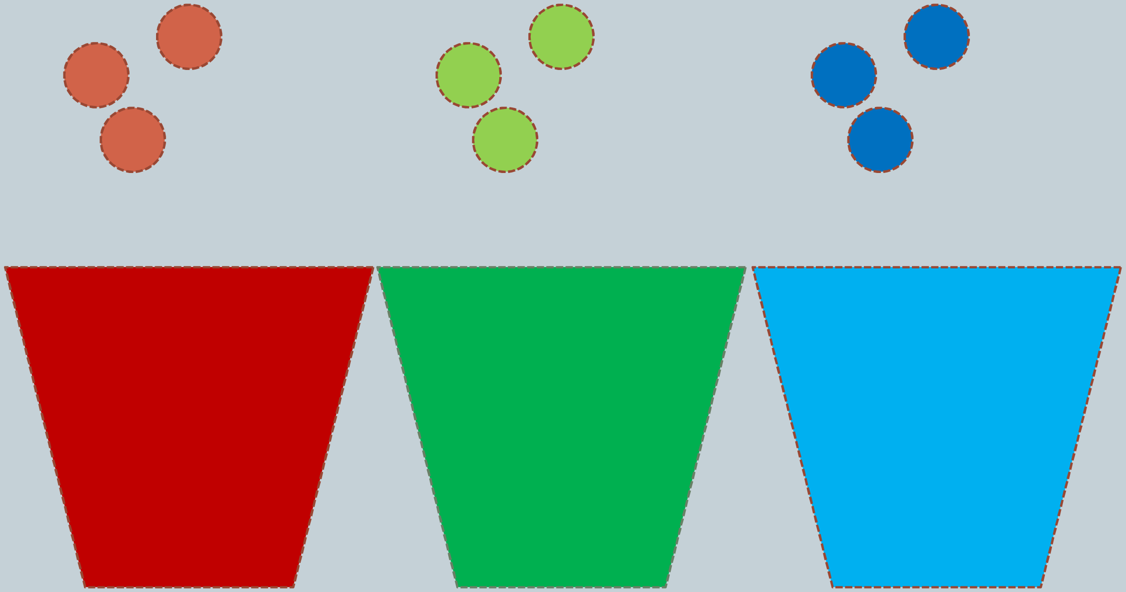
Adaptive Binning



Binning

- Binning is the process of transforming continuous data to discrete or categorical features.
- When we are dealing with continuous data it is helpful when we bin the data for further analysis.
- Binning improves accuracy of the predictive models by reducing the noise or non-linearity in the dataset.
- Binning lets easy identification of outliers, invalid and missing values of numerical variables.

Binning

- We have different names for binning like bucketing, discrete binning, discretization or quantization etc.
- We have two types of binning:
  - Fixed Width Binning
  - Adaptive Binning
- Fixed width binning:
  - In fixed-width binning, we have specific fixed widths for each of the bins which are usually pre-defined by the user analyzing the data. Each bin has a pre-fixed range of values which should be assigned to that bin on the basis of some domain knowledge, rules or constraints.

Binning

- The drawback in using fixed-width binning is that due to us manually deciding the bin ranges, we can end up with irregular bins which are not uniform based on the number of data points or values which fall in each bin.
- Adaptive Binning:
  - Adaptive binning is a binning method where we let data distribution itself to decide our bin ranges.
  - Quantile based binning is a good strategy to use for adaptive binning. Quantiles are specific values or cut-points which help in partitioning the continuous valued distribution of a specific numeric field into discrete contiguous bins or intervals.

Binning with Pandas

- With Pandas library we can do binning using the functions 'cut' and 'qcut'.
- 'cut' can be used to get fixed width bins.
- 'cut' is used to specifically define the bin edges. There is no guarantee about the distribution of items in each bin. In fact, you can define bins in such a way that no items are included in a bin or nearly all items are in a single bin.
- 'qcut' can be used to get adaptive binning as it is a quantile based discretization technique.
- 'qcut' tries to divide up the underlying data into equal sized bins. The function defines the bins using percentiles based on the distribution of the data, not the actual numeric edges of the bins.

Advantages and Disadvantages of Binning

- Advantages:
  - Discretization helps in easy analysis of data.
  - Improves accuracy of the model.
  - It helps in easy identification of outliers, missing value etc of numeric variables.
- Disadvantage:
  - It results in loss of data.
- Advantage of adaptive binning over fixed width binning is that in fixed width binning irregular data will be present in the bins but with adaptive binning we will get equal number of data points in each bin.

Conclusion

- Binning is a discretization method to transform continuous data into discrete data.
- We have seen 2 methods of binning. One is fixed width binning in which we bin the data points according to width that has been decided.
- But here the data points in each bin might not be same. To address this issue we adopt adaptive binning.
- In adaptive binning we bin using quantiles and each bin consist of equal number of data points.
- Lets see the practical application through python codes.