
##   04 [36,45) 100  58  42 -0.5245 0.0300 0.2826 0.0348 0.0199
##  05 [45,Inf)  70  30  40 -1.1350 0.1027 0.2826 0.1327 0.1147
```

6. The smallest difference between `iv.a` and `iv.m` is observed for bins `03 [16,36)` and `04 [36,45)`.

7. The following table presents the updated summary with the bins merged as identified in the previous step.

```
##                         bin  no  ng  nb     woe   iv.b   iv.s
##             1 [01 (-Inf,8)]  87  78   9  1.3122 0.1068 0.2677
##               2 [02 [8,16)] 344 264  80  0.3466 0.0383 0.2677
##  3 [03 [16,36),04 [36,45)] 499 328 171 -0.1959 0.0199 0.2677
##             4 [05 [45,Inf)]  70  30  40 -1.1350 0.1027 0.2677
```