Machine Learning Engineer Nanodegree
Capstone Proposal
Customer Segmentation – Arvato Financial Solutions

Introduction

Arvato Financial Solutions provides a professional B2B (Business to Business) financial services. Few services offered by Arvato are diverse segments, payment processing etc.

From the outcome of this capstone project, Arvato is helping Mail-order companies who sell organic products identify potential customers who are likely to subscribe the product and become loyal customers for the company.

Problem Statement

The problem statement for this project is "How to acquire new clients to sell client's products"[7][8].

The project is sub-divided into following tasks:

1. **Unsupervised learning approach** – In this approach, part/region of population are to be selected who would most likely can be approached as new customers using the given demographic data.
2. **Supervised learning approach** – In this approach, from the above selected region of population which customer segments can be targeted as potential customers for advertising campaign,
3. Kaggle competition, predict if customer would likely convert as company customers or not.

Datasets and Inputs

There are 4 datasets available:

- Udacity_AZDIAS_052018.csv – Demographics data for the general population of Germany; 891,211 persons (rows) X 366 features (columns)
- Udacity_CUSTOMERS_052018.csv - Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns)
- 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)

All the files are provided by Arvato in context of Machine Learning Engineer Nanodegree Program for analysis and customer segmentation project.

Solution Statement

1. For the first part of the problem, which is Unsupervised learning approach, I am planning to use K-means Clustering. As it has higher rate of success in such problems in general.
2. For the second part of the problem, which is the Supervised learning approach. Any supervised algorithm can be implemented, e.g. – SVM, k-Nearest Neighbors, Logistic regression, neural networks, or decision trees and Random Forests. Combining the top performing models will result in a more effective final model for submission.

Benchmark Model

To benchmark the model, I will be using Logistic regression as its easy to train and test with less amount of complexity. The performance and accuracy from this model will be considered as the benchmark to compare against the above trained model(s).

Evaluation Metrics –

The evaluation metrics can be divided, one for each algorithm:

1. **Customer Segmentation using unsupervised algorithm** - To determine which population region to be considered as potential high customer base, I can check if the proportion of customers in that particular region is greater than the general population proportion. And these selected population regions will become the data set for the next problem.
2. **Customer Acquisition using supervised algorithm** - After the population region/area to target is separated. In this section, my algorithm should classify if the customer is likely to become a customer or not. The classification include:
    a. Confusion Matrix – F1 score, Recall, Precision
    b. Area Under the Receiver Operating Curve (AUROC)

The final decision on which evaluation metrics to use highly depend on the information obtained through explanatory data analysis.

Project Design

1. Gather and Explore Data –

a. Collect the data,
b. Sample a data set for data exploration
c. Study the attributes and its characteristics
d. Visualize the data

2. Prepare the data –

a. Fill in the missing values (with zeros, mean, or median) or delete the row.
b. Feature selection - Select those attributes which provide useful information for the task
c. Feature Engineering – Discretize continuous features, decompose features and add any promising transformations of features
d. Feature Scaling – Standardize or normalize features.

3. Model Selection –
    a. Train many models from different categories (i.e., linear, Naïve Bayes, SVM, Neural Net, Random Forest Classifier and XG Boost Classifier etc.)
    b. Measure and compare their performance
    c. Analyze the most significant variables for each algorithm
    d. Analyze types of errors models make.
    e. Short-list the top three or five promising models
4. Model Tuning –
    a. Fine-Tune the hyperparameters using cross-validation
    b. Combine best models to get better results than individual model results
5. Test and Prediction – Test the above selected model and measure its performance on the test set to estimate the generalization error.
6. Kaggle submission – The predictions on the test data will be submitted on Kaggle page for other inspired developers to learn and understand the ML project.

Sources:

1. https://en.wikipedia.org/wiki/Arvato
2. https://classroom.udacity.com/nanodegrees/nd009t/parts/2f120d8a-e90a-4bc0-9f4e-43c71c504879/modules/7e69b87a-bf80-428e-89bf-358b2721fc16/lessons/4f0118c0-20fc-482a-81d6-b27507355985/concepts/e9553619-113b-4565-a34e-a9ef450659de?bounced=1613706125599
3. https://towardsdatascience.com/task-cheatsheet-for-almost-every-machine-learning-project-d0946861c6d0
4. Hands-On Machine Learning with Scikit-Learn & Tensorflow
5. https://www.kdnuggets.com/2018/05/general-approaches-machine-learning-process.html
6. https://github.com/pranaymodukuru/Bertelsmann-Arvato-customer-segmentation/blob/master/proposal.pdf
7. Zeying Li, "Research on customer segmentation in retailing based on clustering model," 2011 International Conference on Computer Science and Service System (CSSS), Nanjing, China, 2011, pp. 3437-3440, doi: 10.1109/CSSS.2011.5974496.
8. (TP044571), N. P., & Jayabalan, M. M. (2018). *Predicting bad loans using machine learning & customer segmentation by visualisation* (Master's thesis, A thesis submitted in fulfillment of the requirements for the award of the degree of M.Sc in Data Science and Business Analytics (UCMF1701DSBA), 2018). Kuala Lumpur: Asia Pacific University.