Data-based Customer Segmentation

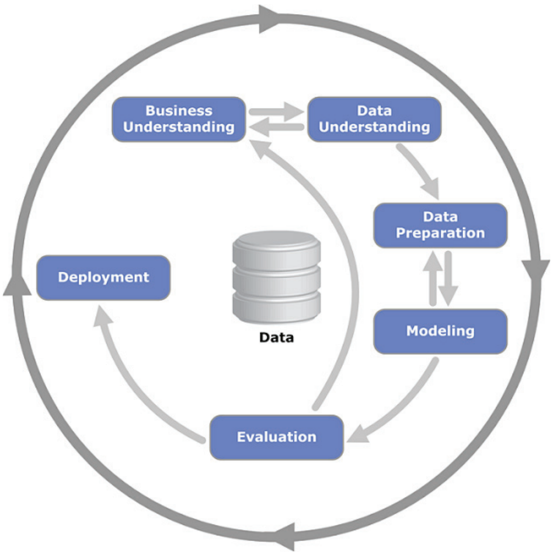
Degree programme: BSc in Industrial Engineering and Management Science | Specialisation: Business Engineering
Thesis advisor: Prof. Dr. Stefan Grösser, Tim Luginbühl
Expert: Prof. Dr. Stefan Grösser

Customer segmentation is one of the core functions of customer relationship management. Customers are divided into several subgroups based on attributes and characteristics. It enables companies to improve their understanding of the customer needs and therefore provides differentiated strategies for each subgroup. The objective of the thesis is to provide a data-based customer segmentation model for a service company using data mining techniques.

Maël Nicolas Droz-dit-Busset
mael.droz@gmail.com

Methodology

This project has been conducted using a cross industry standard process for data mining (CRISP). This process is commonly used for implementing data mining projects. Random forest models are generated from several decision trees. The objective of a decision tree is to create a training model which predicts the class or value of the target variable by learning simple decision rules derived from data. A random forest model consists of many individual decision trees that function as an ensemble. Each decision tree outputs a class prediction, after which the class with the highest accuracy becomes the prediction of the random forest model. In this thesis, random forests were generated to predict the customer's affinity to purchase certain products/services, their affinity to different communication channels and their affnity to up selling and cross selling. Clustering algorithms divide data objects into several groups commonly called clusters. The objective is that data objects that share similarities are grouped together and data objects that diff er from each other are separated. From the predictions of the random forests, a k-means clustering was performed to partition the customers in five distinct segments with the same characteristics.



Capitation: Data mining phases according of CRISP-DM as defined in Chapman et al. (2000). Figure from wikipedia.com

Results

In total, 12 random forests were generated. Three of which are regressions, five are binary classifications and four are non-binary classifications. The regression random forests were used to predict the affinity of customers to several communication channels such as letter, telephone or email. They were evaluated with the R-squared and the Root-Mean-Squared-deviation. The R-squared represents the proportion of variance in the outcome variable which is explained by the predictor variables in the sample. The variation that has been explained by the model is the difference between the total sum of squares and the residual sum of squares, and is called the between groups sum of squares. The R-squared of the different random forests range from 0.914 to 0.924, whereas the Root-Mean-Squared-Error of the different random forests range from 0.087 to 0.101.
The classification random forests predict the affinity of the customers to some products and to up- and cross selling. Their accuracy range from 89.15% to 100%. To ensure that the models do not overfit, a ROC graph was plotted for each random forest. The ROC curve is a probability curve. It reveals how much a model is capable to distinguish between classes and is equal to the accuracy of the model. In this thesis, the ROC graphs proved that the random forests could distinguish the different classes even though some random were imbalanced.
The result of the thesis are five customer segments. For each segment, sociodemographical, product-related and communication attributes were visualized. Furthermore, for each customer segment, a proposal was made about how to contact each customer and what products are the most likely to be sold. Thus, the marketing department can now launch individualized campaigns tailored to customer needs.