Case Study Solution Design

Data Mining Technique

For taking advantage of different classification algorithms and improving performance measures of our classification, we will use multiple classification algorithms including Logistic Regression, K-NN classification and Naïve Bayes Classification.

Variable Selection

For our first part of the analysis, we used data visualization and association rules to understand the characteristics of ‘caravan mobile home insurance’ buyers. The results from these allowed us to state the relationship between existing customers, ‘caravan mobile home insurance’ buyers and some corresponding general characteristics.

Oversampling

For our later part of the analysis, the unbalanced dataset required us to use oversampling techniques to capture the characteristics of the success class (5.9% of the observations). Now we built the above 3 classification techniques on two separate test data frames: the training dataset and the oversampled dataset.

Performance

Since it is critical for our analysis to correctly classify success class observations, the most important performance measures to consider is sensitivity and PPV, we have created different situation based on recommendations associated with different sensitivity and PPV tradeoff values.