Caravan EDA - MA755

Date: 02/27/2018

Author: MIMOZA MARKO (marko_mimo@bentley.edu), SNEHA ARORA (arora_sneh@bentley.edu), TAUSEEF AHMAD (ahmad_taus@bentley.edu)

Objective

The purpose of this notebook is to explain characteristics of customers. The analyses start with exploring the dataset and selecting the 10 most important variables in explaining the response variable CARAVAN. In order to perform the dimensionality reduction we apply the random forest classifier. After acquiring the important variables, we explore the single and multiple variable summaries to find patterns.

1. Dataset Description

The caravan insurance dataset supplied by the Dutch data mining company Sentient Machine Research and is based on a real-world business problem. The purpose of the project is to have a clear insight to why customers have a caravan policy and how these customers are different from the others.

The dataset consists of 87 attributes and 9822 observations. It is further divided into a training set (5822 observations) and a test set (4000 observations). Out of the 86 attributes, 2 are categorical, 83 are numerical and 1 is the response variable (Caravan Insurance Purchased) which indicates whether the customer purchase a caravan insurance policy or not. This is an imbalanced dataset as the target variable Caravan Insurance Purchased has more 0’s i.e. the customers did not purchase the insurance policy as compared to 1’s i.e. the customers did purchase the insurance policy. Futhermore, this dataset is set up as groups and each observation represent a large sample size.