What is the Purpose of Binning Data?

Binning, also called discretization, is a technique for reducing continuous and discrete data cardinality. Binning groups related values together in bins to reduce the number of distinct values.

Example of Binning

Histograms are an example of data binning used to observe underlying distributions. They typically occur in one-dimensional space and equal intervals for ease of visualization.

Data binning may be used when small instrumental shifts in the spectral dimension from mass spectrometry (MS) or nuclear magnetic resonance (NMR) experiments will be falsely interpreted as representing different components when a collection of data profiles is subjected to pattern recognition analysis. A straightforward way to cope with this problem is by using binning techniques. The spectrum is reduced in resolution to a sufficient degree to ensure that a given peak remains in its bin despite small spectral shifts between analyses.

Binning is also used in machine learning to speed up the decision-tree boosting method for supervised classification and regression in algorithms such as Microsoft's LightGBM and scikit-learn's Histogram-based Gradient Boosting Classification Tree.

There are two methods of dividing data into bins and binning data:

1. Equal Frequency Binning: Bins have an equal frequency.

For example, equal frequency:

Input: [5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215]

Output:

[5, 10, 11, 13]

[15, 35, 50, 55]

[72, 92, 204, 215]

Example 2:

Dataset: 10,15, 18, 20, 31, 34, 41, 46, 51, 53, 54, 60

First, we have to sort the data. If we have unsorted data, we need to convert it into sorted data and then apply the second step. In the second step, we have to find the frequency. To calculate the frequency, we can use the formula a **total number of data points/number of bins**.
In this case, the total number of data points is 12, and the number of bins required is 3. Therefore, the frequency comes out to be 4. Now let's add values into the bins.

BIN 1: 10, 15, 18, 20
BIN 2: 31, 34, 41, 46
BIN 3: 51, 53, 54, 60

2. Equal Width Binning: Bins have equal width with a range of each bin are defined as [min + w], [min + 2w] …. [min + nw] where w = (max - min) / (no of bins).

For example, equal Width:

Input: [5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215]

Output:

[5, 10, 11, 13, 15, 35, 50, 55, 72]

[92]

[204, 215]

Example 2:

Dataset points: 10, 15, 18, 20, 31, 34, 41, 46, 51, 53, 54

Step 1: Sort the dataset in ascending order: 10, 15, 18, 20, 31, 34, 41, 46, 51, 53, 54.

Step 2: Determine the width of each bin using the formula: **w = (max-min) / N**, where w is the width, max is the maximum value, min is the minimum value, and N is the number of bins.

In our example, the minimum value is 10, the maximum value is 54, and we choose four bins. So, the width (w) is calculated as (54 - 10) / 4 = 11. Now we have to add value to each bin which would be as follows:

BIN 1 : [lower bound , upper bound] = [(min) , (min + w -1)] = [10, 20]
BIN 2 : [lower bound , upper bound] = [(min + w) , (min + 2w -1)] = [21, 31]

BIN 3 : [lower bound , upper bound] = [(min + 2w) , (min + 3w -1)] = [32, 42]
BIN 4 : [lower bound , upper bound] = [(min + 3w) , (max)] = [43, 54]

So each bin would contain values in between their lower bound and upper bound. So the final bin for this dataset would have values like:

BIN 1 : [10, 15, 18, 20]
BIN 2 : [31]
BIN 3 : [34, 41]
BIN 4 : [46, 51, 53, 54]

What is the Purpose of Binning Data?

The purpose of binning data is to reduce the complexity of data and make it more manageable and easier to analyze. Binning in data mining can be used for both numerical and categorical data and involves grouping data into smaller, more manageable intervals or categories, or bins.

There are several reasons why binning data can be useful -

Simplification of data - Binning reduces the complexity of data by grouping values into a smaller number of categories or intervals, which makes it easier to understand, summarize and visualize.

Reduction of noise - In some cases, binning can help reduce noise in the data by smoothing out variations in individual data points and highlighting larger trends or patterns.

Facilitation of data analysis - Binning can make it easier to perform statistical analysis and create visualizations, such as histograms, by reducing the number of unique values in the data.

Improvement of model performance - Binning can also be used to create new features or input variables for predictive models. By grouping similar values, binning can strengthen the relationship between attributes and improve the performance of machine learning models.