Domain Background

A mail-order sales company in Germany, Arvato Financial Services, is interested in identifying the data to acquire a new client base for their email marketing. One of the challenges for email marketing is that emails are easily ignored by recipients due to a lack of customer base who is potentially interested in their service (Mari, 2019). In order to create a customer base, this project will analyse demographic data for both the general population at large as well as for prior customers for targets of a mail-out marketing campaign.

Problem statement

“How might a mail-order sales company in Germany be able to acquire new clients in an efficient manner?”

Potential solutions:

- Unsupervised learning approach to segment customer into core customer base
- Supervised learning approach to identify target individuals who are most likely to convert into becoming customers

Datasets and inputs

This project will explore 4 datasets, provided by the Arvato in the context of Udacity Machine Learning Engineer Nanodegree course.

- 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).

Also, following datasets are provided to help better mapping of attributes data.

- DIAS Information Levels - Attributes 2017.xlsx: A top-level list of attributes and descriptions, organised by informational category.
- DIAS Attributes - Values 2017.xlsx: A detailed mapping of data values for each feature in alphabetical order.

Solution Statement

This project will firstly use unsupervised learning approach to segment customer base, and then use supervised learning approach to predict potential targets.
For unsupervised learning, feature scaling, PCA, and KMeans clustering are considered to be used for a solution. For supervised learning, GridSearchCV and DecisionTreeRegressor are options to identify targets. Since this is a proposal, methods for each approach might change as the project progresses.

Benchmark Model

This project will use Gradient boosting classifier as a benchmark model for core customer base.

Evaluation Metrics

This project will use metrics such as accuracy, precision, and recall as evaluation metrics for supervised learning. It is also suggested that ROC-AUC is an option sinde this project is a binary classification problem with class imbalances.

Project Design

1. Data cleaning and exploration: This step will explore the data and decide which features to keep or drop for further analysis. This includes examining missing data and outliers to either be filled or dropped based on the feature.
2. Data visualisation: This step will see if there are any patterns and relationships in the data
3. Supervised learning model: This step will build a prediction model based on the findings from prior steps. Currently, GridSearchCV, XGBoost, and LightGBM models are considered to be tried. This step also includes model tuning involving model parameters adjustment to avoid overfitting and data leakage.
4. Testing and evaluation: This step will test the model in competition through Kaggle and evaluate its performance.

Citation

The Rise of Machine Learning in Marketing: Goal, Process, and Benefit of AI-Driven Marketing (Alex Mark, 2019)