Proposal Document

Arvato Financial Solutions

Machine Learning Engineering Nanodegree

Domain Background:

This project is from a company called Arvato Financial Solutions. There will be a blog that I'll write as a final project submission for my nano degree program with Udacity in Machine Learning Engineering. Since I currently work for fintech, this project will help me build a broader perspective of sales optimization in fintech.

Problem Statement:

The main goal of this project is to understand how to acquire more clients or customers efficiently. This was the job of Subject Matter Experts who used to decide on the basis of the experience and business knowledge only. Objective of this project is to help them in making these decisions on the basis of data.

Dataset and Inputs

We were provided with the following 4 data files for this project.

- **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).

Solution Statements

We have attributes and the demographic details of the existing customers. This helps us to identify new potential customers for the company. It helps in bringing efficiency in the customers acquisition process. Instead of reaching out to everyone in the country and blowing the sales budget.

The solution will be comprised of following 3 deliverables:

- **Customer Segmentation Report:** This will be a Jupyter Notebook with markdown notes and visualizations.This notebook will contain all the codes for pre-processing and feature engineering of the data. This will contain a detailed analysis of general population segments or clusters.
- **Supervised Learning Model:** This will be a supervised model object that uses demographics attributes to segment people into potential customers or non-potential customers.
- **Blog post:** This document is the one!

Baseline Model

We have used a boosting algorithm (xgboost) to build our baseline model. All the default parameters were used without any special tuning of hyperparameters. Since we know there is a huge difference in the volume of positive-negative classes, we will handle this as well.

Evaluation Metric

Since it is a classification problem and data is not really very balanced, accuracy may not be the perfect metric to evaluate the performance of the model. We will use AUC(Area Under Curve) ROC CURVE to evaluate the model. It will be done with cross-validation to make sure that there is no high variance in the models.

Project Design

Following is the flow of project execution

1. Join both the base i.e customer and the general population.
2. Pre-process the data like cleaning erroneous data, imputing nulls, scaling, etc.
3. Remove the extra columns present in the customer dataset.
4. Find the optimum K using the elbow method and silhouette analysis.
5. Apply k-means clustering on the whole combined base
6. Cluster analysis.
7. Build the baseline model
8. Hyper-parameter tuning using Bayesian Optimization techniques.
9. Features importance analysis
10. Create prediction file and submit to Kaggle portal.