Leveraging Cross-Validation Approach For Predictive Modeling

Brief Introduction: Regression Analysis

The econometric regression model is often used to study the effect of a set of independent variables on a dependent variable, allowing one to make the inference of the observed variable's statistical significance and the overall model's explanatory power. The price-response function studied considers customer response—quantity demanded—as dependent variable, and price as independent variable. Those other potential factors that could influence the demand decision are mostly captured through the customer segmentation. The exact relationship between the units sold and the price charged is therefore contingent on the product category and the customer segment under consideration.

Brief Discussion: Cross Validation

The external validation, which contrasts with the internal validation that almost ensures the overfitted model, partitions the dataset into train set and test set to assess the out-of-sample predictive ability. Three common cross validation approaches are briefly explained as follows:

Leave-one-out cross validation approach

For each iteration, the model is trained on the $N - 1$ data points and tested against the one data point. The exhaustive train process reduces the potential bias, but also comes with the costs: First, the repetition times (N) make the LOOCV approach computationally expensive, especially when N is large. Second, model validated on only one data point could result in wide variation in prediction error, when some outliers are present.

K-fold cross validation approach

Depending on the number of k, the train set is further splitted into k folds. For each iteration, the k-minus-one folds are used to train the model, and the remaining-one fold is then treated as a validation set to test the model.

Repeated K-fold cross validation approach

Taken the same spirit from the k-fold cross validation, the repeated k-fold cross validation simply repeats the process a number of times to avoid reliance on the certain split.

Approach Adopted

Given the discussion of cross validation approaches, the repeated five-fold cross validation and the for-loop programming construct are used to select the n-th degree polynomial model.

Train set

For each product category and within each of the resulting customer clusters, seventy percent of observations will be randomly assigned to the train set, which is then used to train the model on the five-minus-one folds and validate the performance on the remaining-one fold.

Test set

For each product category and within each of the resulting customer clusters, the remaining thirty percent of observations will be therefore assigned to the test set, which is then used to test the model and make the predictions.