PREDICTIVE MODEL PLAYBOOK

K-CENTROID CLUSTERING

Summary: Cluster analysis identifies cohesive subgroups of observations within a dataset. It allows us to reduce a large number of observations into a smaller number of clusters.

STEP 1: SELECT APPROPRIATE VARIABLES

The first step is to understand the objectives for segmentation. Then, choose the appropriate variables that provide the information needed for clustering. A sophisticated cluster analysis cannot compensate for the poor choice of attributes.

STEP 2: DATA PREPARATION

Numeric data: Cluster analyses requires numeric data. Many non-numeric variables can be converted to numeric ones. Make sure to remove outliers as clustering algorithms are highly sensitive to outliers.

Variable reduction: This step often requires variable reduction techniques to combine variables that revolve around a particular theme. A common method is Principal Component Analysis (PCA), which reduces a set of related variables into few principal components (PCs) that explain most of the variances in the data. Rule of thumb is to use PCs that account for ~80% variance.

Principal Components

Scaling the data: Standardizing each variable using the z-score ensures that the results are not overly sensitive to variables with higher values.

STEP 3: DETERMINE THE NUMBER OF CLUSTERS

Use the AR and CH indices to determine the optimal method and number of clusters. Use a box and whisker plot. The higher the median and smaller the variation the better. Remember, clustering is an iterative process and may require comparing several models to arrive at a good solution.

K-Centroid Cluster Diagnostics

STEP 4: CREATE THE CLUSTERING MODEL

Select the variables, standardization process, clustering method, and number of clusters that gave the best solution. Create the cluster model and append the clusters to the dataset.

K-Centroids Cluster Analysis

STEP 5: VISUALIZE AND VALIDATE RESULTS

Visualization helps us determine the meaning and usefulness of the clustering solution. Use summary statistics to understand difference among clusters.

Validate the results: You can use internal validation and/or external validation. Plot the distribution of the validation variable for each cluster using box and whisker plot to visualize the differences.

Append Clusters