Capstone Project Proposal

Project Overview

It’s crucial for any business to classify the customers as of different segments by specific characteristics. By understanding the behavior and the consumption logics of each segment, a company can target the right consumer groups and make effective advertisement campaign.

In this project, we will use machine learning techniques to analyze demographics data of customers of a mail-order company in Germany. The goal of this project is to create customers segments of population, and to build a model to predict weather a customer will respond to future campaign.

Domain Background

Machine learning methodologies have widely applied in business analysis, such as analyzing customer data and finding insights and patterns. For example, Torizuka [1] used the Random Forest algorithm for segmentation and concluded that the algorithm recognizes data with high accuracy within the presence of noise and outliers. Ezenkwu, Ozuomba and Kalu [2] applied clustering algorithm to discover subtle but tactical patterns for a large unlabeled dataset. In this project, we will also apply artificially intelligent algorithms to analyze the attributes of existing clients and employ machine learning models to help identify new potential clients.

Problem Statement

The main business questions for this project can be divided as follows:

- What are the segments of the general German population and customer population?

We will implement unsupervised machine learning models to perform classification for the customers and the general population.

- How are these segments related to the Arvato customers database?
  By comparing the customer population to the general population, we could identify features that best describe the customers.
- What are the predictions of the targeted customers to the mail advertising campaign?
  A prediction model will then be created to identify individuals who are more likely to respond to the campaign.

Solution Statement

The solutions will be split into the following parts:

1. **Data Preprocessing**
   In this part, we need to perform exploratory data analysis for further analysis. Through EDA, we will be able to
   - Understand features, references, and attributes of data
   - Feature engineering
   - Data Transformation
   - Handling missing values, outliners, and structure conversion
2. **Consumer Segmentation Report**
   We first need to analyze general population (from `’Udacity_AZDIAS_052018.csv’`) and customers of mail-order company (from `’Udacity_CUSTOMERS_052018.csv’`). We will cluster the data using K-mean Algorithm to form segments of the population based on selected features. Since K-mean is exposed to the limitation of high-dimensional data, we will use PCA for dimensional reduction.
3. **Model Training**

We will perform a number of classification algorithms first (e.g. Random Forest, XGBoost, Gradient Boosting, etc.), then compare their performances with predefined metrics (e.g. WSS, ROC, etc.) and select best-performing model to predict potential customers.

Datasets and Inputs

There are four data files associated with this project that are provided by Arvato:

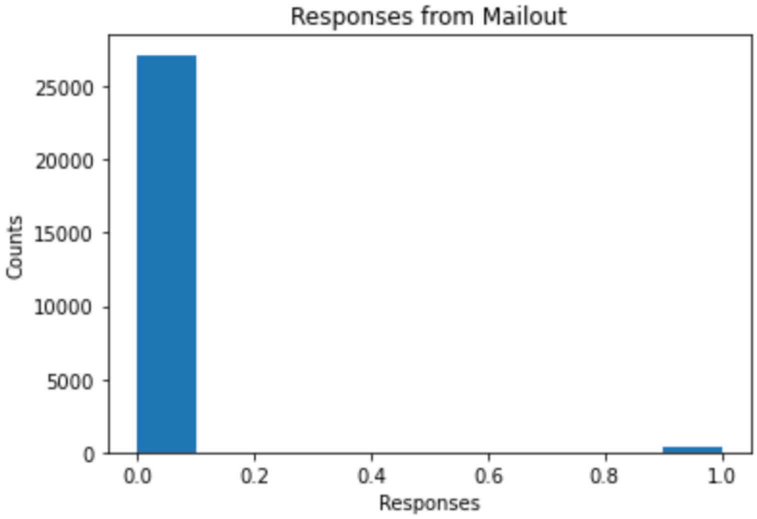
Udacity_AZDIAS_052018.csv: Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (columns).

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).

We noticed a significant imbalance in *Udacity_MAILOUT_052018_TRAIN.csv*. Over 98% of the response was labeled as zero. Under this circumstance, simply using accuracy as a metric might cause over-fitting towards the non-response data.



Evaluation Metrics

One of the most commonly seen metrics for unsupervised learning is Within-cluster sum of square (WSS). WSS calculates sum of squared distance between centroids and data points within each cluster.

Another metric we would like to apply is Area Under Curve (AUC) which specifies how well the models identify each cluster.

Benchmark Model

In Kaggle competition (https://www.kaggle.com/c/udacity-arvato-identify-customers/overview/evaluation), the leaderboard was ranked by AUC for ROC curve. The highest public score was 0.88403.

Project Design

1. Data Preprocessing
    - Understand features, references, and attributes of data.
    - Feature engineering
    - Handling missing values, outliners, and structure conversion.
    - Data Transformation
    - Split the data into training and testing
2. Training, Validating and Testing Data

    The Models will be divided into two parts: customer segmentation and customer prediction. For customer segmentation, we will implement unsupervised algorithm, specifically K-means to group data based on similar attributes. Since there are more than 300 attributes in the dataset, we need to first implement PCA to reduce dimensionality and select the most relevant data for further analysis.

    The other algorithm we will implement for customer prediction is supervised learning. As there are many advanced classifiers we can choose. We will perform each one of them and select the best-performing classifier for this project. Some proposed models are:

    - Gradient Boosting: a generalization of boosting to arbitrary differentiable loss functions. Gradient Boosting can be used for both regression and classification problems.
    - XGBoost: derived from Gradient Boosting but it improved the efficiency of regularization objective function.
    - Random Forest: combines several methods of classification.

Reference

[1] Torizuka, K., Oi, H., Saitoh, H., & Ishizu, S. (2018). Benefit Segmentation of Online Customer Reviews Using Random Forest. 2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 487-491.

[2] Chinedu Pascal Ezenkwu, Simeon Ozuomba and Constance kalu, “Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services” International Journal of Advanced Research in Artificial Intelligence(IJARAI), 4(10), 2015. http://dx.doi.org/10.14569/IJARAI.2015.041007