Caravan Insurance Purchase Prediction

Information about customers consists of 86 variables and includes product usage data and socio-demographic data derived from zip area codes. The data was supplied by the Dutch data mining company Sentient Machine Research and is based on a real world business problem. The training set contains over 5000 descriptions of customers, including the information of whether or not they have a caravan insurance policy. A test set contains 4000 customers of whom only the organizers know if they have a caravan insurance policy.

Evaluation.

Evaluation will be based on:

- Data Preparation (20%)
- Feature Selection & Engineering (20%)
- Model Comparison (35%)
- Model Selection (10%)
- Presentation (15%)

Data Preparation.

The dataset is highly imbalanced and would require upsampling/downsampling strategies.

Feature Selection.

Select the right features based on importance and significance

Feature Engineering.

Apply feature engineering techniques to see how new features can be created to improve the model. Check for Interaction.

Model Comparison.

Apply multiple algorithms and compare results. To try to Decision Tree, Random Forest & Logistic Regression

Model Selection.

Select the best model. Model selection to be based on Kappa value, Sensitivity & Specificity