Customer Segmentation Project Proposal for Arvato Financial Solutions

Domain Background

In this project, the demographics data for customers of a mail-order sales company in Germany will be analyzed and being compared against demographics information for the general population. The unsupervised learning techniques will be used to perform customer segmentation, identifying the parts of the population that best describe the core customer base of the company. Then supervised learning techniques will be used for targeting a marketing campaign for the company to predict which individuals are most likely to convert into customers for the company.

The dataset used has been provided by our partners at Bertelsmann Arvato Analytics. It includes general population dataset, customer segment data set, dataset of mailout campaign with response and test dataset that needs to make predictions.

The techniques in the population segmentation project such as PCA and KMeans, could be used in the unsupervised learning part. The techniques from the detecting payment card fraud such as linear regression, could be used in the supervised part.

Problem Statement

There are two main parts of this project:

1 Customer Segmentation
This is a cluster problem where the model takes all general population datasets/customer datasets as input and produces different clusters. Everyone in the general population/customer dataset should belong to one of the clusters.

2 Supervised Learning Model
This is a binary classification problem where the model takes customer data as input and produces a result (0 or 1) which indicates whether it is worth it to include the person in the campaign.

Datasets and inputs

These datasets are host in Udacity workspace. There are four data files associated with this project:

- Udacity_AZDIAS_052018.csv: Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (columns). Each row of the demographics files represents a single a single person, but also includes information outside of individuals, including information about their household, building, and neighborhood.
- Udacity_CUSTOMERS_052018.csv: Demographics data for customers of a mail-order

company; 191 652 persons (rows) x 369 features (columns). The "CUSTOMERS" file contains three extra columns ('CUSTOMER_GROUP', 'ONLINE_PURCHASE', and 'PRODUCT_GROUP'), which provide broad information about the customers depicted in the file.

- Udacity_MAILOUT_052018_TRAIN.csv: Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns).
- Udacity_MAILOUT_052018_TEST.csv: Demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns). The original 'MAILOUT' file included one additional column, "RESPONSE", which indicated whether each recipient became a customer of the company or not. For the "TRAIN" subset, this column has been retained, but in the "TEST" subset it has been removed; it is against that withheld column that your final predictions will be assessed in the Kaggle competition.

  Otherwise, all the remaining columns are the same between the three data files. For more information about the columns depicted in the files, you can refer to two Excel spreadsheets provided in the workspace. One of them (/DIAS Information Levels - Attributes 2017.xlsx) is a top-level list of attributes and descriptions, organized by informational category. The other (/DIAS Attributes - Values 2017.xlsx) is a detailed mapping of data values for each feature in alphabetical order.

s

Solution Statement

1 Data preprocess
We need replace all missing values with Nan, the DIAS Attributes - Values 2017.xlsx will be used to find all missing values.
Columns with too high Nan ratio will be dropped.
Columns will be divided by types and the remaining Nans will be imputed differently by group.
More transformations (such as one-hot encoding, scale and so on) will be applied to some special columns based on their type and value.
Feature engineer may be needed for some columns.

2 Customer Segmentation
In this part we need to analyze general population and customer segment datasets and use unsupervised learning techniques to perform customer segmentation, identifying the parts of the population that best describe the core customer base of the company.

The **PCA**(principal component analysis) technique for dimensionality reduction. Then, **elbow curve** will be used to select the best number of clusters for **KMean**s algorithms. Finally, the KMeans will be used to make segmentation of population and customers and determine description of target cluster for the company.

3. Supervised Learning Model
In this part we need to build machine learning model using response of marketing campaign and use model to predict which individuals are most likely to convert into becoming customers for the company.

I will use several machine learning classifiers and choose the best using analysis of learning curve. The machine learning classifiers are: Random Forest tree, Adaboost Classifier and Gradient Boosting Classifier. These classifiers are all ensemble classifiers built on top of decision tree model. Then, I will use Grid Search to parametrize the model and make predictions.

4. Kaggle Competition
The results of this part need to be submitted for Kaggle competition

Metrics

Area under the receiver operating characteristic curve (ROC_AUC) from predicted probabilities will be used to evaluate performances of the models. The ROC curve is created by plotting the true positive rate(TPR) against the false positive rate(FPR) at various threshold settings. AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random target person more highly than a random non-target person. Thus, ROC analysis provides tools to select possibly optimal models for customer prediction task.