Arvato Financial Services, Customer Segmentation Use Case

Domain Background: Financial Services, Arvato is an IT company that has different products lines. We will be focusing on a use case that will help their financial group to identify potential customers through their mail order campaign.

Problem Statement: Analyze demographics data for customers of a mail-order sales company (Arvato) in Germany, comparing it against demographics information for the general population. We will use unsupervised learning techniques to perform customer segmentation and identifying parts of the population that best describe the core customer base of the company. Then, we'll apply what we've learned on a dataset, with demographics information for targets of a marketing campaign for the company, and use a model to predict which individuals are most likely to convert into becoming customers for the company.

Key Project Milestones:

- Customer Segmentation using Demographics Data
- Develop a Model to predict potential customer for the company through ad campaign
- Upload the results to Kaggle competition to share and score the performance and analysis

Datasets and Inputs: Arvato is graceful enough to provide us with real demographic and mail order camping dataset to work on the project use case.

There are four data files provided for the use case:

- `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 will use the above datasets to tackle the given problem and arrive at a solution.

Solution Statement: The proposed solution will use the given dataset and potentially identify demographics of the customer base for the company and will use machine learning algorithms and other techniques to identify potential customers for the company, the mail order marketing campaign.

Benchmark Model: The models built for the use case will use the demographics and customer data as a ground truth for evaluation and bench mark of the model and analysis.

Evaluation Metrics: We will use real user data provided by Arvato for the project. We will use Numpy and Pandas libraries to perform data cleanup, EDA and pre-processing activities.

Techniques like PCA will be used to reduce the dimensionality of the data identifying as part of the feature engineering to identify ideal input features. K-means algorithm will be applied on the processed data to identity different population clusters, while fine tuning the algorithm parameters.

The second part of the project involves creating a supervised model to predict potential customers for the company using the given dataset. We will use binary classification and algorithms like Logistic Regression, XGBoost etc to create, train and test at a model and will use metrics like confusion matrix, F2 score, AOC and ROC to validate our model to predict potential customers.

Project Design: We will follow the high-level process flow below for processing data, identifying population cluster and creating a supervised model.

High Level Process Flow:

- Collect Data for the use case
- Clean up and Pre-process the demographics (General & Customer) data
- Feature Engineering
- PCA, reduce input features to identify effective input features
- Split the dataset into test, train and validation
- Generate the shoulder curve to identify the cluster size
- Apply K-means algorithm to the processed dataset to identify clusters
- Identify potential customer base clusters by comparing general vs customer cluster
- Clean up and Pre-process the given train and test customer data
- Identify Input Features
- Create a model
- Train the model
- Test the model with test customer data
- Validate output data and model performance
- Infer insights from the final predicted output