Domain / Description

- Arvato, a services company is looking to help its client, a mail company based in Germany to better understand its customer segments and identify the most probable customers. Dataset is provided by Arvato.
- Company's marketing goal is to target an audience that are most likely to respond to a mail and convert to a customer, by applying customer segmentation techniques such as k-means cluster to identify people that are similar into k number of groups, based on distance between the similarity of the features.

Problem

- Targeting consumers has mainly been driven by business experience and intuition, making it hard to predict or track the success of campaigns. With data, we can make more accurate predictions that lead to a greater outcome (i.e., increase in profitability)
- Classification problem, predicting whether a person will convert to a customer or not

Data

All dataset has been provided under agreement to terms and conditions

- Customer demographic information (191,652 customers, 369 features)
- Population demographic information (891,221 people, 366 features)
- Information describing the demographic metadata and range of values (divided into following categories- person, household, building, Microcell, Grid, Postcode, RR1_ID, PLZ8, Community)
- Training and testing data, including labels for whether a customer converted or not given in training data (training set 42,962 people, 366 feature columns & testing set 42,833 people, 366 feature columns)
- Labels are binary (0 meaning yes, the person converted to a customer or 1 meaning no, the person did not convert)
- Split test 80/20 on training and testing data and use stratifying technique to maintain class balance

Following problems exists within the dataset, which requires additional wrangling:

1. Inconsistency in type and format
2. Unknown values marked as 0's and -1's, need to be converted to nan
3. Sparse dataset- need to drop columns and rows that have majority of nan's and replace the rest with a value
4. Encode categorical variables
5. Too many features- feature selection

Benchmark Model

- Logistic Regression, most foundational and easy to interpret. Logistic regression uses a sigmoid function, which maps the probability between 0 and 1. In this project, I used the

logistic regression to predict whether he/she will convert to a customer (1) or not (0) given the customer's demographics.

Evaluation Metrics

- AUC score

Solution Statement

1. Segment customers based on their attributes using pca analysis and unsupervised machine learning algorithm, such as k means cluster
2. Predict probability of customer making a purchase given mail ad- binary classification problem using models such as logistic regression, which takes in demographic features as inputs and outputs the likelihood that a given person will convert to a purchase (notified as 1) or not convert (notified as 0)
3. Test model on unseen data points (other customers)- predict on fit model. Considering boosting methods to learn from iterations of weaker models
4. Consistent approach to preprocessing data will be taken for all 3 steps and provided as a function. Details to data wrangling is outlined in the Data section.

*Further details and annotations will be available within the jupyter notebook